University of Kent

School of Economics Discussion Papers

Macroeconomic simulation comparison with a multivariate extension of the Markov Information Criterion

Sylvain Barde

June 2019

KDPE 1908

Macroeconomic simulation comparison with a multivariate extension of the Markov Information Criterion

Sylvain Barde[a,*]

[a] *School of Economics, University of Kent*

Abstract

Comparison of macroeconomic simulation models, particularly agent-based models (ABMs), with more traditional approaches such as VAR and DSGE models has long been identified as an important yet problematic issue in the literature. This is due to the fact that many such simulations have been developed following the great recession with a clear aim to inform policy, yet the methodological tools required for validating these models on empirical data are still in their infancy.

The paper aims to address this issue by developing and testing a comparison framework for macroeconomic simulation models based on a multivariate extension of the Markov Information Criterion (MIC) originally developed in Barde (2017). The MIC is designed to measure the informational distance between a set of models and some empirical data by mapping the simulated data to the markov transition matrix of the underlying data generating process, and is proven to perform optimally (i.e. the measurement is unbiased in expectation) for all models reducible to a markov process. As a result, not only can the MIC provide an accurate measure of distance solely on the basis of simulated data, but it can do it for a very wide class of data generating processes.

The paper first presents the strategies adopted to address the computational challenges that arise from extending the methodology to multivariate settings and validates the extension on VAR and DGSE models. The paper then carries out a comparison of the benchmark ABM of Caiani et al. (2016) and the DGSE framework of Smets and Wouters (2007), which to our knowledge, is the first direct comparison between a macroeconomic ABM and a DGSE model.

JEL classification: B41, C15, C52, C63.

Keywords: Model comparison, Agent-based models, Validation methods.

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

The author is extremely grateful to participants at INET Oxford complexity seminar in the fall of 2018 for their helpful suggestions, especially Doyne Farmer, Adrián Carro, Blas Kolic, Luca Fierro and Marco Pangallo. Particular thanks goes to Matteo Richiardi, Jakob Grazzini, Franesco Lamperti, Sander van der Hoog and Marco Gross for very fruitful discussions relating to the multivariate extension strategy during the WEHIA 2017, CEF 2017 and CEF 2018 conferences. Finally, a big thanks goes to Mark Wallis for his support with the ICARUS cluster on which the model comparison exercise was run. The multivariate MIC toolbox can be downloaded from `https://github.com/Sylvain-Barde/mic-toolbox`, while the code base, simulated data and supplementary files required to replicate the results can all be downloaded from the following data repository `https://data.kent.ac.uk/80/`. Any errors in the manuscript remain of course the author's.

Non-Technical Summary

The paper develops and tests a multivariate extension of the Markov Information Criterion (MIC) originally developed in Barde (2017). The main motivation for the MIC is the problem of comparing the distance between a set of models and some empirical data for cases where estimation of the models with traditional methods is not feasible. This is often the case for simulation models such as agent-based models. The MIC performs this measurement by mapping the simulated data to the markov transition matrix of the underlying data generating process, and is proven to perform optimally (i.e. the measurement is unbiased in expectation) for all models reducible to a markov process. As a result, not only can the MIC provide a measure of distance solely on the basis of simulated data, but it can do it for a very wide class of data generating processes. This is illustrated in Barde (2016), which performs a comparison exercise between three agent based models (ABM) of financial markets and a set of ARCH-like models in order to rank them in terms of empirical performance.

The main drawback of the MIC in its original form is that the measurement of the informational distance to the data can only be carried out for univariate models, such as the ABM models of financial series mentioned above. In principle, there is no conceptual problem with extending the MIC to multivariate models, as the state of a markov process can be described by a vector of variables, rather than a single variable. In practice, however, increasing the number of variables needed to describe the state of system leads to a combinatorial explosion in the memory requirement of the context-tree weighting (CTW) algorithm of Willems et al. (1995), which forms the basis of the MIC. As a consequence, a naive extension to multivariate measurements is not possible. Instead, the paper uses a combination of three strategies to overcome this curse of dimensionality and extend the MIC to multiple variables.

- The first uses the fact that the most significant bit (MSB) of a context observation is more informative than the least significant bit (LSB). We therefore start by permuting the bits of the context so that the MSBs are processed first and the LSBs are processed last.
- Following from this, the second strategy is to truncate the context in order to keep the memory requirement bounded to a tractable level, and to prune single observation branches of the context tree, following the suggestion of Willems and Tjalkens (1997), in order to keep the tree as small as possible.
- Finally, because this truncation is expected to worsen the accuracy of the measurement, the final strategy is to take the average of multiple measurements in order to increase the performance. Crucially, this can be done simply by changing the order in variables are conditioned on, and does not require additional simulated or empirical data.

The extended methodology is validated by running two monte carlo model comparisons on VAR and DSGE models in order to evaluate ability of the multivariate MIC to rank these data generating processes relative to traditional methods. These validations establish that the desirable properties of the univariate MIC can be preserved despite the large increase in the state space and the smaller amount of data. Finally, we carry out a proof-of-concept macroeconomic model comparison exercise to demonstrate that the MIC can enable the direct comparison of ABM and DSGE models, which is a crucial step towards increasing the policy relevance of ABMs.

1. Introduction

The last decade has seen a fundamental shift in the 'technological readiness level' of agent-based computational economics (ACE), driven mainly by the joint maturation of agent-based models (ABMs) and the validation methodologies required to bring them to the data. As identified by Grazzini and Richiardi (2015, p.150), ACE is gradually transitioning from the purely qualitative replication of stylised facts and phenomena to a more quantitative replication based on 'sound econometric techniques'. Part of this shift is due to the evolution of ABMs themselves. Following Marks (2013), one should distinguish between two broad types of simulation models. The first kind are "demonstration models", which are typically small models that simply attempt to explore or illustrate a given mechanism. Good examples of such models are the classical Schelling (1971) model of segregation, the Kirman (1993) model of recruitment, the Howitt and Clower (2000) model of emergence of monetary equilibrium or the Abrams and Strogatz (2003) model of language competition. In each case the ABM highlights the effect of a single mechanism or behavioural rule on the aggregate dynamics of the system, and does not aim to provide a realistic simulation of a phenomenon. By contrast, models of the second, "descriptive", type aim to provide a realistic data generating process (DGP) than can replicate historical data, and can therefore be better harnessed for scenario or policy analysis. Fagiolo and Roventini (2012, 2017) provide an excellent set of surveys of such policy-orientated economic ABMs, as well as the potential contribution that ACE can make to informing policy makers. The authors also highlight some outstanding difficulties facing the field, in particular the validation problem we discuss below. Haldane and Turrell (2018) provide a similar survey, and advocate the use of ABM to inform policy-making by identifying types of policy questions that could benefit from ACE.

Several notable macroeconomic ABMs have been developed over the recent years with the explicit aim of providing an integrated framework within which the impact of policies and their interactions can be examined. The Keynes vs Schumpeter (KS) framework of Dosi et al. (2010) combines a Keynesian demand side with Schumpeterian innovation in productive firms and was designed to examine the interaction of supply-side and demand-side policies. This framework has since been used in Dosi et al. (2013) and Dosi et al. (2015) to investigate fiscal and monetary policy. A second example is the Eurace model of Deissenberg et al. (2008) and Teglio et al. (2010), which is a large-scale ABM of the European economy. Dawid et al. (2018) discuss its use for policy making while Dawid et al. (2019) provide a full review of the applications of this model to policy analysis, however a notable example is Raberto et al. (2018), who use the model to investigate the effect of banking and credit regulation on the economy. A third and more recent example of an integrated ABM framework is Caiani et al. (2016) who develop a fully stock-flow consistent model, with the aim of being able to generate the type of deep balance-sheet crisis seen during the great recession. This is used in Caiani et al. (2019) to analyse fiscal policy and in Schasfoort et al. (2017) to investigate monetary policy channels.

Other examples of such policy investigations exist beyond the large integrated frameworks presented above. For example, Baptista et al. (2016) build on the housing ABM model of Axtell et al. (2014) to investigate the impact of macro-prudential policy on housing markets. Ashraf et al. (2017) extend the Howitt and Clower (2000) model to investigate Basel I capital adequacy requirements on the banking sector. While the authors are open about the fact their model remains very stylised, their findings have implications about the tightness of lending standards on the performance of an economy in downturns. This work has since been extended by Popoyan et al. (2017), to investigate the effect of mechanisms in the Basel II and III frameworks.

This shift from demonstration to descriptive models, with a view to inform policy-making, has increased the need to validate the simulations models on empirical data, in order to ensure that they are in fact a suitable description of the phenomena they aim to model. ABM practitioners are acutely aware

of this requirement as well as the challenges involved in doing so. However, as pointed out by Marks (2013, p. 41), at the time "validation of any but very simple simulation models has been slow in appearing in the literature". Validation of an ABM requires overcoming two problems, both of which are complicated by to the fact that ABMs typically do not possess analytical descriptions and properties of the model instead need to be inferred from the simulated data they produce. The first is the estimation of the ABM's parameters from available empirical data, and the second is the comparison of or selection amongst various ABM specifications and traditional models. Several of the models previously mentioned models provide an immediate illustration of this second problem. Ashraf et al. (2017), Popoyan et al. (2017) and Raberto et al. (2018) all investigate the impact of macroprudential banking regulation. A policymaker may well want to know whether the models agree, or which offers the best predictions for a given scenario, or how well they fit the data relative to standard models estimated with traditional techniques.

This pressure to validate models has lead to the development of methodologies that can address both these issues, which is the second important and recent improvement in the field. Fagiolo et al. (2019) provide an excellent review of the problems posed by ABM validation as well as the new methods available to address them. Early approaches to estimation typically rely on the simulated methods of moments (SMM) proposed by Gilli and Winker (2003), a more recent version of which can be found in Grazzini and Richiardi (2015). Other recent developments of interest are the simulated maximum likelihood (SML) of Kukacka and Barunik (2017) as well as Grazzini et al. (2017) and Lux (2018), who investigate Bayesian and state-space estimation methods consistent with those used in more traditional macroeconomic DSGE models. A promising contribution which takes a very different approach to the estimation problem is that of Lamperti et al. (2018), who use machine leaning to build a surrogate model of the ABM in order to efficiently explore its parameter space.

Model comparison methods have seen similar developments. Marks (2013) provides a comparison of three distance measures between two vectors of data. Guerini and Moneta (2017) maps the structure of an ABM to a structural VAR in order to help identify the channels through which shocks are transmitted at the aggregate level. Two recent additions are the Generalized Subtracted L-divergence (GSL-div) of Lamperti (2018b) and the Markov Information Criterion (MIC) of Barde (2017), both of which have been used to perform empirical model comparison exercises. Lamperti (2018a) compares 5 distinct versions of the Brock and Hommes (1998) model of asset pricing with heterogeneous beliefs on the EuroSTOXX 50 and CSI 300, while Barde (2016) compares the performance of the Gilli and Winker (2003), Alfarano et al. (2005) and Franke and Westerhoff (2011) models of recruitment on a range of financial indices against standard ARCH/GARCH processes.

In most cases these new estimation and comparison methods are applied to relatively small scale or univariate models in an empirical setting where data is plentiful, typically a financial ABM applied to market index data. This is understandable given the new and often experimental nature of the methodologies involved. However, validation of the large-scale, policy-relevant ABMs discussed above will require demonstrating that these methodologies can be carried over to typical macroeconomic settings characterised by larger sets of observable variables and a smaller number of observations.[1] The present paper aims to address this issue by extending the univariate ABM comparison exercise of Barde (2016) to the more challenging multivariate macroeconomic setting. This first requires developing and validating an extension of the MIC algorithm of Barde (2017) to a multivariate state space. Once this is done we pro vide a proof of concept for multivariate model comparison by carrying out an ABM - DSGE comparison

[1] Exceptions to this general observation within the papers discussed are Grazzini et al. (2017) and Guerini and Moneta (2017). The former estimate 9 parameters for a macroeconomic ABM, although only 2 observable variables are used (inflation and output gap). In the latter, the causal structure of the Dosi et al. (2015) KS model is explored using a structural VAR on 6 US variables (consumption, investment, unemployment, GDP, inflation and federal funds rate). The comparison does not extend to other macroeconomic models, however.

exercise in the spirit of Fagiolo and Roventini (2012, 2017), using the Caiani et al. (2016) and Smets and Wouters (2007) models.

The remainder of the paper is organised as follows. Section 2 first presents the desirable theoretical properties of the MIC and provides a simple illustration of its effectiveness in univariate settings. Sections 3 and 4 then present the computational strategy used to extend the MIC to multivariate systems and the validation of the strategy, while section 5 presents the comparison exercise itself.

2. Theoretical properties of the univariate MIC

The MIC is a generalisation of the AIC Akaike (1974), in the sense that it provides an unbiased measurement of the cross entropy between the data and a model. Like the AIC, the difference in MIC across models is therefore a measurement of their relative Kullback and Leibler (1951) (KL) divergence. This indirect approach allows the use of the KL divergence to identify the best candidate model while getting around the standard problem that it is not computable in general. While Marks (2013, 2019) correctly points out that the KL divergence is not a true metric, as will be discussed in the first part of this section it possess solid information-theoretical foundations and also maps to likelihoods, and is therefore intimately linked to many model comparison/selection techniques.

The key practical deviation from Akaike (1974) is that instead of using the likelihood to estimate the value of the cross entropy, the MIC measures the cross entropy directly from the empirical and simulated ABM data. This relies on a modified version of the two-pass context tree maximisation (CTM) algorithm of Willems et al. (2006), which itself is an extension of the Willems et al. (1995) one-pass context tree weighting (CTW) algorithm. In the MIC protocol, simulated data is used in the CTM first stage in order to learn the probability structure of the simulation model, which is then used to compress the empirical data in the second stage.[2] The cross-entropy measurement is then simply the length of the compressed empirical data.

The sections below will introduce the notation used throughout the paper, as well as clarify the structure of the CTW/CTM algorithms and explain their desirable theoretical properties, in particular the correction of bias in the cross entropy measurement. This preliminary step is required because this data compression approach and its associated algorithms are likely to be unfamiliar. Furthermore, because the multivariate extension presented in section 3 aims to preserve the univariate MIC's properties, it is important that they be detailed as a first step.

2.1. Conditional likelihood as a biased measurement of cross entropy and Kullback Leibler divergence

Let $X = \{X_t \in S : t \in \mathbb{N}\}$ be a sequence of discrete random variables following a Markov chain of arbitrary order $L \in \mathbb{N}$ over a set of discrete symbols S. In addition, let the number of distinct symbols be fixed at $|S| \in \mathbb{N}$. The dynamics of this Markov chain are completely determined by a transition matrix P of size $|S|^L \times |S|$. Each entry in P is indexed by an ordered pair (Ω_t, x_t), and contains the corresponding probability $p\left(X_t = x_t \mid \Omega_t\right)$ of observing the realisation x_t conditional on $\Omega_t = \{x_{t-1}, x_{t-2}, ..., x_{t-L}\}$, an information set containing the last L symbols in the chain, which identifies the state of the system.

Suppose that the true transition matrix P is unobserved, so that the probabilities must therefore be modelled or approximated. Let $\widehat{P}^i$ be the transition matrix provided by model i, containing transition probabilities $\hat{p}^i\left(X_t = x_t \mid \Omega_t\right)$. As pointed out by Burnham and Anderson (2002), the following KL divergence per transition (Ω_t, x_t) in the chain then provides a theoretical measure of the distance from the model $\widehat{P}^i$ to the truth P.

[2] By contrast, in a pure one-pass data compression application, the data to be compressed is also used for learning the probabilities required for compression, and these are updated after each symbol is compressed. This is because the ultimate goal is to be able to eventually decompress the data: as a given symbol is decompressed, the probabilities can be updated, allowing the decompression of the following symbol.

$$D\big(P\,||\,\widehat{P}^i\big) = E_P\left[\ln\frac{p\left(X_t = x_t \mid \Omega_t\right)}{\hat{p}^i\left(X_t = x_t \mid \Omega_t\right)}\right] \tag{1}$$

Here $E_P\left[\ldots\right]$ indicates that the expectation is taken with respect to the limit distribution of the true process P. The KL divergence's usefulnes for the purpose of model selection is that due to Jensen's inequality and the concavity of the logarithm, $D\big(P\,||\,\widehat{P}^i\big) = 0$, iff $\widehat{P}^i = P$ and otherwise $D\big(P\,||\,\widehat{P}^i\big) > 0$, $\forall\ \widehat{P}^i \neq P$. Unfortunately, because P is not known, the KL divergence cannot be directly calculated. It is possible to make an indirect measurement, however, using cross entropy. Taking advantage of the linearity of the expectations operator to separate the logarithmic ratio in (1) and rearranging one obtains the following expression, which defines the cross entropy rate per transition (Ω_t, x_t):

$$H\big(P\,||\,\widehat{P}^i\big) = H\big(P\big) + D\big(P\,||\,\widehat{P}^i\big) \quad \text{with} \quad \begin{cases} H\big(P\,||\,\widehat{P}^i\big) = E_P\left[\ln\dfrac{1}{\hat{p}^i\left(X_t = x_t \mid \Omega_t\right)}\right] \\ H\big(P\big) = E_P\left[\ln\dfrac{1}{p\left(X_t = x_t \mid \Omega_t\right)}\right] \end{cases} \tag{2}$$

The cross entropy rate between the true P and a model $\widehat{P}^i$, labelled $H\big(P\,||\,\widehat{P}^i\big)$, is the sum of the KL divergence and a term, $H\big(P\big)$, which depends on P only and is therefore constant for all models $\hat{P}^i$. As a result, in the words of Burnham and Anderson (2002, p.58) when taking differences in cross entropy across models "truth drops out as a constant" and the differences in cross entropy across models reflect only differences in the KL divergence of each model to the truth.

$$H\big(P\,||\,\widehat{P}^i\big) - H\big(P\,||\,\widehat{P}^j\big) \quad = \quad D\big(P\,||\,\widehat{P}^i\big) - D\big(P\,||\,\widehat{P}^j\big) \tag{3}$$

This indirect approach to measuring KL divergence across models forms the foundation of many model selection methods. In essence, while it is not possible to know in absolute terms how close a given model $\widehat{P}^i$ is to the truth P, it is possible to compare two models $\widehat{P}^i$ and $\widehat{P}^j$ by calculating their relative distance to the truth.[3] Measurement of cross entropy is feasible because, unlike the KL divergence (1), it does not contain the (unknown) true probabilities P in the argument of the logarithm. Obtaining an accurate measurement, however, remains problematic for two related reasons. The first is that measuring (2) still requires taking an expectation with regards to the truth P. While this can be proxied by the empirical frequencies realised in the data x, this will induce a measurement error. A second problem is the fact that the model probabilities $\widehat{P}^i$ are themselves often not known with certainty, for example due to parameter uncertainty in the underlying model. The existing literature on the measurement of information quantities, such as Basharin (1959), Carlton (1969), Panzeri and Treves (1996) and Roulston (1999), point out that because such a measurement error enters the argument of an expected logarithm, Jensen's inequality will lead to the existence of a bias.

Identifying these biases forms the key goal of model selection methods based on KL divergence. A first important interpretation of cross entropy, which underpins the insight of Akaike (1974) and the AIC, is that it can be estimated using the likelihood function, as their theoretical specifications are related. In the context of the Markov process describes above, let $\tau\left(\Omega_t\right)$ count the number of occurrences of each state Ω_t and $\tau\left(\Omega_t, x_t\right)$ count the occurrences of a transition identified by the ordered pair (Ω_t, x_t). The empirical likelihood of a model $\widehat{P}^i$ given a realisation x of the Markov chain is then provided by the following expression, which is similar to a multinomial likelihood function.

[3] It also forms the basis of the caveat for such methods. The fact that in a comparison exercise one model is selected as the best in a relative sense should not be construed as implying that it is a good model: all the candidate models being compared on a dataset might well be very poor in an absolute sense! This forms one of the main criticisms to the use of KL divergence in Fagiolo et al. (2007) and Marks (2013).

$$\mathcal{L}\big(\widehat{P}^i \mid x\big) = \prod_{t=L}^{T} \hat{p}^i \left(X_t = x_t \mid \Omega_t\right)^{\tau(\Omega_t, x_t)} \tag{4}$$

Taking logarithms to obtain the log likelihood, dividing by the number of observed transitions in the chain $T - L$ and rearranging provides the mean contribution of a single transition to the empirical log likelihood:

$$\overline{\ln \mathcal{L}\big(\widehat{P}^i \mid x\big)} = \sum_{t=L}^{T} \frac{\tau\left(\Omega_t\right)}{T-L} \frac{\tau\left(\Omega_t, x_t\right)}{\tau\left(\Omega_t\right)} \ln \hat{p}^i \left(X_t = x_t \mid \Omega_t\right) \tag{5}$$

The two ratios weighting the logarithm of the model probabilities in the sum are respectively the empirical frequencies for the states and the transitions observed in the data. When taking the limit of this average likelihood as the length of the chain goes to infinity, assuming that the chain X is ergodic, these frequencies will converge almost surely to the limit distribution $\pi\left(\Omega_t\right)$ and the transition probabilities $p\left(X_t = x_t \mid \Omega_t\right)$. Asymptotically the expected contribution to the likelihood of a single transition is therefore the negative of the cross entropy rate (2) of the Markov process.

$$\lim_{T\to\infty} \overline{\ln \mathcal{L}\big(\widehat{P}^i \mid x\big)} = \sum\nolimits_{\Omega_t, x_t} \left[\pi\left(\Omega_t\right) p\left(X_t = x_t \mid \Omega_t\right) \ln \hat{p}^i \left(X_t = x_t \mid \Omega_t\right)\right] = -H\big(P \,||\, \widehat{P}^i\big) \tag{6}$$

For finite sample sizes, however, the empirical likelihood will only provide an approximation, furthermore, as already stated, any error in the model probabilities will also lead to a systematic bias. Suppose that the model probabilities depend on a parameter set θ of dimension $K = |\theta|$, and that the estimated parameters $\hat{\theta}$ are obtained by maximum likelihood (ML) estimation, such that $\widehat{P}^i = P^i\big(\hat{\theta}\big)$. Akaike (1974) establishes that in this case, under the assumption that the candidate models $P^i\big(\hat{\theta}\big)$ are good approximations to the truth P and the sample size is large, a good first-order approximation for the resulting bias is simply the size of the parameter set K. Correcting the empirical log likelihood for the bias provides an estimate of the cross entropy of the $T - L$ observations:[4]

$$H\big(P||P^i(\hat{\theta})\big) \times (T-L) \approx -\ln \mathcal{L}\big(P^i(\hat{\theta}) \mid x\big) + K \tag{7}$$

The requirement that the candidate models be close to the truth was investigated by Takeuchi (1976), who derives a more general specification of the bias estimate that does not require this assumption:

$$H\big(P||P^i(\hat{\theta})\big) \times (T-L) \approx -\ln \mathcal{L}\big(P^i(\hat{\theta}) \mid x\big) + \operatorname{Tr}\left(J(\hat{\theta}) I(\hat{\theta})^{-1}\right) \tag{8}$$

Here $J(\hat{\theta})$ is the outer product of the gradient of the likelihood $\nabla_\theta \ln \mathcal{L}\big(P^i(\hat{\theta}) \mid x\big)$ and $I(\hat{\theta})$ is the Fischer information matrix. Both are $K \times K$ matrices, evaluated at the ML value of the parameters $\hat{\theta}$. Once can see that the AIC is a special case of (8) as $J(\hat{\theta}) = I(\hat{\theta})$ when $P^i\big(\hat{\theta}\big) = P$ and argument of the trace term is the identity matrix. While the TIC is more general than the AIC, Burnham and Anderson (2002) point out that in practice, reliable estimation of $J(\hat{\theta})$ and $J(\hat{\theta})$ is difficult, and K often provides the best estimator of the trace term in (8). Furthermore, when $P^i\big(\hat{\theta}\big)$ is such a poor model of P that K is no longer a good estimate of the bias, the poor fit of the model reflected in $\ln \mathcal{L}\big(P^i(\hat{\theta}) \mid x\big)$ will dominate the resulting information criterion, making the inaccurate measurement of bias irrelevant.

Sugiura (1978) addresses the second core requirement of the AIC, which is the large sample size, and proposes a second order approximation which provides a more reliable estimate of the bias when the sample size is small. This sample corrected AIC ($\mathrm{AIC_c}$) simply adds a correction factor to the bias which depends on the degrees of freedom of the parameter estimation.

[4] In the calculation of the AIC this is multiplied by 2 'for historical reasons'

$$H\big(P||P^i(\hat{\theta})\big) \times (T-L) \approx -\ln \mathcal{L}\big(P^i(\hat{\theta}) \mid x\big) + K\frac{n}{n-K-1} \tag{9}$$

The bias derivations (7) - (9) all rely on the apparatus of ML estimation applied to parametric models $P^i(\theta)$ to provide estimates of the bias between cross entropy and the empirical likelihood. A priori, this makes an extension to simulation models rather complicated: most simulation models, particularly ABMs, do not possess a closed-form specification of the model probabilities as a function of the model parameters, i.e. $P^i(\theta)$ is typically not available. Instead, the model probabilities $\widehat{P}^i$ must be estimated directly from a run of simulated data $x^i(\theta)$, which can be considered as realisations of the Markov chain $X^i(\theta)$ corresponding to the ABM with parameter set θ. Clearly, as is the case for the parametric framework presented above, model probabilities $\widehat{P}^i$ estimated from simulation data $x^i(\theta)$ will contain a measurement error.

Crucially, it should be apparent that the core problem in obtaining an accurate cross entropy measurement for pure simulation models is not the calculation of the empirical cross entropy itself: as pointed by (6) this is directly provided by the empirical likelihood of the model probabilities. Instead, the hurdle is the calculation of $\widehat{P}^i\big(x^i(\theta)\big)$ in a manner that provides a well-behaved expression for the resulting bias. Unfortunately, the absence of a formal specification for these probabilities as a function of either the simulated data or underlying parameters complicates the evaluation of the size and properties of this measurement error, ruling out a systematic derivation of the bias along the lines of (7) - (9).

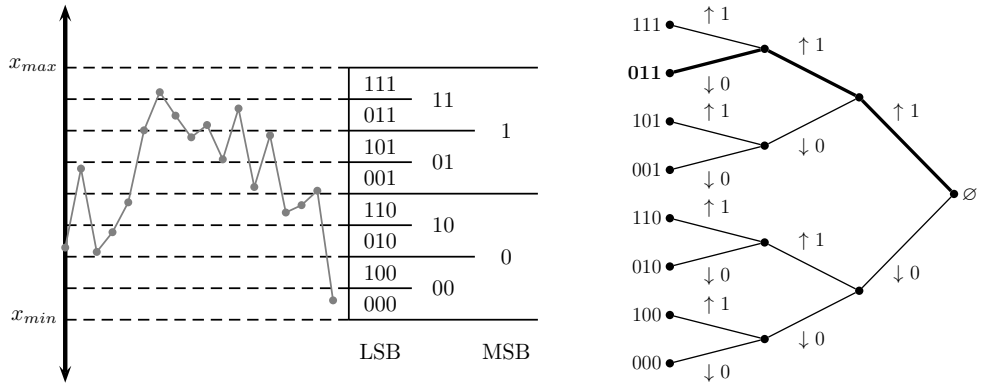
It is at this point that the data compression interpretation of cross entropy becomes helpful. In such a setting, the random variables in the Markov chain X are typically byte values encoding data to be compressed, and the true Markov transition table P governs the dynamics of the DGP.[5] Given this, the cross entropy rate (2) measures the expected lossless compression rate per symbol in the chain that can be achieved by using model probabilities $\widehat{P}^i$. The $H\big(P\big)$ term, known as the Shannon (1948) entropy, is the intrinsic information content of a symbol in the Markov chain and provides the fundamental limit to compression, as it can only be achieved if $\widehat{P}^i = P$. For the general case where $\widehat{P}^i \neq P$, the KL divergence (1) measures the increased size of the compressed data per symbol resulting from having to relying on approximate probabilities provided by a model $\widehat{P}^i$ rather than the truth P.

The practical utility of this interpretation for calculating cross entropy on the basis of simulated data is that in the typical data compression application, the model probabilities for the Markov chain are nearly always determined directly from the data to be compressed, i.e. data compression algorithms have to determine $\widehat{P}\big(x\big)$. The IT revolution, the rise of the internet and the resulting demands on communication bandwidth and data storage have motivated the search for efficient data compression methods, which has led to the development of a large array of methodologies and approaches in the literature, aimed at efficiently determine $\widehat{P}\big(x\big)$.[6] Key to achieving this is modelling general DGPs in a way that produces well-behaved error terms for probability estimates, thus producing small and predictable biases to cross entropy over a large range of data sources.

The CTM and CTW algorithms are chosen because they are proven to perform optimally over all Markov sources, i.e. the compression inefficiency incurred by the measurement error on the probabilities $\widehat{P}\big(x\big)$ is tightly bound above a theoretical bound. Barde (2017) shows that when using data compression as a method of measuring cross entropy between two datasets (empirical and simulated), this translates to providing an unbiased measurement. As will be shown below, this desirable property arises from the use of a context tree structure to learn the model probabilities, and the fact that the data is processed in a binary representation. Unfortunately, it is also the reliance on a context tree which creates computational

[5] The number of symbols in this case is $|S| = 256$, with the bytes encoding characters, numerical data, pixels, etc.

[6] A near exhaustive reference is provided by Salomon and Motta (2010), who ironically comment that the 1307 page handbook is far from a quick reference guide. The authors also review in introduction the reasons why data compression methods remain relevant, despite improvements in data storage capacity and internet connection speeds.



(a) Quantisation of real-valued data, $R = 3$

(b) Context tree of depth $D = 3$ bits

Figure 1: Binary data structures

challenges in extending the MIC to multivariate Markov sources, discussed in section 3. We now formally present the CTW and CTM algorithms and their role in the MIC methodology.

2.2. Structure of the binary discretisation and context trees

Supposing that the state space of the Markov chain X has size $|S| = 2^R, R \in \mathbb{N}$, each state x can be identified using a unique R-length binary representation. This provides a natural encoding scheme for cases where the discrete values of the Markov chain result from the discretisation of a real-valued variable, as illustrated in Figure 1(a) for a 3-bit resolution. Given a closed interval $[x_{min}, x_{max}]$ which bounds the support of the real-valued observations x, the most significant bit (MSB) of the binary representation encodes which half of the interval the observation occupies. Each subsequent bit added to the left of a code similarly locates the observation within the subinterval defined by that code. The last bit used to encode the observation is called the least significant bit (LSB) as it only provides information over a very small range of variation.[7]

This binary encoding scheme allows a Markov chain with arbitrary state spaces to be decomposed into a nested sequence of Bernoulli trials: rather than drawing x_t directly from 2^R possible values, it can be obtained by performing R Bernoulli trials, each with a parameter conditioned on the last L realisations of the chain and on the outcome of the previous trials. Letting $X_{r,t}$ and $x_{r,t}$ represent the r^{th} bit of X_t and x_t respectively, the transition probabilities of the Markov transition matrix can be expressed as follows.

$$p\left(X_t = x_t \mid \Omega_t\right) = \prod_{r=1}^{R} p\left(X_{r,t} = x_{r,t} \mid \Omega_t, \Phi_{r,t}\right) \quad \text{with} \quad \begin{cases} x_{r,t} \in \{0,1\} \\ \Phi_{1,t} = \varnothing \\ \Phi_{r,t} = \{x_{1,t}, x_{2,t}, ..., x_{r-1,t}\} \end{cases} \tag{10}$$

In addition to Ω_t, the information set containing the last L observations in the chain, each of the bit-level probabilities $p\left(X_{r,t} = x_{r,t}\right)$ are now also conditioned on a second information set $\Phi_{r,t}$. This set, which contains the realisations of the first $r-1$ bits of x_t, enables the r^{th} bit of an observation to be conditioned on the bits that precede it. For convenience, let $\Phi_t = \{\Phi_{r,t}\}$ be the super set of the R sets required for conditioning the bits of an observation x_t. In practice, the CTW algorithm uses binary strings, called contexts, to represent the information sets Ω_t and Φ_t. Given 2^R possible distinct values

[7] Clearly, this discretisation discards information through a truncation error, and the choice of R needs to be made carefully. Barde (2017) shows that as long as that this error is i.i.d uniformly distributed and uncorrelated with the discretised variable, it has no effect on the relative score of models. For further details, including which tests to carry out to select R, the reader is referred to Barde (2017).

for x_t and L lags in the Markov process, the information set Ω_t contains RL bits of information. Each possible realisation of Ω_t can therefore be mapped to a unique binary string ω_t of length RL, using a context hashing function $h_c(.)$.[8] Similarly, ϕ_t encodes the information of Φ_t through a trivial observation hashing function $h_o(.)$, which produces all the substrings of x_t of length zero to $R-1$.

$$\begin{cases} \omega_t = h_c\left(\Omega_t\right) \\ \phi_t = h_o\left(x_t\right) \end{cases} \tag{11}$$

The conditioning of a bit's transition probability $p\left(X_{r,t} = x_{r,t}\right)$ on the context information $\{\omega_t, \phi_{r,t}\}$ is carried out by using a set of binary context trees of depth D, an illustration of which is provided in figure 1(b) for $D = 3$ bits. The role of the contexts $\{\omega_t, \phi_{r,t}\}$ is to index a location in the set of trees, similar to the way the ordered pair (Ω_t, x_t) indexes the row and column in $\widehat{P}_i$ from which to draw $\hat{p}^i\left(X_t = x_t \mid \Omega_t\right)$. In fact, the role of the hashing functions (11) is simply to convert the information in the ordered pair (Ω_t, x_t) into a format suitable for indexing in the tree. Assuming that $D = RL$, the number of leaves in the binary tree matches the number of possible histories for the system and each binary string ω_t uniquely maps to a leaf in the tree. One binary tree can thus store the transition probabilities of a single Bernoulli trial, conditioned on all possible realisations of Ω_t. Given a discretisation resolution R for the random variable X_t, $2^R - 1$ such trees are therefore needed to store the full transition table, one per one per possible realisation of $\Phi_{r,t}$. In summary, ϕ_t allows us to select which tree to use and ω_t then identifies the correct leaf in that tree.

Intuitively, the 2^{RL} leaves on $2^R - 1$ trees provide a direct analogue to the Markov transition matrix $\widehat{P}^i$, as they will provide a frequency table for the simulated data x^i over the state space S, for each possible history Ω_t.[9] The benefit of using a set of trees instead of a frequency table is that the branch nodes allow for partial conditioning, in cases where a context ω_t is rarely (or never) observed, and the resulting leaf frequency might therefore contain a large measurement error. Given a context ω_t, one can trace a unique path from the corresponding leaf to the root, shown in bold in figure 1(b) for $\omega_t = 011$. Let $d \leq D$, with $d \in \mathbb{N}$ index the depth of a node on the path, such that $d = 0$ identifies the root of the tree and $d = D$ labels the leaf. Each node in the path corresponds to a partial context $\omega_{d,t}$, which is the d-length suffix of ω_t, and contains a set of counting functions $a^i\left(\omega_{d,t}, \phi_{r,t}\right)$ and $b^i\left(\omega_{d,t}, \phi_{r,t}\right)$, which respectively keep track of the number of zeros and ones observed in the realiation of the model Markov chain X^i for the partial context set $\{\omega_{d,t}, \phi_{r,t}\}$. A direct consequence of the binary structure of the tree is that the zero or one counters for a specific partial context $\omega^*_{d,t}$ can be obtained by summing the counters in the leaves of the subtree rooted in the node identified by $\omega^*_{d,t}$, where [...] is the Iverson bracket:

$$\begin{cases} a^i\left(\omega^*_{d,t}, \phi_{r,t}\right) = \displaystyle\sum_{\omega_t} a^i\left(\omega_t, \phi_{r,t}\right)\left[\omega_{d,t} = \omega^*_{d,t}\right] \\ b^i\left(\omega^*_{d,t}, \phi_{r,t}\right) = \displaystyle\sum_{\omega_t} b^i\left(\omega_t, \phi_{r,t}\right)\left[\omega_{d,t} = \omega^*_{d,t}\right] \end{cases} \tag{12}$$

The counters in nodes corresponding to a partial context $\omega_{d,t}$ truncated at a low depth d will be updated more often than in nodes for which the partial context is truncated at a depth closer to D. At either end of the leaf-to-root path, the counter in the leaves are incremented only for their specific context ω_t, while the root of the tree $\omega_{0,t}$ is included in every path, and its counters will therefore be updated for every full context in the training data. When a leaf is not observed very often, drawing probabilities from

[8] Defining $h_c(.)$ this explicitly may seem like an unnecessary step: after all, if x_{t-1}, x_{t-2}, etc. are already in a binary representation, the context ω_t is simply the concatenation of the conditioning observations. This is indeed true for univariate processes, however, when moving to the multivariate case this natural definition is less evident as there is no obvious way to order the bits from different variables. The choice of $h_c(.)$ will become an important degree of freedom for the researcher.

[9] As explained in section 2.1, the Markov matrix for the system is $2^{RL} \times 2^R$, the binary decomposition relies on the probabilities summing to one to eliminate one column.

nodes closer to the root will reduce the measurement error in the frequency caused by using low counts, at the cost of introducing another error by only partially conditioning on the context. The efficiency of the CTW algorithm stems from its ability to optimally solve the trade-off between these two sources of error.

2.3. Efficient calculation of probabilities and biases

Expressing the transition probabilities of the Markov chain X^i as a set of nested Bernoulli trials (10) and using a context tree to store them is central to the performance of the methodology as this reduces the problem of estimating the multinomial transition probabilities from a realisation x^i to the estimation of a set of Bernoulli parameters. For all possible partial context sets $\{\omega_{d,t}, \phi_{r,t}\}$ the counting functions $a^i(\omega_{d,t}, \phi_{r,t})$ and $b^i(\omega_{d,t}, \phi_{r,t})$ can be used in conjunction with the Krichevsky and Trofimov (1981) (KT) to provide optimal estimates of the corresponding Bernoulli parameter:

$$\begin{cases} \hat{p}^i(X_{r,t} = 0 \mid \omega_{d,t}, \phi_{r,t}) = \dfrac{a^i(\omega_{d,t}, \phi_{r,t}) + \frac{1}{2}}{a^i(\omega_{d,t}, \phi_{r,t}) + b^i(\omega_{d,t}, \phi_{r,t}) + 1} \\ \hat{p}^i(X_{r,t} = 1 \mid \omega_{d,t}, \phi_{r,t}) = \dfrac{b^i(\omega_{d,t}, \phi_{r,t}) + \frac{1}{2}}{a^i(\omega_{d,t}, \phi_{r,t}) + b^i(\omega_{d,t}, \phi_{r,t}) + 1} \end{cases} \tag{13}$$

The KT estimator for Bernoulli parameters (13) is proven to be optimal, in the sense that the measurement error induced by relying on observed frequencies to approximate a real-valued probability is tightly bound. Without loss of generality, let $a^i(.)$ and $b^i(.)$ represent the counting functions for an arbitrary binary sequence x^i generated by a Bernoulli process with parameter θ^i, and let $P^e\big(a^i(.), b^i(.)\big)$ be the one-pass estimated probability of observing this sequence, where $a^i(.)$, $b^i(.)$ and the KT estimate of the corresponding Bernoulli parameter (13) is updated after observing each bit value. This probability is initialised as $P^e(0,0) = 1$ and is updated recursively using (13) after each realisation x_t^i as follows:[10]

$$\begin{cases} P^e\big(a^i(.)+1, b^i(.)\big) = \dfrac{a^i(.) + \frac{1}{2}}{a^i(.) + b^i(.) + 1} P^e\big(a^i(.), b^i(.)\big) \\ P^e\big(a^i(.), b^i(.)+1\big) = \dfrac{b^i(.) + \frac{1}{2}}{a^i(.) + b^i(.) + 1} P^e\big(a^i(.), b^i(.)\big) \end{cases} \tag{14}$$

Intuitively, compressing the sequence x^i using the one-pass KT probabilities (14) will be inefficient compared to what could be achieved using the true, but unobserved, parameter θ^i. Indeed, even if the sequence x^i is long enough to allow the KT estimator (13) to converge to θ^i, performance on the initial observations will be poor. A critical property of the KT estimator is that this difference between the entropy of the binary sequence obtained using (14) and true likelihood (4) based on θ^i is bounded above:

$$\log_2 \frac{1}{P^e\big(a^i(.), b^i(.)\big)} - \log_2 \frac{1}{(1-\theta^i)^{a^i(.)} \theta^{b^i(.)}} \quad \leq \quad \frac{1}{2} \log_2 \big(a^i(.) + b^i(.)\big) + 1 \tag{15}$$

In a series of key contributions, Rissanen (1984, 1986) shows that the compression efficiency cost of updating any estimator of θ^i as the bits are processed possesses a theoretical lower bound in expectation, which for the case of the KT estimator is $\frac{1}{2} \log_2 \big(a^i(.) + b^i(.)\big)$. The fact that upper bound (15) is only one

[10] One can see that this initialisation is equivalent to an uninformative prior, as combined with (14), this leads to $P^e(1,0) = P^e(0,1) = 0.5$

bit above this theoretical lower bound (known as the Rissanen bound) is what makes the KT estimator optimal for binary sources.[11] Furthermore, by recasting a Markov chain as a sequence of Bernoulli trials, the binary encoding scheme (10) extends the optimality of the CTW algorithm to Markov chains with arbitrary state spaces $|S| > 2$. This is confirmed by Begleiter et al. (2004), who show that such an extended CTW algorithm typically outperforms other compression algorithms in practice.

While these tight bounds apply to one-pass compression, where the probabilities are updated as observations are compressed, Barde (2017) shows that in a two-pass application, where the probabilities are first determined from a simulated chain X^i and then applied to compress an observation from an empirical chain X_t, the increase in the Rissanen bound for the extra observation provides a measurement of the bias, allowing the empirical measurement of cross entropy of each individual bit to be corrected. This bias term can be calculated for every node in the context trees identified by its partial context set $\{\omega_{d,t}, \phi_{r,t}\}$:

$$\epsilon^i\left(x_{r,t} \mid \omega_{d,t}, \phi_{r,t}\right) = \frac{1}{2}\log_2 \frac{a^i\left(\omega_{d,t}, \phi_{r,t}\right) + b^i\left(\omega_{d,t}, \phi_{r,t}\right) + 1}{a^i\left(\omega_{d,t}, \phi_{r,t}\right) + b^i\left(\omega_{d,t}, \phi_{r,t}\right)} \tag{16}$$

The second issue to resolve is the choice of depth d from which to calculate $\hat{p}^i\big(X_{r,t} = 1\big)$ given the full context set $\{\omega_t, \phi_{r,t}\}$. As explained in section 2.2, $\{\omega_t, \phi_{r,t}\}$ indexes a unique leaf in a tree, and thus a leaf-to-root path, giving the option of D different nodes to choose (13) from. This can be done by generating a weighted probability in each node, which is a mixture of the KT probabilities (13) of the leaf nodes from the subtree rooted in that node. These weighted probabilities are obtained from the odds ratio of a node $\eta^i\left(\omega_{d,t}, \phi_{r,t}\right)$:

$$\begin{cases} \hat{p}^w\left(X_{r,t} = 0 \mid \omega_{d,t}, \phi_{r,t}\right) = \dfrac{\eta^i\left(\omega_{d,t}, \phi_{r,t}\right)}{\eta^i\left(\omega_{d,t}, \phi_{r,t}\right) + 1} \\ \hat{p}^w\left(X_{r,t} = 1 \mid \omega_{d,t}, \phi_{r,t}\right) = \dfrac{1}{\eta^i\left(\omega_{d,t}, \phi_{r,t}\right) + 1} \end{cases} \tag{17}$$

The odds $\eta^i\left(\omega_{d,t}, \phi_{r,t}\right)$ are themselves calculated recursively on any leaf-to-root path. In the leaf, where $d = D$, the odds are simply:

$$\eta^i\left(\omega_{D,t}, \phi_{r,t}\right) = \frac{\hat{p}^i\left(X_{r,t} = 0 \mid \omega_{D,t}, \phi_{r,t}\right)}{\hat{p}^i\left(X_{r,t} = 1 \mid \omega_{D,t}, \phi_{r,t}\right)} \tag{18}$$

It is obvious from (17) and (18) that $\hat{p}^w\left(X_{r,t} = x_t \mid \omega_{D,t}, \phi_{r,t}\right) = \hat{p}^i\left(X_{r,t} = x_t \mid \omega_{D,t}, \phi_{r,t}\right)$, therefore no weighting occurs in the leaf. For nodes on the branches of the tree, where $d < D$, the odds $\eta^i\left(\omega_{d,t}, \phi_{r,t}\right)$ are a weighted mixture of the KT probabilities (13) in that node and the weighted probabilities (17) of the child node on the path:

$$\eta^i\left(\omega_{d,t}, \phi_{r,t}\right) = \frac{\hat{p}^i\left(X_{r,t} = 0 \mid \omega_{d,t}, \phi_{r,t}\right)\beta^i\left(\omega_{d,t}, \phi_{r,t}\right) + \hat{p}^w\left(X_{r,t} = 0 \mid \omega_{d+1,t}, \phi_{r,t}\right)}{\hat{p}^i\left(X_{r,t} = 1 \mid \omega_{d,t}, \phi_{r,t}\right)\beta^i\left(\omega_{d,t}, \phi_{r,t}\right) + \hat{p}^w\left(X_{r,t} = 1 \mid \omega_{d+1,t}, \phi_{r,t}\right)} \tag{19}$$

The mixture is controlled by $\beta^i\left(\omega_{d,t}, \phi_{r,t}\right)$, which is a probability ratio $\in [0, \infty[$. Willems et al. (2006) refer to this variable as a 'switch', which controls whether a node contributes information to the weighted probability or not. This can be seen by taking the limits of (19) for the extreme values of $\beta^i\left(\omega_{d,t}, \phi_{r,t}\right)$.

[11] As discussed by Willems et al. (1995), such an upper bound does not exist for the more familiar Laplace estimator, which is why the CTW algorithm relies instead on KT probabilities (13) to prove optimality of the CTW algorithm over Markov chains of arbitrary order

$$\begin{cases} \eta^i\left(\omega_{d,t},\phi_{r,t}\right) \to \dfrac{\hat{p}^i\left(X_{r,t}=0 \mid \omega_{d,t},\phi_{r,t}\right)}{\hat{p}^i\left(X_{r,t}=1 \mid \omega_{d,t},\phi_{r,t}\right)} & \text{as } \beta^i\left(\omega_{d,t},\phi_{r,t}\right) \to \infty \\ \eta^i\left(\omega_{d,t},\phi_{r,t}\right) = \dfrac{\hat{p}^w\left(X_{r,t}=0 \mid \omega_{d+1,t},\phi_{r,t}\right)}{\hat{p}^w\left(X_{r,t}=1 \mid \omega_{d+1,t},\phi_{r,t}\right)} & \text{for } \beta^i\left(\omega_{d,t},\phi_{r,t}\right)=0 \end{cases} \tag{20}$$

In the first case, as $\beta^i\left(\omega_{d,t},\phi_{r,t}\right)$ becomes large, the odds are determined by the KT estimator only. The interpretation is that the depth d node can actually be treated as a leaf, as its odds tends towards (18). In the second case, $\eta^i\left(\omega_{d,t},\phi_{r,t}\right) = \eta^i\left(\omega_{d+1,t},\phi_{r,t}\right)$ therefore the node simply transmits the odds it has received and can thus be considered as 'switched off'. The node switches $\beta^i\left(\omega_{d,t},\phi_{r,t}\right)$ associated to a context tree are therefore the crucial component determining the informational structure of that tree. The values of $\beta^i\left(\omega_{d,t},\phi_{r,t}\right)$ are updated recursively every time the node is on the leaf-to-root path corresponding to a transition in the simulated data x^i. Assuming that a node identified by a partial context $\{\omega_{d,t},\phi_{r,t}\}$ was last visited in period $t-k$, the node's β switch is updated using the ratio of the KT probabilities (13) to weighted probabilities (17) for that transition:

$$\beta^i\left(\omega_{d,t},\phi_{r,t}\right) = \beta^i\left(\omega_{d,t-k},\phi_{r,t-k}\right)\frac{\hat{p}^i\left(X_{r,t}=x_{r,t} \mid \omega_{d,t},\phi_{r,t}\right)}{\hat{p}^w\left(X_{r,t}=x_{r,t} \mid \omega_{d+1,t},\phi_{r,t}\right)} \tag{21}$$

This recursive updating rule implies that $\beta^i\left(\omega_{d,t},\phi_{r,t}\right)$ is the relative likelihood of two different models for the binary sequence observed by that node. The use of the KT probabilities (13) in the numerator of the updating rule (21) means that the numerator of $\beta^i\left(\omega_{d,t},\phi_{r,t}\right)$ is equivalent to (14) and thus measures the likelihood of a single Bernoulli source where the parameter is provided by the KT estimator for the node. The denominator is instead the likelihood of a mixture of two Bernoulli sources, where each parameter is estimated using a weighted mixture of all the leaves in the two sub-trees rooted in the child nodes of $\{\omega_{d,t},\phi_{r,t}\}$. The $\beta^i\left(\omega_{d,t},\phi_{r,t}\right)$ ratio therefore tracks the relative performance of a very simple model with a small measurement error (the KT estimator, with aggregated counts) against a more complex mixture model containing more conditioning information. If the simple model dominates the relative likelihood, the value of $\beta^i\left(\omega_{d,t},\phi_{r,t}\right)$ will be large, and as shown by (20), the node will weigh the KT probabilities more heavily. Conversely if the complex model dominates, the value of $\beta^i\left(\omega_{d,t},\phi_{r,t}\right)$ will be small and the node will favour the weighted probability.

Given this informational structure, Willems et al. (1995) show that the best possible performance for the one-pass CTW algorithm is obtained when compressing each observation in a Markov chain X_t by drawing the weighted probability from the root of the tree, i.e. using $\hat{p}^w\left(X_{r,t}=x_{r,t} \mid \omega_{0,t},\phi_{r,t}\right)$, as the observation is used to update the tree. However, the one-pass CTW algorithm is not much use for our purpose. First of all, a two pass algorithm is needed, where the tree is trained using a simulated Markov chain X^i, and the resulting probabilities are scored on the transitions observed in the empirical data X. Secondly, it is important to use KT rather than weighted probabilities, so that the bias incurred in the measurement can be calculated from the $a^i(.)$, $b^i(.)$ counts using (16).

Fortunately, the switch property of the β ratios identified in (20) can be used to identify the best sub-tree from which to draw KT probabilities. First, let us start with the observation that it is very likely that $\beta^i\left(\omega_{0,t},\phi_{r,t}\right) \approx 0$, in other words, the root node is switched off and simply transmits weighted probabilities from the incoming branched of the tree.[12] Finding a good sub-tree simply requires traversing every root-to-leaf path on the tree and stopping at the first node that can be treated as a leaf, as indicated by $\beta^i\left(\omega_{0,t},\phi_{r,t}\right)$. This idea is formalised in Willems et al. (2006), who define $Q^i\left(\omega_{d,t},\phi_{r,t}\right)$ as the maximum a posteriori (MAP) probability that a given tree node is a leaf in the best sub-tree. For the leaf nodes of

[12] If instead the root node is 'switched on', i.e. $\beta^i\left(\omega_{0,t},\phi_{r,t}\right) >> 1$, then this actually implies that the best model for the binary sequence is simply the KT estimator in the root node. In such a case, the entire tree is redundant and conditioning on past observations is not necessary.

the full tree, the MAP probability of being a leaf must be $Q^i(\omega_{D,t}, \phi_{r,t}) = 1$, by construction. For nodes located a depth $d < D$, the MAP probability can be calculated recursively from the relative likelihoods in the node during the first pass of the CTW algorithm, while the set of context trees is being trained on the simulated data X^i:

$$Q^i(\omega_{d,t}, \phi_{r,t}) = \max\left[\frac{Q_0^i(\omega_{d,t}, \phi_{r,t})\, Q_1^i(\omega_{d,t}, \phi_{r,t})}{1 + \beta^i(\omega_{d,t}, \phi_{r,t})}, \frac{\beta^i(\omega_{d,t}, \phi_{r,t})}{1 + \beta^i(\omega_{d,t}, \phi_{r,t})}\right] \tag{22}$$

Here the $Q_0^i(.)$ and $Q_1^i(.)$ notation indicates the MAP probabilities of the two child nodes of the parent node $Q^i(.)$. If the first term is the largest, this indicates that the best performance is achieved by mixing the probabilities of the child nodes, therefore the current node is a branch. If instead the second term is the largest, then the node should be treated as a leaf. This updating rule can be used to identify the optimal nodes from which to draw probabilities when scoring the transitions in the empirical data X in the second pass of the algorithm. Specifically, given a transition identified by a context set $\{\omega_t, \phi_{r,t}\}$, one starts in the root of the tree and traverses the path to the corresponding leaf, comparing the numerators in the argument (22) until one finds the first node for which the $\beta^i(.)$ ratio dominates the product of the child node probabilities $Q_0^m(.)$ and $Q_1^m(.)$:

$$d^* = \min\left\{d : Q_0^i(\omega_{d,t}, \phi_{r,t})\, Q_1^i(\omega_{d,t}, \phi_{r,t}) < \beta^i(\omega_{d,t}, \phi_{r,t})\right\} \tag{23}$$

The CTM empirical cross entropy for a transition (Ω_t, x_t) according to model i is then simply the sum of the R bit-level entropies, using probabilities drawn at the optimal depth (23):

$$\lambda^i(x_t \mid \Omega_t) = -\sum_{r=1}^{R} \log_2\left[\hat{p}^i(X_{r,t} = x_{r,t} \mid \omega_{d^*,t}, \phi_{r,t})\right] \tag{24}$$

The bias induced by using the KT estimates of the unobserved model probabilities is the sum of the bit-level bias (16), with the counts taken at the optimal depth (23).

$$\epsilon^i(x_t \mid \Omega_t) = \frac{1}{2}\sum_{r=1}^{R} \log_2 \frac{a^i(\omega_{d^*,t}, \phi_{r,t}) + b^i(\omega_{d^*,t}, \phi_{r,t}) + 1}{a^i(\omega_{d^*,t}, \phi_{r,t}) + b^i(\omega_{d^*,t}, \phi_{r,t})} \tag{25}$$

As explained in Barde (2017), the expected bias (25) can be substracted from the raw cross-entropy measurement (24) to obtained the corrected MIC measurement for that transition, which is unbiased in expectation.

$$\lambda_c^i(x_t \mid \Omega_t) = \lambda^i(x_t \mid \Omega_t) - \epsilon^i(x_t \mid \Omega_t) \tag{26}$$

Finally, the MIC for the entire empirical sequence x with respect to the simulated training data x^i is simply the sum of the observation-level scores (26).

$$\lambda_c^i(x) = \sum_{t=L}^{T} \lambda_c^i(x_t \mid \Omega_t) \tag{27}$$

The fact that the aggregate MIC (27) is the sum of an observation-level vector (26) means that one can test the relative statistical significance of measurements obtained for several models X^i, for example using the model confidence set (MCS) procedure of Hansen et al. (2011).

2.4. An illustration of the univariate MIC

It is helpful at this point to illustrate the two key properties of the univariate MIC, which are the link between binary cross entropy measurement and the conditional likelihoods as well as the reduction

Table 1: NOLH ARMA-ARCH model parameters

	a_1	a_2	b_1	b_2	c_0	c_1	c_2
Central value	0.5	0.25	0.2	0.2	0.25	0.5	0.3
NOLH shock range	± 0.1	± 0.1	± 0.1	± 0.1	± 0.1	± 0.1	± 0.1

in bias relative to a naïve alternative to determining model probabilities from simulated data, in this case kernel density estimation (KDE). In order to do so, we run a simple Monte Carlo exercise based on the following DGP, which is similar to the one used in Barde (2017) and is composed of two AR lags, two MA lags and two ARCH lags:

$$\begin{cases} X_t = a_1 X_{t-1} + a_2 X_{t-2} + b_1 \sigma_{t-1}\varepsilon_{t-1} + b_2 \sigma_{t-2}\varepsilon_{t-2} + \sigma_t \varepsilon_t \\ \sigma_t^2 = c_0 + c_1 \varepsilon_{t-1}^2 + c_2 \varepsilon_{t-2}^2 \end{cases} \tag{28}$$

The central values for the parameter set $\theta = \{a_1, a_2, b_1, b_2, c_0, c_1, c_2\}$ are provided in table 1. A set of 128 alternative models is generated by varying the parameters within a range of $[-0.1, 0.1]$ around their central values using the nearly orthogonal latin hypercube (NOLH) design of experiment proposed by Cioppa (2002) and Cioppa and Lucas (2007). This provides a 129×7 matrix of shocks $\in [-1, 1]$ which are multiplied by 0.1 and added to the central value above.[13] The NOLH sampling approach provides a setting where all parameters can be varied orthogonally, and the hypercube of the parameter space within which models are drawn is well covered by the sample. Figure A-1 in appendix A shows the two-way scatter plots for the 129 sets of parameter shocks $\Delta\theta^i$ illustrating the good space-filling properties of the NOLH design matrix.

A key motivation for the choice of the ARMA-ARCH specification (28) is that one can easily compute the conditional log likelihood of a particular realisation of x^i, obtained with parameter θ^i, given an alternate parameter vector θ^j:

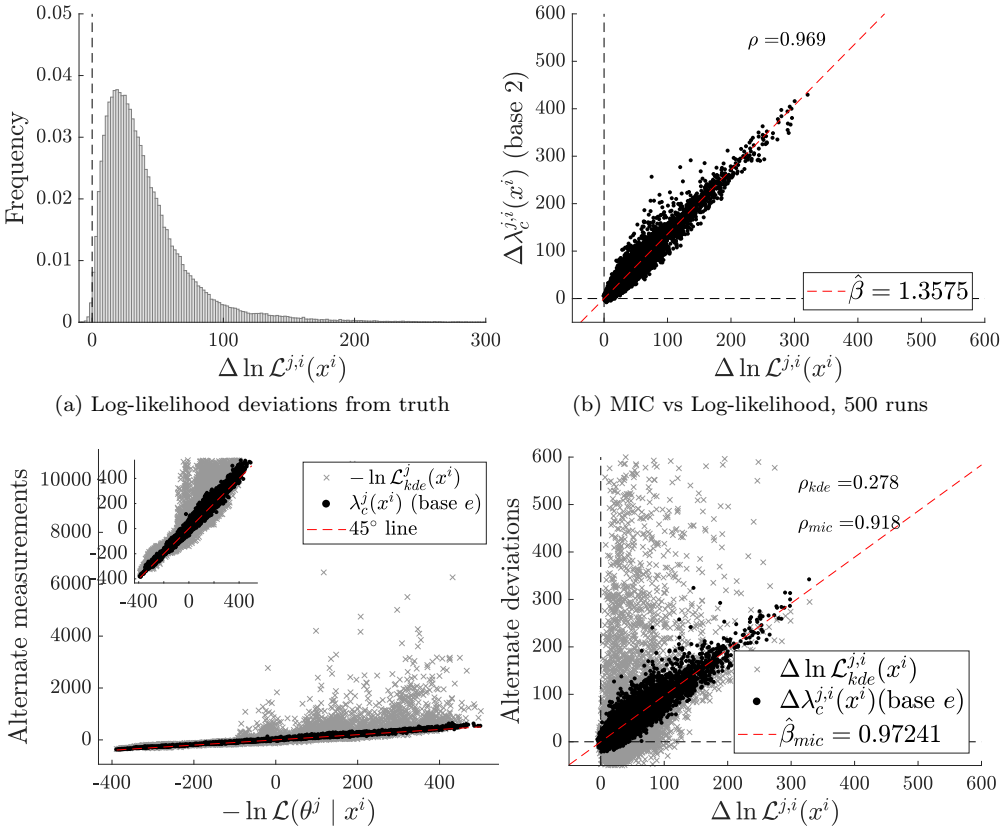
$$\begin{cases} \ln \mathcal{L}\left(\theta^j \mid x^i\right) = -\dfrac{T-2}{2\ln(2\pi)} - \dfrac{1}{2}\sum_{t=3}^{T} \ln(s_t) - \dfrac{1}{2}\sum_{t=3}^{T} \dfrac{u_t}{s_t} \\ u_t = x_t^i - a_1^j x_{t-1}^i - a_2^j x_{t-2}^i - b_1^j u_{t-1} - b_2^j u_{t-2} \\ s_t = c_0^j + c_1^j \left(u_{t-1}\right)^2 + c_2^j \left(u_{t-2}\right)^2 \end{cases} \tag{29}$$

This allows the calculation of the following average log deviation, which measures the expected distance per observation between θ^j and the 'true' data-generating parameter vector θ^i. In line with (6), the negative sign is included to make the likelihood comparable to a cross entropy, as is done when calculating the AIC from a likelihood.

$$\Delta \ln \mathcal{L}^{j,i}(x^i) = -\left[\ln \mathcal{L}\left(\theta^j \mid x^i\right) - \ln \mathcal{L}\left(\theta^i \mid x^i\right)\right] \tag{30}$$

10 simulated realisations x^i of 1000 observations are generated for each of the 129 parametrisations of (28). With $M = 129$ distinct models available, this provides $10 \times M(M-1) = 165120$ possible pairwise comparisons between a data realisation x^i and a model θ^j. Figure 2(a) shows the distribution of these pairwise comparisons according to their log likelihood distance (30), and the aim of the exercise is to establish if alternate measures of the likelihood, obtained directly from simulated data, can replicate this distribution.

[13] By construction, one of the 129 samples has a zero-valued shock for all parameters, and therefore corresponds to the central value in table 1



(a) Log-likelihood deviations from truth

(b) MIC vs Log-likelihood, 500 runs

(c) MIC & KDE vs Log-likelihood, levels

(d) MIC & KDE vs Log-likelihood, 250 runs

Figure 2: Univariate MIC scores versus KDE and log-likelihood

The training data for the MIC algorithm consists of 500 simulated runs of 1000 observations for all M models, which are all discretised to a resolution $R = 7$ on the $[-30, 30]$ interval. Figure 2(b), which shows the scatter plots of the mean relative MIC measurement $\Delta\lambda_c^{j,i}(x^i)$ against $\Delta \ln \mathcal{L}^{j,i}(x^i)$, confirms that the MIC is a reliable, if slightly noisy, measure of the relative conditional likelihood of model parameters θ^j and θ^i given x^i. Interestingly, the high correlation (0.969) between the two enables us to verify that the proportionality ratio between the two is close to the base conversion factor of $1/\ln 2 \approx 1.44$ that one would expect given that the likelihoods (29) are calculated using the natural logarithm, while the MIC (26) uses the binary logarithm.

In order to illustrate the biases generated by using estimates of the model densities with unknown measurement errors, we use 'naïve' KDE to generate an estimated value of the likelihood (29) directly from the training data. The conditional log likelihood is the difference of the log of two kernel density estimates, one based on the estimated probability of observing the current variable and its two lags, the second containing the joint probability of the lags only.

$$\ln \mathcal{L}^j_K(x^i) = \sum_{t=3}^{T} \left[\ln \hat{p}^j_{K^3}(x^i_t, x^i_{t-1}, x^i_{t-2}) - \ln \hat{p}^j_{K^2}(x^i_{t-1}, x^i_{t-2}) \right] \tag{31}$$

Here $\hat{p}^j_{K^3}(.)$ and $\hat{p}^j_{K^2}(.)$ are estimated on the j^{th} training dataset using a 3D and 2D gaussian kernel respectively, and a cross validated bandwidth. It is important to note, finally, that due to computational requirements, only 250 training series were used for the KDEs. The KDE-based likelihood (31) can be used to generate a mean deviation from the true model $\Delta \ln \mathcal{L}^{j,i}_K(x^i)$ in the same manner as the theoretical likelihood (30). Figures 2(c) and 2(d) provide the scatter plots for $\ln \mathcal{L}^j_K(x^i)$ against $\ln \mathcal{L}\left(\theta^j \mid x^i\right)$ and $\Delta \ln \mathcal{L}^{j,i}_K(x^i)$ against $\Delta \ln \mathcal{L}^{j,i}(x^i)$ respectively. For comparison, both plots include the $\lambda^j_c(x^i)$ and $\Delta\lambda^{j,i}_c(x^i)$

measurements obtained using the same 250 series of training data. In order to be able to superimpose these measures in the same diagram, both $\ln \mathcal{L}_K^j(x^i)$ and $\lambda_c^j(x^i)$ are centred using their respective median values, and the MIC measurements are converted to base e. While clearly it is possible to use more efficient methods than KDE to estimate the likelihoods from the training data, the purpose of figure 2(c) is to illustrate the argument made in section 2.1 that naïvely estimating the transition probabilities from simulated data can generate considerable upward bias. Furthermore, as shown in figure 2(d), this bias can be large enough that the relative likelihoods obtained for two models no longer enables the reliable identification of the better model. This is not the case for the MIC measurement, and the main challenge with the extension to the multivariate case is ensuring that these properties are conserved.

3. Extending the MIC to multivariate settings

Let $\mathbf{X}_t = \{X_t^1, X_t^2, ..., X_t^V\}$ be part of a multivariate Markov chain, where each of the V random variables can be used to represent an empirical observable following the notation set up in section 2.2, and let $\mathbf{X}_t^i = \{X_t^{i,1}, X_t^{i,2}, ..., X_t^{i,V}\}$ be a simulated multivariate Markov chain produced by model i for the same V variables. If $\mathbf{r} = (R^1, R^2, ..., R^V)$ is the corresponding vector of discretisation resolutions for these V variables, then the total number of bits required to describe one realisation of either $\mathbf{X}_t$ or $\mathbf{X}_t^i$ is $R^{mv} = \sum_{v=1}^{V} R^v$. From a conceptual point of view, extending the MIC framework presented in section 2 to such a multivariate setting is not a problem, as the transition probabilities to a particular realisation $\mathbf{x}_t$ conditional on the past L realisations $\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, ..., \mathbf{x}_{t-L}$ can still be stored in a suitable context tree.

In practice, however, a direct extension of the univariate MIC framework of section 2.2 to a state space of size R^{mv} is not feasible due to the curse of dimensionality. Because the CTW algorithm relies on binary trees, the memory requirement of the algorithm increases exponentially with the depth of the trees and given L lags in the Markov process, attempting to set the depth of the tree to $D = R^{mv}L$ will lead to intractable requirements. As an illustration, the univariate example in section 2.1, which follows Barde (2016), uses a $R = 7$ bit resolution and $L = 3$ lags of memory, for a context size of $RL = 21$ bits. Keeping the same resolution for a 3-variable system would result in a context size of $R^{mv}L = 63$. While the worst-case memory requirement of 2^{21} nodes per tree is tractable in the univariate case, the 2^{63} requirement for the 3-variable system is clearly not. This is compounded by the fact that the binary decomposition (10) of the current observation $\mathbf{x}_t$ would require $2^{R^{mv}} - 1$ binary trees to provide the conditioning set Φ_t, which is needed to condition on the observation bits. In the univariate case with $R = 7$, 127 trees are needed. In a naïve multivariate extension with $R^{mv} = 21$ this would increase to $2^{21} - 1$ trees, which again is simply intractable.

The computational strategy for extending the MIC to a multivariate setting therefore centres on reducing the dimensionality of Ω_t and Φ_t in order to keep the memory requirements tractable. The first element of the strategy is to use the chain rule for entropy to decompose the multivariate MIC into a sum of univariate cross entropy measurements, each of which is suitably conditioned on the realisations of other variables in the system.

$$\begin{aligned} \lambda_c^i(\mathbf{x}_t \mid \Omega_t) &= \lambda_c^i\left(X_t^1 = x_t^1 \mid x_t^2, ..., x_t^V, \Omega_t\right) + \lambda_c^i\left(X_t^2 = x_t^2 \mid x_t^3, ..., x_t^V, \Omega_t\right) + ... \\ &+ \lambda_c^i\left(X_t^V = x_t^V \mid \Omega_t\right) \end{aligned} \tag{32}$$

This provides two related advantages. First of all, by scoring each variable separately the number of trees required for conditioning on observation bits via Φ_t is dramatically reduced. Rather than requiring $2^{R^{mv}} - 1$ trees for the whole set of V variables, one only requires $2^{R^v} - 1$ trees for any given variable v. The second advantage is that by expressing (32) as a sequence of univariate MIC measurements, the desirable properties outlined in section 2 can be preserved. However, the decomposition (32) comes at the cost of having to perform V distinct runs of the algorithm, as well as conditioning on the realisation

of contemporaneous variables $\mathbf{x}_t$ in addition to the information set Ω_t, which already contains the past L realisations of these variables.

The second element of the extension strategy addresses the increased memory requirement per tree which results from the larger multivariate context ω_t produced by the hashing function (11) from the information set Ω_t. Hardware constraints will impose in practice a de-facto cap $\tilde{D}$ on the depth of the context tree, which will be smaller than the total number of context bits in Ω_t, i.e. $\tilde{D} < R^{mv}L$.[14] Let $\tilde{\omega}_t$ represent the partial context string obtained by truncating the full context ω_t at depth $\tilde{D}$. Here $\tilde{\omega}_t$ represents the maximum amount of conditioning information that one can manage from a computational point of view. It is straightforward to show that using the truncated context $\tilde{\omega}_t$ instead of the full context ω_t in a cross entropy measurement (2) is equivalent to the introduction of a measurement error in the model probabilities, as the argument of the logarithm in the cross entropy term can be expanded to recover and isolate the model probabilities conditioned on ω_t:

$$\begin{aligned} E_P\left[\ln \frac{1}{\hat{p}^i\left(X_{r,t} = x_{r,t} \mid \tilde{\omega}_{d^*,t}, \phi_{r,t}\right)}\right] &= E_P\left[\ln \frac{1}{\hat{p}^i\left(X_{r,t} = x_{r,t} \mid \omega_{d^*,t}, \phi_{r,t}\right)}\right] \\ &+ E_P\left[\ln \frac{\hat{p}^i\left(X_{r,t} = x_{r,t} \mid \omega_{d^*,t}, \phi_{r,t}\right)}{\hat{p}^i\left(X_{r,t} = x_{r,t} \mid \tilde{\omega}_{d^*,t}, \phi_{r,t}\right)}\right] \end{aligned} \tag{33}$$

The second term in the expansion can be expressed as a function of a measurement error $\tilde{\nu}^i_{r,t}$, which captures the percentage difference between the estimated model probabilities conditioned on $\tilde{\omega}_t$ and ω_t.

$$E_P\left[\ln \frac{1}{\hat{p}^i\left(X_{r,t} = x_{r,t} \mid \tilde{\omega}_{d^*,t}, \phi_{r,t}\right)}\right] = E_P\left[\ln \frac{1}{\hat{p}^i\left(X_{r,t} = x_{r,t} \mid \omega_{d^*,t}, \phi_{r,t}\right)}\right] + E_P\left[\ln\left(1 + \tilde{\nu}^i_{r,t}\right)\right] \tag{34}$$

This is essentially the same problem as was discussed in section 2.1, where the presence of measurement error on the model probabilities combined with Jensen's inequality will bias the resulting cross entropy. The complication is that there are now two sources of measurement error, the first being the fact real-valued probabilities are approximated by frequencies, the second being the effect of using conditioning on truncated information. Unfortunately, while the MIC can deal with the former via the bias correction term (25), the properties of the latter are not known, potentially re-introducing the problem of bias in comparison methods for simulated models.

Two properties of the context Ω_t and CTW algorithm can be used to minimise the effect of the conditioning error $\tilde{\eta}_{r,t}$. First, note that if $d^* \leq \tilde{D}$, then it must be that $\tilde{\eta}_{r,t} = 0$, as in this case the two partial context strings $\tilde{\omega}_{d^*,t}$ and $\omega_{d^*,t}$ are identical. Second, the binary discretisation of the context Ω_t is likely to possess a sparse information structure, and not all the bits of the resulting context ω_t string will be equally informative for conditioning the transition probabilities. As is visible from figure 1(a), when attempting to predict the value of X_t from knowledge of x_{t-1}, the MSB of x_{t-1}, which determines if the observation is in the top/bottom half of the range of variation, will be the most informative. Conversely, because the LSB encodes the smallest range of variation measured by the discretisation, it will not provide as much conditioning information when predicting the value of the variable of interest. In addition to this, current and lagged observations from the different variables in the system will not be equally informative in predicting the value of X_t.

Intuitively, these two properties suggest that way to the measurement error $\tilde{\nu}^i_{r,t}$ induced by truncating ω_t at a depth $\tilde{D}$ is to ensure that the most informative bits of context are placed at a depth $d \leq \tilde{D}$, thus only truncating out the bits that contribute the least information to the context. In practice, this

[14] Because the constraint is the available amount of working memory, this cap is hardware-dependant, and one might expect the constraint to relax somewhat with improvements in computers. The applications presented here use $\tilde{D} = 28$.

can be done through the use of a suitable hashing function (11). As mentioned in section 2.2, for the univariate case the hashing function $h_c(\Omega_t)$ is trivial, as it simply concatenates the binary representation of the observations in Ω_t. For the multivariate case, one needs to find a mapping that ensures that the truncated context string $\tilde{\omega}_t$ preserves as much of the useful information in ω_t as possible.

First, during the discretisation process the binary representation of each variable is divided into an informative and an uninformative section, by identifying the resolution at which correlation between the real-valued variable and its discretised version passes 0.95. For example, supposing a variable is discretised to a 7-bit resolution, only the first 3 bits would be treated as informative. Next, before each univariate MIC measurement of a variable x_t^v lagged variables in the information set and the contemporaneous variables required by the use of the (32), are ranked in order of decreasing correlation with x_t^v. The context string ω_t is obtained by first concatenating the informative bits of the variables according to their correlation rank, followed by the uninformative bits, again following the correlation rank. The result is that the most informative bits of the most informative conditioning variable are processed closest to the root of the context trees. The bits in ω_t truncated by the computational resource constraint $\tilde{D}$ are therefore the least informative ones.

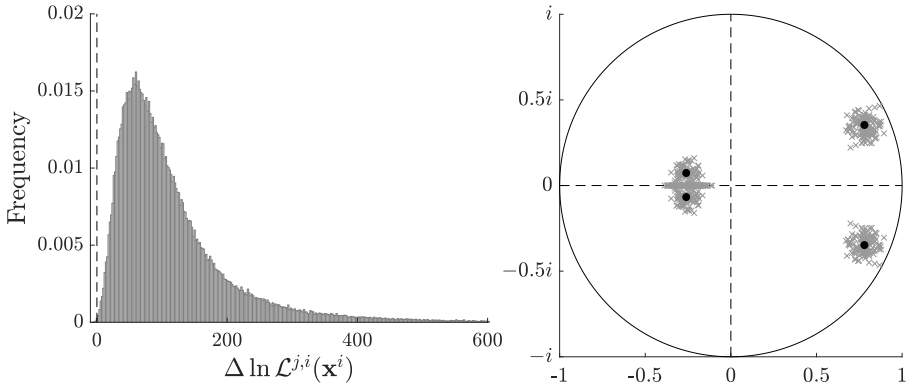
Despite the use of both strategies described above (a chain rule decomposition and the permutation of context bits), it is likely that the accuracy of the MIC will suffer from the large increase in the state space resulting from the shift to a multivariate setting. The final component of the extension strategy is to increase the accuracy of the measurement using a super-resolution approach, which enables multiple measurements with no additional empirical or simulated data. This relies on the fact that the entropy chain rule decomposition (32) is theoretically independent of the conditioning order used in the decomposition. Let a conditioning order o be a permutation of the sequence $\{1, 2, ..., V\}$, and let $\lambda^{i,o}(\mathbf{x})$ be the measurement obtained for model i by conditioning according to o as follows:

$$\begin{aligned} \lambda_c^{i,o}(\mathbf{x}_t \mid \Omega_t) &= \lambda_c^i\left(X_t^{o_1} = x_t^{o_1} \mid x_t^{o_2}, ..., x_t^{o_V}, \Omega_t\right) + \lambda_c^i\left(X_t^{o_2} = x_t^{o_2} \mid x_t^{o_2}, ..., x_t^{o_V}, \Omega_t\right) + ... \\ &+ \lambda_c^i\left(X_t^{o_V} = x_t^{o_V} \mid \Omega_t\right) \end{aligned} \tag{35}$$

In the absence of measurement error, the empirical cross entropy of the multivariate system should be the same regardless of the order in which the chain rule decomposition is carried out. It should therefore be that $\lambda_c^{i,o}(\mathbf{x}_t \mid \Omega_t) = \lambda_c^{i,o'}(\mathbf{x}_t \mid \Omega_t)$ for any two orderings o and o' of the V variables in the system. In practice, however, this is not going to be the case, due to various measurement errors, including (33) induced by truncating the context at depth $\tilde{D}$. It is possible, however, to increase the accuracy of the measurement by averaging multiple MIC measurements (35) obtained on the same empirical and simulated datasets by varying only the conditioning order o. Again the cost of doing so is an increase in the number of times the algorithm has to be run to obtain a final measurement.

4. Validation exercises for the multivariate MIC

Two validation exercises are carried out to evaluate the effectiveness of the extension strategy outlined above. Both follow the general approach of the univariate example in 2.4, using the NOLH design to generate a set of similar parameterisations and taking advantage of the availability of a benchmark measure to test the effectiveness of the multivariate MIC. The first is a bivariate VAR, which forms the simplest possible extension to the univariate setting. The second, more challenging test, applies the MIC to a the Smets and Wouters (2007) DSGE framework.



(a) Log-likelihood deviations (b) VAR eigenvalues

Figure 3: Dynamic properties of the bi-variate VAR(2) model set

4.1. MIC vs. Bivariate VAR likelihood

The bi-variate VAR provides the smallest possible extension of the univariate setting. This follows the strategy used in Barde (2017) for the original validation of the univariate MIC: if the proposed methodology fails even in the simplest and most favourable setting, then the approach is immediately falsified. The DGP is given by the following bi-variate VAR(2):

$$\mathbf{x}_t = \mathrm{A}_1 \mathbf{x}_{t-1} + \mathrm{A}_2 \mathbf{x}_{t-2} + \boldsymbol{\varepsilon}_t \qquad \boldsymbol{\varepsilon}_t \sim N(0, \Sigma) \tag{36}$$

With the following central parametrisation $\Theta = \{A_1, A_2, \Sigma\}$:

$$\mathrm{A}_1 = \begin{bmatrix} 0.20 & 0.15 \\ -0.15 & 0.15 \end{bmatrix} \qquad \mathrm{A}_2 = \begin{bmatrix} 0.50 & 0.30 \\ -0.25 & 0.55 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \tag{37}$$

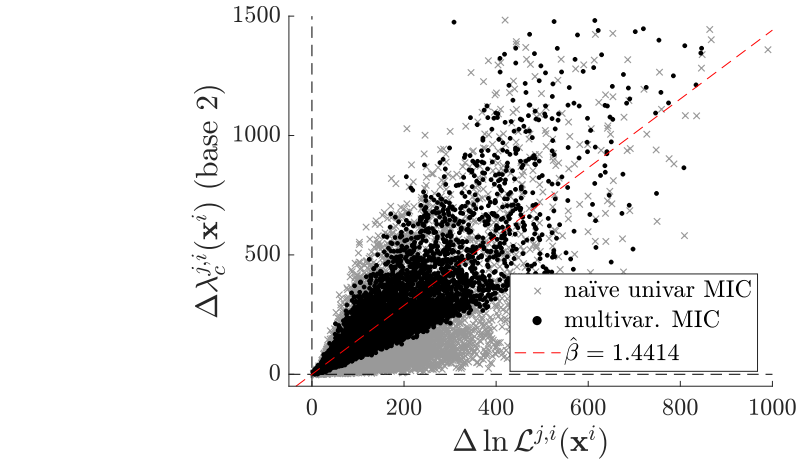
We duplicate the setting used in section 2.4, and vary these central parameters using a 129 by 9 NOLH design. This allows all the parameters in matrices A_1 and A_2, as well as the off-diagonal component of Σ to be modified by a uniform $[-0.1, 0.1]$ shock. Given this, the conditional likelihood of a given parameterisation Θ^j given a realisation $\mathbf{x}^i$ is given by:

$$\begin{cases} \ln \mathcal{L}\left(\Theta^j \mid \mathbf{x}^i\right) = -(T-2)\ln(2\pi) - \dfrac{T-2}{2}\ln|\Sigma| - \dfrac{1}{2}\sum_{t=3}^{T} \mathrm{u}_t' \Sigma^{-1} \mathrm{u}_t \\ \mathbf{u}_t = \mathbf{x}_t^i - \mathrm{A}_1^j \mathbf{x}_{t-1}^i - \mathrm{A}_2^j \mathbf{x}_{t-2}^i \end{cases} \tag{38}$$

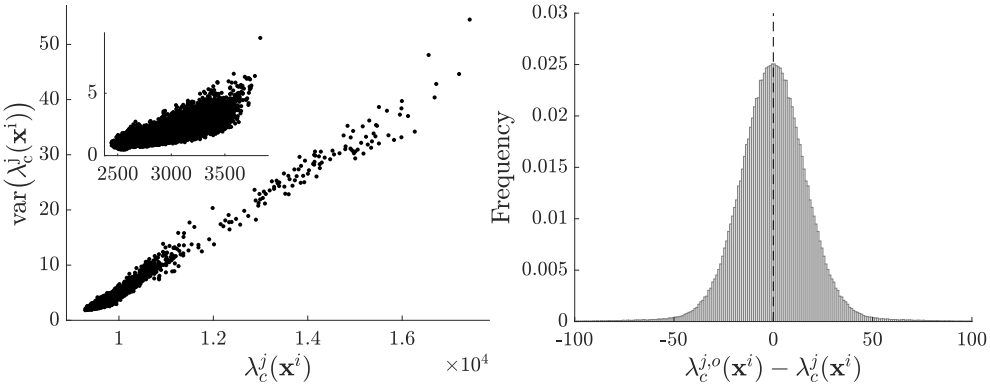
Following (30), these conditional likelihoods are averaged and normalised by the conditional likelihood for the 'true' parameter set Θ^i in order to obtain the relative likelihood $\Delta \ln \mathcal{L}^{j,i}(\mathbf{x}^i)$. Figure 3 confirms that the 129 models generated through the NOLH shocks are again very similar: the mass of the pairwise likelihood deviations from the true DGP resembles the one in figure 2(a) for the univariate setting, and the tight distribution of the eigenvalues visible in figure 3(b) suggests that the models are dynamically very similar. Figure 3(b) also confirms that all 129 models are stable, as their eigenvalues all lie within the unit circle.

Each of the 129 models is used to produce 510 simulated time series of 1000 observations each. The MIC training set consists of 500 of these series, while the remaining 10 series form the 'empirical' data. Both variables are discretised using $r = 7$ bits of resolution over the $[-10, 10]$ interval.[15] The MIC algorithm uses $L = 2$ lags of memory, the depth cap $\tilde{D}$ for the context trees is set at 24 bits and the

[15] See appendix B for discretisation diagnostics.



(a) MIC vs. log-likelihood



(b) MIC heteroscedasticity (univariate inset) (c) Distribution of conditioning order errors

Figure 4: Stability properties of the bi-variate VAR(2)

measurement is obtained by averaging the following two measurements, obtained by taking advantage of the chain rule decomposition (35) for a two-variable system.

$$\begin{cases} \lambda_c^{j,o_1}\left(\mathbf{x}_t^i \mid \Omega_t^i\right) &= \lambda_c^j\left(X_t^{i,1} = x_t^{i,1} \mid x_t^{i,2}, \Omega_t^i\right) + \lambda_c^j\left(X_t^{i,2} = x_t^{i,2} \mid \Omega_t^i\right) \\ \lambda_c^{j,o_2}\left(\mathbf{x}_t^i \mid \Omega_t^i\right) &= \lambda_c^j\left(X_t^{i,2} = x_t^{i,2} \mid x_t^{i,1}, \Omega_t^i\right) + \lambda_c^j\left(X_t^{i,1} = x_t^{i,1} \mid \Omega_t^i\right) \end{cases} \tag{39}$$

We also perform a 'naïve' MIC measurement by taking the sum of two univariate MIC measurements made on each of the variables. This provides a counterfactual by which we can judge the effectiveness of the computational strategy outlined in section 3, while also illustrating the risk involved in ignoring conditioning error in the CTW probabilities when moving to a multivariate setting. Figure 4(a) presents the scatter plot of the resulting $\Delta\lambda_c^{j,i}(\mathbf{x}^i)$ measurements against the relative likelihoods $\Delta \ln \mathcal{L}^{j,i}(\mathbf{x}^i)$ obtained analytically from (38), and suggests that differences in the multivariate MIC across models again consistently tracks the relative likelihood for the same models. As was the case in figure 2(b), the slope of the regression line is approximately $1/\ln 2$. The figure also confirms that univariate MIC performs poorly in this setting, illustrating the importance of taking conditioning error seriously.

One apparent difference between the multivariate performance in figure 4(a) and the univariate version in figure 2(b) is that the multivariate MIC displays heteroscedasticity, with the noise in the measured MIC deviation increasing with the size of the deviation itself. Some heteroscedasticity is to be expected from the fact that the deviations are the difference between two sums of random variables (27) and their variance is therefore given by $\mathrm{var}\left(\Delta\lambda_c^{j,i}(\mathbf{x}^i)\right) = \mathrm{var}\left(\lambda_c^j(\mathbf{x}^i)\right) + \mathrm{var}\left(\lambda_c^i(\mathbf{x}^i)\right) - 2\mathrm{cov}\left(\lambda_c^j(\mathbf{x}^i), \lambda_c^i(\mathbf{x}^i)\right)$. Even if the variance of a measurement $\mathrm{var}\left(\lambda_c^j(\mathbf{x}^i)\right)$ is constant across pairwise comparisons, the covariance will

vary: models that are very close will be highly correlated, which is not the case for models that are more distant. Figure 4(b) reveals however that in addition to this composition effect, the variance raw measurement $\mathrm{var}\left(\lambda_c^j(\mathbf{x}^i)\right)$ is indeed increasing with the size of the measurement. Importantly the figure inset in 4(b) reveals that this is also the case for the univariate MIC, the main difference being that the range of variation of the MIC in the univariate case is much smaller, leading to a smaller variance between models.

The final test of the computational strategy is the histogram in figure 2(c) of the difference between measurements made using the two available orderings to the cross entropy chain rule (39). This is obtained by taking the difference between the values of the individual measurements $\lambda_c^{j,o_1}(\mathbf{x}_t^i)$ and $\lambda_c^{j,o_2}(\mathbf{x}^i)$ from $\lambda_c^j(\mathbf{x}^i)$ for all 165120 pairwise comparisons. The zero mode of the resulting distribution confirms that the cross entropy measurement does not on average depend on the decomposition ordering used in the chain rule. It also confirms that MIC measurements obtained using a single ordering o will contain an error, and that the accuracy of the MIC can be improved by averaging over multiple decomposition orderings.

4.2. MIC vs. Smets and Wouters (2007) marginal densities

Having validated the strategy in a relatively benign setting, we now provide a tougher and more realistic test using the Smets and Wouters (2007) (SW) model, taking advantage of the fact that the ranking of estimated DGSE models can be established using the marginal density. The two factors that make the test tougher are that 7 observable variables are included, greatly increasing the state space of the Markov processes and that the amount of empirical data is much lower that in the univariate ARMA and VAR settings. The comparison framework uses on the standard version of the SW model as well as two additional versions created by imposing parameter restrictions. By fixing the values of parameters that are estimated in the basic model, the aim is to create clearly inferior versions of the model from the point of view of empirical fit. The first set of parameter restrictions involves the utility households derive from consumption, which is embedded in the objective function below. As explained in Smets and Wouters (2007, p 589), setting the habit formation parameter $\lambda = 0$ and the CRRA parameter $\sigma_c = 1$ leads to a strictly forward-looking consumption. The aim here is to remove the smoothing provided by habits, which in practice makes a significant to the goodness of fit.

$$E_t\left[\sum\nolimits_{s=0}^{\infty}\beta^s\left[\frac{\left(C_{t+s}-\lambda C_{t+s-1}\right)^{1-\sigma_c}}{1-\sigma_c}\right]\exp\left(\frac{\sigma_c-1}{1+\sigma_l}(L_{t+s})^{1+\sigma_l}\right)\right] \tag{40}$$

In addition to the restrictions on the utility from consumption, the third version of the model also restricts the following shock processes for spending (ε_t^g), prices (ε_t^p) and wages (ε_t^w). The wage and price shocks are ARMA(1,1) processes based on i.i.d normal innovations η_t^p and η_t^p, while the spending shock includes a passthrough from the innovation in productivity η_t^a which is controlled via ρ_{ga}. The restriction involves setting $\rho_{ga} = \mu_p = \mu_w = 0$, thereby turning all three shocks into AR(1) processes. As for the consumption restrictions, the aim is to target a feature of the model which improves the empirical fit of the model by smoothing capturing the high-frequency component in wages and prices.

$$\begin{cases} \varepsilon_t^g = \rho_g\varepsilon_{t-1}^g + \eta_t^g + \rho_{ga}\eta_t^a \\ \varepsilon_t^p = \rho_p\varepsilon_{t-1}^p + \eta_t^p - \mu_p\eta_{t-1}^p \\ \varepsilon_t^w = \rho_w\varepsilon_{t-1}^w + \eta_t^w - \mu_w\eta_{t-1}^w \end{cases} \tag{41}$$

In addition to this, four of the five calibrated parameters are shocked around their original values using a 65×4 NOLH design, in line with the previous tests. These are the depreciation rate δ, the labour market mark-up rate λ_w and the two curvature parameters for the Kimball (1995) aggregator used in the goods and labour markets, ε_p and ε_w. The fifth calibrated parameter of the SW model, the government spending to GDP ratio g_y, is left unchanged at 0.18 as it is determined from the observed empirical ratio.

Table 2: NOLH parameters for the SW models

	δ	λ_w	ε_p	ε_w
Central value	0.025	1.5	10	10
NOLH shock range	± 0.015	± 0.4	± 5	± 5

Table 2 shows the variation range applied to each of these parameters, as well as the fact that the central values are unchanged from the original setting of Smets and Wouters (2007).

With 65 different calibrations applied to the three versions of the SW model, 195 distinct specifications are available for this comparison exercise. These are estimated on the original Smets and Wouters (2007) dataset, made up of 7 US macroeconomic observables over 160 quarters, from 1965:Q1 to 2004:Q4. The parameter estimates obtained for the three central versions of the SW model are provided in appendix C.[16] Figure 5(b), which shows the histogram of marginal densities obtained for the 195 specifications, confirms that the specifications are clustered in three distinct peaks. Furthermore, it confirms that the loss of flexibility associated with the parameter restrictions significantly worsens the fit of the models. The test is now to check whether training the MIC on simulated data from these 195 specifications and scoring the same 7 US variables results in similar model rankings.

As for the previous exercises, the MIC algorithm is trained using 500 simulated series of 1000 observations each. As is shown in appendix B, all seven variables are discretised to 6 bits of resolution, a single lag of memory is used ($L = 1$) and we set the depth cap $\tilde{D} = 28$. With 7 variables, $7! = 5040$ potential orderings are available for the chain rule decomposition, which is more than is practical. Instead, we average measurements obtained using 21 orderings generated from cyclic permutations of $\{1, 2, 3, 4, 5, 6, 7\}$, $\{7, 6, 5, 4, 3, 2, 1\}$ and $\{1, 7, 2, 6, 3, 5, 4\}$.[17] Finally, in order to verify the robustness of the results obtained, we ran the MIC a second time using a very coarse discretisation of the variables (3 bits) and $L = 3$ lags of memory. The results obtained with this alternate setting are provided in appendix D and very much in line with the those obtained with the 6-bit, 1 lag settings. This also seems to support the findings of Lamperti (2018b), who show that entropic measurements of real-valued series are robust to coarse discretisations.

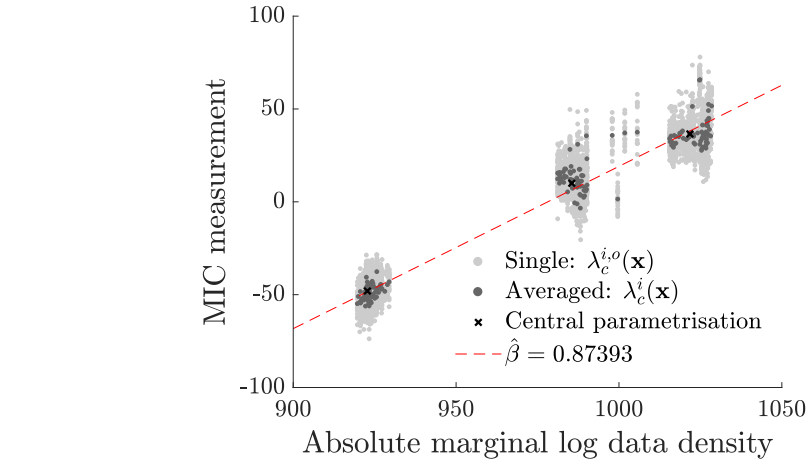
Figure 5(a) provides a scatter plot of the MIC scores obtained for the 195 calibrations against the absolute marginal log densities in figure 5(b), confirming that the MIC can replicate the overall ranking of the three versions of the SW model, despite the larger state space and the much small number of empirical observations than in the VAR example. The figure also reveals that large part of this performance comes from averaging the 21 chain rule decompositions, as the dispersion of these individual measurements is large. It is important to note that in order to be able to meaningfully superpose the scatter plots of the individual orderings, the MIC scores of the 195 models obtained for each ordering are centred on the mean.[18] This enables the calculation of the difference between the measurements obtained using the 21 orderings and their average, $\lambda_c^{i;o}(\mathbf{x}) - \lambda_c^i(\mathbf{x})$. The histogram of these 4095 differences, provided in figure 5(b), reveals a normal distribution, further supporting the strategy of averaging multiple MIC measurements to improve precision.

The top half of table 3 presents the MIC scores obtained for the central parametrisations of the 3 SW models in the comparison exercise. The use of the chain rule (35) means that MIC measurements

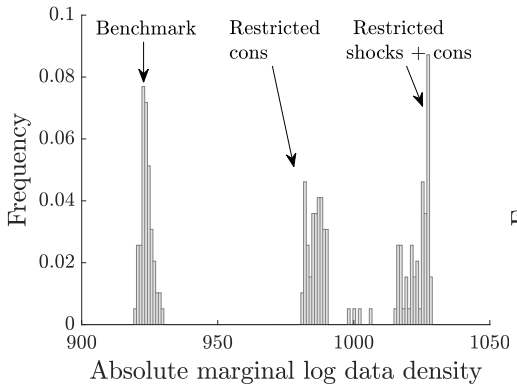
[16] These specifications of the SW model are estimated and simulated using Dynare, based on modified versions of the code originally prepared by Jerome Williams and available through `https://github.com/jeromematthewcelestine/`

[17] The use of cyclic permutations ensures that the 7 measurements obtained from each decomposition ordering are distinct from each other, as there is no duplication of portions of the entropy decomposition (35).

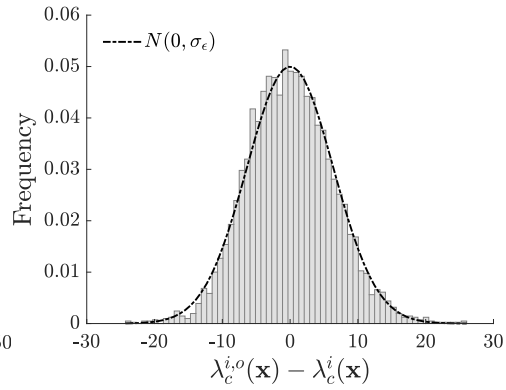
[18] Because this ends up shifting all the 195 model MICs by a constant (which is the average MIC over all 195 models and 21 measurements), this does not affect the relative rankings of the models.



(a) MIC vs. log-likelihood



(b) Distribution of SW models by marginal density



(c) Measurement error

Figure 5: MIC results on SW models, high resolution, 1 lag

are available both at the level of individual variables, as well as on aggregate, although one should note the aggregate MIC is not the sum of the variable-level MIC because of the mutual information between variables. As the MIC (27) is calculated at the level of single observations, it is possible use the MCS procedure to test whether a given model significantly underperforms relative to others. As expected, the two restricted models are significantly excluded from the confidence set on aggregate. Regarding individual variables, it is interesting to note that the impaired performance for the restricted models seems to come from the real variables (Δy, Δc, Δi and to a lesser extend L), while the fit of the nominal variables (r,π and Δw) does not seem affected as badly by the parameter restrictions.

A few additional observations can be made. First, because the original purpose of the CTW/CTM algorithms is data compression, the MIC measurement has a natural interpretation as the number of bits that the discretised data can be compressed to using the model probabilities. Dividing the MIC by the number of bits in the original data provides a compression ratio which can help judge model performance. The compression ratios for the various SW models are provided in the bottom half of table 3, and they reveal that all three models tend to do better on labour hours and the policy rate than it does on output, wage and consumption changes. One should not directly conclude from this that the model is better at predicting the former variables than the latter ones: the better compression rate on L and r can be explained by the fact that these two variables have lower variability and are therefore inherently more predictable.

The compression ratio is particularly useful as it allows the MIC to address the criticism of Marks (2013) and Fagiolo et al. (2007) mentioned in section 2.1. This is because a compression ratio of 1, i.e. the original data is not compressed at all, is what results from attempting to compress data using

Table 3: Variable-level and aggregate MIC for central SW models, high resolution, 1 lag

	L	r	π	Δy	Δc	Δi	Δw	Aggr.
MIC measurements								
Benchmark	600.69	649.61	709.37	810.60	779.33	742.52	861.89	5107.39
	(0.000)	(1.795)	(0.000)	(0.000)	(0.000)	(0.000)	(0.445)	(0.000)
Restrict 1	604.20	634.31	715.33	835.93***	814.50***	754.49**	860.71	5165.42***
	(0.621)	(0.000)	(0.936)	(4.805)	(3.923)	(2.423)	(0.225)	(4.192)
Restrict 2	617.68**	648.59	717.42	828.42***	830.90***	754.15**	859.99	5191.98***
	(2.378)	(1.489)	(1.052)	(4.356)	(9.506)	(2.585)	(0.000)	(4.707)
Compression ratios								
Benchmark	0.626	0.677	0.739	0.844	0.812	0.773	0.898	0.760
Restrict 1	0.629	0.661	0.745	0.871	0.848	0.786	0.897	0.769
Restrict 2	0.643	0.676	0.747	0.863	0.866	0.786	0.896	0.773

- Note: 'Restrict 1' refers to the model with restricted consumption parameters, 'restrict 2' is the model with additional restrictions to the shock processes. MCS t-statistics are provided in parenthesis, with superscripts '*', '**' and '***' indicating that the model is excluded from the confidence set at the 10%, 5% and 1% significance level, based on bootstrapped standard errors.

uniform model probabilities. This provides an anchor point for interpreting the absolute value of the MIC measurement obtained on a model. Following Fagiolo et al. (2007), supposing all the candidate models in the comparison set are flawed, the KL divergence will still aim to identify the best. However, the large compression ratios will unambiguously signal that all the candidate model are worse than an uninformative model. This is not the case here for the SW comparison, but this will be relevant in the following section.

Secondly, it is important to state that the measurements obtained are noisy, as was the case in all previous illustrations. While the MIC can correctly identify which of the 3 versions of the SW model performs best, figure 5(a) suggests that it is difficult to distinguish amongst the local variants generated by the 65 NOLH parameterisations. Running the MCS on all 195 versions of the model confirms that this is the case, as it results in a confidence set of 63 models, which are all NOLH variants of the benchmark SW model. In other words, it is difficult to distinguish models that are intrinsically similar. This is consistent with the performance of the univariate MIC reported in Barde (2017), and underlines the important of testing the statistical significance of any model comparison carried out with the MIC.

5. A macroeconomic agent-based model comparison exercise

As explained in in the introduction, a range of ABM frameworks have been developed with a view to analysing the effect of macroeconomic and macro-prudential policy scenarios. We now show how the MIC can be used to compare such models against more traditional DSGE approaches, using the ABM proposed by Caiani et al. (2016). This choice is motivated by two considerations, first the model was explicitly designed with the aim of addressing some shortcomings of pre-crisis DSGE models, hence justifying a comparison with Smets and Wouters (2007). Second, because it aims to be a 'benchmark model', the source code for the model is publicly available, which greatly facilitates replication and testing of the framework by others in the field.[19]

Before presenting the features of the ABM and the results of the comparison against the SW model, it is important to flag an important issue. The ABM parameter values used in the comparison exercise are unchanged from their original values in Caiani et al. (2016), and the ABM will therefore not be optimised

[19] The model's code is based on the Java Macro Agent Based (JMAB) toolbox, and freely available from `https://github.com/S120/jmab`. The simulation and code for this exercise is available in the supplementary material.

for the datasets used in the comparison, unlike the SW model whose parameters are estimated from the data. While great progress has been made on the development of estimation methods for ABMs, the methods discussed in the introduction have not yet matured to the point where they can easily interface with the existing code base of ABM models. As a result, the comparison exercise can legitimately be seen as unfair, as one would expect the performance the ABM to suffer relative to SW. However, this does replicate the situation that researchers are currently confronted with in attempting to move from demonstration ABMs with ad-hoc parameter calibrations to a descriptive, policy models. We will show that even in such a situation the comparison exercise can be fruitful in identifying the features of the data where the ABM most needs improvement.

5.1. The Caiani et al. (2016) stock-flow consistent macroeconomic model

The agent-based, stock-flow-consistent framework introduced by Caiani et al. (2016) was designed to address several of the shortcomings of DSGE models identified in the aftermath of the 2008 crisis. As was identified in Del Negro and Schorfheide (2013); Del Negro et al. (2016), the forecast performance of the workhorse SW model over this period is relatively poor, and can be greatly improved by introducing information about financial frictions, such as interest rate spreads. While these financial extensions to DSGE models do improve their performance, Lindé et al. (2016) conclude that they may not do enough to allow effective investigations into the effect of unconventional monetary policy or macroprudential instruments, as these require being able to effectively model the interbank network or allow for heterogenous agents. The Caiani et al. (2016) model addresses these concerns by developing a fully-fledged stock-flow consistent banking sector, with endogenous money creation in the form of bank loans to other agents. A key feature is that credit risk is built into the model by assuming a 20 period horizon on loans, while bank liabilities are formed of short-term demand deposits. This, combined with the presence of a bank-firm credit network, discussed below, enables the framework to model banking balance sheet crises of the type encountered in 2008 as well as their contagion to the production sector.

The model contains the following types of agents. Households consume, sell labour to firms, pay taxes and have deposits with banks. They own the firms and banks and receive dividends. The model possesses consumption and capital goods firms in a vertical structure, with upstream capital goods firms producing using labour. Firms invest and finance production from retained profits and by borrowing from banks. Next, banks collect deposits from households and firms, purchase government bonds, and create loans to firms subject to a liquidity ratio. The model possesses a central bank, whose role is to provide advances to banks at a fixed rate of interest, to allow them to meet their liquidity ratio requirement. Finally, the model also includes a government, which employs some public workers, issues benefits to unemployed workers funded by taxes on households and firms, and the issuance of bonds. These agents interact on a set of markets which each generate a network of connections. The consumption goods market matches households and consumption good firms, while the capital goods market matches the consumption firms to capital firms. The labour market is wider, and allocates households to both types of firms, as well as the government. On the financial side of the economy, the credit market connects both types of firms to the banks, and banks collect deposits from households and firms on the deposit market.

Crucially, the model can be used to generate macroeconomic observables from simulated data which are comparable to the seven variables in Smets and Wouters (2007), and can thus form the basis of the comparison exercise. Four of them are directly comparable in their construction: π is the one period log difference in the price index of the consumption good. Δw is the one period log difference in real wages, calculated by deflating the average household nominal wage using the price index. Similarly, Δy is the one period log difference of real GDP and Δi is the one period log difference in real firm investment, both deflated from their nominal values. The remaining three variables are a bit more problematic. The consumption variable, Δc, is determined as an accounting identity from the difference between nominal

Table 4: Parameters used in sensitivity analysis

	Bench.	Low	High
Bank risk aversion:			
Towards consumption firms	3.92	1	8
Towards capital firms	21.51	5	40
Consumption firm investment function:			
Capacity utilization weight	0.02	0	0.04
Cash flow weight	0.01	0	0.04
Precautionary deposits	1	0.5	1.5

- Note: The bank risk aversion parameters with respect to both kinds of firms are varied jointly, which results in 8 alternative parameterisations.

output and investment. While this is appropriate given the closed nature of the ABM, it does highlight the absence of foreign trade in aggregate demand. Second, because households supply their labour inelastically in the ABM, they do not possess an intensive margin. As a result the labour variable L is simply the deviation of the employment rate from the sample mean and does not measure hours worked, as in Smets and Wouters (2007). Finally, the interest rate r is the variable that poses the most problems, as the ABM does not posses an effective counterpart to the empirical US federal funds rate. Banks fund their lending via the collection of deposits, on which they pay interest, and the central bank acts as a lender of last resort for the banking sector, which would suggest that the deposit rate is the natural counterpart we are looking for. However central bank advances are made to banks at a fixed, exogenous interest rate which forms an upper bound to the deposit rate, truncating the range of variation. The simulations reveal that this is indeed the case, as the deposit rate is essentially constant at the exogenous advance rate. This leads us to use the average interest rate on loans to firms, which is the other major interest rate in the model, as the observable rate r. As will be discussed below, this design problem leads to a poor fit on the data.

The training data for the MIC algorithm is generated by running 1000 simulations of the ABM for 800 periods. The first 300 periods of each run are treated as a burn-in period and thus discarded, which results in the same total number of training observations as for the SW case seen in section 4.2.[20] In addition to the main benchmark simulation, a small-scale sensitivity analysis is run along the lines of the one carried out in Caiani et al. (2016) as a robustness check for the comparison. Because the main focus of the paper is contagion of crises in the bank/firm credit network, their analysis varies the parameters related to lending decisions made by banks and investment decisions made by firms. For the former, the parameters that are varied are the risk aversion of lending to consumption and capital firms, while for the latter it is the parameters of the consumption firms' investment functions that are modified. The values values for these are provided in table 4, and correspond to the endpoints of the parameter grid used in Caiani et al. (2016). This restriction is due to the large simulation requirement involved in generating the training data.

5.2. Comparison results against Smets and Wouters (2007)

The comparison exercise is similar to the one used for the Smets and Wouters (2007) validation exercise, in particular the parameter settings for the MIC algorithm are kept the same. Two different datasets are used to compare the models. The first is the original Smets and Wouters (2007) 1965:Q1 - 2004:Q4 dataset, while the second is an extended version of the datset covering the 'crisis period' of

[20] The author is particularly grateful to Luca Fierro for his advice regarding the appropriate length of the burn-in period for the Caiani et al. (2016) model

Table 5: Macroeconomic model comparison, high resolution, 1 lag

	L	r	π	Δy	Δc	Δi	Δw	Aggr.
Smets & Wouters dataset (1965:Q1 - 2004:Q4)								
VAR(1)	590.37	638.03	724.10	818.34***	780.68	745.80	857.83	5115.72
	0.615	0.665	0.754	0.852	0.813	0.777	0.894	0.761
SW	600.69	649.61	709.37	810.60	779.33	742.52	861.89	5107.39
	0.626	0.677	0.739	0.844	0.812	0.773	0.898	0.760
Caiani et al.	894.51***	1715.08***	1817.16**	924.78***	971.27***	918.64***	1892.02***	9368.98***
	0.932	1.787	1.893	0.963	1.012	0.957	1.971	1.394
Crisis period (1997:Q1 - 2017:Q2)								
VAR(1)	267.94	164.48	306.11	369.95	356.09	346.64	446.47	2269.31
	0.551	0.338	0.630	0.761	0.733	0.713	0.919	0.667
SW	300.62***	234.81***	321.26*	381.68	369.23***	347.93	444.81	2396.26***
	0.619	0.483	0.661	0.785	0.760	0.716	0.915	0.704
Caiani et al.	511.67***	563.54**	407.00***	417.21***	457.10**	463.95***	1081.27***	4139.31***
	1.053	1.160	0.837	0.858	0.941	0.955	2.225	1.217

- Note: Because the MIC algorithm parameters are the same as in section 4.2, the scores obtained for the Smets & Wouters model over the original dataset are the same as in table 3.
- A MCS analysis was carried out but only the significance of the test is reported in order to save space. As before, superscripts '*', '**' and '***' indicating that the model is excluded from the confidence set at the 10%, 5% and 1% significance level, based on bootstrapped standard errors. Compression ratios are included below the score.

1997:Q1 - 2017:Q2. Because this second dataset contains the 1997 Asian financial crisis, the 2000 dot-com crash and 2008 great recession it should enable us to test the relative performance of the ABM and SW models in crisis periods. In addition to the Smets and Wouters (2007) and Caiani et al. (2016) models, we include a VAR(1) model estimated on the relevant empirical data set. Because the MIC can only assess the relative performance of models, this provides a benchmark by which to assess the performance of the DSGE model itself. As previously explained, the SW and VAR models are estimated on both datasets, while the ABM calibration is the same benchmark calibration as Caiani et al. (2016) in both cases.

The results of the main comparison are provided in table 5, with the top half of the table providing the scores for the original dataset. The immediate observation here is that the performance of the ABM model is poor across the board relative to the other two models, and it is clearly rejected from the confidence set. In addition, both the SW model and VAR(1) are included in the confidence set on aggregate, as well as for all individual variables except output, where the SW model seems to do slightly better. Given the noise present in the MIC measurement, this is consistent with the findings in table 2 of Smets and Wouters (2007), who also run several VAR specifications and find that the unconstrained VAR(1) is outperformed by their model. The poor aggregate performance of the ABM model relative to SW is expected, given the fact it is not calibrated on the data, therefore what is more interesting from the point of view of the modeler is the breakdown by variable. This reveals that the performance on the real variables (L, Δy, Δc and Δi), while poor, is not too far from the VAR and SW models, and is not unreasonable considering the lacking calibration. In particular, the compression ratio for these variables is below 1, with the exception of consumption, which is borderline, indicating that the model has explanatory power on the data even in its uncalibrated state. Conversely, the ABM clearly performs poorly on the nominal variables (π, r and Δw), with compression rates significantly above 1, the threshold which indicates that a particular model is outperformed by uninformative uniform transition probabilities. These findings carry across to the results of the sensitivity analysis, provided in table A-1 in appendix A: all calibrations possess similar aggregate scores and tend to display the same pattern on the real and nominal variables. The only significant deviation relates to the cash flow weight parameter, where a low value greatly worsens the performance, while a high value improves well it beyond that of the main calibration in table 5. In both cases, this is linked to the π and Δw and to a lesser extent L and r.

Before discussing the origin of this difference in performance across variables and its potential design implications for the Caiani et al. (2016) model, we briefly discuss the performance of the models on the crisis sample, the in the bottom half of table 5. A first observation is that the compression rates of all models improve, reflecting the fact that several variables exhibit lower variance over the period and become easier to predict, most notably r and π due to the lower zero bound. A second important aspect is the clear worsening of the SW model's performance relative to the VAR, to the point where the SW model is excluded from the confidence set on aggregate and also for most variables. This is in line with the findings of the literature mentioned in section 5.1 regarding the poor performance of DSGE models that do not incorporate financial frictions over the crisis period. While ABM still performs poorly, the 12.7% improvement in the aggregate compression ratio is in line with that of the VAR (12.4%), and much larger than the 7.4% improvement seen for SW. This improvement is mainly driven by large improvements on the inflation and interest rate variables, which provides support to the claim in Caiani et al. (2016) that their ABM is designed and calibrated towards replicating the crisis period. Interestingly, the labour market variables (wages and employment) see instead a slight degradation in performance compared to original dataset for the ABM. This is not the case for the VAR and SW models, who see slight improvements in compression ratios for these variables.

The MIC scores obtained by the ABM on the nominal variables (π, r and Δw) are poor enough that they can be explained by a failure of the ABM to match the unconditional distributions of those variables, let alone the conditional transition probabilities. In order to illustrate this, figure A-2 in appendix A provides the unconditional probability distributions of the observable variables for the ABM and Smets and Wouters models. These are obtained simply by taking the histograms of the respective training datasets over the discretisation range used by the MIC algorithm.[21] The plot confirm that the distributions of the policy rate and inflation in particular are very sharply peaked and inconsistent with the ones generated by SW on the original period. Similarly, the shifts in the SW distributions for these variables over the crisis period explain the great improvement in the ABM performance for this second sample. For the real variables (L, Δy, Δc, Δi), the ABM is much closer to the SW unconditional distributions.

Several observations relating to the design of the ABM can be made at this point. First, the combined lack of an active monetary policy and policy rate in the modelling of the central bank is probably a major factor in the inability of the model to replicate the interest rate and inflation data. Second, the poor performance on wages seem to stem mainly from the fact that changes in real wages are centred about zero. While the distribution is not fully symmetric, it does suggest that reductions in the real wages are relatively common in the ABM, which is not the case in the real data, where downward rigidities mean negative real wage growth is relatively rare. Finally, even though the fit for the real variables (Δy, Δc, Δi) is good considering the lack of calibration, the unconditional distributions are also centred close to zero, and are thus shifted to the left compared to the SW model. This can be linked to the fact that the ABM has a fixed number of workers and constant labour/capital productivities and therefore does not possess a trend growth component. This design implication has been addressed in Caiani et al. (2019) in the form R & D investment in by the capital goods firms which will improve the productivity of machines over time.

6. Conclusion

The central aim of this paper is to provide a proof of concept for a macroeconomic model comparison framework based on simulated data alone. In order to ensure accuracy in the comparison, the key difficulty

[21] The unconditional distributions for the VAR are not provided for reasons of clarity, as they overlap the Smets and Wouters distribution for most variables.

that needs to be overcome is the evaluation of the bias generated by errors in the conditional probabilities that are estimated directly from the simulated data. The algorithms underpinning the univariate version of the MIC in Barde (2017, 2016) are chosen specifically for their proven ability to minimise the bias incurred and calculate its expected value. The challenge in extending this to a multivariate setting is in managing the increased computational requirements without loosing these desirable theoretical properties. The first contribution we make is therefore to show that the MIC can indeed be extended to multivariate systems and successfully discriminate models on relatively short time series. Several validation exercises are carried out to establish that the extension strategy is effective. The multivariate extension of the MIC enables allows us to demonstrate how a VAR, an ABM and a DSGE model can be compared on the same empirical data on the basis of simulated data alone. The exercise is clearly able to rank the relative performance of these models, both at the aggregate level and on individual variables. This provides the proof of concept that forms the main contribution of the paper.

There are, however, several weaknesses or potential criticisms of this exercise that need to be discussed. First, as was previously highlighted, the ABM is not calibrated on the empirical data used in the comparison, which certainly puts it at an unfair disadvantage in a comparison with the SW model. While this is unfortunate, it does reflect the current state of play in the ACE field, which has developed both rich, policy-relevant large-scale models and advanced validation techniques, but has yet to properly connect the two. This highlights the need for further research and development of platforms where ABM development and validation can be better integrated. In addition, despite the clear disadvantage of the ABM, the comparison does provide useful insights and directions for developing the model to improve its empirical performance. For example, considering the lack of an effective calibration, the model performs well on the real variables, and it is instead the monetary policy variables that require reworking.

Conversely, one could argue that because the scores obtained on the ABM are much greater than those of the SW model, the comparison exercise is too easy, and as a result is not a good test of the MIC as a criterion for model selection. It is partly to address such concerns that two Monte Carlo validation exercises are run on the multivariate MIC prior to the comparison exercise itself. Both of these establish that the criterion can indeed distinguish between multivariate models, even with a relatively low number of observations. This is confirmed by the inclusion of a VAR(1) in the macroeconomic comparison, which correctly identifies the degradation in the performance of the SW model during the crisis period, in line with the standard finding of the literature. This effect is much more subtle that the large gap between the ABM and the SW model, and the MIC is able to detect it using only 80 empirical observations.

Finally a more general discussion relates to what variables are appropriate for the comparison of frameworks with such different ontologies as DGSE models and ABMs. To some extent this more general debate goes beyond the scope of this paper, however it can help illustrate the limits of the comparison exercise carried out here. ABM researchers would rightly point out that Caiani et al. (2016) model many more phenomena and variables than standard DGSE models. Indeed their paper not only displays the time evolution of aggregate variables, but also micro-stylised facts such as firm size or income distributions, connectivity in the various inter-agent networks, etc. As explained by Fagiolo and Roventini (2017), these micro-level variable, which are typical in many ABM frameworks, simply have no counterpart in many DSGE models, including Smets and Wouters (2007) due to simplifying assumptions. One might then legitimately argue that the ABM is the better model, as it encompasses more phenomena. As a practical example, suppose that empirical data on firm size or income distributions is included in the dataset used to compare the models in section 5.2. The Caiani et al. (2016) model would be able to produce simulated counterparts to this new empirical data, however the use of representative agents in Smets and Wouters (2007) would lead to degenerate distributions. Attempting to score these with the MIC would then lead to infinite scores for the SW model, which in some sense is entirely correct as it reflects the fact that one should reject a model which cannot account for the existence of an observed phenomenon. However, it is

also unhelpful, as any model able to produce a non-degenerate distribution on these observables would be preferable to the SW model, even if its performance on the macroeconomic variables is unacceptably poor. The premise used in this comparison exercise is therefore that models should be ranked on the basis of how well they explain the variables they have in common. Indeed, if a model aims to target more empirical variables than an alternative model, it can only be judged to encompass the alternative model if it offers at least equivalent performance on the variables they possess in common. In practice, however, it may well be that trade-offs exist between ABM and DSGE modelling approaches, much in the way Del Negro and Schorfheide (2006) show the DSGE as a restricted version of a VAR. We hope that the increased availability of reliable estimation and comparison tools will enable a better understanding of this important area of research.

References

Abrams, D. M., Strogatz, S. H., 2003. Linguistics: Modelling the dynamics of language death. Nature 424 (6951), 900.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC-19, 716–723.

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

Ashraf, Q., Gershman, B., Howitt, P., 2017. Banks, market organization, and macroeconomic performance: an agent-based computational analysis. Journal of Economic Behavior & Organization 135, 143–180.

Axtell, R., Farmer, D., Geanakoplos, J., Howitt, P., Carrella, E., Conlee, B., Palmer, N., 2014. An agent-based model of the housing market bubble in metropolitan washington dc. In: Whitepaper for Deutsche Bundesbank's Spring Conference on "Housing markets and the macroeconomy: Challenges for monetary policy and financial stability.".

Baptista, R., Farmer, J. D., Hinterschweiger, M., Low, K., Tang, D., Uluc, A., 2016. Macroprudential policy in an agent-based model of the uk housing market. Bank of England Staff Working papers619.

Barde, S., 2016. Direct comparison of agent-based models of herding in financial markets. Journal of Economic Dynamics and Control 73, 329–353.

Barde, S., 2017. A practical, accurate, information criterion for nth order markov processes. Computational Economics 50 (2), 281–324.

Basharin, G. P., 1959. On a statistical estimate for the entropy of a sequence of independent random variables. Theory of Probability & Its Applications 4 (3), 333–336.

Begleiter, R., El-Yaniv, R., Yona, G., 2004. On prediction using variable order markov models. Journal of Artificial Intelligence Research 22, 385–421.

Brock, W. A., Hommes, C. H., 1998. Heterogeneous beliefs and routes to chaos in a simple asset pricing model. Journal of Economic dynamics and Control 22 (8-9), 1235–1274.

Burnham, K. P., Anderson, D. R., 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media.

Caiani, A., Godin, A., Caverzasi, E., Gallegati, M., Kinsella, S., Stiglitz, J. E., 2016. Agent based-stock flow consistent macroeconomics: Towards a benchmark model. Journal of Economic Dynamics and Control 69, 375–408.

Caiani, A., Russo, A., Gallegati, M., 2019. Does inequality hamper innovation and growth? an ab-sfc analysis. Journal of Evolutionary Economics 29, 177–228.

Carlton, A., 1969. On the bias of information estimates. Psychological Bulletin 71 (2), 108 – 109.

Cioppa, T. M., 2002. Efficient nearly orthogonal and space-filling experimental designs for high-dimensional complex models. Doctoral Dissertation.

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

Dawid, H., Gemkow, S., Harting, P., van der Hoog, S., Neugart, M., 2018. Agent-Based Macroeconomic Modeling and Policy Analysis: The Eurace@Unibi Model. In: Chen, S.-H., M., K. (Eds.), The Oxford Handbook of Computational Economics and Finance. Oxford University Press.

Dawid, H., Harting, P., van der Hoog, S., Neugart, M., 2019. Macroeconomics with heterogeneous agent models: fostering transparency, reproducibility and replication. Journal of Evolutionary Economics 29 (1), 467–538.

Deissenberg, C., Van Der Hoog, S., Dawid, H., 2008. Eurace: A massively parallel agent-based model of the european economy. Applied Mathematics and Computation 204 (2), 541–552.

Del Negro, M., Hasegawa, R. B., Schorfheide, F., 2016. Dynamic prediction pools: An investigation of financial frictions and forecasting performance. Journal of Econometrics 192 (2), 391–405.

Del Negro, M., Schorfheide, F., 2006. How good is what you've got? dgse-var as a toolkit for evaluating dsge models. Economic Review-Federal Reserve Bank of Atlanta 91, 21–337.

Del Negro, M., Schorfheide, F., 2013. Dsge model-based forecasting. In: Handbook of economic forecasting. Vol. 2. Elsevier, pp. 57–140.

Dosi, G., Fagiolo, G., Napoletano, M., Roventini, A., 2013. Income distribution, credit and fiscal policies in an agent-based keynesian model. Journal of Economic Dynamics and Control 37 (8), 1598–1625.

Dosi, G., Fagiolo, G., Napoletano, M., Roventini, A., Treibich, T., 2015. Fiscal and monetary policies in complex evolving economies. Journal of Economic Dynamics and Control 52, 166–189.

Dosi, G., Fagiolo, G., Roventini, A., 2010. Schumpeter meeting keynes: A policy-friendly model of endogenous growth and business cycles. Journal of Economic Dynamics and Control 34 (9), 1748–1767.

Fagiolo, G., Guerini, M., Lamperti, F., Moneta, A., Roventini, A., 2019. Validation of agent-based models in economics and finance. In: Computer Simulation Validation. Springer, pp. 763–787.

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

Fagiolo, G., Roventini, A., 2012. Macroeconomic policy in dsge and agent-based models. Revue de l'OFCE (5), 67–116.

Fagiolo, G., Roventini, A., 2017. Macroeconomic policy in dsge and agent-based models redux: New developments and challenges ahead. Journal of Artificial Societies and Social Simulation 20 (1).

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

Gilli, M., Winker, P., 2003. A global optimization heuristic for estimating agent based models. Computational Statistics & Data Analysis 42 (3), 299–312.

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

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

Guerini, M., Moneta, A., 2017. A method for agent-based models validation. Journal of Economic Dynamics and Control 82, 125–141.

Haldane, A. G., Turrell, A. E., 2018. Drawing on different disciplines: macroeconomic agent-based models. Journal of Evolutionary Economics, 1–28.

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

Howitt, P., Clower, R., 2000. The emergence of economic organization. Journal of Economic Behavior & Organization 41 (1), 55–84.

Kimball, M. S., 1995. The quantitative analytics of the basic neomonetarist model. Journal of Money, Credit and Banking 27 (4), 1241–1277.

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

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

Kukacka, J., Barunik, J., 2017. Estimation of financial agent-based models with simulated maximum likelihood. Journal of Economic Dynamics and Control 85, 21–45.

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.

Lamperti, F., 2018a. Empirical validation of simulated models through the gsl-div: an illustrative application. Journal of Economic Interaction and Coordination 13 (1), 143–171.

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

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

Lindé, J., Smets, F., Wouters, R., 2016. Challenges for central banks' macro models. In: Handbook of macroeconomics. Vol. 2. Elsevier, pp. 2185–2262.

Lux, T., 2018. Estimation of agent-based models using sequential monte carlo methods. Journal of Economic Dynamics and Control 91, 391–408.

Marks, R. E., 2013. Validation and model selection: Three similarity measures compared. Complexity Economics 2 (1), 41–61.

Marks, R. E., 2019. Validation metrics: A case for pattern-based methods. In: Computer Simulation Validation. Springer, pp. 319–338.

Panzeri, S., Treves, A., 1996. Analytical estimates of limited sampling biases in different information measures. Network: Computation in Neural Systems 7, 87–107.

Popoyan, L., Napoletano, M., Roventini, A., 2017. Taming macroeconomic instability: Monetary and macro-prudential policy interactions in an agent-based model. Journal of Economic Behavior & Organization 134, 117–140.

Raberto, M., Ozel, B., Ponta, L., Teglio, A., Cincotti, S., 2018. From financial instability to green finance: the role of banking and credit market regulation in the eurace model. Journal of Evolutionary Economics, 1–37.

Rissanen, J., 1984. Universal coding, information, prediction and estimation. IEEE Transactions on Information Theory

IT-30, 629–636.

Rissanen, J., 1986. Complexity of strings in the class of markov sources. IEEE Transactions on Information Theory IT-32, 526–532.

Roulston, M. S., 1999. Estimating the errors on measured entropy and mutual information. Physica D 125, 285–294.

Salomon, D., Motta, G., 2010. Handbook of data compression. Springer Science & Business Media.

Schasfoort, J., Godin, A., Bezemer, D., Caiani, A., Kinsella, S., 2017. Monetary policy transmission in a macroeconomic agent-based model. Advances in Complex Systems 20 (8).

Schelling, T. C., 1971. Dynamic models of segregation. Journal of mathematical sociology 1 (2), 143–186.

Shannon, C. E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423.

Smets, F., Wouters, R., 2007. Shocks and frictions in us business cycles: A bayesian dsge approach. American economic review 97 (3), 586–606.

Sugiura, N., 1978. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. Communications in Statistics-Theory and Methods 7 (1), 13–26.

Takeuchi, K., 1976. The distribution of information statistics and the criterion of goodness of fit of models. Suri-Kagaku (Mathematical Science) 153, 12–18.

Teglio, A., Raberto, M., Cincotti, S., 2010. Balance sheet approach to agent-based computational economics: The eurace project. In: Combining Soft Computing and Statistical Methods in Data Analysis. Springer, pp. 603–610.

Willems, F. M. J., Shtarkov, Y. M., Tjalkens, T. J., 1995. The context-tree weighting method: Basic properties. IEEE Transactions on Information Theory IT-41, 653–664.

Willems, F. M. J., Tjalkens, T. J., 1997. Complexity reduction of the context-tree weighting algorithm: A study for kpn research. EIDMA ReportRS.97.01.

Willems, F. M. J., Tjalkens, T. J., Ignatenko, T., 2006. Context-tree weighting and maximizing: Processing betas. Proc. of the Inaugural Workshop of the ITA (Information Theory and its Applications).

A. Supplementary results

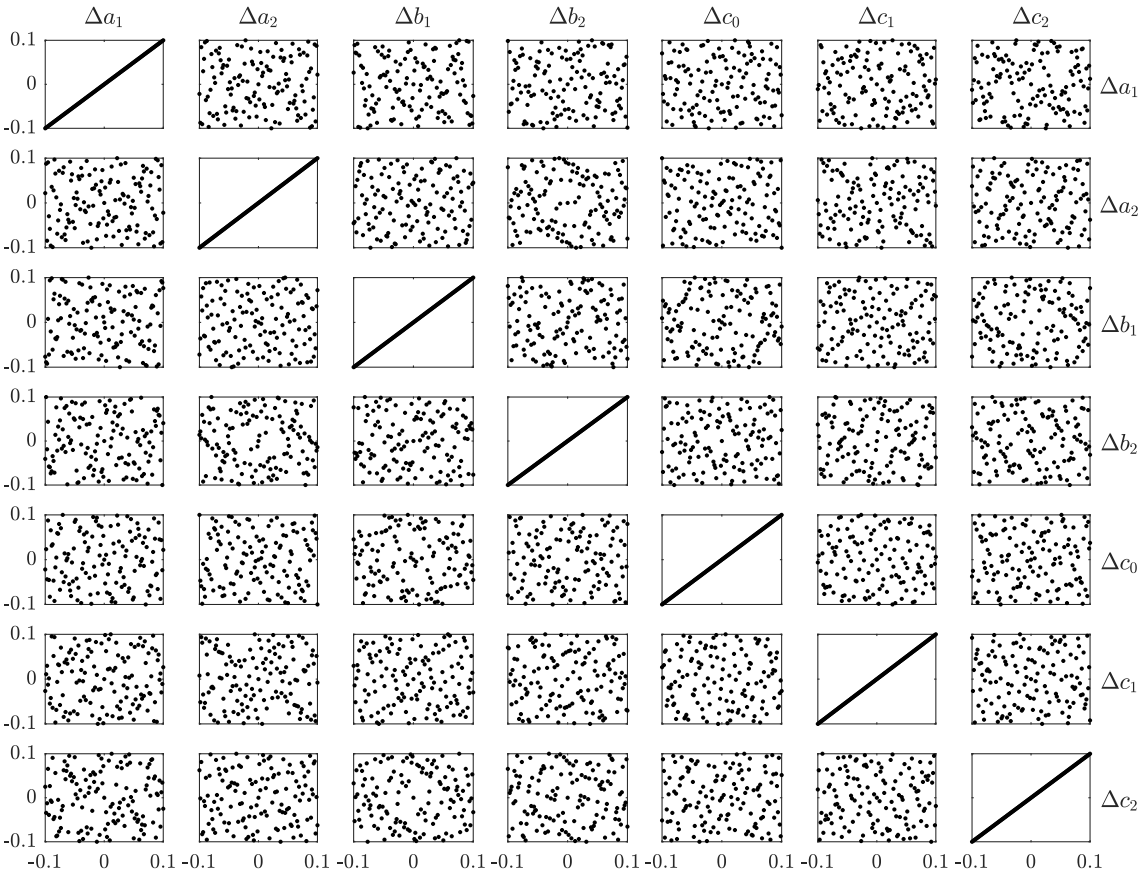
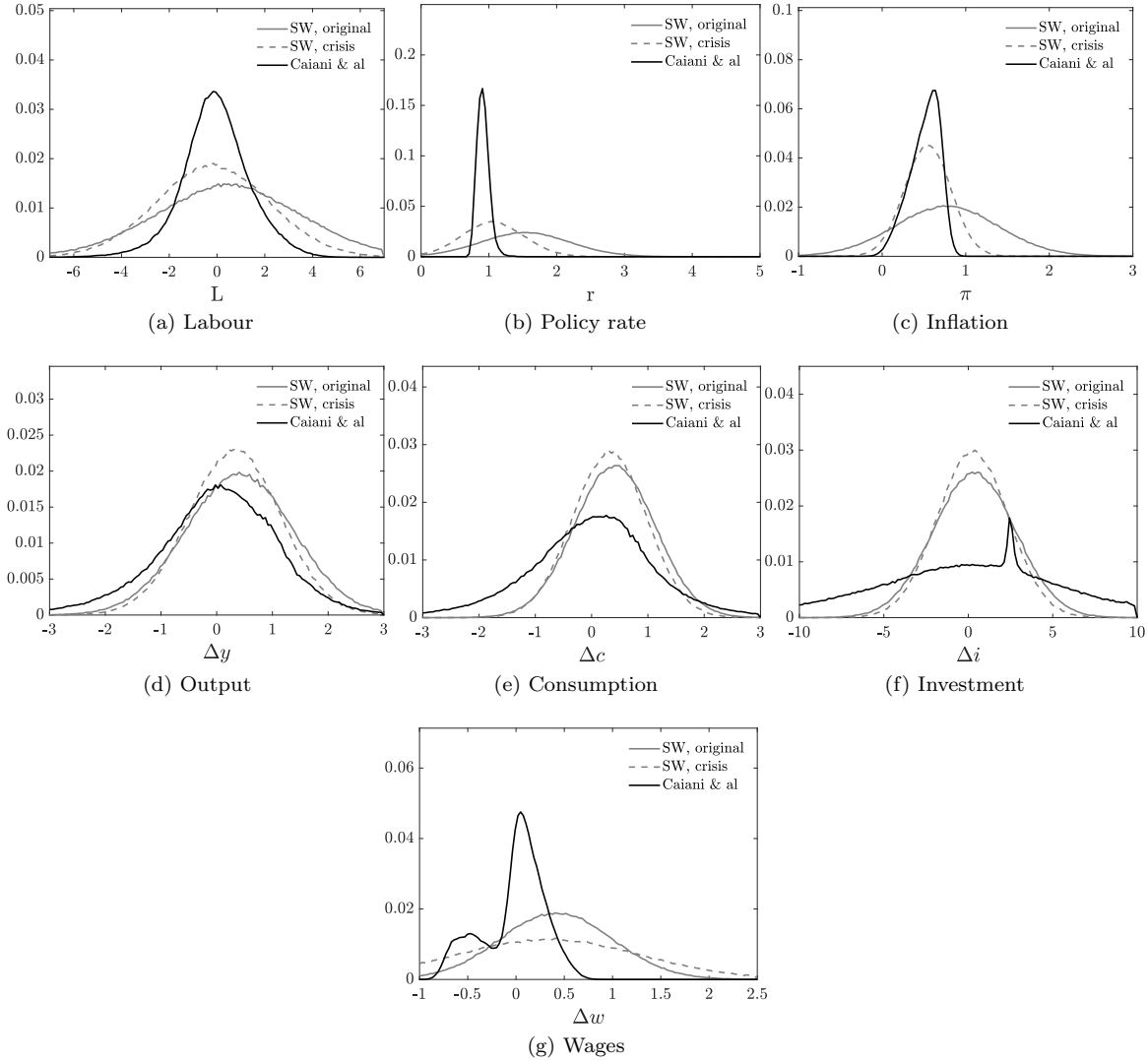


Figure A-1: Two-way scatter diagram of $\Delta\theta^i$ NOLH parameter shocks, 129 samples



(a) Labour (b) Policy rate (c) Inflation

(d) Output (e) Consumption (f) Investment

(g) Wages

Figure A-2: Unconditional probability distributions

Table A-1: Caiani et al. (2016) sensitivity analysis, high resolution, 1 lag

	L	r	π	Δy	Δc	Δi	Δw	Aggr.
Smets & Wouters dataset (1965:Q1 - 2004:Q4)								
Low bank risk av.	851.52	1704.36	1818.66	905.03	925.81	909.99	1885.79	9337.17
	0.887	1.775	1.894	0.943	0.964	0.948	1.964	1.389
High bank risk av.	865.52	1843.42	1633.00	905.93	932.65	930.71	1882.54	9325.12
	0.902	1.920	1.701	0.944	0.972	0.969	1.961	1.388
Low cap. util. weight	854.77	1767.54	1195.28	941.73	913.07	906.42	1703.99	8769.86
	0.890	1.841	1.245	0.981	0.951	0.944	1.775	1.305
High cap. util. weight	854.18	1809.27	1823.76	897.99	891.19	940.66	1916.59	9420.62
	0.890	1.885	1.900	0.935	0.928	0.980	1.996	1.402
Low prec. deposit	845.22	1860.19	1881.58	901.52	902.35	913.27	1790.61	9442.60
	0.880	1.938	1.960	0.939	0.940	0.951	1.865	1.405
High prec. deposit	867.68	1717.41	1811.28	881.09	913.23	897.18	1862.45	9271.72
	0.904	1.789	1.887	0.918	0.951	0.935	1.940	1.380
Low cash flow weight	2152.35	2912.44	2230.79	1075.58	1002.68	949.69	2031.12	12501.56
	2.242	3.034	2.324	1.120	1.044	0.989	2.116	1.860
High cash flow weight	828.42	1754.15	955.41	924.26	869.66	961.95	1130.82	8129.37
	0.863	1.827	0.995	0.963	0.906	1.002	1.178	1.210
Crisis period (1997:Q1 - 2017:Q2)								
Low bank risk av.	756.43	595.63	371.42	410.10	448.26	463.51	1067.37	4373.74
	1.556	1.226	0.764	0.844	0.922	0.954	2.196	1.286
High bank risk av.	644.83	565.00	418.54	419.39	410.17	456.32	1045.95	4107.15
	1.327	1.163	0.861	0.863	0.844	0.939	2.152	1.207
Low cap. util. weight	602.49	585.94	392.60	441.09	418.30	458.73	953.19	4066.97
	1.240	1.206	0.808	0.908	0.861	0.944	1.961	1.195
High cap. util. weight	580.09	538.98	425.69	408.38	411.88	459.79	1100.61	4055.01
	1.194	1.109	0.876	0.840	0.847	0.946	2.265	1.192
Low prec. deposit	502.68	567.05	433.57	422.94	398.35	463.97	1089.57	4077.72
	1.034	1.167	0.892	0.870	0.820	0.955	2.242	1.199
High prec. deposit	479.20	521.88	437.37	414.80	414.26	457.16	1093.32	3980.33
	0.986	1.074	0.900	0.854	0.852	0.941	2.250	1.170
Low cash flow weight	1380.74	941.45	622.76	477.20	463.22	436.08	768.60	5655.57
	2.841	1.937	1.281	0.982	0.953	0.897	1.581	1.662
High cash flow weight	604.10	600.62	380.84	470.28	396.12	496.22	686.49	3782.52
	1.243	1.236	0.784	0.968	0.815	1.021	1.413	1.112

- Note: A MCS analysis was carried out but in order to save space no results are included in this table, as all the Caiani et al. (2016) variants are excluded from the confidence set. Compression ratios are included below the score.

B. MIC discretisation diagnostics

Table A-2: Quantisation diagnostics for Monte Carlo validation exercises

	ARMA-ARCH	VAR(2)	
	X	X^1	X^2
Lower bound	-30	-10	-10
Upper bound	30	10	10
Smallest observation	-256.118	-17.243	-14.996
Largest observation	147.040	16.387	15.599
Proportion out of bounds	0.050	0.012	0.009
Proportion of KS fails	0.002	0.002	0.002
Proportion of LB fails	0.058	0.051	0.054
Proportion of Spearman fails	0.788	0.119	0.143

- N = 1000, 1290 series (129 parametrisations, 10 repetitions each)

Table A-3: Quantisation diagnostics for the Smets and Wouters (2007) dataset

	L	r	π	Δy	Δc	Δi	Δw
Support diagnostics							
Lower bound	−7	0	−1	−3	−3	−10	−1
Upper bound	7	5	3	3	3	10	2.5
Min obs.	−7.749	0.209	−1.101	−3.080	−2.957	−9.386	−1.097
Max obs.	5.745	4.445	3.661	3.728	4.763	9.869	2.661
N° out of bounds	4	0	2	4	1	0	3
Resolution (bits)	6	6	6	6	6	6	6
Kolmogorov-Smirnov (KS) tests for uniformity of discretisation errors							
Test statistic	0.061	0.052	0.035	0.057	0.039	0.078	0.070
P-value	0.775	0.906	0.999	0.846	0.994	0.466	0.619
Ljung-Box (LB) tests for autocorrelation of discretisation errors							
Test statistic	20.276***	4.200	2.628	1.781	7.689	3.476	3.726
P-value	0.002	0.650	0.854	0.939	0.262	0.747	0.714
Spearman correlation of discretisation error with discretised data							
Correlation	0.045	−0.000	0.040	0.123*	−0.100	−0.066	−0.100
P-value	0.499	0.998	0.549	0.064	0.131	0.319	0.131

- Tests are run on the full 1947:Q3 - 2004:Q4 Smets and Wouters (2007) dataset, so N = 230.
- KS test H_0: Discretisation errors are uniformly distributed
- LB test H_0: Discretisation errors are independently distributed. $\chi^2(0.05, 6) = 12.592$.

C. Central Smets & Wouters estimates

Table A-4: Prior & posterior distribution of structural parameters for the central SW models

	Prior			Basic			Restricted cons.			Restricted cons. and shocks		
Parameter	Type	Mean	St. Dev.	Mean	90% Interval		Mean	90% Interval		Mean	90% Interval	
φ	Normal	4.00	1.50	5.72	3.97	7.28	0.81	0.54	1.10	0.72	0.48	0.99
σ_c	Normal	1.50	0.38	1.39	1.17	1.59	-	-	-	-	-	-
λ	Beta	0.70	0.10	0.71	0.64	0.78	-	-	-	-	-	-
ξ_w	Beta	0.50	0.10	0.70	0.60	0.81	0.69	0.57	0.81	0.88	0.84	0.91
σ_l	Normal	2.00	0.75	1.82	0.89	2.69	1.51	0.70	2.31	2.36	1.49	3.28
ξ_p	Beta	0.50	0.10	0.65	0.56	0.75	0.66	0.57	0.76	0.67	0.60	0.74
ι_w	Beta	0.50	0.15	0.58	0.39	0.79	0.55	0.34	0.75	0.72	0.57	0.87
ι_p	Beta	0.50	0.15	0.24	0.09	0.38	0.22	0.09	0.37	0.15	0.05	0.24
ψ	Beta	0.50	0.15	0.55	0.38	0.74	0.60	0.39	0.82	0.61	0.43	0.81
ϕ_p	Normal	1.25	0.13	1.61	1.48	1.74	1.39	1.27	1.51	1.41	1.30	1.52
r_π	Normal	1.50	0.25	2.04	1.75	2.34	2.23	1.94	2.49	2.12	1.84	2.40
ρ	Beta	0.75	0.10	0.81	0.77	0.85	0.79	0.74	0.83	0.80	0.76	0.84
r_y	Normal	0.13	0.05	0.09	0.05	0.13	0.12	0.07	0.16	0.13	0.08	0.17
$r_{\Delta y}$	Normal	0.13	0.05	0.23	0.18	0.27	0.30	0.25	0.34	0.29	0.25	0.34
$\bar{\pi}$	Gamma	0.63	0.10	0.78	0.60	0.96	0.79	0.63	0.95	0.88	0.74	1.02
$\tilde{\beta}$	Gamma	0.25	0.10	0.17	0.07	0.26	0.19	0.10	0.30	0.20	0.09	0.31
$\bar{l}$	Normal	0.00	2.00	0.53	-1.36	2.33	0.32	1.63	1.05	-1.09	-2.08	-0.02
$\bar{\gamma}$	Normal	0.40	0.10	0.43	0.41	0.46	0.42	0.39	0.45	0.38	0.36	0.41
α	Normal	0.30	0.05	0.19	0.16	0.22	0.19	0.16	0.22	0.20	0.17	0.24

- The log marginal densities are -922.76 for the basic model -985.51 for the restricted consumption model and -1021.78 for the model with restricted consumption and shocks model.
- Note: β denotes the modified discount rate $\tilde{\beta} = 100(\beta^{-1} - 1)$

Table A-5: Prior & posterior distribution of shock processes for the central SW models

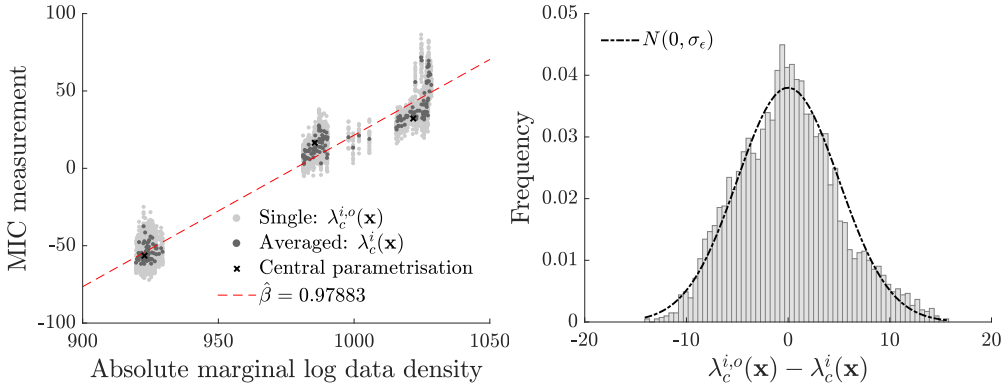
	Prior			Basic			Restricted cons.			Restricted cons. and shocks		
Parameter	Type	Mean	St. Dev.	Mean	90% Interval		Mean	90% Interval		Mean	90% Interval	
σ_a	Invgamma	0.10	2.00	0.46	0.41	0.50	0.49	0.43	0.54	0.49	0.43	0.53
σ_b	Invgamma	0.10	2.00	0.24	0.20	0.28	0.29	0.24	0.33	0.29	0.24	0.33
σ_g	Invgamma	0.10	2.00	0.53	0.48	0.58	0.51	0.46	0.56	0.58	0.53	0.64
σ_i	Invgamma	0.10	2.00	0.45	0.37	0.54	0.68	0.48	0.86	0.86	0.59	1.13
σ_r	Invgamma	0.10	2.00	0.25	0.22	0.27	0.28	0.24	0.31	0.27	0.24	0.30
σ_p	Invgamma	0.10	2.00	0.14	0.11	0.17	0.15	0.13	0.18	0.11	0.08	0.14
σ_w	Invgamma	0.10	2.00	0.24	0.21	0.28	0.27	0.23	0.32	0.24	0.19	0.28
ρ_a	Beta	0.50	0.20	0.96	0.94	0.98	0.96	0.94	0.97	0.95	0.92	0.97
ρ_b	Beta	0.50	0.20	0.21	0.07	0.35	0.82	0.76	0.88	0.84	0.78	0.89
ρ_g	Beta	0.50	0.20	0.98	0.96	0.99	0.96	0.94	0.98	0.95	0.93	0.97
ρ_i	Beta	0.50	0.20	0.71	0.61	0.81	0.95	0.92	0.99	0.98	0.96	0.99
ρ_r	Beta	0.50	0.20	0.15	0.05	0.26	0.05	0.01	0.09	0.05	0.01	0.09
ρ_p	Beta	0.50	0.20	0.89	0.80	0.97	0.89	0.80	0.97	0.63	0.50	0.76
ρ_w	Beta	0.50	0.20	0.97	0.95	0.99	0.95	0.91	0.98	0.32	0.19	0.46
μ_p	Beta	0.50	0.20	0.69	0.53	0.86	0.71	0.57	0.86	-	-	-
μ_w	Beta	0.50	0.20	0.84	0.74	0.94	0.86	0.77	0.94	-	-	-
ρ_{ga}	Normal	0.50	0.25	0.52	0.36	0.67	0.55	0.42	0.68	-	-	-

D. Results for alternate MIC settings

Table A-6: Variable-level and aggregate MIC for SW models, low resolution, 3 lags

	L	r	π	Δy	Δc	Δi	Δw	Aggr.
MIC measurements								
Benchmark	184.75	196.20	242.81	353.74	322.94	306.00	384.65	1952.63
	(0.028)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.114)	(0.000)
Restrict 1	184.61	208.90	249.19	373.84***	346.99***	318.45**	384.38	2025.56***
	(0.000)	(0.849)	(1.126)	(4.136)	(2.795)	(2.493)	(0.000)	(4.214)
Restrict 2	189.36	204.79	261.38**	364.71***	360.46***	322.58***	384.71	2041.26***
	(1.802)	(0.604)	(2.804)	(2.736)	(6.690)	(3.524)	(0.129)	(4.373)
Compression ratios								
Benchmark	0.385	0.409	0.506	0.737	0.673	0.637	0.801	0.581
Restrict 1	0.385	0.435	0.519	0.779	0.723	0.663	0.801	0.603
Restrict 2	0.395	0.427	0.545	0.760	0.751	0.672	0.801	0.608

- Note: 'Restrict 1' refers to the model with restricted consumption parameters, 'restrict 2' is the model with additional restrictions to the shock processes. MCS t-statistics are provided in parenthesis, with superscripts '*', '**' and '***' indicating that the model is excluded from the confidence set at the 10, 5 and 1% significance level, based on bootstrapped standard errors.



(a) MIC vs. log-likelihood (b) Measurement error

Figure A-3: MIC results on SW models, Low resolution, 3 lags

Table A-7: Macroeconomic model comparison, low resolution, 3 lags

	L	r	π	Δy	Δc	Δi	Δw	Aggr.
Smets & Wouters dataset (1965:Q1 - 2004:Q4)								
VAR(1)	183.57	196.33	261.88***	353.77	322.32	308.42	384.18	1968.34
	0.382	0.409	0.546	0.737	0.671	0.643	0.800	0.586
SW	184.75	196.20	242.81	353.74	322.94	306.00	384.65	1952.63
	0.385	0.409	0.506	0.737	0.673	0.637	0.801	0.581
Caiani et al.	351.71***	923.42**	1192.71**	490.27***	392.96***	430.59***	1211.44***	5158.22***
	0.733	1.924	2.485	1.021	0.819	0.897	2.524	1.535
Crisis period (1997:Q1 - 2017:Q2)								
VAR(1)	67.97	47.41	94.47	137.58	127.25	109.37	226.99	798.27
	0.287	0.200	0.399	0.580	0.537	0.461	0.958	0.481
SW	75.04	53.04	99.29	145.74	133.14	115.53	225.68	831.66***
	0.317	0.224	0.419	0.615	0.562	0.487	0.952	0.501
Caiani et al.	282.93***	137.99*	122.98	178.81**	183.94***	208.87***	650.55***	1979.83***
	1.194	0.582	0.519	0.754	0.776	0.881	2.745	1.193

- Note: A MCS analysis was carried out but only the significance of the test is reported in order to save space. As before, superscripts '*', '**' and '***' indicating that the model is excluded from the confidence set at the 10%, 5% and 1% significance level, based on bootstrapped standard errors. Compression ratios are included below the score.

Table A-8: Caiani et al. (2016) sensitivity analysis, low resolution, 3 lags

	L	r	π	Δy	Δc	Δi	Δw	Aggr.
Smets & Wouters dataset (1965:Q1 - 2004:Q4)								
Low bank risk av.	477.37	857.75	1198.56	511.94	371.23	441.37	1193.35	5286.38
	0.995	1.787	2.497	1.067	0.773	0.920	2.486	1.573
High bank risk av.	292.91	951.41	956.36	442.26	390.93	441.13	1253.55	4943.56
	0.610	1.982	1.992	0.921	0.814	0.919	2.612	1.471
Low cap. util. weight	327.08	940.26	547.32	503.41	426.84	393.51	1086.94	4517.15
	0.681	1.959	1.140	1.049	0.889	0.820	2.264	1.344
High cap. util. weight	280.19	953.15	931.00	430.36	379.32	449.20	1287.33	4916.17
	0.584	1.986	1.940	0.897	0.790	0.936	2.682	1.463
Low prec. deposit	281.66	963.10	912.50	434.54	386.53	439.38	1249.89	4868.70
	0.587	2.006	1.901	0.905	0.805	0.915	2.604	1.449
High prec. deposit	266.78	920.71	941.42	433.34	379.51	432.24	1254.54	4873.31
	0.556	1.918	1.961	0.903	0.791	0.900	2.614	1.450
Low cash flow weight	1478.62	1936.21	1627.57	581.38	470.34	373.19	1451.49	8114.01
	3.080	4.034	3.391	1.211	0.980	0.777	3.024	2.415
High cash flow weight	307.39	952.01	437.98	515.59	422.94	518.57	670.77	4238.57
	0.640	1.983	0.912	1.074	0.881	1.080	1.397	1.261
Crisis period (1997:Q1 - 2017:Q2)								
Low bank risk av.	392.97	175.16	132.52	176.66	175.14	214.78	653.05	2165.89
	1.658	0.739	0.559	0.745	0.739	0.906	2.755	1.306
High bank risk av.	195.84	132.23	134.15	166.95	158.84	209.17	650.63	1845.62
	0.826	0.558	0.566	0.704	0.670	0.883	2.745	1.112
Low cap. util. weight	159.10	148.63	125.83	182.37	197.60	194.84	613.93	1828.72
	0.671	0.627	0.531	0.769	0.834	0.822	2.590	1.102
High cap. util. weight	168.51	134.29	130.68	164.41	159.91	217.24	681.02	1835.83
	0.711	0.567	0.551	0.694	0.675	0.917	2.873	1.107
Low prec. deposit	158.09	142.43	135.75	162.78	163.66	209.18	656.35	1804.77
	0.667	0.601	0.573	0.687	0.691	0.883	2.769	1.088
High prec. deposit	169.78	156.77	124.51	162.22	157.00	213.19	684.83	1837.64
	0.716	0.661	0.525	0.684	0.662	0.900	2.890	1.108
Low cash flow weight	782.74	340.14	275.90	202.78	172.53	173.31	649.69	3064.28
	3.303	1.435	1.164	0.856	0.728	0.731	2.741	1.847
High cash flow weight	151.58	162.45	106.90	203.30	196.70	244.75	412.43	1612.84
	0.640	0.685	0.451	0.858	0.830	1.033	1.740	0.972

- Note: A MCS analysis was carried out but in order to save space no results are included in this table, as all the Caiani et al. (2016) variants are excluded from the confidence set. Compression ratios are included below the score.

Recent Kent Discussion Papers in Economics

19/07: Ethnic Identities, Public Spending and Political Regimes', Sugata Ghosh and Anirban Mitra

19/06: 'The Effect of 9/11 on Immigrants' Ethnic Identity and Employment: Evidence from Germany', Isaure Delaporte

19/05: 'Social Effects of the Vote of the Majority: A Field-Experiment on the Brexit-Vote', Fernanda L. Lopez de Leon and Markus Bindemann

19/04: 'Ethnic Identity and the Employment Outcomes of Immigrants: Evidence from France', Isaure Delaporte

19/03: 'The Determinants of Tax Revenue and Tax Effort in Developed and Developing Countries: Theory and New Evidence 1995-2015', Marcelo Piancastelli and A.P.Thirlwall

19/02: 'Make Yourselves Scarce: The Effect of Demographic Change on the Relative Wages and Employment Rates of Experienced Workers', Michael J. Böhm and Christian Siegel

19/01: 'Engines of Sectoral Labor Productivity Growth', Zsófia L. Bárány and Christian Siegel

18/14: 'The Effects of Risk and Ambiguity Aversion on Technology Adoption: Evidence from Aquaculture in Ghana', Christian Crentsil, Adelina Gschwandtner and Zaki Wahhaj

18/13: 'A Simple Model of Growth Slowdown', Katsuyuki Shibayama

18/12: 'Measured Productivity with Endogenous Markups and Economic Profits', Anthony Savagar

18/11: 'Endogenous Time-Varying Volatility and Emerging Market Business Cycles', Jan-Philipp Dueber

18/10: 'Marriage, Work and Migration: The Role of Infrastructure Development and Gender Norms', Amrit Amirapu, M Niaz Asadullah and Zaki Wahhaj

18/09: 'Population Aging, Government Policy and the Postwar Japanese Economy', Keisuke Otsu and Katsuyuki Shibayama

18/08: 'The Missing Link: Monetary Policy and The Labor Share', Cristiano Cantore, Filippo Ferroni and Miguel A. León-Ledesma

18/07: 'University vice-chancellor pay, performance and (asymmetric) benchmarking', Adelina Gschwandtner and Richard McManus

18/06: 'Constrained public goods in networks', Nizar Allouch and Maia King

18/05: 'The Fall in German Unemployment: A Flow Analysis', Carlos Carrillo-Tudela, Andrey Launov and Jean-Marc Robin

18/04: 'Labor Responses, Regulation and Business Churn in a Small Open Economy', Marta Aloi, Huw Dixon and Anthony Savagar