Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Bayesian herd detection for dynamic data

Jussi Keppo [a], Ville A. Satopää [b,*]

[a] NUS Business School and Institute of Operations Research and Analytics, National University of Singapore, Singapore
[b] Technology and Operations Management, INSEAD, France

ARTICLE INFO

Keywords:
Bayesian statistics
Dynamic modeling
Gaussian process
Judgmental forecasting
Imperfect information game

ABSTRACT

This article analyzes multiple agents who forecast an underlying dynamic state based on streams of (partially overlapping) information. Each agent minimizes a convex combination of their forecasting error and deviation from the other agents' forecasts. As a result, the agents exhibit herding behavior, a bias well-recognized in the economics and psychology literature. Our first contribution is a general framework for analyzing agents' forecasts under different levels of herding. The underlying state dynamics can be non-linear with seasonality, trends, shocks, or other time series components. Our second contribution describes how models within our framework can be estimated from data. We apply our estimation procedure to surveys of equity price forecasts and find that the agents concentrate 37% of their efforts on making similar forecasts on average. However, there is substantial variation in the level of herding over time; even though herding fell substantially during the 2007–2008 financial crisis, it rose after the crisis.

> *The herd instinct among forecasters makes sheep look like independent thinkers.*
>
> [– Edgar R. Fiedler, June 1977.]

1. Introduction

Decision makers optimize their actions based on available forecasts of quantities, such as future demand for a new product, a competitor's price, raw material costs, and asset returns. In many settings, there are multiple agents providing forecasts, and these forecasts may be influenced by each other. For instance, the agents may seek to avoid exceptionally deviant forecasts due to career or reputation concerns. Thus, agents may consider each other's forecasts, which leads to herding among the agents.

Morris and Shin (2002) develop a structural model of forecasts under herding. Their analysis is motivated by social planners, such as central banks or government agencies, who can influence the agents' behavior by disclosing public information. These planners face the challenge of deciding how much and how often information should be disclosed. The latter question, however, is difficult to study with their model because it is static and, hence, only applies in one-off settings in which each agent makes a single forecast. Furthermore, given that one forecast (or action in general) per agent does not provide enough information to estimate their model (Bergemann & Morris, 2013), it cannot be calibrated to the context at hand and hence offers limited utility to social planning. In many applications, however, agents update their forecasts over time as more information becomes available. An apparent solution is to develop a dynamic extension of the model in Morris and Shin (2002) and estimate it based on forecasts made over time.

Motivated by this, we introduce a framework for estimable dynamic forecasting models under herding. The underlying state is modeled with a Gaussian process, allowing the potential for seasonality, trends, shocks, noise, or almost any other time series component deemed appropriate for the application. Each agent has private and public information to forecast the underlying state while considering each other's forecasts. By second-guessing

[*] Corresponding author.
E-mail addresses: keppo@nus.edu.sg (J. Keppo), ville.satopaa@insead.edu (V.A. Satopää).

https://doi.org/10.1016/j.ijforecast.2023.03.001
0169-2070/© 2023 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

each other's forecasts, they minimize a convex combination of the expected forecasting error and the expected deviation from the other agents' forecasts. The relative importance of the second objective determines the level of herding among the agents.

After discussing the theoretical properties of agents' equilibrium forecasts, we turn to model estimation. Specifically, we develop a Bayesian estimation procedure for a model under which the agents get to observe each other's past forecasts, and the underlying state follows a simple autoregressive process. These choices are parsimonious and often suitable for modeling forecasts of equity returns or inflation (e.g., Cutler, Poterba, and Summers 1991, Gavin and Kliesen 2006, Giacomini, Skreta, and Turen 2020, Moskowitz, Ooi, and Pedersen 2012). To illustrate, we first simulate data and show that the true parameter values can be estimated accurately based on the public signals and agents' forecasts.

Having confirmed this, we apply our model in a case study of analysts' one-year-ahead forecasts of the equity price of EOG Resources, including between 2001 and 2015. Consistent, for example, with Hong, Kubik, and Solomon (2000) and Jegadeesh and Kim (2010), who use reduced-form models to provide evidence of herding among security analysts, we find a significant amount of herding. In particular, our analysis suggests that the analysts concentrate about 37% of their efforts on making similar forecasts, resulting in around a 5% loss in forecasting accuracy. However, there is substantial variation in the level of herding over time. The herding level was high before and after the 2007–2008 financial crisis but fell rapidly during the crisis. For robustness, we repeat our analysis for 150 companies in different industries and find qualitatively similar results.

1.1. Related literature

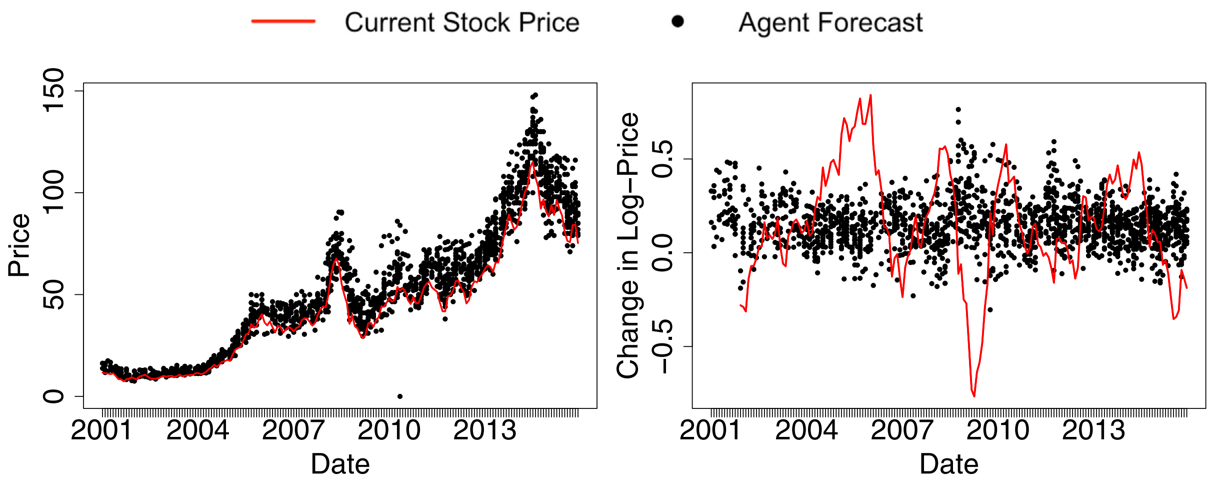
Agents may herd for many reasons. For instance, they may not have confidence in their private information (Bikhchandani, Hirshleifer, & Welch, 1992); they find the task unusually difficult (Kim & Pantzalis, 2003); or their private signal is pessimistic (Olsen, 1996). Another source of herding is the agents' reputation or career concerns that arise because their superiors can learn about an exogenous characteristic, such as their forecasting ability, through their reported forecasts (e.g., Hong et al. 2000). Similarly, Graham (1999) and Lamont (2002) find empirically that strong public information and career concerns increase herding. Herding may also be culturally motivated (Ashiya & Doi, 2001) or stem from social psychological or even biological foundations as discussed, for example, in Baddeley (2010), Raafat, Chater, and Frith (2009), and Stallen, De Dreu, Shalvi, Smidts, and Sanfey (2012).

Herding has been subject to many modeling efforts. Ottaviani and Sørensen (2006) derive a static model for the strategic behavior of agents by considering a forecasting competition between the agents and their reputational cheap talk aimed at convincing the market that they are well informed. Cheap talk induces the agents to give similar forecasts (for related models, see, e.g., Prendergast and Stole 1996, Scharfstein and Stein 1990, Trueman 1994, Zwiebel 1995). Banerjee, Guo, and Wang (2005) and Bikhchandani et al. (1992) model informational cascades, where it is optimal for individuals to ignore their private information and simply follow the behavior of the preceding individuals that they observe. Thus, in these models, myopic individuals do not account for the value of signaling information to successors, and a herd arises, i.e., everyone eventually settles on the same choice. Smith and Sørensen (2000) show that, in this setting, a herd appears almost definitely, and forecasts reflect no private information. Smith, Sørensen, and Tian (2021) study the inefficiency of the resulting equilibrium and show that socially optimal behavior demands that everyone errs towards choosing more informative forecasts in every period. Çelen, Kariv, and Schotter (2010), however, find that subjects in a laboratory social-learning situation appear more willing to follow the advice given to them by their predecessor than to copy their forecast, and that the presence of advice increases subjects' welfare. In the present paper, we do not model herding that arises from cascading information. Instead, each agent participates in a game and, by second-guessing other agents' forecasts, tries to make forecasts similar to the upcoming consensus.

Whether the agent's herd, anti-herd (i.e., report overly deviant forecasts), or act non-strategically is likely to depend largely on the context. One of the most studied applications is financial analysts' forecasts of future earnings. For this context, past literature offers mixed results. Clements (2018) summarizes recent literature supporting either herding or anti-herding. Bernhardt, Campello, and Kutsoati (2006) develop a test of herding and provide evidence that financial analysts anti-herd. In contrast, (Trueman, 1994) show that low-skill analysts may seek to mimic high-skill analysts because this may allow them to boost their compensation. Kim, Kim, and Shim (2019) develop a new test of herding and find strong evidence of herding in the analyst forecasts in the Institutional Brokers Estimate System (I/B/E/S) database that we use in this paper. As discussed before, Hong et al. (2000) provide empirical evidence linking career concerns and herding behavior among security analysts.

Previous empirical work on herding has primarily focused on analyzing forecast clustering, which is then interpreted as herding (Clement & Tse, 2005; Grinblatt, Titman, & Wermers, 1995; Lakonishok, Shleifer, & Vishny, 1992; Wermers, 1999). The authors, however, emphasize that clustering is not sufficient for herding. For instance, non-herding agents may cluster because they use common information. To distinguish spurious herding from true herding behavior, Cipriani and Guarino (2014) estimate a structural model of herd behavior. Their model is specific to financial transaction data, and they define herding as placing too much emphasis on past market trends instead of private information. Other empirical models have also defined herding as an over-tendency towards the (current or past) consensus forecast that is assumed to be known to the agents (e.g., Giacomini et al. 2020).

However, the agents do not know the current consensus in many contexts. For instance, they may make



(a) Current Prices and 1-Year-Ahead Forecasts

(b) Annual Change in Log-Price

Fig. 1. Forecasts of the future stock price of EOG Resources, Inc. This particular company was chosen because, in our I/B/E/S dataset, it is the company for which we have the largest number of forecasts.

forecasts simultaneously or may not have access to others' forecasts. Therefore, if they do not want to deviate too far from others, they must second guess where the consensus will be during their next forecast. Recognizing this, Jia, Keppo, and Satopää (2022) extend the model in Morris and Shin (2002) from point to probabilistic forecasts (i.e., each agent's forecast takes the form of a probability distribution) and show that their model can be estimated based on one-off probabilistic forecasts. In the present paper, the agents continue to second-guess the consensus. Still, instead of considering more complex formats of forecasting, we find more statistical power from multiple-point predictions made over time.

1.2. Application: Forecasts of stock returns

The main application of this paper considers agents who forecast a stock's future returns. By the momentum studies in Cutler et al. (1991), Fama and French (1988), Lo and MacKinlay (1988), and Moskowitz et al. (2012), the current and past prices act as public signals here. In addition, the agents can acquire private signals by running in-house analytics. Based on this information, the agents estimate expected future returns and, due to, for example, their career or reputation concerns, might herd (e.g., Hong et al. 2000).

In this paper, we illustrate our model on analyst forecasting data from the Institutional Brokers Estimate System (I/B/E/S).[1] This is one of the leading analyst forecast databases in finance and accounting research. Our dataset involves thousands of analysts who made one-year-ahead forecasts of different stock prices between the beginning of 2001 and the end of 2015. Throughout the paper, we refer to the forecasts about the stock price of EOG Resources, Inc (shortly just EOG) – an American energy company specializing in hydrocarbon exploration. This particular stock was chosen as our running example because it has the largest number of forecasts among all the stocks in our dataset. Fig. 1(a) illustrates the forecasts with black circles and the current price with the red line. In this example, 108 unique analysts made, on average, 15.7 forecasts per month between December 2001 and December 2015. Fig. 1(b) transforms the forecasts and the current prices in terms of annual changes in log prices. These transformed values are more amenable to modeling because they can take on values in the entire real line and appear stationary over time. Even though our discussion focuses largely on EOG, in Supplementary Material S4, we check robustness, repeat our analysis for 150 companies in different industries, and find results that are qualitatively similar to those in our analysis of EOG. For other potential applications of our modeling framework, see Section S1 of the Supplementary Materials.

1.3. Organization of the paper

The rest of the paper is organized as follows. Section 2 briefly reviews the static model in Morris and Shin (2002). Section 3 extends this model to a dynamic setting. Section 4 describes and illustrates how the model parameters can be estimated from data. Section 5 first evaluates the model on synthetic data and then applies it to analysts' stock price forecasts in the I/B/E/S database. Section 6 concludes the paper. Supplementary Materials discuss alternative applications of our model, consider agents who do not observe each other's past forecasts, test the robustness of our estimation procedure when the agents have heterogeneous levels of private information, present a computationally efficient way to calculate the agents' forecast weights, describe the prior distributions and the likelihood function used in model estimation, apply our model to analyst forecasts of the equity prices

[1] Full details and the dataset can be found at https://www.refinitiv.com/en/financial-data/company-data/institutional-brokers-estimate-system-ibes.

Table 1
Central notation and formulas in our dynamic model. Non-bold symbols are scalars. Lowercase and bold symbols are column vectors. Upper-case and bold symbols are matrices.

$K \in \mathbb{N}$	Number of agents making forecasts
$T \in \mathbb{N}$	Number of time points in the data
$\alpha(t), \beta(t) \in \mathbb{R}_{>0}$	Precision of the public and private signals at time t
$r(t) \in [0, 1]$	Level of herding among agents at time t
$\boldsymbol{\theta} \in \mathbb{R}^T$, $\boldsymbol{\mu}_\theta$, $\boldsymbol{\Sigma}_\theta$	Underlying state vector, its mean and covariance matrix
$\boldsymbol{y}(t), \boldsymbol{x}_k(t) \in \mathbb{R}^t$	Public signals and agent k's private signals at time t
$\boldsymbol{z}_k(t)$	All agent k's signals at time t
$\tilde{\boldsymbol{y}}(t), \tilde{\boldsymbol{x}}_k(t)$	Signals known by everyone and signals known only by agent k at time t
$\boldsymbol{\mu}_z(t)$, $\boldsymbol{\Sigma}_z(t)$, $\boldsymbol{\Sigma}_{zz}(t)$, $\boldsymbol{\Sigma}_{\theta z}(t)$	$\mathbb{E}[\boldsymbol{z}_k(t)]$, $\mathrm{Cov}(\boldsymbol{z}_k(t))$, $\mathrm{Cov}(\boldsymbol{z}_k(t), \boldsymbol{z}_j(t))$ with $k \neq j$, and $\mathrm{Cov}(\boldsymbol{\theta}, \boldsymbol{z}_k(t))$
$\boldsymbol{a}_k^*(t), \boldsymbol{a}_k(t) \in \mathbb{R}^T$	Agent k's optimal and herding forecasts at time t
$q(t)$	Variance inflation factor in the herding forecast
$\boldsymbol{W}_z^*(t)$, $\boldsymbol{W}_x^*(t)$, $\boldsymbol{W}_y^*(t)$	Optimal weights of $\boldsymbol{z}_k(t)$, $\boldsymbol{x}_k(t)$, and $\boldsymbol{y}(t)$
$\boldsymbol{W}_z(t)$, $\boldsymbol{W}_x(t)$, $\boldsymbol{W}_y(t)$	Herding weights of $\boldsymbol{z}_k(t)$, $\boldsymbol{x}_k(t)$, and $\boldsymbol{y}(t)$
$E_k(t)$, $\boldsymbol{E}_k(t)$	Inner and outer products of agent k's errors, $[\boldsymbol{a}_k(t) - \boldsymbol{\theta}]'[\boldsymbol{a}_k(t) - \boldsymbol{\theta}]$ and $[\boldsymbol{a}_k(t) - \boldsymbol{\theta}][\boldsymbol{a}_k(t) - \boldsymbol{\theta}]'$, at time t
$\boldsymbol{V}(t)^*$, $\boldsymbol{V}(t)$	$\mathbb{E}[\boldsymbol{E}_k(t)]$ under no herding and herding
$D_k(t)$, $\boldsymbol{D}_k(t)$	Average inner and outer products of agent k's distance to all forecasts, $\frac{1}{K}\sum_{j=1}^K [\boldsymbol{a}_k(t) - \boldsymbol{a}_j(t)]'[\boldsymbol{a}_k(t) - \boldsymbol{a}_j(t)]$ and $\frac{1}{K}\sum_{j=1}^K [\boldsymbol{a}_k(t) - \boldsymbol{a}_j(t)][\boldsymbol{a}_k(t) - \boldsymbol{a}_j(t)]'$, at time t
$\boldsymbol{M}[i, j]$, $\boldsymbol{M}[j, \cdot]$, $\boldsymbol{M}[\cdot, j]$	Top left $i \times j$ submatrix, j top rows, and j left-most columns of matrix $\boldsymbol{M}$
$\boldsymbol{m}[i]$	First i elements of vector $\boldsymbol{m}$

of 150 companies, and provide proof of our mathematical statements. Finally, for the reader's convenience, Table 1 summarizes the central notation of the paper.

2. Static model of herding agents

Morris and Shin (2002) consider a continuum of agents indexed by the unit interval $[0, 1]$. The agents estimate an unknown variable $\theta \in \mathbb{R}$ based on private and public information. Specifically, agent k observes a public signal $y = \theta + \alpha^{-1/2}\epsilon$ and a private signal $x_k = \theta + \beta^{-1/2}\epsilon_k$, where ϵ and ϵ_k are independent standard Gaussian random variables and the constants $\alpha > 0$ and $\beta > 0$ denote the precisions of the two signals, respectively. Agent k's Bayesian *optimal forecast* then is the precision weighted average of the private and public signals:

$$a_k^* = \mathbb{E}[\theta | y, x_k] = \frac{\alpha y + \beta x_k}{\alpha + \beta}, \qquad (1)$$

where the superscript * represents optimality in terms of forecasting accuracy. Specifically, a_k^* makes optimal use of agent k's information and minimizes the expected (squared) forecasting error among all functions of agent k's signals.

Morris and Shin (2002) do not model herding arising from cascading information (see our discussion in Section 1.1). Instead, each agent participates in a game and, by second-guessing the other agents' forecasts, tries to make a forecast similar to the upcoming consensus. This is modeled with a dual objective between forecasting error and deviation from the other agents' forecasts. The relative weight between these two objectives parametrizes the herding behavior. More specifically, agent k's *herding forecast* a_k is

$$\operatorname{argmin}_{a_k \in \mathbb{R}} \mathbb{E}\Big[(1-r)\underbrace{(a_k - \theta)^2}_{\text{Error}} + r\underbrace{\left(D_k - \bar{D}\right)}_{\text{Deviation}} \Big| y, x_k\Big], \qquad (2)$$

where $r \in [0, 1]$ controls the level of herding, $D_k = \int_0^1 \left(a_k - a_j\right)^2 dj$ is the average (squared) distance from agent k's herding forecast to other agents' herding forecasts, and $\bar{D} = \int_0^1 D_k dk$ is the average of these distances. The error component in (2) captures the agents' utility from forecasting the state of the economy accurately. The deviation component, on the other hand, represents the game of guessing well what the other agents will forecast. The parameter r controls the relative priority of each objective.

Morris and Shin (2002) solve (2) and show that agent k's herding forecast is

$$a_k = \frac{\alpha y + \beta(1-r)x_k}{\alpha + \beta(1-r)}.$$

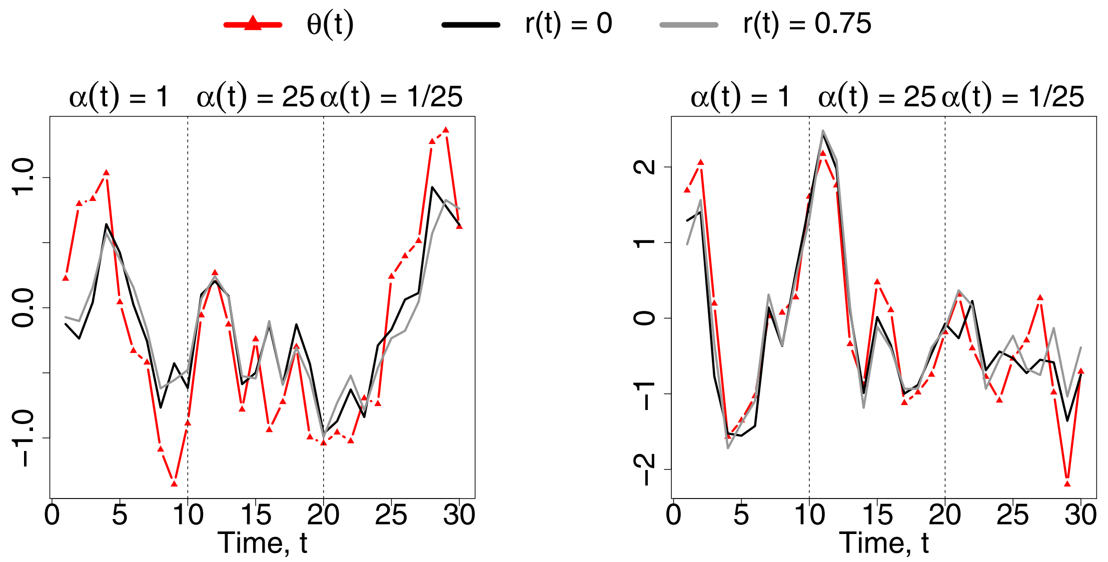
This forecast differs from agent k's optimal forecast a_k^* as long as $r > 0$. In particular, if $r > 0$, the agents herd and, compared to their optimal forecasts, place less weight on their private signals and more on the public signal. This discrepancy rises in r. In the extreme case where $r = 1$, agents use only the public signal, and their forecasts are identical, $a_k = y$ for all $k \in [0, 1]$.

3. Dynamic model of herding agents

In this section, we introduce a new framework that extends the model in Morris and Shin (2002) to a dynamic setting, where agents' information accumulates and the underlying variable of interest changes over time.

3.1. Underlying state process

In our dynamic model, the underlying state is dynamic and assumed to follow a Gaussian process $\theta(t) \sim \mathcal{GP}(\mu, \phi)$ with some mean and covariance functions $\mu(\cdot)$ and $\phi(\cdot, \cdot)$, respectively. Specifically, let $\boldsymbol{\theta} = (\theta(1), \ldots, \theta(T))'$ be a vector of the underlying state variables at



(a) Example 1: AR(1) with $\mu_0 = 0$, $\theta_0 = 0$, $\rho = 0.75$, and $\sigma = 0.5$.

(b) Example 2: Seasonality with a non-linear trend, $\mu(t) = 0$, $\ell_P = 1$, $\sigma_P = 1$, $p = 10$, $\ell_S = 1$, and $\sigma_S = 1$.

Fig. 2. The optimal forecast, the herding forecast, and the underlying process $\theta(t)$ follows Examples 1 and 2 in Section 3.1. The agents observe each other's past forecasts. Other parameters: $\beta(t) = 1$ for all t and $K = 5$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

times $t = 1, \ldots, T$.[2] Then $\boldsymbol{\theta}$ follows a multivariate Gaussian distribution:

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta), \tag{3}$$

where the tth element of $\boldsymbol{\mu}_\theta$ is $\mu(t)$ and the (t, t')th element of $\boldsymbol{\Sigma}_\theta$ is $\phi(t, t')$. Even though the mean and covariance functions could depend on other known covariates besides time, we assume, for the sake of simplicity, that the expected behavior of $\theta(t)$ is entirely characterized by time.

Overall, formulation (3) is very general and captures many well-known processes. This is illustrated with the following two examples.

Example 1. Changes in inflation (e.g., Gavin and Kliesen 2006, Giacomini et al. 2020) or log stock prices (e.g. Cutler et al. 1991, Fama and French 1988, Lo and MacKinlay 1988, Moskowitz et al. 2012) are often modeled with an AR(1) process:

$$\theta(t+1) = \mu_0 + \rho\theta(t) + \sigma\epsilon(t+1), \tag{4}$$

where $\theta(0) = \theta_0 \in \mathbb{R}$, $\mu_0 \in \mathbb{R}$, $\rho > 0$, and $\sigma > 0$ are finite constants, and $\epsilon(t)$ is an independent standard Gaussian random variable. This is equivalent to the following mean and covariance functions:

$$\mu_{AR}(t) = \mathbb{E}[\theta(t)] = \mu_0 \sum_{m=0}^{t-1} \rho^m + \rho^t \theta_0 \quad \text{and}$$

$$\phi_{AR}(t, t') = \mathrm{Cov}(\theta(t), \theta(t')) = \rho^{t-t'} \sigma^2 \sum_{m=0}^{t'-1} \rho^{2m},$$

where $t' \leq t$. The derivation of this result is in Supplementary Material S6. Fig. 2(a) shows (in red) one draw of $\boldsymbol{\theta}$ from this process with $\mu_0 = 0$, $\theta_0 = 0$, $\rho = 0.75$, and $\sigma^2 = 0.25$.

Example 2. In Gaussian process modeling, time-dependent properties of the underlying process can be captured by adding or multiplying different covariance functions. For instance, in many applications, the underlying state process has seasonality (e.g., electricity demand and price modeling in Keppo, Audet, Heiskanen, and Vehviläinen 2004, Taylor 2003). Seasonality can be introduced with the periodic kernel:

$$\phi_P(t_1, t_2) = \sigma_P^2 \exp\left(-\frac{2\sin^2(\pi|t_1 - t_2|/p)}{\ell_P^2}\right),$$

where $p > 0$ is the period, the *length-scale* $\ell_P > 0$ determines how "wiggly" the process is, and $\sigma_P^2 > 0$ describes how far, on average, the process is from its mean. A non-linear trend, on the other hand, can be introduced by the squared exponential kernel:

$$\phi_{SE}(t_1, t_2) = \sigma_S^2 \exp\left(-\frac{(t_1 - t_2)^2}{2\ell_S^2}\right),$$

where, similarly to the periodic kernel, the *length-scale* $\ell_S > 0$ determines how "wiggly" the process is, and $\sigma_S^2 > 0$ describes how far, on average, the process is from its mean. The above two kernels can be added, $\phi_P(t_1, t_2) + \phi_{SE}(t_1, t_2)$, to describe a seasonal process with a non-linear

[2] For simplicity, we assume that the time points are equally spaced between 1 and T. An extension to unequally spaced time points, however, is straightforward.

trend. Fig. 2(b) shows (in red) one draw of $\boldsymbol{\theta}$ from this process with $\mu(t) = 0$ for all t, $p = 10$, $\sigma_P^2 = 1$, $\ell_P = 1$, $\sigma_S = 1$, and $\ell_S = 1$.

By changing the mean and covariance functions, $\mu(\cdot)$ and $\phi(\cdot, \cdot)$, it is possible to incorporate seasonality, trends, different levels of smoothness, noise, shocks, and almost any other aspect that is deemed appropriate for the application at hand. For more information, see Williams and Rasmussen 2006 for an excellent introduction to Gaussian processes. This theory-focused section does not assume a particular form for the mean or covariance function. Therefore, all theoretical results hold under any underlying Gaussian process.

3.2. Agents' forecasts

Suppose there are K agents forecasting the value of the underlying state. At time t, the system generates a public signal $y(t)$ and a private signal $x_k(t)$ for all $k \in \{1, 2, \dots, K\}$. Similarly to the static model in Section 2, these signals are drawn from a Gaussian distribution centered at the underlying state process:

$$y(t) = \theta(t) + \alpha(t)^{-1/2}\epsilon_y(t) \quad \text{and}$$
$$x_k(t) = \theta(t) + \beta(t)^{-1/2}\epsilon_{x,k}(t),$$

where $\alpha(t) > 0$ and $\beta(t) > 0$ are the respective precisions of the public and private signals at time t, and $\epsilon_y(t)$ and $\epsilon_{x,k}(t)$ are independent standard Gaussian random variables.

The precisions of the signals, $\alpha(t)$ and $\beta(t)$, are assumed to be the same for all the agents. This type of *symmetry* assumption is common in the analysis of agent forecasting (Gaba, Popescu, and Chen 2018, Kim, Lim, and Shaw 2001, Lichtendahl Jr, Grushka-Cockayne, and Pfeifer 2013, Morris and Shin 2002, Ottaviani and Sørensen 2006, Palley and Soll 2019, Pfeifer, Grushka-Cockayne, and Lichtendahl Jr 2014, Satopää, Pemantle, and Ungar 2016, and many others) because it offers a tractable model of agents with both private and public information. In our case, the symmetry assumption allows us to find the equilibrium forecast in closed form and express the likelihood of data, making model estimation possible. This assumption also allows us to analyze data from large multiyear surveys where agents rarely make forecasts systematically at regular time intervals and may enter or exit in the middle of the survey. For instance, in the EOG example illustrated in Fig. 1, the time horizon is $T = 169$ and there are 110 unique agents, but the median number of forecasts made by an agent is only 9. Therefore, an agent does not predict most time points. In Supplementary Material S5, we perform robustness checks and demonstrate that our approach can estimate the level of herding $r(t)$ accurately even when the agents have heterogeneous levels of private information.

Signals generated by time T can be summarized with vectors $\boldsymbol{y} = (y(1), \dots, y(T))'$ and $\boldsymbol{x}_k = (x_k(1), \dots, x_k(T))'$ for each agent $k \in \{1, 2, \dots, K\}$. The signals and the underlying state vector $\boldsymbol{\theta}$ follow a multivariate Gaussian distribution:

$$\begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{y} \\ \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_K \end{pmatrix} \sim \mathcal{N}$$
$$\times \left(\begin{pmatrix} \boldsymbol{\mu}_\theta \\ \boldsymbol{\mu}_\theta \\ \boldsymbol{\mu}_\theta \\ \vdots \\ \boldsymbol{\mu}_\theta \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_\theta & \boldsymbol{\Sigma}_\theta & \boldsymbol{\Sigma}_\theta & \dots & \boldsymbol{\Sigma}_\theta \\ \boldsymbol{\Sigma}_\theta & \boldsymbol{\Sigma}_\theta + \boldsymbol{I}_\alpha^{-1} & \boldsymbol{\Sigma}_\theta & \dots & \boldsymbol{\Sigma}_\theta \\ \boldsymbol{\Sigma}_\theta & \boldsymbol{\Sigma}_\theta & \boldsymbol{\Sigma}_\theta + \boldsymbol{I}_\beta^{-1} & \dots & \boldsymbol{\Sigma}_\theta \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_\theta & \boldsymbol{\Sigma}_\theta & \boldsymbol{\Sigma}_\theta & \dots & \boldsymbol{\Sigma}_\theta + \boldsymbol{I}_\beta^{-1} \end{pmatrix} \right), \tag{5}$$

where $\boldsymbol{I}_\alpha = \text{diag}(\alpha(1), \dots, \alpha(T))$ and $\boldsymbol{I}_\beta = \text{diag}(\beta(1), \dots, \beta(T))$ are diagonal matrices of the signal precisions.

At time t the system generates a public signal $y(t)$ and a private signal $x_k(t)$ for each agent $k \in \{1, 2, \dots, K\}$. The agents may have access to each other's past private signals and current private signals. More generally, collect all (past and current) signals that agent k has access to at time t into a vector $\boldsymbol{z}_k(t)$. Due to our symmetry assumption, each signal is known either by one agent or by everyone, allowing us to write $\boldsymbol{z}_k(t) = (\tilde{\boldsymbol{y}}(t)', \tilde{\boldsymbol{x}}_k(t)')'$, where $\tilde{\boldsymbol{y}}(t)$ is a vector of signals known by everyone and $\tilde{\boldsymbol{x}}_k(t)$ is a vector of signals known only by agent k at time t. Then, based on $\boldsymbol{z}_k(t)$, agent k forecasts $\boldsymbol{a}_k(t) \in \mathbb{R}^T$ for the value of the underlying process $\boldsymbol{\theta}$. This forecast is a vector containing backcasts of past values $\theta(t')$ for $t' < t$, a nowcast of the current value $\theta(t)$, and forecasts of future values $\theta(t')$ for $t' \geq t$. This way, at each time point, as new signals arrive, the agents update their beliefs about the past and future values of the underlying state.

Similarly to Morris and Shin (2002), we model herding with a dual objective. Our model, however, is dynamic, so agent k chooses the final forecast by considering the current and future values of the dual objective. Given that the agents directly observe different subsets of system-generated public and private signals, their forecasts do not influence the stream of information. Furthermore, by distribution (3), the agents' forecasts do not affect the underlying state dynamics. Therefore, the current forecast of agent k does not influence their or the other agents' future forecasts, and they can focus on minimizing the flow objective. Specifically, at time t, agent k's herding forecast $\boldsymbol{a}_k(t)$ for the state vector $\boldsymbol{\theta}$ is given by

$$\text{argmin}_{\boldsymbol{a}_k(t) \in \mathbb{R}^T} \mathbb{E}\left\{ (1 - r(t)) \underbrace{E_k(t)}_{\text{Error}} + r(t) \underbrace{D_k(t)}_{\text{Deviation}} \middle| \boldsymbol{z}_k(t) \right\}, \tag{6}$$

where $r(t) \in [0, 1]$ controls the level of herding, the error $E_k(s) = [\boldsymbol{a}_k(s) - \boldsymbol{\theta}]' [\boldsymbol{a}_k(s) - \boldsymbol{\theta}]$ is the sum of squared errors in agent k's forecasts, the deviation $D_k(s) = \frac{1}{K} \sum_{j=1}^K \left[\boldsymbol{a}_k(s) - \boldsymbol{a}_j(s)\right]' \left[\boldsymbol{a}_k(s) - \boldsymbol{a}_j(s)\right]$ is the average of the squared distances between agent k's forecast and all other forecasts (including their own)[3] at time t.

[3] Our definition of deviation in (6) differs slightly from the one in Morris and Shin (2002). In the static model (2), deviation measures

Before we present the solution for the agents' herding forecasts (6), we consider the Bayesian optimal forecast $\boldsymbol{a}_k^*(t)$ and its error. These expressions follow directly from distribution (5) and the well-known results on conditional distributions of a multivariate Gaussian random variable (e.g., Ravishanker and Dey 2001).

Theorem 1 (*Optimal Forecast*)**.** *Conditional on $\boldsymbol{z}_k(t)$, agent k's optimal forecast and error covariance matrix are*

$$\begin{aligned}\boldsymbol{a}_k^*(t) &= \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{z}_k(t)] = \boldsymbol{\mu}_\theta + \boldsymbol{W}_z^*(t)[\boldsymbol{z}_k(t) - \boldsymbol{\mu}_z(t)] \quad \textit{and}\\ \boldsymbol{V}^*(t) &= \mathbb{E}\{[\boldsymbol{\theta} - \boldsymbol{a}_k^*(t)][\boldsymbol{\theta} - \boldsymbol{a}_k^*(t)]'|\boldsymbol{z}_k(t)\}\\ &= \boldsymbol{\Sigma}_\theta - \boldsymbol{W}_z^*(t)\boldsymbol{\Sigma}_{\theta z}(t)',\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\mu}_z(t) &= \mathbb{E}[\boldsymbol{z}_k(t)], & \boldsymbol{W}_z^*(t) &= \boldsymbol{\Sigma}_{\theta z}(t)\boldsymbol{\Sigma}_z^{-1}(t),\\ \boldsymbol{\Sigma}_{\theta z}(t) &= Cov(\boldsymbol{\theta}, \boldsymbol{z}_k(t)), \ \textit{and} & \boldsymbol{\Sigma}_z(t) &= Cov(\boldsymbol{z}_k(t)).\end{aligned}$$

To solve the agents' herding forecasts, we use the first order condition of (6) and express agent k's herding forecast $\boldsymbol{a}_k(t)$ as follows:

$$\boldsymbol{a}_k(t) = (1 - r(t))\mathbb{E}\left[\boldsymbol{\theta}|\boldsymbol{z}_k(t)\right] + r(t)\mathbb{E}\left[\frac{1}{K}\sum_{j=1}^{K}\boldsymbol{a}_j(t)\middle|\boldsymbol{z}_k(t)\right]. \tag{7}$$

The agents' herding forecast $\boldsymbol{a}_k(t)$ is then the weighted average of agent k's optimal forecast $\boldsymbol{a}_k^*(t)$ and agent k's belief about the average herding forecast among all agents (including themselves). This shows that if $r(t) = 0$, the agents' herding forecasts coincide with their optimal forecasts. If $r(t) > 0$, the agents try to second-guess each other's forecasts, participate in a dynamic game, and herd. In the limit $r(t) = 1$, the agents simply forecast their best estimate about the crowd's average forecast. The following theorem solves (7) and, similarly to the optimal forecast $\boldsymbol{a}_k^*(t)$, expresses agent k's herding forecast as a linear combination of $\boldsymbol{z}_k(t)$.

Theorem 2 (*Herding Forecast*)**.** *If $r(t) \in [0, 1)$, then agent k's herding forecast at time t is given by*

$$\boldsymbol{a}_k(t) = \boldsymbol{\mu}_\theta + \boldsymbol{W}_z(t)[\boldsymbol{z}_k(t) - \boldsymbol{\mu}_z(t)],$$

where

$$\begin{aligned}\boldsymbol{W}_z(t) &= \boldsymbol{\Sigma}_{\theta z}(t)\left[\boldsymbol{\Sigma}_z(t) + q(t)\,\mathrm{diag}\left(\boldsymbol{0}'_{|\tilde{\boldsymbol{y}}(t)|}, 1/\tilde{\boldsymbol{\beta}}(t)'\right)\right]^{-1},\\ q(t) &= r(t)(K-1)/[K(1-r(t))] \geq 0,\end{aligned}$$

and $1/\tilde{\boldsymbol{\beta}}(t)$ is the vector of variances of the private information $\tilde{\boldsymbol{x}}_k(t)$.

Several details of this theorem should be emphasized. First, the herding forecasts align with the optimal forecasts if the agents do not herd $r(t) = 0$ or there is only $K = 1$ agent because one agent cannot herd alone. Second, compare the form of the herding forecast $\boldsymbol{a}_k(t)$ with the optimal forecast $\boldsymbol{a}_k^*(t)$. In both cases, the forecasts are linear in the (centered) signals, and the weights, $\boldsymbol{W}_z^*(t)$ and $\boldsymbol{W}_z(t)$, are deterministic. The only difference is that the herding forecast $\boldsymbol{a}_k(t)$ adds $q(t)\left(\boldsymbol{0}'_{|\tilde{\boldsymbol{y}}(t)|}, 1/\tilde{\boldsymbol{\beta}}(t)'\right)$ to the diagonal of $\boldsymbol{\Sigma}_z(t)$. Given that $\boldsymbol{z}_k(t)$ has been ordered such that private information comes last, the herding weights $\boldsymbol{W}_z(t)$ are equal to the optimal weights $\boldsymbol{W}_z^*(t)$ under the false belief that the variances of the private information $\tilde{\boldsymbol{x}}_k(t)$ are $(q(t) + 1) \geq 1$ times larger than what they truly are. A herding agent then behaves as if the private information $\tilde{\boldsymbol{x}}_k(t)$ are less informative than they truly are. As a result, herding agents underweight their private information and, consequently, overweight their public information. Given that $q(t)$ increases in K, this discrepancy strengthens as the crowd size K grows larger. For a large crowd ($K \to \infty$), this inflation factor coincides with the inflation factor in the static model of Section 2.

Third, the level of herding must be strictly less than one in Theorem 2. In particular, given that $r(t) = 1$ implies $q(t) = \infty$, extreme herding agents treat private information as if it had infinite variance. Even though the matrix inversion in $\boldsymbol{W}_z(t)$ cannot be solved in this extreme case, it suggests that as $r(t) \to 1$, the agents exclude private information and, hence, forecast only based on public information.

The next corollary offers further intuition into the agents' trade-off between herding and private information.

Corollary 1 (*Private Information and Herding Trade-Off*)**.** *Let $\tilde{\boldsymbol{\alpha}}(t)$ and $\tilde{\boldsymbol{\beta}}(t)$ be vectors of precisions of the shared $\tilde{\boldsymbol{y}}_k(t)$ and private information $\tilde{\boldsymbol{x}}_k(t)$, respectively. Then, conditional on $\boldsymbol{\Sigma}_\theta$ and $\tilde{\boldsymbol{\alpha}}(t)$, the herding weights $\boldsymbol{W}_z(t)$ are invariant to changes in $r(t)$ and $\tilde{\boldsymbol{\beta}}(t)$ if the new $r'(t)$ and $\tilde{\boldsymbol{\beta}}'(t)$ satisfy*

$$\tilde{\boldsymbol{\beta}}'(t)\frac{K - r(t)}{1 - r(t)} = \frac{K - r'(t)}{1 - r'(t)}\tilde{\boldsymbol{\beta}}(t).$$

Therefore, at time t, there are infinitely many $(\beta(1), \ldots, \beta(t), r(t))$-tuples that give the same coefficients $\boldsymbol{W}_z(t)$. This is illustrated in Fig. 3, where we set $\beta(t) = \beta$ for all t and plot the (log) ratio $\frac{K-r(t)}{(1-r(t))\beta}$ under different values of β and $r(t)$. The coefficients $\boldsymbol{W}_z(t)$ remain the same over each contour. This shows that $r(t)$ and β can increase or decrease together without affecting $\boldsymbol{W}_z(t)$. Intuitively this happens because weakly herding agents, who believe their private information to have low precision, treat their signals similarly to strongly herding agents, who believe their private information to have high precision.

The following corollary explains the effects of this trade-off on accuracy and forecast heterogeneity.

Corollary 2 (*Forecast Behavior*)**.** *If $r(t) \in (0, 1)$, then the herding forecasts are, in expectation, more similar to each other and less accurate than the corresponding optimal forecasts.*

According to (7) and Corollary 2, herding agents sacrifice accuracy to make predictions that are more likely

the distance between agent k's deviation and the average deviation in the crowd. Therefore, a herding agent does not want to appear more deviant than the average agent. On the other hand, in our dynamic model, a herding agent wants to make a forecast that is close to the forecasts of others. Even though these interpretations are slightly different, they are both plausible models of herding and lead to identical forecasting behavior.

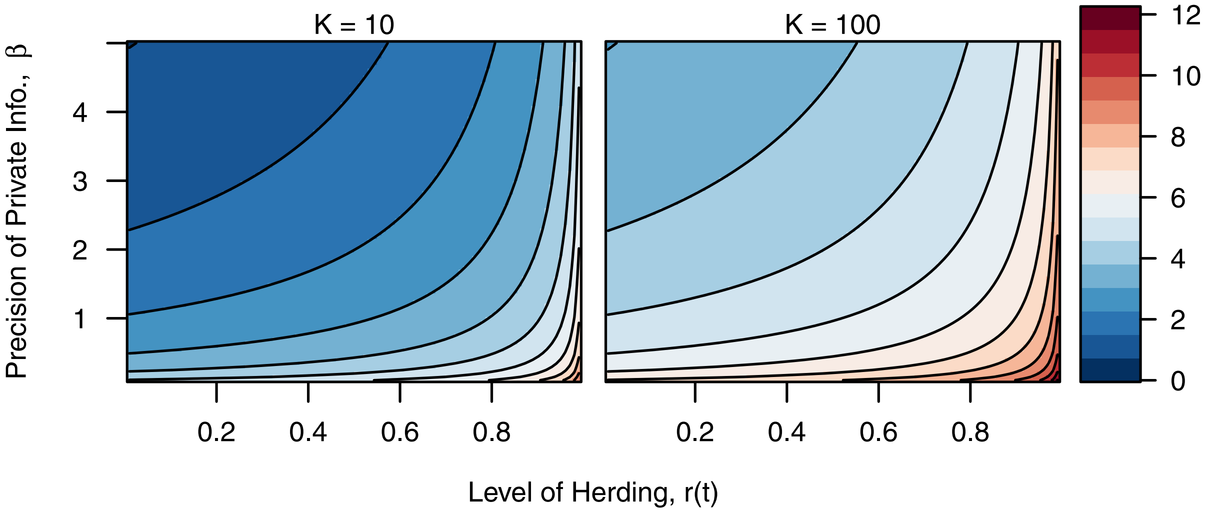


Fig. 3. The log of the quantity presented in Corollary 1. The precision of the private signal $\beta(t) = \beta$ for all t. The herding weights $\boldsymbol{W}_z(t)$ remain the same over each contour. This illustrates how the effects of increasing or decreasing both β and r can cancel each other out in the weights.

to resemble the crowd's average prediction. This seems to contradict recent research on intersubjective scoring, which shows that agents whose forecasts are, on average, closer to the crowd consensus tend to be more accurate (see, e.g., Karger, Monrad, Mellers, and Tetlock 2021, Kurvers et al. 2019, Witkowski, Atanasov, Ungar, and Krause 2017). To reconcile this seeming contradiction, we must observe that the models behind intersubjective scoring are typically formulated such that accurate forecasting of the underlying state and crowd consensus coincide. For instance, a typical assumption considers the agents' information sets independent conditional on the underlying state (Liu, Wang, & Chen, 2020). Then, the agents' only point of coordination is the underlying state, and forecasting the state accurately becomes their only means to make similar predictions. Of course, the agents may have some "side information", which can introduce a mismatch between forecasting the underlying state and the crowd consensus (Liu et al., 2020). This is what drives the result in Corollary 2: In our model, the agents can coordinate by placing more weight on such "side information", namely the public signals and the prior belief (i.e., $\boldsymbol{\mu}_\theta$), and less weight on private information.

To counter these negative effects of herding, the central planner can encourage the agents to emphasize their private signals more. For instance, the planner can reward each agent based on how much they can improve the crowd's average prediction (Budescu & Chen, 2015) or by eliciting their predictions through a winner-takes-all or, better yet, an incentive compatible forecasting competition (Lichtendahl Jr et al., 2013; Witkowski, Freeman, Vaughan, Pennock, & Krause, 2022). Based on our model, the planner can eliminate herding by subtracting the term $r(t)D_k(t)$ from the agents' objective function (6) (for a related result, see Smith et al. 2021). Given that (6) is a minimization problem, this corresponds to paying a bonus of $r(t)D_k(t)$ to agent k, which means the planner compensates for deviant forecasts. This bonus is dynamic and agent-specific, but it depends on the level of herding $r(t)$ and deviation $D_k(t)$ that must be estimated from forecasting data.

Unfortunately, the level of herding is notoriously difficult to estimate (Bergemann & Morris, 2013). In fact, at first, Corollary 1 seems to make model estimation challenging or even impossible. Observe, however, that Corollary 1 does not consider the likelihood of the underlying signals. In particular, even though changing β and $r(t)$ along the lines in Fig. 3 does not change $\boldsymbol{W}_z(t)$, changing β influences the behavior of the private signal $\tilde{\boldsymbol{x}}_k(t)$. For instance, decreasing β makes $\tilde{\boldsymbol{x}}_k(t)$ and, hence, $\boldsymbol{a}_k(t)$ more volatile. As shown in the following sections, this allows us to estimate the level of herding in our dynamic context.

4. Model estimation

4.1. Preliminaries

Before the model presented in Section 3 can be estimated from data, we must make some specific choices that seek to strike a balance between computational tractability and realism: First, model estimation is performed by a planner with access to the public signals and the agents' forecasts over the time interval $t \in \{1, \ldots, T\}$. Second, the underlying state $\theta(t)$ follows the AR(1) process (á la Example 1). Third, the agents only forecast the scalar value $\theta(t)$ at time t. Finally, the agents observe each other's past forecasts, not current ones. Even though, in principle, their current forecasts could influence each other's future information, we suppose the agents have no such strategic incentives.[4] The next theorem shows

[4] Even though there are many reasons why agents may act myopically, one justification is that reasoning about one's informational effects on others requires a level of sophistication that is unrealistic in most applications (Harel, Mossel, Strack, & Tamuz, 2021). Indeed, most of the learning literature assumes myopic agents either explicitly (e.g., Bala & Goyal, 1998; Keppo, Smith, & Davydov, 2008; Parikh & Krasucki, 1990; Sebenius & Geanakoplos, 1983) or implicitly by

how the observed past forecasts give the agents more information.

Theorem 3 (*Augmented Information*). *Denote the first n_0 elements of an n_1-vector $\boldsymbol{m}$ with the sub-vector $\boldsymbol{m}[n_0] = (v_1, \ldots, v_{n_0})'$. If at time t the agents observe each other's previous forecasts up to time $t-1$, then agent k observes and forecasts based on $\boldsymbol{y}[t]$, $\boldsymbol{x}_1[t-1], \ldots, \boldsymbol{x}_K[t-1]$, and $x_k(t)$.*

After observing others' forecasts, agent k can solve for the private signals that led to those forecasts (for a related model, see Ottaviani & Sørensen, 2006). In terms of our earlier notation, agent k then forecasts at time t based on $\boldsymbol{z}_{k,pb}(t) = (\boldsymbol{y}[t]', \boldsymbol{x}_1[t-1]', \ldots, \boldsymbol{x}_K[t-1]', x_k(t))'$, where the subscript *pb* stands for "public" and reminds us that the agents under this specification observe each other's past forecasts. See Supplementary Material S2 for a specification of private forecasts.

To estimate our model, we refer to Bayesian statistics that treat parameters as random variables (e.g., Gelman et al. 2013). Recently Bayesian methods have become increasingly popular in empirical studies with small sample data because they do not rely on asymptotic approximations – a property that can be a hindrance when employing frequentist methods on small data. Unlike more classical techniques, a Bayesian model requires two components: the *prior* distribution that captures the planner's uncertainty about the parameters of the model before observing the data, and the *likelihood* that specifies the probability of the data as a function of the parameters. The Bayes' rule is then used to update the prior with the observed data. The updated distribution, known as the *posterior*, describes all uncertainties in the parameters after accounting for the planner's prior beliefs and the data.

Supplementary Material S3.1 describes our prior distributions. In short, these are proper and weakly informative so that they do not dominate the construction of the posterior distribution. The likelihood of the data (i.e., the public signals and the forecasts) is given by a multivariate Gaussian distribution with dimension $T + T \times K$ (see Supplementary Material S3.2 for details). Even though this likelihood can be expressed in closed form, it is non-standard and does not permit a closed-form expression of the posterior distribution. Therefore, we estimate the posterior distribution numerically with a generic approach called *slice sampling* (Neal, 2003).[5] This is a state-of-the-art Markov Chain Monte Carlo (MCMC) technique that moves across the parameter space iteratively, each time proposing a new value of the parameters based on a criterion that depends on the most recently accepted value of the parameters. After running the sampler for sufficiently long, the chain converges, and the sequence of accepted values represents an i.i.d. sample from the true posterior distribution.[6] Compared to other MCMC techniques, one of the advantages of slice sampling is that it does not require the user to tune the algorithm before using it. However, before it can be applied in practice, we must address practical matters, namely the proliferation of parameters and infrequent predictions.

4.2. Proliferation of parameters

In our framework, the level of herding and precisions of the private and public signals can vary freely over time. This introduces, in total, $3T$ parameters into our model, which can make estimation infeasible even for moderately large T. In addition, at each time point, each agent's forecast incorporates all past signals and their precisions. Unfortunately, if the signals at different time points are allowed to have different precisions, the likelihood cannot be simplified and computed efficiently. Previous work has assumed time-independent parameters to simplify the estimation of dynamic game models (e.g., Bajari, Benkard, & Levin, 2007; Egesdal, Lai, & Su, 2015; Sieg & Yoon, 2017, and Salz & Vespa, 2015). Therefore, we let private signal precision $\beta(t) = \beta$ and public signal precision $\alpha(t) = \alpha$ for all t so that the agents receive a stationary flow of private and public information.

Even though the agents' forecasts depend on the precision of all past signals, they do not depend on past levels of herding. This observation allows us to estimate a dynamic level of herding without complicating the likelihood calculation. Paralleling our approach to modeling the underlying process $\theta(t)$, we use a Gaussian process prior on the (log-odds of) herding parameter $r(t)$:

$$R(t) = \log(r(t))/(1 - r(t)) \sim \mathcal{GP}(0, \phi_r) \qquad (8)$$

for all $t \in \{1, 2, \ldots, T\}$ and some covariance function $\phi_r(\cdot, \cdot)$ (see Section 3.1). The prior mean of the Gaussian process is set to zero (as is typically done in the literature; see, for example, Williams and Rasmussen 2006), and the covariance function should be chosen based on how the level of herding is expected to evolve over time.

The log-odds of the herding parameter $\boldsymbol{R} = (R(1), \ldots, R(T))'$ follow a multivariate Gaussian distribution. These latent variables must be estimated jointly with the rest of the model. Unfortunately, this entails that the number of parameters grows in the time horizon T, making the estimation computationally expensive even for a moderately large T. To avoid this, we take an approach similar to the projected process method (see, e.g., Seeger, Williams, and Lawrence 2003) and consider a sparse approximation to the latent Gaussian process. Specifically, suppose there are some $T^+ < T$ number of *inducing time points* $\{t_1^+, \ldots, t_{T^+}^+\}$. Denote the covariance matrix of the corresponding $\boldsymbol{R}^+ = (R(t_1^+), \ldots, R(t_{T^+}^+))'$ with $\boldsymbol{\Sigma}_{T^+T^+}$. If

modeling a continuum of agents (e.g., Duffie, Giroux, & Manso, 2010; Duffie, Malamud, & Manso, 2009; Duffie & Manso, 2007; Gale & Kariv, 2003; Vives, 1993).

[5] Slice sampling is motivated by the observation that sampling from the posterior distribution is equivalent to sampling uniformly from the space under the posterior density function. The literature has proposed many different slice-sampling approaches based on this general idea. Specifically, our implementation adopts the oblique hyperrectangle method (Thompson, 2011) that adaptively considers the parameters' posterior dependence, allowing the sampler to explore the parameter space more efficiently.

[6] We monitor convergence by tracking the potential scale reduction factor for each parameter (see, e.g., subsection 11.4. in Gelman et al. 2013).

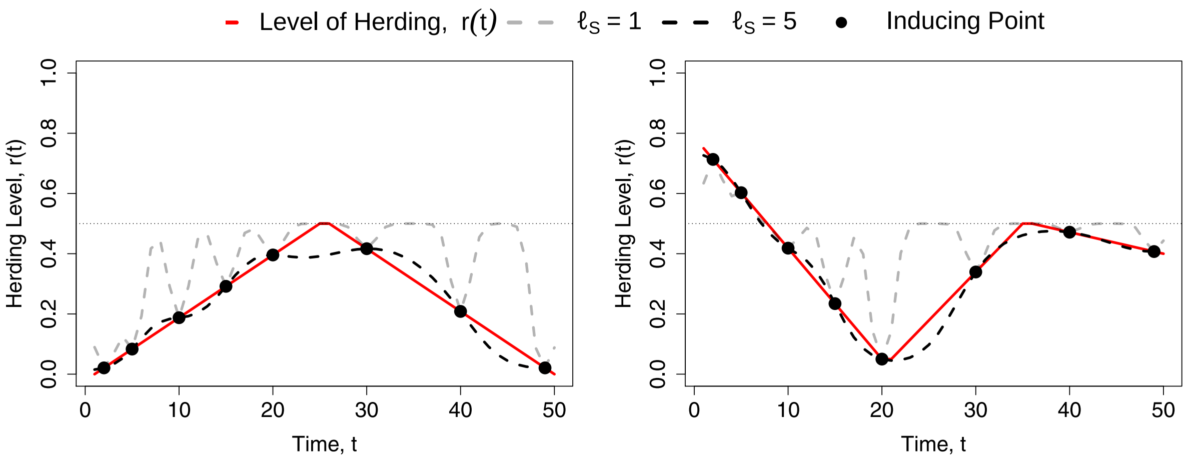


Fig. 4. Illustration of the sparse approximation of the Gaussian Process. The (log-odds of the) level of herding $R(t) = \log(r(t))/(1-r(t)) \sim \mathcal{GP}(0, \phi_{SE})$, where the covariance function ϕ_{SE} is described by the exponential kernel of Example 2 in Section 3.1. There are eight inducing points, shown as solid black circles. The inducing points are projected to the entire time interval using (9), and the projection is converted back to the scale of $r(t)$, leading to a sparse approximation of $r(t)$. The dashed lines show two sparse approximations under two different values of the length-scale ℓ_S. Given that σ_S^2 cancels out in the projection (9), it does not affect the approximation once the inducing points $\boldsymbol{R}^+$ have been fixed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the cross-covariance matrix between $\boldsymbol{R}^+$ and $\boldsymbol{R}$ is $\boldsymbol{\Sigma}_{T^+T}$, the latent values $\boldsymbol{R}$ can be estimated by projecting the inducing points to the entire time range as follows:

$$\hat{\boldsymbol{R}} = \boldsymbol{\Sigma}_{T^+T}\boldsymbol{\Sigma}_{T^+T^+}^{-1}\boldsymbol{R}^+. \tag{9}$$

In this projection approach, the number of parameters does not grow in T. Notably, T^+ can be chosen to match the computational resources available and, hence, does not need to grow in the dataset size. Ideally, the inducing points would be placed around times when the level of herding is expected to change substantially. However, without any such prior knowledge available, we place the points uniformly over the interval $\{1, \ldots, T\}$.

Fig. 4 illustrates the approximation for $T = 50$ times points. $T^+ = 8$ inducing points are shown as solid black circles. Here the covariance function is described by the exponential kernel of Example 2 (Section 3.1). The solid red line shows the true level of herding $r(t)$, and the dashed lines show the projected approximation under different values of the length-scale ℓ_S. The darker dashed line corresponds to a higher value of ℓ_S and hence a less "wiggly" approximation of the true level of herding. When ℓ_S is small, the sparse approximation quickly reverts to 0.5 as it moves away from the inducing points. This happens because, under small ℓ_S, each inducing point $R^+(t^+)$ does not correlate strongly with distant values of $R(t)$. Then, in the absence of information from the inducing points, the approximation relies on prior information. As mentioned, the prior mean of $R(t)$ is fixed to zero. Given that $R(t)$ represents the log-odds of $r(t)$, a prior mean of zero corresponds to 0.5 in the scale of $r(t)$. Under this parsimonious approximation, the unknown parameters of our model specification are $(\theta_0, \mu_0, \rho, \sigma, \alpha, \beta, \phi_r, \boldsymbol{R}^+)'$.

4.3. *Infrequent predictions*

In real-life surveys, not all agents make forecasts systematically at every time point, and some agents may enter or exit the system in the middle of the survey. The result can be a highly sparse dataset with different numbers and time points of forecasts per agent. To illustrate, in our EOG example in Fig. 1, there are $T = 169$ time points and $K = 110$ agents. If each agent predicted at each time point, there would be a total of $TK = 18{,}590$ forecasts. However, given that the actual dataset involves only 1,698 forecasts, around 90.1% of the potential data points are missing. If our procedure is to have real practical significance, it must be able to handle such sparse data.

To explore ways to do this, collect all agent k's forecasts over time points $t \in \{1, 2, \ldots, T\}$ into a vector $\boldsymbol{a}_k$ and let $\boldsymbol{a} = (\boldsymbol{a}_1', \ldots, \boldsymbol{a}_K')'$ be a concatenation of all agents' vectors. If $\boldsymbol{y}$ collects all public signals in our dataset, then one approach to handling missing values recognizes that our complete dataset $(\boldsymbol{y}, \boldsymbol{a})$ follows a multivariate Gaussian distribution. Any non-missing subset of these data also follows a multivariate Gaussian distribution. Given that missing values do not typically occur systematically, the covariance matrix of the non-missing values is unlikely to have a structure (such as block compound symmetry) that can be exploited for efficient computation of the likelihood. Therefore, this approach is feasible only if there are few forecasts. Alternatively, one can treat the missing values as unknown variables that are estimated together with all other parameters (Gelman et al., 2013). Unfortunately, given that this introduces one additional unknown variable per missing value, it becomes computationally infeasible if the number of missing values is moderately large.

Therefore, standard approaches are helpful when there are either very few non-missing values (i.e., the dataset is small) or very few missing values (i.e., the dataset is almost complete). Unfortunately, many real-world datasets resemble our EOG example, where there is a large number of both non-missing and missing values. To handle such common cases, we introduce an alternative approach. First, denote all non-missing forecasts and public signals with the vectors $\breve{\boldsymbol{a}}$ and $\breve{\boldsymbol{y}}$, respectively. Collect all non-missing forecasts given at time point t into the vector $\breve{\boldsymbol{a}}(t)$ and let $K(t) = |\breve{\boldsymbol{a}}(t)|$ denote the number of forecasts at time t. The likelihood is computed in its factored form, $\pi(\breve{\boldsymbol{a}}, \breve{\boldsymbol{y}}) = \pi(\breve{\boldsymbol{a}}|\breve{\boldsymbol{y}})\pi(\breve{\boldsymbol{y}})$. If T is not very large or only a few public signals are missing, the marginal likelihood of the public signals $\pi(\breve{\boldsymbol{y}})$ can be calculated using the second standard approach described above. However, in practice, market forces or a central planner often give the public signal, such as a central bank. For instance, in our EOG example, the public signal is given by the current market price. Therefore, public signals are unlikely to be missing, and structure can be exploited to calculate the marginal likelihood $\pi(\breve{\boldsymbol{y}})$ efficiently, as described in Supplementary Material S3.2.

The main challenge is typical with infrequent forecasts. Given that the agents share all past signals except the most recent private signals, forecasts within the same time point are exchangeable. The conditional likelihood $\pi(\breve{\boldsymbol{a}}|\breve{\boldsymbol{y}})$ then has a block structure, where the forecasts in a block have the same mean, variance, and covariance. Specifically, all unique elements of the blocked mean vector, the within-block variance and covariance, and the cross-block variance and covariance are given, respectively, by

$$\breve{\boldsymbol{\mu}} = \mathbb{E}(\boldsymbol{a}_k|\breve{\boldsymbol{y}}) = \boldsymbol{\mu}_{\boldsymbol{a}_k} - \boldsymbol{\Sigma}_{\boldsymbol{a}_k,\breve{y}}\boldsymbol{\Sigma}_{\breve{\boldsymbol{y}}}^{-1}(\breve{\boldsymbol{y}} - \boldsymbol{\mu}_{\breve{\boldsymbol{y}}}), \tag{10}$$

$$\breve{\boldsymbol{\Sigma}}_{kk} = \text{Cov}(\boldsymbol{a}_k, \boldsymbol{a}_k|\breve{\boldsymbol{y}}) = \boldsymbol{\Sigma}_{\boldsymbol{a}_k} - \boldsymbol{\Sigma}_{\boldsymbol{a}_k,\breve{y}}\boldsymbol{\Sigma}_{\breve{\boldsymbol{y}}}^{-1}\boldsymbol{\Sigma}'_{\boldsymbol{a}_k,\breve{y}}, \text{ and} \tag{11}$$

$$\breve{\boldsymbol{\Sigma}}_{kj} = \text{Cov}(\boldsymbol{a}_k, \boldsymbol{a}_j|\breve{\boldsymbol{y}}) = \boldsymbol{\Sigma}_{\boldsymbol{a}_k,\boldsymbol{a}_j} - \boldsymbol{\Sigma}_{\boldsymbol{a}_k,\breve{\boldsymbol{y}}}\boldsymbol{\Sigma}_{\breve{\boldsymbol{y}}}^{-1}\boldsymbol{\Sigma}'_{\boldsymbol{a}_k,\breve{\boldsymbol{y}}}, \tag{12}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{a}_k,\breve{\boldsymbol{y}}} = \text{Cov}(\boldsymbol{a}_k, \breve{\boldsymbol{y}})$ and $\boldsymbol{\Sigma}_{\boldsymbol{a}_k} = \text{Cov}(\boldsymbol{a}_k, \boldsymbol{a}_k)$. We can take advantage of this structure by computing the conditional likelihood in a factored form

$$\pi(\breve{\boldsymbol{a}}|\breve{\boldsymbol{y}}) = \prod_{t:K(t)>0} \pi(\breve{\boldsymbol{a}}(t)|\breve{\boldsymbol{y}}, \breve{\boldsymbol{a}}(t_0), t_0 < t), \tag{13}$$

because each factor is the density of exchangeable Gaussian random variables and hence can be evaluated efficiently.

The computation proceeds sequentially from $t = 1$ to $t = T$, each time updating the quantities (10)–(12) by conditioning on all public signals and previous forecasts. Denote the updated versions at iteration t with $\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}$, $\breve{\boldsymbol{\Sigma}}_{kk}^{(t)}$, and $\breve{\boldsymbol{\mu}}^{(t)}$. First, we initialize $\breve{\boldsymbol{\Sigma}}_{kk}^{(1)} = \breve{\boldsymbol{\Sigma}}_{kk}$, $\breve{\boldsymbol{\Sigma}}_{kj}^{(1)} = \breve{\boldsymbol{\Sigma}}_{kj}$, and $\breve{\boldsymbol{\mu}}^{(t)} = \breve{\boldsymbol{\mu}}$. Then, for $t \in \{1, 2, \dots, T\}$, iteration t begins by computing $\pi(\breve{\boldsymbol{a}}(t)|\breve{\boldsymbol{y}}, \breve{\boldsymbol{a}}(t_0), t_0 < t)$, which is a multivariate Gaussian with constant mean and compound symmetric covariance matrix. Specifically, the constant mean $\breve{\mu}^{(t)}$ is the tth element of $\breve{\boldsymbol{\mu}}^{(t)}$. The diagonal and off-diagonal elements, $a^{(t)}$ and $b^{(t)}$, of the covariance matrix are the (t, t)th elements of $\breve{\boldsymbol{\Sigma}}_{kk}^{(t)}$ and $\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}$, respectively. By the properties of a compound symmetric matrix (e.g., Rao 2009 and Theorem 2 in the supplementary material of Dobbin and Simon 2005), the determinant of this covariance matrix is $(a^{(t)} - b^{(t)})^{K(t)-1}[a^{(t)} + (K(t) - 1)b^{(t)}]$ and its inverse is also compound symmetric with off-diagonal element $b^{(t)'} = -b^{(t)}/\{(a^{(t)} - b^{(t)})[(a^{(t)} - b^{(t)}) + K(t)b^{(t)}]\}$ and diagonal element $1/(a^{(t)} - b^{(t)}) + b^{(t)'}$. These results lead to a simple closed form expression of $\pi(\breve{\boldsymbol{a}}(t)|\breve{\boldsymbol{y}}, \breve{\boldsymbol{a}}(t_0), t_0 < t)$. The iteration ends by updating $\breve{\boldsymbol{\mu}}^{(t)}$, $\breve{\boldsymbol{\Sigma}}_{kk}^{(t)}$, and $\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}$ in terms of $\breve{\boldsymbol{a}}(t)$. By the conditional distributions of Gaussian random variables (e.g., Rao 2009), the updating equations are

$$\breve{\boldsymbol{\mu}}^{(t+1)} = \breve{\boldsymbol{\mu}}^{(t)} + \rho(t)\left[\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}\right]_{\cdot t}\left(\frac{1}{K(t)}\mathbf{1}'_{K(t)}\breve{\boldsymbol{a}}(t) - \breve{\mu}^{(t)}\right),$$

$$\breve{\boldsymbol{\Sigma}}_{kk}^{(t+1)} = \breve{\boldsymbol{\Sigma}}_{kk}^{(t)} - \rho(t)\left[\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}\right]_{\cdot t}\left[\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}\right]_{t\cdot}, \text{ and}$$

$$\breve{\boldsymbol{\Sigma}}_{kj}^{(t+1)} = \breve{\boldsymbol{\Sigma}}_{kj}^{(t)} - \rho(t)\left[\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}\right]_{\cdot t}\left[\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}\right]_{t\cdot},$$

where $\rho(t) = K(t)(a^{(t)'} + [K(t) - 1]b^{(t)'})$, and the tth row and column of $\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}$ are denoted with $\left[\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}\right]_{t\cdot}$ and $\left[\breve{\boldsymbol{\Sigma}}_{kj}^{(t)}\right]_{\cdot t}$, respectively. After the final iteration $t = T$, the conditional likelihood $\pi(\breve{\boldsymbol{a}}|\breve{\boldsymbol{y}})$ is calculated using factored form (13).

5. Application

5.1. Synthetic data

Before applying our estimation procedure to real-world forecasting data, we evaluate its potential on synthetically generated data. Specifically, the underlying state variable $\theta(t)$ follows the AR(1) process (4) in Example 1 (Section 3.1) with $\mu_0 = 10$, $\theta_0 = 0.0$, $\rho = 0.95$, and $\sigma^2 = 25$. The precisions of public and private information are fixed to $\alpha = 0.05$ and $\beta = 0.1$.

Fig. 5 illustrates the estimated level of herding $r(t)$ on a single simulated dataset with $T = 50$ time points and $K = 50$ agents.[7] In this time interval, the underlying state is given by an upward trending process with a large random variation. The solid red line is the true level of herding $r(t)$, the solid black line is the posterior mean, and the shaded region gives the (pointwise) 95% credible intervals. Estimation of the herding level uses ten inducing points placed evenly over the time interval. The covariance structure (8) is described by a simple squared exponential kernel (recall Example 2 in Section 3.1). Finally, even though our primary focus is on the herding parameter, Table 2 gives the posterior means and 95% credible intervals for all other parameters. Overall, our procedure captures the true parameter values reasonably well. The private signal precisions are slightly underestimated. This analysis, however, considers only a single simulated dataset. Therefore, we can only expect a reasonable alignment between our estimates and the truth.

A more comprehensive evaluation can be conducted by simulating many datasets under different forecasting environments. Fig. 6 describes the performance of our estimation procedure over a grid of different time horizons $T \in \{10, 20, \dots, 50\}$ and numbers of agents $K \in \{10, 20, \dots, 50\}$. Besides T and K, the other parameters

[7] In this section, each model is estimated by running our slice sampler for a total of 10,000 iterations. The first 5,000 iterations are discarded for burn-in.

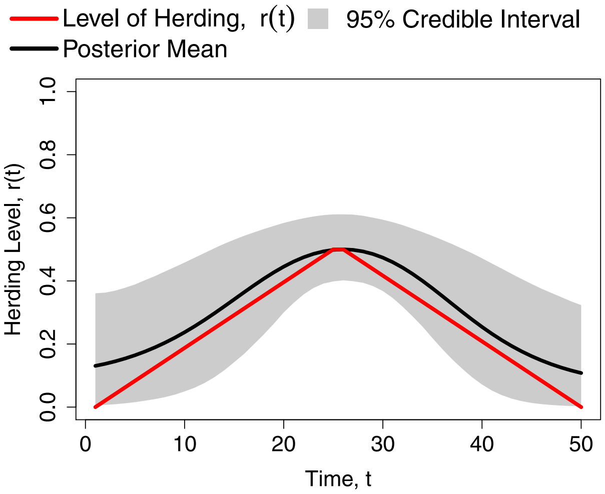


Fig. 5. The estimated level of herding $r(t)$ with 10 inducing points. There are $T = 50$ time points, $K = 50$ agents, and the underlying state variable $\theta(t)$ follows the AR(1) process (4) in Example 1 with $\mu_0 = 10$, $\theta_0 = 0.0$, $\rho = 0.95$, $\sigma^2 = 25$, $\alpha = 0.05$, and $\beta = 0.1$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Posterior means and 95% credible intervals of other parameters besides the level of herding.

Parameter	μ_0	θ_0	ρ	σ^2	α	β
True value	10	0.0	0.95	25	0.05	0.1
Posterior mean	8.56	−0.67	0.96	28.56	0.05	0.07
95% interval	[4.9, 11.3]	[−6.7, 7.5]	[0.94, 0.99]	[24.2, 32.8]	[0.04, 0.06]	[0.06, 0.10]

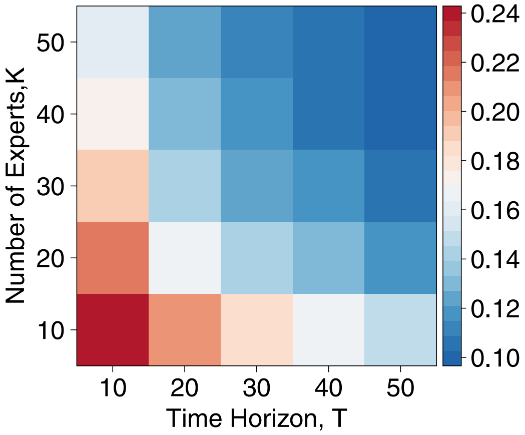


Fig. 6. The root-mean-squared-error in estimating the level of herding $r(t)$ over the entire time horizon $t \in \{1, \ldots, T\}$. Each value represents the average error over 1,000 simulated datasets. Besides T and K, the other parameters are the same as in Fig. 5.

are the same as before. In particular, in each case, the true level of herding increases linearly from $r(1) = 0$ to $r(T/2) = 0.5$ and then decreases linearly to $r(T) = 0$, as in Fig. 5. Under each (T, K)-pair, a total of 1,000 datasets were generated, leading to 1,000 estimates of $r(t)$.

Fig. 6 plots the average root-mean-squared-error in estimating $r(t)$ over the entire time horizon. For instance, when $T = 10$ and $K = 10$, the error of our estimation procedure is around 0.25, suggesting that the estimated $r(t)$ is, on average, around 0.25 from the true level of herding. This error decreases quickly in T and K. The rate at which this happens is very similar for both T and K, suggesting that asking more agents for their forecasts or asking the current agents to make more forecasts over a longer time period are both effective ways to improve the estimation accuracy. Furthermore, this illustrates how our Bayesian estimation procedure does not require unreasonably large amounts of data.

5.2. Case study: Future stock prices of EOG resources, Inc.

In this subsection, we apply our estimation technique to the analyst forecasts of year-over-year differences in the log prices shown in Fig. 1(b). First, we choose a suitable covariance function for the latent Gaussian process of the level of herding $r(t)$. Specifically, we consider a nested sequence of models with increasing complexity and explore how much complexity the data support (see Section 3.1):

(a) No herding: $r(t) = 0$ for all $t \in \{1, \ldots, T\}$.
(b) Constant level of herding: $r(t) = r$ for all $t \in \{1, \ldots, T\}$.
(c) Long-term trend in $r(t)$, modeled with a squared exponential kernel (SE).
(d) Long-term trend with an annual periodic pattern in $r(t)$, modeled with a sum of a squared exponential kernel and a periodic kernel (SE + P). Periodicity can occur if, for instance, the agents always go through performance evaluations in the same month of the year.

(*e*) Long term trend with an annual quasi-periodic pattern (that can change over time) in $r(t)$, modeled with a sum of a squared exponential kernel and a product of another squared exponential kernel and a periodic kernel (SE + SE×P). This is similar to (d), but the seasonal component can evolve over time.

The final three models $(c) - (e)$ depend on the number and location of inducing points of the latent Gaussian process (recall Section 4.2). In each case, we consider $T^+ \in \{10, 15, \ldots, 30\}$ inducing points placed uniformly over the time interval $[1, T]$. Systematically comparing models, ranging from simple to complex, allows us to balance the expressiveness of the model with what is supported by the data. See Gigerenzer (2018) for further discussion on systematic model comparison and selection. Of course, many other models could be considered by mixing and matching more kernels. However, each kernel brings in more parameters, increases the computational burden, and reduces the interpretability of the model.

For comparing models, the literature recommends using the *expected log pointwise predictive density* (ELPD). This is because, unlike the mean square error or absolute error that only measure how accurately the models make point forecasts, the log-likelihood evaluates the entire predictive distribution (see, e.g., Williams and Rasmussen 2006). In general, ELPD can be estimated with *leave-one-out cross-validation* (LOO-CV), where data points are left out one at a time and then predicted with a model estimated based on the remaining data. Unfortunately, this procedure is computationally very demanding because the model must be estimated as many times as there are observations in the dataset. This is infeasible for many complex Bayesian models that can take hours or even days to estimate using sampling-based techniques. However, if the data follows a multivariate Gaussian distribution, as is the case in our dynamic model of herding, Pareto smoothed importance sampling can be used to approximate LOO-CV based on a single model estimated on the full data (Bürkner, Gabry, and Vehtari 2020, Gelfand, Dey, and Chang 1992, Vehtari, Gelman, and Gabry 2017).

Table 3 presents these estimates $\widehat{\text{ELPD}}$ under each model in our comparison.[8] The first column shows that under the no herding model, $\widehat{\text{ELPD}}$ is 1,199.2. The other columns present the estimated (absolute) changes in ELPD from the no-herding model to increasingly more complex models. Larger values indicate better model fit in the log scale. Therefore, for instance, a difference of $\log(2) \approx 0.69$ implies that the likelihood is two times larger under the improved model. To gauge statistical significance, the values in parentheses provide the standard errors of the corresponding changes in ELPD.

The comparison gives us several observations. First, we compare the two simple models with no herding $(r(t) = 0)$ and constant herding $(r(t) = r)$. They both assume forecast heterogeneity to remain constant over time. However, our structural model of herding $(r(t) = r)$ is a better fit for the data than the classical model of Bayesian rational agents who simply forecast the outcome as accurately as possible $(r(t) = 0)$. Furthermore, given that the improvement in $\widehat{\text{ELPD}}$ is more than two times the associated standard error, we consider this improvement statistically significant. Second, the models with dynamic $r(t)$ fit the data significantly better than the two models with constant $r(t)$. Third, the models with different Gaussian process formulations for $r(t)$ do not show large differences in model fit. This, in particular, suggests no evidence of annual periodicity in the agents' herding. Fourth, the fit of our model shows little sensitivity to the exact number of inducing points.

Given that the data do not show strong support for a large number of inducing points or a periodic kernel, we select and, for the rest of this subsection, analyze the simplest model with dynamic $r(t)$, namely the SE model with ten inducing points. First, Table 4 shows the posterior means and 95% credible intervals of the parameters under this model. The posterior mean of the intercept term μ_0 is 0.11, reflecting an upward trend in the EOG stock price. The value of ρ is estimated to be small, which suggests a strong mean-reversion of the underlying process. By comparing the estimates of α and β, the agents' private information is more precise than the public information. Finally, we consider the parameters σ_R and ℓ_R that control the level of herding $r(t)$. Given that σ_R is estimated to be different from zero, we can expect variation in the level of herding over time. Furthermore, given that ℓ_R is large, we expect this variation to be captured with a slowly moving function.

Fig. 7(a) shows the posterior mean and 95% credible interval of the level of herding $r(t)$ for all $t \in \{1, \ldots, T\}$. The horizontal dashed line around 0.36 is the posterior mean of herding under the constant $r(t) = r$ model. This can be viewed as a constant approximation of the dynamic $r(t)$. The dynamic level of herding is estimated to be high (around 0.4–0.5) in the early 2000s. Then, during the 2007–2008 financial crisis (highlighted with a diagonally dashed rectangle), the level of herding dropped to 0.1. It is possible that, before the crises, strongly herding analysts were operating under similar information. This indicates a lack of risk management in decision-making, which can cause the market to be more vulnerable. Eventually, the level of herding $r(t)$ drops, suggesting the analysts lose trust in the public information. Even though this explanation may seem plausible, based on our analysis alone, we cannot know whether this is the true driver of the drop in $r(t)$. A similar dynamic pattern was found by Jia et al. (2022), who analyze herding in probabilistic forecasts of inflation, unemployment, and gross domestic product.

Strong levels of herding, however, do not need to result in large decreases in the agents' expected forecasting accuracy. For instance, if the public information is much more precise than the agents' private information, overemphasizing the public signal due to herding is unlikely to cause high accuracy losses. However, in the current case study, the agents' private information is estimated to be

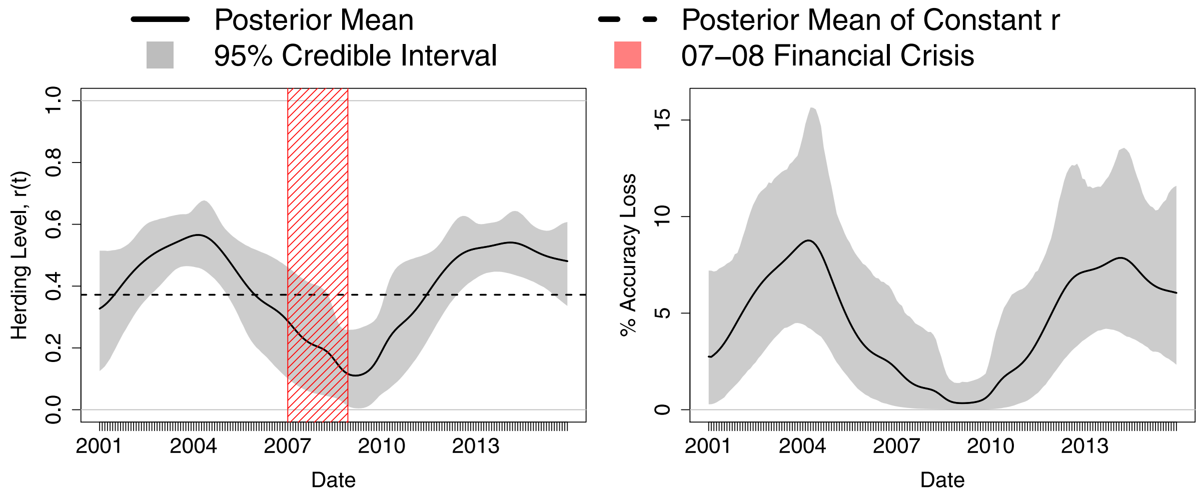
[8] We only leave out and predict the agents' forecasts because our main interest is on modeling agent behavior – not the process of public signals. Each model is estimated by running our slice sampler for a total of 500,000 iterations. The first 250,000 iterations are discarded for burn-in. We monitor convergence by repeating each estimation twice from two different starting points and ensuring that the two chains converge to similar estimates of the model parameters and ELPD.

Table 3
Model comparison on the EOG forecasting data. Estimates of the expected log pointwise predictive density $\widehat{\text{ELPD}}$, obtained using Pareto smoothed importance sampling. The first column shows $\widehat{\text{ELPD}}$ for the no herding ($r(t) = 0$) model, and the other columns show how increasingly more complex models improve this $\widehat{\text{ELPD}}$. T^+ is the number of inducing points, SE is the squared exponential kernel, and P is the periodic kernel. Higher scores represent better model fit in the log scale. To gauge statistical significance, the values in the parentheses provide the standard errors of the corresponding changes in ELPD.

	$r(t) = 0$	$r(t) = r$	T^+	SE	SE + P	SE + SE × P
$\widehat{\text{ELPD}}$	1,199.2	+20.8 (7.6)	10	+92.3 (19.2)	+92.1 (19.4)	+92.4 (19.6)
			15	+90.9 (19.3)	+89.9 (19.4)	+94.5 (19.6)
			20	+100.3 (19.7)	+97.5 (19.7)	+91.0 (19.0)
			25	+99.9 (20.0)	+97.4 (19.7)	+91.2 (19.1)
			30	+101.1 (20.2)	+97.1 (19.6)	+94.3 (19.5)

Table 4
Parameter estimates under the SE model in Table 3 with 10 inducing points.

Parameter	μ_0	θ_0	ρ	σ	$\alpha^{-1/2}$	$\beta^{-1/2}$	σ_R	ℓ_R
Posterior mean	0.11	−1.14	0.22	0.39	0.78	0.56	2.05	25.28
2.5% quantile	0.08	−2.56	0.16	0.35	0.67	0.45	1.21	5.22
97.5% quantile	0.13	0.13	0.29	0.43	0.90	0.68	3.64	50.29



(a) Level of Herding, $r(t)$

(b) Percent Loss in Expected Accuracy.

Fig. 7. Illustration of the estimated model (SE model in Table 3 with ten inducing points) on 1-year ahead forecasts of the stock price of EOG Resources Inc. Fig. 7(a) shows the posterior mean and 95% credible interval of the level of herding. The horizontal dashed line is the posterior mean of herding under the constant $r(t) = r$ model. Fig. 7(b) reports the percentage loss in accuracy as the agents' level of herding rises from zero to the estimated level of $r(t)$.

more precise than the public information (see Table 4). To explore the effect of herding on the accuracy, we compute the percentage change in the expected accuracy as the agents' level of herding rises from zero to the estimated level of $r(t)$. Fig. 7(b) summarizes the resulting posterior distributions of the accuracy loss by plotting their posterior means and (pointwise) 95% credible intervals. As expected, the agents' accuracy decreases proportionally to their level of herding. Before and after the financial crisis of 2007–2008, the loss of accuracy is as high as 7%. During the crisis, however, when the herding is low, the loss is near zero.

For robustness, Supplementary Material S4 repeats this analysis for 150 other companies from three industry divisions: (i) Finance, Insurance, and Real Estate, (ii) Services, and (iii) Manufacturing. Overall, the results align well with those described in our EOG case study. In 136/150 and 150/150 cases, our model with constant or dynamic herding (i.e., the SE model with $T^* = 10$ inducing points), respectively, is a better fit to data than the baseline model of no herding. Furthermore, in 84/150 and 148/150 cases, respectively, these differences are statistically significant. The patterns in the estimated level of herding and expected accuracy loss exhibit high variability across stocks but, on average, resemble those shown in Fig. 7 for EOG. In particular, the average level of herding is estimated to remain around 0.4–0.5, corresponding to an accuracy loss of around 10%. Across industries, herding is the lowest in the manufacturing division and highest in the finance, insurance, and real estate divisions. Furthermore, the drop during the financial crisis is the sharpest in the finance, insurance, and real estate divisions. This can be expected because, after all, this was a financial crisis.

6. Conclusion

The main contributions of this paper are the proposed framework of agent forecasting under herding and its estimation procedure. Our framework is flexible and can be adapted to a wide range of dynamic contexts. In particular, the user of our framework can alter the agents' information structure by conditioning on different signals and changing the underlying dynamics with the mean and covariance functions. Our estimation procedure uses numerical Bayesian techniques that, relative to their frequentist counterparts, often perform better in small samples and can incorporate prior information. This is important in practice because many forecasting data sets are small. Estimation efficiency can be improved by choosing an analytically tractable covariance function. For instance, in our AR(1) formulation, both the covariance matrix of the state process and its inverse have simple well-known forms (see Supplementary Material S6). If this is not possible, it may be possible to approximate the likelihood with standard techniques designed for Gaussian process modeling (Williams & Rasmussen, 2006).

To illustrate our estimation framework, we applied it to agent forecasts of future stock prices. The empirical results indicate substantial herding among the forecasters before and after the 2007–2008 financial crisis. Herding is estimated to decrease the agents' forecasting accuracy by around 5%. This illustrates the negative effects of herding: it decreases agents' accuracy and forecast heterogeneity, which, in turn, decreases welfare and the value of a single agent.

A direction of future work could consider the optimal aggregation of forecasts made by a herding crowd. This is equivalent to using our model to estimate the distribution of the underlying state conditional on the agents' current and past forecasts. Such a Bayesian approach would consider the agents' cognitive bias, namely herding, seeking to "undo" it, and revealing the agents' underlying signals that can be combined into an accurate consensus forecast. Even though cognitive models have been leveraged in aggregation before (e.g., Lee & Danileiko, 2014), most work has focused on static forecasting environments. There is very little work on dynamic forecast aggregation (e.g., Satopää et al., 2014) – not to mention aggregation of forecasts of agents who participate in a dynamic game. This is surprising because, in many crowd forecasting applications, agents consider each other's forecasts and make their forecasts over time (see, e.g., Atanasov et al. 2017, Atanasov, Witkowski, Ungar, Mellers, and Tetlock 2020, Ungar, Mellers, Satopää, Tetlock, and Baron 2012).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Jussi Keppo acknowledges research support from National Research Foundation (Singapore) and A*STAR (Singapore), under RIE 2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) grant call (Grant no. I2001E0059).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2023.03.001.

References

Ashiya, M., & Doi, T. (2001). Herd behavior of Japanese economists. *Journal of Economic Behaviour and Organization*, *46*(3), 343–346.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., et al. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, *63*(3), 691–706.

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, *160*, 19–35.

Baddeley, M. (2010). Herding, social influence and economic decision-making: socio-psychological and neuroscientific analyses. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *365*(1538), 281–290.

Bajari, P., Benkard, C. L., & Levin, J. (2007). Estimating dynamic models of imperfect competition. *Econometrica*, *75*(5), 1331–1370.

Bala, V., & Goyal, S. (1998). Learning from neighbours. *Review of Economic Studies*, *65*(3), 595–621.

Banerjee, A., Guo, X., & Wang, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, *51*(7), 2664–2669.

Bergemann, D., & Morris, S. (2013). Robust predictions in games with incomplete information. *Econometrica*, *81*(4), 1251–1308.

Bernhardt, D., Campello, M., & Kutsoati, E. (2006). Who herds? *Journal of Financial Economics*, *80*(3), 657–675.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.

Bürkner, P.-C., Gabry, J., & Vehtari, A. (2020). Efficient leave-one-out cross-validation for Bayesian non-factorized normal and student-t models. *Computational Statistics*, 1–19.

Çelen, B., Kariv, S., & Schotter, A. (2010). An experimental test of advice and social learning. *Management Science*, *56*(10), 1687–1701.

Cipriani, M., & Guarino, A. (2014). Estimating a structural model of herd behavior in financial markets. *American Economic Review*, *104*(1), 224–251.

Clement, M. B., & Tse, S. Y. (2005). Financial analyst characteristics and herding behavior in forecasting. *The Journal of Finance*, *60*(1), 307–341.

Clements, M. P. (2018). Do macroforecasters herd? *Journal of Money, Credit and Banking*, *50*(2–3), 265–292.

Cutler, D. M., Poterba, J. M., & Summers, L. H. (1991). Speculative dynamics. *Review of Economic Studies*, *58*(3), 529–546.

Dobbin, K., & Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, *6*(1), 27–38.

Duffie, D., Giroux, G., & Manso, G. (2010). Information percolation. *American Economic Journal: Microeconomics*, *2*(1), 100–111.

Duffie, D., Malamud, S., & Manso, G. (2009). Information percolation with equilibrium search dynamics. *Econometrica*, *77*(5), 1513–1574.

Duffie, D., & Manso, G. (2007). Information percolation in large markets. *American Economic Review*, *97*(2), 203–209.

Egesdal, M., Lai, Z., & Su, C.-L. (2015). Estimating dynamic discrete-choice games of incomplete information. *Quantitative Economics*, *6*(3), 567–597.

Fama, E. F., & French, K. R. (1988). Permanent and temporary components of stock prices. *Journal of Political Economy*, *96*(2), 246–273.

Gaba, A., Popescu, D. G., & Chen, Z. (2018). Assessing uncertainty from point forecasts. *Management Science*, *65*(1), 90–106.

Gale, D., & Kariv, S. (2003). Bayesian learning in social networks. *Games and Economic Behavior*, *45*(2), 329–346.

Gavin, W. T., & Kliesen, K. L. (2006). *Forecasting inflation and output: Comparing data-rich models with simple rules*. Federal Reserve Bank of St. Louis.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Department of Statistics, Stanford University.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. CRC Press.

Giacomini, R., Skreta, V., & Turen, J. (2020). Heterogeneity, inattention, and Bayesian updates. *American Economic Journal: Macroeconomics*, *12*(1), 282–309.

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*(3–4), 303–336.

Graham, J. R. (1999). Herding among investment newsletters: Theory and evidence. *The Journal of Finance*, *54*(1), 237–268.

Grinblatt, M., Titman, S., & Wermers, R. (1995). Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. *American Economic Review*, *85*, 1088–1105.

Harel, M., Mossel, E., Strack, P., & Tamuz, O. (2021). Rational groupthink. *Quarterly Journal of Economics*, *136*(1), 621–668.

Hong, H., Kubik, J. D., & Solomon, A. (2000). Security analysts' career concerns and herding of earnings forecasts. *Rand Journal of Economics*, *31*, 121–144.

Jegadeesh, N., & Kim, W. (2010). Do analysts herd? An analysis of recommendations and market reactions. *The Review of Financial Studies*, *23*(2), 901–937.

Jia, Y., Keppo, J., & Satopää, V. (2022). Herding in probabilistic forecasts. *Management Science*.

Karger, E., Monrad, J., Mellers, B., & Tetlock, P. (2021). Reciprocal scoring: A method for forecasting unanswerable questions. Available At SSRN.

Keppo, J., Audet, N., Heiskanen, P., & Vehvilainen, I. (2004). Modelling electricity forward curve dynamics in the nordic market. In *Modelling prices in competitive electricity markets* (pp. 251–264). Wiley Series in Financial Economics.

Keppo, J., Smith, L., & Davydov, D. (2008). Optimal electoral timing: Exercise wisely and you may live longer. *Review of Economic Studies*, *75*(2), 597–628.

Kim, J. Y., Kim, Y., & Shim, M. (2019). Do financial analysts herd? Available At SSRN: https://Ssrn.Com/Abstract=3372445.

Kim, O., Lim, S. C., & Shaw, K. W. (2001). The inefficiency of the mean analyst forecast as a summary forecast of earnings. *Journal of Accounting Research*, *39*(2), 329–335.

Kim, C., & Pantzalis, C. (2003). Global/industrial diversification and analyst herding. *Financial Analysts Journal*, *59*(2), 69–79.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., et al. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, *5*(11), eaaw9011.

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1992). The impact of institutional trading on stock prices. *Journal of Financial Economics*, *32*(1), 23–43.

Lamont, O. A. (2002). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behaviour and Organization*, *48*(3), 265–280.

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*(3), 259.

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., & Pfeifer, P. E. (2013). The wisdom of competitive crowds. *Operations Research*, *61*(6), 1383–1398.

Liu, Y., Wang, J., & Chen, Y. (2020). Surrogate scoring rules. In *Proceedings of the 21st acm conference on economics and computation* (pp. 853–871).

Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, *1*(1), 41–66.

Morris, S., & Shin, H. S. (2002). Social value of public information. *American Economic Review*, *92*(5), 1521–1534.

Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. *Journal of Financial Economics*, *104*(2), 228–250.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 705–741.

Olsen, R. A. (1996). Implications of herding behavior for earnings estimation, risk assessment, and stock returns. *Financial Analysts Journal*, *52*(4), 37–41.

Ottaviani, M., & Sørensen, P. N. (2006). The strategy of professional forecasting. *Journal of Financial Economics*, *81*(2), 441–466.

Palley, A. B., & Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, *65*(5), 2291–2309.

Parikh, R., & Krasucki, P. (1990). Communication, consensus, and knowledge. *Journal of Economic Theory*, *52*(1), 178–189.

Pfeifer, P. E., Grushka-Cockayne, Y., & Lichtendahl Jr, K. C. (2014). The promise of prediction contests. *American Statistician*, *68*(4), 264–270.

Prendergast, C., & Stole, L. (1996). Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning. *Journal of Political Economy*, *104*(6), 1105–1134.

Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, *13*(10), 420–428.

Rao, C. R. (2009). *Wiley series in probability and statistics*: *vol. 22*, *Linear statistical inference and its applications*. New York, New York: John Wiley & Sons.

Ravishanker, N., & Dey, D. K. (2001). *A first course in linear model theory*. CRC Press.

Salz, T., & Vespa, E. (2015). Estimating dynamic games of oligopolistic competition: An experimental investigation. Available At SSRN: https://Ssrn.Com/Abstract=2621270.

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., Ungar, L. H., et al. (2014). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *Annals of Applied Statistics*, *8*(2), 1256–1280.

Satopää, V. A., Pemantle, R., & Ungar, L. H. (2016). Modeling probability forecasts via information diversity. *Journal of the American Statistical Association*, *111*(516), 1623–1633.

Scharfstein, D. S., & Stein, J. C. (1990). Herd behavior and investment. *American Economic Review*, 465–479.

Sebenius, J. K., & Geanakoplos, J. (1983). Don't bet on it: Contingent agreements with asymmetric information. *Journal of the American Statistical Association*, *78*(382), 424–426.

Seeger, M., Williams, C., & Lawrence, N. (2003). Fast forward selection to speed up sparse Gaussian process regression.

Sieg, H., & Yoon, C. (2017). Estimating dynamic games of electoral competition to evaluate term limits in us gubernatorial elections. *American Economic Review*, *107*(7), 1824–1857.

Smith, L., & Sørensen, P. (2000). Pathological outcomes of observational learning. *Econometrica*, *68*(2), 371–398.

Smith, L., Sørensen, P., & Tian, J. (2021). Informational herding, optimal experimentation, and contrarianism. *Review of Economic Studies*, Forthcoming.

Stallen, M., De Dreu, C. K., Shalvi, S., Smidts, A., & Sanfey, A. G. (2012). The herding hormone: oxytocin stimulates in-group conformity. *Psychological Science*, *23*(11), 1288–1292.

Taylor, J. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, *54*(8), 799–805.

Thompson, M. B. (2011). *Slice sampling with multivariate steps*. University of Toronto Toronto, Canada.

Trueman, B. (1994). Analyst forecasts and herding behavior. *The Review of Financial Studies*, *7*(1), 97–124.

Ungar, L., Mellers, B., Satopää, V., Tetlock, P., & Baron, J. (2012). The good judgment project: A large scale test of different methods of combining expert predictions. Association for the Advancement of Artificial Intelligence Technical Report FS-12-06.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vives, X. (1993). How fast do rational agents learn? *Review of Economic Studies*, *60*(2), 329–347.

Wermers, R. (1999). Mutual fund herding and the impact on stock prices. *The Journal of Finance*, *54*(2), 581–622.

Williams, C. K., & Rasmussen, C. E. (2006). *vol. 2*, *Gaussian processes for machine learning*. (3), MIT press Cambridge, MA.

Witkowski, J., Atanasov, P., Ungar, L. H., & Krause, A. (2017). Proper proxy scoring rules. In *Thirty-first aaai conference on artificial intelligence*.

Witkowski, J., Freeman, R., Vaughan, J. W., Pennock, D. M., & Krause, A. (2022). Incentive-compatible forecasting competitions. *Management Science*.

Zwiebel, J. (1995). Corporate conservatism and relative compensation. *Journal of Political Economy*, *103*(1), 1–25.