Optimal moment set selection for the SMM using machine learning

April 13, 2022

Eric Zila[b,a], Jiri Kukacka[a,b,*]

[a] *Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodarenskou vezi 4, 182 00 Prague 8, Czechia*
[b] *Charles University, Faculty of Social Sciences, Institute of Economic Studies, Opletalova 26, 110 00 Prague 1, Czechia*

Abstract

This paper addresses the moment selection issue of the simulated method of moments, which is an estimation technique commonly applied to analytically intractable agent-based models. To reduce the arbitrariness of the choice of moments, we develop a simple machine learning extension on top of this method that offers a novel automated approach to the optimal moment set selection. The idea beyond the extension follows the stepwise regression framework. Two algorithms are proposed: forward stepwise moment selection and backward stepwise moment elimination. The methodology is tested on three models with increasing complexity and a number of estimated parameters. We use one standard time series model and two popular financial agent-based models, and we employ three sets of moments from the literature as benchmarks. We find that both selection algorithms consistently identify multiple moment sets that outperform all benchmark sets for every model. Moreover, for financial agent-based models, where one needs to address moment set selection in practice, we achieve considerable estimation precision enhancements between 9 and 67 percent. Finally, we demonstrate the overidentification issue of too many moments decreasing estimation performance.

JEL: C13, C15, C22, C51, C58, G40.
Keywords: agent-based model, machine learning, simulated method of moments, stepwise selection.

*Corresponding author
Email addresses: 10026254@fsv.cuni.cz (Eric Zila), jiri.kukacka@fsv.cuni.cz (Jiri Kukacka)

1. Introduction

Since the 1960s, researchers have been increasingly aware of the stylized facts observable in empirical financial data. The absence of autocorrelation, clustered volatility, and heavy tails and the emergence of speculative bubbles are among the essential empirical patterns identified in data across all types of financial assets (Cont, 2001). In the search for an explanation of these robust empirical regularities, the field of finance turned to commonly used asset pricing models. However, the traditional approaches have not succeeded in this respect (Hong and Stein, 1999). For instance, traditional capital asset pricing models are unsuccessful at explaining the reoccurring clusters of increased volatility initially described in Ding et al. (1993). Similarly, these models do not reflect the abnormally large number of extreme returns registered on the market observed initially by Fama (1965).

Beginning with Kirman (1993), the field of financial agent-based modeling emerged to face this challenge. Agent-based models in which boundedly rational agents interact and respond to their environment are capable of both replicating the stylized empirical facts and explaining them endogenously (Cont, 2007; Chen et al., 2012). These models have attracted considerable attention in reaction to the global financial crisis between late 2007 and 2009. The mainstream approaches that assume perfect rationality and aggregation of all market participants to a single 'representative' agent were unable to predict the looming collapse of the financial system. More importantly, they failed to provide policy recommendations to reduce the contagion effects and the devastating impact of the housing and stock market drops on the real economy. Since then, many economists have turned to agent-based models for answers regarding sudden downturns and crashes with an immense global impact.

The major challenge for the field of financial agent-based models appears when researchers face their empirical estimation (Lux and Zwinkels, 2018; Fagiolo et al., 2019). Due to their complexity, standard analytical parameter estimation methods, such as the generalized method of moments and the maximum likelihood method, often cease to be viable options. This drawback greatly complicates the testing of statistical hypotheses and the selection between various models of the same phenomenon. Because closed-form solutions are not feasible, more complex models often require computationally demanding simulation-based approaches. In the behavioral finance econometric literature, the most prominent alternative to analytical methods is the simulated method of moments (SMM), which was originally proposed by McFadden (1989) and Pakes and Pollard (1989). Its main advantages are a straightforward and transparent application without restrictive theoretical assumptions, customizability to various types of models, and well-defined mathematical properties.

In essence, the SMM repeatedly simulates the estimated model to obtain its parameters such that the simulated time series and the studied empirical time series are as close as possible. The closeness between the two series is defined as the weighted sum of distances between some of their carefully chosen characteristics, such as the mean and variance, or the first-order autocorrelation. These characteristics can be associated with moments of the data generating processes. The necessity of specifying the moments used in the estimation process and the potential arbitrariness of this choice are among the most concerning problems of the method (Platt, 2020). Additionally, selection of the moments can markedly affect the method's performance. In the literature, the moment set is typically constructed intuitively based on expert knowledge of the estimated model, following a deep study of its dynamics. The moments that, at least in theory, describe the dynamics of the model such that it reflects the stylized empirical facts are selected and used during

estimation. However, no guarantee of optimality of a moment set constructed in this manner can be provided, and the moment set may not contain sufficient moments to completely capture the model's dynamics. Additionally, an overabundance of moments might lead to difficulties in the estimation process because the impact of crucial moments on the minimized function is diminished. Furthermore, some important moments may be overlooked and ignored due to limited knowledge of the characteristics of a model or due to arbitrary decisions of a researcher.

The first simplified application of the SMM to an agent-based model of financial markets can be traced to Gilli and Winker (2003) applying it to the Kirman (1993) model. Complete specification of the method enabling the application to other models is then presented in Winker et al. (2007) and Franke (2009), who applies an alternative version of the method to the Manzan and Westerhoff (2007) model. Many other estimation attempts have followed these initial contributions. For instance, the SMM finds its use in a series of papers starting with the initial model presentation in Franke and Westerhoff (2011), its expansion and generalization in Franke and Westerhoff (2012), and detailed exploration in Franke and Westerhoff (2016). Fabretti (2013) applies the same method to the Farmer and Joshi (2002) model. Other examples of papers studying the application of the SMM to financial agent-based models are Grazzini and Richiardi (2015), Chen and Lux (2018), and Tubbenhauer et al. (2021); we refer to recent excellent surveys on financial agent-based modeling and econometrics by Dieci and He (2018); Lux and Zwinkels (2018) for an overview.

To reduce the arbitrariness of the moments' choice, this paper proposes a novel automated approach to moment set selection. It is a straightforward machine learning method applied on top of the SMM. The extension is based on the stepwise regression framework. Although stepwise regression is outdated as a model selection tool, the simplistic machine learning technique fits the problem at hand perfectly. The proposed methodology should identify an optimal moment set given an arbitrary model in practice. It also fosters an econometric analysis of models for which no intuitive moment set is known. Subsequently, such a moment set can be employed to estimate the model using empirical data.

After introducing the method, its performance is studied using three models with increasing complexity and a number of estimated parameters. These include a simple random walk with a structural break, the popular financial agent-based model of herding by Alfarano et al. (2008), and the Franke and Westerhoff (2012) model, which is a more complex financial agent-based model for studying interactions of fundamental and speculative traders in an asset market. The first model should scrutinize the proposed methodology's functionality and enable an analysis of its general properties. The other two agent-based models are standardly estimated using the SMM in practice. Therefore, the application of the methodology has a direct practical implication, especially for the Franke and Westerhoff (2012) model, because no closed-form solution of the model exists, which means that standard estimation approaches are infeasible. Finally, three moment sets from the literature are included in the testing as benchmarks.

Recently, following a seminal paper by Salle and Yıldızoğlu (2014) on the kriging interpolation method, Lamperti et al. (2018); Bargigli et al. (2020); Zegadło (2021); Zhang et al. (2021); Chen and Desiderio (2022) were among the first to directly introduce machine learning methods to financial agent-based modeling and validation. However, they primarily take advantage of these approaches to simplify complex models, a process generally called meta-modeling or surrogate modeling, while employing intelligent/efficient sampling, e.g., for parameter space exploration and calibration purposes, not for methodological advancements of existing estimation methods. Platt (2022) then extends the approach of Grazzini et al. (2017) by applying deep neural networks to

approximate the likelihood within a Bayesian estimation of the Brock and Hommes (1998) and Franke and Westerhoff (2012) models.

The paper proceeds as follows. The next Section 2 introduces the SMM and implementation details. Section 3 then presents the proposed machine learning extension, the selection criterion, and the full moment set. In Section 4, the three testbed models together with their parameterizations are described, and Section 5 summarizes our numerical setup and important pre-estimation analyses. Next, Section 6 presents the results of a Monte Carlo simulation analysis and quantifies performance improvements over the benchmark moment sets, while Section 7 provides important sensitivity checks. In Section 8, we summarize the general properties of the estimation methodology and discuss the patterns observable across all models. Finally, Section 9 concludes the paper. Technical details and additional results are relegated to the Appendix.

2. Simulated method of moments

The simulated method of moments (SMM, also referred to as the method of simulated moments in the literature) was initially introduced by McFadden (1989) and Pakes and Pollard (1989) as a modification of the traditional method of moments. It enables the estimation of models to which analytical estimation approaches and numerical integration methods cannot be applied, such as multinomial discrete response models and a large portion of complex agent-based models. While the method applies to any model, it is used primarily in situations where closed-form solutions do not exist due to the high computational requirements. Lee and Ingram (1991) expand on the estimation framework by utilizing the method in time series model estimation and describing asymptotic properties of the estimator given additional constraints on the error term. The method is extended in Duffie and Singleton (1993) for the estimation of panel data models. As two examples of the more recent development of the SMM, Franke (2009) proposes abandoning the weighting matrix in the objective function and substituting it with the t-statistics of the moments. Franke and Westerhoff (2012) then introduce the so-called joint moment coverage ratio, i.e., the percentage of simulation runs of which all moments lie within their empirical confidence intervals, as an alternative evaluation criterion.

The objective of the method is to match the sample counterparts of the selected moments of the simulated data as closely as possible to the sample counterparts of the moments of the empirical data by optimizing the model parameters. The basic principle is that the method yields a point estimate of the vector of parameters of interest by minimizing the weighted sum of distances between a selected set of moments of simulations approximating the population and the empirical moments. Two approaches can be taken to obtain the standard errors of the point estimate. It is possible to use approximation methods to compute the point estimate's standard errors (Franke, 2009). Alternatively, Monte Carlo simulations can be employed on top of the SMM, and a large number of point estimates can be generated and then used to estimate the standard errors. This approach ensures that a large amount of additional information about the dynamics of the model and the estimator's performance is generated. We take advantage of this newly collected information in the proposed machine learning-inspired methodology.

2.1. Formal definition

Consider an empirical time series $\{y_t\}$, $t = 1, \ldots, T_{emp}$ stored in the form of a column vector denoted y. We can define a function $m(y)$ for computing the sample counterpart of a moment of

choice m for the empirical time series vector. To give an example, if m represents the mean, the function would be defined as follows:

$$m(y) = \frac{1}{T_{emp}} \sum_{t=1}^{T_{emp}} y_t, \tag{1}$$

where y_t represents the t-th element of vector y. In this case, we apply the fact that the arithmetic average is commonly used as a sample counterpart to the mean. For more complex moments, such as the 4^{th}-order autocorrelation of squared values, we define the function as follows:

$$m(y) = \frac{\sum_{t=5}^{T_{emp}} \left[(y_t^2 - \overline{y^2})(y_{t-4}^2 - \overline{y^2}) \right]}{\sum_{t=1}^{T_{emp}} (y_t^2 - \overline{y^2})^2}, \tag{2}$$

where $\overline{y^2}$ is the arithmetic average of the series with each element squared. Assume we choose a total of D moments of interest and denote their sample counterpart functions as $m_d(y),\ d = 1, \ldots, D$. Subsequently, we can apply these functions to the empirical time series and organize the results into a $D \times 1$ moment vector of empirical data as follows:

$$m^{emp} = [m_1(y^{emp}), \ldots, m_D(y^{emp})]'. \tag{3}$$

Suppose we have a fully specified, stochastic model $f(\cdot)$ that, according to our beliefs, represents the true description of some aspect of the real world that generated the empirical time series $\{y_t\}$. Assume we can use the model to generate a simulated time series $\{y_t(\theta)\},\ t = 1, \ldots, T_{sim}$, where θ is a vector of true parameters of the model we are trying to estimate stored in a column vector denoted $y^{sim}(\theta)$. To moderate the effect of randomness when calculating the moment vector of simulated data, we use multiple random draws of the stochastic part of the model, each resulting in one simulated time series vector. We can organize these vectors into a matrix $Y^{sim}(\theta) = \left[y_1^{sim}(\theta), \ldots, y_N^{sim}(\theta) \right]$, where N is the total number of simulated time series and $y_i^{sim}(\theta)$ is the i-th simulated time series vector. $Y^{sim}(\theta)$ is then a $T_{sim} \times N$ matrix. Subsequently, we can define a function for computing the simulated sample counterpart of a moment for each of the D moments of interest as follows:

$$m_d^{sim}(Y^{sim}(\theta)) = \frac{1}{N} \sum_{n=1}^{N} m_d(Y_{\cdot n}^{sim}(\theta)),\ d = 1, \ldots, D, \tag{4}$$

where $Y_{\cdot n}^{sim}(\theta)$ represents the n-th column of matrix $Y^{sim}(\theta)$. Similar to the moment vector of the empirical data, the $D \times 1$ moment vector of the simulated data becomes as follows:

$$m^{sim}(\theta) = \left[m_1^{sim}(Y^{sim}(\theta)), \ldots, m_D^{sim}(Y^{sim}(\theta)) \right]'. \tag{5}$$

The process of matching moments is based on minimizing the distance between m^{emp} and $m^{sim}(\theta)$. We can define a function for computing a vector of differences between the two for each pair of moments as follows:

$$h(\theta) = m^{emp} - m^{sim}(\theta). \tag{6}$$

The distance is calculated as a weighted sum of squares of elements of the vector. In matrix form, we can write this as follows:

$$J(\theta) = h(\theta)' W h(\theta), \tag{7}$$

where W is a positive definite weighting matrix of size $D \times D$. Finally, the SMM estimator is defined as follows:

$$\hat{\theta} = \arg\min_{\theta} J(\theta), \tag{8}$$

where $\theta \in \Theta$, Θ is the parameter space of the search.

2.2. Weighting matrix

There are multiple approaches to the calculation of weighting matrix W. In all cases, however, the aim is to assign greater importance to moments that are stable across samples. For a moment with a high sampling variability, a high difference between the empirical and simulated values of the moment should be considered less critical than if such a difference occurs for a moment with low variability. In addition, information about possible correlations between individual moments should be included in the matrix. In the ideal case scenario, we construct the matrix from the variance-covariance matrix Σ of the moments such that $W = \Sigma^{-1}$.

However, since the true values of the matrix elements are unknown, it is necessary to estimate Σ. In our implementation, we follow Franke and Westerhoff (2011, 2012). Their model represents the most challenging testbed framework on which we test the methodology proposed in this paper. Specifically, we utilize a continued block-bootstrap approach to create new samples from the original pseudo-empirical sample generated within our Monte Carlo simulation analysis. We use these new bootstrapped samples to estimate the frequency distributions of the moments, which we subsequently apply to estimate the variances of individual moments and covariances between them. It is necessary to select data blocks instead of individual observations because many moments traditionally used for the SMM estimation of financial models are lag dependent. Assume that we create a total of B samples using the block-bootstrap approach. Using each sample, we can generate a total of B moment vectors of bootstrapped data m^b, $b = 1, \ldots, B$. Then, we can create an estimate of the variance-covariance matrix, as follows:

$$\hat{\Sigma} = \frac{1}{B} \sum_{b=1}^{B} (m^b - \overline{m})(m^b - \overline{m})', \tag{9}$$

where $\overline{m} = \frac{1}{B} \sum_{b=1}^{B} m^b$. As outlined previously, we can then compute the weighting matrix W by taking the inverse of $\hat{\Sigma}$.

3. Moment selection using machine learning

We now provide a complete specification of the proposed machine learning-based approach to selecting the moment set used for the SMM. In the past, researchers addressed the problem intuitively based on their expert knowledge of the estimated model and the data generating process it represents. This approach, however, raises questions regarding the optimality of a moment set selected in this manner and the application of the SMM itself (Platt, 2020). Jalali et al. (2015) note that the number of moments should be at least as large as the number of estimated parameters for a given model. Furthermore, the authors suggest that any other moments above this lower boundary should allow for better identification; thus, their presence should improve the parameter estimates. However, Chen and Lux (2018) find that more extensive sets of moments may result in worse fits than their subsets. In general, the set should include moments that describe all the

Algorithm 1 Forward stepwise moment selection

$f(\cdot) \sim$ estimated model
$\theta_{true} \sim$ pseudo-true parameters selected for the estimated model
$y \sim$ pseudo-empirical time series generated by applying parameters θ_{true} to model $f(\cdot)$
$M_F \sim$ full set $\{m_1, \ldots, m_D\}$ of all moment functions

procedure FSMS
    $M_S \leftarrow$ empty set $\varnothing$ of moment functions
    while $M_S \neq M_F$ **do**
        for all moments m **in** $M_F \setminus M_S$ **do**
            $M_T \leftarrow M_S \cup \{m\}$
            conduct SMM for time series y using model $f(\cdot)$ and moment functions in M_T
            compute quality measure using results of SMM and parameters θ_{true}
        end for
        $M_S \leftarrow M_T$ such that quality measure is optimized across all **for loop** cycles
    end while
end procedure

essential dynamics of the model. However, for complex models, finding and defining such moments are highly complicated tasks.

Stepwise regression is one of the primary algorithms used for model selection. Furthermore, it represents one of the earliest and most straightforward uses of machine learning techniques in econometrics. Two basic approaches lie at its core: forward stepwise selection and backward stepwise elimination. The former relies on an iterative addition of variables to a regression model initially composed of only the intercept. The latter successively excludes variables from a model that initially contains all variables at its disposal. Here, we propose a similar application of stepwise techniques to the moment set choice in the SMM estimation. In both cases, the procedure minimizes a metric representing the overall goodness of fit along its path. The immediate success of these two approaches spurred their further development and improvement in several directions. Among the most important is a bidirectional technique choosing from all possible variable expansions and eliminations at each step. Another augmentation allows for multiple variables to be added to or excluded from the model in each step (Hastie et al., 2009).

3.1. Forward stepwise moment selection

The first algorithm proposed is forward stepwise moment selection (FSMS). The intuition behind it closely follows the forward stepwise selection of variables in stepwise regression. Here, the subject of iterative expansion is the moment set used for the SMM estimation. We begin with an empty moment set. In each step, we try to expand it by every moment that is not already in the collection. The expansion yielding the most precise estimate given a pseudo-empirical time series is added. The final step involves all the moments used in the estimation process at once. Subsequently, the evolution of fits and the moment additions can be studied to choose the optimal moment set for the model at hand. Concerning Chen and Lux (2018), the main expectation is that many subsets outperform the full set of moments.

The algorithm written in pseudo-code can be found in Algorithm 1. A total of four structures must be present in the environment where the algorithm is executed. First, a function $f(\cdot)$ for

generating a time series using the model must be at its disposal (line 1). We would not be able to analyze the model without such a function that accepts the parameters of the model as its input. Second, we have to choose pseudo-true parameter values θ_{true} of the estimated model that we subsequently use to generate pseudo-empirical data (line 2). The pseudo-true parameter values of the data-generating process must be known to establish the quality of each estimation result. The selection may be arbitrary; however, parameterizations from the literature are preferred, as they ensure that the model generates desired stylized facts. Third, the pseudo-empirical dataset $y = f(\theta_{true})$ must be generated before engaging in the procedure (line 3). Last, a set of D moment functions $M_F = \{m_1, \ldots, m_D\}$ from which the procedure chooses is required (line 4). Furthermore, it is necessary to execute the SMM estimation process using any specified set of moment functions.

At the beginning of the procedure, we initiate M_S as an empty set (line 7). The set is subsequently extended by one additional moment function in each selection round of the algorithm. In essence, we consider a selection round to be each round of the `while` cycle representing the foundation of the algorithm (line 8). The `while` cycle continues for as long as the set M_S does not contain every single moment from the full moment function set M_F. In other words, the `while` cycle does not stop until the set M_S contains D elements. Within each selection round, the procedure iterates over all moment functions from the full set M_F that are not already contained in the set M_S using a `for` cycle (line 9). In each iteration, a temporary set M_T is created such that it contains all moment functions from the set M_S and one additional moment function from the full set M_F not contained in the set M_S that the `for` cycle currently iterates over (line 10). Next, the parameters of the model $f(\cdot)$ are estimated using the SMM applied to the pseudo-empirical time series y while utilizing moment functions contained in M_T (line 11). Based on the estimation results and the true parameter values θ_{true}, a measure of quality of estimation is then computed (line 12). Finally, the moment function set M_T generating the best results in terms of the estimation quality measure is selected and used to replace the set M_S for the next selection round of the procedure (line 14).

3.2. Backward stepwise moment elimination

The second algorithm proposed is backward stepwise moment elimination (BSME). Overall, the algorithm works in the same manner as FSMS; however, in this case, the initial moment set contains all the available moments. In each iteration, one moment is removed from the moment set such that the fit achieved for the pseudo-empirical time series is the best possible. The logic behind this choice is that the moment removed adds the smallest amount of information for the estimation method. Again, the procedure should reveal moments with the highest value for estimation.

Pseudo-code of BSME is presented in Algorithm 2. Only subtle differences exist between the two algorithms. The first difference is that the environment where the algorithm is executed does not need to contain the full set M_F of moment functions. Instead, the set M_S is initiated to contain all the moment functions (line 6). Second, the `while` cycle represents elimination rounds of the algorithm cycles until the size of the set M_S is smaller than or equal to one (line 7). In other words, the algorithm continues for as long as there is at least one moment to be left out of the set M_S while not causing the set to become empty. Additionally, the `for` cycle of each selection round iterates over all moments in the set M_S (line 8). By doing so, the procedure establishes the impact of the elimination of each moment function on the overall estimation performance. Finally, the temporary set M_T is created by removing the moment that the `for` cycle currently iterates from the set M_S (line 9). Apart from these few differences, the algorithm remains the same.

Algorithm 2 Backward stepwise moment elimination

$f(\cdot) \sim$ estimated model
$\theta_{true} \sim$ pseudo-true parameters selected for the estimated model
$y \sim$ pseudo-empirical time series generated by applying parameters θ_{true} to model $f(\cdot)$

procedure BSME
    $M_S \leftarrow$ full set $\{m_1, \ldots, m_D\}$ of all moment functions
    while SIZE$(M_S) > 1$ **do**
        for all moments m **in** M_S **do**
            $M_T \leftarrow M_S \setminus \{m\}$
            conduct SMM for time series y using model $f(\cdot)$ and moment functions in M_T
            compute quality measure using results of SMM and parameters θ_{true}
        end for
        $M_S \leftarrow M_T$ such that quality measure is optimized across all **for loop** cycles
    end while
end procedure

3.3. Selection criterion

While describing the algorithms, we do not elaborate on the metric used to differentiate between the results of estimations employing different sets of moments. At the same time, it is clear that it has a crucial impact on the selection process. We use the widely accepted root mean square error (RMSE). Traditionally, the metric is calculated in the context of a single parameter. Therefore, we make a slight adjustment to the RMSE that accommodates the necessity of involving multiple parameters in the metric simultaneously.

Assume we have a model with a total of K estimated parameters. Given this model, we may generate a simulated time series using a concrete set of parameters. We organize these parameters into the form of a vector $\theta = [\theta_1, \ldots, \theta_K]$. Subsequently, we apply the SMM to the simulated time series. Depending on the number of Monte Carlo replications, we obtain a certain number of point estimates for each parameter. From these point estimates, we compute the parameter estimates as their median values across all replications. The parameter estimates form a vector $\hat{\theta} = \left[\hat{\theta}_1, \ldots, \hat{\theta}_K\right]$. Additionally, we can use the point estimates to estimate the standard error of each parameter. The single parameter normalized RMSE_{norm} is then calculated as follows:

$$RMSE_{norm} := RMSE_{norm}(\hat{\theta}_i) = \sqrt{\frac{(\hat{\theta}_i - \theta_i)^2 + \hat{\sigma}_i^2}{|\theta_i|}}, \tag{10}$$

where $i = 1, \ldots, K$ and ${\hat{\sigma}_i}^2$ is the estimated variance of the SMM estimator of the parameter θ_i. Traditionally, only the nominator containing the bias term $(\hat{\theta}_i - \theta_i)^2$ and the variance term $\hat{\sigma}_i^2$ is used as the metric. Division of the metric by the true parameter in absolute value represents a normalization step. As a result, it ensures that parameters of different magnitudes can be compared and aggregated into a single value. Subsequently, we can compute the metric for multiple parameters as follows:

$$RMSE_{\hat{\theta}} := RMSE_{norm}(\hat{\theta}) = \frac{1}{K} \sum_{i=1}^{K} RMSE_{norm}(\hat{\theta}_i). \tag{11}$$

In terms of the practical application of the methodology, the specification of the selection criterion may be adjusted such that penalization follows individual goals. To provide an example, if lowering the bias was the primary task of the process, the impact of the bias term on the metric could be amplified by adding weights to the two terms involved in $\text{RMSE}_{\hat{\theta}}$.

3.4. Full moment set

The presented approaches need to be supplied with an initial full set of moments used in the selection/elimination process. Theoretically, there is an infinite number of moments that can be used. However, due to the computational complexity of the SMM, the initial set should be of a reasonable size. Assume that one wants to use D_F initial moments. Then, for FSMS, there are a total of $D_F + (D_F - 1) + \cdots + 2 + 1 = \frac{D_F(D_F+1)}{2}$ moment sets for which the estimation is conducted. For BSME, the total number is one less because the initial full set with D_F moments is not evaluated. Thus, the computation of results for a vast number of moment sets is avoided, which is in stark contrast to the naive procedure of 'best-subset selection,' which considers all possible combinations of the moment set. This is because 'greedy' stepwise procedures can be expected to choose only among the most promising moment sets. Given D_F moments, there are a total of $\binom{D_F}{1} + \binom{D_F}{2} + \cdots + \binom{D_F}{D_F-1} + \binom{D_F}{D_F} = 2^{D_F} - 1$ different moment sets that would be computed within the best-subset selection algorithm. For $D_F = 15$ moments, the FSMS procedure computes

Table 1: Moments sets: The full set and benchmarks

Moment function	*Label*	*CHL4*	*FW9*	*CHL15*
$m_1 = E(r_t^2)$	RAW-VAR	✓		✓
$m_2 = E(r_t^4)$	RAW-KURT	✓		✓
$m_3 = E(r_t r_{t-1})$	RAW-AC1	✓	✓	✓
$m_4 = E(\|r_t\|)$	ABS-MEAN		✓	
$m_5 = Hill(\|r_t\|, 5)$	ABS-HILL		✓	
$m_6 = E(\|r_t\|\|r_{t-1}\|)$	ABS-AC1		✓	✓
$m_7 = E(\|r_t\|\|r_{t-5}\|)$	ABS-AC5		✓	✓
$m_8 = E(\|r_t\|\|r_{t-10}\|)$	ABS-AC10		✓	✓
$m_9 = E(\|r_t\|\|r_{t-15}\|)$	ABS-AC15			✓
$m_{10} = E(\|r_t\|\|r_{t-20}\|)$	ABS-AC20			✓
$m_{11} = E(\|r_t\|\|r_{t-25}\|)$	ABS-AC25		✓	✓
$m_{12} = E(\|r_t\|\|r_{t-50}\|)$	ABS-AC50		✓	
$m_{13} = E(\|r_t\|\|r_{t-100}\|)$	ABS-AC100		✓	
$m_{14} = E(r_t^2 r_{t-1}^2)$	SQR-AC1	✓		✓
$m_{15} = E(r_t^2 r_{t-5}^2)$	SQR-AC5			✓
$m_{16} = E(r_t^2 r_{t-10}^2)$	SQR-AC10			✓
$m_{17} = E(r_t^2 r_{t-15}^2)$	SQR-AC15			✓
$m_{18} = E(r_t^2 r_{t-20}^2)$	SQR-AC20			✓
$m_{19} = E(r_t^2 r_{t-25}^2)$	SQR-AC25			✓

Note: Benchmark naming is based on the underlying papers: Chen and Lux (2018) for *CHL4* and *CHL15*; Franke and Westerhoff (2012) for *FW9*.

the results for only 120 of 32,767 total available moment sets. The comparison becomes even more stark as the number of moments increases. Assuming $D_F = 50$ moments, the results for only $1,275$ moment sets must be calculated out of a total of 1.1×10^{15} possible combinations.

We turn to two papers from the financial agent-based models literature to choose the moments included in the initial full set. In Franke and Westerhoff (2012), the SMM is applied to the model described in Subsection 4.3. In Chen and Lux (2018), the estimation procedure is studied in connection to a simple financial agent-based model introduced in Alfarano et al. (2008). Both models capture the main stylized facts of financial data, specifically the absence of autocorrelation of raw returns, heavy tails of the distribution, clustered volatility, and potentially long memory patterns. In both papers, the moment sets used for the SMM are constructed to capture the dynamics behind these stylized facts. However, even though the same stylized empirical patterns are supposed to be captured by the sets, they differ significantly between the papers. Three moment sets, one from Franke and Westerhoff (2012) and two from Chen and Lux (2018), are selected as benchmarks for the proposed SMM extension. Most recently, identical sets of moments were applied for the approximate Bayesian computation method to the Alfarano et al. (2008) model in Lux (2021a). The full set provided to the algorithms is then the union of the moment sets used in the two papers. It is fully specified in Table 1.

Table 1 thus defines the complete set of moment functions m_i, $i = 1, \ldots, 19$ used to test the machine learning extension of the SMM. The moment selection includes the unconditional second and fourth moments of raw returns, the first-order autocorrelation of raw returns, the unconditional first moment of absolute returns, the Hill estimator of the tail index of absolute returns based on 5% of extreme observations, i-th-order autocorrelations, $i \in \{1, 5, 10, 15, 20, 25, 50, 100\}$ of absolute returns, and j-th-order autocorrelations, $j \in \{1, 5, 10, 15, 20, 25\}$ of squared returns.

4. Testbed models

This section presents three models to which we apply the FSMS and BSME algorithms to identify optimal moment sets for the SMM. Due to the need for substantial computing power, the SMM is used predominantly in situations where analytical estimation techniques are infeasible. Therefore, to provide evidence of the functionality of the proposed methodology and to observe its general properties, three models with an increasing level of complexity (Lee et al., 2015; Mandes and Winker, 2017; Kukacka and Kristoufek, 2020, 2021) and an increasing number of estimated parameters are selected. The first testbed model is a simple random walk model with a structural break. As the proposed algorithms should mostly find their use in the estimation of analytically unsolvable agent-based models, a popular simple financial model of herding by Alfarano et al. (2008) with three estimated parameters and a more complex structural stochastic volatility model by Franke and Westerhoff (2012) with seven estimated parameters are also considered.

4.1. Random walk with a structural break

The baseline model used to test the fundamental functionality of the proposed methodology is an extension of the random walk model with drift that enables replication of a simple structural break in both the drift and volatility occurring at a certain period in time. The price evolution is defined as follows:

$$p_t = p_{t-1} + d_t + \varepsilon_t, \;\; \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \tag{12}$$

where d_t is the drift term, as follows:

$$d_t, \sigma_t = \begin{cases} d_1, \sigma_1 & \text{if } t \leq \tau, \\ d_2, \sigma_2 & \text{if } t > \tau \end{cases} \tag{13}$$

and τ is the time period in which the structural break occurs.

In contrast to the two agent-based models described below, a random walk with a structural break is not commonly used to depict a real-world system. Rather, it is used to provide a challenge for newly introduced methodologies and metrics. Its specification presented here follows Lamperti (2018), who uses the model as a testing exercise for GLS-div, a model selection criterion. The author appreciates the extremely low computational burden of the model along with its ability to represent stochastic properties of financial and macroeconomic time series. Moreover, the model mimics instability, one of the traditional characteristics of agent-based models often involving sudden shifts and tipping points in their systems. For the same reasons, we see the model being used as one of the benchmarks for the likelihood-based Bayesian estimation in Platt (2020) and its extension utilizing neural networks in Platt (2022).

The parameterization and optimization constraints follow Platt (2022, 2020). The true values of the two estimated parameters are $d_1 = 0.4$ and $d_2 = 0.7$. The other parameters of the model are set such that $\tau = 0.7 \cdot T_{sim}$, $\sigma_1 = 1$, and $\sigma_2 = 2$. Finally, price differences are used as the model output for the estimation routine.

4.2. Alfarano, Lux, and Wagner (2008) model

Originally developed and applied through a series of papers (Alfarano et al., 2005, 2006, 2007), the second framework on which we test the methodology is the model of herding by Alfarano et al. (2008) built in the tradition of the 'ant dynamics' introduced by Kirman (1991, 1993). This model was selected following the recent estimation efforts by Ghonghadze and Lux (2016); Chen and Lux (2018); Lux (2021a, 2018, 2021b), where it standardly serves as a testbed for the performance of the SMM, generalized method of moments (GMM), approximate Bayesian computation, and new advanced methods based on the Markov chain Monte Carlo approach, such as sequential Monte Carlo based on a particle filter or the adaptive particle Markov chain Monte Carlo.

The model assumes two subpopulations of market participants: N_f fundamentalists and N_c noise traders and a log fundamental value F_t of the underlying asset. The average trading volume of fundamentalists is V_f, and their excess demand thus follows $D^f = N_f V_f (F_t - p_t)$, where p_t is the equilibrium market log-price at time t. The behavior of noise traders is determined by an opinion dynamics process, where an optimistic/pessimistic state of a given agent implies buying/selling V_c units of the asset at time t. The authors define the population sentiment index x_t as equal to zero for balanced sentiment, while it reaches positive/negative values when noise traders are dominantly optimistic/pessimistic: $x_t = \frac{2n_{o,t}}{n_c} - 1 \in \langle -1, 1 \rangle$, where $n_{o,t}$ is the number of optimistic noise traders at time t.

The opinion dynamics process is associated with the subpopulation of noise traders who enter pairwise communication and switch between the two states. The Poisson intensity of autonomous switches $a \geq 0$ and the herding rate $b \geq 0$ drive the dynamics of the model. Setting $b >> a$ leads to a bimodal distribution of the sentiment index x_t, i.e., one observes a strong and persistent majority opinion with an episodic mean-reverting tendency among noise traders over time, as pairwise communication governs sentiment dynamics. Conversely, setting $a >> b$ leads to a unimodal unconditional distribution of x_t centered around 0 as autonomous opinion switches dominantly

govern the sentiment with a stronger mean-reverting tendency. The model further assumes a Walrasian price adjustment mechanism with instantaneous market clearing. The continuous-time version of the model is summarized as follows:

$$\begin{aligned} \mathrm{d}F_t &= \sigma_f \mathrm{d}B_{1,t}, && (14)\\ \mathrm{d}x_t &= -2ax_t\mathrm{d}t + \sqrt{2b(1-x_t^2)+\frac{4a}{n_c}}\mathrm{d}B_{2,t}, && (15)\\ p_t &= F_t + \frac{N_c V_c}{N_f V_f}x_t, && (16)\\ r_t &= p_t - p_{t-1} \equiv \sigma_f e_t + \frac{N_c V_c}{N_f V_f}(x_t - x_{t-1}), && (17) \end{aligned}$$

where F_t follows standard Brownian motion without drift, $\sigma_f \geq 0$ represents the standard deviation of its innovations, $B_{1,t}$ and $B_{2,t}$ are independent Wiener processes, r_t is the log-return at time t, and $e_t \sim \mathcal{N}(0,1)$ due to the fundamental value for a unit time change $F_{t+1} \sim \mathcal{N}(F_t, \sigma_f)$.

The parametrization follows Ghonghadze and Lux (2016); Chen and Lux (2018); Lux (2018, 2021b,a) and the Langevin equation approximation of the model. The optimization constraints follow Lux (2021a). The true values of the three estimated parameters are $a = 0.0003$, $b = 0.0014$, and $\sigma_f = 0.03$. The other parameters of the model are set such that $n_c = 100$ and $\frac{n_c v_c}{n_f v_f} = 1$. Naturally, log-returns are used as the model output in our estimation exercises.

4.3. Franke and Westerhoff (2012) model

As the last model, we consider a structural stochastic volatility model originally presented in Franke and Westerhoff (2011) and further developed by Franke and Westerhoff (2012). It describes a market where two population fractions of agents, chartists and fundamentalists, interact and respond to their dynamic environment by switching between the two groups. Since its adoption, the model has been used in many estimation exercises (Franke and Westerhoff, 2016; Lux, 2021b; Platt, 2022, to name few).

The price of an asset traded on the market is governed by demand. Specifically, a market maker changes the log-price p_t from one period to the next by observing the excess demand for the asset and proportionately adjusting the price such that $p_t = p_{t-1} + \mu(n_{t-1}^f d_{t-1}^f + n_{t-1}^c d_{t-1}^c)$, where d_t^f and d_t^c describe the excess demands of fundamentalists and chartists, respectively, n_t^f and $n_t^c = 1 - n_t^f$ represent the corresponding population share of each group, and μ stands for the adjustment rate of the asset price by the market maker. Depending on the current trading approach of a trader, their excess demand is governed by one of the following equations:

$$\begin{aligned} d_t^f &= \phi(p^* - p_t) + \varepsilon_t^f \qquad \varepsilon_t^f \sim \mathcal{N}(0, \sigma_f^2), && (18)\\ d_t^c &= \chi(p_t - p_{t-1}) + \varepsilon_t^c \qquad \varepsilon_t^c \sim \mathcal{N}(0, \sigma_c^2), && (19) \end{aligned}$$

where p^* is the fundamental log-value of the asset, ε_t^f and ε_t^c represent the noise terms for fundamentalists and chartists, respectively, σ_f and σ_c determine the corresponding volatilities of the noise terms, and ϕ and χ are the respective adjustment parameters of the demands. The excess demands are governed by simplistic rules used throughout the literature. The excess demand of fundamentalists is driven by the deviation of the current asset price from its fundamental value;

whereas the excess demand of chartists recognizes the change in price of the asset from the previous period to the current period. Adjustments of population shares are governed by the index a_t indicating the propensity to switch such that:

$$a_t = \alpha_n(n_t^f - n_t^c) + \alpha_0 + \alpha_p(p_t - p^*)^2, \tag{20}$$

$$n_t^f = \frac{1}{1 + \exp(-\beta a_{t-1})}, \tag{21}$$

where $\alpha_n \geq 0$ is a 'herding' parameter, α_0 is a 'predisposition' parameter representing a general trading preference towards one of the strategies, and $\alpha_p \geq 0$ is a price 'misalignment' parameter measuring the impact of mispricing from the fundamental log-value. Finally, β is the intensity of choice within the popular binomial logistic discrete choice approach.

Franke and Westerhoff (2012) present an entire family of models varying in the approach to population fraction switching and its determinants. Regarding the former, the transition probability approach (TPA) of Franke and Westerhoff (2011) is complemented by the discrete choice approach (Brock and Hommes, 1998, DCA) presented in (21). As for the determinants of the switching index, accumulated past profits of alternative trading strategies called wealth (W) are included among herding (H), predisposition (P), and misalignment (M) terms that we use in (20). Consequently, various subsets of the four determinants can be utilized to create multiple models. Since the authors indicate that the DCA-HPM setup combination is superior to the rest, we use this combination in this paper.

The parameterization, which follows Franke and Westerhoff (2012), is also used in Platt (2022), whom we also follow in terms of the optimization constraints: $\phi = 0.12$, $\chi = 1.5$, $\sigma_f = 0.758$, $\sigma_c = 2.087$, $\alpha_n = 1.79$, $\alpha_0 = -0.327$, $\alpha_p = 18.43$. The remaining parameters are set such that $\mu = 0.01$, $p^* = 0$, and $\beta = 1$. Again, log-returns are used as the model output for the estimation routine.

5. Monte Carlo simulation study

Although all three testbed models and their parameterizations are well known and verified in the recent literature, we first perform their basic time-series analysis. Next, we conduct a detailed preliminary analysis to find a suitable simulation setup that optimizes computational time without compromising the estimation precision of the SMM. This is evaluated via $\text{RMSE}_{\hat{\theta}}$ for the most challenging estimation exercise with the Franke and Westerhoff (2012) model, which we can also directly compare to the performance of the likelihood-based Bayesian method utilizing neural networks by Platt (2022).

5.1. Analysis of the warm-up period

First and foremost, we conduct a test of the length of the warm-up period to ensure sufficient discarding of the initial observations to eliminate any influence of the initial conditions. We follow Welch's method (Welch, 1983) recently discussed in Vandin et al. (2021). Based on 1,000 independent realizations of the process, a warm-up period of length 100 appears sufficient to provide both the Alfarano et al. (2008) and Franke and Westerhoff (2012) models enough time to stabilize their dynamics. A random walk with a structural break naturally cannot pass this procedure, but for its trivialized version without/until the structural break, a warm-up of 100 initial observations would also be sufficient. For cautionary reasons, we implement a multiple-length warm-up period of 2,000 initial observations in the subsequent analysis.

5.2. General simulation setup for all models

The analysis is executed in `Julia 1.6.1`, and we take advantage of the `Distributed.jl` package for parallel computing. The length of the analyzed time series is always $T_{emp} = T_{sim} = 6,000$ for all three models after discarding 2,000 initial warm-up periods. Six thousand observations are intended to represent more than two decades of daily financial time series data. Interestingly, while for the double length $T_{emp} = T_{sim} = 12,000$ the computational time also almost doubles, the SMM estimation precision increases by only approximately 3%.

5.3. Descriptive statistics and the analysis of stationarity and ergodicity

Typical time series outputs of the testbed models are depicted in Appendix Figure A.4. Table A.8 further displays aggregate descriptive statistics, their 95% sample confidence intervals, and rejection rates of six essential time-series statistical tests based on 1,000 random runs and the general simulation setup. A random walk model with drift but no structural break is then added for comparison to represent a 'well-behaving' normally distributed process. As expected, all three models violate normality, demonstrated through excess kurtosis and confirmed statistically by 100% rejection rates in the Jarque-Bera test. For all three models, stationarity is supported by the augmented Dickey-Fuller test. Nevertheless, only for the two agent-based models is it also not rejected by the Kwiatkowski-Phillips-Schmidt-Shin test. This is, again, an expected result, as the structural break in the random walk process naturally contradicts the definition of stationarity.

Finally, we conduct two ergodicity tests according to Grazzini (2012) and Delli Gatti and Grazzini (2020). While the latter supports ergodicity for all models, the 'older' Grazzini (2012) test clearly rejects the null hypothesis for the random walk model without/until the structural break. This is due to the differences in the technical implementation of the tests: while in the Grazzini (2012) test, one extremely large time series is cut to many subsamples, uncentered moments of which are compared to the moments of originally generated short time series using the Wald-Wolfowitz runs test, in the 'new' version of the test, the generated time series are all the same size. Thus, only Delli Gatti and Grazzini (2020) can handle the structural break, while for the Grazzini (2012) test, the existence of the break directly corrupts the idea of the test.

5.4. General Monte Carlo and SMM estimation setup

We evaluate each Monte Carlo experiment based on 96 independent replications. While the parallel server capacities at our disposal allow for a largely sufficient computation budget for most of the standard tasks, the simulation setup is dictated by the enormously computationally demanding Monte Carlo executions of the FSMS and BSME machine learning algorithms. Nevertheless, while for 192 runs the computational time naturally doubles, the SMM estimation precision increases by less than 1%.

The SMM is implemented with the continued block-bootstrapped weighting matrix $W = \hat{\Sigma}^{-1}$ with potentially overlapping blocks of 250 observations and the bootstrap size $B = 5,000$ (Franke and Westerhoff, 2011, 2012). For the SMM optimizations, simulated moments are computed as averaged values from 100 independent simulations. Interestingly, again, while for more simulations, the computational time increases linearly, for 250 simulations, the estimation precision increases by not even 3%. Finally, a default recommended differential evolution optimizer from the `BlackBoxOptim.jl` package is used. The optimization algorithm was tested against competing optimizers available in other packages, such as the standard `Optim.jl`, and it delivers the best performance for the given optimization problem. Technically, the optimization is constrained by

4,000 functional evaluations. Based on our preliminary analysis, allowing for more evaluations only linearly increases the computational time without delivering better estimation precision for the SMM.

6. Analysis of results

6.1. Random walk with a structural break

As the baseline model, we first interpret the results for the random walk with a structural break. The full specification of the model along with its parameterization can be found in Subsection 4.1. We estimate two parameters, d_1 and d_2, representing the drift before and after the structural break. Similar to Lamperti (2018) and Platt (2020, 2022), we apply the proposed methodology to this model as a challenge because the structural break extension notably increases the estimation difficulty of the model. At the same time, it shares an important commonality with agent-based models in the occurrence of large systemic shifts in model dynamics.

Two machine learning processes: Figure 1, panels (a) and (b), present the results of the two machine learning processes, FSMS and BSME. In each selection round of the FSMS algorithm, one moment is added to the initially empty moment set such that the set achieving the lowest $\text{RMSE}_{\hat{\theta}}$ in the given round determines the expansion. Therefore, the number of moments included in a set coincides with the selection round for FSMS. For BSME, this relationship is reversed, i.e., the moments are gradually eliminated. Thus, the first round of the procedure includes sets containing the largest number of moments, namely, 18, and each subsequent round lowers the number of moments in the corresponding moment sets by one.

Three benchmark moment sets: Immediately, it is possible to see that the performance of all three benchmark sets taken from Franke and Westerhoff (2012); Chen and Lux (2018) is not nearly as good as that of most of the moment sets identified by both algorithms. The further inferiority of the *CHL4* and *CHL15* benchmark sets compared to *FW9* is caused by the absence of the mean of absolute returns in the two sets. This is because for a simpler model without the structural break, i.e., the random walk with drift, the drift parameter should be fully explained by the mean, which asymptotically approaches its value. As the full moment set contains the unconditional first moment of absolute returns (ABS-MEAN according to Table 1), it can be expected that FSMS chooses this moment in the early stages of the selection process and BSME avoids removing it until the end, also for the extended version with a structural break. Moreover, the benchmarks from Chen and Lux (2018) can be expected to underperform due to the absence of this moment. Additionally, the *CHL15* benchmark achieves only slightly better $\text{RMSE}_{\hat{\theta}}$ compared to *CHL4*, while containing many more moments, in stark contrast to the results for the other two models presented in panels (c)–(f). This means either that the additional moments contribute no additional explanatory power or that the impact of the important moments is diminished by the unimportant ones.

FSMS, BSME: Looking at panel (a), one can notice a non-negligible variance in moment set performance in the first round of FSMS, with set #4 strongly outperforming the other moment sets. Set #4, with a trivial size of one and the $\text{RMSE}_{\hat{\theta}}$ slightly below 0.4, represents exactly the unconditional mean of absolute returns (ABS-MEAN), as expected above. For the second round, which considers 18 different moment sets involving two moments, their dispersion even increases, but the chosen set #30 is only slightly superior to some of the other sets. Furthermore,

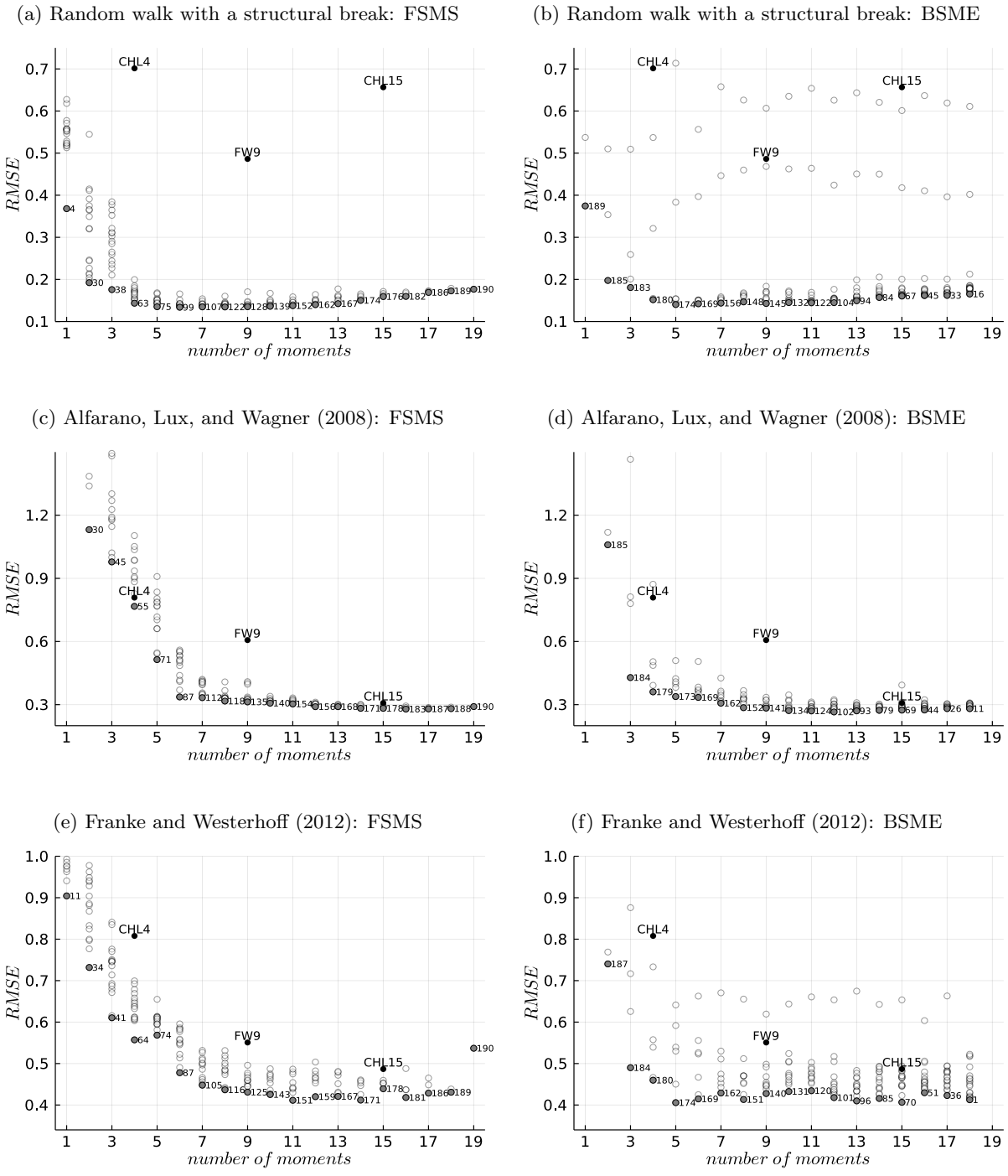

Figure 1: Results for FSMS and BSME. *Note:* Transparent points depict results for all moment sets involved in the selection process, numbered gray points depict results for the best performing moment sets of each selection round, and named black points depict results for the benchmark moment sets (see Table 1). For visualization purposes, moment sets with high $\text{RMSE}_{\hat{\theta}}$ compared to the benchmark sets are excluded. The number of moments included in a set coincides with the selection round for FSMS; for BSME, this relationship is reversed: the results of the first BSME round are associated with 18 moments, the second round with 17 moments, etc.

Table 2: Estimation improvement measured by $\text{RMSE}_{\hat{\theta}}$: Overall best sets

		CHL4	*FW9*	*CHL15*
Random walk with a structural break				
Benchmark	$\text{RMSE}_{\hat{\theta}}$	0.701	0.486	0.656
FSMS—Set #99 (6)	0.134	80.9%	72.4%	79.6%
BSME—Set #174 (5)	0.140	80.0%	71.2%	78.7%
Alfarano, Lux, and Wagner (2008)				
Benchmark	$\text{RMSE}_{\hat{\theta}}$	0.809	0.607	0.308
FSMS—Set #183 (16)	0.279	65.5%	54.0%	9.4%
BSME—Set #102 (12)	0.265	67.2%	56.3%	14.0%
Franke and Westerhoff (2012)				
Benchmark	$\text{RMSE}_{\hat{\theta}}$	0.808	0.551	0.487
FSMS—Set #151 (11)	0.411	49.1%	25.4%	15.6%
BSME—Set #174 (5)	0.406	49.8%	26.3%	16.6%

Note: Numbers in parentheses after set numbers indicate the set size. Estimation improvements compared to the benchmark moment sets are reported in %.

its estimation precision increases markedly as the $\text{RMSE}_{\hat{\theta}}$ falls below 0.2. The next round follows the same trend, with selected set #38 clearly outperforming most of the other sets. However, from round no. 4, which considers moment sets with four moments onward, the $\text{RMSE}_{\hat{\theta}}$ values of estimates for different moment sets become very similar and almost indistinguishable for the larger moment sets. This follows the notion that the unconditional mean of absolute returns together with a few other moments capture the dynamics of the model very well, while other moments provide almost no added value to the SMM estimation.

In a similar fashion, the results in panel (b) for BSME exhibit low variation in the $\text{RMSE}_{\hat{\theta}}$ values of the estimates. However, two moment sets in each selection round perform much worse than all the other sets. These are depicted by the two curve patterns of transparent points: the upper pattern sets follow the $\text{RMSE}_{\hat{\theta}}$ performance of the *CHL4* and *CHL15* benchmark sets, and the bottom pattern follows the performance of the *FW9* benchmark. Importantly, the upper pattern represents moment sets for which the mean of absolute returns (ABS-MEAN) is eliminated. The bottom patterns correspond in all cases to moment sets, excluding the second moment of raw returns (RAW-VAR).

U-shaped pattern for the best performing sets: The performance of the best sets in each selection round is logically tied to the number of included moments. Consequently, the best moment sets form a flat but noticeable U-shaped pattern, especially for FSMS in panel (a). Therefore, adding moments to the moment set makes the estimation less efficient from a certain point on. Such a phenomenon is supported by the increased variation in the $\text{RMSE}_{\hat{\theta}}$ values in panel (b) for

larger moment sets. The observed behavior follows the intuition that for many moments, the impact of the most important ones on the minimized function is diminished, which causes difficulties in the optimization process.

Machine learning evolution and the crucial moments: The complete evolution of the selected moment sets in each round of the selection/elimination process is as follows:

- FSMS: ABS-MEAN → ABS-AC50 → RAW-VAR → ABS-AC25 → ABS-AC5 → SQR-AC25 (end of the best set #99) → ABS-AC100 → SQR-AC10 → ABS-AC1 → ABS-AC10 → SQR-AC5 → SQR-AC20 → ABS-AC20 → SQR-AC1 → RAW-KURT → ABS-HILL → ABS-AC15 → SQR-AC15 → RAW-AC1,
- BSME: SQR-AC10 → SQR-AC1 → ABS-AC10 → SQR-AC5 → SQR-AC20 → ABS-AC20 → ABS-HILL → ABS-AC100 → ABS-AC25 → SQR-AC25 → RAW-AC1 → RAW-KURT → SQR-AC15 → ABS-AC15 → (beginning of the best set #174) ABS-AC50 → ABS-AC1 → RAW-VAR → ABS-AC5 → ABS-MEAN.

The presented order, coded according to labels in Table 1, shows that while for FSMS the mean of absolute returns (ABS-MEAN) is selected first (moment set #4), for BSME, it also survives until the very end of the BSME elimination process (moment set #189). A more detailed comparison of the evolutions for the two selection approaches extends our discussion above of the two remarkable patterns in panel (b) and confirms that in addition to the mean, the variance of raw returns (RAW-VAR) has crucial importance for the SMM estimation precision. This moment is the third one included in the set by FSMS and among the last three moments surviving BSME. This choice is not random but rather reflects the impact of the other two calibrated parameters of the model, and the volatilities of the noise terms σ_1 and σ_2, on the dynamics of the model. Finally, some autocorrelations of absolute returns, namely, ABS-AC50 and ABS-AC5, appear to be the most important for capturing the model dynamics, as they are selected among the first five within FSMS and survive until the last five in BSME.

Estimation improvement for the overall best sets: Table 2 compares the performance of the benchmark moment sets and the overall best moment sets found by each machine learning algorithm across all selection/elimination rounds. Based on the previous interpretation, these sets clearly outperform all benchmark sets. The $\text{RMSE}_{\hat{\theta}}$ corresponding to the *CHL4* and *CHL15* benchmark sets reach 0.701 and 0.656, respectively, and the *FW9* benchmark achieves a $\text{RMSE}_{\hat{\theta}}$ value of 0.486. We observe the most significant improvement of approximately 80% against the *CHL4* and *CHL15* sets because of their problematic applicability to the model due to the absence of the mean of absolute returns. In the case of the *FW9* benchmark set, the improvements are lower but still noticeably large, above 70%. As our full moment set is derived from the estimation literature on financial agent-based models, it is possible that better results for this standard time series model could be achieved with an expanded full moment set. Overall, FSMS successfully finds a slightly better performing moment set #99 ($\text{RMSE}_{\hat{\theta}}$=0.134) than does BSME, for which the elimination procedure leads to #174 ($\text{RMSE}_{\hat{\theta}}$=0.140).

Interestingly, for the identical model setup, Platt (2022, Tab. 3) achieves the best estimation performance with $\text{RMSE}_{\hat{\theta}}$=0.221, recomputed according to (11), using the likelihood-based Bayesian method utilizing a mixture density neural network.

6.2. Alfarano, Lux, and Wagner (2008) model

The second exercise involves the simple financial agent-based Alfarano et al. (2008) model of herding. A detailed description and parameterization of the model can be found in Subsection 4.2. In the exercise, we estimate all three model parameters that are typically of interest: the intensity of autonomous switches $a \geq 0$, the herding rate $b \geq 0$, and the standard deviation of fundamental innovations $\sigma_f \geq 0$. The model successfully captures key stylized facts of financial data, namely, the absence of autocorrelation, clustered volatility, and heavy-tailed distributions (Chen et al., 2012). As opposed to the random walk with a structural break, since various moment conditions for this model can be derived analytically, one can potentially apply standard GMM estimation (Ghonghadze and Lux, 2016). However, the complexity of the model due to interactions between an ensemble of individual agents calls for a more flexible SMM approach. Thus, the results obtained for this model have practical implications, as one has to specify the moment set in empirical practice (Chen and Lux, 2018).

Two machine learning processes: The results shown in Figure 1, panels (c) and (d), partially resemble those obtained for the random walk with a structural break. However, there is a sizable difference in the quality of individual estimates. Especially for FSMS, the increased complexity of the model requires more crucial moments to be included to sufficiently describe the dynamics of the model. This results in an initial steep decline in $\text{RMSE}_{\hat{\theta}}$ in panel (c).

Three benchmark moment sets: Compared to the previous model, the performance of benchmarks is clearly positively related to the number of parameters they contain. While the *CHL4* benchmark appears strongly inferior, the *CHL15* set achieves a performance comparable to that of the best moment sets identified by our method. This result is not surprising, as the *CHL15* benchmark set has been suggested and thoroughly analyzed by one of the authors of the model (Chen and Lux, 2018). Still, the inferiority of *CHL4* is markedly larger than reported in their paper, which might be due to differences in the estimation procedure. The performance of the *FW9* benchmark lies between these two, but, by far, falls short of the performance of the best moment sets. Again, this is an expected result, as the *FW9* moment set has been designed for a different model. Finally, there is a seemingly disproportionately large difference between the performance of the *CHL4* benchmark and the best moment sets identified by the FSMS and BSME procedures. While for FSMS the *CHL4* benchmark set is comparable to the best set of the same size and outperforms other combinations of four moments, for BSME, all best sets of length greater than or equal to three clearly outperform *CHL4*. The explanation of this difference lies in the inherent differences between the two procedures, the selection vs. elimination, and the potential interplay between different moments. This phenomenon is further elaborated in Subsections 8.2 and 8.3.

FSMS, BSME: As shown in Table 2, an overall better comparative performance of the benchmark sets to other moment sets does not mean that there are no improvements to be made, as the best identified moment sets always attain better results. Focusing on panel (c) for FSMS, it is clear that sets containing up to five moments do not offer nearly enough information for the estimation of the model. While the addition of the sixth moment produces a severe performance enhancement, the improvement curve achieves relative flatness in the subsequent steps. Nevertheless, one can observe slight performance increases for the best performing moment sets up to round no. 16. Starting from selection round no. 7, a sudden decrease in the volatility of estimation performance for different moment sets is also observed. In addition, from round no. 10 on, the volatility almost disappears. Moreover, as noted from the first elimination round of the BSME process, there is no

curve pattern of transparent points with high $\text{RMSE}_{\hat{\theta}}$, as observed in panel (b) for the random walk model. This finding suggests that, instead of a few parameters crucial to the SMM performance for the previous model, there is no dominant moment for the Alfarano et al. (2008) model. Interactions of multiple moments are, thus, necessary for proper estimation. This claim is further supported by the performance of the benchmark moment set *CHL15* taken from Chen and Lux (2018), which almost reaches the estimation performance of the best moment sets identified by the two machine learning algorithms.

Machine learning evolutions and crucial moments:

- FSMS: ABS-AC100 → ABS-AC25 → ABS-AC10 → RAW-VAR → RAW-KURT → ABS-MEAN → SQR-AC25 → ABS-AC20 → SQR-AC20 → ABS-AC15 → SQR-AC15 → ABS-HILL → SQR-AC5 → ABS-AC1 → ABS-AC50 → SQR-AC1 (end of the best set #83) → SQR-AC10 → RAW-AC1 → ABS-AC5,
- BSME: ABS-AC25 → ABS-AC5 → ABS-AC10 → SQR-AC20 → ABS-AC50 → ABS-AC20 → RAW-AC1 → (beginning of the best set #102) SQR-AC25 → SQR-AC10 → ABS-AC15 → SQR-AC1 → SQR-AC15 → SQR-AC5 → ABS-HILL → ABS-AC1 → ABS-AC100 → RAW-VAR → RAW-KURT → ABS-MEAN.

A detailed focus on the machine learning evolutions of selected/eliminated moments reveals the inherent differences between the two pairwise algorithms. It also explains the absence of clearly dominant moments. For FSMS, three high-order autocorrelations of absolute returns, namely, ABS-AC100, AC25 and AC10, are selected first (moment sets #13 [not displayed due to high $\text{RMSE}_{\hat{\theta}}$ compared to the benchmark sets], #30, and #45), thereby reflecting a more complex nature of the model output and the importance of volatility clustering and potentially long memory for model dynamics. The variance and heavy tail measures follow these, the variance and kurtosis of raw returns (RAW-VAR and RAW-KURT) for #65 and #71, respectively. The addition of the sixth moment, the mean of absolute returns (ABS-MEAN) in set #87, then leads to the last considerable performance boost, which is followed by only tiny improvements as the moment set grows further.

In contrast, the BSME algorithm first eliminates exactly those moments selected first by FSMS: ABS-AC25 and ABS-AC10, together with the other four high-order autocorrelations in the first six rounds (ABS-AC25 and AC10; SQR-AC20). ABS-AC100, RAW-VAR, RAW-KURT, and ABS-MEAN, however, survive the elimination process until the last rounds (set #179), together with another autocorrelation of absolute returns, namely, ABS-AC1, and another measure of heavy tails, namely, the Hill estimator (set #169).

Estimation improvement for the overall best sets: According to Table 2, the overall best moment sets again outperform all benchmark sets, although the *CHL15* benchmark's estimation performance is only slightly worse. The $\text{RMSE}_{\hat{\theta}}$ corresponding to the benchmarks reaches 0.809, 0.607, and 0.308 for *CHL4*, *FW9*, and *CHL15*, respectively, for the Alfarano et al. (2008) model. The best sets #183 ($\text{RMSE}_{\hat{\theta}}$=0.279) and #102 ($\text{RMSE}_{\hat{\theta}}$=0.265) improve the performance over the *CHL4* benchmark by 65.5% and 67.2% for FSMS and BSME, respectively, while for *CHL15*, the performance is improved by only 14.0% and 9.0%. The performance against the *FW9* benchmark lies between these two, and the improvements against it reach 54.0% and 56.3%.

Interestingly, Lux (2018, Tab. 1) achieves the best performance with $\text{RMSE}_{\hat{\theta}}$=0.229, re-computed according to (11), using sequential Monte Carlo based on a particle filter, while Lux (2021a, Tab. 1, Method 3, Method 2) achieves the best performance with $\text{RMSE}_{\hat{\theta}}$=0.188 and $\text{RMSE}_{\hat{\theta}}$=0.260 using the approximate Bayesian computation method.

6.3. Franke and Westerhoff (2012) model

The final model to which our selection/elimination procedures are applied is the Franke and Westerhoff (2012) model described in Subsection 4.3. In the literature, various subsets of the complete parameter set containing seven parameters are typically estimated. However, our ultimate ambition is to estimate the complete parameter set at once. Again, in practice, the model uses the SMM for its estimation and, consequently, must address the problem of moment set specification. Additionally, the performance enhancement over the *FW9* benchmark moment set provides valuable information about the capabilities of the proposed methodology since this benchmark set was explicitly constructed to estimate the given seven parameters of the model using the SMM. As the analysis naturally shares commonalities with the results for the Alfarano et al. (2008) model, in this section, we primarily focus on them, as well as on the differences between the two.

Two machine learning processes: According to Figure 1, panels (e) and (f), patterns presented within the analysis for the previous model seem to be intensified. This is likely due to the increased complexity of the estimation process for a more complex model with a larger number of estimated parameters. Specifically, the moment sets estimated as part of the BSME procedure have considerable variation that is stable across all elimination rounds. Furthermore, the volatility of the estimation performance in the selection rounds of FSMS, although also markedly large compared to that of the Alfarano et al. (2008) model, follows a decreasing trend with each additional moment. Especially for FSMS, it becomes necessary for more moments to gradually come into play to explain the dynamics of the model, resulting in a relatively smooth decrease in $\text{RMSE}_{\hat{\theta}}$ in panel (e).

Three benchmark moment sets: Again, all three benchmarks are noticeably outperformed by most moment sets identified by the selection procedures. Surprisingly, this also holds for moment sets containing a relatively low number of moments, even 4 or 5. This result is emphasized graphically as the benchmarks almost form an 'upper envelope' above all the estimated sets of sizes three and more. This is because while the *CHL15* benchmark set continues to outperform the other benchmarks, it performs much worse than for the previous model. Similarly, for the *CHL15* vs. *FW9* benchmark set, the improvement is much smaller, primarily because the *FW9* benchmark was explicitly designed for the given model.

FSMS, BSME: Panel (e) for FSMS reveals an apparent smooth decrease in $\text{RMSE}_{\hat{\theta}}$ with additional moments up to the moment set #151 consisting of 11 moments and a slight increase in the metric thereafter. This means that multiple moments gradually add to the explanatory power of the expanding set up to some saturation level from which the redundant ones diminish the impact of the important ones. In contrast, the best sets in the corresponding elimination rounds of BSME manifest much more erratic behavior with several ups and downs. While panel (e) shows a declining trend of the volatility of estimation performance with additional moments, in panel (f), as opposed to all previous results, no pattern of volatility is immediately visible for the BSME procedure. Overall, both approaches manifest a higher degree of estimation volatility in the selection and elimination rounds compared to both previously analyzed models, which is attributed to the significantly increased complexity of the model (Kukacka and Kristoufek, 2020) and the estimation procedure for the seven parameters.

Finally, panel (f) for BSME contains a curve pattern of inferior transparent points with high $\text{RMSE}_{\hat{\theta}}$, similar to those observed in panel (b) for the random walk model. However, there is a

slightly different cause for the pattern this time, as it does appear before the variance of raw returns (RAW-VAR) is eliminated in the first BSME round. Then, from the second elimination round on, the pattern appears based on the additional elimination of the mean of absolute returns (ABS-MEAN). These are the exact moments, as in the random walk case. Nevertheless, in contrast, sole elimination of ABS-MEAN does not cause such apparent inferiority of the moment set performance due to its evident interplay with RAW-VAR.

U-shaped pattern for the best performing sets: For the random walk model, the results for FSMS in panel (e) reveal a tendency for a U-shaped pattern, especially up to selection round no. 15. Concretely, adding moments after round no. 11 increases the $\text{RMSE}_{\hat{\theta}}$. The intuition behind this pattern is the same as before. The impact of the important moments is diminished by redundant moments, which decrease the efficiency of the estimation process.

Machine learning evolution and the crucial moments:

- FSMS: ABS-AC25 → SQR-AC10 → ABS-MEAN → ABS-AC50 → ABS-HILL → RAW-KURT → ABS-AC15 → ABS-AC5 → RAW-VAR → SQR-AC15 → SQR-AC1 (end of the best set #151) → ABS-AC100 → SQR-AC5 → ABS-AC1 → ABS-AC20 → RAW-AC1 → SQR-AC20 → SQR-AC25 → ABS-AC10,
- BSME: RAW-VAR → SQR-AC20 → SQR-AC5 → SQR-AC25 → SQR-AC15 → ABS-AC50 → RAW-AC1 → ABS-AC20 → ABS-AC15 → ABS-AC5 → ABS-AC25 → SQR-AC10 → SQR-AC1 → ABS-AC10 → (beginning of the best set #174) ABS-AC100 → ABS-AC1 → ABS-HILL → ABS-MEAN → RAW-KURT.

Similar to the Alfarano et al. (2008) model, the selection process of FSMS starts with two high-order autocorrelations of returns, namely, ABS-AC25 and SQR-AC10, being selected first (moment sets #11 and #34). This preference highlights the impact of volatility clustering and potentially long memory on the dynamics of the model output. These two are followed by the mean of absolute returns (ABS-MEAN, set #41), another high-order autocorrelation (ABS-AC50, set #64), and two measures of heavy tails, the Hill estimator and kurtosis of raw returns (ABS-HILL and RAW-KURT) for #74 and #87, respectively. Interestingly, the selection of the Hill estimator worsens the estimation performance until combined with the kurtosis in the next selection step. The further additions are followed by small but gradual improvements in the $\text{RMSE}_{\hat{\theta}}$ up to round no. 11. Importantly, in contrast to the previous model, the variance RAW-VAR does not appear to be of crucial importance.

This is confirmed by the BSME process, which first eliminates precisely this moment (RAW-VAR, set #1), followed by two autocorrelations of squared returns that also remain excluded from the relatively large overall best set #151 based on the FSMS procedure, SQR-AC20 and SQR-AC5, for #36 and #51, respectively. Similar to the behavior observed for the Alfarano et al. (2008) model, BSME further leaves SQR-AC25 (set #70), which is, conversely, selected first by FSMS, and the same holds for the eliminated ABS-AC50 after the next two rounds (set #96). Almost identically to the previous model, ABS-AC100, ABS-AC1, ABS-HILL, ABS-MEAN, and RAW-KURT, in that order, make it to the end of the elimination process. The only difference is that the variance (RAW-VAR) is now naturally missing because it was eliminated first.

Estimation improvement for the overall best sets: All overall best moment sets outperform all benchmark sets to a large extent. The $\text{RMSE}_{\hat{\theta}}$ corresponding to the benchmarks reaches 0.808, 0.551, and 0.487 for *CHL4*, *FW9*, and *CHL15*, respectively. FSMS identifies set #151

(RMSE$_{\hat{\theta}}$=0.411) as the best moment set, and BSME chooses set #174 (RMSE$_{\hat{\theta}}$=0.406) as the overall best set. The differences in the RMSE$_{\hat{\theta}}$ values across selection processes are thus small, even for this more complex model. Both algorithms achieve a considerable improvement of approximately 26% over the *FW9* benchmark set and approximately 16% over the *CHL15* set. Naturally, the performance of the *CHL4* benchmark is inferior to that of these two.

Interestingly, Platt (2022, Tab. 7) achieves the best performance with RMSE$_{\hat{\theta}}$=0.404, recomputed according to (11), using the likelihood-based Bayesian method utilizing a neural network.

7. Sensitivity checks

This section presents three sensitivity checks challenging our general setups. First, we add a total of 11 new redundant moments to the original set and let our methodology face this unnecessarily large set of moments. Second, we compare the impact of alternative weighting matrices on the estimation performance under consideration of our benchmarks and overall best moment sets. Finally, we focus on the performance of the best sets containing the same number of moments as each benchmark set.

7.1. Machine learning effectiveness under 30 moments

The 11 new uncontributive moments are presented in Table 3. We mostly select moments that are 'localized' between already utilized moments; thus, their explanatory power to a large extent overlaps with that of our original full moment set. This characteristic holds especially for the profile of the autocorrelation function of absolute returns, which is regularly slowly decaying. Thus, the original range should already sufficiently represent its shape. The new moments include the unconditional third, fifth, and sixth moments of raw returns, i-th-order autocorrelations, $i \in \{40, 60, 80\}$, of absolute returns, and j-th-order autocorrelations, $j \in \{40, 50, 60, 80, 100\}$, of squared returns. The moment selection exercise becomes extremely computationally intensive because the machine learning process must consider the new full set consisting of 30 moments. Thus, we compute results for only the Franke and Westerhoff (2012) model depicted in Figure 2.

Table 3: Uncontributive moments

Moment function	*Label*
$m_{20} = E(r_t^3)$	RAW-SKEW
$m_{21} = E(r_t^5)$	RAW-MOM5
$m_{22} = E(r_t^6)$	RAW-MOM6
$m_{23} = E(\|r_t\|\|r_{t-40}\|)$	ABS-AC40
$m_{24} = E(\|r_t\|\|r_{t-60}\|)$	ABS-AC60
$m_{25} = E(\|r_t\|\|r_{t-80}\|)$	ABS-AC80
$m_{26} = E(r_t^2 r_{t-40}^2)$	SQR-AC40
$m_{27} = E(r_t^2 r_{t-50}^2)$	SQR-AC50
$m_{28} = E(r_t^2 r_{t-60}^2)$	SQR-AC60
$m_{29} = E(r_t^2 r_{t-80}^2)$	SQR-AC80
$m_{30} = E(r_t^2 r_{t-100}^2)$	SQR-AC100

(a) Franke and Westerhoff (2012): FSMS

(b) Franke and Westerhoff (2012): BSME

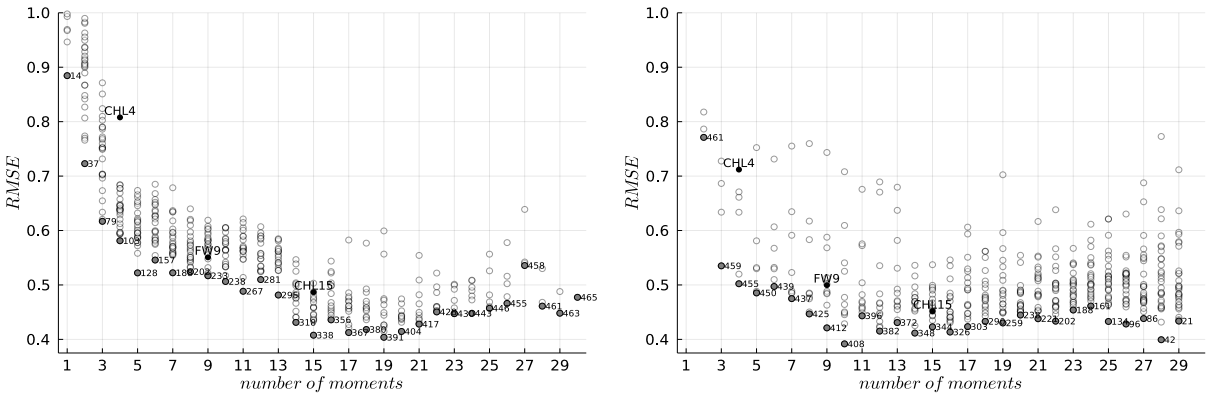

Figure 2: Results for the Franke & Westerhoff (2012) model with 30 moments. *Note:* Transparent points depict results for all moment sets involved in the selection process, numbered gray points depict results for the best performing moment sets of each selection round, and named black points depict results for the benchmark moment sets. For visualization purposes, moment sets with $\text{RMSE}_{\hat{\theta}} > 1$ are excluded.

Machine learning evolution and crucial moments:

- FSMS: ABS-AC25 → ABS-MEAN → SQR-AC10 → ABS-AC80 → ABS-AC50 → SQR-AC5 → SQR-AC80 → ABS-AC40 → SQR-AC60 → RAW-MOM5 → ABS-AC60 → ABS-HILL → RAW-VAR → ABS-AC5 → ABS-AC100 → SQR-AC20 → ABS-AC15 → ABS-AC10 → RAW-AC1 (end of the best set #391) → ABS-AC20 → SQR-AC40 → SQR-AC50 → RAW-SKEW → SQR-AC25 → RAW-MOM6 → SQR-AC100 → SQR-AC1 → ABS-AC1 → RAW-KURT → SQR-AC15,
- BSME: SQR-AC5 → ABS-AC15 → SQR-AC80 → ABS-AC1 → SQR-AC15 → SQR-AC40 → SQR-AC60 → ABS-AC40 → ABS-AC5 → RAW-MOM5 → RAW-MOM6 → SQR-AC10 → ABS-AC25 → SQR-AC20 → SQR-AC50 → RAW-KURT → SQR-AC1 → ABS-AC50 → ABS-AC80 → ABS-AC100 → (beginning of the best set #408) RAW-SKEW → ABS-AC10 → SQR-AC100 → RAW-AC1 → SQR-AC25 → ABS-AC60 → ABS-AC20 → ABS-MEAN → RAW-VAR → ABS-HILL.

In general, the main patterns observed for the original FSMS selection in Figure 1, panels (e) and (f), hold. The variation of estimation performance in individual selection rounds has naturally increased as the range of moment combinations is considerably higher. The key moments selected at the beginning of the original FSMS selection (ABS-AC25, SQR-AC10, ABS-MEAN, ABS-AC50) maintain their importance and are selected within the first five rounds. Interestingly, the second measure of heavy tails, RAW-KURT, which shares importance with the Hill estimator in the initial analysis, is selected second to last. For BSME, only two of the key moments survive to the end of the elimination procedure (ABS-HILL, ABS-MEAN), combined with the variance of raw returns (RAW-VAR) and a collection of autocorrelations.

As a key result, the tendency for a U-shape from the original selection process is now fully manifested for FSMS. After the overall best set #391 ($\text{RMSE}_{\hat{\theta}}$=0.404) is found in selection round

no. 19, the best $\text{RMSE}_{\hat{\theta}}$ in the next rounds almost continuously increases up to round no. 27. This graphic pattern strongly supports our intuition about the suboptimal performance of the SMM under large sets of mutually overlapping moments already discussed for the random walk model. It now appears clear that the impact of the important moments detected in the first half of the FSMS selection process is diminished by the other moments, reducing the estimation performance. For BSME, the phenomenon is less manifested via a U-shape of the best sets. Instead, it is demonstrated via a gradual reduction in the $\text{RMSE}_{\hat{\theta}}$ dispersion for individual sets towards the overall best set #408 ($\text{RMSE}_{\hat{\theta}}$=0.392).

7.2. Alternative weighting matrices

While the utilized weighting matrix can be any positive-definite matrix of given dimensions, estimation efficiency is highly dependent on its selection. At the same time, only a limited number of studies have been conducted to compare various approaches to its construction. We thus study and quantify the impact of different weighting matrices for the most challenging Franke and Westerhoff (2012) model. We take the overall best moment sets obtained for this model in the primary analysis and the three benchmarks and check their performance under alternative weighting matrices in individual estimation exercises. We always compare the performance of the SMM estimator considering the alternative construction of the weighting matrix to our standard setup, the continued block-bootstrap weighting matrix.

The three alternative approaches comprise the trivial identity matrix, which essentially neglects the importance of moments sampling variability; the simplified diagonal weighting matrix (Franke, 2009; Franke et al., 2015; Franke, 2018; Jang and Sacht, 2016, 2021, who provide an excellent discussion on practical and economic rationale of this simplification) in which the off-diagonal components of the continued block-bootstrap matrix are ignored; and the averaged block-bootstrap matrix (Franke and Westerhoff, 2016), again with the individual block size set to 250 observations and the bootstrap size of $B = 5,000$. As opposed to the continued block-bootstrap matrix, the blocks of data drawn from the pseudo-empirical time series are not pasted together to form a bootstrapped series of the original length T_{sim}. Thus, one avoids the potential problem of disregarding the artificial structure of bootstrapped time series when calculating autocorrelation moments near the points where the individual blocks are joined. However, the absence of the long-bootstrapped time series for the computation of simulated moments may outweigh this advantage.

Table 4 is to be compared with the bottom part of Table 2. One directly observes that the trivialized setup with the identity matrix delivers largely inferior results. The $\text{RMSE}_{\hat{\theta}}$ for all benchmarks and the best sets are up to several times larger than our initial results based on the continued block-bootstrap matrix. The difference is pronounced for the smallest benchmark set *CHL4*. Importantly, negative estimation improvements suggest that the other two benchmarks deliver better estimation performance compared to that of the best sets.

On the other hand, while delivering slightly worse performances, the diagonal and averaged block-bootstrap matrices are generally comparable to the continued block-bootstrap matrix. The $\text{RMSE}_{\hat{\theta}}$ values for the *CHL4* and *FW9* benchmarks are very similar; the continued block-bootstrap matrix (Table 2) seems worth strong recommending only for the *CHL15* moment set. The best sets always deliver slightly worse performance. Nevertheless, the estimation improvements are positive in all cases but smaller than those in Table 2 and follow the same decreasing tendency with an expanding set of moments for the Franke and Westerhoff (2012) model. The only exception for the diagonal matrix and *CHL15* is due to unusually high $\text{RMSE}_{\hat{\theta}}$ for the given benchmark.

Table 4: Franke and Westerhoff (2012): Alternative weighting matrices

		CHL4	*FW9*	*CHL15*
Identity matrix				
Benchmark	$\text{RMSE}_{\hat{\theta}}$	2.453	0.728	0.793
FSMS—Best set #151	0.867	64.7%	-19.1%	-9.3%
BSME—Best set #174	1.119	54.4%	-53.7%	-41.1%
Diagonal continued block-bootstrap matrix				
Benchmark	$\text{RMSE}_{\hat{\theta}}$	0.820	0.551	0.659
FSMS—Best set #151	0.501	38.9%	9.1%	24.0%
BSME—Best set #174	0.456	44.4%	17.2%	30.8%
Averaged block-bootstrap matrix				
Benchmark	$\text{RMSE}_{\hat{\theta}}$	0.803	0.586	0.544
FSMS—Best set #151	0.536	33.3%	8.5%	1.5%
BSME—Best set #174	0.542	32.5%	7.5%	0.4%

Note: Estimation improvements compared to the benchmark moment sets are reported in %.

7.3. Performance of the equal-sized sets

In terms of the $\text{RMSE}_{\hat{\theta}}$, Section 6 shows that both of the proposed algorithms successfully find multiple moment sets of various sizes that outperform all benchmark sets for every model. In Table 5, we focus on the best sets containing the same number of moments as each of the benchmark sets. This comparison highlights the potential for estimation performance enhancement for moment sets of limited size.

The random walk model with a structural break shows that the general applicability of the benchmark moment sets for financial models does not always hold. Specifically, if certain assumptions are made during the construction of the moment sets, their usefulness might be limited. The unconditional first moment of absolute returns contained in the *FW9* benchmark set ensures a comparably better fit to the *CHL4* and *CHL15* benchmarks, in which the given moment is absent. Additionally, none of the benchmarks can compare to any of the estimates achieved using the moment sets identified by the selection algorithms, even if only the equal-sized sets are considered. This can be explained by the fact that none of the benchmarks contain both crucial moments for estimating the random walk model. Specifically, the absence of a joint pair of the unconditional first moment of absolute returns (ABS-MEAN) and the variance of raw returns (RAW-VAR), selected/eliminated among the first/last three using FSMS and BSME, respectively, is the likely cause of the generally unsatisfactory estimation performance of the benchmarks. Numerically, the $\text{RMSE}_{\hat{\theta}}$ of the equal-sized best sets achieves only slightly worse estimation improvements compared to the overall best sets identified in Section 6.

For the Alfarano et al. (2008) and Franke and Westerhoff (2012) models, a similar numerical

Table 5: Estimation improvement measured by $\text{RMSE}_{\hat{\theta}}$: Equal-sized sets

		CHL4	*FW9*	*CHL15*
Random walk with a structural break				
Benchmark	$\text{RMSE}_{\hat{\theta}}$	0.701	0.486	0.656
FSMS		0.143	0.135	0.159
		79.6%	72.2%	75.8%
BSME		0.152	0.143	0.160
		78.3%	70.6%	75.6%
Alfarano, Lux, and Wagner (2008)				
Benchmark	$\text{RMSE}_{\hat{\theta}}$	0.809	0.607	0.308
FSMS		0.767	0.313	0.283
		5.2%	48.4%	8.1%
BSME		0.361	0.284	0.274
		55.4%	53.2%	11.0%
Franke and Westerhoff (2012)				
Benchmark	$\text{RMSE}_{\hat{\theta}}$	0.808	0.551	0.487
FSMS		0.557	0.431	0.439
		31.1%	21.8%	9.9%
BSME		0.460	0.428	0.407
		43.1%	22.3%	16.4%

Note: Estimation improvements compared to the benchmark moment sets are reported in %.

conclusion can be made. The similarity is due to the generally flat shape of the curves of best sets in later stages of the selection rounds or from the beginning of the elimination rounds, as observed in Figure 1. Nevertheless, the better performance of the overall best sets than equal-sized sets is more pronounced for these models. However, two exceptions appear. First, the Alfarano et al. (2008) best set of four moments #55 identified by FSMS only slightly outperforms the *CHL4* benchmark, which, on the other hand, surpasses all different sets of four moments found using the machine learning algorithm. This is a surprising result, as the given best set shares only one moment (RAW-VAR) with the *CHL4* benchmark set. Second, for the Franke and Westerhoff (2012) model, FSMS equal-sized sets perform worse than those identified by BSME. We discuss this general phenomenon in Subsection 8.2.

8. Discussion

We now summarize and discuss the general properties of the proposed machine-learning-inspired methodology and the main conclusions of our analysis. We first discuss the overidentifica-

tion pitfalls and benefits for the SMM. Next, we consider the relationship between the complexity of a model and the size of the overall best moment sets identified by the FSMS and BSME algorithms and compare and contrast the specifics of the two algorithms. Finally, we study particular moments prioritized by the selection algorithms across models and provide our final recommendations for practical SMM applications.

8.1. Overidentification

In the GMM framework, overidentification represents a situation in which the number of unconditional moment restrictions surpasses the number of estimated parameters. According to Hansen (1982), overidentification allows for the construction of more efficient estimators and the application of additional specification tests. However, the findings of Chen and Lux (2018) show that this rule might not always hold for the simulated version of this estimation method. The authors employ four moment sets of increasing size for estimation of the Alfarano et al. (2008) model using the SMM. These moment sets are constructed such that each set is a subset of all larger sets. In the smallest set identical to our benchmark set *CHL4*, only the most crucial moments are included. To these moments, other moments are added iteratively to construct larger sets topped with the collection of 15 moments that are identical to our benchmark set *CHL15*. The authors find that well-constructed smaller moment sets sometimes perform better than larger ones. These findings suggest that excessive overidentification hurts the estimation process of the SMM. In other words, too many moments included in the moment set may result in worse estimation performance. This is likely caused by the abundance of other moments diminishing the impact of the crucial moments on the minimized function and, thus, creating difficulties for the optimization algorithm.

Figure 3 represents a graphical simplification of Figure 1 and Figure 2 and depicts all moment sets with the best performance in their respective round. If excessive overidentification causes inefficiencies in the estimation process, we expect U-shaped curves to appear. Such a pattern follows the notion that while an individual moment identified in the first round of FSMS or eliminated in the last round of BSME alone is naturally completely insufficient for SMM purposes, a few subsequent additions to the moment set considerably improve the estimation performance. This tendency gradually decreases to a point where the improvements become stale. Finally, the performance begins to worsen with each addition due to an excessive number of moments being used. This shape can be easily identified in the selection processes of the random walk with a structural break, as already highlighted in Section 6. Nevertheless, the same shape is not present for the Alfarano et al. (2008) model and for the BSME elimination for the Franke and Westerhoff (2012) model. Furthermore, it mostly also holds in these cases that the overall best moment set is found somewhere in the middle of the pack rather than being the full moment set. However, a slight upwards trend following round no. 11 containing the overall best set can be observed in the results for the Franke and Westerhoff (2012) model under the FSMS selection in panel (e). After introducing 11 additional moments within the sensitivity checks in Subsection 7.1, the U-shape becomes fully manifested, as depicted in panel (g). These results clearly show that moment overidentification may indeed lead to estimation inefficiencies when applying the SMM. Our default full moment set contains too many elements for the random walk model with the overall best sets of sizes 6 and 5. Conversely, for the Alfarano et al. (2008) model with the largest overall best sets containing 16 and 12 moments, the overidentification tendency does not manifest itself within the full set of size 19. Finally, for the Franke and Westerhoff (2012) model with the overall best sets of sizes 11 and 6, there is already an indication of an overidentification issue for the default

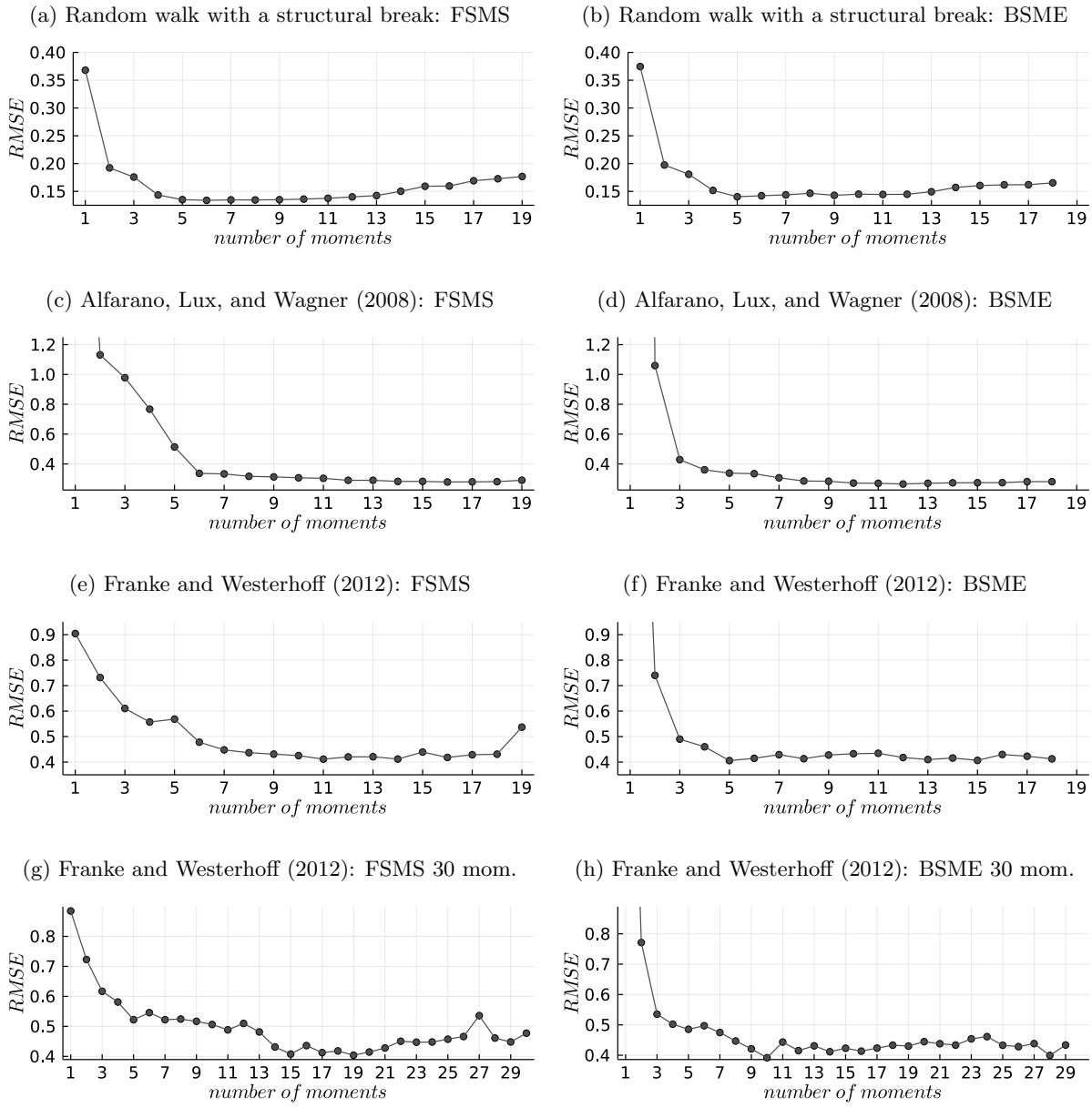

Figure 3: Best moment sets: Improvement patterns. *Note:* The left half of the figure displays the evolution of $\text{RMSE}_{\hat{\theta}}$ of the best sets of each selection round for FSMS. The right half of the figure displays the evolution of $\text{RMSE}_{\hat{\theta}}$ of the best sets of each elimination round for BSME. Moment sets with high $\text{RMSE}_{\hat{\theta}}$ are excluded for visualization purposes.

full moment set, especially for the FSMS selection, and after the addition of 11 uncontributive moments, the issue becomes obvious.

Table 6: Overall best moment sets: Sizes, intersections, unions

	Par.	FSMS	BSME	$\bigcap$	$\bigcup$
Random walk with a structural break	2	6	5	4 (0.153)	7 (0.138)
Alfarano, Lux, and Wagner (2008)	3	16	12	11 (0.268)	17 (0.275)
Franke and Westerhoff (2012)	7	11	5	3 (0.508)	13 (0.422)

Note: 'Par.' corresponds to the number of estimated parameters, FSMS and BSME to the size of the best moment set identified by the algorithm, and $\bigcap$ and $\bigcup$ to the number of moments of the intersection and union of the best moment sets. Numbers in parentheses indicate the $\text{RMSE}_{\hat{\theta}}$ of the given intersection/union.

8.2. Sizes of the overall best sets and differences between algorithms

In the previous Subsection 8.1, we focus on the link between the number of moments in a moment set and the estimation performance. However, it is interesting to consider the connection between the number of model parameters and the size of the overall best moment sets identified by the proposed selection procedures. Furthermore, the optimality of the identified best moment sets can be challenged if they differ significantly across the two approaches.

Table 6 displays the size of the overall best moment sets identified by FSMS and BSME, respectively, for each model. One immediately notices that there is no clear relationship between the number of estimated parameters and the size of the overall best set, but the sets chosen by BSME are consistently smaller than those identified by FSMS. The difference stems from the very nature of the algorithms and their potential ability to take advantage of the interplay between specific moments. While in the first rounds of the selection process, FSMS is forced to choose the moments according to their solo performance, for BSME, such a complex interplay is considered from the very beginning of the process of gradual elimination of the full moment set. In other words, parameters identified as crucial in the first rounds of the FSMS procedure can be found to be less important or even redundant from the perspective of the BSME search if a combination of some already included moments largely compensates for their absence.

The same argumentation explains the different patterns in the left and right columns of Figures 1 and 3. For FSMS, the expanding set leads to a hyperbolic-like shape of the imaginary efficiency curve, which gradually decreases and reaches relative flatness only in the later stages of the selection process. In contrast, the step-by-step elimination process of BSME results in a greater estimating efficiency until the last rounds, followed by large jumps in the $\text{RMSE}_{\hat{\theta}}$ for the smallest moment sets. These results also rule out a potential considered disadvantage of BSME that the lack of clarity about importance of the individual overlapping moments under the full moment set might lead to elimination of some critical moments in the first rounds, mainly due to randomness.

The set for the random walk model chosen by FSMS as the overall best includes six moments, while the best set identified by BSME consists of five moments. The two sets share a total of four moments: ABS-MEAN, RAW-VAR, ABS-AC50, and ABS-AC5. For the Alfarano et al. (2008) model, the content of the two overall best sets differs markedly. The FSMS algorithm leads to a set consisting of 16 moments, while BSME produces an overall best set with 12 moments, achieving slightly better $\text{RMSE}_{\hat{\theta}}$. The two sets almost perfectly overlap and share a total of 11 moments: only SQR-AC10, which is included in the smaller set #102, is excluded from the larger set #183.

A potential absence of some specific moments, thus, does not appear to matter crucially for this model due to the interplay between other moments that fully compensate for it. The composition of the two overall best sets also differs considerably for the Franke and Westerhoff (2012) model. The FSMS algorithm detects a more extensive set with 11 moments, while BSME produces a parsimonious set containing five moments while still achieving the same level of $\text{RMSE}_{\hat{\theta}}$. The two sets share only three moments, the indicator of the overall volatility and measures of heavy tails (ABS-MEAN, RAW-KURT, and ABS-HILL), because the autocorrelation composition differs according to the interpretation above.

By and large, the number of moments intersecting the two overall best sets indicates that the selection processes prioritize the same moments. With this in mind, we argue that directly taking the moment sets obtained from the two algorithms and employing them in the SMM without giving it a second thought might lead to failure to take full advantage of our machine learning procedure. Thus, a recommended approach in the final stage is to consider the outputs of the two automated selection processes in a 'manual' manner, both independently and concerning each other. In an ideal case scenario, one would first consider particular moments driving the individual selection processes. Next, the intersecting moments in both model sets are identified, thereby crucially contributing to the estimation routine. Subsequently, these moments can be used to potentially prepare an even better moment set in an additional manual trial-and-error Monte Carlo comparison experiment. Its performance can be tested against the previously achieved results of the FSMS and BSME algorithms.

Two possible moment sets to try first are the intersection and union of the two overall best moment sets. As such, even if the machine-learning automated procedures' outputs were not used as a final moment set for the resulting empirical SMM directly, they would still represent a crucial tool in the process. The numbers in parentheses in the last two columns of Table 6 indicate the $\text{RMSE}_{\hat{\theta}}$ of the given intersection/union, which is to be compared to $\text{RMSE}_{\hat{\theta}}$ reported in Table 2. Importantly, none of these combinations outperform the better ones from the two overall best sets. On the other hand, the union of seven moments achieves a slightly better $\text{RMSE}_{\hat{\theta}}$ than that of the final BSME output for the random walk model, and the intersection of 11 moments reaches a slightly better $\text{RMSE}_{\hat{\theta}}$ than that of the final FSMS output for the Alfarano et al. (2008) model.

8.3. Moment selection across models

The previous Subsection 8.2 concludes that the overall best moment sets share a large portion of elements for each model, even though their sizes vary markedly. However, we have yet to consider the particular composition of the overall best sets across all models. Instead of taking these sets into scope individually as in Section 6, the initial selections of FSMS and the final eliminations of BSME are considered as corresponding to the crucial moments for the respective model. Additionally, these selections should allow us to analyze the moment preference stability across the two approaches. Table 7 summarizes the developments of the composition of the overall best moment sets. For FSMS, the depicted number corresponds to the round in which the respective moment was selected; for BSME, the logic is reversed: the most important moment eliminated last is marked by '1', the second-last by '2', etc.

Beginning with the most straightforward random walk with a structural break, we find that the four intersecting moments also largely overlap in terms of their importance. Overall, both approaches seem to identify the crucial moments similarly. This indicates a high degree of reliability of the identified sets and a clear order of importance of specific moments. Both algorithms give the highest priority to the unconditional mean of absolute returns (ABS-MEAN) and the third-highest

Table 7: Moments sets: Evolution of the overall best sets

		RWwSB		ALW(2008)		FW(2012)	
Moment function	*Label*	FSMS	BSME	FSMS	BSME	FSMS	BSME
$m_1 = E(r_t^2)$	RAW-VAR	3	3	4	3	9	
$m_2 = E(r_t^4)$	RAW-KURT			5	2	6	1
$m_3 = E(r_t r_{t-1})$	RAW-AC1						
$m_4 = E(\|r_t\|)$	ABS-MEAN	1	1	6	1	3	2
$m_5 = Hill(\|r_t\|, 5)$	ABS-HILL			12	6	5	3
$m_6 = E(\|r_t\|\|r_{t-1}\|)$	ABS-AC1		4	14	5		4
$m_7 = E(\|r_t\|\|r_{t-5}\|)$	ABS-AC5	5	2			8	
$m_8 = E(\|r_t\|\|r_{t-10}\|)$	ABS-AC10			3			
$m_9 = E(\|r_t\|\|r_{t-15}\|)$	ABS-AC15			10	10	7	
$m_{10} = E(\|r_t\|\|r_{t-20}\|)$	ABS-AC20			8			
$m_{11} = E(\|r_t\|\|r_{t-25}\|)$	ABS-AC25	4		2		1	
$m_{12} = E(\|r_t\|\|r_{t-50}\|)$	ABS-AC50	2	5	15		4	
$m_{13} = E(\|r_t\|\|r_{t-100}\|)$	ABS-AC100			1	4		5
$m_{14} = E(r_t^2 r_{t-1}^2)$	SQR-AC1			16	9	11	
$m_{15} = E(r_t^2 r_{t-5}^2)$	SQR-AC5			13	7		
$m_{16} = E(r_t^2 r_{t-10}^2)$	SQR-AC10				11	2	
$m_{17} = E(r_t^2 r_{t-15}^2)$	SQR-AC15			11	8	10	
$m_{18} = E(r_t^2 r_{t-20}^2)$	SQR-AC20			9			
$m_{19} = E(r_t^2 r_{t-25}^2)$	SQR-AC25	6		7	12		

Note: 'RWwSB' is the random walk with a structural break, 'ALW(2008)' is the Alfarano et al. (2008) model, and 'FW(2012)' is the Franke and Westerhoff (2012) model. For FSMS, the depicted number corresponds to the round in which the respective moment was selected; for BSME, the logic is reversed: the most important moment eliminated last is marked by '1', the second-last by '2', etc.

priority to the unconditional variance of raw returns (RAW-VAR). The two autocorrelations of absolute returns (ABS-AC5 and ABS-AC-50) show flipped priorities having one other unshared autocorrelation between them in the fourth position.

Similar to the random walk model, both procedures prioritize ABS-MEAN and RAW-VAR for the two financial agent-based models. Within the most significant moments, both approaches for both models also pinpoint the kurtosis of raw returns (RAW-KURT) and the Hill estimator of the tail index of absolute returns (ABS-HILL). However, the prioritization is ordered differently and subsequently follows different paths for the two procedures and the two models. This might be caused by the difficulty of identifying moment couples/groups that yield strong estimation performance in combination, while inferior results individually, as discussed in Subsection 8.2.

The commonalities and differences between the compositions for the FSMS and BSME approaches lead us to three important conclusions for the Alfarano et al. (2008) model. First, autocorrelations of absolute returns are generally more critical than those of squared returns. Second, measures of the most proclaimed stylized facts, together with standard 'scale' moments capturing overall volatility (ABS-MEAN, RAW-VAR), are crucial for the SMM performance. Third, while

some high-order autocorrelations of returns are essential for capturing the phenomenon of volatility clustering, their actual composition is not of central importance due to a complex interplay between other autocorrelations and additional moments being able to fully compensate for the absence of one particular moment condition.

The results for the Franke and Westerhoff (2012) model lead us to very similar overall conclusions. First, autocorrelations of absolute returns seem, again, more important than those of squared returns. Second, measures of heavy tails and volatility clustering together with the typically important indicator of overall volatility (ABS-MEAN) are crucial for the SMM performance. Third, the actual composition of high-order autocorrelations does not matter crucially due to the mutual interactions between the moments that compensate for the impact of one particular. The conclusion on the effect of the difference between the nature of the selection and eliminations process also holds. Finally, in regard to the importance of individual moments, we observe a very similar pattern for both financial agent-based models; only an additional measure of heavy tails, the Hill estimator, is emphasized at the expense of variance (RAW-VAR), which is no longer critical for the Franke and Westerhoff (2012) model.

An additional observation valid for all models is that the measure of one of the basic stylized facts, the absence of autocorrelation in returns, the first-order autocorrelation of raw returns (RAW-AC1), is generally unimportant for the SMM, as it never makes it to the composition of the overall best sets and is consistently selected among the last by FSMS. One could, however, already have expected this behavior based on the respective descriptive statistics regarding AR(1) dynamics in Table A.8, where autocorrelation coefficients on the first lag are negligible. Thus, the value of RAW-AC1 also contributes negligibly to the overall criterion distance $J(\theta)$ (7). We hypothesize that it would gain more importance if the absence of autocorrelation is violated in some other financial model under the SMM estimation.

Overall, for simpler models, our extension appears to identify almost the same sets of moments using both algorithms. For more complex models, the processes ultimately lead to markedly similar moment sets in terms of moment intersections. Nevertheless, the selections follow different paths, and BSME tends to prefer parsimonious sets, a feature stemming from the differences between the two algorithms. Other specifications of the selection techniques might be used to potentially address this ambiguity. For instance, stepwise selection of moment couples could help; however, the considerably increased computational burden must be taken into account.

9. Conclusion

This paper introduces machine learning methods of feature selection to agent-based econometrics. We propose a machine learning extension of the SMM to expand the estimation methodology for analytically intractable financial agent-based models. Their complexity often prevents the application of standard econometric tools, such as the generalized method of moments and the maximum likelihood method. Therefore, simulated counterparts are traditionally employed instead. The most prominent alternative is the SMM, which, however, faces several difficulties.

The necessity of specifying the moments used in the estimation process and the potential arbitrariness of this choice are among the most concerning problems of the method (Platt, 2020). The moment set is traditionally constructed intuitively concerning the estimated model and the stylized facts it should replicate. However, an insufficient number of moments may be provided to the estimation method with vital parts of the model dynamics not captured by the set. On the other hand, the impact of important moments may be diminished if too many moments are included

in the set, which would render the method less efficient. The proposed extension automates an appropriate moments' choice. The main aim is to identify optimal moment sets for a specific model yielding better estimation results than moment sets traditionally used in the literature.

The intuition behind the extension follows one of the most straightforward and earliest applications of machine learning in the field of econometrics. Specifically, it revolves around the methods defined within the stepwise regression framework. Instead of model selection, those are applied to select the set of moments used by the SMM. There are two algorithms proposed and described, forward stepwise moment selection (FSMS) and backward stepwise moment elimination (BSME). Subsequently, these algorithms are tested on three models with an increasing level of complexity.

The testbed models include the random walk with a structural break and two popular financial agent-based frameworks: the Alfarano et al. (2008) model and the Franke and Westerhoff (2012) model. The set of employed testbed models enables robust testing of the functionality of the proposed methodology. Additionally, it allows for a detailed study of its general properties. The limited complexity and relative estimation simplicity of the random walk model combined with its ability to represent stochastic properties of financial time series thus tests and challenges the applicability of the proposed methodology, as inspired by Lamperti (2018) and Platt (2020, 2022). The results are compared with estimates generated using three benchmark moment sets from Franke and Westerhoff (2012) and Chen and Lux (2018) that have explicitly been constructed with the empirical stylized facts replicated by each model in mind.

We find that the two algorithms consistently detect the most important moments for every model. The moment sets identified by the selection methods achieve a substantial performance gain over the benchmark sets. Moreover, the fact that both agent-based frameworks are designed to explain similar financial stylized facts creates an opportunity for more general conclusions regarding the SMM estimation of this type of financial models. Although the benchmark moment sets were specifically designed for the given models, all three are considerably outperformed by many moment sets identified by the proposed methodology. Consequently, these sets could be directly applied to estimate the model using empirical data. As a follow-up to this work, if the full moment set was appropriately expanded, then a study of the application of the methodology to small-scale macroeconomic agent-based models could be conducted as well.

Acknowledgments

This paper benefited greatly from Reiner Franke's detailed comments and research discussions. We are also indebted to the participants of the EAEPE Conference 2020, the ERFIN Workshop 2021, the ICNTEF conference 2020 and 2021, and the EcoSta conference 2021 for their many useful suggestions, as well as to Sergey Bolshakov and Jan Hanzal for their excellent research assistantship.

Jiri Kukacka gratefully acknowledges the financial support from the Czech Science Foundation under the project 'Linking financial and economic agent-based models: An econometric approach' [grant number 20-14817S], from the Charles University PRIMUS program [grant number PRIMUS/19/HUM/17], and from the Charles University UNCE program [grant number UNCE/HUM/035]. This work was supported by the Cooperatio Program at Charles University, research area Economics.

Declarations

Declaration of competing interest: None.

Authors' contributions: Both authors contributed equally to this manuscript.

CRediT author statement: **Eric Zila:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Jiri Kukacka:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

References

Alfarano, S., T. Lux, and F. Wagner (2005). Estimation of agent-based models: The case of an asymmetric herding models. *Computational Economics 26*, 19–49.

Alfarano, S., T. Lux, and F. Wagner (2006). Estimation of a simple agent-based model of financial markets: An application to australian stock and foreign exchange data. *Physica A: Statistical Mechanics and its Applications 370*(1), 38–42.

Alfarano, S., T. Lux, and F. Wagner (2007). Empirical validation of stochastic models of interacting agents. *The European Physical Journal B 55*(2), 183–187.

Alfarano, S., T. Lux, and F. Wagner (2008). Time variation of higher moments in a financial market with heterogeneous agents: An analytical approach. *Journal of Economic Dynamics and Control 32*(1), 101–136.

Bargigli, L., L. Riccetti, A. Russo, and M. Gallegati (2020). Network calibration and metamodeling of a financial accelerator agent based model. *Journal of Economic Interaction and Coordination 15*(2), 413–440.

Brock, W. A. and C. H. Hommes (1998). Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics & Control 22*, 1235–1274.

Chen, S. and S. Desiderio (2022). A regression-based calibration method for agent-based models. *Computational Economics 59*(2), 687–700.

Chen, S.-H., C.-L. Chang, and Y.-R. Du (2012, 6). Agent-based economic models and econometrics. *The Knowledge Engineering Review 27*, 187–219.

Chen, Z. and T. Lux (2018). Estimation of sentiment effects in financial markets: A simulated method of moments approach. *Computational Economics 52*(3), 711–744.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance 1*(2), 223–236.

Cont, R. (2007). Volatility clustering in financial markets: Empirical facts and agent-based models. In G. Teyssiere and A. Kirman (Eds.), *Long Memory in Economics*, pp. 289–309. Springer Berlin Heidelberg.

Delli Gatti, D. and J. Grazzini (2020). Rising to the challenge: Bayesian estimation and forecasting techniques for macroeconomic agent based models. *Journal of Economic Behavior & Organization 178*, 875–902.

Dieci, R. and X.-Z. He (2018). Heterogeneous agent models in finance. In C. Hommes and B. LeBaron (Eds.), *Handbook of Computational Economics*, Volume 4 of *Handbook of Computational Economics*, Chapter 5, pp. 257–328. Elsevier.

Ding, Z., C. W. Granger, and R. F. Engle (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance 1*(1), 83–106.

Duffie, D. and K. J. Singleton (1993). Simulated moments estimation of Markov models of asset prices. *Econometrica 61*(4), 929–952.

Fabretti, A. (2013). On the problem of calibrating an agent based model for financial markets. *Journal of Economic Interaction and Coordination 8*(2), 277–293.

Fagiolo, G., M. Guerini, F. Lamperti, A. Moneta, and A. Roventini (2019). Validation of agent-based models in economics and finance. In C. Beisbart and N. J. Saam (Eds.), *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*, pp. 763–787. Springer International Publishing.

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business 38*(1), 34–105.

Farmer, J. D. and S. Joshi (2002). The price dynamics of common trading strategies. *Journal of Economic Behavior & Organization 49*(2), 149–171.

Franke, R. (2009). Applying the method of simulated moments to estimate a small agent-based asset pricing model. *Journal of Empirical Finance 16*, 804–815.

Franke, R. (2018). Competitive moment matching of a New-Keynesian and an Old-Keynesian model. *Journal of Economic Interaction and Coordination 13*(2), 201–239.

Franke, R., T.-S. Jang, and S. Sacht (2015). Moment matching versus Bayesian estimation: Backward-looking behaviour in a New-Keynesian baseline model. *The North American Journal of Economics and Finance 31*, 126–154.

Franke, R. and F. Westerhoff (2011). Estimation of a structural stochastic volatility model of asset pricing. *Computational Economics 38*(1), 53–83.

Franke, R. and F. Westerhoff (2012). Structural stochastic volatility in asset pricing dynamics: Estimation and model contest. *Journal of Economic Dynamics and Control 36*(8), 1193–1211.

Franke, R. and F. Westerhoff (2016). Why a simple herding model may generate the stylized facts of daily returns: Explanation and estimation. *Journal of Economic Interaction and Coordination 11*(1), 1–34.

Ghonghadze, J. and T. Lux (2016). Bringing an elementary agent-based model to the data: Estimation via GMM and an application to forecasting of asset price volatility. *Journal of Empirical Finance 37*, 1–19.

Gilli, M. and P. Winker (2003). A global optimization heuristic for estimating agent based models. *Computational Statistics & Data Analysis 42*, 299–312.

Grazzini, J. (2012). Analysis of the emergent properties: Stationarity and ergodicity. *Journal of Artificial Societies and Social Simulation 15*(2), 7.

Grazzini, J. and M. Richiardi (2015). Estimation of ergodic agent-based models by simulated minimum distance. *Journal of Economic Dynamics & Control 51*, 148–165.

Grazzini, J., M. G. Richiardi, and M. Tsionas (2017). Bayesian estimation of agent-based models. *Journal of Economic Dynamics and Control 77*, 26–47.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning, Second Ed.* Springer Series in Statistics, New York.

Hong, H. and J. C. Stein (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance 54*(6), 2143–2184.

Jalali, M., H. Rahmandad, and H. Ghoddusi (2015). *Using the method of simulated moments for system identification.* MIT Press: Cambridge, MA.

Jang, T.-S. and S. Sacht (2016). Animal spirits and the business cycle: Empirical evidence from moment matching. *Metroeconomica 67*(1), 76–113.

Jang, T.-S. and S. Sacht (2021). Forecast heuristics, consumer expectations, and New-Keynesian macroeconomics: A Horse race. *Journal of Economic Behavior & Organization 182*, 493–511.

Kirman, A. (1991). Epidemics of opinion and speculative bubbles in financial markets. In M. Taylor (Ed.), *Money and Financial Markets*, pp. 354–368. Macmillan, New York, USA.

Kirman, A. (1993). Ants, rationality, and recruitment. *The Quarterly Journal of Economics 108*(1), 137–156.

Kukacka, J. and L. Kristoufek (2020). Do 'complex' financial models really lead to complex dynamics? Agent-based models and multifractality. *Journal of Economic Dynamics and Control 113*, 103855.

Kukacka, J. and L. Kristoufek (2021). Does parameterization affect the complexity of agent-based models? *Journal of Economic Behavior & Organization 192*, 324–356.

Lamperti, F. (2018). An information theoretic criterion for empirical validation of simulation models. *Econometrics and Statistics 5*, 83–106.

Lamperti, F., A. Roventini, and A. Sani (2018). Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control 90*, 366–389.

Lee, B.-S. and B. F. Ingram (1991). Simulation estimation of time-series models. *Journal of Econometrics 47*(2), 197–205.

Lee, J. S., T. Filatova, A. Ligmann-Zielinska, B. Hassani-Mahmooei, F. Stonedahl, I. Lorscheid, A. Voinov, G. Polhill, Z. Sun, and D. C. Parker (2015). The complexities of agent-based modeling output analysis. *The Journal of Artificial Societies and Social Simulation 18*(4).

Lux, T. (2018). Estimation of agent-based models using sequential Monte Carlo methods. *Journal of Economic Dynamics and Control 91*, 391–408.

Lux, T. (2021a). Approximate Bayesian inference for agent-based models in economics: A case study. Technical report, University of Kiel.

Lux, T. (2021b). Bayesian estimation of agent-based models via adaptive particle Markov Chain Monte Carlo.

Computational Economics, 1–27.

Lux, T. and R. C. Zwinkels (2018). Empirical validation of agent-based models. In C. Hommes and B. LeBaron (Eds.), *Handbook of Computational Economics*, Volume 4 of *Handbook of Computational Economics*, Chapter 8, pp. 437–488. Elsevier.

Mandes, A. and P. Winker (2017). Complexity and model comparison in agent based modeling of financial markets. *Journal of Economic Interaction and Coordination 12*(3), 469–506.

Manzan, S. and F. H. Westerhoff (2007). Heterogeneous expectations, exchange rate dynamics and predictability. *Journal of Economic Behavior & Organization 64*, 111–128.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica 57*(5), 995–1026.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*(5), 1027–1057.

Platt, D. (2020). A comparison of economic agent-based model calibration methods. *Journal of Economic Dynamics and Control 113*, 103859.

Platt, D. (2022). Bayesian estimation of economic simulation models using neural networks. *Computational Economics 59*(2), 599–650.

Salle, I. and M. Yıldızoğlu (2014, Dec). Efficient sampling and meta-modeling for computational economic models. *Computational Economics 44*(4), 507–536.

Tubbenhauer, T., C. Fieberg, and T. Poddig (2021). Multi-agent-based var forecasting. *Journal of Economic Dynamics and Control 131*, 104231.

Vandin, A., D. Giachini, F. Lamperti, and F. Chiaromonte (2021). Automated and distributed statistical analysis of economic agent-based models. Technical report.

Welch, P. D. (1983). The statistical analysis of simulation results. In *The Computer Performance Modeling Handbook*, Volume 22, Chapter 6, pp. 268–328.

Winker, P., M. Gilli, and V. Jeleskovic (2007). An objective function for simulation based inference on exchange rate data. *Journal of Economic Interaction and Coordination 2*(2), 125–145.

Zegadło, P. (2021). Efficient calibration of a financial agent-based model using the method of simulated moments. In M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. Sloot (Eds.), *Computational Science – ICCS 2021*, Cham, pp. 316–329. Springer International Publishing.

Zhang, W., A. Valencia, and N.-B. Chang (2021). Synergistic integration between machine learning and agent-based modeling: A multidisciplinary review. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21.

Appendix A. Model outputs and descriptive statistics

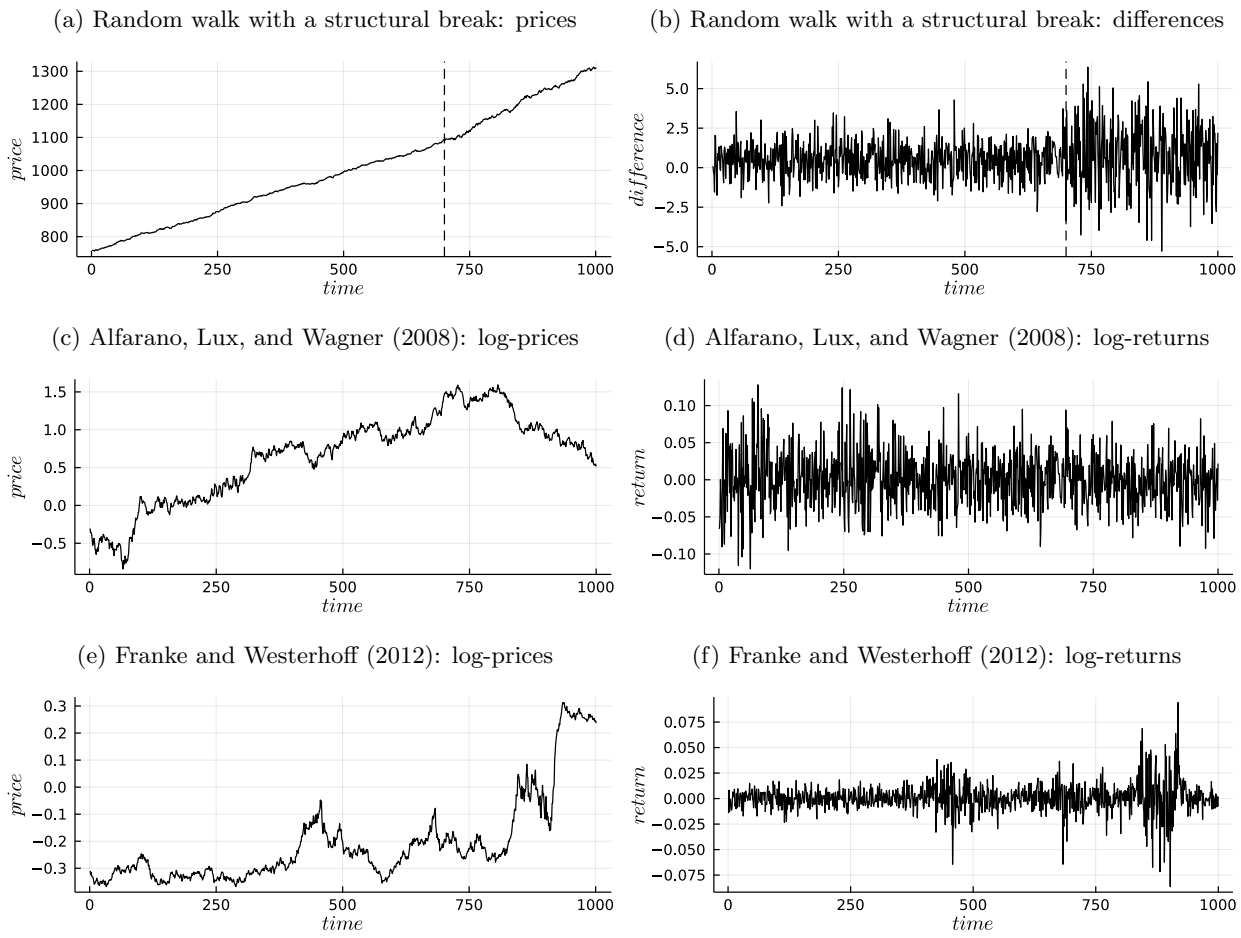

Figure A.4: Model outputs. *Note:* The figure displays typical time series outputs of the testbed models under the general simulation setup. For visual clarity, only a time period of length 1,000 is used for generation. For the random walk with a structural break, the dashed vertical line depicts the occurrence of the structural break.

Table A.8: Descriptive statistics

	RWwD	RWwSB	ALW (2008)	FW (2012)
Mean	0.40	0.49	0.00	0.00
	(0.37;0.42)	(0.45;0.52)	(0.00;0.00)	(0.00;0.00)
SD	1.00	1.38	0.04	0.01
	(0.98;1.02)	(1.35;1.41)	(0.04;0.05)	(0.01;0.01)
Skew	0.00	0.21	0.00	0.00
	(-0.06;0.06)	(0.08;0.34)	(-0.11,0.11)	(-0.60;0.56)
Ex Kurt	0.00	1.58	0.88	10.08
	(-0.12;0.13)	(1.24;1.95)	(0.47;1.27)	(6.87;14.68)
$\widehat{AR}(1)$	0.00	0.01	0.00	0.01
	(-0.03;0.03)	(-0.02;0.04)	(-0.03;0.03)	(-0.04;0.06)
L-B	5.8%	17.2%	8.0%	35.6%
J-B	5.4%	100.0%	100.0%	100.0%
ADF	100.0%	100.0%	100.0%	100.0%
KPSS	5.4%	100.0%	3.3%	0.1%
G2012	5.8;5.0;4.8;4.5%	100.0%	6.3;5.4;5.5;4.0%	4.7;4.6;5.2;5.6%
DGG2020	2.2%	4.9%	4.6%	6.2%

Note: 'RWwD' is the random walk with a drift, 'RWwSB' is the random walk with a structural break, 'ALW(2008)' is the Alfarano et al. (2008) model, and 'FW(2012)' is the Franke and Westerhoff (2012) model. The averaged results are based on 1,000 random runs and the general simulation setup. Resulting 95% sample confidence intervals are reported in parentheses. $\widehat{AR}(1)$ reports the estimated AR(1) coefficient based on an ARIMA (1,0,0) model. Rejection rates based on a 5% significance level are reported for the statistical tests: H_0 for the Ljung-Box test (L-B) is 'no autocorrelation on the first lag,' H_0 for the Jarque-Bera test (J-B) is 'normality,' H_0 for the augmented Dickey-Fuller test (ADF) with constant is 'unit root presence/covariance nonstationarity,' H_0 for the Kwiatkowski-Phillips-Schmidt-Shin test (KPSS) is 'level stationarity,' H_0 for the Grazzini (2012) test (G2012) is 'ergodicity' based on the first four moments, and H_0 for the Delli Gatti and Grazzini (2020) test (DGG2020) is 'ergodicity'. The figures are rounded to one or two decimal places.