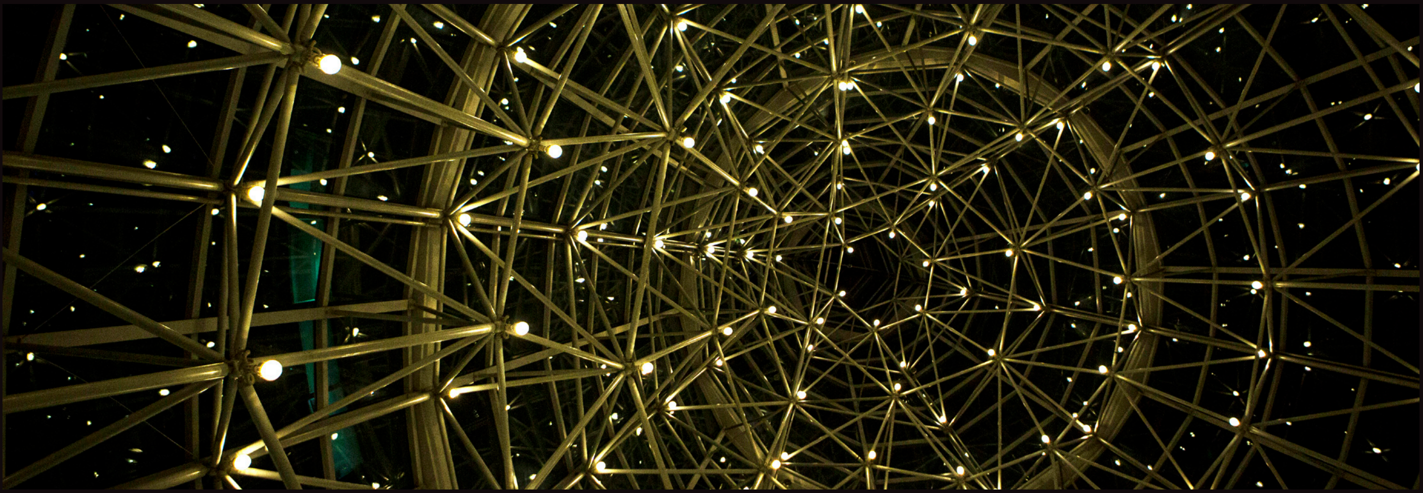
IPR JOURNALS

WINTER 2019 VOLUME 1 NUMBER 1

the journal of financial data science

JFDS.iprjournals.com



the journal of
financial data science
VOLUME 1, NUMBER 1 WINTER 2019

Until recently, the depth and breadth of datasets available to financial researchers was, to put it mildly, extremely shallow. Some exchanges did not record volume information until the early 2000s. The wide adoption of time stamping with millisecond resolution took even longer. Outside exchange trade records and infrequent government statistics, alternative data sources were rare. The implication is that financial researchers conducted the large majority of their analyses on daily price series. This state of data paucity set a hard limit on the sophistication of the techniques that financial researchers could use. In that financial paleo-data age, the linear regression method was a reasonable choice, even though most of us suspected that the linearity assumption may not provide a realistic representation of a system as complex and dynamic as modern financial markets.

Today, we live in a different era, the age of financial Big Data. Researchers have at their disposal datasets that only a few years ago were unimaginable: Satellite images, credit card transactions, sensor data, web scrapes, sentiment from news and tweets, recordings from speeches, geolocation of cargos crossing the oceans, web searches, supply-chain statistics, and the like. The size, quality, and variety of these sources of information, combined with the power of modern computers, allow us to apply more sophisticated mathematical techniques.

However, the adoption of these new techniques is not straightforward. It requires researchers to abandon the comfort of closed-form solutions and embrace the flexibility of numerical and nonparametric methods. The goal of this journal is to facilitate this transition among academics and practitioners. We, the editors, felt that the established journals were not ready to serve this goal for multiple reasons. Our readers will find in this journal high-quality academic articles that are applicable to the practical problems faced by asset managers. These articles present fresh ideas that challenge the traditional way of thinking about finance, the economy, and investing. Through case studies, we offer a front-row view of the cutting-edge of empirical research in financial economics.

In the first article, two of the co-editors, Joseph Simonian and Frank J. Fabozzi, position financial data science within the broader history of econometrics. They explain why its ascendance marks a re-orientation of the field toward a more empirical and pragmatic stance, and that due to the unique nature of financial information, financial data science should be considered a field in its own right and not just an application of data science methods to finance.

Ashby Monk, Marcel Prins, and Dane Rook explain how in finance as alternative data become mainstream, institutional investors may benefit from rethinking how they engage with alternative datasets. By rethinking their approaches to alternative data as the authors suggest, institutional investors can select alternative datasets that better align with their organizational resources and contexts.

As a remedy to the shortcomings of traditional factor models, Joseph Simonian, Chenwei Wu, Daniel Itano, and Vyshaal Narayanam describe a machine learning approach to factor modeling based on the random forests algorithm. As a case study, the authors apply random forests to the well-known Fama-French-Carhart factors and analyze the major equity sectors, showing that compared to a traditional regression-based factor analysis, the random forests algorithm provides significantly higher explanatory power, as well as the ability to account for factors' nonlinear behavior and interaction effects. In addition to providing evidence that the random forests framework can enhance ex post risk analysis, the authors also demonstrate that combining the random forest algorithm with another machine learning framework, association rule learning, can also help produce useful ex ante trading signals.

It is well-known that the classic mean-variance portfolio framework generates weights for the optimized portfolios that are directly proportional to the inverse of the asset correlation matrix. However, most of contemporary portfolio optimization research focuses on optimizing the correlation matrix itself, and not its inverse. Irene Aldridge demonstrates that this is a mistake, specifically from a Big Data perspective. She demonstrates that the inverse of the correlation matrix is much more unstable and sensitive to random perturbations than the correlation matrix itself. The results she reports are novel in the Data Science space, extending far beyond financial data, and are applicable to any data correlation matrices and their inverses.

Although machine learning offers a set of powerful tools for asset managers, one crucial limitation involves data availability. Because machine learning applications typically require far more data than are available, especially for longer-horizon investing, it is important for asset managers to select the right application before applying the tools. Rob Arnott, Campbell Harvey, and Harry Markowitz provide a research checklist that can be used by asset managers and quantitative analysts to select the appropriate machine learning applications as well as, more generally, providing a framework for best practices in quantitative investment research.

Applying a machine learning technique that is new to finance called independent Bayesian classifier combination, David Bew, Campbell Harvey, Anthony Ledford, Sam Radnor, and Andrew Sinclair test whether valuable information can be extracted from analysts' recommendations of stock performance. The technique provides a way to weight analysts forecasts based on their performance in rating a particular stock as well as their performance rating other stocks. Their results show that a combination of their machine learning recommendations along with the analysts' ratings leads to excess returns in their sample suggesting this new technique could be useful for active investors.

Thousands of journal articles have claimed to have discovered a wide range of risk premia. Most of these discoveries are false, as a result of selection bias under multiple testing. Using a combination of extreme value theory and unsupervised learning, Marcos López de Prado proposes a practical method to discount the inflationary effect that selection bias has on a particular discovery.

Ananth Madhavan and Aleksander Sobczyk employ data science to create an investible, dynamic portfolio to mimic the factor characteristics of private equity. Using textual analysis, they first identify firms taken private and then use a multifactor model to measure the cross-sectional factor exposures of firms immediately prior to the announcement that they were being acquired by a private equity firm. Then the authors use holdings-based optimization to build a liquid, investible, long-only portfolio that dynamically mimics the factor characteristics of the portfolio of stocks that were taken private.

Julia Klevak, Joshua Livnat, and Kate Suslava illustrate how the utilization of text mining and scoring of an unstructured data can add information to investors beyond structured data. They demonstrate how the application to the analysis of earnings conference call transcripts produces a signal that is incrementally additive to earnings surprises and the short-term returns around the earnings announcement.

In their article, Sidney C. Porter and Sheridan Porter contribute two new fundamental properties of indexes—similarity and stability—to indexing theory, made practical by advances in data science technology. In the application of the theory, they introduce a framework for a repeatable decomposition of private equity returns that disambiguates the quantification of manager skill.

A graph-theoretic framework for monitoring system-wide risk by extending methods widely deployed in social networks is provided by Sanjiv R. Das, Seoyoung Kim, and Daniel N. Ostrov. They introduce desired properties for any systemic risk measure and provide a novel extension of the well-known Merton credit risk model to a generalized stochastic network-based framework across large financial institutions.

The problem of optimally hedging an options book in a practical setting, where trading decisions are discrete and trading costs can be nonlinear and difficult to model. Using reinforcement learning, a well-established machine learning technique, Petter Kolm and Gordon Ritter propose a flexible, accurate and very promising model for solving this problem.

Frank J. Fabozzi,
Marcos López de Prado,
Joseph Simonian
Editors

the journal of financial data science

The Journal of Financial Data Science®, (ISSN# 2640-3943) is published quarterly by Pageant Media, Ltd., 41 Madison Avenue, 20th Floor, New York, NY 10010. Phone (646) 891-2110. Periodical postage paid at New York, NY and at additional mailing offices. Copyright © 2019 Pageant Media, Ltd. POSTMASTER: Send address changes for *The Journal of Financial Data Science* to Pageant Media, Ltd., 41 Madison Avenue, 20th Floor, New York, NY 10010 USA, Att: Journals Department. Reproduction in whole or in part without written permission is prohibited.

Manuscript: *The Journal of Financial Data Science* considers but assumes no responsibility for unsolicited manuscripts. Unsolicited manuscripts will not be returned. Instructions for submission of manuscripts are printed at the end of the journal.

Subscriptions: For subscription enquiries please contact a member of the team at +1 646 931 9045/+44 20 7139 1600 or at iprjournal@pageantmedia.com. Printed by The Manson Group, St Albans, UK.

The opinions expressed are those of the individual authors. IPR Journals is not responsible for the accuracy, completeness, or timeliness of the information contained in the articles herein. If you require expert advice, you should seek the services of a competent professional. No statement in this journal is to be construed as a recommendation to buy or sell securities. Product names mentioned in the articles in this journal may be trademarks or service marks of their respective owners.

the journal of financial data science

The authors situate financial data science within the broader history of econometrics and argue that its ascendance marks a reorientation of the field toward a more empirical and pragmatic stance. They also argue that owing to the unique nature of financial information, financial data science should be considered a field in its own right and not just an application of data science methods to finance.

As alternative data steadily become mainstream in finance, institutional investors may benefit from rethinking how they engage with alternative datasets. Specifically, they could gain from rethinking (1) alternative data's value proposition, (2) how they characterize alternative data, and (3) how they access alternative data. Rethinking their approaches to alternative data in these ways can help investors select alternative datasets that better align with their organizational resources and contexts. Such rethinking offers the greatest advantages when it focuses on building *defensive* and *defensible* strategies around alternative data, rather than prioritizing quicker exploitation of short-lived opportunities. Rethinking alternative data will require institutional investors to investigate new partnering possibilities, which should help them weather (and even thrive during) the escalating arms race among financial-market participants for alternative data. Building capacity for alternative data in these ways could also help investors accelerate innovation.

Factor models are by now ubiquitous in finance and form an integral part of investment practice. The most common models in the investment industry are linear, a development that is no doubt the result of their familiarity and relative simplicity. Linear models, however, often fail to capture important information regarding asset behavior. To address the latter shortcoming, the authors show how to use random forests, a machine learning algorithm, to produce factor frameworks that improve upon more traditional models in terms of their ability to account for nonlinearities and interaction effects among variables, as well as their higher explanatory power. The authors also demonstrate, by means of a simple example, how combining the random forest algorithm with another machine learning framework known as association rule learning can produce viable trading strategies. Machine learning methods thus show themselves to be effective tools for both ex post risk decomposition and ex ante investment decision-making.

In the classic mean–variance portfolio theory as proposed by Harry Markowitz, the weights of the optimized portfolios are directly proportional to the inverse of the asset correlation matrix. However, most contemporary portfolio optimization research focuses on optimizing the correlation matrix itself, and not its inverse. In this article, the author demonstrates that this is a mistake. Specifically, from the Big Data perspective, she proves that the inverse of the correlation matrix is much more unstable and sensitive to random perturbations than is the correlation matrix itself. As such, optimization of the inverse of the correlation matrix adds more value to optimal portfolio selection than does optimization of the correlation matrix. The author further shows the empirical results of portfolio reallocation under different common portfolio composition scenarios. The technique outperforms traditional portfolio allocation techniques out of sample, delivering nearly 400% improvement over the equally weighted allocation over a 20-year investment period on the S&P 500 portfolio with monthly reallocation. In general, the author demonstrates that the correlation inverse optimization proposed in this article significantly outperforms the other core portfolio allocation strategies, such as equally weighted portfolios, vanilla mean–variance optimization, and techniques based on the spectral decomposition of the correlation matrix. The results presented in this article are novel in the data science space, extend far beyond financial data, and are applicable to any data correlation matrixes and their inverses, whether in advertising, healthcare, or genomics.

A Backtesting Protocol in the Era of Machine Learning 64

Rob Arnott, Campbell R. Harvey, and Harry Markowitz

Machine learning offers a set of powerful tools that holds considerable promise for investment management. As with most quantitative applications in finance, the danger of misapplying these techniques can lead to disappointment. One crucial limitation involves data availability. Many of machine learning's early successes originated in the physical and biological sciences, in which truly vast amounts of data are available. Machine learning applications often require far more data than are available in finance, which is of particular concern in longer-horizon investing. Hence, choosing the right applications before applying the tools is important. In addition, capital markets reflect the actions of people, who may be influenced by the actions of others and by the findings of past research. In many ways, the challenges that affect machine learning are merely a continuation of the long-standing issues researchers have always faced in quantitative finance. Although investors need to be cautious—indeed, more cautious than in past applications of quantitative methods—these new tools offer many potential applications in finance. In this article, the authors develop a research protocol that pertains both to the application of machine learning techniques and to quantitative finance in general.

Modeling Analysts' Recommendations via Bayesian Machine Learning 75

David Bew, Campbell R. Harvey, Anthony Ledford, Sam Radnor, and Andrew Sinclair

Individual analysts typically publish recommendations several times per year on the handful of stocks they follow within their specialized fields. How should investors interpret this information? How can they factor in the past performance of individual analysts when assessing whether to invest long or short in a stock? This is a complicated problem to model quantitatively: There are thousands of individual analysts, each of whom follows only a small subset of the thousands of stocks available for investment. Overcoming this inherent sparsity naturally raises the question of how to learn an analyst's forecasting ability by integrating track-record information from all the stocks the analyst follows; in other words, inferring an analyst's ability on Stock X from track records on both Stock X and stocks other than X. The authors address this topic using a state-of-the-art computationally rapid Bayesian machine learning technique called independent Bayesian classifier combination (IBCC), which has been deployed in the physical and biological sciences. The authors argue that there are many similarities between the analyst forecasting problem and a very successful application of IBCC in astronomy, a study in which it dominates heuristic alternatives including simple or weighted averages and majority voting. The IBCC technique is ideally suited to this particularly sparse problem, enabling computationally efficient inference, dynamic tracking of analyst performance through time, and real-time online forecasting. The results suggest the IBCC

technique holds promise in extracting information that can be deployed in active discretionary and quantitative investment management.

Most discoveries in empirical finance are false, as a consequence of selection bias under multiple testing. Although many researchers are aware of this problem, the solutions proposed in the literature tend to be complex and hard to implement. In this article, the author reduces the problem of selection bias in the context of investment strategy development to two sub-problems: determining the number of essentially independent trials and determining the variance across those trials. The author explains what data researchers need to report to allow others to evaluate the effect that multiple testing has had on reported performance. He applies his method to a real case of strategy development and estimates the probability that a discovered strategy is false.

In this article, the authors use textual analysis to create an investable, dynamic portfolio to mimic the factor characteristics of private equity. First, using textual analysis, they identify firms taken private by those firms in the 10-year period ending June 2018. Second, they use a multifactor model to measure the cross-sectional factor exposures of firms immediately prior to the announcement that they were being acquired by a private equity firm. Finally, they use holdings-based optimization to build a liquid, investible, long-only portfolio that dynamically mimics the factor characteristics of the portfolio of stocks that were taken private. Practitioner applications include interim beta solutions for investors (including venture capital and private equity firms) seeking to deploy excess cash, mitigate underfunding risk, and manage capital calls.

The purpose of the study is to illustrate one application of unstructured data analysis in finance: the scoring of a text document based on its tone (sentiment) and specific events that are important for the end user. The methodology begins with the well-known practice of counting positive and negative words and progresses to illustrate the construction of relevant events. The authors show how the application of this methodology to the analysis of earnings conference call transcripts produces a signal that is incrementally additive to earnings surprises and the short-term returns around the earnings announcement. An interesting feature of the tone change extracted from the conference calls is that it has a relatively low correlation with both earnings surprises and the short-term return around the earnings announcement. This indicates how use of text mining and scoring of unstructured data can add information to investors beyond structured data.

Private-equity asset owners seeking to reduce downside risk and increase upside probability would logically benefit from indexing prospective asset managers by their skill. However, theoretical deficiencies and a lack of rigorous market calibration prevent the metrics and techniques commonly used in private equity from isolating manager skill. In this article, the authors introduce a new conceptual framework for a repeatable decomposition of private equity returns that disambiguates the quantification of manager skill. Modern proxy benchmarks are a key component of the framework for their definition of systemic returns specific to the target asset. They satisfy the fundamental properties of an index (systematic, transparent, and investable) suggested by Andrew Lo and the CFA Institute's SAMURAI criteria for a valid benchmark. However, the authors propose that the integrity of the decomposition requires that the benchmark's similarity (to target) and its stability be systematically derived, measured quantities. The authors discuss these two new properties in conjunction with the technology that enables the construction of modern proxy benchmarks and their active management over time. With systemic returns thus defined, excess returns against the modern proxy benchmark are attributed to dynamic elements under the control of the manager, which the authors define as manager alpha. Systemic returns in excess of a broad/policy benchmark are deemed static elements. Static elements measure the portion of returns attributable to size and sector selection, in which a manager tends to specialize and which are known to the limited partner investor prior to investment. Although both static and dynamic elements contribute active returns to the investment, it is the dynamic elements—alpha—that should merit attention (and high fees) from limited partners.

In this article, the authors propose a theory-driven framework for monitoring system-wide risk by extending data science methods widely deployed in social networks. Their approach extends the one-firm Merton credit risk model to a generalized stochastic network-based framework across all financial institutions, comprising a novel approach to measuring systemic risk over time. The authors identify four desired properties for any systemic risk measure. They also develop measures for the risks created by each individual institution and a measure for risk created by each pairwise connection between institutions. Four specific implementation models are then explored, and brief empirical examples illustrate the ease of implementation of these four models and show general consistency among their results.

The authors of this article address the problem of how to optimally hedge an options book in a practical setting, where trading decisions are discrete and trading costs can be nonlinear and difficult to model. Based on reinforcement learning (RL), a well-established machine learning technique, the authors propose a model that is flexible, accurate and very promising for real-world applications. A key strength of the RL approach is that it does not make any assumptions about the form of trading cost. RL learns the minimum variance hedge subject to whatever transaction cost function one provides. All that it needs is a good simulator, in which transaction costs and options prices are simulated accurately.

Triumph of the Empiricists: *The Birth of Financial Data Science*

Joseph Simonian and Frank J. Fabozzi

Joseph Simonian
is the director of quantitative research at Natixis Investment Managers in Boston, MA.
joseph.simonian@natixis.com

Frank J. Fabozzi
is a professor of finance at EDHEC Business School in Nice, France, and the editor of *The Journal of Portfolio Management.*
frank.fabozzi@edhec.edu

The methodological foundations of contemporary econometrics were laid in the aftermath of a debate that was epitomized by Tjalling Koopmans' (1947) critical review of Arthur Burns and Wesley Mitchell's (1946) *Measuring Business Cycles*. In the review, "Measurement Without Theory," Koopmans, who was a member of the theory-focused Cowles Commission, argued that economic data cannot be properly interpreted without the benefit of well-hewn economic assumptions. The target of the review was not only Burns and Mitchell's book, but also the empiricist econometric methodology employed by the National Bureau of Economic Research, which Koopmans felt was overly preoccupied with devising techniques for measuring economic data at the expense of the development of the theory necessary to draw robust economic conclusions. In the review of Burns and Mitchell's book, Koopmans defines empiricism as a scientific methodology in which decisions about "what economic phenomena to observe, and what measures to define and compute, are made with a minimum of assistance from theoretical conceptions or hypotheses regarding the nature of the economic processes…" The motivating belief that drives Koopmans' argument is a committed philosophical realism regarding economic phenomena. Just as natural science assumes that physical and biological phenomena are regulated by natural laws, Koopmans assumes that economic phenomena are governed by their own set of immutable laws. If this is the case, then the job of the economist is to discover truths about economic reality in the same way that a physicist discovers (or is often assumed to discover) truths about physical reality.[1]

In the years following the theory versus measurement debate, as economics' theoretical footing was being reified, the field was also being increasingly formalized—to the extent that, by the early 1980s, the philosopher of science Alexander Rosenberg could confidently state that economic theory is most appropriately viewed not as a science, but as a branch of mathematics (Rosenberg 1983). In Rosenberg's characterization, economics abstracts away from actual human interaction and posits a set of basic assumptions from which it derives a formally impressive yet empirically empty set of conclusions. He ultimately argued that economics should be treated as "somewhere on the intersection between pure and applied axiomatic systems," whose findings may not correspond to any facts in the world but that are

[1] Koopmans forcefully argued his case and gave the impression that it is impossible to justify an empirically robust theory at all, given that you seemingly need to have a well-grounded theory to comprehend empirical evidence in the first place.

nevertheless interesting from an intellectual standpoint. Although Rosenberg's account of economics may be viewed as somewhat extreme and not reflective of how most economists view themselves and their profession, it nevertheless brings to the fore the extent to which economics is viewed by many as a largely theoretical endeavor. That being said, we do not need to subscribe in whole to Rosenberg's argument to recognize that a broadly theoretical approach has become dominant in economics—econometrics in particular—so much so that today we may succinctly summarize the primary beliefs that drive contemporary econometric practice as follows:

1. The goal of econometrics is to discover well-defined economic processes, mechanisms, and structures.[2]
2. Modern probability theory and statistical inference are indispensable tools in the definition and discovery of economic phenomena.[3]
3. An econometric methodology founded on points 1 and 2 can produce reliable economic forecasts, which can fruitfully be applied in business and policymaking.

Although econometrics is anchored toward the ideology of philosophical realism and strict adherence to the tenets of probability theory, as the quotation from Koopmans indicates, at any given time, the degree to which scientific methodologies are theory-laden may vary. Moreover, scientific frameworks in practice generally differ not by their choice of *either* a purely empiricist or realist methodology, but by the degree to which a given methodological program is guided by empirical considerations. Where economics has erred, in our opinion, is in allowing the pendulum to swing too far in favor of theory. In the physical sciences, although the tension between more theoretically and more empirically inclined methodologies exists, experiments are nevertheless considered indispensable tools for validating or invalidating theories. Experimental tools in the physical sciences are of course better developed than in economics, for a variety of reasons.[4] With the advent of data science, however, we believe that economics now possesses a tool with which economic theories can be tested in a more robust manner, using new and richer datasets. Accordingly, financial data science is well positioned to reorient financial econometrics toward a more empirical stance, a methodological position that was in fact advocated in an argument almost as old as Koopmans'.

THEORY, SHMEORY: AN INSTRUMENTALIST VIEW OF ECONOMICS

At around the same time that Koopmans was arguing in defense of economic realism and the importance of theory, another well-known economist, Milton Friedman (1953), presented an argument in favor of an empirical approach to economics. The strain of empiricism Friedman defended is usually labeled *instrumentalism* (although Friedman never mentioned the term) and emphasizes the predictive role of science, downplaying science's role as an unassailable arbiter of "reality."[5]

In the 20th century, different forms of instrumentalism were championed by a wide variety of thinkers, from Pierre Duhem (1914) to John Dewey (1916, 1938). Today, instrumentalism is an influential methodology in the physical sciences (Torretti 1999). In contrast to Koopmans, Friedman viewed assumptions as tools to be employed in the production of reliable forecasts. As such, in Friedman's instrumentalism, a theory need only be sufficiently coherent if it leads to successful predictions. This view of theory thus eschews, or at least radically downplays, its explanatory role and instead relegates it to a device to frame and guide the process of prediction, or as Friedman put it, "to serve as a filing system for organizing empirical material." Indeed, under an instrumentalist view of economics, the truth or falsity of the axioms and postulates of an economic theory is less relevant than the degree to which a theory facilitates successful prediction.

[2] This belief is perhaps best exemplified by econometrics' preoccupation with causality, a highly complex, not to mention metaphysical, concept that has been a major focus of philosophical analysis for centuries. For a sample of some of the extensive literature on causality in economics, see Haavelmo (1943), Simon (1953), Granger (1969), Hicks (1979), and Hoover (2001).

[3] For a classic statement and argument of this view, see Haavelmo (1944).

[4] To name just two, the ability to conduct closed experiments and to study subject matter that behaves more or less mechanistically gives the physical sciences the ability to confidently draw conclusions from experiments in a way that has hitherto been impossible in economics.

[5] For a review of Friedman's instrumentalism and its critics, see Boland (2016).

Contemporary econometrics' emphasis on theory versus prediction has been detrimental to the ability of the field to produce models with reliable forecasting ability, an outcome that explains its relative lack of influence on economics as a whole versus more "pragmatic" empirical work that generally proceeds with a broad theoretical stance but without a theoretical "straightjacket" (Summers 1991). This paucity of influence on the field as a whole is true even for some of the most popular econometric models, such as the Dynamic Stochastic General Equilibrium (DSGE) class of models, which have shown themselves to be unexceptional forecasting tools both in absolute terms and in relation to much simpler frameworks (Edge and Gurkaynak 2010; Edge, Kiley, and Laforte 2010). Why is this the case? It is surely not due to lack of sophistication on the part of the builders of these elaborate models. To the contrary, it may be due to their "square-in-the-circle" attempts to build predictive models within the strict confines of often complex econometric theories, rather than conforming theories to empirical findings. This approach to model building is, in addition to being less useful from a practical standpoint, also the antithesis of scientific practice; natural scientists, in general, evaluate and refine theories through empirical observation, not the other way around.[6]

FINANCIAL ECONOMETRICS FOR THE 21ST CENTURY

We believe that financial data science represents an advancement over the traditional econometrics toolbox. As a scientific endeavor, data science combines statistics and computing in an effort to uncover patterns in information that can then be used to assist decision-making. Although data science employs statistical concepts, its methodological approach is decidedly instrumentalist and is open to using any type of quantitative method, heuristic, or technique in so far as it is useful in producing accurate predictions and informed decisions, regardless of strict adherence to the tenets of any theory. The instrumentalist orientation of data science is precisely what makes it so useful for applications to investment research, a pursuit that is valuable only if it leads to practical results, namely the improvement of individuals' and institutions' financial well-being.

That said, we believe that financial data science is a discipline in its own right, and not merely the application of data science methods to finance. We hold this view for at least three reasons. First, finance brings with it a unique set of problems and puzzles that distinguish it from standard applications of data science, especially those in the natural sciences. The challenges that practitioners face in devising trading strategies, asset allocation, and financial risk management, for example, all require specific solutions. Second, financial time series possess unique characteristics that reflect their origins in human action and intentionality. The defining properties of financial time series such as volatility clustering, momentum, and mean reversion are prime examples of this. Third, modeling agents, especially the collective agents that constitute "the market," is an extremely challenging problem that demands specialized techniques. For these three reasons, we believe it would be a mistake to think that financial data science is merely one area of applied data science.[7]

Just as we believe it is a mistake to consider financial data science as simply a subset of data science, we likewise believe that it is a mistake to consider financial data science as a branch of financial econometrics. Rather, it would be more accurate to describe financial data science as encompassing traditional financial econometrics and expanding it with new techniques and a new orientation. Although financial data science brings its own set of formal tools to the analysis of time-series, cross-sectional, and panel data, it also brings with it a mathematical arsenal capable of dealing with disparate types of data—both structured data, which is the terrain of traditional econometrics, and unstructured data, such as textual and visual information. Moreover, financial data science has a distinctly applied and hence empirical orientation, dispensing with unnecessary theoretical machinery and abstraction in favor of methods designed to adequately frame and solve real-life problems. Its methodological orientation thus places it, and by extension finance as a whole, closer to engineering than to pure science.

From a historical standpoint, the emergence of financial data science represents both a resurgence of

[6] A classic example of this process is given by the set of experiments designed to verify the theory of special relativity (see, e.g., Robertson 1949).

[7] For a similar argument, see López de Prado and Israel (forthcoming).

instrumentalism as a scientific methodology in financial econometrics and, because of the introduction of a multitude of new analytical techniques, an enhancement of the pragmatic empiricism mentioned earlier. By prioritizing successful prediction and usable results, financial data science promises to bring financial econometrics more in line with mainstream scientific practice and, in doing so, takes up the mantle in defending it and economics as a whole against critics who charge that economics is not a "real science." That financial data science is being increasingly recognized as an indispensable part of investment research is a testament to its practical value and a triumph of the empiricism on which it is founded.

ACKNOWLEDGMENTS

The authors would like to thank Marcos López de Prado for his valuable comments on an earlier draft of this article.

REFERENCES

Boland, L. A. "Reading and Misreading Friedman's 1953 Methodology Essay." In *Milton Friedman: Contributions to Economics and Public Policy*, edited by R. Cord and D. Hammond, pp. 541–560. Oxford: Oxford University Press, 2016.

Burns, A., and W. Mitchell. *Measuring Business Cycles*. New York: National Bureau of Economic Research, 1946.

Dewey, J. *Essays in Experimental Logic*. Chicago: The University of Chicago Press, 1916.

——. *Logic: The Theory of Inquiry*. Oxford: Holt, 1938.

Duhem, P. *The Aim and Structure of Physical Theory*. Translated from the second edition by P. P. Wiener. Princeton: Princeton University Press, 1914 (1954).

Edge, R. M., and R. S. Gurkaynak. 2010. "How Useful Are Estimated DSGE Model Forecasts for Central Bankers?" *Brookings Papers on Economic Activity* 41 (2): 209–259.

Edge, R. M., M. T. Kiley, and J. Laforte. 2010. "A Comparison of Forecast Performance Between Federal Reserve Staff Forecasts, Simple Reduced Form Models, and a DSGE Model." *Journal of Applied Econometrics* 25 (4): 720–754.

Friedman, M. "The Methodology of Positive Economics." In *Essays in Positive Economics*, pp. 3–43. Chicago: University of Chicago Press, 1953.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37 (3): 424–438.

Haavelmo, T. 1943. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica* 11 (1): 1–12.

——. 1944. "The Probability Approach in Econometrics." *Econometrica* 12 (supplement).

Hicks, J. *Causality in Economics*. Oxford: Basil Blackwell, 1979.

Hoover, K. D. *Causality in Macroeconomics*. Cambridge: Cambridge University Press, 2001.

Koopmans, T. C. 1947. "Measurement Without Theory." *The Review of Economics and Statistics* 29 (3): 161–172.

López de Prado, M., and R. Israel. 2019. "Beyond Econometrics: A Roadmap Towards Financial Machine Learning." *The Journal of Financial Data Science* 1 (1): 99–110.

Robertson, H. P. 1949. "Postulate versus Observation in the Special Theory of Relativity." *Reviews of Modern Physics* 21 (3): 378–382.

Rosenberg, A. 1983. "If Economics Isn't Science, What Is It?" *Philosophical Forum* 14 (3/4): 296–314.

Simon, H. "Causal Ordering and Identifiability." In *Studies in Econometric Method*, edited by W. C. Hood and T. C. Koopmans. New Haven, CT: Yale University Press, 1953.

Summers, L. H. 1991. "The Scientific Illusion in Empirical Macroeconomics." *The Scandinavian Journal of Economics* 93 (2): 129–148.

Torretti, R. *The Philosophy of Physics*. Cambridge: Cambridge University Press, 1999.

Disclaimer

This editorial reflects the current opinions of the authors, which are not those of Natixis Investment Managers.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

Rethinking Alternative Data in Institutional Investment

Ashby Monk, Marcel Prins, and Dane Rook

Ashby Monk
is executive director of the Global Projects Center at Stanford University in Palo Alto, CA.
amonk@stanford.edu

Marcel Prins
is chief operating officer at APG Asset Management in Amsterdam, The Netherlands.
marcel.prins@apg-am.nl

Dane Rook
is a postdoctoral researcher in the Global Projects Center at Stanford University in Palo Alto, CA.
rook@stanford.edu

Alternative datasets (alt-datasets) appear to be entering the financial mainstream. Alternative data (alt-data) have always occupied a crucial role in financial markets, but, until recently, cultivation and use of alt-data were largely seen as niche activities for specialist players (e.g., hedge funds with esoteric strategies). Yet, the number and diversity of readily accessible alt-datasets has ballooned in the past decade. This proliferation now confronts institutional investors (Investors)—such as public pension funds, endowments, and sovereign wealth funds—with a dilemma: How can they responsibly choose which alt-datasets are most likely to be sources of significant value for their investment objectives? This article's main goal is to help Investors properly address that question.

Within the financial community, alt-data are widely understood to be datasets that are not conventionally used in investment decision making.[1] A few archetypal (and well-hyped) examples of alt-data have emerged in recent years. These include

- satellite imagery of commercial or economic activity (e.g., the number of cars in parking lots of major retailers, ships passing through ports, and agricultural or mining operations);
- social-media streams, from which consumer, political, or other sentiment may be gauged;
- microdata about consumers' shopping activities (e.g., credit card transactions or in-app purchases on smartphones);
- data scraped from the internet (e.g., job postings to track corporate hiring patterns); and
- data exhaust—the assortment of log files, cookies, and other digital footprints created by people's online browsing (including geolocation data from searches on mobile devices).

These diverse examples are united by a common value proposition for alt-data: market participants can extract an informational edge from some alt-datasets and use it to beat competitors when identifying trading opportunities.[2] This opportunistic,

[1] Some examples of conventional financial datasets include asset prices and trading volumes; corporate earnings reports; economic forecasts of employment, inflation, housing starts, and consumer spending; exchange rates; and yield curves.

[2] Less commonly, some Investors are beginning to view various alt-datasets as sources of insight for responsible investing (e.g., as providing information about environmental, social, or governance impacts of investable companies). As we discuss later in this article, the value proposition of such uses for alt-data does not rely on speed. Nevertheless, such applications largely remain viewed as (at best) secondary applications for alt-datasets by most market players currently active in the alt-data space (although some experts expect it to become more primary over the coming years).

speed-centric perspective on alt-data's value is pervasive and neatly captured by the tagline of a leading alt-data platform operator: "Alternative data is untapped alpha."[3]

We argue that alt-data's core value proposition is, however, meaningfully different for Investors than that slogan would suggest. Investors (as defined earlier) have a distinct comparative advantage over other market participants: patience. Because of their long operating horizons, Investors can pursue investment strategies unavailable to other market players. This comparative advantage is more aligned with defensive and defensible approaches to alt-data than it is with the exploitative strategies that short-horizon investors tend to pursue. That is, Investors will likely be better off using alt-data in ways that are unharmed by competition over alt-data (i.e., nonrivalrous) or for activities others cannot easily replicate (i.e., excludable).[4] In rethinking how alt-data will be most valuable to long-term strategies, we contend that Investors must also rethink how they evaluate and characterize alt-data, along with whom they should partner in gaining access to alt-datasets.

Rethinking these three issues could guide Investors in selecting alt-datasets, and strategies for analyzing and acting on them, that better fit with their organizational contexts. We seek to help Investors re-examine how alt-data could best serve their needs and offer recommendations that are informed by both formal empirical findings and our own close interactions with Investors. We also explore examples of how alt-data can be creatively used in defensive or defensible strategies.

Although building capacity around alt-data is strategically valuable in its own right, doing so has the added benefit of promoting innovation. Using alt-data demands (almost by definition) that Investors depart from the status quo in their decision making. As such, thoughtful design of an alt-data program can drive innovation in all aspects of an Investor's business (e.g., creative improvements in processes, people's skill sets, and technology). Finding partnerships that facilitate, rather than forfeit, opportunities to innovate and learn from alt-data is therefore a key issue we address and one that is likely to materially affect Investors' success (with alt-data and beyond).

The rest of this article is organized as follows. We first make the case that Investors are better off designing their alt-data strategies around defensive and defensible approaches to using alt-data than aiming to use it for alpha-oriented, opportunistic purposes. We provide examples of creative uses of alt-datasets under these strategies. These examples emphasize how alt-data can be used for deeper understanding of risk and generating operational alpha. We then cover why existing systems for characterizing alt-datasets do not fit Investors' needs. We consider a replacement system that could improve the appraisal of alt-datasets in terms of how well their characteristics align with an Investor's specific objectives and capabilities. Next, we distill our empirical findings about Investors' organizational attitudes on, and capacities for, alt-data. Our analysis concludes that Investors will generally need to partner for access to alt-data and to realize efficiencies in organizing and (pre-)processing alt-datasets. We detail the benefits and costs of partnering with different types of entities and remark on how opportunities for innovation may be a core consideration in selecting alt-data partners. We then describe how the growing arms race around alt-data could affect Investors. Finally, we close by summarizing our findings and highlight additional facets of alt-data strategies that Investors might wish to rethink in the future.

RETHINKING ALT-DATA'S VALUE PROPOSITION

Although alt-data have garnered increased attention in recent years, their use in finance is not new. Alt-data have played an integral role in investing ever since humans first began keeping records of trade: They deepen the connections between financial valuations and real-world sources of value. For instance, some enterprising merchants in ancient Babylon used measurements of the Euphrates' depth and flow to gain an informational edge in trading various commodities (because they realized that these variables were correlated with market supply) (Lo and Hasanhodzic 2010).

What has recently changed about alt-data's role in finance is its degree of accessibility. Perhaps the most

[3] See: https://www.quandl.com/alternative-data.

[4] An example of a nonrivalrous application of alt-data is in screening public equities based on sustainability criteria for the underlying companies. An example of excludable alt-data use would be for due diligence on direct investments in startup companies to which an Investor has privileged access (e.g., a university endowment having first access to funding spinouts from its research laboratories). In this sense, an Investor benefits not from an alt-dataset being excludable but from its own ability to use that data being an excludable (i.e., not easily repeated or imitated) capability.

recited example of alt-data in finance is hedge funds counting cars in retailers' parking lots (which supposedly is a leading indicator of sales performance). In the past, such counts had to be made manually, with analysts physically located in or near cars they tracked. Apart from a small number of well-resourced hedge funds, few financial organizations could devote sufficient resources to such a narrow endeavor. Currently, however, these data are accessible through a subscription service to any investment organization inclined to purchase it (thanks to lower costs of satellite imagery).

More generally, the number and diversity of alt-data sources that are readily accessible to financial entities has mushroomed. The tally of large-scale alt-data vendors who specifically cater to investment organizations has gone from a few dozen to several hundred in less than half a decade.[5] The total alt-data sources potentially relevant to investment decision making that can be cheaply and easily accessed is in the many millions. Furthermore, tools for acquiring and processing these plentiful datasets are increasingly user friendly.[6] Alt-data are steadily becoming mainstream.

As a result, the rate at which any one type or source of alt-data becomes conventional—and therefore ceases to be alt-data—is likely to increase. If the value of alt-data is premised on their conferring advantages in faster exploitation of trading opportunities (as is the case for many financial-market participants), then this means the value of any given alt-dataset will probably deteriorate at an accelerating rate because both alt-data and their value are relatively determined. Notice that data may qualify as alternative at any of three levels: the firm, the industry, and the financial ecosystem as a whole. For example, a dataset may be unconventional for a given hedge fund, but not for other funds in the hedge-fund industry. Likewise, some data may be conventional for a given firm, yet be unconventional for most organizations in the wider financial system. When enough organizations make use of any alt-dataset, it stops being alternative at a system-wide level.

Similarly, two relative dimensions help determine the value of any alt-dataset: *rivalry* and *excludability*.[7] Rivalry is the extent to which one entity's use of a resource diminishes its value for another entity.[8] Excludability is the degree to which one entity can prevent another from using a resource. When alt-data's value is premised on allowing market players to better exploit trading opportunities, then alt-datasets will tend to exhibit high rivalry. Moreover, rising accessibility of many alt-datasets is tending to lead to lower excludability.[9] These trends suggest the shelf lives for alt-datasets may be shortening if their value comes solely from helping to exploit opportunities.[10]

Defensive and Defensible Value

When an alt-dataset's value is premised on it improving a market participant's ability to speedily seize trading opportunities, there is an embedded assumption that the participant will need to act quicker than others to realize that value. This value proposition for alt-data implies that alt-datasets should be more useful for financial organizations with comparative advantages in rapid execution.

Speed is, in general, not a comparative advantage for Investors, and for sound reasons: They are long-lived organizations whose success is mission critical for their beneficiaries. Building an investment strategy around speed can greatly increase the risk of

[5] Here, we make a meaningful distinction between providers or sources of alt-data (*point vendors*) and alt-data access providers (*platform vendors*). Later, we discuss why this distinction is relevant. For now, we simply note that the number of alt-data vendors vastly exceeds the number of platforms, and this gap is only likely to widen in the future.

[6] These tools may be standalone (e.g., http://scikit-learn.org/stable/) or part of the suite of offerings from an alt-data platform (i.e., an entity that offers not just alt-datasets but also additional support or tools for working with them).

[7] The dimensions of rivalry and excludability are conventionally used to classify economics goods as private, public, club, or common pool. For such purposes, rivalry and excludability are usually treated as binary categories (i.e., something is either rivalrous or nonrivalrous and excludable or not). We see them here as continuous properties.

[8] Rivalry is a *congestion effect*, which is the opposite of a *network effect* (i.e., a resource's value grows with popularity).

[9] This decreasing excludability may become more prevalent as methods for dataset emulation and replication (e.g., statistically synthesizing better proxy datasets) techniques improve. Likewise, the bigger the market for alt-data becomes, the less incentivized many vendors are likely to be, given that they may be able to maximize revenue by selling their datasets to a wider demand base.

[10] A plausible circularity may exacerbate the shrinking shelve lives of alt-datasets: As the number of alt-datasets grows, more value accrues to those market participants that build alt-data capacity, which makes providing alt-datasets that much more appealing for vendors, who then increase market supply further, and so on.

losing unacceptable amounts of capital. Because most speed-oriented strategies are expensive to implement (e.g., they usually require specialized infrastructure or talent), they are often only efficient to deploy when large amounts of capital can be allocated to them. This risk profile for speed-based investing makes it unpalatable for most long-term Investors to undertake directly. In contrast, many asset-management firms (e.g., hedge funds, active mutual funds, or other organizations that extract management fees) can be relatively short-lived entities (i.e., they may not exist after their founders leave), and their failure would usually be less socioeconomically disastrous than it would be for Investors; thus, their cost of allocating most of their capital to speed-driven strategies is far lower.

Investors are also comparatively disadvantaged in terms of agility. As noted, rising rivalry and declining excludability of many alt-datasets means that market participants who attempt to use alt-data to exploit opportunities must be somewhat flexible to succeed; when some alt-datasets lose value from becoming more conventional, others must be sought. Because alt-datasets are largely heterogeneous, organizations that design investment strategies around them need to be agile. The level of agility required for this purpose would overwhelm the data-management and governance systems of many Investors. Although it can be argued that Investors should strive to improve such systems, in many cases it is more pragmatic to align their use of alt-data with their native strengths.

Perhaps the most powerful comparative strength that Investors have is patience. Their long horizons of operation mean that Investors can reap greater gains than other market participants by being more methodical and disciplined in their investment activities. Accordingly, we assert that the deepest value proposition alt-data has for Investors entails *defensive* and *defensible* strategies.

Defensive strategies prioritize capital preservation and prudent risk-taking over speedily exploiting opportunities. Hence, defensive strategies that incorporate alt-data should be centered on pursuits such as advanced risk analysis and management or improving operating efficiencies. Done correctly, these strategies can substantially decrease the degree of rivalry over an alt-dataset (i.e., one Investor building a defensive strategy around an alt-dataset need not reduce the value to another Investor of doing likewise). Risk management and exclusionary screening in responsible/sustainable investing are quintessential examples of defensively applying alt-data: Alt-data can be an invaluable source of intelligence on environmental, social, governance, and other factors that are germane to responsible/sustainable investment decisions, and use of an alt-dataset for exploring those factors does not necessarily degrade its value for use in the same type of decisions by others.

Defensible alt-data strategies, meanwhile, can help Investors increase the excludability of an alt-dataset by either restricting access to it (e.g., via making it proprietary) or by developing execution capabilities around it that are not replicable by other market participants (e.g., through having privileged access to infrastructure deals via special relationships with local governments).

In this article, we concentrate on defensive alt-data strategies because we believe these are most broadly applicable across various Investor types and circumstances. We cover defensible strategies briefly in the final section of this article, and we reserve a detailed treatment for a companion article.[11] From what we see, the two clearest categories of defensive alt-data strategies for Investors are deeper understanding of risk (to better allocate and manage it) and driving operational alpha.

Understanding Risk

Modern efforts in risk management largely emphasize simplifying risk over deeply comprehending its sources. Put differently, such risk-management paradigms are better at detecting that specific risks have materialized in the past than revealing why they have done so. For example, they may uncover how price movements for a given basket of securities correlate when responding to some event, but they deliver scant insight into why the event transpired in the first place. For market participants that operate over short horizons, knowing the correlation may suffice for managing risk, but for Investors to better leverage their capacity for patience, understanding the reasons why can be essential.

This need to more deeply probe causality is due to the fact that correlations in conventional datasets often break down over longer horizons and typically do not reflect the entire spectrum of events that could occur over long periods of time. Alt-data can (partly) mitigate these shortcomings by supplying more context about how events in the wider world drive downside moves

[11] See Monk, Prins, and Rook (2018).

in markets. Although it is true that rapid detection of such events might allow Investors to exploit opportunities, a less rivalrous (and more durable) benefit of early detection is that it allows more time for Investors to respond to downside events once they are flagged as likely. Moreover, added context can help warn about unprecedented downside events. When more variables are tracked, there is a higher likelihood of catching anomalous behavior that heralds highly atypical events, even if the precise impacts of such events might not be immediately apparent.[12] The ability to be alerted about unusual events is of prime importance to Investors. Large market crashes practically never play out in the same ways their predecessors did, but a single crash can fully nullify many years of outstanding performance.[13]

The purpose of defensive alt-data strategies is not to totally eliminate risk exposure for Investors but more to distribute it selectively.[14] Selective risk exposure is the chief idea behind smart-beta investment strategies, which seek to control exposures by holding positions in assets that are not necessarily proportional to their respective market capitalizations. Today, many Investors pursue smart-beta investing through purposed exchange-traded funds (ETFs), but smart-beta ETFs often lack fine control over risk exposure. For one, such ETFs are usually only ever composed of public securities and thus are not helpful for controlling private-asset exposure. Second, the asset weightings for the vast majority of ETFs are based on factors derived from conventional data (e.g., company size, dividends, or price momentum). These factors mostly fail to reflect risk in any nuanced way. For finer control over risk exposures through smart-beta ETFs, Investors must purchase shares in niche ETFs that can have high liquidity risk and management fees. Finally, the programmatic rebalancing rules for passive (and many semiactive) smart-beta ETFs can create unintended—and severely disadvantageous—consequences when abrupt market downturns occur.

Judicious use of alt-data may allow Investors to deploy smart-beta (or similar) strategies in ways that avoid these shortfalls. A suitable supply of alt-data could allow Investors to design index-construction methods for public (or private) assets that create tailored, controlled risk exposures.[15]

The use of alt-data to more deeply understand risk is not confined to portfolio construction. Indeed, alt-data have applications in other areas of risk management, such as in asset oversight and due-diligence processes, especially in private markets. For example, if an Investor directly owns a real-estate development project in an emerging market, it may hire a local manager to oversee that asset's construction. However, this delegation can generate agency problems, such as when the Investor must rely primarily on the local manager's reports about the project's progress. A form of alt-data that might lessen such problems is images of shadow lengths from the project's construction site (e.g., taken from aircraft or satellites). Algorithms such as those developed by Orbital Insight are capable of converting the lengths in such images into calculations of the pace of projects so that an Investor might enjoy greater clarity about whether its local manager is providing valid reports.[16]

An example of alt-data's use for deeper understanding of risk in due diligence involves the analysis of a venture capitalist's networks in determining whether to invest in one of its funds. The relevant networks might be derived from alt-data sources, such as LinkedIn (for general partners' professional and social networks), or

[12] Consider a parable example: An island civilization that never has witnessed (or even heard of) a tsunami may nonetheless get advanced warning of an impending anomalous event because of the sudden, dramatic recession of shoreline that characteristically precedes a tsunami.

[13] Long-lived entities are more likely to encounter such crashes, so being able to not do too badly during these crashes is as good as, if not better than, exploitation speed. Investors cannot just shut their doors if they do poorly.

[14] That is, by augmenting information sets with alt-data, Investors may reduce unwanted exposures (e.g., to climate change or reputational risk of investee companies) in a more controlled way, while increasing their desired exposures.

[15] In practice, such methods might be similar to those used by Kensho Technologies to construct its *New Economy Indices*, which capture public companies' degrees of involvement in thematic technological trends, such as artificial intelligence, autonomous vehicles, or drones. To derive its indexes, Kensho uses natural-language processing to identify a company's exposure to a given trend by parsing its public filings (e.g., 10-Ks, 20-Fs) for information on (for example) product lines, supply chains, or planned capital expenditures. Although such filings do not qualify as alt-data, this approach could be applied on other, less-conventional text documents to construct indexes (e.g., sustainability reports).

[16] Another example that may materialize in the future could involve Investors using internet-of-things data feeds from their investee companies or assets. Such data could be used in risk management, help in monitoring human work patterns and information flow, give greater clarity on microjudgments, and help make valuation more real time.

built from scraping websites or digital newsfeeds (to capture what other funds were co-investors on specific deals). Because relationships are integral to most venture capitalists' success, understanding the strength or weakness of a fund manager's networks can be a crucial variable for deciding whether an Investor should allocate capital to that manager.[17]

Some other examples of how alt-data may be used defensively for understanding risk include the following:

- harvesting dynamic pricing information from online sources to garner a clearer, more real-time picture of inflation (and draw on wider or more targeted sources of pricing information than are usual in generic consumer-price indexes);
- aggregating label information (e.g., nutrition facts, ingredients lists) from food-product companies' offerings to see how they may be vulnerable to shifting dietary trends or new warnings by health agencies (Investors may then be able to compel company managers to alter their offerings—e.g., through shareholder activism for publicly traded companies);
- assembling online price and ratings histories of possible competitors (e.g., from Airbnb, TripAdvisor, or Yelp) or price series of airfares to that locale when doing due diligence on candidate direct investments in leisure-related properties (e.g., hotels or casinos);
- using microsensors (or other remote sensors) to track fluctuations in soil moisture for determining what plants are best suited to intercropping in a plantation-forestry investment; and
- controlling reputational risk from investee companies by monitoring controversies about them that arise in social-media posts (or other localized or unconventional news outlets).

Generating Operational Alpha

Alongside deeper understanding of risk, Investors can also use alt-datasets in defensive ways by turning them into sources of *operational alpha*. The chief idea behind operational alpha is to better align operating resources with investment strategies by eliminating internal inefficiencies in how investment processes are executed. This concept is (loosely) related to investment alpha, which is the generation of returns in excess of some benchmark, after adjusting for the riskiness of the assets used to generate the excess returns. Although operational alpha has a secondary benefit of (potentially) improving gross investment returns, its chief aim is to improve net returns by reducing unneeded operating costs. Because such reductions are often risk free, operating alpha can substitute for, and in many instances is superior to, investment alpha.[18] It can also complement investment alpha because it frees up room in the risk budget and thus allows pursuit of strategies with higher upside.

Alt-data can aid Investors in driving operational alpha. Perhaps surprisingly, most Investors already possess large volumes of alt-data within their own organizations. Because alt-data are defined as data not conventionally used in decision making, novel forms of internal data count as alt-data.

Aggregation and disaggregation are key to converting conventional internal data into alt-data. For instance, inventive collation and synthesis of documents (e.g., e-mails, investment memos, and contracts) can uncover precious metadata that is able to provide insights for enhancing communication, culture, negotiation, time allocation, benchmarking, and diligence. Likewise, the disaggregation of collective processes into individual contributions can give a clearer picture of where latent organizational resources—and opportunities to improve them—reside. For example, by tracking how individual internal users query and access documents in organizational databases, an Investor can construct a map of intraorganizational knowledge flows and examine the typical approaches its analysts use in problem solving. More granular visibility of these individual activities can not only expose areas for improvement but also help better identify best practices.[19]

[17] More specifically, an Investor may have little ex ante clarity about the specific startup companies in which a venture capitalist will invest (and no control over how it does so once capital is pledged). The quality of the venture capitalist's likely co-investors, however, may be easier to discern and serve as an indicator of the ultimate riskiness of its portfolio.

[18] Notably, operational alpha can be (almost or fully) market agnostic.

[19] Such added visibility of internal processes also has a potential risk-management benefit in the form of compliance. Newly legislated requirements for data handling (e.g., the European Union's General Data Protection Regulation) mandate that users be made aware of how their personal data are being treated. In the case of

Implications of a Changed Value Proposition

In rethinking the value proposition of alt-data, Investors will need to re-examine other views and approaches they have regarding alt-data. Specifically, in pursuing defensive or defensible alt-data strategies, Investors will likely need to alter how they characterize and access alt-datasets. In the next two sections of this article, we discuss pragmatic paths for addressing each of these matters.

RETHINKING HOW ALT-DATA IS CHARACTERIZED

Because the number and diversity of alt-datasets is enormous, Investors need to be discriminating when selecting which alt-datasets deserve resources (e.g., money to acquire; time to store, prepare, and analyze; and capacity to be governed). Such selectivity requires characterizing alt-datasets to establish which will be most valuable for organizational needs. As the value any dataset has to an Investor lies in the questions it can help answer, there is a need for data-characterization methods that can reflect the question-answering capabilities of datasets (alternative or otherwise).

Alt-data are defined in an exclusionary way—by stating what they are not (conventional). However, unlike alt-data's definition, a characterization system for alt-data should not be constructed around exclusion: It is more reasonable to characterize an alt-dataset by those properties that it verifiably exhibits, rather than those it does not. Problematically, however, few Investors—or, for that matter, financial organizations in general—have any such system for alt-data characterization. In fact (and as we will detail later), Investors rarely have any formal criteria for establishing whether a dataset is indeed alternative (i.e., a threshold that divides conventional from unconventional data on the basis of scarcity, novelty, or another relevant quantitative or qualitative dimension).[20]

Unsurprisingly, because few Investors have any systems for distinguishing or characterizing alt-data, few use any consistent process for valuing its worth in advancing organizational objectives. Undoubtedly, rigorous valuation of alt-data (or any data, for that matter) is a difficult undertaking and subject to wide error margins.[21] Characterization is a more achievable step: It at least facilitates judgments about whether a given alt-dataset aligns with organizational capabilities and strategic priorities. Lack of characterization systems, however, invites the expenditure of resources on alt-datasets that do not fit with organizational priorities and resources and promotes avoidable waste.

Apart from being wasteful, not having characterization systems can challenge an Investor's fulfillment of its fiduciary duties or regulatory compliance: Investors may be hard-pressed to claim that they are engaging in responsible decision making when decisions are made based on data that are not well understood (e.g., in terms of blind spots it may create). Suitably understanding data (whether alternative or conventional) in any consistent way requires a means of characterizing it.

Existing Characterization Systems

Existing systems for characterizing alt-datasets are not suitably aligned with the value propositions we have described. These existing systems either ignore the ways in which an alt-dataset is likely to create value for an Investor (and so neglect organizational context) or assume that any dataset's main use will be driving investment alpha (or a similar short-term, opportunistic pursuit).

For example, Kolanovic and Krishnamachari (2017) posited a characterization system for alt-data that focuses on the origins of datasets (Exhibit 1). This system is not ideal for Investors' purposes for several reasons. First, although it encompasses many sources of alt-datasets, it is not necessarily exhaustive. Second, it gives no indication of how valuable a given alt-dataset is to an Investor. Taxonomical schemas such as this are not best suited to help Investors evaluate alt-data.[22]

Investors, these users can be their employees. Because the definition of what constitutes personal data is evolving, Investors stand better chances of remaining compliant if they already have developed processes and systems for tracking diverse forms of internal data in their organization.

[20] More generally, many Investors have no formalized models or system for characterizing data or judging data quality.

[21] Inarguably, an alt-dataset's value should be positively related to its quality. Yet no quality metrics exist that are universally applicable across datasets or free of restrictive assumptions. We must resort to using properties of data that can serve as context-appropriate proxies for quality. It is on these properties that alt-data should be characterized.

[22] Taxonomical systems are characterization systems that are (or attempt to be) mutually exclusive and collectively exhaustive—that is, the items they characterize must fit into one, and only one, classification category within the system.

EXHIBIT 1

Kolanovic and Krishnamachari's Characterization System for Alt-Data

Source Category	Specific Alt-Data Source
Individual Processes	Social media, news and reviews, web searches, personal data
Business Processes	Transaction data, corporate data, government agency data
Sensors	Satellites, geolocation, other sensors

Source: Kolanovic and Krishnamachari (2017).

Kolanovic and Krishnamachari (2017) proposed another taxonomical schema for alt-data characterization that does embed a value proposition and strives to indicate the usefulness of alt-datasets in relation to use cases based on asset class and investing style. Unfortunately, that system is premised on investment-alpha generation, and so it does not cover defensive or defensible uses, which thus undercuts its relevance for Investors (which is further lowered by being taxonomical).

Dannemiller and Kataria (2017) avoided the taxonomical approach and instead suggested that alt-data be characterized on a "continuum … from structured to unstructured." For the purposes of indicating the likely value of an alt-dataset, using continuums, and not discrete categories, makes sense, but whether a dataset is structured or unstructured does not immediately reflect its value for an Investor. It is true that more effort may be required to extract insight from unstructured datasets (which makes them more expensive from an organizational-resource perspective), but this does not necessarily reflect the full value that an alt-dataset holds. For example, both unstructured and structured alt-datasets may be relevant (or not) for defensive or defensible approaches by Investors.

Although big data and alt-data are not perfectly identical, there are cases in which alt-data qualify as big data. It may thus be hoped that characterization schemas for big data could sometimes be applicable to alt-data. The most prevalent such schema is the *3 Vs* of big data: volume, velocity, and variety. IBM's Big Data unit suggests a further dimension: *veracity* (i.e., the degree of uncertainty around a dataset).[23] These systems are a step in the right direction because veracity, velocity (the rate at which new data arrive), and volume (the size of a dataset) could all potentially add to a dataset's value for an Investor.[24] Yet these dimensions by themselves are incomplete, and none seem to squarely encapsulate how specific properties of an alt-dataset should translate into value. For example, velocity may be important for assets that have value-determining properties, which change frequently, but not so important for those without such properties (e.g., many private assets).[25] Thus, freshness—how well a dataset reflects the most recent changes that are material for decision making—might be more appropriate. Likewise, volume seems to be less important for Investors than whether a dataset is comprehensive. That is, a dataset may contain many items (i.e., have high volume) from only a narrow number of categories of interest. In such a case, a dataset that has smaller volume, but encompasses more categories (i.e., is more comprehensive), would likely have higher value. We thus need a different characterization scheme.

The system devised by Kitchin (2015) comes closest to what Investors need. It builds upon the 3-Vs setup (but is still intended for characterizing big data, rather than alt-data) by adding four additional dimensions: *comprehensiveness*, *granularity* (how fine- or coarse-scaled the data are), *relationality* (how many fields a dataset shares with other datasets of interest), and *flexibility* (how easily new fields can be added to a dataset).[26] Comprehensiveness and granularity seem to be apt fits for Investors' purposes, but it is less clear that relationality or flexibility are pertinent concerns. Furthermore, Kitchin's scheme gives no explicit consideration to the known quality (i.e., reliability) of data. Knowing how reliable a dataset is can be essential for Investors to decide how it can be used.

Six Dimensions of Alt-Data

We adapt Kitchin's (2015) system by replacing relationality, flexibility, variety, and volume with the dimensions of *reliability*, *actionability*, and *scarcity* (and replacing the velocity dimension with the more fitting notion of *freshness*). Reliability (which covers the

[23] See: http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

[24] Velocity may concern the rate at which new datasets are onboarded or the rate at which existing ones are refreshed.

[25] Velocity may also be valuable (for example) in rapidly detecting reputational risks for Investors in social-media activity.

[26] Kitchin actually uses "exhaustivity" and "resolution" in place of comprehensiveness and granularity, respectively.

Exhibit 2
Six-Dimensional Characterization of Alt-Data

Dimension	Explanation
Reliability	How accurate, precise, and verifiable the data are (e.g., error-free, unbiased, checkable)
Granularity	The scale covered by specific data points or entries (e.g., continental, industry-wide)
Freshness	Age of the data (i.e., when collected/generated) relative to the phenomena they reflect
Comprehensiveness	What portion of a given domain the data cover (e.g., 25% of households in Canada)
Actionability	Degree to which significant actions or decisions can be made based on the data
Scarcity	How widely or readily available the data are to other (especially competing) organizations

Source: Authors.

accuracy, precision, and verifiability of a dataset) seems to us a more fitting concept than IBM's veracity. Reliability essentially equates with the known quality of a dataset.[27] Actionability and scarcity are loosely related to, but distinct from, the ideas of rivalry and excludability. In a sense, actionability and scarcity are primitives of rivalry and scarcity. For rivalry to matter, an alt-dataset must be actionable (i.e., it needs to be usable for decisions that lead to actions). Likewise, when rivalry is a concern, it is valuable to have access to scarce (albeit relevant) datasets. Excludability refers to scarcity that is (semi-)permanent. Hence, this characterization schema helps clarify not only what kinds of questions can be answered by a particular alt-dataset but also what kinds of strategies that an alt-dataset may usefully inform (see Exhibit 2 for further details on each of these dimensions).

How do these six characterization dimensions meaningfully contribute to an alt-dataset's potential value in defensive or defensible strategies? The first three dimensions' contributions are relatively clear-cut (although they are also relevant for opportunistic strategies). Because alt-data's purpose is to guide decisions, it should be trustworthy (and, in some cases, transparently verifiable). Similarly, decisions can be made at different levels, and those made at highly specific levels often require very fine data, whereas high-level decisions can usually be made on less granular (or, at least, more highly condensed) data. Lastly, decisions should not be made on stale data for which more recent versions exist. High freshness is thereby desirable in most cases.[28] What qualifies as *high*, however, can vary with the nature of the decisions that are made based on the dataset in question.

The chief way our proposed characterization is more applicable to defensive and defensible alt-data strategies than it is to opportunistic strategies is in importance of comprehensiveness.[29] For opportunistic uses, alt-data need not be comprehensive: They can encompass narrow ranges of instances or categories and still deliver genuine advantages. Although narrow alt-data can still be useful for defensive or defensible purposes, comprehensive datasets are generally more valuable because they give more complete visibility and scope. This greater breadth of coverage is useful for a deeper understanding of risk situations or internal inefficiencies (for defensive strategies), as well as for more exhaustive awareness of ways in which defensible advantages might be vulnerable.

Actionability is highly important for both defensive and defensible approaches because alt-data that

[27] Reliability includes how verifiable a dataset is. Verifiability here has two aspects: (1) how readily a dataset's accuracy can be confirmed by using other datasets and (2) the clarity of its provenance. Also note that the first four elements (reliability, granularity, freshness, and comprehensiveness) may be seen as referring to a dataset's richness. Importantly, these four dimensions appear to be the most objective and universally applicable across Investors (it can be argued that scarcity depends on substitutability, which may differ for some Investors—depending on their specific organizational contexts): These dimensions could therefore potentially be standardized to some degree to allow faster assessment of alt-datasets. This might be a useful enterprise for some commercial organization (e.g., an alt-data platform vendor) to undertake in the near future (it also is one that could generate considerable efficiency gains for Investors).

[28] Desirability of low latency does not mean longer time series of alt-data are less valuable. Latency in the case of time series refers to the most recent record in a series. The length of the time series instead reflects its comprehensiveness.

[29] Although we expect that this characterization will likely be useful for many financial-market participants, we realize that the relative importance of each dimension will likely differ across entities (or different types of financial entity).

cannot be translated into proactive or reactive actions are of little (or no) practical value to any Investor. Scarcity has different bearings for defensible and defensive strategies. For the former, its value is more directly connected to excludability. For the latter, scarcity is more related to the rate at which alt-data spread to different financial organizations. If some alt-dataset is very accessible (e.g., public information) and many organizations begin noticing and acting on it at once, there can be systematic effects, which can be troublesome from a risk-management standpoint. Meanwhile, alt-datasets whose scarcity declines slowly can enable more considered and advantageous reaction.

RETHINKING ACCESS TO ALT-DATA

In addition to rethinking the value proposition of alt-data and how they are characterized, Investors might need to rethink how they access alt-data. Indeed, the first two reconsiderations are irrelevant if Investors cannot access alt-data. How any Investor should appropriately access alt-data is a joint function of (1) what entities can provide it and how they go about doing so and (2) what the Investor's current organizational capabilities in and attitudes toward alt-data are. Answers to these questions will necessarily vary to some degree across Investors. Our research indicates, however, that some generalizations can be made so that a typical recommendation can be safely made to Investors. Succinctly, we find evidence that Investors are eager to tap the potential benefits of alt-data but are, on average, not (yet) adequately equipped to independently source, process, and maintain alternative-data resources. However, these current circumstances do not suggest that Investors should abandon efforts to build internal alt-data capabilities by surrendering all alt-data functions to third parties—especially to external asset managers. Instead, we find it reasonable that Investors should prioritize partnerships with platform providers of alt-data (at least for the near-term future).

In the remainder of this section, we first explore empirical evidence on Investors' current capabilities in, and organizational stances on, alt-data. We then turn to how these findings intersect with the different alt-data access modes available to Investors. A focal component of our analysis here is how alt-data can be used as an accelerant for various forms of organizational innovation.

Empirical Findings on Alt-Data in Institutional Investment

The findings reported in this subsection are drawn from extensive interviews with senior decision makers across a diverse sample of institutional-investment organizations, along with results from a survey of Investors. We describe these studies more extensively later, but we first give an overall synopsis.

Succinctly, Investors' current relationships with the rise of alt-data can be described as considerably interested yet significantly underprepared. More fully, we observe the following:

- Investors pervasively believe that alt-data can be used to improve net investment returns, but many are unconvinced that their organization is well equipped to use alt-data to do so.
- Few Investors have a formalized strategy regarding alt-data or are actively developing one.
- Many Investors worry about alt-data costs, specifically to develop in-house capability.
- Investors widely view building or acquiring proprietary alt-datasets as a way to succeed with alt-data and feel that the most valuable use of alt-data is in identifying opportunities.

Both survey evidence and content from interviews provide rationale for, and additional details on, these summary findings. Regarding the former, our survey instrument was completed in February 2018 by senior decision makers (i.e., chief executive officer, chief information officer, chief technology officer) from 22 leading institutional-investment organizations. Collectively, respondent organizations manage over US$1 trillion; they represent a diverse mix of geographies (Australasia, Europe, Middle East, and North America), fund types (sovereign wealth funds, endowments, public pension funds), and fund sizes.

Although 70% of respondents feel that alt-data could help improve risk-adjusted returns in their organization, 90% state that their organization has no "defined alternative-data strategy" (of the 10% that do have alt-data strategies, all admitted that these strategies are "not well developed"). Furthermore, less than 15% claim their organization is "equipped to handle" multiple forms of alt-data (nearly 30% strongly disagree that they are equipped). Less than one-third report that alt-data are a "priority" for senior management, although 60% of respondents note that their organization is actively monitoring developments in alt-data or considering creating capacity in alt-data.

In aggregate, these response patterns depict an uneasy tension. Investors are clearly aware of alt-data's potential benefits but are not situating themselves strategically to reap these benefits. This awareness-without-progress could drive a reactive need to catch up in the future and cause alt-data strategies to be less carefully designed than they might have been with proactive planning.

Respondents also believe speed and quality are significantly more important properties for alt-data than are granularity or volume.[30] Over 80% claim "opportunity identification" to be the capability that alt-data could improve most within their organizations ("risk management" was selected by less than 10% of respondents). These answers indicate a view that the primary beneficial application of alt-data is in allowing rapid detection of mispriced assets (e.g., arbitrages).

Finally, among survey respondents, a lack of "suitable ways to invest" (i.e., actionability) is stated to be the "biggest challenge" to their effective use of alt-data (32%), followed by the state of their existing technology (23%), analytic capability (23%), organizational culture (18%), and trust in alt-data from key decision makers (4%). We comment on the gravity of these challenges shortly.

To validate our survey findings and probe the situations behind them, we conducted a series of in-depth, semistructured interviews with seven of the respondents (one-third of the full sample). Interviews were conducted by telephone and lasted between 30 and 45 minutes. Overall, these interviews not only confirmed results from the survey but also provided additional details germane to Investors' perspectives about alt-data. First, none of the interviewee organizations have formal definitions for what constitutes alt-data. Such definitions are, arguably, a prerequisite for prudent alt-data strategies. Second, interviewees voiced concern over both the costliness of acquiring alt-data and their organizations' ability to be competitive in their usage of alt-data. Worries about cost fixate on how expensive interviewees think it will be to conduct alt-data operations in-house. Relatedly, although respondents generally feel that they could become as capable as their peers in developing alt-data functions, they are unsure about whether they can compete with other entities (especially hedge funds) when it comes to their ability to use (i.e., analyze) alt-datasets. Third, interviewees confirmed the survey finding that rapid identification of mispriced assets is the application for alt-datasets with which Investors are most (and, for some, exclusively) familiar. Fourth, a consensus emerged among interviewees that alt-data are most valuable if they are proprietary.

Two other notable points arose in the interviews, concerning (1) data provenance and (2) cooperation. For some Investors, a key stall point is how transparent an alt-dataset's lineage is (i.e., how clear is knowledge of its source, what transformations have been performed on it, and who performed them). Several interviewees noted that their organizations would have reservations about making decisions based on alt-data of uncertain provenance and that murky provenance could dissuade or prevent them from using third-party alt-data. Furthermore, most interviewees agree that their organizations would very likely cooperate with peers in building alt-data capacity.

Modes for Accessing Alt-Data's Benefits

In sum, the preceding observations strongly demonstrate that Investors do not appear prepared to go it alone in sourcing, processing, or maintaining either a wide or deep array of alt-data. Yet, the results also indicate that Investors seem sufficiently interested in alt-data to be unlikely to ignore it altogether. Nor should they. For reasons already mentioned, alt-data could serve Investors as a crucial resource. The question then surfaces of how Investors should access alt-data.

We see two assisted paths Investors might follow in accessing alt-data. The first involves trusting external third parties (including asset managers) to provide indirect access. That is, those access providers take care of the difficult tasks of sourcing, managing, and acting on alt-data, and Investors reap some of the benefits that they may have otherwise received from handling the alt-data themselves. This path addresses the realization that accessing alt-data should not be an end goal in its own right for Investors. Instead, they should aim to maximize the benefits from alt-data.

[30] The fact that respondents do not feel alt-dataset size is of primary importance is reinforced by the fact that a majority (72%) answered that alt-data are not "essentially the same as big data."

Nonetheless, offloading alt-data responsibilities onto access providers deprives Investors of a pivotal benefit that building alt-data capacity could provide to them: accelerating innovation. Investors generally struggle with innovation (Monk and Rook 2018). Alt-data, however, supply a springboard for innovation. By definition, the use of alt-data in decision making requires at least some innovation by Investors. In many cases, the amount of innovation itself may be modest, but the amount of learning from it (which could drive future innovation) can be significant.

Moreover, alt-data is a topic that invites considerable excitement and stirs imaginations: It is a sexy concept in finance. Investors can often struggle with innovation simply because they lack internal agreement (within their organizations) about what resources deserve innovation. Alt-data's allure could make it a common point of agreement for coalescing support for innovation.

As we elaborate later, outsourcing alt-data capabilities—such as relying exclusively upon external third parties for indirect access to alt-data—could cause a sizable sacrifice in innovation capabilities for Investors. We believe that many, if not most, Investors should be thinking about how to build in-house capacity around alt-data, especially for defensive and defensible strategies.[31] The degree and nature of this capacity will need to vary with each Investor's own organizational context, but every Investor is indeed capable of building such capacity—to at least a minor extent.

The drive to build some internal alt-data capacity—coupled with the fact that Investors are not ready, by and large, to undertake the sourcing and management of alt-data all by themselves—suggests the second assisted path by which Investors may feasibly access alt-data: alt-data vendors. Two main types of alt-data vendors can be distinguished: point vendors, who offer either a single or limited number and type of alt-dataset; and platform vendors, who tend to offer wider selections of alt-datasets and may additionally offer integration or analytical tools that aid use of alt-datasets.

In the following, we compare prospects and demerits of Investors seeking alt-dataset access through both kinds of vendor, in relation to one another and in relation to external access providers. On the latter, we focus on the impacts of Investors relying on external asset managers for alt-data.

External Asset Managers as Access Providers

Some external asset managers (e.g., some hedge funds) have enjoyed relatively lengthy experience in working with alt-data—at least when compared to Investors. Given Investors' widespread desire to gain exposure to the benefits of alt-data but lack of full capacity to do so at present, it may seem advisable that they seek indirect access through such managers. If doing so came only at the cost of forfeiting some experience with learning to innovate, this option might be recommendable. However, there are at least three additional reasons why it is not. The first stems from the opportunistic nature of most external asset managers. In general, external managers are less incentivized to be concerned about capital preservation and are more motivated to fixate upon investment alpha than are Investors. These differences are not by themselves inherently problematic, given that external managers often are able to build stronger comparative advantages in generating investment alpha than are many Investors (although such advantages are routinely on a gross basis and may not hold once costs are fully considered). What is troublesome, however, is the fact that this emphasis on alt-data for opportunity identification and exploitation predisposes external asset managers to becoming engulfed in an escalating arms race around alt-data. We discuss the drivers, dynamics, and likely implications for Investors of that arms race in the next section.

A second major reason why it might not be recommendable for Investors to rely too heavily on external managers for alt-data access involves transparency and provenance. When Investors outsource their alt-data efforts to external managers, they lose the ability to inspect, verify, and otherwise work with the data on which those managers are basing decisions. Not only does this loss translate into opportunity costs from forgone innovation opportunities, it also creates issues around lack of visibility and verifiability. In not directly accessing alt-data used by their external managers, Investors are forced to rely on those managers to establish and maintain their quality. As we explain in the next section, however, heightening competition over

[31] If Investors are electing not to build in-house capacity, then we recommend that the decision result from thorough analysis of long-term trade-offs to the organization (e.g., from loss in potential innovation versus resource absorption).

alt-data may well push external managers to accept and execute investment decisions on alt-datasets of increasingly lower quality, which can inject unforeseen (and sometimes unidentifiable) risk into Investors' portfolios. The likelihood of transparency problems will probably worsen as competition over alt-data grows; managers should then tend to be more secretive about their processes around and sources of alt-data.

Another major reason why Investors should restrict reliance on external asset managers in accessing alt-data is the subsidization of a capability gap. That is, although some external managers presently possess some comparative advantages over Investors when it comes to alt-data, those advantages need not be permanent. Whenever an Investor contracts an external manager to invest on its behalf, and the manager makes use of alt-data to do so, that Investor is effectively subsidizing the manager in improving its capacity for alt-data relative to the Investor's own capacity. This subsidization thus increases both the manager's comparative advantage and the Investor's reliance on external parties for alt-data capacity, reducing the Investor's future strategic flexibility around alt-data.

Access through Alternative-Data Vendors

The path of building increasing internal capacity around alt-data through partnering with vendors mitigates or eliminates many of the aforementioned problems with relying on external managers. First, vendors are (usually) just providers of alt-data, the use of which is determined by Investors. Hence, vendor-supplied alt-data do not necessarily expose Investors to problems connected with opportunistic usage of alt-data (although, as mentioned earlier, many vendors do stress the alpha-generating merits of their datasets). Second, concerns about transparency are partly lessened when Investors access alt-data directly through vendors rather than indirectly through external managers; in the former instance, Investors are actually able to examine the alt-datasets. To be clear, being able to actually work with the data directly does not eliminate the possibility of errors or other quality problems in the data. Yet such possibilities are typically more investigable (i.e., Investors may be able to request assurances about the secure provenance of the alt-datasets) than they are with external managers. Furthermore, because quality and trustworthiness are dimensions on which vendors compete with one another, many are incentivized to remain highly transparent.

A third concern that is alleviated by partnering with vendors rather than asset managers is that of subsidization. It is true that whenever an Investor subscribes to or buys an alt-dataset from a third-party vendor, it is subsidizing that vendor's comparative advantage in sourcing alt-data (and possibly cleaning or preprocessing alt-data, depending on the services that vendor provides). When creating defensible strategies around proprietary alt-datasets, this subsidization may be problematic. However, we expect that most Investors will instead favor defensive applications of alt-data, in which case such subsidization would actually tend to be helpful for Investors: It would help fund the vendor's provision of additional alt-datasets, and so would further benefit Investors.

Additionally—depending on its infrastructure and particular method of accessing alt-data from vendors—experimenting with different forms of alt-data may be substantially easier through vendors than through external asset managers. That is, switching between vendor subscriptions is, in many situations, likely to be less arduous than switching allocations to different external asset managers. Thus, partnering with vendors may allow Investors to try out more configurations of alt-data when attempting to incorporate it into their strategies, thus increasing their odds of finding a good fit.

Still, the path of accessing alt-data via vendors is not without its downside.[32] The foremost of these is the low degree of excludability for vendor-supplied alt-data. Of course, when Investors' use cases for alt-data are predominantly defensive, excludability becomes less worrying. Likewise, when Investors use alt-data to build capabilities that are defensible (even when the alt-data upon which they are based are not), such as privileged access to deal flows, excludability is not a concern.

Moreover, higher (if not total) excludability can often be achieved at higher cost: Vendors may be willing to provide more exclusive access to alt-datasets for premium prices. In many cases, therefore, Investors that

[32] One particular challenge that Investors may face in relying on platform vendors to access alt-data is whether external data provided by the vendor can be easily integrated with the Investor's internal data—without Investors losing control over their internal data or giving others access to it. Tackling this challenge could help vendors distinguish themselves.

access alt-data through vendors can balance dataset cost against scarcity. Striking such a balance may frequently entail working with multiple vendors. In so doing, any Investor should consider the relative advantages and disadvantages of point and platform vendors.

Point vendors tend to be more specialized than platform vendors.[33] The former therefore can often provide more novel, differentiated alt-datasets. Moreover, because point vendors have fewer product offerings than platform vendors, they may be able to verify a larger fraction of their data more intensively than platform vendors (although this need not always be true). Point vendors, however, often have smaller markets for their offerings than do platform vendors, which can bundle together multiple datasets to broaden their appeal. This narrower market for many point vendors means that their costs can be higher than their platform counterparts and so put them out of reach for smaller Investors (or those with less budgetary room for alt-data). Also, point vendors can face diseconomies of scope that are less severe for platform vendors. For instance, it is typical that platform vendors can deliver alt-datasets in a single format or offer more streamlined integration (through, e.g., standardized APIs).[34] Doing so simplifies access for Investors—relative to having to integrate multiple, distinct formats from point vendors.[35]

We anticipate that many Investors who partner with third-party vendors to serve their alt-data ambitions will select a limited (e.g., one or two) number of platform partners and supplement the alt-datasets offered by these platform vendors with specific alt-datasets accessed through point vendors.

THE ESCALATING ALTERNATIVE-DATA ARMS RACE

Rethinking alt-data—in terms of its value proposition, characterization, and access—will almost surely be a strenuous process for most Investors. Might it not be better for some to avoid involvement in alt-data altogether? We think not. As we explain here, an arms race around alt-data is underway and gathering momentum across financial markets. The ways in and extent to which we foresee this race escalating lead us to believe that Investors will not be able to escape becoming meaningfully affected by it. We advise that they try to proactively engage with alt-data by building defensive and defensible alt-data strategies, rather than being dragged along in a reactive manner.

Arms-Race Logic

In an elegant application of formal economic logic, Grossman and Stiglitz (1980) proved that the persistence of efficient equilibria is impossible in financial markets. They did so by highlighting a fundamental paradox. Market efficiency is driven by profit-motivated market participants who aim to exploit the mispricing of financial assets through transacting, based on information they possess. In transacting, they jointly increase market efficiency and decrease the value of their information. In (the strongest forms of) equilibrium, there is no unexploited information and so no incentives for participants to either transact or seek out additional information to exploit. However, because the wider world is never in stasis—new information is arriving all the time—markets cannot be permanently in equilibrium. If they were, then there would be no (nonrandom) transacting, which would permit existence of unexploited information and thereby mean that no equilibrium existed, by definition.

Paradoxically, competition is a force that makes markets more efficient but also ensures that they cannot become entirely efficient. An unceasing inflow of new information and data is the key to this seeming contradiction. If no new data about the wider world were to be created, then markets would (hypothetically) settle into equilibrium, but because the world is ever changing, there is continual production of new information.[36]

[33] Examples of platform providers that specialize in alt-data include Neudata and Quandl. More traditional financial-data platforms, such as Bloomberg and FactSet, also are increasing their alt-data offerings. Interestingly, a new type of alt-data entity also seems to be emerging that offers analysis of specific types of alt-datasets, rather than just providing access to them (in some instances such entities do not provide access to the alt-datasets themselves). Examples of these new kinds of entity include Orbital Insight (for satellite-image analysis) and Predata (for social-media analytics).

[34] Integration difficulties may (initially) favor platform providers that specialize in conventional data but offer alt-datasets as an additional service. Increasingly, incumbent providers of conventional data are also offering alt-data.

[35] Platforms may also prove a more efficient way for Investors to keep pace with changing data regulations, under the assumption that the chosen platform can be trusted to stay current with data legislation and related compliance issues.

[36] More than 90% of all digital data that have ever existed was created in the last two years (see, e.g., Henke, Libarikian, and Wiseman 2016).

Ongoing competition among market participants to exploit this new information and data squarely qualifies as an *arms race*, which is definable as a situation in which parties are locked in perpetual efforts to outcompete one another, without a defined endpoint. Thus, any effort at active investing amounts to participating in a data arms race. Still, this race is useful: If all participants were passive, then markets would not function.

Every Investor is therefore directly affected by active investing, even if its own strategies are fully passive. By merely deploying capital in public markets (which every Investor does), they are exposed to the active-investing activities of other parties, which affect the volatility and liquidity of their own portfolios. Much of this active investing is done by non-Investor asset managers, who are either hired by or compete with Investors. Thus, all Investors are directly affected by the arms race for data in general (not just alt-data) that is continually underway in public markets. To better understand the consequences of an alt-data arms race for Investors, we should understand what drives the intensity of data arms races more broadly. To that end, rivalry and excludability are core forces.

Role of Rivalry and Excludability

The intensity of data arms races is fueled by the rivalry and excludability of the datasets based on which their participants aim to make investment decisions. Practically all data in finance are rivalrous in the sense that any use of data for transacting reduces (or even eliminates) the value in executing similar transactions thereafter, regardless of who conducts them. This property means any (profitable) actionability of data is eventually self-eliminating so that the value of a dataset decreases by acting on it. This self-eroding value of data's actionability can, however, be partly offset by scarcity. The fewer entities that have access to a dataset, the more proportional value can be kept by those with access. Scarcity is a crucial reason why alt-datasets can be so precious. Most conventional datasets in finance are nonexcludable.[37] Entities with them cannot readily bar others from getting them, and when they transact on these datasets, others can better divine their content, which devalues them more (i.e., they devalue when first transacted on as a result of decreased actionability, and then again from reduced scarcity).

Alt-data, meanwhile, are typically more excludable—and so any specific alt-dataset tends to be scarcer—than are conventional data. Some alt-datasets can be *permanently excludable*: Those who create or acquire them first can prevent all others from possessing and transacting on them. More typically, alt-datasets are *limitedly excludable*: Entities with them can only exclude others from acquiring them (or replicating them, to some approximation) for a limited time or else can only restrict the number of others who obtain them to a limited extent. Consequently, excludability of many alt-datasets means substantial value can be realized by being first to capture a dataset, even if it cannot be immediately acted on (i.e., scarcity might offset low near-term actionability).[38]

This interplay among competition, rivalry, and excludability underpins the intensity of current land grabs for alt-data (i.e., an alt-data arms race) in global financial markets. Moreover, the combination of these factors creates perverse incentives for market participants to (1) overweight specific facets of alt-datasets when evaluating them, (2) focus on short horizons, and (3) potentially overprice alt-datasets of undetermined value. Valuing an alt-dataset is an uncertain business. Its full richness (i.e., comprehensiveness, reliability, granularity, and freshness) is often hard to establish without spending much time working with it. Likewise, the complete set of ways in which it is actionable may not come into focus until it is more thoroughly processed and analyzed. These layers of uncertainty mean that a hierarchy often emerges for alt-datasets, whereby scarcity and immediate actionability trump other characteristics.

The primacy of these two factors, in light of the limited excludability of many alt-datasets, means

[37] This low excludability is increasing the need for financial organizations to conceal their digital activities (i.e., reduce their digital footprints) so that their data and information inputs are less inferable by other, competing organizations.

[38] Alt-data that concern sustainable/responsible investing may be somewhat different from other forms of alt-data in this respect. Investors may well benefit from reducing the excludability of alt-data that are relevant for sustainable/responsible investing (e.g., that relate to environmental, social, or governance factors or sustainable development); in doing so, they might benefit from the emergence of stronger standards and norms regarding sustainable/responsible investment practices.

short timeframes can easily become overemphasized.[39] First, accentuation of datasets that have immediate actionability naturally biases use of them toward the short term. Second, limited excludability creates an impetus to act before others are able to acquire or create substitute datasets. Third, the outsized value of scarcity can encourage *data hoarding*, whereby entities leap before looking and obtain alt-datasets that promise high scarcity and excludability but only minimally consider the actionability of such alt-datasets upfront. Data hoarding can lead to *strategic misfits*, that is, alt-datasets that are poorly aligned with organizational capabilities or priorities and so have low long-term strategic value. Datasets with sufficiently low value can drive pursuit of shorter payback periods to offset their costs and thus compress the time horizons of decisions made with them.

For entities that can cope with, or even excel at, concentrating on short horizons (e.g., some hedge funds), the current intensity of the alt-data arms race may be meaningfully beneficial and increase rewards for their comparative advantages in speed or agility. In general, Investors are not in this group. By and large, their foremost advantage is patience and the ability to operate over long timescales. Unfortunately, Investors' involvement in this arms race is not readily avoidable, which is a problem because the race shows no sign of abating soon. On the supply side, an increasing number of sources and formats for new data continues to emerge. Meanwhile, proliferation of advanced analytic tools, such as deep-learning platforms, are stoking fiercer competition over alt-data.

Sticky Consequences for Investors

Few, if any, Investors will be able to successfully decouple themselves from the alt-data arms race. Its stickiness will mean that Investors cannot insulate themselves from it and still achieve current risk and performance targets. A pivotal realization here is that market competition makes alt-data a moving target. In not using alt-data, market participants handicap themselves by limiting any informational edge that they can possess over other participants. As more participants begin to acquire and transact on any specific type of alt-data (if not the same alt-dataset), however, that type starts to become conventional data, which then lifts the net value of other unconventional datasets. In short, opportunity costs for many market participants, especially non-Investor asset managers, become too great to not seek and use alt-data. As more market participants embrace alt-datasets, markets (especially public markets) will be more affected by them, until they affect even passive investing.

A vital question for Investors engaged in predominantly passive strategies is how alt-data's increased influence over market activity will change the character of that activity itself. How will the rising intensity of the alt-data arms race alter the nature of risk in markets? There is a reasonable case to be made that the increased intensity of this race will not lower volatility in public markets. Indeed, the opposite appears to us more probable, due to (at least) three factors. For one, pressures toward short-termism that we discussed earlier bias decisions toward action rather than inaction. More market activity means greater volatility. Furthermore, intensified competition over alt-data means that there is pressure not only to act fast but also to act big because of fleeting actionability. Possessing a unique and excludable alt-dataset does not block other entities from eroding its actionability by acting on it first: There is reason to act not only swiftly but also extensively to prevent actionability from evaporating. More extensive activity also increases volatility. Finally, the increasing use of algorithmic methods for trading based on alt-data will likely contribute to higher market volatility. Increased volatility will probably raise costs of passive investing through a combination of higher transaction costs (because of faster turnover), hedging costs, liquidity threats, and cash drag.[40] Whether these negative possibilities might push more Investors away from passive strategies is not yet clear.

[39] Alt-datasets that are perfectly excludable can still create bias toward short-term action. In contrast, alt-datasets with limited excludability carry additional pressure because of their wasting nature, which can encourage use-it-or-lose-it mentalities. Furthermore, many limitedly excludable alt-datasets are cheaper and quicker to capture than are perfectly excludable ones.

[40] One way to temper risk in passive investing is to increase portfolio allocations to cash, versus the market portfolio. Because the return on cash will not necessarily be increased because of higher volatility in the wider market portfolio, there will be likely be opportunity costs in gross (and possibly net) returns when cash allocations increase (i.e., cash drag).

But if the alt-data arms race succeeds in shifting more capital to active-investment strategies, then a circularity might arise: More money pumped into active investing would raise the value in using alt-data for active investors, which would increase the intensity of the arms race around alt-datasets. This is a perilous treadmill for Investors and threatens their interests.

We have already asserted one way to avoid stepping onto that treadmill: concentrating on cultivating defensive and defensible alt-data strategies. Such approaches could partly immunize Investors against the arms race over alt-data but, by themselves, may not be sufficient. To properly insulate themselves from the alt-data arms race, Investors might need to bolster their capabilities in real-asset investments, such as natural resources and infrastructure. These types of investment have risk profiles distinct from public securities and naturally lend themselves to more defensive applications of alt-data. Moreover, real-asset investments generally allow Investors to more fully exercise their comparative advantages in long-term investing. Rethinking the value proposition for alt-data could therefore go hand in hand with rethinking the composition of long-term portfolios.

SUMMARY

The rising accessibility and diversity of and competition over alt-datasets presents Investors with novel challenges. We believe that these challenges give Investors cause to rethink how they will strategically engage with alt-data. Escalating competition for alt-datasets means that Investors are unlikely to remain unaffected by alt-data and that their strategic planning should take this fact into account. Potential opportunities afforded by defensive and defensible alt-data strategies give Investors ample reasons not only to seek access to alt-datasets but to build internal capacity for working with and acting on them. Cultivating such capacity could be a key engine for innovation.

Although we see many merits for Investors in directly engaging with alt-data, we point out that not all Investors should do so in the same ways or to equal degrees. Defensive and defensible alt-data strategies should be designed in ways that respect the specific resources and organizational contexts of individual Investors, which necessarily means that such strategies will differ from one Investor to the next. However, they need not differ so extensively that Investors cannot beneficially work together in growing their capacities for alt-data, including collaborating to generate and share alt-datasets with one another. We investigate these collaborative opportunities in a companion article.

In closing, we remind Investors of the advantages in being open-minded about alt-data and specifically about taking a wide view on how they can leverage alt-data that already exist in their own organizations. Such data need not be exotic or complicated to be valuable. Indeed, the rising sophistication, but user friendliness, of many data-science tools should cause an increasing number of internal alt-datasets to be significant sources of operational alpha within the immediate future. Moreover, Investors should bear in mind that alt-datasets that relate to internal operations have a very valuable property: They are maximally excludable and thus a fully defensible form of data.

ACKNOWLEDGMENTS

We thank survey and interview participants for their valuable contributions.

REFERENCES

Dannemiller, D., and R. Kataria. 2017. "Alternative Data for Investment Decisions: Today's Innovation Could Be Tomorrow's Requirement." Deloitte Center for Financial Services, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-fsi-dcfs-alternative-data-for-investment-decisions.pdf.

Grossman, S., and J. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70, no. 3 (June): 393–408.

Henke, N., A. Libarikian, and B. Wiseman. 2016. "Straight Talk about Big Data." *McKinsey Quarterly*, October, https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/straight-talk-about-big-data.

Kitchin, R. 2015. "The Opportunities, Challenges and Risks of Big Data for Official Statistics." *Statistical Journal of the IAOS* 31 (3): 471–481.

Kolanovic, M., and R. Krishnamachari. 2017. "Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing." Report, J.P. Morgan, May 18.

Lo, A., and J. Hasanhodzic. *The Evolution of Technical Analysis: Financial Prediction from Babylonian Tablets to Bloomberg Terminals*. New York: Bloomberg Press, 2010.

Monk, A., M. Prins, and D. Rook. 2018. "Playing Data Defense in Institutional Investing." Mimeo.

Monk, A., and D. Rook. 2018. "The Technological Investor: Deeper Innovation through Reorientation." SSRN eLibrary, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3134078.

A Machine Learning Approach to Risk Factors: *A Case Study Using the Fama–French–Carhart Model*

JOSEPH SIMONIAN, CHENWEI WU, DANIEL ITANO, AND VYSHAAL NARAYANAM

JOSEPH SIMONIAN
is the director of quantitative research at Natixis Investment Managers in Boston, MA.
joseph.simonian@natixis.com

CHENWEI WU
is a quantitative analyst at Natixis Investment Managers in Boston, MA.
chenwei.wu@natixis.com

DANIEL ITANO
is a senior quantitative analyst at Natixis Investment Managers in Boston, MA.
daniel.itano@natixis.com

VYSHAAL NARAYANAM
is a data science co-op at Natixis Investment Managers in Boston, MA.
vnarayanam@natixis.com

Perhaps the most important defining characteristic of factor models is that they must explain asset behavior to a sufficient degree given a restricted set of explanatory variables. Given this, the primary challenge for anyone building a factor model is to settle on a set of factors that on one hand can adequately explain portfolio behavior over time, and on the other is simple enough to remain computationally tractable. In this way, the challenge faced in building a factor model is the same faced by scientists when building theories to explain natural phenomena, in which the trade-off between informative power and simplicity is also a fundamental consideration.

Although it is generally accepted that factor models should be built based on the foregoing principles, we often see practitioners developing and using factor models that deviate from them in significant ways. This is especially the case with the models that underlie many commercially available risk platforms, which often include hundreds of correlated variables that are presented as factors. The reason why commercial risk platforms take a maximalist approach to factor modeling is likely rooted in their motivation to provide a comprehensive picture of the risk exposures driving portfolio behavior. It is also rooted in their use of linear models. A linear factor model that restricts itself to a small number of factors faces the risk of providing an inadequate picture of portfolio behavior over a given measurement period. As a result, commercial risk platforms try to cover their bases by including a multitude of factors so that no exposure is seemingly unaccounted for. Despite this technical maneuver, the resulting frameworks are usually not genuine linear factor models, because of their size and the presence of correlated variables, nor are they maximally informative, because of their inability to account for the nonlinear behavior of and/or interaction effects among factors.

A natural response to the shortcomings of linear factor models is to recommend the use of nonlinear factor models; however, parametric nonlinear models have a number of shortcomings. First, the structure of the latter models is often heavily dependent on the sample data. As the sample expands or contracts, we at times find that the function specified by the model changes, sometimes dramatically. Second, unlike linear models, parameter estimates cannot always be derived analytically. Rather, solutions are often found using iterative methods, in which initial values are posited for each unknown target variable before various optimization techniques are invoked to home in on a solution. Although iterative methods can be useful, the optimizations that drive them may ultimately fail to converge if the initial values are too distant from possible solution values. Initial values that are remote from optimal values can also cause convergence to a local solution rather than a global solution.

As a remedy to the drawbacks of both linear models and parametric nonlinear models, in this article the authors present a factor framework based on a machine learning algorithm known as *random forests* (RFs) (Ho 1995, 1998; Breiman 2001). The authors show how to use the RF algorithm to produce models that, within a single framework, provide information regarding the sensitivity of assets to factors broadly analogous to those generated by more commonly used frameworks, but with a significantly higher level of explanatory power. Moreover, RF-based factor models are able to account for the nonlinear relationships, discontinuities (e.g., threshold correlations), and interactions among the variables, while dispensing with the need for complex functional forms or additional interaction terms (thus remaining in harmony with the principle of parsimony). In the last section of the article, the authors demonstrate how the framework can be combined with another machine learning algorithm known as *association rule learning* (ARL) to build effective trading strategies, using a sector rotation strategy as an example.

BASIC FEATURES OF FACTOR MODELS

Investment factor models are supposed to provide insight into the primary drivers of portfolio behavior. Formally, there are various ways to build a factor model (for a basic overview, see Connor 1995). Perhaps the simplest way is via an ordinary least squares (OLS) regression, in which the portfolio return is the dependent variable, and the risk factors are the independent variables. As long as the independent variables have sufficiently low correlation, different models will be statistically valid and explain portfolio behavior to varying degrees. In addition to revealing what percentage of a portfolio's behavior is explained by the model in question, a regression will also reveal the sensitivity of a portfolio's return to each factor's behavior. These sensitivities are expressed by the beta coefficient attached to each factor.

Factor sensitivities and measures of explanatory power are the defining characteristics of factor models and are present in other common frameworks, such as those based on principal component analysis (PCA) (see Jolliffe 2002). As we show later, factor models based on machine learning can also describe the sensitivity of variables to the factors that explain them and provide information relating to the overall explanatory power of a given model. However, as previously mentioned, they also offer some distinct advantages over more traditional frameworks, such as the ability to capture nonlinear behavior and the interaction effects between factors. Additionally, RF models are generally less influenced by correlations between variables. Indeed, the question of multicollinearity does not enter into the picture when building an RF model in the way it does in an OLS regression.[1] One reason for this is that unlike OLS regression, RF models are estimated without requiring the inversion of a covariance matrix. Another distinct advantage of RF models is that they do not have strict parametric assumptions, nor do they rely on other time series assumptions such as homoskedasticity or independence of errors. Nevertheless, although RF models are relatively rule-free, it is our view that a fair amount of pre-model work should be done to ensure that the inputs into the model make sense from the standpoint of both investment relevance and economic coherence and possess a sufficient level of factor uniqueness to produce models that are both practical and free from explanatory redundancies. Although factor selection is an important aspect of building any factor model, it is especially critical when using machine learning-based methods.

MACHINE LEARNING AND THE RANDOM FOREST ALGORITHM

Machine learning refers to a collection of computational techniques that facilitate the automated learning of patterns and the formation of predictions from data. As such, machine learning methods can be used to build models with minimal human intervention and pre-programmed rules. Machine learning algorithms are (very) broadly classified as either *supervised* or *unsupervised*

[1] Those concerned with multicollinearity may benefit from using PCA or LASSO (least absolute shrinkage and selection operator) in the pre-model stage of an analysis to aid in generating factors that are unique.

learning algorithms. Unsupervised learning algorithms include those encompassing clustering and dimension reduction, in which the goal is to draw inferences and define hidden structures from input data. Unsupervised algorithms are distinguished by the fact that the input data are not categorized or classified. Rather, the algorithm is expected to provide a structure for the data. A well-known example of an unsupervised learning algorithm is *k*-means clustering (Lloyd 1982). In contrast, supervised learning (including reinforcement learning) algorithms use input variables that are clearly demarcated. With supervised learning, the goal is to produce rules and/or inferences that can be reliably applied to new data, whether for classification or regression-type problems. The RF algorithm used in this article is an example of a supervised learning algorithm and has been shown to be extremely effective in a variety of scientific applications, such as medical diagnosis, genome research, and cosmology.

Some machine learning algorithms, including RF, incorporate decision trees, a tool that is helpful in analyzing and explaining complex datasets. For regression-type problems, decision trees start from a topmost or *root node* and proceed to generate *branches*, with each branch containing a condition, and a prediction in the form of a real-valued number, given the condition in question. Trees are composed of a series of conditions attached to *decision nodes*, which ultimately arrive at a *leaf* or *terminal node* whose value is a real number.[2] The latter value represents a predicted value for a target variable given a set of predictor values. In Exhibit 1, we show a simple example of a decision tree that analyzes the relationship between monetary policy and macroeconomic conditions.

Decision trees can be constructed using various procedures (e.g., ID3, CHAID, MARS). In this article, we use a procedure known as CART (classification and regression tree, a methodology developed by Breiman et al. 1984). CART uses an algorithm called *binary recursive partitioning*, which divides the input space into binary decision trees. In this procedure, features are evaluated using all sample values, and the feature that minimizes the cost function at a specific value is chosen as the best split. Recursive partitioning takes place at each level down the tree, and the value at each leaf of the tree is the average of all the resulting observations.

In Exhibits 2, 3, and 4, we proceed to describe binary recursive partitioning and the RF algorithm in formal detail. In doing so, we use (with some modification) the descriptions provided by Cutler, Cutler, and Stevens (2012). We begin with the definition of binary recursive partitioning in Exhibit 2.

RF uses an ensemble of decision trees in conjunction with the CART technique. Each tree in the ensemble is constructed via bootstrapping, which involves resampling from the data with replacement to build a unique dataset for each tree in the ensemble. The trees in the ensemble are then averaged (in the case of regression), resulting in a final model. The bootstrap aggregation of a large number of trees is called *bagging*.[3] We describe the RF algorithm formally in Exhibit 3.

Predicted values of the response variable for regression[4] at a given point x are given by

$$\hat{f}(x) = \frac{1}{J}\sum_{j=1}^{J}\hat{h}_j(x)$$

where $\hat{h}_j(x)$ is the prediction of the response variable at x using the jth tree (This formula concludes the algorithm in Exhibit 3).

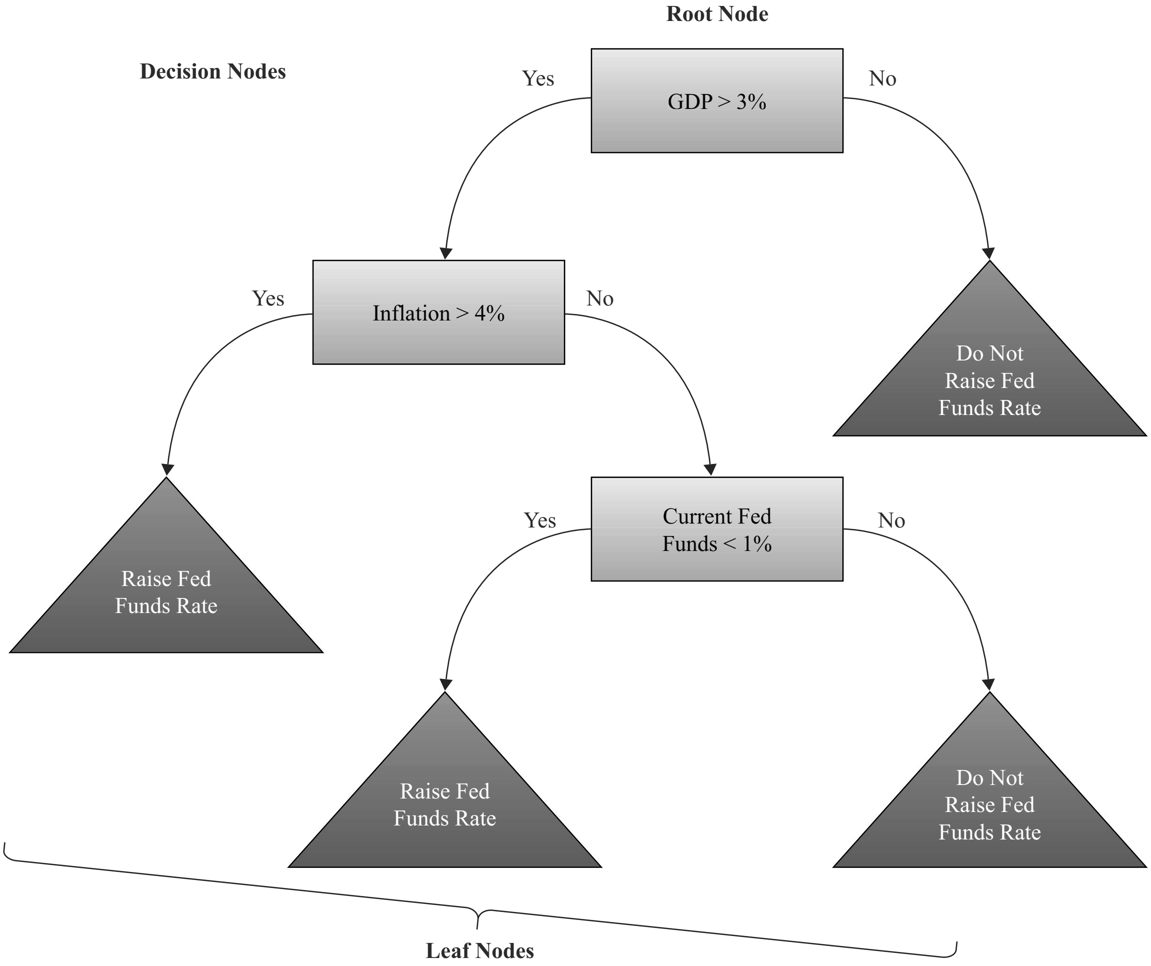
When a bootstrap is conducted, some observations are left out of the bootstrap. These are called *out-of-bag* (OOB) data and are used for measuring estimation error and variable importance. If trees are large, using all the trees may produce a false level of confidence in the predictions of the response variable for observations in the training set $\mathcal{D}$. To remedy this risk, the prediction of the response variable for training set observations is done exclusively with trees for which the observation is OOB. The resulting

[2] Several stopping criteria can be used to halt the tree-building process—for example, a minimum number of samples in a leaf, the depth of the tree, and the total number of leaves.

[3] Bagging is useful because it generally reduces overfitting and has a lower variance when compared to processes that only use individual decision trees. An individual tree may end up learning highly idiosyncratic relationships among the data and hence may end up overfitting the model. Averaging ensembles of trees provides a better opportunity to uncover more general patterns and relationships between variables. Overfitting can also be addressed by using simpler trees (i.e., those with a lower number of splits).

[4] For classification, the prediction values are given by $\hat{f}(x) = \text{argmax}_y \sum_{j=1}^{J} I(\hat{h}_j(x) = y)$.

EXHIBIT 1
Example of a Decision Tree



predictions, fittingly labeled *out-of-bag predictions*, are defined in Exhibit 4.

For regression[5] with squared error loss, generalization error is generally measured using the OOB mean squared error (MSE): $MSE_{OOB} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}_{OOB}(x_i))^2$.

An RF analysis produces two basic outputs. The first output is simply a set of conditional values—for example, a set of factor returns and a predicted value for a dependent variable such as a portfolio return, given the posited factor returns. The second output is something called *feature importance* (FI). As its name implies, FI indicates the importance of each explanatory variable in contributing to the predicted value of the dependent variable in question.

We calculate FI using *mean decrease accuracy*, which measures the degree to which the predictive power of the model would be diluted if the values for the explanatory variable in question were randomly changed. The mechanics of FI measurement work as follows: Once the *j*th tree is generated, the values for the predictor variables are randomly permuted in the bootstrapped sample, and the prediction accuracy is recalculated. For regression, the FI for the observation is calculated as the difference between the MSE of the predictions using the permuted

[5] For classification with zero-one loss, the generalization error rate is given by $E_{OOB} = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{f}_{OOB}(x_i))$.

Exhibit 2

Algorithm for Binary Recursive Partitioning

```
Let $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ denote the training data, with $x_i = (x_{i,1}, ..., x_{i,p})^T$.
    1. Start with all observations $(x_1, y_1), ..., (x_n, y_n)$ in a single node.
    2. Repeat the following steps recursively for each unsplit node until the stopping criterion is met:
        a. Find the best binary split among all binary splits on all $p$ predictors.
        b. Split the node into two descendant nodes using the best split (step 2a).
    3. For prediction at $x$, pass $x$ down the tree until it lands in a terminal node. Let $K$ denote the terminal node, and let $y_{k_1}, ..., y_{k_n}$ denote the response values of the training data in node $k$. Predicted values of the response variable for regression are given by $\hat{h}(x) = \overline{y}_k = \frac{1}{n}\Sigma_{i=1}^{n} y_{k_i}$
```

Note: For classification, the prediction values are given by $\hat{h}(x) = \text{argmax}_y \sum_{i=1}^{n} I(y_{ki} = y)$ *where* $I(y_{ki} = y) = 1$ *if* $y_{ki} = y$ *and 0 otherwise.*
Source: Cutler, Cutler, and Stevens (2012).

Exhibit 3

Algorithm for Random Forests

```
Let $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ denote the training data, with $x_i = (x_{i,1}, ..., x_{i,p})^T$.
For $J = 1$ to $j$:
    1. Draw a bootstrap sample $\mathcal{D}_j$ of size $N$ from $\mathcal{D}$.
    2. Using the bootstrap sample $\mathcal{D}_j$ as the training data, fit a tree using binary recursive partitioning (Exhibit 2):
        a. Start with all observations in a single node.
        b. Repeat the following steps recursively for each unsplit node until the stopping criterion is met:
              i. Select $m$ predictors at random from the $p$ available predictors.
             ii. Find the best binary split among all binary splits on the $m$ predictors from step i.
            iii. Split the node into two descendant nodes using the split from step ii.
```

Source: Cutler, Cutler, and Stevens (2012).

Exhibit 4

Algorithm for Out-of-Bag Predictions

```
Let $\mathcal{D}_j$ denote the $j$th bootstrap sample and $\hat{h}_j(x)$ denote the prediction $x$ from the $j$th tree, for $j = 1, ..., J$. For $i = 1$ to $N$:
1. Let $\mathcal{J}_i = \{j: (x_i, y_i) \notin \mathcal{D}_j\}$, and let $J_i$ be the cardinality of $\mathcal{J}_i$ (Exhibit 3).
2. Define the OOB prediction for regression[6] at $x_i$ to be $\hat{f}_{\text{OOB}}(x_i) = \frac{1}{J_i} \sum_{j \in \mathcal{J}_i} \hat{h}_j(x_i)$.
```

Note: For classification, the OOB prediction is given by $\hat{f}_{OOB}(x_i) = \text{argmax}_y \sum_{j \in \mathcal{J}}^{j} I(\hat{h}_j(x_i) = y)$, *where* $\hat{h}_j(x_i)$ *is the prediction of the response variable at* x_i *using the jth tree.*
Source: Cutler, Cutler, and Stevens (2012).

data and the MSE of the predictions using the original data: $FI_j(i) = MSE_{OOB_{Permuted}} - MSE_{OOB}$.[6]

Next, a normalization is generally conducted to allow an assignment of a relative FI (RFI) value to each feature. The normalization is accomplished by adding the FI values for each factor in a single tree and dividing that value into the FI value for each factor. This will yield a cross section of FI values that sum to unity. This operation is repeated for each tree, and the normalized FI (NFI) values are then averaged across all the generated trees to produce an RFI value for a given feature *k*—that is, $RFI(k) = \frac{\sum_{j=1}^{J} NFI_j(k)}{J}$. The RFIs will also fall in the range [0,1] and sum to unity. As we shall see in the forthcoming sections of the article, the RFI measure plays a pivotal role in building and interpreting RF-based approaches to factor modeling.

BUILDING FACTOR MODELS USING RANDOM FORESTS

Factor models are generally articulated as linear models despite the drawbacks highlighted earlier. Linear models are preferred by practitioners because they generally present readily understandable and interpretable analysis. In contrast, machine learning approaches, although useful in uncovering the nonlinear behavior of and interaction relationships among variables, are often articulated in a way that makes their output unintuitive, and hence unattractive, to many investment professionals. Nonetheless, as we shall demonstrate, it is possible to interpret the results of an RF factor analysis in a way that is both tractable and practical.

To frame our discussion, we use a variant of the well-known Fama–French–Carhart (FFC) equity factor model (Fama and French 1992, 1993; Carhart 1997). The FFC model is a multifactor extension of the capital asset pricing model (CAPM) (Treynor 1961; Sharpe 1964; Lintner 1965; and Mossin 1966), where the market represents the sole source of systemic risk. The FFC model extends the CAPM framework by introducing three new factors in addition to the market factor: the size factor (small-cap stock returns minus large-cap stock returns), value (high book-to-price stock returns minus low book-to-price stock returns), and momentum (high-returning stocks minus low-returning stocks).[7] Both the CAPM and the FFC are typically expressed as linear models. Thus, the RF variant of the FFC presented here provides a counterpoint to its traditional representation.

We use the FFC model to explain the performance of the 10 primary sectors of the stock market, with each sector represented by its respective Dow Jones index. In an RF model, the FFC factors function as features that we use to predict return values for each of our sectors. It is thus possible to examine how various factors influence the predicted values for a target variable when the latter takes on different values. A natural way to do this is to divide the predicted sector returns into percentiles and observe the value that factor returns take at each of them. Here we select observations that map to the 10th, 25th, 50th, 75th, and 90th percentile values of the target variable, then observe how each factor's influence on the predicted value of the target variable differs at each percentile.[8] Doing this produces

[6] For classification, $FI(i) = E_{OOB_{Permuted}} - E_{OOB}$. We note that although we have chosen to use MSE as our operative measure of feature importance, it is not the only one available. Other commonly used metrics include mean absolute error, the Gini index, and entropy.

[7] The following are detailed factor descriptions obtained from Ken French's website (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html):

- $R_m - R_f$, the excess return on the market, is the value-weighted return of all CRSP firms incorporated in the United States and listed on the NYSE, AMEX, or NASDAQ that have a CRSP share code of 10 or 11 at the beginning of month *t*, good shares and price data at the beginning of *t*, and good return data for *t* minus the one-month Treasury bill rate.
- SMB (small minus big) is the average return on three small portfolios minus the average return on three big portfolios: $$SMB = \frac{Small\ value + Small\ neutral + Small\ growth}{3} - \frac{Big\ value + Big\ neutral + Big\ growth}{3}$$
- HML (high minus low) is the average return on two value portfolios minus the average return on two growth portfolios: $$HML = \frac{Small\ value + Big\ value}{2} - \frac{Small\ growth + Big\ growth}{2}$$
- Momentum is the average return on two high prior return portfolios minus the average return on two low prior return portfolios: $Mom = \frac{Small\ high + Big\ high}{2} - \frac{Small\ low + Big\ low}{2}$

[8] It is also possible to organize the explanatory variables into percentiles and investigate how the predicted values for the target variable change in response to significant shifts in the values of the predictors.

information regarding the sensitivity of sector returns to factor returns that is similar to that provided by a quantile regression (Koenker 2005). Observing variable behavior across percentiles is useful because doing so often reveals asymmetric relationships between factors and target variables within a set of observations.

In Exhibit 5, we show the returns for each FFC factor at different percentiles, as well as the predicted equity sector return. We also show the RF model R^2 value and the OLS R^2 value for each sector. As the exhibit shows, the R^2 produced by the RF model for each sector is, in general, significantly higher than that produced using an OLS regression. We also see that examining sector returns at different percentiles allows us to observe the varying influence of the FFC factors as the level of the predicted sector returns changes. In some cases, we see significant divergences between sector and factor returns. For example, for the consumer staples sector, at the 10th and 25th percentiles, all of the FFC factor returns are significantly more negative than the sector's predicted returns. One can interpret this result as reinforcing the sector's reputation as a defensive "low beta" sector. The opposite is true for the financials and materials sectors, whose predicted 90th percentile returns are significantly higher than the FFC factor returns.

In Exhibit 6, we show the RFI of each factor. Again, a factor's RFI indicates its importance in predicting sector returns when compared to the other factors in a set. Because RFI values sum to unity, they are naturally viewed as weights. As such, RFI values can plausibly be used to offer guidance in portfolio construction along the lines of a traditional returns-based style analysis (RBSA).[9] Assuming that investible proxies are available for the factors used in a given model, RFI values can be used to inform the weighting of the proxies used as constituents in a portfolio seeking to mimic the behavior of a target strategy. The RF model, however, possesses an advantage over a standard RBSA in generally providing a much better fit, as evidenced by the R^2 values it produces.

USING FEATURE IMPORTANCES TO DERIVE PSEUDO-BETAS

Because the RF model captures hierarchical (non-geometric) relationships between factors, it cannot be understood as a direct analog of an OLS regression or PCA because it does not convey the individual directional relationships between factors and assets. It is nevertheless possible to provide an interpretation of the RF model output so that the influence of the predictors can be understood in a way that is similar to traditional models. Previous attempts to "beta-ize" tree-based predictors have, for the most part, been of a more formal nature (e.g., Friedman 2001). Here we take a more conceptual approach because our goal is merely to provide a translation of the RF model output to individuals who are more familiar with linear models. We do not recommend using the results of the translation for trading applications, but simply as a communication device.

Recall that a widely accepted definition of *beta* is the elasticity of one variable to another. If we assume factor independence, then as a first step we can simply divide the predicted target variable return by each predictor return to gain a raw elasticity value for each factor. For example, let us consider the returns at the median for the industrials sector in Exhibit 5 referenced earlier. In the following, we list the sector and factor returns, along with the raw factor elasticity values in parentheses next to each.

The raw elasticity values indicate a sort of ceteris paribus degree of target variable sensitivity to each predictor; however, the raw values provide an incomplete picture of the relationship between target and predictor variables because they do not account for each factor's importance as a predictor, something expressed by RFI values. As such, our second step is to weight each factor's

Industrials	Rm-Rf	SMB	HML	MoM
1.3%	0.9% (1.44)	–0.1% (–13.0)	–0.3% (–4.33)	0.3% (4.33)

[9] Returns-based style analysis was introduced by Sharpe (1988, 1992). It is a way of analyzing and replicating investment strategies by means of investable proxies. The analysis is regression based, expressed formally as $R_t^m = \alpha + \sum_{i=1}^{I} \beta_i R_t^i + \epsilon_t$, where R_t^m is the return stream for the investment strategy to be replicated, R_t^i is the set of return streams for the proxy returns, I is the number of investable proxies, and ϵ_t is the error term. Two important constraints are put in place to produce a combination of investable proxies suitable for a long-only implementation. First, each beta coefficient is constrained to be greater than zero—that is, $\beta_i > 0$, $\forall i$. Second, the sum of the betas is constrained to sum to unity—that is, $\sum_{i=1}^{I} \beta_i = 1$. As such, each beta is interpreted as a weight assigned to a particular investable proxy in a replication portfolio.

EXHIBIT 5
Factor Percentile Returns and Equity Sector Predicted Values (monthly returns, Jan 1991 to Aug 2018)

Percentile	Consumer Discretionary	Rm-Rf	SMB	HML	MoM	Consumer Staples	Rm-Rf	SMB	HML	MoM
10th	–5.2%	–4.7%	–3.3%	–3.0%	–4.7%	–3.3%	–12.4%	–10.8%	–7.6%	–21.2%
25th	–1.7%	–1.9%	–1.8%	–1.5%	–1.4%	–0.9%	–5.3%	–3.6%	–3.6%	–5.4%
50th	1.3%	1.4%	0.3%	0.2%	1.0%	1.0%	2.2%	1.0%	1.0%	1.9%
75th	4.4%	3.7%	2.4%	2.1%	3.3%	2.9%	5.1%	3.4%	3.4%	4.6%
90th	6.7%	5.9%	3.9%	3.9%	5.2%	4.1%	11.4%	22.1%	12.9%	18.3%
R^2	0.96					0.91				
OLS R^2	0.81					0.44				
Percentile	**Energy**	**Rm-Rf**	**SMB**	**HML**	**MoM**	**Financials**	**Rm-Rf**	**SMB**	**HML**	**MoM**
10th	–4.6%	–4.5%	–3.3%	–2.9%	–4.2%	–5.9%	–3.2%	–2.8%	–2.4%	–2.9%
25th	–2.1%	–1.2%	–1.3%	–1.1%	–0.8%	–2.4%	–1.2%	–1.2%	–1.1%	–0.6%
50th	0.8%	0.8%	–0.2%	–0.3%	0.2%	1.6%	1.4%	0.3%	0.1%	0.9%
75th	4.4%	3.8%	2.5%	2.2%	3.4%	4.1%	3.9%	2.5%	2.3%	3.5%
90th	6.5%	7.0%	5.0%	5.5%	7.0%	7.9%	4.8%	3.2%	3.2%	4.1%
R^2	0.90					0.96				
OLS R^2	0.40					0.84				
Percentile	**Health Care**	**Rm-Rf**	**SMB**	**HML**	**MoM**	**Industrials**	**Rm-Rf**	**SMB**	**HML**	**MoM**
10th	–5.5%	–17.2%	–17.3%	–11.1%	–34.4%	–4.6%	–4.8%	–3.5%	–3.2%	–4.9%
25th	–1.3%	–3.2%	–2.4%	–2.0%	–2.8%	–1.4%	–1.4%	–1.4%	–1.2%	–1.0%
50th	1.2%	1.4%	0.3%	0.3%	0.9%	1.3%	0.9%	–0.1%	–0.3%	0.3%
75th	3.3%	3.1%	1.5%	1.3%	2.6%	3.7%	3.2%	1.7%	1.5%	2.6%
90th	3.3%	3.1%	1.5%	1.3%	2.6%	6.4%	5.2%	3.5%	3.5%	4.6%
R^2	0.91					0.97				
OLS R^2	0.48					0.83				
Percentile	**Information Technology**	**Rm-Rf**	**SMB**	**HML**	**MoM**	**Materials**	**Rm-Rf**	**SMB**	**HML**	**MoM**
10th	–6.3%	–5.1%	–3.6%	–3.2%	–5.0%	–4.9%	–3.7%	–3.0%	–2.6%	–3.5%
25th	–2.9%	–3.0%	–2.7%	–2.3%	–2.7%	–2.7%	–2.0%	–1.9%	–1.6%	–1.6%
50th	1.6%	0.6%	–0.3%	–0.3%	0.2%	0.8%	0.6%	–0.4%	–0.4%	0.1%
75th	5.5%	5.2%	3.5%	3.5%	4.6%	3.2%	3.0%	1.6%	1.7%	2.6%
90th	9.0%	7.1%	5.4%	5.8%	7.6%	7.6%	4.7%	3.3%	3.2%	4.5%
R^2	0.96					0.94				
OLS R^2	0.77					0.66				
Percentile	**Telecom Services**	**Rm-Rf**	**SMB**	**HML**	**MoM**	**Utilities**	**Rm-Rf**	**SMB**	**HML**	**MoM**
10th	–6.7%	–9.2%	–5.4%	–6.2%	–9.2%	–5.1%	–12.7%	–11.1%	–7.9%	–21.5%
25th	–2.1%	–2.6%	–2.2%	–1.8%	–2.1%	–1.6%	0.3%	1.3%	0.6%	1.0%
50th	1.0%	1.4%	0.2%	0.1%	0.8%	1.1%	1.5%	0.5%	0.4%	1.0%
75th	3.8%	6.1%	5.9%	4.8%	7.0%	3.2%	4.0%	2.3%	2.3%	3.5%
90th	5.0%	7.9%	6.8%	7.5%	9.9%	5.5%	6.5%	4.5%	4.8%	6.2%
R^2	0.89					0.87				
OLS R^2	0.40					0.21				

Source: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html and Natixis Investment Managers.

Exhibit 6

Relative Feature Importance of Fama–French–Carhart Factors (Jan 1991 to Aug 2018)

	Rm-Rf	SMB	HML	MoM
Consumer Discretionary	0.84	0.04	0.05	0.07
Consumer Staples	0.50	0.20	0.15	0.15
Energy	0.52	0.13	0.18	0.17
Financials	0.72	0.07	0.16	0.05
Health Care	0.53	0.19	0.16	0.13
Industrials	0.84	0.05	0.07	0.04
Information Technology	0.73	0.05	0.15	0.08
Materials	0.69	0.07	0.11	0.13
Telecom Services	0.49	0.18	0.17	0.16
Utilities	0.37	0.25	0.20	0.18

Source: Natixis Investment Managers.

respective raw elasticity by its RFI to obtain a set of importance-adjusted elasticity values or *pseudo-betas*:

Rm-Rf	SMB	HML	MoM
1.21	–0.65	–0.30	0.17

We formally express the entire operation as

$$\frac{\text{Target variable value}}{\text{Predictor value}} \times \text{Feature importance}$$

Again, it is important to keep in mind that the intent here is not to discard the actual results of the analysis but to provide a simple way to facilitate communication with investment professionals who are accustomed to OLS betas, PCA loadings, and the like.

TRADING APPLICATION: BUILDING A SECTOR ROTATION STRATEGY USING THE RF FFC MODEL AND ASSOCIATION RULE LEARNING

In the previous sections of the article, we have shown how to use the RF algorithm to decompose risk ex post. In what follows, we adapt the framework for its use ex ante in trading applications. In particular, we apply our RF variant of the FFC model to build a sector rotation strategy. In doing so, we demonstrate how combining the output of an RF model with a simple, almost primitive signal can generate tradable information and provide the rudiments to developing a more sophisticated investment strategy. We do this to demonstrate the power of the RF model and to show that its effectiveness as an alpha generation tool does not necessarily depend on a complicated implementation.

We develop our trading strategy with the help of another machine learning methodology known as association rule learning (ARL) (Agrawal, Imieliński, and Swami 1993). ARL is a framework originally developed for discovering the relationships between sets of variables in a database. It can alternatively be viewed as a framework for deriving (learning) deductive inference rules from empirical data. In our example, we use ARL to establish a relationship between a pair of signals and the one-month-ahead return for a given sector over 18-month rolling windows. The signals are the RF-predicted return of a sector and the ratio of shorter-term to longer-term realized volatility (24-month vs. 36-month).[10] If (1) a positive relationship has been established between our signals and the one-month-ahead returns over the preceding 18-month window, (2) the ratio of shorter-term to longer-term volatility is less than one, and (3) the RF-predicted return for next month is greater than a designated threshold value, then we will own the sector for the month. Otherwise, the portfolio will carry a zero weight in the sector. The sectors that are owned will be equally weighted. We describe the association, trading, and portfolio construction rules that frame the strategy in formal detail in Exhibit 7.

We display out-of-sample backtest (Panel A) and bootstrap[11] results (Panel B) in Exhibit 8, comparing both unconstrained and constrained versions of our active strategy with a passive equal-weight portfolio.[12] In Exhibit 9, we show the cumulative out-of-sample backtest performance of each strategy. As we see in each exhibit, the active strategy outperforms the "no information" equal-weight portfolio, both in unconstrained form and with turnover constraints. The active strategy also exhibits respectable values for the

[10] The ratio of longer-term to shorter-term volatility has also been shown to reinforce other types of market signals (e.g., momentum). See Wang and Xu (2015) and Simonian et al. (2018).

[11] For the bootstrap, we use the stationary bootstrap approach described by Politis and Romano (1994), with an average block size of six months. The values in the exhibit are obtained by averaging 500 bootstrap samples.

[12] For each asset, $\textit{Constrained active weight} = \textit{Unconstrained active weight} \times 30\% + \frac{70\%}{\textit{\# Assets}}$.

Exhibit 7
Sector Rotation Strategy Rules

Association Rules[a]

Association rule: $X \Rightarrow Y$

Support of factors X in the set of observations T: $supp\ (X) = \dfrac{|\{t \in T; X \subseteq t\}|}{|T|}$

Lift of a rule, $X \Rightarrow Y$: $lift\ (X \Rightarrow Y) = \dfrac{supp\ (X \cup Y)}{supp\ (X) \times supp\ (Y)} = \dfrac{|\{t \in T; X \cup Y \subseteq t\}| \times |T|}{|\{t \in T; X \subseteq t\}| \times |\{t \in T; Y \subseteq t\}|}$

Trading Rules[b]

Association rule: $(VR_t < 1, \hat{r}_{t+1} > \alpha) \Rightarrow r_{t+1} > \beta$

where

$\hat{r}_{t+1}$ = RF prediction for one-month-ahead sector return

VR_t = *Volatility ratio at time t* = $\dfrac{Volatility\ (t-24,\ t)}{Volatility\ (t-36,\ t)}$

$VR_t < 1$: Near-term volatility is smaller than long-term volatility.

α and β are non-negative thresholds. In our example, we posit a value of 0 for α and 0.02 for β.

Support of the signal: $supp\ \big((VR_t < 1,\ \hat{r}_{t+1} > \alpha)\big) = \dfrac{\#\ of\ months\ (VR_t < 1,\ \hat{r}_{t+1} > \alpha)}{Window\ Length}$

Lift of the rule: $lift\ \big((VR_t < 1,\ \hat{r}_{t+1} > \alpha) \Rightarrow r_{t+1} > \beta\big) =$

$$\frac{\dfrac{\#\ of\ months\ (VR_t < 1,\ \hat{r}_{t+1} > \alpha,\ r_{t+1} > \beta)}{Window\ Length}}{\dfrac{\#\ of\ months\ (VR_t < 1,\ \hat{r}_{t+1} > \alpha)}{Window\ Length} \times \dfrac{\#\ of\ months\ (r_{t+1} > \beta)}{Window\ Length}}$$

Trading and Portfolio Construction Rules

1. *IF*, at the end of month t:
   a. A sector's $lift_{\ (t-18,t)} > 1.1$
   b. $VR_t < 1$
   c. $\hat{r}_{t+1} > \alpha$
2. *THEN* $w_i > 0$
3. *ELSE* $w_i = 0$
4. *Portfolio construction Rule:* $w_i = \dfrac{100\%}{N}$ if $\boldsymbol{N} \neq 0$; $w_i = \dfrac{100\%}{\#\,Assets}$ if $N = 0$

where w_i is the portfolio weight of asset i, and N is the number of sectors with a positive weight in the portfolio.

[a] *An additional way of measuring rule strength is via the confidence of a rule, X ⇒ Y, where*

$$conf\,(X \Rightarrow Y) = \frac{supp\ (X \bigcup Y)}{supp\ (X)} = \frac{|\{t \in T\}; X \bigcup Y \subseteq t|}{|\{t \in T\}; X \subseteq Y|}$$

[b] *It is also possible to construct our trading rule using the confidence of a rule:*

$$conf\ \ ((VR_t < 1, \hat{r}_{t+1} > \alpha) \Rightarrow r_{t+1} > \beta) = \frac{\dfrac{\#\ of\ months\,(VR_t < 1, \hat{r}_{t+1} > \alpha, r_{t+1} > \beta)}{Window\ length}}{\dfrac{\#\ of\ months\,(VR_t < 1, \hat{r}_{t+1} > \alpha)}{Window\ length}}$$

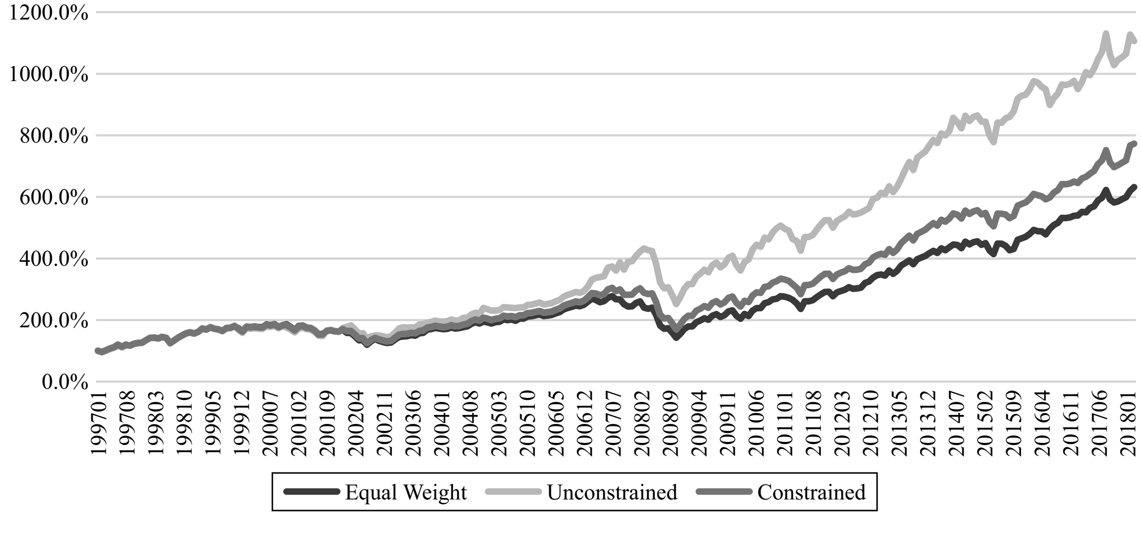
Source: Agrawal, Imieliński, and Swami (1993).

EXHIBIT 8
Sector Rotation Strategy Backtest and Bootstrap Results (Jan 1997 to Aug 2018)

	Annualized Return	Annualized Volatility	PSharpe 0.0	PSharpe 0.1	PSharpe 0.2	PIR 0.0	PIR 0.1	PIR 0.2	Turnover	Total Cumulative Return
Panel A: Out-of-Sample Backtest Results										
Equal-Weight	10.0%	13.8%	99.2%	78.8%	23.4%	N/A	N/A	N/A	0.0%	632%
Unconstrained Strategy	12.9%	14.2%	99.9%	94.9%	53.8%	96.9%	60.3%	10.3%	74.6%	1106%
Constrained Strategy	11.0%	13.6%	99.7%	87.3%	35.0%	97.8%	66.1%	13.6%	22.4%	773%
Panel B: Bootstrap Performance Results										
Equal-Weight	11.0%	13.4%	96.7%	78.9%	43.0%	N/A	N/A	N/A	0.0%	982%
Unconstrained Strategy	12.2%	14.4%	97.6%	82.2%	45.8%	66.7%	25.7%	4.2%	85.4%	1202%
Constrained Strategy	11.4%	13.4%	97.3%	81.3%	45.9%	66.7%	25.8%	4.3%	25.6%	1041%

Source: Natixis Investment Managers.

EXHIBIT 9
Cumulative Out-of-Sample Backtest Performance of Sector Rotation Strategy vs. Equal-Weight Portfolio (Jan 1997 to Aug 2018)



probabilistic Sharpe ratio (PSharpe) introduced by Bailey and López de Prado (2012)[13] and favorable values for the information ratio variant of the PSharpe (PIR), where the equal-weighted portfolio is used as the benchmark. The PSharpe measure is designed to show the probability of a strategy achieving a given Sharpe ratio threshold given a specific track record or backtest length and the

[13] The PSharpe is defined as

$$\widehat{PSR}(SR^*) = Z\left[\frac{(\widehat{SR}-SR^*)\sqrt{n-1}}{\sqrt{1-\hat{\gamma}_3 SR^* + \frac{\hat{\gamma}_4-1}{4}\widehat{SR}^2}}\right]$$

where Z [·] is the cumulative distribution function of a standard normal distribution, and $\widehat{SR}$ is the observed Sharpe ratio. SR^* is the predefined benchmark Sharpe ratio (ex ante Sharpe ratio), n is the number of periods over which the strategy's performance is tested, and $\hat{\gamma}_3$ and $\hat{\gamma}_4$ are the respective observed skewness and kurtosis values of the strategy.

presence of non-normal returns. The comparatively favorable results for our active investment strategy demonstrate that even with the barest of inputs, machine learning methods—and the RF and ARL frameworks in particular—provide powerful means to uncover useful patterns in investment data. It is a given that the strategy presented here could be built up and improved upon, with the introduction of new factors and/or a more nuanced treatment of existing inputs. Nevertheless, the results here convincingly speak to the investment insights that can be gained by practitioners willing to incorporate machine learning methods into their investment process.

CONCLUSION

Machine-learning approaches to risk factor modeling offer investment practitioners the ability to enrich their analysis by providing insight into relationships between variables that are unaccounted for in more traditional models such as OLS regression. By means of the RF algorithm, the authors uncover nonlinear relationships and interaction effects between the well-known FFC factors and show how to translate the output from the RF model so that it has the basic form of a more traditional factor model. In the last section of the article, the authors combine the RF algorithm with another machine learning framework, association rule learning, to build a sector rotation strategy. The article thus demonstrates that machine learning approaches can inform both risk analysis and portfolio management, providing readily usable output that can be communicated in a straightforward manner.

REFERENCES

Agrawal, R., T. Imieliński, and A. Swami. "Mining Association Rules Between Sets of Items in Large Databases." In *Proceedings of the 1993 ACM–SIGMOD '93*, Washington, D.C., May 1993, pp. 207–216. New York: ACM Press.

Bailey, D. H., and M. López de Prado. 2012. "The Sharpe Ratio Efficient Frontier." *Journal of Risk* 15 (2): 3–44.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

Carhart, M. M. 1997. "On Persistence in Mutual Fund Performance." *The Journal of Finance* 52 (1): 57–82.

Connor, G. 1995. "The Three Types of Factor Models: A Comparison of Their Explanatory Power." *Financial Analysts Journal* 51 (3): 42–46.

Cutler, A., D. R. Cutler, and J. R. Stevens. "Random Forests." In *Ensemble Machine Learning*, edited by C. Zhang and Y. Ma, pp. 157–175. Boston: Springer, 2012.

Fama, E. F., and K. R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–465.

——. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189–1232.

Ho, T. K. "Random Decision Forests." *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14–16 August 1995, pp. 278–282.

——. 1998. "The Random Subspace Method for Constructing Decision Forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8): 832–844.

Jolliffe, I. T. *Principal Component Analysis*. New York: Springer-Verlag, 2002.

Koenker, R. *Quantile Regression*. Cambridge: Cambridge University Press, 2005.

Lintner, J. 1965. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics* 47 (1): 13–37.

Lloyd, S. P. 1982. "Least Squares Quantization in PCM." *IEEE Transactions on Information Theory* 28 (2): 129–137.

Mossin, J. 1966. "Equilibrium in a Capital Asset Market." *Econometrica* 34 (4): 768–783.

Politis, D. N., and J. P. Romano. 1994. "The Stationary Bootstrap." *Journal of the American Statistical Association* 89 (428): 1303–1313.

Sharpe, W. F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk." *The Journal of Finance* 19 (3): 425–442.

——. 1988. “Determining a Fund’s Effective Asset Mix.” *Investment Management Review* 2 (6): 59–69.

——. 1992. “Asset Allocation: Management Style and Performance Measurement.” *The Journal of Portfolio Management* 18 (2): 7–19.

Simonian, J., O. Sosa, E. Heilbron, M. Senoski, and T. McFarren. 2018. “Capital-Market-Aware LDI: Actively Navigating the De-Risking Journey.” *The Journal of Portfolio Management* 44 (2): 130–135.

Treynor, J. L. “Market Value, Time, and Risk.” Unpublished manuscript (1961), No. 95–209.

Wang, K. Q., and J. Xu. 2015. “Market Volatility and Momentum.” *Journal of Empirical Finance* 30 (1): 79–91.

Disclaimer

This editorial reflects the current opinions of the authors, which are not those of Natixis Investment Managers.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

Big Data in Portfolio Allocation: *A New Approach to Successful Portfolio Optimization*

IRENE ALDRIDGE

IRENE ALDRIDGE
is managing director of research at AbleMarkets in New York, NY, and visiting professor at Cornell University in New York, NY.
irene@ablemarkets.com

According to DeMiguel, Garlappi, and Uppal (2009, p. 1915), the idea of diversifying one's financial portfolio dates back at least to the fourth century AD, when Rabbi Issac bar Aha documented a rule for asset allocation in the Babylonian Talmud (Tractate Baba Mezi'a, folio 42a): "One should always divide his wealth into three parts: a third in land, a third in merchandise, and a third ready to hand."

Modern portfolio theory originated from Markowitz (1952), and the body of work suggested not only diversifying assets and asset classes but also finessing portfolio composition by taking into account mutual co-movement of returns. Investments, the theory goes, should be diversified so that if or when one investment heads south, the others rise or at least counterbalance the total value of the portfolio. Co-movement of returns is often proxied by correlation matrixes. The optimal portfolio weights are computed to be directly proportional to the correlation matrix inverse.

When the number of positions is relatively small and stable, the classic Markowitz framework may work well. For larger portfolios, such as mutual funds and hedge funds with assets valued in the billions of US dollars, diversification suffers with unstable variance–covariance matrixes, costly reallocation requirements, and some illiquid positions. Exchange-traded funds further complicate the situation by providing a low-cost universe of potentially redundant securities that did not exist during Markowitz era, as described by Aldridge and Krawciw (2017). The correlation matrixes become very large. Big data techniques become necessary to intelligently reduce the size of the correlation matrixes, to select the key drivers in portfolios, and to remove redundant securities. Doing so helps portfolio managers improve transaction costs, stability of portfolio weights, and liquidity. With the advent of MiFID II and streamlined, potentially flat transaction fees per financial instrument, the smaller universe of financial instruments traded may be particularly beneficial to institutional investors.

Another benefit of reducing portfolio selection is the shortened history required for a robust performance estimation. As illustrated by DeMiguel, Garlappi, and Uppal (2009), increasing the number of instruments in the portfolio requires a significant increase in the length of historical data. Specifically, DeMiguel, Garlappi, and Uppal (2009) found that a portfolio of 25 assets with monthly reallocation requires a 250-year estimation window (across all positions) to reliably outperform the equally weighted (EW) strategy. This is a difficult requirement to fulfill considering reliable daily records have been kept for less than 70 years. DeMiguel, Garlappi, and Uppal (2009) also showed that the required estimating window scales linearly with the number of assets in the portfolio. Thus, a portfolio with five assets requires only 50 years of monthly data for reliable estimation.

Several techniques have been proposed over the years to mitigate the issues surrounding the Markowitz model. At the core of portfolio management is the following question: Which instruments should be removed and which ones kept? The decision is hardly trivial. Big data techniques do help to pinpoint the keepers in a reasonable time.

Traditional, not–big data solutions to the problem of optimal portfolio allocation fall roughly into two categories: Bayesian and non-Bayesian. Bayesian approaches include statistical, diffuse-priors, shrinkage estimators, and asset-pricing model priors. The diffuse-priors approach was pioneered by Barry (1974) and Bawa, Brown, and Klein (1979). The original shrinkage estimators date back to Jobson, Korkie, and Ratti (1979); Jobson and Korkie (1980); and Jorion (1985, 1986). The original asset-pricing models for establishing a prior were discussed by Pastor (2000) and Pastor and Stambaugh (2000) and, more recently, Brandt et al. (2005). They developed, for example, a simulation-based approach using recursion of approximations to the portfolio policy. Garlappi and Skoulakis (2008) simulated optimal portfolio choices using recursion of approximations to the portfolio value function.

Non-Bayesian non–big data approaches to minimizing estimation errors are similarly numerous. Goldfarb and Iyengar (2003) and Garlappi, Uppal, and Wang (2007) proposed robust portfolio optimization to deal with estimation errors using uncertainty structures and confidence intervals, respectively. MacKinlay and Pastor (2000) restricted the moments of returns by imposing factor dependencies. Best and Grauer (1992); Chan, Karceski, and Lakonishok (1999); and Ledoit and Wolf (2004a, 2004b) proposed methods for reducing the errors in the estimation of variance–covariance matrixes. Frost and Savarino (1988), Chopra and Ziemba (1993), and Jagannathan and Ma (2003) introduced short-selling constraints.

A separate stream of literature considers different portfolio optimization frameworks that depend on the concurrent market regime (i.e., bull versus bear market). For example, Ang and Bekaert (2002) used the Markov regime-switching model to show that regime-switching strategies that rely on macro factors as states outperform static portfolio allocation strategies out of sample.

Optimization problems from other disciplines with similarities to portfolio management and optimal asset allocation have been successfully studied in great detail in the field of big data analytics, and big data has been making inroads in portfolio management. Partovi and Caputo (2004) were the first to apply principal component analysis (PCA) to the portfolio choice problem to decompose principal portfolios uncorrelated by construction. Meucci (2009) followed up on the idea with the creation of maximum entropy portfolios. Garlappi and Skoulakis (2008) applied singular value decomposition (SVD) to solving several portfolio optimization problems in the context of the investor utility maximization. To do so, they deployed SVD to decompose state variables into fundamental drivers and shocks. The highest singular values or eigenvalues portray the drivers, whereas the lowest identify the shocks. Garlappi and Skoulakis (2008) applied the technique to solving the classic portfolio choice problem first proposed by Samuelson (1970) and extended by Hakansson (1971) and, later, Loistl (1976); Pulley (1981, 1983); Kroll, Levy, and Markowitz (1984); and Markowitz (1991), among others. A relatively recent stream of literature applies eigenvalue techniques to covariance matrixes to create eigenportfolios from any set of assets chosen by a researcher or a portfolio manager by some other evaluation criteria (see, for example, Steele (1995), Partovi and Caputo (2004), Avellaneda and Lee (2010), and Boyle (2014)).

The covariance and correlation matrixes, however, have been known to evolve, presenting a challenge to portfolio managers and researchers. Allez and Bouchaud (2012) studied eigenvalue evolution in covariance matrixes and attempted to find a time-based pattern of covariance evolution. They found that the covariance eigenvalues evolve over time, as expected. To deal with the estimation errors in the forward-looking correlation and covariance matrixes, Ledoit and Wolf (2017) proposed shrinking the sample covariance matrix toward a multiple of the identity matrix to push sample eigenvalues toward their mean. They proposed shrinking covariance matrixes by sampling eigenvalues in a nonlinear manner. Fan, Liao, and Mincheva (2013) developed a principal orthogonal complement thresholding method to estimate a high-dimensional covariance matrix with a conditional sparse structure and fast-diverging eigenvalues.

In this article I provide the first study of the big data properties of the inverse of the correlation matrix and show that the inverse is much more informative than the correlation matrix itself, from the big data perspective. Subsequently, the article proposes big data approaches to harness the correlation inverse and to deliver superior out-of-sample returns. The three key advantages of the method proposed in this article are conceptual simplicity, analytically tractable performance improvements, and empirically verified portfolio gains.

BIG DATA OVERVIEW

Many big data techniques, such as spectral decomposition, first appeared in the 18th century when researchers grappled with solutions to differential equations in the context of wave mechanics and vibration physics. Fourier has furthered the field of eigenvalue applications extensively with partial differential equations and other work.

At the heart of many big data models is the idea that the properties of every dataset can be uniquely summarized by a set of values, called *eigenvalues*. An eigenvalue is a total amount of variance in the dataset explained by the common factor. The bigger the eigenvalue, the higher the proportion of the dataset dynamics that eigenvalue captures.

Eigenvalues are obtained via either PCA or SVD. The latter technique is discussed in the following. The eigenvalues and related eigenvectors describe and optimize the composition of the dataset, perhaps best illustrated with an example of an image.

Consider the black-and-white image shown in Exhibit 1. It is a set of data points, *pixels* in computer lingo, whereby each data point describes the color of that point on a 0–255 scale, where 0 corresponds to pure black, 255 to pure white, and all other shades of gray lie in between. This particular image contains 960 rows and 720 columns.

To perform spectral decomposition on the image, I use SVD, a technique originally developed by Beltrami (1873).[1] PCA is a related technique that produces eigenvalues and eigenvectors identical to those produced by SVD when PCA eigenvalues are normalized. Raw, non-normalized, PCA eigenvalues can be negative or positive and do not equal the singular values produced by SVD. For the purposes of the analysis presented here, we assume that all the eigenvalues are normalized, equal to singular values, and we will use the terms *singular values* and *eigenvalues* interchangeably throughout this article because the results presented can be developed using SVD and PCA techniques.

EXHIBIT 1
Original Sample Image

Source: Courtesy Dr. Frank Fabozzi, 2018.

In SVD, a matrix X is decomposed into three matrixes: U, S, *and* V

$$X = USV' \tag{1}$$

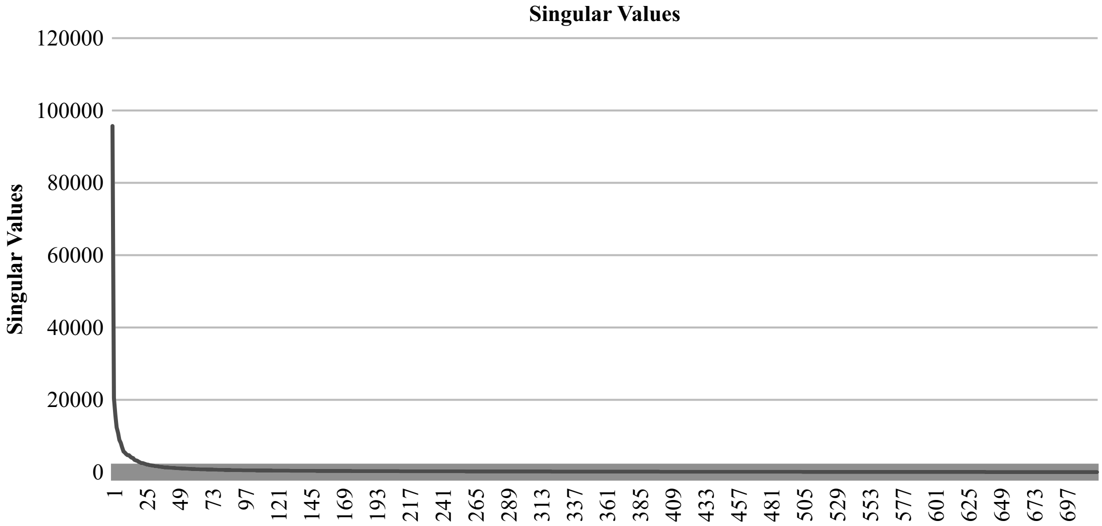
where X is the original $n \times m$ matrix; S is an $m \times m$ diagonal matrix of singular values or eigenvalues sorted from the highest to the lowest on the diagonal; V' is a transpose of the $m \times m$ matrix of so-called singular vectors, sorted according to the sorting of S; and U is an $n \times n$ user matrix containing characteristics of rows vis-a-vis singular values.

SVD delivers singular values sorted from largest to smallest. The plot of the singular values corresponding to the image in Exhibit 1 is shown in Exhibit 2. The plot of singular values is known as a *scree plot* because it resembles a real-life scree, a rocky mountain slope.

[1] For a detailed history of SVD, please see Stewart (1993).

EXHIBIT 2

Scree Plot Corresponding to the Image in Exhibit 1



A scree plot is a simple line segment plot that shows the fraction of total variance in the data as explained or represented by each singular value (eigenvalue). The singular values are ordered and are assigned a number label by decreasing order of contribution to total variance.

To reduce the dimensionality of a dataset, we select k singular values. If we were to use the most significant of the singular values, typically containing macroinformation common to the dataset, we would select the first k values. However, in applications involving idiosyncratic data details, we may be interested in the last k values (e.g., when we need to evaluate the noise in the system). A rule of thumb dictates breaking the eigenvalues into sets before the elbow and after the elbow sets in the scree plot.

What is the perfect number of singular values to keep in the image of Exhibit 1? An experiment presented in the seven panels in Exhibit 3 shows the evolution of the data with varying number of eigenvalues included. The eigenvalues and the corresponding eigenvectors composed of linear combinations of the original data create new dimensions of data. As the seven panels in the Exhibit 3 show, as few as 10 eigenvalues allow a human eye to identify the content of the image, effectively reducing dimensionality of the image from 720 columns to 10.

However, the guesswork is not at all needed because the optimal method of discarding the eigenvectors associated with the smallest eigenvalues has already been developed (see, for example, Carrasco, Florens, and Renault 2007). The method is known as the *spectral cutoff method*. Carrasco and Noumon (2011) further proposed a data-driven method to select the optimal number of principal components to be kept in the spectral cutoff method.

To create the reduced dataset, we restrict the number of columns in the S and V matrixes to k by selecting k first elements, determined by the spectral cutoff method. The resulting matrix $X_{reduced}$ has dimensions n rows and k columns, where

$$X_{reduced,nxk} = U_{nxk}\ S_{kxk}\ V_{kxk}^{T} \tag{2}$$

TRADITIONAL PORTFOLIO OPTIMIZATION AND BIG DATA APPLICATIONS

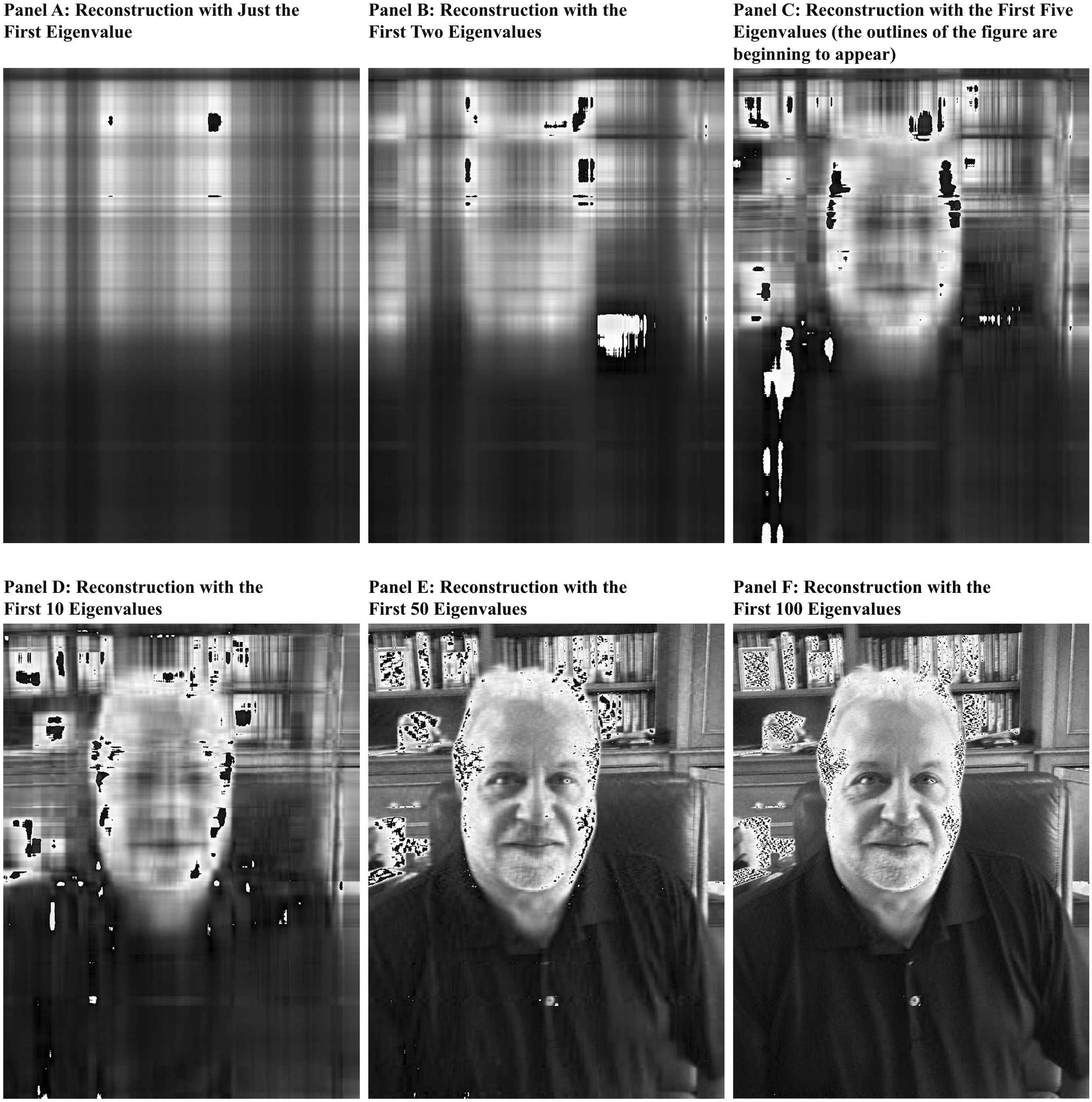
Markowitz-style portfolio optimization is often known as *mean–variance optimization* (MVO) because it seeks to increase mean returns while simultaneously decreasing variance in portfolios. Denoting the beginning prices of each asset i X_i, $i = 0, 1, \ldots, n$, we can express the investment portfolio as

$$w_0X_0 + w_1X_1 + \cdots + w_nX_n \tag{3}$$

where w_i, $i = 0, 1, \ldots, n$ are portfolio weights: the proportion of the total portfolio wealth that is invested in

Exhibit 3

Reconstruction of the Image of Exhibit 1



Panel A: Reconstruction with Just the First Eigenvalue

Panel B: Reconstruction with the First Two Eigenvalues

Panel C: Reconstruction with the First Five Eigenvalues (the outlines of the figure are beginning to appear)

Panel D: Reconstruction with the First 10 Eigenvalues

Panel E: Reconstruction with the First 50 Eigenvalues

Panel F: Reconstruction with the First 100 Eigenvalues

the asset *i*. The sum of the weights of the portfolio assets is then equal to 1, and $w_0 + w_1 + \cdots + w_n = 1$. The asset with $i = 0$ is often assumed to be the prevailing risk-free rate, denoted r_0.

Denoting risk aversion as γ, we now express the traditional MVO as follows:

$$max_{w,w_0}(w_0 r_0 + w'\mu - \gamma w'\Sigma w) \qquad s.t.\ w_0 + w'1 = 1 \quad (4)$$

where Σ represents the variance–covariance matrix of the returns of the n assets under consideration.

Subtracting the risk-free rate, the maximization problem can be rewritten as follows:

$$max_{w,w_0}(w'(\mu - r_0 1) - \gamma w' \Sigma w) \qquad s.t.\ w_0 + w'1 = 1 \quad (5)$$

Equation 2 then leads to the following optimal solution:

$$w = \frac{1}{2\gamma} \sum^{-1} (\mu - r_0 1) \quad (6)$$

Although the vector of returns is typically assumed to be the long-running average of returns on assets under consideration (see, for example, Jegadeesh and Titman (1993)), the covariance matrix presents several challenges to researchers and practitioners. Specifically, the covariance matrix can in turn be decomposed into variance and correlation matrixes, although variances tend to be sticky and reasonably predictable by techniques such as generalized autoregressive conditional heteroskedasticity[2] and correlations of asset returns are notoriously volatile.[3] It is the properties of correlation matrixes that induce two key problems portfolio managers encounter when implementing MVO:

1. Possibly extreme positions in selected assets (i.e., a large proportion of the portfolio) resulting in liquidity constraints and violating the economic equilibrium of the portfolio allocation. To solve the issue, Black and Litterman (1993) and others proposed a blended solution between economic equilibrium and MVO.
2. Possibly extreme changes in portfolio weights from one investment period to the next, resulting in large transaction costs. Bertsimas and Lo (1998), Liu (2004), Muthuraman and Kumar (2006), Lynch and Tan (2008), and Mei, DeMiguel, and Nogales (2016), for example, propose penalizing the MVO function with transaction costs as the remedy to the problem. However, such methods often tend to be opaque in practice.

Big data techniques, such as spectral decomposition, have appealed to researchers for their data size reduction and stabilization properties but have produced variable results to date. Several techniques have been developed and popularized over the years, all deploying big data on the correlation matrix or, worse, on the covariance matrix itself, instead of tackling the root of the portfolio management woes: the correlation matrix inverse.

Reduction of the covariance matrix can be considered erroneous for the following reasons: The volatility properties have been well studied and can be successfully modeled independently of the correlation framework. As a result, including variances in the optimization bag together with the correlations prevents the researchers from finessing the optimization with the independent volatility properties.

The prevalent techniques for the stand-alone correlation optimization suffer from an even bigger flaw. The classic foundation technique, known as PCA, is at the heart of most current optimization frameworks for the correlation matrix. The technique decomposes the correlation matrix into its eigenvalue-related principal components and then shrinks the correlation matrix by setting the eigenvalue tail to zeros. The technique follows the principles of big data optimization discussed in the previous section.

Two immediate issues arise. First, the largest eigenvalue of the correlation matrix has long been known to be a market portfolio, whereas the eigenvalue tail corresponds to the idiosyncratic properties of the assets under consideration. Retaining the dominant market portfolio while discarding the idiosyncratic pieces goes completely against the spirit of the classical Markowitz optimization, which seeks instead to diversify away from the market. Second, setting eigenvalues to zero prior to matrix inversion renders matrixes singular and, therefore, noninvertible. In other words, reducing the spectral dimensionality of the correlation matrixes and subsequent inversion blow up the outcome. To overcome the issue, researchers often use *whitening*—replacing set-to-zero eigenvalues with white noise $N(0, 1)$ to allow matrix invertibility. The process introduces noise into the system, resulting in classic "garbage in, garbage out" situations well known in engineering disciplines.

Most models, such as shrinkage operators and Bayesian optimization frameworks, use the described faulty PCA as their underlying core, producing suboptimal

[2] See Engle (1982), Bollerslev (1986), and Andersen et al. (2006).

[3] See Davis and Mikosch (1998), Gourieroux (1997), and Cont (2001).

results. The same argument applies to recently popular eigenportfolios and other techniques that apply spectral decomposition or PCA to correlation or covariance matrixes, instead of correlation matrix inverses.

BIG DATA WITH THE INVERSE OF THE CORRELATION MATRIX: A NOVEL APPROACH

In contrast to the established techniques tackling the correlation matrix, big data application to the inverse of the correlation matrix appears to be more promising and robust. The eigenvectors of an invertible matrix are also the eigenvectors of the matrix's inverse. To show this, consider an invertible matrix A. Matrix A is invertible if and only if its determinant is not zero (Lipschutz 1991, p. 45), which in turn implies that matrix A columns are linearly independent, further implying that its eigenvalue λ is not zero. Suppose that matrix A has eigenvectors $\boldsymbol{v}$. By definition of eigenvectors, $A\boldsymbol{v} = \lambda\boldsymbol{v}$. Multiplying by A^{-1} from the left, we obtain

$$v = A^{-1}\lambda\boldsymbol{v} \tag{7}$$

$$A^{-1}v = (1/\lambda)\boldsymbol{v} \tag{8}$$

Another solution is to exploit the fact that singular values of a matrix may be found and the dimensions reduced after the inversion with equal success and without sacrificing data precision.

$$(AB)^{-1} = B^{-1}A^{-1} \tag{9}$$

More generally,

$$\left(\prod_{k=0}^{N} A_k\right)^{-1} = \prod_{k=0}^{N} A_{N-k}^{-1} \tag{10}$$

So, for SVD

$$A(p) = U(p)S(p)V'(p) \tag{11}$$

the inverse becomes

$$(A(p))^{-1} = (V')^{-1}S^{-1}U^{-1} \tag{12}$$

SVD of the inverse of the correlation matrix is, therefore, much more precise because no data are lost as a result of the poorly specified input to the inversion process that occurs with whitening methodology. Accordingly, the SVD in the case of the matrix inversion can be performed as follows: The spectral decomposition can be performed after the matrix inversion without sacrificing results.

If SVD decomposes a correlation matrix C into $C = USV^T$, then the inverse of the matrix C can be written as $C^{-1} = (V^T)^{-1}S^{-1}U^{-1}$, where S^{-1} is the inverse of the diagonal matrix S

$$S = \begin{matrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & \cdots & \lambda_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & \lambda_n \end{matrix}$$

$$S^{-1} = \begin{matrix} 1/\lambda_n & 0 & 0 & \cdots & 0 \\ 0 & 1/\lambda_{n-1} & 0 & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & 1/\lambda_2 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1/\lambda_1 \end{matrix}$$

Inverting the correlation matrix first and then spectrally decomposing it to retrieve eigenvalues $\{\lambda\}$ therefore allows researchers to retain much more precision. Instead of replacing the irrelevant eigenvalues with noise to allow inversion, the proposed process is to replace the eigenvalues directly with 0 postinversion.

Which eigenvalues should you keep or discard? This, once again, is a nontrivial question. Spectral decomposition of the original, noninverted correlation matrix results in principal components or portfolios sorted according to their universality vis-a-vis all assets considered. Thus, the largest component often represents the global macro portfolio factor driving most of the performance and typically reflecting the broad market movement. Several of the following eigenvalues deliver portfolios that induce synchronized fluctuations of groups of stocks; these can be, for example, factors driving industries. The remaining small components are idiosyncratic in nature. Spectral decomposition of the inverted correlation matrix produces eigenvalues sorted in the opposite order: from smallest to the largest.

Numerous big data techniques have been developed to help us understand the information content of the matrix under consideration—in our case, the inverse of the correlation matrix. Here, we develop and prove a conjecture that the top eigenvalue information content of the inverse of the correlation matrix always exceeds that of the correlation matrix itself. As a result, the big data analysis pertaining to the optimal portfolio allocation should be carried out on the correlation matrix inverse, not on the correlation matrix as is done at present. The invert-then-optimize methodology proposed in this article, and diametrically opposite to established methodologies, not only delivers superior results but also delivers explicit tractable solutions to the most-cited woes of existing portfolio optimization methodologies: correlation instability and extreme portfolio weights.

In short, the proposed methodology is to retain the largest eigenvalues in the inverse of the correlation matrix. These eigenvalues correspond to the smallest eigenvalues of the original correlation matrix, the values discarded in traditional analyses. We show that these values, long known to contain idiosyncratic properties of assets, are indeed key to successful portfolio optimization.

CORRELATION MATRIXES VERSUS INVERSES: STABILITY AND SENSITIVITY TO PERTURBATIONS

Given that the main problems associated with large-scale portfolio optimization revolve around the instability of the resulting portfolio weights, the objective of the decomposition should be to preserve the most stable components and to discard the least stable ones. Much of the traditional literature interprets this as retaining the top eigenvalues of the correlation matrix and discarding the smallest values. However, this does not make sense given that the final portfolio weights are proportional to the inverse of the correlation matrix instead. As this section shows, the inverse of the correlation matrix is necessarily less stable than the correlation matrix itself; to stabilize portfolios, one needs to stabilize the inverse of the correlation matrix, not the correlation matrix itself.

A vast stream of literature focusing on the stability of matrixes and their sensitivity to perturbations dates back to Gershgorin (1931). Gershgorin circles allow us to identify the span of possible values for eigenvalues in our system. The Gershgorin circles define the radii around each a_{ii} in a matrix A, within which lies eigenvalue i

$$\left|\lambda_i - a_{ii}\right| = \sum_{i \neq j} \left|a_{ij}\right| \tag{13}$$

The tighter the Gershgorin circle around i, the more stable the eigenvalue i to small perturbations in the matrix under consideration. Correspondingly, the larger the Gershgorin circle around i, the less stable the ith eigenvalue and the more sensitive the matrix is to even the smallest changes in the underlying data.

Gershgorin circles form a convenient visual representation of the sensitivity of data to small perturbations. As an example, consider just five equities (A, AA, AAL, AAMC, and AAN) over a three-week period ending October 27, 2017, with the summary statistics shown in Exhibit 4. Exhibit 5 shows the normalized eigenvalues of the correlation matrix, the respective Gershgorin radii of the correlation matrix, the eigenvalues of the inverse of the correlation matrix, and the Gershgorin radii of the inverse of the correlation matrix.

The two panels in Exhibit 6 represent the resulting Gershgorin circles visually. As this exhibit shows, the Gershgorin circles of the inverse are much larger, indicating that the inverse of the matrix is much more unstable than the sample correlation matrix itself.

Similar empirical results can be obtained with the Bauer–Fike theorem (Bauer and Fike, 1960) and other methods, such as the Robinson and Wathen (1992) method. The Bauer–Fike theorem proposes comparing operator vector norms of eigenvectors. The vector norms serve as upper bounds for perturbations for respective eigenvectors. Exhibit 7 shows the upper bounds for matrix perturbations for vanilla correlation and correlation inverse matrixes for data in Exhibit 1. As shown in Exhibit 7, the bounds on the inverse of the correlation matrix are considerably higher than that on the correlation matrix itself, implying once again that the inverse of the correlation matrix is much less stable than the correlation matrix.

Similar results can be obtained using key relation between matrixes, ordered eigenvalues {□}, and matrix inverses derived by Robinson and Wathen (1992)

Exhibit 4

An Illustration of Gershgorin Circles on Sample Correlation Matrixes

	Mean	St Dev	Corr A	AA	AAL	AAMC	AAN
A	0.001384	0.008118	1.000	0.391	–0.024	0.315	–0.150
AA	0.003440	0.019856	0.391	1.000	0.344	0.365	–0.194
AAL	–0.00064	0.017241	–0.024	0.344	1.000	0.123	–0.125
AAMC	–0.00400	0.023821	0.315	0.365	0.123	1.000	0.693
AAN	–0.00390	0.014468	–0.150	–0.194	–0.125	0.693	1.000

$$\frac{1}{\lambda_1}+\frac{(\lambda_1-1)^2}{\lambda_1(\lambda_1-s_{ii})}\le (A^{-1})_{ii}\le \frac{1}{\lambda_n}-\frac{(\lambda_n-1)^2}{\lambda_n(-\lambda_n+s_{ii})},$$

where $s_{ii}=\sum_k a_{ik}^2$.

A formal theoretical conclusion showing the higher instability of the correlation inverse is as follows: The largest eigenvalue of the inverse of the correlation matrix is always larger than the largest eigenvalue of the correlation matrix itself. The proof of this theoretical conclusion is provided in the online supplement.

The obtained results are independent of the underlying distribution of returns. Indeed, the result accommodate Gaussian, leptokurtic, and other distributions with equal effect, making the strategy robust to a variety of financial return models. Furthermore, the result of the theoretical conclusion extends far beyond financial data and is applicable to any datasets, whether advertising, healthcare, or genomics.

SENSITIVITY OF CORRELATION MATRIXES VERSUS THEIR INVERSES: SIMULATION

To ascertain the validity of our conjecture, we perform 10,000 experiments of the following nature:

1. We create a random symmetric 100 × 100 matrix $\{A_{ij}\}$ simulating the real-life correlation structure: All the values on the diagonal are set to 1.0, and all other values for $i \neq j$ range in the interval [−1.0, 1.0], with entries $a_{ij} = a_{ji} = \forall i, j$.
2. We compute and document the eigenvalues of the correlation matrix and its inverse.

As the results presented in Exhibit 8 illustrate, the top eigenvalue of the inverse is considerably higher than the top eigenvalue of the correlation matrix itself. As the simulation results show, it is the inverse of the correlation matrix that is the unstable component of the portfolio optimization puzzle. Because the portfolio weights are directly proportional to the inverse of the correlation

Exhibit 5

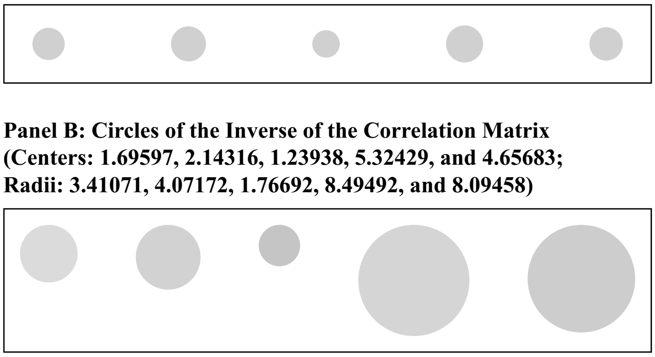
Comparative Dispersion of Eigenvalues via Gershgorin Radii for the Correlation Matrix of Exhibit 4 and the Inverse

Normalized Eigenvalues of the Correlation Matrix	Gershgorin Radii of the Correlation Matrix	Normalized Eigenvalues of the Inverse of the Correlation Matrix	Gershgorin Radii of the Inverse of the Correlation Matrix
0.09290	0.880	0.54085	3.41071
0.46752	1.294	0.63756	4.07172
1.02214	0.616	0.97834	1.76692
1.56849	1.496	2.13897	8.49492
1.84896	1.162	10.7639	8.09458

Exhibit 6

Gershgorin Circles of the Sample Correlation Matrix of Exhibit 4 and the Inverse of the Matrix, Graphical Representation

Panel A: Circles of the Correlation Matrix All Centered on 1; Sizes Are Comparable and Close to 1



Panel B: Circles of the Inverse of the Correlation Matrix (Centers: 1.69597, 2.14316, 1.23938, 5.32429, and 4.65683; Radii: 3.41071, 4.07172, 1.76692, 8.49492, and 8.09458)

Exhibit 7

Bauer–Fike Norms for Eigenvectors of Correlations and Inverse Correlations of Data in Exhibit 4

	||V|| Corr	||V–1|| Corr	||V|| Corr Inverse	||V–1|| Corr Inverse
1	2.10	2.11	2.10	2.11
2	**1.31**	**1.24**	**1.45**	**1.60**
∞	2.11	2.10	2.11	2.10

Note: Bolded value show much higher inverse dispersion.

Exhibit 8

Summary Statistics for Eigenvalues of 10,000 Simulated Correlation Matrixes and Their Inverses

	Top 1	Bottom 1	Top 1 Inverse	Bottom 1 Inverse
Mean	11.60537822	0.05523877	**312.46715013**	0.08622607
St Dev	0.30493585	0.04221966	**8,508.6027317**	0.00225207
Skew	0.30964001	1.00654776	**50.50362873**	–0.14982966
Kurt	0.17120805	0.88971901	**2,776.6559658**	0.02559518
Max	12.83546300	0.25704200	500,000.00000	0.09435499
99%	12.39773146	0.18056210	1,545.2583525	0.09108325
95%	12.12671910	0.13857045	262.28504037	0.08983642
90%	12.00535900	0.11442080	121.92152061	0.08909438
75%	11.80207150	0.08023350	46.19737847	0.08777734
50%	11.59161850	0.04672950	21.39976314	0.08626923
25%	11.39246225	0.02164625	12.46362244	0.08473089
10%	11.22405270	0.00820200	8.73966974	0.08329613
5%	11.13134355	0.00381265	7.21654767	0.08246254
1%	10.97896710	0.00064715	5.53826083	0.08065992
Min	10.59827400	0.00000200	3.89041480	0.07790915

Note: Bolded value show much higher inverse dispersion.

matrix, stabilization and other optimization of the inverse of the correlation matrix—not the correlation matrix itself—are critical for successful portfolio allocation.

OUT-OF-SAMPLE APPLICATIONS TO FINANCIAL DATA

I next test the theory (the importance of the optimization of the correlation matrix inverse) on the historical financial data. I performed two experiments:

1. Comparison of the core portfolio management techniques on the S&P 500 data for the 20-year period from 1998 through 2017, with monthly reallocation
2. Comparison of the portfolio management techniques on 1,000 portfolios with 50 or more stocks each, the constituents of which were randomly drawn from the S&P 500 from 1998 through 2017, with monthly reallocation

Both experiments show that regardless of portfolio composition, the correlation inverse optimization proposed in this article significantly outperforms the other core portfolio allocation strategies.

Out-of-Sample Application to the S&P 500

The test uses daily closing price data for the S&P 500 constituents for the 20-year period spanning 1998–2017 and obtained from Yahoo!. We assume monthly portfolio reallocation and test the following strategies on the S&P 500 data: EW, vanilla MVO, PCA with the top eigenvalues retained, and PCA_Inverse with the bottom eigenvalues of the inverse taken into account and the bottom eigenvalues discarded.

To compute strategy performance, the lognormal daily returns from the price data are first determined

$$r_t = log(P_t) - log(P_{t-1}) \tag{14}$$

Next the monthly correlation matrixes using the returns falling into each calendar month in the 1998–2017 span are computed. Each correlation matrix then serves as an input to the strategy evaluation over the following month. For example, the correlation matrix computed on January 30, 1998 serves as the input for portfolio selection for February 1998.

Monthly performance of the strategies is next measured using the strategy weights computed on the last day of the previous month using the daily returns for the previous month. For analytical tractability, the risk aversion coefficient is chosen to be 1; however, it can be easily scaled up or down because the portfolio weights of the MVO, PCA of MVO, and PCA_Inverse of MVO strategies are directly proportional to the risk aversion coefficient. The performance evaluation applies the weights to the returns observed on the last trading day of the following month vis-a-vis the price levels observed on the last day of the portfolio creation month. Thus, the performance of portfolios created on January 30, 1998 is tested by returns observed from the closing price on January 30, 1998 to the closing price observed on February 27, 1998.

The four panels in Exhibit 9 document the performance of the monthly reallocation of the strategies. As the exhibit shows, the PCA_Inverse strategy outperforms the other strategies when the number of selected eigenvalues is small, such as the top one eigenvalue selected in the PCA_Inverse strategy shown in Panel A (with outliers) and Panel B (outliers removed for clarity) of Exhibit 9. As the number of retained eigenvalues increases, the PCA_Inverse strategy loses its power and eventually yields to the EW strategy.

Exhibit 10 shows the Sharpe ratios from the obtained strategies. As the exhibit shows, the PCA_Inverse strategy consistently outperforms other portfolio management strategies, particularly when the outliers, such as extreme one-time returns, are discarded from the data. Exhibit 11 presents average monthly returns for each strategy computed over the 1998–2017 period. As shown in the exhibit once again, the PCA_Inverse strategy delivers superior results when a concentrated number of eigenvalues is deployed to create an optimal portfolio allocation.

The results of the analysis so far show that just the top eigenvalue of the inverse of the correlation matrix contains enough portfolio information to outperform the other strategies. Just how many instruments does such a strategy contain? Exhibits 12 and 13 help answer this question. The number of positions with the absolute value greater than or equal to 2% of the total portfolio value varied throughout the 20-year period; the number of stocks was significantly smaller than that of other strategies, pointing to a smart diversification portfolio selection of the PCA_Inverse strategy.

Bootstrapping the S&P 500: Technique Comparison on Randomly Selected Subportfolios over 1998–2017 Period

To anticipate the objections of researchers and portfolio managers dealing with assets other than the prim and proper S&P 500 and to showcase the strength and capability of the correlation inverse optimization proposed in this article, the following tests were conducted:

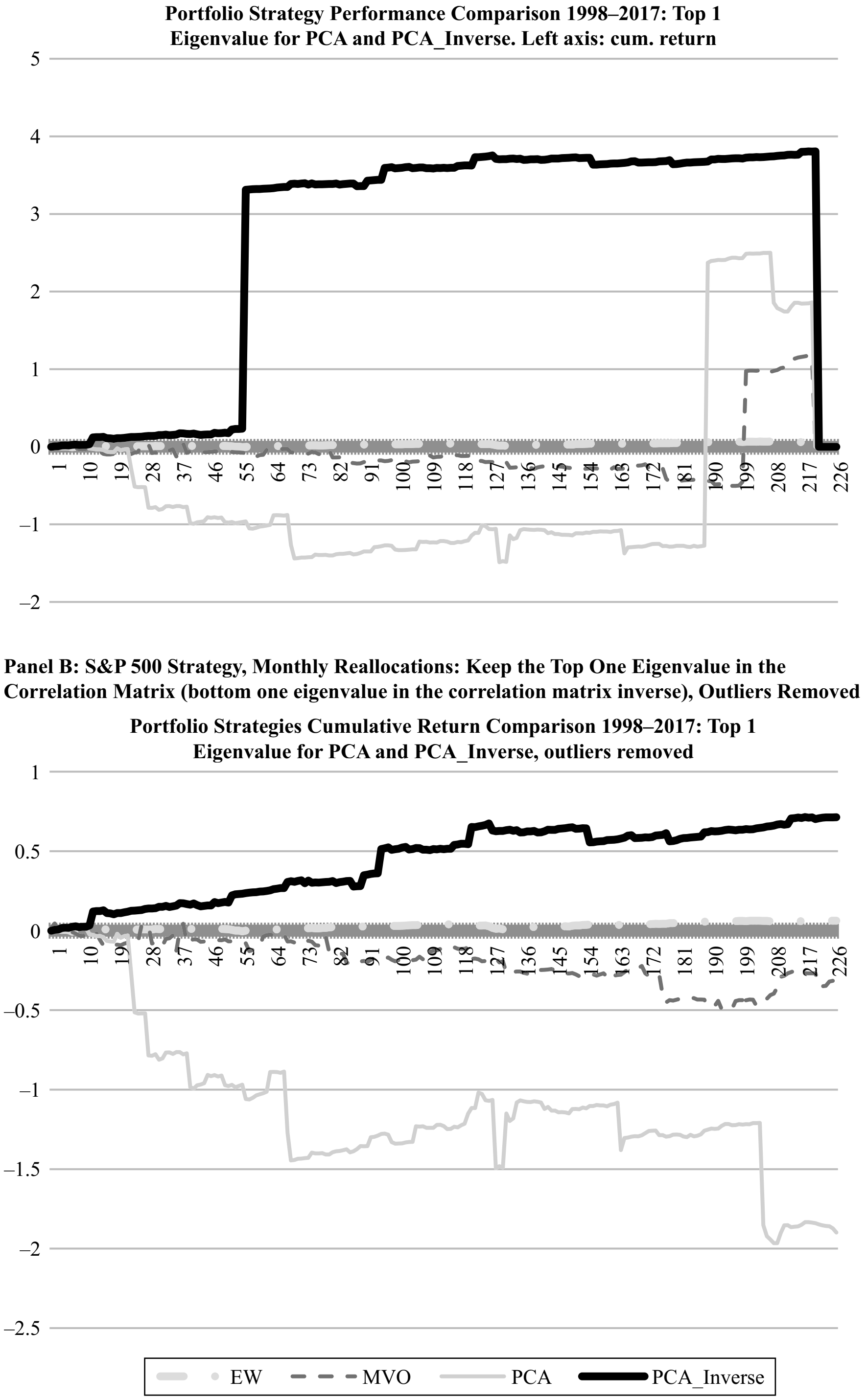
1. On January 1, 1998, we randomly select 50 or more names from the S&P 500. There are no restrictions on the name selections or their quantity, other than the randomly chosen portfolio must include at least 50 names. As noted earlier in this article, portfolios of fewer than 50 names are considered suitable for vanilla MVO and may not be as interesting for our purposes.
2. The four core portfolio management strategies with monthly reallocation on the portfolio randomly chosen in Step 1 were then run: (a) EW; (b) MVO; (c) spectral decomposition and optimization via PCA of the asset correlation matrix (PCA), retaining the top eigenvector only; and (d) the methodology proposed in this article, spectral decomposition and optimization of the inverse of the asset correlation matrix (PCA_Inverse), again, retaining only the top eigenvector, this time of the inverse.

The portfolio compositions do not change from 1998 through 2017. The portfolio weights are computed on the last trading day of each month. The EW weights do not change, unless the originally chosen stock is no longer trading. For MVO, PCA, and PCA_Inverse, the correlation matrixes used to set portfolio weights for the following month are computed on the last day of each trading month using daily log returns based on closing prices for the past month. Thus, the correlation matrix used to compute the weights for March 2005 is determined on the last trading day of February 2005 using all the closing daily returns for February 2005, including the first and the last trading days.

The traditional PCA approach to the correlation matrix is analogous to the eigenportfolio selection. As our analysis shows, the methodology on the correlation inverse PCA (PCA_Inverse) proposed in this article is far superior to the plain eigenportfolio construction. Panel A in Exhibit 14 shows the cumulative returns of the four core strategies averaged by month across 30 random draws of 50 or more securities comprising the S&P 500. Panel B shows standard deviations of the 30 independent repetitions by month from 1998 through 2017, illuminating outliers. As the two panels of Exhibit 14 show, even with severe outliers, the proposed methodology significantly outperforms other methods, regardless of portfolio construction. Panel C of Exhibit 14 shows the cumulative returns of PCA_Inverse over the 20-year period from 1998 through 2017.

Exhibit 9

Portfolio Strategy Performance Comparison, S&P 500, 1998–2017



Panel A: S&P 500 Strategy, Monthly Reallocations: Keep the Top One Eigenvalue in the Correlation Matrix (bottom one eigenvalue in the correlation matrix inverse)

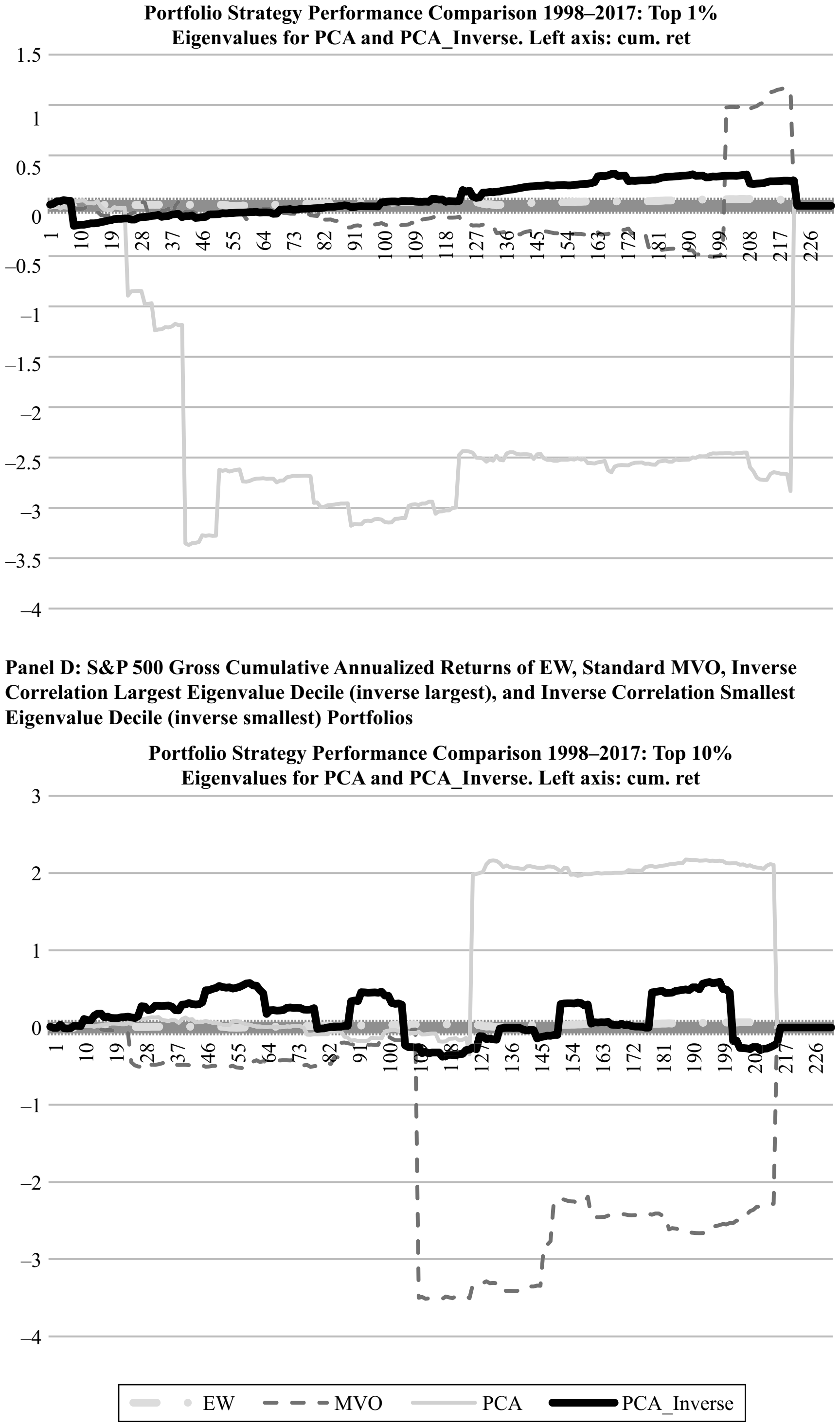
Panel B: S&P 500 Strategy, Monthly Reallocations: Keep the Top One Eigenvalue in the Correlation Matrix (bottom one eigenvalue in the correlation matrix inverse), Outliers Removed

(continued)

Exhibit 9 *(continued)*

Portfolio Strategy Performance Comparison, S&P 500, 1998–2017

Panel C: S&P 500 Strategy, Monthly Reallocations: Keep the Top 1% of Eigenvalues in the Correlation Matrix (bottom 1% of eigenvalues in the correlation matrix inverse)



Panel D: S&P 500 Gross Cumulative Annualized Returns of EW, Standard MVO, Inverse Correlation Largest Eigenvalue Decile (inverse largest), and Inverse Correlation Smallest Eigenvalue Decile (inverse smallest) Portfolios

Notes: Gross cumulative annualized returns of EW, standard MVO, inverse correlation largest eigenvalue deciles (inverse largest), and inverse correlation smallest eigenvalue decile (inverse smallest) portfolios.

EXHIBIT 10

Sharpe Ratios on Strategy Performance, S&P 500, 1998–2017, Monthly Reallocation

	EW	MVO	PCA	PCA_Inverse
10% Eigenvalues	0.4398529652	0.1660338977	0.2175831572	–0.05572290167
1% Eigenvalues	0.4398529652	0.1660338977	–0.2620669154	0.1854018356
1 Eigenvalue, with outliers	0.4398529652	0.1660338977	0.1121639833	0.2838331832
1 Eigenvalue, outliers removed	0.4398529652	–0.1695154284	–0.3799479568	0.6117174285

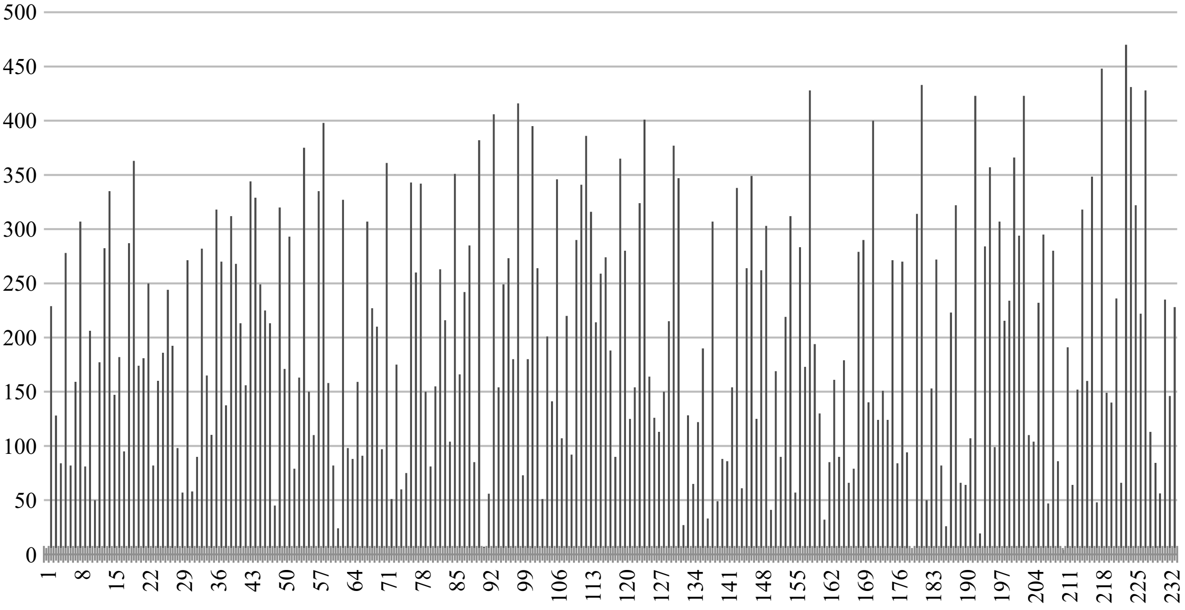
EXHIBIT 11

Average Monthly Returns per Strategy, S&P 500, 1998–2017, Monthly Reallocation

	EW	MVO	PCA	PCA_Inverse
10% Eigenvalues	0.0002994329004	0.004832313043	0.01287782073	–0.001001351801
1% Eigenvalues	0.0002994329004	0.004832313043	–0.01278558696	0.001204575758
1 Eigenvalue, with outliers	0.0002994329004	0.004832313043	0.007869265217	0.01649431169
1 Eigenvalue, outliers removed	0.0002994329004	–0.001353022026	–0.008363651982	0.003141938326

EXHIBIT 12

Number of Securities Selected Each Month from the S&P 500 by PCA_Inverse Method, 1998–2017



Notes: Using only the top eigenvalue of the inverse. It is common for the algorithm to deliver a single-digit number of names under this portfolio construction.

CONCLUSIONS

In this article, I demonstrate that the inverse of the correlation matrix is inherently more sensitive to perturbations than the correlation matrix itself, affecting the Markowitz portfolio allocation strategies. To harness the power of big data analytics to capitalize on this information content, I propose a big data refinement to portfolio selection: applying spectral decomposition to the inverse of the correlation matrix, instead of to the correlation matrix. The proposed methodology is tested on the S&P 500 Index and random subportfolios of the

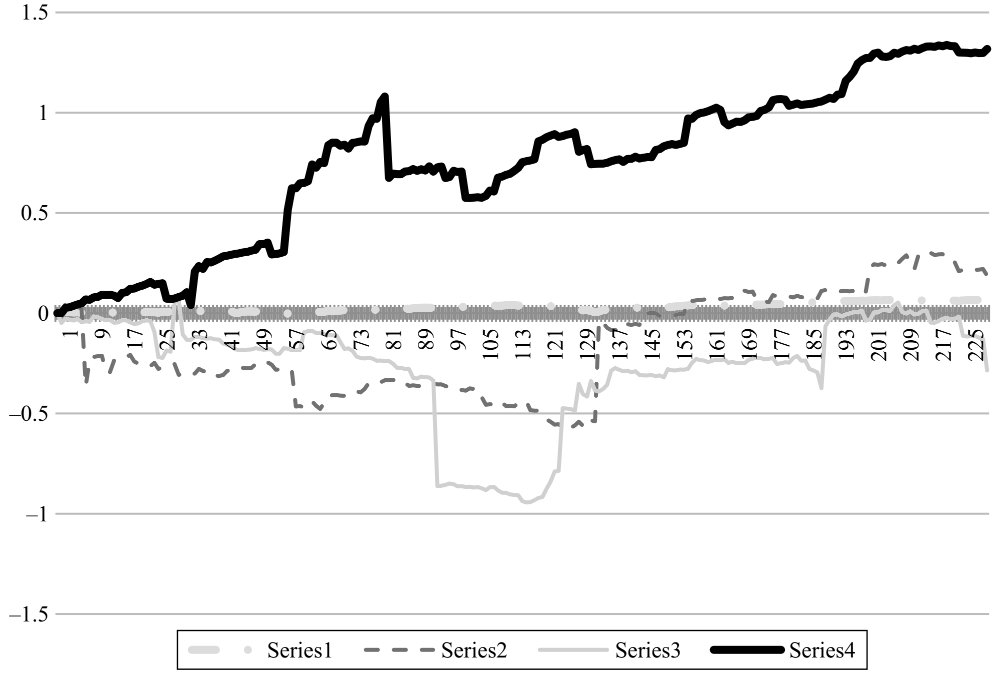
EXHIBIT 13
Mean and Standard Deviation (in parentheses) for the Number of Equities from the S&P 500 with Absolute Values of Weights Exceeding 1% or 2% of the Entire Portfolio Selected Monthly by Vanilla MVO, PCA, and PCA_Inverse Methods for Different Eigenvalue Cutoffs

		MVO	PCA	PCA_Inverse
Top 1 Eigenvalue	|weight| > 1%	375.49 (82.16)	375.15 (82.94)	192.44 (117.07)
	|weight| > 2%	343.57 (88.33)	343.40 (89.77)	112.64 (116.94)
Top 1% of Eigenvalues	|weight| > 1%	375.49 (82.16)	376.49 (84.30)	190.90 (115.95)
	|weight| > 2%	343.57 (88.33)	346.27 (91.65)	109.98 (115.37)
Top 10% of Eigenvalues	|weight| > 1%	375.69 (84.82)	377.83 (85.41)	372.13 (82.73)
	|weight| > 2%	341.54 (93.03)	343.97 (93.17)	335.50 (89.24)

Note: Data: 1998–2017, monthly portfolio rebalancing.

EXHIBIT 14
Performance of Portfolio Randomly Selected from the S&P 500 Constituents, 1998–2017

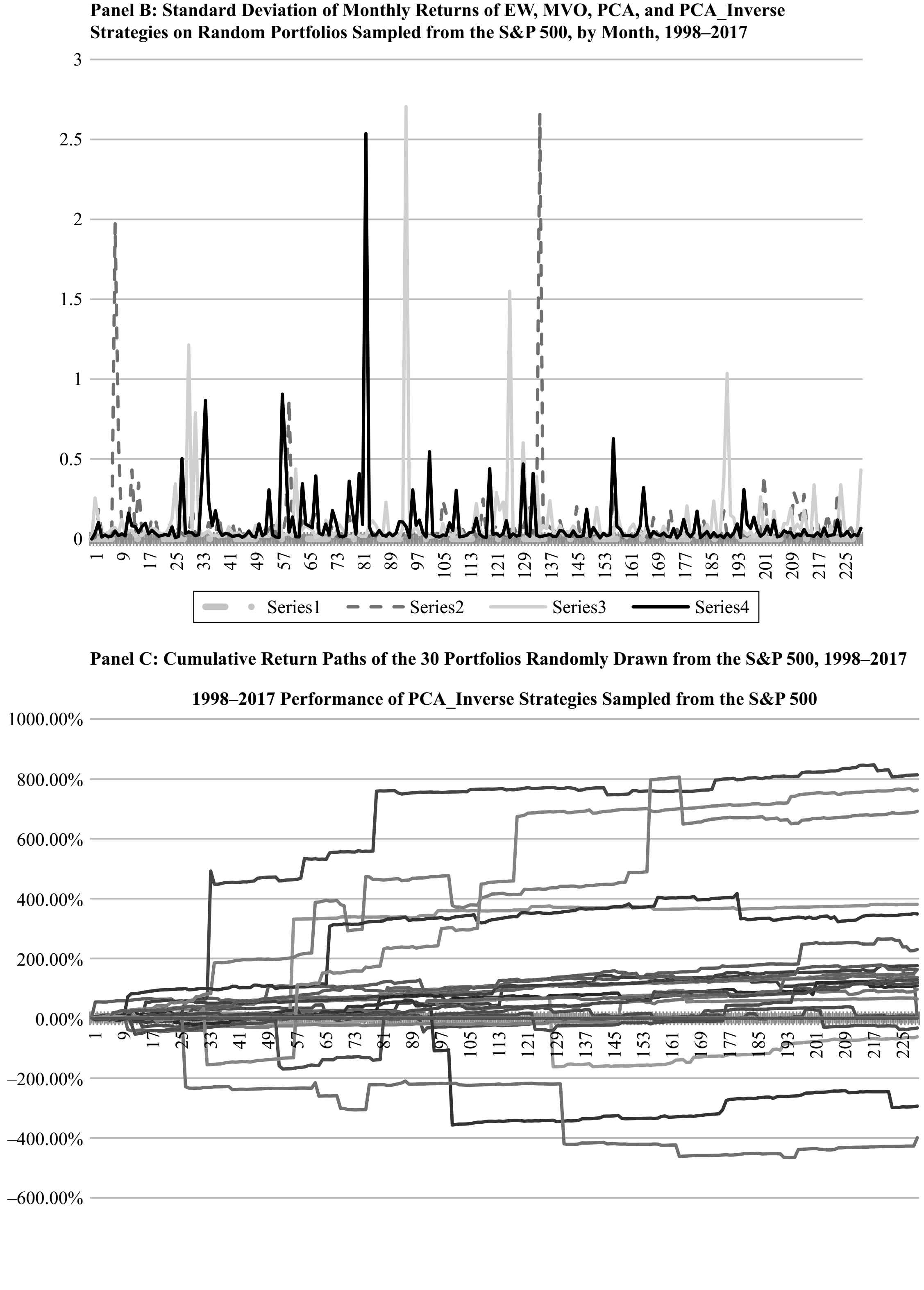
Panel A: Cumulative Returns of EW, MVO, PCA, and PCA_Inverse Strategies on Random Portfolios Sampled from the S&P 500, Averaged by Month, 1998–2017



(continued)

EXHIBIT 14 *(continued)*

Performance of Portfolio Randomly Selected from the S&P 500 Constituents, 1998–2017



Panel B: Standard Deviation of Monthly Returns of EW, MVO, PCA, and PCA_Inverse Strategies on Random Portfolios Sampled from the S&P 500, by Month, 1998–2017

Panel C: Cumulative Return Paths of the 30 Portfolios Randomly Drawn from the S&P 500, 1998–2017

1998–2017 Performance of PCA_Inverse Strategies Sampled from the S&P 500

S&P 500 from 1998 through 2017. Out of sample, the methodology consistently outperforms other common methods, such as EW portfolio allocation, plain MVO, and previously suggested big data portfolio optimization methodologies.

ACKNOWLEDGMENTS

I am grateful to Marco Avellaneda, Frank Fabozzi, Sasha Stoikov, and Victoria Averbukh; participants of the 2018 Sixth Annual Big Data Finance Conference; and my students at Cornell University for suggestions and advice.

REFERENCES

Aldridge, I., and S. Krawciw. *Real-Time Risk: What Investors Should Know about Fintech, High-Frequency Trading and Flash Crashes*. Hoboken, NJ: Wiley, 2017.

Allez, R., and J. P. Bouchaud. 2012. “Eigenvector Dynamics: General Theory and Some Applications.” *Physical Review E* 86: 46202.

Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold. “Volatility and Correlation Forecasting.” In *Handbook of Economic Forecasting*, edited by G. Elliot, C.W.J. Granger, and A. Timmermann, pp. 778–878. Amsterdam: North-Holland; 2006.

Ang, A., and G. Bekaert. 2002. “International Asset Allocation with Regime Shifts.” *Review of Financial Studies* 15 (4): 1137–1187.

Avellaneda, M., and J. H. Lee. 2010. “Statistical Arbitrage in the US Equities Market.” *Quantitative Finance* 10 (7): 761–782.

Barry C. 1974. “Portfolio Analysis under Uncertain Means, Variances, and Covariances.” *The Journal of Finance* 29 (2): 515–522.

Bauer, F. L., and C. T. Fike. 1960. “Norms and Exclusion Theorems.” *Numerische Matematik* 2: 137–141.

Bawa, V. S., S. J. Brown, and R. W. Klein. *Estimation Risk and Optimal Portfolio Choice*. Amsterdam: North-Holland; 1979.

Beltrami, E. 1873. “Sulle funzioni bilineari.” *Giornale di Matematiche ad Uso degli Studenti Delle Universita* 11: 98–106.

Bertsimas, D., and A. W. Lo. 1998. “Optimal Control of Execution Costs.” *Journal of Financial Markets* 1 (1): 1–50.

Best, M. J., and R. R. Grauer. 1992. “Positively Weighted Minimum-Variance Portfolios and the Structure of Asset Expected Returns.” *Journal of Financial and Quantitative Analysis* 27 (4): 513–537.

Black, F., and R. Litterman. 1993. “Global Portfolio Optimization.” *Financial Analysts Journal* 48: 28–43.

Bollerslev, T. 1986. “Generalized Autoregressive Conditional Heteroskedasticity.” *Journal of Econometrics* 31 (3): 307–327.

Boyle, P. 2014. “Positive Weights on the Efficient Frontier.” *North American Actuarial Journal* 18 (4): 462–477.

Brandt, M. W., A. Goyal, P. Santa-Clara, and J. R. Stroud. 2005. “A Simulation Approach to Dynamic Portfolio Choice with an Application to Learning about Return Predictability.” *Review of Financial Studies* 18: 831–873.

Carrasco, M., J. P. Florens, and E. Renault. 2007. “Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization.” *Handbook of Econometrics* 6: 5633–5751.

Carrasco, M., and N. Noumon. 2011. “Optimal Portfolio Selection Using Regularization.” Discussion paper.

Chan, L. K. C., J. Karceski, and J. Lakonishok, 1999. “Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model.” *Review of Financial Studies* 12 (5): 937–974.

Chopra, V., and W. Ziemba. 1993. “The Effect of Errors in Means, Variances, and Covariances on Optimal Portfolio Choice.” *The Journal of Portfolio Management* 19: 6–12.

Cont, R. 2001. “Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues.” *Quantitative Finance Volume* 1: 223–236.

Davis, R. A., and T. Mikosch. 1998. “Limit Theory for the Sample ACF of Stationary Process with Heavy Tails with Applications to ARCH.” *The Annals of Statistics* 26: 2049–2080.

DeMiguel, V., L. Garlappi, and R. Uppal. 2009. “Optimal versus Naive Diversification: How Inefficient Is the 1/*N* Portfolio Strategy?” *Review of Financial Studies* 22: 1915–1953.

Engle, R. F. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica* 50 (4): 987–1008.

Fan, J., Y. Liao, and M. Mincheva. 2013. "Large Covariance Estimation by Thresholding Principal Orthogonal Complements." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (4): 603–680.

Frost, P. A., and J. E. Savarino. 1988. "For Better Performance: Constrain Portfolio Weights." *The Journal of Portfolio Management* 15 (1): 29–34.

Garlappi, L., and G. Skoulakis. 2008. "A State Variable Decomposition Approach for Solving Portfolio Choice Problems." *Review of Economic Studies* 23: 3346–3400.

Garlappi, L., R. Uppal, and T. Wang. 2007. "Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach." *The Review of Financial Studies* 20 (1): 41–81.

Gerschgorin, S. 1931. "Über die Abgrenzung der Eigenwerte einer Matrix." Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk 7, 749–754.

Goldfarb, D., and G. Iyengar. 2003. "Robust Portfolio Selection Problems." *Mathematics of Operations Research* 28 (1): 1–38.

Gourieroux, C. *ARCH Models and Financial Applications*. Berlin: Springer, 1997.

Hakansson, N. H. 1971. "Capital Growth and the Mean–Variance Approach to Portfolio Selection." *Journal of Financial and Quantitative Analysis* 6: 517–557.

Jagannathan, R., and T. Ma. 2003. "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps." *The Journal of Finance* 58 (4): 1651–1683.

Jegadeesh, N., and S. Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *The Journal of Finance* 48 (1): 65–91.

Jobson, J. D., and B. Korkie. 1980. "Estimation for Markowitz Efficient Portfolios." *Journal of the American Statistical Association* 75: 544–554.

Jobson, J. D., R. Korkie, and V. Ratti. 1979. "Improved Estimation for Markowitz Portfolios Using James-Stein Type Estimators." *Proceedings of the American Statistical Association* 41: 279–292.

Jorion, P. 1985. "International Portfolio Diversification with Estimation Risk." *The Journal of Business* 58 (3): 259–278.

——. 1986. "Bayes-Stein Estimation for Portfolio Analysis." *The Journal of Financial and Quantitative Analysis* 21 (3): 279–292.

Kroll, Y., H. Levy, and H. M. Markowitz. 1984. "Mean–Variance versus Direct Utility Maximization." *The Journal of Finance* 39: 47–75.

Ledoit, O., and M. Wolf. 2004a. "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis* 88 (2): 365–411.

——. 2004b. "Honey, I Shrunk the Sample Covariance Matrix." *The Journal of Portfolio Management* 30 (4): 110–119.

——. 2017. "Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks." *The Review of Financial Studies* 30 (12): 4349–4388.

Lipschutz, S. *Schaum's Outline: Linear Algebra*. New York: McGraw-Hill; 1991.

Liu, H. 2004. "Optimal Consumption and Investment with Transaction Costs and Multiple Risky Assets." *The Journal of Finance* 59 (1): 289–338.

Loistl, O. 1976. "The Erroneous Approximation of Expected Utility by Means of a Taylor's Series Expansion: Analytic and Computational Results." *The American Economic Review* 66: 904–910.

Lynch, A. W., and S. Tan. 2008. "Multiple Risky Assets, Transaction Costs and Return Predictability: Allocation Rules and Implications for US Investors." *Journal of Financial and Quantitative Analysis* 45 (4): 1015.

MacKinlay, A. C., and L. Pastor. 2000. "Asset Pricing Models: Implications For Expected Returns and Portfolio Selection." *Review of Financial Studies* 13 (4): 883–916.

Markowitz, H. M. 1952. "Portfolio Selection." *The Journal of Finance*, 7 (1): 77–91.

—— 1991. "Foundations of Portfolio Theory." *The Journal of Finance* 46: 469–477.

Mei, X., V. DeMiguel, and F. Nogales. 2016. "Multiperiod Portfolio Optimization with Multiple Risky Assets and General Transaction Costs." *Journal of Banking and Finance* 69: 108–120.

Meucci, A. "Managing Diversification." *Risk,* May 1, 2009.

Muthuraman, K., and S. Kumar. 2006. "Multidimensional Portfolio Optimization with Proportional Transaction Costs." *Mathematical Finance* 16 (2): 301–335.

Partovi, M. H., and M. Caputo. 2004. "Principal Portfolios: Recasting the Efficient Frontier." *Economics Bulletin* 7 (3): 1–10.

Pastor, L. 2000. "Portfolio Selection and Asset Pricing Models." *The Journal of Finance* 55 (1): 179–223.

Pastor, L., and R. Stambaugh, 2000. "Comparing Asset Pricing Models: An Investment Perspective." *Journal of Financial Economics* 56 (3): 335–381.

Pulley, L. B. 1981. "A General Mean–Variance Approximation to Expected Utility for Short Holding Periods." *Journal of Financial and Quantitative Analysis* 16: 361–373.

——. 1983. "Mean–Variance Approximations to Expected Logarithmic Utility." *Operations Research* 31: 685–696.

Robinson, P. D., and A. J. Wathen. 1992. "Variational Bounds on the Entries of the Inverse of a Matrix." *IMA Journal of Numerical Analysis* 12: 463–486.

Samuelson, P. A., 1970. "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances and Moments." *Review of Economic Studies* 37: 537–542.

Steele, A. 1995. "On the Eigen Structure of the Mean Variance Efficient Set." *Journal of Business Finance & Accounting* 22 (2): 245–255.

Stewart, G. W. 1993. "On the Early History of the Singular Value Decomposition." *SIAM Review* 35 (4): 551–566.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

A Backtesting Protocol in the Era of Machine Learning

Rob Arnott, Campbell R. Harvey, and Harry Markowitz

Rob Arnott
is chairman and founder of Research Affiliates, LLC, in Newport Beach, CA.
arnott@rallc.com

Campbell R. Harvey
is a professor of finance at Duke University in Durham, NC, and a partner and senior advisor at Research Affiliates, LLC, in Newport Beach, CA.
cam.harvey@duke.edu

Harry Markowitz
is founder of Harry Markowitz Company in San Diego, CA.
harryhmm@aol.com

Data mining is the search for replicable patterns, typically in large sets of data, from which we can derive benefit. In empirical finance, data mining has a pejorative connotation. We prefer to view data mining as an unavoidable element of research in finance. We are all data miners, even if only by living through a particular history that shapes our beliefs. In the past, data collection was costly, and computing resources were limited. As a result, researchers had to focus their efforts on the hypotheses that made the most sense. Today, both data and computing resources are cheap, and in the era of machine learning, researchers no longer even need to specify a hypothesis—the algorithm will supposedly figure it out.

Researchers are fortunate today to have a variety of statistical tools available, among which machine learning, and the array of techniques it represents, is a prominent and valuable one. Indeed, machine learning has already advanced our knowledge in the physical and biological sciences and has also been successfully applied to the analysis of consumer behavior. All of these applications benefit from a vast amount of data. With large data, patterns will emerge purely by chance. One of the big advantages of machine learning is that it is hardwired to try to avoid overfitting by constantly cross-validating discovered patterns. Again, this advantage serves well in the presence of a large amount of data.

In investment finance, apart from tick data, the data are much more limited in scope. Indeed, most equity-based strategies that purport to provide excess returns to a passive benchmark rely on monthly and quarterly data. In this case, cross-validation does not alleviate the curse of dimensionality. As a noted researcher remarked to one of us:

> [T]uning 10 different hyperparameters using k-fold cross-validation is a terrible idea if you are trying to predict returns with 50 years of data (it might be okay if you had millions of years of data). It is always necessary to impose structure, perhaps arbitrary structure, on the problem you are trying to solve.

Machine learning and other statistical tools, which have been impractical to use in the past, hold considerable promise for the development of successful trading strategies, especially in higher-frequency trading. They might also hold great promise in other applications, such as risk management. Nevertheless, we need to be careful in applying these tools. Indeed, we argue that given the limited nature of the standard data that we use in finance, many of the challenges we face in the era of machine learning are very similar to the issues we have long faced in quantitative finance in general. We want to avoid backtest overfitting of investment strategies, and we want a robust environment to maximize the discovery of new (true) strategies.

We believe the time is right to take a step back and to re-examine how we do our research. Many have warned about the dangers of data mining in the past (e.g., Leamer 1978; Lo and MacKinlay 1990; and Markowitz and Xu 1994), but the problem is even more acute today. The playing field has leveled in computing resources, data, and statistical expertise. As a result, new ideas run the risk of becoming very crowded, very quickly. Indeed, the mere publishing of an anomaly may well begin the process of arbitraging the opportunity away.

Our article develops a protocol for empirical research in finance. Research protocols are popular in other sciences and are designed to minimize obvious errors, which might lead to false discoveries. Our protocol applies to both traditional statistical methods and modern machine learning methods.

HOW DID WE GET HERE?

The early days of quantitative investing brought many impressive successes. Severe constraints on computing and data led research to be narrowly focused. In addition, much of the client marketplace was skeptical of quantitative methods. Consequently, given the limited capital deployed on certain strategies, the risk of crowding was minimal. Today, however, the playing field has changed. Now almost everyone deploys quantitative methods—even discretionary managers—and clients are far less averse to quantitative techniques.

The pace of transformation is striking. Consider the Cray 2, the fastest supercomputer in the world in the late 1980s and early 1990s (Bookman 2017). It weighed 5,500 pounds and, adjusted for inflation, cost over US$30 million in 2019 dollars. The Cray 2 performed an extraordinary (at the time) 1.9 billion operations per second (Anthony 2012). Today's iPhone Xs is capable of 5 trillion operations per second and weighs just six ounces. Whereas a gigabyte of storage cost $10,000 in 1990, it costs only about a penny today. Furthermore, a surprising array of data and application software is available for free, or very nearly free. The barriers to entry in the data-mining business, once lofty, are now negligible.

Sheer computing power and vast data are only part of the story. We have witnessed many advances in statistics, mathematics, and computer science, notably in the fields of machine learning and artificial intelligence. In addition, the availability of open-source software has also changed the game: It is no longer necessary to invest in (or create) costly software. Essentially, anyone can download software and data and potentially access massive cloud computing to join the data-mining game.

Given the low cost of entering the data-mining business, investors need to be wary. Consider the long–short equity strategy whose results are illustrated in Exhibit 1. This is not a fake exhibit.[1] It represents a market-neutral strategy developed on NYSE stocks from 1963 to 1988, then validated out of sample with even stronger results over the years 1989 through 2015. The Sharpe ratio is impressive—over a 50-year span, far longer than most backtests—and the performance is both economically meaningful, generating nearly 6% alpha a year, and statistically significant.

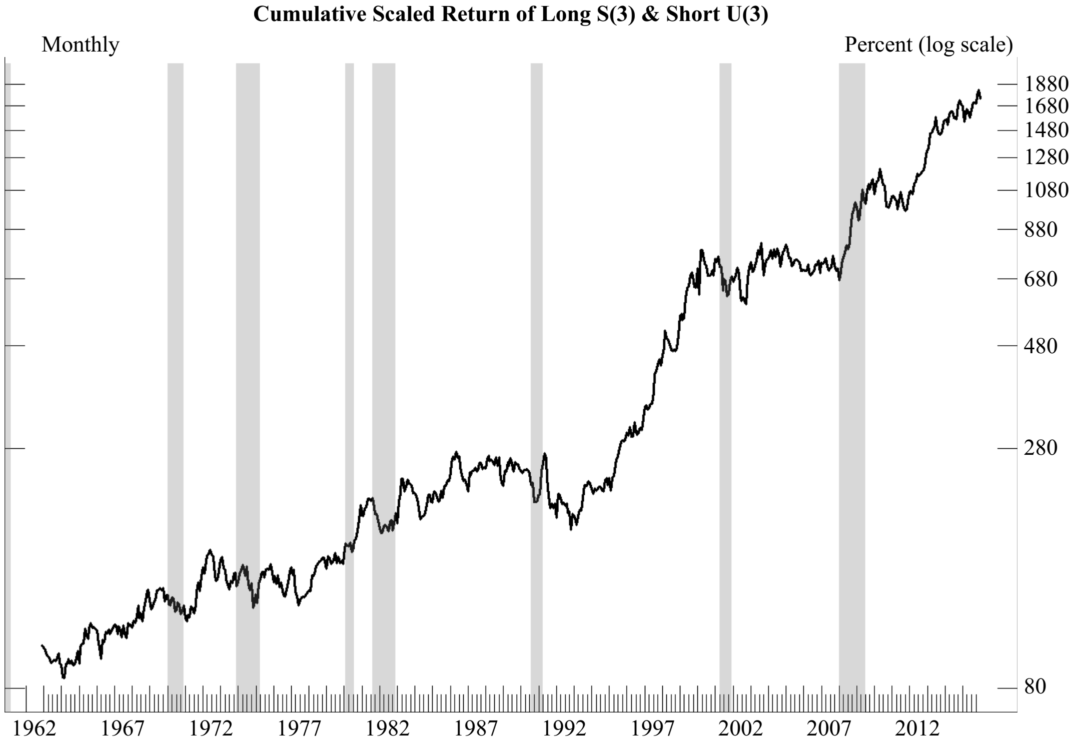
Better still, the strategy has five very attractive practical features. First, it relies on a consistent methodology through time. Second, performance in the most recent period does not trail off, indicating that the strategy is not crowded. Third, the strategy does well during the financial crisis, gaining nearly 50%. Fourth, the strategy has no statistically significant correlations with any of the well-known factors, such as value, size, and momentum, or with the market as a whole. Fifth, the turnover of the strategy is extremely low, less than 10% a year, so the trading costs should be negligible.

This strategy might seem too good to be true. And it is. This data-mined strategy forms portfolios based on letters in a company's ticker symbol. For example, A(1)−B(1) goes long all stocks with "A" as the first letter of their ticker symbol and short all stocks with "B" as the first letter, equally weighting in both portfolios. The strategy in Exhibit 1 considers all combinations of the first three letters of the ticker symbol, denoted as S(3)−U(3). With 26 letters in the alphabet and with two pairings on three possible letters in the ticker symbol, thousands of combinations are possible. In searching

[1] Harvey and Liu (2014) presented a similar exhibit with purely simulated (fake) strategies.

Exhibit 1

Long–Short Market-Neutral Strategy Based on NYSE Stocks, January 1963 to December 2015



Notes: Gray areas denote NBER recessions. Strategy returns scaled to match S&P 500 T-bill volatility during this period.
Source: Campbell Harvey, using data from CRSP.

all potential combinations,[2] the chances of finding a strategy that looks pretty good are pretty high.

A data-mined strategy that has a nonsensical basis is, of course, unlikely to fool investors. We do not see exchange-traded funds popping up that offer "alphabets," each specializing in a letter of the alphabet. Although a strategy with no economic foundation might have worked in the past by luck, any future success would be the result of equally random luck.

The strategy detailed in Exhibit 1, as preposterous as it seems, holds important lessons in both data mining and machine learning. First, the S(3)−U(3) strategy was discovered by brute force, not machine learning. Machine learning implementations would carefully cross-validate the data by training the algorithm on part of the data and then validating on another part of the data. As Exhibit 1 shows, however, in a simple implementation when the S(3)−U(3) strategy was identified in the first quarter-century of the sample, it would be "validated" in the second quarter-century. In other words, it is possible that a false strategy can work in the cross-validated sample. In this case, the cross-validation is not randomized; as a result, a single historical path can be found.

The second lesson is that the data are very limited. Today, we have about 55 years of high-quality equity data (or less than 700 monthly observations) for many of the metrics in each of the stocks we may wish to consider. This tiny sample is far too small for most machine learning applications and impossibly small for advanced approaches such as deep learning. Third, we have a strong prior that the strategy is false: If it works, it is only because of luck. Machine learning, and particularly unsupervised machine learning, does not impose

[2] Online tools, such as those available at http://datagrid.lbl.gov/backtest/index.php, generate fake strategies that are as impressive as the one illustrated in Exhibit 1.

economic principles. If it works, it works in retrospect but not necessarily in the future.

When data are limited, economic foundations become more important. Chordia, Goyal, and Saretto (2017) examined 2.1 million equity-based trading strategies that use different combinations of indicators based on data from Compustat. They carefully took data mining into account by penalizing each discovery (i.e., by increasing the hurdle for significance). They identified 17 strategies that "survive the statistical and economic thresholds."

One of the strategies is labeled (dltis-pstkr)/mrc4. This strategy sorts stocks as follows: The numerator is long-term debt issuance minus preferred/preference stock redeemable. The denominator is minimum rental commitments four years into the future. The statistical significance is impressive, nearly matching the high hurdle established by researchers at CERN when combing through quintillions of observations to discover the elusive Higgs boson (ATLAS Collaboration 2012; CMS Collaboration 2012). All 17 of the best strategies Chordia, Goyal, and Saretto identified have a similarly peculiar construction, which—in our view and in the view of the authors of the paper—leaves them with little or no economic foundation, even though they are based on financial metrics.

Our message on the use of machine learning in backtests is one of caution and is consistent with the admonitions of López de Prado (2018). Machine learning techniques have been widely deployed for uses ranging from detection of consumer preferences to autonomous vehicles, all situations that involve big data. The large amount of data allows for multiple layers of cross-validation, which minimizes the risk of overfitting. We are not so lucky in finance. Our data are limited. We cannot flip a 4TeV switch at a particle accelerator and create trillions of fresh (not simulated) out-of-sample data. But we are lucky in that finance theory can help us filter out ideas that lack an ex ante economic basis.[3]

We also do well to remember that we are not investing in signals or data; we are investing in financial assets that represent partial ownership of a business, or of debt, or of real properties, or of commodities. The quantitative community is sometimes so focused on its models that we seem to forget that these models are crude approximations of the real world and cannot possibly reflect all nuances of the assets that actually comprise our portfolios. The amount of noise usually dwarfs the signal. Finance is a world of human beings, with emotions, herding behavior, and short memories, and market anomalies—opportunities that are the main source of intended profit for the quantitative community and their clients—are hardly static. They change with time and are often easily arbitraged away. We ignore the gaping chasm between our models and the real world at our peril.

THE WINNER'S CURSE

Most in the quantitative community will acknowledge the many pitfalls in model development. Considerable incentives exist to beat the market and to outdo the competition. Countless thousands of models are tried. In contrast to our example with ticker symbols, most of this research explores variables that most would consider reasonable. An overwhelming number of these models do not work and are routinely discarded. Some, however, do appear to work. Of the models that appear to work, how many really do, and how many are just the product of overfitting?

Many opportunities exist for quantitative investment managers to make mistakes. The most common mistake is being seduced by the data into thinking a model is better than it is. This mistake has a behavioral underpinning. Researchers want their model to work. They seek evidence to support their hypothesis—and all of the rewards that come with it. They believe if they work hard enough, they will find the golden ticket. This induces a type of selection problem in which the models that make it through are likely to be the result of a biased selection process.

Models with strong results will be tested, modified, and retested, whereas models with poor results will be quickly expunged. This creates two problems. One is that some good models will fail in the test period, perhaps for reasons unique to the dataset, and will be forgotten. The other problem is that researchers seek a narrative to justify a bad model that works well in the test period, again perhaps for reasons irrelevant to the future efficacy of the model. These outcomes are false negatives and false positives, respectively. Even more common

[3] Economists have an advantage over physicists in that societies are human constructs. Economists research what humans have created, and as humans, we know how we created it. Physicists are not so lucky.

than a false positive is an *exaggerated* positive, an outcome that seems stronger, perhaps much stronger, than it is likely to be in the future.

In other areas of science, this phenomenon is sometimes called the *winner's curse*. This is not the same winner's curse as in auction theory. The researcher who is first to publish the results of a clinical trial is likely to face the following situation: Once the trial is replicated, one of three different outcomes can occur.[4] First (sadly the least common outcome), the trial stands up to many replication tests, even with a different sample, different time horizons, and other out-of-sample tests, and continues to work after its original publication roughly as well as in the backtests. Second, after replication, the effect is far smaller than in the original finding (e.g., if microcap stocks are excluded or if the replication is out of sample). The third outcome is the worst: There is no effect, and the research is eventually discredited. Once published, models rarely work as well as in the backtest.[5]

Can we avoid the winner's curse? Not entirely, but with a strong research culture, it is possible to mitigate the damage.

AVOIDING FALSE POSITIVES: A PROTOCOL

The goal of investment management is to present strategies to clients that perform, as promised, in live trading. Researchers want to minimize false positives but to do so in a way that does not miss too many good strategies. Protocols are widely used both in scientific experiments and in practical applications. For example, every pilot is now required to go through a protocol (sometimes called a checklist) before takeoff, and airline safety has greatly improved in recent years. More generally, the use of protocols has been shown to increase performance standards and prevent failure, as tasks become increasingly complex (e.g., Gawande 2009). We believe that the use of protocols for quantitative research in finance should become de rigueur, especially for machine learning–based techniques, as computing power and process complexity grow. Our goal is to improve investor outcomes in the context of backtesting.

Many items in the protocol we suggest are not new (e.g., Harvey 2017, Fabozzi and López de Prado 2018, and López de Prado 2018), but in this modern era of data science and machine learning, we believe it worthwhile to specify best research practices in quantitative finance.

CATEGORY #1: RESEARCH MOTIVATION

Establish an Ex Ante Economic Foundation

Empirical research often provides the basis for the development of a theory. Consider the relation between experimental and theoretical physics. Researchers in experimental physics measure (generate data) and test the existing theories. Theoretical physicists often use the results of experimental physics to develop better models. This process is consistent with the concept of the scientific method: A hypothesis is developed, and the empirical tests attempt to find evidence inconsistent with the hypothesis—so-called falsifiability.[6]

The hypothesis provides a discipline that reduces the chance of overfitting. Importantly, the hypothesis needs to have a logical foundation. For example, the "alpha-bet" long–short trading strategy in Exhibit 1 has no theoretical foundation, let alone a prior hypothesis. Bem (2011) published a study in a top academic journal that "supported" the existence of extrasensory perception using over 1,000 subjects in 10 years of experiments. The odds of the results being a fluke were 74 billion to 1. They were a fluke: The tests were not successfully replicated.

The researcher invites future problems by starting an empirical investigation without an ex ante economic hypothesis. First, it is inefficient even to consider models or variables without an ex ante economic hypothesis (such as scaling a predictor by rental payments due in the fourth year, as in Exhibit 1). Second, no matter the outcome, without an economic foundation for the

[4] In investing, two of these three outcomes pose a twist to the winner's curse: private gain and social loss. The investment manager pockets the fees until the flaw of the strategy becomes evident, and the investor bears the losses until the great reveal that it was a bad strategy all along.

[5] See McLean and Pontiff (2016). Arnott, Beck, and Kalesnik (2016) examined eight of the most popular factors and showed an average return of 5.8% a year in the span before the factors' publication and a return of only 2.4% after publication. This loss of nearly 60% of the alpha on a long–short portfolio before any fees or trading costs is far more slippage than most observers realize.

[6] One of the most damning critiques of theories in physics is to be deemed unfalsifiable. Should we hold finance theories to a lesser standard?

model, the researcher maximizes the chance that the model will not work when taken into live trading. This is one of the drawbacks of machine learning.

One of our recommendations is to carefully structure the machine learning problem so that the inputs are guided by a reasonable hypothesis. Here is a simple example: Suppose the researcher sets a goal of finding a long–short portfolio of stocks that outperforms on a risk-adjusted basis, using the full spectrum of independent variables available in Compustat and I/B/E/S. This is asking for trouble. With no particular hypothesis, and even with the extensive cross-validation done in many machine learning applications, the probability of a false positive is high.

Beware an Ex Post Economic Foundation

It is also almost always a mistake to create an economic story—a rationale to justify the findings—after the data mining has occurred. The story is often flimsy, and if the data mining had delivered the opposite result, the after-the-fact story might easily have been the opposite. An economic foundation should exist first, and a number of empirical tests should be designed to test how resilient that foundation is. Any suspicion that the hypothesis was developed *after* looking at the data is an obvious red flag.

Another subtle point: In other disciplines such as medicine, researchers often do not have a prespecified theory, and data exploration is crucial in shaping future clinical trials. These trials provide the researcher with truly out-of-sample data. In finance and economics, we do not have the luxury of creating a large out-of-sample test. It is therefore dangerous to appropriate this exploratory approach into our field. We may not jeopardize customer health, but we will jeopardize their wealth. This is particularly relevant when it comes to machine learning methods, which were developed for more data-rich disciplines.

CATEGORY #2: MULTIPLE TESTING AND STATISTICAL METHODS

Keep Track of What Is Tried

Given 20 randomly selected strategies, one strategy will likely exceed the two-sigma threshold (*t*-statistic of 2.0 or above) purely by chance. As a result, the *t*-statistic of 2.0 is not a meaningful benchmark if more than one strategy is tested. Keeping track of the number of strategies tried is crucial, as is measuring their correlations (Harvey 2017; López de Prado 2018). A bigger penalty in terms of threshold is applied to strategies that are relatively uncorrelated. For example, if the 20 strategies tested had a near 1.0 correlation, then the process is equivalent to trying only one strategy.

Keep Track of Combinations of Variables

Suppose the researcher starts with 20 variables and experiments with some interactions, say (variable 1 × variable 2) and (variable 1 × variable 3). This single interaction does not translate into only 22 tests (the original 20, plus two additional interactions) but into 190 possible interactions. Any declared significance should take the full range of interactions into account.[7]

Beware the Parallel Universe Problem

Suppose a researcher develops an economic hypothesis and tests the model once; that is, the researcher decides on the data, variables, scaling, and type of test—all in advance. Given the single test, the researcher believes the two-sigma rule is appropriate, but perhaps it is not. Think of being in 20 different parallel universes. In each, the researcher chooses a different model informed on the identical history. In each, the researcher performs a single test. One of them works. Is it significant at two sigma? Probably not.

Another way to think about this is to suppose that (in a single universe) the researcher compiles a list of 20 variables to test for predictive ability. The first one "works." The researcher stops and claims to have done a single test. True, but the outcome may be lucky. Think of another researcher with the same 20 variables who tests in a different order, and only the last variable "works." In this case, a discovery at two sigma would be discarded because a two-sigma threshold is too low for 20 different tests.

CATEGORY #3: SAMPLE CHOICE AND DATA

Define the Test Sample Ex Ante

The training sample needs to be justified in advance. The sample should never change after the research begins. For example, suppose the model

[7] There are *20 choose 2* interactions, which is 20!/(18!2!).

"works" if the sample begins in 1970 but does not work if the sample begins in 1960—in such a case, the model does not work. A more egregious example would be to delete the global financial crisis data, the tech bubble, or the 1987 market crash because they hurt the predictive ability of the model. The researcher must not massage the data to make the model work.

Ensure Data Quality

Flawed data can lead researchers astray. Any statistical analysis of the data is only as good as the quality of the data that are input, especially in the case of certain machine learning applications that try to capture nonlinearities. A nonlinearity might simply be a bad data point.

The idea of garbage in/garbage out is hardly new. Provenance of the data needs to be taken into account. For example, data from CRSP, Compustat, or some other "neutral" provider should have a far higher level of trust than raw data supplied by some broker. In the past, researchers would literally eyeball smaller datasets and look for anomalous observations. Given the size of today's datasets, the human eyeball is insufficient. Cleaning the data before employing machine learning techniques in the development of investment models is crucial. Interestingly, some valuable data science tools have been developed to check data integrity. These need to be applied as a first step.

Document Choices in Data Transformations

Manipulation of the input data (e.g., volatility scaling or standardization) is a choice and is analogous to trying extra variables. The choices need to be documented and ideally decided in advance. Furthermore, results need to be robust to minor changes in the transformation. For example, given 10 different volatility-scaling choices, if the one the researcher chose is the one that performed the best, this is a red flag.

Do Not Arbitrarily Exclude Outliers

By definition, outliers are influential observations for the model. Inclusion or exclusion of influential observations can make or break the model. Ideally, a solid economic case should be made for exclusion—*before* the model is estimated. In general, no influential observations should be deleted. Assuming the observation is based on valid data, the model should explain all data, not just a select number of observations.

Select Winsorization Level before Constructing the Model

Winsorization is related to data exclusion. Winsorized data are truncated at a certain threshold (e.g., truncating outliers to the 1% or 2% tails) rather than deleted. Winsorization is a useful tool because outliers can have an outsize influence on any model, but the choice to winsorize, and at which level, should be decided before constructing the model. An obvious sign of a faulty research process is a model that "works" at a winsorization level of 5% but fails at 1%, and the 5% level is then chosen.

CATEGORY #4: CROSS-VALIDATION

Acknowledge Out of Sample Is Not Really Out of Sample

Researchers have lived through the hold-out sample and thus understand the history, are knowledgeable about when markets rose and fell, and associate leading variables with past experience. As such, no true out-of-sample data exist; the only true out of sample is the live trading experience.

A better out-of-sample application is on freshly uncovered historical data; for example, some researchers have tried to backfill the historical database of US fundamental data to the 1920s. It is reasonable to assume these data have not been data mined because the data were not previously available in machine readable form. But beware: Although these data were not previously available, well-informed researchers are aware of how history unfolded and how macroeconomic events were correlated with market movements. For those well versed on the history of markets, these data are in sample in their own experience and in shaping their own prior hypotheses. Even for those less knowledgeable, today's conventional wisdom is informed by past events.

As with deep historical data, applying the model in different settings is a good idea but should be done with caution because correlations exist across countries. For example, a data-mined (and potentially fake) anomaly that works in the US market over a certain sample may also work in Canada or the United Kingdom over the same time span, given the correlation between these markets.

Recognize That Iterated Out of Sample Is Not Out of Sample

Suppose a model is successful in the in-sample period but fails out of sample. The researcher observes that the model fails for a particular reason. The researcher modifies the initial model so it then works both in sample and out of sample. This is no longer an out-of-sample test. It is overfitting.

Do Not Ignore Trading Costs and Fees

Almost all of the investment research published in academic finance ignores transactions costs.[8] Even with modest transactions costs, the statistical significance of many published anomalies essentially vanishes. Any research on historical data needs to take transactions costs and, more generally, implementation shortfall into account in both the in-sample and out-of-sample analysis (Arnott 2006).

CATEGORY #5: MODEL DYNAMICS

Be Aware of Structural Changes

Certain machine applications have the ability to adapt through time. In economic applications, structural changes—or nonstationarities—exist. This concern is largely irrelevant in the physical and biological sciences. In finance, we are not dealing with physical constants; we are dealing with human beings and with changing preferences and norms. Once again, the amount of available data is limiting, and the risk of overfitting the dynamics of a relation through time is high.

Acknowledge the Heisenberg Uncertainty Principle and Overcrowding

In physics, the Heisenberg uncertainty principle states that we cannot know a particle's position and momentum simultaneously with precision. The more accurately we know one characteristic, the less accurately we can know the other. A similar principle can apply in finance. As we move from the study of past data into the live application of research, market inefficiencies are hardly static. The cross-validated relations of the past may seem powerful for reasons that no longer apply or may dissipate merely because we are now aware of them and are trading based on them.

Indeed, the mere act of studying and refining a model serves to increase the mismatch between our expectations of a model's efficacy and the true underlying efficacy of the model—and that is before we invest live assets, moving asset prices and shrinking the efficacy of the models through our own collective trading.

Refrain from Tweaking the Model

Suppose the model is running but not doing as well as expected. Such a case should not be a surprise because the backtest of the model is likely overfit to some degree. It may be tempting to tweak the model, especially as a means to improve its fit in recent, now in-sample, data. Although these modifications are a natural response to failure, we should be fully aware that they will generally lead to further overfitting of the model and may lead to even worse live-trading performance.

CATEGORY #6: MODEL COMPLEXITY

Beware the Curse of Dimensionality

Multidimensionality works against the viability of machine learning applications; the reason is related to the limitations of data. Every new piece of information increases dimensionality and requires more data. Recall the research of Chordia, Goyal, and Saretto (2017), who examined 2.1 million equity models based on Compustat data. There are orders of magnitude more models than assets. With so many models, some will work very well in sample.

Consider a model to predict the cross section of stock prices. One reasonable variable to explore is past stock prices (momentum), but many other variables, such as volume, trailing volatility, bid–ask spread, and option skew, could be considered. As each possible predictor variable is added, more data are required, but history is limited and new data cannot be created or simulated.[9]

[8] See Asness and Frazzini (2013). Hou, Xue, and Zhang (2017) showed that most anomaly excess returns disappear once microcaps are excluded.

[9] Monte Carlo simulations are part of the toolkit, perhaps less used today than in the past. Of course, simulations will produce results entirely consonant with the assumptions that drive the simulations.

Macroeconomic analysis provides another example. Although most believe that certain economic state variables are important drivers of market behavior and expected returns, macroeconomic data, generally available on a monthly or quarterly basis, are largely offside for most machine learning applications. Over the post-1960 period,[10] just over 200 quarterly observations and fewer than 700 monthly observations exist.

Although the number of historical observations is limited for each time series, a plethora of macroeconomic variables is available. If we select one or two to be analyzed, we create an implicit data-mining problem, especially given that we have lived through the chosen out-of-sample period.

Pursue Simplicity and Regularization

Given data limitations, regularizing by imposing structure on the data is important. Regularization is a key component of machine learning. It might be the case that a machine learning model decides that a linear regression is the best model. If, however, a more elaborate machine learning model beats the linear regression model, it had better win by an economically significant amount before the switch to a more complex model is justified.

A simple analogy is a linear regression model of Y on X. The in-sample fit can almost always be improved by adding higher powers of X to the model. In out-of-sample testing, the model with the higher powers of X will often perform poorly.

Current machine learning tools are designed to minimize the in-sample overfitting by extensive use of cross-validation. Nevertheless, these tools may add complexity (which is potentially nonintuitive) that leads to disappointing performance in true out-of-sample live trading. The greater the complexity and the reliance on nonintuitive relationships, the greater the likely slippage between backtest simulations and live results.

Seek Interpretable Machine Learning

It is important to look under the hood of any machine learning application. It cannot be a black box. Investment managers should know what to expect with any machine learning–based trading system. Indeed, an interesting new subfield in computer science focuses on interpretable classification and interpretable policy design (e.g., Wang et al. 2017).

CATEGORY #7: RESEARCH CULTURE

Establish a Research Culture That Rewards Quality

The investment industry rewards research that produces backtests with winning results. If we do this in actual asset management, we create a toxic culture that institutionalizes incentives to hack the data to produce a seemingly good strategy. Researchers should be rewarded for good science, not good results. A healthy culture will also set the expectation that most experiments will fail to uncover a positive result. Both management and researchers must have this common expectation.

Be Careful with Delegated Research

No one can perform every test that could potentially render an interesting result, so researchers will often delegate. Delegated research needs to be carefully monitored. Research assistants have an incentive to please their supervisor by presenting results that support the supervisor's hypothesis. This incentive can lead to a free-for-all data-mining exercise that is likely to lead to failure when applied to live data.

Exhibit 2 condenses the foregoing discussion into a seven-point protocol for research in quantitative finance.

CONCLUSIONS

The nexus of unprecedented computing power, free software, widely available data, and advances in scientific methods provide us with unprecedented opportunities for quantitative research in finance. Given these unprecedented capabilities, we believe it is useful to take a step back and reflect on the investment industry's research process. It is naïve to think we no longer need economic models in the era of machine learning. Given that the quantity (and quality) of data is relatively limited in finance, machine learning applications face many of the same issues quantitative finance researchers have struggled with for decades.

[10] Monthly macroeconomic data generally became available in 1959.

EXHIBIT 2

Seven-Point Protocol for Research in Quantitative Finance

1. Research Motivation

a. Does the model have a solid economic foundation?

b. Did the economic foundation or hypothesis exist before the research was conducted?

2. Multiple Testing and Statistical Methods

a. Did the researcher keep track of all models and variables that were tried (both successful and unsuccessful), and are the researchers aware of the multiple-testing issue?

b. Is there a full accounting of all possible interaction variables if interaction variables are used?

c. Did the researchers investigate all variables set out in the research agenda, or did they cut the research as soon as they found a good model?

3. Data and Sample Choice

a. Do the data chosen for examination make sense? And, if other data are available, is it reasonable to exclude these data?

b. Did the researchers take steps to ensure the integrity of the data?

c. Do the data transformations, such as scaling, make sense? Were they selected in advance? And are the results robust to minor changes in these transformations?

d. If outliers are excluded, are the exclusion rules reasonable?

e. If the data are winsorized, was there a good reason to do it? Was the winsorization rule chosen before the research was started? Was only one winsorization rule tried (as opposed to many)?

4. Cross-Validation

a. Are the researchers aware that true out-of-sample tests are only possible in live trading?

b. Are steps in place to eliminate the risk of out-of-sample iterations (i.e., an in-sample model that is later modified to fit out-of-sample data)?

c. Is the out-of-sample analysis representative of live trading? For example, are trading costs and data revisions taken into account?

5. Model Dynamics

a. Is the model resilient to structural change, and have the researchers taken steps to minimize the overfitting of the model dynamics?

b. Does the analysis take into account the risk/likelihood of overcrowding in live trading?

c. Do researchers take steps to minimize the tweaking of a live model?

6. Complexity

a. Is the model designed to minimize the curse of dimensionality?

b. Have the researchers taken steps to produce the simplest practicable model specification?

c. Has an attempt been made to interpret the predictions of the machine learning model rather than using it as a black box?

7. Research Culture

a. Does the research culture reward the quality of the science rather than the finding of a winning strategy?

b. Do the researchers and management understand that most tests will fail?

c. Are expectations clear (that researchers should seek the truth, not just something that works) when research is delegated?

In this article, we have developed a research protocol for investment strategy backtesting. The list is applicable to most research tools used in investment strategy research—from portfolio sorts to machine learning. Our list of prescriptions and proscriptions is long, but hardly exhaustive.

Importantly, the goal is not to eliminate all false positives. Indeed, that is easy—just reject every single strategy. One of the important challenges we face is satisfying the dual objectives of minimizing false strategies but not missing too many good strategies at the same time. The optimization of this trade-off is the subject of ongoing research (see Harvey and Liu 2018).

At first reading, our observations may seem trivial and obvious. Importantly, our goal is not to criticize quantitative investing. Our goal is to encourage humility, to recognize that we can easily deceive ourselves into thinking we have found the Holy Grail. Hubris is our enemy. A protocol is a simple step. Protocols can improve outcomes, whether in a machine shop, an airplane cockpit, a hospital, or for an investment manager. For the investment manager, the presumptive goal is an investment process that creates the best possible opportunity to match or exceed expectations when applied in live trading. Adopting this process is good for the client and good for the reputation of the investment manager.

ACKNOWLEDGMENTS

We have benefited from the comments of Frank Fabozzi, Marcos López de Prado, and Joseph Simonian.

REFERENCES

Anthony, S. "The History of Supercomputers." ExtremeTech.com, April 10, 2012.

Arnott, R. 2006. "Implementation Shortfall." *Financial Analysts Journal* 62 (3) (May/June): 6–8.

Arnott, R., N. Beck, and V. Kalesnik. "Timing 'Smart Beta' Strategies? Of Course! Buy Low, Sell High!" Research Affiliates Publications, September 2016.

Asness, C., and A. Frazzini. 2013. "The Devil in HML's Details." *The Journal of Portfolio Management* 39 (4): 49–68.

ATLAS Collaboration. 2012. "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC." *Physics Letters B* 716 (1): 1–29.

Bem, D. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personality and Social Psychology* 100 (3): 407–425.

Bookman, S. "15 Huge Supercomputers That Were Less Powerful Than Your Smartphone." TheClever.com, April 18, 2017.

Chordia, T., A. Goyal, and A. Saretto. 2017. "*p*-Hacking: Evidence from Two Million Trading Strategies." Swiss Finance Institute Research Paper No. 17–37, SSRN.

CMS Collaboration. 2012. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC." *Physics Letters B* 716 (1): 30–61.

Fabozzi, F., and M. López de Prado. 2018. "Being Honest in Backtest Reporting: A Template for Disclosing Multiple Tests." *The Journal of Portfolio Management* 45 (1): 141–147.

Gawande, A. *The Checklist Manifesto: How to Get Things Right.* New York: Henry Holt and Sons, 2009.

Harvey, C. R. 2017. "Presidential Address: The Scientific Outlook in Financial Economics." *The Journal of Finance* 72: 1399–1440.

Harvey, C. R., and Y. Liu. 2014. "Evaluating Trading Strategies." *The Journal of Portfolio Management* 40 (5): 108–118.

——. 2018. "False (and Missed) Discoveries in Financial Economics." SSRN, https://ssrn.com/abstract=3073799.

Hou, K., C. Xue, and L. Zhang. 2017. "Replicating Anomalies." SSRN, https://www.ssrn.com/abstract=2961979.

Leamer, E. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: John Wiley & Sons, 1978.

Lo, A., and A. C. MacKinlay. 1990. "Data-Snooping Biases in Tests of Financial Asset Pricing Models." *Review of Financial Studies* 3 (3): 431–467.

López de Prado, M. 2018. "The 10 Reasons Most Machine Learning Funds Fail." *The Journal of Portfolio Management* 44 (6): 120–133.

Markowitz, H., and B. L. Xu. 1994. "Data Mining Corrections." *The Journal of Portfolio Management* 21 (1): 60–69.

McLean, R. D., and J. Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *The Journal of Finance* 71 (1): 5–32.

Wang, T., C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille. 2017. "A Bayesian Framework for Learning Rule Sets for Interpretable Classification." *Journal of Machine Learning Research* 18: 1–37.

Modeling Analysts' Recommendations via Bayesian Machine Learning

David Bew, Campbell R. Harvey, Anthony Ledford, Sam Radnor, and Andrew Sinclair

David Bew
is principal engineer at Man-AHL in Oxford, United Kingdom.
david.bew@man.com

Campbell R. Harvey
is professor of finance at Duke University in Durham, NC, and a research associate at the National Bureau of Economic Research in Cambridge, MA.
cam.harvey@duke.edu

Anthony Ledford
is chief scientist at Man-AHL and an associate at Oxford-Man Institute of Quantitative Finance in Oxford, United Kingdom.
anthony.ledford@man.com

Sam Radnor
is senior vice president at quant PORT in London, United Kingdom.
sradnor@quantport-am.com

Andrew Sinclair
is senior quantitative analyst at Realindex Investments in Sydney, Australia.
andrew.sinclair@realindex.com.au

In 2009, a unique citizen science project called the Galaxy Zoo Supernovae project was launched.[1] One of the goals of the project was to identify new supernovae (SN)—and to recruit the help of thousands of amateur astronomers. The astronomers were asked to give three levels of classification: very likely SN object, possible SN object, and not likely SN object. Determination of the true classification came from the spectrographic analysis of Caltech's Palomar Transient Factory.[2]

The problem arose as to how to combine the classifications. At any point in time, many astronomers may be scoring a particular object. Should we look at the average classification? Obviously the classifications are imperfect, and an average may reduce the noise. A simple majority vote (yes or no) is another possibility. However, both the majority and the average do not allow for differential skill among the classifiers. Is there a way to build a system that takes the track record of the astronomer into account? Importantly, the quality of the track record should be dynamic to allow for both improvement through time as well as fatigue.

Such a task is an ideal application of a type of machine learning called *independent Bayesian classifier combination*[3] (IBCC), originally defined by Ghahramani and Kim (2003). The Galaxy Zoo data were analyzed by Simpson et al. (2013) with impressive results. They found that their probabilistic model for the IBCC technique led to dramatic improvements in classification. For example, allowing for a 10% error rate, the rate of correct classification went from approximately 65% using the average to 97% using IBCC.

What does the classification of SN have to do with finance? It turns out that there are striking similarities to the problem facing an investment manager in evaluating analysts' recommendations: As in the Galaxy Zoo project, there are thousands of objects (companies) and thousands of astronomers (analysts). In both cases, the subjects do not cover all the objects (companies), but only a subset (sparsity). The classification mechanism in the Galaxy Zoo project (very likely, possible, and not likely) has an uncanny resemblance to buy, hold, or sell. In addition, it is reasonable to assume a differential degree of skill among analysts; hence, the IBCC method, given its track record in the physical and biological sciences, is a logical place to start.

The goal of our article is to apply IBCC to the I/B/E/S forecast universe to determine

[1] See Lintott (2012).

[2] https://www.ptf.caltech.edu/iptf.

[3] Despite its name, the IBCC model does not assume independence but, instead, assumes conditional independence, which is discussed later.

whether the classifier provides information that may lead to improved investment management. We are fully aware that analysts' forecasts are a well-researched area in the academic finance and accounting literature. Indeed, Brown (2000) detailed 575 studies, many of which are focused on analysts' forecasts—and that article is 20 years out of date. A search of SSRN's Financial Economics Network and Accounting Research Network reveals over 1,000 papers dealing with analysts' forecasts.[4]

Despite the large quantity of research, ours is the first article (that we know of) to apply IBCC to the important problem of how to combine analysts' recommendations. Previous applications of IBCC in economics include work by Levenberg et al. (2013), who focused on forecasting the trend of the US nonfarm payrolls, and by Levenberg et al. (2014), who incorporated sentiment measures obtained using sentence-level language analysis. The popularity of IBCC in large-scale machine learning applications is largely due to it providing a scalable multidimensional inference procedure for combining arbitrary groups of simultaneous recommendations from multiple sources. It does this while requiring only univariate classifier learning, thereby allowing the set of sources to be easily extended. These features also make it ideal for combining analysts' forecasts.

With the potential for incorporating so many classifier sources, avoiding overfitting becomes an important consideration. Bayesian models are not as prone to overfitting as are models that require point estimates to be specified for large numbers of parameters; uncertainty about all the unknowns in a Bayesian model is described using their joint posterior probability distribution. Prediction requires integrating over this distribution, a procedure that properly accounts for diffuse knowledge about all parameters rather than requiring point values to be ascribed. The primary drawback of Bayesian models, which automatically account for parameter uncertainty, is that their use can be computationally demanding, often making them unsuitable or even impossible for real-time use. In contrast, our inference approach uses a state-of-the-art Bayesian technique called *variational approximation*, and it is extremely efficient computationally. The model we present here can be applied to learn about each analyst individually or groups of analysts. Restrictions currently in place require that we only report at the broker level.

We realize that predicting financial outcomes remains difficult, even when expansive datasets and sophisticated machine learning models are available. Our primary aim is not to identify the best analyst or broker but to make a coherent ensemble forecast in which the weight given to each broker is driven by the length and quality of the broker's track record. In our application, the best results arise when there is agreement between broker recommendations and the forecasts obtained using IBCC. This confirmation (or reinforcement) effect, which pervades our long-only, long–short, and short-only portfolios and the various robustness analyses we perform, suggests intriguing ways for machine learning to enhance the investment processes of both quantitative and discretionary fund managers.

Our article is organized as follows. The second section discusses the data, focusing on nonstandard features such as their categorical nature, dependence structure, and sparsity (i.e., characteristics that necessitate a bespoke modeling treatment). The third section details the IBCC model and discusses important choices about priors and hyperparameters within our Bayesian framework. The fourth section explains how inference is undertaken using a state-of-the-art computationally efficient technique called variational approximation. Empirical results are presented in the fifth section, together with a range of robustness checks. Concluding remarks and some suggestions for further research are offered in a final part.

DESCRIPTION OF THE DATA MODELING PROBLEM

Our study falls within the area of machine learning known as *supervised learning.* The input data are categorical analyst recommendations about individual companies and are obtained from a large, publicly available database. Associated with each analyst recommendation is a categorical outcome variable (sometimes called a target, or *truth*, within the IBCC literature) that describes the directional price movement of the company's stock (relative to a benchmark) subsequent to the recommendation. We aim to use a modern Bayesian machine learning method to learn the relationship between these input and target data and thereby predict future price movements based on current recommendations data.

[4] Early reviews of the literature were conducted by Givoly and Lakonishok (1984), Schipper (1991), and Brown (1993). A more recent treatment was done by Bradshaw (2011).

Input Data: I/B/E/S Broker Recommendations

A vast amount of analyst data are available on both the individual stocks and the various subsectors within international equities markets. Our focus here is on recommendation data from the Thomson Reuters I/B/E/S database, a data source that covers nearly all analysts within their respective geographies and provides analyst-by-analyst recommendations for individual securities.

A *recommendation* is simply an analyst's rating for a particular company, and because different analysts use a variety of ratings schemes, each recommendation received from a contributing analyst is mapped by Thomson Reuters to one of five Standard Ratings: strong buy, buy, hold, underperform, and sell.

Several factors distinguish such data from those typically encountered in mainstream financial forecasting applications. First, unlike in standard time-series forecasting, recommendations are not observed at a fixed frequency but are event based; that is, they are observed irregularly and at largely unpredictable discrete dates. Second, instead of being quantitative forecasts on some continuous-valued scale, recommendations are categorical. This makes them better suited to a classification-based analysis than to a standard regression approach. Additionally, the recommendation database we examine has the following characteristics:

1. **Very high dimensionality:** Recommendations are received on thousands of stocks from thousands of individual analysts.
2. **Extreme sparsity:** Typically only a small number of analysts issue recommendations on any particular stock on any particular day; the rest say nothing.
3. **Dependence:** We expect analyst recommendations to be statistically dependent for a number of reasons:
   A. **Cross-sectional dependence:** Contributing analysts often have exposure to correlated information sets and therefore reach the same or similar conclusions even though their decision processes are otherwise independent. This is an example of an important special case in statistics: When a multivariate random variable, $(X_1, X_2, \ldots, X_m, Z)$ say, is such that $\Pr(X_1, X_2, \ldots, X_m | Z) = \Pr(X_1 | Z) \times \Pr(X_2 | Z) \times \cdots \times \Pr(X_m | Z)$, then the Xs are said to be conditionally independent given Z, or equivalently, the Xs are independent conditional on Z. The IBCC model makes extensive use of such a conditional independence structure (see the third section).
   B. **Temporal dependence:** Analyst views typically update gradually, and analysts often restate their previous recommendations. This leads to serial correlation. Group behavior among analysts can also generate serial correlation (e.g., some analysts leading opinion and others following consensus).
4. **Lack of consistency:** Although the analyst recommendations provided by I/B/E/S are recorded on the common five-category scale given earlier, for many analysts, only two of these categories are populated. Other analysts may use three of the available categories and still others all five. Although it is quite possible to deal with this inconsistency using all five categories within the IBCC model, there is little practical gain in doing so here. Thus, we group together the first two and last two Thomson Reuters Standard Ratings and relabel the original I/B/E/S analyst recommendations as buy, hold, and sell. For each I/B/E/S recommendation, we artificially label each analyst not issuing a recommendation for that stock-day pair with the category label "Missing." This means that recommendations are recorded on the following four-category scale: missing, buy, hold, and sell. Finally, we note that the distribution of buys and sells can be extremely uneven reflecting inherent biases in broker behavior.

Accounting for any one of these four characteristics within a Bayesian analysis requires detailed probabilistic modeling. Our IBCC methodology deals with all of them simultaneously and does so with a computationally rapid approach that allows the resulting system to calibrate dynamically to the prevailing environment. We also require that the prediction computations required for forecasting be feasible in real time so incoming recommendations can be responded to with minimal delay. Our Bayesian approach also allows prior beliefs to be accommodated so that the system can be guided by information from outside the observed data, should that be required.

Outcome Data: Post-Recommendation Price Movements

Unlike the input recommendations data, which are intrinsically categorical, the outcome data we seek to predict are price movements of the underlying company's stock over some future time horizon. Such price movements arise on an essentially continuous rather than categorical scale, whereas the IBCC model, which we seek to apply here, requires categorical targets. Our first step is therefore to create these categorical targets for the historical recommendation data. For consistency with the IBCC literature, these targets will be referred to as *truths*.

We first need to choose the time horizon, $\Delta\tau$, over which we are interested in predicting the movement of the stock price; for the majority of this study, we use $\Delta\tau = 60$ business days. For each analyst recommendation, we note the day it became public, s, and calculate $r_{(s,\Delta\tau)}$, which is defined as the excess return of the relevant stock over the $\Delta\tau$ period starting the next business day after s and measured relative to our benchmark return. We use this together with a relative measure of index volatility to define a categorical truth variable t for each recommendation according to

$$t = \begin{cases} 0, & \text{if } r_{(s,\Delta\tau)} \leq -5\% \times RVol_s, \\ 2, & \text{if } r_{(s,\Delta\tau)} \geq \ \ 5\% \times RVol_s, \\ 1, & \text{otherwise} \end{cases}$$

where $RVol_s$ denotes an estimator of index volatility scaled to have unit mean. Given their obvious interpretations, we refer interchangeably to the truth states $\{0, 1, 2\}$ as Price_Down, Price_Flat, and Price_Up, respectively. Clearly, the truth variable defined here has nothing to do with any broker recommendation being correct or incorrect; it is determined solely by the subsequent performance of the stock relative to the index after the recommendation. Many reasonable extensions of this truth variable definition are possible—for example, one could incorporate the market β of each underlying stock.

Exhibit 1
The Structure of the Dataset

Stock ID	Date	Truth	Broker 1	Broker 2	...	Broker N
#1234	July 10, 2008	0	3	0		0
#5678	Feb 7, 2012	0	0	1		2
#5678	July 1, 2012	2	2	0		0
#5678	Mar 14, 2012	1	0	3		1
	⋮	⋮	⋮	⋮	⋮	⋮

Notes: Integer codes {0,1,2} are used to denote the truth outcomes {Price_Down, Price_Flat, Price_Up}, respectively. The artificial recommendation label "Missing" is encoded as 0 for each noncontributing broker in each row, and the resulting recommendation set {Missing, Hold, Sell, Buy} is encoded as {0,1,2,3}, respectively. Each row contains at least one nonzero recommendation code. A very high proportion of recommendations is recorded as 0, corresponding to "Missing," because only a small number of brokers within each row issues hold, sell, or buy recommendations.

We restrict our attention to the period January 1, 2004, to January 1, 2013, and include only the pan-European region comprising Austria, Belgium, the Czech Republic, Cyprus, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, the Netherlands, Norway, Poland, Portugal, Russia, Spain, Sweden, Switzerland, Turkey, and the United Kingdom. Our benchmark is the Dow Jones Euro Stoxx Index.[5] Additionally, at the announcement time of each recommendation, we apply a filter to the stock universe to ensure our results are free of survivorship bias.

We group analysts by their stated corporate employer, henceforth *broker*, which gives 347 separate brokers. To be clear, the IBCC technique can be applied at the level of individual analysts or at the broker level. Because of reporting restrictions, we focus this article at the broker level.

Aggregating recommendations about the same stock that arise on the same day, we obtain the combined recommendations and truths dataset described in Exhibit 1. The dataset has 105,319 rows.[6] If recommendations were recorded for all 347 brokers for each of the 105,319 rows there would be 36,545,693 nonzero recommendation codes, corresponding to combinations of the labels hold, sell, and buy. However, the reality is that only 116,220 of the recommendation codes in Exhibit 1 are nonzero, meaning 99.7% correspond to the label "Missing." This demonstrates the extreme sparsity of the data object at the heart of our IBCC analysis.

[5] Bloomberg ticker: SXXE Index.

[6] This choice of a one-day aggregation period is arbitrary and is something we return to later. From the previous discussion about group behavior, we would expect statistical dependence between rows at this aggregation were analysts to issue recommendations on a stock prompted by others doing so.

THE IBCC MODEL: PROBABILISTIC SPECIFICATION AND CONSTRUCTION OF THE POSTERIOR

The IBCC model is a fully probabilistic model that relates a constellation of categorical inputs—in our case, the constellation of broker recommendations within each row of the data object described in Exhibit 1—and a categorical truth variable associated with those inputs.[7]

We start by specifying a probabilistic model over the categorical truth variable T. In our IBCC implementation, T takes values over states $\{0, 1, 2\}$ corresponding to Price_Down, Price_Flat, and Price_Up, respectively, and is assumed to have probability mass function

$$\Pr(T = t | \boldsymbol{\kappa}) = \kappa_t \text{ for } t \in \{0, 1, 2\} \tag{1}$$

where the parameter $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$ denotes a three-vector of probabilities so that $\kappa_0 + \kappa_1 + \kappa_2 = 1$. This specification is simply saying the truths $\{0, 1, 2\}$ occur with probabilities $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$ respectively, and that no other truth outcomes are possible. The conditioning notation in Equation 1 makes explicit that the parameter $\boldsymbol{\kappa}$ is assumed to be known at this stage.

The next step is to specify, for each broker, three separate distributions to describe their recommendation behavior given each possible truth. More explicitly, letting $B_k \in \{0, 1, 2, 3\}$ denote the recommendation of broker k corresponding to missing, hold, sell, and buy, respectively, for each $k \in \{1, \ldots, N\}$ we require distributions for the following three conditional random variables: $B_k|T = 0$, $B_k|T = 1$, and $B_k|T = 2$. Writing T_j for the truth in row j, the IBCC model assumes, conditionally on $T_j = t$, that the B_k are independent and have probability mass functions given by

$$\Pr(B_k = b_{kj} | T_j = t, \boldsymbol{\pi}_t^{(k)}) = \pi_{t\ b_{kj}}^{(k)} \quad \text{for } b_{kj} \in \{0,1,2,3\} \tag{2}$$

where, for each truth $t \in \{0, 1, 2\}$, the parameter $\boldsymbol{\pi}_t^{(k)} = [\pi_{t0}^{(k)}, \pi_{t1}^{(k)}, \pi_{t2}^{(k)}, \pi_{t3}^{(k)}]$ denotes a four-vector of probabilities for broker k and so satisfies $\pi_{t0}^{(k)} + \pi_{t1}^{(k)} + \pi_{t2}^{(k)} + \pi_{t3}^{(k)} = 1$. This conditional specification looks complicated, but all we are doing is defining three separate four-dimensional multinomial distributions for each broker, one for each of the possible truth outcomes. Thus, for each broker k, we have parameters $\boldsymbol{\pi}_0^{(k)}$, $\boldsymbol{\pi}_1^{(k)}$, and $\boldsymbol{\pi}_2^{(k)}$. Again, the conditioning notation in Equation 2 makes explicit that the parameters $\boldsymbol{\pi}_t^{(k)}$ are assumed known at this point.

The assumption that the broker recommendations within row j are independent conditionally on $T_j = t_j$ allows the likelihood contribution for row j to be constructed, giving

$$\Pr(T = t_j, B_1 = b_{1j}, B_2 = b_{2j}, \ldots, B_N = b_{Nj}) \\ = \kappa_{t_j} \ \pi_{t_j b_{1j}}^{(1)} \ \pi_{t_j b_{2j}}^{(2)} \cdots \pi_{t_j b_{Nj}}^{(N)} = \kappa_{t_j} \prod_{l=1}^{N} \pi_{t_j b_{lj}}^{(l)}$$

The IBCC model assumes all rows in the data object described in Exhibit 1 are independent, so the full likelihood, over its n distinct rows, is given by

$$\Pr(\boldsymbol{t}, \boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N) = \prod_{j=1}^{n} \left(\kappa_{t_j} \prod_{l=1}^{N} \pi_{t_j b_{lj}}^{(l)} \right) \tag{3}$$

where for notational brevity we have written $\boldsymbol{t} = (t_1, \ldots, t_n)$ for the column of n truths in Exhibit 1 and $\boldsymbol{b}_k = (b_{k1}, \ldots, b_{kn})$ for the recommendation column for broker k, for each $k \in \{1, \ldots, N\}$.

So far we have treated the parameters of the truths and broker distributions—that is, $\boldsymbol{\kappa}$ and $(\boldsymbol{\pi}_0^{(k)}, \boldsymbol{\pi}_1^{(k)}, \boldsymbol{\pi}_2^{(k)})$ for $k \in \{1, \ldots, N\}$, respectively—as fixed parameters. In a frequentist analysis, we would need to estimate these (e.g., by maximum likelihood and its well-established asymptotic theory) to obtain point estimates and confidence intervals. This is not the approach we adopt here. Our Bayesian analysis requires that we treat all these quantities probabilistically so that each is described according to its own prior probability distribution. Formulation of the posterior distribution then proceeds via the product of these prior distributions and the likelihood given in Equation 3, and inferences are made based on the posterior distribution alone (see Lee 2012).

Thus, we must specify priors over $\boldsymbol{\kappa}$ and $\boldsymbol{\pi}_0^{(k)}$, $\boldsymbol{\pi}_1^{(k)}$, and $\boldsymbol{\pi}_2^{(k)}$ for each $k \in \{1, \ldots, N\}$. Because the truth and broker recommendation distributions are all examples of multinomial distributions, we choose to use the family of Dirichlet distributions as priors because the

[7] Code for IBCC is available at https://github.com/edwinrobots/pyIBCC. This is not the code that we used for our research.

Dirichlet family is the conjugate[8] family of priors for the multinomial distribution (for details, see Bishop 2006).

For the truth probabilities $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$, we assume a three-dimensional Dirichlet distributed prior, that is, the continuous distribution with probability density function over domain of support $D = \{(\kappa_0, \kappa_1, \kappa_2);\ 0 \leq \kappa_j \leq 1, \Sigma_{t=0}^{2}\kappa_t = 1\}$ given by $\Pr(\boldsymbol{\kappa}|\boldsymbol{\nu}) = C(\boldsymbol{\nu})\Pi_{t=0}^{2}\kappa_t^{\nu_{0t}-1}$, where $C(\boldsymbol{\nu}) = \Gamma(\nu_{00} + \nu_{01} + \nu_{02})/\{\Gamma(\nu_{00}) \times \Gamma(\nu_{01}) \times \Gamma(\nu_{02})\}$; the three-vector $\boldsymbol{\nu} = (\nu_{00}, \nu_{01}, \nu_{02})$ denotes a so-called *hyperparameter* (i.e., a parameter of the prior); and $\Gamma(\cdot)$ is the gamma function. Note that substituting $\nu_{0t} \equiv 1$ into this probability density function for each $t \in \{0, 1, 2\}$ yields a flat prior for $\boldsymbol{\kappa}$ over D. Similarly, for each broker recommendation B_k for $k \in \{1, \ldots, N\}$ and conditional on truth $t \in \{0, 1, 2\}$ we assume $\{\pi_{t0}^{(k)}, \pi_{t1}^{(k)}, \pi_{t2}^{(k)}, \pi_{t3}^{(k)}\}$ has a four-dimensional Dirichlet prior with hyperparameters $\{\alpha_{0,t0}^{(k)}, \alpha_{0,t1}^{(k)}, \alpha_{0,t2}^{(k)}, \alpha_{0,t3}^{(k)}\}$. To condense the notation, we denote the complete set of broker recommendation probabilities conditional on each truth by $\boldsymbol{\Pi} = [\{\pi_{t0}^{(k)}, \pi_{t1}^{(k)}, \pi_{t2}^{(k)}, \pi_{t3}^{(k)}\}: t = 0, 1, 2; k = 1, \ldots, N]$ and their corresponding hyperparameters by $\mathbf{A}_0 = [\{\alpha_{0,t0}^{(k)}, \alpha_{0,t1}^{(k)}, \alpha_{0,t2}^{(k)}, \alpha_{0,t3}^{(k)}\}: t = 0, 1, 2; k = 1, \ldots, N]$.

Having now fully specified both the likelihood and the prior, we are equipped to construct the posterior distribution, which is proportional to their product, and hence satisfies

$$\begin{aligned}&\Pr(\boldsymbol{\kappa}, \boldsymbol{\Pi}, \boldsymbol{t}, \boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N \mid \mathbf{A}_0, \boldsymbol{\nu}) \\ &\quad \propto \prod_{j=1}^{n}\left(\kappa_{t_j}\prod_{l=1}^{N}\pi_{t_j b_{lj}}^{(l)}\right)\Pr(\boldsymbol{\kappa}|\boldsymbol{\nu})\Pr(\boldsymbol{\Pi}|\mathbf{A}_0). \qquad (4)\end{aligned}$$

This is a joint distribution in over 4,000 dimensions[9] and incorporates information about $\boldsymbol{\kappa}$ and $\boldsymbol{\Pi}$ from both the observed data and the prior. In some IBCC applications, it is common to choose informative priors; however, we deliberately choose priors that are flat over their respective domains. This is achieved by setting the hyperparameters $\nu_{0t} \equiv 1$ for each $t \in \{0, 1, 2\}$, and $\alpha_{0,tj}^{(k)} \equiv 1$ for each $j \in \{0, 1, 2, 3\}$ where $\{k \in 1, \ldots, N;\ t \in 0, 1, 2\}$. These flat priors ensure that only information learned from the observed truths and recommendations data, and not our choice of priors, is driving the trading signals and allows straightforward assessment of the efficacy of our learning framework.

For the avoidance of doubt, we note that the IBCC model incorporates no sense of ordering within the category labels for either the truths or the broker recommendations. Its fundamental job is simply to learn how one set of labels (the broker recommendations) relates to the other set (the truths, which encode subsequent price outcome). Indeed, a broker that always recommends buy when the truth is Price_Down is just as informative within our IBCC implementation as a broker that always recommends sell in such cases.

The high dimensionality and data sparsity of our application mean using alternative dependence models (e.g., copulas) to capture the dependence between different brokers is computationally infeasible. The IBCC model deals with this limitation by assuming conditional independence and thereby provides a scalable and computationally efficient multidimensional inference procedure over arbitrary groups of classifiers that requires only univariate classifier learning. This key feature of the IBCC model is one of the reasons it has become popular for large-scale Bayesian machine learning applications.

VARIATIONAL BAYESIAN INFERENCE

In this section, we introduce variational Bayesian inference, an approach sometimes termed *variational Bayes*, or simply VB. See Bishop (2006, Chapter 10) and Blei, Kucukelbir, and McAuliffe (2018) for detailed treatments and Fox and Roberts (2011) for a tutorial.[10] We then provide the key results of applying VB to our IBCC model. The theory is elegant, but its mathematical derivation can obscure the simplicity of the underlying approach: We approximate a multivariate distribution by a product of simpler distributions that we update iteratively to obtain the best overall approximation. In what follows, all logarithms are natural logs, that is, $\log_e(\cdot)$.

Let $\boldsymbol{X}$ denote a set of observed data and $\boldsymbol{Z}$ a combined set of latent (i.e., unobserved) parameters and variables. We use the generic shorthand $p(\cdot)$ to denote the probabilistic model governing whatever quantities appear inside the parentheses; for example, the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Z}$ is written $p(\boldsymbol{X}, \boldsymbol{Z})$. Our goal is to find a good approximation, $q(\boldsymbol{Z})$ say, for the posterior

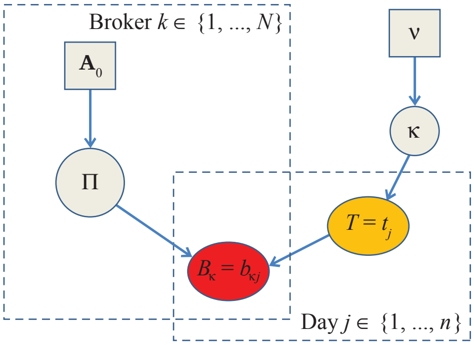
[8] A conjugate prior is one that leads to a posterior distribution that is within the same parametric family as the prior, which therefore leads to greatly simplified Bayesian analysis. See Bishop (2006).

[9] There are 347 brokers, each requiring three separate four-dimensional distributions, plus the three-dimensional truth distribution. In all, this makes 347 × 4 × 3 + 3 = 4,167 dimensions.

[10] Also, see https://staff.aist.go.jp/bevan.jones/vb-tutorial-slides.pdf.

EXHIBIT 2

Graphical Model of Our IBCC Implementation


Broker $k \in \{1, ..., N\}$
$\mathbf{A}_0$
ν
κ
Π
$T = t_j$
$B_\kappa = b_{\kappa j}$
Day $j \in \{1, ..., n\}$

Notes: Elliptical/circular nodes are variables with a distribution, whereas rectangular nodes represent hyperparameter variables that are instantiated with fixed values. The red shaded node represents recommendations, which are observed during both training and prediction. The orange shaded node represents truths, which are observed during training but have to be inferred during prediction.

$p(\boldsymbol{Z}|\boldsymbol{X})$. In our IBCC implementation, $\boldsymbol{Z}$ will include the truth outcome we seek to predict (i.e., Price_Up, Price_Down, or Price_Flat; see Exhibit 2).

Noting that $q(\boldsymbol{Z})$ represents a probability model and therefore integrates to one, we may always write $\log p(\boldsymbol{X}) = \int q(\boldsymbol{Z}) \log p(\boldsymbol{X}) d\boldsymbol{Z}$, where $p(\boldsymbol{X})$ denotes the so-called *model evidence*. Furthermore, because the definition of conditional probability gives $p(\boldsymbol{X}) = p(\boldsymbol{X},\boldsymbol{Z})/p(\boldsymbol{Z}|\boldsymbol{X})$, we may substitute for $p(\boldsymbol{X})$ in this integral to obtain

$$\log p(\boldsymbol{X}) = \int q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X},\boldsymbol{Z})}{p(\boldsymbol{Z} \mid \boldsymbol{X})} \right\} d\boldsymbol{Z}$$

$$= \int q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})} \times \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} \right\} d\boldsymbol{Z}$$

This can be written as $\log p(\boldsymbol{X}) = L(q) + KL(q,p)$, where $KL(q,p) = -\int q(\boldsymbol{Z}) \log\{p(\boldsymbol{Z}|\boldsymbol{X})/q(\boldsymbol{Z})\} d\boldsymbol{Z}$ denotes the Kullback–Leibler[11] divergence (KL-divergence) between $q(\boldsymbol{Z})$ and $p(\boldsymbol{Z}|\boldsymbol{X})$, and $L(q) = \int q(\boldsymbol{Z}) \log\{p(\boldsymbol{X},\boldsymbol{Z})/q(\boldsymbol{Z})\} d\boldsymbol{Z}$ is the negative of a quantity called the *variational free energy*.[12] Standard properties of the KL-divergence include that it is always nonnegative and that $KL(q,p) = 0$ if and only if $q(\boldsymbol{Z})$ equals $p(\boldsymbol{Z}|\boldsymbol{X})$. This implies $L(q)$ is a lower bound for $\log p(\boldsymbol{X})$ and furthermore that this lower bound can be maximized by minimizing the KL-divergence, $KL(q,p)$, with respect to the distribution $q(\boldsymbol{Z})$. This is a *calculus of variations* problem.[13]

VB considers a restricted but tractable family of distributions to represent $q(\boldsymbol{Z})$ and then seeks the element of that family that maximizes $L(q)$. The approach we adopt involves partitioning $\boldsymbol{Z}$ into m groups of variables and assuming that $q(\boldsymbol{Z})$ can be approximated by the factorized structure $q(\boldsymbol{Z}) = \Pi_{i=1}^{m} q_i(\boldsymbol{Z}_i)$. This factorized version of variational approximation has its origins in physics, where it is called *mean field theory*.[14] Thus, among all distributions of the form $q(\boldsymbol{Z}) = \Pi_{i=1}^{m} q_i(\boldsymbol{Z}_i)$, we seek the distributions $q_i^*(\boldsymbol{Z}_i)$ that jointly maximize $L(q)$. To be clear, other than the assumed factorization structure $q(\boldsymbol{Z}) = \Pi_{i=1}^{m} q_i(\boldsymbol{Z}_i)$, no further assumptions about $q(\boldsymbol{Z})$ are required.

Substituting our assumed factorization $q(\boldsymbol{Z}) = \Pi_{i=1}^{m} q_i(\boldsymbol{Z}_i)$ into the definition of $L(q)$ given earlier and adopting the notation $q_i = q_i(\boldsymbol{Z}_i)$, we obtain $L(q) = \int \Pi_i q_i \{\log p(\boldsymbol{X},\boldsymbol{Z}) - \Sigma_i \log q_i\} d\boldsymbol{Z}$. We now rewrite this expression to make clear how it depends on one of the individual factors, $q_j(\boldsymbol{Z}_j)$ say, noting that any terms not involving q_j may be treated as constant with respect to $\boldsymbol{Z}_j$. We thereby obtain

$$L(q) = \int q_j \left\{ \int \log p(\boldsymbol{X},\boldsymbol{Z}) \; \Pi_{i \neq j} q_i \, d\boldsymbol{Z}_i \right\} d\boldsymbol{Z}_j$$

$$- \int q_j \log \, q_j \, d\boldsymbol{Z}_j + \text{Constant} \tag{5}$$

We now define the new distribution $\tilde{p}(\boldsymbol{X},\boldsymbol{Z}_j)$ by $\log \tilde{p}(\boldsymbol{X},\boldsymbol{Z}_j) = E_{i \neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})] + c$, where c is a normalization constant and $E_{i \neq j}[\cdot]$ denotes expectation with respect to all q_i distributions for $i \neq j$ so that

[11] The KL-divergence between the two probability distributions f and g is a global measure of their dissimilarity and is defined by $KL(f,g) = -\int f(\boldsymbol{x}) \log\{g(\boldsymbol{x})/f(\boldsymbol{x})\} d\boldsymbol{x}$. It is called a divergence, rather than a distance, because it is not symmetric; that is, $KL(f,g) \neq KL(g,f)$. Standard properties include that $KL(f,g) \geq 0$ always and that $KL(f,g) = 0$ if and only if $f = g$.

[12] To avoid the possibility of misinterpretation, for clarity we remark that $L(q)$ is not the likelihood function. Writing $-L(q) = -\int q(\boldsymbol{Z}) \log p(\boldsymbol{X},\boldsymbol{Z}) d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \log\{1/q(\boldsymbol{Z})\} d\boldsymbol{Z}$, we obtain an energy term minus an entropy term, which is why it is called the free energy. See Sato (2001).

[13] Standard calculus allows functions to be optimized, where a function is a map that takes the value of some variable as input and returns the value of the function as output. Calculus of variations allows functionals to be optimized rather than functions, where functionals are maps that take functions as inputs.

[14] See Parisi (1988).

$E_{i\neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})] = \int \log p(\boldsymbol{X},\boldsymbol{Z})\Pi_{i\neq j} q_i d\boldsymbol{Z}_i$. Careful inspection of Equation 5 now shows that $L(q)$ is simply the negative KL-divergence between $q_j(\boldsymbol{Z}_j)$ and $\tilde{p}(\boldsymbol{X},\boldsymbol{Z}_j)$, which is minimized by taking $q_j(\boldsymbol{Z}_j) = \tilde{p}(\boldsymbol{X},\boldsymbol{Z}_j)$. Thus, keeping q_i constant for each $i \neq j$, we have that maximizing $L(q)$ over all possible distributions $q_j(\boldsymbol{Z}_j)$ is achieved by taking $\log q_j^*(\boldsymbol{Z}_j) = E_{i\neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})] + c$, where c denotes a normalizing constant. This key result provides the basis for application of variational methods.

The set of equations $\log q_j^*(\boldsymbol{Z}_j) = E_{i\neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})] + c$ for each $j \in \{1,\ldots,m\}$ provides conditions for the maximum of $L(q)$ subject to the assumed factorization $q(\boldsymbol{Z}) = \Pi_{i=1}^{m} q_i(\boldsymbol{Z}_i)$. However, these equations do not provide an explicit solution because the expression for each $q_j^*(\boldsymbol{Z}_j)$ involves taking expectation with respect to the other $q_i^*(\boldsymbol{Z}_i)$ distributions for $i \neq j$. To solve these equations, we proceed iteratively. First, each $q_i(\boldsymbol{Z}_i)$ distribution is initiated—for example, with parameters chosen broadly to match moments of the observed data. Then we cycle through each $j \in \{1, \ldots, m\}$, updating $q_j(\boldsymbol{Z}_j)$ by evaluating $E_{i\neq j}[\log p(\boldsymbol{X},\boldsymbol{Z})]$ using the current estimates of $q_i(\boldsymbol{Z}_i)$ for each $i \neq j$. Convergence to a local maximum is guaranteed because of certain convexity properties of $L(q)$ with respect to the factors $q_i(\boldsymbol{Z}_i)$ (see Boyd and Vendenberghe 2004). Furthermore, in the particular case of our IBCC implementation, because all the factors we have chosen are of the exponential family type (Bernardo and Smith 1994), this maximum can be shown to be the global maximum within the family of factorized distributions.

Variational Inference for Our IBCC Implementation

Our IBCC application deviates from those of Kim and Ghahramani (2012) and Simpson et al. (2013) in three key ways. First, we intend to perform online forecasting, so temporal consistency requires running the model using only information that is already available at the time of each forecast. Second, with the exception of truths corresponding to recommendations made within the most recent $\Delta\tau$ period, all truths within our training data are completely observed because they are based on publicly available price data. In contrast, for the Galaxy Zoo project, the truth data were largely missing. Finally, our primary interest is the predictive distribution $\Pr(T = t | B_1 = b_1, B_2 = b_2, \ldots, B_N = b_N)$, rather than the posterior, because we wish to forecast the truth outcome conditional on, for example, today's constellation of broker recommendations.[15]

Although it is possible to extend the IBCC model to include explicit temporal structure, as done by Simpson et al. (2013), our approach is based on calibrating their simpler static model to a dataset that updates as time evolves. Specifically, we truncate the observed data, comprising the time-stamped recommendations and truths, at a sequence of evaluation dates, ensuring additionally that a buffer of duration $\Delta\tau$ is incorporated between the last admitted training observations and the onset of prediction. For each training data set so created, we seek to calculate the predictive distribution $\Pr(T = t | B_1 = b_1, B_2 = b_2, \ldots, B_N = b_N)$ for each constellation of broker recommendations that arises until the next evaluation date. Learning remains halted over this prediction phase, so each constellation of analyst recommendations we use in prediction is treated individually. All our findings are obtained using this rolling out-of-sample scheme.

For each evaluation date, we undertake both expanding-window and moving-window analyses. The expanding-window analysis admits all data from January 1, 2004, up to the evaluation date, whereas the moving-window analysis admits only data within a three-year lookback from each evaluation date. In principle, the evaluation dates could be chosen to index each business day; however, for practical reasons[16] we set them quarterly, to the first day of March, June, September, and December.

Let index $i \in (1, \ldots, n_r)$ denote the rows of the training data, renumbered as required for the rolling window case. Because all the recommendations and truths are observed for these training data and because we chose conjugate Dirichlet priors for both $\boldsymbol{\kappa}$ and $\boldsymbol{\Pi}$, standard properties of the multinomial-Dirichlet family (see Bishop 2006) give the following:

1. The posterior of $\boldsymbol{\kappa}$ is a Dirichlet distribution with parameter $\boldsymbol{\nu}^* = (\nu_0^*, \nu_1^*, \nu_2^*)$, where $\nu_j^* = \nu_{0j} + N_j$

[15] The predictive distribution we seek, sometimes called the *posterior predictive distribution*, is defined by the multivariate integral $\int \Pr(T = t | B_1 = b_1, B_2 = b_2, \ldots, B_N = b_N, \boldsymbol{\kappa}, \boldsymbol{\Pi}) \Pr(\boldsymbol{\kappa}, \boldsymbol{\Pi}) d\boldsymbol{\kappa} d\boldsymbol{\Pi}$, where $\Pr(\boldsymbol{\kappa}, \boldsymbol{\Pi})$ denotes the posterior distribution of $(\boldsymbol{\kappa}, \boldsymbol{\Pi})$, which depends implicitly on the training data.

[16] Risk managers tend to have a preference for models with parameters that remain static for reasonable periods rather than models in which parameters change on a daily basis.

and N_j denotes the number of occurrences of truth j in the training data for $j \in \{0, 1, 2\}$. The $\nu_j^* = \nu_{0j} + N_j$ formula is often referred to as the *prior counts plus data counts* updating relationship for the multinomial-Dirichlet family.

2. The posterior of $\boldsymbol{\pi}_t^{(l)}$ is a Dirichlet distribution with parameters $(\alpha_{t0}^{(l*)}, \alpha_{t1}^{(l*)}, \alpha_{t2}^{(l*)}, \alpha_{t3}^{(l*)})$, where $\alpha_{tb}^{(l*)} = N_{tb}^{(l)} + \alpha_{0,tb}^{(l)}$, and $N_{tb}^{(l)}$ denotes the number of recommendations of type $b \in \{0, 1, 2, 3\}$ made in the training data by broker $l \in \{1,\ldots, N\}$ for each truth $t \in \{0, 1, 2\}$. We let $\mathbf{A}^*$ denote the collection of all these posterior parameters.

Our procedure for approximating the predictive distribution $\Pr(T = t|B_1 = b_1, B_2 = b_2,\ldots, B_N = b_N)$ starts by considering $\Pr(\boldsymbol{\kappa}, \boldsymbol{\Pi}, t, b_1, b_2,\ldots, b_N|\mathbf{A}^*, \boldsymbol{\nu}^*)$. This has the same structure as the individual data terms in Equation 4 except that now the truth t is unobserved, and $(\mathbf{A}^*, \boldsymbol{\nu}^*)$ denotes the ensemble of posterior parameters given earlier. Thus, $\log\Pr(\boldsymbol{\kappa}, \boldsymbol{\Pi}, t, b_1, b_2, \ldots, b_N|\mathbf{A}^*, \boldsymbol{\nu}^*)$ is of the form

$$\sum_{j=0}^{2} I(t = j)\left(\log \kappa_j + \sum_{l=1}^{N} \log \pi_{jb_l}^{(l)}\right) + \log \Pr(\boldsymbol{\kappa}|\boldsymbol{\nu}^*) + \log \Pr(\boldsymbol{\Pi}|\mathbf{A}^*) + \text{Constant}. \tag{6}$$

Here, we have introduced the indicator function $I(\cdot)$, defined by $I(t = j) = 1$ if $t = j$ and $I(t = j) = 0$ otherwise, because it will be convenient later. To reduce clutter, we drop the dependence on $(b_1,\ldots, b_N\, \mathbf{A}^*, \boldsymbol{\nu}^*)$ from our notation. We therefore represent the latent variables and parameters by $\boldsymbol{Z} = (t, \boldsymbol{\kappa}, \boldsymbol{\Pi})$. We assume $q(\boldsymbol{Z})$ factorizes as $q(t, \boldsymbol{\kappa}, \boldsymbol{\Pi}) = q(t)q(\boldsymbol{\kappa}, \boldsymbol{\Pi})$. This is the only assumption we need to make; several further simplifications arise because of the structure of the IBCC model. For example, Equation 6 shows that the terms involving $\boldsymbol{\kappa}$ and $\boldsymbol{\Pi}$ can be separated, which implies the additional factorization $q^*(\boldsymbol{\kappa}, \boldsymbol{\Pi}) = q^*(\boldsymbol{\kappa})q^*(\boldsymbol{\Pi})$.

We start by initializing the distributions for $\boldsymbol{\kappa}$ and $\boldsymbol{\pi}_t^{(l)}$ with their posterior distributions, that is, the Dirichlet distributions with parameters $\boldsymbol{\nu}^*$ and $\mathbf{A}^*$ given earlier. To obtain $q^*(t)$, we need to evaluate $\log q^*(t) = E_{\boldsymbol{\kappa},\boldsymbol{\Pi}}[\log p(t, \boldsymbol{\kappa}, \boldsymbol{\Pi})] + \text{Constant}$. Extracting the relevant terms from Equation 6, we obtain $\log q^*(t) = E_{\boldsymbol{\kappa}} \log \kappa_t + \Sigma_{l=1}^N E_{\boldsymbol{\pi}_t^{(l)}} \log \pi_{tb_l}^{(l)} + \text{Constant}$. Standard properties of the Dirichlet distribution (e.g., Bishop 2006) give $E_{\boldsymbol{\kappa}} \log \kappa_t = \Psi(\nu_{\text{t}}^*) - \Psi(\Sigma_{j=0}^2 \nu_j^*)$ and $E_{\boldsymbol{\pi}_t^{(l)}} \log \pi_{tb_l}^{(l)} = \Psi(\alpha_{tb_l}^{(l*)}) - \Psi(\Sigma_{s=0}^3 \alpha_{ts}^{(l*)})$, where $\Psi(\cdot)$ denotes the DiGamma function.[17] Next, defining the terms $\log \rho_t = \Psi(\nu_t^*) - \Psi(\Sigma_{j=0}^2 \nu_j^*) + \Sigma_{l=1}^N[\Psi(\alpha_{tb_l}^{(l*)}) - \Psi(\Sigma_{s=0}^3 \alpha_{ts}^{(l*)})]$, where $b_1,\ldots, b_N$ denote the observed broker recommendations for the prediction, we therefore obtain $q^*(t) = \rho_t / (\rho_0 + \rho_1 + \rho_2)$. This expression for $q^*(t)$ provides our initial estimate of $\Pr(T = t| B_1 = b_1,\ B_2 = b_2,\ldots, B_N = b_N)$.

Deriving $q^*(\boldsymbol{\kappa})$ and $q^*(\boldsymbol{\Pi})$ requires taking expectations with respect to this newly calculated $q^*(t)$ distribution. We start by extracting the terms involving $\boldsymbol{\kappa}$ from Equation 6. Recalling that for any event X, the expectation of $I(X)$ is $\Pr(X)$, we obtain $\log q^*(\boldsymbol{\kappa}) = \Sigma_{j=0}^2 q(t = j)\log \kappa_j + \Sigma_{j=0}^2(\nu_j^* - 1)\log \kappa_j + \text{Constant}$. Gathering together the $\log\kappa_j$ terms in this expression shows $q^*(\boldsymbol{\kappa})$ to be Dirichlet distributed with parameters $\nu_j = \nu_j^* + q(t = j)$ for $j \in \{0, 1, 2\}$. This formula for iterating the $\boldsymbol{\kappa}$ distribution is similar in structure to the prior counts plus data counts relation noted previously, except that now the prior over each forecasting period is the posterior obtained at the relevant evaluation date, and the counts for each truth class are replaced with their expected values; that is, $E_t I(t = j) = q(t = j)$ for $j \in \{0,1,2\}$.

We essentially repeat this argument to obtain the update equations for $q^*(\boldsymbol{\Pi})$. First, because the $\boldsymbol{\pi}_j^{(l)}$ terms in Equation 6 are separate for each truth $j \in \{0, 1, 2\}$ and each broker $l \in \{1,\ldots, N\}$, we obtain the further factorization $q^*(\boldsymbol{\Pi}) = \Pi_{l=1}^N \Pi_{j=0}^2 q^*(\boldsymbol{\pi}_j^{(l)})$. Extracting the $\boldsymbol{\pi}_j^{(l)}$ terms and taking expectation with respect to $q^*(t)$ thereby yields $\log q^*(\boldsymbol{\pi}_j^{(l)}) = \Sigma_{j=0}^2 q^*(t = j)\Sigma_{l=1}^N \log \pi_{tb_l}^{(l)} + \Sigma_{s=0}^3(\alpha_{js}^{(l*)} - 1)\log \pi_{js}^{(l)} + \text{Constant}$. Gathering together terms in $\log \pi_{jb}^{(l)}$ now shows $q^*(\boldsymbol{\pi}_j^{(l)})$ to be Dirichlet distributed with parameters $\alpha_{jb}^{(l)} = q(t = j)I(b = b_l) + \alpha_{jb}^{(l*)}$ for $b \in \{0,1,2,3\}$ and $l \in \{1,\ldots, N\}$. As before, these equations for iterating the $\boldsymbol{\Pi}$ distributions have the same prior counts plus expected counts interpretation.

Having updated both $q^*(\boldsymbol{\kappa})$ and $q^*(\boldsymbol{\Pi})$, we now use these distributions to obtain the next update of $q^*(t)$, and the whole scheme is iterated until convergence is obtained. The truth distribution that results is the VB

[17] If the d-dimensional variable $\boldsymbol{X} = (X_1,\ldots, X_d)$ is Dirichlet distributed with parameter $(\mu_1,\ldots, \mu_d)$, then $E(\log X_i) = \Psi(\mu_i) - \Psi(\Sigma_{j=1}^d \mu_j)$ for each $i \in \{1,\ldots, d\}$, where $\Psi(\cdot)$ denotes the DiGamma function, which is defined as $\Psi(z) = \frac{d}{dz}\log\Gamma(z)$, where $\Gamma(\cdot)$ denotes the gamma function.

approximation to $\Pr(T = t | B_1 = b_1,\ B_2 = b_2, \ldots, B_N = b_N)$. In practice, convergence is achieved rapidly.

Although we have expressed the method in terms of a single prediction, in practice the calculations can be undertaken in parallel, allowing efficient prediction of the truth distribution for multiple constellations of broker recommendations. We remark that although the VB iteration scheme is operationally similar to the update procedure of the expectation-maximization (EM) algorithm,[18] the VB and EM algorithms do very different things: EM obtains the maximum likelihood (i.e., point) estimate of a parameter, whereas VB provides a global approximation of the distribution.

From Predictive Probabilities to Decisions

The outputs of the previous procedure are the estimated truth probabilities, (q_0, q_1, q_2) say, for Price_Down, Price_Flat, and Price_Up, respectively, for each out-of-sample constellation of broker recommendations. Even when these predictive probabilities have been calculated, one still requires a decision rule—that is, a rule to decide what, if any, action to take.

We restrict our attention to the discrete set of actions Go_Short, No_Trade, and Go_Long.[19] It is tempting to choose one of these actions according to whichever of Price_Down, Price_Flat, or Price_Up has the highest predictive probability (HPP). Unfortunately, this HPP rule, which chooses Go_Short if $q_0 > \max(q_1, q_2)$, Go_Long if $q_2 > \max(q_0, q_1)$, and No_Trade otherwise, is not selective enough and results in too many Go_Long actions. This behavior is unsurprising because the underlying training dataset contains unadjusted biases; analysts typically issue more buy recommendations than hold or sell, and there are more Price_Up labels than Price_Down or Price_Flat.[20]

Recalling that q_t is an estimate of the conditional probability $\Pr(T = t | B_1 = b_1,\ B_2 = b_2, \ldots,\ B_N = b_N)$, our preferred decision rule is to take the HPP action only when q_t exceeds the current estimate of the unconditional probability of $T = t$, which is κ_t. This simple extension of the HPP rule ensures a Go_Long (Go_Short) decision arises only when knowledge of the observed constellation of broker recommendations $B_1 = b_1$, $B_2 = b_2, \ldots$, $B_N = b_N$ boosts the estimated probability of Price_Up (Price_Down) relative to the background level observed within the training data.

Our default decision rule is the $c = k = 1$ case of the more general decision rule summarized as follows:

Decision	Trigger Condition
Go_Short	$q_0/\kappa_0 > c$ and $q_0 > k \max(q_1, q_2)$
No_Trade	otherwise
Go_Long	$q_2/\kappa_2 > c$ and $q_2 > k \max(q_0, q_1)$

Both parameters, c and k, affect the selectivity of this trading rule, but their effects are different and somewhat complementary. The parameter c relates to comparison of the conditional and unconditional probabilities of each truth outcome. Thus, increasing c while keeping $k = 1$ fixed means the value of the information imparted by the broker recommendations needs to be higher for a Go_Long (Go-Short) decision to arise. In contrast, the condition involving parameter k relates to the relationship among the three conditional truth probabilities, q_0, q_1, and q_2, but does not involve the unconditional probabilities. Thus, increasing k while keeping $c = 1$ fixed raises the threshold required for HPP decision making to produce a Go_Long (Go-Short) outcome; simply being the largest value of q_0, q_1, and q_2 is no longer sufficient.

EMPIRICAL RESULTS AND ROBUSTNESS CHECKS

The results are based on grouping the analysts by broker (i.e., their stated corporate employers or affiliation). Learning is undertaken at this broker level and is achieved by integrating information over all the stocks and all the analysts affiliated with that broker. It is possible to implement IBCC on different types of groupings—or even by individual analysts. Such information pooling is a powerful feature of the IBCC model and Bayesian approaches more generally (e.g., providing protection against overfitting). Finer aggregations than this are possible; for example, learning could be

[18] See Dempster, Laird, and Rubin (1977) and Tanner (1996).

[19] Many alternatives to our discrete choice rule are possible here. For example, the calculated (q_0, q_1, q_2) probabilities could be used to derive weights on a continuous long–short scale.

[20] In the Galaxy Zoo project, Simpson et al. (2013) subsampled to adjust for class imbalance. We chose not to do this, instead developing a model that reflects the probabilistic structure of the observed dataset, including its biases, and dealing with these biases using an extension of the HPP decision rule.

undertaken at the Global Industry Classification Standard sector or subsector level within each broker or even at the individual analyst level, where sufficiently detailed tracking information exists to follow an analyst's career between brokers. There is, of course, a complexity penalty for finer aggregations—more model components to infer based on the same amount of data. We do not report on such aggregations here.

Another feature of our IBCC implementation is its ability to combine multiple simultaneous recommendations for each stock without the need for extra parameters. To exploit this, in the backtest simulations reported later, recommendations are aggregated over a lookback of 30 calendar days, a process that increases the number of concurrent recommendations within the rows of the training data. This procedure is best understood by considering a single stock: When a new recommendation appears, we simply look back and find the latest recommendations from the other brokers within a 30-day window and group them together in a single row of the data. Further examinations (not reported) show the impact of this choice of lookback window to be minimal.

Our standard approach is to estimate the IBCC model on a three-year period of in-sample data and then apply it out of sample to the recommendations that arise over the subsequent quarter. We then either expand or roll forward the in-sample period to include the next quarter, always applying the new fit out of sample to the following unused quarter of data. The default decision rule we use is the $c = k = 1$ case of the rule given previously. The impact of varying the parameters c and k is examined later.

We benchmark IBCC performance against a scheme that does no learning but simply aims to follow each broker's recommendations. This broker-following benchmark is referred to as *Brok_Flw* in the exhibits that follow and allows assessment of the value added by IBCC.

The Brok_Flw benchmark is constructed as follows:

1. For every buy recommendation, we create a signal of +1 that lasts from the day following the recommendation for 60 business days.
2. Likewise, for every sell recommendation we create a signal of −1.
3. These signals are summed within a stock, both across the multiple brokers and across multiple recommendations from the same broker.
4. The resulting signal is capped/floored at ±10.
5. For long-only portfolios, only underlying long recommendations are included, and conversely for short-only portfolios.
6. Each portfolio's positions are rebalanced on a daily basis to maintain a gross exposure of \$100; that is, $position_{it} = signal_{it} / \Sigma_i \left| signal_{it} \right|$, where the sum in this normalization is across all contemporaneous positions, both long and short.

The following nomenclature is used in presenting the results:

- **Brok_Flw_LS**: This is the broker-following benchmark described previously. We ignore recommendations in which there are simultaneous buys and sells for the same stock from different brokers.
- **IBCC_Rol_LS**: Here we apply the IBCC algorithm, fitting on a three-year rolling window, with both long and short positions.
- **IBCC_Exp_LS**: As noted earlier, but now the estimation is performed on an expanding window.
- **Both_Rol_LS**: *Both* here denotes that we only take a position if the IBCC recommendation and the raw Brok_Flw signal agree at the individual broker level. This prevents IBCC from reversing broker recommendations. Estimation is performed on a three-year rolling window.
- **Both_Exp_LS**: As noted earlier, but using the IBCC model on an expanding window.

Here, L (S) is used in place of LS when only long (short) positions are allowed. In all cases, the gross exposure is normalized to \$100.[21] This means that net exposure for the LS portfolio is time varying according to the relative number of long and short recommendations. In particular, the LS results in the exhibits cannot be imputed from the separate L and S short results.

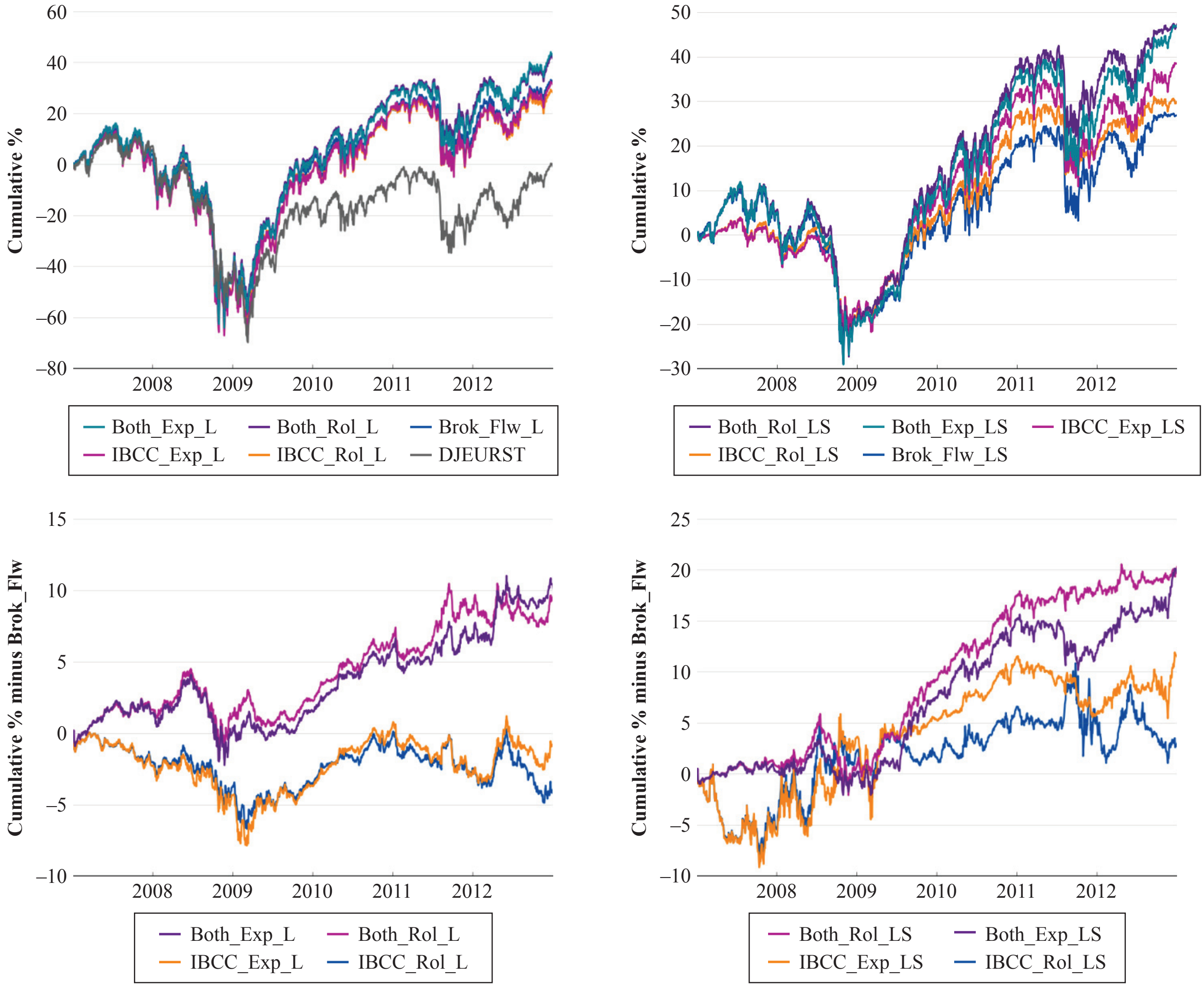
The reference index used for the intercept and slope estimates, α and β, reported in the following is the Euro Stoxx,[22] the same index we used in defining

[21] Gross exposure is defined as $\Sigma \left| pos_i \right|$, where pos_i is the position in the i^{th} market, in US dollars.

[22] DJEURST in Thomson Reuters notation.

Exhibit 3

Performance of Long-Only Models (left) and Long–Short Models (right)



Notes: Outright performance is shown in the top panels, whereas the bottom panels show performance relative to the relevant Brok_Flw_* benchmark. Note the vertical axes do not share a common scale.

the truths for the training data. This is based on a liquid subset of around 300 Eurozone stocks from the STOXX Europe 600. This index had an average return close to zero over the 2007–2012 period, so the reported alphas are similar to the outright returns. Returns on short portfolios are reported assuming that all stocks have been borrowed and sold short; however, transaction and borrowing costs are not included in the results.

Exhibit 3 shows the performance of the long-only and long-short portfolios, both in terms of their outright performance and their performance relative to the relevant Brok_Flw_* benchmark. All long portfolios struggled during the global financial crisis (GFC) but comfortably outperformed the DJEURST index from 2009 onward. The long IBCC strategies remain broadly in line with the Brok_Flw_L benchmark, with best long performance arising for the strategies labeled Both.

Exhibit 4

Performance Statistics for the Various Long-Only, Short-Only, and Long–Short Models for the Period 2007–2012

					Alpha		Beta	
Side	**Model**	**Mean**	**Vol**	**Alpha**	***t*-Stat**	**Beta**	***t*-Stat**	**Turnover**
Long-Only	Brok_Flw_L	5.43	24.18	5.47	2.73	1.01	26.97	5.75
	IBCC_Rol_L	4.77	24.66	4.91	2.09	1.02	23.36	5.74
	IBCC_Exp_L	5.30	24.89	5.39	2.27	1.03	23.63	5.68
	Both_Rol_L	6.99	24.51	7.06	2.75	1.00	20.18	6.13
	Both_Exp_L	7.13	24.71	7.28	2.84	1.01	20.47	6.07
Short-Only	Brok_Flw_S	–0.51	24.96	–0.13	–0.05	–1.03	–29.42	6.38
	IBCC_Rol_S	–3.38	25.24	–3.23	–1.46	–1.05	–34.83	6.15
	IBCC_Exp_S	–3.54	24.85	–3.53	–1.60	–1.03	–34.76	6.18
	Both_Rol_S	2.99	25.98	3.45	1.06	–1.03	–21.81	7.12
	Both_Exp_S	2.11	25.71	2.46	0.79	–1.03	–25.39	7.04
Long–Short	Brok_Flw_LS	4.54	13.92	4.65	2.09	0.52	10.88	6.50
	IBCC_Rol_LS	5.07	11.01	5.30	2.69	0.39	10.63	7.23
	IBCC_Exp_LS	6.50	12.66	6.64	3.35	0.47	13.46	7.06
	Both_Rol_LS	7.99	15.43	8.15	3.18	0.56	11.11	6.32
	Both_Exp_LS	7.88	16.00	8.09	3.12	0.59	11.69	6.29

Notes: The reference index used for the α *and* β *calculations is the Euro Stoxx, the same index used for defining the truths in the training data. The alpha values are annualized. Turnover denotes a measure of the volume traded by each portfolio on a standardized scale that allows meaningful comparison between portfolios.*

For the long–short portfolios, there is no corresponding LS index, but all IBCC portfolios outperform the Brok_Flw_LS benchmark. Again, the portfolios labeled Both provide the strongest performance. Investing only when both the IBCC model and the underlying broker recommendations agree suggests a straightforward and intriguing way this machine learning application may assist investment management. No consistent benefit of fitting with rolling or expanding data windows is observed in these results.

Results for all the long-only, long–short, and short-only portfolios are tabulated in Exhibit 4, and a yearly breakdown is provided in Exhibit 5. The Brok_Flw_S benchmark and both of the short IBCC strategies are loss making, so we do not focus on their outright performance. The more interesting point is that the short portfolios labeled Both again perform better, repeating the outperformance pattern seen earlier in the long-only and long–short cases. The relative performance chart for the short-only portfolios is given in Exhibit 6 and shows the outperformance of the Both portfolios to be reasonably consistent over the post-GFC period.

Robustness Checking—Impact of Firm Liquidity

A potentially serious concern is that our IBCC procedure might be favoring recommendations from brokers who recommend smaller, less well-known stocks and thus may be inadvertently accessing a size bias. A quick check of Exhibit 7, for example, shows that Brok_Flw_L holds more stocks over $25 billion than does IBCC.

In an attempt to control for this effect, we split the stock universe in half by market capitalization. We rank the original universe of liquid stocks by market capitalization and form a large-half backtest by including only the largest half of these stocks; in the small-half backtest, we only include the smallest half. This determination is made each month and is implemented with a five-business-day lag in an effort to reduce short-term timing effects. In the subsequent backtesting, we use these reduced universes both for the fitting of the IBCC models and subsequently for their assessment on the usual rolling out-of-sample basis. The overall number of recommendations in the two backtests is shown in Exhibit 8. The split is surprisingly even.

EXHIBIT 5

Calendar Year Performance for Long-Only, Short-Only, and Long–Short Portfolios from 2007 to 2012 Inclusive (expressed as percentage)

Side	Year	Brok_Flw	IBCC_Rol	IBCC_Exp	Both_Rol	Both_Exp	Euro Stoxx
Long-Only	2007	4.37	2.37	2.11	6.07	5.83	7.51
	2008	–47.35	–49.06	–49.38	–47.82	–48.56	–51.09
	2009	44.20	44.45	44.80	45.25	45.47	28.84
	2010	20.29	23.92	24.79	25.00	25.04	5.84
	2011	–9.99	–12.79	–13.56	–8.00	–9.33	–11.91
	2012	20.69	19.42	22.65	20.94	23.81	20.63
Short-Only	2007	5.39	–0.32	–0.99	7.07	8.83	7.51
	2008	46.59	41.07	40.97	41.64	41.26	–51.09
	2009	–42.88	–46.24	–43.83	–39.54	–38.96	28.84
	2010	–12.20	–13.29	–12.73	–7.41	–5.05	5.84
	2011	20.76	18.86	18.04	33.24	25.15	–11.91
	2012	–20.70	–20.14	–22.42	–17.26	–18.74	20.63
Long–Short	2007	5.04	–0.36	–0.99	5.87	5.29	7.51
	2008	–24.52	–16.82	–14.78	–24.42	–24.59	–51.09
	2009	22.66	21.94	24.22	30.99	29.98	28.84
	2010	16.58	21.62	22.81	24.83	24.33	5.84
	2011	–4.83	–6.51	–10.37	–4.03	–7.83	–11.91
	2012	12.00	10.18	17.62	14.12	19.51	20.63

Notes: The figures quoted are the sum of each year's daily returns. For reference, Euro Stoxx returns are shown in the right-hand column.

Backtest performance is shown in Exhibit 9 for the long-only and short-only cases,[23] and the distributions of market capitalization for the two subportfolios are shown in Exhibit 10. We conclude that IBCC is able to add value to plain I/B/E/S estimates in both large- and small-capitalization subportfolios and that the efficacy of the algorithm is not driven by a size bias.

Robustness Checking—Selectivity of the Trading Rule

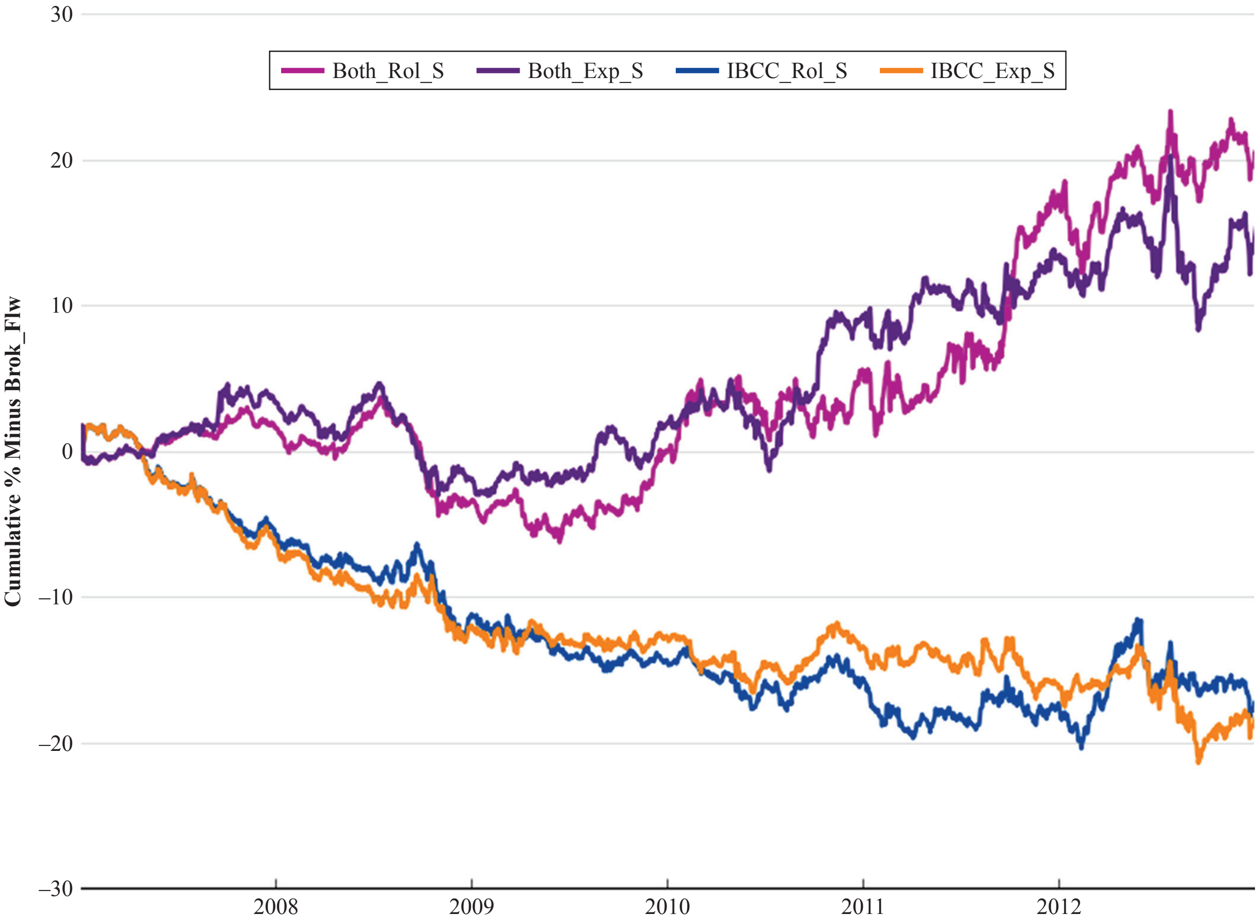
We examine the impact of changing the selectivity of the trading rule so that only recommendations with progressively higher levels of conviction produce trades. The IBCC procedure remains identical to that used before; the only changes are to the values of the parameters c and k within the decision rule. This also provides a principled way to control the number of open positions. Recall that c may be interpreted as a threshold on the information content needed within the observed constellation of broker recommendations to generate a trade. In contrast, $k > 1$ raises the threshold required for HPP decision making to produce a Go_Long (Go-Short) outcome; simply being the largest value of q_0, q_1, and q_2 is no longer sufficient.

We examine the impact of varying c and k separately; for space considerations, we examine only the long-only portfolios. Exhibit 11 shows the results of varying c while holding $k = 1$, and Exhibit 12 shows the results of varying k while holding $c = 1$. The results for varying c while holding $k = 1$ suggest some strengthening of both the alpha and beta as c is raised, in particular for the Both_Exp_L results. The results for varying k while holding $c = 1$ show a milder effect. As might be expected, we observed that the turnover increases as the decision rules become more selective, although the effect is mild compared to the baseline $c = k = 1$ case.

[23] We do not expect bottom-up broker recommendations to yield effective market-timing portfolios, so we do not explore the long–short case here for brevity.

EXHIBIT 6

Performance of the Short-Only Models Relative to the Brok_Flw_S Benchmark



Note: The portfolios labeled Both provide the best performance, as was the case for the long-only and long–short portfolios.

EXHIBIT 7

Size Tilts for the Different Portfolios

Model	Mega Cap >$25 Billion	Large Cap $10 Billion to $25 Billion	Mid Cap $2 Billion to $10 Billion	Small Cap $250 Million to $2 Billion	Micro Cap <$250 Million	Missing Data
Brok_Flw_L	23.1	20.8	48.7	7.2	0.0	0.0
IBCC_Rol_L	17.2	19.9	53.0	9.9	0.0	0.0
IBCC_Exp_L	16.3	19.8	53.8	9.9	0.0	0.0
Both_Rol_L	18.3	19.3	53.2	9.2	0.0	0.0
Both_Exp_L	17.5	19.3	54.0	9.2	0.0	0.0
Brok_Flw_S	15.9	20.6	52.5	10.9	0.0	0.1
IBCC_Rol_S	23.5	19.7	47.7	8.8	0.0	0.0
IBCC_Exp_S	22.8	19.6	48.7	8.6	0.1	0.1
Both_Rol_S	14.8	19.6	53.4	11.7	0.0	0.1
Both_Exp_S	14.2	19.9	54.2	11.2	0.1	0.1

Note: Exhibit shows the sum of absolute positions by market capitalization bucket, averaged across time.

EXHIBIT 8

Number of Recommendations after Bisecting the Universe of Stocks by Market Capitalization

Universe	Number of Recommendations	As a Percentage
Large Half	58,466	56%
Small Half	45,316	44%

Robustness Checking—Sensitivity to Truth Threshold

A threshold of 5% was used in the truth definition given by

$$t = \begin{cases} 0, & \text{if } r_{(s,\Delta\tau)} \le -5\% \times RVol_s, \\ 2, & \text{if } r_{(s,\Delta\tau)} \ge 5\% \times RVol_s, \\ 1, & \text{otherwise} \end{cases}$$

This value has been used throughout. Here we explore varying this parameter between 1% and 10%, keeping everything else the same. Results are summarized in Exhibit 13 for the long-only and short-only portfolios, where for brevity we quote results only for the Both portfolios. Unreported results show that IBCC_* consistently underperforms Brok_Flw and Both_*, consistent with our previous findings.

We find that

- As before, the Both portfolios outperform the relevant Brok_Flw_* benchmark at all threshold settings for both long-only and short-only.
- There seems to be a sweet spot for thresholds within the 4%–6% range for the long-only portfolios, particularly in terms of the *t*-statistic for alpha, which broadly measures the consistency of the outperformance.
- In the case of short-only portfolios, a tighter threshold of around 2%–3% gives slightly better results, although nothing obtains statistical significance. One possible explanation is that the smaller number of short recommendations leads to greater sampling error in assessing a broker's short efficacy, and allocating more Price_Down truths may mitigate this.

Robustness Checking—Sensitivity to Holding Period

Here we explore the sensitivity to the arbitrary 60-day holding period that has been used throughout. For brevity, we quote results for just the long-only portfolios.

From Exhibit 14 it is reasonably clear that

- Shorter holding periods give stronger performance.
- Shorter holding periods increase turnover.
- The Both portfolios again are the strongest performers for all horizons.
- Pure IBCC underperforms the Brok_Flw_L benchmark.

MACHINE LEARNING IN ACTION

An unhelpful aspect of machine learning systems is their reputation for being *black boxes* that users cannot understand. Whether or not one subscribes to this point of view, it is important to have easily interpreted diagnostic tools available that allow inspection of the model's internal components, especially as these evolve through time. In what follows, we provide two such tools.

Broker-Level Diagnostics

The first is an animated visualization that displays the evolution of a broker's recommendation distributions[24] conditional on each truth $t = 0, 1, 2$. These distributions are precisely what the system has learned about that broker's recommendation behavior up to each evaluation date. A snapshot of the animation for one broker (the one with broker code IBES_207) is given in Exhibit 15; the full animated version, which depicts the evolution of these distributions for four different brokers (IBES_199, IBES_207, IBES_410 and IBES_1296), is available online.[25]

[24] Each conditional recommendation distribution is actually four-dimensional, not three-dimensional. In each case, we have marginalized over the label corresponding to Missing to obtain a three-dimensional distribution. It is these that we have plotted as triangular heatmaps.

[25] https://faculty.fuqua.duke.edu/~charvey/JFDS_2018/IBCC_Animation.mpeg.

Exhibit 9
IBCC Results with the Stock Universe Split into Two by Market Capitalization

Side	Size	Simulation Name	Return Mean	Vol	Alpha	Alpha *t*-Stat	Beta	Turnover
Long-Only	Large Half	Brok_Flw_L	5.95	23.93	5.96	3.74	1.01	5.79
		IBCC_Rol_L	6.96	24.57	7.22	3.01	1.01	5.75
		IBCC_Exp_L	7.70	24.50	7.98	3.44	1.01	5.71
		Both_Rol_L	7.77	24.75	7.98	3.14	1.01	6.40
		Both_Exp_L	8.11	24.51	8.38	3.43	1.00	6.32
	Small Half	Brok_Flw_L	4.70	25.60	4.76	1.63	1.03	6.00
		IBCC_Rol_L	4.15	25.98	4.54	1.73	1.06	6.08
		IBCC_Exp_L	2.42	26.08	2.76	1.06	1.07	6.03
		Both_Rol_L	5.89	25.61	6.30	2.13	1.03	6.33
		Both_Exp_L	4.66	25.78	4.98	1.72	1.04	6.27
Short-Only	Large Half	Brok_Flw_S	–2.43	24.81	–2.33	–1.08	–1.03	6.55
		IBCC_Rol_S	–5.69	25.70	–5.30	–2.93	–1.08	6.37
		IBCC_Exp_S	–5.22	25.54	–5.03	–2.46	–1.07	6.41
		Both_Rol_S	–2.53	27.90	–2.13	–0.61	–1.11	7.60
		Both_Exp_S	4.35	28.30	4.59	1.22	–1.12	7.58
	Small Half	Brok_Flw_S	1.98	27.23	2.54	0.77	–1.09	6.51
		IBCC_Rol_S	0.08	26.24	0.23	0.07	–1.05	6.54
		IBCC_Exp_S	–2.68	26.14	–2.13	–0.68	–1.05	6.51
		Both_Rol_S	8.68	27.32	9.49	2.22	–1.02	7.25
		Both_Exp_S	3.23	27.41	4.63	1.02	–1.01	7.16

Note: The alpha values are annualized.

Exhibit 10
Distribution of Market Capitalization after Bisecting the Universe

Universe	Mktcap Bucket	Sum Position (%)	Number of Stocks	Return (% p.a.)	Risk (% p.a.)
Large half	Mega Cap (>US$25 billion)	41.5	75.2	4.2	6.7
	Large Cap (US$10 billion to US$25 billion)	33.4	70.6	3.1	6.1
	Mid Cap (US$2 billion to US$10 billion)	22.2	53.4	–0.3	6.9
	Small Cap (US$250 million to US$2 billion)	2.8	7.8	–1.4	1.9
Small half	Micro Cap (<US$250 million)	0.0	0.1	0.0	0.1
	Mega Cap (>US$25 billion)	0.0	0.1	0.0	0.0
	Large Cap (US$10 billion to US$25 billion)	5.1	9.9	1.7	0.9
	Mid Cap (US$2 billion to US$10 billion)	80.2	149.8	6.8	15.2
	Small Cap (US$250 million to US$2 billion)	14.5	30.7	–3.8	8.5
	Micro Cap (<US$250 million)	0.1	0.2	–0.2	0.3

Notes: Here the positions for Brok_Flw_L are summarized. The numbers shown in this exhibit are time series averages 2007–2012.

The vertices of the triangles represent the three different recommendations hold, buy (here labeled Go_L), and sell (here, Go_S). Each point within a triangle corresponds to a three-vector of probabilities over these recommendations, with the color of each point depicting its posterior probability. If the three heatmaps in Exhibit 15 were identical, then knowledge of that broker's recommendation would impart no information

Exhibit 11

Varying *c* to Change the Conviction Level Needed to Initiate a Trade for the Long-Only Models for the Period 2007–2012

					Alpha		Beta	
Model	***c***	**Mean**	**Vol**	**Alpha**	***t*-Stat**	**Beta**	***t*-Stat**	**Turnover**
Brok_Flw_L	–	5.43	24.18	5.47	2.73	1.01	26.97	5.75
IBCC_Exp_L	1.0	5.30	24.89	5.39	2.27	1.03	23.63	5.68
	1.1	5.13	25.38	5.43	2.06	1.04	20.67	5.81
	1.2	4.77	26.31	5.42	1.82	1.06	17.78	5.96
	1.3	6.46	26.23	6.49	2.04	1.05	16.58	6.06
	1.4	6.37	26.86	6.15	1.78	1.06	14.70	6.17
	1.5	5.89	27.34	5.97	1.59	1.06	13.13	6.33
IBCC_Rol_L	1.0	4.77	24.66	4.91	2.09	1.02	23.36	5.74
	1.1	4.82	25.14	5.12	1.98	1.03	20.62	5.83
	1.2	4.66	25.60	5.33	1.90	1.04	18.41	5.97
	1.3	5.14	26.04	5.20	1.74	1.05	16.95	5.95
	1.4	5.21	26.48	5.18	1.63	1.06	16.03	6.18
	1.5	5.55	27.01	5.56	1.61	1.07	14.56	6.38
Both_Exp_L	1.0	7.13	24.71	7.28	2.84	1.01	20.47	6.07
	1.1	7.10	25.15	7.59	2.67	1.01	18.17	6.23
	1.2	6.85	26.36	7.22	2.18	1.04	15.14	6.33
	1.3	8.73	26.67	8.90	2.50	1.04	14.26	6.49
	1.4	8.33	27.63	8.40	2.14	1.06	12.99	6.56
	1.5	8.52	28.15	8.85	2.10	1.07	11.85	6.73
Both_Rol_L	1.0	6.99	24.51	7.06	2.75	1.00	20.18	6.13
	1.1	6.98	24.86	7.40	2.63	1.00	17.93	6.28
	1.2	7.12	25.44	7.48	2.37	1.01	15.31	6.38
	1.3	7.36	26.14	7.54	2.18	1.03	13.81	6.37
	1.4	6.34	27.09	6.56	1.74	1.05	12.89	6.58
	1.5	6.97	27.76	7.12	1.76	1.06	11.91	6.78

Notes: Results are based on varying c while holding $k = 1$ in the decision rule. Recommendations were aggregated within the usual 30-day window when combining brokers. The alpha values are annualized.

about the truth outcome. In this exhibit, the three heatmaps are not identical, but the differences are subtle. This broker also displays the typical broker characteristic of having a low probability of issuing sell (Go_S) recommendations whatever the observed truth outcome.

Stock-Level Diagnostics

The focus of the previous section was visualizing what the model learns about a particular broker from the ensemble of their recommendations across a multiplicity of stocks. Here we fix our attention on a particular stock and visualize information from the multiplicity of brokers that make recommendations on that stock.

Exhibit 16 shows our visual diagnostic for the stock with identifier AST14822 (an internal code that is unimportant). The top-panel shows the time series of recommendations for the five most prolific brokers that comment on that stock; the green and red symbols represent buy and sell, respectively, and the black symbol represents hold (labeled here as *filtered*, equivalently). The second panel lists the same information but is more cluttered because it now includes the recommendations of all brokers commenting on that stock. The third panel shows the actions that result from the predicted truths, obtained using our rolling out-of-sample process with the $c = k = 1$ case of the decision rule discussed previously. No predictions are made during the initial three-year in-sample period, so the panel is blank at the start. The fourth panel shows the positions obtained from these actions for the Both portfolios in the expanding-window long-only and long–short cases, together with

EXHIBIT 12

Varying *k* to Change the Conviction Level Needed to Initiate a Trade for the Long-Only Models for the Period 2007–2012

Model	*k*	Mean	Vol	Alpha	Alpha *t*-Stat	Beta	Beta *t*-Stat	Turnover
Brok_Flw_L	–	5.43	24.18	5.47	2.73	1.01	26.97	5.75
IBCC_Exp_L	1.0	5.30	24.89	5.39	2.27	1.03	23.63	5.68
	1.1	5.11	24.96	5.22	2.11	1.03	22.25	5.73
	1.2	5.22	25.20	5.40	2.07	1.03	20.90	5.81
	1.3	5.45	25.78	5.44	1.91	1.04	18.22	5.92
	1.4	5.30	26.25	5.22	1.74	1.06	17.36	6.16
	1.5	5.50	26.51	5.47	1.72	1.06	16.30	6.10
IBCC_Rol_L	1.0	4.77	24.66	4.91	2.09	1.02	23.36	5.74
	1.1	4.88	24.70	5.05	2.07	1.02	22.00	5.83
	1.2	4.83	24.84	5.06	1.97	1.02	20.19	5.85
	1.3	5.24	25.21	5.28	1.94	1.03	18.80	5.97
	1.4	5.23	25.57	5.21	1.83	1.04	17.73	6.12
	1.5	5.37	25.73	5.46	1.83	1.04	16.95	6.05
Both_Exp_L	1.0	7.13	24.71	7.28	2.84	1.01	20.47	6.07
	1.1	7.02	24.78	7.19	2.70	1.01	19.38	6.09
	1.2	7.24	24.89	7.48	2.69	1.01	18.64	6.17
	1.3	7.44	25.65	7.60	2.48	1.03	16.61	6.34
	1.4	7.21	26.30	7.27	2.21	1.04	15.56	6.42
	1.5	7.62	26.67	7.72	2.23	1.05	14.85	6.47
Both_Rol_L	1.0	6.99	24.51	7.06	2.75	1.00	20.18	6.13
	1.1	7.22	24.60	7.32	2.78	1.00	19.44	6.20
	1.2	7.07	24.69	7.29	2.62	1.00	18.06	6.25
	1.3	7.21	24.95	7.40	2.50	1.00	16.94	6.34
	1.4	6.35	25.36	6.48	2.06	1.01	15.54	6.37
	1.5	6.86	25.61	7.03	2.10	1.01	14.65	6.45

Notes: Results are based on varying k while holding c = 1 in the decision rule. Recommendations were aggregated within the usual 30-day window when combining brokers. The alpha values are annualized.

various Brok_Flw_* benchmarks. The final panel shows the truth (target) outcomes for each recommendation.

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

We have demonstrated a computationally efficient practical approach for combining analysts' forecasts using a probabilistic machine learning model called IBCC combining it with a state-of-the-art approximate inference technique called VB. Throughout our results, the best outcomes were obtained when there was agreement between the broker recommendations and the machine learning–based forecasts obtained using IBCC. These findings echo important current research in the area of human–computer interaction, where decision making based on inputs from artificial intelligence and other sources is used to assist human decision making. It also suggests some intriguing research directions for enhancing the investment processes and performance of both quantitative and discretionary fund managers.

An important advantage of the IBCC model is its scalability compared to other multivariate dependence techniques (e.g., copula models). Our application integrated recommendations from 347 brokers; however, IBCC has been successfully used in applications involving many thousands of individual classifiers, so there is ample scope for extension. For example, we could look at individual analysts, or more refined groups of analysts, rather than brokers.[26] In addition, it may

[26] We are unable to report on this because of current restrictions.

EXHIBIT 13

Varying the Truth Boundary Parameter over the Period 2007–2012 for Long-Only and Short-Only Portfolios

Model	Truth (%)	Mean	Vol	Alpha	Alpha *t*-Stat	Beta	Beta *t*-Stat	Turnover
Brok_Flw_L	–	5.43	24.18	5.47	2.73	1.01	26.97	5.75
Both_Exp_L	1	6.07	24.70	6.04	2.55	1.02	21.53	6.00
	2	6.09	24.72	6.24	2.59	1.02	21.34	6.03
	3	6.39	24.65	6.56	2.68	1.01	21.24	6.00
	4	6.57	24.63	6.70	2.68	1.01	20.93	6.03
	5	7.13	24.71	7.28	2.84	1.01	20.47	6.07
	6	7.61	24.86	7.81	2.90	1.01	19.82	6.25
	8	5.94	26.31	6.24	1.86	1.04	15.25	6.56
	10	6.66	26.95	6.74	1.84	1.05	14.93	6.77
Both_Rol_L	1	6.29	24.39	6.30	2.66	1.00	21.82	6.00
	2	6.10	24.39	6.26	2.62	1.00	21.41	6.03
	3	6.10	24.42	6.26	2.59	1.00	21.34	6.00
	4	6.32	24.43	6.46	2.61	1.00	20.69	6.05
	5	6.99	24.51	7.06	2.75	1.00	20.18	6.13
	6	6.97	24.60	7.29	2.75	1.00	19.98	6.29
	8	7.17	26.19	7.95	2.33	1.03	14.78	6.67
	10	7.74	27.00	7.59	2.13	1.06	15.09	6.87
Brok_Flw_S	–	–0.51	24.96	–0.13	–0.05	–1.03	–29.42	6.38
Both_Exp_S	1	2.10	25.76	2.31	0.83	–1.05	–26.19	6.91
	2	3.36	25.56	3.55	1.27	–1.04	–26.51	6.94
	3	3.13	25.45	3.38	1.23	–1.04	–26.27	6.97
	4	2.17	25.53	2.35	0.81	–1.03	–25.31	7.00
	5	2.11	25.71	2.46	0.79	–1.03	–25.39	7.04
	6	4.09	26.51	4.02	1.05	–1.02	–19.30	7.20
	8	0.67	28.84	1.06	0.21	–1.04	–15.77	7.50
	10	–1.15	31.44	–1.28	–0.22	–1.09	–14.66	7.76
Both_Rol_S	1	2.73	25.68	2.92	1.09	–1.05	–25.52	6.79
	2	2.75	25.43	2.93	1.08	–1.04	–24.57	6.78
	3	3.28	25.51	3.49	1.27	–1.04	–24.65	6.84
	4	1.80	25.69	1.98	0.69	–1.04	–24.71	6.99
	5	2.99	25.98	3.45	1.06	–1.03	–21.81	7.12
	6	2.71	26.76	3.07	0.88	–1.05	–23.33	7.23
	8	0.36	27.63	–0.24	–0.06	–1.05	–20.90	7.50
	10	–1.18	29.56	–1.81	–0.36	–1.08	–20.01	7.62

Notes: Unreported results show that IBCC_* consistently underperforms Brok_Flw and Both_*, consistent with our previous findings. The alpha values are annualized.

be useful to combine the recommendation data examined here with categorical sentiment measures extracted using a range of different natural language interpreters on both mainstream and financial news sources. There is scope to obtain an order of magnitude more classifiers. The computational efficiency of our implementation would enable such data to be handled without issue and real time forecasting to be undertaken.

Although the VB implementation of the IBCC holds promise, it also has limitations. In the Galaxy Zoo experiment that we used to motivate the research application, several distinct issues make the application to analysts different from the application to astronomers. First, it is reasonable to assume that the astronomers are operating independently (not collaborating with each other). However, it is likely that analysts are aware of

EXHIBIT 14

Varying the Holding Period for the Long-Only Models for Period 2007–2012

Model	Holding Period	Mean	Vol	Alpha	Alpha *t*-Stat	Beta	Beta *t*-Stat	Turnover
Brok_Flw_L	10	10.04	23.99	10.01	4.55	0.99	29.08	28.20
	20	7.39	24.16	7.27	3.51	1.01	28.17	14.89
	30	6.82	24.33	6.71	3.23	1.02	26.99	10.35
	45	6.05	24.27	6.01	2.98	1.01	26.48	7.31
	60	5.43	24.18	5.47	2.73	1.01	26.97	5.75
	90	5.46	23.94	5.32	2.75	1.00	29.61	4.31
IBCC_Exp_L	10	7.09	24.01	7.08	2.92	0.98	25.66	28.73
	20	5.93	24.81	5.76	2.29	1.02	21.80	15.19
	30	5.15	24.77	5.18	2.13	1.02	21.93	10.55
	45	5.61	24.79	5.76	2.43	1.02	23.20	7.29
	60	5.30	24.89	5.39	2.27	1.03	23.63	5.68
	90	4.95	24.62	5.10	2.23	1.02	26.34	4.20
IBCC_Rol_L	10	8.02	24.01	7.84	3.15	0.98	22.83	28.76
	20	5.75	24.76	5.64	2.22	1.01	21.09	15.24
	30	6.13	24.55	6.20	2.56	1.01	22.49	10.53
	45	5.60	24.46	5.79	2.52	1.01	24.75	7.31
	60	4.77	24.66	4.91	2.09	1.02	23.36	5.74
	90	5.12	24.58	5.28	2.31	1.02	25.71	4.21
Both_Exp_L	10	14.05	23.97	14.21	5.51	0.97	25.38	30.18
	20	9.76	24.75	9.80	3.76	1.01	20.78	15.99
	30	8.42	24.79	8.24	3.18	1.01	20.87	11.03
	45	7.98	24.91	7.93	3.04	1.02	20.73	7.74
	60	7.13	24.71	7.28	2.84	1.01	20.47	6.07
	90	6.63	24.39	6.65	2.66	1.00	22.05	4.71
Both_Rol_L	10	13.38	23.87	13.36	5.00	0.96	22.10	30.28
	20	9.44	24.35	9.62	3.63	0.99	20.48	16.06
	30	9.27	24.66	9.11	3.41	1.00	21.07	11.03
	45	7.71	24.51	7.75	3.07	1.00	21.43	7.73
	60	6.99	24.51	7.06	2.75	1.00	20.18	6.13
	90	6.48	24.27	6.61	2.66	0.99	21.83	4.67

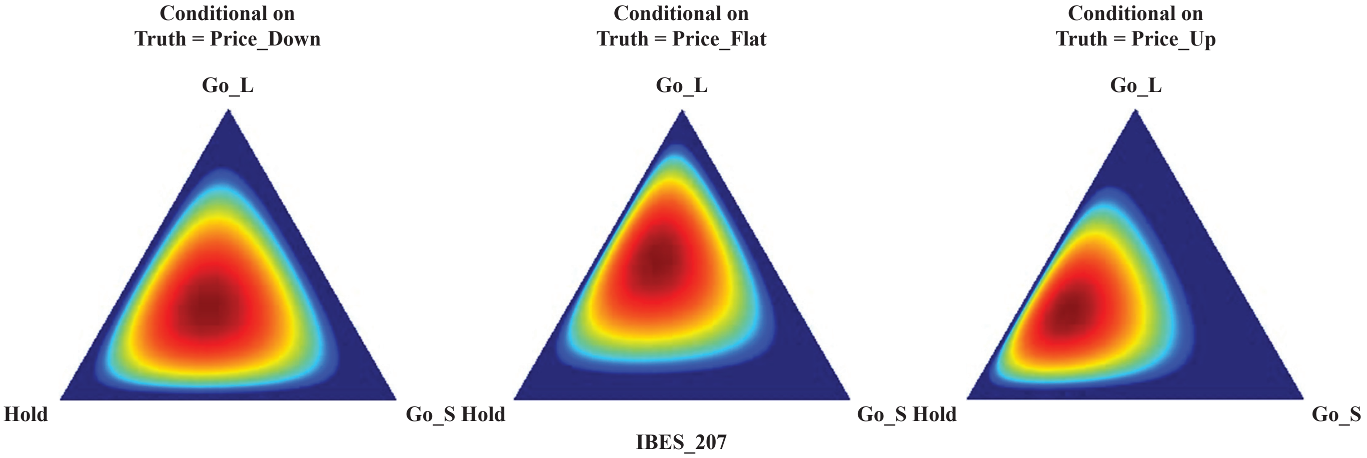
Notes: Here the trade holding period and the holding period for assessing truths are constrained to be equal. The ±5% threshold for converting stock returns to truths is scaled to yield a similar number of truths for each horizon, using the usual random walk property that $\sigma(X_t) \propto \sqrt{t}$*, which gives threshold =* $\sqrt{t / t_0} \times 5\% = \sqrt{holding\ \ period/60} \times 5\%$*. As elsewhere, recommendations are aggregated with a lookback of up to 30 days when combining brokers. The alpha values are annualized.*

other analysts' forecasts—and this could affect their forecasts. Second, the quality of all analysts' forecasts could be affected by common events such as a recession, global sentiment, and common market factors that may affect their sector or region. Such common factors do not apply in the Galaxy Zoo experiment.

There are also areas for methodological consideration within the current implementation. For example, the IBCC model has no concept of ordering within the truth outcomes or the recommendations; they are simply sets of categorical labels. Perhaps more importantly, IBCC has no concept of parity between the recommendations and truths. Maybe it is therefore only to be expected that our strongest results arose when we looked for reinforcement between the raw broker recommendations and the IBCC predictions. Changing the model to incorporate some parity effect would make it less general but would likely boost performance in our application. On the other hand, if sufficient data were available to learn the parity relationship with the

Exhibit 15

Screenshot of the Website Animation



Conditional on Truth = Price_Down

Conditional on Truth = Price_Flat

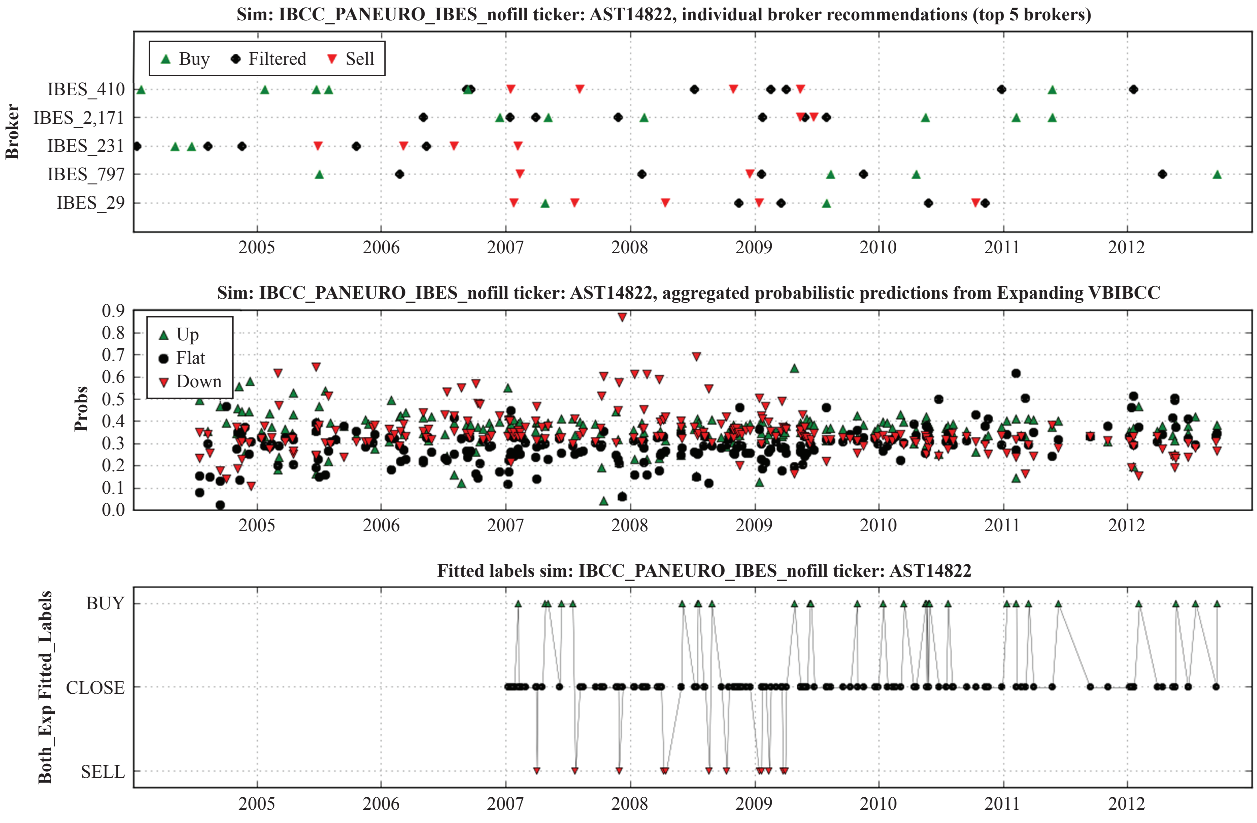
Conditional on Truth = Price_Up

IBES_207

Notes: This exhibit shows the evolving recommendation distributions for the broker with identifier IBES_207, conditional on truth=Price_Down (left), truth=Price_Flat (middle), and truth=Price_Up (right). Within each triangle, each pixel represents a three-vector of probabilities over the recommendations hold, buy (Go_L), and sell (Go_S). The color of each pixel represents the posterior probability of this corresponding three-vector; blue pixels have very low posterior probability, and dark red pixels have the highest posterior probability.

Exhibit 16

Diagnostic Panel for the Stock with Identifier AST14822



Sim: IBCC_PANEURO_IBES_nofill ticker: AST14822, individual broker recommendations (top 5 brokers)

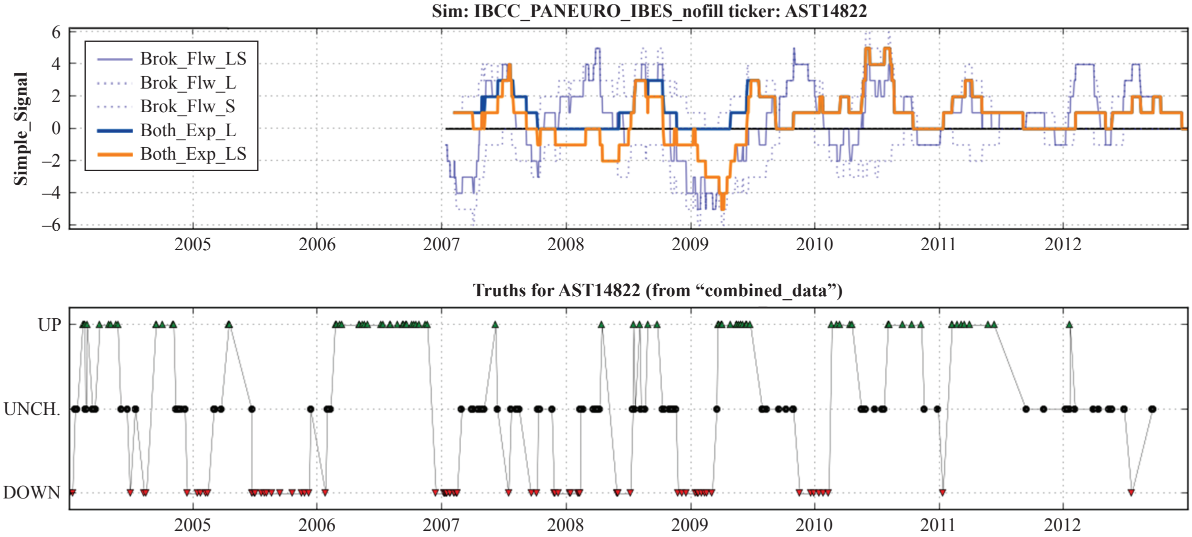
Sim: IBCC_PANEURO_IBES_nofill ticker: AST14822, aggregated probabilistic predictions from Expanding VBIBCC

Fitted labels sim: IBCC_PANEURO_IBES_nofill ticker: AST14822

(continued)

Exhibit 16 *(continued)*

Diagnostic Panel for the Stock with Identifier AST14822



Notes: The top panel shows recommendations for the five most prolific brokers that comment on the stock; green and red represent buy and sell, respectively, and the black symbol represents hold (labeled here as filtered, equivalently). The second panel lists the same information and now includes the recommendations of all brokers commenting on that stock. The third panel shows the actions that result from the predicted truths, obtained using our rolling out-of-sample process with the c = k = 1 case of the decision rule. No predictions are made during the initial three-year in-sample period. The fourth panel shows the positions obtained from these actions for the Both portfolios in the expanding-window long-only and long–short cases, together with various Brok_Flw_* benchmarks. The final panel shows the truth (target) outcomes for each recommendation.

original IBCC model, then there would be no issue. Our practical experience is that there are never sufficient data available compared to what we would like, so working with flexible but not completely general models gives the best results.

ACKNOWLEDGMENT

We thank Edwin Simpson for supporting discussions.

REFERENCES

Bernardo, J. M., and A. F. M. Smith. *Bayesian Theory*. Hoboken: Wiley, 1994.

Bishop, C. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe. 2018. "Variational Inference: A Review for Statisticians." arXiv:1601.00670v9.

Boyd, S., and L. Vandenberghe. *Convex Optimization*. Cambridge: Cambridge University Press, 2004.

Bradshaw, M. T. 2011. "Analysts' Forecasts: What Do We Know after Decades of Work?" SSRN, June 30, https://ssrn.com/abstract=1880339.

Brown, L. 1993. "Earnings Forecasting Research: Its Implications for Capital Markets Research." *International Journal of Forecasting* 9: 295–320.

——, ed. 2000. *I/B/E/S Research Bibliography*. 6th ed. New York: I/B/E/S International Inc., 2000.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B (Methodological)* 39 (1): 1–38.

Fox, C. W., and S. J. Roberts. 2011. "A Tutorial on Variational Bayesian Inference." *Artificial Intelligence Review* 38 (2): 85–95.

Ghahramani, Z., and H. C. Kim. 2003. "Bayesian Classifier Combination." Gatsby Computational Neuroscience Unit technical report no. GCNU-T. London, UK.

Givoly, D., and J. Lakonishok. 1984. "Properties of Analysts' Forecasts of Earnings: A Review and Analysis of the Research." *Journal of Accounting Literature* 3: 117–152.

Kim, H. C., and Z. Ghahramani. 2012. "Bayesian Classifier Combination." Proceedings of the 15th AISTATS Conference.

Lee, P. M. *Bayesian Statistics: An Introduction*. Chichester, UK: John Wiley, 2012.

Levenberg, A., S. Pulman, K. Moilanen, E. Simpson, and S. Roberts. 2014. "Predicting Economic Indicators from Web Text Using Sentiment Composition." *International Journal of Computer and Communication Engineering* 3 (2): 109–115.

Levenberg, A., E. Simpson, S. Roberts, and G. Gottlob. "Economic Prediction Using Heterogeneous Data Streams from the World Wide Web." In *Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE), Proceedings of ECML/PKDD Workshop.* New York: Springer, 2013.

Lintott, C. 2012. "I, for One, Welcome Our New Machine Collaborators." August 3, https://blog.zooniverse.org/2012/08/03/i-for-one-welcome-our-new-machine-collaborators.

Parisi, G. *Statistical Field Theory*. Boston: Addison-Wesley, 1988.

Sato, M. A. 2001. "Online Model Selection Based on the Variational Bayes." *Neural Computation* 13: 1649–1681.

Schipper, K. 1991. "Commentary on Analysts' Forecasts." *Accounting Horizons* 5 (4): 105–121.

Simpson, E., S. Roberts, I. Psorakis, and A. Smith. "Dynamic Bayesian Combination of Multiple Imperfect Classifiers." In *Decision Making and Imperfection. Intelligent Systems Reference Library Series,* Vol. 474. New York: Springer, 2013.

Tanner, M. A. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* New York: Springer, 1996.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

A Data Science Solution to the Multiple-Testing Crisis in Financial Research

MARCOS LÓPEZ DE PRADO

MARCOS LÓPEZ DE PRADO
is principal and head of machine learning at AQR Capital Management in Greenwich, CT, and an adjunct professor at Cornell University in Ithaca, NY.
marcos.lopezdeprado@aqr.com

Academics and investors often compute the performance of an investment strategy or factor to determine whether such strategy or factor profits beyond what could be considered "luck." By far the most commonly used investment performance statistic is the Sharpe ratio (SR), first introduced by Sharpe (1966) and further studied by Sharpe (1975, 1994). The probability distribution of this statistic is well known under a variety of assumptions (Lo 2002; Bailey and López de Prado 2012). Using those distributions, it is possible to derive the probability that the observed SR exceeds a given threshold. Under this framework, an investment strategy with a low SR based on a long backtest or track record may be preferred to an alternative strategy with a high SR computed on a short backtest or track record. One problem with this approach is that it does not account for *selection bias under multiple testing* (SBuMT).

In 1933, Jerzy Neyman and Egon Pearson developed the standard hypothesis test used in most scientific applications. These authors did not consider the possibility of performing multiple tests on the same dataset and selecting the most favorable outcome (the one that rejects the null with the lowest false positive probability). At that time, the absence of powerful computers made SBuMT unlikely. Bonferroni (1935) was among the first to recognize that the probability of obtaining a false positive would increase as a test is repeated multiple times over the same dataset. Ever since, statisticians have taken the problem of multiple testing seriously (Gelman and Locken 2013). In its ethical guidelines,[1] the American Statistical Association warns that "failure to disclose the full extent of tests and their results in such a case would be highly misleading" (American Statistical Association 1999).

Given this background, it is surprising to find that practically all papers in empirical finance fail to disclose the number of trials involved in a discovery. Virtually every paper reports a result as if it were the only trial attempted. This is, of course, rarely the case, and it is common for economists to conduct millions of regressions or simulations before finding a result striking enough to merit publication (Sala-i-Martin 1997; Leinweber 2007). Researchers in other fields have taken steps to control for and prevent SBuMT (e.g., visit www.alltrials.net, or see Szucs and Ioannidis 2017). Unlike physics, finance does not have laboratories in which false claims can be easily debunked based on independent tests: All we count on are the same time series used to overfit the backtest, and gathering out-of-sample evidence will take decades (López de Prado 2017).

[1] See Ethical Guideline A.8: http://community.amstat.org/ethics/aboutus/new-item.

A very common misconception is that the problem of SBuMT only affects historical simulations (backtesting). In fact, this problem encompasses any situation in which we select one outcome without controlling for the totality of alternative outcomes from which we choose. For example, a hedge fund may want to hire a portfolio manager with an SR of 2. To that purpose, the fund may interview multiple candidates, not realizing that they should adjust the SR higher with every additional interview. The fact that the SR is computed on an actual track record does not mean that SBuMT will not take place. We could interview a series of dart-throwing monkeys, and eventually we would find one with an SR of 2.

There is nothing wrong with carrying out multiple tests. Researchers should perform multiple tests and report the results of all trials; however, when the extent of the tests carried out is hidden from journal referees, readers, and investors, it is impossible for them to assess whether a particular result is a false positive (Bailey et al. 2014, 2017). For this reason, Harvey, Liu, and Zhu (2016) concluded that "most claimed research findings in financial economics are likely false."

Yet, there is hope. SBuMT can be prevented and corrected in financial economics. Nothing forbids financial researchers from joining the ranks of researchers from other fields who control for SBuMT. Accordingly, the main goal and contribution of this article is to provide a template for how the results from multiple trials could be reported in financial publications. The information regarding all trials could be disclosed in a separate section or an appendix to a publication, while the focus remains on explaining the selected finding. Ideally, the author would report the performance of a proposed investment strategy or factor adjusted for SBuMT. In this particular article we apply the deflated SR (DSR) method (Bailey and López de Prado 2014; López de Prado and Lewis 2018) to control for the effects of SBuMT, non-normality, and sample length. It is not the goal of this article to present a financial discovery or promote an investment strategy, even though the results presented in this publication correspond to an actual investment mandate.

In the following sections, we provide a template for how authors and journals could expose to referees and readers critical information concerning all trials involved in a discovery.

EXHIBIT 1

Performance Statistics for the Index and the Selected Strategy

Statistic	iBoxxIG	Strategy
Start date	1/21/2010	1/21/2010
End date	5/1/2018	5/1/2018
aRoR (Total)	4.90%	9.35%
Avg AUM (1E6)	1,000.00	1,506.43
Avg Gini	0.29	0.88
Avg Duration	7.88	0.08
Avg Default Prob	1.36%	1.58%
An. Sharpe ratio	0.99	2.00
Turnover	0.64	5.68
Effective Number	1034.87	186.26
Correl. to Ix	1.00	0.48
Drawdown (95%)	3.17%	2.89%
Time Underwater (95%)	0.23	0.20
Leverage	1.00	3.59

DISCLOSURE OF ALL TRIALS

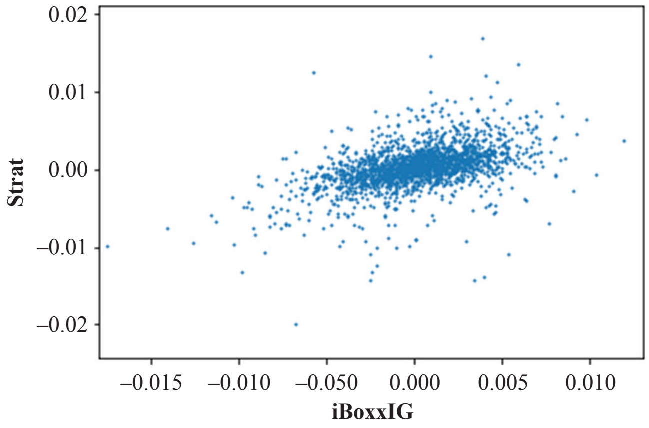
We have developed a market-neutral strategy that invests in liquid high-grade corporate bonds denominated in US dollars. The investment universe is taken from the history of constituents of the Markit iBoxx IG USD index. At each point in time, the strategy may invest in bonds included in the coetaneous index definition, so as to prevent survivorship bias and other forms of information leakage. Although the target portfolio aims at being market neutral, market frictions may prevent all intended trades from being executed. When that happens, the residual risk is hedged with bond futures.

Exhibit 1 lists some statistics associated with the selected strategy. As a reference, it also provides the same information for the index, although results from a long-only index are not directly comparable to those of a market-neutral strategy. Exhibit 2 shows a scatter plot of index returns against strategy returns. The Appendix provides a definition for each of these statistics.

Performance incorporates transaction costs and slippage, based on real transaction cost information collected for this universe over the years. An SR of 2.0 is generally considered high, because the probability of observing that SR after a single trial is infinitesimal, under the null hypothesis that the true SR is zero (see Bailey and López de Prado 2012 for the estimation of such probability).

EXHIBIT 2

Scatter Plot of iBoxx IG Returns (x-axis) against Strategy Returns (y-axis)



Other specifics about the strategy, such as the underlying principle exploited or predictive features, belong to a different discussion. As explained earlier, our key concern is to provide a template for reporting the information from all trials conducted so that journal referees and investors may evaluate the probability that the discovered strategy is a false positive as a result of SBuMT.

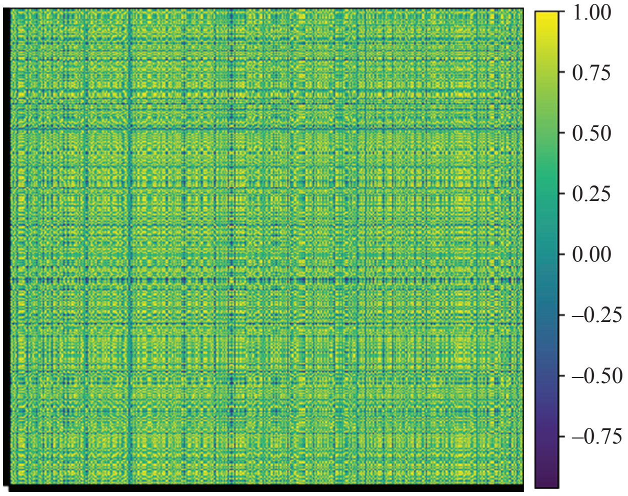
Unlike the practical totality of publications in finance, we begin by acknowledging that the results presented in Exhibits 1 and 2 are not the outcome of a single trial. Because more than one trial took place, the reader must assume that this result is the best out of many alternative ones, and therefore SBuMT is present. By disclosing the information associated with those alternative outcomes, we allow referees and investors to adjust for the inflationary effect of SBuMT.

Exhibit 3 plots the heatmap of return correlations among the 6,385 trials that have taken place before the selection of this investment strategy. This set of trials satisfies the following properties:

- **Complete**
  - The set includes every backtest computed by any of the authors for this or similar investment mandates.
  - Researchers do not have the ability to delete trials, and they are not allowed to backtest outside the official research platform.

EXHIBIT 3

Heatmap of the Correlation Matrix of the Returns of All 6,385 Trials



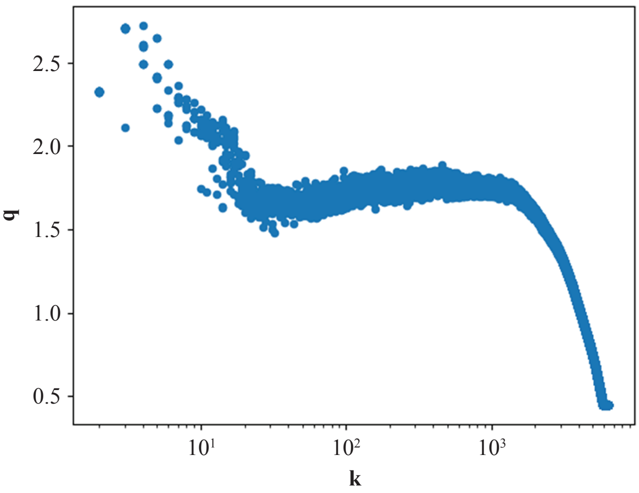
- **Coerced**
  - Researchers do not choose what to log or present. Terabytes of intermediate research metadata are automatically recorded and curated by research surveillance systems.
- **Untainted**
  - Every batch of backtests must be preapproved by the research committee to prevent that externally preselected trials contaminate the internal trials.

External trials are those that have been executed by other authors, outside the control of our research framework. They may have been preselected; hence, they are likely to be biased. To reduce the likelihood of external trials, ideally the research committee may require that trials be justified by *a priori* economic or mathematical theories (e.g., arbitrage-free pricing equations) rather than *a posteriori* empirical theories (e.g., conjectures based on empirical studies).

As is customary in machine learning applications, the main diagonal crosses the Cartesian product from the bottom left to the top right. A light color indicates that the correlation between the returns of two trials was high. The predominance of light colors suggests that the number of uncorrelated trials may be relatively low.

Exhibit 4

Quality of Clusters (y-axis) for a Varying Number of Clusters (x-axis, in logarithmic scale)



To assess whether the strategy reported in Exhibit 1 is a false investment strategy, we need to discount the inflationary effect caused by all the trials displayed in Exhibit 3. The first step is to determine the number of effectively uncorrelated clusters of trials.

CLUSTERING OF TRIALS

In this section, we apply the *optimal number of clusters* (ONC) algorithm introduced by López de Prado and Lewis (2018) to the correlation matrix plotted in Exhibit 3. Exhibit 4 plots the measure of the quality of clusters q_k that results from producing k clusters, where $k = 2, \ldots, 6385$. The quality of the clusters seems to collapse beyond $k = 1{,}000$. The higher quality levels are observed for $k < 10$, with the maximum reached by $k = 4$.

Exhibit 5 shows the clustered correlation matrices derived for $k \leq 10$. A visual inspection of these heatmaps seems to confirm that the best clustering is achieved by $k = 4$. For instance, the heatmaps for $k \geq 5$ show multiple large, off-diagonal blocks of highly correlated trials. These off-diagonal blocks appear when very similar trials belong to different (and nonconsecutive) clusters, indicating that the correlation matrix has been overclustered. In contrast, no such off-diagonal blocks can be appreciated in the heatmap for $k = 4$.

One explanation for the low number of clusters is that the researchers tried only strategy configurations that had a rigorous theoretical foundation, derived from mathematical bond pricing equations. The search region was narrowly constrained by predefined mathematical theories. The number of clusters would have been much larger, perhaps in the hundreds, if researchers had tried less mathematical (more arbitrary) configurations.

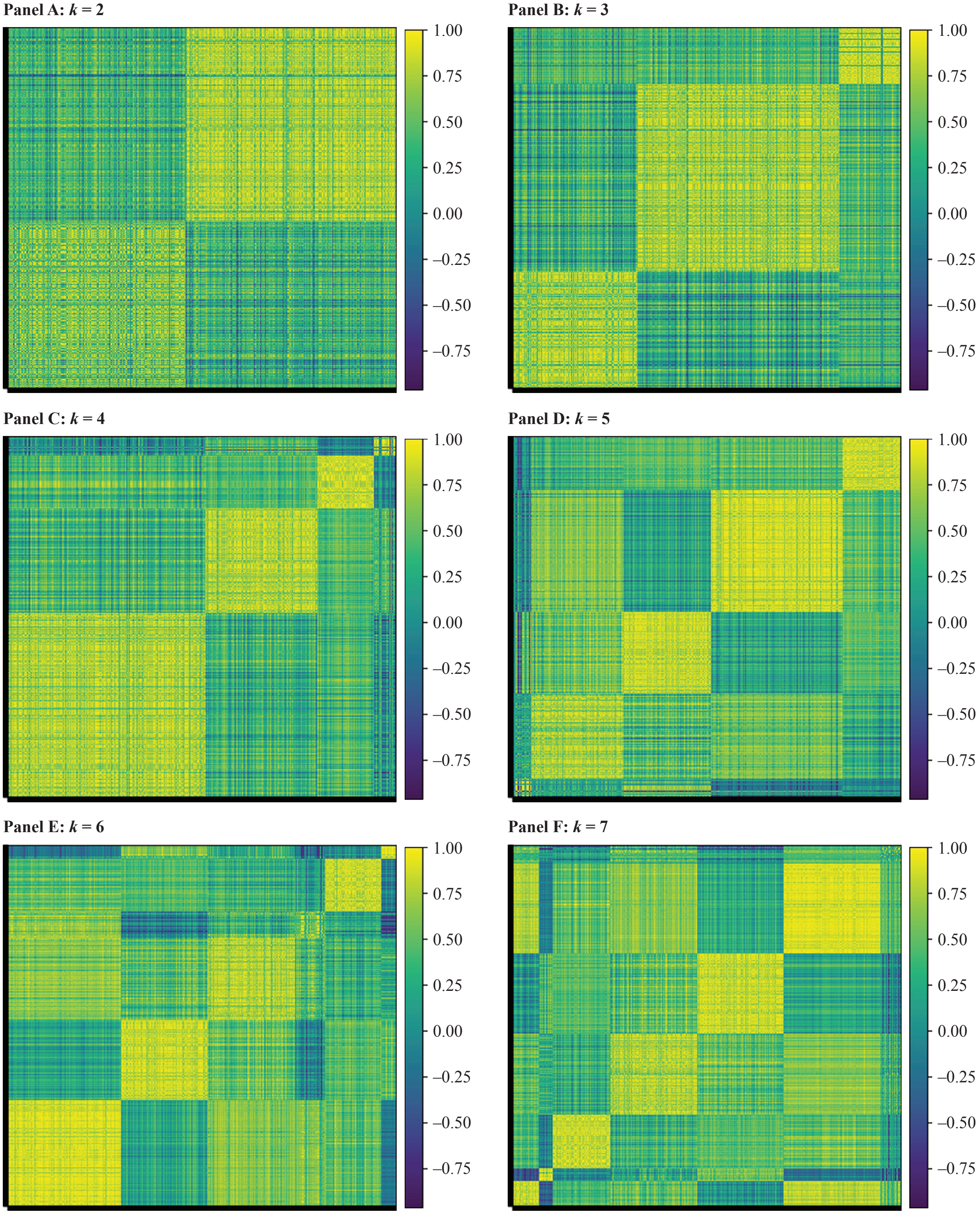
CLUSTER STATISTICS

Following López de Prado and Lewis (2018), we have computed one return series for each cluster; each cluster's composition was determined in the previous section. Forming one time series per cluster further reduces the bias caused by selecting outliers: We do not evaluate the strategy based on a single (potentially "lucky") trial, but based on a large collection of similar trials. In particular, we compute each cluster's returns applying the minimum variance allocation so that highly volatile trials do not dominate the returns time series. Otherwise, a single volatile trial might bias the time series of returns that characterize the entire cluster. Exhibit 6 reports the statistics computed on the clusters' returns series.

For each cluster, we report the following information: (1) *Strat Count* is the number of trials included in a cluster; (2) *aSR* is the annualized SR; (3) *SR* is the nonannualized SR (computed on the same sampling frequency of the original observations; in this case, daily); (4) *Skew* is the skewness of the returns (in the original frequency); (5) *Kurt* is the kurtosis of the returns (in the original frequency); (6) *T* is the number of observations in the returns series; (7) *StartDt* is the date of the first observation in the returns series; (8) *EndDt* is the date of the last observation in the returns series; (9) *Freq* is the average number of observations per year, used to annualize the SR; (10) *sqrt*(*V*[*SR_k*]) is the standard deviation of the SRs across clusters, expressed in the frequency of the cluster; (11) *E*[*max SR_k*] is the expected maximum SR, derived from the false strategy theorem; and (12) *DSR* is the deflated SR—that is, the probability that the true SR exceeds zero after controlling for SBuMT. For the cluster that contains the selected strategy, we have highlighted the *SR* and *E*[*max SR_k*] so that the reader can appreciate the inflationary effect caused by

Exhibit 5

Heatmap of the Clustered Correlation Matrix for $k = 2, \ldots, 10$



Panel A: $k = 2$

Panel B: $k = 3$

Panel C: $k = 4$

Panel D: $k = 5$

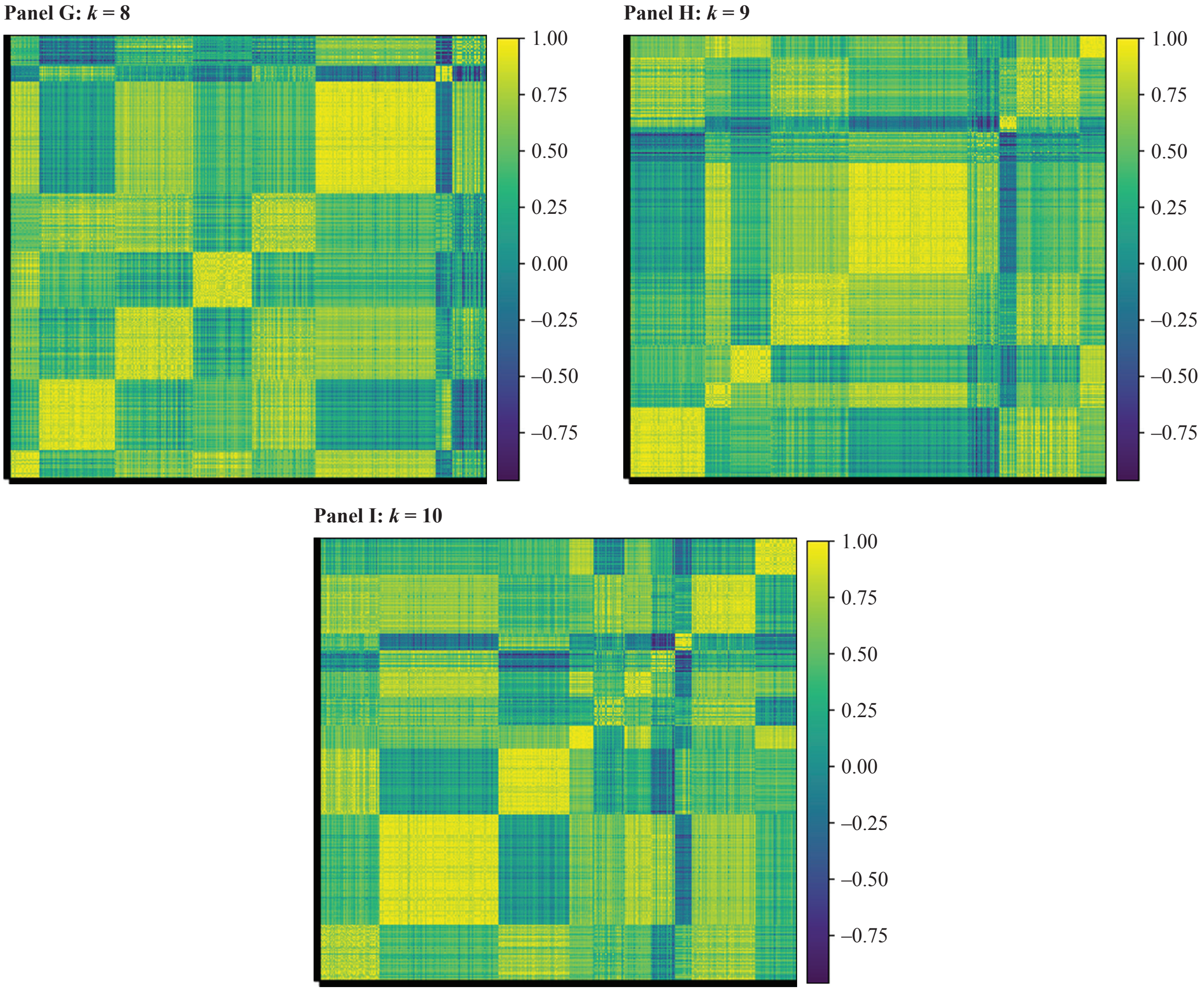
Panel E: $k = 6$

Panel F: $k = 7$

(continued)

Exhibit 5 *(continued)*

Heatmap of the Clustered Correlation Matrix for k = 2, …, 10



Panel G: k = 8

Panel H: k = 9

Panel I: k = 10

multiple testing. Also highlighted is *DSR*, which corrects for the aforementioned inflation.

Cluster 2 of Exhibit 6 contains the strategy reported in Exhibit 1. The aSR for Cluster 2 is 2.0275, in line with the aSR reported in Exhibit 1. The nonannualized SR is 0.1255, which is consistent with the aSR ($2.0275 \approx 0.1255\sqrt{261.1159}$). Given the number of clusters, and the variance of the cluster SRs, the expected maximum SR (nonannualized) is 0.027, which is significantly lower than 0.1255. Consequently, the DSR is very close to 1. Hence, the probability that the selected strategy is a false positive is virtually zero.

ROBUSTNESS OF THE FINDING

Even though the empirical evidence strongly indicates that $k = 4$ is the optimal clustering, we choose to provide full results for all k = 2, …, 10. In this way, referees and readers can evaluate the robustness of the conclusions under alternative scenarios, as unlikely as those scenarios might be. Exhibit 7 displays the cluster statistics for k = 2,3,5, …, 10, in the same format we previously used for $k = 4$. For each clustering, we have highlighted the cluster that contains the strategy reported in Exhibit 1.

EXHIBIT 6

Statistics Computed on Clusters' Returns (k = 4, q = 2.7218)

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Strat Count	3,265	1,843	930	347
aSR	1.5733	1.4907	2.0275	1.0158
SR	0.0974	0.0923	0.1255	0.0629
Skew	–0.3333	–0.4520	–0.4194	0.8058
Kurt	11.2773	6.0953	7.4035	14.2807
T	2,172	2,168	2,174	2,172
StartDt	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	05-01-2018	04-25-2018	05-03-2018	05-01-2018
Freq	261.0474	261.0821	261.1159	261.0474
sqrt(V[SR_k])	0.0257	0.0256	0.0256	0.0257
E[max SR_k]	0.0270	0.0270	0.0270	0.0270
DSR	0.9993	0.9985	**1.0000**	0.9558

Note: Results for the cluster containing the chosen strategy are shaded.

Results are robust and consistent across all the studied clusterings. The lowest DSR takes place when $k = 10$, where $DSR = 0.9995$. This DSR level is well above the common confidence levels of 0.95 or 0.975 used in most publications. In any event, this DSR corresponds to a very unlikely scenario, given the relatively low quality of the $k = 10$ clustering, compared to the quality achieved by the $k = 4$ clustering. Under these circumstances, we conclude that the strategy underlying these performance results is unlikely to be a false positive caused by SBuMT.

The reader should not infer from this analysis that the strategy will never lose money. All investments involve risk, even those with an SR that almost certainly is positive (see Exhibit 6). The purpose of this analysis was to determine whether the strategy appears to be profitable because of the inflationary effects of SBuMT. Even though the strategy is unlikely to be a false positive, no risky investment can guarantee a positive outcome.

IMPLICATIONS FOR AUTHORS, JOURNALS, AND FINANCIAL FIRMS

The research crisis that afflicts financial economics is not unsolvable. In this article we have presented a template of how this problem can be addressed in practical terms. If the publication of future discoveries could be accompanied with information regarding all the trials involved in those discoveries, financial economics would be able to overcome this crisis and reassert its credibility.

In particular, authors could (1) add to every publication an appendix explaining why the purported discovery is not a false positive caused by SBuMT; (2) certify that they have logged and recorded all the trials that took place during their research; and (3) provide to journal referees the outcomes from all trials. Journals could publish the outcomes from all trials in their websites so that researchers can evaluate the totality of the evidence, not only the trials handpicked by the authors or referees.

Journals could demand that authors (1) disclose all trials; (2) report the extent to which their findings are affected by SBuMT; and (3) evaluate the robustness of their findings to alternative scenarios of SBuMT, as shown in this article.

Financial firms could (1) avoid the practice of optimizing backtests (i.e., picking the winners while ignoring the losers); (2) implement research surveillance frameworks that record, store, and curate every single research trial that takes place within the organization; and (3) estimate the probability of a false positive, objectively controlling for SBuMT.

We believe that adopting these or similar controls for SBuMT would significantly improve the quality of financial journals.

EXHIBIT 7
Statistics Computed on Clusters' Returns

Panel A: *k* = 2, *q* = 2.3274

Stats	Cluster 0	Cluster 1
Strat Count	2,937	3,448
aSR	1.7707	1.6023
SR	0.1096	0.0992
Skew	–0.5780	–0.3351
Kurt	6.5878	11.3212
T	2,174	2,172
StartDt	01-04-2010	01-04-2010
EndDt	05-03-2018	05-01-2018
Freq	261.1159	261.0474
sqrt(V[SR_k])	0.0074	0.0074
E[max SR_k]	0.0038	0.0038
DSR	**1.0000**	1.0000

Panel B: *k* = 3, *q* = 2.7068

Stats	Cluster 0	Cluster 1	Cluster 2
Strat Count	2,063	3,329	993
aSR	1.4411	1.5780	2.0638
SR	0.0892	0.0977	0.1277
Skew	–0.4310	–0.3357	–0.4137
Kurt	5.8606	11.2267	7.3681
T	2,170	2,172	2,174
StartDt	01-04-2010	01-04-2010	01-04-2010
EndDt	04-27-2018	05-01-2018	05-03-2018
Freq	261.1507	261.0474	261.1159
sqrt(V[SR_k])	0.0202	0.0203	0.0202
E[max SR_k]	0.0173	0.0173	0.0173
DSR	0.9995	0.9999	**1.0000**

Panel C: *k* = 5, *q* = 2.6517

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Strat Count	317	1,524	1,434	2,169	941
aSR	0.9690	1.4664	1.4065	1.5272	2.0319
SR	0.0600	0.0907	0.0870	0.0945	0.1257
Skew	2.2161	–0.3286	–0.4864	–0.4086	–0.4172
Kurt	41.2726	9.7988	5.4162	12.1809	7.4552
T	2,172	2,170	2,168	2,172	2,174
StartDt	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	05-01-2018	04-27-2018	04-25-2018	05-01-2018	05-03-2018
Freq	261.0474	261.1507	261.0821	261.0474	261.1159
sqrt(V[SR_k])	0.0234	0.0234	0.0234	0.0234	0.0234
E[max SR_k]	0.0279	0.0279	0.0279	0.0279	0.0279
DSR	0.9418	0.9979	0.9964	0.9987	**1.0000**

(continued)

EXHIBIT 7 *(continued)*

Statistics Computed on Clusters' Returns

Panel D: *k* = 6, *q* = 2.4919

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Strat Count	1,873	1,418	1,447	476	935	236
aSR	1.5205	1.4034	1.4580	1.3853	2.0296	0.4322
SR	0.0941	0.0869	0.0902	0.0857	0.1256	0.0267
Skew	–0.4254	–0.4872	–0.3458	0.5432	–0.4188	0.1344
Kurt	13.0185	5.4077	9.9281	16.1401	7.4308	5.6976
T	2,170	2,168	2,170	2,172	2,174	2,170
StartDt	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	04-27-2018	04-25-2018	04-27-2018	05-01-2018	05-03-2018	04-27-2018
Freq	261.1507	261.0821	261.1507	261.0474	261.1159	261.1507
sqrt(V[SR_k])	0.0321	0.0321	0.0321	0.0321	0.0321	0.0321
E[max SR_k]	0.0417	0.0418	0.0417	0.0418	0.0417	0.0417
DSR	0.9909	0.9797	0.9862	0.9807	**0.9999**	0.2421

Panel E: *k* = 7, *q* = 2.3650

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Strat Count	443	232	940	1,436	1,418	1,591	325
aSR	1.4985	0.4229	2.0314	1.4566	1.4034	1.4816	1.2380
SR	0.0927	0.0262	0.1257	0.0901	0.0869	0.0917	0.0766
Skew	–0.4098	0.1355	–0.4174	–0.3447	–0.4872	–0.4488	10.2898
Kurt	10.4565	5.6820	7.4499	9.9064	5.4077	13.8743	295.3934
T	2,170	2,170	2,174	2,169	2,168	2,170	2,172
StartDt	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	04-27-2018	04-27-2018	05-03-2018	04-26-2018	04-25-2018	04-27-2018	05-01-2018
Freq	261.1507	261.1507	261.1159	261.1164	261.0821	261.1507	261.0474
sqrt(V[SR_k])	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298
E[max SR_k]	0.0413	0.0413	0.0413	0.0413	0.0413	0.0413	0.0413
DSR	0.9901	0.2403	**0.9999**	0.9868	0.9807	0.9884	0.9799

Panel F: *k* = 8, *q* = 2.2822

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Strat Count	411	1,021	1,037	794	846	1,606	228	442
aSR	1.8643	1.3267	1.4133	1.9881	1.5228	1.4607	0.3817	1.3586
SR	0.1154	0.0821	0.0875	0.1230	0.0942	0.0904	0.0236	0.0841
Skew	–0.2217	–0.4884	–0.3657	–0.4156	–0.3822	–0.4481	0.1270	1.6051
Kurt	13.2850	5.1541	10.3922	6.7874	7.4346	12.7538	5.3075	34.8674
T	2,170	2,167	2,169	2	2,168	2,170	2,170	2,172
StartDt	01-04-2010	01-05-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	04-27-2018	04-25-2018	04-26-2018	05-03-2018	04-25-2018	04-27-2018	04-27-2018	05-01-2018
Freq	261.1507	261.0477	261.1164	261.1159	261.0821	261.1507	261.1507	261.0474
sqrt(V[SR_k])	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298	0.0298
E[max SR_k]	0.0435	0.0435	0.0435	0.0435	0.0435	0.0435	0.0435	0.0435
DSR	0.9994	0.9606	0.9772	**0.9998**	0.9895	0.9829	0.1774	0.9754

(continued)

EXHIBIT 7 *(continued)*

Statistics Computed on Clusters' Returns

Panel G: *k* = 9, *q* = 2.2594

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Strat Count	1,021	352	536	1,037	1,593	440	228	846	332
aSR	1.3267	1.8185	1.8971	1.4133	1.4578	1.3482	0.3817	1.5228	1.9497
SR	0.0821	0.1125	0.1174	0.0875	0.0902	0.0834	0.0236	0.0942	0.1207
Skew	–0.4884	–0.2077	–0.3769	–0.3657	–0.4467	2.2752	0.1270	–0.3822	–0.4008
Kurt	5.1541	13.3085	6.1852	10.3922	12.7629	49.3210	5.3075	7.4346	10.0715
T	2,167	2,170	2,160	2,169	2,170	2,172	2,170	2,168	2,171
StartDt	01-05-2010	01-04-2010	01-22-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010
EndDt	04-25-2018	04-27-2018	05-03-2018	04-26-2018	04-27-2018	05-01-2018	04-27-2018	04-25-2018	04-30-2018
Freq	261.0477	261.1507	260.9792	261.1164	261.1507	261.0474	261.1507	261.0821	261.0131
sqrt(V[SR_k])	0.0290	0.0290	0.0290	0.0290	0.0290	0.0290	0.0290	0.0290	0.0290
E[max SR_k]	0.0441	0.0441	0.0441	0.0441	0.0441	0.0441	0.0441	0.0441	0.0441
DSR	0.9580	0.9990	**0.9995**	0.9755	0.9813	0.9736	0.1696	0.9886	0.9997

Panel H: *k* = 10, *q* = 2.2211

Stats	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Strat Count	806	1,596	948	332	409	353	327	227	851	536
aSR	1.5222	1.4586	1.3083	1.9497	1.3378	1.8174	1.2172	0.3787	1.4057	1.8971
SR	0.0942	0.0903	0.0810	0.1207	0.0828	0.1125	0.0753	0.0234	0.0870	0.1174
Skew	–0.3953	–0.4461	–0.4847	–0.4008	–0.1356	–0.2065	4.5167	0.1274	–0.4064	–0.3769
Kurt	6.9109	12.7512	5.1189	10.0715	7.4999	13.3321	108.1831	5.3035	10.9871	6.1852
T	2,168	2,170	2,167	2,171	2,170	2,170	2,172	2,170	2,169	2,160
StartDt	01-04-2010	01-04-2010	01-05-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-04-2010	01-22-2010
EndDt	04-25-2018	04-27-2018	04-25-2018	04-30-2018	04-27-2018	04-27-2018	05-01-2018	04-27-2018	04-26-2018	05-03-2018
Freq	261.0821	261.1507	261.0477	261.0131	261.1507	261.1507	261.0474	261.1507	261.1164	260.9792
sqrt(V[SR_k])	0.0278	0.0278	0.0278	0.0279	0.0278	0.0278	0.0278	0.0278	0.0278	0.0279
E[max SR_k]	0.0438	0.0438	0.0439	0.0439	0.0438	0.0438	0.0439	0.0438	0.0438	0.0439
DSR	0.9889	0.9819	0.9544	0.9997	0.9636	0.9990	0.9483	0.1706	0.9748	**0.9995**

Note: Results for the cluster containing the chosen strategy are shaded.

ACKNOWLEDGMENTS

Special thanks to Prof. Riccardo Rebonato for his valuable comments and to Dr. Michael J. Lewis for his assistance with the exhibits.

APPENDIX

PERFORMANCE STATISTICS

aRoR (Total)

Total return obtained by annualizing the geometrically linked total daily returns. This includes returns due to income from coupons, clean price changes, and financing.

Avg AUM (1E6)

Average of the daily assets under management of the long portfolio, expressed in millions of US dollars.

Avg Gini

Average of the daily Gini coefficients. The daily Gini coefficient is the ratio (1) and (2), where (1) is the area between the Lorenz curve and the line of equality and (2) is the area under the line of equality. The input is the vector of allocations (*w*) for the ISINs in the index at that moment.

```
def getGiniCoeff(w):
    w=w/w.sum()
    N=len(w)
    Ideal=(N+1)/2.
    lorenz=np.sum(np.cumsum(np.sort(w)))
    return (ideal-lorenz)/ideal
```

Avg Duration

Average of the daily weighted average durations of the portfolio (includes long, short, and futures positions), where the weights are derived from market value allocations. The daily weighted average duration δ_t is computed as

$$\delta_t = \frac{\sum_{k=0}^{n} \omega_{t,n}\delta_{t,n}}{\sum_{k=0}^{n} \left|\omega_{t,n}\right|}$$

Avg Default Prob

Average of the daily weighted average default probabilities of long positions. Weights are derived from market value allocations. A default on a short position is favorable; hence, only long positions are included in the calculation.

An. Sharpe Ratio

Annualized Sharpe ratio computed from daily total returns.

Turnover

Annualized turnover measures the ratio of the average dollar amount traded per year to the average annual assets under management.

Effective Number

The effective number of positions in the portfolio, controlling for concentration of allocations. For a detailed explanation, see López de Prado (2018), Chapter 18, Section 18.7.

```
def getEffNum(w):
    w=w.replace(0,np.nan)
    return np.exp(-(w*np.log(w)).sum())
```

Correl to Ix

Correlation of daily returns relative to the index.

Drawdown (95%)

The 95th percentile across all drawdowns. Drawdowns are computed using the following function.

```
def computeDD_TuW(series,dollars=False):
    df0=series.to_frame(‘pnl’)
    df0[‘hwm’]=series.expanding().max()
    df1=df0.groupby(‘hwm’).min().reset_index()
    df1.columns=[‘hwm’,’min’]
    df1.index=df0[‘hwm’].drop_duplicates \
        (keep=‘first’).index # time of hwm
    df1=df1[df1[‘hwm’]>df1[‘min’]]
    if dollars:dd=df1[‘hwm’]−df1[‘min’]
    else:dd=1−df1[‘min’]/df1[‘hwm’]
    tuw=((df1.index[1:]−df1.index[:−1])/ \
        np.timedelta64(1,‘Y’)).values # in years
    tuw=pd.Series(tuw,index=df1.index[:−1])
    return dd,tuw
```

Time Underwater (95%)

The 95th percentile across all time underwater. The series of time underwater is computed using the above function.

Leverage

Average of the daily leverage. Daily leverage is defined as the ratio between the market value of the long positions and the assets under management.

REFERENCES

American Statistical Association. 1999. “Ethical Guidelines for Statistical Practice.” Committee on Professional Ethics, approved by the Board of Directors (August 7, 1999). http://community.amstat.org/ethics/aboutus/new-item.

Bailey, D., J. Borwein, M. López de Prado, and J. Zhu. 2014. “Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance.” *Notices of the American Mathematical Society* 61 (5): 458–471.

——. 2017. “The Probability of Backtest Overfitting.” *Journal of Computational Finance* 20 (4): 39–70.

Bailey, D., and M. López de Prado. 2012. "The Sharpe Ratio Efficient Frontier." *Journal of Risk* 15 (2): 3–44.

——. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *The Journal of Portfolio Management* 40 (5): 94–107.

Bonferroni, C. E. 1935. "Il calcolo delle assicurazioni su gruppi di teste." Tipografia del Senato.

Gelman, A., and E. Locken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'p-hacking' and the Research Hypothesis Was Posited Ahead of Time." Working paper, http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Harvey, C., Y. Liu, and C. Zhu. 2016. "…and the Cross-Section of Expected Returns." *Review of Financial Studies* 29 (1): 5–68.

Leinweber, D. 2007. "Stupid Data Miner Tricks: Overfitting the S&P 500." *The Journal of Investing* 16 (1): 15–22.

Lo, A. 2002. "The Statistics of Sharpe Ratios." *Financial Analysts Journal* (July): 36–52.

López de Prado, M. 2017. "Finance as an Industrial Science." *The Journal of Portfolio Management* 43 (4): 5–9.

——. *Advances in Financial Machine Learning*, 1st ed. Hoboken, NJ: Wiley, 2018.

López de Prado, M., and M. Lewis. 2018. "Detection of False Investment Strategies Using Unsupervised Learning Methods." Working paper, https://ssrn.com/abstract=3167017.

Sala-i-Martin, X. 1997. "I Just Ran Two Million Regressions." *American Economic Review* 87 (2).

Sharpe, W. 1966. "Mutual Fund Performance." *The Journal of Business* 39 (1): 119–138.

——. 1975. "Adjusting for Risk in Portfolio Performance Measurement." *The Journal of Portfolio Management* 1 (2): 29–34.

——. 1994. "The Sharpe Ratio." *The Journal of Portfolio Management* 21 (1): 49–58.

Szucs, D., and J. Ioannidis. 2017. "When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment." *Frontiers in Human Neuroscience* 11 (Article 390).

Disclaimer

AQR Capital Management is a global investment management firm and may or may not apply investment techniques or methods of analysis similar to those described herein. The views expressed here are those of the authors and not necessarily those of AQR.

Fine-Tuning Private Equity Replication Using Textual Analysis

Ananth Madhavan and Aleksander Sobczyk

Ananth Madhavan
is managing director at BlackRock, Inc., in San Francisco, CA.
ananth.madhavan@blackrock.com

Aleksander Sobczyk
is director at BlackRock, Inc., in San Francisco, CA.
aleksander.sobczyk@blackrock.com

Private equity is an attractive asset class to investors seeking superior returns and low correlation to public equity markets. By *private equity*, we refer to buyouts of mature companies and growth equity—that is, established and expanding firms.[1] Private equity investors make money through management fees and performance-based fees. The illiquid nature of private equity and asymmetric information may drive risk premiums. In comparison to private markets, public markets are more transparent, and asymmetric information risk is mitigated by accounting that abides by generally accepted accounting principles, company regulation, and the attention of financial analysts and short sellers. Illiquidity is a major consideration. Private equity investments typically require a long-term commitment of 10 years or more, with the first 2 to 4 years being investment years (capital calls or periodic draw) and subsequent years the harvest period. Although the investment period is known up front to investors, capital calls are unpredictable and can vary significantly in size. The schedule of capital calls may be estimated by managers at launch (and through annual updates) but for many private equity investors in the early years the experience is quite variable.

These considerations imply that private equity investors and venture capitalists have a need for an interim beta solution to mitigate cash drag and the risk of underperformance arising from large and unpredictable capital calls. The need for interim exposure applies even to investors who have a full private equity allocation (beyond the early years) who face reinvestment risk with distributions (e.g., return of capital, realized gains, interest income) or liquidity needs for various expenses. Note the interim solution is not intended to permanently replace private equity, given the unique risk premiums associated with this asset class, but rather to supplement an actual private equity position within a portfolio.

What would such an interim portfolio solution look like? First, it should be liquid because the funds could be needed at short notice, and it should be long-only because there may be investment constraints on short positions. The solution should also be unlevered; any desired leverage can be added on top of the portfolio. Finally, and perhaps most importantly, the interim beta solution should provide dynamic economic exposure to the asset class. A dynamic portfolio is important because private equity opportunities vary over time with changes in regulation,

[1] Buyouts are often structured as leveraged buyouts but can also refer to so-called turnaround approaches through earnings growth, often accompanied by cost-reduction strategies. By contrast, we think of venture capital as more early-stage investment in firms that have yet to go public.

technology, and the business cycle. In this article, we develop a *holdings-based* methodology using modern data science to create a liquid investable portfolio to mimic dynamically the factor characteristics of private equity over time.

The alternative to a holdings-based approach is to use reported returns. Indeed, a substantial literature on hedge fund replication looks to create an investable liquid proxy for this asset class by regressing fund returns on factors associated with public equities. The returns-based regression approach has the distinct advantage of requiring only return data, but it faces some particularly difficult challenges with private equity, as we describe in detail later, because of return smoothing and other distortions.

It is generally recognized that holdings-based data provide a more accurate way to measure performance and gauge a fund's exposure to factors.[2] But how do we measure at a point in time holdings that are, by definition, not public? The approach taken here involves several steps using modern data science techniques. First, using textual analysis, we create a dictionary of private equity firms from a variety of sources. We then identify firms taken private by those private equity firms in the 10-year period ending June 2018. This step is needed because there is no explicit flag for private equity transactions in the data. Previous analyses (see, e.g., Stafford 2017) used a combination of methods and heuristics. Our approach allows us to create a *dynamic* portfolio that resembles that of private equity at a point in time.

Next, using a multifactor risk model, we measure on a quarterly basis the cross-sectional factor exposures of firms immediately prior to the *announcement* (not *effective*) date when the firms were being acquired by a private equity firm.[3] This analysis is of interest in itself because it complements the growing literature (see, e.g., Kinlaw, Kritzman, and Mao 2014) on the factor characteristics of private equity. We show that the private equity deal portfolio looks quite different from other transactions. Finally, we use holdings-based optimization to build a liquid, investable, unlevered, long-only portfolio that mimics the factor characteristics of the stocks taken private. This portfolio evolves dynamically on a quarter-by-quarter basis and, overall, has risk–return characteristics that are similar to those of reported private equity returns. Interestingly, the mimicking portfolio does not load heavily on small size or the broader market, and the factor loadings vary substantially over time. Value and minimum volatility are important attributes overall, indicative of a preference for cheaper, more stable firms, but traditional quality metrics such as profitability are not preferred, perhaps because private equity firms seek turnaround companies.

PREVIOUS LITERATURE

The article is related to several distinct areas of the literature. First, we complement previous empirical studies of the characteristics of private equity. In particular, Stafford (2017) noted that private equity funds tend to select small firms with low multiples of price to earnings before interest, tax, depreciation, and amortization (EBITDA) (i.e., smaller, value firms). He found that a passive portfolio of small, low-EBITDA-multiple stocks with modest leverage and hold-to-maturity accounting produces an unconditional return distribution that is highly consistent with that of the pre-fee aggregate private equity index.[4] This passive replicating strategy represents an economically large improvement in risk- and liquidity-adjusted returns over direct allocations to private equity funds. Franzoni, Nowak, and Phalippou (2012) estimated a four-factor model to private equity returns and reported significant exposure to factors for the market, liquidity, and value, but not size. The four-factor alpha is zero, and the liquidity risk premium is about 3% annually.

[2] Use of stock-level information in return analysis dates back to at least Brinson and Fachler (1985) and Brinson, Hood, and Beebower (1995) in holdings-based return attribution. Chen, Forsberg, and Gallagher (2016) used institutional holdings data and concluded that hedge funds are superior to other institutional investors at security selection, and hedge funds, mutual funds, and pension funds are able to successfully time the market. Lo (2008) and Hsu, Kalesnik, and Myers (2010) showed how to identify the factor and nonfactor components of active returns using security-level holdings. Grinold (2006) proposed a holdings-based attribution method using characteristic portfolios. When managers dynamically change factor loadings in response to changing economic environments, regression-based approaches may result in excessively smoothed coefficients.

[3] This approach is well established in asset pricing; an early approach was used by Ferson and Harvey (1991), who assumed that betas are a linear function of characteristics.

[4] Indexes are unmanaged, and it is not possible to invest directly in an index.

Our work is also related to efforts to capture the returns of private equity using liquid assets. Kinlaw, Kritzman, and Mao (2014) used a proprietary database of private equity returns to measure the excess return of private equity over public equity and to partition it into two components: an asset class alpha and compensation for illiquidity. They found that private equity managers generate alpha by anticipating the relative performance of economic sectors, consistent with our notion that capturing the dynamics of the opportunity set is important. They interpreted the balance of excess return as a premium for illiquidity. They also noted that their results suggest that investors can capture the asset class alpha of private equity by using liquid assets such as exchange-traded funds (ETFs) to match the sector weights of private equity investors.

The approach is also related to a large literature on hedge fund replication that seeks to create an investable liquid proxy for an asset class by regressing returns of that asset class on factors associated with public equities. Private equity investment returns are reported only quarterly and appear to provide high levels of return with only modest amounts of volatility, as we show later. Return series available to researchers are typically smoothed, the result of appraisal-based valuations.[5] As a result of artificial smoothing, which we investigate empirically in the following, investment returns exhibit high levels of autocorrelation and understate true volatility. Leverage further enhances returns, but transaction costs are understated because of the lack of liquidity. These factors tend to produce high Sharpe ratios that are difficult to proxy with public companies.

By working with unlevered, whitened returns, we can potentially approximate the true unobserved returns to private equity. One approach was illustrated by Pedersen, Page, and He (2014), who employed a lagged factor model to describe the performance of a variety of alternative and illiquid asset classes. The authors described how to estimate risk factor exposures when the available asset return series may be smoothed (owing to the difficulty of obtaining market-based valuations). They showed that private equity has exposure to beta, size, value, and liquidity factors.

An alternative source of return data is use of the returns of publicly traded firms that have private equity portfolios. Although approximately 60 global companies that invest in private equity are publicly traded (including well-known firms such as Apollo Global Management, Blackstone Group, and KKR), many private equity companies (such as Bain Capital) are structured as private partnerships. It is not clear that the returns of public companies are necessarily a representative proxy for the returns to the asset class in general.[6] However, even if we have accurate return information for private equity, the challenge of regression-based time-series coefficients being the (weighted) average over the particular sample period remains, and any attempt to replicate their exposures is inherently static.

Finally, it is worth noting that the holdings-based approach taken here is consistent with the literature showing that factor loadings vary over time. Indeed, conditional factor models, beginning with the conditional capital asset pricing model, predict that betas are a function of the economic environment, time-varying company characteristics, or the changing risk aversion of economic agents.[7] The time variation of factor loadings also significantly affects the interpretation and estimation of econometric and statistical models, including those for private equity. For example, Jagannathan and Wang (1996) showed that conditional betas are an omitted state variable, and failing to take this into account causes other coefficients, including factor loadings and alphas, to be biased. These private equity factor benchmarks we construct at each point in time are dynamic, investable, and without look-ahead bias.

EMPIRICAL ANALYSIS OF RETURNS

Before we turn to our holdings-based approach, it is useful to provide some evidence on returns to motivate the analysis to follow. We gathered quarterly total return data for private equity (Cambridge Associates US Private Equity) and two small-capitalization public equity proxies (see, e.g., Stafford 2017), namely the

[5] Returns reflect management fees, which can range widely (Stafford 2017 estimated fees of 3.5% to 5% annually), and performance fees on the profits (up to 20%). Returns may reflect unrealized and realized gains from the investments, as well as income from the investment in credit instruments. See also Ang et al. (2018).

[6] As an aside, several ETFs hold public companies that invest in private equity, an indicator of interest in this asset class.

[7] See Ang (2014) for a comprehensive review of the major literature in this area.

EXHIBIT 1
Summary Statistics on Quarterly Returns, January 1999 to April 2018

Statistics	Private Equity Returns	Russell 2000 Index	S&P 600 Index
Mean	**3.18**	**2.61**	**2.98**
Std. Dev.	5.19	10.07	9.36
Min.	–15.44	–26.12	–25.17
First Quartile	0.61	–3.78	–1.25
Median	3.88	3.21	3.66
Third Quartile	5.45	8.89	8.73
Max.	18.26	23.42	21.06

Note: Statistics are returns, in percent, observed on a quarterly basis.

Source: Based on data from Bloomberg, FactSet, and Cambridge Associates.

Russell 2000 Index and the S&P Small Cap 600 Index, for the period of January 1999 to April 2018, for a total of 78 quarters. Summary statistics on all three return series are presented in Exhibit 1. Consistent with the previous literature, we find the following:

- Private equity average returns are higher than both public indexes.
- Private equity returns are less variable in terms of measures such as the interquartile range (third quartile less first quartile) or the (quarterly) standard deviation of returns. There is little difference between the two small-cap public equity return series in terms of the measures of central tendency and dispersion, and the correlation in the two public return series is approximately 0.98.
- The risk–return trade-off is seemingly quite favorable to private equity. Approximate annual returns are 12.7% for private equity, and the annualized standard deviation is 10.4%, a ratio of return to risk of 1.22. By contrast, for the Russell 2000 Index, the approximate corresponding figures are 10.4% and 20.1%, respectively, for a return to risk ratio of 0.52.

Private equity reported returns are strongly statistically related to contemporaneous US small-cap equity returns. How well can we model the time-series pattern of returns to private equity in Exhibit 1 using a public small-cap equity index? We regress the quarterly reported private equity returns on the Russell 2000 quarterly return. The coefficient on small-cap returns is only 0.36, significantly less than 1, and the intercept (or Jensen's alpha)[8] is 2.23% per *quarter*; both are highly significant (*t*-values of 5.17 and 8.72, respectively).[9] Although the R^2 is relatively high at 0.49, this simple model illustrates that small-cap returns alone are unable to explain the risk and return characteristics of reported private equity returns.

Unlike the two public equity indexes of Exhibit 1, neither of which have autocorrelations at any lag that are statistically significantly different from zero, the reported returns of private equity show complicated dynamics possibly reflective of interpolation, smoothing, and appraisal-based valuations. We fit an autoregressive moving-average model to the private equity returns to model these dynamics parsimoniously. The model for reported private equity returns is given by:

$$r_t = \mu + \sum_{i=1}^{p} \rho_i r_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \qquad (1)$$

Here r_t is the reported private equity return in quarter t, μ is a constant term, and $\{\epsilon_t\}$ is a weak white noise process with expectation zero and constant variance. In the $ARMA(p, q)$ model of Equation 1, returns are a function of p quarters of past returns through the autoregressive coefficients $\{\rho_i\}$ and q quarters of moving average terms via the coefficients $\{\theta_i\}$. The estimated autoregressive $\{\rho_i\}$ and moving average coefficients $\{\theta_i\}$ are shown in Exhibit 2 for $(p, q) = (5, 5)$. Not only are the autoregressive elements important and significant, it is also clear that all the moving-average terms up to quarterly lag 5 are highly statistically significant.

Recall that the autocorrelation function of an $ARMA(p, q)$ process exhibits exponential decay toward zero, but with possibly damped oscillations. The conclusion that emerges from our analysis of the time series of reported private equity returns is that even with 78 quarters of data, return dynamics are very complex. Although it is certainly possible to try to correct for the impact of smoothing, staleness, interpolation, and

[8] See, for example, Jensen (1968). We get very similar results when using the S&P 600 series instead of the Russell 2000.

[9] Note also that the low beta coefficient on small-cap returns is to be expected if the proxy (i.e., the Russell 2000) is very noisy because of a well-known errors-in-variables problem.

EXHIBIT 2

ARMA(p, q) **Model of Reported Private Equity Returns**

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	θ_1	θ_2	θ_3	θ_4	θ_5
Coefficients	**0.98**	0.11	**0.69**	**–1.14**	0.27	**–1.79**	**0.50**	**–0.56**	**1.79**	**–0.94**
Std. Errors	0.15	0.12	0.07	0.11	0.14	0.11	0.19	0.18	0.14	0.09

Notes: Coefficients that are statistically significant are marked in bold. The log likelihood is −222.48, and the Akaike information criterion is 466.95. Technically, we estimate an ARIMA(5, 1, 5) model to handle nonstationarity.

Source: Based on return data from Cambridge Associates, in percent, observed on 78 quarters.

appraisal-based valuations (e.g., whitening), this is not an easy task given the complexity shown in Exhibit 2. Accordingly, we turn to a different approach, based on holdings, that allows us to dynamically model the factor attributes of private equity.

DYNAMIC HOLDINGS-BASED LIQUID ALTERNATIVES MODELING

Private Equity Acquisitions of Public Companies

Our holdings-based approach to liquid alternatives modeling consists of three elements:

- Use public equity markets, identify companies that were acquired by private equity firms;
- Use a multifactor model to measure cross-sectional characteristics of those firms prior to the announcement; and
- Use holdings-based factor characteristic mapping to build liquid, investable, long-only portfolios that vary over time to mimic the dynamic target private equity portfolio.

Our sample covers the 10-year period between June 2008 and June 2018. We use the FactSet M&A database and the Bloomberg CAX database to identify 1,107 mergers and acquisitions (M&A) events that resulted in a public US company being delisted (approximately 110 instances per calendar year). Unfortunately, there is no easy way to detect which of these events involved private equity. Although it is possible to make such a determination by hand, this is not a scalable approach—certainly not globally or on a going-forward basis for a possible product or a client portfolio. To develop a systematic approach to identify private equity deals, we used textual identification against custom word dictionaries (based on relevant textual sources for broad industry designations and including firm-specific words) to identify acquirers that are private equity firms or part of private equity–led consortia or private groups. This exercise resulted in 159 events. Exhibit 3 provides a sample textual analysis of an event; Panel A shows the dictionary, and Panel B shows how we distinguish between two M&A events, one of which involves private equity whereas the other is a within-industry biopharma acquisition.

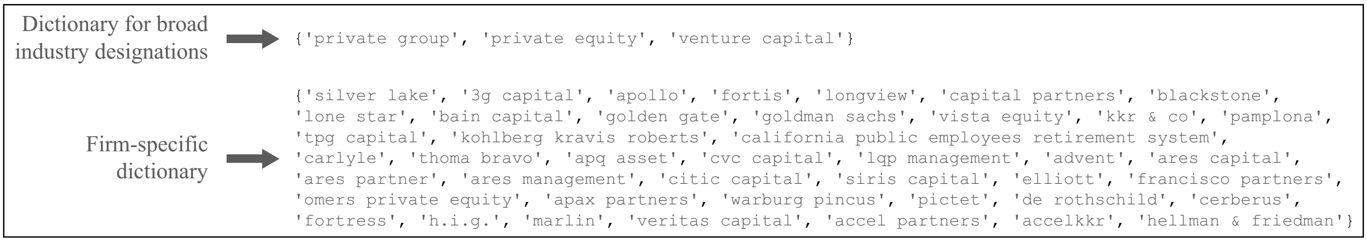
Not surprisingly, the great majority of private equity deals (*deal portfolio*) involve cash (97%), as shown in Exhibit 4. Furthermore, the majority of private equity acquirers are private (86%), with some notable exceptions (e.g., Blackstone, Apollo). Private equity firms target smaller deals (average of $2.5 billion compared to $3.8 billion for non–private equity acquirers), but there are some notable exceptions. This finding suggests that using the returns of public companies that invest in private equity as proxies for private equity returns may not be representative. We note in passing that the sector and industry composition of the deal portfolio changes over time, as one would expect given a time-varying opportunity set, again motivating the need for a dynamic approach.

Mapping Style Factor Exposures

The next step is to determine the factor exposures of the deal portfolio at each point in time, as a prelude to mapping them to investable, long-only factors. For individual stocks, we collect risk characteristics produced by an industry multifactor risk model, BlackRock's Fundamental US Equity Risk Model (BFRE USAM), that includes style characteristics such as momentum, volatility, size, value, and trading, as well as individual sector

Exhibit 3
Sample Textual Analysis

Panel A: Custom Dictionary



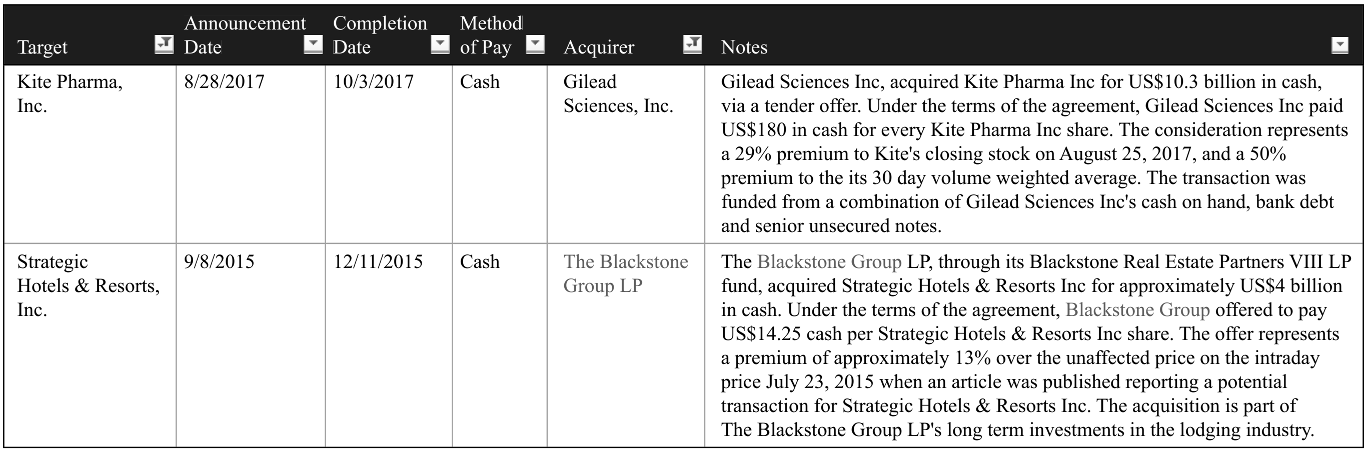
Dictionary for broad industry designations

```
{'private group', 'private equity', 'venture capital'}
```

Firm-specific dictionary

```
{'silver lake', '3g capital', 'apollo', 'fortis', 'longview', 'capital partners', 'blackstone',
'lone star', 'bain capital', 'golden gate', 'goldman sachs', 'vista equity', 'kkr & co', 'pamplona',
'tpg capital', 'kohlberg kravis roberts', 'california public employees retirement system',
'carlyle', 'thoma bravo', 'apq asset', 'cvc capital', 'lqp management', 'advent', 'ares capital',
'ares partner', 'ares management', 'citic capital', 'siris capital', 'elliott', 'francisco partners',
'omers private equity', 'apax partners', 'warburg pincus', 'pictet', 'de rothschild', 'cerberus',
'fortress', 'h.i.g.', 'marlin', 'veritas capital', 'accel partners', 'accelkkr', 'hellman & friedman'}
```

Panel B: Distinguishing between M&A Events



Target	Announcement Date	Completion Date	Method of Pay	Acquirer	Notes
Kite Pharma, Inc.	8/28/2017	10/3/2017	Cash	Gilead Sciences, Inc.	Gilead Sciences Inc, acquired Kite Pharma Inc for US$10.3 billion in cash, via a tender offer. Under the terms of the agreement, Gilead Sciences Inc paid US$180 in cash for every Kite Pharma Inc share. The consideration represents a 29% premium to Kite's closing stock on August 25, 2017, and a 50% premium to the its 30 day volume weighted average. The transaction was funded from a combination of Gilead Sciences Inc's cash on hand, bank debt and senior unsecured notes.
Strategic Hotels & Resorts, Inc.	9/8/2015	12/11/2015	Cash	The Blackstone Group LP	The Blackstone Group LP, through its Blackstone Real Estate Partners VIII LP fund, acquired Strategic Hotels & Resorts Inc for approximately US$4 billion in cash. Under the terms of the agreement, Blackstone Group offered to pay US$14.25 cash per Strategic Hotels & Resorts Inc share. The offer represents a premium of approximately 13% over the unaffected price on the intraday price July 23, 2015 when an article was published reporting a potential transaction for Strategic Hotels & Resorts Inc. The acquisition is part of The Blackstone Group LP's long term investments in the lodging industry.

Source: BlackRock, based on FactSet M&A and Bloomberg CAX data.

Exhibit 4
M&A-Driven Delistings of US Public Companies, 2008 to 2018

Acquirer	Number of Deals	Public Acquirer (%)	Cash Deal (%)	Deal Size ($Millions) Average	Deal Size ($Millions) Median
Private Equity	159	14	97	2,514	795
Other	948	83	61	3,794	1,318
All	1,107	73	66	3,610	1,259

Source: BlackRock, based on FactSet M&A and Bloomberg CAX data.

exposures. The choice of risk model has little impact on the results, given the similarity in models across different providers. Our set of tradable factors are MSCI single factor indexes, which offer a variety of targeted factor exposures and can be traded by an investor at low cost through ETFs. The individual factors, beyond the broad market (Russell 3000), are size, value, momentum, minimum volatility, and quality.

When to Map Factors?

We have a choice in the date of factor mapping. One approach is to use the last calendar month-end date before delisting (i.e., the effective date). Thus, if a company was taken private and delisted from an exchange on June 13, 2001, we could use factor exposures on May 31, 2001. An alternative is to use the style factor characteristics prior to the deal announcement date—that is, if a company was delisted from an exchange on June 13, 2001, but the deal was announced on March 15, 2001, we could use factor exposures on February 28, 2001. The advantage of using the deal announcement date is that this analysis controls for any post-announcement price movement (i.e., from the market price before announcement to the target price). Indeed, we find that the factor

EXHIBIT 5
Comparison of US Public Companies Delisted, 2008 to 2018

	Private Equity Deals			
	Announcement	**Effective**	**Other Deals**	***t*-Statistic**
Momentum	–0.57	0.20	0.63	**–12.26**
Volatility	0.69	0.69	0.79	–1.15
Earnings Yield	–0.15	–0.24	–0.56	**4.22**
Size	–2.00	–1.96	–1.80	**–2.46**
Growth	–0.16	–0.19	–0.06	–1.12
Leverage	0.01	–0.02	–0.06	0.74
Reversal	0.22	0.16	0.25	–0.33
Value	0.53	0.28	–0.16	**7.77**
Dividend Yield	–0.46	–0.55	–0.41	–0.65
Small Cap	2.04	1.93	1.61	**3.80**
Liquidity	–1.27	–1.99	–1.45	1.64
Profitability	–0.25	–0.25	–0.57	**3.27**

Notes: The column "Other" represents all 948 nonprivate deals out of a total of 1,107 deals. The t-statistic refers to a two-tailed test of pre-announcement date factor loadings for private versus other deals, with significant differences marked in bold.

Source: BlackRock, based on FactSet M&A and Bloomberg CAX data.

exposures of private equity targets are significantly more differentiated when mapping their style exposures on the announcement date than on the effective date. Not surprisingly, the effects are strongest for the momentum factor. Just prior to the announcement, the momentum Z-score of the deal portfolio averages −0.57 versus 0.20 on the effective date. Price appreciation to the new target price leads to momentum and a reversal toward the effective date. There is also an evident effect on the liquidity factor and, through price appreciation, on the value factor. In what follows, we will use the (pre) announcement date for factor matching.

Comparison of Deals

We assume a standard multifactor model in which the returns of stock i at time t, $r_{i,t}$, are a linear function of K factors with betas that vary over time:

$$r_{i,t} = \alpha_i + \sum_{k=1}^{K} \beta_{i,k,t-1} F_{k,t} + \varepsilon_{i,t}, \tag{2}$$

where $\beta_{i,k,t-1}$ denotes the exposure of stock i to factor k at time t. Note that timing is explicit in the subscripts. We instrument the beta $\beta_{i,k,t-1}$ for returns at time t to emphasize that it is measurable with respect to information at time $t-1$. We summarize this information in a characteristics vector, denoted by $z_{i,t-1}$. The constant or alpha is not time subscripted, meaning it does not itself vary with time, but returns in excess of the time-varying factor exposures are subject to stochastic shocks, $\varepsilon_{i,t}$.

We estimate factor loadings, $\beta_{i,k,t-1}$, using cross-sectional information, $z_{i,t-1}$, by assuming that various sets of factors are functions of security-level risk characteristics. Exhibit 5 shows the mean Z-score by factor model (using BFRE USAM) for the sample of companies identified as private equity deals based on two dates, announcement and effective. The analysis in Exhibit 5 also shows the changes in average style factor exposure between the deal announcement date and the effective (delisting) date for private equity targets. The column "Other" represents other deals.

Of special interest is the comparison of the deal portfolio to other M&A transactions. Exhibit 6 shows the t-statistic for a two-tailed test of a difference between the mean (pre-announcement date) of the private deal sample versus the "other" deal category. There are marked differences in the factor characteristics of the deal portfolio compared with other public M&A targets. Consistent with the previous literature, we find that compared to all other public M&A targets, targets of private equity firms are

- smaller (size and small cap; also less liquid)
- cheaper (value and earnings yield)
- higher quality (profitability)

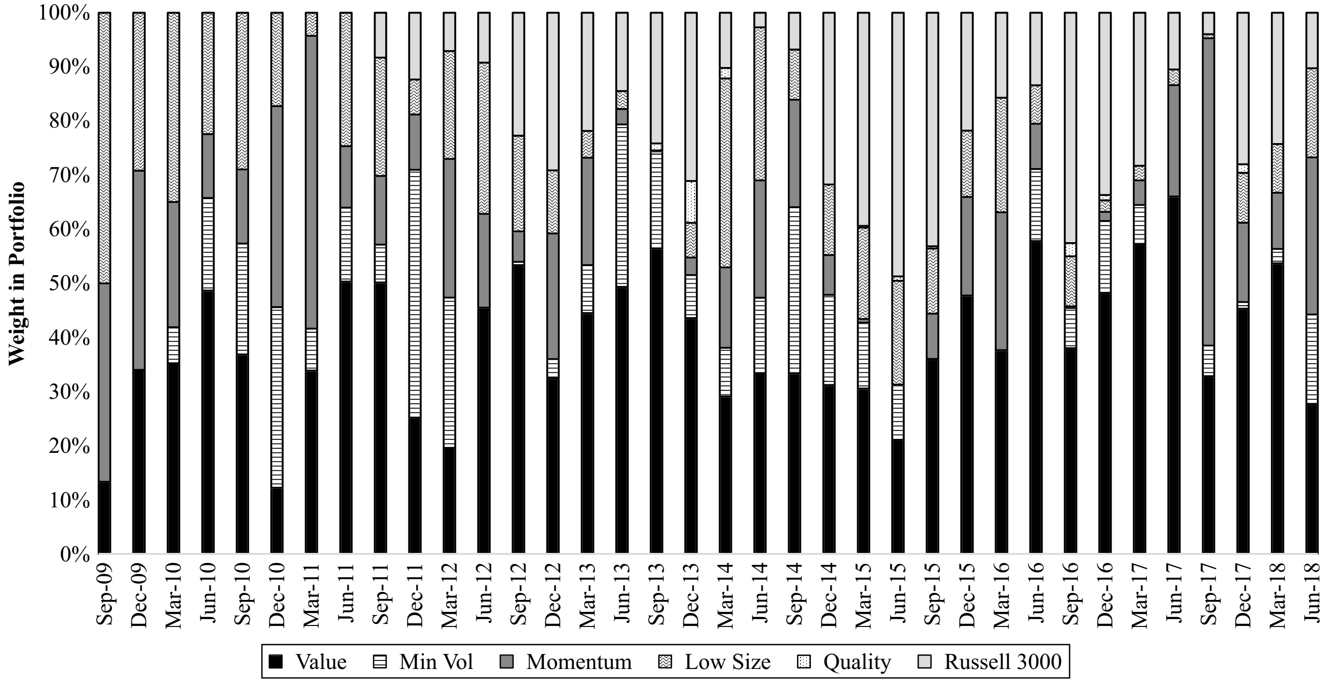
Important differences between the deal and other portfolios relate to factors such as momentum (private equity deals have significantly negative momentum relative to other deals when estimated pre-announcement). There are also important and statistically significant differences for value, quality, yield, and profitability.

Factor Index Portfolios—Intuition

The next step is to translate the private equity portfolio's cross-sectional risk characteristics into investable

Exhibit 6

Investable Private Equity Mimicking Portfolio



Source: BlackRock, based on FactSet M&A and Bloomberg CAX data from September 30, 2009, to June 30, 2018.

factor index portfolios at each point in time. We follow the approach of Ang, Madhavan, and Sobczyk (2017):

- We start with risk characteristics: variables such as beta or book-to-price for stocks, but also sectors, countries, and currencies.
- The risk model maps securities onto risk characteristics such as value and momentum, as described by Equation 2 earlier.
- At each point in time, optimally match the risk characteristics of a given company to the risk characteristics of a set of third-party long-only style indexes using optimization.
- The resulting portfolio is dynamic, investable, and without any look-ahead bias.

We assume that traded securities have security-level risk attributes. Some of these characteristics, such as valuation ratios or past returns, are sometimes directly used to form style (or smart beta) factors, following Fama and French (1993). We compute factor loadings for our proxy private equity fund at a given time by finding the combination of factors with the closest match, in terms of characteristics, to that fund's holdings. The formal optimization problem is laid out in the next section, but the intuition is quite simple. Suppose that the private equity portfolio at a particular point in time has a value Z-score (e.g., using metrics such as earnings/price or book/price) of 0.40 and a momentum (e.g., trailing-12-month returns omitting the most recent month) Z-score of 0.25. Suppose a long-only, investable value index has Z-scores to value and momentum of 0.90 and −0.10, respectively. Furthermore, suppose a long-only momentum index has Z-scores to value and momentum of −0.10 and 0.70, respectively. It is easy to see then that the investable factor-mimicking portfolio is composed of 50% value index and 50% momentum index. A formal exposition follows.

Formal Optimization Objective

The formal objective is to translate cross-sectional risk characteristics (exposures) into investable factor index exposures. At the start of period t, for any given fund, define an *index factor portfolio* comprising weights $w_{j,t-1}^{IND}$ in an investable index factor $j = 1 \ldots M$, where the number of investable funds (e.g., ETFs) does not exceed the number of possible risk factors in Equation 1 (i.e., $M \leq K$). We require the weights in the index portfolio to satisfy $0 \leq w^{IND}_{j,t-1} \leq 1$ and $\Sigma^{M}_{j=1} w^{IND}_{j,t-1} = 1$ (i.e., the portfolio is long-only and fully invested). Denote by $\hat{\beta}^{IND}_{j,k,t}$ the exposure of investable fund j to risk factor k in period t. It follows that the expected return of the private equity *factor portfolio* with weights $w^{IND}_{j,t-1}$ (where $j = 1 \ldots M$) in t is:

$$E[R^{IND}_{t}] = \sum_{j=1}^{M} w^{IND}_{j,t-1} \left(\sum_{k=1}^{K} E(\hat{\beta}^{IND}_{j,k,t}) E(F_{k,t}) \right) \quad (3)$$

The difference between the fund's expected total return attributable to static exposures to the K risk factors (from Equation 2) and the expected return of the index factor portfolio (from Equation 3) is denoted by $\hat{\eta}_t$, where

$$\hat{\eta}_t = \sum_{k=1}^{K} E(\hat{\beta}_{k,t}) E(F_{k,t}) - \sum_{j=1}^{M} w^{IND}_{j,t-1} \left(\sum_{k=1}^{K} E(\hat{\beta}^{IND}_{j,k,t}) E(F_{k,t}) \right) \quad (4)$$

The ordinary least squares estimate for the index factor portfolio at time t is the set of M weights $w^{IND}_{j,t-1}$ that minimizes the squared residual in Equation 4, subject to the following constraints:

$$\sum_{j=1}^{M} w^{IND}_{j,t-1} = 1,$$
$$\text{and} \quad 0 \leq w^{IND}_{j,t-1} \leq 1, \text{ for each } j = 1 \ldots M. \quad (5)$$

In other words, we require full investment and long-only positions in the factor indexes. This approach could be used more broadly for alpha capture with other return drivers.

Investable Liquid Portfolios

Following the methodology described earlier, we constructed an investable factor-mimicking portfolio for public companies that were targets of private equity acquisitions. The mimicking portfolio is rebalanced quarterly (although we could use an alternative frequency such as monthly, albeit with fewer constituents) and created using the following liquid public instruments:

- MSCI USA Enhanced Value Index (Value)
- MSCI USA Minimum Volatility Index (Min Vol)
- MSCI USA Momentum Index (Momentum)
- MSCI USA Risk Weighted Index (Low Size)
- MSCI USA Sector Neutral Quality Index (Quality)
- Russell 3000 Index (Market)

Exhibit 6 shows the composition of the investable mimicking portfolio based on private deals over the sample period of 10 years, from Q3 2009 through Q2 2018, with no look-forward bias. The factor-mimicking portfolio is dynamic, changing with the latest quarter's private deal portfolio.

Exhibit 6 shows that the mimicking portfolio is not simply composed of the broader market (e.g., the Russell 3000 Index has a median weight of only 12.9% over the whole period) but reflects time-varying factor attributes. Nor is the mimicking portfolio completely dominated by small-cap proxies. Consistent with our earlier regression results of private equity returns on small-cap indexes, low size is an important (median of 15.5%; range of 0.2%–50.0%), but by no means dominant, factor. Rather, value is the largest component, with a median allocation of 37.3% over the entire period, but it also exhibits significant time-series variation. The quality factor generally has a very low or zero weight in most quarters (the median weight is zero, the lowest possible given the long-only constraint), possibly because private equity firms seek out companies they can turn around that score low on quality metrics such as return on equity and profitability. By contrast, minimum volatility has a median weight of 8.4%, consistent with the notion that private equity firms prefer companies with stable cash flows over the business cycle that have the capacity to carry additional debt.

Recall that the private equity mimicking portfolio represented in Exhibit 6 has no look-ahead bias. It is worth emphasizing that we do not use reported private equity returns (which could reflect smoothing and leverage) to construct this portfolio. Assuming quarterly reconstitution and no leverage, the annualized return for the mimicking portfolio in the period beginning September 30, 2009, and ending June 30, 2018, based on the returns to the factors, is 15.1%, with a volatility of 11.6%.[10] By contrast, from Exhibit 1, the reported

[10] The computation takes the sum of factor weights times the returns to the associated factors. Indexes are unmanaged, and it is not possible to invest directly in an index. In the last three years, there are ETFs that offer low-cost proxies (15 basis points) for the factor indexes considered here.

(approximate) annual private equity returns are 12.7% with an annualized standard deviation of 10.4%. It is important to note, though, that the investable mimicking portfolio is not intended to be a substitute for private equity (which offers liquidity and other risk premiums), but rather is a way to deploy excess cash in anticipation of possible future capital calls.

CONCLUSIONS

Private equity is of considerable interest as an asset class. In the early years, private equity investors are committed, but they face challenges because capital calls are unpredictable and can vary significantly in size, risking a performance shortfall. Consequently, private equity investors and venture capitalists have a need for liquid solutions that provide economic exposure to the asset class (so-called *interim beta*) to deploy excess cash, mitigate the risk of underfunding, and manage large and unpredictable capital calls.

In this article, we develop a *holdings-based* methodology using modern data science to create an investable portfolio to dynamically replicate the factor characteristics of private equity. The alternative to a holdings-based approach is to use reported returns, following a large literature on hedge fund replication that seeks to create an investable liquid proxy by regressing hedge fund returns on factors associated with public equities. The returns-based regression approach has the distinct advantage of requiring only return data, but it faces some particularly difficult challenges in private equity, as we showed. Specifically, using 78 quarters of data from January 1999 to April 2018, we find significant evidence of autocorrelation and moving average terms at up to five quarterly lags. These complex dynamics reflect effects such as smoothing, interpolation, and appraisal-based valuations, but perhaps also cyclical factors or error correction. Further research into the dynamics of reported returns is clearly important to investors seeking to understand the diversification benefits of private equity across the business cycle and as part of a public portfolio.

As an alternative to using reported private equity returns, we explore a holdings-based approach to mimic the factor characteristics of a private equity portfolio dynamically. Using textual analysis, we create a dictionary of private equity firms and then identify firms taken private by those firms in the 10-year period ending June 2018. We measure the cross-sectional factor exposures of firms immediately prior to the announcement that they were being acquired by a private equity firm using a risk model. We show the importance of measuring factor exposures of the private equity deal portfolio prior to the announcement date and demonstrate significant changes in momentum and value between the announcement and effective dates. Private equity portfolios look different from other deals: They are smaller (size and small cap; also, less liquid), cheaper (value and earnings yield), and higher quality (profitability).

Finally, we use holdings-based robust optimization to build a portfolio of factor indexes that replicate the factor characteristics of the stocks taken private. This exercise can be repeated at any interval. It would be interesting to understand better how the dynamic holdings-based approach described here compares and contrasts with a returns-based approach and whether both could be integrated in some fashion. In recent years, the mimicking portfolio loads not only on small size but on value and other factors. From a practical perspective, the ability to create a factor-mimicking portfolio that is liquid, investable, and long-only offers a valuable way for private equity investors to maintain exposure and control the risks associated with unpredictable capital calls.

ACKNOWLEDGMENT

This work reflects considerable discussion with and insights from Bart Sikora and Ravi Goutam.

REFERENCES

Ang, A. *Asset Management: A Systematic Approach to Factor Based Investing.* New York, NY: Oxford University Press, 2014.

Ang, A., B. Chen, W. Goetzmann, and L. Phalippou. 2018. “Estimating Private Equity Returns from Limited Partner Cash Flows.” *The Journal of Finance* 78 (4): 1751–1783.

Ang, A., A. Madhavan, and A. Sobczyk. 2017. “Estimating Time-Varying Factor Exposures.” *Financial Analysts Journal* 73 (4): 41–54.

Brinson, G., and N. Fachler. 1985. “Measuring Non-US Equity Portfolio Performance.” *The Journal of Portfolio Management* 11 (3): 73–76.

Brinson, G., L. R. Hood, and G. Beebower. 1995. "Determinants of Portfolio Performance." *Financial Analysts Journal* 51 (1): 133–138.

Chen, Z., D. Forsberg, and D. Gallagher. 2016. "Which Institutional Investor Types Are the Most Informed?" SSRN, https://ssrn.com/abstract=2840549.

Fama, E., and K. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

Ferson, W., and C. Harvey. 1991. "The Variation of Economic Risk Premiums." *Journal of Political Economy* 99 (2): 385–415.

Franzoni, F., E. Nowak, and L. Phalippou. 2012. "Private Equity Performance and Liquidity Risk." *The Journal of Finance* 67 (6): 2341–2373.

Grinold, R. 2006. "Attribution." *The Journal of Portfolio Management* 32 (2): 9–22.

Hsu, J., V. Kalesnik, and B. Myers. 2010. "Performance Attribution: Measuring Dynamic Allocation Skill." *Financial Analysts Journal* 66 (6): 17–26.

Jagannathan, R., and Z. Wang. 1996. "The Conditional CAPM and the Cross-Section of Expected Returns." *The Journal of Finance* 51 (1): 3–53.

Jensen, M. 1968. "The Performance of Mutual Funds in the Period 1945–1964." *The Journal of Finance* 23 (2): 389–416.

Kinlaw, W., M. Kritzman, and J. Mao. 2014. "The Components of Private Equity Performance: Implications for Portfolio Choice." Working paper 5084-14, MIT Sloan School.

Lo, A. 2008. "Where Do Alphas Come From? A Measure of the Value of Active Investment Management." *Journal of Investment Management* 6 (2): 1–29.

Pedersen, N., S. Page, and F. He. 2014. "Asset Allocation: Risk Models for Alternative Investments." *Financial Analysts Journal* 70 (3): 34–45.

Stafford, E. 2017. "Replicating Private Equity with Value Investing, Homemade Leverage, and Hold-to-Maturity Accounting." Working paper, Harvard Business School.

Disclaimer

The views expressed here are those of the authors alone and not of BlackRock, Inc., its officers, or its directors.

A Practical Approach to Advanced Text Mining in Finance

JULIA KLEVAK, JOSHUA LIVNAT, AND KATE SUSLAVA

JULIA KLEVAK
is a director of systems development at QMA in Newark, NJ.
julia.klevak@qma.com

JOSHUA LIVNAT
is a professor emeritus of accounting in the Stern School of Business Administration at New York University in New York, NY and the head of research at QMA in Newark, NJ.
jlivnat@stern.nyu.edu

KATE SUSLAVA
is an assistant professor of management in the Freeman College of Management at Bucknell University in Lewisburg, PA.
kate.suslava@bucknell.edu

The sheer volume of data available for analysis can be a daunting prospect for an investor. Some of these data are structured and normalized—for example, numeric information from financial statements or analysts' earnings forecasts. However, a large body of data is available in the form of text, such as Securities and Exchange Commission (SEC) filings, news, materials scraped from the web, and so on. The construction of an investment signal from text is a complex task because it requires the ability to identify positive and negative parts of the text, weight the different parts, and construct a final score. In this study, we briefly review the evolution of the analysis of text in finance and accounting and provide a concrete example through the analysis of earnings conference call transcripts.

The early literature in finance and accounting used a simplistic way to score a text: counting the number of positive and negative words in the text to determine the overall tone of the text. Researchers initially identified positive and negative words using general dictionaries developed in psychology, such as the *General Inquirer* (Tetlock 2007). It soon became clear that the business use of words is different from the general use, and some words that are generally considered negative may not be so in a business context. For example, *liability* has a negative connotation in general use; however, in the language of business it simply refers to the company's debts. The word *sinking* normally has a negative sentiment, but in business communication, using this word in the phrase *sinking fund* refers to a regular financing practice. On the other hand, although *reconciliation* is generally a positive word, in business a *bank reconciliation* is a regular accounting procedure that does not carry any sentiment.

Some authors reacted to this challenge by constructing their own dictionary of positive and negative words (Henry 2008). A more comprehensive effort was made by Loughran and McDonald 2011; they used Form 10-K SEC filings, the annual form publicly listed companies are required to file, to construct a comprehensive dictionary from words that were frequently used by firms. This list of words still needed to be classified into positive and negative categories, as well as uncertainty, litigious, modal, and constraining categories.[1] This became the golden standard in academic studies afterward, and many studies used it to analyze various channels of financial disclosures: the management discussion and analysis (MD&A) section of 10-Q and 10-K (Feldman et al. 2010), earnings conference calls (Brochet, Loumioti, and Serafeim 2015; Suslava 2016), loan agreements (Bozanic, Cheng, and Zach 2018),

[1] See: https://sraf.nd.edu/textual-analysis/resources/.

and initial public offering prospectuses (Fishe, North, and Smith 2014).

The most common approach in text mining has been to count the number of positive and negative words in a text and construct an overall score based on these quantities, which then defines the tone or sentiment of the text. For example, one can use the number of positive words minus the number of negative words, scaled by the total number of words (or the sum of positive and negative words) as a measure of tone. Some studies use only the proportion of negative words in a text to measure tone because text generated by a company, such as press releases, tend to have a positive bias. Using word lists to identify the tone of a document is essentially a blunt tool and may be used as a cursory and superficial instrument, similar to taking body temperature to diagnose an illness. Realizing this bluntness, some authors focused on comparisons across time for the same company, which can mitigate the positive bias inherent in self-reporting.

Another text mining approach is to use *classification*, which can be performed at the level of the entire document, the individual paragraph, or even the specific sentence. Typically, the classification process begins with training data (i.e., a preliminary set of annotated examples that provide the basis for classification of future documents). Of course, to obtain decent accuracy with future classification, the training set is crucially important. Using more annotated examples typically leads to a more accurate classification. Additionally, the training set should be annotated in a similar manner. Having several annotators, which is often the case in obtaining a large training set, leads to inconsistency across annotators, which will likely reduce the future accuracy of the classification. This approach is similar to the doctor who searches for symptoms that may indicate the types of illnesses the patient may be suffering.

As an example of the classification approach, Li (2010) used phrases that discuss future events in the MD&A sections of SEC filings to assess the quality of earnings and improve the predictability of future earnings. His approach was based on identifying words that denote the future (e.g., *expect*) and assessing whether the entire sentence was positive or negative, using a large sample of examples to classify future phrases. Using a large dataset of annotated sentences to classify future text is also the basis of some commercial work in text mining (e.g., the news sentiment approach used by RavenPack).

It should be noted that in addition to measuring the tone or sentiment of a text, one can use a classification technique to identify events that are present in the text. For example, using training data on acquisition announcements by companies, one can develop a classification tool that will seek to identify future occurrences of acquisitions in other documents. What reduces the accuracy of future identification of events is the similarity in reporting of events that involve more than one company. For example, language used to describe an acquisition may be similar to the language used to report merger and acquisition activity, joint ventures, alliances, and so on. Thus, the accuracy of event extraction from a text is dependent on the number of extracted events, the size of the training set, and the correct annotation of the training set.

A third and more sophisticated text mining approach is based on writing natural language processing (NLP) rules to extract specific events from text. These rules use the semantic structure of a sentence to report the event, and in some cases this approach involves writing numerous NLP rules for a single event. This approach usually yields more accurate identification of events in a text, but it requires expending significant effort for each event. However, NLP offers the rule creator an opportunity to continuously improve the rules when extracted events are wrong or when omitted events have not been extracted by the rules. This approach is more specific and targeted and is similar to various medical tests the doctor orders to better diagnose a patient's condition.

Other, more mechanical types of textual analyses include comparing texts from the same company across time to identify differences (Cohen, Malloy, and Nguyen 2015) or counting the number of words or numbers in a document (Zhou 2018). Similarly, several authors examined how text clarity and readability affect investors (Li 2008; Biddle, Hilary, and Verdi 2009; Lehavy, Li, and Merkley 2011; De Franco et al. 2015). Others use the business description in the 10-K to examine the similarity of firms' products and operations (Hoberg, Phillips, and Prabhala 2014). Such approaches, although worthwhile from the perspective of research topics, are less interesting as text mining endeavors because they are based on well-known mechanical tools for processing text.

The purpose of this study is to illustrate an advanced approach to text mining that is based on writing specific NLP rules to identify events without advanced knowledge of NLP theory. The approach uses software that allows a domain expert to easily write NLP rules that identify events of interest for the expert in that domain, quickly test the rules by searching for other examples within a small corpus of documents from that domain, and then decide whether to use that rule on the entire domain or to modify it. We show in the following the application of the approach to earnings conference call transcripts. We should emphasize that the approach we describe in this article is one efficient way of writing rules but not necessarily the only way.

AN EASY PROCESS TO RULE WRITING

The approach we describe is based on software developed by Amenity Analytics, Inc. (Amenity).[2] Amenity's software is used to process a text and to extract from it both the tone (sentiment) and events for which Amenity has developed specific rules. The extracted tone is based on expanded word lists that the founders have accumulated through their long experience working with financial literature. In addition to words, Amenity's software has cataloged phrases that are often used in financial texts and has carefully written rules to identify sentences and even complete paragraphs as positive or negative. Amenity has also written rules around events it considers important and assigned weights to the various events, reflecting its views on the significance of these events. Thus, a document is given a numeric score based on a combination of sentiment and events captured from the text.

For advanced users, the real benefit of the software is in the user's ability to write additional rules to identify events without having any NLP expertise. This is accomplished through a few steps that we will outline. The first step involves creating a small sample of documents from the corpus. These documents are then processed through the software and are available for viewing. The text of each document is annotated in color to reflect the words and phrases used for extracted sentiment and events; the user has the ability to indicate whether these annotations are correct or not. The second step in rule writing is to highlight a sentence or a phrase of interest, which is used by the program to parse the sentence and create a graph of the parsed sentence. The user can examine the graph and then easily modify it to make it more general, so the program can identify similar structures in other documents. For example, the user can omit company-specific identifications ("The agricultural products segment"), specific numerical values ("generated growth of 18%"), and auxiliary words ("*has* generated") and use synonyms for the key words ("generated" and "growth"). The software automatically creates an NLP equivalent of the modified parsed graph. The next step is to determine the efficacy of the rule for other documents. This is done by clicking on a "Find Matches" button, which retrieves other examples from the sample corpus. If only the original example is retrieved, it is likely that the rule is not general enough and additional parts may be omitted. If a sufficient number of other examples are correctly extracted and the user is satisfied with the rule, the user assigns a name to the event (e.g., Revenue Growth), determines its polarity (e.g., positive), and saves the rule. The user can also determine the weight that should be attached to the event, depending on its perceived importance in predicting future returns. On reprocessing the documents in the corpus, all identified Revenue Growth events should now be highlighted and color-coded for their polarity. For example, if the manager said, "Sales were affected negatively because our Florida stores did not generate the expected growth," the software will flag the event Revenue Growth as negative.

Note that the software does all the heavy lifting of writing NLP rules. The user is only required to home in on a sentence or a phrase that is of interest and then easily construct a rule that addresses what the user wanted and is sufficiently general to capture other examples in the corpus. The software allows people who are domain experts to write rules, thereby making the task of rule writing more efficient and streamlined and more efficiently utilizing the domain expert's perspective on events that typically affect security prices. However, it should be stressed that having people who are both domain experts and proficient in NLP is likely to yield even greater benefits.

AN APPLICATION TO EARNINGS CONFERENCE CALL TRANSCRIPTS

A few weeks after the fiscal quarter ends, most companies issue a press release that provides preliminary

[2] The software was licensed from the vendor through a paid subscription. There are no beneficial agreements between the authors and Amenity.

details about their performance in the prior quarter. Many companies will also host a conference call with analysts to provide additional color on the press release. The call typically consists of two parts: a scripted part during which management discusses the newly issued release, which then often opens the floor to a question and answer discussion with analysts. The transcripts of these conference calls are available for consumption through several vendors.

Earnings conference call transcripts are of particular interest to quantitative investors. First, these transcripts are available for a large proportion of the investable universe (typically over 80% of the 3,000 largest companies in the United States). Second, this source of data becomes available at regular quarterly intervals. Third, like earnings surprises and estimate revisions, which were shown to exhibit investors' underreaction and autocorrelations, conference call tone signals are also likely to be autocorrelated from quarter to quarter. Finally, quantitative portfolios are typically broad, with many positions and with small bets on those positions. This is ideal for a text-mining tool, which is likely to make errors in identifying the precise tone or tone change of a specific firm but is likely to be correct more often than not for a broad portfolio, if constructed appropriately. We should note, however, that the conference call occurs immediately after earnings are released. To the extent that investors react to the earnings and the information contained in the earnings press release, the stock price may have already incorporated the information discussed during the call. Thus, in all our tests we control for the earnings surprise and the abnormal return around the earnings announcement, which captures other information in the earnings press release.

Analysis

For this analysis, we obtained conference call transcripts from Thomson Reuters for the period 2002–2016. We restricted our sample to earnings conference calls of US companies that had preliminary earnings information in the Compustat Point-in-Time database and returns in the Center for Research in Security Prices (CRSP) database. For each conference call, we first calculated the earnings surprise (*SUE*) as the earnings per share (EPS) reported in the earnings release minus the EPS reported in the same quarter of the prior year, and minus the average same-quarter EPS differences in the prior eight quarters.[3] We scale this earnings surprise by the standard deviation of the same-quarter EPS differences during the prior eight quarters. We then rank all the earnings surprises during a calendar quarter into quintiles (0 through 4), divide by 4, and subtract 0.5. We use this transformed variable as an independent variable in quarterly regressions of the abnormal future return on various signals. Its coefficient is equivalent to the return on a hedge portfolio that has a long position in the top quintile (4, the largest positive earnings surprises) and a short position in the bottom quintile (0, the most negative earnings surprises).

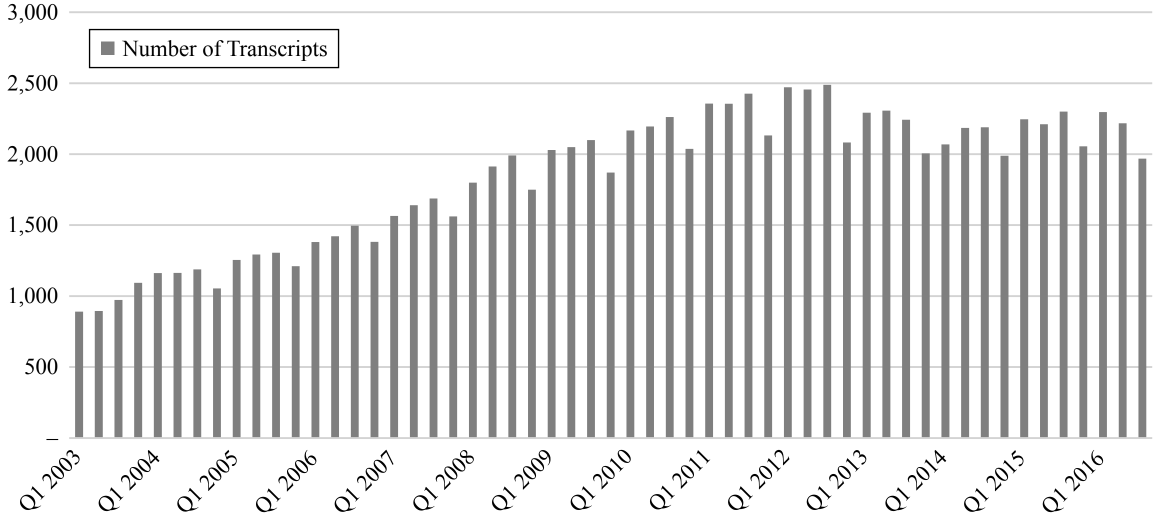
We use two abnormal return windows in this study. The first is a short window around the earnings release date [−1, +1], where day 0 is the earnings release date (*XRET_PRELIM*). The second begins on day +2 through one day after the earnings announcement date of the subsequent quarter (*XRET_DRIFT*). We use *XRET_PRELIM* to complement the earnings surprise in case additional information is released in the preliminary earnings announcement. As we did for *SUE*, we rank *XRET_PRELIM* within a calendar quarter into quintiles, divide the rank by 4, and subtract 0.5. The longer return window is a standard definition of the drift return. We calculate abnormal return as the buy-and-hold return on the stock minus the value-weighted buy-and-hold return on all stocks of the same size (three groups), book/market ratio (B/M; three groups), and 11-month momentum (three groups).

The initial analysis we performed on the conference call transcripts involved counting the number of positive words (POS) and negative words (NEG) according to Loughran and McDonald (2011). For each transcript, we calculated the word count tone as (POS − NEG)/(POS + NEG). We then calculated the word-count tone change variable as the transcript tone minus the average tone of all available transcripts for this company in the prior 370 days (*TONE_CH_L&M*). Thus, the tone change was a number in the range of [−2, +2]. In the following, we provide evidence about the incremental contribution of *TONE_CH_L&M* to the drift return beyond the earnings surprise and the short-window return around the earnings announcement.

To assess the contribution of using Amenity's software plus our rule writing, we began by writing

[3] The subtraction of the average differences adjusts for cases in which earnings grow (or decline) by a constant amount each period.

Exhibit 1
Quarterly Earnings Conference Call Transcripts



Source: Thomson Reuters conference call transcripts, Compustat Point-in-Time data, CRSP return data, and authors' analysis.

additional rules on six areas of interest to us. One of these areas included operational issues discussed by management or analysts. For example, we identified any problems in distributing products, sourcing raw materials, labor strikes, and so on and created specific rules to identify such events under the heading of *operational problems*. We added approximately 500 rules to the roughly 3,600 rules that Amenity already had already written to capture events. Using our own weights for these rules, we obtained a new tone score for each transcript based on a weighted combination of sentiment scores and event scores. In addition, we compiled a list of euphemisms that management or analysts used on the conference call, such as *headwinds*, *speedbumps, and hiccups,* (Suslava 2016), and created specific rules to identify those. We added the euphemisms score to the combined sentiment and events score and calculated a total tone score as (POS − NEG)/(POS + NEG). As before, we focused on the tone change variable by subtracting the average tone of all available earnings transcripts in the prior 370 days (*TONE_CH_AM*).

Results

Exhibit 1 shows the number of transcripts per quarter for the period of our analysis, where we have earnings surprise, returns, and tone change variables. We begin with about 890 transcripts in 2003, exceed 2,000 in 2009, and remain at that level through 2016, with slight variations. Thus, we have good representation of many firms in the investable universe.

Exhibit 2 reports summary statistics on our main variables. It shows that firms with conference calls tend to be larger, have comparable B/M ratios to the universe, have median abnormal returns that are negative, and have median earnings surprises and tone change variables that are positive.

Exhibit 3 provides the correlations among the variables used in the cross-sectional regressions in the following. Recall that *SUE* and *XRET_PRELIM* were transformed to variables between −0.5 and +0.5, so their coefficients in the regression will reflect the return on the hedge portfolio that is long the top (most positive) quintile minus the bottom (most negative quintile). We follow a similar procedure for the two tone change variables. As can be seen from the exhibit, the drift return is positively and significantly associated with all the independent variables, but its highest correlation is with *TONE_CH_AM*, followed by *TONE_CH_L&M*, *XRET_PRELIM*, and *SUE*. Note the high correlation (50%) between the *TONE_CH_AM* and *TONE_CH_L&M* variables. However, the *TONE_CH_AM* variables did not have high correlations with the earnings surprises or the short-window abnormal returns,

EXHIBIT 2
Descriptive Statistics

Variable	*N*	Mean	Median	Std. Dev.	Q1	Q3
TONE_CH_L&M	101,125	0.006	0.011	0.156	–0.092	0.109
TONE_CH_AM	101,125	0.000	0.007	0.196	–0.119	0.126
SUE	101,125	–9.102	–0.026	2,836.380	–0.717	0.637
MKT	101,125	6,756.79	1,211.860	23,759.85	393.02	3,829.41
BM	101,125	0.610	0.481	0.615	0.284	0.760
XRET_PRELIM	101,125	0.241	0.110	8.626	–3.760	4.189
XRET_DRIFT	101,125	0.611	–0.102	20.129	–9.235	9.142

Notes: Exhibit 2 reports summary statistics for variables used in subsequent tests. TONE_CH_L&M *is the difference between the L&M tone in a company's conference call and the mean L&M tone in the company's conference calls held within the preceding 370 calendar days. L&M tone is a sentiment signal based on the Loughran and McDonald dictionary and calculated as the difference between the positive sentiment score and the negative sentiment score, scaled by the sum of the positive and the negative sentiment score.* TONE_CH_AM *is calculated in the same manner as* TONE_CH_L&M *but is based on the Amenity score, which consists of Amenity sentiment and events score, including events identified by the authors.* SUE *is calculated as the EPS reported in the earnings release minus the EPS reported in the same quarter of the prior year and minus the average same-quarter EPS differences in the prior eight quarters, scaled by the standard deviation of the same-quarter EPS differences during the prior eight quarters.* MKT *is the market value of equity at the conference call date.* BM *is shareholders' equity divided by pre-earnings announcement market value.* XRET_PRELIM *is the buy-and-hold return on a stock minus the average return on a matched size–B/M–momentum portfolio in the interval [–1, +1], where day 0 is the preliminary earnings announcement date.* XRET_DRIFT *is the buy-and-hold return on a stock minus the average return on a matched size–B/M–momentum portfolio from two days after the preliminary earnings announcement date through one day after the subsequent quarter's preliminary earnings announcement.*

Source: Thomson Reuters conference call transcripts, authors' analysis using Amenity software, CRSP returns, and Compustat Point-in-Time earnings.

EXHIBIT 3
Pearson Correlations

	XRET_DRIFT	*SUE*	*XRET_PRELIM*	*TONE_CH_L&M*	*TONE_CH_AM*
XRET_DRIFT	1				
SUE	0.027***	1			
XRET_PRELIM	0.033***	0.160***	1		
TONE_CH_L&M	0.048***	0.212***	0.228***	1	
TONE_CH_AM	0.036***	0.157***	0.176***	0.503***	1

Notes: This exhibit reports Pearson correlations for our testing variables. All variables are defined in the footnotes to Exhibit 2.

**** denotes significance at the 1% level.*

Source: Thomson Reuters conference call transcripts, authors' analysis using Amenity software, CRSP returns, and Compustat Point-in-Time earnings.

indicating a sufficiently different potential source of information.

Exhibit 4 contains the results of quarterly cross-sectional regressions of the abnormal drift returns on various independent variables in the manner of Fama and MacBeth (1973). We report the average quarterly coefficient and its *t*-statistic over the 55 quarters used in the study.

The first specification shows the contribution of earnings surprises and the short-window abnormal returns around the earnings announcement to explain the drift return. Both are positive and statistically significant, as we expected based on prior studies. Together they yield about 3% of abnormal return per quarter, again in line with prior studies. In the second specification, we added the L&M tone change to the prior two explanatory variables. All three independent variables are still positive and significant, but now the word-count tone change contributes the most to the drift, with a quarterly hedge return of 1.66%, bringing the total drift return to 4.19%, a significant increase in return. Our third specification introduces the Amenity tone change

EXHIBIT 4

Fama–MacBeth Regressions of Excess Returns

Variables	*XRET_ PRELIM*	*XRET_ PRELIM*	*XRET_ DRIFT*
INTERCEPT	0.6631**	0.6626**	0.6622**
	(2.10)	(2.10)	(2.10)
TONE_CH_L&M		1.6567***	
		(7.65)	
TONE_CH_AM			2.2657***
			(10.18)
SUE	1.3282***	1.1039***	0.9071**
	(3.70)	(3.09)	(2.57)
XRET_PRELIM	1.6763***	1.4319***	1.2561***
	(5.67)	(4.87)	(4.31)
No. Obs.	101,125	101,125	101,125
No. Regressions	55	55	55
Quarter FE	YES	YES	YES
R^2	0.54%	0.71%	0.78%

*Notes: The exhibit presents the results of Fama–MacBeth style quarterly cross-sectional regressions of the excess drift buy-and-hold return (*XRET_DRIFT*) after the conference call dates. For regression analysis* SUE *and* XRET_PRELIM *are normalized between −0.5 and 0.5 by ranking them into the quintiles every fiscal quarter, dividing the rank by 4, and subtracting 0.5. All unscaled variables are defined in the footnotes to Exhibit 2. The t-statistics are reported in parentheses.*

******* *denotes significance at the 1% level.*

*** denotes significance at the 5% level.*

Source: Thomson Reuters conference call transcripts, authors' analysis using Amenity software, CRSP returns, and Compustat Point-in-Time earnings.

variable (*TONE_CH_AM*), which shows a hedge return of 2.27% per quarter, bringing the total contribution of the three variables to 4.45% per quarter. Thus, our effort of using Amenity's software and writing additional rules to capture our own events and euphemisms added another 26 bps of abnormal return per quarter.[4]

CONCLUSIONS

This study proposes an advanced approach to analyzing text and converting it into a numerical score, which is easy to implement with the right software. This approach determines the sentiment in the document, but more importantly, it identifies additional events of importance to the user. The software allows the writing of NLP rules by individuals who are not data scientists or NLP experts but are experts in the particular domain from which the text originated. This reduces the cost of writing rules to capture relevant events and makes the text mining process more efficient. Another advantage of this approach is that each firm can create its own rules to capture specific events from the same text used by other firms, but each will have its own secret sauce. Furthermore, each firm that uses such software will have incentive to continuously engage in writing new rules to capture new events of interest and improve its future returns. This same process can also be used to generate rules that capture specific items of interest, such as newly imposed tariffs (see Klevak et al. 2018) or sensitivity to operations in a specific country.

ACKNOWLEDGMENT

We thank Kuberre Systems for providing us the Charter Oak Point-in-Time data.

REFERENCES

Biddle, G. C., G. Hilary, and R. S. Verdi. 2009. "How Does Financial Reporting Quality Relate to Investment Efficiency?" *Journal of Accounting and Economics* 48: 112–131.

Bozanic, Z., L. Cheng, and T. Zach. 2018. "Soft Information in Loan Agreements." *Journal of Accounting, Auditing & Finance* 33 (1): 40–71.

Brochet, F., M. Loumioti, and G. Serafeim. 2015. "Speaking of the Short-Term: Disclosure Horizon and Managerial Myopia." *Review of Accounting Studies* 20 (3): 1122–1163.

Cohen, L., C. J. Malloy, and Q. Nguyen. 2015. "Lazy Prices." Working paper, SSRN, https://ssrn.com/abstract=1658471.

De Franco, G., O. K. Hope, D. Vyas, and Y. Zhou. 2015. "Analyst Report Readability." *Contemporary Accounting Research* 32: 76–104.

Fama, E. F., and J. D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81: 607–636.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal. 2010. "Management's Tone Change, Post Earnings Announcement Drift and Accruals." *Review of Accounting Studies* 15: 915–953.

[4] Controlling for size and B/M in the regressions did not change the relative order and the significance of the contributions.

Fishe, R. P. H., D. S. North, and A. Smith. 2014. "Words that Matter for Asset Pricing: The Case of IPOs." Working paper, SSRN, https://ssrn.com/abstract=2413934.

Henry, E. 2008. "Are Investors Influenced by How Earnings Press Releases Are Written?" *Journal of Business Communication* 45 (4): 363–407.

Hoberg, G., G. Phillips, and N. Prabhala. 2014. "Product Market Threats, Payouts, and Financial Flexibility." *The Journal of Finance* 69 (1): 293–324.

Klevak, J., J. Livnat, D. Pei, and K. Suslava. 2018. "'Fake' Tariff News: Is Corporate America Concerned with Trade Wars?" Working paper, SSRN, http://dx.doi.org/10.2139/ssrn.3207725.

Lehavy, R., F. Li, and K. Merkley. 2011. "The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts." *The Accounting Review* 86: 1087–1115.

Li, F. 2008. "Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45: 221–247.

——. 2010. "The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach." *Journal of Accounting Research* 48 (5): 1049–1102.

Loughran, T., and B. McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (1): 35–65.

Suslava, K. 2016. "Stiff Business Headwinds and Unchartered Economic Waters: The Use of Euphemisms in Earnings Conference Calls." Working paper, Rutgers Business School, SSRN, https://ssrn.com/abstract=2876819.

Tetlock, P. C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance* 63 (3): 1139–1167.

Zhou, D. 2018. "Do Numbers Speak Louder Than Words?" Working paper, City University of New York, Baruch College, SSRN, https://ssrn.com/abstract=2898595.

Disclaimer

QMA is the primary business name of Quantitative Management Associates, LLC. This article represents the views and opinions of the authors regarding the economic conditions, asset classes, or financial instruments referenced herein and not necessarily the views of QMA.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

Introducing Objective Benchmark-Based Attribution in Private Equity

SIDNEY C. PORTER AND SHERIDAN PORTER

SIDNEY C. PORTER
is the chief data scientist
at FEV Analytics Corp
in Kirkland, WA.
sid@FEVanalytics.com

SHERIDAN PORTER
is the chief product officer
at FEV Analytics Corp
in Kirkland, WA.
sheridan@FEVanalytics.com

In this article, we propose that in private equity, measurement of asset manager (general partner [GP]) skill should begin with a repeatable benchmark-based performance attribution, which is then extended to explicitly quantify sources of alpha. Furthermore, in this article, we lay out a framework for repeatable measurement of performance attribution. Modern proxy benchmarks form a key component of this framework by enabling public market information to systematically inform private equity performance.

For manager evaluation, benchmarks serve as a standard against which past performance is measured and compared. In private equity, creating such a standard is confounded by the absence of price for extended periods and by the vicissitude of exposures and cash flows a private equity fund will experience over its lifetime. Recent advances in data science technology now support a host of indexing capabilities that work within these constraints, allowing fully modernized benchmarking and performance evaluation.

Modern indexes, as described by Lo (2016), are similar in concept to the modern proxy benchmark by their systematic construction and ability to highlight systemic behavior of a set of stocks against a given factor(s). Such indexes are diverse, at the forefront of financial innovation, and—notable to this article—include non-market-cap-weighted (fundamental value) varieties.

Technology can now systematically and objectively determine a suitably stable set of stocks based on their similarity to a target asset, providing the basis for quantifying *systemic* elements of performance. The custom index, thus composed, closely resembles the size and sector style returns described by Sharpe (1992) and can be used to define the highly prized excess returns produced by the manager relative to that of a passive mix with the same style (see Exhibit 1).

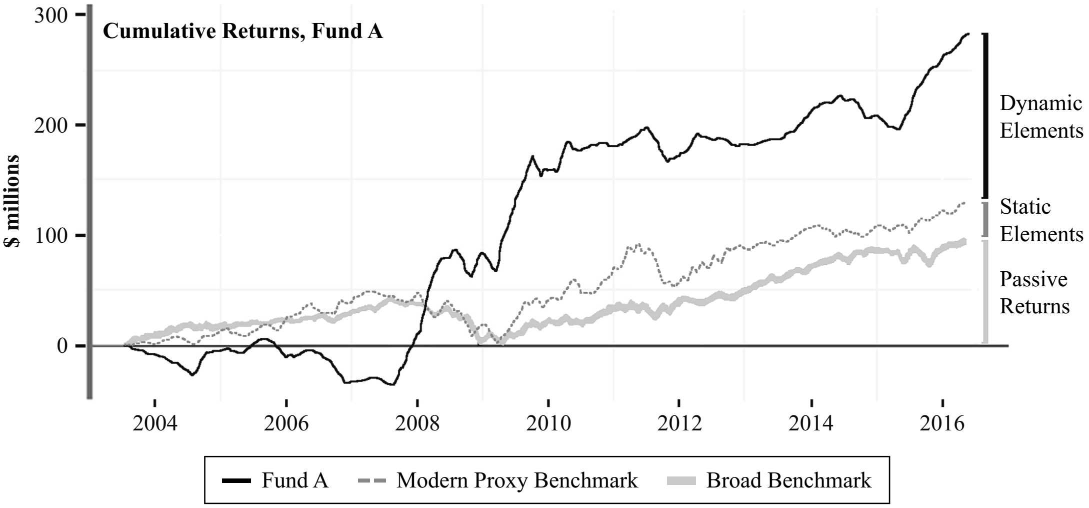
PROPERTIES OF MODERN PROXY BENCHMARKS

Modern proxy benchmarks are constructed as custom modern indexes but with two distinct (additional) properties: They satisfy fully the CFA Institute's criteria for a valid benchmark ("SAMURAI"),[1] and they are functionally ideal for performance comparison because they are objective, actually investable, and transparent. These properties are nontrivial to achieve and introduce two requisite quantities to the construction process: technical similarity and stability.

[1] "SAMURAI" stands for Specified in advance; Appropriate; Measurable; Unambiguous; Reflective of manager's universe; Accountable; Investable.

Exhibit 1

The Role of Modern Proxy Benchmarks in the Decomposition of Fund Performance



Notes: The modern proxy benchmark is a public peer set, adjusted for liquidity and control, reflective of the target entity's industry, economic size (not market cap), and interplay of financial ratios. It has 50–150 companies, objectively composed by technology for optimal specificity and stability.

Technical Similarity

Within the attribution taxonomy described by Sharpe (1992), and within our own related approach, *similarity* between the target asset and the benchmark is germane to its economic meaning. But what is similarity? Arguably, the most enduring markers of similarity are also the most fundamental: financial quantities and business descriptors of the companies within the funds (Arnott, Hsu, and Moore 2005).

The second question then becomes how it is measured. In this work, we develop a similarity measurement based on a combination of a company's fundamental components: its economic size[2] (as opposed to market cap or price), industry, and interplay of financial ratios. Similarity can then be implemented as a distance function, in which components are mapped along these three axes, with closer companies being more similar to the target company. This approach, which measures at the underlying company level as opposed to the fund level, is made possible by data science technology.

Technical similarity radically transforms similarity from a subjective notion (inevitably biased toward familiar companies[3]) into an objective empirical quantity.

Stability

The purpose of the benchmark is to capture systemic behavior against which idiosyncratic behavior (i.e., excess returns) may be measured. It is critical then that its constituents be sufficiently numerous to aggregate to a systemic representation and attenuate idiosyncratic behavior. A benchmark's ability to reliably capture systemic behavior describes the property of *stability.*

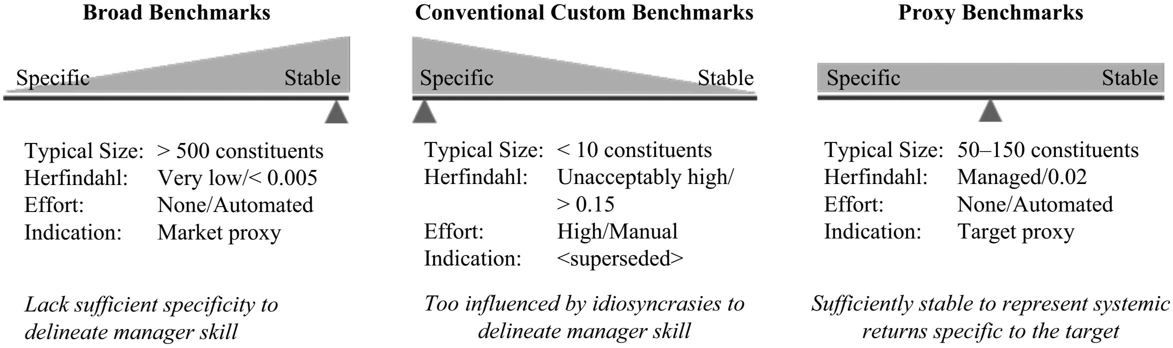
We propose that stability is an essential property of a benchmark, central to the integrity of an excess returns measure. An unstable benchmark can be unduly influenced by a subset (or one) of its constituents, misrepresenting systemic returns and causing unavoidable confusion of constituent and target idiosyncrasies. An unstable benchmark therefore makes a repeatable definition of manager alpha unattainable.

[2] Economic size—what we call *fundamental economic value*—is itself a mathematically valid indexation of a company's intrinsic value as measured by statistical models. The predictive accuracy (to market value) of these statistical models is 0.813 (the R^2 statistic) and is equally accurate across the public–private divide.

[3] For further reading on familiarity bias, see Heath and Tversky (1991).

EXHIBIT 2

Benchmarking Private Equity—Balancing a Fundamental Trade-Off between Specificity and Stability (and effort)



	Broad Benchmarks	Conventional Custom Benchmarks	Proxy Benchmarks
Typical Size:	> 500 constituents	< 10 constituents	50–150 constituents
Herfindahl:	Very low/< 0.005	Unacceptably high/ > 0.15	Managed/0.02
Effort:	None/Automated	High/Manual	None/Automated
Indication:	Market proxy	<superseded>	Target proxy

Lack sufficient specificity to delineate manager skill

Too influenced by idiosyncrasies to delineate manager skill

Sufficiently stable to represent systemic returns specific to the target

Appropriating the Herfindahl Index to Measure Stability

The Herfindahl (also called the Herfindahl–Hirschman Index, or HHI) is a commonly accepted device to measure market concentration.[4] However, the Herfindahl can be appropriated to calibrate stock exposures in benchmark construction and determine the number of constituents needed to control for benchmark stability. Higher concentrations are represented by a larger Herfindahl. For example, a benchmark containing fewer than 10 constituents—not uncommon for custom benchmarks in private equity—approximates to a Herfindahl between 0.10 and 0.25, sometimes even larger. At a Herfindahl of 0.05—far more stable than 0.15—a benchmark produces unacceptable variances and is at high risk of being materially skewed by idiosyncratic behavior of its constituents.

Our research finds evidence of concentration thresholds in a benchmark that provide guidance for a stability range. The upper threshold corresponds to a Herfindahl of 0.05, as previously stated; however, the lower threshold should be considered in the context of what the benchmark is trying to capture: systemic behavior specific to the target entity.

The Specificity–Stability Spectrum

In general, by making the benchmark more stable, we are trading off specificity to the target asset(s). Conventional practice in private equity bunches benchmarks at either end of this spectrum, in which a custom benchmark with (typically) fewer than 10 constituents is at one end, and a broad benchmark like the Russell 3000 is at the other (see Exhibit 2). Although practitioners may be aware that neither case is suitable for performance evaluation, gaining consensus on a larger set of subjectively chosen stocks and weighting them in an index is an extremely difficult and universally tiresome process with little clarity in the end.

The compelling "middle ground" of this spectrum, in which a benchmark is stable *and* specific to its target, typically requires between 50 and 150 constituents that are each monitored to continuously satisfy the requirement of technical similarity.

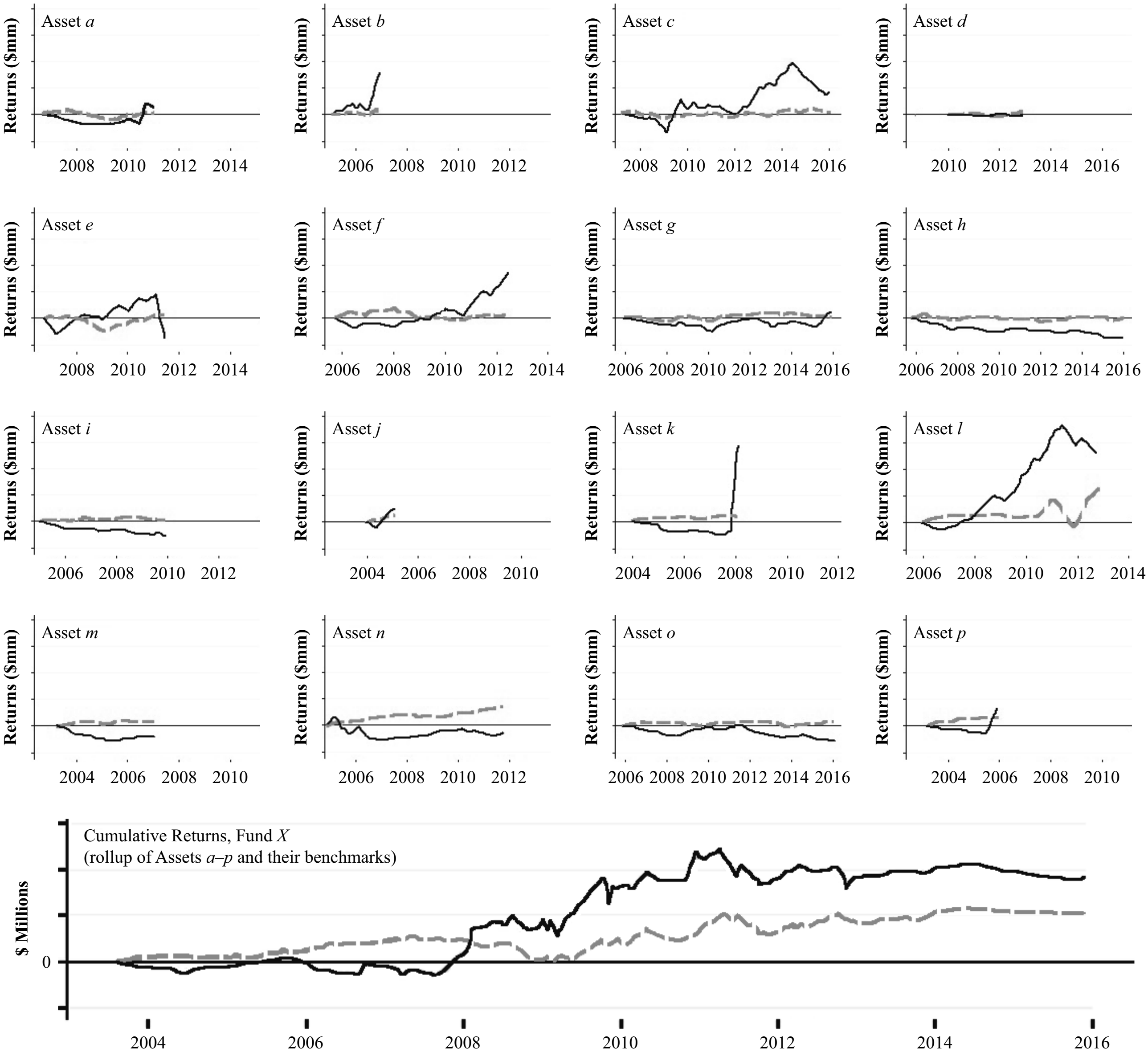
BENCHMARKING TECHNOLOGY

Practitioners will recognize that, in the absence of benchmarking technology, benchmark construction is an intensely subjective stock-picking process. Because of this difficulty, a (conventionally constructed) custom benchmark is typically aimed at the fund or manager level and composed of far fewer than the 50 constituents per target company needed for benchmark stability; however, this crude approach is a practical workaround rather than a theoretical consideration—a

[4] For example, the United States Department of Justice and the Federal Trade Commission consider the Herfindahl a measure of market concentration: https://www.justice.gov/atr/herfindahl-hirschman-index.

Exhibit 3

Cumulative Returns of Assets (dark line) and Their Modern Proxy Benchmarks (dashed), Rolled Up to Their Fund



constraint readily removed by automated benchmarking technology.

The efficiency of technology makes robust benchmarking at the company level feasible and excess returns calculable on a per company basis (see Exhibit 3). Alpha can be "rolled up" outside of the fund wrapper and examined in compelling ways (i.e., by industry, company size, and even deal team). As represented by Korteweg and Sørensen (2014), this intelligence—afforded by technology—may unlock greater predictive power in the manager evaluation process.

Data science technology is already behind remarkable innovation in finance and investing, so it is perhaps unsurprising that it also has the power to unlock new capabilities in private equity. Modern proxy benchmarks and benchmark-based attribution are one such example, with precision and objectivity offering new capabilities, such as

- Quantification and indexation (i.e., direct comparison) of manager skill
- Consistent delineation of systemic returns (i.e., company size, industry, region)
- Emergence of risk metrics and portfolio construction
- Consideration of dynamic and static elements of active returns in fee structures.

Maintenance of the Modernized Benchmark

The changing composition of a private equity fund as assets enter and exit is further complicated (in terms of performance attribution) by asset bolt-ons, divestments, and restructuring during the holding period. The changing nature of a fund's exposures and the economic size of its assets is the result of GP operational control and is constitutive to private equity investing; however, despite these changes being implicit to GP skill, the fixed benchmark typical in private equity is indifferent to it.

Through its efficiency, benchmarking technology makes it possible to preserve specificity and stability over time in a systematic manner. In so doing, the GP's skill is more accurately captured because it is always being gauged appropriately. To illustrate, imagine a merger of two similar assets. The strategy is to drive operating efficiency, strengthen exit multiples, and generate higher total returns. If the benchmark did not change to reflect the merged entity, then the GP would be gauged against companies that had lower operating efficiencies, with no understanding of the potential returns or the opportunity as indicated by its similarly sized peers. If the benchmark is adjusted to reflect the larger entity, then the opportunity (and by extension the opportunity cost) is made measurable.

To capture the more predictive components of performance (i.e., GP skill), the benchmark must be meticulously maintained to reflect the actions of the GP as it changes the composition and nature of the fund.

APPROACH

The exactitude of the approach enabled by technology starts with measuring at the holdings or asset level. For every company inside the fund, a proxy benchmark consisting of between 50 and 150 technically similar companies (peers) is systematically constructed. With today's computing power, technology can evaluate the similarity of 6,000 companies in less than six seconds, making a scan of all public exchanges (including over-the-counter stocks) for fundamentally similar constituents a rapid process.

Constituents with greater similarity to the target company are weighted accordingly by the number of their shares held in the benchmark. Therefore, for each private company in a fund, benchmark technology creates a mathematically precise cloud of public peers. The company-level public benchmarks are then rolled up to the fund level to create the equivalent of a virtual synthetic investment.

A private equity fund can be thought of as a series of stakes in various companies over time. Change in fund composition or the nature of any one component company is generally accompanied by a cash flow event; a stake is the period between cash flow events (see Exhibit 4). Each stake is treated as its own virtual synthetic investment that is completely cashed out and immediately reinvested as a new stake at cash flow events. The synthetic fund is actively managed by the technology to mirror the changed nature of every stake for the life of the fund.

In this way, the GP's actions are precisely replicated in the benchmark. Furthermore, implied returns may be captured without forcing the benchmark to engineer its way around going short—a nontrivial issue that various public market equivalent (PME[5]) methods handle differently.

Much like the S&P 500 is not a fixed basket of stocks, technically similar companies are actively moved in and out of the benchmark based on a consistent methodology that is specified in advance. Although computationally intensive, rigorous ex post analysis of a private equity fund for the purpose of manager evaluation can be completed in seconds.

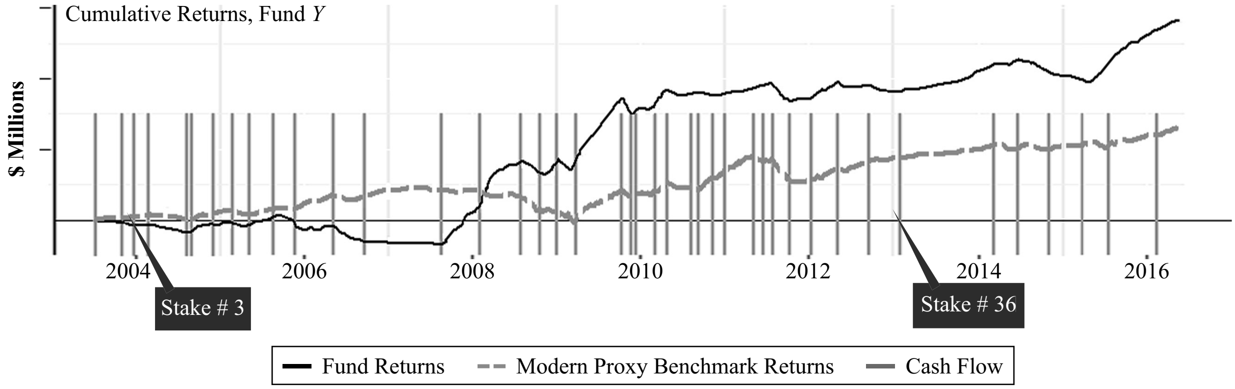
FRAMEWORK FOR DECOMPOSITION OF RETURNS

An asset's cumulative returns are modeled from unsmoothed cash flows and explicitly account for

[5] PME—also known as the Index Comparison Method, or ICM—analysis was first proposed by Long and Nickels (1996); subsequent flavors have worked toward minimizing technical issues, including the public index going short.

EXHIBIT 4

The Benchmark Is Adjusted Synchronous with Every Cash Flow to Reflect the Actions of the GP



Note: Each vertical line represents a cash flow event, indicative of a change in the fund's exposures or component company size.

economic value growth and market fluctuations with an objective and measurably accurate continuous pricing mechanism. Data science technology provides such a mechanism.[6]

Active/Passive Decomposition

Market returns are indicated by a broad index such as the ACWI or Russell 3000 or by returns of the investor's own public equities portfolio. The market, so defined, marks the active/passive decomposition of the private asset's returns (see Exhibit 5).

Limitations of the PME in Performance Attribution

Broad benchmarks differ substantially in similarity to a given fund and do not offer a meaningful marker of manager skill. For example, consider the Russell 3000 gaining 20% while the technology sector goes down 15%. A fund (or GP) weighted in technology assets would appear unskilled but when the technology sector recovers it would look highly skilled.

This simple example illustrates the major limitation of the PME approach to performance attribution: It is simultaneously driven and limited by its benchmark. It is the benchmark that drives the legitimacy of performance attribution (Cumming, Hass, and Schweizer 2013) in the dimension of time and in the dimension of similarity/stability. As previously discussed, a meaningful measure of alpha is dependent on a systematic and robust quantification of systemic performance that is *always* technically similar to the target asset.

Decomposition of Active Returns

From the active/passive decomposition, active returns are further decomposed into static and dynamic elements by the modern proxy benchmark.

Static Elements of Active Returns

Modern proxy benchmarks capture the systemic portion of active returns, namely the contribution of industry, asset size range, and region. Therefore, cumulative returns of the proxy benchmark in excess of passive returns approximate the economic value of a GP's exposure to these systemic factors.

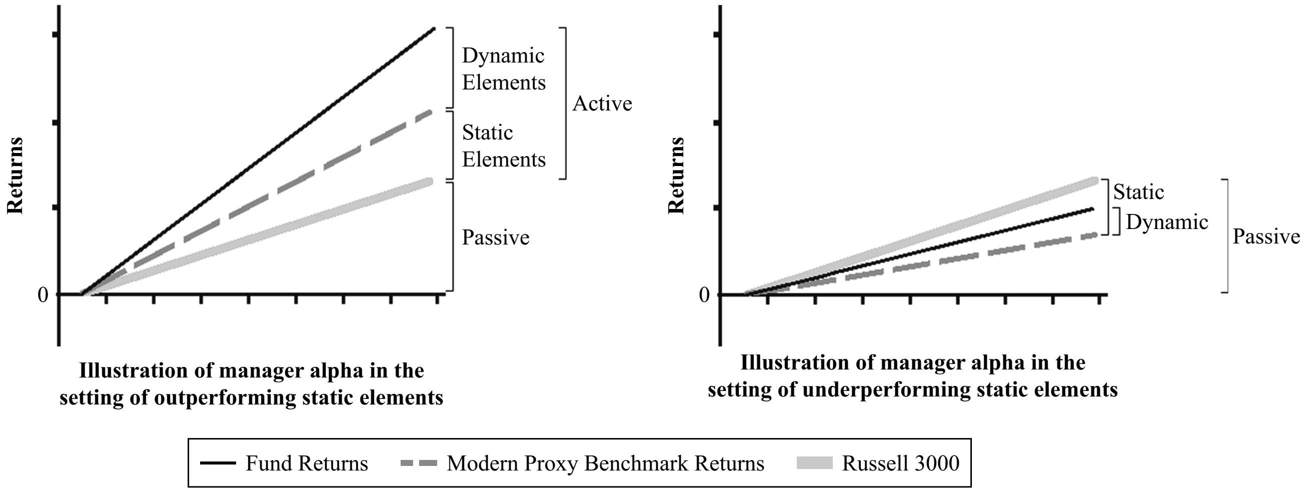
For most GPs, the systemic factors to which their funds provide limited partners (LPs) access are relatively static over time. It could be argued then that the value added by static elements reflects a strategic bet on beta made *explicitly by the LP*. Certainly, an objective segmentation of systemic returns gives perspective to the concern of many LPs that they may be paying high fees for beta.

Nevertheless, as framed by Lo (2007), there exists a potential economic value in a GP's weighting of certain

[6] See the section "Objective Measurement of Interim Valuations" later in this article for elaboration on the continuous pricing mechanism.

Exhibit 5

Framework for Decomposition of Returns



Notes: Advanced applications of data science technology allow modeling of private equity performance in time series and the componentization of returns. The structure of these components, shown in the exhibit, allows independent consideration and comparison of market returns, the investor's choice of systemic exposures, and the GP's skill in producing excess returns. Within this framework, manager alpha is excess return from the proxy benchmark, provided the proxy benchmark exhibits the properties of stability and technical similarity and is actively maintained to continuously mirror the fund.

exposures within a given fund. In other words, the degree to which a GP uses prior information to "launch boats onto rising tides" may be considered the private equity equivalent of timing. Strategic weighting of static elements by the GP is an interesting source of potential value, further examination of which is supported by our attribution framework.

Dynamic Elements of Active Returns: Manager Skill

If the modern proxy benchmark represents systemic returns, then excess returns are idiosyncratic in origin. The GP's operational control of the asset makes these idiosyncrasies a genuine approximation of GP skill. Therefore, within an attribution framework, excess returns to the proxy benchmark equals manager alpha, which equals manager skill.

The control levers available to a GP—namely asset growth, financial engineering, and transaction premiums[7]—are applied dynamically in the marketplace by the GP to maximize yield given contemporaneous market conditions and opportunities. The GP's ability to forecast and execute winning strategies underwrites how it works the control levers, and alpha is its measure.

Because of their predictive nature, dynamic elements are worthy of greater examination. Modern proxy benchmarks support an attribution framework capable of extending to a novel quantification of contributions to alpha by source. These sources are described fully in a forthcoming paper by Porter and Porter on performance attribution measurement in private equity.

APPLICATION OF MODERN BENCHMARK-BASED ATTRIBUTION TO MANAGER EVALUATION

Benchmark-based attribution using technology provides an objective, consistent decomposition of performance ex post that allows simultaneous comparison between GPs and to the market (see Exhibit 6). This allows the LP to index investment proposals by the

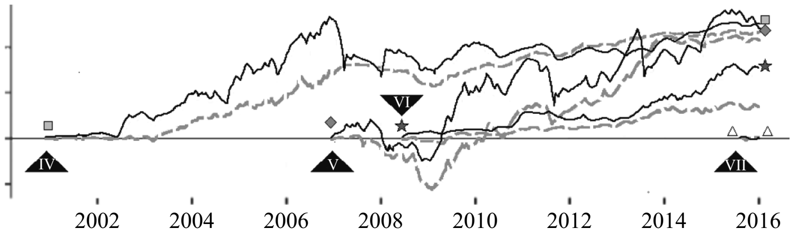
[7] We measure the premium paid on the asset (company) at entry and exit as per $Market\ \ premium = \frac{(Price - FEV)}{FEV}$.

The asset's premium is mapped to a distribution of the modern proxy benchmark's component firm premiums at the exact corresponding times. This determines whether the GP paid a higher or lower premium relative to the market at each transaction (i.e., buy low and sell high) and whether the premium percentile moved favorably between transactions.

Exhibit 6

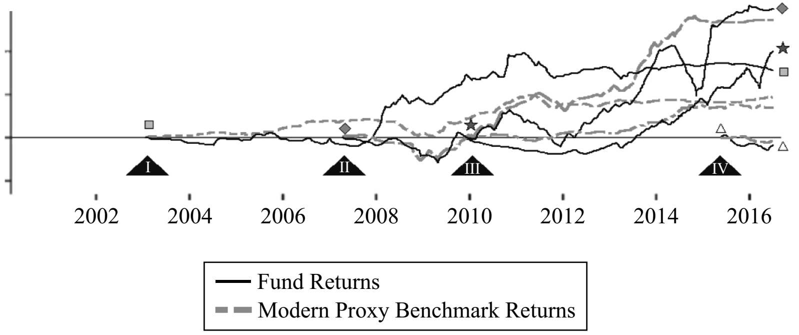
Comparing Alpha across Funds and Between GPs

Cumulative Returns, GP(A)



Fund	Vintage	% Realized	Fund Alpha	Overall Alpha
IV □	2001	93%	1.3%	1.8 %
V ◆	2006	71%	0.8%	
VI ★	2008	47%	11.6%	
VII △	2015	0%	21.9%	

Cumulative Returns, GP(B)



Fund	Vintage	% Realized	Fund Alpha	Overall Alpha
I □	2003	68%	5.5%	4.4 %
II ◆	2007	60%	1.8%	
III ★	2009	26%	10.1%	
IV △	2015	0%	–12.1%	

factor that is widely accepted as most predictive of future performance: GP skill.

Reducing the Influence of Gameable Metrics

The cash flow–weighted metrics that dominate the industry impose severe limitations on GP evaluation because, fundamentally, they do not offer predictive intelligence (Porter and Porter 2018). In the absence of viable alternative performance metrics, the industry has nonetheless drawn conclusions from, and made inferences based on, internal rates of return (IRRs) to inform investment decision-making.

Unfortunately, the gameable nature of IRR inputs has opened the door to it being used for the purpose of *mis*information. For instance, peer ranking is most commonly based on fund-level IRR as compiled by numerous third parties. A "top quartile" fund has an IRR in the top 25% of its vintage year cohort.[8] Aside from issues pertaining to data completeness and selection bias of the proprietary database, the IRR cannot determine manager skill or rank performance. Research by Gottschalg and Phalippou (2007) illustrated this. They found that the IRR materially misstated returns, even on fully realized funds, and that rankings were not preserved when reinvestment rates were adjusted using the modified IRR function (MIRR).[9] For example, the top two ranked funds by the MIRR did not make an appearance in the top 10 funds ranked by the IRR.

The potential for distortion, even misrepresentation, is increased when either IRR or MIRR analysis involves current assets (i.e., interim funds) because of the subjectivity of underlying asset valuations. Discussed widely in the literature and echoed by the Securities and Exchange Commission repeatedly from 2013 (Karpati 2013; Bowden 2014), GPs generally inflate valuations to peak IRR at the time of fundraising (Barber and Yasuda 2016). Subjectivity means that valuations can be (and generally are) manipulated for the express purpose of influencing the investor

[8] The definition of *vintage year* varies among researchers. It may mean the date of fund closing, the date of first capital call, or the fund's date of first entry. This wiggle room can potentially be used to identify a fund with a cohort in which it ranks more favorably.

[9] $$MIRR = \sqrt[n]{\frac{FV\ (Positive\ cashflows,\ Reinvestment\ rate)}{-PV\ (Negative\ cashflows,\ Finance\ rate)}} - 1.$$

(Barber and Yasuda 2016)—despite the fact that interim valuations are typically produced by independent valuation experts.[10]

Against a backdrop of at-times flagrant gaming of the IRR,[11] the industry has nonetheless been driven to infer that successive top-quartile funds indicate GP skill (Sensoy, Wang, and Weisbach 2014). As an illustration of the degree to which this persistence is prized: A GP with an existing top-quartile buyout fund can raise a follow-on fund 5.7 times faster than a bottom-quartile counterpart (Barber and Yasuda 2016). However, the IRR cannot discern between skill and luck. It *is* possible to be lucky on multiple occasions, especially if the market is generally rising and competition does not increase. These conditions unfortunately do not describe the current or likely future scenario of private equity, and at least for buyouts, persistence as an indicator of future performance has already evaporated (Braun, Jenkinson, and Stoff [2015]).

BENCHMARK-BASED ATTRIBUTION WITHIN AN OBJECTIVE MEASUREMENT FRAMEWORK

The decomposition of returns discussed in this article provides a clarity of analysis because it is created by repeatable measurement. Creation of an objective measurement framework necessarily begins with high-accuracy valuation technology.

Objective Measurement of Interim Valuations

The approach to interim valuation underpinning the framework described herein involves the capture of three sets of information: the economic size of the target company, company idiosyncratic factors, and market movements.

1. The *economic size* of the target company is measured by its fundamental economic value (FEV). The FEV is a unique size measure produced by data science technology, with certain properties:
   - i. High predictive accuracy of market price ($R^2 = 0.813$)
   - ii. Objective and systematic, requiring no forward-looking or subjective quantities
   - iii. Simultaneously measures public and private companies with equal accuracy on a standardized basis
   - iv. Fundamentals-driven
   - v. Automated
2. *Company idiosyncratic factors* are estimated by computing the premium of each company from its entry price. By also computing the premium distribution of the proxy benchmark at that date, the distance of the company's premium from the median premium can be attributed to company idiosyncratic factors, which might include brand power, customer base, assets, growth potential, and so on. The premium percentile of the company purchase price represents the overall impact of idiosyncratic factors at entry. Under the assumption that these idiosyncratic sources of value vary slowly, the premium percentile can be used as a proxy for them.
3. *Market movements* are captured by the modern proxy benchmark. Together with the FEV, it is possible to simultaneously measure pricing changes and growth in the economic ecosystem of each company and roll these up to the fund and GP levels.

Interim valuations are calculated by computing the FEV of each company's most recently available financial information, calculating the premium distribution of the proxy benchmark on the valuation date, using the

[10] According to the (formerly named) Financial Services Authority (2006) (FSA Discussion Paper 06/06), private equity firms number among the largest clients for most big financial intermediaries—banks, lawyers, accountants, management consultants—creating the potential for moral hazard. In particular, the revenue stream a service provider receives from a private equity firm "may cause them to consider actions that they would normally discount." For example, one (unnamed) bank earned almost €900 million from its private equity–related activities in a year, whereas another bank was shown to generate over 50% of its income from private equity.

[11] In addition to manipulating valuations, the IRR can be gamed by delaying capital calls. GPs can finance acquisitions using short-term debt, delaying capital calls by months or potentially years; however, if the debt is secured against LP commitments, the LP bears default risk (off balance sheet). Although this is not an illegal practice, it nonetheless has the effect of juicing the IRR, which is then used to substantiate the skill and fees of the GPs in their marketing process.

distribution and the premium percentile to estimate the market premium of the company, and then deriving *Fair value* = *FEV* × (1 + *Estimated market premium*).

This interim valuation method is wholly objective and has been tested on more than 100,000 public company quarterly valuation estimates. The tests show that the method is unbiased, with a median absolute percent error of less than 0.05.

Outputs of the Objective Modern Framework Beneficial to GP Performance Attribution Analysis

- An objectively constructed and investable benchmark (full SAMURAI compliance)
- A high-integrity benchmark in terms of specificity and robustness—that is, a robust public proxy
- An efficient and exacting mechanism for maintaining benchmark integrity over time
- An elegant method for calculating implied returns from the benchmark
- Rigorous delineation of dynamic and static elements of active returns
- The means to align performce compensation with the investment decision-making process

In addition, the objective basis of this framework permits an approach of continuous improvement and accountability to the manager evaluation process itself. Whether corroborating or challenging an investment narrative, objectivity enhances the overall probity of the process.

CONCLUSION

Benchmark-based attribution as described throughout this article disambiguates the quantification and comparison of manager skill.

Technology efficiencies radically change our understanding of what is practical; LPs can measure and index manager skill on a far broader scale than what has been previously possible. For instance, early-stage rigorous analysis that incurs less cost, time, and effort may be implemented by LPs as a screening mechanism. Collectively, these analyses can inform an LP's understanding of the changing fund-raising climate over time, which in turn informs investment discipline and tactical excellence alongside pacing plans or other forms of investment pressure. The evaluation process can take on more of a funnel shape with a significantly wider catchment than is currently normal and eliminate the LP's equivalent of potentially harmful sample selection bias.

Finally, the systematic separation of active returns into static and dynamic elements by the modern proxy benchmark allows LPs to actively seek improved performance from both.

Although the performance metrics of public equities have surged in dimensionality and predictive power, a similar scientific quest has not flourished in private equity. The reason might have cultural underpinnings, but the absence of a continuous price mechanism has historically posed an intractable technical barrier. Advances in data science—a combination of computational power, statistical programming language, and the development of advanced mathematical models—have now allowed that barrier to be broken, engendering exciting opportunities for private equity, risk management, and portfolio management.

REFERENCES

Arnott, R., J. Hsu, and P. Moore. 2005. "Fundamental Indexation." *Financial Analysts Journal* 61 (2): 83–99.

Barber, B., and A. Yasuda. 2016. "Interim Fund Performance and Fundraising in Private Equity." SSRN, https://ssrn.com/abstract=2357570.

Bowden, A. 2014. "Spreading Sunshine in Private Equity." Speech at *PEI Private Fund Compliance Forum*, SEC, http://www.sec.gov/News/Speech/Detail/Speech/1370541735361.

Braun, R., T. Jenkinson, and I. Stoff. 2015. "How Persistent Is Private Equity Performance? Evidence from Deal-Level Data." SSRN, https://ssrn.com/abstract=2314400.

Cumming, D., L. Hass, and D. Schweizer. 2013. "Private Equity Benchmarks and Portfolio Optimization." *Journal of Banking and Finance* 37 (9): 3515–3528.

Financial Services Authority. 2006. "Private Equity: A Discussion of Risk and Regulatory Engagement." Discussion paper, http://fsa.gov.uk/pubs/discussion/dp06_06.pdf.

Gottschalg, O., and L. Phalippou. 2007. "The Truth About Private Equity Performance." Harvard Business Review, https://hbr.org/2007/12/the-truth-about-private-equity-performance.

Heath, C., and A. Tversky. 1991. "Preference and Belief: Ambiguity and Competence in Choice under Uncertainty." *Journal of Risk and Uncertainty* (4): 5–28.

Karpati, B. 2013. "Private Equity Enforcement Concerns." Speech at *Private Equity International Conference*, SEC, http://www.sec.gov/News/Speech/Detail/Speech/1365171492120.

Korteweg, A., and M. Sørensen. 2014. "Researchers: How Do You Find the Best Private Equity Funds?" Insights by Stanford Business, http://www.gsb.stanford.edu/insights/researchers-how-do-you-find-best-private-equity-funds.

Lo, A. 2007. "Where Do Alphas Come From? A New Measure of the Value of Active Investment Management." *Journal of Investment Management* 6 (3): 6–34.

——. 2016. "What Is an Index?" *The Journal of Portfolio Management* 42 (2): 21–36.

Long, A., and C. Nickels. 1996. "A Private Investment Benchmark." Presentation given at *AIMR Conference on Venture Capital Investing*, The University of Texas System, http://dns1.alignmentcapital.com/pdfs/research/icm_aimr_benchmark_1996.pdf.

Porter, S., and S. Porter. 2018. "The Many Theoretical Deficiencies of Performance Evaluation in Private Equity." SSRN, https://ssrn.com/abstract=3243147.

Sensoy, B., Y. Wang, and M. Weisbach. 2014. "Limited Partner Performance and the Maturing of the Private Equity Industry." *Journal of Financial Economics* 112: 320–343.

Sharpe, W. 1992. "Asset Allocation: Management Style and Performance Measurement." *The Journal of Portfolio Management* 18 (2): 7–19.

Dynamic Systemic Risk: *Networks in Data Science*

Sanjiv R. Das, Seoyoung Kim, and Daniel N. Ostrov

Sanjiv R. Das
is the William and Janice Terry professor of finance and data science at Santa Clara University in Santa Clara, CA.
srdas@scu.edu

Seoyoung Kim
is an associate professor of finance at Santa Clara University in Santa Clara, CA.
srkim@scu.edu

Daniel N. Ostrov
is a professor of mathematics at Santa Clara University in Santa Clara, CA.
dostrov@scu.edu

Systemic risk arises from the confluence of two effects. First, individual financial institutions (FIs) experience increases in the likelihood of default. Second, these degradations in credit quality are transmitted through the connectedness of these institutions. The framework in this article explicitly models the contributions of both of these drivers of systemic risk. By embedding these constructs in a data science model drawn from the field of social networks, we are able to construct a novel measure of systemic risk.

The Dodd–Frank Act (2010) defined a *systemically important FI* (SIFI) as any FI that is (1) large, (2) complex, (3) connected to other FIs, and (4) critical, in that it provides hard-to-substitute services to the financial system.[1] The Act did not recommend a systemic risk-scoring approach. This article provides objective models to determine SIFIs and to calculate a composite systemic risk score.

The Merton (1974) model provides an elegant way to use option pricing theory to determine the credit quality of a single firm (i.e., its term structure of credit spreads and the term structure of the probability of default [PD] for different horizons). We demonstrate how the model may be extended to a network of connected FIs, including a metric for the systemic risk of these firms that evolves over time. Therefore, this article provides an example of the power of combining mathematical finance with network science.

Our systemic risk measure has two primary attributes: (1) aggregation—that is, our metric combines risk across all firms and all connections between firms in the system to produce a summary systemic risk number that may be measured and tracked over time; and (2) attribution—how systemic risk can be mathematically analyzed to measure the sources that contribute to overall system risk. The primary way we want to understand attribution is through an *institution risk measure*, which determines the risk contributions from each firm so that the extent to which a single firm contributes to systemic risk at any point in time is quantifiable. A secondary way to look at attribution is to compute a *connectedness risk measure*, which determines the risk contributions from each pairwise link between two firms at any point in time.

CONTRAST WITH EXTANT APPROACHES

Current approaches to measuring systemic risk include the systemic expected shortfall (SES) measure of Acharya et al. (2017)[2]; the conditional value at risk (CoVaR)

[1] See also the literature analysis of Silva, Kimura, and Sobreiro (2017) for a conceptual overview and definition of systemic financial risk.

[2] See the extensive research in this class of models at Rob Engle's V-Lab at NYU: https://vlab.stern.nyu.edu/.

measure of Adrian and Brunnermeier (2016); the construction of FI networks using bivariate Granger causality regressions from Billio et al. (2012) (and a more general framework from Merton et al. 2013); the distressed insurance premium measure of Huang, Zhou, and Zhu (2012) and Black et al. (2016); the absorption ratio of Kritzman et al. (2011); the system value at risk of Bluhm and Krahnen (2014); the credit default swap (CDS)-based metric of interconnectedness used by Abbass et al. (2016); and the calculation of capital charges required to insure against unexpected losses as from Avramidis and Pasiouras (2015).

These approaches predominantly employ the correlation matrix of equity returns to develop their measures. A recent comprehensive article by Giglio, Kelly, and Pruitt (2016) examines 19 systemic risk metrics for the US economy and finds that these measures collectively are predictive of heightened left-tail economic outcomes. Furthermore, a dimension reduction approach creates a composite systemic risk measure that performs well in forecasts. Unlike the measure in this article, these 19 metrics do not exploit network analysis. All measures cited are mostly return based, and these have been criticized by Löffler and Rapauch (2018) as being subject to gaming in that a bank may cause the systemic risk measure to rise, while, at the same time, having its own contribution fall. These spillover issues do not appear to be a problem in this article.

In contrast, Burdick et al. (2011) used semistructured archival data from the Securities and Exchange Commission and Federal Deposit Insurance Corporation to construct a co-lending network and then used network analysis to determine which banks pose the greatest risk to the system. Finally, Das (2016) combined credit and network information to construct aggregate systemic risk metrics that are decomposable and may be measured over time. The unifying theme across these models is to offer static snapshots of the network of FIs at various points in time. This article is a stochastic dynamic extension of the Das (2016) model.

STOCHASTIC DYNAMICS IN A NETWORK MODEL

We extend these static network models by including stochastic dynamics for the assets of the financial firms in the model. This is where the Merton (1974) model becomes useful. We give this model the moniker *Merton on a network*. This model uses geometric Brownian motion as the stochastic process for each FI's underlying assets. That is, for the *n* FIs in the system, we have

$$da_i = \mu_i a_i \, dt + v_i a_i \, dB_i, \qquad i = 1, 2, \ldots, n \tag{1}$$

$$da_i \, da_j = \rho_{ij} \, dt, \qquad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, n \tag{2}$$

Here μ_i is the *i*th FI's expected growth rate, and v_i is its volatility (both annualized). The asset movement of FIs *i* and *j* are correlated through the coefficient ρ_{ij}.

Assuming that the *i*th FI has a face value of debt D_i with maturity *T*, Merton's model established that the FI's equity, E_i, is a call option on the assets:

$$E_i = a_i \Phi(d_{1,i}) - D_i e^{-r_f T} \Phi(d_{2,i}) \tag{3}$$

$$d_{1,i} = \frac{\ln(a_i / D_i) + (r_f + v_i^2/2)T}{v_i \sqrt{T}} \tag{4}$$

$$d_{2,i} = d_1 - v_i \sqrt{T} = \frac{\ln(a_i / D_i) + (r_f - v_i^2/2)T}{v_i \sqrt{T}} \tag{5}$$

where r_f is the risk-free rate of interest (annualized), and $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx$ is the cumulative standard normal distribution function. Merton's model also shows that the volatility of equity is

$$\sigma_i = v_i \frac{\partial E_i}{\partial a_i} \frac{a_i}{E_i} \tag{6}$$

Because a_i and v_i are not directly observable in the market, but E_i and σ_i are, the pair of Equations 3 and 6 may be solved simultaneously to determine the values of a_i and v_i for each *i* at any time, *t*. These values, as we will see later, allow us to obtain the one-year probability of default (PD) for each financial firm, denoted λ_i, at any given point in time.[3]

Our measure for systemic risk captures the size and PD of all FIs (from the Merton model) and combines this with a network of FI connectedness to construct

[3] In implementing our model as Merton on a network, our approach is distinct from those that infer risk-neutral PDs from CDS spreads on the referenced banks (e.g., as by Huang, Zhou, and Zhu 2012). We are also afforded greater flexibility in inferring what the PDs may be under varying market conditions.

one composite system-wide value. We exploit the stochastic structure of the asset movements of all FIs via Equations 1 and 2 to create a variety of constructions of the connectedness (network) matrix. Because the underlying assets are stochastic and correlated, so is the network; as a consequence, the systemic risk score is dynamic. In sum, we have a systemic risk measure that captures, over time, the size, risk, and connectedness of firms in the financial system.

The contagion literature has attempted to capture stochastic systemic risk by other means. Simulation of contagion networks is one approach; see Espinosa-Vega and Sole (2010), Upper (2011), and Hüser (2015). Bivalent networks of banks and assets have been simulated on data from Venezuela in another approach by Levy-Carciente et al. (2015). In our complementary approach, network and firm risk are endogenously generated through the underlying Merton (1974) model, which also offers a direct empirical implementation. To illustrate, we will later provide an example using a 20-year data sample from large, publicly traded FIs.

PRACTICAL VALUE OF THE MODEL

The models developed here have many features of interest to risk managers and regulators. First, each model produces a single number for the systemic risk in the economy. Second, the risk contribution of each institution in the system enables a risk ranking of these institutions. This ranking and the measures that determine them can help determine whether an institution is systemically important, the extent of additional supervision the institution should require, and how much the capital charge should be for the risks the institution poses to the system. Third, the risk contribution of each pairwise connection between two FIs can be measured. This allows regulators to determine which relationships between FIs are of greatest concern to the overall health of the system. Fourth, the models display several useful mathematical properties that we develop to indicate a good measure of systemic risk, as discussed in the next section. Fifth, the model's rich comparative statics may be used to examine various policy prescriptions for mitigating systemic risk.

In the next section, we introduce our general framework for systemic risk and the institution risk measure. This section also introduces four desirable properties for a systemic risk model. The following section introduces three models within the general framework that have similar structures. We discuss the institution risk measure for the three models and then show that each model possesses all four desirable properties. In the next section, we introduce our fourth model, which takes a different, although intuitive, structure from the first three models. Here we discuss both the institution risk measure and the connectedness risk measure for the model, although in this case we show that the model possesses only three of the four desirable properties. The data section provides a discussion of the data, spanning two decades (from 1995 to 2015), to which we apply our four models. The empirical section describes applications of our four models and demonstrates the general consistency of their results. We close with a concluding discussion and extensions.

A GENERAL FRAMEWORK FOR SYSTEMIC RISK

Dependence

For our general framework, the systemic risk, $\mathcal{S}$, for a system of n FIs depends on the following three sets of variables:

1. $\boldsymbol{\lambda}$, an n-vector whose components, λ_i, represent the annual probability that the ith FI will default.
2. $\mathbf{a}$, an n-vector whose components, a_i, represent the market value of assets in the ith FI.
3. $\boldsymbol{\Sigma}$, an $n \times n$ matrix whose components, Σ_{ij}, represent the financial connection from the ith FI to the jth FI. Depending on the model for these connections, $\boldsymbol{\Sigma}$ may or may not be symmetric.

In other words, our systemic risk measures take the following functional form

$$\mathcal{S} = f(\boldsymbol{\lambda}, \mathbf{a}, \boldsymbol{\Sigma}) \tag{7}$$

where a specific systemic risk model corresponds to a specific function f and specific definition for the connection matrix $\boldsymbol{\Sigma}$.

Our approach complements the ideas laid out by De Nicolo, Favara, and Ratnovski (2012), who offered a class of externalities that lead to systemic risk. First, externalities from strategic complementarities are captured through asset ($\mathbf{a}$) correlations in our model.

Second, externalities related to fire sales are embedded in the default probabilities ($\boldsymbol{\lambda}$). Third, externalities from interconnectedness are captured through network structures ($\boldsymbol{\Sigma}$) in the model. These features connect the financial sector to systemic risk and the macroeconomy.

THE INSTITUTION RISK MEASURE, CONNECTEDNESS, AND THE CONNECTEDNESS RISK MEASURE

It is important that the impact of each institution on the overall systemic risk, $\mathcal{S}$, can be measured. For example, consider the case in which $\mathcal{S}$ is homogeneous in its default risks, $\boldsymbol{\lambda}$, which means, for any scalar $\alpha > 0$,

$$\alpha f(\boldsymbol{\lambda}, \mathbf{a}, \Sigma) = f(\alpha\boldsymbol{\lambda}, \mathbf{a}, \Sigma) \tag{8}$$

In this case one way to measure the impact of each institution on $\mathcal{S}$ is to decompose $\mathcal{S}$ into the sum of n components by differentiating Equation 8 with respect to α, yielding the result of Euler's theorem

$$\mathcal{S} = \frac{\partial \mathcal{S}}{\partial \boldsymbol{\lambda}} \boldsymbol{\lambda} = \sum_{i=1}^{n} \frac{\partial \mathcal{S}}{\partial \lambda_i} \lambda_i \tag{9}$$

This result clearly suggests using each component, $\frac{\partial \mathcal{S}}{\partial \lambda_i}\lambda_i$, of the sum to define the corresponding *institution risk measure* of institution i.

Systemic risk is also impacted by the *connectedness* of the institutions via pairwise links between the institutions. These links may be directed or undirected, depending on the model. One way to measure the connection from institution i to institution j is to use Σ_{ij}. In this case, if $\boldsymbol{\Sigma}$ is symmetric, it corresponds to undirected links; otherwise, there is at least one $\Sigma_{ij} \neq \Sigma_{ji}$, which corresponds to a directed link. Graphically, these links can be shown for a directed or undirected network by using a binary network adjacency matrix $\mathbf{B}$ whose components, B_{ij}, are derived from Σ_{ij} by selecting a threshold value K and then defining $B_{ij} = 1$ if $\Sigma_{ij} > K$ and $i \neq j$; otherwise, $B_{ij} = 0$. Links are then shown in an edge graph only when $B_{ij} = 1$, noting that the threshold value K can be altered as desired.

The strength of the connections described in the last paragraph do not necessarily correspond to measurements of the risk that the connection from institution i to institution j poses to the overall systemic risk. In the cases in which it does, we can refer to the strength of the connection as the *connectedness risk measure* from institution i to institution j. Connectedness risk measures are important to regulators who wish to determine which relationships between institutions are of primary concern to the overall health of the system.

FOUR FINANCIAL PROPERTIES

Ideally, from a practical viewpoint, the definition of $\boldsymbol{\Sigma}$ and the definition of the function f that defines systemic risk, $\mathcal{S}$, conforms to the following four financial properties:

- **Property 1: All other things being equal, $\mathbb{S}$ should be minimized by dividing risk equally among the n FIs and maximized by putting all the risk into one institution.** That is, the more the risk is spread out, the lower $\mathcal{S}$ should be. The definition of risk will depend on the model. This is a standard property emanating from diversification but is also applicable in the case of contagion. If all risk is concentrated in one entity, then contagion is instantaneous; therefore, if risk is spread out, a useful property is that the systemic score should be correspondingly lower.
- **Property 2: $\mathcal{S}$ should increase as the FIs become more entwined.** That is, if any of the off-diagonal elements of $\boldsymbol{\Sigma}$ increase, then $\mathcal{S}$ should increase. The more connected the institutions are, the greater the likelihood of contagion and systemic risk.
- **Property 3: If all the assets, a_i, are multiplied by a common factor, $\alpha > 0$, they should have no effect on $\mathcal{S}$.** If a country's FIs' assets all grow or all shrink in the same way, it should not affect the systemic risk of the country's financial system. That is, we want $f(\boldsymbol{\lambda}, \alpha\mathbf{a}, \boldsymbol{\Sigma}) = f(\boldsymbol{\lambda}, \mathbf{a}, \boldsymbol{\Sigma})$. This property is useful because it enables comparison of systemic risk scores across countries, and even for the same country, across time.
- **Property 4: Substanceless partitioning of a bank into two banks has no effect on $\mathcal{S}$.** If institution i's assets are artificially divided into two institutions of size γa_i and $(1-\gamma)a_i$ for some $\gamma \in [0, 1]$, where both of these new institutions are completely connected to each other and both have the same connections with the other banks that the original institution did, then this division is without substantive meaning, so it should not affect the value of $\mathcal{S}$. Splitting a large bank into two fully connected components with the same

connections as before should not change $\mathcal{S}$ because such a split is mere window dressing. To bring down the value of $\mathcal{S}$ by breaking up a bank, the metric states that it is important to either disconnect the two components or reduce the connectivity for each one. In fact, the metric $\mathcal{S}$ enables a regulator to assess the effect of different kinds of bank splits on reducing systemic risk.

SYSTEMIC RISK NETWORK MODELS THAT ARE HOMOGENOUS IN DEFAULT RISKS

We first examine three models that are homogenous in default risks, each using different empirical approaches and notions of risk. All three of these models satisfy all four of the financial properties listed earlier. The proof that they are satisfied is contained in the Appendix.

Models C, D, and G

We define $\mathbf{\Sigma} = \mathbf{M}$, an $n \times n$ matrix where $M_{ij} \in [0, 1]$ for all i and j and $M_{ii} = 1$ for all i. We consider three examples of $\mathbf{M}$ matrices with this property:

1. Model **C**, a correlation-based model. In this case, $M_{ij} = \frac{1}{2}(\rho_{ij} + 1)$, where ρ_{ij} is the correlation between the daily asset returns of institutions i and j. Here, **M** defines an undirected network for connectedness.
2. Model **D**, a conditional default model. In this case, M_{ij} is the annual conditional probability that institution j defaults if institution i fails. In this case, **M** defines a directed network. We note that even though the model is composed of default probabilities, we are using the Merton model only to define connectedness over the long term and thereafter assume this is independent of day-to-day changes in default risk.
3. Model **G**, a Granger causality model. This model is based on the methodology in Billio et al. (2012). For each pair of FIs (i, j), a pair of lagged value regressions of daily asset returns, r, is run to determine whether i Granger causes j and whether j Granger causes i.

$$r_i(t) = \delta_1 + \delta_2 \cdot r_i(t-1) + \delta_3 \cdot r_j(t-1) + \epsilon_i$$

$$r_j(t) = \delta_4 + \delta_5 \cdot r_j(t-1) + \delta_6 \cdot r_i(t-1) + \epsilon_j$$

The connectedness matrix is defined as follows: $M_{ij} = 1 - p(\delta_6)$ and $M_{ji} = 1 - p(\delta_3)$, where $p(x)$ is the p-value for the hypothesis that the coefficient $x = \delta_6$ or δ_3 is equal to zero in the regressions. When $i = j$, we set $M_{ii} = 1$. In this case, **M** defines a directed network.

Next, define **c** to be the n-vector whose components, c_i, represent institution i's credit risk. Specifically, we define

$$\mathbf{c} = \mathbf{a} \circ \lambda$$

where $\circ$ represents the Hadamard (or Schur) product, meaning that we have element-wise multiplication: $c_i = a_i\lambda_i$.[4]

With these definitions of **M** and **c**, we can define the *systemic risk*, $\mathcal{S}$, by

$$\mathcal{S} = \frac{\sqrt{\mathbf{c}^T\mathbf{Mc}}}{\mathbf{1}^T\mathbf{a}} \tag{10}$$

where **1** is an n-vector of ones, and the superscript T denotes the transpose of the vector. Note that the numerator is the weighted norm of the vector **c**, and the denominator $\mathbf{1}^T\mathbf{a} = \sum_{i=1}^{n} a_i$ represents the total assets in the n FIs. Also note that **M** is unitless in models **C**, **D**, and **G**; therefore, because of the presence of assets, both the numerator and denominator in Equation 10 have monetary units that cancel each other, so $\mathcal{S}$ is a unitless measure of systemic risk.

The Institution Risk Measure and Connectedness

Our model is homogeneous in $\boldsymbol{\lambda}$, so, from Equation 9, we have that

$$\mathcal{S} = \frac{\partial\mathcal{S}}{\partial\lambda}\,\lambda = \sum_{i=1}^{n}\frac{\partial\mathcal{S}}{\partial\lambda_i}\lambda_i$$

where, from differentiating our system risk definition in Equation 10, we obtain the n-dimensional vector

$$\frac{\partial\mathcal{S}}{\partial\lambda} = \frac{1}{2}\frac{\mathbf{a}\circ[(\mathbf{M}+\mathbf{M}^T)\mathbf{c}]}{\mathbf{1}^T\mathbf{a}\sqrt{\mathbf{c}^T\mathbf{Mc}}} \tag{11}$$

[4] We note that this definition of credit risk is qualitatively similar in nature to replacing **a** with the quantity of debt. That is, FIs tend to uniformly maximize along the imposed capital adequacy ratio, which results in the low cross-sectional variation in leverage across the institutions in question. Exhibit 1 presents various examples of the range in leverage across institutions at different points in time.

This decomposition of $\mathcal{S}$ gives the risk measure of each institution. The off-diagonal elements of $\mathbf{M}$ give the connectedness, although this notion of connectedness is not a connectedness risk measure.

A SYSTEMIC RISK NETWORK MODEL THAT IS NOT HOMOGENOUS IN DEFAULT RISKS

The network model in this section corresponds to a different financial view of constituting risk. As explained in the Appendix, this section's model satisfies our first three financial properties, but not the fourth.

Model R (Internal Risk Plus External Risks Model)

For this model we define $\mathbf{\Sigma} = \mathbf{M}$, where M_{ij} is the annual probability that FIs *i* and *j* both default. Next, we consider the following view of defining the risk to the system from institution *i*: Institution *i* has internal risk, which measures the chance that it will collapse and via the impact of that collapse, hurts the system directly; and it has external risk, the chance that its collapse will cause other FIs to collapse, hurting the system further. The internal risk for FI *i* is defined simply as the credit risk, $c_i = \lambda_i a_i$, that we had previously. Note that we can also write this as $c_i = M_{ii} a_i$ because, by definition, $M_{ii} = \lambda_i$. The external risk from FI *i* to FI *j* is defined as the probability that FI *i* will default multiplied by the probability that FI *j* will default given that FI *i* defaults multiplied by the assets in FI *j*. Because this is equal to the probability that both FI *i* and FI *j* default multiplied by the assets in FI *j*, we can write this as $M_{ij}a_j$.

We thus can define ρ_i, which is the internal risk from FI *i* plus the sum of the external risks from FI *i* to each of the other FIs, by

$$\rho_i = \sum_{j=1}^{n} M_{ij} a_j \tag{12}$$

Defining $\boldsymbol{\rho}$ to be the *n*-vector with components ρ_i, we can define the systemic risk to be

$$\mathcal{S} = \frac{\sqrt{\boldsymbol{\rho}^T \boldsymbol{\rho}}}{\mathbf{1}^T \mathbf{a}} \tag{13}$$

Note again that $\mathcal{S}$ is unitless, as was the case in the previous section when we defined $\mathcal{S}$ in Equation 10 for models **C**, **D**, and **G**.

The Institution Risk Measure and the Connectedness Risk Measure

These measures are straightforward. Institution *i*'s risk measure in this case is the value of ρ_i defined earlier. Note here that $\Sigma_{i=1}^{n} \rho_i \neq \mathcal{S}$, unlike the case in which $\mathcal{S}$ is homogeneous in $\boldsymbol{\lambda}$, for which this equality holds because of Equation 9. This model, unlike the three models from the previous section, has a connectedness risk measure from bank *i* to bank *j*, which is the external risk, $M_{ij}a_j$.

DATA SOURCES AND DESCRIPTION OF VARIABLES

All four models are easy to implement using publicly available data. We describe our data sources and present key summary statistics. The data used are extensive and publicly available. Hence, the approach is amenable to many data science methods applied to big data.

Sources

Our sample period spans January 1992 to December 2015 and consists of publicly traded FIs under major Standard Industrial Classification (SIC) groups 60 (depository institutions), 61 (nondepository credit institutions), and 62 (security and commodity brokers, dealers, exchanges, and services).[5] We obtain daily stock returns, stock prices, and shares outstanding for each of these firms, as well as the daily market returns, from the Center for Research in Securities Prices. We obtain applicable Treasury rates (i.e., the constant-maturity rates) on a monthly basis from the Federal Reserve Bank reports, and we obtain quarterly balance-sheet and income-statement data from Compustat. Our final sample consists of a panel dataset of 2,066,868 firm-days for 1,171 distinct FIs, from which we select the 20 largest institutions by total assets at various points across time. Working with more institutions does not pose computational difficulty; we choose only 20 institutions

[5] For a detailed breakdown of the SIC division structure, see https://www.osha.gov/pls/imis/sic_manual.html.

for clarity. The top 20 institutions consistently represent over 70% of the total worth of the assets in the 1,171 FIs.

Key Definitions and Data-Generating Computations

We solve for the *i*th FI's market value of assets, $a_i(t)$, and the annualized volatility of asset returns, $v_i(t)$ on day t, based on the Merton (1974) model for calculating equity value and equity return volatility. Recall Equations 3 and 6. Given market capitalization, $E_i(t)$; annualized equity return volatility,[6] $\sigma_i(t)$; total face value of debt, $D_i(t)$; and the annualized risk-free rate of return,[7] $r_f(t)$, we can use a simultaneous nonlinear equation root finder to simultaneously solve Equations 3 and 6 and determine the values of $a_i(t)$ and $v_i(t)$ for any i and t.[8]

Once we have our panel of daily asset values, $a_i(t)$, and volatilities, $v_i(t)$, we can calculate the daily asset returns, $r_i(t)$. The daily asset returns allow us to run the Granger regressions that determine M_{ij} in model **G** and to determine ρ_{ij}, the correlation of the daily asset returns of institutions i and j, which defines M_{ij} in model **C**. Furthermore, the daily asset returns allow us to compute asset betas, $\beta_i(t)$, which we do on a daily, rolling basis, based on a three-year (i.e., 750-day) lookback period for $r_i(t)$. Using this information, we can then calculate expected asset returns, $\mu_i(t)$, using the capital asset pricing model as follows

$$\mu_i(t) = \beta_i(t) \cdot (\mu_{MKT}(t) - r_f(t)) + r_f(t) \tag{14}$$

where $\mu_{MKT}(t)$ represents the annualized expected return on the market portfolio on day t. For the illustrative purposes of this article, we simply set $\mu_{MKT}(t)$ equal to a constant value of 10%.

The expected asset returns are used to determine $\lambda_i(t)$, the annualized PD, which is the probability that the market value of the FI's assets, a_i, governed by the geometric Brownian motion in Equation 1, will become smaller than the FI's current debt, D_i, in a year. That is

$$\lambda_i(t) = \Phi\left(-\hat{d}_{2,i}\right) \tag{15}$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function,

$$\hat{d}_{2,i} = \frac{\ln\left(\frac{a_i(t)}{D_i(t)}\right) + \left(\mu_i(t) - \left(\frac{(v_i(t))^2}{2}\right)\right)T}{v_i(t)\sqrt{T}}, \tag{16}$$

and $T = 1$ year. Note that $\hat{d}_{2,i}$ has the same definition as $d_{2,i}$ in Equation 5, but with $r_f(t)$ in that equation replaced by $\mu_i(t)$. That is, $\hat{d}_{2,i}$ corresponds to $d_{2,i}$ in the physical, instead of the risk-neutral, measure.

To determine the joint probability that both FIs i and j will default, which is the M_{ij} for model **R**, we have that

$$M_{ij} = \Phi_2(-\hat{d}_{2,i}, -\hat{d}_{2,j}, \rho_{ij})$$

where $T = 1$ year and $\Phi_2(\cdot, \cdot, \cdot)$ is the bivariate cumulative standard normal distribution function defined by

$$\Phi_2(z_1, z_2, \rho) = \int_{-\infty}^{z_1}\int_{-\infty}^{z_2} \frac{1}{2\pi\sqrt{\det(\mathbf{S})}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{S}^{-1}\mathbf{x}\right) dx_1 dx_2$$

where $\mathbf{x}$ is a column vector with entries x_1 and x_2, and $\mathbf{S}$ is a 2 × 2 matrix with ones on the diagonal and ρ in the two off-diagonal entries.[9] Finally, to determine the conditional default probability M_{ij} for model **D**, we simply divide the M_{ij} for model **R** by λ_i.

Exhibit 1 shows the evolution of these basic summary statistics over time. We note as a reality check for our calculations that the total book value of assets tracks our calculated implied market value of assets in each exhibit. For instance, as of the end of June 1995, we see that our 20 FIs held an average of approximately \$120.1 billion in total assets, which grows considerably to \$354.3 billion by the end of June 2000 and then grows further to \$1,313 billion by the end of June 2007. However, as a result of the financial crisis of 2008, this average is only moderately greater, at \$1,546 billion, by the end of June 2015. The average leverage stays approximately constant at 0.9407, 0.9475, and 0.9521 in June of 1995, 2000, and 2007, respectively. Some deleveraging to an average ratio of 0.9265 happens by the end of June 2015. The dominance of the 20 largest FIs over the field of all FIs fluctuates over the years, from

[6] We calculate equity-return volatility based on a 130-day (i.e., six-month) lookback period, which we then multiply by $\sqrt{252}$.

[7] We use the three-month constant maturity T-bill rate.

[8] We use the multiroot function for finding roots, which is included in R's rootSolve package.

[9] We use the pmvnorm function, which is included in R's mvtnorm package, to calculate $\Phi_2(\cdot, \cdot, \cdot)$.

Exhibit 1
20 Largest FIs at Various Times

	Mean	P25	P50	P75
Panel A: June 30, 1995				
Book Value of Assets	120,061	66,302	107,548	125,522
Leverage	0.9407	0.9290	0.9390	0.9529
Market Capitalization, *E*	7,622	2,718	6,814	9,548
Equity Volatility, σ	0.2370	0.1941	0.2230	0.2711
Implied Market Value of Assets, *a*	114,788	65,456	108,927	124,465
Implied Volatility of Assets, *v*	0.0171	0.0101	0.0175	0.0224
% of all FIs' Total Assets:	77.34%			
Panel B: June 30, 2000				
Book Value of Assets	354,319	235,274	276,039	417,851
Leverage	0.9475	0.9342	0.9518	0.9655
Market Capitalization, *E*	40,353	11,354	26,272	58,439
Equity Volatility, σ	0.4485	0.3522	0.4662	0.4999
Implied Market Value of Assets, *a*	357,125	243,392	265,194	418,995
Implied Volatility of Assets, *v*	0.0531	0.0159	0.0394	0.0922
% of all FIs' Total Assets:	73.83%			
Panel C: June 29, 2007				
Book Value of Assets	1,313,221	796,235	1,191,233	1,541,156
Leverage	0.9521	0.9443	0.9558	0.9673
Market Capitalization, *E*	61,228	3,472	39,833	88,437
Equity Volatility, σ	0.1956	0.1646	0.1999	0.2240
Implied Market Value of Assets, *a*	1,254,163	777,503	1,132,813	1,468,445
Implied Volatility of Assets, *v*	0.0092	0.0006	0.0073	0.0181
% of all FIs' Total Assets:	77.51%			
Panel D: June 30, 2015				
Book Value of Assets	1,546,362	1,068,044	1,525,649	1,883,981
Leverage	0.9265	0.9063	0.9393	0.9432
Market Capitalization, *E*	70,998	1,823	52,161	88,712
Equity Volatility, σ	0.2287	0.1969	0.2156	0.2583
Implied Market Value of Assets, *a*	1,500,492	1,093,359	1,433,546	1,826,187
Implied Volatility of Assets, *v*	0.0096	0.0004	0.0104	0.0191
% of all FIs' Total Assets:	76.83%			

Note: All dollar amounts are all in millions.

77.34% of all FIs' total assets in June 1995, to 73.83% in June 2000, and then to 77.51% in June 2007. Interestingly, even with global concern over FIs deemed too big to fail during the financial crisis of 2008, this number only dips slightly to 76.83% by June 2015.

Summary Statistics

We present basic summary statistics for the 20 largest FIs at various points in time. These summary statistics, given in Exhibit 1, consist of

1. *Book value of assets*, the total book value of each of the 20 FI's assets (in millions of dollars).
2. *Leverage*, the total face value of debt scaled by the total book value of the assets.
3. *Market capitalization*, *E*, the total market value of equity (in millions), calculated as the price per share times the number of shares outstanding.
4. *Equity volatility*, σ, the equity-return volatility based on a 130-day (i.e., six-month) lookback period.

5. *Implied market value of assets*, *a*, the implied market value of assets (in millions) based on the Black–Scholes formula for options valuation.
6. *Implied volatility of assets*, ν, the implied assets' return volatility based on the Black–Scholes formula for options valuation.
7. The total book value of the assets held by the 20 largest FIs as a percentage of the total book value of the assets held by all FIs.

EMPIRICAL ILLUSTRATIONS

We test our network risk framework on the financial data mined in the previous section. Recall that we have four models for systemic risk (models **C**, **D**, **G**, and **R**) within our overall framework. We compare these models in this section.

We determine systemic risk under each of our four models every six months (at the end of June and December) between 1995 and 2015. At each of these six-month intervals, we extract and analyze data for the top 20 FIs by total book value of assets, which, as we have noted, consistently accounts for approximately 75% of the aggregate assets of the more than 1,000 FIs we had available. For each of the four models, we plot the value of systemic risk over time, with each time series normalized to be in the range [0, 1], in Exhibit 2. First, this plot confirms that systemic risk spiked in the financial crisis of 2008. We also see smaller conflagrations of systemic risk in 2000 and 2011. Second, we see that all the models generate time series that track each other closely, with pairwise correlations ranging from 90%–97% (with a mean of 95%). Therefore, even though the four models are derived in uniquely different ways, time variation in the systemic risk score in these models is very much the same, implying that our systemic risk framework is robust to model choice.

It is also useful to look at the institution risk measure to see which FIs contributed the most to systemic risk. This is shown in Exhibit 3 using model **G** in 2007 and 2014. We can see that in 2007 mortgage-related FIs such as RBS Holdings (discontinued ticker ABNYY), Banco Santander (SAN), Federal Home Loan Mortgage Corp (FMCC), Fannie Mae (FNMA), Mitsubishi Trust (MTU), and Lehman Brothers (LEHMQ) were the top systemically risky firms. In 2014, the top systemic risk contributors were Mizuho Financial Group (ticker MFG), Lloyds Banking Group (LYG), Royal Bank of Scotland (RBS), Mitsubishi Trust (MTU), Sumitomo Mitsui Financial Group (SMFG), and Barclays (BCS). From both plots, we see that risk contributions are concentrated in a few banks. Furthermore, mortgage-related firms were more systemically risky in 2007, whereas in 2014, the traditional large banks were salient contributors of systemic risk.

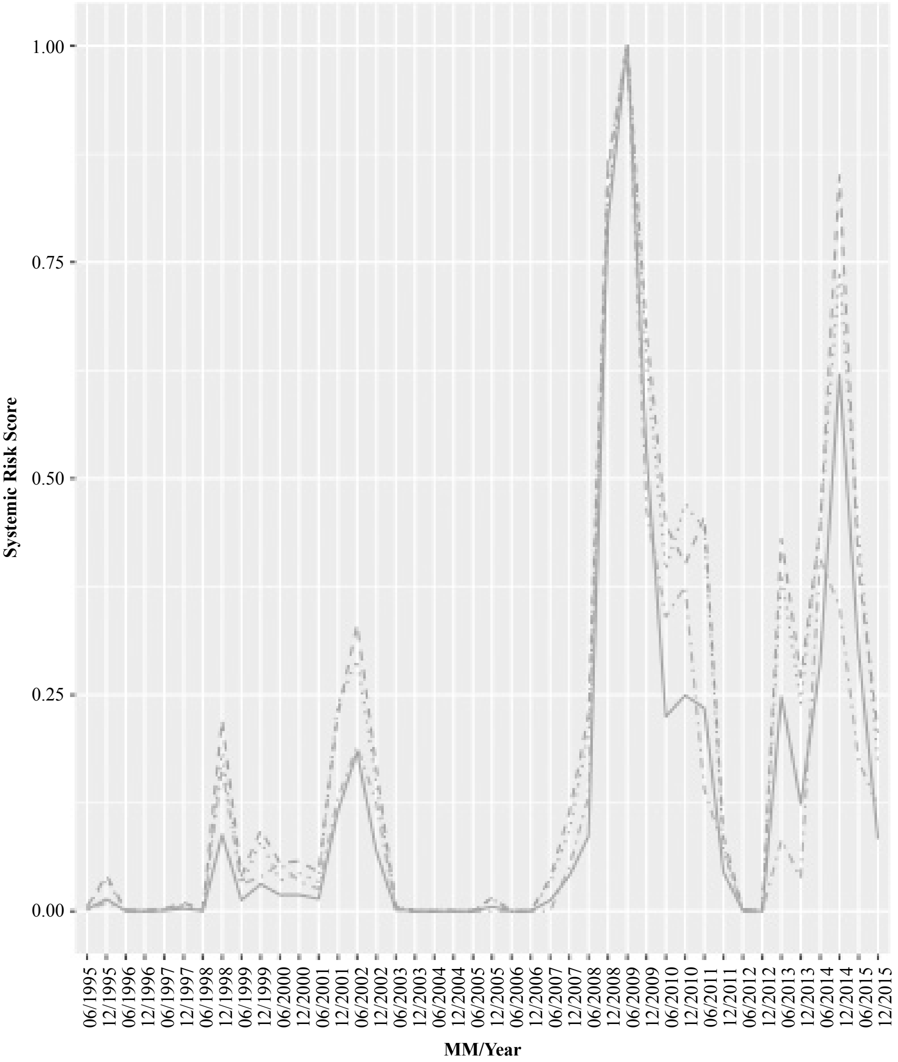
We checked that the institution risk measure rankings are similar across the four models. The top few names remain very much the same, irrespective of which model is used. In particular, the top five systemically risky FIs are the same in all four models, although not in the same order. These are Royal Bank of Scotland (RBS), Lloyds (LYG), Mizuho (MFG), Mitsubishi (MTU), and Sumitomo Mitsui (SMFG). Thus, there are two UK banks and three Japanese banks. Post-crisis measures in the United States may have reduced these banks' systemic risk levels.

Exhibit 4 extends this consistency check by displaying the union of the four models' top five risky institutions in each six-month interval. We note that in each interval there are between 5 and 13 FIs, where 5, of course, represents complete agreement among the four models and 20, of course, is the maximum possible number of FIs in the union. The average number of FIs is 6.45, showing considerable consistency among the four models in determining the top risky FIs.

We see Lehman Brothers (LEHMQ) appear consistently as a top systemically risky institution up until its demise in 2008. Around the time of the financial crisis in 2008, we also see Fannie Mae (FNMA) and the Federal Mortgage Credit Corporation (FMCC) show up as key contributors to systemic risk. Interestingly, though, these institutions were beginning to appear in the top risky list in 2003, suggesting that our methodology may have been able to provide an early warning about these mortgage-related institutions and their role in the systemic risk of the financial system.

In the latter time periods from our sample, we see Lloyds (LYG), Royal Bank of Scotland (RBS), Bank of America (BAC), and Deutsche Bank (DB) appear consistently, reflecting the fact that these institutions have been troubled in the last few years. Other large US banks that appear regularly, as is to be expected, are Citigroup (C), J.P. Morgan (JPM), and Morgan Stanley (MS). Many Japanese banks also appear, such

Exhibit 2
Systemic Risk over Time (1995–2015)

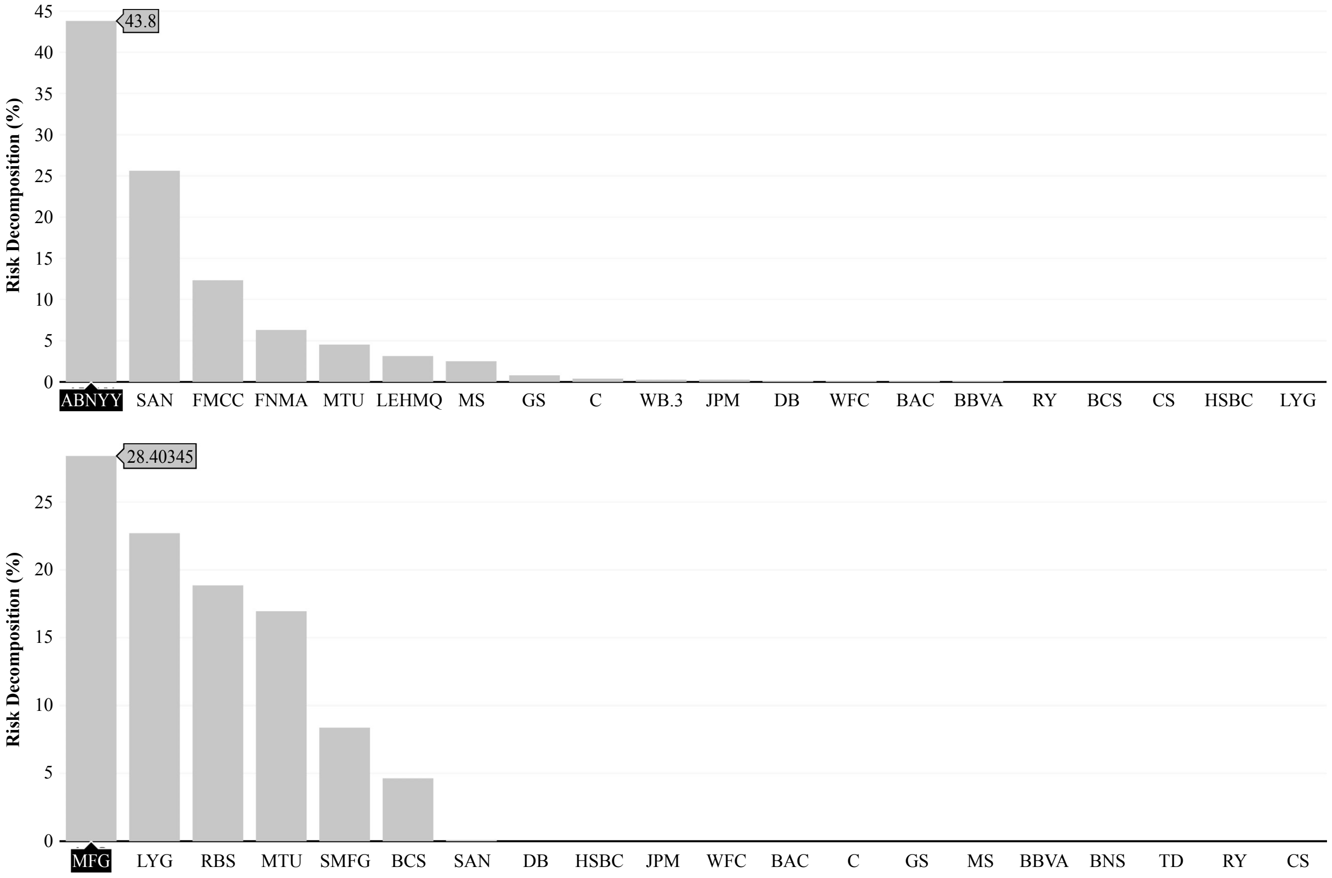


Notes: The plot shows systemic risk computed from data for the top 20 FIs (by assets). All four models, C (dashed line), D (dotted line), G (dotted dashed line), and R (solid line), are represented. The average correlation between all four models' time series is 95%.

as Mitsubishi (MTU), Mizuho (MFG), and Sumitomo Mitsui (SMFG).

We can also investigate the links between institutions that contribute the most to systemic risk in each six-month interval. Exhibit 5 illustrates this for model R. We see the same SIFIs that show up in Exhibit 4, but in this graphic, we show links (pairs of FIs) rather than individual FIs. As expected, up to the crisis we see Lehman (LEHMQ) appear on a regular basis, both as affecting other FIs and being affected by others. Santander (SAN) appears on both sides of links throughout the sample. Morgan Stanley (MS) seems to be at the receiving end of most links in which it appears. In the latter third of the sample, Mitsubishi (MTU) and Mizuho (MFG), both Japanese banks, demonstrate mutual systemic spillover risk to each other. They are also connected to another Japanese FI, Sumitomo Mitsui (SMFG). These examples illustrate that, in addition to designating individual SIFIs,

Exhibit 3
Institution Risk Measures



Notes: We display the institution risk measure using model G. This decomposes the systemic risk by institution. The upper plot is for December 2007, and the lower is for December 2014.

our model may also be used to designate systemically risky relationships.

We may wish to explore how sensitive the systemic risk measure is both, to changes in the financial strength of the FIs and to changes in the strength of the connections between the FIs. Specifically, we explore the changes in our systemic risk measures when we impose a blanket-wide increase in all the PD values (i.e., all PDs, λ_i) and when we impose a blanket-wide decrease or increase in all the pairwise correlations (i.e., all the ρ_{ij}, subject, of course, to remaining within the interval [−1,1]). In Exhibit 6, we demonstrate the effect of these changes at two snapshots in time: December 29, 2000 and December 31, 2007. We see from the exhibit that reasonable changes in either the PD values or in the correlation values affect the systemic risk score, mirroring the importance of considering the strength of both the individual FIs and the interconnections between the FIs in calculating systemic risk.

Finally, we consider the deficiency of return-based models highlighted by Löffler and Rapauch (2018). They showed that many of these popular models permitted banks to take on more risk, thereby raising overall systemic risk but at the same time reducing their own risk contribution relative to others, sometimes to the extent that their systemic risk contribution would even decline. We examine whether our model suffers from such a deficiency by increasing an FI's PD by 1% while holding all the other FIs' PD values frozen and then calculating how much the FI's institution risk

EXHIBIT 4

Top SIFIs

19950630	C	BT.2	FFB	CMB.1	GWF.	BK	MS						
19951229	BPC.2	C	CMB.1	JPM	GWF.								
19960628	C	FNMA	MS	JPM	FMCC	GWF.							
19961231	FNMA	MS	LEHMQ	C	FMCC								
19970630	GWF.	LEHMQ	MS	C	FMCC								
19971231	NW.1	LEHMQ	MS	SAN	C	BSC.1							
19980630	SAN	C	BSC.1	LEHMQ	BMO								
19981231	NW.1	LEHMQ	SAN	MS	JPM	C							
19990630	SAN	LEHMQ	MS	BSC.1	C								
19991231	NW.1	SAN	LEHMQ	BSC.1	FMCC	MS							
20000630	SAN	AFS	LEHMQ	FMCC	MS	FNMA							
20001229	SAN	LEHMQ	MS	BSC.1	JPM								
20010629	SAN	MS	LEHMQ	BSC.1	C	BAC							
20011231	ABNYY	SAN	MS	LEHMQ	AXP								
20020628	IMI.2	ABNYY	SAN	JPM	MS	C							
20021231	ABNYY	SAN	IMI.2	JPM	C								
20030630	ABNYY	JPM	MS	FNMA	LEHMQ	C	GS						
20031231	FMCC	FNMA	MS	JPM	GS	C							
20040630	DB	LEHMQ	BBVA	IMI.2	SAN	FNMA	BCS	ABNYY	MS	GS	C	JPM	
20041231	MS	NMR	BBVA	ABNYY	CS	C	FNMA	JPM	DB	LEHMQ	GS	FMCC	BCS
20050630	FNMA	FMCC	MS	BCS	ABNYY	C	JPM	LEHMQ	GS				
20051230	ABNYY	FNMA	FMCC	BCS	SAN	CS	LEHMQ	GS	C	JPM			
20060630	FNMA	LEHMQ	DB	GS	SAN	C	JPM						
20061229	LEHMQ	BSC.1	ABNYY	SAN	BBVA	MS	C	FNMA	DB	BAC	BCS		
20070629	MTU	LEHMQ	MS	GS	ABNYY	FNMA							
20071231	ABNYY	SAN	FNMA	FMCC	MTU								
20080630	MTU	FNMA	FMCC	DB	C	MS							
20081231	MTU	FMCC	FNMA	C	BAC	DB							
20090630	MFG	FNMA	MTU	C	BAC	DB							
20091231	MFG	MTU	NMR	FNMA	FMCC	C							
20100630	MTU	MFG	NMR	CS	BBVA	DB	SAN						
20101231	MFG	MTU	BBVA	SAN	CS	DB							
20110630	MFG	MTU	MS	BBVA	SAN								
20111230	BAC	C	MS	DB	BBVA	WFC							
20120629	BAC	MS	C	JPM	BBVA								
20121231	BBVA	DB	BAC	C	MS	SAN							
20130628	LYG	MFG	SMFG	RBS	MS	BBVA							
20131231	LYG	SMFG	MFG	SAN	BCS	HSBC	JPM	DB	BAC	C			
20140630	LYG	BCS	SMFG	BBVA	BAC								
20141231	RBS	LYG	MFG	MTU	SMFG								
20150630	RBS	LYG	SAN	MTU	SMFG								
20151231	RBS	SAN	BCS	DB	MS	BAC							

Notes: The graphic shows the FIs that contribute the most to systemic risk every half year in the sample across all four models. Each row displays the union of each of the four models' top five FIs that contribute the most risk. If the FIs are the same across all models, we will see exactly five FIs listed in a row; if not, then a few more will appear. One can see high agreement across models because the average number of firms in the rows is only 6.45.

EXHIBIT 5
Top Risky Links

19950630	BT.2:C	C:BT.2	FFB:C	C:FFB	C:CMB.1
19951229	BPC.2:C	C:BPC.2	BPC.2:CMB.1	BPC.2:JPM	CMB.1:BPC.2
19960628	C:FNMA	FNMA:C	C:JPM	JPM:C	FMCC:JPM
19961231	LEHMQ:MS	MS:LEHMQ	C:FNMA	FMCC:FNMA	FNMA:C
19970630	GWF.:C	GWF.:FMCC	C:FNMA	GWF.:FNMA	BK:FNMA
19971231	LEHMQ:MS	MS:LEHMQ	SAN:NW.1	NW.1:SAN	BSC.1:SAN
19980630	SAN:C	BSC.1:C	C:SAN	BSC.1:FMCC	FMCC:C
19981231	LEHMQ:NW.1	SAN:NW.1	NW.1:LEHMQ	NW.1:SAN	NW.1:MS
19990630	BSC.1:SAN	LEHMQ:SAN	LEHMQ:MS	SAN:LEHMQ	SAN:C
19991231	SAN:NW.1	NW.1:SAN	LEHMQ:NW.1	NW.1:LEHMQ	BSC.1:NW.1
20000630	AFS:SAN	SAN:FNMA	SAN:FMCC	LWHMQ:MS	FMCC:FNMA
20001229	LEHMQ:MS	LEHMQ:SAN	SAN:LEHMQ	BSC.1:SAN	SAN:JPM
20010629	SAN:MS	MS:SAN	LEHMQ:MS	LEHMQ:SAN	SAN:LEHMQ
20011231	SAN:ABNYY	ABNYY:SAN	ABNYY:MS	MS:ABNYY	SAN:MS
20020628	IMI.2:ABNYY	SAN:ABNYY	ABNYY:SAN	IMI.2:SAN	ABNYY:IMI.2
20021231	SAN:ABNYY	IMI.2:SAM	ABNYY:SAN	SAN:JPM	SAN:IMI.2
20030630	ABNYY:JPM	JPM:ABNYY	ABNYY:FNMA	FNMA:ABNYY	LEHMQ:MS
20031231	FMCC:FNMA	FNMA:FMCC	JPM:FNMA	GS:MS	FNMA:JPM
20040630	DB:ABNYY	LEHMQ:MS	FNMA:DB	DB:FNMA	MS:LEHMQ
20041231	MS:GS	MS:LEHMQ	LEHMQ:MS	GS:MS	MS:FNMA
20050630	FMCC:FNMA	FNMA:FMCC	LEHMQ:MS	MS:GS	MS:LEHMQ
20051230	FMCC:ABNYY	ABNYY:FMCC	ABNYY:BCS	ABNYY:JPM	ABNYY:CS
20060630	FNMA:DB	DB:FNMA	LEHMQ:GS	GS:LEHMQ	FNMA:C
20061229	BSC.1:LEHMQ	LEHMQ:BSC.1	LEHMQ:MS	DB:C	DB:ABNYY
20070629	LEHMQ:MTU	MS:MTU	MTU:MS	LEHMQ:GS	MTU:LEHMQ
20071231	SAN:ABNYY	ABNYY:SAN	FMCC:ABNYY	FNMA:ABNYY	FNMA:SAN
20080630	FNMA:MTU	FMCC:MTU	MTU:FNMA	MTU:FMCC	FNMA:FMCC
20081231	FMCC:MTU	FNMA:MTU	C:MTU	MTU:DB	FMCC:DB
20090630	MFG:MTU	FNMA:MTU	FNMA:MFG	MTUlMFG	C:MTU
20091231	MFG:MTU	NMR:MTU	FNMA:MTU	FMCC:MTU	MTU:MFG
20100630	MFG:MTU	MTU:MFG	NMR:MTU	NMR:MFG	BBVA:MTU
20101231	MFG:MTU	MTU:MFG	BBVA:MTU	BBVA:MFG	SAN:MTU
20110630	MFG:MTU	MTU:MFG	MS:MTU	MS:MFG	BBVA:MTU
20111230	C:BAC	BAC:C	MS:BAC	BAC:DB	MS:DB
20120629	MS:BAC	C:BAC	BAC:C	BAC:JPM	JPM:BAC
20121231	BBVA:BAC	C:DB	MS:BAC	DB:C	BAC:BBVA
20130628	LYG:MFG	SMFG:MFG	MFG:SMFG	LYG:SMFG	MFG:LYG
20131231	LYG:SMFG	SMFG:LYG	SMFG:MFG	LYG:MFG	MFG:SMFG
20140630	LYG:BCS	SMFG:BCS	LYG:SMFG	BCS:SMFG	BCS:LYG
20141231	RBS:MTU	LYG:MTU	MFG:MTU	SMFG:MTU	RBS:MFG
20150630	LYG:RBS	SAN:RBS	RBS:SAN	LYG:SAN	RSB:LYG
20151231	RBS:SAN	SAN:RBS	SAN:BCS	RBS:BCS	BCS:SAN

Notes: The graphic shows the five links with the highest connectedness risk measure in each six-month interval according to model R. The links are listed in the form i:j for a directed link from institution i to institution j.

Exhibit 6

Percentage Changes in Systemic Risk Measures

20 Largest FIs: December 29, 2000

Weighted Average Pd = 1.83%

Weighted Average Pairwise Correlation (Corr) = 0.2286

	PD + 0.1%	PD + 1%	Corr – 0.2	Corr + 0.2
Model R	+4.91%	+55.55%	–7.18%	+9.23%
Model C	+4.24%	+42.51%	–6.68%	+3.24%
Model E	+2.62%	+27.87%	–3.37%	+3.91%
Model G	+9.62%	+84.81%	–	–

20 Largest FIs: December 31, 2007

Weighted Average Pd = 3.76%

Weighted Average Pairwise Correlation (Corr) = 0.3065

	PD + 0.1%	PD + 1%	Corr – 0.2	Corr + 0.2
Model R	+3.10%	+36.64%	–4.70%	+5.01%
Model C	+2.22%	+22.29%	–5.98%	+1.77%
Model E	+1.46%	+15.24%	–1.86%	+1.74%
Model G	+8.57%	+73.18%	–	–

Notes: This exhibit demonstrates how the systemic risk score changes with changes in the PD or changes in the strength of the network structure.

measure changes compared to each of the other FIs. Exhibit 7 shows this effect for the top 20 FIs in 2007 and for the top 20 FIs in 2014. At both times, for each of the 20 FIs, we see from the exhibit that the FI's own institution risk measure increases more than that of the other FIs, because, for each row, the values on the diagonal are higher than the other values. A closer analysis of the data used to create the exhibit shows that the other FIs' institution risk measure actually decreases generally, and the highest increase in the data is only about half of the increase of the FI whose PD is increased. This indicates that our metric is not susceptible to gaming by any one bank.

CONCLUDING COMMENTS

Using data science and modeling tools from the social networks arena, we capture the systemic risk of a financial system in a Merton-on-a-network model that includes three important determining elements: (1) connectedness (via banking networks), (2) joint default risk (from an extension of the Merton 1974 model), and (3) size (i.e., the market value of a bank's assets, also implied from the Merton model). We define and analyze four important properties of our systemic risk measure and develop four different models that generally have these properties.

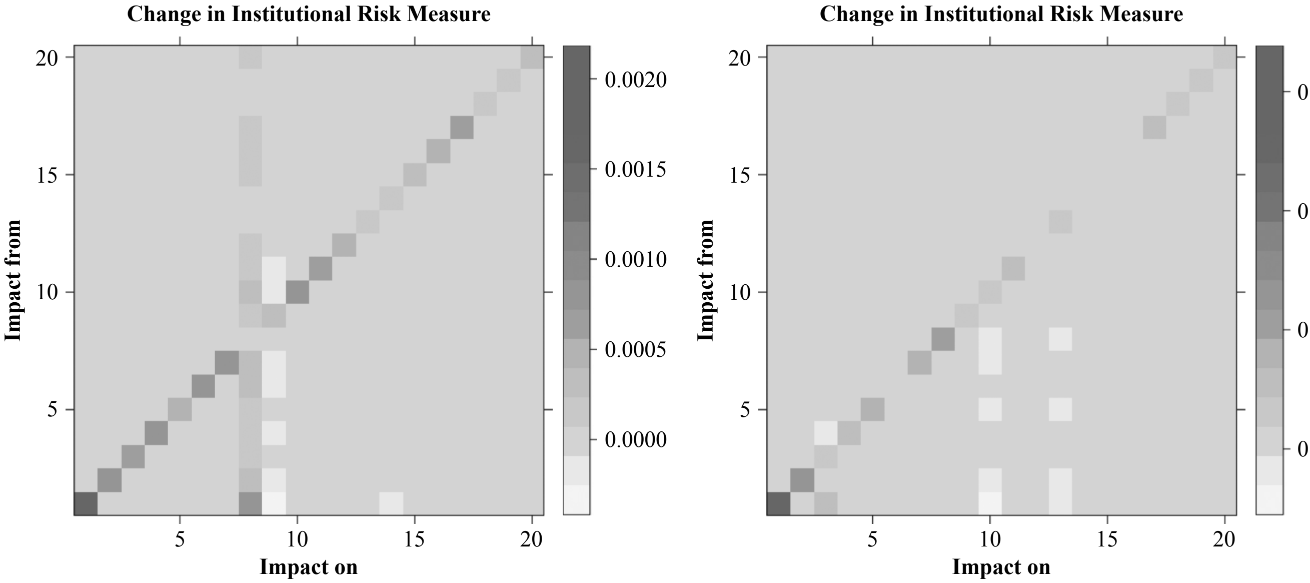
Empirical examination demonstrates that systemic risk, as well as the risk assigned to individual banks within the system, are similar across these four models, suggesting that the framework is robust to implementation design, in contrast to conflicting findings about other systemic risk measures, as shown by Benoit et al. (2013).[10] The metric also does not appear to suffer from the deficiency noted by Löffler and Rapauch (2018).

The current model supports many theoretical and empirical extensions. For example, whereas the model setting is that of the financial system, we may embed this model within a broader general equilibrium model of the entire economy, either by adding other sectors or by making the financial system variables functions of the broader macroeconomy.

[10] This article found systemic risk results to vary markedly across the four models they surveyed, namely marginal expected shortfall and SES, both from Acharya et al. (2017); the systemic risk measure from Acharya, Engle, and Richardson (2012) and Brownlees and Engle (2012); and the ΔCoVaR from Adrian and Brunnermeier (2016).

EXHIBIT 7

Spillover Risk–Change in Institutional Risk Measures



Notes: We see how much a single bank's increase in its PD affects its institution risk measure (i.e., its contribution to systemic risk) in comparison to that of the other banks. The left panel is for 2007 and the right for 2014. This experimental analysis was done for the case of model G. The largest numbers are on the diagonal, indicating that an increase to a bank's own PD increases its institution risk measure more than it increases any of the other 19 banks' institution risk measures. The diagonal values are higher than the off-diagonal values, which are mostly indistinguishable from zero. Also note that the difference in increases are more marked for 2007 before the crisis than they were for 2014.

Furthermore, we are able to extract the time series for systemic risk, which may be related to macroeconomic variables and events. Our framework supports objective real-time measurement of systemic risk, identification of SIFIs, and identification of systemically important connections between FIs so that the system may be analyzed, monitored, and controlled by regulators. The article demonstrates the efficacy of open big data in conjunction with data science techniques in risk management.

APPENDIX

PROOFS OF MODEL PROPERTIES

Financial Properties for the Homogenous Models C, D, and G

All four desired financial properties for $\mathcal{S}$ hold in models **C**, **D**, and **G**, as we next proceed to establish.

Property 1: All other things being equal, $\mathbb{S}$ is minimized by dividing the credit risk equally among the n FIs and is maximized by putting all the credit risk into one institution. To make all other things be equal, we set the total assets, $\Sigma_{i=1}^{n} a_i = \mathbf{1}^T \mathbf{a}$, constant; set the total credit risk, $\Sigma_{i=1}^{n} c_i = \mathbf{1}^T \mathbf{c}$, equal to a constant, c_{total}; and set M_{ij} equal to the same number, m, if $i \neq j$ while, of course, keeping $M_{ii} = 1$ for all i. For the singular case in which $m = 1$, all the institutions act like a single institution, and so it makes no difference to $\mathcal{S}$ how the credit risk is spread among the institutions. For the general case in which $m < 1$, from the definition of $\mathcal{S}$ in Equation 10, we see that maximizing or minimizing $\mathcal{S}$ now corresponds to maximizing or minimizing $\mathbf{c}^T \mathbf{M} \mathbf{c} = \Sigma_{i=1}^{n} c_i^2 + m\Sigma_{i=1}^{n} \Sigma_{j \neq i} c_i c_j$, subject to the restriction that $\mathbf{1}^T \mathbf{c} = \Sigma_{i=1}^{n} c_i = c_{total}$.

Because $m < 1$, it is clear that $\Sigma_{i=1}^{n} c_i^2 + m\Sigma_{i=1}^{n} \Sigma_{j \neq i} c_i c_j \leq \Sigma_{i=1}^{n} c_i^2 + \Sigma_{i=1}^{n} \Sigma_{j \neq i} c_i c_j = (\Sigma_{i=1}^{n} c_i)^2 = c_{total}^2$. However, if all the credit risk is put into one institution, we have $\Sigma_{i=1}^{n} c_i^2 + m\Sigma_{i=1}^{n} \Sigma_{j \neq i} c_i c_j = c_{total}^2$, the highest possible value, and so $\mathcal{S}$ is maximized when all the credit risk is concentrated into one FI.

On the other hand, the Lagrange multiplier method tells us that we have minimized $\Sigma_{i=1}^{n} c_i^2 + m\Sigma_{i=1}^{n} \Sigma_{j \neq i} c_i c_j$ subject to the restriction $\Sigma_{i=1}^{n} c_i = c_{total}$ when (denoting the Lagrange multiplier by λ),

$$\frac{\partial}{\partial c_k}\left(\sum_{i=1}^{n} c_i^2 + m\sum_{i=1}^{n}\sum_{j\neq i}^{n} c_i c_j\right) = \lambda \frac{\partial}{\partial c_k}\sum_{i=1}^{n} c_i \quad \text{where } k = 1, 2, \ldots\ n$$

and

$$\sum_{i=1}^{n} c_i = c_{total}$$

The first n equations give us that $c_1 = c_2 = \ldots = c_n = \frac{\lambda - 2mc_{total}}{2(1-m)}$. That is, when $\mathcal{S}$ is minimized, all c_i have the same value. The second equation then tells us that each $c_i = \frac{c_{total}}{n}$, and so we have that $\mathcal{S}$ is minimized by dividing the credit risk equally among the n institutions.

Property 2: $\mathbb{S}$ should increase as the institutions' defaults become more connected. Consider the case in which $\mathbf{a}$ and $\mathbf{c}$ are both held constant so that $\mathcal{S}$ only depends on $\mathbf{M}$, specifically through the expression

$$\mathbf{c}^T\mathbf{Mc} = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i M_{ij} c_j$$

in the numerator of our model's definition of $\mathcal{S}$. Clearly, the bigger the values of M_{ij}, the larger $\mathcal{S}$ becomes. Because M_{ii} must always equal 1, $\mathcal{S}$ is minimized when $\mathbf{M} = \mathbf{I}$, the identity matrix, and is maximized when the components of the $\mathbf{M}$ matrix are all ones. We note that when $\mathbf{M} = \mathbf{I}$ $\sqrt{\mathbf{c}^T\mathbf{Mc}} = \sqrt{\Sigma_{i=1}^{n} c_i^2} = ||c||_2$, the 2-norm of the vector $\mathbf{c}$, whereas when $\mathbf{M}$ is all ones, $\sqrt{\mathbf{c}^T\mathbf{Mc}} = \Sigma_{i=1}^{n} c_i = ||c||_1$, the 1-norm of the vector $\mathbf{c}$.

Property 3: If all the assets, a_i, are multiplied by a common factor, $\alpha > 0$, it should have no effect on $\mathbb{S}$. In our model, if we replace each a_i with αa_i, we then replace $\sqrt{\mathbf{c}^T\mathbf{Mc}}$ by $\alpha\sqrt{\mathbf{c}^T\mathbf{Mc}}$ and replace $\mathbf{1}^T\mathbf{a}$ with $\alpha\mathbf{1}^T\mathbf{a}$. Because the α then cancel in the expression for $\mathcal{S}$ from Equation 10, we have the desired property that systemic risk is unchanged.

Property 4: Substanceless partitioning of an institution into two institutions should have no effect on $\mathbb{S}$. If institution i's assets are artificially divided into two institutions of size γa_i and $(1 - \gamma)a_i$ for some $\gamma \in [0, 1]$, where both of these new institutions are completely connected to each other and both have the same connections with the other banks that the original institution did, then this division is without substantive meaning and should not affect the value of $\mathcal{S}$. Without loss of generality, we can let the index of the divided institution $i = n$, so, in our model, the new $(n + 1)$-vector $\mathbf{c}$ is

$$\mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_{n-1} \\ \gamma c_n \\ (1-\gamma)c_n \end{bmatrix}$$

and the new $(n + 1) \times (n + 1)$ matrix $\mathbf{M}$ is

$$\mathbf{M} = \begin{bmatrix} 1 & M_{12} & \cdots & M_{1(n-1)} & M_{1n} & M_{1n} \\ M_{21} & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & M_{(n-2)(n-1)} & \vdots & \vdots \\ M_{(n-1)1} & \cdots & M_{(n-1)(n-2)} & 1 & M_{(n-1)n} & M_{(n-1)n} \\ M_{n1} & \cdots & \cdots & M_{n(n-1)} & 1 & 1 \\ M_{n1} & \cdots & \cdots & M_{n(n-1)} & 1 & 1 \end{bmatrix}$$

where we note that $M_{(n+1)n} = M_{n(n+1)} = 1$ to reflect the fact that both of the new institutions are completely connected to each other. A quick computation shows that the new $\sqrt{\mathbf{c}^T\mathbf{Mc}}$ is equal to the old $\sqrt{\mathbf{c}^T\mathbf{Mc}}$, and because $a_1 + \ldots + a_n = a_1 + \ldots + a_{(n-1)} + \gamma a_n + (1 - \gamma)a_n$, we also have that the new $\mathbf{1}^T\mathbf{a}$ is equal to the old $\mathbf{1}^T\mathbf{a}$. Therefore, the value of $\mathcal{S}$ in Equation 10 is unchanged, and our model has this desired property.

Financial Properties for the Nonhomogeneous Model R

Property 1: All other things being equal, $\mathbb{S}$ is minimized by dividing the risk equally among the n FIs and is maximized by putting all the risk into one institution. Paralleling our approach in the previous section, we hold the total assets, $\Sigma_{i=1}^{n} a_i = \mathbf{1}^T\mathbf{a}$, constant and hold the total risk, $\Sigma_{i=1}^{n}\rho_i = \mathbf{1}^T\boldsymbol{\rho}$, equal to a constant. If we replace $\mathbf{c}$ and $\mathbf{M}$ in the model from the previous section for $\mathcal{S}$ given in Equation 10 with $\boldsymbol{\rho}$ and the identity matrix $\mathbf{I}$, we get our new model for $\mathcal{S}$ in Equation 13. Therefore, the proof of Property 1 from the previous section with $m = 0$ also establishes Property 1 for the model of $\mathcal{S}$ in Equation 13.

We note that if the numerator in the definition of $\mathcal{S}$ in Equation 13 were $\Sigma_{i=1}^{n}\rho_i$, the 1-norm of $\boldsymbol{\rho}$, instead of $\sqrt{\boldsymbol{\rho}^T\boldsymbol{\rho}}$, the 2-norm of $\boldsymbol{\rho}$, we would lose Property 1.

Property 2: $\mathbb{S}$ should increase as the institutions' defaults become more connected. An increasing connection means M_{ij} is increasing, which, from Equation 12, means that ρ_i increases. As any ρ_i increases, we have from Equation 13 that $\mathcal{S}$ increases, assuming, as we also did in the previous section, that $\mathbf{a}$ is held constant.

Property 3: If all the assets, a_i, are multiplied by a common factor, $\alpha > 0$, it should have no effect on $\mathcal{S}$. In our model, if we replace each a_i with αa_i, we replace $\sqrt{\boldsymbol{\rho}^T\boldsymbol{\rho}}$ by $\alpha\sqrt{\boldsymbol{\rho}^T\boldsymbol{\rho}}$, and we replace $\mathbf{1}^T\mathbf{a}$ with $\alpha\mathbf{1}^T\mathbf{a}$. Because the α then cancel in the expression for $\mathcal{S}$ given in Equation 13, we have the desired property that systemic risk is unchanged.

Property 4: Substanceless partitioning of an institution into two institutions should have no effect on $\mathbb{S}$. This property does not hold. Let's say we artificially divide institution n's assets into two institutions, call them institution n_{new} and institution $(n + 1)_{new}$, of size γa_n and $(1 - \gamma)a_n$. Because the division is artificial, $M_{n_{new}n_{new}} = M_{(n_{new}+1)n_{new}} = M_{n_{new}(n_{new}+1)} = M_{(n_{new}+1)(n_{new}+1)}$, which all equal M_{nn}, where n again represents the divided institution before it was divided, and, for any $i < n$, $M_{n_{new}i} = M_{(n_{new}+1)i} = M_{in_{new}} = M_{i(n_{new}+1)}$ equals $M_{ni} = M_{in}$.

From Equation 12, we see that the ρ_i are unchanged for $i = 1, 2, ..., n$. However, an extra $(n + 1)$th component now has been added to the vector $\boldsymbol{\rho}$, where $\rho_{n+1} = \rho_n$, which must increase the norm of $\boldsymbol{\rho}$, which must increase the systemic risk $\mathcal{S}$ in Equation 13. Therefore, artificial division of a FI increases $\mathcal{S}$ instead of having no effect on it.

ACKNOWLEDGMENTS

We are grateful for comments and input from Frank Fabozzi; Deborah Lucas; Robert Merton; Hersh Shefrin; and participants in talks at MIT's Center for Financial Policy, the Big Data in Finance Conference at University of Michigan, Box, CapitalOne, BlackRock, Berkeley's Center for Data Analytics and Risk, RFinance Conference Chicago, EDHEC Business School, McGill University, University of Delaware, AI and Machine Learning Conference (London), and the Open Data Science Conference (ODSC West 2017, San Francisco).

REFERENCES

Abbass, P., C. Brownlees, C. Hans, and N. Polich. 2016. "Credit Risk Interconnectedness: What Does the Market Really Know." *Journal of Financial Stability* 29: 1–12.

Acharya, V., R. F. Engle, and M. Richardson. 2012. "Capital Shortfall: A New Approach to Ranking and Regulating Systemic Risks." *American Economic Review* 102 (3): 59–64.

Acharya, V., L. H. Pedersen, T. Philippon, and M. Richardson. 2017. "Measuring Systemic Risk." *The Review of Financial Studies* 30 (1): 2–47.

Adrian, T., and M. Brunnermeier. 2016. "CoVaR." *American Economic Review* 106 (7): 1705–1741.

Avramidis, P., and F. Pasiouras. 2015. "Calculating Systemic Risk Capital: A Factor Model Approach." *Journal of Financial Stability* 16: 138–150.

Benoit, S., J. E. Colliard, C. Hurlin, and C. Perignon. 2013. "A Theoretical and Empirical Comparison of Systemic Risk Measures." *Working Paper #FIN-2014-1030*, HEC Paris.

Billio, M., M. Getmansky, A. Lo, and L. Pelizzon. 2012. "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors." *Journal of Financial Economics* 104 (3): 536–559.

Black, L., R. Correa, X. Huang, and H. Zhou. 2016. "The Systemic Risk of European Banks during the Financial and Sovereign Debt Crises." *Journal of Banking and Finance* 63: 107–125.

Bluhm, M., and J. P. Krahnen. 2014. "Systemic Risk in an Interconnected Banking System with Endogenous Asset Markets." *Journal of Financial Stability* 13: 75–94.

Brownlees, T. C., and R. F. Engle. 2012. "Volatility, Correlation and Tails for Systemic Risk Measurement." Working paper, New York University.

Burdick, D., S. Das, M. A. Hernandez, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, and S. Vaithyanathan. 2011. "Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study." *IEEE Data Engineering Bulletin* 34 (3): 60–67.

Das, S. 2016. "Matrix Metrics: Network-Based Systemic Risk Scoring." *The Journal of Alternative Investments* 18 (4): 33–51.

De Nicolo, G., G. Favara, and L. Ratnovski. 2012. "Externalities and Macroprudential Policy." Discussion note 12/05, International Monetary Fund.

Espinosa-Vega, M., and J. Sole. 2010. "Cross-Border Financial Surveillance: A Network Perspective." Working paper WP/10/105, International Monetary Fund.

Giglio, S., B. Kelly, and S. Pruitt. 2016. "Systemic Risk and the Macroeconomy: An Empirical Evaluation." *Journal of Financial Economics* 119 (3): 457–471.

Huang, X., H. Zhou, and H. Zhu. 2012. "Systemic Risk Contributions." *Journal of Financial Services Research* 42 (1): 55–83.

Hüser, A. C. 2015. "Too Interconnected to Fail: A Survey of the Interbank Networks Literature." *Journal of Network Theory in Finance* 1 (3): 1–50.

Kritzman, M., Y. Li, S. Page, and R. Rigobon. 2011. "Principal Components as a Measure of Systemic Risk." *The Journal of Portfolio Management* 37 (4): 112–126.

Levy-Carciente, S., D. Y. Kenett, A. Avakian, H. E. Stanley, and S. Havlin. 2015. "Dynamical Macroprudential Stress Testing Using Network Theory." *Journal of Banking and Finance* 59: 164–181.

Löffler, G., and P. Rapauch. 2018. "Pitfalls in the Use of Systemic Risk Measures." *Journal of Financial and Quantitative Analysis* 53 (1): 269–298.

Merton, R. C. 1974. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *The Journal of Finance* 29 (2): 449–470.

Merton, R. C., M. Billio, M. Getmansky, D. Gray, A. W. Lo, and L. Pelizzon. 2013. "On a New Approach for Analyzing and Managing Macrofinancial Risks." *Financial Analysts Journal* 69 (2): 22–33.

Silva, W. H. Kimura, and V. A. Sobreiro. 2017. "An Analysis of the Literature on Systemic Financial Risk: A Survey." *Journal of Financial Stability* 28: 91–114.

Upper, C. 2011. "Simulation Methods to Assess the Danger of Contagion in Interbank Markets." *Journal of Financial Stability* 7: 111–125.

Dynamic Replication and Hedging: *A Reinforcement Learning Approach*

PETTER N. KOLM AND GORDON RITTER

PETTER N. KOLM
is a clinical professor and director of the Mathematics in Finance Master's Program at NYU's Courant Institute of Mathematical Sciences in New York, NY.
petter.kolm@nyu.edu

GORDON RITTER
is adjunct professor at NYU's Courant Institute of Mathematical Sciences, NYU's Tandon School of Engineering, and the Department of Mathematics of Baruch College, and is a Professor of Practice in Rutgers University's Department of Statistics in New York, NY.
wgr2@nyu.edu

The problem of replicating and hedging an option position is fundamental in finance. Since the publication of the seminal work of Black and Scholes (1973) and Merton (1973) on option pricing and dynamic hedging (jointly referred to as BSM), a substantial number of articles have addressed the problem of optimal replication and hedging. The core idea of BSM is that in a complete and frictionless market there is a continuously rebalanced dynamic trading strategy in the stock and riskless security that perfectly replicates the option.

However, in practice, continuous trading of arbitrarily small amounts of stock is infinitely costly. Instead, the portfolio replicating the option is adjusted at discrete times to minimize trading costs. Consequently, perfect replication is impossible, and an optimal hedging strategy will depend on the desired trade-off between replication error and trading costs. In other words, the hedging strategy chosen by an agent depends on the agent's risk aversion.

Although a number of articles have considered discrete time hedging or transaction costs alone, Leland (1985) was first to address discrete hedging under transaction costs. His work was followed by others.[1] The majority of these studies treat proportionate transaction costs. More recently, several studies have considered option pricing and hedging subject to both permanent and temporary market impact in the spirit of Almgren and Chriss (1999), including Rogers and Singh (2010); Almgren and Li (2016); Bank, Soner, and Vob (2017); and Saito and Takahashi (2017).

In this article, we show how to build a system that can learn how to optimally hedge an option (or other derivative security) in a fully realistic setting. Our method applies to the real-world engineering problem faced daily by trading and risk management desks at investment banks. In such situations, continuous-time theory is only a guide. Portfolio rebalance decisions must be made in discrete time and in markets with frictions, in which liquidity is not guaranteed and the market impact of the hedge could be substantial if not managed carefully. Almgren and Chriss (1999) showed that executing a large trade in a single stock is a multiperiod planning problem that can be solved by mean–variance optimization. The option hedging problem is similar but more complex. In most cases, the hedge itself is not static but needs to be continuously readjusted. Nonetheless, both problems are related in the sense that one wishes to minimize (1) all forms of cost and (2) the deviation from the optimal hedge.

[1] See, for example, Figlewski (1989), Boyle and Vorst (1992), Henrotte (1993), Grannan and Swindle (1996), Toft (1996), Whalley and Wilmott (1997), and Martellini (2000).

This article contributes to the literature in several ways. First, our method is quite general. In particular, given any derivative security that we know how to price (even if that pricing is done by Monte Carlo), our method will quickly produce an autonomous agent who knows how to optimally trade off trading costs versus hedging variance for that security. The relative importance of cost versus variance is determined by the agent's risk-aversion parameter.

Second, our method is based on reinforcement learning (RL). Although RL is well known in its own right, to the best of our knowledge this form of machine learning technique has previously not been applied to discrete replication and hedging subject to nonlinear transaction costs. It is worthwhile to note that with the flexibility of the technique presented in this article, it is a straightforward process to extend the model with additional features and constraints such as round-lotting and position-level constraints. Although Halperin (2017) applied RL to options, the methods therein appear very specific to the BSM model, whereas our method allows the user to plug in any option pricing and simulation library and then train the system with no further modifications. Note also that Halperin (2017) did not consider transaction costs. Our article is also related to work by Buehler et al. (2018), who evaluated neural network–based hedging under convex risk measures subject to proportional transaction costs.

Third, our method is based on a continuous state space, and the training neither uses finite-state-space methods nor does it use or require a (necessarily arbitrary) selection of basis functions (as semigradient methods from Sutton and Barto (2018) would). Rather, we introduce a training method that has not been applied to derivatives hedging problems previously. Our training method relies on applying nonlinear regression techniques to the *sarsa targets* (Equation 6) derived from the Bellman equation.

Methods that require finite state spaces fail for larger problems, due to the curse of dimensionality. The state vector must contain all variables that are relevant to making a decision. For example, suppose there are k such variables, and each variable is allowed to have 10 possible values. The resulting state space has 10^k elements. Of course, this leads to insurmountable problems, such as (1) the fact that the training process can never visit most of the states; (2) there is no guarantee that the value function will be continuous, let alone smooth; (3) a vector containing all such states cannot fit in computer memory; and (4) one must estimate millions of independent parameters from relatively fewer data points. By using a continuous state space, we avoid the curse of dimensionality and are thereby able to apply our method to higher-dimensional problems.

Fourth, the method extends in a straightforward way to arbitrary portfolios of derivative securities. For example, envision a trader who has inherited a derivative security that he or she must hold to expiration because of some exogenous constraint. The trader has no directional view on the derivative or its underlier. With the method proposed in this article, the trader can essentially press a button to train an algorithm to hedge the position. The algorithm can then handle the hedging trades until expiration with no further human intervention.

REINFORCEMENT LEARNING

RL[2] has been developed largely independently from classical utility theory in finance. It provides a way to train artificial agents that learn through positive reinforcement to interact with an environment, with the goal of optimizing a reward over time. The learning agent does this through simple trial and error by receiving feedback on the amount of reward that a particular action yields. In contrast to supervised learning, an RL agent is not trained on labeled examples to optimize its actions. In addition, RL is not trying to find a hidden structure in unlabeled data and hence is different from unsupervised learning.

Mathematically speaking, RL is a way to solve multiperiod optimal control problems. The agent's policy typically consists of explicitly maximizing the action-value function for the current state. This value function is an approximation of the true value function of the multiperiod optimal control problem. *Training* refers to the process of improving on the approximation of the value functions as more training examples are made available.

Following the notation of Sutton and Barto (2018), the sequence of rewards received after time step

[2] See Sutton and Barto (2018) and Kaelbling, Littman, and Moore (1996) for an introduction to RL.

t is denoted R_{t+1}, R_{t+2}, R_{t+3}, …. The agent's goal is to maximize the expected cumulative reward, denoted by

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \qquad (1)$$

The agent then searches for policies that maximize $\mathbb{E}[G_t]$. The sum in Equation 1 can be either finite or infinite. The constant $\gamma \in [0, 1]$ is known as the discount rate. If rewards are bounded, then $\gamma < 1$ ensures convergence of the infinite sum.

A policy, denoted π, is a way of choosing an action a_t, conditional on the current state s_t. A policy is allowed to be stochastic. For example, choosing a random action is also a policy.

There are principally two kinds of value functions; at optimality, one is a maximization of the other. The action-value function expresses the value of starting in state s, taking an arbitrary action a, and then following policy π thereafter

$$q_\pi(s,a) := \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \qquad (2)$$

where $\mathbb{E}_\pi$ denotes the expectation under the assumption that policy π is followed. The state-value function is the action-value function, where the first action also comes from the policy π

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = q_\pi(s, \pi(s))$$

Action-value functions are, for most practical purposes, more useful than state-value functions because any action-value function immediately gives rise to a natural policy: If $\hat{q}$ is any action-value function, the $\hat{q}$-greedy policy is to choose the action a, in state s, that maximizes $\hat{q}(s, a)$.

Policy π is defined to be at least as good as π' if $v_\pi(s) \geq v_{\pi'}(s)$ for all states s. An optimal policy is defined to be one that is at least as good as any other policy. There need not be a unique optimal policy, but all optimal policies share the same optimal state-value function $v_*(s) = \sup_\pi v_\pi(s)$ and optimal action-value function $q_*(s, a) = \sup_\pi q_\pi(s, a)$. Note also that v_* is the supremum over a of q_*. In particular, $v_*(s)$ is the expected gain (under any optimal policy), given that one started from state s. Colloquially, one might then refer to $v_*(s)$ as the value of being in state s.

The search for an optimal policy reduces to the search for the optimal action-value function q_* because the q_*-greedy policy is optimal. The typical way of searching for q_* is to produce a sequence of iterates that approximates q_* with increasing accuracy. Methods for producing those iterates are based on the Bellman equations, which we now recall.

Let $p(s', r \mid s, a)$ denote the probability that the process transitions to state s' and the agent receives reward r, conditional on the event that the process was previously in state s, and in that state, the agent choses action a. The optimal state-value function and action-value function satisfy the Bellman equation

$$v_*(s) = \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')] \qquad (3)$$

$$q_*(s,a) = \sum_{s',r} p(s', r \mid s, a)[r + \gamma \max_{a'} q_*(s', a')] \qquad (4)$$

where the sum over s', r, denotes a sum over all states s', and all rewards r.

The intuition for Equation 3 is that the value of being in state s equals the average, over all possible next states s', of the value of being in s' plus the reward associated with making the transition $s \to s'$. The intuitive interpretation for Equation 4 is very similar; indeed $\max_{a'} q_*(s', a') = v_*(s')$, so the bracketed quantities are the same in both equations.

The state-value function $v_*(s)$ has a natural interpretation in derivatives pricing theory. Specifically, in continuous time and frictionless markets, the optimal value function of the dynamic replicating strategy is obviously equal to the no-arbitrage price of the option. This is the value function that solves the Hamilton–Jacobi–Bellman partial differential equation, as shown by Merton and Samuelson (1992). Thus, it is natural that RL, in which value functions organize the search for optimal policies, should apply to pricing and, by extension, hedging of derivatives.

TRAINING VIA SIMULATION AND BATCH LEARNING

Although the state of the art is still evolving, the vast majority of the most successful applications of RL in recent years use a simulation of the environment to generate training data (as opposed to, say, training on historical data).

In a famous example from Mnih et al. (2013, 2015), a deep RL system learned to play video games on a superhuman level. According to the authors, the network was not provided with any game-specific information or hand-designed features and was not privy to the internal state of the emulator. It simply learned from nothing but the video input, the reward and terminal signals, and the set of possible actions.

In another famous example, Silver et al. (2017) created the best Go player in the world "based solely on RL, without human data, guidance, or domain knowledge beyond game rules." The associated system, termed AlphaGo Zero "is trained solely by self-play RL, starting from random play, without any supervision or use of human data."

In these cases (and many simpler ones—see Sutton and Barto (2018) for examples), the agents are trained in a simulated environment, as opposed to being trained on historical data. This has an advantage: Millions of training examples can be generated, limited only by computer hardware capabilities. The examples in the present article follow the same pattern: The system is trained by interacting with a simulator.

We now provide more details about how the training procedure works. We start with an estimate $\hat{q}$ of the optimal action-value function. This estimate is often initialized to be the zero function and is refined as the algorithm continues.

All RL systems must balance exploration and exploitation in the training process. They must sometimes take random actions to explore new areas of state space and action space—this is exploration. However, ultimately they must use their experience to concentrate the search around strategies that are likely to be optimal and refine the estimate of the value function on those areas of state space. We follow standard practice, which is to force exploration during training by using an ϵ-greedy policy relative to $\hat{q}$

$$\pi_{\epsilon\text{-greedy}}(s) = \begin{cases} \tilde{a} & u < \epsilon \\ \operatorname{argmax}_a \hat{q}(s,a) & u \geq \epsilon \end{cases} \tag{5}$$

where ϵ is a real number between 0 and 1, u is a uniformly distributed random variable on $(0,1)$, and ã is sampled uniformly from the action space. As is standard in RL and necessary to ensure convergence, we decrease the value of ϵ as training progresses.

Let s_t be the state at the t-th step in the simulation, and let $a_t = \pi_{\epsilon\text{-greedy}}(s)$ be the associated ϵ-greedy action. Let

$$X_t := (s_t, a_t)$$

be the resulting state-action pair. The update target Y_t is defined to be any valid approximation of $q_\pi(s_t, a_t)$. In this article we use the one-step sarsa target, which approximates $q_\pi(s_t, a_t)$ as follows

$$Y_t = r_{t+1} + \gamma \hat{q}(s_{t+1}, a_{t+1}) \tag{6}$$

Intuitively, Equation 6 resembles part of the Bellman equation

$$q_*(s,a) = \sum_{s',r} p(s', r \mid s, a)[r + \gamma \max_{a'} q_*(s', a')] \tag{7}$$

Indeed, if $a_{t+1} = \operatorname{argmax}_{a'} q_*(s', a')$, then Equation 6 would be a sample of the random variable in brackets in Equation 7. Thus, Equation 6 may be viewed as an approximation of $q_\pi(s_t, a_t)$.

We shall define a batch to be a collection of pairs of the form (X_t, Y_t) where $X_t: = (s_t, a_t)$ is a state-action pair, and Y_t is the corresponding update target (Equation 6). A batch is typically obtained by running the simulator for the required number of time steps and choosing the actions via some policy π that is being evaluated.

Suppose we are going to run B different batches, indexed by $b = 1, \ldots, B$. We assume there is a nonlinear regression learner available that can learn a function of the form $Y = \hat{q}^{(b)}(X)$ using all of the samples in the batch. Suitable nonlinear regression learners are a topic of frequent study in the statistical learning literature (see Friedman, Hastie, and Tibshirani (2001) for an overview). They include random forests, Gaussian process regression, support vector regression, and artificial neural networks.

The fitted model $\hat{q}^{(b)}$ will then be used to improve the model current $\hat{q}$ by model averaging. We then generate batch $b + 1$, using the updated/improved $\hat{q}$ to calculate the Y_t, and repeat until we have B batches and $\hat{q}$ has been updated B times. Alternating between generation of batches and fitting models continues until some convergence criterion is reached. The simulations in this article used $B = 5$ batches each consisting of 750,000 (X, Y) pairs.

AUTOMATIC HEDGING IN THEORY

We define automatic hedging to be the practice of using trained RL agents to handle the hedging of certain derivative positions. The agent has a long option position that cannot be traded. The agent is only allowed to trade any other nonoption positions that would be used for replication. In a world with no trading frictions and where continuous trading is possible, there may be a dynamic replicating portfolio that hedges the option position perfectly, meaning that the overall portfolio (option minus replication) has zero variance. In our setting in this article, we will consider frictions and where only discrete trading is possible. Here the goal becomes minimization of variance and cost.

We will derive the precise form of the reward function, assuming our agent has a quadratic utility.[3] In particular, the agent's optimal portfolio is given by the solution to a mean–variance optimization problem with risk-aversion κ

$$\max\left(\mathbb{E}[w_T]-\frac{\kappa}{2}\,\mathbb{V}[w_T]\right) \tag{8}$$

where the final wealth w_T is the sum of individual wealth increments δw_t

$$w_T = w_0 + \sum_{t=1}^{T}\delta w_t$$

and so $\mathbb{E}[w_T] = w_0 + \sum_t \mathbb{E}[\delta w_t]$. The variance term involves cross-covariances of the form $cov(\delta w_t, \delta w_s)$ for $s \neq t$, but if we are willing to assume independence of wealth increments across time, that is

$$cov(\delta w_t, \delta w_s) = 0 \text{ for } s \neq t$$

then $\mathbb{V}[w_T] = \sum_t \mathbb{V}[\delta w_t]$.[4]

In complete markets, options are redundant instruments. They can be exactly replicated (with zero variance) by a continuous-time dynamic trading strategy that trades infinitely often in infinitesimal increments. In the real world, the profit and loss (P&L) variance of an option minus its offsetting replicating portfolio is not zero. In the spirit of Almgren and Chriss (2001), our hedging agent would like to solve a simplified version of Equation 8, namely

$$\min_{\text{strategies}} \sum_{t=0}^{T}\left(\mathbb{E}[-\delta w_t]+\frac{\kappa}{2}\,\mathbb{V}[\delta w_t]\right) \tag{9}$$

where the minimum is computed across all permissible trading strategies. What is different in our work as compared to that of Almgren and Chriss (2001) is that a machine will learn the optimal strategy by simulating a financial market and applying RL to the simulation results.

If the log price process is a random walk, then wealth increments can be decomposed as

$$\delta w_t = q_t - c_t$$

where q_t is random walk term, and c_t is the total trading cost paid in period t (including commissions, bid–offer spread cost, market impact cost, and other sources of slippage). In the random walk case, the expected wealth increment is therefore just −1 times the expected cost

$$\mathbb{E}[-\delta w_t] = \mathbb{E}[c_t]$$

In other words, in this case the problem (Equation 9) becomes a trade-off of cost versus variance. The agent can hedge more frequently to reduce the variance of the hedged position, but at increased trading costs.

As shown by Ritter (2017), with an appropriate choice of the reward function, the problem of maximizing $\mathbb{E}[u(w_T)]$ can be recast as a RL problem. The reward in each period corresponding to Equation 9 is approximately

$$R_t := \delta w_t - \frac{\kappa}{2}(\delta w_t)^2$$

By plugging each one-period reward into the cumulative reward (Equation 1), we obtain an approximation of the mean–variance objective. Thus, training reinforcement learners with this kind of reward function amounts to training expected-utility maximizers. In the context of option hedging, it amounts to training automatic hedgers that are prepared to optimize the trade-off of costs versus variance from being out of hedge.

[3] See Ritter (2017) for a discussion of how the mean–variance assumption fits in within a general utility framework.

[4] The independence assumption will be violated in a number of interesting examples, such as assets with long-lived transient market impact.

In the next section, we shall show that automatic hedging is indeed possible using RL training methods.

AUTOMATIC HEDGING IN PRACTICE

We look at the simplest possible example: a European call option with strike price K and expiry T on a non-dividend-paying stock. We take the strike and maturity as fixed, exogenously given constants. For simplicity, we assume the risk-free rate is zero. The agent we train will learn to hedge this specific option with this strike and maturity. It is not being trained to hedge any option with any possible strike/maturity.[5]

The agent comes into the current period with a fixed option position of L contracts. We assume for simplicity that this option position will stay the same until the option either is exercised or expires—we are training an agent to be an optimal hedger of a given contract, not an agent that can decide not to hold the contract at all.

Each period, the agent observes a new state and then can decide on an action. Available actions always include trading shares of the underlying, with bounds dictated by the economics of the problem. For example, with L contracts, each for 100 shares, one would not want to trade more than $100 \cdot L$ shares. If the option is American, then there is an additional action, which is to exercise the option and hence buy or sell shares at the strike price K.

In any successful application of RL, the state must contain all of the information that is relevant for making the optimal decision. Information that is not relevant to the task at hand, or which can be derived directly from other variables of the state, does not need to be included. For European options, the state must minimally contain the current price S_t of the underlying and the time $\tau := T - t > 0$ still remaining to expiry, as well as our current position of n shares. The state is thus naturally an element of

$$\mathcal{S} := \mathbb{R}_+^2 \times \mathbb{Z} = \{(S, \tau, n) \mid S > 0, \tau > 0, n \in \mathbb{Z}\}$$

If the option is American, then it may be optimal to exercise early just before an ex-dividend date. In this situation, the state must be augmented with one additional variable: the size of the anticipated dividend in period $t + 1$.

The state does not need to contain the option Greeks because they are (nonlinear) functions of the variables the agent has access to via the state. We expect agents, given enough simulations, to learn such nonlinear functions on their own as needed. This has the advantage of not requiring any special, model-specific calculations that may not extend beyond BSM models.

Practitioners often compute the delta of an option position, for hedging purposes, using the BSM formula:

$$\begin{aligned} \Delta &= \frac{\partial C}{\partial S} = N(d_1), \\ d_1 &= \frac{\ln \frac{S_t}{K} + \frac{\tau\sigma^2}{2}}{\sigma\sqrt{\tau}}, \\ \tau &:= T - t > 0 \end{aligned} \tag{11}$$

but with σ replaced by the implied volatility. This is referred to as *practitioner delta* by Hull and White [2017]. Note that parameters such as K and σ^2 are not provided to the agent, although they are used in constructing the simulation under which the agent is trained.

The agent will learn the properties of the stochastic world it inhabits by means of a large number of simulations of such world, as described. Nonlinear functions such as Δ as given by Equation 11, insofar as they affect the optimal strategy, will become part of the agent's learned action-value function (Equation 2).

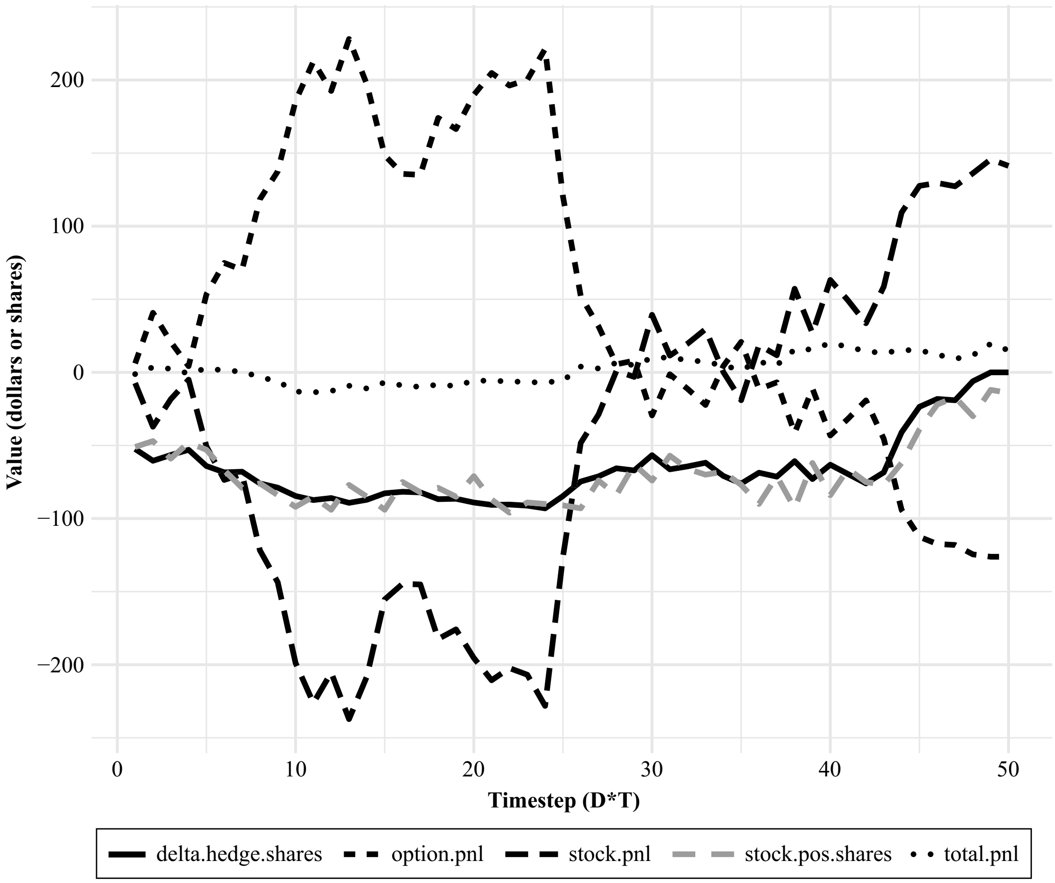
We simulate a BSM world but modified to reflect the realities of trading: discrete time and space. We consider a stock whose price process is a geometric Brownian motion (GBM) with initial price S_0 and daily lognormal volatility of σ/day. We consider an initially at-the-money European call option (struck at $K = S_0$) with T days to maturity. We discretize time with D periods per day; hence each episode has $T \cdot D$ total periods. We require trades (hence also holdings) to be integer numbers of shares. We assume that our agent's job is to hedge one contract of this option. In the following specific examples, the parameters are $\sigma = 0.01$, $S_0 = 100$, $T = 10$, and $D = 5$. In addition, we set the risk aversion $\kappa = 0.1$.

We first consider a frictionless world without trading costs and answer the question of whether it is possible for a machine to learn what we teach students

[5] However, we note that this is possible on an extended state space.

Exhibit 1

Out-of-Sample Simulation of a Trained RL Agent



Notes: We depict cumulative stock, option, and total P&L; RL agent's position in shares (stock.pos.shares); and −100·Δ (delta.hedge.shares). Observe that (1) cumulative stock and options P&L roughly cancel one another to give the (relatively low variance) total P&L, and (2) the RL agent's position tracks the delta position even though it was not provided with it.

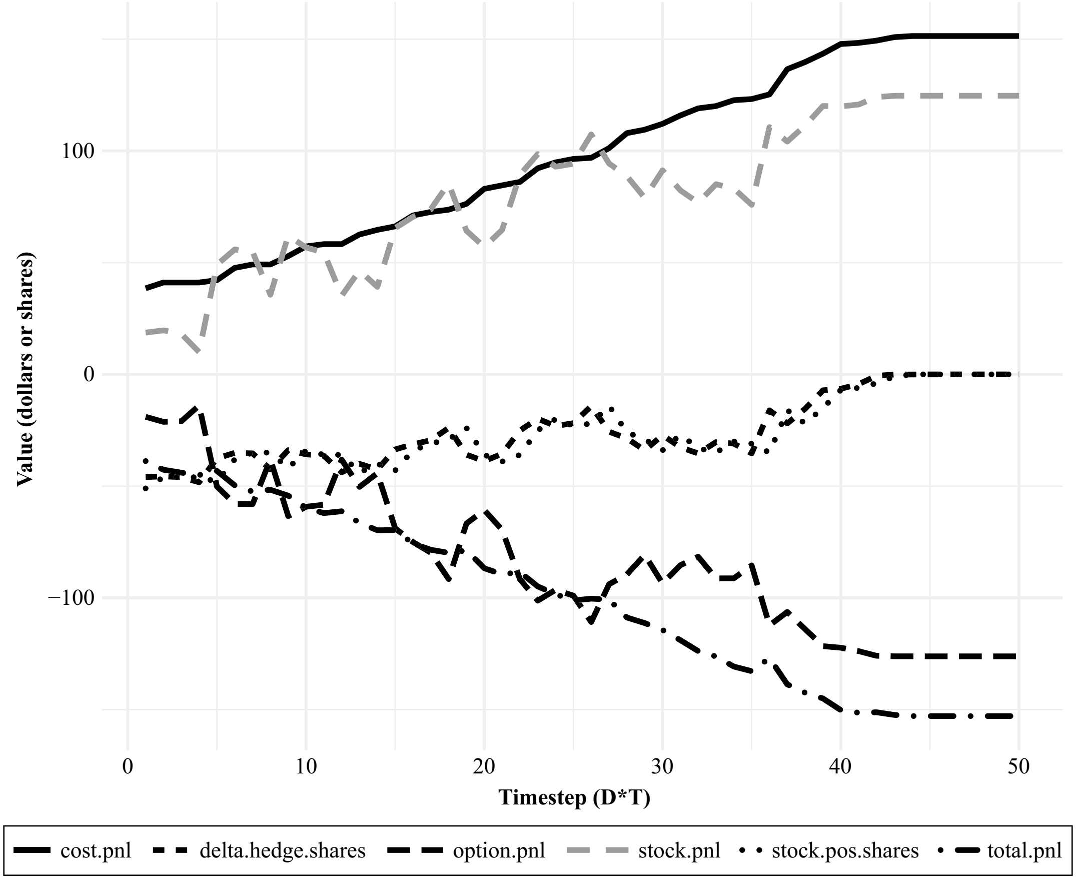
in their first semester of business school: formation of the dynamic replicating portfolio strategy. Unlike our students, the machine can only learn by observing and interacting with simulations.

The RL agent is at a disadvantage, initially. Recall that it does not know any of the following pertinent pieces of information: (1) the strike price K, (2) the fact that the stock price process is a GBM, (3) the volatility of the price process, (4) the BSM formula, (5) the payoff function $(S - K)_+$ at maturity, and (6) any of the Greeks. It must infer the relevant information from these variables, insofar as it affects the value function, by interacting with a simulated environment.[6] Each out-of-sample simulation of the GBM is different, but we show a typical example of the trained agent's performance in Exhibit 1.

[6] One could try to help the algorithm by providing the BSM delta as part of the state variable, hence allowing the reinforcement learner to use that directly, but we deliberately chose not to include any of the option Greeks as state variables. Giving the system access to the option Greeks is sure to improve its performance because the function being learned is closer to linear. We chose not to do this to make the problem as hard as possible and to see if RL is up to the challenge. However, in a real-world production scenario, we recommend making the problem as easy as possible by including certain option Greeks in the state variable, unless they are prohibitively hard to calculate.

Exhibit 2

Out-of-Sample Simulation of a Baseline RL Agent That Uses Policy Delta or π_{DH}, Defined in Equation 12



Notes: We show cumulative stock P&L and option P&L, which roughly cancel one another to give the (relatively low variance) total P&L. We show the position, in shares, of the agent (stock.pos.shares). The agent trades so that the position in the next period will be the quantity −100·Δ rounded to shares.

Because the examples of Exhibit 1 were generated in a frictionless simulation, why is the total P&L not exactly zero? The answer is discretization error. Time is discretized (to five periods per day), so continuous hedging is not possible. Moreover, the simulation requires trading an integer number of shares, which introduces further discretization error.

Any complex model should be tested against a simpler model as a baseline. To justify its additional complexity, the more complex model should be able to do something that the simpler model cannot. Along these lines, let us define a simple policy, π_{DH}, as a baseline for the more complex policy learned by RL methods.

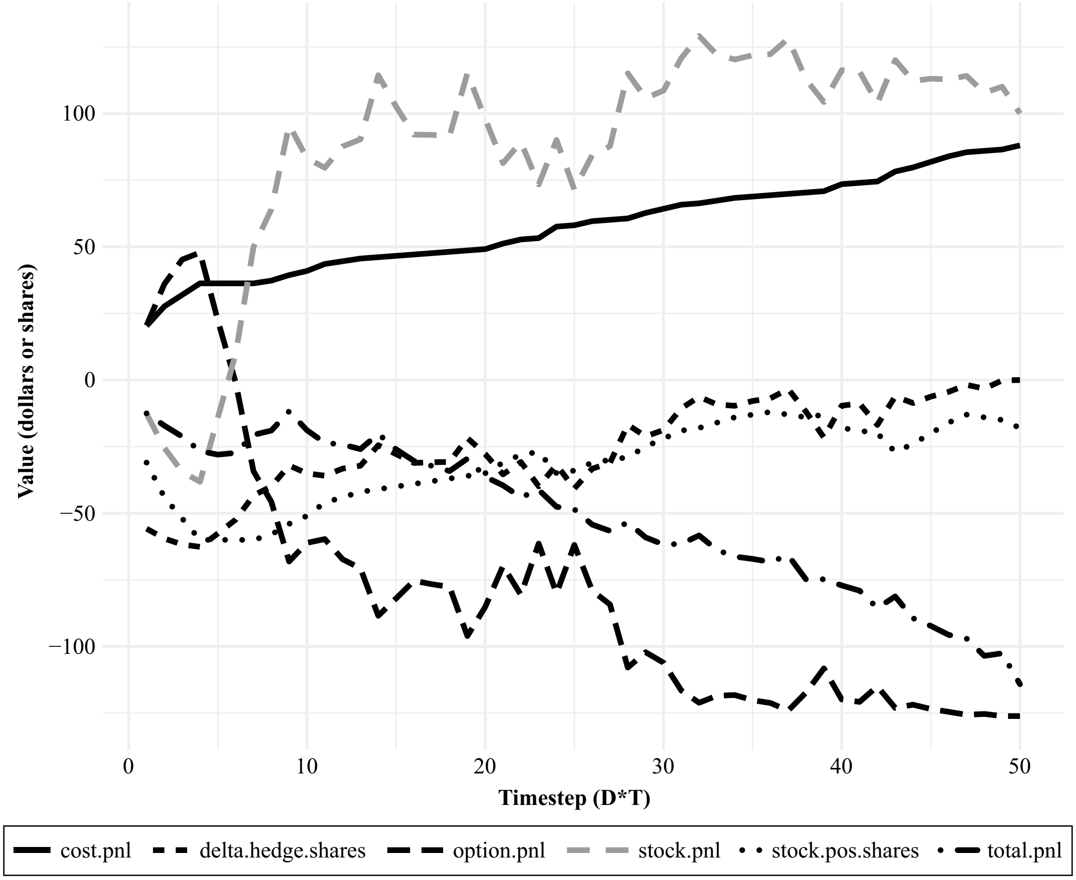
As in Equation 11, let $\Delta(p_t, \tau)$ denote the delta as computed from the price p_t at time t and the time-to-expiry $\tau = T - t$. The full state variable is then $s_t = (p_t, \tau, n_t)$, where n_t denotes the agent's current holding, in shares, at time t. Our simple baseline policy must output an action, which is just a number of shares to trade, given this state vector. Define

$$\pi_{DH}(s_t) = \pi_{DH}(p_t, \tau, n_t) := -100 \cdot \text{round}(\Delta(p_t, \tau)) - n_t \quad (12)$$

where the round function returns the closest integer to the argument.

The policy π_{DH}, without rounding, is optimal in a hypothetical trading-cost-free world, where the number of time steps goes to infinity and where one can trade fractional numbers of shares. There is, however, no reason to expect that π_{DH} would solve the utility-maximization problem (Equation 9) in a simulation with

Exhibit 3
Out-of-Sample Simulation of Our Trained RL Agent



Note: The curve representing the agent's position (stock.pos.shares) controls trading costs and is hence much smoother than the value of −100·Δ (called delta.hedge.shares), which naturally fluctuates along with the GBM process.

nontrivial trading costs or, for that matter, in the real world (where we know trading costs are nontrivial).

For a trade size of n shares we define

$$\text{cost}(n) = \text{multiplier} \times \text{TickSize} \times (|n| + 0.01n^2) \qquad (13)$$

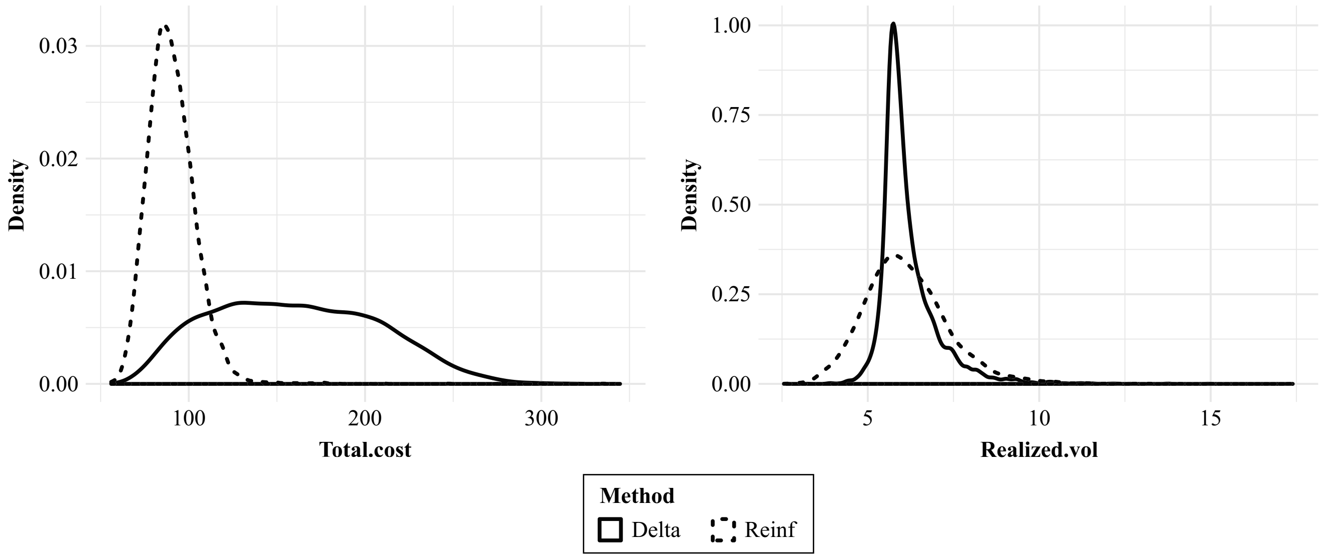
where we take TickSize = 0.1. With multiplier = 1, the term TickSize × $|n|$ represents a cost, relative to the midpoint, of crossing a bid–offer spread that is two ticks wide. The quadratic term in Equation 13 is a simplistic model for market impact. Exhibit 1 has multiplier = 0.

A key strength of the RL approach is that it does not make any assumptions about the form of the cost function (Equation 13). It will learn to optimize expected utility under whatever cost function is provided. In Exhibit 1, we had taken multiplier = 0 in the function cost(n), representing no frictions. We now take multiplier = 5, representing a high level of friction. Our intuition is that in high-trading-cost environments (which would always be the case if the position being hedged were large relative to the typical volume in the market), the simple policy π_{DH} trades too much. One could perhaps save a great deal of cost in exchange for a slight increase in variance.

Given the mean–variance utility function in Equation 9, we expect RL to learn the trade-off between variance and cost. In other words, we expect it to realize lower cost than π_{DH}, possibly coming at the expense of higher variance, when averaged across a sufficiently large number of out-of-sample simulations (i.e., simulations that were not used during the training phase in any way).

Exhibit 4

Kernel Density Estimates for Total Cost (left panel) and Volatility of Total P&L (right panel) from N = 10,000 Out-of-Sample Simulations



Notes: Policy delta is π_{DH}*, while policy reinf is the greedy policy of an action-value function trained by RL. The reinf policy achieves much lower cost (t-statistic = −143.22) with no significant difference in volatility of total P&L.*

We trained the agent using five batches with 15,000 episodes per batch, each episode having $D \cdot T = 50$ time steps, as before. This means that each call to the nonlinear regression learner involves 750,000 (X_t, Y_t) pairs. The training procedure took one hour on a single CPU. After training, we ran $N = 10{,}000$ out-of-sample simulations. Using the out-of-sample simulations, we ran a horse race between the baseline agent that uses just delta-hedging and ignores cost and the RL-trained agent that trades cost for realized volatility.

Exhibit 2 shows one representative out-of-sample path of the baseline agent. We see that the baseline agent is overtrading and paying too much cost. Exhibit 3 shows the RL agent on the same path. We see that, while maintaining a hedge, the agent is trading in a cost-conscious way. The curves in Exhibit 2, representing the agent's position (stock.pos.shares), are much smoother than the value of $-100{\cdot}\Delta$ (called delta.hedge.shares in Exhibit 2), which naturally fluctuates along with the GBM process.

Exhibit 3 consists of only one representative run from an out-of-sample set of $N = 10{,}000$ paths. To summarize the results from all runs, we computed the total cost and standard deviation of total P&L of each path. Exhibit 4 shows kernel density estimates (basically, smoothed histograms) of total costs and volatility of total P&L of all paths. In each case, we performed a Welch two-sample *t*-test to determine whether the difference in means was significant. The difference in average cost is highly statistically significant, with a *t*-statistic of −143.22. The difference in vols, on the other hand, was not statistically significant at the 99% level.

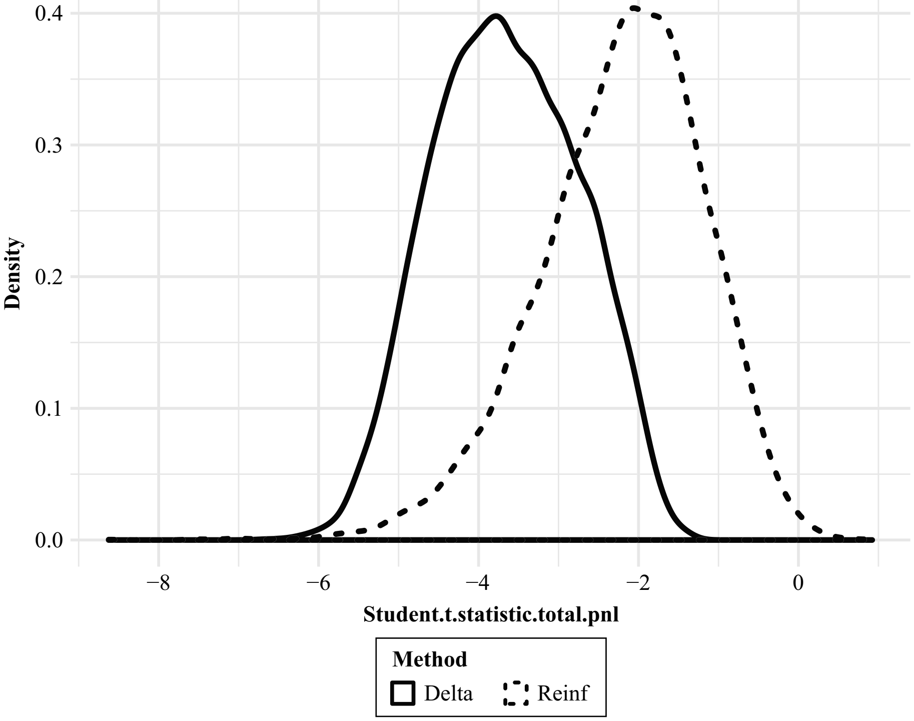
One can also gauge the efficacy of an automatic hedging model by how often the total P&L (including the hedge and all costs) is significantly less than zero. For both policies (delta and reinf), we computed the *t*-statistic of total P&L for each of our out-of-sample simulation runs and constructed kernel density estimates (see Exhibit 5). The reinf method is seen to outperform: Its *t*-statistic is much more often close to zero and insignificant.

CONCLUSIONS

The main contribution of this article is to show that with RL one can train a machine learning algorithm to hedge an option under realistic conditions. Somewhat remarkably, it accomplishes this without the user providing any of the following pertinent pieces of information:

EXHIBIT 5

Kernel Density Estimates of the t-Statistic of Total P&L for Each of Our Out-of-Sample Simulation Runs and for Both Policies Represented Previously (delta and reinf)



Note: The reinf method is seen to outperform in the sense that the t-statistic is much more often close to zero and insignificant.

(1) the strike price K, (2) the stock price process, (3) the volatility of the price process, (4) the BSM formula, (5) the payoff function $(S - K)_+$ at maturity, and (6) any of the Greeks. This is the financial derivatives analogue of the examples of Mnih et al. (2013) and Mnih et al. (2015), wherein computers learned to play games without knowing the rules.

A key strength of the RL approach is that it does not make any assumptions about the form of trading cost. RL learns the minimum variance hedge subject to whatever transaction cost function one provides. All it needs is a good simulator in which transaction costs and options prices are simulated accurately. This has the interesting implication that any option that can be priced can also be hedged, whether or not the pricing is done by explicitly constructing a replicating portfolio—whether or not a replicating portfolio even exists among the class of tradable assets.

Our approach does not depend on the existence of perfect dynamic replication. It will learn to optimally trade off variance and cost using whatever assets it is given as potential candidates for inclusion in a hedging portfolio. In other words, it will find the minimum-variance dynamic hedging strategy, whether or not the minimum variance is actually zero (as it typically is in derivatives pricing, where one needs perfect replication to derive a no-arbitrage price). This is important because, in many realistic cases, markets are not complete and hence some of the assets required for perfect replication may not exist.

Another advantage of this approach is that it can deal automatically with position-level constraints. It is part of the structure of any RL problem that, for each possible state s of the environment, the agent has a (potentially state-dependent) list of possible actions. In the examples given, the list of possible actions was taken to be buying or selling up to 100 shares in integer numbers of shares. We note that other trade or position constraints could be incorporated in a straightforward way, simply by modifying the state-dependent list of available actions.

In this article, we leave open several avenues for further research. One obvious point of interest would be to train agents like ours on more sophisticated hardware and hence to take advantage of many more simulations and finer discretization of time. Silver et al. (2017) described various Go players that were trained on clusters with up to 176 GPUs and/or 48 TPUs, with training times ranging from 3 to 40 days. For reference, all of the examples in this article were trained on a single CPU, and the longest training time allowed was one hour.

Transaction costs are not static. The intraday term structure of trading volume has a well-known smile shape (documented by Chan, Christie, and Schultz 1995), with a nontrivial fraction of US equity trading volume occurring in the close and closing auction. Our RL system should handle this sort of complication very well. For instance, the simulator could be augmented with a nuanced cost function that depends on the time of day and add a discrete time-of-day indicator to the state vector.

Another interesting line of research would be to investigate optimal hedging strategies for portfolios of options in the presence of trading costs. Obviously, for low-gamma portfolios, delta-hedging would not be needed so frequently, thus naturally reducing the trading costs for that kind of portfolio. In general, the most cost-effective way to reduce variance is likely to use other options rather than a replicating portfolio of the underlier.

REFERENCES

Almgren, R., and N. Chriss. 1999. “Value under Liquidation.” *Risk* 12 (12): 61–63.

——. 2001. “Optimal Execution of Portfolio Transactions.” *Journal of Risk* 3: 5–40.

Almgren, R., and T. M. Li. 2016. “Option Hedging with Smooth Market Impact.” *Market Microstructure and Liquidity* 2 (1): 1650002.

Bank, P., H. M. Soner, and M. Vob. 2017. “Hedging with Temporary Price Impact.” *Mathematics and Financial Economics* 11 (2): 215–239.

Black, F., and M. Scholes. 1973. “The Pricing of Options and Corporate Liabilities.” *Journal of Political Economy* 81 (3): 637–654.

Boyle, P. P., and T. Vorst. 1992. “Option Replication in Discrete Time with Transaction Costs.” *The Journal of Finance* 47 (1): 271–293.

Buehler, H., L. Gonon, J. Teichmann, and B. Wood. 2018. “Deep Hedging.” *arXiv* 1802.03042.

Chan, K. C., W. G. Christie, and P. H. Schultz. 1995. “Market Structure and the Intraday Pattern of Bid–Ask Spreads for NASDAQ Securities.” *The Journal of Business* 68 (1): 35–60.

Figlewski, S. 1989. “Options Arbitrage in Imperfect Markets.” *The Journal of Finance* 44 (5): 1289–1311.

Friedman, J., T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Berlin: Springer, 2001.

Grannan, E. R., and G. H. Swindle. 1996. “Minimizing Transaction Costs of Option Hedging Strategies.” *Mathematical Finance* 6 (4): 341–364.

Halperin, I. 2017. “QLBS: Q-Learner in the Black–Scholes (–Merton) Worlds.” *arXiv* 1712.04609.

Henrotte, P. “Transaction Costs and Duplication Strategies.” Graduate School of Business, Stanford University, 1993.

Hull, J., and A. White. 2017. “Optimal Delta Hedging for Options.” *Journal of Banking & Finance* 82: 180–190.

Kaelbling, L. P., M. L. Littman, and A. W. Moore. 1996. “Reinforcement Learning: A Survey.” *Journal of Artificial Intelligence Research* 4: 237–285.

Leland, H. E. 1985. “Option Pricing and Replication with Transactions Costs.” *The Journal of Finance* 40 (5): 1283–1301.

Martellini, L. 2000. “Efficient Option Replication in the Presence of Transactions Costs.” *Review of Derivatives Research* 4 (2): 107–131.

Merton, R. C. 1973. “Theory of Rational Option Pricing.” *The Bell Journal of Economics and Management Science* 4 (1): 141–183.

Merton, R. C., and P. A. Samuelson. *Continuous-Time Finance*. Boston: Blackwell Boston, 1992.

Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. 2013. "Playing Atari with Deep Reinforcement Learning." *arXiv* 1312.5602.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. 2015. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518 (7540): 529.

Ritter, G. 2017. "Machine Learning for Trading." *Risk* 30 (10): 84–89.

Rogers, L. C. G., and S. Singh. 2010. "The Cost of Illiquidity and Its Effects on Hedging." *Mathematical Finance* 20 (4): 597–615.

Saito, T., and A. Takahashi. 2017. "Derivatives Pricing with Market Impact and Limit Order Book." *Automatica* 86: 154–165.

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. 2017. "Mastering the Game of Go without Human Knowledge." *Nature* 550 (7676): 354–359.

Sutton, R. S., and A. G. Barto. *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, MA: MIT Press, 2018 (in progress).

Toft, K. B. 1996. "On the Mean–Variance Tradeoff in Option Replication with Transactions Costs." *Journal of Financial and Quantitative Analysis* 31 (2): 233–263.

Whalley, A. E., and P. Wilmott. 1997. "An Asymptotic Analysis of an Optimal Hedging Model for Option Pricing with Transaction Costs." *Mathematical Finance* 7 (3): 307–324.

ACTIVE THINKING®