An Information Theoretic Criterion for Empirical Validation of Time Series Models

Francesco Lamperti*

Sant'Anna School of Advanced Studies
Institute of Economics

February 27, 2015

Abstract

Simulated models suffer intrinsically from validation and comparison problems. The choice of a suitable indicator quantifying the distance between the model and the data is pivotal to model selection. However, how to validate and discriminate between alternative models is still an open problem calling for further investigation, especially in light of the increasing use of simulations in social sciences. In this paper, we present an information theoretic criterion to measure how close models' synthetic output replicates the properties of observable time series without the need to resort to any likelihood function or to impose stationarity requirements. The indicator is sufficiently general to be applied to any kind of model able to simulate or predict time series data, from simple univariate models such as Auto Regressive Moving Average (ARMA) and Markov processes to more complex objects including agent-based or dynamic stochastic general equilibrium models. More specifically, we use a simple function of the L-divergence computed at different block lengths in order to select the model that is better able to reproduce the distributions of time changes in the data. To evaluate the L-divergence, probabilities are estimated across frequencies including a correction for the systematic bias. Finally, using a known data generating process, we show how this indicator can be used to validate and discriminate between different models providing a precise measure of the distance between each of them and the data.

JEL codes: C15, C52, C63

Keywords: Simulations, Empirical Validation, Time Series, Agent Based Models

*email address: f.lamperti@sssup.it

The author is indebted with Andrea Roventini and Giovanni Dosi for incredibly valuable support and discussions. He would also thank Sylvain Barde, Gianbiagio Curato, Daniele Giachini, Mattia Guerini, Marco LiCalzi, Fabrizio Lillo, Marco Lippi, Alessio Moneta, Matteo Sostero and all the participants to the 11^{th} ESSA Annual Meeting in Barcelona and 8^{th} CFE International Conference in Pisa for useful comments and suggestions. He also acknowledges financial support from European Unions 7th FP for research, technological development and demonstration under G.A. No 603416 - Project IMPRESSIONS (Impacts and risks from high-end scenarios: Strategies for innovative solutions).

1 Introduction

In this paper we introduce a new information-theoretic criterion, called *Generalized Subtracted L-divergence* (*GSL-div*), to measure the distance between the dynamics of time series produced by a model and the empirically observable counterpart. The *GSL-div* can be used to quantitatively establish the empirical validity of a model and to discriminate between sets of competing ones. Various properties well suited to the scope of model validation are introduced and discussed. Unlike many other indicators, the *GSL-div* relies only on the synthetic data generated by simulations and does not impose any additional assumptions on the underlying stochastic processes. Its flexibility allows applications to every model involving production of time series, admitting also a direct comparison between classes (e.g. macro Agent Based Models, Dynamical Stochastic General Equilibrium models and System Dynamics approaches). Interestingly, it does not require knowledge of the probabilistic structure (or likelihood) of the series it uses. This feature makes the *GSL-div* especially well suited for validating Agent Based Models (ABM), where the statistical properties of aggregate variables are *a priori* unknown.

A common protocol for empirical validation is still an open problem for simulation studies (see [63, 22] and more recently [44]). Finding tools that are appropriate to the scope is crucial both for the scientific debate and for policy analysis; the academia needs to develop theories whose implications fit with empirical evidences, and policy makers need information coming from reliable models. Assessing the fit of different models with empirical data is exactly what we are doing in this paper. Manson ([59]) distinguishes between *output validation* and *structural validation*. The latter asks how well the simulation model represents the (prior) conceptual model of the real-world system, while the former asks how successfully the simulations' output exhibits the historical behaviours of the real-world target system. Output validation can be directly related to what Leombruni et al. ([53]) define as *empirical validity* of a model, i.e. validity of the empirically occurring true value relative to its indicator. Leombruni et. al introduce other four validity concepts that simulation studies must consider: theory (the validity of the theory relative to the simuland), model (the validity of the model relative to the theory), program (the validity of the simulating program relative to the model), operational (the validity of the theoretical concept to its indicator or measurement). Any-time simulations exhibit lacks with respect to one or more of these validities, empirical validity is in turn affected and thereby reduced. Following [52], it is useful to think of two parallel unfolding: the evolution of the system (an economy, a market or whatever) and the evolution of the model of the system. If the model is properly specified and calibrated, the simulation should mirror the historical evolution of the real-world system with respect to the variables of interest.

Empirical validation is crucial for all modelling efforts that attempts at providing support to policy decisions, independently of their theoretical background[1]. In macroeconomics, for ex-

[1]With reference to the ABM community see the special issue on agent based models for economic policy edited

ample, [28, 11, 49] provide details about how to estimate and validate DSGEs models. However, their approach cannot be extended to different settings where agents' heterogeneity or non-linear relationships are part of the picture, which is the typical case in ABMs and Systems Dynamics respectively. The community will certainly benefit from the construction of common grounds where different models can be compared in terms of their empirical validity (see [48] and [34]). Moreover, ABMs are sometimes considered as candidates to substitute or complement DSGEs as the standard tool for macroeconomic analysis ([24]). One of the major critiques to these approaches is a substantial lack of a sound empirical grounding ([26]). With respect to this issue, one of the problems with ABM is that the statistical properties of policy-relevant variables (e.g. inflation, GDP, employment, rate of adoption of a new technology) are *a priori* unknown, even to the modeller. The reason is that they *emerge* indirectly from the repeated interactions among ecologies of heterogeneous, boundedly rational and adapting agents. These interactions give rise to the emergence of properties that cannot be simply deduced by aggregating those of micro variables ([5],[61], [33], [27]). Therefore, the synthetic output generated by the simulation constitutes the only available source of information about the aggregate behaviour of the model. Our approach is tailored on this constriction: simulated and observed data are the only input we require.

We tackle directly the fourth issue raised in [63], i.e. validating agent based models using historical data. However, since the *GSL-div* can be applied to any kind of model producing time series, it also provides a framework to compare models' relative performances: going beyond validation we offer a tool for model selection. Therefore, we integrate both the literature addressing empirical validity of simulation models, and ABMs in particular, (see [22] and references therein, [45], [10], [12], [35], [34][2] and that on model selection[3] ([1] and subsequent developments, [62, 46, 14]). In particular, our criterion allows to quantify the distance between the *true* probabilistic dynamics of models' output and the data even when the former is unknown *a priori*. This result is obtained via the simulation of an ensemble of independent runs and the analytical correction of a systematic bias arising in the estimation process. In particular, we define and compute an information theoretic criterion based on a simple function of the L-divergence ([43]). Validation is achieved capturing the ability of a given model to reproduce the distributions of time changes (that is, changes in the process' values from time to time) in the real-world observed process, without the need to resort to any likelihood function or to

by Fagiolo and Dawid ([16]).

[2]See also the website maintained by Leigh Tesfatsion (`http://www2.econ.iastate.edu/tesfatsi/empvalid.htm` for additional interesting material about empirical validation. More specifically, the majority of models in the literature has undertaken either a simple output validation process based on stylized facts replication ([19, 17, 18]) or an input validation process based on parameters' estimation ([29, 3, 4, 2, 9, 10]). Our criterion both provide a more detailed measure for output validation and can be adapted as a distance to be minimized in the context input validation, as briefly outlined below.

[3]The literature on model selection is incredibly vast. Here we refer to the case of establishing which model produces time series that are closest to that it is intended to model.

impose any stationarity requirements.

In a similar context, [34] studies the estimation of ergodic ABM by simulated minimum distance (SMD). It is relevant to remark that in this paper we refer to models having already undertaken a calibration procedure: parameters' values have been already assigned an estimate. However, our approach is complementary. First, it can be used to test how close models estimated via SMD reproduce the behaviour of the data; secondly, it offers a nice object to compare data and simulations in the time dimension. In many applications, SMD minimizes a quadratic form specifying the distance between sample moments and predicted moments from simulations of a given model ([41], [20], [36] and more recently [35] where the procedure is applied to a simple ABM). When the distance is taken between coefficients of meta-parameters of a given meta-model we get indirect inference ([30]). Relevant shortcomings of these techniques are the usual requirement to deal with stationary and ergodic series[4] and, to the purposes of validation, that moments are poor representations of time series' behaviour. Our approach allows to partially overcome these limitations and offers a meaningful object to be used in SMD estimation in the place of standard longitudinal moments. In addition, the small sample problem identified in [34] can be solved or attenuated. Moreover, our exercise is similar to some of those carried out in [10]: model's output is directly compared with empirical data. What we add is something that seems missing in this literature: a precise quantification of the distance between the model and data with respect to their dynamics in the time domain. In this respect, our work builds on [46] and extend it by capturing the dynamical nature of time series models. With an underlying purpose similar to ours, [6] has recently developed an interesting generalization of the Akaike Information Criterion ([1]) that can be used for model selection. It assumes each model to produce an n^{th}-order Markov process and computes the intrinsic complexity of each series on the basis on the conditional probabilistic structure of the output, which has been previously converted into binary strings. Our approach, which builds on frequencies, is less computational intense and give more emphasis to similarities in the dynamical behaviour of the data, rather then their complexity.

For illustrative purposes, we use the *GSL-div* to compare univariate time series mainly produced by the simulation of different ARMA processes with very similar parametric representations. The choice of such stochastic processes is instrumental: knowing everything about their stochastic behaviour we are able to test the ability of the *GSL-div* in distinguishing processes we are sure to be different, but showing dynamics that are difficult (or even impossible) to be discerned by inspection. We obtain that the *GSL-div* is very precise in discerning the dynamics of such models and we show its robustness with respect to the only two free parameters of our approach. As a remark, even if the we use univariate settings, the approach can be extended straightforwardly to multivariate data structures.

[4] this is a strong limitation in ABM where for multiple regimes and tipping points might generate, by definition, non-ergodic series.

The paper is organized as follows. Section 2 introduces and defines the *Generalized Subtracted L-divergence* between models and data; section 3 provides a set of properties characterizing the *GSL-div* and emphasizes their role to the purposes of model validation. Section 4 presents a simple application to *ad-hoc* chosen time series. Section 5 provides the main evidences of robustness for this new criterion, showing its ability to distinguish between very similar stochastic processes. Finally, section 6 concludes the paper.

2 A novel approach to model validation: the GSL-div

In this section we present a novel approach to measure the distance between time series. It consists on the development of a new information theoretic criterion, called Generalized Subtracted L-divergence. This criterion is used to provide an answer to the problems of model selection and model validation outlined in the previous section.

In our context, a *model* is broadly defined as a representations of a system which is able to produce some synthetic output tracking the evolution of the system. For example, both macro ABMs and DSGEs are models of the economic system or parts of it[5]. In addition, we define *real world data* as the empirically observable elements of the system. Establishing the *empirical validity* of a model concerns the evaluation of the relationship of similarity between models' output and the real world data. To be concrete, consider for example Figure 1. It shows a plot of the rate of growth of quarterly GDP during the period 2001-2011 for the US. Suppose we have a properly calibrated model trying to explain its behaviour. How can one claim that the output of the model is empirically valid? As our argument goes, it has to show a reasonable degree of similarity with the empirically observable counterpart. In particular, we want the model to reproduce similar *trends* (if any), similar patterns of *concave* and *convex behaviour* (with respect to time), similar *persistence* over some interval (for instance, from Figure 1 we see that the rate of growth of the US GDP falls between 0 and 4% most of the times) and similar *sequences of time to time changes* in the series.

Conversely, there are features of the real world series we do not strictly require models to capture. This is due both to the fact that models are, by definition, only stylized representations of a system, and to the intrinsic complexity of establishing an exact mapping between periods in a simulation and time in the observed world. The last point is touched in [63] with reference to the difficulty of selecting the starting and ending point of a simulation run. We argue that this problem is even deeper and affects each period of a simulation. In particular, if the series display multiple regimes, it might not only be that mis-specifying the initial point of the simulation we compare different regimes even though the model perfectly reflects the data, but it could happen that the length of regimes differs from real data and simulations. This introduces a sequence

[5]For a comprehensive treatment of similarities and differences between macro ABM and DSGE models see [24]

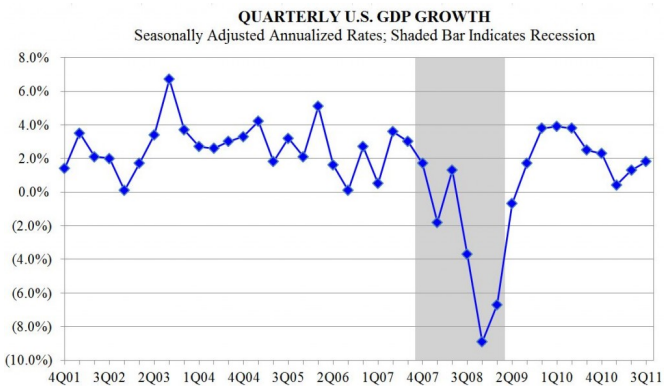

Figure 1: US quarterly GDP growth rate. Source: U.S. Bureau of Economic Analysis.

of mis-matches between contemporaneous time changes in the two series (simulated and real) even though the qualitative nature of regimes and their order of appearance is the same. The question concerns how relevant this issue could be for empirical validity. Our answer is that it depends. Having an exact time correspondence should be rewarded, but validating models on this ground (for example, using some quadratic error measure) could lead to misleading conclusions.

For instance, we can consider the Great Moderation and the Great Recession as two different regimes for the US' quarterly GDP growth rate. There is little concern in saying that the Great Moderation started around mid-80s ([60]) and concluded with the last quarter of 2007: this implies a total length of around 88 quarters. Assume we have two models (A and B) and both are calibrated with quarterly data. Model A simulates a time span where the economy experiences a reduction in volatilities, but it takes around 100 periods to show a pattern similar to the Great Recession. Each time we compare a model's run with the real series we find that the starting point of the Great Recession is not matched. Assume instead that B identifies the regime shift perfectly but thereafter exhibit an expansionary regime. If the latter is is smooth enough and the two series are compared using some distance function between contemporaneous observations of the two series, it might result that model A deviates from real data more than model B does, regardless of the fact that the latter never produces the Great Recession regime. In addition, it is worth to note that our criterion is designed to capture similarities in the behaviour of the time series, and not in their levels. This reflects the opinion that is not relevant for a simulation to mirror the same values of the real data but to display the same shape in terms of trends, variabilities and trajectories. Furthermore, given two series sharing the same dynamics but different levels, it is sufficient to change initial conditions to notice they are in effect identical. A simple addition or subtraction of a constant to each variable in one of the two processes could serve the scope. Finally, levels largely depend on the unit of measure used by different models, while series' behaviour does not.

To sum up in few words, we argue that empirical validity should be assessed on the ground of similarity between the dynamic behaviour of a system and that of the model of the system. What matters are patterns, not absolute errors.

2.1 The Theoretical Background

As well explained in [46], using the glasses of information theory rather than statistics, the observed data contain information, and the (descriptive) models we develop (from our theoretical understanding of the underlying processes generating the observed data) can be seen of as attempts to reproduce the highest possible fraction of these information, in the most compact way. When several models referring to the same phenomenon are available, empirical validation should be able to point out the "best" model, that is the model whose output lose the least amount of information with respect to the real-world data.

Distance or divergence measures are widely used in a number of theoretical and applied statistical inference and data processing problems, including estimation, detection, compression and model selection ([8]). Most of them rely largely on the concept of Shannon's entropy ([58]), which expresses the amount of uncertainty associated with a random variable. Among these measures, one of the best known is the Kullback-Leibner divergenge (*KL-div)* between two distributions, $D(\mathbf{p}||\mathbf{q})$, or *relative entropy* ([40]). It is a measure of the inefficiency of assuming that the distribution is $\mathbf{q}$ when the true one is $\mathbf{p}$. The following discussion will be limited to discrete probability distributions, but results can be generalized to probability density functions. Consider a discrete random variable with support indicated by S and probability mass function $p(s)$, $s \in S$. If $q(s)$ is another probability mass function defined on the same support S, the *KL-div* is defined as

$$D_{KL}(\mathbf{p}||\mathbf{q}) = \sum_{s \in S} p(s) \log \left(\frac{p(s)}{q(s)} \right), \tag{1}$$

where the logarithm is, usually, in base 2. Throughout the paper the following conventions will be used: $0 \log(0/0) = 0$ and, on the basis of continuity arguments, $0 \log(0/q) = 0$, independently of the logarithm's base. It is immediate to see that if there exist any symbol $s \in S$ such that $p(s) > 0$ and $q(s) = 0$ then, $D_{KL}(\mathbf{p}||\mathbf{q})$ is undefined. This means that distribution $\mathbf{p}$ has to be absolutely continuous with respect to $\mathbf{q}$ for the *KL-div* to be defined [39]. In addition, the $D_{KL}(\mathbf{p}||\mathbf{q})$ is non-negative, additive but not symmetric. In order to overcome these problems Lin ([41]) defined a new symmetric measure, called L-divergenge, shortly *L-div*:

$$D_L(\mathbf{p}||\mathbf{q}) = D_{KL}(\mathbf{p}||\mathbf{m}) + D_{KL}(\mathbf{q}||\mathbf{m}), \tag{2}$$

where $\mathbf{m} = (\mathbf{p} + \mathbf{q})/2$ is the *mean* probability mass function. As the names suggest the *L-div* is the basic building block we will use to construct the *GSL-div*. It is immediate to see that

$D_L(\mathbf{p}||\mathbf{q})$ vanishes only if $\mathbf{p} = \mathbf{q}$ and that the L-divergence is bounded above by two. This is more evident when expressing the L-divergence in terms of the Shannon entropy, that is

$$D_L(\mathbf{p}||\mathbf{q}) = 2H\left(\frac{\mathbf{p}+\mathbf{q}}{2}\right) - H(\mathbf{p}) - H(\mathbf{q}), \tag{3}$$

i.e. the difference between twice the mean distribution and the sum of the entropies of $\mathbf{p}$ and $\mathbf{q}$. The generalization of the *L-div* is the Jensen-Shannon divergence (see [43]), defined as

$$Div_{JS}(\mathbf{p}, \mathbf{q}) = H(\pi_1\mathbf{p} + \pi_2\mathbf{q}) - \pi_1 H(\mathbf{p}) - \pi_2 H(\mathbf{q}), \tag{4}$$

where the weights π_1 and π_2 must satisfy $\pi_1, \pi_2 \geq 0$ and $\pi_1 + \pi_2 = 1$. It is straightforward that $D_L(\mathbf{p}||\mathbf{q}) = 2Div_{JS}(\mathbf{p}, \mathbf{q})$ for $\pi_1 = \pi_2 = 1/2$. It is to be noticed that the *KL-div*, and consequently the *L-div*, does not satisfy the triangle inequality, and hence cannot be considered a proper metric[6]. With reference to the use of these measures as tools for model validation and selection Marks ([46]) outlines their inadequacy due to the previous problem. However, if models' data-distributions (say $\mathbf{q}$) are always compared directly with the real data-distribution (say $\mathbf{p}$), and not among themselves, model selection does not need a metric satisfying triangle inequality. Moreover, Endres et al. ([21]) found that the square root of the *L-div* is a metric and they called it the Jensen-Shannon distance.

We use the *L-div* as a measure that captures the distance between the distributions of time-changes in the real-world process and those generated by the synthetic output of simulated, competing models. Time-windows of different lengths are taken into consideration for the generation of the state space, which is represented by the set of values the series might take at each instant of time. The *L-div* is estimated for all available lengths of the time-window and results are finally aggregated into a single information criterion, the Generalized Subtracted L-divergence (shortly *GSL-div*). Our approach can be seen as an extension of the work provided in [15], where the Jensen-Shannon divergence is used to measure the distance between distributions of single observations at different aggregation scales, and [42], where time series are symbolized in a similar way.

2.2 The *GSL-divergence*

Before introducing the functional form of the *GSL-div* it is necessary to describe the preliminary procedures time series undergo before being analysed. In order to be manageable for similarity assessment, they are *symbolized* and *sub-divided* in successive blocks. Hence, frequency distributions over alphabets of symbols are estimated and used to fed the *GSL-div*.

The procedure starts with the symbolization of the series, which is carried out to constrain series to take only a finite set of values. In particular, consider a time series $\{x(t)\}_{t=1}^{T}$, where each

[6] A metric is a distance function which must satisfy non-negativity, symmetry, coincidence and triangle inequality (see Chapter 2 in [55]).

$x(t)$ is a real number. To symbolize it, we firstly take the real interval $[x_{min}; x_{max}]$ and partition it in $b \in \mathbf{N}_0$ subintervals, each of equal length. These intervals are numbered increasingly from 1 to b, with 1 assigned to $[x_{min}; x_{min} + \frac{(x_{max} - x_{min})}{b})$. The parameter b controls for the precision of the symbolization: for $b = 1$ the symbolized series takes one and only one value (namely 1) while for $b \to \infty$ we are back to the (scaled) real-valued process. The symbolization is simple and works as follows: each $\{x(t)\}_{t=1}^{T}$ is mapped into the natural number corresponding to the partition interval where it falls. For example, consider the following realization of the stochastic process $x(t)$ with $t = 3$: $\{0; 0.65; 1\}$. Choosing $b = 2$, the symbolized series will be $x^s(t) = \{1, 2, 2\}$, while choosing $b = 10$ the symbolized series becomes $x^s(t) = \{1, 7, 10\}$. It is immediate to see that increasing b the information loss about the behaviour of the stochastic process due to the symbolization becomes smaller and smaller. However, as it typically happens, increasing the precision of the symbolization has a cost: higher b translates also in higher size of the alphabet, that is the total number of words that could be created using symbols $\{1, ..., b\}$. The size of the alphabet corresponds to the cardinality of the state space, and increasing it might require larger time series to conveniently estimate frequency distributions over the alphabet. However, as will be shown, the *GSL-div* does not suffer from the use of low values of b. High precisions of the symbolization procedure can be used with no worries when large amounts of data are available, for example in high frequency models of financial markets (see, among others, these recent contributions [14] and [23], where our methodology is directly applicable to the time series of stock prices). A more detailed discussion about the partitioning of the state space when dealing with information theoretic functional is provided in [50]. To the purposes of our analysis we recall that it is important to use symbolized time series obtained applying the same precision (b).

Once the time series are properly symbolized a second procedure applies. They are subdivided in blocks of equal length l, each block corresponding to a symbol of the available alphabet. This operation is recursive for $l = 1, ..., L$, where L is the maximum block's length (time-window) considered. The alphabet is composed by all possible combinations of the first b natural numbers taken l at the time.

Definition 2.1 *(Alphabet)*
Let $b \in \mathbf{N}_0$ be the precision of the symbolization process and $L \in \mathbf{N}_0$ the maximum block length. Then, for each $l < L$, we define the corresponding alphabet as

$$S_{l,b} = \binom{A_b}{l} \tag{5}$$

where $A_b = \{1, 2, ..., b\}$. That is, $S_{l,b}$ is the set of all the l-combinations of A_b.

The cardinality of the alphabet is defined as

$$a_{l,b} = 2^{S_i} = b^l, \ \forall l = 1, ..., L \tag{6}$$

and corresponds to the number of different symbols the series' blocks of length l might be associated to once b is chosen. It is relevant to see that, once l is fixed, blocks might overlap. For instance, consider $x^s(t) = \{1, 2, 2\}$. For $l = 2$, it is possible to identify two blocks: $\{1, 2\}$ and $\{2, 2\}$, which correspond to symbols (12) and (22) respectively. Since there are T observations for each series, $T-l+1$ blocks will be obtained for each value of l. L represents the maximum length of the windows which are used to compare the behaviour of the real data with the synthetic ones. It has to be chosen considering both (i) the nature of the phenomenon of interest and (ii) the size of the available real-world time series which can be used to validate the models. The first criterion reflects the time-horizon one considers when analysing a given phenomenon. For example, when the focus is on business cycles, data will be typically quarterly, series will cover around 200 periods and time-window of interests will lasts around eight or twelve periods; conversely, in case one considers economic growth in the long run, data will be annual and the window considerably enlarged. The second criterion puts a constraint on the comparability of real and simulated data: when a real-world time series of length T is the only available source of information about the phenomenon under study, it makes a non-sense to compare it with a double-length simulated series. On the other hand it could be perfectly reasonable to take an ensemble of replicated series each of length T, both to wash away across-simulation variability and to solve the small sample problem ([38]).

Now, frequency of symbols in each series are estimated. Let $x^s(t)$ and $y^s(t)$ be two symbolized time series. $S_{b,l}$ is the alphabet and $\mathbf{f}, \mathbf{f}'$ are vectors collecting the occurrence frequencies of all available symbols. Even though $\mathbf{f}$ and $\mathbf{f}'$ are very rough estimates of the probability the two original processes ($x^s(t)$ and $y^s(t)$) assign to symbols, particularly when processes are not stationary and ergodic, it will be showed that taking an ensemble of independent runs of the same processes[7], frequency vectors converges to a particular probability distribution, which is well suited to compare behaviours of simulated data.

Once frequency vectors are estimated, obtaining the *GSL-div* is straightforward. For each value of l, a subtracted version of the *L-div* is estimated from the data. It provides a measure of how close the behaviour of synthetic data replicates the real one when the series are studied along windows of length l. The *GSL-div* aggregates subtracted *L-div* values using weights increasing in l. In general, the choice of weights is up to the modeller. In the next subsection we introduce three different set of weights and discuss in more details the rationale of assigning greater importance to *GSL-div*s obtained comparing longer time-windows. With these premises, the *GSL-div* between two distributions can be defined as follows.

Definition 2.2 *(GSL-div between distributions)*
Assume that b and L have been fixed. Let $x^s(t)$ and $y^s(t)$ be two symbolized time series exhibiting respectively frequencies $\boldsymbol{f}$ and $\boldsymbol{f}'$ over the alphabet S_l. We define the Generalized Subtracted L-

[7]This is possible since we are using simulated models.

divergence between the two distributions as

$$
\begin{aligned}
D_{GSL}(\boldsymbol{f}||\boldsymbol{f}') &= \sum_{i=1}^{L} w_i \left(-2 \sum_{s \in S_i} m_i(s) \log_{a_l} m_i(s) + \sum_{s \in S_i} f_i'(s) \log_{a_l} f_i'(s) \right) \\
&= \sum_{i=1}^{L} w_i \left(2H^{S_i}(\boldsymbol{m}_i) - H^{S_i}(\boldsymbol{f}_i') \right), \qquad (7)
\end{aligned}
$$

where the symbol $H^{S_i}(\cdot)$ indicates the Shannon entropy of a distribution over the state space S_i.

On the right hand side of the first line of (7) the big square brackets contain the subtracted L-divergence computed at different block lengths l. In particular we take the *L-div* ([43]) and subtract the entropy for the frequency vector corresponding to $x^s(t)$. This can be justified in two ways. On the one hand it is due to the fact that $x^s(t)$ is always taken to be the real-world time series and it can be observed only once. This means it is not possible to replicate this series and create an ensemble, as it will be done for the time series produced by models. As a consequence, it cannot be corrected for the systematic bias stemming from the fact that its entropy is computed using an estimator (the frequency over the state space) and not the true probabilistic structure (see [7] and [38]). On the other hand, being the *GSL-div* always applied to real data against models' output, when one compares the distance between simulated output from different models and the real counterpart, the entropy of the latter is always washed away.

The logarithm is always in base a_l with $i = 1, ..., L$, which corresponds to the cardinality of the alphabet available at length $l = i$. It is worth recalling that, in equation (7), $\mathbf{m}_i$ indicates the *mean* distribution between $\mathbf{f}_i$ and $\mathbf{f}_i'$:

$$
m_i(s) = \frac{f_i(s) + f_i'(s)}{2}, \quad \forall\, s \in S_l \qquad (8)
$$

where $f_i(s)$ is the frequency such that symbol s appears in the first series, $x^s(t)$, while f_i' is the counterpart for the second series, $y^s(t)$.

2.3 Ensembles of Runs, *GSL-div* and Probabilities

Previous section discussed the *GSL-div* between distributions obtained from two series. Now, we extend it to accomplish our main task: assess the similarity in the dynamics expressed by a model (and not a single run) and that of real data. In principle one might want to estimate the distance between the data and her own model relying on the probabilistic structure of the latter. For example, using the *GSL-div*, one would like to fed it with the *true* probability that the model assign to each sequence of symbols rather than its frequency. However, in many cases this is difficult or even not possible. Using micro-founded models like ABM, for instance, the statistical structure of the aggregate variables is not known in advance and *emerge* from the

interactions of an ecology of heterogeneous agents ([5, 61]). The only information about the stochastic processes underlying the aggregate behaviour of these models is available through the synthetic series they produce. Along this argument, [33] agues that investigating the properties of these processes is essential in order to adequately understand both the model and the real system it is intended to represent[8]. Moreover, it underlines that quantitative analysis of the artificial data helps to acquire such knowledge.

In this section we show that averaging the *GSL-div* over an ensemble of independent runs of the same calibrated model and correcting for a systematic bias stemming from Jensen inequality, we are able to infer the distance between the distribution of time changes observed in the data and the *true* distribution generated by the probabilistic structure of the model, which is unknown *a priori*. To do this we introduce a specific probability function. Its main feature is that it avoids the problem of exact time correspondence between the simulation and the data: these probabilities are called *time-average probabilities*. Let $\mathcal{M}$ be the set of all models that simulate the (real valued) time series $x(t)$.

Definition 2.3 *(Time-average probability) Consider a model $\mu \in \mathcal{M}$ and the symbolized time series $\{x_t^s\}_{t=1}^T$ taking values in the alphabet $S_{l,b}$. Assume that according to model μ the unconditional probability of each symbol $s \in S_l$ is allowed to vary over time: $p_\mu(x_t^s = s \,|\, t = 1) \neq p_\mu(x_t^s = s \,|\, t = 2) \neq ... \neq p_\mu(x_t^s = s \,|\, t = T)$. The time-average probability assigned by μ to each $s \in S_l$ is defined as*

$$\overline{p}_\mu(s) = \frac{1}{T} \sum_{t=1}^{T} p_\mu(x_t^s = s). \tag{9}$$

Remark 1 *Let $x(t)$ be the output of model μ and assume it is (strongly) stationary, then*

$$\overline{p}_\mu(s) = p(x^s = s).$$

It is easy to see that remark 1 follows immediately from the definition of (strong) stationarity. Time-average probabilities are particularly well suited for our purposes. They express the average probability of observing a given symbol in the time interval $[1; T]$. This implies that comparing time-average probabilities with the distribution of symbols observed in real data, by means of the *GSL-div*, we are measuring the similarity of the behaviours predicted by the model and observed in the data, without taking into account whether they are synchronous or not. Indeed, this is exactly what we would request to assess the empirical validity of a set models.

[8] In particular [33] focuses on the analysis of stationarity and ergodicity of simulated data and offers an algorithm to test these properties.

Now let us introduce a proposition showing that frequencies of symbols observed in a time series are good estimates of time average probabilities of a model μ when a sufficiently large ensemble is considered. Let $f_\mu(s)$ be the frequency of symbol s observed in a given run of μ.

Proposition 2.4 *Consider an ensemble of R independent runs of μ, then $\frac{1}{R}\sum_r f_{\mu,r}(s)$ is an unbiased and consistent estimator of $\overline{p}_\mu(s)$ for all $s \in S_{l,b}$.*

Proof of proposition (2.4) can be found in Appendix A.
Despite our estimator is a reasonable one, if we were to use it to compute the *GSL-div* we would underestimate the distance between model μ and the data. This is due to the systematic bias in the computation of information theoretic quantities that stems from Jensen inequality (see next subsection for a complete discussion). As showed in [54], this bias is relatively large when time series are short, which might well be the case in many practical applications[9]. However, thanks to the possibility of correcting for the bias, we can obtain a nice estimate of the *GSL-div* between the unknown probabilistic structure of the model and the distribution of observed time changes. In particular, let $x(t)$ be the real world time series and $p(s)$ be frequency of symbol s observed in $x(t)$. Moreover, consider an ensemble of independent runs of size R and let $f(s)_r$ be the frequency of s in the r^{th} run. Finally assume that each series have been symbolized with equal precision b.

Proposition 2.5 *The GSL-div between $\overline{p}_\mu$ and $p(s)$ is given by*

$$\begin{aligned} GSL(p(s)\,||\,\overline{p}_\mu(s)) &= \sum_{i=1}^{L} w_i\,\mathbf{E}\left(-2\sum_{s\in \mathrm{S}_i} m_i(s)\log_{a_i} m_i(s) + \sum_{s\in \mathrm{S}_i} f(s)\log_{a_i} f(s)\right) \\ &+ \sum_{i=1}^{L} w_i\left(\frac{B_i^m - C^m}{4T_i} - \frac{B_i^{\overline{p}_\mu} - C^{\overline{p}_\mu}}{2T_i}\right) + O(T^{-2}). \end{aligned} \tag{10}$$

where $\mathbf{E}(\cdot)$ is the expectation over the ensemble, w_i are arbitrary positive weights such that $\sum_i w_i = 1$, B^j is the cardinality of the support of $j = \{m, \overline{p}_\mu\}$ and $B^j \geq C^j \geq 1$.

The proof of proposition (2.5) follows straightforwardly from what is contained in next subsection and Appendix B. Some attention is deserved by the term C^j; in Appendix B. a direct computation and a more detailed discussion is offered; here we recall that for practical applications we set $C^j = 1$ (see also next subsection). Proposition (2.5) constitutes one of the main results of the paper: it shows that simulating an ensemble of independent runs of model μ and correcting for the systematic bias up to the second order, it is possible to obtain the

[9] In general, in economics or econometric studies it is difficult to deal with series lasting more than 500 periods, which is the threshold below which [54] claims biases are not negligible. For instance, think about macroeconomic issues or technology diffusion problems.

GSL-div between the unknown *true* distribution of time changes produced by a model and that observed in the real world series.

There is also an additional reason that makes this approach particularly well suited for validation. In particular, the use of a sufficiently large ensemble of runs to validate each model allow to capture the overall degree of similarity between data and the model, washing away run-specific effects. This is true in particular dealing with agent based and complex systems approaches: they deal with chaotic dynamics, abrupt changes and tipping points. The system they model is often out of equilibrium and governed by stochastic shocks and feedbacks loops. The consequence is that one run of the model might be completely different from the others[10]. Therefore, to adequately explore its behaviour, a relatively large number of runs have to be considered and, when it comes to validate the model [11], different runs should ideally exhibit relatively similar dynamics.

2.3.1 The systematic bias and its correction

When an information theoretic function is computed without knowing the exact probability of each symbol, a systematic error might arise. In particular this the case when the true probabilistic structure of a process has to be estimated from a finite sequence of observations (see [7, 38, 31, 32, 57]). Even knowing the true distribution $\mathbf{p}$ of a time series $x(t)$ over some state space S and knowing it is strongly stationary, when one computes any of the *KL-div*, *JS-div*, *L-div* or *GSL-div* between $\mathbf{p}$ and $\mathbf{f}$ estimated from $\{x(t)\}_{t=1}^{T}$ with $T < \infty$, the result would be larger than zero. Obviously, the bias is also present when computing the distance between two frequency vectors that are estimated from two realizations of the same stochastic process.

The concept of systematic bias for the numerical values of information theoretic functional is well known in the literature and follows directly from Jensen inequality (see [13]). In particular, if g is a concave function, the bias is identified with the expectation value $E[g(\mathbf{f})]$ being lower than $g(\mathbf{p})$, where $\mathbf{f}$ is an estimator of the true probability distribution $\mathbf{p}$. Applying this result to the Shannon entropy one obtains

$$H(\mathbf{p}) = H(E(\mathbf{f})) \geq E[H(\mathbf{f})], \tag{11}$$

where the expectation is defined over an ensemble of independent finite-length sequences generated by the probability distribution $\mathbf{p}$.
Following [54] it can be shown that the expected value of the observed entropy is systematically

[10] See the discussion in section 3.3 of [51].
[11] After having conveniently calibrated and/or estimated it.

biased downwards from the true entropy and that a correction term can be found:

$$E[H(\mathbf{f})] = H(\mathbf{p}) - \frac{B-1}{2T} + O(T^{-2}), \tag{12}$$

where T is the length of each time series and B is the number of states $s \in S$ such that $p(s) > 0$. This result was originally obtained by Basharin ([7]) and Herzel ([38]) who also found also that, up to the first order $O(T^{-1})$, the bias is independent of the actual distribution $\mathbf{p}$. However, as showed in appendix B, this result does not hold when probabilities are allowed to vary over time and we are interested in their average: in this case the correction term will be lower than with time invariant processes. The term $O(T^{-2})$ contains unknown probabilities $\mathbf{p}$ and cannot be estimated in general (see [31, 32, 54]). It is to be noticed that the estimation of B is, in general, non-trivial [50]. In [32, 54, 15] the number of states in the support of $\mathbf{p}$ are approximated by the number of states occurring with non-zero frequency. However this approach could be totally misleading, in particular if the series are non-stationary. [64] shows how to compute any function of the true probability distribution starting with a limited number of samples relying on Bayes rule. In our case the problem could be solved more easily: as shown by proposition 2.4, we are able to obtain nice estimates of the underlying time average probabilities; therefore, we can use the cardinality of the support of our estimates as a reasonable value for B. A more detailed discussion and the analytical computation of the correction term is contained in Appendix B[12]. In what remains we discuss aggregation weights, some interesting property of the *GSL-div* and provide some examples of its ability to detect similarities in time series dynamics.

2.4 Aggregation weights

As introduced above, each of the subtracted L-divergences entering the *GSL-div* is assigned an increasing weight. This reflects the grater importance assigned to the ability of the simulated data to match the behaviour of the real process over a longer time-window and, additionally, it compensates for the increasing value of the logarithms' base a_l. Since the only condition we require is that weights normalize to 1, an infinite number of increasing weights could be chosen a priori for the aggregation of single L-divergences. In this paper we consider and discuss three different sets of weights: *additively progressive weights*, such that the difference between successive weights is positive and constant, *geometrically progressive weights*, such that the ratio between successive weights is positive and constant and *uniform weights*, which assign each L-divergence the same weight.

In particular, *additively progressive weights* are chosen to guarantee that their first differences are constant; that is, the weight assigned at a given length of the time window is equal to the one assigned at the previous length plus a constant term. More formally, they satisfy:

[12]One can notice that there is an alternative way to compute the correction term ([[31]); however the leading $O(1/N)$ term is exactly the same as in equation 12

a) $\sum_{l=1}^{L} w_l = 1$

b) $w_0 = 0$

c) $w_{l+1} - w_l = k, \quad k \in \mathcal{R}^+$.

The following proposition provides the unique formulation of *additively progressive weights*.

Proposition 2.6 *The only set of weights* $\{w_l\}_{l=1}^{L}$ *satisfying a), b) and c) is given by*

$$w_{l+1} = w_l + \frac{2}{L(L-1)}. \tag{13}$$

The proof is included in Appendix C. An alternative specification is provided by *geometrically progressive weights*. They are defined by the same set of properties as additively progressive ones but for c), which is substituted by

d) $\frac{w_{l+1}}{w_l} = c, \quad \forall\, l \geq 1 \quad \text{with}\, c > 1.$

Geometrically progressive weights are not unique, in the sense that they depend on the choice of c. However they allow for an interesting property of the *GSL-div* that will be characterized in the next section. The idea behind the use of this set of aggregation weight is that there is a constant rate of growth in the importance assigned to differences in the distributions of time changes when the length of blocks increases.

Finally, a benchmark choice is considered: *uniform weights* assign the same importance to each L-div in the aggregation process, independently of the length of blocks. It is immediate to see that they satisfy a), b) and

e) $w_{i+1} = w_i, \quad \forall\, l \geq 1.$

The choice of the set of weights to use relies upon the researcher. However, our results will show that the *GSL-div* is robust to this choice: when more than one model for the same phenomenon is available, models' ordering is largely independent from aggregation weights. However some remarks applies. On one side, the ability to reproduce time changes over longer windows should be rewarded. Consider for example a set of models addressing some macroeconomic issues; when $l = 1$ the *GSL-div* captures the difference in models' ability to reproduce the persistence of observable data on single states (e.g. low, moderate or high inflation). Passing to $l = 2$, the *GSL-div* aggregates the outcome of models' evaluation at $l = 1$ and their ability to reproduce the distribution of shifts from one regime of inflation to the other (e.g from low to moderate inflation). When $l = 3$ it extends the aggregation to longer patterns (e.g. from low inflation to moderate and back to low). Having the goal of validating models against observable data it seems natural to give more importance to the ability of reproducing larger portions of series'

dynamics: trends and trajectories are more relevant than persistence. There is an additional reason progressive (i.e. increasing) weights should be preferred. It is simple but more technical. In order to aggregate single L-divergences into the *GSL-div* we use logarithms' bases in (7) that depends on the length, l, of the blocks. Having l growing, it also increases the cardinality of the state space, which corresponds exactly to the base of the logarithm we use. Being the logarithm decreasing in its base, this implies that the *L-div* between two distributions decreases with l as well (see property 6. in next section). The use of progressive weights compensate for this effect.

3 Properties of the *GSL-div*

The *GSL-div* exhibits a set of interesting properties that make it particularly well suited to compare empirical validity of a set $\mathcal{M}$ of models. For illustrative purposes, let us simplify our notation by removing some dependences: p_μ indicates time average probabilities for model $\mu \in \mathcal{M}$ and p indicates frequencies observed in the real data. As usual, b is the precision of the symbolization and l the length of time-windows or, equivalently, the length of symbols. Proofs of all properties is contained in Appendix D.

1. The *GSL-div* is well defined for all p and $\overline{p}_\mu$.

This property guarantees that for all probability vectors the *GSL-div* between pairs of them exists and can be computed independently of their support, which might be the same or not. This is a direct advantage with respect to the *KL-div* and comes from the definition of the *L-div*.

2. $0 < GSL(p \,||\, \overline{p}_\mu) < 2$

The *GSL-div* is bounded both from above and below. This is relevant to the purposes of model selection and validation since makes it possible to compare the estimated value with bounds. However, to have a meaningful comparison, bounds have to be explored. The lower one is only theoretical: since we consider a subtracted version of the *L-div* the effective lower bound corresponds to the entropy of the data generating process, but this will be more evident discussing next property. The upper bound, instead, is reached when the model and the real series are persistently stacked in two different states (corresponding to different symbols). For example, from a practical point of view, a model producing only hyperinflation is maximally distant from an observed time series of persistent deflation.

3. $GSL(p \,||\, \overline{p}_\mu) = H(p) \iff p = \overline{p}_\mu \;\; \forall\, l = 1, ..., L \text{ and } s \in S_{l,b}$

This property indicates the effective lower bound for the *GSL-div*: it is equal to the entropy of the real time series if and only if for every block length model μ assigns each symbol the same

probability the observed time series does. This would be the case of perfect matching between the dynamics of the simulated series and the data, whose patterns are mirrored exactly in all model's runs[13].

4. The non-subtracted version of *GSL-div* is symmetric:

$$GL(p \,||\, \overline{p}_\mu) \,=\, GL(\overline{p}_\mu \,||\, p). \tag{14}$$

This property stems directly from the fact that the L-div is symmetric. It refers to a version of the *GSL-div* where we do not subtract the entropy of the real series. In practice one could use this measure as well; implicitly, it would amount to assume that, with reference to the real data, the ensemble is composed by identical series corresponding exactly to the observed one. This assumption might be justified by acknowledging that when it comes to observe the world, it can observed only once. As a matter of fact, the real series is the only source of information we can compare models with.

5. For all $\{\overline{p}_{\mu,i}\}_{i=1}^N$ and $\{\pi_i\}_{i=1}^N$ such that $\sum_i \pi_i = 1$,

$$GSL(p_i \,||\, \sum_i \pi_i \overline{p}_{\mu,i}) \,\geq\, \sum_i \pi_i GSL(p \,||\, \overline{p}_{\mu,i}). \tag{15}$$

In words, given a set of $N+1$ probability vectors, the *GSL-div* between the first and any convex combination of the remaining N is greater than the convex combination of the *GSL-div*s between the first and each of the others. This property is relevant when the modeller is uncertain about the parametrization of the model. It could be the case, for example, when experts are asked to express their preferences and views about a relevant parameter[14] Let $\{\pi_i\}_{i=1}^N$ be the *belief* vector assigning probabilities to each of the N possible parametrizations of the model. From the point of view of the modeller, the expected probabilistic structure should be $\sum_{i=1}^N \pi_i \overline{p}_{\mu,i}$ where $\overline{p}_{\mu,i}$ corresponds to the set of parameters $i \in \{1, .., N\}$. Property 4. states that the average distance between data and the each configuration of the parameters of the model constitutes a lower bound for the distance between the expected probabilistic structure and the observed data.

6. Consider two distributions which are unaffected by a marginal change in b or l, then the *GSL-div* between the two is decreasing both in the precision of the symbolization (b) and the length of blocks (l).

This is a direct consequence of the particular basis of the logarithm we use in definition 2.2. This choice is necessary to aggregate comparable quantities: it guarantees that estimated entropies

[13] Accounting for the fact the movements might not be synchronous up to a delay of L.
[14] Think about climate sensitivity in climate change models.

fall within the $[0;1]$ interval independently from symbolization precision and block length. Moreover, property 6. helps justify the use of aggregation weights, w_i, that are increasing in l.

7. If $x(t)$ and $y(t)$ are strongly stationary and the *GSL-div* is evaluated with geometrically progressive weights such that $c = f(L)$ with $f'(L) > 0$ and logarithms are in base 2, then the quantity $\frac{1}{T}$*GSL-div* converges to a subtracted version of the asymptotic L-div.

When two series are strongly stationary, the *GSL-div* between the two computed with a particular class of geometrically progressive weights (for instance with $c = L$) converges to the asymptotic value of the *L-div* when both T and L approaches infinity. It is relevant to recall that, as almost each information theoretic quantity (see [13]), even the *L-div* has been defined using logarithms in base 2. Given a series, the asymptotic entropy is defined as the limit of the ratio between the entropy of the sample and the length of the series[15]: it expresses the average information contained in a single random variable of the process $x(t)$. Property 6. highlights the bridge linking the *GSL-div* between distributions (definition 2.2) and the famous Akaike Information Theoretic Criterion (Akaike, [1]). The AIC is derived as an asymptotically unbiased estimator of a function used to rank candidate models that is a variant of the Kullback-Leibler divergence between the true model and the approximating candidate model. When stationarity requirements are met, the *GSL* is an asymptotically consistent estimator of a symmetric version of the *KL-div* between the symbolized series.

4 A simple example

In this section we show the performance and the precision of the *GSL-div* criterion in distinguishing between three ad hoc created time series. $x(t)$ is chosen to be the observed series while $x_A(t)$ and $x_B(t)$ are to be intended as the output of two models (A and B respectively) trying to simulate $x(t)$. These series are consciously chosen to have $x_A(t)$ much more close to the behaviour of $x(t)$ with respect to $x_B(t)$. Their plot is reported in Figure 2.

Clearly, we expect the *GSL-div* criterion to show a lower distance between the observed time series coming from the unknown data generating process and model A's output. Before showing the results we present the symbolization process. The three series take values in the real interval $[0, 1]$ and a very small sample consisting of six observations is chosen:

$$x(t) = \{0.2; 0.3; 0.8; 0.4; 0.45; 0.15\},$$

$$x_A(t) = \{0.1; 0.25; 0.72; 0.45; 0.5; 0.35\},$$

$$x_B(t) = \{0.05; 0.15; 0.65; 0.9; 0.4; 0, 25\}.$$

[15] More precisely, $H_\infty = \lim_{T\to\infty} H[(x_T, x_{T-1}, ..., x_1)]/T$, with logarithms in base 2

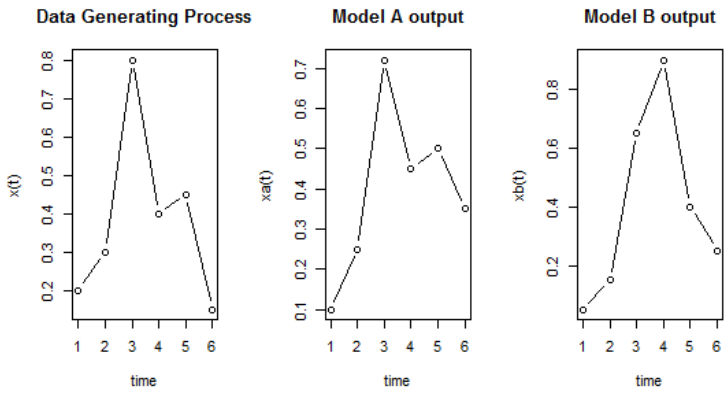

Figure 2: Behaviour of three selected time series

The precision of the symbolization is set to $b = 3$; this choice leads to the following partition of the original state space: $[0; 0.33); [0.33; 0.66); [0.66; 1]$. Despite the choice of b is arbitrary results are robust to changes in the value of this parameter (see next section for a complete robustness exercise). The use of a low value of b can be justified here by the fact that the time series are very short, which implies there are few observations to estimate frequency vectors; in addition, representing it the the precision of the symbolization process, the use of a low value for b makes it more difficult to distinguish between the series. The ability of the *GSL-div* to recognize the most similar even when the symbolization is relatively imprecise would confirm the power of this new criterion.

According to the chosen parametrization, the three symbolized time series are:

$$x(t) = \{1, 1, 3, 2, 2, 1\},$$

$$x_A(t) = \{1, 1, 3, 2, 2, 2\},$$

$$x_B(t) = \{1, 1, 2, 3, 2, 1\}.$$

By inspection it is possible to notice that x_A is much more closer to x than x_B: while the former exhibits the same behaviour of the real data apart form the very last period, the latter displays twice the opposite one (it increases from $t = 3$ to $t = 4$ when $x(t)$ is decreasing and vice-versa in the following period).

Given the use of short time series, the maximum value of the time-window's length along which the three processes are compared cannot be set above $L = 5$; otherwise one and only one block would be available, the probability distribution over the alphabet would appear constant and its entropy pushed to zero. We set $L = 3$; this is, again, a conservative choice: direct inspection reveals that when $l = 4$ is chosen only $x_A(t)$ exhibits sequences of symbols retrievable in $x(t)$ and when $l = 5$ both $x_A(t)$ and $x_B(t)$. Hence the choice of $L \leq 3$ would further increase

the discrepancy between model A and B. Therefore, blocks and corresponding alphabets for $l = 1, 2, 3$ are analysed.. Respectively, six, five and four observations are obtained and used to estimate the frequencies. As it is obvious, these are very rough estimates of the probabilities the three process assign to symbols $x \in S_{l,b}$.

Notwithstanding this limitation, the performance of the *GSL-div* in selecting model A and validating its output against real data is excellent. Table 1 (with additively progressive weights) and Table 2 (with uniform weights) provide evidence of this result. Two observations deserve

			Subtracted L-div	
	block length	weights	model A	model B
	1	0.17	0.951167	0.920620
	2	0.33	0.894378	1.151525
	3	0.50	0.609108	1.174573
GSL-div			**0.761397**	**1.123795**

Table 1: GSL-div for $x(t)$ and both $x_A(t)$ and $x_B(t)$ with progressive weights

			Subtracted L-div	
	block length	weights	model A	model B
	1	0.33	0.951167	0.920620
	2	0.33	0.894378	1.151525
	3	0.33	0.609108	1.174573
GSL-div			**0.818218**	**1.082239**

Table 2: GSL-div for $x(t)$ and both $x_A(t)$ and $x_B(t)$ with uniform weights

attention. First, the subtracted L-divergence at blocks' length equal to one is lower for model B's than for model A's series. This is driven by the fact that $x_B(t)$ and $x(t)$ have been chosen to exhibit the same frequency distribution over the alphabet available for $l = 1$, $S_1 = \{1, 2, 3\}$, while $x_A(t)$ has not. This means that it becomes relatively more difficult to recognise $x_A(t)$ as the series most similar to $x(t)$. However, the distribution of time-changes is completely different between $x(t)$ and $x_B(t)$. The result is that when one move to $l = 2, 3$, corresponding to capture longer trends and trajectories, $x_A(t)$ equals and overcome $x_B(t)$'s performance in simulating the behaviour of $x(t)$. In addition, this justifies the choice of using progressive weights in the definition of the *GSL-div*: a model matching the distribution of changes for a longer time window should always be preferred and selected.

Secondly, the three time series have been selected ad hoc to show the performance of the *GSL-div*. Not having a proper model it is not possible to replicate simulations and correct for the systematic bias[16]. This implies the use of definition 2.2.

In the next section we move away from this example and we show the precision of the *GSL-*

[16]The only meaningless solution would be assuming deterministic models producing always the same realization.

div in validating and selecting the most appropriate among 9 univariate stochastic models; the correction term for the systematic bias is added to the estimation of the criterion.

5 Selecting and Validating ARMA models

This section presents the application of the *GSL-div* and shows its power. The *GSL-div* is used to compare a set of models with the data. The choice of which models to consider is of primary importance here. We wanted to show the ability of our criterion to discriminate between models exhibiting very similar behaviours, but having a different underlying stochastic structure. In addition, we looked for something simple, easily tractable and not computational intensive. These requirements prevented the use of ABMs, even of simple ones.[17] and pushed us to the choice of Auto Regressive Moving Average (ARMA) models with Gaussian innovations[18]. Despite being extremely simple, this class of models can generate a variety of very different behaviours, ranging from random walks to stationary patterns to "explosive" dynamics. In addition, their response to small changes in the parametrization could be deduced in advance.

The Data Generating Process (DGP) is selected to be a Gaussian $AR(1)$ process with autoregressive order-one parameter $\phi_1 = 0.1$. A realization of length $T = 1000$ is taken to be the real-world data. Figure 3 provides a plot of this time series. Trying to produce a similar behaviour, a set of nine ARMA models with a parametric structure very similar to the DGP has been chosen. The *GSL-div* is then used to select the model which minimizes the distance with respect to the distribution of time changes in the real data.

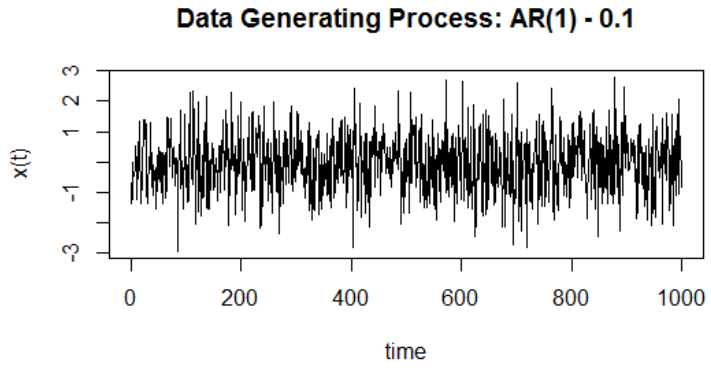

Figure 3: The real-world time series

Table 3 summarizes the main features of the models which are considered for replicating the

[17] Even in simple ABMs it is relatively easy to have bifurcations and chaotic behaviours (see for example some models in this review [56]). In addition, the stochastic processes generating aggregate outcomes are not known in general.

[18] See [37] for details.

behaviour of the real data. All of them are ARMAs with Gaussian $N(0,1)$ innovations and are used to produce an ensemble of $M = 1000$ Monte Carlo replications, each of length $T = 1000$. These series are symbolized using precision $b = 5$.

		parameters		**properties**	
	model	ϕ	θ	stationary	invertible
1	AR(1)	0.1	0	yes	yes
2	AR(1)	0.2	0	yes	yes
3	AR(1)	0.5	0	yes	yes
4	AR(1)	0.01	0	yes	yes
5	AR(1)	0.9	0	yes	yes
6	ARMA(1,1)	0.2	0.9	yes	yes
7	ARMA(1,1)	0.5	2	yes	no
8	AR(1)	1	0	no	yes
9	AR(1)	2	0	no	yes

Table 3: Main features of the nine models considered

The majority of the models we consider are stationary and, even if not reported directly, they are also causal. In addition, most of them are invertible. Together with the use of Gaussian random errors, this allows to conclude that six out of nine are unique, meaning that there is a one-to-one correspondence between the family of the finite dimensional distributions of the process and its finite parametric representation (see [37][19]). The same applies also to the DGP. A direct consequence is that, playing with values assigned to parameters, we obtain very similar patterns of behaviour (it can be seen by inspection, see top and middle panels in Figure 4) coming from surely different stochastic processes.

The *GSL-div* is expected to recognize the model which is most similar to the *DGP*: model 1 exhibits exactly the same parametric representation of the data generating process from which $x(t)$ is taken. In addition one should ask the *GSL-div* to identify models producing series completely inconsistent with the real world data $x(t)$: model 9 is strongly non-stationary and exhibits an explosive behaviour. Therefore, within the class of models considered here, we expect the *GSL-div* to reach a minimum when model 1 is evaluated and a maximum when model 9 is compared to observed data. Figures 4 provides a plot of a realization of model 1 (top) and 9 (bottom).

Results of the estimation of the distance between data and models are collected in Tables 4 and 5, which show the performance of the *GSL-div* after correcting for the systematic bias. The maximum length of the time-window (or block-length), L, is chosen to be six. Both additively progressive and uniform weights are considered[20]

[19]Here you find both definitions of causal processes, invertible processes and conditions for the one-to-one correspondence between a finite dimensional distributions of the process and the finite parametric representation.

[20]We omit geometrically progressive weights since they are not unique: they depend on an additional free parameter that has to be chosen by the modeller.

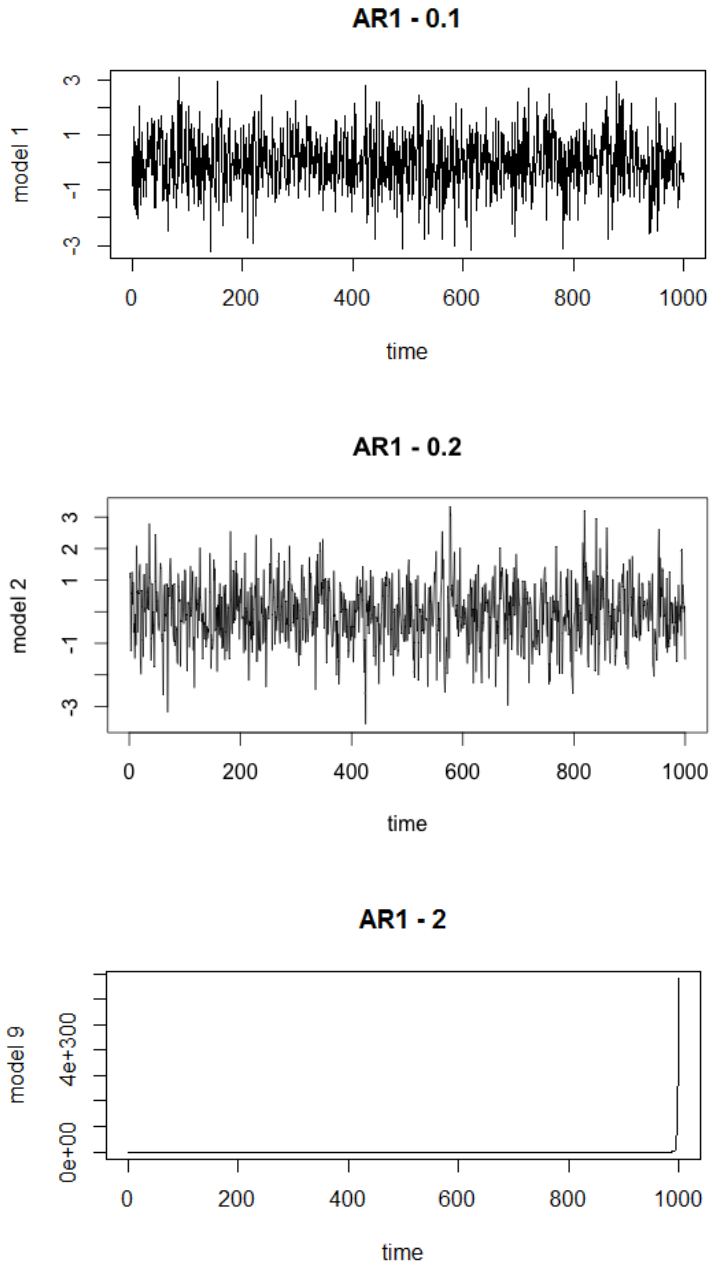

Figure 4: A realization of model 1 (top), model 2 (middle), model 9 (bottom).

		Subtracted L-div times the corresponding weight								
block length	*weights*	*AR1 - 0.1*	*AR1 - 0.2*	*AR1 - 0.01*	*AR1 - 0.5*	*ARMA11 - 0.2&0.9*	*ARMA11 - 0.5&2*	*AR1 - 0.9*	*AR1 - 1*	*AR1 - 2*
1	0.05	0.739961	0.740193	0.739479	0.740022	0.740705	0.741579	0.746528	0.847038	1.519963
2	0.10	0.737830	0.737918	0.738141	0.746422	0.753683	0.766014	0.802260	0.912441	1.156505
3	0.14	0.736086	0.736167	0.736762	0.745997	0.759310	0.770605	0.811238	0.915131	1.033641
4	0.19	0.732324	0.732547	0.733376	0.740736	0.758631	0.770255	0.813153	0.924642	1.005211
5	0.24	0.721106	0.721464	0.722128	0.726598	0.752120	0.763155	0.813535	0.943901	1.013572
6	0.29	0.702880	0.703000	0.703406	0.718427	0.774408	0.782027	0.875425	0.988041	1.054068
GSL-div		**0.722666**	**0.722859**	**0.723363**	**0.732255**	**0.760361**	**0.770209**	**0.826552**	**0.941125**	**1.064143**

Table 4: *GSL-div* for $x(t)$ and nine ARMA models with progressive weights

		Subtracted L-div times the correspondig weight								
block length	*weights*	*AR1 - 0.1*	*AR1 - 0.2*	*AR1 - 0.01*	*AR1 - 0.5*	*ARMA11 - 0.2&0.9*	*ARMA11 - 0.5&2*	*AR1 - 0.9*	*AR1 - 1*	*AR1 - 2*
1	0.17	0.739961	0.740193	0.739479	0.740022	0.740705	0.741579	0.746528	0.847038	1.519963
2	0.17	0.737830	0.737918	0.738141	0.746422	0.753683	0.766014	0.802260	0.912441	1.156505
3	0.17	0.736086	0.736167	0.736762	0.745997	0.759310	0.770605	0.811238	0.915131	1.033641
4	0.17	0.732324	0.732547	0.733376	0.740736	0.758631	0.770255	0.813153	0.924642	1.005211
5	0.17	0.721106	0.721464	0.722128	0.726598	0.752120	0.763155	0.813535	0.943901	1.013572
6	0.17	0.702880	0.703000	0.703406	0.718427	0.774408	0.782027	0.875425	0.988041	1.054068
GSL-div		**0.742932**	**0.743119**	**0.743460**	**0.751094**	**0.771606**	**0.780918**	**0.826564**	**0.940303**	**1.153103**

Table 5: *GSL-div* for $x(t)$ and nine ARMA models with uniform weights

Expectations are perfectly confirmed: model 1 turns out be the closest to the real data while model 9 the most distant. In particular, the chain of models' proximity to the data identified by the *GSL-div* reveals:

$$\begin{aligned} AR(0.1) \succ\ & AR(0.2) \succ AR(0.01) \succ AR(0.5) \succ ARMA(0.2, 0.9) \succ ARMA(0.5, 2) \succ \\ \succ\ & AR(0.9) \succ AR(1) \succ AR(2) \end{aligned}$$

where numbers in parenthesis indicate parameters' values and the expression $x \succ y$ means that model x is preferred to y given its closeness to the data. In general, the *GSL-div* is shown to distinguish clearly among models: non-stationary processes are the most distant from the real data and when a MA component is added to the process the distance from the real data increases. This is true especially when the MA part is non-invertible (model 7). Moreover, among the same class of processes (AR(1)) the criterion is able to recognize those having a parametric representation which is closer to the *DGP*.

It is worth noticing that such results are robust to the choice of the weights in the functional representation of the *GSL-div*. Finally, the correction term for the systematic bias is, in absolute value, considerably low with respect to the estimated value of the *GSL-div* criterion, and it becomes even smaller the longer the time series. In particular, the correction never affects results and the ranking of models' distance from the real world data. (see appendix B).

5.1 Robustness Checks

In this subsection we briefly present a robustness analysis for the previous exercise. It shows that models' ranking is largely independent from the choice of the free parameters in the *GSL-div* estimation. Precisely, there are only two free parameters to be set by the modeller: b, the precision of the symbolization process and L, the maximum length of time windows used to compare the series.

We performed the same exercise as the one above by changing the values of both b and L. In particular, we test how many of the binary relations identified by the estimated *GSL-div* between models and data are robust to changes in the two parameters. Let $\mathcal{M}$ be a set of J competing models. Then, each total order (that is, the ranking of models with respect to the data) established by the *GSL-div* on $\mathcal{M}$ entails $\binom{J}{2}$ relations. Since we have 9 ARMA models there are 36 binary relations. Table 6. reports the total number of changes in the set of relations between models due to a change in the value assigned to free parameters. Results are the same both with additively progressive and uniform weights.

Whatever the chosen couple (b, L), the overall number of changes is extremely low compared to the total number of unaltered relations and, in general, it never exceeds three. It also appears clear that, increasing either the precision of the symbolization, the maximum length of time-windows or both, the number of changes in the orderings is almost always decreasing and reaches

Table 6: Overall number of changed binary relations

b\L	**3**	**4**	**5**	**6**	**7**
3	2	2	2	1*	2
4	3	2	2	0*	0*
5	0*	0*	0*	-	0*
6	0*	0*	0*	0*	0*

zero even for low parameters values. Symbol $*$ indicates that using the corresponding values of b and L model 1 is the closest to the data, models 8 and 9 are the most distant and adding a Moving Average component to the process always increases the estimate of the *GSL-div*. Furthermore, it is worth reporting that models 5,8 and 9 are the most distant from the real series (specifically, this is the right tail of the chain: mod5 $\succ$ mod8 $\succ$ mod9) in each of the 20 experiments we carried out using different values of the two free parameters. The robustness of our criterion and its ability to provide good results even with relatively low precisions is remarkable. Increasing b and L leads to an explosion of the cardinality of the potential alphabet of symbols which, in turns, requires much more computational time to machines to estimate frequencies, especially in presence large ensembles, which, to the contrary, are needed for guaranteeing consistency of probability estimates. Noteworthy, our *GSL-div* indicator is very precise even with relatively low values of both b and L and, in addition, the informational gains obtained increasing these values saturates quickly. This means that the user can save a great amount of computational time by specifying low b and L and, at the same time, be confident in the results delivered by the *GSL-div*.

6 Conclusions and Further Research

Validation of simulated models is still an open issue. One way of tackling this problem is via the identification of a measure of the distance between simulated and real-world data. This paper provides an information theoretic criterion, the *GSL-div*, which captures this distance without any requirement of stationarity nor the need to resort to any likelihood function. This constitutes a direct advantage with respect to other approaches aimed at characterizing times series and their behaviour. Our criterion applies to any model able to produce time series as output. In particular, simulation models and ABM especially, are likely to produce series whose statistical properties cannot be derived directly from models' assumptions. That is, the probabilistic behaviour of aggregated variables is unknown to the modeller. Using reasonably large ensembles and correcting for a classic systematic bias, the *GSL-div* allows to infer the distance between the unknown probabilistic dynamics of the model and observed data.

The *GSL-div* leaves two free parameters: the precision of the symbolization process, namely b, and the maximum length of the time-window used to identify blocks of the time series, namely

L. Both can be increased when the size of real time series against which models are evaluated becomes larger and larger; however, we showed that using relatively low parameters' values ($b = 5$, $l = 6$) the *GSL-div* is extremely precise in selecting and ordering the models which are better able to reproduce the distributions of time-changes observed in real data.

Different developments and applications of the *GSL-div* are possible. First of all, it represents an innovative tool to study the similarity of time series dynamics in general, and can be applied to single series, without generating ensembles and correcting for the systematic bias. For example, it could be employed to assess interdependences and contagion phenomena in financial markets (see [25]). In addition, extensions to multivariate settings are possible. In such a context applications are numberless, going from empirical validation of complex macro ABMs, to horse races among different models trying to reproduce the same observed patterns (e.g. housing market bubbles or crashes). Finally, it can be used for the comparison of model projections under policy interventions with baseline scenarios. A notable application within this domain refers to climate policy evaluation. Therein, the estimation of the *GSL-div* allows to answer the question of how far models' predictions about the evolution of climate variables move away from the baselines cases, under different policy interventions.

References

[1] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[2] Simone Alfarano, Thomas Lux, and Florian Wagner. Empirical validation of stochastic models of interacting agents. *The European Physical Journal B-Condensed Matter and Complex Systems*, 55(2):183–187, 2007.

[3] Simone Alfarano, Thomas Lux, and Friedrich Wagner. Estimation of agent-based models: the case of an asymmetric herding model. *Computational Economics*, 26(1):19–49, 2005.

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

[5] Philip W Anderson et al. More is different. *Science*, 177(4047):393–396, 1972.

[6] S. Barde. A practical, universal, information criterion over nth order markov processes. Kent Discussion Paper 15/04, 2015.

[7] G.P. Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Prob. App.*, 4:333–338, 1959.

[8] M. Basseville. Review: Divergence measures for statistical data processing-an annotated bibliography. *Signal Process.*, 93(4):621–633, April 2013.

[9] Carlo Bianchi, Pasquale Cirillo, Mauro Gallegati, and Pietro A Vagliasindi. Validating and calibrating agent-based models: a case study. *Computational Economics*, 30(3):245–264, 2007.

[10] Carlo Bianchi, Pasquale Cirillo, Mauro Gallegati, and Pietro A Vagliasindi. Validation in agent-based models: An investigation on the cats model. *Journal of Economic Behavior & Organization*, 67(3):947–964, 2008.

[11] Fabio Canova and Luca Sala. Back to square one: identification issues in dsge models. *Journal of Monetary Economics*, 56(4):431–449, 2009.

[12] Pasquale Cirillo and Mauro Gallegati. The empirical validation of an agent-based model. *Eastern Econ J*, 38(4):525–547, 09 2012.

[13] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons: Hoboken, New Jersey, 2006.

[14] G. Curato and F. Lillo. Modeling the coupled return-spread high frequency dynamics of large tick assets. *ArXiv e-prints*, October 2013.

[15] G. Curato and F. Lillo. Multiscale model selection for high-frequency financial data of a large tick stock by means of the jensenshannon metric. *Entropy*, 16(1):567–581, 2014.

[16] Herbert Dawid and Giorgio Fagiolo. Agent-based models for economic policy design: Introduction to the special issue. *Journal of Economic Behavior & Organization*, 67(2):351–354, 2008.

[17] Giovanni Dosi, Giorgio Fagiolo, Mauro Napoletano, and Andrea Roventini. Income distribution, credit and fiscal policies in an agent-based keynesian model. *Journal of Economic Dynamics and Control*, 37(8):1598–1625, 2013.

[18] Giovanni Dosi, Giorgio Fagiolo, Mauro Napoletano, Andrea Roventini, and Tania Treibich. Fiscal and monetary policies in complex evolving economies. *Journal of Economic Dynamics and Control*, 52(0):166 – 189, 2015.

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

[20] Darrell Duffie and Kenneth J Singleton. Simulated moments estimation of markov models of asset prices. *Econometrica*, 61(4):929–952, 1990.

[21] D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *Information Theory, IEEE Transactions on*, 49(7):1858–1860, July 2003.

[22] G. Fagiolo, C. Birchenhall, and P. Windrum. Empirical validation in agent-based models: Introduction to the special issue. *Computational Economics*, 30(3):189–194, 2007.

[23] G. Fagiolo, Jacob Leal, M. S. Napoletano, and A. Roventini. Rock around the Clock: An Agent-Based Model of Low- and High-Frequency Trading. LEM Papers Series 2014/03, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy, 2014.

[24] Giorgio Fagiolo and Andrea Roventini. Macroeconomic policy in dsge and agent-based models. *Revue de l'OFCE*, (5):67–116, 2012.

[25] Kristin J Forbes and Roberto Rigobon. No contagion, only interdependence: measuring stock market comovements. *The journal of Finance*, 57(5):2223–2261, 2002.

[26] M. Gallegati and M. Richiardi. *Agent based modelling in economics and complexity*, volume In: Meyer, R.A. (Ed.), Encyclopedia of Complexity and Sistem Science. Springer, New York, USA, 2009.

[27] Mauro Gallegati and Alan Kirman. Reconstructing economics. *Complexity Economics*, 1(1):5–31, 2012.

[28] Domenico Giannone, Lucrezia Reichlin, and Luca Sala. Vars, common factors and the empirical validation of equilibrium business cycle models. *Journal of Econometrics*, 132(1):257–279, 2006.

[29] Manfred Gilli and Peter Winker. A global optimization heuristic for estimating agent based models. *Computational Statistics & Data Analysis*, 42(3):299–312, 2003.

[30] Christian Gourieroux, Alain Monfort, and Eric Renault. Indirect inference. *Journal of applied econometrics*, 8(S1):S85–S118, 1993.

[31] P. Grassberger. Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A*, 128:369–373, 1988.

[32] P. Grassberger and T. Schurmann. Entropy estimation of symbol sequencences. *Chaos*, 6, 1996.

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

[34] Jakob Grazzini and Matteo Richiardi. Estimation of ergodic agent-based models by simulated minimum distance. *Journal of Economic Dynamics and Control*, 2014.

[35] Jakob Grazzini, Matteo Richiardi, and Lisa Sella. Small sample bias in msm estimation of agent-based models. In Andrea Teglio, Simone Alfarano, Eva Camacho-Cuena, and Miguel Gins-Vilar, editors, *Managing Market Complexity*, volume 662 of *Lecture Notes in Economics and Mathematical Systems*, pages 237–247. Springer Berlin Heidelberg, 2013.

[36] George Hall and John Rust. Simulated minimum distance estimation of a model of optimal commodity price speculation with endogenously sampled prices. *manuscript, Yale University*, 2003.

[37] J.D. Hamilton. *Time Series Analysis*. Princeton University Press: Princeton, New Jersey, 1994.

[38] H. Herzel, A.O. Schmitt, and W. Ebeling. Finite sample effects in sequence analysis. *Chaos, Solitons & Fractals*, 4(1):97 – 113, 1994. Chaos and Order in Symbolic Sequences.

[39] S. Kullback. *Information Theory and Statistics*. Dover Publications, New York, 1968.

[40] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.

[41] Bong-Soo Lee and Beth Fisher Ingram. Simulation estimation of time-series models. *Journal of Econometrics*, 47(2):197–205, 1991.

[42] Jessica Lin and Yuan Li. Finding structural similarity in time series data using bag-of-patterns representation. In *Scientific and Statistical Database Management*, pages 461–477. Springer, 2009.

[43] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37:145–151, 1991.

[44] C.M. Macan. On validating multi-agent system applications. In *14th International Workshop on MultiAgent-Based Simulation*, 2013.

[45] R.E. Marks. Validating simulation models: A general framework and four applied examples. *Computational Economics*, 30(3):265–290, 2007.

[46] R.E. Marks. Validation and model selection: Three similarity measures compared. *Complexity Economics*, 2(1), 2013.

[47] A.C.G. Mennucci and S.K. Mitter. *Probabilit e informazione*. Edizioni della Normale, 2008.

[48] Poudyal N. and Spanos A. Confronting theory with data: Model validation and dsge modeling. Working Paper, mimeo, 2013.

[49] A. Paccagnini. Model validation in the dsge approach: A survey. Working Paper, mimeo, 2009.

[50] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: comput. in Neur. Sys.*, 7:87–107, 1996.

[51] Andreas Pyka and Giorgio Fagiolo. 29 agent-based modelling: a methodology for neo-schumpeterian economics'. *Elgar companion to neo-schumpeterian economics*, 467, 2007.

[52] Rosen R. *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. Oxford: Pergamon, 1985.

[53] Matteo Richiardi, Roberto Leombruni, Nicole J. Saam, and Michele Sonnessa. A common protocol for agent-based social simulation. *Journal of Artificial Societies and Social Simulation*, 9(1):15, 2006.

[54] M.S. Roulston. Estimating errors on measured entropy and mutual information. *Physica D*, 125:285–294, 1999.

[55] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[56] Egle Samanidou, Elmar Zschischang, Dietrich Stauffer, and Thomas Lux. Agent-based models of financial markets. *Reports on Progress in Physics*, 70(3):409, 2007.

[57] I. Samengo. Estimating probabilities from experimental frequencies. *ArXiv: cond. math. stat. mech.*, 2002.

[58] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

[59] Manson S.M. *Validation and verification of multi-agent systems, in Complexity and Ecosystem Management*. Cheltenham: Edward Elgar, edited by m.a. janssen edition, 2002.

[60] James H. Stock and Mark W. Watson. Has the Business Cycle Changed and Why? NBER Working Papers 9127, National Bureau of Economic Research, Inc, August 2002.

[61] Leigh Tesfatsion and Kenneth L Judd. *Handbook of computational economics: agent-based computational economics*, volume 2. Elsevier, 2006.

[62] Tina Toni and Michael P. H. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, 2010.

[63] P. Windrum, G. Fagiolo, and A. Moneta. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10(2):8, 2007.

[64] David H Wolpert and David R Wolf. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841, 1995.

Appendix

A. Proof: Consistency and unbiasedness of frequency estimator

Recall that $\overline{p}_\mu(s) = \frac{1}{T}\sum_{t=1}^{T} p_\mu(x_t^s = s)$.

Assume that the ensemble we consider is composed by R independent runs of a given model, each of length T. Since we keep the model fixed, to ease notation we suppress the dependences from μ. Let $x_r^s(t)$ the symbolized version of the r^{th} run. For all $r = 1, ..., R$ the observed frequency of symbol s is

$$f_{\mu,r}(s) = \frac{1}{T}\sum_{t=1}^{T} \#[x_r^s(t) = s], \tag{16}$$

where $\#(\cdot)$ is the Dirac measure. Hence

$$\begin{aligned} \frac{1}{R}\sum_r f\mu, r &= \frac{1}{R}\sum_r \frac{1}{T}\sum_{t=1}^{T} \#[x_r^s(t) = s] & (17)\\ &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{R}\sum_r \#[x_r^s(t) = s]\right). & (18) \end{aligned}$$

Let us fix t and concentrate on the term $\frac{1}{R}\sum_r \#[x_r^s(t) = s]$. It is the frequency symbol s appears with across the ensemble in period t. First we observe that, by the weak law of large numbers, frequencies converges almost surely to probabilities

$$\frac{1}{R}\sum_r \#[x_r^s(t) = s] \xrightarrow{a.s.} p[x^s(t) = s]. \tag{19}$$

Since equation 19 holds for each $t = 1, ..., T$, we see that our estimator $\frac{1}{R}\sum_r f_{\mu,r}(s)$ is consistent. To better see this and its unbiasedness let up proceed as follows.

Keep fixed t and consider the random variable $n_r = \sum_r \#[x_r^s(t) = s]$. It is distributed as a Binomial($p[x^s(t) = s]$, R). Its expected value corresponds to $Rp[x^s(t) = s]$ which implies that n_r/R is an unbiased estimator of the true probability. Hence $\frac{1}{R}\sum_r f_{\mu,r}(s)$ is unbiased as well. Finally

$$\lim_{R\to\infty} Var(n_r/R) = \lim_{R\to\infty} \frac{p[x^s(t) = s](1 - p[x^s(t) = s])}{R} = 0 \tag{20}$$

confirms our estimator is consistent. Being $\frac{1}{R}\sum_r f_{\mu,r}(s)$ a linear combination of n_r/R for all available periods, it is consistent as well.

B. A direct computation of the systematic bias

In many cases a closed form computation of the systematic bias is possible, but it typically requires some assumption about the stochastic structure of the processes generating the time series. Here show how to compute it analytically under a quite general setup: we assume that the unconditional probabilities of observing a given symbol might vary over time and that realizations of the stochastic processes are time independent in the sense that

$$p(y_t^s = s_i \mid t' < t) = p(y_t^s = s_i) = p_t(s_i), \tag{21}$$

where y_t^s is the observed symbol at time step t in the symbolized time series. Throughout this section the independence assumption is relevant since allows to treat frequency counts as random variables with well known distributions and well defined moments.

Consider two symbolized time series x_t^s and y_t^s both of total length T and a given length of the blocks l. Within our framework x_t^s can be thought as the real world time series and y_t^s as one of the M runs of a given model of $y(t)$. For simplicity assume $l = 1$ so that the number of time windows coincides with the length of the series, that is $N_l = T$. Under a precision of the symbolization of b we recall that there are b^l symbols. The unconditional probability of observing a symbol $s_i \in S$ is allowed to vary over time. Hence

$$p(y_t^s = s_i \,|\, t = 1) \neq p(y_t^s = s_i \,|\, t = 2) \neq ... \neq p(y_t^s = s_i \,|\, t = T) \tag{22}$$

and the same holds for x_t^s.

Let n_i be the number of times the symbolized stochastic process y_t^s takes value s_i:

$$n_i = \sum_{t=1}^{T} I(y_t^s = s_i), \tag{23}$$

where I is an indicator function taking value 1 when $y_t^s = s_i$ and 0 otherwise, and let n_i' be the same quantity for the process x_t^s. If the realizations of each of the two processes are independent[21], n_i and n_i' are distributed as Poisson Binomial random variables with mean and variance respectively given by

[21] Realizations of x_t^s are independent from $x_{t'}^s$ for all $t' < t$ and the same holds for y_t^s.

$$E(n_i) = \sum_{t=1}^{T} p_t(s_i)$$
$$V(n_i) = \sum_{t=1}^{T} (1 - p_t(s_i)) p_t(s_i). \tag{24}$$

and

$$E(n'_i) = \sum_{t=1}^{T} p'_t(s_i)$$
$$V(n'_i) = \sum_{t=1}^{T} (1 - p'_t(s_i)) p'_t(s_i). \tag{25}$$

Not to abuse notation let $p_t(s_i) = p_{it}$ and $p'_t(s_i) = p'_{it}$. Now introduce $v_i = (n_i + n'_i)/2$; it satisfies

$$E(v_i) = \frac{\sum_{t=1}^{T} p_{it} + \sum_{t=1}^{T} p'_{it}}{2} \tag{26}$$

and

$$V(v_i) = \frac{\sum_{t=1}^{T} p_{it}(1 - p_{it}) + \sum_{t=1}^{T} p'_{it}(1 - p'_{it})}{4} \tag{27}$$

under the assumption that n_i and n'_i are independent. Note that this assumption is extremely reasonable in our context: models' output at time t is independent from the real world's corresponding observation at the same time.

Let $\mathbf{f} = \{f_1, ..., f_S\} = \{n_1/N_l, ..., n_S/N_l\}$ be a vector expressing the occurrence frequency of any symbol s_i according to the process y_t^s, and let $\mathbf{f'}$ be the counterpart for x_t^s. Then we define $m_i = v_i/N_l$ and we note that $m_i = (f_i + f'_i)/2$. Recall that the true unconditional probability assigned by a given model to symbol s_i in the time span $[t_0; T]$ is defined by (2.2) and that for R going to infinity the time frequency estimator converges almost-surely. Following [54], to compute the systematic bias let us introduce two new variables ϵ_i and ϵ'_i defined as

$$\epsilon_i = \frac{f_i - p_i}{p_i}, \tag{28}$$

and

$$\epsilon_i' = \frac{m_i - p_i''}{p_i''} \tag{29}$$

where p and p'' are defined according to (2.2) [22]. Using equation (7), it is possible to express the observed *GSL-div*[23] between x_t^s and y_t^s as

$$\begin{aligned} GSL(x(t)\,|\,y(t))_{obs} &= \sum_{l=1}^{L} w_i(-2\sum_{s_i\in S} m_i(1+\epsilon_i')\log_{a_l} m_i(1+\epsilon_i') \\ &+ \sum_{s_i\in S} f_i(1+\epsilon_i)\log_{a_l} f_i(1+\epsilon_i)). \end{aligned} \tag{30}$$

We separate the computation of the systemic bias for the two entropic functional in (30). Let us start with the observed entropy of $\mathbf{m}$:

$$\begin{aligned} H_{obs}(\mathbf{m}) &= -\sum_{s_i\in S} m_i(1+\epsilon_i')\log_{a_l} m_i(1+\epsilon_i') \\ &= -\sum_{s_i\in S} m_i(1+\epsilon_i')[\log_{a_l} m_i + \log_{a_l}(1+\epsilon_i')]. \end{aligned} \tag{31}$$

If Tp_i'' is large then ϵ_i' is small and the logarithm in (31) can be approximated in a Taylor series to give

$$\begin{aligned} H_{obs}(\mathbf{m}) &= -\sum_{s_i\in S} p_i''\log_{a_l} p_i'' + \epsilon_i' p_i''(1+\log_{a_l} p_i'') + \frac{(\epsilon_i')^2 p_i''}{2} + O[(\epsilon_i')^3] \\ &= H_\infty(\mathbf{m}) - \sum_{s_i\in S} \epsilon_i' p_i''(1+\log_{a_l} p_i'') + \frac{(\epsilon_i')^2 p_i''}{2} + O[(\epsilon_i')^3], \end{aligned}$$

where the true entropy is indicated as $H_\infty(m) = -\sum p_i''\log_{a_l} p''$. Since the expectation of the error term ϵ_i' is zero, the ensemble average of the observed entropy up to the second order can be expressed as

$$\langle H_{obs}(\mathbf{m})\rangle \approx H_\infty(\mathbf{m}) + \sum_{s_i\in S} \frac{\langle(\epsilon_i')^2\rangle p_i''}{2},$$

where the second term corrects for the systematic bias. In order to evaluate the correction one needs to compute the expectation of $(\epsilon_i')^2$. Recalling that, being X a random variable,

[22] In particular notice that $p'' = (p_i + p_i')/2$, where p_i and p_i' are time average probabilities in the sense of (2.2)
[23] Recall that the observed *GSL* is the *GSL* computed using frequencies instead of probabilities

$E[X^2] = V(X) + E^2[X]$ and that $E(\epsilon'_i) = 0$, one obtains

$$\langle(\epsilon'_i)^2\rangle = \frac{V(v_i)}{4T^2} \cdot \frac{1}{p''_i} \tag{32}$$

$$= \frac{\sum_{t=1}^{T} p_{it} - {p_{it}}^2 + p'_{it} - {p'_{it}}^2}{(\sum_{t=1}^{T} p_{it} + p'_{it})^2} \tag{33}$$

$$= \frac{\sum_{t=1}^{T} p_{it} + p'_{it} - \sum_{t=1}^{T} {p_{it}}^2 + {p'_{it}}^2}{(\sum_{t=1}^{T} p_{it} + p'_{it})^2} \tag{34}$$

if $p_i \neq 0$ or $p'_i \neq 0$. This leads to the following expression for the correction term

$$\sum_{s_i \in S} \frac{\langle(\epsilon'_i)^2\rangle p''_i}{2} = \frac{1}{4T} \sum_{s_i \in S} \left[1 - \frac{\sum_{t=1}^{T} {p_{it}}^2 + {p'_{it}}^2}{\sum_{t=1}^{T} p_{it} + p'_{it}}\right] \tag{35}$$

$$= \frac{B^m - \sum_{s_i \in S} \frac{\sum_{t=1}^{T} {p_{it}}^2 + {p'_{it}}^2}{\sum_{t=1}^{T} p_{it} + p'_{it}}}{4T} \leq \frac{B^m - 1}{4T} \tag{36}$$

where B^m is the number of symbols $s_i \in S$ such that $p''(s_i) \neq 0$[24]. We notice that the correction term is increasing in B^m. The use of large alphabets of symbols entails the risk of increasing the correction term in a way that it becomes too large relative to estimated entropies, therefore loosing its rationale. Equation (36) represents the correction term for the observed entropy of the mean frequency distribution between the two processes. Two important remarks apply. First, we notice that $B^m \geq \sum_{s_i \in S} \frac{\sum_{t=1}^{T} {p_{it}}^2 + {p'_{it}}^2}{\sum_{t=1}^{T} p_{it} + p'_{it}} \geq 1$, meaning that when we use the correction term in equation (10) we might overestimate the bias. The choice of using $\frac{B^m - 1}{4T}$ as correction term is due to the fact that the true bias depends on the unknown probabilistic structure of the real series. However, this choice provides additional robustness to our results: even using larger correction terms than we should, results are good and the estimated downward bias is very small with respect to the observed entropies even for relatively large alphabets (see Table 7 and 8 below). Moreover, using a larger first order correction term might compensate for higher order ones. Second, if we simplify our setting by considering i.i.d processes with time invariant probabilities ($p_t(s_i) = p(s_i)\ \forall t$), we end up with the same estimate of the systematic bias as in [7], [38] and [54].

Applying the same procedure to the observed entropy of model's output, $H_{obs}(\mathbf{f})$, it is immediate to obtain the following correction term

$$\sum_{s_i \in S} \frac{\langle\epsilon_i^2\rangle p'_i}{2} \leq \frac{B^f - 1}{2T} \tag{37}$$

[24] Recall that $p''(s_i)$ can be reasonably estimated through independent Montecarlo runs.

where B^f is the number of states such that $p'(s_i)$ is different from zero. Combining (36) and (37) with the functional form the *GSL-div*, (7), one obtains exactly the estimate of the systematic bias in (10).

Finally it is important to recall that the time independence assumption is useful to derive analytical estimates of the systematic bias; however, since the correction term is not affected by the assumption and B is obtained through ensemble averages of frequencies, which converge towards unconditional probabilities by the law of large numbers, the conditional structure of the processes does not affect the bias' estimate.

block length	*alphabet size*	*weights*	*AR1 - 0.1*	*AR1 - 0.2*	*AR1 - 0.01*	*AR1 - 0.5*	*ARMA11 - 0.2&0.9*	*ARMA11 - 0.5&2*	*AR1 - 0.9*	*AR1 - 1*	*AR1 - 2*
1	5	0.05	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	25	0.10	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000714
3	125	0.14	0.000000	0.000000	0.000000	0.000000	0.000000	0.000214	0.000929	0.003500	0.004786
4	625	0.19	0.000000	0.000000	0.000000	0.000000	0.000762	0.001333	0.004381	0.013619	0.017524
5	3125	0.24	0.000000	0.000000	0.000000	0.000119	0.003690	0.005119	0.012738	0.033333	0.041429
6	15625	0.29	0.000857	0.000714	0.000286	0.004857	0.019000	0.020571	0.044000	0.068000	0.080714

Table 7: Correction terms for each ARMA model studied in section 5. for each block length, under additively progressive weights.

block length	*alphabet size*	*weights*	*AR1 - 0.1*	*AR1 - 0.2*	*AR1 - 0.01*	*AR1 - 0.5*	*ARMA11 - 0.2&0.9*	*ARMA11 - 0.5&2*	*AR1 - 0.9*	*AR1 - 1*	*AR1 - 2*
1	5	0.17	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	25	0.17	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.001250
3	125	0.17	0.000000	0.000000	0.000000	0.000000	0.000000	0.000250	0.001083	0.004083	0.005583
4	625	0.17	0.000000	0.000000	0.000000	0.000000	0.000667	0.001167	0.003833	0.011917	0.015333
5	3125	0.17	0.000000	0.000000	0.000000	0.000083	0.002583	0.003583	0.008917	0.023333	0.029000
6	15625	0.17	0.000500	0.000417	0.000167	0.002833	0.011083	0.012000	0.025667	0.039667	0.047083

Table 8: Correction terms for each ARMA model studied in section 5. for each block length, under uniform weights.

C. Proof: Uniqueness of additively progressive weights

In this section we show that *additively progressive weights* are unique in the sense that (13) is the only formulation satisfying

a) $\sum_{l=1}^{L} w_l = 1$

b) $w_0 = 0$

c) $w_{l+1} - w_l = k, \quad k \in \mathcal{R}^+$.

To see this we start by writing weights recursively:

$$\begin{aligned} w_0 &= 0 \\ w_1 &= w_0 + k = k \\ w_2 &= w_1 + k = 2k \\ &... \\ w_L &= w_{L-1} + k = Lk. \end{aligned}$$

Using a) and b) we obtain

$$\sum_{l=1}^{L} w_l = k(1 + 2 + ... + L) = 1$$

whose unique solution is $k = \frac{2}{L(L+1)}$. Direct substitution in c) gives (13).

D. Proofs: Properties of the *GSL-div*

In this section we offer a proof or a sketch of it for each property outlined in section 3. As a starting point we recall that, for many of them, it is useful to rewrite the *GSL-div* as the following difference

$$GSL(p \,||\, \overline{p}_\mu) = \sum_i w_i \left[2H_i(m) - H_i(\overline{p}_\mu)\right], \tag{38}$$

where $m = \frac{p+\overline{p}_\mu}{2}$. This, together with the convention $0 \log(\frac{0}{0}) = 0$, is sufficient to guarantee that property 1. holds everywhere.

Then, given the bases of logarithms we use, it always holds $0 \leq H(\cdot) \leq 1$. Since $H(m) \neq 0$ if $H(\overline{p}_\mu) = 1$ property 2. follows immediately.

The proof of the third property is as simple as the second. Let us consider the $\Rightarrow$ direction. If $p = \overline{p}_\mu$ for all $j = 1, ..., L$ and $s \in S_{j,b}$, then $m = p = \overline{p}_\mu$ and hence $GSL(p \,||\, \overline{p}_\mu) = H(p) =$

$H(\overline{p}_\mu)$. Now let us move to the $\Leftarrow$ part. Assume

$$GSL(p\,||\,\overline{p}_\mu) = \sum_j w_j \left(2H_j(m) - H_j(\overline{p}_\mu)\right) = H(p).$$

Then we can write $H(p) = \sum_j w_j H_j(p) = \sum_j w_j H(p)$. Substituting in the expression above and rearranging one obtains

$$\sum_j w_j \left(2H_j(m) - H_j(\overline{p}_\mu) - H_j(p)\right) = 0.$$

Each term in the sum is non-negative by property 2. To have the equality holding each term has to be equal to zero:

$$2H_j(m) - H_j(\overline{p}_\mu) - H_j(p) = 0 \quad \forall\, j.$$

By the concavity of Shannon entropy and Jensen inequality it implies that m has to be *constant* and hence $p = \overline{p}_\mu$ for all $j = 1, ..., L$.

Property 4. and 5. follow immediately from the definition of *L-div* and the observation that the Generalized Jensen Shannon divergence (Lin 1991, [43]) is larger or equal to zero, with equality satisfied either if all the distributions analysed are equal one to other everywhere or they are independent. In particular, to see property 5. consider

$$GSL(p\,||\,\sum_i \pi_i \overline{p}_{\mu,i}) = \sum_j w_j [2H(\frac{p + \sum_i \pi \overline{p}_{\mu,i}}{2}) - H(\sum_i \pi_i \overline{p}_{\mu,i})] \tag{39}$$

and

$$\begin{aligned} \sum_i \pi_i GSL(p\,||\,\overline{p}_{\mu,i}) &= \sum_i \pi_i \left[\sum_j w_j [2H(\frac{p + \overline{p}_{\mu,i}}{2} - H(\overline{p}_{\mu,i})]\right] &(40)\\ &= \sum_j w_j \left[\sum_i \pi_i [2H(\frac{p + \overline{p}_{\mu,i}}{2}) - H(\overline{p}_{\mu,i})]\right] &(41) \end{aligned}$$

where the dependence of $H(\cdot)$ on j is omitted not to abuse notation. Now take the difference between 39 and 41 and forget about the summation over j for a while. Let us observe that

$$H(\frac{p + \sum_i \pi_i \overline{p}_{\mu,i}}{2}) - \sum_i \pi_i H(\frac{p + \overline{p}_{\mu,i}}{2}) \geq 0$$

by Jensen inequality and the concavity of entropy. The same holds for the difference between $H(\sum_i \pi_i \overline{p}_{\mu,i}$ and $\sum_i \pi_i H(\overline{p}_{\mu,i})$. Since this is true for all $j = 1, ..., L$ proposition 5. follows immediately. As a remark one can notice that equality holds in two cases: the trivial case

where all probability vectors are equal (and hence the *GSL-div* is zero) and where $\overline{p}_{\mu,i}$s and p are mutually independent. In the latter case, property 5. follows from the *additive* property of Shannon entropy (see [13]).

Let us move to property 6. To prove it we start observing that the base of logarithms we use, $a_{l,b} = b^l$, is increasing both in l and b. Thus, it suffices to prove that the *GSL-div* is decreasing in $a_{l,b}$ to obtain property 6. immediately. Let us avoid to report dependences from l and b to ease notation

$$\begin{aligned} GSL(p\,||\,\overline{p}_\mu) &= \sum w\left[2H(m) - H(\overline{p}_\mu)\right] && (42)\\ &= \sum w\left[-2\sum m\log_a m + \sum \overline{p}_\mu \log_a \overline{p}_\mu\right] && (43)\\ &= \sum w\left[\frac{1}{\log a}(-2\sum m\log m + \sum \overline{p}_\mu \log \overline{p}_\mu)\right] && (44) \end{aligned}$$

where the logarithm in last line is in base e. Let $D(\cdot)$ indicate the derivative operator. Since $a > 0$,

$$D\left(\frac{1}{\log a}\right) = (-1)\cdot\frac{1}{(\log a)^2}\cdot\frac{1}{a} < 0.$$

Finally we offer a proof of property 7. Let us start by focusing on weights we use, that is, geometrically progressive ones (see section 2.4). Let L be the maximum blocks' length we consider. Since $\sum_i w_i = 1$,

$$\begin{aligned} w_1 &= \frac{1}{1 + c + c^2 + ..., + c^{L-1}}\\ w_2 &= \frac{c}{1 + c + c^2 + ... + c^{L-1}}\\ &...\\ w_L &= \frac{c^{L-1}}{1 + c + c^2 + ... + c^{L-1}}. \end{aligned}$$

Collecting c^{L-1} in each denominator, they can be rewritten as

$$
\begin{aligned}
w_1 &= \frac{1}{c^{L-1}\left[(\frac{1}{c})^{L-1}+(\frac{1}{c})^{L-2}+...+(\frac{1}{c})+1\right]} = \frac{(\frac{1}{c})^{L-1}}{\left[(\frac{1}{c})^{L-1}+(\frac{1}{c})^{L-2}+...+(\frac{1}{c})+1\right]} \\
w_2 &= \frac{c}{c^{L-1}\left[(\frac{1}{c})^{L-1}+(\frac{1}{c})^{L-2}+...+(\frac{1}{c})+1\right]} = \frac{(\frac{1}{c})^{L-2}}{\left[(\frac{1}{c})^{L-1}+(\frac{1}{c})^{L-2}+...+(\frac{1}{c})+1\right]} \\
&... \\
w_L &= \frac{c^{L-1}}{c^{L-1}\left[(\frac{1}{c})^{L-1}+(\frac{1}{c})^{L-2}+...+(\frac{1}{c})+1\right]} = \frac{1}{\left[(\frac{1}{c})^{L-1}+(\frac{1}{c})^{L-2}+...+(\frac{1}{c})+1\right]}
\end{aligned}
$$

It is relevant to observe that if $c = f(L)$ with f increasing in L, when $L \to \infty$ all weights but the last one converges to zero and $w_L \to 1$. Now, let us study the limit

$$
\lim_{T\to\infty}\lim_{L\to\infty}\frac{1}{T}GSL(p\,||\,\overline{p}_\mu) = \lim_{T\to\infty}\lim_{L\to\infty}\frac{1}{T}\left\{\sum_{j=1}^{L} w_j[2H_j(m) - H_j(\overline{p}_\mu)]\right\}, \tag{45}
$$

where logarithms in entropies are in base 2. Previous result about weights ensures that $\lim_{L\to\infty}\left\{\sum_{j=1}^{L} w_j[2H_j(m) - H_j(\overline{p}_\mu)]\right\}$ behave asymptotically as the last term of the sum over j. Recalling that entropies of distributions are bounded from above for any finite support, the limit in 45 can be re written as

$$
\lim_{T\to\infty}\lim_{L\to\infty}\frac{1}{T}[2H_L(m) - H_L(\overline{p}_\mu)] \tag{46}
$$

which is equivalent as

$$
\lim_{T\to\infty}\frac{1}{T}[2H_T(m) - H_T(\overline{p}_\mu)] \tag{47}
$$

if $x(t)$ and $y(t)$ are strongly stationary. Moreover, when series are stationary, the limit exists finite (see pp. 255 of [47]) and

$$
\lim_{T\to\infty}\frac{1}{T}[2H_T(m) - H_T(\overline{p}_\mu)] = 2H_\infty(m) - H_\infty(\overline{p}_\mu). \tag{48}
$$

where $H_\infty(\cdot)$ is the asymptotic entropy ([13, 47]). Therefore, we have obtained the convergence of a particular specification of the *GSL-div* to a subtracted version of the asymptotic *GSL-div*, which can interpreted as a symmetric and well defined version of the asymptotic *KL-div* and, therefore, of the AIC ([1]).