University of Kent

School of Economics Discussion Papers

Direct calibration and comparison of agent-based herding models of financial markets

Sylvain Barde

April 2015

KDPE 1507

Direct calibration and comparison of agent-based herding models of financial markets*

Sylvain Barde[†‡]

April 2015

Abstract

The present paper aims to test a new model comparison methodology by calibrating and comparing three agent-based models of financial markets on the daily returns of 18 indices. The models chosen for this empirical application are the herding model of Gilli & Winker, its asymmetric version by Alfarano, Lux & Wagner and the more recent model by Franke & Westerhoff, 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 clear and consistent ranking of the three models. More importantly, it also reveals that the best performing model, Franke & Westerhoff, is generally not distinguishable from an ARCH-type process, suggesting their explanatory power on the data is similar.

JEL classification: C15, C52, G12.

Keywords: Model selection, Agent-based models, herding behaviour.

*The author is extremely grateful to Sandrine Jacob-Léal for her advice on suitable agent-based models of financial markets and to James Holdsworth for his help in maintaining the computer cluster on which the model comparison exercise was run. Any errors in the manuscript remain of course the author's.

†School of Economics, Keynes College, University of Kent, Canterbury, CT2 7NP, UK
tel : +44 (0)1 227 824 092, email: s.barde@kent.ac.uk

‡affiliate, Observatoire Français des Conjonctures Economiques

Non-Technical Summary

The present paper provides the first empirical application of a novel and innovative model comparison methodology designed to provide an information criterion on a given set of data for any model that is reducible to a Markov process. The rationale behind the development of this methodology is to allow the explanatory power of simulation models to be compared to more traditional modeling approaches. This has been identified as one of the main hurdles to the development of simulation methods, particulary agent-based models, in the field of economics.

The empirical exercise carried out calibrates and compares three agent-based models of financial markets on the daily returns of 18 financial indices, the end goal being to establish both the robustness of their respective calibrations and their explanatory power on the data. 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, 2015). Two justifications explain the specific choice of these three models amongst the many available in the literature on agent-based models of finance:

- All three models have already been calibrated on financial market data, and all replicate the stylised facts of financial markets such as volatility clustering and fat tails in the distribution of returns.
- All three share a common theoretical lineage to the herding mechanism introduced by Kirman (1993). Each of them modifies this mechanism is some way, but the basic herding mechanism exists in all three.

The intention is to produce a setting where the structure and predictions of the models are similar, providing a challenging test when ranking the models. In addition to these three models, a set of standard ARCH/GARCH processes are estimated and included for each financial series. The purpose of including these econometric specifications is firstly to provide a benchmark for the explanatory power of the agent based models and secondly to demonstrate the ability of the methodology to compare very different modeling approaches.

The main findings are that the methodology provides a clear and consistent ranking of the three herding models over the 18 data series. The model of Gilli and Winker (2003) performs the worst on the 18 data series, and is beaten by Alfarano et al. (2005), itself beaten by Franke and Westerhoff (2011, 2015). More importantly, the ranking exercise also reveals that Franke & Westerhoff is generally not distinguishable from the best ARCH process, suggesting their explanatory power on the data is similar. More detailed analysis using a narrow observation window also reveals that all three models can dramatically outperform the ARCH processes during periods of financial turmoil, strongly suggesting that the herding mechanism explains these periods in time better than the standard econometric specifications. Because this could be due to levels of noise in the demand functions of agents or to differences in the herding process itself, more analysis is needed to identify the precise reason for the better performance of the agent-based models during these periods.

Finally, beyond the importance of these findings for the literature on herding in agent-based models of financial markers, the empirical exercise confirms the methodology's ability to rigorously compare the explanatory power of very different modeling approaches on a given set of data.

1 Introduction

The emergence of agent-based modeling as an alternative to the more traditional fully rational representative agent approach has enabled the integration 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.

The first hurdle is the use of data to estimate the parameters that govern the simulation. 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. Solutions to this problem have been found which 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) advocates 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 selction/comparison methods. Using the estimation methods mentioned above to compare models is possible but potentially problematic, as they often require tailoring to the model specification being estimated, making direct comparison across specification 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 requirement of deciding which moments to reproduce complicates the problem of comparison across models. 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 (2015), are estimated using the MSM, the former uses two moments, the latter nine, none of which are the same.[2]

Recent work by Barde (2015) and Lamperti (2015) argues 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) divergence between model and data. The methodological approach is different in both cases, 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. For large scale model comparisons, such as the one carried out here, Barde (2015) has the additional benefit that it scores each model at the level of individual data observations, which means that it integrates seamlessly with the model comparison set (MCS) approach developed by Hansen et al. (2011). The central implication is that not only can models be ranked in terms of their explanatory power on the data, this ranking can be testing statistically to determine the subset of candidate models that can be rejected at a chosen confidence level.

[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 (2015) 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

This paper aims to run a full-scale model comparison test of Barde (2015) in order to evaluate the methodology's potential for model selection. The aim and setting is similar to the recent model contest of Franke and Westerhoff (2012), which to the author's knowledge is the only existing example of a direct contest between agent-based models. The comparison exercise carried out here aims to go beyond Franke and Westerhoff (2012), however, by directly incorporating models calibrated by different authors as well as econometric ARCH/GARCH specifications to serve as a benchmark for explanatory power. The chosen setting, agent-based models of herding in financial markets, possesses several ideal characteristics for the exercise. First of all, there is an established literature on the herding mechanism which crucially offers several models that have been calibrated but not yet systematically compared. The second desirable characteristic is that because these models typically attempt to explain the dynamics of stock market returns, they offer a univariate setting with plentiful data which simplifies the problem of comparison. Finally, the focus on stock market 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, 2015). 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, which means that the 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 and discussed in section 4, while section 5 draws the main conclusions.

2 Agent-based models of herding in financial markets

The three models in the comparison exercise from the lineage of 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

Table 1: Calibrated values of model parameters

Parameter	Interpretation	Value
	Gilli & Winker (GW)	
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
ϵ	Probability of random switch	0.0001
ρ	Probability of direct recruitment	0.264
	Alfarano, Lux & Wagner (ALW)	
ϵ_1	Propensity for fundamentalist → chartist switch	16
ϵ_2	Propensity for chartist → fundamentalist switch	4.9
ρ	Herding tendency	0.0025
	Franke & Westerhoff (FW)	
ϕ	Agressiveness of fundamentalists	0.198
χ	Aggressiveness of chartists	2.263
σ_f	Noise in fundamentalist demand	0.782
σ_c	Noise in chartist demand	1.851
μ	Market impact factor of demand	0.01*
p^*	Log of fundemental value	0*
ν	Flexibility in population dynamics	0.05*
α_0	Predisposition parameter	-0.155
α_x	Herding parameter	1.299
α_m	Misalignment parameter	12.648

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 (2015) is calibrated on the S&P500 index.
'*' indicates the parameter value is assumed by the authors, not calibrated.

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 $\tilde{x}_t \sim \mathrm{N}\left(x_t, \sigma_x^2\right)$ 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 $\bar{p}$ at a given rate ϕ, while chartists simply extrapolate from past price movements.

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

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

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

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 τ^{th} observation from the raw interaction-level series (4).

2.2 The Alfarano, Lux and Wagner (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:

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

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

[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

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 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.

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

The drift term of this time evolution depends on $\bar{x} = \epsilon_1/(\epsilon_1 + \epsilon_2)$, 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 $\bar{p}$. 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:

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

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 equations (6) and (8) with $\Delta t = 1$.

[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.

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 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 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 $\bar{p}$, 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.

$$s_t = \alpha_0 + \alpha_x x'_{t-1} + \alpha_m (p_{t-1} - \bar{p})^2 \tag{10}$$

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

$$x'_t = x'_{t-1} + \left(1 - x'_{t-1}\right) P_{t-1}^{cf} - \left(1 + x'_{t-1}\right) P_{t-1}^{fc} \tag{11}$$

[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.

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 $u_t^c \sim \mathrm{N}\left(0, \sigma_c^2\right)$.

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

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:

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

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.

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

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.

All the ARCH models of daily returns r_t in the benchmark set have the same AR(2) mean equation, 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.

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

Several specifications are included, in order to provide as wide a target as possible. The first and most basic specification for the time-varying variance is the ARCH model. A single version is included, with $p = 5$ lags. It is expected that this will be misspecified for most of the data series, however the intention is to provide some 'low hanging fruit' for the agent-based models in the comparison exercise.

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

The second specification included is a standard GARCH model. As for the ARCH specification, it is not expected to provide the best specification for the daily returns series, but instead 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 ARCH 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 differently 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).

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

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

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

Three versions of each of the specifications (17) to (20) are estimated, corresponding to one, two and three (p, q, r) lags. This results in 13 ARCH family models for each data series, which were estimated using Eviews 8. Two examples of the estimation results are provided in as an illustration, in Tables 8 for the FSTE index and 9 for the S&P500 index. The full set of 18 estimations, which is not included here in the interest of brevity, is available as supplementary material to the paper.

3.2 The model comparison methodology

Before discussing the data used to evaluate these agent-based models of herding, it is important to briefly review 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 (2015) and its supplementary material.

The general spirit of the methodology is to map the data-generating processes of a set of candidate models $\{M_1, M_2, ..., M_m\}$ 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 $y_{t-1}, y_{t-2}, ..., y_{t-L}$. This full set of conditional probabilities forms the transition matrix of the L^{th} order Markov process underlying M_i.

$$P_{M_i}\left(Y_t = y_t \middle| \, y_{t-1}, y_{t-2}, ..., y_{t-L}\right) \tag{21}$$

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

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

Barde (2015) 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 universal information criterion (UIC) similar in spirit to the Akaike information criterion or a log-likelihood:

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

Differences in (23) across models M_i, M_j directly reflect differences in the Kullback and Leibler (1951) divergence between model 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 $\{M_1, M_2, ..., M_m\}$ 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, the context tree weighting (CTW) algorithm of Willems et al. (1995), which is specifically 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, in that the inefficiency cost incurred by having to determine the transition matrix from the data is proven to achieve the theoretical lower bound. As shown by Barde (2015), this central property, referred to as *universality*, allows for the correction of the resulting measurement error in the observation score (22). This justifies the choice of the algorithm as the basis of the methodology and also underpins the choice of name for the criterion.

The CTW algorithm's proven optimal performance stems from the fact that it operates on a binary representation of the data series $\{y_1, y_2, ...y_N\}$, 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.

This is made possible by the way the state space is discretised. Given a choice of bounds

$[b_l, b_u]$ and resolution r, observations can take 2^r distinct states spanning the support determined by the bounds. 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. Each 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.[6]

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 and the probability of an observation being in any of the 2^r states can be 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 a 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.[7] These cover markets in 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 crash, the asian crisis of the late 1990's, the dot-com bubble

[6] 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.

[7] The index series used here are publicly available from the historical prices section of finance.yahoo.com.

of the early 2000s up to the turmoil following the fall of Lehman Brothers in late 2008. 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 events. Table 6 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 a 7-bit resolution, i.e. 128 discrete bins, within the bounds [-0.15,0.15], any observations outside of those bounds being truncated to the bound itself. As seen from the 5^{th} column of Table 6, out-of-bounds observations are not a major problem. The more important aspect is the choice of resolution for the data, i.e. 7 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 (2015), 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. 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.

The discretisation diagnostics are reported in the last 3 columns of Table 6 and show that the 7-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.

In practice, the model comparison exercise takes the form of a sensitivity analysis of

Table 2: Parameter values used in sensitivity analysis

Parameter	Values tested	N° models
	Gilli & Winker (GW)	
σ_x	0.20 - 0.22 - 0.24 - 0.26 - 0.28 - 0.30 - 0.32 - 0.34 - 0.36	
ϵ	(0.5 - 1 - 1.5 - 2 - 2.5 - 3 - 3.5 - 4 - 4.5) $\times 10^{-3}$	729
ρ	0.05 - 0.10 - 0.15 - 0.20 - 0.25 - 0.30 - 0.35 - 0.40 - 0.45	
	Alfarano, Lux & Wagner (ALW)	
ϵ_1	2 - 4 - 6 - 8 - 10 - 12 - 14 - 16 - 18	
ϵ_2	2 - 4 - 6 - 8 - 10 - 12 - 14 - 16 - 18	729
ρ	(1 - 1.5 - 2 - 2.5 - 3 - 3.5 - 4 - 4.5 - 5) $\times 10^{-3}$	
	Franke & Westerhoff (FW)	
χ	0.263 - 1.263 - 2.263 - 3.263 - 4.263	
σ_f	0.282 - 0.532 - 0.782 - 1.032 - 1.332	625
σ_c	1.851 - 2.851 - 3.851 - 4.851 - 5.851	
α_x	0.599 - 0.799 - 0.999 - 1.299 - 1.399	

the GW, ALW and FW models using the parameter values in table 2. Parameters are varied around the calibrated values in table 1 with each possible combination of parameters generating a candidate model for the comparison exercise. For the GW and ALW models, three parameters are varied with seven values each, leading to $3^7 = 729$ candidate models. For the case of the FW model, four parameters are tested with five possible values (in order to keep the overall number of models tractable) resulting in $4^5 = 625$ combinations. The values of those parameters of the GW and FW models not present in Table 2 remain identical to the ones in Table 1. Similarly, 13 sets of simulated series are generated for the ARCH benchmark using the specifications (15) - (20) and the parameter estimates corresponding to each data series. This implies that the simulated data for the ARCH benchmark is specific to each of the 18 data series, which is not the case of the agent-based models, where the parameterisation resulting from the combinations in Tables 1 and 2 are the same for all data series.

Each of the 2096 candidate models is used to produce a simulated series with 500000 observations, which is discretised to a 7-bit resolution 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 18 data series in the second stage of the methodology.

Table 3: UIC scores on financial data series

	GW		ALW		FW		ARCH	
Index	min(UIC)	id	min(UIC)	id	min(UIC)	id	min(UIC)	id
AOI	4.2112	661	3.9141	705	**3.8836**	482	**3.8793***	5
NIKKEI	4.7434	573	4.6175	657	**4.5297**	122	**4.5196***	7
KOSPI	4.7733	407	4.6024	655	**4.5231**	111	**4.5201***	7
ST	4.4938	385	4.2498	722	**4.2029***	235	**4.2198**	3
HS	4.9870	431	4.8461	722	**4.7214**	122	**4.7000***	5
NIFTY	5.0241	421	4.8701	656	**4.7736**	122	**4.7323***	7
DAX	4.7310	573	4.5998	656	**4.4936***	123	**4.4945**	7
CAC	4.7431	573	4.6294	722	**4.5099**	122	**4.5053***	7
FTSE	4.3917	581	4.1450	723	**4.1077***	368	**4.1095**	13
IBEX	4.7660	571	4.6190	576	**4.5211**	123	**4.5155***	7
AEX	4.5845	385	4.3511	713	**4.2915***	368	**4.2975**	3
STOXX	4.5936	500	4.4063	722	**4.3335***	238	**4.3428**	5
IPC	4.8869	431	4.7093	722	**4.6002***	122	**4.6075**	5
DJ	4.3313	385	4.0684	641	**4.0269***	360	**4.0360**	5
S&P 500	4.3772	581	4.1469	723	**4.0830***	365	**4.0976**	3
NASDAQ	4.9462	421	4.8239	655	**4.7167**	122	**4.6952***	1
OEX	4.4145	581	4.2113	723	**4.1512**	368	**4.1473***	7
GSPTSE	4.1547	611	3.8249	722	**3.7996**	479	**3.7947***	5

Bold indicates the best model is in the model confidence set at the 90% level
'*' indicates the best overall model for the series

4 Results

The aggregate results of the model comparisons exercise on each of the 18 series is displayed in Table 3. Its main finding, which is relatively consistent across 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, although there are differences between series. 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 can approach, and in some cases exceed, the performance of the ARCH benchmarks.

The parameter values for the best models identified in Table 3 are provided in Table 4, for those models selected at least twice in the exercise, and Table 7 in the appendix for the full set of identified models. Because the model comparison exercise is a sensitivity

Table 4: MCS model parameters

Gilli & Winker (GW)

σ_q	**0.22**	0.28	0.3	0.3	0.34	0.34
ρ	**0.25**	0.35	0.35	0.4	0.3	0.25
ϵ	**0.0005**	0.0035	0.001	0.002	0.0005	0.001
model id	**86**	385	421	431	573	581
N° series	**0**	3	2	2	3	3

Alfarano, Lux & Wagner (ALW)

ϵ_1	**16**	2	2	18	18
ϵ_2	**4**	14	16	4	6
ρ	**0.0025**	0.005	0.005	0.005	0.005
model id	**308**	655	656	722	723
N° series	**0**	2	2	6	3

Franke & Westerhoff (FW)

χ	1.263	2.263	2.263	**2.263**
σ_f	1.332	1.332	1.032	**0.782**
σ_c	5.851	5.851	5.851	**1.851**
α_x	0.599	0.599	0.999	**1.299**
model id	122	123	368	**388**
N° series	7	2	3	**0**

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

analysis rather than a full calibration, it is important to be cautious in the interpretation of the parameters for any specific series. Even though the original calibrations (in bold) are never matched exactly, for most parameters there is either broad agreement with, or variation around, the original values. The only systematic deviations seem to be in the noise parameters, which are higher than the original calibrations. For GW, $\sigma_s > 0.3$ in nearly every series, above the original value of 0.22. The ρ parameter of ALW, which enters the diffusion term of the fundamentalist share (6) is also higher than the original value. Finally, the clearest example is provided by the noise component of chartist demand σ_c in the FW model, which is much larger than the value obtained in Franke and Westerhoff (2015). The implication of this is discussed below.

As explained in section 3.2, one of the benefits of using the UIC (23) to score the models is that the methodology returns an observation-level vector of scores (22), which can be used to test the statistical significance of these rankings. This is done with the MCS procedure of Hansen et al. (2011) using 1000 replications of the Politis and Romano

Table 5: Size of model confidence set per class of model

Index	$\lvert\mathcal{M}_{90}\rvert$	GW	ALW	FW	ARCH
AOI	336	0	0	327	9
NIKKEI	134	0	0	124	10
KOSPI	235	0	0	225	10
ST	300	0	0	294	6
HS	16	0	0	12	4
NIFTY	14	0	0	9	5
DAX	210	0	0	202	8
CAC	98	0	0	86	12
FTSE	353	0	0	340	13
IBEX	162	0	0	149	13
AEX	364	0	0	354	10
STOXX	320	0	0	310	10
IPC	96	0	0	90	6
DJ	255	0	0	247	8
S&P 500	360	0	0	356	4
NASDAQ	118	0	0	107	11
OEX	353	0	0	343	10
GSPTSE	140	0	0	133	7
N° of models:	2096	729	729	625	13

(1994) block bootstrap.[8] The results, shown in Table 5, confirm that none of the ALW and GW calibrations make it into the confidence set at the 90% confidence level, which is restricted to a subset of the FW models and the ARCH benchmarks.

The second benefit of having an observation-level vector of scores is that the relative performance of models over relatively short time-lengths can also be examined, as illustrated in Figure 1 for three of the series where FW is identified as the best model.[9] The three line plots show the relative scores $\Delta\bar{\lambda}_{i,arch}(r_t)$ 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 $\lambda_i(r_t)$ to evaluate the significance of the resulting average $\bar{\lambda}_i(r_t)$. In order to avoid complicating the figures further and given the rankings in Table 3, the test is only carried out on the FW/ARCH head-to-head comparison and the resulting MCS composition is displayed using the vertical banding.

The observation-level plots reveal two important features not discernable from the aggregate rankings in Table 3. The first is that while they confirm the relative rankings

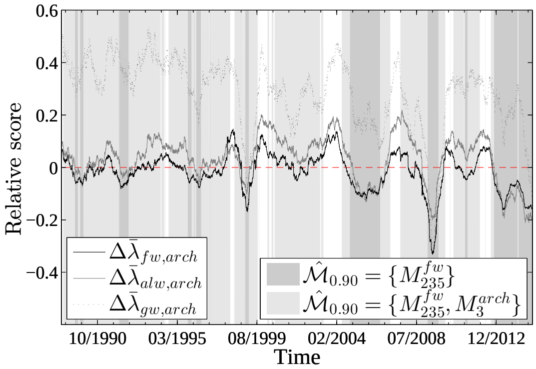
[8] The optimal block length for each series was determined by running the Politis and White (2004) algorithm on the scores prior to performing the bootstrapped analysis

[9] As for the series-specific ARCH estimations discussed in section 3.1, the complete set of plots covering all 18 series is not included here to save space, but is available in the supplementary material.

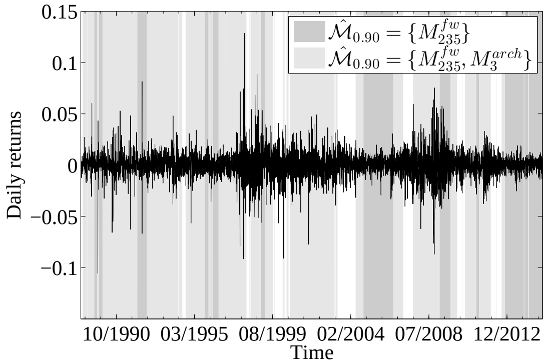
of the three models, in particular the poor performance of GW, they also reveal that the performance of the FW and ALW models is close and their scores relative to the best ARCH model often co-move. This suggests that they both explain similar features of the data and capture similar mechanisms. The second important feature is the presence of clear spikes in performance around turbulent events, in particular the 2008 crash. 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.[10] Crucially, in most cases both the ALW and GW models exhibit similar spikes over the same periods.

Let us summarise at this stage what these findings imply for the various herding mechanisms discussed in section 2. First of all, the spikes in relative scores identified in Figure 1 around critical market events and the co-movement GW and ALW with FW relative to ARCH strongly suggest that herding models offer an important explanation for the dynamics of conditional heteroscedasticity during turbulent events. The poor performance of the GW model compared to ALW and FW, however, indicates that the basic herding mechanism (1) of Kirman (1993) is probably too simplistic. The superior performance of ALW and FW as well as their co-movement on the data series support the hypothesis of Alfarano et al. (2005) that asymmetry is an important part of the herding story, and the lack of this mechanism in GW may explain its performance. Finally, allowing for noise in the chartist demand function rather than pure momentum trading also seems to improve performance. 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 and is entirely driven by noise trading, while in the FW model the chartist demand (12) contains both momentum and noise trading, with Table 4 suggesting a 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.

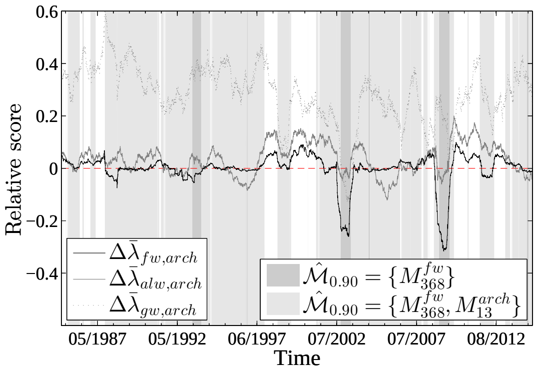
[10]It is also important to note that for most of those series in Table 3 where the ARCH benchmarks beat FW the spike 'events' are absent, which probably explains why FW does not perform better. These diagrams are available in the supplementary material.



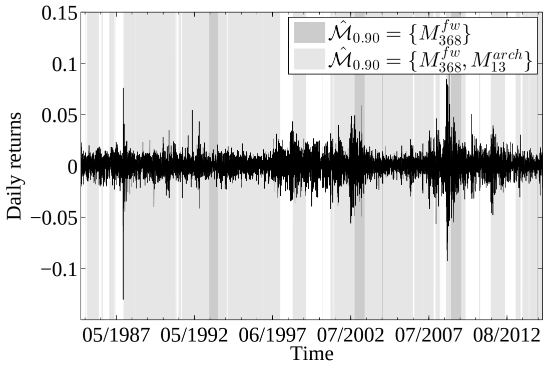
(a) ST - Relative scores



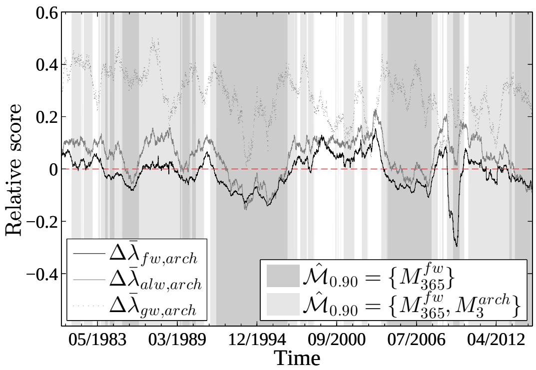
(b) ST - Scored time-series



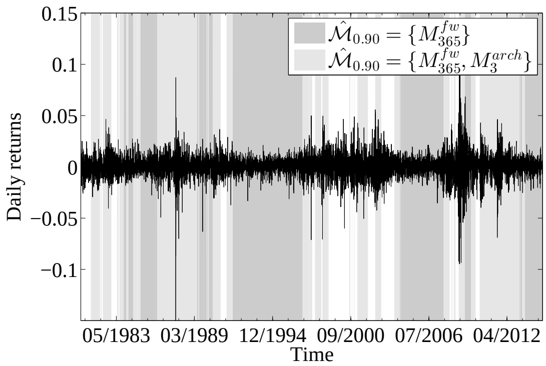
(c) FTSE - Relative scores



(d) FTSE - Scored time-series



(e) S&P500 - Relative scores



(f) S&P500 - Scored time-series

Figure 1: Agent based model scores relative to ARCH benchmark for ST, FTSE and S&P 500

5 Conclusion

Two types of conclusions can be drawn from the model comparison exercise carried out in this paper. The first relates to the agent-based models of herding in financial markets that were being compared, with the results suggesting 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 herding models is better than ARCH-type models on key events where one would expect these stylised facts to be prominent. The results also suggest 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 as 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 is 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 second set of conclusions relates to the model comparison methodology itself. The exercise demonstrates that is possible to effectively compare large numbers of agent-based simulation models to more traditional regression models on the the basis of their simulated output alone. In itself this is not necessarily surprising as a body of work already exists using simulation-based estimation, however, the potential strength of this new methodology rests in the fact that it 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 desirable 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. These aspects are both 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 methodology does possess some limits, which should be the focus of future development work. The first, as pointed out in Barde (2015) is that in its current version the methodology is not designed for estimation, but for comparison of models that already possess calibrated parameters, even if the calibration is poor. This should be visible in the protocol used for comparison, as for each of the three models a 'brute force' grid search is used to test the sensitivity of parameters around existing calibrations. Unsurprisingly, this grid search rapidly encounters the curse of dimensionality, as the FW models can only test five values on four parameters compared to seven values per parameter on the ALW and GW models. The methodology can rank the performance of different models and perhaps perform some basic sensitivity testing, but if the researcher's goal is simply to calibrate a single model, it currently cannot compare for example with the Nelder-Mead simplex search used in Gilli and Winker (2003). 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, 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 indices, but is problematic if multivariate models need to be tested. 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.

References

Alfarano, Simone, Thomas Lux, and Friedrich Wagner (2005) "Estimation of agent-based models: the case of an asymmetric herding model," *Computational Economics*, Vol. 26, pp. 19–49.

Barde, Sylvain (2015) "A Practical, Universal, Information Criterion over Nth Order Markov Processes," *University of Kent School of Economics Discussion Paper 15/04*.

Bianchi, Carlo, Pasquale Cirillo, Mauro Gallegati, and Pietro A. Vagliasin (2007) "Validating and Calibrating Agent-Based Models: A Case Study," *Computational Economics*, Vol. 30, pp. 245–264.

Dawid, Herbet and Giorgio Fagiolo (2008) "Agent-based models for economic policy design: Introduction to the special issue," *Journal of Economic Behavior and Organization*, Vol. 67, pp. 351–354.

Durlauf, Steven N (2005) "Complexity and Empirical Economics," *The Economic Journal*, Vol. 115, pp. F225–F243.

Fagiolo, Giorgio, Alessio Moneta, and Paul Windrum (2007) "A Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems," *Computational Economics*, Vol. 30, pp. 195–226.

Franke, Reiner and Frank Westerhoff (2011) "Estimation of a structural stochastic volatility model of asset pricing," *Computational Economics*, Vol. 38, pp. 53–83.

——— (2012) "Structural stochastic volatility in asset pricing dynamics: Estimation and model contest," *Journal of Economic Dynamics and Control*, Vol. 36, pp. 1193–1211.

——— (2015) "Why a simple herding model may generate the stylized facts of daily returns: explanation and estimation," *Journal of Economic Interaction and Coordination*, Vol. Forthcoming.

Gilli, Manfred and Peter Winker (2003) "A global optimization heuristic for estimating agent based models," *Computational Statistics & Data Analysis*, Vol. 42, pp. 299–312.

Gouriéroux, Christian and Alain Monfort (1993) "Simulation based inference : a survey with special reference to panel data models," *Journal of Econometrics*, Vol. 59, pp. 5–33.

——— (1996) *Simulation-based Econometric Methods*: Oxford University Press.

Hansen, Peter R., Asger Lunde, and James M. Nason (2011) "The Model Confidence Set," *Econometrica*, Vol. 79, pp. 453–497.

Hommes, Cars (2011) "The heterogeneous expectations hypothesis: Some evidence from the lab," *Journal of Economic Dynamics and Control*, Vol. 35, pp. 1–24.

Hommes, Cars H (2006) "Heterogeneous agent models in economics and finance," *Handbook of computational economics*, Vol. 2, pp. 1109–1186.

Kirman, Alan (1993) "Ants, Rationality and Recruitment," *Quarterly Journal of Economics*, Vol. 108, pp. 137–156.

Krichevsky, Raphail E. and Victor K. Trofimov (1981) "The Performance of Universal Encoding," *IEEE Transactions on Information Theory*, Vol. IT-27, pp. 629–636.

Kullback, S. and R. A. Leibler (1951) "On Information and Sufficiency," *Annals of Mathematical Statistics*, Vol. 22, pp. 79–86.

Lamperti, F. (2015) "An Information Theoretic Criterion for Empirical Validation of Time Series Models," *LEM working paper series*, Vol. 2015/02.

Politis, Dimitris N. and Joseph P. Romano (1994) "The Stationary Bootstrap," *Journal of the American Statistical Association*, Vol. 89, pp. 1303–1313.

Politis, Dimitris N. and Halbert White (2004) "Automatic Block-Length Selection for the Dependent Bootstrap," *Econometric Reviews*, Vol. 23, pp. 53–70.

Tesfatsion, Leigh (2006) *Handbook of Computational Economics, Vol. 2: Agent-Based Computational Economics*, Chap. Agent-Based Computational Economics: A Constructive Approach to Economic Theory, pp. 831–880: North-Holland.

Westerhoff, Frank (2009) "Exchange Rate Dynamics: A nonlinear survey," *Handbook of Research on Complexity*, pp. 287–325.

White, Halbert (2000) "A Reality Check For Data Snooping," *Econometrica*, Vol. 68, pp. 1097–1126.

Willems, Frans M. J., Yuri M. Shtarkov, and Tjalling J. Tjalkens (1995) “The Context-tree Weighting Method: Basic Properties,” *IEEE Transactions on Information Theory*, Vol. IT-41, pp. 653–664.

A Extended tables

Table 6: Descriptive statistics and discretisation tests on financial series

Index	Start date	Obs.	Zeros	$\notin$ $[-0.15, 0.15]$	Test 1 KS Stat.	Test 2 LB Stat.	Test 3 LB Stat.
AOI	03/08/1984	7684	60	1	0.0090	23.0947	25.2689
					(0.9149)	(0.5720)	(0.4474)
NIKKEI	04/01/1984	7619	22	3	0.0098	22.8884	25.0043
					(0.8525)	(0.5841)	(0.4621)
KOSPI	04/01/1980	9511	36	0	0.0146	36.1121	23.3394
					(0.2598)	(0.0699)	(0.5578)
ST	31/12/1985	6770	54	0	0.0121	28.5121	24.1152
					(0.7009)	(0.2848)	(0.5127)
HS	02/01/1980	8734	144	4	0.0185*	17.4125	13.4567
					(0.0980)	(0.8663)	(0.9704)
NIFTY	03/07/1990	5891	17	1	0.0100	32.4911	15.4849
					(0.9280)	(0.1443)	(0.9293)
DAX	26/11/1990	6091	20	0	0.0087	27.9607	23.4626
					(0.9747)	(0.3097)	(0.5506)
CAC	01/03/1990	6280	18	0	0.0091	34.8334	22.3324
					(0.9574)	(0.0913)	(0.6165)
FTSE	02/04/1984	7756	19	0	0.0182	19.8943	35.3451
					(0.1525)	(0.7524)	(0.0821)
IBEX	06/09/1991	5863	16	0	0.0109	25.9464	22.6266
					(0.8742)	(0.4105)	(0.5994)
AEX	25/11/1988	6757	180	0	0.0269**	23.6209	20.4041
					(0.0146)	(0.5414)	(0.7253)
STOXX	31/12/1986	7200	22	0	0.0075	22.1629	15.5075
					(0.9871)	(0.6263)	(0.9287)
IPC	08/11/1991	5776	7	0	0.0132	29.3852	19.4519
					(0.6967)	(0.2481)	(0.7750)
DJ	01/10/1928	21665	102	1	0.0060	44.8689***	19.8889
					(0.8292)	(0.0087)	(0.7526)
S&P 500	31/12/1979	9143	207	2	0.0271***	17.5026	19.1225
					(0.0023)	(0.8628)	(0.7912)
NASDAQ	01/10/1985	7363	9	2	0.0125	23.2413	20.2894
					(0.6109)	(0.5635)	(0.7315)
OEX	02/08/1982	8164	28	1	0.0075	15.6071	17.9312
					(0.9762)	(0.9260)	(0.8453)
GSPTSE	31/12/1976	9551	20	0	0.0084	32.0618	34.1488
					(0.8898)	(0.1562)	(0.1048)

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 series, is the 12th of December 2014.

Table 7: MCS model parameters (Full table)

Gilli & Winker (GW)

σ_q	**0.22**	0.28	0.3	0.3	0.3	0.32	0.34	0.34	0.34	0.34	0.36
ρ	**0.25**	0.35	0.1	0.35	0.4	0.25	0.2	0.3	0.25	0.4	0.2
ϵ	**0.0005**	0.0035	0.0005	0.001	0.0015	0.0001	0.0005	0.0005	0.001	0.0025	0.001
model id	**86**	385	407	421	431	500	571	573	581	611	661
N° series	**0**	3	1	2	2	1	1	3	3	1	1

Alfarano, Lux & Wagner (ALW)

ϵ_1	**16**	2	18	2	2	2	14	16	18	18
ϵ_2	**4**	18	4	14	16	18	6	4	4	6
ρ	**0.0025**	0.0045	0.0045	0.005	0.005	0.005	0.005	0.005	0.005	0.005
model id	**308**	576	641	655	656	657	705	713	722	723
N° series	**0**	1	1	2	2	1	1	1	6	3

Franke & Westerhoff (FW)

χ	0.263	1.263	2.263	4.263	2.263	4.263	4.263	2.263	**2.263**	3.263	1.263
σ_f	0.782	1.332	1.332	0.532	0.782	0.532	0.782	1.032	**0.782**	0.282	0.532
σ_c	5.851	5.851	5.851	5.851	5.851	5.851	5.851	5.851	**1.851**	5.851	5.851
α_x	0.599	0.599	0.599	0.799	0.799	0.999	0.999	0.999	**1.299**	1.299	1.299
model id	111	122	123	235	238	360	365	368	**388**	479	482
N° series	1	7	2	1	1	1	1	3	**0**	1	1

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

Table 8: ARCH models estimation for the London Stock Exchange FTSE 100 Index (FTSE)

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	0.0496***	0.0458***	0.0459***	0.046***	0.0249***	0.025***	0.0269***	0.0209**	0.0209**	0.0229**	0.0225**	0.0228**	0.0242**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0098)	(0.0096)	(0.0055)	(0.0253)	(0.0256)	(0.0141)	(0.0195)	(0.0181)	(0.0128)
a_1	0.0201*	0.0117	0.0121	0.0121	0.0158	0.0164	0.0174	0.0168	0.0166	0.0166	0.0177	0.0166	0.0198
	(0.0790)	(0.3315)	(0.3044)	(0.3076)	(0.1867)	(0.1628)	(0.1490)	(0.1466)	(0.1519)	(0.1514)	(0.1381)	(0.1571)	(0.1003)
a_2	−0.0129	−0.0066	−0.0077	−0.0080	−0.0024	−0.0033	−0.0025	0.0006	−0.0004	−0.0043	−0.0004	−0.0021	−0.0012
	(0.2778)	(0.5773)	(0.5281)	(0.5133)	(0.8351)	(0.7817)	(0.8344)	(0.9610)	(0.9720)	(0.7227)	(0.9748)	(0.8550)	(0.9179)
σ_0	0.3457***	0.017***	0.0322***	0.0278***	0.0184***	0.034***	0.001***	−0.126***	−0.2512***	−0.0144***	0.0193***	0.0193**	0.0013***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0008)	(0.0000)	(0.0175)	(0.0032)
α_1	0.1108***	0.0942***	0.0789***	0.0816***	0.0294***	0.0166**	0.0103	0.1575***	0.1572***	0.1153***	0.0788***	0.0516***	0.066***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0491)	(0.1832)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0020)	(0.0000)
α_2	0.1757***		0.0961***	0.0098		0.0387***	0.0094		−0.1072	0.0241		0.0134	−0.0054
	(0.0000)		(0.0000)	(0.4674)		(0.0001)	(0.5209)		(0.0000)	(0.5905)		(0.9966)	(0.9285)
α_3	0.1473***			0.0628***			−0.0156			−0.1213***			−0.0518**
	(0.0000)			(0.0000)			(0.1376)			(0.0008)			(0.0473)
α_4	0.1504***												
	(0.0000)												
α_5	0.1241***												
	(0.0000)												
β_1		0.8912***	0.1031	1.0975***	0.9040***	0.1254	1.1979***	0.9813***	−0.015***	1.4965***	0.9091***	0.933**	1.4485***
		(0.0000)	(0.4298)	(0.0000)	(0.0000)	(0.3858)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0195)	(0.0000)
β_2			0.6943***	−1.0275***		0.6982***	0.2971***		0.9776***	−0.1834		−0.0243	−0.1874
			(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.6964)		(0.9466)	(0.7503)
β_3				0.752***			−0.5019***			−0.3146			−0.2692
				(0.0000)			(0.0000)			(0.1585)			(0.3161)
γ_1					0.0954***	0.0912***	0.1142***	−0.0729***	−0.0753***	−0.0993***	0.4217***	0.9005***	0.6185***
					(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0078)	(0.0000)
γ_2						0.0811***	0.0141		−0.0698***	0.0448		−1.0000	0.9790
						(0.0000)	(0.6478)		(0.0000)	(0.1490)		(0.9975)	(0.9382)
γ_3							−0.1246***			0.0487*			0.6506***
							(0.0000)			(0.0560)			(0.0014)
δ											1.4795***	1.4544***	1.4254***
											(0.0000)	(0.0000)	(0.0000)
AIC	2.7352	2.6917	2.6915	2.6917	2.6771	2.6773	2.6707	2.6780	2.6781	2.6727	2.6759	2.6760	2.6701
BIC	2.7433	2.6970	2.6986	2.7007	2.6834	2.6863	2.6824	2.6842	2.6871	2.6844	2.6831	2.6859	2.6827

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

Table 9: ARCH models estimation for the Standard & Poor's 500 Index (S&P 500)

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	−0.0624***	−0.0574***	−0.0568***	−0.0561***	−0.0787***	−0.0789***	−0.0781***	−0.0873***	−0.0861***	−0.0858***	−0.0844***	−0.0852***	−0.0842***
	(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_1	0.0070	−0.0007	0.0003	0.0000	0.0062	0.0074	0.0079	0.0132	0.0164	0.0163	0.0102	0.0093	0.0096
	(0.5054)	(0.9531)	(0.9758)	(0.9973)	(0.5883)	(0.4993)	(0.4735)	(0.2364)	(0.1057)	(0.1122)	(0.3733)	(0.3949)	(0.3888)
a_2	−0.0125	−0.0103	−0.0116	−0.0115	−0.0113	−0.0117	−0.0107	−0.0111	−0.0117	−0.0118	−0.0099	−0.0114	−0.0107
	(0.2216)	(0.3461)	(0.2862)	(0.2765)	(0.2919)	(0.2837)	(0.3060)	(0.2969)	(0.2669)	(0.2590)	(0.3636)	(0.2986)	(0.3100)
σ_0	0.3756***	0.0134***	0.0095***	0.0095***	0.0075***	0.0058***	0.0062***	−0.101***	−0.0717***	−0.0578**	0.008***	0.0062***	0.007***
	(0.0000)	(0.0000)	(0.0000)	(0.0004)	(0.0000)	(0.0002)	(0.0000)	(0.0000)	(0.0000)	(0.0214)	(0.0000)	(0.0015)	(0.0000)
α_1	0.0896***	0.0804***	0.0401***	0.0447***	0.0045*	−0.0141***	−0.0119***	0.1355***	0.0504***	0.0568***	0.0631***	0.0361	0.0419
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0619)	(0.0008)	(0.0064)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.7943)	(0.9258)
α_2	0.1288***		0.0113	−0.0277*		0.0204***	0.0110		−0.1072	−0.0092		0.0067	−0.0495***
	(0.0000)		(0.2933)	(0.0582)		(0.0000)	(0.2241)		(0.0272)	(0.7488)		(0.9812)	(0.0002)
α_3	0.1414***			0.0367***			0.0064			0.0298			0.0625***
	(0.0000)			(0.0060)			(0.3135)			(0.3235)			(0.0000)
α_4	0.145***												
	(0.0000)												
α_5	0.1711***												
	(0.0000)												
β_1		0.9086***	1.4118***	1.823***	0.9362***	1.3105***	1.9028***	0.9913***	1.3457***	1.7857***	0.9385***	1.3201***	1.8423***
		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
β_2			−0.471***	−1.3057***		−0.3581**	−1.6271***		−0.3523***	−1.1013**		−0.3659**	−1.509***
			(0.0000)	(0.0002)		(0.0145)	(0.0000)		(0.0025)	(0.0149)		(0.0410)	(0.0000)
β_3				0.4214***			0.6706***			0.3103*			0.613***
				(0.0022)			(0.0000)			(0.0578)			(0.0000)
γ_1					0.1157***	0.0933***	0.1029***	−0.089***	−0.1215***	−0.1217***	0.6643***	1.0000	1.0000
					(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.8642)	(0.9507)
γ_2						−0.0137	−0.0999***		0.0632***	0.1223***		−0.9999	0.7716***
						(0.5418)	(0.0000)		(0.0000)	(0.0030)		(0.9877)	(0.0069)
γ_3							0.0912***			−0.0487**			0.5069***
							(0.0000)			(0.0414)			(0.0000)
δ											1.3863***	1.3031***	1.3253***
											(0.0000)	(0.0000)	(0.0000)
AIC	2.7238	2.6737	2.6716	2.6717	2.6455	2.6449	2.6443	2.6438	2.6416	2.6421	2.6431	2.6421	2.6417
BIC	2.7309	2.6784	2.6779	2.6796	2.6511	2.6527	2.6545	2.6494	2.6495	2.6524	2.6494	2.6507	2.6527

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

Recent Kent Discussion Papers in Economics

15/06: 'Efficiency in a forced contribution threshold public good game', Edward Cartwright and Anna Stepanova

15/05: 'Military Aid, Direct Intervention and Counterterrorism', María D.C. García-Alonso, Paul Levine and Ron Smith

15/04: 'A Practical, Universal, Information Criterion over Nth Order Markov Processes', Sylvain Barde

15/03: 'Public Good Provision in Indian Rural Areas: the Returns to Collective Action by Microfinance Groups', Paolo Casini, Lore Vandewalle and Zaki Wahhaj

15/02: 'Fiscal multipliers in a two-sector search and matching model', Konstantinos Angelopoulos, Wei Jiang and James Malley

15/01: 'Military Aid, Direct Intervention and Counterterrorism', María D.C. García-Alonso, Paul Levine and Ron Smith

14/18: 'Applying a Macro-Finance Yield Curve to UK Quantitative Easing', Jagjit S. Chadha and Alex Waters

14/17: 'On the Interaction Between Economic Growth and Business Cycles', Ivan Mendieta-Muñoz

14/16: 'Is there any relationship between the rates of interest and profit in the U.S. economy?', Ivan Mendieta-Muñoz

14/15: 'Group Lending and Endogenous Social Sanctions', Jean-Marie Baland, Rohini Somanathan and Zaki Wahhaj

14/14: 'Tax Reforms in Search-and-Matching Models with Heterogeneous Agents', Wei Jiang

14/13: 'A Fair Wage Explanation of Labour Market Volatility', Robert Jump

14/12: 'Explaining Differences in the Productivity of Capital Across Countries in the Context of 'New' Growth Theory', Kevin S. Nell and A.P. Thirlwall

14/11: 'The Organic Food Premium: A Canterbury Tale', Adelina Gschwandtner

14/10: 'Contrasting the Perception and Response of Domestic Manufacturing Firms to FDI in Sub-Saharan Africa', Penélope Pacheco-López

14/09: 'Euro- US Real Exchange Rate Dynamics: How Far Can We Push General Equilibrium Models?', Aydan Dogan

14/08: 'The Role of Conferences on the Pathway to Academic Impact: Evidence from a Natural Experiment', Fernanda L. L. de Leon and Ben McQuillin

14/07: 'Optimal taxation and labour wedge in models with equilibrium unemployment', Wei Jiang