Direct comparison of agent-based models of herding in financial markets

Sylvain Barde, Ofce Observatoire Français Des Conjonctures Économiques

Journal of Economic Dynamics & Control 73 (2016) 329–353

Contents lists available at ScienceDirect

Journal of Economic Dynamics & Control

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

Direct comparison of agent-based models of herding in financial markets

Sylvain Barde [a,b,*]

[a] School of Economics, Keynes College, University of Kent, Canterbury CT2 7NP, UK
[b] Observatoire Français des Conjonctures Economiques (affiliate), France

ARTICLE INFO

Article history:
Received 26 April 2016
Received in revised form
20 September 2016
Accepted 10 October 2016
Available online 19 October 2016

JEL:
C15
C52
G12

Keywords:
Model selection
Agent-based models
Herding behaviour

ABSTRACT

The present paper tests a new model comparison methodology by comparing multiple calibrations of three agent-based models of financial markets on the daily returns of 24 stock market indices and exchange rate series. The models chosen for this empirical application are the herding model of Gilli and Winker (2003), its asymmetric version by Alfarano et al. (2005) and the more recent model by Franke and Westerhoff (2011), which all share a common lineage to the herding model introduced by Kirman (1993). In addition, standard ARCH processes are included for each financial series to provide a benchmark for the explanatory power of the models. The methodology provides a consistent and statistically significant ranking of the three models. More importantly, it also reveals that the best performing model, Franke and Westerhoff, is generally not distinguishable from an ARCH-type process, suggesting their explanatory power on the data is similar.

1. Introduction

The emergence of agent-based modeling as an alternative to the more traditional fully rational representative agent approach has enabled the integration of many new mechanisms and behaviours into economic analysis. As pointed out by Tesfatsion (2006), such models allow for bounded rationality, learning, switching, etc. and typically offer great flexibility for investigating the emergence of aggregate equilibria from the interaction of often simple behaviours at the individual level. This increase in modeling flexibility has come at a cost, however. Even as the methodology increased in popularity over the last decade and a half, concerns were being voiced about the issue of calibrating and validating these models, as well as comparing their predictions to more traditional approaches. Durlauf (2005, p. F241) for instance, finding weaknesses in the existing empirical literature of the time relating to the analysis of complexity, warned that “it will not become a major component of economic reasoning until a tight connection between theoretical work and empirics is developed. Unless such a connection is achieved, even an open-minded complexity advocate will be justified in taking the Scottish legal option of concluding that the importance of complexity in understanding socioeconomic phenomena is ‘not proven’.” Fagiolo et al. (2007) and Dawid and Fagiolo (2008) also identify this issue of validation as the main open question facing the agent-based simulation community.

* Correspondence address: School of Economics, Keynes College, University of Kent, Canterbury CT2 7NP, UK.
E-mail address: s.barde@kent.ac.uk

http://dx.doi.org/10.1016/j.jedc.2016.10.005
0165-1889/

The first hurdle is the estimation of the parameters that govern the simulation from available data. This process is often complicated by the presence of nonlinearity and/or emergence of complexity from simple rules, which makes the inference of parameter values with traditional statistical methods difficult. Existing solutions to this problem rely on simulation methods such as simulated maximum likelihood (SML) or the method of simulated moments (MSM), both reviewed in Gouriéroux and Monfort (1993). More recently, Bianchi et al. (2007) advocate the use of the indirect inference approach of Gouriéroux and Monfort (1996), which generalises MSM by using a binding function rather than directly selecting the moments that need matching.

A second issue, the comparison of agent-based models against each other and against other approaches, still remains somewhat of a problem today. As pointed out in Hommes (2011, p. 2), one of the problems with agent-based models is the great number of degrees of freedom they offer for modeling agent behaviour, leading to a “wilderness” where a large number of models can coexist that all seem to replicate the stylised facts.[1] Addressing this issue requires not only estimation methods but also dedicated model selection/comparison methods. Using the estimation methods mentioned above to compare models is possible, as shown by the estimation in Winker et al. (2007) of the Gilli and Winker (2003) and Lux and Marchesi (2000) models and their comparison to a GARCH process. Nevertheless, doing so can be potentially problematic, as this often requires tailoring the model specification being estimated, thus making a direct comparison across specifications difficult. For example, Franke and Westerhoff (2012, p. 1208) argue that one advantage of using the MSM to estimate models it that “it is […] a very transparent method as it requires the researcher to make up his or her mind about the stylized facts that a model should be able to reproduce, and to set up the precise summary statistics (the moments) by which he or she wants to quantify them”. However, this requires deciding which moments to reproduce and complicates the problem of comparison across models that have already been calibrated in previous works. A simple illustration of this is that while two of the agent-based models of financial markets used in this paper, Gilli and Winker (2003) and Franke and Westerhoff (2016), are estimated using the MSM, the former uses two moments, the latter nine, none of which are the same.[2]

More recently, dedicated model comparison methodologies have been developed in order to provide more reliable tools for comparing and selecting amongst agent-based models. In particular, Barde (2016) and Lamperti (2015) argue that the comparison of simulation models is best carried out with standardised criteria based on accepted information-theoretical measures such as the Kullback and Leibler (1951) (KL) divergence between model and data. Their technical implementation differ, however, given a data set and a simulated series produced by a model, both produce an information criterion which scores the performance of the model on the data. The main advantage of these methods is that by relying on an information measure such as the KL divergence one no longer needs to select which moments to match. This is because in both cases the models are scored on the basis of their full, simulated, conditional densities. In addition to this, Barde (2016) possesses two key benefits. Firstly, in line with the arguments of Mandes and Winker (2015), it allows the comparison of the candidate models on the basis of their complexity, in addition to more standard goodness-of-fit considerations. This is done by measuring the stochastic complexity of the simulated data series, following the minimum description length principle of Grünewald (2007). The second benefit is that Barde (2016) scores each model at the level of individual empirical observations, which means that it integrates seamlessly with the model confidence set (MCS) approach developed by Hansen et al. (2011). The central implication is that not only can models be formally ranked in terms of their explanatory power on the data, this ranking can be tested statistically to determine the subset of best candidate models that cannot be rejected at a chosen confidence level.

This paper aims to run a full-scale model comparison test of Barde (2016) in order to evaluate the methodology's potential for model selection, illustrate its flexibility and provide further insights into how herding mechanisms explain the features of financial data. The basic setting is similar to the recent model contest of Franke and Westerhoff (2012), however the comparison exercise carried out here aims to go beyond that framework by integrating two key elements from Winker et al. (2007). The first is that the analysis will directly incorporate models calibrated by different authors, the second is that it will also include econometric ARCH/GARCH specifications to serve as a benchmark for explanatory power. The analysis focuses on agent-based models of herding in financial markets due to the desirable characteristics this setting displays. First of all, there is an established literature on herding mechanisms which crucially offers several recent models that have been calibrated but not yet systematically compared. The second desirable characteristic is that because these models typically attempt to explain financial market returns, they offer a univariate setting with plentiful data which simplifies the problem of comparison. Finally, this focus on daily financial returns also means that standard econometric models of conditional heteroscedasticity can be used to provide a reliable benchmark for comparison.

The three models selected for the comparison exercise are those of Gilli and Winker (2003), Alfarano et al. (2005) and Franke and Westerhoff (2011, 2016), which will be referred to in the rest of the paper as GW, ALW and FW respectively. The very thorough reviews of Hommes (2006) and Westerhoff (2009) show that the literature on agent-based models of financial markets is extensive and contains many different behavioural mechanisms. The first reason for this specific selection of models is that they have all been calibrated on empirical data using either SML or MSM. This means that the

[1] Hommes (2011) refers to 1000 papers over 20 years on learning and bounded rationality mechanisms alone, one can only suppose that this problem has since worsened.

[2] Gilli and Winker (2003) match the ARCH(1) parameter estimate and the kurtosis of the raw simulated returns, Franke and Westerhoff (2016) use the mean of the absolute returns, the first order autocorrelation of the raw returns, six lags of the autocorrelation function of the absolute returns and the Hill estimator of the tail index of the absolute returns.

robustness of each calibration can be evaluated and compared to the others. A second consideration is that all three share a common theoretical lineage with the herding mechanism initiated by Kirman (1993), which should hopefully make it more difficult to separate them empirically, thus offering the model comparison methodology a decent challenge.

The remainder of the paper is organised as follows. Section 2 starts by reviewing the three candidate models examined in the comparison exercise and presents their respective herding mechanisms. Section 3 then details the econometric benchmark, data and comparison protocol used to assess performance. The results of the comparison exercise are presented in Section 4 and discussed in Section 5, while Section 6 draws the main conclusions.

2. Agent-based models of herding in financial markets

The three models in the comparison exercise share a common lineage with Kirman (1993), which sets up a basic recruitment framework where two populations of agents coexist, and members of one category can recruit members from the other. The framework assumes a population of $N \in \mathbf{N}$ agents, divided into two strategy types: "fundamentalists" and "chartists". Describing the state of the system is simple: at any point in time, let n_t be the number of fundamentalist agents in the market, the remaining $N-n_t$ agents being the number of chartists. In the following discussion, it will be convenient to refer to $x_t = n_t/N$ as the fundamentalist share of the population, with $1-x_t = (N-n_t)/N$ as the share of chartists.

As pointed out in Kirman (1993), agents can change strategy over time, either spontaneously or because they are recruited by an agent using the other strategy. If ϵ is the probability of an agent spontaneously changing strategy and ρ the probability of a successful recruitment following an encounter between agents using two different strategies, then the dynamic evolution of the system is governed by the following transition probabilities, where superscripts *fc* and *cf* respectively indicate the case where a fundamentalist agent switches to chartist strategies and the reverse case where a chartist becomes a fundamentalist:

$$\begin{cases} P_t^{cf} = (1-x_t)(\epsilon+\rho x_t) \\ P_t^{fc} = x_t(\epsilon+\rho(1-x_t)) \end{cases} \tag{1}$$

It is important to point out that the notation used below has been harmonised and is somewhat different from that used in each of the three papers. This has been done in order to facilitate the exposition of the mechanisms and their comparison across models.

2.1. The Gilli and Winker (2003) model of herding

The Gilli and Winker (2003) model follows the literal interaction mechanism described by Kirman (1993) to produce a system whose time-evolution is governed by the transition probabilities (1). Essentially, it directly simulates the interactions of a population of N agents: at each point in time, three agents are randomly selected from the population, with the first convincing the second to switch to his strategy with probability ρ and the third spontaneously switching strategy with probability ϵ.

An important aspect of the model is that the current population share of fundamentalists is imperfectly evaluated by agents, who receive a signal $\tilde{x}_t \sim \mathrm{N}(x_t, \sigma_x^2)$ containing a measurement error. This reflects the fact that the beliefs of traders and the strategies they use are likely to be private information and leads to the following expected population share:

$$\omega_t = P\left(\tilde{x}_t > \frac{1}{2}\right) \tag{2}$$

Chartist and fundamentalist agents differ in the way they form price expectations, and therefore in their excess demand functions. Fundamentalists expect future prices to correct to their fundamental value $\overline{p}$ at a given rate ϕ, while chartists simply extrapolate from past price movements:

$$\begin{cases} E^f(\Delta p_t) = d_t^f = \phi(\overline{p}-p_{t-1}) \\ E^c(\Delta p_t) = d_t^c = p_{t-1}-p_{t-2} \end{cases} \tag{3}$$

Combining the expected share (2) with the excess demands (3) and adding an exogenous perturbation $u_t \sim \mathrm{N}(0, \sigma_s^2)$ provides the equation determining the evolution of the price at each point in time:

$$p_t = p_{t-1} + \omega_t \phi(\overline{p}-p_{t-1}) + (1-\omega_t)(p_{t-1}-p_{t-2}) + u_t \tag{4}$$

Gilli and Winker (2003) intend this process to describe the evolution of the price p_t and agents share x_t at the time scale of individual interactions. In order to produce a daily series, which is the typical time frequency used for empirical applications, one must specify a parameter τ for the number of interactions per trading day, and sample each τth observation from the raw interaction-level series (4).

2.2. The Alfarano et al. (2005) model of asymmetric herding

The Alfarano et al. (2005) model of herding similarly embeds the Kirman (1993) mechanism, but describes the time evolution of the state x_t directly from the transition probabilities rather than simulating the agent-level interactions. The transition probabilities for their model are given by:

$$\begin{cases} P_t^{cf} = (N-n_t)(\epsilon_1+n_t)\rho \\ P_t^{fc} = n_t(\epsilon_2+(N-n_t))\rho \end{cases} \tag{5}$$

While slightly different in appearance, this system is nevertheless broadly equivalent to (1), as redefining $\rho=\rho'/N^2, \epsilon_* = (\epsilon'_* N)/\rho'$ and setting $x_t = n_t/N$ recovers the specification of the Kirman (1993) transition probabilities.[3] The first difference with Kirman (1993) and Gilli and Winker (2003) is the fact that Alfarano et al. (2005) allow for different autonomous probabilities of switching, governed by ϵ_1 and ϵ_2, which may not be equal to each other, allowing for asymmetry in the herding mechanism.

The second difference is that rather than simulating the agent interactions, Alfarano et al. (2005) provide an analytical solution to the time evolution of the system by solving the Fokker–Plank approximation in continuous time to obtain the Master equation generated by the transition probabilities (5) for large N. This results in the following time evolution of the population share for an arbitrary time increment Δt:

$$x_{t+\Delta t} = x_t + \rho(\epsilon_1+\epsilon_2)(\overline{x}-x_t)\Delta t + \lambda_t\sqrt{2\rho(1-x_t)x_t\Delta t} \tag{6}$$

The drift term of this time evolution depends on $\overline{x} = \epsilon_1/(\epsilon_1+\epsilon_2)$, which is the mean population share of fundamentalists over time, while the second part is a diffusion term determined by λ_t, which follows an i.i.d. standard normal distribution.

As is the case in Gilli and Winker (2003), the two types of agents differ in their demand functions. Fundamentalists are defined similarly as expecting log prices p_t to revert to their fundamental level $\overline{p}$. Chartists, on the other hand, are essentially noise traders whose demands are determined by a random variable η_t, which is uniformly distributed over $[-1,1]$ and a scaling parameter r_0 which governs the expected size of the price fluctuations.[4] Given population sizes n_t and $N-n_t$, the excess demands are given by:

$$\begin{cases} d_t^f = n_t(\overline{p}-p_t) \\ d_t^c = (N-n_t)r_0\eta_t \end{cases} \tag{7}$$

Setting the sum of excess demands (7) equal to zero directly leads to the following expression for the value of the log returns r_t:

$$r_t = r_0\frac{x_t}{1-x_t}\eta_t \tag{8}$$

This results in a very elegant and parsimonious model, which only requires the two autonomous switching parameters ϵ_1, ϵ_2 and the direct recruitment parameter ρ. Returns can be simulated by drawing a set of standard normal variables λ_t and a set of uniformly distributed variables η_t and using them in Eqs. (6) and (8) with $\Delta t = 1$.

2.3. The Franke and Westerhoff (2011) structural stochastic volatility model

The model proposed by Franke and Westerhoff (2011) also uses the basic herding mechanism of Kirman (1993), but expresses that the state variable slightly differently. The population state is defined as $x'_t = (2n_t - N)/N$, leading to $x'_t = -1$ if all the population is chartist ($n_t = 0$) and $x_t{}' = 1$ if all the population is fundamentalist ($n_t = N$).[5] This is done to facilitate the exposition of the herding mechanism in the transition probabilities, which relies on the exponential of a switching propensity s_t:

$$\begin{cases} P_t^{cf} = \nu\exp(s_t) \\ P_t^{fc} = \nu\exp(-s_t) \end{cases} \tag{9}$$

The propensity to switch s_t is determined by several factors. The first is an exogenous effect α_0, which aims to capture the existence of autonomous switching, similar to the ϵ parameter of the previous models. The second term, which depends on the population state $x_t{}'$, encapsulates the herding concept, increasing the probability of switching to a strategy based on the

[3] Should one try to perform this reparametrisation with the values shown in Table 1 however, one would find they do not agree across models. This is because of the different time scales involved: the Gilli and Winker (2003) parameters are calibrated for τ transitions per daily return, while the Alfarano et al. (2005) model parameters embed a single transition per daily return.

[4] Alfarano et al. (2005) offer two options for the chartist noise specification, "spin" noise, which takes values $\{-1,+1\}$ with equal probability, and uniform noise, which is used here. They also show that choosing a scaling parameter $r_0 = (\epsilon_2-1)/\epsilon_1$ results in a unit-variance daily returns series.

[5] The x' notation is used to emphasise this difference from the other models. Setting $x_t = (1+x_t{}')/2$ in the model equations recovers the standard share variable $x_t = n_t/N$ used in the previous two models.

current popularity of that strategy. Should the two populations be balanced ($n_t = N/2$), one has $x_t' = 0$ and there is no herding effect. The final term, which depends on the squared deviation of the log price p_t from its fundamental value $\overline{p}$, is designed to encourage switching to fundamentalism when the price deviates from the fundamental value. Given that such deviations tend to occur mainly when a large share of the population uses chartist strategies, this feedback mechanism will generate asymmetry in the switching process:

$$s_t = \alpha_0 + \alpha_x x_{t-1}' + \alpha_m \left(p_{t-1} - \overline{p}\right)^2 \tag{10}$$

The transition probabilities (9) lead to the following population dynamics for the model:

$$x_t' = x_{t-1}' + (1 - x_{t-1}')P^{cf}_{t-1} - (1 + x_{t-1}')P^{fc}_{t-1} \tag{11}$$

The excess demand functions of the fundamentalists and chartists, below, mirror (3) as used by Gilli and Winker (2003), with two exceptions. First of all, the price expectations of chartists now also have an adjustment parameter χ, similar to the ϕ controlling the fundamentalist adjustment. Secondly, both excess demands now incorporate a noise component $u_t^f \sim \mathrm{N}\left(0, \sigma_f^2\right)$ and $u_t^c \sim \mathrm{N}(0, \sigma_c^2)$:

$$\begin{cases} d_t^f = \phi(\overline{p} - p_t) + u_t^f \\ d_t^c = \chi(p_t - p_{t-1}) + u_t^c \end{cases} \tag{12}$$

Given the evolution of the population index (11) and the demand functions (12), the time evolution of log price is described by the following equation:

$$p_t = p_{t-1} + \mu\left(\frac{(1 + x_{t-1}')}{2}\phi(\overline{p} - p_t) + \frac{(1 - x_{t-1}')}{2}\chi(p_t - p_{t-1}) + u_t\right) \tag{13}$$

The noise term $u_t \sim \mathrm{N}(0, \sigma_t^2)$ forms the structural stochastic volatility component of the model, as the variance of this noise is governed by the population-weighted average of the fundamentalist and chartist noise terms:

$$\sigma_t^2 = \tfrac{1}{2}\left((1 + x_t')^2\sigma_f^2 + (1 - x_t')^2\sigma_c^2\right) \tag{14}$$

3. The model comparison protocol

3.1. The ARCH family benchmark

As stated previously, a set of ARCH models is included as part of the model comparison exercise. The purpose of this is twofold: firstly, to provide a reliable benchmark for the explanatory power of the agent-based models and secondly, to demonstrate the ability of the methodology described in Section 3.2 to cope with a wide range of modeling approaches, from agent-based simulations to more traditional regression methods.

Because the aim of the analysis is to compare model performance across a wide set of empirical data series, it is important that this benchmark set be standardised in order to allow such a comparison. This means that although their value will change from series to series, the number of estimated parameters (and therefore lags) in a given specification will be the same for all data series. Similarly, given that each specification is estimated on all 24 data series, the lag order is always set to $p = q = r$, thus ensuring that the total number of estimations to be carried out remains tractable.

All the ARCH-type models in the benchmark set have the same AR(2) mean equation (15) for the daily returns r_t, and only differ in the specification of the time-varying volatility σ_t in the error term $\varepsilon_t = \sigma_t z_t$, where z_t is a standard normal variable. While the second AR lag is insignificant for most series, it is significant at the 5% level for some and is therefore included in the standard specification.[6] A third AR lag was tested for in exploratory estimations of the mean equation but was insignificant for all series, and therefore not included:

$$r_t = c + a_1 r_{t-1} + a_2 r_{t-2} + \sigma_t z_t \tag{15}$$

Several specifications are included for the time-varying variance σ_t, in order to provide as wide a target as possible for the comparison exercise. The first and most basic specification is the ARCH model (16). A single version is included with $p = 5$ lags, corresponding to the length of the average working week. The AIC and BIC results from the estimations show that this naive specification performs poorly for most of the data series compared to the more refined specifications below. The intention, however, was to provide some "low hanging fruit" for the agent-based models in the comparison exercise:

$$\sigma_t^2 = \sigma_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 \tag{16}$$

[6] This is most notably the case for the IPC, NASDAQ and DJ indices. For the latter, this can be seen in Table A6 in the appendix.

The second basic specification included is a set of three standard GARCH models (17), with lags $p=q\in\{1,2,3\}$. A maximum of 3 (p,q) lags are used, as preliminary analysis revealed that the lowest BIC values are reached with this range. Setting the lag order to 4 or higher lead to systematic increases in the BIC for every series. As for the ARCH specification, this is not expected to provide the best specification for the daily returns series, but instead to provide a reasonable target for the agent-based models:

$$\sigma_t^2=\sigma_0+\sum_{i=1}^{p}\alpha_i\varepsilon_{t-i}^2+\sum_{j=1}^{q}\beta_j\sigma_{t-j}^2 \tag{17}$$

An important consideration in choosing the benchmark specifications is that both the ALW and FW models allow for asymmetry in herding. The preferred ARCH family specifications are therefore the three following models, which all include asymmetry terms γ_k allowing positive and negative lags of the error term to have different effects on the volatility. These are the threshold GARCH (TGARCH) specification (18), with negative lags identified by the indicator variable I_{t-k}, the exponential GARCH (EGARCH) specification (19) and finally the power GARCH (PGARCH) specification (20):

$$\sigma_t^2=\sigma_0+\sum_{i=1}^{p}\alpha_i\varepsilon_{t-i}^2+\sum_{j=1}^{q}\beta_j\sigma_{t-j}^2+\sum_{k=1}^{r}\gamma_k I_{t-k}\varepsilon_{t-k}^2 \tag{18}$$

$$\ln(\sigma_t^2)=\sigma_0+\sum_{i=1}^{p}\alpha_i\left|\frac{\varepsilon_{t-i}}{\sigma_{t-i}}\right|+\sum_{j=1}^{q}\beta_j\ln\left(\sigma_{t-j}^2\right)+\sum_{k=1}^{r}\gamma_k\frac{\varepsilon_{t-k}}{\sigma_{t-k}} \tag{19}$$

$$\sigma_t^\delta=\sigma_0+\sum_{i=1}^{p}\alpha_i\left(|\varepsilon_{t-i}|-\gamma_i\varepsilon_{t-i}\right)^\delta+\sum_{j=1}^{q}\beta_j\sigma_{t-j}^\delta \tag{20}$$

As for the GARCH models (17) , three versions of specifications (18), (19) and (20) are estimated, with lags $p=q=r\in\{1,2,3\}$. As for the GARCH case, setting the (p,q,r) lag order above 3 does not lead to an improvement in the BIC values (with the exception of the Hang Seng and Straight Times indices), and in several cases leads instead to convergence failures for the PGARCH estimates.

Combined with the ARCH (16) and the three GARCH (17) specifications, this results in 13 ARCH family models for each data series, which were estimated in Eviews 9 using the "legacy" option. As shown in Table A4, the estimates converged rapidly in the vast majority of cases, and only a few PGARCH and TGARCH estimations failed to achieve convergence. Three examples of the estimation results are provided in Appendix A as an illustration, in Tables A5 and A6 for the DAX and DJ indices respectively and Table A7 for the US dollar to Yen exchange rate.[7]

3.2. The model comparison methodology

Before discussing the data used to evaluate these agent-based models of herding, it is important to review briefly the main aspects of the methodology that will be used for the model comparison exercise, as the purpose of the paper is as much to evaluate the methodology as it is to evaluate the models themselves. The implementation details of the methodology and a proof-of-concept are provided in Barde (2016) and its supplementary material.

The general spirit of the methodology is to map the data-generating processes of a set of candidate models $\{M_1,M_2,\ldots,M_m\}$ to a corresponding set of standardised Markov processes (or equivalently finite state machines). Let us assume for the moment that a discrete random variable Y_t describes the time evolution of a system, and that the formal structure of a model M_i enables the researcher to calculate the following conditional probabilities for every possible history of the system $y_{t-1},y_{t-2},\ldots,y_{t-L}$. This full set of conditional probabilities forms the transition matrix of the Lth order Markov process underlying M_i:

$$P_{M_i}(Y_t|y_{t-1},y_{t-2},\ldots,y_{t-L}) \tag{21}$$

If a data series $\{y_1,y_2,\ldots y_N\}$ is available, the researcher can very easily obtain a score for each observation by taking the logarithm of the reciprocal model probabilities (21) for the state configuration $\{y_{t-L},\ldots,y_{t-2},y_{t-1},y_t\}$ corresponding to each observation:

$$\lambda_i(y_t)=\ln\frac{1}{P_{M_i}(Y_t=y_t|y_{t-1},y_{t-2},\ldots,y_{t-L})} \tag{22}$$

Barde (2016) shows that the mean value of the observation-level score (22) is an estimate of cross entropy of the real data with the model M_i, providing a Markov information criterion (MIC) similar in spirit to the Akaike information criterion

[7] The full set of 312 estimates corresponding to the 13 specifications on the 24 data series is not included here in the interest of brevity, however it is available as supplementary material to the paper.

or to a log-likelihood:

$$\text{MIC}_i = \frac{1}{N-L}\sum_{t=L+1}^{N}\lambda_i(y_t) \tag{23}$$

Differences in (23) across models M_i, M_j directly reflect differences in the KL divergences between the models and data, with the best model identified as the one with the lowest score, or equivalently (taking the negative) the highest log-likelihood. Thus, while the method used to obtain the measurement (23) might be new or unfamiliar, the nature of the measurement itself should not be.

The technical challenge resides in efficiently mapping the simulated data produced by the set of models $\{M_1, M_2, \ldots, M_m\}$ to their underlying Markov process, i.e. efficiently obtaining the conditional probabilities (21) from the simulated data produced by each model M_i. This is achieved by relying on a universal data compression algorithm, specifically the context tree weighting (CTW) algorithm of Willems et al. (1995), which is designed to determine the Markov transition matrix of a data generating process directly from the data. The central justification for choosing this technique is that the CTW algorithm's mapping of the data to the transition matrix is optimal on all Markov processes of arbitrary order. Specifically, this means that the inefficiency cost incurred by having to learn the transition matrix from the data is proven to achieve the theoretical lower bound. As shown by Barde (2016), this central property, referred to as *universality*, allows for the correction of the resulting measurement error in the observation score (22) and justifies this choice of the algorithm as the basis of the model comparison methodology.

The CTW algorithm's proven optimal performance stems from the fact that it operates on a binary representation of the data series $\{y_1, y_2, \ldots y_N\}$, where each observation is treated as the result of a set of Bernoulli trials. The only variables that need to be estimated are the set of Bernoulli parameters that determine the probability of a given observation bit being "1" conditional on a particular system history. The CTW algorithm obtains these using the Krichevsky and Trofimov (1981) estimator, which is proven to possess the tightest possible bound on its inefficiency.

Converting real-valued data to its binary representation requires a specific discretisation strategy. Given a choice of bounds $[b_l, b_u]$ and resolution r, the real-valued observations are binned into 2^r distinct states spanning the support determined by the bounds. Each bin (or state) is identified with a distinct r-bit representation where each 1/0 value indicates if the observation is in the top/bottom half of the subset of the support determined by the previous bits.[8] Given an additional choice of L lags of time dependence, this means that the CTW algorithm produces a standardised transition matrix of size $2^{rL} \times 2^r$ for each model M_i.

The crucial benefit of this binary representation is that even large state spaces, with relatively high values of the resolution r, can be represented as a sequence of chained Bernoulli trials. In this setting the probability of an observation being in any of the 2^r states can be simply reconstructed by chaining the probabilities that each successive trial results in the value given by the r-bit representation. Similarly, the score for a given observation (22) is simply the sum of the binary log scores for each of the r bits that make up the observation. This produces an observation-level vector of scores, which sums up to the aggregate score for the model, as can be seen from (23).

The availability of a vector of observation-level scores (22) has two crucial benefits compared to alternative methods of evaluating models. The first is the ability to use the variance in scores at the observation level to test the statistical significance of the aggregate criterion (23) in any model comparison exercise, using the data snooping procedure of White (2000) or the model confidence set of Hansen et al. (2011). The second benefit is the ability to evaluate the relative explanatory power of models over subsets of the data. Both these aspects are illustrated in the comparison exercise.

3.3. The stock market index data and comparison protocol

The data used for the model comparison exercise are the daily logarithmic returns for a set of 18 stock market indices and 6 exchange rates series.[9] The stock market data covers the major time-zones of Asia, Europe and the Americas, with 6 indices selected from each of these zones. Furthermore, most series consist of over five thousand daily observations since the mid-1980s, capturing key events such as the 1987 Black Monday crash, the Asian crisis of the late-1990s, the dot-com bubble of the early 2000s up to the turmoil following the fall of Lehman Brothers in late 2008. In addition to this, because both the Gilli and Winker (2003) and Franke and Westerhoff (2011) studies estimate their models on the US/Deutschmark exchange rate as well as stock market data, the data set also includes the exchange rate series for the US dollar against six major currencies. Because, as pointed out by Andersen et al. (2000), exchange rate returns are more symmetric than stock market returns, this will also provide a test of the asymmetric herding mechanisms in the ALW and FW models.[10]

This wide geographical selection and long time period is intended to provide a broad test of the explanatory power of the agent based models of herding by enabling evaluation of the models both at the aggregate level as well as on individual

[8] As an example, a resolution $r=3$ indicates observations can take 8 distinct values. Given an observation of "101", the first bit indicates the observation is in bins 5, 6, 7 or 8, the second indicates that the observation is in either the 5th or 6th bin, and the final bit determines that the observation is in the 6th bin.

[9] The stock market indices used here are publicly available from the historical prices section of http://finance.yahoo.com while the exchange rate data was taken from the Federal Reserve historical foreign exchange rate releases (H.10) at www.federalreserve.gov/econresdata.

[10] The author is grateful to one of the referees for suggesting the use of exchange rate data as a way of checking the importance of asymmetry.

Table 1
Calibrated values of model parameters.

Param.	Interpretation	Value	Bounds	
Gilli and Winker (GW) – US → DM exchange rate				
N	Number of agents	100*	–	–
τ	Number of interactions per trading day	50*	–	–
ϕ	Adjustment speed in fundamentalist expectations	0.0225*	–	–
σ_s	Standard deviation of price shocks	0.25*	–	–
σ_x	Standard deviation noise in majority assessment	0.219	0.05	0.35
ϵ	Probability of random switch	0.0001	0	0.0002
ρ	Probability of direct recruitment	0.264	0.05	0.45
Alfarano, Lux and Wagner (ALW) – DAX index				
ϵ_1	Propensity for fundamentalist → chartist switch	16	2	18
ϵ_2	Propensity for chartist → fundamentalist switch	4.9	2	18
ρ	Herding tendency	0.0025	0.001	0.004
Franke and Westerhoff (FW) – S&P 500				
ϕ	Aggressiveness of fundamentalists	0.198	0.05	0.35
χ	Aggressiveness of chartists	2.263	0.1	4.5
σ_f	Noise in fundamentalist demand	0.782	0.1	1.5
σ_c	Noise in chartist demand	1.851	1	6
μ	Market impact factor of demand	0.01*	–	–
p^*	Log of fundamental value	0*	–	–
ν	Flexibility in population dynamics	0.05*	–	–
α_0	Predisposition parameter	−0.155	−0.25	−0.05
α_x	Herding parameter	1.299	0.3	2.3
α_m	Misalignment parameter	12.648	10	15

Gilli and Winker (2003) is calibrated on the DM/US-$ exchange rate.
Alfarano et al. (2005) is calibrated on the DAX index.
Franke and Westerhoff (2016) is calibrated on the S&P500 index.
* The parameter value is assumed by the original authors, not calibrated.

events. Tables A1 and A3 in the appendix provides greater detail such as the starting date, the number of observations and the results of the diagnostic tests for each data series.

Because the model comparison methodology operates on discrete variables, the raw logarithmic returns are discretised to an 8-bit resolution, i.e. grouped in 256 discrete bins, within the bounds $[-0.3, 0.3]$. Any observations falling outside of those bounds are truncated to the bound itself, however as seen from the 5th column of Tables A1 and A3, there is only a single out-of-bound observation, for the Hang Seng index.[11] The more important aspect is the choice of resolution for the data, i.e. 8 bits. While the discretisation of the returns is required by the methodology, the procedure inevitably discards information and it is important to ensure that this does not affect the measurement. As explained in Barde (2016), the resolution r should be large enough to ensure that the discretisation error is i.i.d. uniform and uncorrelated with the discretised variable. When this is the case, any extra bit of resolution will take value 0 or 1 with equal probability 0.5, regardless of any conditioning on the past values of the variable. At that point, a larger choice of resolution r will simply increase the resulting information criterion by a constant for all models in the comparison set and will therefore not affect comparisons made by using differences in the information criterion across models.[12]

The discretisation diagnostics are reported in the last 3 columns of Tables A1 and A3 and show that the 8-bit discretisation of the data is sufficient for most series. Uniformity of the discretisation error is rejected for the HS, AEX and S&P500 indices, but this is due to the presence of a relatively large number of zero returns (144, 180 and 207 respectively) created by reported closing index values that are unchanged over two or more days. These exact zeros create a systematic spike in the discretisation error, leading to the rejection of uniformity, but are not a major concern for the methodology. The only slight concern is for the DJ industrial index, for which the autocorrelation in the error term cannot be rejected. The discretisation error, however, seems uniform and uncorrelated with the discretised variable, which suggests that the problem is not critical.

The model comparison exercise is based on a sensitivity analysis of the GW, ALW and FW models around their calibrated parameter values, shown in the third column of Table 1. This is motivated by two considerations: first of all, it is necessary to allow flexibility in the parameter set, as the calibrated values in Table 1 were each obtained on a specific data series and will not be optimal over all 24 series. Secondly, using a sensitivity analysis enables an evaluation of whether the MIC methodology presented above can effectively pin down the best parameter values for the three models over the data series.

[11] This corresponds to the Black Monday crash of 19/10/1987, where the logarithmic return for the day was -0.405. This is truncated to the -0.3 bound for the purposes of the analysis.

[12] As a robustness check the analysis was also carried out using a lower resolution $r = 7$ on the $[-0.15, 0.15]$ interval. This setting produces essentially the same results, which are also available in the supplementary material.

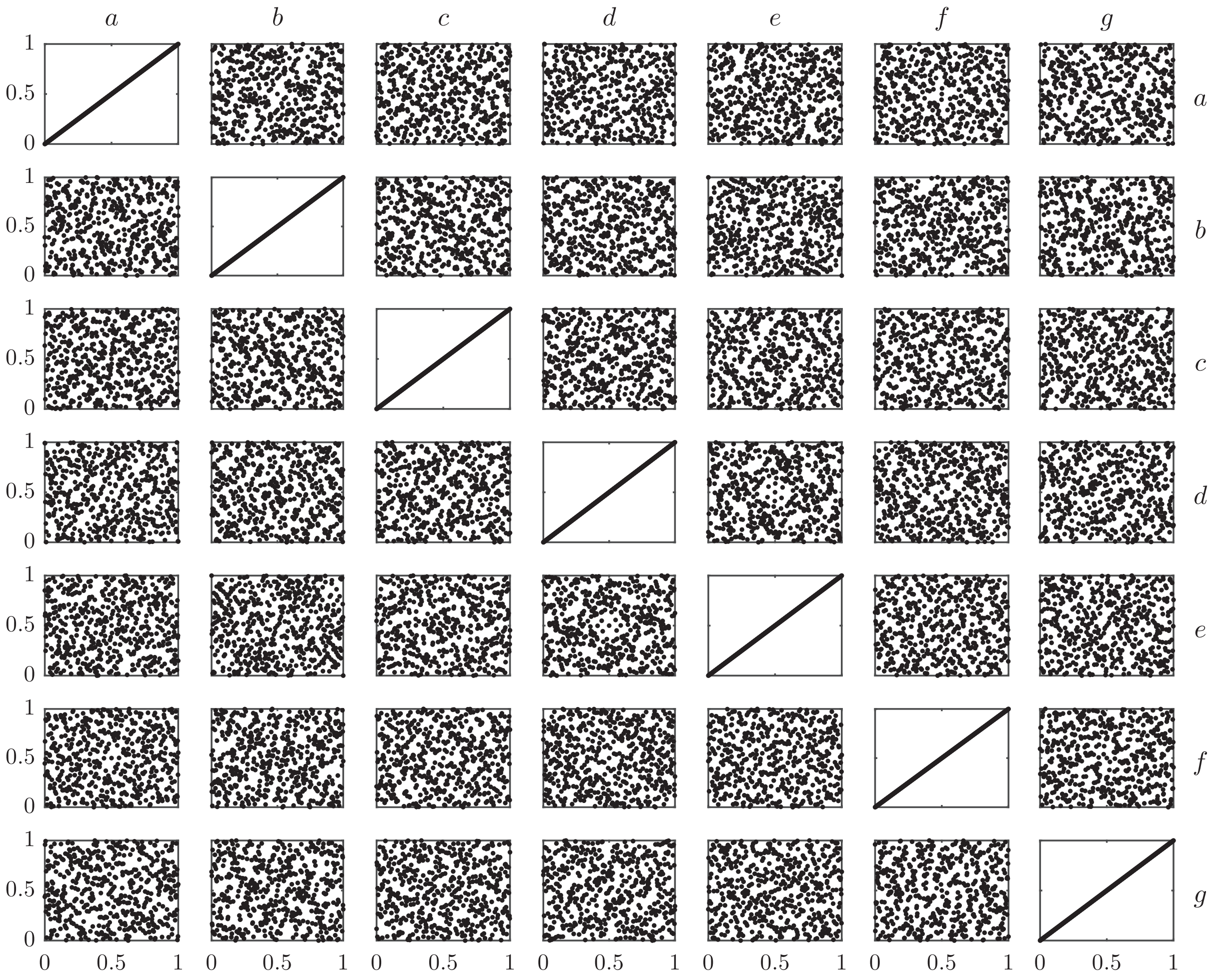

Fig. 1. Two-way scatter plot of the 513×7 NOLH sampling matrix.

In order to ensure that the sensitivity analysis is as efficient as possible, we follow the suggestion of Salle and Yildizoğlu (2014) and use the Nearly Orthogonal Latin Hypercube (NOLH) approach of Cioppa and Lucas (2007) to generate 513 sample points in the unit hypercube. A two-way scatter plot of the resulting 7-parameter sampling matrix is provided in Fig. 1. The NOLH approach is chosen due to the two key benefits it possesses. First of all, because the resulting sample forms a Latin hypercube in the parameter space, every sample point possesses unique values in each dimension. This ensures that the design possesses both a high resolution and good space-filling properties, as is visible in Fig. 1. The second key property is that the 7 resulting parameter vectors (labelled a–g in Fig. 1) are nearly orthogonal to each other. As pointed out by Salle and Yildizoğlu (2014), combined with the space-filling characteristics, this enables an increase in efficiency of the design by reducing the number of sampling points required to identify the relationships between the input parameters and the model outputs. Finally, given that the number of sample points in the NOLH design is equal to 2^n+1 for integer n, the choice of a sample size of 513 is a compromise between providing good coverage of the parameter space and keeping the overall analysis tractable.

In order to provide model-specific parameter samples, the columns in the sampling matrix are mapped from the $[0, 1]$ interval to the parameter ranges in the last two columns of Table 1. For the GW and ALW models, only 3 parameters are required therefore only columns a, b and c are used, while for FW the full sampling matrix is used.[13] Similarly, 13 simulated series are generated for the ARCH benchmark set using the specifications (15)–(20) and the parameter estimates corresponding to each data series. This implies that the simulated data for the set of ARCH benchmarks is specific to each of the 24 data series, which is not the case of the agent-based models, where the parameterisation resulting from the combinations in Table 1 are the same for all data series.

[13] While Table 1 shows that the GW model has 7 parameters, the first 4 are assumed by the authors and not calibrated. The same procedure is used in this analysis.

Table 2
MIC scores on financial data series.

Index	GW		ALW		FW		ARCH	
	min(MIC)	id	min(MIC)	id	min(MIC)	id	min(MIC)	id
AOI	4.3077	11	3.9255	112	**3.8885**	432	**3.8829***	5
NIKKEI	4.7543	248	4.6321	220	**4.5294**	470	**4.5200***	7
KOSPI	4.7947	190	4.6055	54	**4.5226**	279	**4.5193***	4
ST	4.5352	75	4.2617	54	**4.2089***	88	**4.2156**	7
HS	4.9898	327	4.8677	54	**4.7087**	284	**4.7031***	5
NIFTY	5.0210	327	4.8851	494	**4.7555**	284	**4.7367***	6
DAX	4.7431	248	4.6003	75	**4.4925***	470	**4.4953**	1
CAC	4.7528	248	4.6438	75	**4.5142**	470	**4.5062***	6
FTSE	4.4579	248	4.1605	112	**4.1084***	320	**4.1095**	2
IBEX	4.7786	453	4.6277	75	**4.5219**	470	**4.5179***	4
AEX	4.6156	75	4.3607	112	**4.2935***	288	**4.2969**	3
STOXX	4.6194	11	4.4183	54	**4.3353***	243	**4.3422**	5
IPC	4.8886	190	4.7224	112	**4.5968***	284	**4.6073**	5
DJ	4.4018	11	4.0871	112	**4.0270***	246	**4.0345**	5
S&P 500	4.4536	248	4.1685	112	**4.0839***	320	**4.0996**	3
NASDAQ	4.9563	418	4.8321	54	**4.7014**	284	**4.6957***	12
OEX	4.4758	248	4.2270	112	**4.1501**	320	**4.1495***	6
GSPTSE	4.2438	11	3.8396	220	**3.8059**	432	**3.7958***	5
USD → GBP	3.7749	11	3.4136	315	**3.3139***	37	**3.3231**	7
USD → EUR	3.9183	11	3.5638	124	**3.4647**	353	**3.4580***	7
USD → YEN	3.9087	11	3.5542	220	**3.4529***	109	3.4676	5
USD → CHF	4.0319	11	3.6552	120	**3.5588***	28	3.5819	1
USD → AUD	4.1675	11	3.8178	177	**3.7515***	145	**3.7557**	4
USD → MXN	3.6602	11	3.3549	104	**3.3331**	37	**3.3160***	6

"id" identifies the sample in either the NOLH sampling matrix or the ARCH benchmark set.
Bold indicates that the best model is in the model confidence set at the 90% level.
* The best overall score on each series.

With 513 candidate parameterisations for the GW, ALW and FW models and an additional 13 ARCH benchmark models, this results in a total of 1552 candidate models for each data series. Each of the candidate models is used to produce a simulated series with 500,000 observations, which is discretised to an 8-bit resolution on the $[-0.3, 0.3]$ support in accordance with the discretisation tests run on the data series and mentioned above. In the first stage of the methodology, these discretised series are processed by the CTW algorithm using $L=3$ lags of memory to recover their Markov transition matrix, which scored against the 24 data series in the second stage of the methodology.

4. Results

The combination of NOLH sampling, MIC scoring and MCS testing in the design is intended to offer a large degree of versatility for the analysis. This framework enables a ranking of the classes of models being compared both at the local and aggregate levels, as well as an analysis of parameter sensitivity within each class of model. All three types of analysis can be tested statistically and are illustrated here.

4.1. Relative aggregate performance of model classes

The aggregate MIC (23) is simply the mean of the observation-level vector of scores (22) obtained for each of the 1552 candidate models on the 24 series. Table 2 displays the identifier and value of the smallest MIC score for each data series and model class. The main finding, which is consistent across all series, is that the best GW calibration displays the highest score and is systematically beaten by the best ALW model, which in turn is systematically outperformed by the best FW calibration. Interestingly, the results also reveal that the latter model is comparable to the best ARCH-type model in terms of overall explanatory power. Because the parameters used for the ARCH simulations, obtained by estimation, are specific to each series while the parameter values used for the three sets of agent-based simulations are fixed *ex ante*, it was reasonable to expect the ARCH benchmark to outperform the agent-based models. It is therefore interesting to note that despite this potential handicap the best FW calibration approaches, and in half the cases exceed, the performance of the ARCH benchmarks. Even more significantly, looking at the exchange rate series suggests that the performance of the FW model relative to the ARCH models is even better, as it performs best on four of the six series.

As explained in Section 3.2, the fact that the methodology returns an observation-level vector of scores (22) can be used to test the statistical significance of the relative scores in Table 2. This is done by running the MCS procedure of Hansen et al.

Table 3
Size of model confidence set per class of model.

Index	$\lvert\mathcal{M}_{90}\rvert$	GW	ALW	FW	ARCH
AOI	101	0	0	91	10
NIKKEI	55	0	0	47	8
KOSPI	81	0	0	74	7
ST	91	0	0	83	8
HS	18	0	0	12	6
NIFTY	18	0	0	13	5
DAX	99	0	0	91	8
CAC	72	0	0	60	12
FTSE	127	0	0	114	13
IBEX	77	0	0	66	11
AEX	131	0	0	121	10
STOXX	98	0	0	89	9
IPC	45	0	0	38	7
DJ	64	0	0	56	8
S&P 500	109	0	0	107	2
NASDAQ	74	0	0	64	10
OEX	109	0	0	100	9
GSPTSE	48	0	0	44	4
USD → GBP	104	0	0	95	9
USD → EUR	103	0	0	96	7
USD → YEN	27	0	0	27	0
USD → CHF	52	0	0	52	0
USD → AUD	199	0	0	191	8
USD → MXN	48	0	0	37	11
N° of models:	1552	513	513	513	13

Table 4
MCS model parameters.

Gilli and Winker (GW)						
σ_x	0.298	**0.200**	0.308	0.316	0.330	0.326
ρ	0.069	**0.250**	0.056	0.083	0.114	0.051
ϵ	1.953e−4	**1.000e−4**	1.844e−4	1.586e−4	1.695e−4	6.133e−5
id	11	**65**	75	190	248	327
N°	10	**0**	2	2	6	2
Alfarano, Lux and Wagner (ALW)						
ϵ_1	15.750	2.250	17.375	**16.375**	15.813	
ϵ_2	3.875	16.750	5.375	**4.500**	4.188	
ρ	3.906e−3	3.578e−3	3.953e−3	**3.039e−3**	3.637e−3	
id	54	75	112	**124**	220	
N°	5	3	7	1	3	
Franke and Westerhoff (FW)						
ϕ	0.066	0.347	0.321	0.285	**0.257**	0.287
χ	1.509	1.312	3.804	3.752	**2.377**	4.027
σ_f	0.111	1.104	0.841	0.349	**0.901**	1.065
σ_c	4.867	5.990	5.639	4.721	**2.006**	5.580
α_0	−0.102	−0.239	−0.099	−0.117	**−0.137**	−0.213
α_x	1.222	0.538	1.132	1.218	**1.124**	0.359
α_m	12.617	11.533	11.143	10.420	**11.943**	14.893
id	37	284	320	432	**451**	470
N°	2	4	3	2	**0**	4

Bold indicates the model id corresponding to the calibrations of the original works.
"N°" counts a given "id" that occurs in Table 2.

(2011) on each data series. Starting with the full set of 1552 candidate models, the procedure identifies a subset $\mathcal{M}_\alpha$ composed of the best-performing models whose aggregate scores cannot be distinguished at a given level of statistical significance $1-\alpha$.[14] The results, shown in Table 3, confirm that none of the ALW and GW calibrations make it into the aggregate confidence set at the 90% confidence level, which is restricted to a subset of the FW calibrations and the ARCH benchmarks only. More importantly, it also confirms that none of the ARCH models are included in the confidence set for

[14] The MCS analysis carried out here used 1000 replications of the Politis and Romano (1994) block bootstrap. The optimal block length for each series was determined by running the Politis and White (2004) algorithm on the scores prior to performing the bootstrapped analysis.

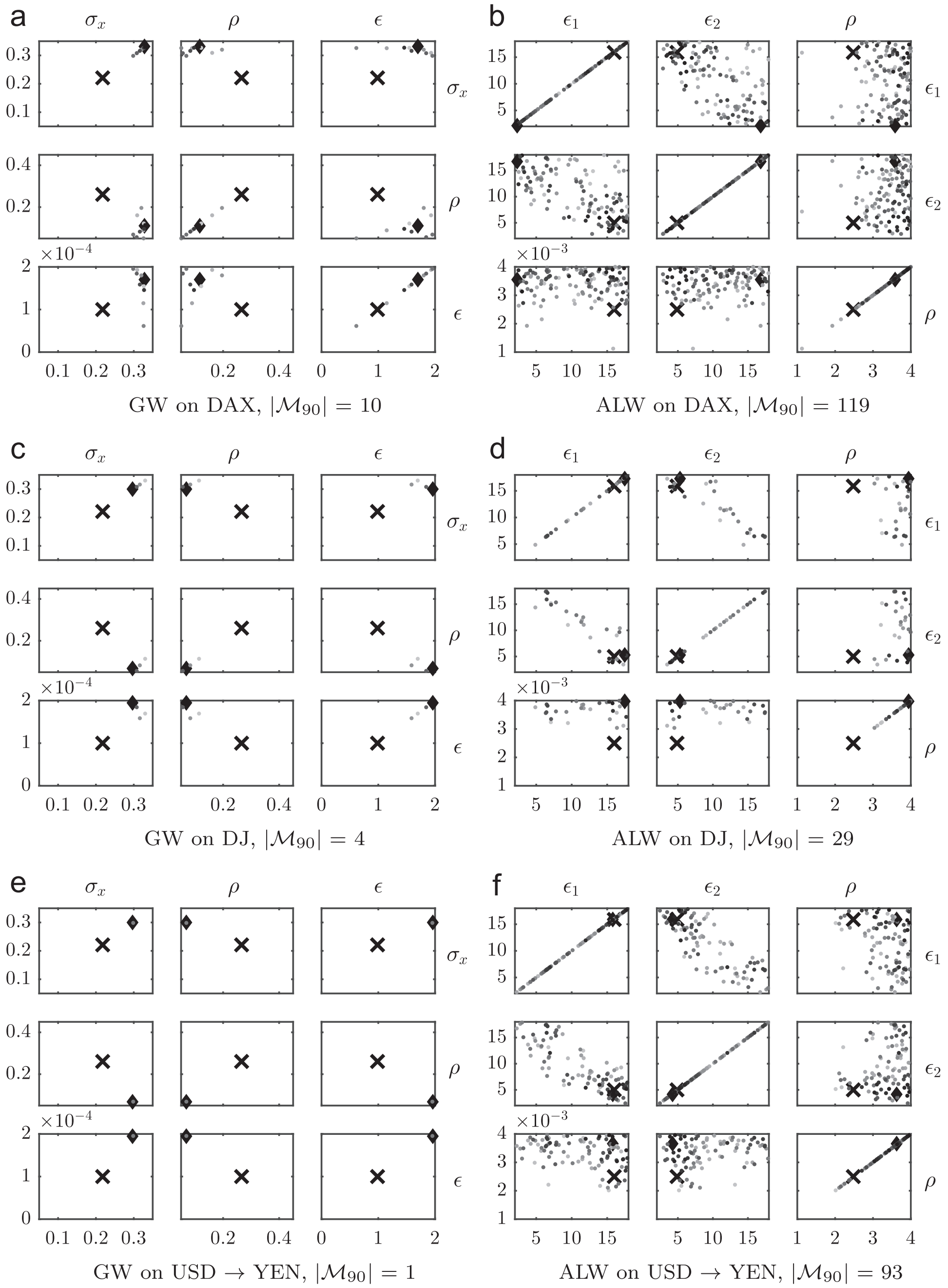

Fig. 2. Two way sensitivity scatter plots for GW and ALW. × is the original calibration, ◆ the best model, and lighter points perform worse.

both the US dollar/Yen and Swiss franc exchange rates, suggesting that the FW model performs particularly well on the exchange rate data.

4.2. Parameter sensitivity in calibrations

The parameter values corresponding to the calibrations identified in Table 2 are provided in Table A3 in the appendix. For convenience, Table 4 displays only those calibrations identified as best at least twice in the exercise. Both tables also include for purposes of comparison those calibrations that come closest to the original calibrations displayed in Table 1. Because the model comparison design uses an *ex-ante* sampling of the parameter space rather than a full re-calibration, it is important to check the sensitivity of these calibrations in order to evaluate their robustness.

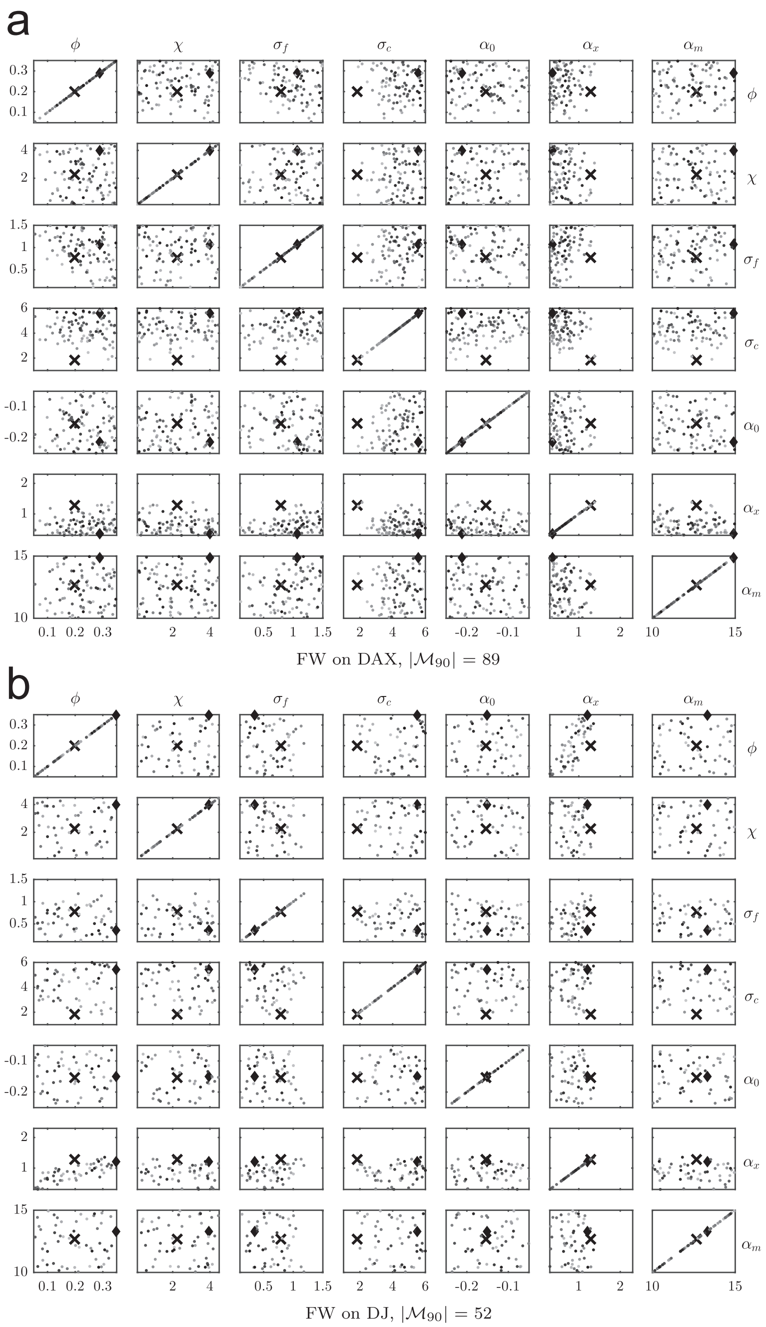

Fig. 3. Two way sensitivity scatter plots for FW – Part 1. × is the original calibration, ♦ the best model, and lighter points perform worse.

For each class of model, this is achieved by running the MCS analysis over the 513 candidate calibrations. Taking advantage of the orthogonal structure and space filling properties of the NOLH sampling, one can use the resulting confidence set $\mathcal{M}_\alpha$ to identify those regions in the parameter space that provide good calibrations. Fig. 2 displays the results of this analysis for the GW and ALW models for three of the series where FW is identified as the best model, while Fig. 3 does the same for the FW model.[15] Both figures show a scatter plot of those calibrations in the 90% confidence set and identify the

[15] Due to space constraints, it is not possible to display here the full set of figures corresponding to the 3 model classes on the 24 data series. These are however freely available as supplementary material.

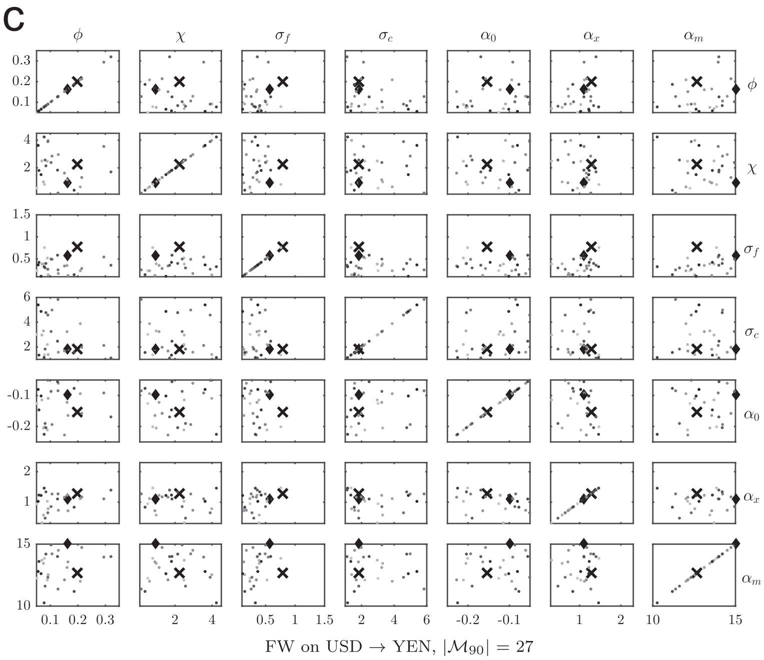

Fig. 3. Continue.

original and best calibrations with a cross ($\times$) and a diamond ($\diamond$) respectively. The greyscale is used to give information about the ranking of the calibration within the confidence set, with darker points identifying higher performing calibrations.

For the GW model, first of all, it is interesting to note that the confidence set is typically very small and concentrated around the best-performing calibration. For all six exchange rates it is even the case that $|\mathcal{M}_{90}| = 1$, implying that the confidence set is simply reduced to the best calibration. Combined with the fact that the location of the best calibration in the parameter space is very consistent across series, this suggests that only a very narrow range of parameter values enable good performance from the GW model and that the methodology is able to pin them down effectively. In terms of the values themselves, the best calibrations deviate strongly from the values in Table 1, with higher values of σ_x and ϵ and a lower value of ρ. This will be discussed further in Section 5.

For the ALW model, on the other hand, the sensitivity analysis suggests that the best calibrations identified are in line with the original estimates from Table 1, especially for ϵ_1 and ϵ_2. Although the original value of ρ is often below the best-fitting value, it is also often within or on the edge of the region of the parameter space containing the confidence set. The $\{\epsilon_1, \epsilon_2\}$ scatter plots in particular reveal an interesting symmetry in the confidence set, as two "good" regions seem to coexist. One is centred on the original calibration $\{\epsilon_1 = 16, \epsilon_2 = 4.9\}$. The other is centred on a zone where these two values are interchanged, as is the case in Fig. 2(b) for the best DAX calibration. While this second region does not provide the best calibration for the DJ and Dollar/Yen series, Fig. 2(d) and (f) do show that it is included in the confidence set. At this stage, it is important to point that this symmetry is in line with the original Alfarano et al. (2005) results for the DAX series. While they do find $\{\epsilon_1 = 16, \epsilon_2 = 4.9\}$ when estimating the uniform noise model, the estimates flip over to $\{\epsilon_1 = 1.37, \epsilon_2 = 14.0\}$ when estimating the spin noise model, which is in line with the values in Fig. 2(b). Crucially, the fact that the combination of NOLH sampling, MIC comparison and MCS testing is able to identify this symmetry in the $\{\epsilon_1, \epsilon_2\}$ space suggests that this protocol is capable of reliably identifying good parameterisations.

Finally the analysis for the FW model, shown in Fig. 3, shows that the spread of the $\mathcal{M}_{90}$ confidence set over the parameter space is much wider than was the case for GW and ALW in Fig. 2. This is in line with the findings in Tables 2 and A3, where the parameter values for the best-performing calibrations fluctuate quite widely around the original calibrations, in bold. Overall, this suggests that parameter values for the FW model are much less tightly pinned down than for the other two. As will be discussed below, this is probably due to the larger number of parameters available in the model, which would also explain why the FW model is consistently ranked better than ALW and GE. Fig. 3 does show, nevertheless, that for α_x, σ_c and to a lesser extent σ_f, the methodology is able to exclude portions of the parameter space. For α_x, the original value identified by $\times$ is at the lower end of the range of the best-fitting values while for σ_c the opposite is true. Again, the implications of this will be discussed below.

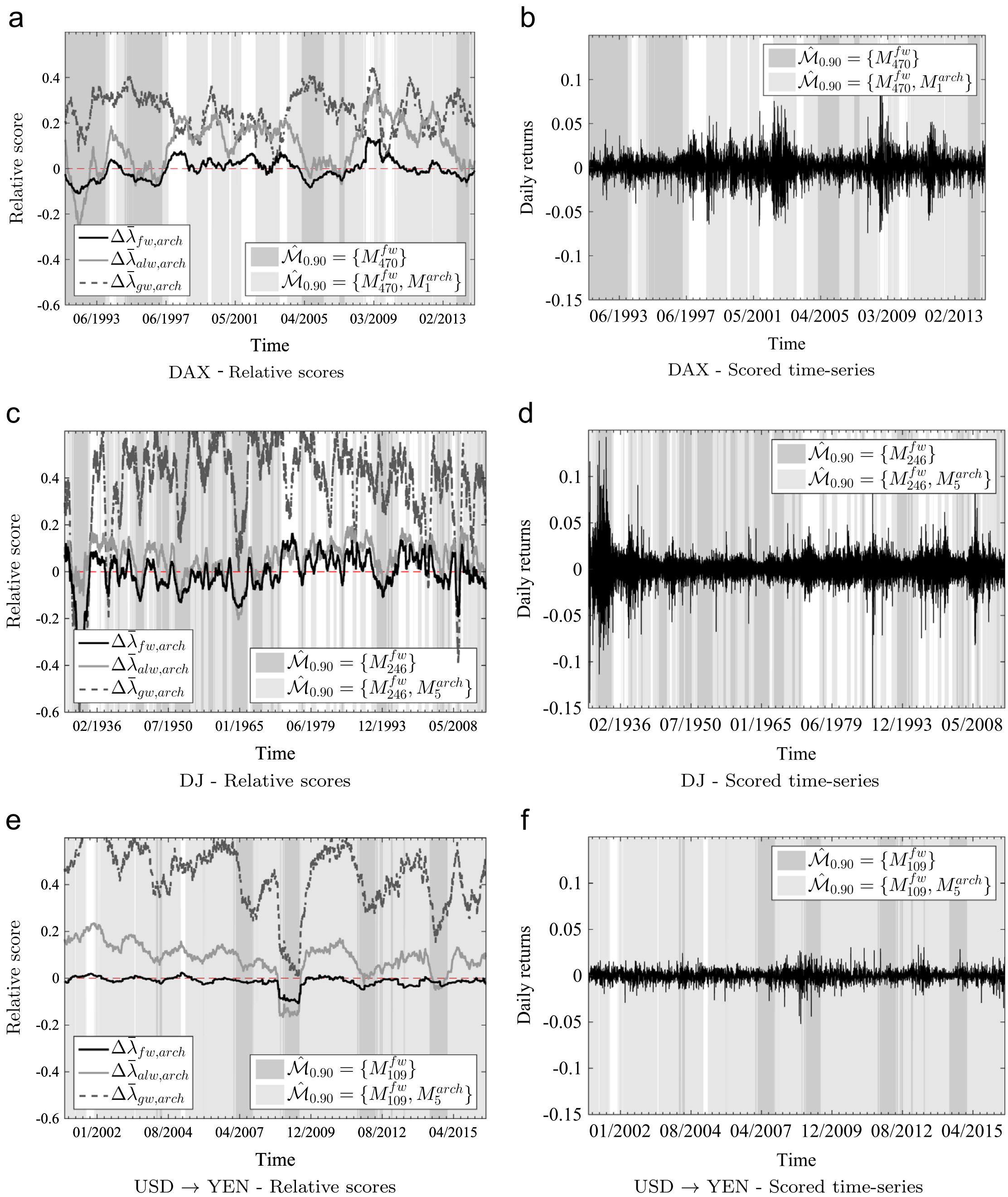

Fig. 4. Model scores relative to ARCH benchmark for DAX, DJ and USD → YEN.

4.3. *Local analysis*

Another benefit of having an observation-level vector of scores is that the relative performance of models can also be examined over relatively short time-lengths, as illustrated in Fig. 4, for the same three series as Figs. 2 and 3.[16] The three line plots show the relative scores $\Delta\overline{\lambda}_{i,arch}(r_t)$ of the three agent based models against the ARCH benchmark, smoothed using a moving average window of 200 observations. Using smoothed scores means that an MCS test can be carried out on the 200 individual observation scores $\lambda_i(r_t)$ in the window to evaluate the significance of the resulting average $\overline{\lambda}_i(r_t)$. In order to avoid complicating the figures further and given the clear rankings in Table 2, the test is only carried out on the FW/ARCH head-to-head comparison

[16] As for the sensitivity scatter plots 2 and 3, the complete set of plots covering all 24 series is not included here to save space, but is available in the supplementary material.

and the resulting MCS composition is displayed using the vertical banding. Dark grey bands reveal the time-periods for which the FW model beats ARCH significantly and white bands show the reverse. The lighter grey identifies time periods where the performance of the best FW and best ARCH specifications are statistically indistinguishable.

First of all, these observation-level plots confirm the relative rankings of the GW, ALW and FW models displayed in Table 2, in particular that the best GW calibration performs poorly relative to ARCH for all series. More importantly, the figures also reveal important features that are not discernable from the aggregate rankings in Table 2. A critical aspect from the point of view of agent-based modelling is that the performance of the FW and ALW models is often very close and their scores relative to the best ARCH model often co-move. This is most visible in Fig. 4(c) and suggests that they both explain similar features of the data and capture similar mechanisms. A second important feature is the presence of clear spikes in performance of all 3 models around turbulent events, in particular the 1929 and 2008 crashes for the DJ. These spikes, along the vertical bands where the MCS is restricted to the FW model alone, indicate that over these time periods, the FW model drastically outperforms the ARCH benchmark. Crucially, in most cases both the ALW and GW models also exhibit similar spikes over the same periods, which suggests that even if they may not be the best model to fit the data, their structure can capture the herding dynamics of turbulent periods just as well as the FW model relative to the benchmark.

5. Discussion

Let us discuss at this stage what these findings imply for the herding mechanisms discussed in Section 2. First of all, the fact that the FW model can offer statistically equivalent performance to ARCH-like specifications is not, by itself, a validation of these mechanisms. Instead, it is the spikes in relative scores around critical market events, identified in Fig. 4, which strongly suggest that herding offers an important explanation for the dynamics of conditional heteroscedasticity. A key aspect of this conclusion is the co-movement of GW and ALW with FW relative to ARCH on these events, regardless of the relative ranking of the three models on the overall data. The most salient example of this is provided in Fig. 4(c) and (e), where the performance of the GW model improves drastically at the same points in time as ALW and FW, despite a clearly poor overall performance.

The question then turns to explaining the relative rankings of the three models, and what can be inferred from this in terms of the design of their specific herding mechanisms. The poor performance of the GW model compared to ALW and FW suggests that the basic herding mechanism (1) of Kirman (1993) is probably too simplistic. Two explanations are possible, the first of which is the hypothesis of Alfarano et al. (2005) that asymmetry is an important part of the herding story. Because the GW model lacks this component while the ALW and FW models possess it, this would explain the difference in performance. However, the fact that the overall rankings on the exchange rate data are comparable to the stock market data suggests that this is not the main explanation, even if it cannot be excluded completely.

A second, more probable, explanation is that the presence of noise in the chartist demand function improves the performance of a model. In GW, the demand function of chartists (3) is purely momentum driven, with no specific noise component other than the overall exogenous error in (4) controlled by σ_s. On the other hand, the chartist demand of the ALW model (7) has no momentum component at all and is entirely driven by noise trading. Similarly, in the FW model the chartist demand (12) contains both momentum and noise trading, with Table 4 suggesting a relatively large noise component σ_c. This more flexible specification for chartist demand probably contributes to the higher performance of FW (and to a lesser extent ALW) compared to the basic momentum trading of GW. As shown in Fig. 3, however, while this increased parametric flexibility can help explain the better performance of FW relative to the other two models, it probably comes at the cost of not being able to pin down the parameter values as effectively.

A final aspect to consider is the sensitivity of the parameter values obtained in Tables 2 and in Fig. 2 and how they relate to the values obtained in the original studies. A key finding is that the methodology is able to narrow down effectively the parameter values for the GW and ALW models, although it cannot do as well for the FW model, probably due to the higher dimensionality. Interestingly, for the ALW model the parameter values obtained are in line with those of Alfarano et al. (2005), which is not the case for the GW model parameter values, which deviate strongly from Gilli and Winker (2003). A possible explanation for this discrepancy lies in the difference in estimation methods used by those two studies. As mentioned in the introduction, Gilli and Winker (2003) use MSM and choose to match the ARCH(1) parameter estimate and the kurtosis of the returns. Alfarano et al. (2005), instead, use maximum likelihood to estimate the parameters from the analytical solution resulting from the Fokker–Plank equation. This is conceptually similar to the MIC methodology, which produces a log score (22) of the conditional probability structure of a model. It is not entirely surprising, therefore, that the best MIC parameterisations obtained here agree with the maximum likelihood estimates of Alfarano et al. (2005), but not with the MSM estimates of Gilli and Winker (2003). A compounding factor is also the fact that Gilli and Winker (2003) only use two moments in their MSM objective function, while by construction the MIC integrates the entire conditionally density. This implies that if some moments of the conditional density left out by Gilli and Winker (2003) are informative for parameter values, there will be a systematic deviation in the parameters between the two methodologies. This offers support to the argument, mentioned in introduction, that the choice of methodology to be used when comparing models is not neutral, and needs to be carefully investigated.

6. Conclusion

The central aim of the paper was to test a new model comparison methodology for agent-based models by carrying out a horse race on three agent-based models of herding in financial markets and evaluate their performance relative to standard ARCH/GARCH benchmarks. The novelty of the approach stems from the ability of the MIC comparison methodology to directly compare any class of model (agent based simulations or traditional regressions) on an equal footing, by relying on their simulated output alone to produce a standardised criterion. The methodology first uses the simulated data to build a Markov transition matrix for the model, and then uses the empirical data to produce a log score for the model on the data. The rankings obtained can be tested statistically both at the aggregate and local level, thus providing a better understanding of the relative strengths and weaknesses of the candidate models on the data.

Two types of conclusions can be drawn from the results obtained in this paper. The first relates to the methodology itself, as the exercise demonstrates that it is indeed possible to compare effectively large numbers of agent-based simulation models to more traditional regression models on the basis of their simulated output. In itself, this is not necessarily surprising, as a body of work already exists using simulation-based estimation, however, this new methodology provides several desirable characteristics not present in alternative methods. The first is that while the algorithms used to obtain the measurement might seem unfamiliar, its nature – a log score of the conditional probability structure of a model – directly links up with the standard literature on maximised likelihoods and information criteria, making the interpretation of the results straightforward. The second, and most important, characteristic is the fact that the log score is produced at the observation level, enabling model comparison over sub-sets of the data and allowing the use of the Hansen et al. (2011) MCS methodology to provide statistical confidence when comparing models. This can be carried out across model classes, in order to test the rankings obtained for them, or within a model class in order to test the sensitivity of the best performing calibration. This latter analysis can be enhanced with a careful choice of sampling design. These aspects are all illustrated in the comparison exercise and reveal much more information about the relative performance of the three models against the benchmark than would be available from the set of aggregate rankings alone.

The second set of conclusions relates to the agent-based models of herding in financial markets that were being compared. The results suggest that population switching is an important factor for explaining the stylised facts of financial markets such as volatility clustering and fat tails. This is not only supported by the aggregate ranking of the models, but also by the plots of sub-sample relative scores, which indicate that the performance of all three herding models improves against ARCH-type models on key events where one would expect these stylised facts to be prominent. The results also suggest, however, that the herding mechanism is more complex than the basic framework of Kirman (1993) and better captured by richer mechanisms building in asymmetries in the propensity to switch, or feedback effects where the probability of switching is also determined by the perceived deviation from the fundamentals. Similarly, the findings suggest that the traditional division of population into "fundamentalists" driven purely by reversion to fundamentals and "chartists" driven purely by momentum strategies is also over-simplistic. Indeed, the results support the idea that noise traders, either independently or as a component of chartist demand, play a role in explaining the dynamics of these markets. The limitations of the exercise, discussed below, mean that it is not really possible to identify whether the better performance of the ALW and FW models compared to the basic GW model is due to the richer herding model or the noisier chartist demand, leaving this as an open question for future research.

The methodology does possess some limitations, which should be the focus of future development work. The first, as pointed out in Barde (2016) and mentioned in the introduction, is that in its current version the methodology is not designed for estimation, but instead for comparison of models that already possess calibrated parameters, even if this calibration is poor. This should be visible in the protocol used for comparison, as for each of the three models a preexisting set of samples is used to test the sensitivity of parameters around existing calibrations. To some extent, this problem can be mitigated by a careful choice of sampling design. This is illustrated here by using the NOLH design to provide a sample with good coverage and orthogonality properties. The results obtained show that given this choice of sampling design, the methodology can indeed perform some sensitivity testing for each model. Nevertheless, if the researcher's goal is simply to calibrate a single model, it currently cannot compare with simulation-based estimation methods. An objective for future development is therefore to investigate if the methodology can be used as the loss function in a more refined search algorithm, for example along the lines of the Nelder–Mead simplex search used in Gilli and Winker (2003), which would allow for more effective parameter calibration. The other current limitation of the methodology is its univariate nature, which is perfectly appropriate for models of financial data, but is problematic if multivariate models need to be compared. There is no theoretical hurdle stopping the CTW algorithm from being extended to multivariate settings, however, and developing such an extension is a second important development goal for the future.

Acknowledgements

The author is extremely grateful to participants at the WEHIA 2015 and CEF 2015 conferences, especially Blake LeBaron, Philipp Harting, Cars Hommes, Thomas Lux and Sander van der Hoog for their helpful suggestions. The paper benefited greatly from the suggestions of Dr John Peirson and two anonymous referees, particularly with respect to the sampling approach used. The author is particularly grateful to Sandrine Jacob-Léal for her advice on suitable agent-based models of financial markets and to James Holdsworth, Steve Sanders and Mark Wallis for their help in maintaining the SAL and PHOENIX computer clusters on which the model comparison exercise was run. Any errors in the paper remain of course the author's.

Appendix A. Extended tables

See Tables A1–A7.

Table A1
Descriptive statistics and discretisation tests on financial series.

Index	Start date	Obs.	Zeros	$\notin$ [−0.3,0.3]	Test 1 KS Stat.	Test 2 LB Stat.	Test 3 LB Stat.
AOI	03/08/1984	7684	60	0	0.0081 (0.9632)	27.5517 (0.3289)	24.8426 (0.4712)
NIKKEI	04/01/1984	7619	22	0	0.0105 (0.7930)	19.0253 (0.7959)	23.1469 (0.5690)
KOSPI	04/01/1980	9511	36	0	0.0127 (0.4226)	34.0359 (0.1071)	23.0432 (0.5750)
ST	28/12/1987	6770	54	0	0.0134 (0.5709)	20.3379 (0.7289)	18.6876 (0.8118)
HS	02/01/1980	8734	144	1	0.0210** (0.0426)	20.5453 (0.7177)	11.9306 (0.9871)
NIFTY	03/07/1990	5891	17	0	0.0081 (0.9894)	36.2685* (0.0676)	13.3362 (0.9721)
DAX	26/11/1990	6091	20	0	0.0099 (0.9279)	24.4933 (0.4910)	17.7300 (0.8536)
CAC	01/03/1990	6280	18	0	0.0116 (0.7877)	26.5557 (0.3784)	15.9344 (0.9168)
FTSE	02/04/1984	7756	19	0	0.0187 (0.1315)	17.3303 (0.8695)	36.0054* (0.0715)
IBEX	06/09/1991	5863	16	0	0.0075 (0.9964)	17.8435 (0.8490)	18.3471 (0.8273)
AEX	25/11/1988	6757	180	0	0.0268** (0.0154)	27.9252 (0.3113)	19.2415 (0.7854)
STOXX	31/12/1986	7200	22	0	0.0069 (0.9949)	26.4649 (0.3831)	19.0646 (0.7940)
IPC	08/11/1991	5776	7	0	0.0083 (0.9881)	34.1984 (0.1037)	15.9858 (0.9152)
DJ	01/10/1928	21665	102	0	0.0059 (0.8496)	45.4703*** (0.0074)	25.2449 (0.4487)
S&P 500	31/12/1979	9020	207	0	0.0263*** (0.0039)	26.3490 (0.3892)	17.7093 (0.8545)
NASDAQ	01/10/1985	7363	9	0	0.0114 (0.7219)	33.3841 (0.1217)	25.6411 (0.4269)
OEX	02/08/1982	8164	28	0	0.0069 (0.9905)	11.5033 (0.9901)	13.4947 (0.9698)
GSPTSE	31/12/1976	9551	20	0	0.0065 (0.9876)	35.9749* (0.0720)	30.2662 (0.2145)

Test 1 – Kolmogorov–Smirnov test on discretisation error.
H_0: Discretisation error is uniformly distributed over [0,1].
Test 2 – Ljung–Box test on 25 lags of the discretisation error.
H_0: Discretisation error is independently distributed (no autocorrelation).
Test 3 – Ljung–Box test of the discretisation error against 25 lags of the discretisation series
H_0: Discretisation error is not correlated with discretised series.
P-values are in parenthesis, * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.
The last observation, for all indices, is the 12 December 2014.

Table A2
Descriptive statistics and discretisation tests on exchange rate series.

Series	Start date	Obs.	Zeros	$\notin$ $[-0.3,0.3]$	Test 1 KS Stat.	Test 2 LB Stat.	Test 3 LB Stat.
USD → GBP	04/01/2000	4155	28	0	0.0094	19.2831	16.6553
					(0.9928)	(0.7833)	(0.8939)
USD → EUR	03/01/2000	4156	36	0	0.0152	23.1126	31.4125
					(0.7231)	(0.5710)	(0.1757)
USD → YEN	03/01/2000	4156	32	0	0.0262	36.0190*	26.3771
					(0.1129)	(0.0713)	(0.3877)
USD → CHF	03/01/2000	4156	34	0	0.0154	28.8017	28.7722
					(0.7048)	(0.2723)	(0.2735)
USD → AUD	04/01/2000	4155	41	0	0.0128	24.2354	39.3309**
					(0.8860)	(0.5058)	(0.0341)
USD → MXN	03/01/2000	4156	19	0	0.0185	20.4445	21.6351
					(0.4701)	(0.7231)	(0.6567)

Test 1 – Kolmogorov–Smirnov test on discretisation error.
H_0: Discretisation error is uniformly distributed over [0,1].
Test 2 – Ljung–Box test on 25 lags of the discretisation error.
H_0: Discretisation error is independently distributed (no autocorrelation).
Test 3 – Ljung–Box test of the discretisation error against 25 lags of the discretisation series.
H_0: Discretisation error is not correlated with discretised series.
P-values in parenthesis, * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.
The last observation, for all exchange rate series, is the 15 July 2016.

Table A3
MCS model parameters (full table).

Gilli and Winker (*GW*)															
σ_x	0.298	**0.200**	0.308	0.316	0.330	0.326	0.320	0.327							
ρ	0.069	**0.250**	0.056	0.083	0.114	0.051	0.196	0.096							
ϵ	1.953e−4	**1.000e−4**	1.844e−4	1.586e−4	1.695e−4	6.133e−5	1.801e−4	1.457e−4							
id	11	**65**	75	190	248	327	418	453							
N°	10	**0**	2	2	6	2	1	1							
Alfarano, Lux and Wagner (*ALW*)															
ϵ_1	15.750	2.250	12.875	17.375	14.875	**16.375**	14.438	15.813	15.469	17.281					
$\epsilon 2$	3.875	16.750	2.875	5.375	7.000	**4.500**	6.938	4.188	3.531	3.469					
ρ	3.906e−3	3.578e−3	3.742e−3	3.953e−3	3.812e−3	**3.039e−3**	3.332e−3	3.637e−3	3.643e−3	3.115e−3					
id	54	75	104	112	120	**124**	177	220	315	494					
N °	5	3	1	7	1	**1**	1	3	1	1					
Franke and Westerhoff (*FW*)															
ϕ	0.348	0.066	0.289	0.162	0.164	0.201	0.346	0.265	0.347	0.211	0.321	0.187	0.285	**0.257**	0.287
χ	4.225	1.509	1.338	0.925	1.355	3.211	3.967	2.945	1.312	3.220	3.804	0.384	3.752	**2.377**	4.027
σ_f	0.177	0.111	0.275	0.570	0.816	0.926	0.346	0.114	1.104	1.180	0.841	0.146	0.349	**0.901**	1.065
σ_c	5.063	4.867	5.805	1.859	5.434	5.512	5.473	5.600	5.990	4.779	5.639	1.557	4.721	**2.006**	5.580
α_0	−0.128	−0.102	−0.230	−0.098	−0.243	−0.165	−0.151	−0.175	−0.239	−0.236	−0.099	−0.077	−0.117	**−0.137**	−0.213
α_x	1.316	1.222	0.706	1.112	1.464	0.667	1.198	0.554	0.538	1.085	1.132	0.835	1.218	**1.124**	0.359
α_m	13.555	12.617	12.344	15.000	14.785	12.168	13.301	11.338	11.533	10.049	11.143	11.182	10.420	**11.943**	14.893
id	28	37	88	109	145	243	246	279	284	288	320	353	432	**451**	470
N °	1	2	1	1	1	1	1	1	4	1	3	1	2	**0**	4

Bold indicates the model id corresponding to the calibrations of the original works.

Table A4
Number of iterations required for convergence of ARCH models estimations.

Type	ARCH	GARCH			TGARCH			EGARCH			PGARCH		
Lags	5	1	2	3	1	2	3	1	2	3	1	2	3
id	1	2	3	4	5	6	7	8	9	10	11	12	13
AOI	85	89	104	208	39	31	74	175	187	189	76	111	242
NIKKEI	31	44	75	104	29	35	72	48	67	54	51	43	447
KOSPI	11	42	82	73	48	67	54	64	46	37	48	42	232
ST	30	40	32	31	51	50	182	60	52	173	50	81	128
HS	23	26	24	47	21	48	39	94	100	53	29	114	–
NIFTY	17	17	18	27	14	18	46	18	24	170	20	98	198
DAX	20	55	68	38	38	77	130	41	102	10	50	–	196
CAC	11	19	28	80	22	29	43	27	26	80	29	41	61
FTSE	9	12	53	32	12	37	27	22	36	31	13	52	164
IBEX	12	13	30	22	15	20	60	21	113	45	19	35	55
AEX	12	13	38	37	21	24	24	19	34	17	17	32	
STOXX	16	45	59	60	29	34	192	36	10	36	52	59	–
IPC	11	13	22	34	14	16	36	16	11	27	15	25	41
DJ	16	24	39	76	19	33	34	44	101	75	31	61	291
S&P 500	13	11	20	17	12	23	34	24	50	45	16	44	58
NASDAQ	13	22	35	74	21	23	31	29	16	15	22	38	77
OEX	25	35	86	75	25	27	9	40	61	32	30	28	33
GSPTSE	18	28	68	46	38	24	36	42	45	124	41	87	67
USD → GBP	23	30	29	35	11	40	53	33	71	60	67	94	134
USD → EUR	11	12	14	192	16	32	87	21	11	81	23	–	143
USD → YEN	13	13	27	22	17	17	22	15	23	91	20	37	57
USD → CHF	83	138	235	139	123	–	16	117	111	242	273	22	26
USD → AUD	19	14	25	23	18	32	52	29	30	37	21	40	–
USD → MXN	17	13	11	36	14	21	–	22	26	28	18	34	30

"–" indicates that the estimation did not converge after 500 iterations.

Table A5
ARCH models estimation for the Frankfurt Stock Exchange DAX index (DAX).

Type	ARCH	GARCH			TGARCH			EGARCH			PGARCH		
Lags	5	1	2	3	1	2	3	1	2	3	1	2	3
id	1	2	3	4	5	6	7	8	9	10	11	12	13
c	0.0696***	0.0651***	0.0621***	0.0647***	0.0344**	0.0370***	0.0388***	0.0304**	0.0304**	0.0266*	0.0281*	0.0284**	0.0320**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0184)	(0.0085)	(0.0052)	(0.0329)	(0.0324)	(0.0589)	(0.0500)	(0.0438)	(0.0243)
a_1	0.0058	−0.0009	0.0001	0.0004	0.0030	0.0040	0.0075	−0.0003	−0.0041	−0.0006	0.0016	−0.0022	0.0024
	(0.6449)	(0.9519)	(0.9959)	(0.9732)	(0.8367)	(0.7789)	(0.5801)	(0.9817)	(0.7296)	(0.9580)	(0.9082)	(0.8719)	(0.8647)
a_2	−0.0071	0.0020	0.0013	0.0013	0.0109	0.0150	0.0132	0.0029	−0.0047	−0.0001	0.0065	0.0001	0.0020
	(0.5790)	(0.8846)	(0.9243)	(0.9224)	(0.4145)	(0.2516)	(0.3385)	(0.8212)	(0.7030)	(0.9952)	(0.6213)	(0.9955)	(0.8745)
σ_0	0.5532***	0.0313***	0.0375***	0.0003***	0.0334***	0.0003***	0.0004***	−0.0838***	−0.0537***	−0.0328***	0.0290***	0.0269**	0.0153***
	(0.0000)	(0.0000)	(0.0000)	(0.0050)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0496)	(0.0000)
α_1	0.0459***	0.0824***	0.0190**	0.0266***	0.0142***	−0.0017	−0.0308***	0.1199***	−0.0140	−0.0294	0.0662***	0.0500**	0.0496
	(0.0000)	(0.0000)	(0.0179)	(0.0000)	(0.0005)	(0.7139)	(0.0000)	(0.0000)	(0.4561)	(0.1334)	(0.0000)	(0.0289)	(0.3878)
α_2	0.1917***		0.0727***	0.0317***		0.0031	0.0763***		−0.1072	0.1705***		0.0154	−0.0517***
	(0.0000)		(0.0004)	(0.0000)		(0.4954)	(0.0000)		(0.0000)	(0.0000)		(0.8533)	(0.0058)
α_3	0.1879***			−0.0570***			−0.0437***			−0.0955***			0.0413***
	(0.0000)			(0.0000)			(0.0000)			(0.0003)			(0.0003)
α_4	0.1642***												
	(0.0000)												
α_5	0.1599***												
	(0.0000)												
β_1		0.9006***	1.0769***	2.1913***	0.9101***	1.8660***	1.5825***	0.9799***	1.4357***	1.9144***	0.9234***	1.1426***	2.0695***
		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0048)	(0.0000)
β_2			−0.1885	−1.5170***		−0.8676***	−0.3471***		−0.4480***	−1.2042***		−0.2175	−1.7604***
			(0.2082)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.5543)	(0.0000)
β_3				0.3242***			−0.2373***			0.2838***			0.6461***
				(0.0000)			(0.0000)			(0.0018)			(0.0000)
γ_1					0.1099***	0.1608***	0.1387***	−0.0801***	−0.1983***	−0.1701***	0.6707***	1.0000	1.0000
					(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.1667)	(0.6159)
γ_2						−0.1608***	−0.0767***		0.1616***	0.1702***		−1.0000	0.6296*
						(0.0000)	(0.0086)		(0.0000)	(0.0000)		(0.9104)	(0.0779)
γ_3							−0.0624***			−0.0201			0.1324
							(0.0002)			(0.5014)			(0.5727)
δ											1.1790***	1.0886***	1.1300***
											(0.0000)	(0.0000)	(0.0000)
AIC	3.2651	3.2272	3.2227	3.2188	3.2069	3.1922	3.1900	3.2060	3.1939	3.1914	3.2009	3.1982	3.1937
BIC	3.2750	3.2338	3.2315	3.2299	3.2146	3.2032	3.2043	3.2137	3.2049	3.2057	3.2097	3.2103	3.2091

P-values in parenthesis, * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Table A6
ARCH models estimation for the New York Stock Exchange Dow Jones Industrial Average (DJ).

Type	ARCH	GARCH			TGARCH			EGARCH			PGARCH		
Lags	5	1	2	3	1	2	3	1	2	3	1	2	3
id	1	2	3	4	5	6	7	8	9	10	11	12	13
c	0.0542***	0.0435***	0.0453***	0.0451***	0.0229***	0.0253***	0.0243***	0.0215***	0.0216***	0.0211***	0.0206***	0.0214***	0.0251***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0001)	(0.0002)	(0.0003)	(0.0002)	(0.0000)
a_1	0.0917***	0.0856***	0.0880***	0.0863***	0.0903***	0.0884***	0.0884***	0.0873***	0.0871***	0.0879***	0.0894***	0.0866***	0.0876***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
a_2	−0.0275***	−0.0233***	−0.0248***	−0.0228***	−0.0145**	−0.0193***	−0.0192***	−0.0125*	−0.0119*	−0.0125*	−0.0129*	−0.0149**	−0.0168**
	(0.0000)	(0.0009)	(0.0004)	(0.0011)	(0.0404)	(0.0061)	(0.0065)	(0.0705)	(0.0886)	(0.0734)	(0.0683)	(0.0376)	(0.0167)
σ_0	0.2810***	0.0101***	0.0001***	0.0206***	0.0117***	0.0003***	0.0003***	−0.1124***	−0.2262***	−0.3894***	0.0128***	0.0140***	0.0003
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.4937)
α_1	0.1303***	0.0824***	0.0947***	0.0833***	0.0281***	0.0040	−0.0009	0.1445***	0.1441***	0.1462***	0.0741***	0.0514	0.0734***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.1516)	(0.8380)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.8878)	(0.0000)
α_2	0.1502***		−0.0934***	0.0096***		−0.0013	0.0098		−0.1072	0.2095***		0.0142	−0.0671
	(0.0000)		(0.0000)	(0.0000)		(0.6471)	(0.2585)		(0.0000)	(0.0000)		(0.9904)	(0.6019)
α_3	0.1708***		0.0827***		−0.0063		0.1454***		−0.0004				
	(0.0000)			(0.0000)			(0.1947)			(0.0000)			(0.9994)
α_4	0.1855***												
	(0.0000)												
α_5	0.1656***												
	(0.0000)												
β_1		0.9103***	1.8547***	0.7938***	0.9149***	1.8166***	1.8017***	0.9867***	−0.0098***	−0.4477***	0.9219***	0.8167***	1.8082
		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.2256)
β_2			−0.8562***	−0.8929***		−0.8203***	−0.7912***		0.9830***	0.4153***		0.0967	−0.7831
			(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.5759)	(0.7754)
β_3				0.9094***			−0.0141***			0.9865***			−0.0280
				(0.0000)			(0.0000)			(0.0000)			(0.9822)
γ_1					0.0918***	0.1611***	0.1623***	−0.0699***	−0.0692***	−0.0711***	0.4472***	1.0000	0.8970***
					(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.9194)	(0.0000)
γ_2						−0.1596***	−0.1605***		−0.0716***	−0.1019***		−1.0000	0.9982
						(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.9931)	(0.3406)
γ_3							−0.0005			−0.0704***			−0.9966
							(0.8052)			(0.0000)			(0.9996)
δ											1.4489***	1.4304***	1.2790***
											(0.0000)	(0.0000)	(0.0000)
AIC	2.6517	2.5873	2.5818	2.5852	2.5704	2.5564	2.5566	2.5717	2.5718	2.5712	2.5681	2.5667	2.5526
BIC	2.6551	2.5895	2.5847	2.5889	2.5730	2.5601	2.5614	2.5743	2.5755	2.5760	2.5710	2.5708	2.5577

P-values in parenthesis, * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Table A7
ARCH models estimation for the Japanese Yen per US Dollar exchange rate (USD → YEN).

Type	ARCH	GARCH			TGARCH			EGARCH			PGARCH		
Lags	5	1	2	3	1	2	3	1	2	3	1	2	3
id	1	2	3	4	5	6	7	8	9	10	11	12	13
c	0.0062	0.0084	0.0084	0.0097	0.0052	0.0055	0.0067	0.0040	0.0078	0.0050	0.0038	0.0041	0.0054
	(0.5101)	(0.3300)	(0.3363)	(0.2636)	(0.5633)	(0.5445)	(0.4601)	(0.6406)	(0.3479)	(0.5597)	(0.6576)	(0.6405)	(0.5305)
a_1	−0.0161	−0.0152	−0.0153	−0.0142	−0.0144	−0.0140	−0.0126	−0.0157	−0.0185	−0.0151	−0.0162	−0.0162	−0.0153
	(0.3437)	(0.3425)	(0.3388)	(0.4095)	(0.3741)	(0.3797)	(0.4628)	(0.3218)	(0.2272)	(0.3594)	(0.3101)	(0.3056)	(0.3504)
a_2	−0.0087	−0.0129	−0.0129	−0.0111	−0.0123	−0.0115	−0.0150	−0.0095	−0.0130	−0.0152	−0.0096	−0.0091	−0.0118
	(0.6153)	(0.4178)	(0.4251)	(0.5098)	(0.4418)	(0.4813)	(0.3829)	(0.5432)	(0.3957)	(0.3111)	(0.5420)	(0.5690)	(0.4863)
σ_0	0.3046***	0.0045***	0.0091***	0.0056***	0.0057***	0.0105***	0.0073***	−0.0923***	−0.1830***	−0.0122***	0.0098***	0.0183***	0.0161***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
α_1	0.0778***	0.0352***	0.0350***	0.0653***	0.0263***	0.0197*	0.0399***	0.1023***	0.1121***	0.1299***	0.0516***	0.0469***	0.0605***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0594)	(0.0026)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
α_2	0.0950***		0.0353***	−0.0343**		0.0280**	−0.0127		−0.1072	−0.2264***		0.0492***	0.0349**
	(0.0000)		(0.0000)	(0.0461)		(0.0110)	(0.5287)		(0.0000)	(0.0000)		(0.0005)	(0.0333)
α_3	0.0173			0.0112			0.0047			0.1100***			−0.0116
	(0.1000)			(0.3384)			(0.7230)			(0.0000)			(0.3507)
α_4	0.0253**												
	(0.0104)												
α_5	0.0658***												
	(0.0000)												
β_1		0.9544***	−0.0342	1.3689***	0.9494***	0.1539	1.1167***	0.9832***	−0.0129***	2.2813***	0.9449***	0.1126	0.9230***
		(0.0000)	(0.5821)	(0.0000)	(0.0000)	(0.6659)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.6542)	(0.0000)
β_2			0.9430***	−1.1723***		0.7535**	−0.7530***		0.9804***	−1.6956***		0.7848***	−0.7487***
			(0.0000)	(0.0000)		(0.0260)	(0.0000)		(0.0000)	(0.0000)		(0.0009)	(0.0000)
β_3				0.7483***			0.5728***			0.4121***			0.7367***
			(0.0000)			(0.0000)			(0.0000)				(0.0000)
γ_1					0.0213***	0.0233**	0.0455***	−0.0258***	−0.0212***	−0.0747***	0.2611***	0.2357**	0.4018***
					(0.0000)	(0.0347)	(0.0082)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0432)	(0.0008)
γ_2						0.0162	0.0252		−0.0250***	0.1090***		0.2936***	0.1925
						(0.2523)	(0.3911)		(0.0000)	(0.0001)		(0.0083)	(0.4334)
γ_3							−0.0427**			−0.0368**			0.7736
							(0.0278)			(0.0136)			(0.3685)
δ											1.1003***	1.0964***	1.0487***
											(0.0000)	(0.0000)	(0.0000)
AIC	1.9416	1.8924	1.8934	1.8896	1.8910	1.8923	1.8878	1.8872	1.8818	1.8839	1.8871	1.8884	1.8838
BIC	1.9553	1.9016	1.9055	1.9049	1.9016	1.9075	1.9076	1.8978	1.8970	1.9037	1.8993	1.9052	1.9052

P-values in parenthesis, * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Appendix B. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jedc.2016.10.005.

References

Alfarano, S., Lux, T., Wagner, F., 2005. Estimation of agent-based models: the case of an asymmetric herding model. Comput. Econ. 26 (1), 19–49.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2000. Exchange Rate Returns Standardized by Realized Volatility Are (Nearly) Gaussian. Technical Report. National Bureau of Economic Research.

Barde, S., 2016. A practical, universal, information criterion over nth order Markov processes. Comput. Econ. http://dx.doi.org/10.1007/s10614-016-9617-9. (Forthcoming).

Bianchi, C., Cirillo, P., Gallegati, M., Vagliasin, P.A., 2007. Validating and calibrating agent-based models: a case study. Comput. Econ. 30, 245–264.

Cioppa, T.M., Lucas, T.W., 2007. Efficient nearly orthogonal and space-filling latin hypercubes. Technometrics 49, 45–55.

Dawid, H., Fagiolo, G., 2008. Agent-based models for economic policy design: introduction to the special issue. J. Econ. Behav. Organ. 67, 351–354.

Durlauf, S.N., 2005. Complexity and empirical economics. Econ. J. 115 (504), F225–F243.

Fagiolo, G., Moneta, A., Windrum, P., 2007. A critical guide to empirical validation of agent-based models in economics: methodologies, procedures, and open problems. Comput. Econ. 30, 195–226.

Franke, R., Westerhoff, F., 2011. Estimation of a structural stochastic volatility model of asset pricing. Comput. Econ. 38 (1), 53–83.

Franke, R., Westerhoff, F., 2012. Structural stochastic volatility in asset pricing dynamics: estimation and model contest. J. Econ. Dyn. Control 36 (8), 1193–1211.

Franke, R., Westerhoff, F., 2016. Why a simple herding model may generate the stylized facts of daily returns: explanation and estimation. J. Econ. Interact. Coord. 11 (1), 1–34.

Gilli, M., Winker, P., 2003. A global optimization heuristic for estimating agent based models. Comput. Stat. Data Anal. 42 (3), 299–312.

Gouriéroux, C., Monfort, A., 1993. Simulation-based inference: a survey with special reference to panel data models. J. Econom. 59, 5–33.

Gouriéroux, C., Monfort, A., 1996. Simulation-Based Econometric Methods. Oxford University Press, Oxford.

Grünewald, P.D., 2007. The Minimum Description Length Principle. MIT Press, Cambridge, Massachusetts.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79, 453–497.

Hommes, C., 2011. The heterogeneous expectations hypothesis: some evidence from the lab. J. Econ. Dyn. Control 35 (1), 1–24.

Hommes, C.H., 2006. Heterogeneous agent models in economics and finance. In: Handbook of Computational Economics, vol. 2, pp. 1109–1186.

Kirman, A., 1993. Ants, rationality and recruitment. Q. J. Econ. 108, 137–156.

Krichevsky, R.E., Trofimov, V.K., 1981. The performance of universal encoding. IEEE Trans. Inf. Theory IT-27, 629–636.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.

Lamperti, F., 2015. An Information Theoretic Criterion for Empirical Validation of Time Series Models. LEM Working Paper Series 2015/02.

Lux, T., Marchesi, M., 2000. Volatility clustering in financial markets: a microsimulation of interacting agents. Int. J. Theor. Appl. Financ. 3 (04), 675–702.

Mandes, A., Winker, P., 2015. Complexity and model comparison in agent based modeling of financial markets. J. Econ. Interact. Coord., 1–38.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. J. Am. Stat. Assoc. 89, 1303–1313.

Politis, D.N., White, H., 2004. Automatic block-length selection for the dependent bootstrap. Econ. Rev. 23, 53–70.

Salle, I., Yildizoğlu, M., 2014. Efficient sampling and meta-modeling for computational economic models. Comput. Econ. 44 (1), 507–536.

Tesfatsion, L., 2006. Agent-based computational economics: a constructive approach to economic theory. In: Handbook of Computational Economics, vol. 2. North-Holland, Amsterdam pp. 831–880.

Westerhoff, F., 2009. Exchange rate dynamics: a nonlinear survey. In: Handbook of Research on Complexity, pp. 287–325.

White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.

Willems, F.M.J., Shtarkov, Y.M., Tjalkens, T.J., 1995. The context-tree weighting method: basic properties. IEEE Trans. Inf. Theory IT-41, 653–664.

Winker, P., Gilli, M., Jeleskovic, V., 2007. An objective function for simulation based inference on exchange rate data. J. Econ. Interact. Coord. 2 (2), 125–145.