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. 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.

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 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

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.

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.

[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.

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]

[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).

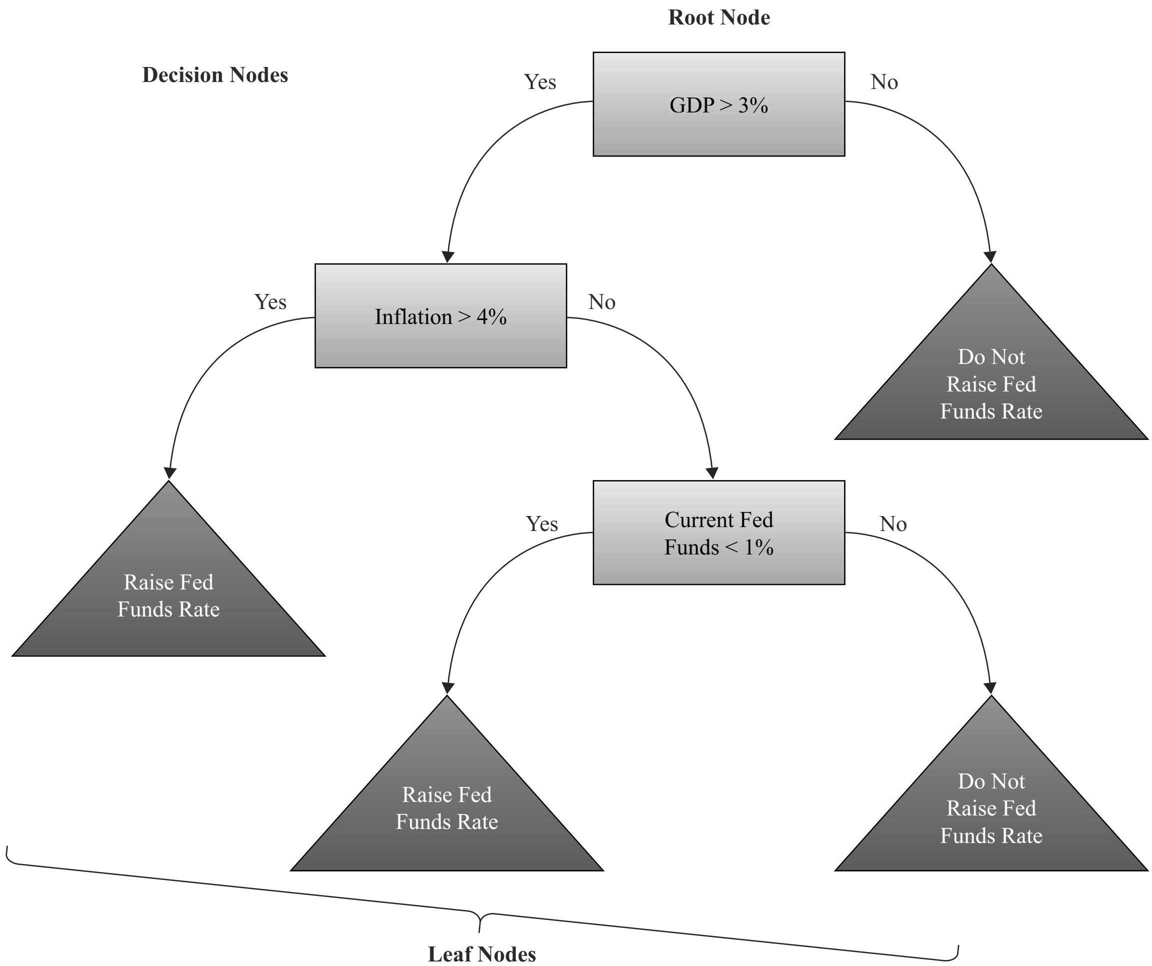
When a bootstrap is conducted, some observations are left out of the bootstrap. These are called *out-of-bag* (OOB) data and are used for measuring estimation error and variable importance. If trees are large, using all the trees may produce a false level of confidence in the predictions of the response 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 jth bootstrap sample and $\hat{h}_j(x)$ denote the prediction x from the jth 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 $\hat{f}_{\text{OOB}}(x_i) = \frac{1}{J_i} \Sigma_{j \in \mathcal{J}_i} \hat{h}_j(x_i)$.

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

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

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

[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: $supp\ \left((VR_t < 1,\ \hat{r}_{t+1} > \alpha)\right) = \dfrac{\#\ of\ months\ (VR_t < 1,\ \hat{r}_{t+1} > \alpha)}{Window\ Length}$

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

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

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 $N \neq 0$; $w_i = \dfrac{100\%}{\#\ Assets}$ 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, $X \Rightarrow Y$, where*

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

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

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

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

Exhibit 8

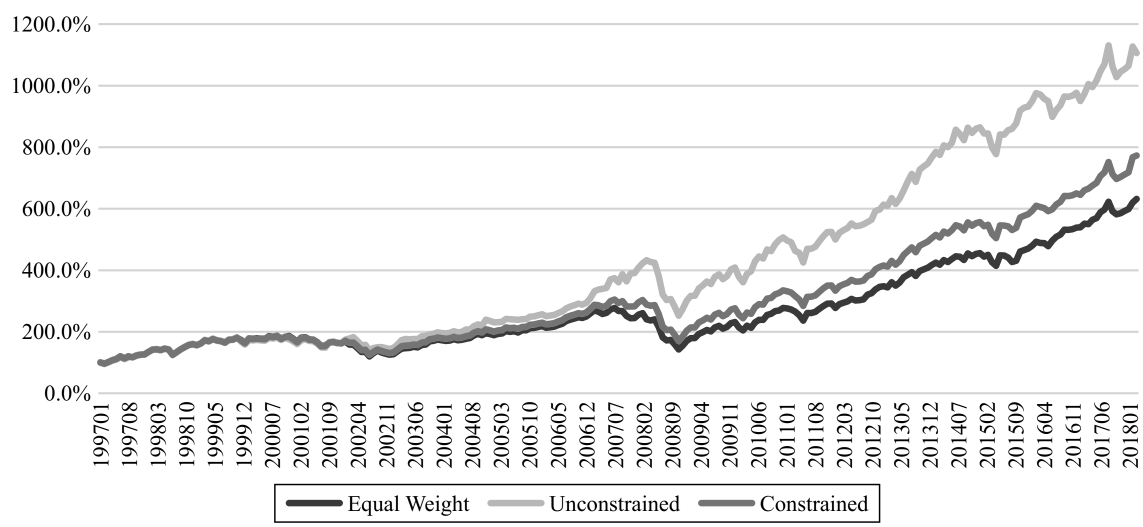
Sector Rotation Strategy Backtest and Bootstrap Results (Jan 1997 to Aug 2018)

	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.

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

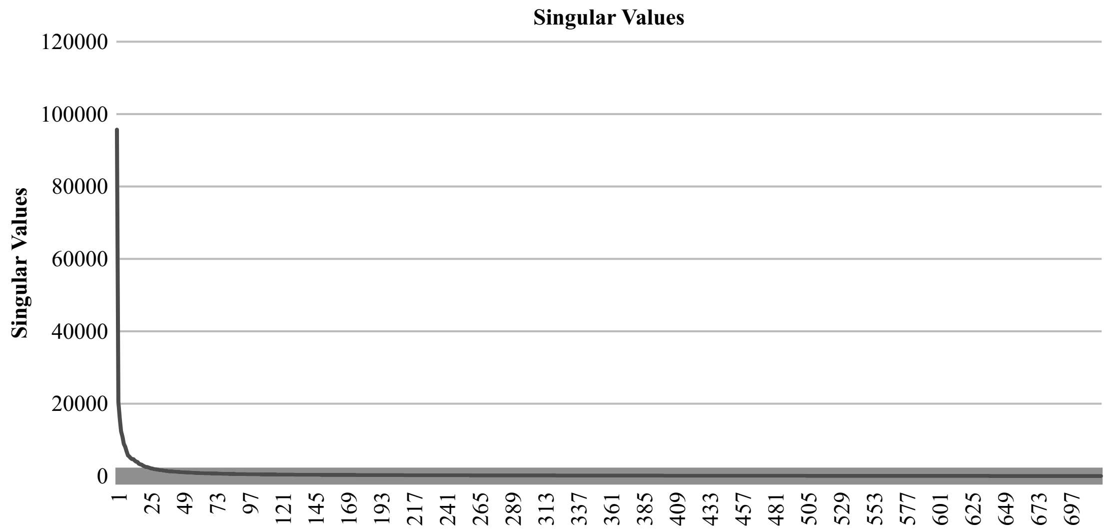
$$X = USV' \tag{1}$$

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

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

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, $i = 0, 1, ..., n$, we can express the investment portfolio as

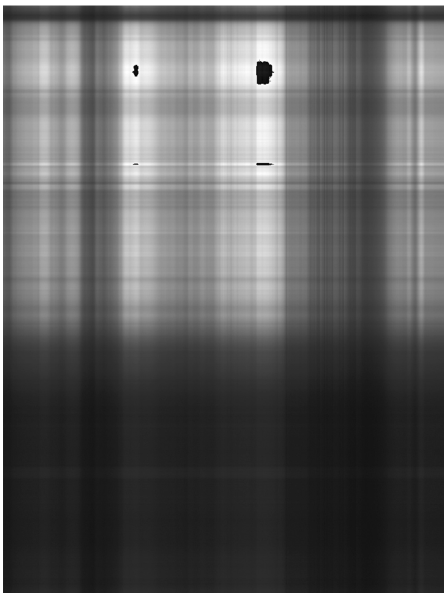
$$w_0X_0 + w_1X_1 + \cdots + w_nX_n \tag{3}$$

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

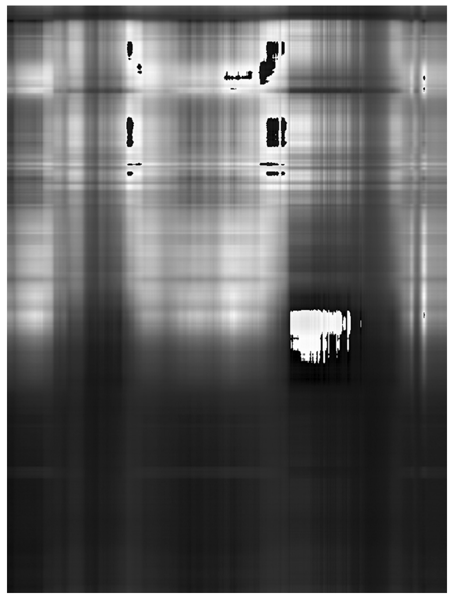
EXHIBIT 3

Reconstruction of the Image of Exhibit 1

Panel A: Reconstruction with Just the First Eigenvalue



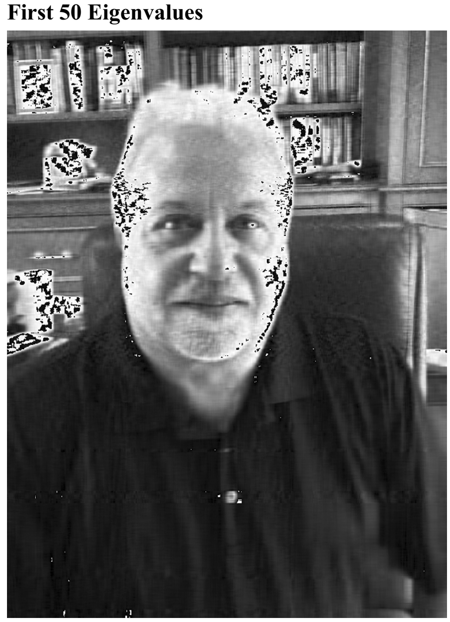
Panel B: Reconstruction with the First Two Eigenvalues



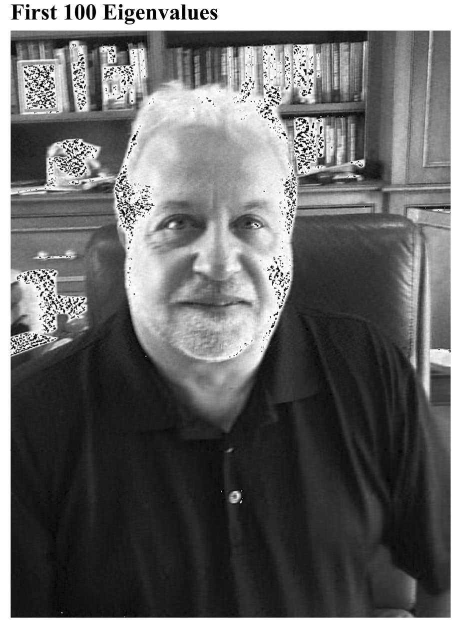
Panel C: Reconstruction with the First Five Eigenvalues (the outlines of the figure are beginning to appear)

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:

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

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

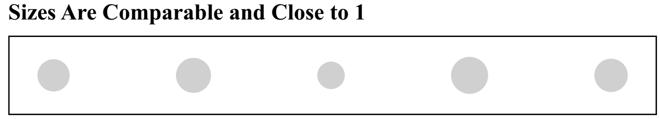
Comparative Dispersion of Eigenvalues via Gershgorin Radii for the Correlation Matrix of Exhibit 4 and the Inverse

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)

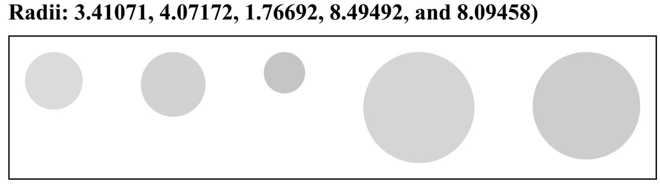


EXHIBIT 7

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.

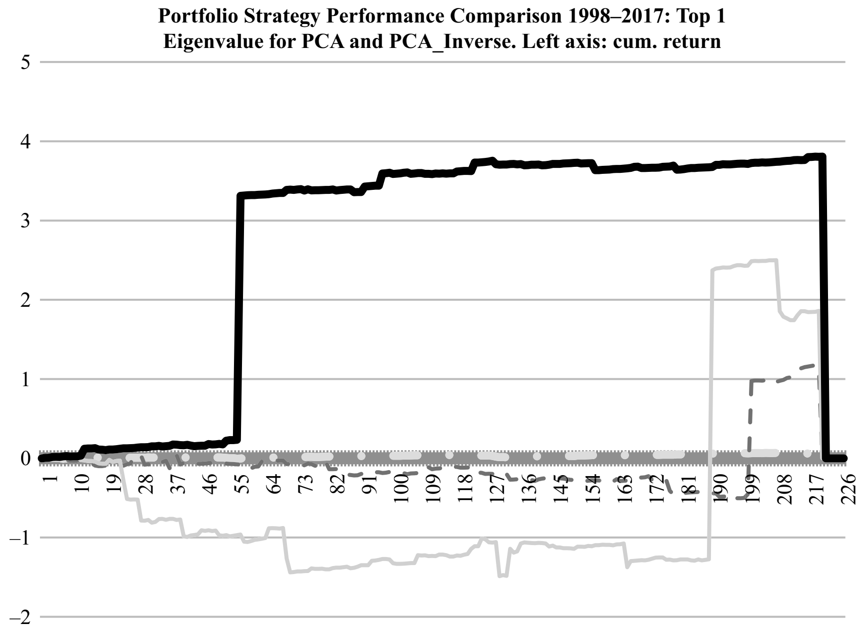
The portfolio compositions do not change from 1998 through 2017. The portfolio weights are computed on the last trading day of each month. The EW weights do not change, unless the originally chosen stock is no longer trading. For MVO, PCA, and PCA_Inverse, the correlation matrixes used to set portfolio 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.

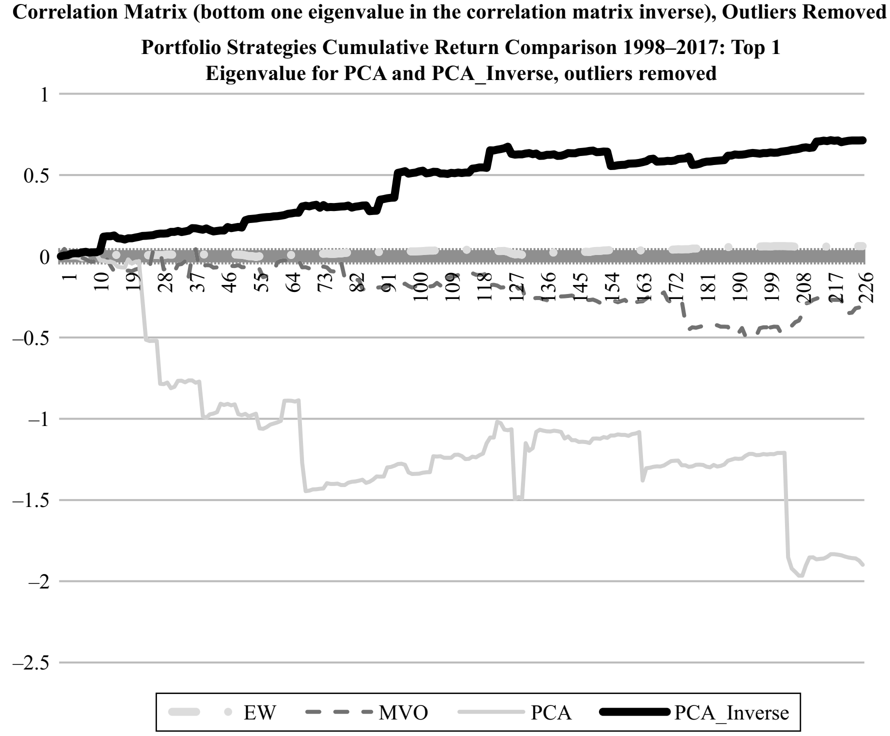
EXHIBIT 9

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



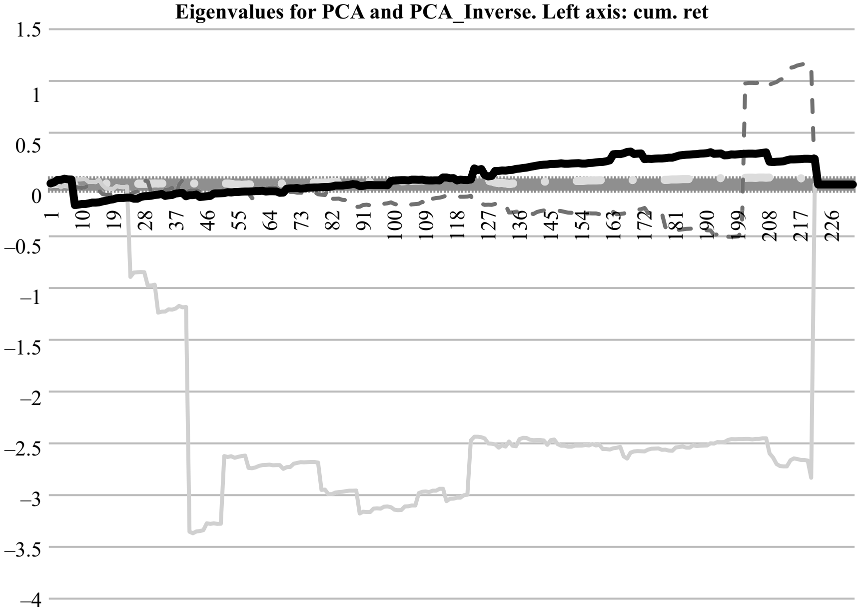
(continued)

Exhibit 9 *(continued)*

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

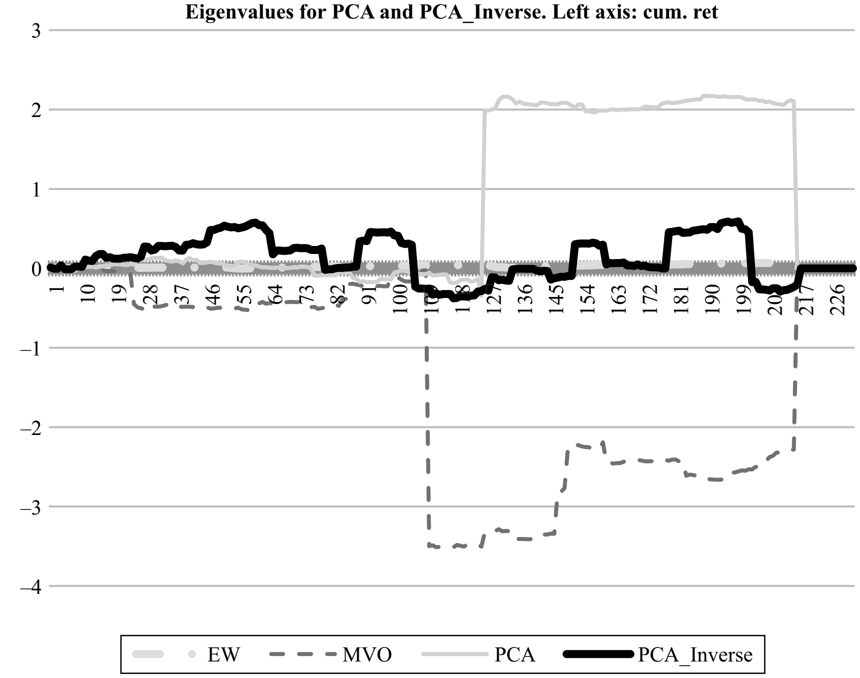
Panel C: S&P 500 Strategy, Monthly Reallocations: Keep the Top 1% of Eigenvalues in the Correlation Matrix (bottom 1% of eigenvalues in the correlation matrix inverse)

Portfolio Strategy Performance Comparison 1998–2017: Top 1% Eigenvalues for PCA and PCA_Inverse. Left axis: cum. ret



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

Portfolio Strategy Performance Comparison 1998–2017: Top 10% Eigenvalues for PCA and PCA_Inverse. Left axis: cum. ret



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.

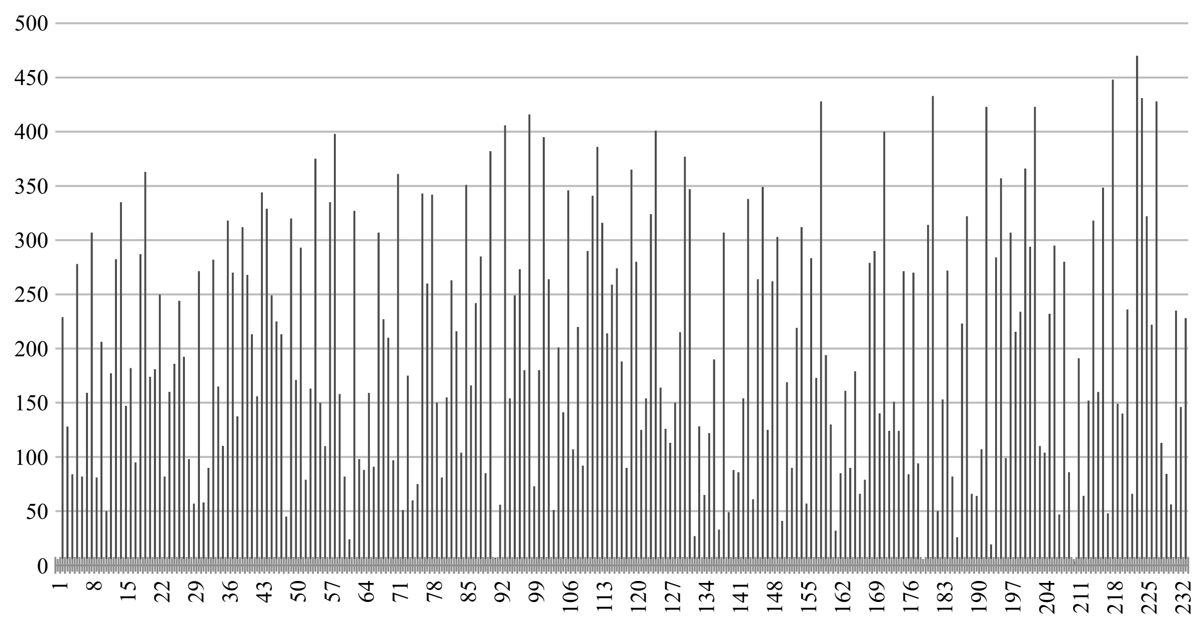
EXHIBIT 10
Sharpe Ratios on Strategy Performance, S&P 500, 1998–2017, Monthly Reallocation

	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

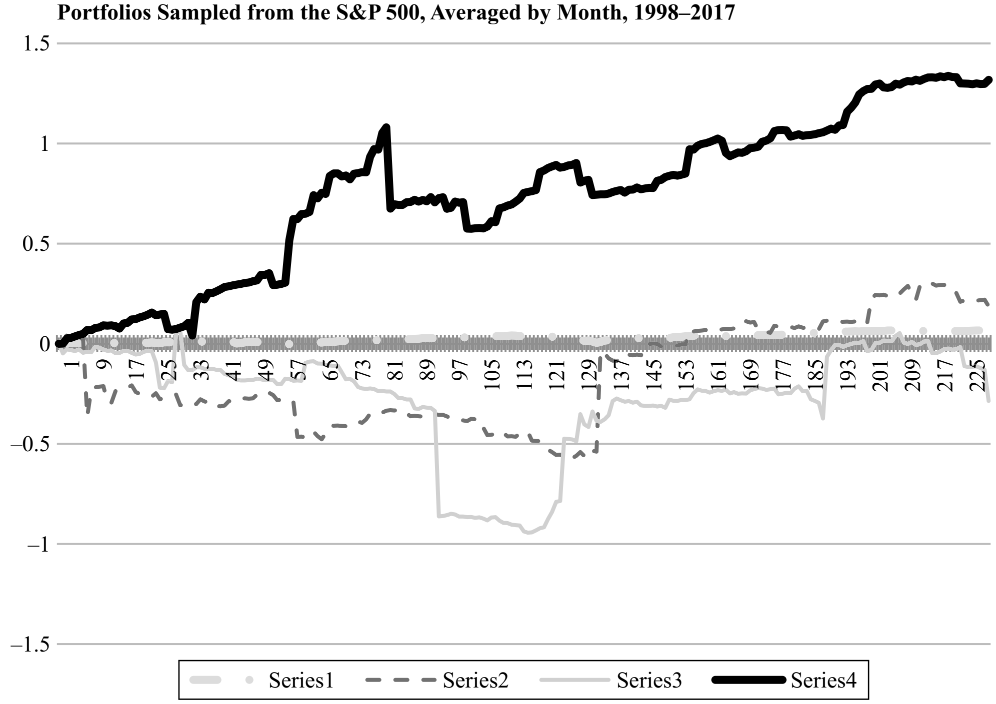
		MVO	PCA	PCA_Inverse
Top 1 Eigenvalue	|weight| > 1%	375.49 (82.16)	375.15 (82.94)	192.44 (117.07)
	|weight| > 2%	343.57 (88.33)	343.40 (89.77)	112.64 (116.94)
Top 1% of Eigenvalues	|weight| > 1%	375.49 (82.16)	376.49 (84.30)	190.90 (115.95)
	|weight| > 2%	343.57 (88.33)	346.27 (91.65)	109.98 (115.37)
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

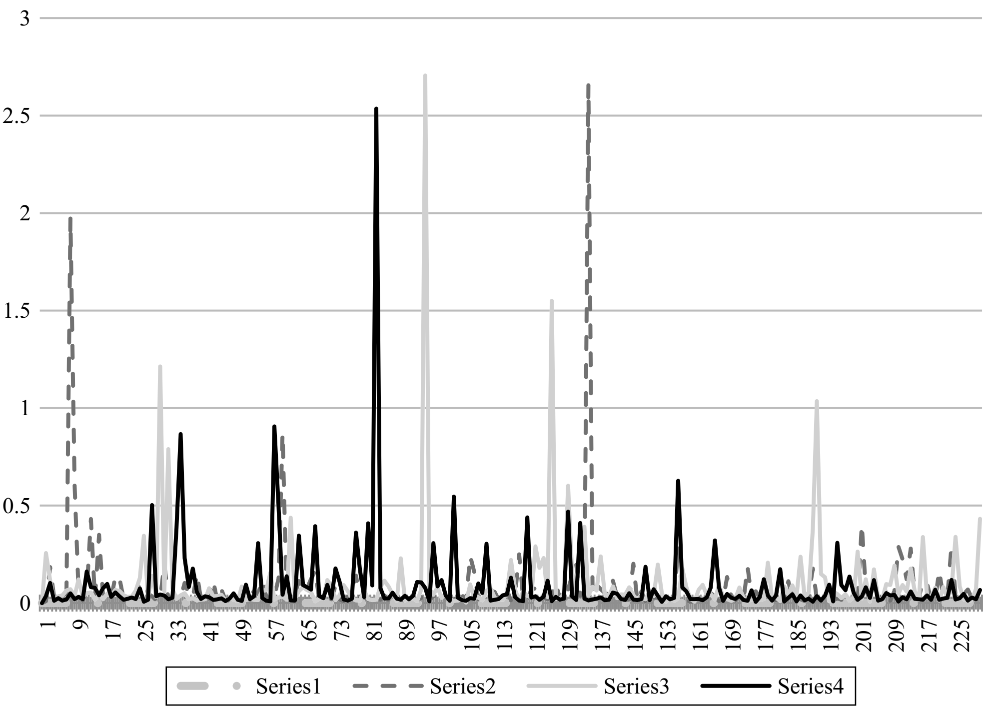


(continued)

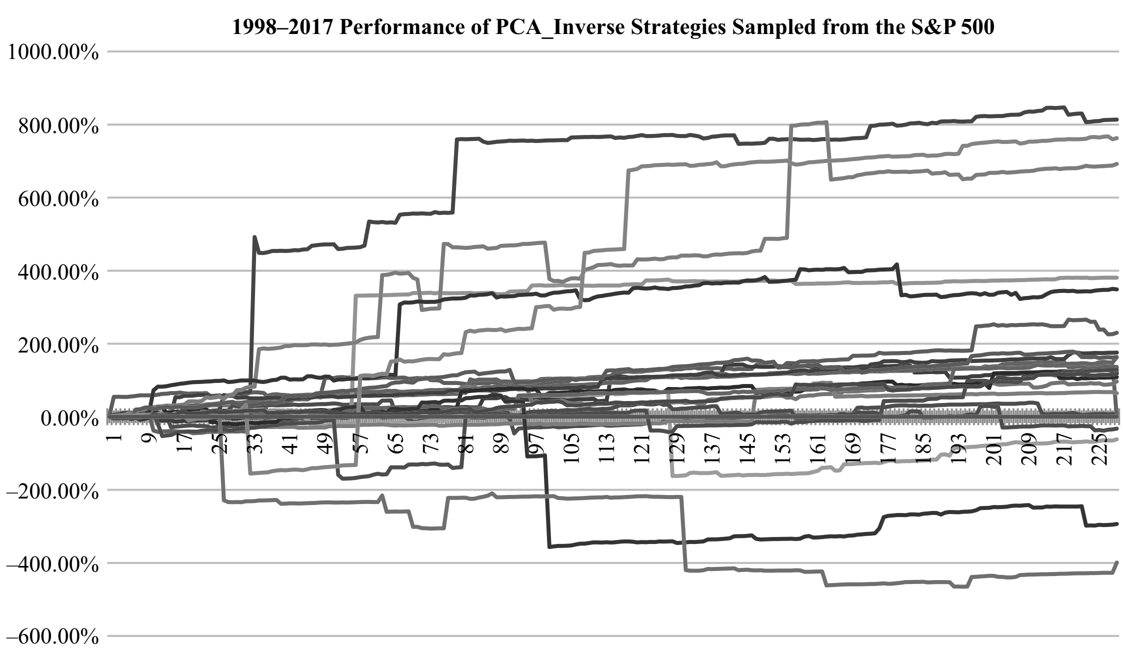
E X H I B I T 1 4 *(continued)*

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



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.

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.

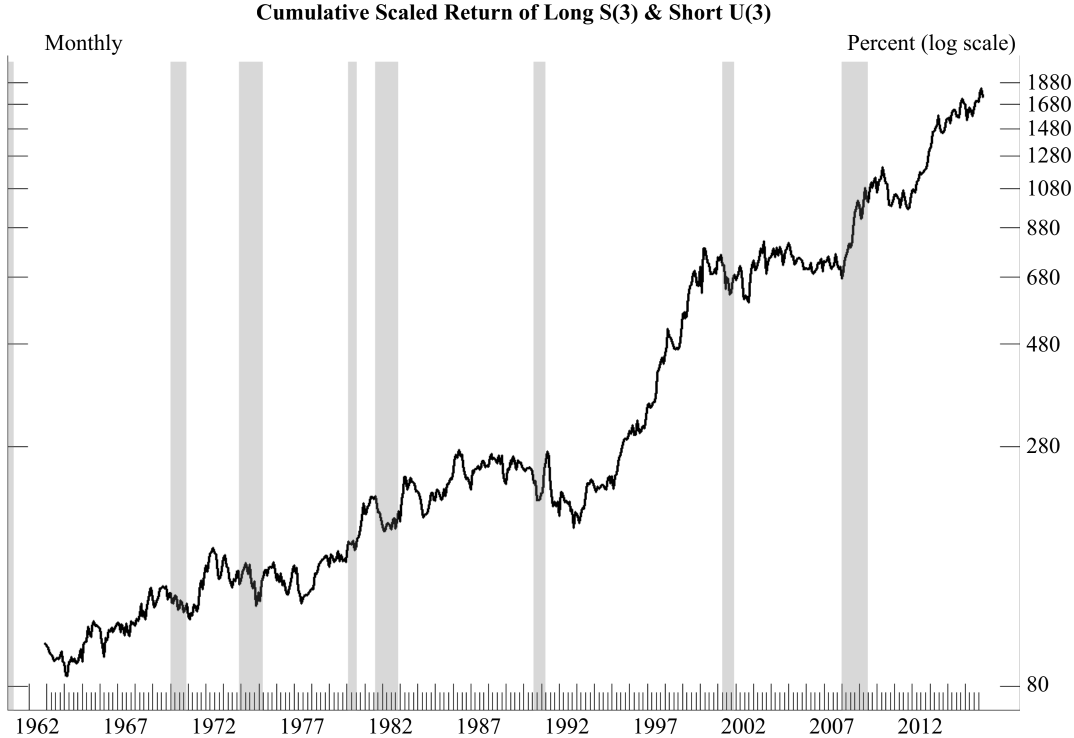
Given the low cost of entering the data-mining business, investors need to be wary. Consider the long–short equity strategy whose results are illustrated in Exhibit 1. This is not a fake exhibit.[1] It represents a market-neutral strategy developed on NYSE stocks from 1963 to 1988, then validated out 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).

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

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.

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.

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.

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.

[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

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

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 $D = \{(\kappa_0, \kappa_1, \kappa_2);\ 0 \le \kappa_j \le 1, \Sigma_{t=0}^{2} \kappa_t = 1\}$ 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

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

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

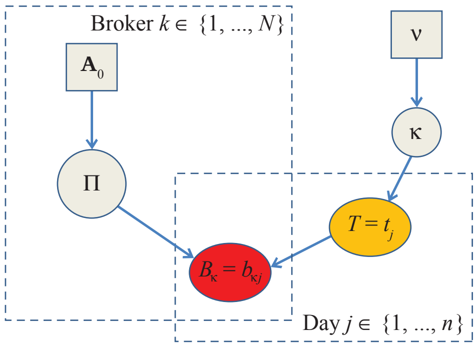
[8] A conjugate prior is one that leads to a posterior distribution that is within the same parametric family as the prior, which therefore leads to greatly simplified Bayesian analysis. See Bishop (2006).

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

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

Exhibit 2

Graphical Model of Our IBCC Implementation



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 $\pi_t^{(l)}$ 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 $\pi_t^{(l)}$ 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 $\log q^*(t) = E_{\boldsymbol{\kappa}} \log \kappa_t + \Sigma_{l=1}^{N} E_{\pi_t^{(l)}} \log \pi_{tb_l}^{(l)} + \text{Constant}$. Standard properties of the Dirichlet distribution (e.g., Bishop 2006) give $E_{\boldsymbol{\kappa}} \log \kappa_t = \Psi(\nu_t^*) - \Psi(\Sigma_{j=0}^{2} \nu_j^*)$ and $E_{\pi_t^{(l)}} \log \pi_{tb_l}^{(l)} = \Psi(\alpha_{tb_l}^{(l*)}) - \Psi(\Sigma_{s=0}^{3} \alpha_{ts}^{(l*)})$, 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 $\log q^*(\boldsymbol{\kappa}) = \Sigma_{j=0}^{2} q(t = j) \log \kappa_j + \Sigma_{j=0}^{2} (\nu_j^* - 1) \log \kappa_j$ + 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 $\pi_j^{(l)}$ 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 $q^*(\boldsymbol{\Pi}) = \Pi_{l=1}^{N} \Pi_{j=0}^{2} q^*(\pi_j^{(l)})$. Extracting the $\pi_j^{(l)}$ terms and taking expectation with respect to $q^*(t)$ thereby yields $\log q^*(\pi_j^{(l)}) = \Sigma_{j=0}^{2} q^*(t = j) \Sigma_{l=1}^{N} \log \pi_{tb_l}^{(l)} + \Sigma_{s=0}^{3} (\alpha_{js}^{(l*)} - 1) \log \pi_{js}^{(l)}$ + Constant. Gathering together terms in $\log \pi_{jb}^{(l)}$ now shows $q^*(\pi_j^{(l)})$ 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, $position_{it} = signal_{it} / \Sigma_i |signal_{it}|$, 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.

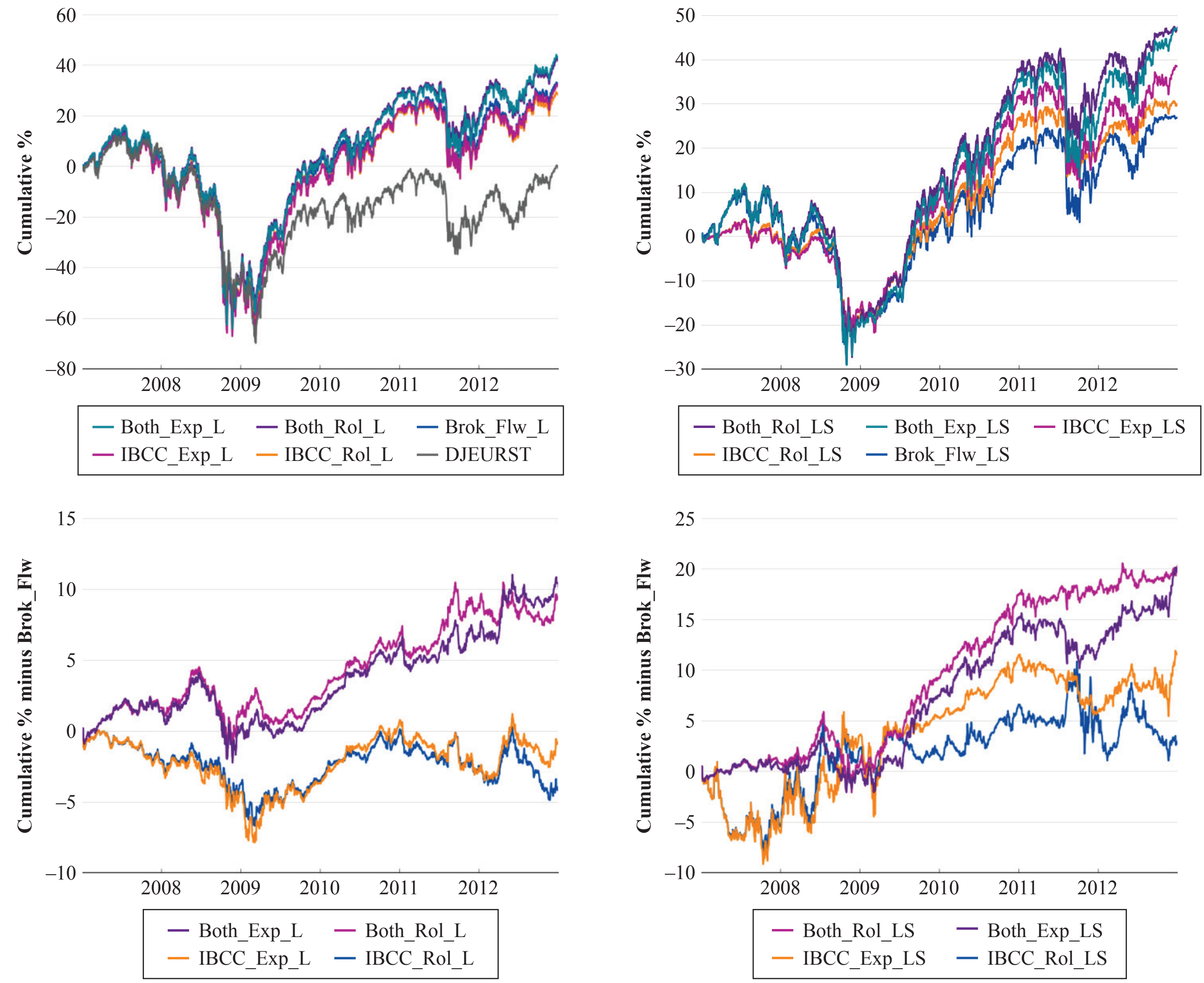
The reference index used for the intercept and slope estimates, α and β, reported in the following is the Euro Stoxx,[22] the same index we used in defining

[21] Gross exposure is defined as $\Sigma|pos_i|$, 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

Side	Model	Mean	Vol	Alpha	Alpha *t*-Stat	Beta	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.

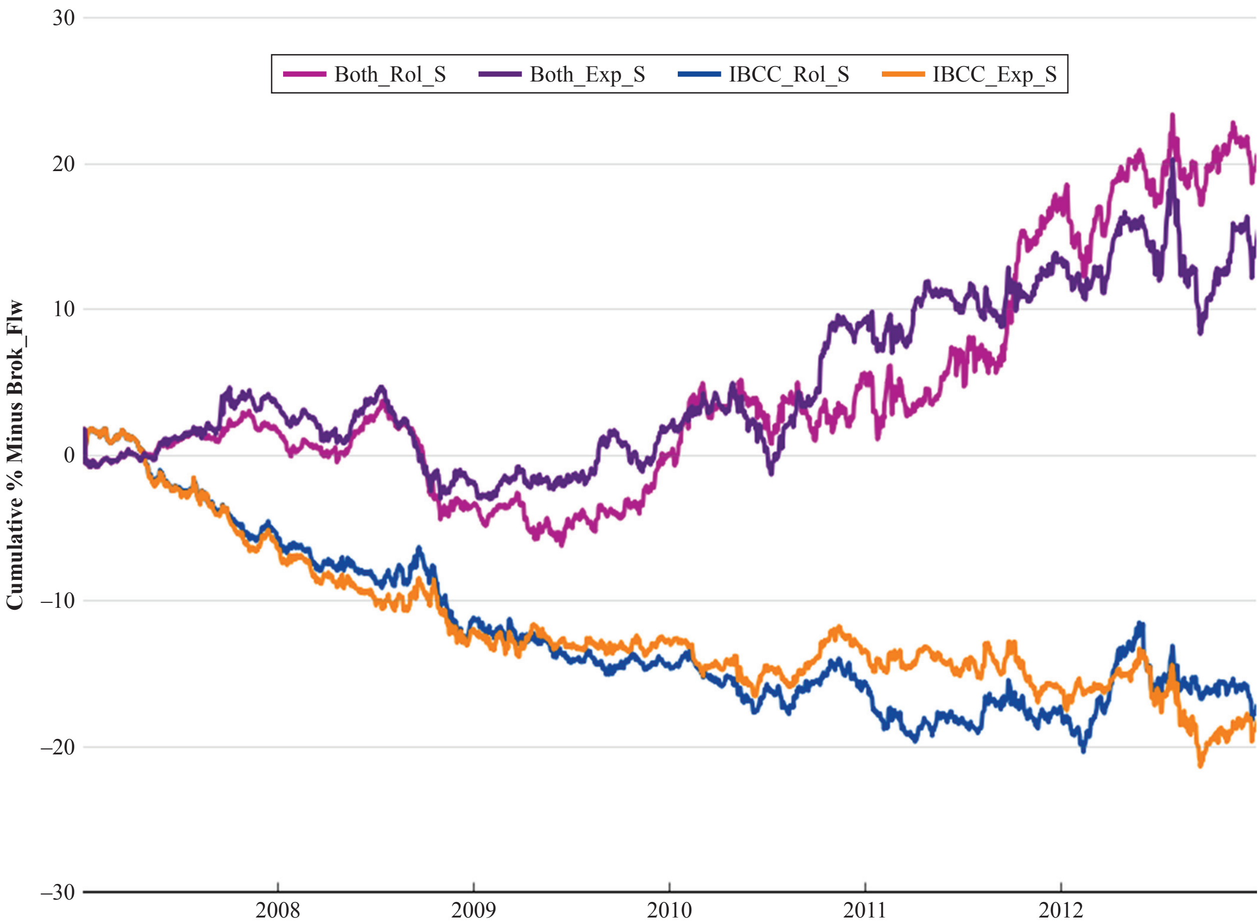
Robustness Checking—Selectivity of the Trading Rule

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

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

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

Model	c	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	5.13	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

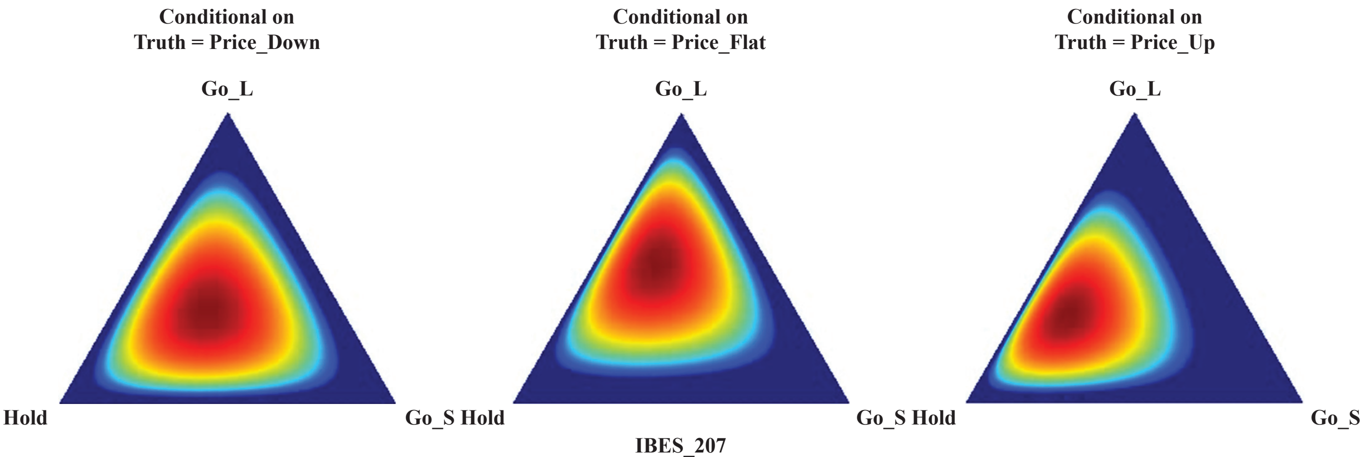
Notes: Here the trade holding period and the holding period for assessing truths are constrained to be equal. The ±5% threshold for converting stock returns to truths is scaled to yield a similar number of truths for each horizon, using the usual random walk property that $\sigma(X_t) \propto \sqrt{t}$*, 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



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.

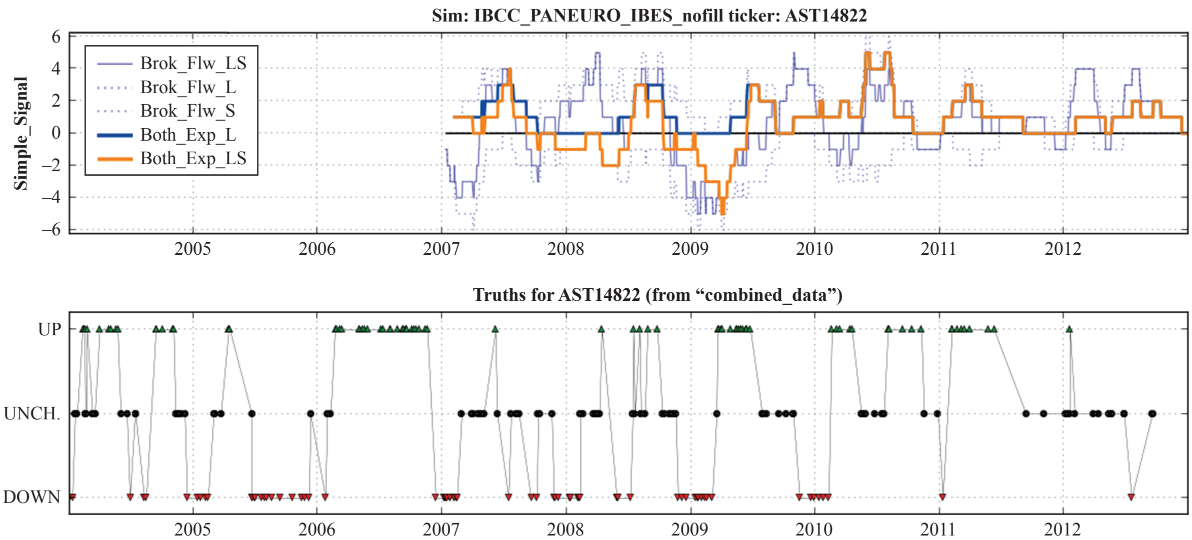
Exhibit 16

Diagnostic Panel for the Stock with Identifier AST14822

(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.