Modelling Analysts' Recommendations via Bayesian Machine Learning*

David Bew
Man-AHL, Oxford, OX2 6ED, UK

Campbell R. Harvey
Fuqua School of Business, Duke University, Durham, NC 27708, USA
National Bureau of Economic Research, Cambridge, MA 02912, USA

Anthony Ledford
Man-AHL, Oxford, OX2 6ED, UK
Oxford-Man Institute of Quantitative Finance, University of Oxford, Oxford, OX2 6ED, UK

Sam Radnor
Independent

Andrew Sinclair
Independent

ABSTRACT

We apply state-of-the-art Bayesian machine learning to test whether we can extract valuable information from analysts' recommendations of stock performance. We use a probabilistic model for independent Bayesian classifier combination that has been successfully applied in both the physical and biological sciences. The technique is ideally suited for the particular problem where any individual analyst only focuses on a handful of the thousands of companies and it allows for dynamic inference as we track the performance of the analysts through time. The results suggest this technique holds promise in extracting information that can be deployed in active investment management.

JEL: G11, G14, G17, C11, C58, M41

Keywords: *Variational Bayes, VB, IBCC, Machine Learning, Data Science, Analysts' Forecasts, IBES, Analysts' Recommendations, Forecast Fatigue, Forecast Combination, Dirichlet Distributed Prior, Galaxy Zoo.*

---

* Version October 23, 2018. Corresponding author: Campbell R. Harvey (cam.harvey@duke.edu).

1. Introduction

In 2009, a unique citizen science project was launched called the Galaxy Zoo Supernovae project.[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, there may be many astronomers 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], or IBCC, originally defined by Ghahramani and Kim (2003). The Galaxy Zoo data was analyzed by Simpson et al. (2013) with striking 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 supernovae have to do with finance? It turns out that there are striking similarities to the problem facing an investment manager in evaluating analysts' recommendations. Similar to 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) – 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 the analysts and, hence, the IBCC method, given its track record in the physical and biological sciences, is a logical place to start.

The goal of our paper is to apply IBCC to the I/B/E/S forecast universe to determine 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) details 575 studies many of which are focused on analysts' forecasts – and this article is 20 years out of date. A search of SSRN's Financial Economics Network and Accounting Research Network reveals over one thousand papers dealing with analysts' forecasts.[4]

Despite the large quantity of research, ours is the first paper (that we know of) to apply IBCC to the important problem of how to combine analysts' recommendations. Previous applications of IBCC in

[1] See, https://blog.zooniverse.org/2012/08/03/i-for-one-welcome-our-new-machine-collaborators/

[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.

[4] Early reviews of the literature are found in Givoly and Lakonishok (1983), Schipper (1991) and Brown (1993). A more recent treatment is in Bradshaw (2011).

economics include Levenberg *et al.* (2013), where the focus is on forecasting the trend of the U.S. Nonfarm Payrolls, and Levenberg *et al.* (2014) where sentiment measures obtained using sentence-level language analysis is incorporated. 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 over-fitting becomes an important consideration. Bayesian models are not as prone to over-fitting as models which require point estimates to be specified for large numbers of parameters. This is because 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 which 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 using them 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 to groups of analysts. Restrictions currently in place require that we only report at the broker level.

We realize that predicting financial outcomes remains hard, even when expansive datasets and sophisticated machine-learning models are available. Our primary aim is not about identifying the best analyst or broker, but making a coherent ensemble forecast where the weight given to each broker is driven by the length and quality of their 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 paper is organized as follows. The second section discusses the data, focusing on nonstandard features such as its categorical nature, dependence structure and sparsity, i.e., characteristics which necessitate a bespoke modelling treatment. The third part 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.

2. Description of the data modelling 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) which describes the directional price movement of the company's stock 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.

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

A vast amount of analyst data are available on both the individual stocks and the various sub-sectors within international equities markets. Our focus here is on recommendation data from the Thomson Reuters I/B/E/S database, a data source which 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 since different analysts use a variety of ratings schemes, each recommendation received from a contributing analyst is mapped by Thomson Reuters to one of their five Standard Ratings:

Strong Buy, Buy, Hold, Underperform, Sell.

There are several factors which 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 a standard regression approach. Additionally, the recommendation database we examine has the following characteristics:

A. **Very high dimensionality:** Recommendations are received on thousands of stocks from thousands of individual analysts.

B. **Extreme sparsity:** Typically only a small number of analysts issue recommendations on any particular stock on any particular day; the rest say nothing.

C. **Dependence:** We expect analyst recommendations to be statistically dependent for a number of reasons.
   i. **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, $(X_1, X_2, \dots, X_m, Z)$ say, is such that $\Pr(X_1, X_2, \dots, X_m|Z) = \Pr(X_1|Z) \times \Pr(\mathrm{X}_2|\mathrm{Z}) \times \cdots \times \Pr(X_m|Z)$, then the $X's$ are said to be conditionally independent given Z, or equivalently, the $X's$ are independent conditional on Z. The IBCC model makes extensive use of such a conditional independence structure, see Section 3.
   ii. **Temporal dependence:** Analyst views typically update gradually, and analysts often restate their previous recommendations. This leads to serial correlation. Group behaviour

between analysts can also generate serial correlation, for example some analysts leading opinion and others following consensus.

D. **Lack of consistency**: Although the analyst recommendations provided by I/B/E/S are recorded on the common five-category scale given above, for many analysts only two of these categories are populated. Other analysts may use three of the available categories, and others all five. While it is quite possible to deal with this inconsistency using all 5-categores 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 re-label the original I/B/E/S analyst recommendations as Buy, Hold, and Sell. Finally, 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 extension results in recommendations being recorded on the following four-category scale: Missing, Buy, Hold, and Sell.

Accounting for any one of these characteristics within a Bayesian analysis requires detailed probabilistic modelling. 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 are 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.

2.2 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, ΔT say, over which we are interested in predicting the movement of the stock price; for the majority of this study we use $\Delta T = 60$ business days. For each analyst recommendation, we note the day it became public, s say, and calculate $r_{(\mathrm{s},\Delta T)}$ which is defined to be the excess return of the relevant stock over the ΔT period starting the next business day after s and measured relative to the Dow Jones Euro Stoxx Index.[5] We use this together with a relative measure of index volatility to define a categorical truth variable t for each recommendation according to

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

where $RVol_{(s,\Delta T)}$ denotes a unit mean normalized estimator of the volatility of the Dow Jones Eurostoxx Index over the $(\mathrm{s},\Delta T)$ period. Given their obvious interpretations, we refer interchangeably to the truth

[5] Bloomberg ticker: SXXE Index.

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 attention to January 1, 2004 to January 1, 2013 and include only the pan-European region comprising Austria, Belgium, Czech Republic, Cyprus, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Russia, Spain, Sweden, Switzerland, Turkey, and the UK. Additionally, at the announcement time of each recommendation, we apply a filter to the stock universe in order 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. Due to reporting restrictions, we focus this research paper on a broker level.

Aggregating recommendations about the same stock arising on the same day, we obtain the combined recommendations and truths dataset described in Exhibit 1, which 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 non-zero 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 Table 1 are non-zero, meaning 99.7% correspond to the label "Missing". This demonstrates the extreme sparsity of the data object at the heart of our IBCC analysis.

Stock ID	**Date**	**Truth**	**Broker 1**	**Broker 2**	$\cdots$	**Broker N**
#1234	20080710	0	3	0		0
#5678	20120207	0	0	1		2
	20120701	2	2	0		0
	20120314	1	0	3		1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Exhibit 1: The structure of the dataset. 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 non-contributing broker in each row, and the resulting recommendation set {Missing, Hold, Sell, Buy} is encoded $\{0, 1, 2, 3\}$ respectively. Each row contains at least one non-zero recommendation code. A very high proportion of recommendations are recorded as 0, corresponding to "Missing", since only a small number of brokers within each row issue Hold, Sell or Buy recommendations.

[6] This choice of a 1-day aggregation period is arbitrary, and is something we return to later. From the previous discussion about group behaviour, we would expect statistical dependence between rows at this aggregation were analysts to issue recommendations on a stock prompted by others doing so.

3. The IBCC model: probabilistic specification and construction of the posterior

The independent Bayesian classifier combination (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 Table 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 3-vector of probabilities so that $\kappa_0 + \kappa_1 + \kappa_2 = 1$. All this specification is saying is that 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 known at this stage.

The next step is to specify, for each broker, three separate distributions to describe their recommendation behaviour 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 $k \in \{1, \dots, N\}$ 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

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

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 4-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 4-dimensional multinomial distributions for each broker, one for each of the possible truth outcomes. Thus for each broker k, we have parameters $\boldsymbol{\pi}_0^{(k)}$, $\boldsymbol{\pi}_1^{(k)}$ and $\boldsymbol{\pi}_2^{(k)}$. Again, the conditioning notation in Equation 2 makes explicit that the parameters $\boldsymbol{\pi}_t^{(k)}$ are assumed known at this point.

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

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

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

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

$$\Pr(\boldsymbol{t}, \boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_N}) = \prod_{j=1}^{n} \left(\kappa_{t_j} \prod_{l=1}^{N} \pi_{t_j b_{lj}}^{(l)} \right), \tag{3}$$

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

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 $k \in \{1, \dots, N\}$ respectively, as fixed parameters. In a frequentist analysis, we would need to estimate these, for example 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 treating 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 $k \in \{1, \dots, N\}$. Since the truth and broker recommendation distributions are all examples of multinomial distributions, we choose to use the family of Dirichlet distributions as priors, as the 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 $D = \left\{(\kappa_0, \kappa_1, \kappa_2);\ 0 \le \kappa_j \le 1, \sum_{t=0}^{2} \kappa_t = 1\right\}$ given by $\Pr(\boldsymbol{\kappa} | \boldsymbol{\nu}) = C(\boldsymbol{\nu}) \prod_{t=0}^{2} \kappa_t^{\nu_{0t}-1}$, where $C(\boldsymbol{\nu}) = \Gamma(\nu_{00} + \nu_{01} + \nu_{02}) / \{\Gamma(\nu_{00}) \times \Gamma(\nu_{01}) \times \Gamma(\nu_{02})\}$, the 3-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 $k \in \{1, \dots, N\}$ and conditional on truth $t \in \{0, 1, 2\}$, we assume $\{\pi_{to}^{(k)}, \pi_{t1}^{(k)}, \pi_{t2}^{(k)}, \pi_{t3}^{(k)}\}$ has a four-dimensional Dirichlet prior with hyper-parameters $\{\alpha_{0,t0}^{(k)}, \alpha_{0,t1}^{(k)}, \alpha_{0,t2}^{(k)}, \alpha_{0,t3}^{(k)}\}$. In order to condense the notation, we denote the complete set of broker recommendation probabilities conditional on each truth by $\boldsymbol{\Pi} = [\{\pi_{to}^{(k)}, \pi_{t1}^{(k)}, \pi_{t2}^{(k)}, \pi_{t3}^{(k)}\} : t = 0, 1, 2;\ k = 1, \dots, N]$ and their corresponding hyperparameters by $\mathbf{A_0} = [\{\alpha_{0,t0}^{(k)}, \alpha_{0,t1}^{(k)}, \alpha_{0,t2}^{(k)}, \alpha_{0,t3}^{(k)}\} : t = 0, 1, 2;\ k = 1, \dots, N]$.

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:

$$\Pr(\boldsymbol{\kappa}, \boldsymbol{\Pi}, \boldsymbol{t}, \boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_N} | \boldsymbol{A_0}, \boldsymbol{\nu}) \propto \prod_{j=1}^{n} \left(\kappa_{t_j} \prod_{l=1}^{N} \pi_{t_j b_{lj}}^{(l)} \right) \Pr(\boldsymbol{\kappa} | \boldsymbol{\nu}) \Pr(\boldsymbol{\Pi} | \mathbf{A_0}). \tag{4}$$

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 usual to choose informative priors, however we deliberately choose priors that are flat over their respective domains. This is achieved by

[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 4-dimensional distributions, plus the 3-dimensional truth distribution. In all, this makes $347 \times 4 \times 3 + 3 = 4{,}167$ dimensions.

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 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.

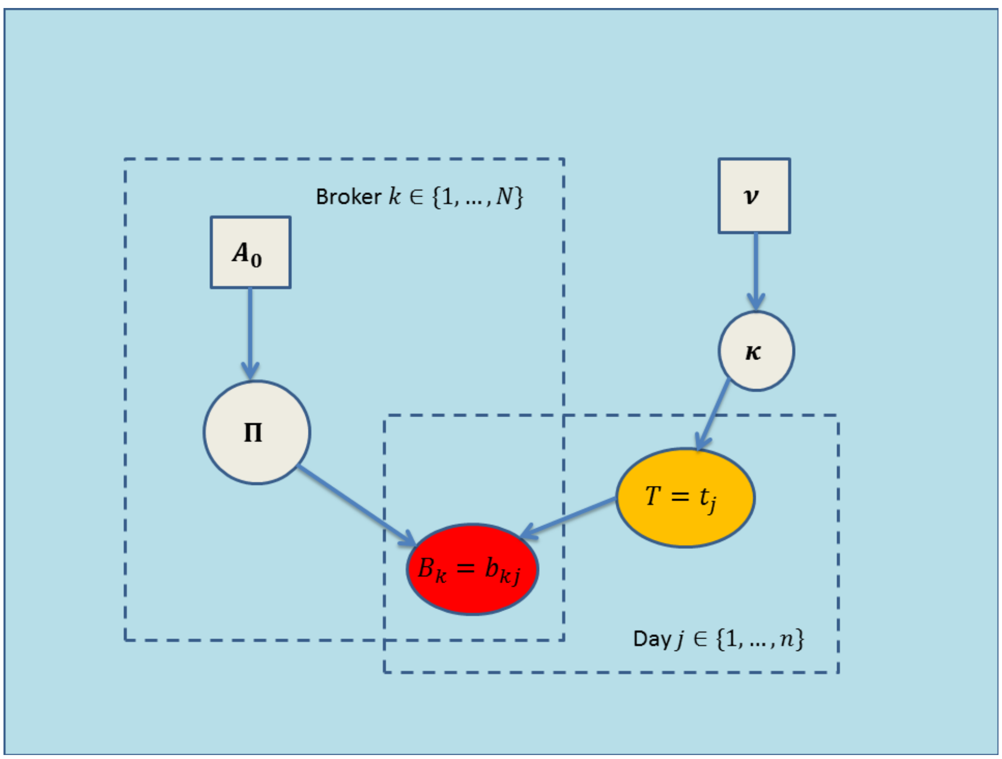

Exhibit 2: Graphical model of our IBCC implementation. Elliptical/circular nodes are variables with a distribution whereas rectangular nodes represent hyperparameter variables that are instantiated with fixed values. The red shaded node represents recommendations, which are observed during both training and prediction. The orange shaded node represents truths, which are observed during training, but have to be inferred during prediction.

4. 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 *et al.* (2018) for detailed treatments, and Fox and Roberts (2011) for a tutorial. 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, i.e., $\log_e(\cdot)$.

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 $p(\boldsymbol{Z}|\boldsymbol{X})$. 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, since 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:

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

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[10] divergence (KL-divergence for short) 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[11]. Standard properties of the KL-divergence include that it is always non-negative, 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 variation*s problem[12].

Variational Bayes considers a restricted but tractable family of distributions to represent $q(\boldsymbol{Z})$, and then seeks the element of that family which 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 $q(\boldsymbol{Z}) = \prod_{i=1}^{m} q_i(\boldsymbol{Z}_i)$. This factorized version of variational approximation has its origins in physics where

[10] The Kullback-Leibler 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, i.e., $KL(f,g) \neq KL(g,f)$. Standard properties include that $KL(f,g) \geq 0$ always, and that $KL(f,g) = 0$ if and only if $f = g$.

[11] To avoid the chance 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).

[12] 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.

it is called *mean field theory*[13]. Thus among all distributions of the form $q(\boldsymbol{Z}) = \prod_{i=1}^{m} q_i(\boldsymbol{Z}_i)$, we seek the distributions $q_i^*(\boldsymbol{Z}_i)$ that jointly maximize $L(q)$. To be clear, other than the assumed factorization structure $q(\boldsymbol{Z}) = \prod_{i=1}^{m} q_i(\boldsymbol{Z}_i)$, no further assumptions about $q(\boldsymbol{Z})$ are required.

Substituting our assumed factorization $q(\boldsymbol{Z}) = \prod_{i=1}^{m} q_i(\boldsymbol{Z}_i)$ into the definition of $L(q)$ given above, and adopting the notation $q_i = q_i(\boldsymbol{Z}_i)$, we obtain $L(q) = \int \Pi_i q_i \{\log p(\boldsymbol{X}, \boldsymbol{Z}) - \sum_i \log q_i\}\, d\boldsymbol{Z}$. We now re-write 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 $E_{i \neq j}[\log p(\boldsymbol{X}, \boldsymbol{Z})] = \int \log p(\boldsymbol{X}, \boldsymbol{Z})\ \Pi_{i \neq j} q_i\ d\boldsymbol{Z}_i$. 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 $j \in \{1, \dots, m\}$ provides conditions for the maximum of $L(q)$ subject to the assumed factorization $q(\boldsymbol{Z}) = \prod_{i=1}^{m} q_i(\boldsymbol{Z}_i)$. 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 $j \in \{1, \dots, m\}$, 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 exponential family type (Bernardo and Smith, 1994), this maximum can be shown to be the global maximum within the family of factorized distributions.

4.1 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 Δt-period, all truths within our training data are completely observed as 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 $\Pr(T = t | B_1 = b_1, B_2 = b_2, \dots, B_N = b_N)$, rather than the posterior, as we

[13] See Parisi (1988).

wish to forecast the truth outcome conditional on, for example, today's constellation of broker recommendations[14].

Although it is possible to extend the IBCC model to include explicit temporal structure as in 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 census dates, ensuring additionally that a buffer of duration Δt 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 $\Pr(T = t|B_1 = b_1, B_2 = b_2, \dots, B_N = b_N)$ for each constellation of broker recommendations that arises until the next census date. Learning remains halted over this prediction phase, so each constellation of analyst recommendations we use in prediction gets 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 3-year lookback from each evaluation date. In principle, the evaluation dates could be chosen to index each business day, however for practical reasons[15] we set them quarterly, to the first day of March, June, September and December.

Let index $i \in (1, \dots, n_r)$ denote the rows of the training data, renumbered as required for the rolling window case. Since 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 that:

1. The posterior of $\boldsymbol{\kappa}$ is a Dirichlet distribution with parameter $\boldsymbol{\nu}^* = (\nu_0^*, \nu_1^*, \nu_2^*)$, where $\nu_j^* = \nu_{0j} + N_j$, and N_j denotes the number of occurrences of truth j in the training data, for $j \in \{0,1,2\}$. The $\nu_j^* = \nu_{0j} + N_j$ formula is often referred to as the 'prior counts plus data counts' updating relationship for the multinomial-Dirichlet family.
2. The posterior of $\boldsymbol{\pi}_t^{(l)}$ is a Dirichlet distribution with parameters $\left(\alpha_{t0}^{(l*)}, \alpha_{t1}^{(l*)}, \alpha_{t2}^{(l*)}, \alpha_{t3}^{(l*)}\right)$, 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 analyst $l \in \{1, \dots, N\}$ for each truth $t \in \{0,1,2\}$. We let $\boldsymbol{A}^*$ denote the collection of all these posterior parameters.

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

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

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

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

Here, we have introduced the indicator function $I(\cdot)$, defined by $I(t=j)=1$ if $t=j$ and $I(t=j)=0$ otherwise, as it will be convenient later. To reduce clutter, we drop the dependence on $(b_1,\ldots,b_N,\boldsymbol{A}^*,\boldsymbol{\nu}^*)$ 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 as several further simplifications arise due to the structure of the IBCC model. For example, Equation 6 shows that the terms involving $\boldsymbol{\kappa}$ and $\boldsymbol{\Pi}$ can be separated, which implies the additional factorization $q^*(\boldsymbol{\kappa},\boldsymbol{\Pi})=q^*(\boldsymbol{\kappa})q^*(\boldsymbol{\Pi})$.

We start by initializing the distributions for $\boldsymbol{\kappa}$ and $\boldsymbol{\pi}_t^{(l)}$ with their posterior distributions, i.e., the Dirichlet distributions with parameters $\boldsymbol{\nu}^*$ and $\boldsymbol{A}^*$ given above. 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 $\log q^*(t)=E_{\boldsymbol{\kappa}}\log\kappa_t+\sum_{l=1}^N E_{\boldsymbol{\pi}_t^{(l)}}\log\pi_{tb_l}^{(l)}+\text{constant}$. Standard properties of the Dirichlet distribution (e.g., Bishop, 2006) give $E_{\boldsymbol{\kappa}}\log\kappa_t=\Psi(\nu_t^*)-\Psi(\sum_{j=0}^2\nu_j^*)$ and $E_{\boldsymbol{\pi}_t^{(l)}}\log\pi_{tb_l}^{(l)}=\Psi\left(\alpha_{tb_l}^{(l*)}\right)-\Psi\left(\sum_{s=0}^3\alpha_{ts}^{(l*)}\right)$ where $\Psi(\cdot)$ denotes the DiGamma function[16]. Defining $\log\rho_t=\Psi(\nu_t^*)-\Psi\left(\sum_{j=0}^2\nu_j^*\right)+\sum_{l=1}^N[\Psi(\alpha_{tb_l}^{(l*)})-\Psi\left(\sum_{s=0}^3\alpha_{ts}^{(l*)}\right)]$, 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 $\log q^*(\boldsymbol{\kappa})=\sum_{j=0}^2 q(t=j)\log\kappa_j+\sum_{j=0}^2(\nu_j^*-1)\log\kappa_j+\text{constant}$. Gathering together the $\log\kappa_j$ terms in this expression shows $q^*(\boldsymbol{\kappa})$ to be Dirichlet distributed with parameters $\nu_j=\nu_j^*+q(t=j)$ for $j\in\{0,1,2\}$. This formula for iterating the $\boldsymbol{\kappa}$ distribution is similar in structure to the 'prior counts plus data counts' relation noted above, except that now the prior over each forecasting period is the posterior obtained at the relevant census date, and the counts for each truth class are replaced with their expected values $E_tI(t=j)=q(t=j)$, which are fractional, rather than integer, valued.

We essentially repeat this argument to obtain the update equations for $q^*(\boldsymbol{\Pi})$. First, since the $\boldsymbol{\pi}_j^{(l)}$ terms in Equation 6 are separate for each truth $j\in\{0,1,2\}$ and each broker $l\in\{1,\ldots,N\}$, we obtain the further factorization $q^*(\boldsymbol{\Pi})=\prod_{l=1}^N\prod_{j=0}^2 q^*(\boldsymbol{\pi}_j^{(l)})$. Extracting the $\boldsymbol{\pi}_j^{(l)}$ terms and taking expectation with respect to $q^*(t)$, we obtain $\log q^*(\boldsymbol{\pi}_j^{(l)})=\sum_{j=0}^2 q^*(t=j)\sum_{l=1}^N\log\pi_{tb_l}^{(l)}+\sum_{s=0}^3(\alpha_{js}^{(l*)}-1)\log\pi_{js}^{(l)}+\text{constant}$. Gathering together terms in $\log\pi_{jb}^{(l)}$ shows $q^*\left(\boldsymbol{\pi}_j^{(l)}\right)$ is 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

[16] If the d-dimensional variable $\boldsymbol{X}=(X_1\ldots,X_d)$ is Dirichlet distributed with parameter $(\mu_1,\ldots,\mu_d)$, then $E(\log X_i)=\Psi(\mu_i)-\Psi(\sum_{j=1}^d\mu_j)$ 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.

VB approximation to $\Pr(T = t|B_1 = b_1, B_2 = b_2, \dots, B_N = b_N)$. 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[17], 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.

4.2 From predictive probabilities to decisions

The outputs of the above 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 attention to the discrete set of actions Go_Short, No_Trade and Go_Long.[18] 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), respectively. 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 behaviour is unsurprising, as the underlying training dataset contains unadjusted biases, since analysts typically issue more Buy recommendations than Hold or Sell, and there are more Price_Up labels than Price_Down or Price_Flat[19].

Recalling that q_t is an estimate of the conditional probability $\Pr(T = t|B_1 = b_1, B_2 = b_2, \dots, B_N = b_N)$, 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 $B_1 = b_1, B_2 = b_2, \dots, B_N = b_N$ 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 below:

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 $q_2 > k \max(q_0, q_1)$

[17] See Dempster *et al.* (1977) and Tanner (1996).

[18] 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.

[19] 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 which reflects the probabilistic structure of the observed dataset, including its biases, and dealing with these biases using an extension of the HPP decision rule.

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, whilst keeping $k = 1$ fixed, means the value of the information imparted by the broker recommendations needs to be higher in order for a Go_Long (Go-Short) decision to arise. In contrast, the condition involving parameter k relates to the relationship between the three conditional truth probabilities, q_0, q_1 and q_2, but does not involve the unconditional probabilities. Thus increasing k, whilst 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.

5. 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, for example providing protection against over-fitting. Finer aggregations than this are possible, for example learning could be undertaken at the GICS sector or sub-sector 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 below, recommendations are aggregated over a lookback of 30 calendar days, a process which 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 above. 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 tables and figures which 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; 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 $position_{it} = signal_{it} / \sum_t |signal_{it}|$ 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 above. We ignore recommendations where 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 3-year rolling window, with both long and short positions.
- **IBCC_Exp_LS**: As above, 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 3-year rolling window.
- **Both_Exp_LS**: As above, but using the IBCC model on an expanding window.

Where "L" or "S" is used in place of "LS" then only long (respectively short) positions are allowed. In all cases, the gross exposure is normalised to $100[20]. 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 tables 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 below is the EURO STOXX[21], the same index we used in defining the truths for the training data. This is based on a liquid subset of around 300 Euro-zone 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 onwards. The long IBCC strategies remain broadly in line with the Brok_Flw_L benchmark, with best long performance arising for the strategies labelled "Both". For the long-short portfolios, there is no corresponding LS-index, but all IBCC portfolios outperform the Brok_Flw_LS benchmark. Again, the portfolios labelled "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

[20] Gross exposure is defined as $\sum | pos_i|$ where pos_i is the position in the i-th market, in USD.
[21] DJEURST in Thomson Reuters notation.

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 labelled "Both" again perform better, repeating the outperformance pattern seen above 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.

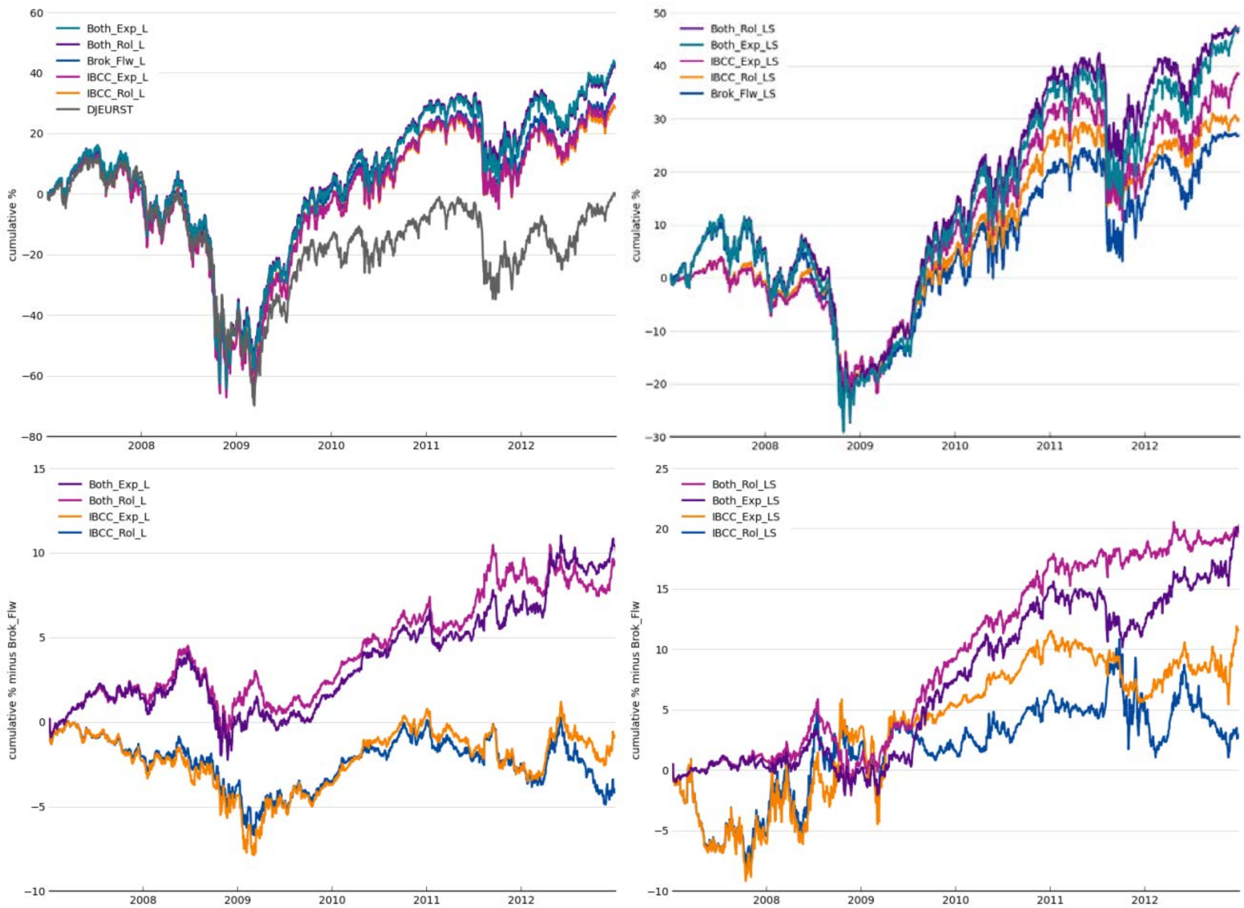

Exhibit 3: Performance of long-only models (left) and long-short models (right). Outright performance is shown in the top panels, whereas the bottom panels show performance relative to the relevant Brok_Flw_* benchmark. Note the vertical axes do not share a common scale.

side	model	mean	vol	alpha	alpha tstat	beta	beta tstat	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

Exhibit 4: Performance statistics for the various long-only, short-only and long-short models for the period 2007-2012. 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.

		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

Exhibit 5: Calendar year performance for long-only, short-only and long-short portfolios from 2007 to 2012 inclusive, expressed as %. The figures quoted are the sum of each year's daily returns. For reference, EURO STOXX returns are shown in the right-hand column.

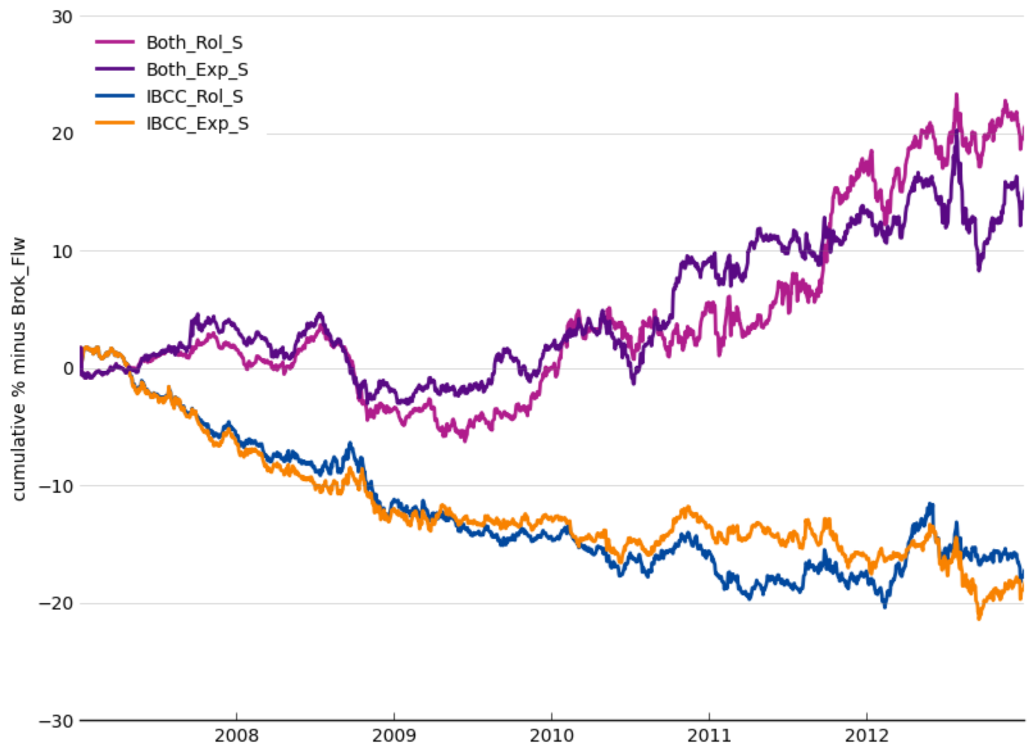

Exhibit 6: Performance of the short-only models relative to the Brok_Flw_S benchmark. The portfolios labelled "Both" provide the best performance, as was the case for the long-only and long-short portfolios.

5.1 Robustness checking - impact of firm liquidity

A potentially serious concern is that our IBCC procedure might be favouring 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 $25bn than 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.

Backtest performance is shown in Exhibit 9 for the long-only and short-only cases[22], and the distributions of market capitalisation for the two sub-portfolios 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 sub-portfolios, and that the efficacy of the algorithm is not driven by a size bias.

	Mega Cap >$25bn	Large Cap $10bn-$25bn	Mid Cap $2bn-$10bn	Small Cap $250M-$2bn	Micro Cap <$250M	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

Exhibit 7: Size tilts for the different portfolios, showing the sum of absolute positions by market capitalization bucket, averaged across time.

universe	number of recommendations	as a percentage
large half	58,466	56%
small half	45,316	44%

Exhibit 8: Number of recommendations after bisecting the universe of stocks by market capitalization.

[22] 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.

side	size	sim_name	return_mean	vol	alpha	alpha_tstat	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

Exhibit 9: IBCC results with the stock universe split into two by market capitalization. The alpha values are annualized.

universe	mktcap bucket	sum position (%)	nstocks	return (%p.a.)	risk (%p.a.)
large half	Mega Cap (USD >25bn)	41.5	75.2	4.2	6.7
	Large Cap (USD 10bn-25bn)	33.4	70.6	3.1	6.1
	Mid Cap (USD 2bn-10bn)	22.2	53.4	-0.3	6.9
	Small Cap (USD 250M-2bn)	2.8	7.8	-1.4	1.9
	Micro Cap (USD <250M)	0.0	0.1	-0.0	0.1
small half	Mega Cap (USD >25bn)	0.0	0.1	0.0	0.0
	Large Cap (USD 10bn-25bn)	5.1	9.9	1.7	0.9
	Mid Cap (USD 2bn-10bn)	80.2	149.8	6.8	15.2
	Small Cap (USD 250M-2bn)	14.5	30.7	-3.8	8.5
	Micro Cap (USD <250M)	0.1	0.2	-0.2	0.3

Exhibit 10: Distribution of market capitalization after bisecting the universe. Here the positions for Brok_Flw_L are summarized. The numbers shown in this table are time series averages 2007–2012.

5.2 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 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 in order to generate a trade. In contrast, $k > 1$ raises the threshold required for highest predictive probability (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, and for space considerations 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, and in particular for the Both_Exp_L results. The results for varying k whilst 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.

model	c	mean	vol	alpha	alpha tstat	beta	beta tstat	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

Exhibit 11: Varying the conviction level needed to initiate a trade for the long-only models for the period 2007-2012. Results are based on varying c whilst holding $k = 1$ in the decision rule. Recommendations were aggregated within the usual 30-day window when combining brokers. The alpha values are annualized.

model	k	mean	vol	alpha	alpha tstat	beta	beta tstat	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

Exhibit 12: Varying the conviction level needed to initiate a trade for the long-only models for the period 2007-2012. Results are based on varying k whilst holding $c = 1$ in the decision rule. Recommendations were aggregated within the usual 30-day window when combining brokers. The alpha values are annualized.

5.3 Robustness checking - sensitivity to truth threshold

A threshold of 5% was used in the truth definition given by

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

This value has been used throughout. Here we explore varying this parameter between 1% and 10%, keeping everything else the same. Results are summarised in Exhibit 13 below 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.

model	truth (%)	mean	vol	alpha	alpha tstat	beta	beta tstat	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

Exhibit 13: Varying the truth boundary parameter over the period 2007 - 2012, for long-only and short-only portfolios. Unreported results show that "IBCC_*" consistently underperforms "Brok_Flw" and "Both_*", consistent with our previous findings. The alpha values are annualized.

5.4 Robustness checking - sensitivity to holding period

Here we explore the sensitivity to the arbitrary 60-day holding period which has been used throughout. For brevity, we quote results for just the long-only portfolios.

From below 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.

model	holding period	mean	vol	alpha	alpha tstat	beta	beta tstat	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

Exhibit 14: Varying the holding period for the long-only models for period 2007 - 2012. Here the trade holding period and the holding period for assessing truths are constrained to be equal. The $\pm 5\%$ threshold for converting stock returns to "truths" is scaled to get 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.

6. 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.

6.1 Broker level diagnostics

The first is an animated visualization that displays the evolution of a broker's recommendation distributions[23] 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 4 different brokers, (IBES_199, IBES_207, IBES_410 and IBES_1296) is available from the [Journal website.](https://faculty.fuqua.duke.edu/~charvey/JFDS_2018/IBCC_Animation.mpeg)[24]

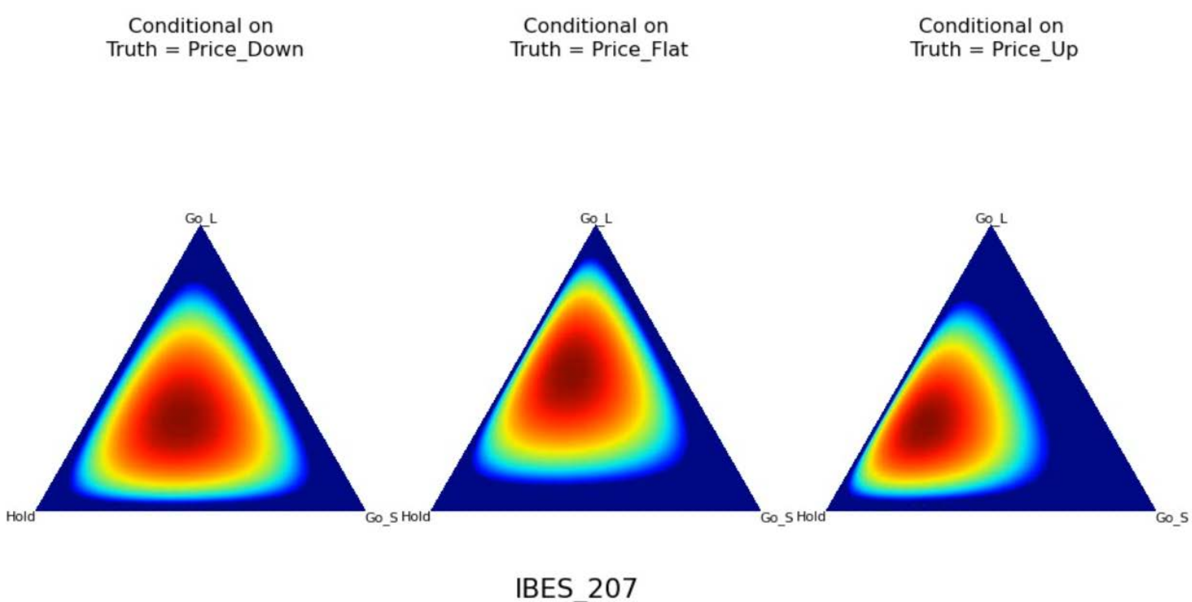

Exhibit 15: Screenshot of the website animation showing the evolving recommendation distributions for the broker with identifier IBES_207, conditional on truth=Price_Down (left), truth=Price_Flat (middle) and truth=Price_Up (right). Within each triangle, each pixel represents a 3-vector of probabilities over the recommendations Hold, Buy (Go_L) and Sell (Go_S). The color of each pixel represents the posterior probability of this corresponding 3-vector; blue pixels have very low posterior probability, and dark red pixels have the highest posterior probability.

The vertices of the triangles represent the three different recommendations Hold, Buy (here labelled Go_L) and Sell (here, Go_S). Each point within a tringle corresponds to a 3-vector of probabilities over

[23] Each conditional recommendation distribution is actually 4-dimensional, not 3-dimensional. In each case, we have marginalized over the label corresponding to "Missing" to obtain a 3-dimensional distribution. It is these that we have plotted as triangular heatmaps.

[24] The journal website is https://faculty.fuqua.duke.edu/~charvey/JFDS_2018/IBCC_Animation.mpeg [change later]

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 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.

6.2 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 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 5 most prolific brokers that comment on that stock; the green and red symbols represent buy and sell respectively, and the black symbol represents hold (labelled here as *filtered,* equivalently). The second panel lists the same information but is more cluttered as it now includes the recommendations of all brokers commenting on that stock. The third panel shows the actions which result from the predicted truths, obtained using our rolling out-of-sample process with the $c = k = 1$ case of the decision rule discussed above. No predictions are made during the initial 3-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 various Brok_Flw_* benchmarks. The final panel shows the truth (target) outcomes for each recommendation.

Exhibit 16: Diagnostic panel for the stock with identifier AST14822. The top-panel shows recommendations for the 5 most prolific brokers that comment on the stock; green and red represent buy and sell, respectively, the black symbol represents hold (labelled here as *filtered,* equivalently). The second panel lists the same information now includes the recommendations of all brokers commenting on that stock. The third panel shows the actions which 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 3-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.

7. Conclusions and future research directions

We have demonstrated a computationally efficient practical approach for combining analysts' forecasts using a probabilistic machine-learning model called independent Bayesian Classifier Combination, or IBCC, combining it with a state-of-the-art approximate inference technique called variational Bayes, or 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. In addition, combining 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.

While the variational Bayes implementation of the IBCC holds promise, there are also limitations. In the Galaxy Zoo experiment that we used to motivate the research application, there are several distinct issues that make the application to analysts different than to astronomers. First, it is reasonable to assume that the astronomers are operating independently, and not influencing each other even though their interpretations of the same image may be correlated (i.e., they behave independently conditional on a given image). However, it is possible that analysts are aware of other analysts' forecasts – and this could impact their forecasts. Second, the quality of all analysts' forecasts could be impacted 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 original IBCC model then there would be no issue. Our practical experience is that there is never sufficient data available compared to what we would like, so working with flexible but not completely general models gives the best results.

Acknowledgement

AL thanks Edwin Simpson for supporting discussions.

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley.

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2018). Variational inference: a review for statisticians. arXiv:1601.00670v9.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press.

Bradshaw, Mark T, Analysts' Forecasts: What Do We Know after Decades of Work? (June 30, 2011). Available at SSRN: https://ssrn.com/abstract=1880339

Brown, L. 1993. Earnings forecasting research: Its implications for capital markets research." *International Journal of Forecasting* 9: 295-320.

Brown, L., ed. 2000. *I/B/E/S Research Bibliography*: Sixth Edition. I/B/E/S International Inc., New York, NY.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (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 Roberts, S. J. (2011). A tutorial on variational Bayesian inference. *Artificial Intelligence Revue. DOI 10.1007/s10462-011-9236-8. ISSN 0269-2821.*

Ghahramani, Z. and Kim, H-C. (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 Ghahramani. Z (2012). Bayesian Classifier Combination. Proc. 15th AISTATS Conference.

Lee, P. M. (2012). *Bayesian Statistics: An Introduction*. John Wiley, Chichester.

Levenberg, A., Simpson, E., Roberts, S. and Gottlob, G. (2013). Economic prediction using heterogeneous data streams from the world wide web. *Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE), Proceedings of ECML/PKDD Workshop*. Springer.

Levenberg, A., Pulman, S., Moilanen, K., Simpson, E. and Roberts, S. (2014). Predicting economic indicators from web text using sentiment composition. *International Journal of Computer and Communication Engineering*. IACSIT Press.

Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley

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., Roberts, S., Psorakis, I. and Smith A. (2013). Dynamic Bayesian Combination of Multiple Imperfect Classifiers. *Decision Making and Imperfection. Intelligent Systems Reference Library series, Vol 474*. Springer.

Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* Springer, New York.