Econometrics and Statistics 20 (2021) 62–86

Contents lists available at ScienceDirect

Econometrics and Statistics

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

Model calibration and validation via confidence sets

Raffaello Seri[a,b], Mario Martinoli[a,*], Davide Secchi[b], Samuele Centorrino[c]

[a] Department of Economics, Università degli Studi dell'Insubria, Varese, Italy
[b] Centre for Computational & Organisational Cognition (CORG), Department of language and communication, University of Southern Denmark, Slagelse, Denmark
[c] Economics Department, Stony Brook University, Stony Brook, USA

ARTICLE INFO

Article history:
Received 29 April 2019
Revised 27 January 2020
Accepted 29 January 2020
Available online 18 February 2020

Keywords:
Calibration
Validation
Simulated models
Model confidence set
Large deviations

ABSTRACT

The issues of calibrating and validating a theoretical model are considered, when it is required to select the parameters that better approximate the data among a finite number of alternatives. Based on a user-defined loss function, Model Confidence Sets are proposed as a tool to restrict the number of plausible alternatives, and measure the uncertainty associated to the preferred model. Furthermore, an asymptotically exact logarithmic approximation of the probability of choosing a model via a multivariate rate function is suggested. A simple numerical procedure is outlined for the computation of the latter and it is shown that the procedure yields results consistent with Model Confidence Sets. The illustration and implementation of the proposed approach is showcased in a model of inquisitiveness in ad hoc teams, relevant for bounded rationality and organizational research.[1]

1. Introduction

In the social sciences, there are usually two methods for determining the parameters of a specific model: calibration and estimation. While the latter is more appealing in a number of cases, it usually requires making additional, and sometimes arbitrary, assumptions about the way data were generated. Exogenous shocks may be added to the model in an ad hoc fashion, and point identification and estimation of the model's parameters are often an issue for complicated, nonlinear and nonseparable structural equations. Examples of the latter instances are applied general equilibrium analysis, in which hypotheses for identification and estimation of parameters are often difficult (see Canova and Sala, 2009, for a discussion on the issues of identification in dynamic general equilibrium models), and agent-based models (ABM hereafter), that are often very hard to describe within a unified statistical framework (see, e.g., Gilbert and Terna, 2000; Fagiolo et al., 2007; Gilbert, 2007; Windrum et al., 2007; Di Molfetta, 2016; Guerini and Moneta, 2017; Fagiolo et al., 2019).

Therefore, calibration has evolved to be a very popular method for parameter determination. Calibration allows one to set model parameters to certain values that are chosen according to prevailing theoretical and/or empirical evidence (see Brenner, 2006; Crooks et al., 2008). This definition of calibration is consistent with most of the economic literature and goes back at least to the seminal business cycle paper of Kydland and Prescott (1982) (see also Cooley, 1997 and Romer, 2012, Sec. 5.8, p. 217). Calibration is followed by a step that is referred to as model validation (or verification, see Oreskes et al., 1994, and Boero and Squazzoni, 2005), in which moments of random sequences generated by the model are compared

* Corresponding author.
E-mail address: m.martinoli@uninsubria.it (M. Martinoli).

[1] Code and data are available as an online supplement.

https://doi.org/10.1016/j.ecosta.2020.01.001

(or eye-balled) with sample moments to show that, given the parameters' value, the economic model provides outcomes consistent with observable data.

From a statistical perspective, drawbacks of this particular definition of calibration and validation are that it does not usually provide a quantitative measure of distance of the chosen model to the data, and that it does not take into account the uncertainty related to the choice of a specific vector of parameters.

This definition has been challenged by Hansen and Heckman (1996). The authors have noticed that the distinction between calibration and estimation is "artificial at best" (see also Richiardi et al., 2006). They argue that "[calibration] looks remarkably like a way of doing estimation without accounting for sampling errors in the sample mean". They further claim that the validation step used by Kydland and Prescott (1982) is tantamount to testing. They argue that both steps are coming from the minimization of an implicit loss function which, if made explicit, would make the principle by which a particular model is chosen easier to understand.

While there is little agreement as to what "calibration" is, there is even more confusion about what are "verification" and "validation". Some authors (Oreskes et al., 1994, pp. 642-643 or Hansen and Heckman, 1996, pp. 91-92) use "verification" for the comparison of the output of a calibrated model with the real world. Other authors (North and Macal, 2007, pp. 30-31, or Crooks et al., 2008, p. 419) use "verification" as the process of making sure that the implementation is coherent with the structure of a model and "validation" as the process of making sure that the implementation is coherent with the real world, thus leading to an overlap between "validation" and "calibration" (Crooks et al., 2008, p. 419). In Richiardi et al. (2006), "validation" is used as a check of the adequacy of the simulated model with another model it is intended to portray. At last, Boero and Squazzoni (2005, Sec. 2.1-2.4) use "verification" as in North and Macal (2007), Crooks et al. (2008), "calibration" as in Oreskes et al. (1994), Hansen and Heckman (1996) and "validation" for the comparison of the output of a calibrated model with an external source of data (see also Fagiolo et al., 2019).

In this paper, we follow the approach suggested in Hansen and Heckman (1996) and consider the issue of calibrating and validating an economic model by minimization of a user-specified loss function, when researchers need to choose among several (albeit finite; see also Boero and Squazzoni, 2005, Sec. 2) combinations of the parameters. The underlying economic model does not have to admit a closed-form solution, or to be point identified in a statistical sense. It should, however, be possible to generate series of simulated data from the model itself, for specific configurations of parameters. While the restriction to a finite number of configurations of the parameters may appear to be a limitation in some cases, it is, in our opinion, a minor one. First, the number of configurations may be increased at a relatively low computational burden. Second, most optimization algorithms are based on a finite number of evaluations and on stopping rules that are not guaranteed to reach the true optimum. Possible extensions to the infinite case are deferred to future work.

We thus consider the issue of this theoretical model matching some benchmark data. It is important to note that the benchmark is not necessarily composed of real-world observations: it could also be a function, or results from another simulation, for example. In some cases, one looks for values of parameters that provide simulations that replicate a deterministic target or achieve a fixed goal (see, e.g., Crooks et al., 2008). What is more relevant instead is that the real data should contain patterns, i.e. "observations of any kind showing nonrandom structure and therefore containing information on the mechanisms from which they emerge" (see Grimm et al., 2005, p. 991). The match between the benchmark and the simulated samples is assessed through a distance. In the following it will be clear that our developments do not rely on any specific property of a distance. However, we prefer to use this name as most applications will deal with this case. Given a distance, we propose to construct Model Confidence Sets (MCS) in the spirit of Hansen et al. (2011). This technique not only allows us to rank models based on their closeness to the benchmark, but also to construct a plausible set of alternative specifications of parameters that cannot be distinguished from the chosen one, at least not in a statistical sense. In this respect our procedure can be used in a first calibration step, to select the vectors of parameters that are closer to the benchmark data, and, being related to a testing procedure through the duality between confidence sets and tests (see, e.g., Bickel and Doksum, 2015, Sec. 4.5, p. 241), it can also be used for validation in a second step. Notice that we do not see the requirement that parameters have discrete support as a drawback. As a matter of fact, most simulation-based estimators are obtained by sampling from conveniently selected points in the sample space (as sampling from the entire space is often impossible in practice). Moreover, when used for calibration, our procedure can help identify multiple local optima of the objective function, when direct minimization would only identify one of those local optima. This feature is especially useful for highly parametrized models where point identification is often impossible to show (see Crooks et al., 2008, p. 419). Along the lines of Hansen et al. (2011), we show that MCS are valid confidence sets when the number of simulated samples, n, diverges.

We complement this analysis by characterizing the model's choice as an estimation problem in which the parameters have discrete support (see Choirat and Seri, 2012). We show that, despite the computation of the probability of choosing a combination of parameters over the others is probably out of reach, one can still provide, using large deviations theory, a measure of the rate of decrease of this probability with n through a multivariate rate function, which depends on the vector of observables. The finite sample approximation of this rate function is interesting for at least two reasons. On the one hand, it can be used empirically as an alternative metric to establish the plausibility of our set of models. However, its computation is not straightforward. On the other hand, its theoretical properties have not been explored thus far, to the best of our knowledge, as published papers only consider rate functions in the univariate context (see, e.g., Duffield et al., 1995; Duffy and Metcalfe, 2005a; 2005b; Rohwer et al., 2015; Duffy and Williamson, 2015). In the univariate case, the rate function needs to be calculated at one point only and this does not involve particular computational issues. In our case,

however, the measure of the decrease rate of the probability arises from the optimization of a function over a multivariate domain and this requires high accuracy in the computation of the whole function to be minimized.

First, we propose a method based on the *linear-time Legendre Transform* (LLT) to numerically compute the rate function. As the name suggests, the numerical complexity of this method only grows linearly with the number of data points, and it thus provides a computationally efficient method, even when the number of observations is large. Second, we show that this rate function approximation is consistent, and we obtain its rate of convergence. We then perform a short simulation study to demonstrate the validity of this theoretical result in finite samples.

We conclude the paper with an application of this methodology to an ABM of inquisitiveness (henceforth, *inquisitiveness ABM*) (see Bardone and Secchi, 2017). Inquisitiveness is a development of "docility", as a strategy that decision makers use to cope with their bounded rationality (Simon, 1990; 1993) and it has been applied to agent-based simulations recently for its cognitive links to behavior (Secchi, 2016; Secchi and Gullekson, 2016; Miller and Lin, 2010). A docile individual is someone who leans on other people's advice to make decisions. In particular, when docile individuals operate inside a team (or any community of reference) to solve a problem, trust is placed on the advice of team members, while data coming from outside the team are mistrusted. An inquisitive individual, instead, does not operate on a team basis but s/he is rather a problem solver, who may look for help outside the team. We show that both Model Confidence Sets and the approach based on a rate function approximation yield similar results. This ABM has been selected for three reasons: (a) theoretical relevance, (b) applicability to other economic domains, and (c) computational simulation robustness. The first criterion entails selecting a model that is anchored to a solid theoretical background while, at the same time, it should not be a mere replication of other models. The inquisitiveness ABM is based on Simon's bounded rationality (Simon, 1955; 1979; 1997). This is a theory with high relevance, given the value that its theoretical and empirical applications have had in economics in the recent decades (e.g., Kahneman, 2003; Thaler and Sunstein, 2008; Camerer and Fehr, 2006). Inquisitiveness exemplifies the tension toward finding more suitable theoretical conceptualizations of bounded rationality. At the same time, it has wide applicability to organizations of various sizes as well as wider economic systems. Finally, the code has been openly peer reviewed in a certifiable and documentable way, as it is uploaded into OpenABM. This platform is supported by the Network for Computational Modeling in Social and Ecological Sciences (CoMSES) whose scholars offer peer review specifically on the code, when modelers ask for it. Among the (very few) models that had been peer reviewed, the one on the inquisitiveness ticked also all of our other criteria for selection.

The inquisitiveness model shows high parametrization and nonlinearities in both variables and parameters, and calls for a rigorous calibration procedure, as it is widely done in ABM research. Alternative procedures have also been extensively explored (see Gilli and Winker, 2003; Windrum et al., 2007; Winker et al., 2007; Fabretti, 2013; Grazzini and Richiardi, 2015; Recchioni et al., 2015; Grazzini et al., 2017; Guerini and Moneta, 2017; Kukacka and Barunik, 2017; Lamperti et al., 2018; Fagiolo et al., 2019). The most recent trend in calibration of ABM is to apply sound econometric techniques to select appropriate parameter values (see Chen et al., 2012). Methods such as Indirect Inference (Gilli and Winker, 2002; 2003), Simulated Method of Moments (Winker et al., 2007), Simulated Maximum Likelihood (Kukacka and Barunik, 2017), Simulated Minimum-Distance (Grazzini and Richiardi, 2015; Recchioni et al., 2015; Lamperti, 2018a; 2018b), and Approximate Bayesian Computation (Grazzini et al., 2017) have emerged as calibration techniques in ABM. We believe there are three main drawbacks of performing calibration in this way. First, its results may vary widely depending on sample selection, as models are not necessarily identified, and therefore uniqueness of the optimum is not guaranteed (not even asymptotically). Similarly, there has been little effort to formally characterize the uncertainty associated to a specific choice of parameters (Seri and Secchi, 2017). Finally, given the high parametrization of ABM, computational costs can be extremely high (e.g., Thiele et al., 2015; Lorscheid and Meyer, 2016). Our technique is instead computationally fast, and effectively uses the sample to validate one or more choices of parameters.

Our procedure bears a resemblance with other ones introduced in the literature. For instance, Lamperti et al. (2018) use Surrogate Meta Models and Machine Learning to search the parameter space. Their iterative technique is based on a surrogate model which learns from an initial random subset of parameter combinations. We perceive that our method is similar in spirit, although one learns from a finite number of combinations of parameters that are directly specified by users, instead of being determined iteratively from a (finite-dimensional) subset of the parameter space. Similarly, Barde (2016) applies the methodology developed by Barde (2017) to select among different models, scoring each one of them at the level of individual empirical observations. Differently from Barde (2016), we rigorously study the construction of a Model Confidence Set and we provide the statistical properties of the rate function.

Our approach can complement existing ones, and be used in a validation step to compare estimates obtained by any of the aforementioned estimation methods. By appropriately splitting the data into a fitting sample and a training sample, one can obtain several plausible parameter values from an initial estimation step, and then compare them in a testing (or validation) step using MCS.

The paper is structured as follows. The mathematical notation that is needed for the discussion of the model is introduced in Section 2. Section 3 presents the statistical framework and provides the asymptotic relations between the probabilities of choosing a model and the rate functions. Section 4 discusses the construction of Model Confidence Sets for a problem of calibration. Section 5 outlines the numerical approximation of the rate function and provides its statistical properties. Readers more interested in the application of our procedure can directly move to Section 6. We provide an outline of our findings in Section 7. All technical proofs are relegated to the Appendix.

2. Notations

This section introduces the notation that will be used throughout the paper. Capital bold letters, such as $\mathbf{A}$, denote matrices while lowercase bold letters, such as $\mathbf{a}$, usually denote vectors. The $i-$th element of vector $\mathbf{a}$ is generally denoted a_i. $\mathbf{u}_n$ is a n-vector composed of ones. $\mathbf{I}_n$ is the $(n \times n)$-identity matrix. $\mathbf{U}_n$ is a $(n \times n)$-matrix composed of ones. $\mathbf{e}_{i,n}$ is a n-vector of zeros with a one in the i-th position; when the length is clear from the context we simply use $\mathbf{e}_i$. $\mathbf{0}_{m,n}$ is a $(m \times n)$-matrix composed of zeros. We do not indicate the dimensions when they are clear from the context. diag($\mathbf{a}$) is a diagonal matrix with $\mathbf{a}$ on its diagonal. $\mathbf{A}'$ and $\mathbf{A}^{-1}$ are respectively the transpose and the classical inverse of the matrix $\mathbf{A}$, provided they exist.

For a set A, the notations int$A$, $\overline{A}$, ∂A, co$A$ and $\overline{\text{co}}A$ respectively denote the interior, the closure, the boundary, the convex hull and the closed convex hull of A. The positive half-line is $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$. The positive orthant is $\mathbb{R}^d_+ := (\mathbb{R}_+)^d$. Hyperplanes are indicated with the point-normal notation, i.e. as $\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}'\boldsymbol{\alpha} = \mathbf{v}'\boldsymbol{\alpha}\}$ where $\boldsymbol{\alpha}$ is a normal vector and $\mathbf{v}$ is a point of the hyperplane. The same notation is used for a half-space as $H^+(\boldsymbol{\alpha}, \mathbf{v}) := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}'\boldsymbol{\alpha} \geq \mathbf{v}'\boldsymbol{\alpha}\}$. For a scalar β, $\beta A := \{\beta \mathbf{x} \colon \mathbf{x} \in A\}$.

We introduce the definition of the *effective domain* of a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ as $\mathcal{D}(f) := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) < \infty\}$ (see Dembo and Zeitouni, 2010, p. 4). The *Legendre* or *Legendre-Fenchel transform* of f is the function $f^\star : \mathbb{R}^d \to \overline{\mathbb{R}}$ defined by the variational formula $f^\star(\mathbf{y}) := \sup_{\mathbf{x} \in \mathbb{R}^d}\{\mathbf{y}'\mathbf{x} - f(\mathbf{x})\}$. $f^\star$ is also said to be the *convex conjugate* of f. χ_A is the (convex analysis) *characteristic function* taking the value 0 in A and $+\infty$ outside.

Expectations are denoted as $\mathbb{E}$. When the integration variable is not clear from the context, we indicate it explicitly as a subscript as in $\mathbb{E}_x$. We will do the same for the variance $\mathbb{V}_x$. For a random vector $\mathbf{X}$ taking values in $\mathbb{R}^d$, we define the *moment generating function* (*MGF*) of $\mathbf{X}$ as $M(\mathbf{u}) := \mathbb{E}\exp\{\mathbf{u}'\mathbf{X}\}$ and the *cumulant generating function* (*CGF*) as $\Lambda(\mathbf{u}) := \ln M(\mathbf{u})$.

3. Framework and preliminary results

Let $\mathbf{y}$ be a vector of benchmark observations, that can contain individual level data, a time series, etc. It can be determined by a deterministic or a stochastic mechanism, but in any case it is supposed to be fixed in the subsequent analysis. Fixing the outputs $\mathbf{y}$ may come from the difficulty in modelling the real-world phenomenon leading to the data, or from an interest in the real-world data rather than in the data generating process (DGP) leading to them.

Example 1 (Multiple outputs)**.** Suppose that the simulation model describes a situation characterized by several outputs $\mathbf{y} = (y_1, \ldots, y_p)$ measured at the same time. The researcher may be interested in looking for the parameters of the simulation model yielding outputs that are as similar as possible to these observations, without the need to model the DGP leading to these data.

Example 2 (Time trend)**.** Consider a simulation model describing the evolution of a system over time. In this case $\mathbf{y} = (y_1, \ldots, y_t)$ may be a time series.

Example 3 (Spatial patterns)**.** In some cases, the aim of the simulation is to simulate the spatial patterns arising from a complex system. This situation takes place in the study of animal diffusions, geographical plant distributions, etc. In this case $\mathbf{y}$ is a curve in a space or a spatial distribution of points.

We assume to have m_0 configurations of parameters, say θ_i for $i = 1, \ldots, m_0$. We will not need any special structure on the set of possible parameters. The set of all parameters is denoted by $\mathcal{M}^0 := \{1, \ldots, m_0\}$. For each one of them, we simulate n realizations $\mathbf{z}_j(\theta_i)$, for $j = 1, \ldots, n$. We compute the distances $d(\mathbf{y}, \mathbf{z}_j(\theta_i))$ for $i \in \mathcal{M}^0$ and $j = 1, \ldots, n$. Here and in the following we will use the term distance loosely; as an example, we will never use any specific property of a distance, such as non-negativity or the triangle inequality.

Example 4 (Multiple outputs - Example 1 continued)**.** A reasonable choice of a distance is:

$$d(\mathbf{y}, \mathbf{z}_j(\theta_i)) = \sum_{h=1}^{p} a_h |y_h - z_{jh}(\theta_i)|,$$

where the a_h's, for $h = 1, \ldots, p$, are constants keeping into account the different contributions of the observations to the overall distance.

Example 5 (Time trend - Example 2 continued)**.** The time series case admits several distances, with different interpretations. In the ergodic case, when a single instance of a time series is sufficient to estimate probabilities of events, one can consider distances between the probability measures that generated the data. In this case it is quite natural to keep into account the DGP of $\mathbf{y}$ when drawing inferences. For a review of metrics or similarity measures for statistical data processes see Basseville (2013) and Marks (2013). Two further interesting contributions are provided by Barde (2017) and Lamperti (2018b). Other distances cannot be cast as metrics between the probability measures that generated the data, but only as distances between the trajectories over time. These distances are probably more interesting for the following as they apply to trending non-ergodic time series (e.g., Liao, 2005; Fu, 2011; Wang et al., 2013).

Example 6 (Spatial patterns - Example 3 continued)**.** If one considers an animal roaming in a certain area, several distances can be considered, such as the Fréchet distance, used for dynamic time warping (see Kruskal, 1983; Berndt and Clifford, 1994; Gudmundsson et al., 2008).

We introduce the following notations. The expected distance relative to the i-th combination of parameters is $\overline{D}_i := \mathbb{E}_{\mathbf{z}} d(\mathbf{y}, \mathbf{z}(\theta_i))$. We denote as $\overline{\mathbf{D}} := (\overline{D}_1, \ldots, \overline{D}_{m_0})'$ the vector containing the average distances. The distance corresponding to the j-th realization and the i-th combination of parameters is $D_{j,i} := d(\mathbf{y}, \mathbf{z}_j(\theta_i))$, and the sample average distance is $\overline{D}_{n,i} := \frac{1}{n}\sum_{j=1}^{n} D_{j,i}$. Let us define the m_0-vector:

$$\overline{\mathbf{D}}_n := \left(\overline{D}_{n,1}, \ldots, \overline{D}_{n,m_0}\right)'.$$

If we denote:

$$\mathbf{D}_k := \left(D_{k,1}, \ldots, D_{k,m_0}\right)'$$

for $k = 1, \ldots, n$, we can write $\overline{\mathbf{D}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{D}_k$.

We want to identify the values i corresponding to the smallest expected distance $\overline{D}_i$. Other criterion choices may be more relevant in some situations but in this case, we consider only the expected distance. The set of parameters achieving the minimal distance is denoted as $\mathcal{M}^\star := \{j \in \mathcal{M}^0 : \overline{D}_j = \min_{i\in\mathcal{M}^0} \overline{D}_i\}$. It is clear that a choice would be to take the value $\widehat{i}_n$ minimizing $\overline{D}_{n,i}$ for $i \in \mathcal{M}^0$. We suppose that $\widehat{i}_n$ is always single valued while $\mathcal{M}^\star$ may not be a singleton.

The choice of a value from a finite set can be characterized as a discrete-parameter estimation problem (see Choirat and Seri, 2012) or as a classification problem. In the following we characterize the behavior of $\widehat{i}_n$ as in Choirat and Seri (2012). We can write:

$$\begin{aligned}\mathbb{P}\left\{\widehat{i}_n = j\right\} &= \mathbb{P}\left\{\widehat{\theta}_n = \theta_j\right\} = \mathbb{P}\left\{\overline{D}_{n,j} \le \overline{D}_{n,i}, 1 \le i \le m_0, i \ne j\right\} \\ &= \mathbb{P}\left\{\overline{\mathbf{D}}_n \in \mathcal{P}_j\right\}\end{aligned}$$

where $\mathcal{P}_j$ is the (unbounded) polytope characterized by the inequalities:

$$\mathcal{P}_j := \left\{\mathbf{x} \in \mathbb{R}^{m_0} : x_j \le \min_{1\le\ell\le m_0, \ell\ne j} x_\ell\right\}.$$

Fig. 1 provides a graphical representation of $\mathcal{P}_3 = \{\mathbf{x} \in \mathbb{R}^3 : x_3 \le \min\{x_1, x_2\}\}$ in $\mathbb{R}^3$.

Note that, in the previous definition, there is no special reason to use strict inequalities instead of non-strict ones. As the polytopes are a partition of the whole space, $\overline{\mathbf{D}}$ can belong to one and only one of the polytopes. But in some cases, $\overline{\mathbf{D}}$ may be on the face between two or more polytopes and, as the attribution of the faces to one polytope or another is arbitrary, we prefer to explicitly allow $\mathcal{M}^\star$ not to be a singleton. For any $j \notin \mathcal{M}^\star$, our assumptions guarantee that the probability $\mathbb{P}\{\overline{\mathbf{D}}_n \in \partial\mathcal{P}_j\}$ is asymptotically negligible, i.e. that $\mathbb{P}\{\overline{\mathbf{D}}_n \in \text{int}\mathcal{P}_j\}$ and $\mathbb{P}\{\overline{\mathbf{D}}_n \in \overline{\mathcal{P}}_j\}$ behave similarly for any $j \notin \mathcal{M}^\star$. This makes immaterial whether the faces of each $\mathcal{P}_j$ are attributed to a polytope or the other.

Once the attribution of the faces is solved, the polytopes $\{\mathcal{P}_j, j \in \mathcal{M}^0\}$ form a partition of $\mathbb{R}^{m_0}$. Under appropriate assumptions (see A1 below), Khintchin's WLLN shows that $\overline{\mathbf{D}}_n$ converges in probability to $\overline{\mathbf{D}}$. This implies that:

$$\mathbb{P}\left\{\widehat{i}_n \in \mathcal{M}\right\} = \mathbb{P}\left\{\overline{\mathbf{D}}_n \in \bigcup_{j\in\mathcal{M}} \mathcal{P}_j\right\} \to \mathbb{P}\left\{\overline{\mathbf{D}} \in \bigcup_{j\in\mathcal{M}} \mathcal{P}_j\right\}$$

for any $\mathcal{M} \subset \mathcal{M}^0$. It is clear that:

$$\mathbb{P}\left\{\widehat{i}_n \in \mathcal{M}^\star\right\} = \mathbb{P}\left\{\overline{\mathbf{D}}_n \in \bigcup_{j\in\mathcal{M}^\star} \mathcal{P}_j\right\} \to 1$$

while $\mathbb{P}\{\widehat{i}_n = j\} = \mathbb{P}\{\overline{\mathbf{D}}_n \in \mathcal{P}_j\} \to 0$ for any $j \notin \mathcal{M}^\star$. In the following we show that the probabilities like $\mathbb{P}\{\widehat{i}_n = j\}$ for any $j \notin \mathcal{M}^\star$ converge to 0 exponentially fast and we characterize the rate of exponential convergence of these probabilities.

This shows that our estimator $(\widehat{i}_n)$ is consistent, i.e. it takes a value in the set of minimizers of the average distance $(\mathcal{M}^\star)$ with probability converging to 1 as the number of replications (n) diverges. Moreover, the probability of the estimator $(\widehat{i}_n)$ taking any value outside the set of minimizers of the average distance $(\mathcal{M}^\star)$ converges to 0 exponentially. As in Choirat and Seri (2012), the estimator has an exponential convergence rate. In the following, we try to characterize, and then estimate, the rate of convergence to 0 of this probability. This is done through large deviations principles (LDP).

As they do not belong to the toolbox of the average econometrician, it is useful to provide a brief explanation of LDP in the case that is most relevant for our purposes (Corollary 6.1.6 in Dembo and Zeitouni, 2010, p. 253). Let $\{\mathbf{X}_1, \ldots\}$ be a sequence of independent and identically distributed random vectors. If $\mathbb{E}\mathbf{X}$ exists, $\overline{\mathbf{X}}_n := \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k$ converges almost surely to $\mathbb{E}\mathbf{X}$. For a given set $A \subset \mathbb{R}^d$, $\mathbb{E}\mathbf{X} \in A$ implies that $\mathbb{P}\{\overline{\mathbf{X}}_n \in A\} \to 1$. Instead, if $\mathbb{E}\mathbf{X} \notin A$, $\mathbb{P}\{\overline{\mathbf{X}}_n \in A\} \to 0$ exponentially fast. In particular, $\frac{1}{n}\ln\mathbb{P}\{\overline{\mathbf{X}}_n \in A\}$ is asymptotically bracketed between two bounds depending on the distribution of $\mathbf{X}_1$ and on the set A. These bounds can respectively be expressed as the infimum over int$A$ and $\overline{A}$ of a so-called rate function $\Lambda^\star$ that is the

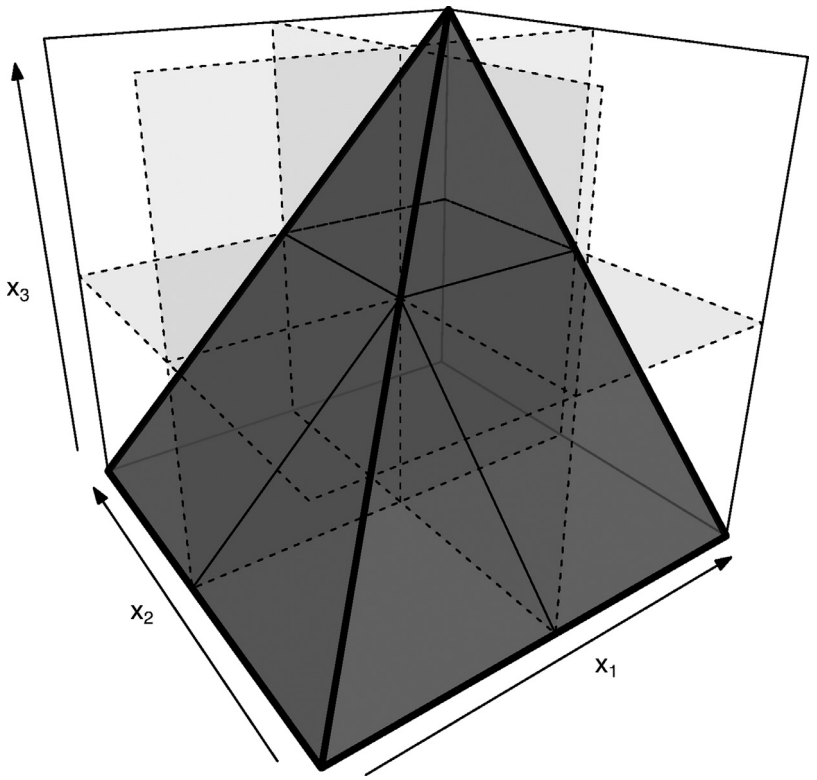

Fig. 1. Polytope $\mathcal{P}_3$ in $\mathbb{R}^3$: the dark grey surfaces delimit $\mathcal{P}_3$ from above; the light grey rectangles represent the three planes $x_1 = 0$, $x_2 = 0$ and $x_3 = 0$; the dashed lines are the intersections of the planes $x_1 = 0$, $x_2 = 0$ and $x_3 = 0$; the solid thin lines are the intersections of the previous planes with the surfaces delimiting $\mathcal{P}_3$.

Legendre-Fenchel transform of the CGF Λ of $\mathbf{X}_1$. When the set A and the random vector $\mathbf{X}_1$ are sufficiently well behaved, the infima over int$A$ and $\overline{A}$ coincide and one can identify a limit of $\frac{1}{n}\ln\mathbb{P}\{\overline{\mathbf{X}}_n \in A\}$. This limit can be expressed as the infimum of the rate function over the set A, where the infimum is generally located on the boundary of A. The point in which the infimum is reached is called a dominating point (see Ney, 1983; 1984) and it can be used to express the large deviations result in an alternative way.

The role played by the CGF and by its Cramér transform in these bounds can be justified considering the upper bound. The derivation of the lower bound is too complicated to be explained here. We consider the scalar case with $A = [x, \infty)$. For every $\lambda \geq 0$, the Chernoff bound yields:

$$\begin{aligned}\mathbb{P}\left\{\overline{X}_n \in [x,\infty)\right\} &= \mathbb{E}1\left\{\overline{X}_n - x \geq 0\right\} \leq \mathbb{E}\exp\left\{\lambda\left(\overline{X}_n - x\right)\right\} \\ &= \exp\{-\lambda x\}\prod_{k=1}^{n}\mathbb{E}\exp\left\{\frac{\lambda}{n}X_k\right\} = \exp\{-\lambda x\}\left[M\left(\frac{\lambda}{n}\right)\right]^n \\ &= \exp\left\{-\lambda x + n\Lambda\left(\frac{\lambda}{n}\right)\right\}\end{aligned}$$

where we have used the inequality $1\{x \geq 0\} \leq \exp\{\lambda x\}$ for $\lambda \geq 0$. As $\lambda \geq 0$ is arbitrary, we have:

$$\begin{aligned}\mathbb{P}\left\{\overline{X}_n \in [x,\infty)\right\} &\leq \exp\left\{\inf_{\lambda\geq 0}\left[-\lambda x + n\Lambda\left(\frac{\lambda}{n}\right)\right]\right\} = \exp\left\{-\sup_{\lambda\geq 0}\left[\lambda x - n\Lambda\left(\frac{\lambda}{n}\right)\right]\right\} \\ &= \exp\left\{-n\sup_{\eta\geq 0}\left[\eta x - \Lambda(\eta)\right]\right\} = \exp\{-n\Lambda^\star(x)\}\end{aligned}$$

or $\frac{1}{n}\ln\mathbb{P}\{\overline{X}_n \in [x,\infty)\} \leq -\Lambda^\star(x)$. If $\eta^\star$ is the dominating point, i.e. the point $\eta^\star$ at which $\eta^\star x - \Lambda(\eta^\star) = \sup_{\eta\geq 0}[\eta x - \Lambda(\eta)]$, we have:

$$\mathbb{P}\left\{\overline{X}_n \in [x,\infty)\right\} \leq \exp\{-n[\eta^\star x - \Lambda(\eta^\star)]\}.$$

This justifies the logarithmic asymptotics of the probabilities and explains the rationale behind the assumptions we are going to put forward. Indeed, they guarantee that a dominating point exists or, equivalently, that $\Lambda(\cdot)$ is sufficiently well behaved that $\sup_{\eta\geq 0}[\eta x - \Lambda(\eta)]$ takes a finite value.

Now we turn to the assumptions. The following one contains the basic properties of the distances. It may be useful to explain its rationale. Each simulation run $\mathbf{z}_j(\theta_i)$ is independent of all the other runs for each $j \in \{1,\dots,n\}$ and $i \in \mathcal{M}_0$. For fixed $i \in \mathcal{M}_0$, the simulation runs $\mathbf{z}_j(\theta_i)$ are also identically distributed across $j \in \{1,\dots,n\}$. As a result, the distances $D_{j,i}$ share the same independence properties. This is sufficient to yield consistency and measurability of the estimator $\widehat{i}_n$ (see Choirat and Seri, 2012, Proposition 1, p. 280 for a slightly different result).

A1 For $k = 1,\dots,n$ the vectors $\mathbf{D}_k$ are independent and identically distributed. For each $j \in \mathcal{M}^0$, the distances $D_{k,j}$ are independent. The mean $\mathbb{E}D_{k,j}$ exists and is finite for each $j \in \mathcal{M}^0$.

As briefly explained above, in order to obtain a LDP we need some functions related to the vector $\mathbf{D}_k$. We define the MGF and CGF of $\mathbf{D}_k$:

$$M(\mathbf{u}) = \mathbb{E}\exp\left\{\mathbf{u}'\mathbf{D}_k\right\} = \mathbb{E}\exp\left\{\sum_{\ell=1}^{m_0} u_\ell D_{k,\ell}\right\}$$
$$= \prod_{\ell=1}^{m_0} \mathbb{E}\exp\{u_\ell D_{k,\ell}\} = \prod_{\ell=1}^{m_0} M_\ell(u_\ell)$$

and:

$$\Lambda(\mathbf{u}) = \sum_{\ell=1}^{m_0} \ln M_\ell(u_\ell) = \sum_{\ell=1}^{m_0} \Lambda_\ell(u_\ell).$$

We introduce the Legendre-Fenchel or Cramér transform of Λ as:

$$\Lambda^\star(\mathbf{y}) := \sup_{\mathbf{u}\in\mathbb{R}^{m_0}} \left\{\mathbf{y}'\mathbf{u} - \Lambda(\mathbf{u})\right\} = \sup_{\mathbf{u}\in\mathbb{R}^{m_0}} \sum_{\ell=1}^{m_0} \{y_\ell u_\ell - \Lambda_\ell(u_\ell)\}$$
$$= \sum_{\ell=1}^{m_0} \sup_{u_\ell\in\mathbb{R}} \{y_\ell u_\ell - \Lambda_\ell(u_\ell)\} = \sum_{\ell=1}^{m_0} \Lambda_\ell^\star(y_\ell).$$

We now state three technical assumptions. The first one (A2) guarantees that the level sets of $\Lambda^\star$, i.e. the sets $\{\mathbf{u} \in \mathbb{R}^d : \Lambda^\star(\mathbf{u}) \leq \alpha\}$, are compact, a property that is useful in order to compute the infimum of the rate function over a set (see Theorem 1). The second one (A3) forces the upper and the lower bounds to be equal. The third one (A4) guarantees that, for any two parameter values θ_j and θ_i, the supports of the distances $D_{k,j}$ and $D_{k,i}$ intersect, otherwise the probabilities involving them may end up to be identically equal to 0 or 1. More complete discussions of these assumptions are contained in the paragraphs following them.

A2 There exists $\delta > 0$ such that, for any $\eta \in (-\delta, +\delta)$, $\Lambda_\ell(\eta) < \infty$ for any $\ell \in \mathcal{M}^0$.

It may be interesting to discuss this assumption in greater detail. It is well known that the CGF at the origin is 0, as the MGF is equal to 1. This assumption requires something more, i.e. that the MGF exists in a neighborhood of the origin, and can be shown (see, e.g., Choirat and Seri, 2012, Lemma 1, p. 281) to be equivalent to the requirement that the so-called Cramér condition $\mathbf{0} \in \text{int}\mathcal{D}(\Lambda)$ holds true. This ensures that the level sets of $\Lambda^\star$, i.e. the sets $\{\mathbf{u} \in \mathbb{R}^d : \Lambda^\star(\mathbf{u}) \leq \alpha\}$, are compact. See Ney and Robinson (1995) for large deviations principles without this assumption.

A3 Let $\mathcal{D}_\ell := \text{int}\{u \in \mathbb{R} : \Lambda_\ell(u) < \infty\}$. For any $\ell \in \mathcal{M}^0$, the function Λ_ℓ is *steep*, i.e. for any sequence $\{u_m\}$ in $\text{int}\mathcal{D}_\ell$ converging to a boundary point of $\mathcal{D}_\ell$:

$$\lim_m \left|\Lambda'_\ell(u_m)\right| = \infty.$$

If the aim of the analysis is to bracket asymptotically the quantity $\frac{1}{n}\ln\mathbb{P}\{\widehat{\theta} = \theta_j\}$ between two constants, one can dispense with the previous assumption. Otherwise, if the aim is to obtain a more precise limit, this assumption is not necessary but very comfortable, as the results that do not use it are generally quite complex (see, e.g., Comman, 2009). This can be linked to another problem arising in large deviations. Consider the scalar case and suppose that Λ is not steep, i.e. $\lim_m |\Lambda'(u_m)| = \lambda < \infty$ for $u_m \to u_\infty$ with $u_\infty \in \partial\mathcal{D}(\Lambda)$. Convex conjugacy implies that $\Lambda'(u) = y$ is equivalent to $u = \Lambda^{\star\prime}(y)$. This implies that $\Lambda^\star$ is linear with slope u_∞ for $y > \lambda$, if u_∞ is the right endpoint of $\mathcal{D}(\Lambda)$, or for $y < -\lambda$, if u_∞ is the left endpoint of $\mathcal{D}(\Lambda)$. Therefore, the rate function $\Lambda^\star$ is not strictly convex. In this case, large deviations principles are not guaranteed to hold (see, e.g., De Marco et al., 2016).

A4 Let $[L_h, U_h]$ be the closure of the convex hull of the support of the law of $D_{k,h}$ for $h \in \mathcal{M}^0$. Let j be the index of the parameter under scrutiny, i.e. $\mathbb{P}\{\widehat{\theta} = \theta_j\}$. Then $U_\ell > L_j$ for any $\ell \in \mathcal{M}^0$.

In order to see what can go wrong when this assumption is not verified, consider the case when $\mathcal{M}^0 = \{1, 2\}$. Suppose that $L_2 > U_1$ or, equivalently, that $[U_2, L_2] \cap [U_1, L_1] = \emptyset$. Then, for any n:

$$\mathbb{P}\left\{\widehat{i}_n = 1\right\} = \mathbb{P}\left\{\widehat{\theta}_n = \theta_1\right\} = \mathbb{P}\left\{\overline{D}_{n,1} \leq \overline{D}_{n,2}\right\} = 1$$

and $\mathbb{P}\{\widehat{i}_n = 2\} = 0$. This implies that $\frac{1}{n}\ln\mathbb{P}\{\widehat{i}_n = 2\} = -\infty$ and the LDP does not apply. This also suggests that combinations of parameters for which the hypothesis is not verified can be safely removed from $\mathcal{M}^0$ as they dominate or are dominated by the other ones.

Theorem 1. *Suppose that $j \notin \mathcal{M}^\star$. Then, under A1:*

$$\liminf \frac{1}{n}\ln\mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} \geq -\inf_{\mathbf{y}\in\mathrm{int}\mathcal{P}_j} \Lambda^\star(\mathbf{y}).$$

Under A1-A2:

$$\limsup \frac{1}{n}\ln\mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} \leq -\inf_{\mathbf{y}\in\overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y}).$$

Under A1-A4:

$$\lim \frac{1}{n}\ln\mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} = -\inf_{\mathbf{y}\in\mathrm{int}\mathcal{P}_j} \Lambda^\star(\mathbf{y}) = -\inf_{\mathbf{y}\in\overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y}).$$

Under A1-A4:

$$\inf_{\mathbf{y}\in\overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y}) = \widetilde{\mathbf{u}}'\widetilde{\mathbf{y}} - \Lambda(\widetilde{\mathbf{u}})$$

where:

- $\widetilde{\mathbf{y}} \in \partial\mathcal{P}_j$;
- *the equation* $(\Lambda')^{-1}(\widetilde{\mathbf{y}}) = \widetilde{\mathbf{u}}$ *has a unique solution* $\widetilde{\mathbf{u}}$;
- $\mathcal{P}_j \subset H^+(\widetilde{\mathbf{u}}, \widetilde{\mathbf{y}})$.

Remark 1. (i) The requirement that $j \notin \mathcal{M}^\star$ is quite natural. Indeed, if $j \in \mathcal{M}^\star$, $\overline{D}_j = \min_{i\in\mathcal{M}^0}\overline{D}_i$ and $\overline{\mathbf{D}} \in \overline{\mathcal{P}}_j$. Now, $\Lambda^\star(\overline{\mathbf{D}}) = 0$ and $\Lambda^\star(\mathbf{y}) > 0$ for any $\mathbf{y} \neq \overline{\mathbf{D}}$. Therefore, if $j \in \mathcal{M}^\star$:

$$\lim \frac{1}{n}\ln\mathbb{P}\{\widehat{\theta} = \theta_j\} = -\inf_{\mathbf{y}\in\overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y}) = -\Lambda^\star\left(\overline{\mathbf{D}}\right) = 0.$$

(ii) It is easy to see why $\inf_{\mathbf{y}\in\mathrm{int}\mathcal{P}_j} \Lambda^\star(\mathbf{y}) = \inf_{\mathbf{y}\in\overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y})$ provides a measure of improbability. Each probability $\mathbb{P}\{\widehat{\theta} = \theta_j\}$ is associated with the infimum of the same function $\Lambda^\star(\cdot)$ over a different set $\mathcal{P}_j$. Now, the function $\Lambda^\star(\cdot)$ is strictly convex, positive and has a single zero in $\overline{\mathbf{D}}$. When j is such that $\mathcal{P}_j$ is far away from $\overline{\mathbf{D}}$, $\inf_{\mathbf{y}\in\mathcal{P}_j} \Lambda^\star(\mathbf{y})$ will be larger than 0. The farther away from $\overline{\mathbf{D}}$ is $\mathcal{P}_j$, the larger is $\inf_{\mathbf{y}\in\mathcal{P}_j} \Lambda^\star(\mathbf{y})$, and the faster is the convergence of $\mathbb{P}\{\widehat{\theta} = \theta_j\}$ to 0. This establishes a relation connecting the distance between $\mathcal{P}_j$ and $\overline{\mathbf{D}}$, on the one hand, and the rate of decrease of $\mathbb{P}\{\widehat{\theta} = \theta_j\}$.
(iii) This kind of large deviations result provides a formula for the logarithm of the probability. In a similar context, it was shown by Choirat and Seri (2012) that exact or sharp large deviations results and saddlepoint approximations were able to provide valid approximations for the probabilities (and not only for their logarithms). However, in this paper we do not investigate exact large deviations or saddlepoint approximations. The reason is that we aim at using them for the computation of some measures of precision, and exact large deviations are not suitable for this purpose (see the end of Section 5).

Example 7. In order to illustrate the principle behind the formulas, we consider the following example. We suppose that $D_{k,1}$ is a sample of n exponential random variables with parameter 2 (and mean 1/2) and $D_{k,2}$ is a sample (independent of the previous one) of n exponential random variables with parameter 1 (and mean 1). We want to study:

$$\mathbb{P}\left\{\overline{D}_{n,1} \geq \overline{D}_{n,2}\right\} = \mathbb{P}\left\{\sum_{k=1}^n D_{k,1} \geq \sum_{k=1}^n D_{k,2}\right\} = \mathbb{P}\left\{\sum_{k=1}^n \mathbf{D}_k \in \mathcal{P}_2\right\},$$

where $\mathbf{D}_k := (D_{k,1}, D_{k,2})'$. We have:

$$M(\mathbf{u}) = \frac{2}{2-u_1}\cdot\frac{1}{1-u_2},$$
$$\Lambda(\mathbf{u}) = \ln 2 - \ln(2-u_1) - \ln(1-u_2)$$

and

$$\Lambda^\star(\mathbf{y}) = 2y_1 - \ln y_1 + y_2 - \ln y_2 - \ln 2e^2.$$

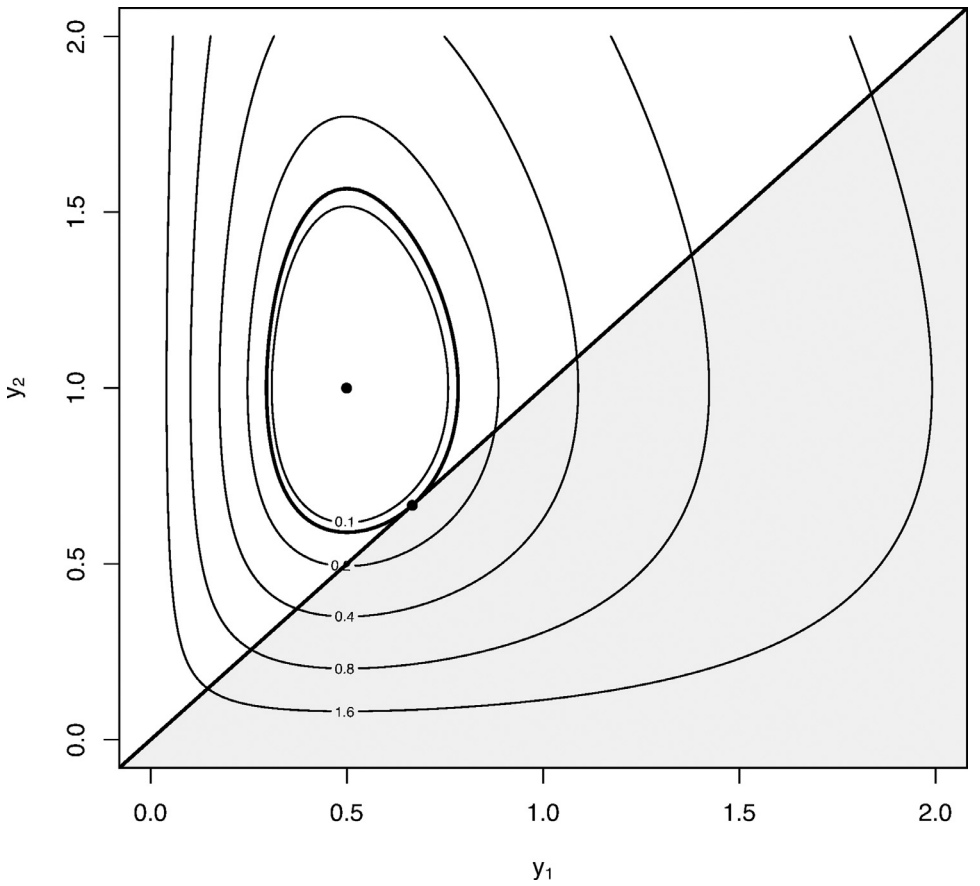

Fig. 2. Example 7: level curves of the function $\Lambda^\star$ (in thin solid lines), level curve corresponding to $\Lambda^\star(\cdot) = \ln(9/8)$ (in thick solid line), points $\widetilde{\mathbf{y}} = (3/2, 3/2)'$ and $\overline{\mathbf{D}} = (1/2, 1)'$, half-plane $\mathcal{P}_2$ (in shaded grey area).

The equation $\widetilde{\mathbf{y}} = \Lambda'(\widetilde{\mathbf{u}})$ is:

$$\begin{cases} \widetilde{y}_1 = \frac{1}{2-\widetilde{u}_1}, \\ \widetilde{y}_2 = \frac{1}{1-\widetilde{u}_2}. \end{cases}$$

As $\widetilde{\mathbf{y}} \in \partial\mathcal{P}_2$, $\widetilde{y}_1 \equiv \widetilde{y}_2$ and $\widetilde{u}_2 = \widetilde{u}_1 - 1$. As $\mathcal{P}_2 \subset H^+(\widetilde{\mathbf{u}}, \widetilde{\mathbf{y}})$, the vector $\widetilde{\mathbf{u}}$ must be normal to the line $\widetilde{y}_1 \equiv \widetilde{y}_2$, or $\widetilde{u}_1 = -\widetilde{u}_2$. The final solution is $\widetilde{u}_1 = 1/2$, $\widetilde{u}_2 = -1/2$, and $\widetilde{y}_1 = \widetilde{y}_2 = \frac{2}{3}$. As a result:

$$\inf_{\mathbf{y} \in \overline{\mathcal{P}}_2} \Lambda^\star(\mathbf{y}) = \Lambda^\star(\widetilde{\mathbf{y}}) = \widetilde{\mathbf{u}}'\widetilde{\mathbf{y}} - \Lambda(\widetilde{\mathbf{u}}) = \ln(9/8) \doteq 0.1177830357.$$

This means that:

$$\lim \frac{1}{n} \ln \mathbb{P}\left\{\overline{D}_{n,1} \geq \overline{D}_{n,2}\right\} = -\ln(9/8).$$

Fig. 2 provides a graphical representation of the function $\Lambda^\star$.

4. The model confidence set

In this section we investigate the construction of a Model Confidence Set, i.e. a subset of $\mathcal{M}^0$ containing the models minimizing the average distance with prescribed probability. We will investigate the properties of the procedure under the assumption that n, the number of runs per each configuration of parameters, diverges.

4.1. The general procedure

The method starts from a set $\mathcal{M}^0 := \{1, \ldots, m_0\}$. In the following, $\mathcal{M} \subset \mathcal{M}^0$ denotes a generic set of discrete parameters. The iterative procedure is based on an equivalence test $\delta_{\mathcal{M}}$ and a selection rule $e_{\mathcal{M}}$, that are associated to the set $\mathcal{M}$.

The test $\delta_{\mathcal{M}}$ is used to test the null hypothesis:

$$\mathsf{H}_{0,\mathcal{M}} : \overline{D}_i = \overline{D}_j, \forall i, j \in \mathcal{M}.$$

We defined above $\mathcal{M}^\star := \{j \in \mathcal{M}^0 : \overline{D}_j = \min_{i\in\mathcal{M}^0} \overline{D}_i\}$. Note that $\mathsf{H}_{0,\mathcal{M}^\star}$ is true while $\mathsf{H}_{0,\mathcal{M}}$ is false whenever $\mathcal{M} \not\subseteq \mathcal{M}^\star$. We also introduce the alternative hypothesis:

$$\mathsf{H}_{A,\mathcal{M}} : \exists i, j \in \mathcal{M} \text{ such that } \overline{D}_i \neq \overline{D}_j.$$

We say that $\delta_{\mathcal{M}} = 1$ when the test rejects the null hypothesis and $\delta_{\mathcal{M}} = 0$ when it does not reject it.

The elimination rule $e_{\mathcal{M}}$ is used to delete an element from $\mathcal{M}$ when $\delta_{\mathcal{M}} = 1$. We suppose that $e_{\mathcal{M}}$ takes its values in $\mathcal{M}$, so that the result of the elimination rule is to pass from $\mathcal{M}$ to $\mathcal{M}\backslash e_{\mathcal{M}}$.

One starts from the set $\mathcal{M} = \mathcal{M}^0$ and performs the test $\delta_{\mathcal{M}} = \delta_{\mathcal{M}^0}$. If the test is rejected, then an elimination step $e_{\mathcal{M}} = e_{\mathcal{M}^0}$ is performed to get a new set $\mathcal{M}_1$. The process is repeated until the test $\delta_{\mathcal{M}}$ does not reject the null hypothesis. The final set of models is called $\widehat{\mathcal{M}}^\star$. If all the tests are performed at the same significance level α, one can explicitly write $\widehat{\mathcal{M}}^\star = \widehat{\mathcal{M}}^\star_{1-\alpha}$.

4.2. The implementation

In order to build a MCS, we have to choose a test procedure $\delta_{\mathcal{M}}$ and an elimination procedure $e_{\mathcal{M}}$.

As a test procedure $\delta_{\mathcal{M}}$, we suppose to estimate the mean $\overline{D}_i := \mathbb{E}_{\mathbf{z}} d(\mathbf{y}, \mathbf{z}(\theta_i))$ and the variance $\sigma_i^2 := \mathbb{V}_{\mathbf{z}}[d(\mathbf{y}, \mathbf{z}(\theta_i))]$ corresponding to each value of θ_i, through Gaussian quasi-likelihood. We will need the following assumptions.

A5 The variances σ_i^2 are finite for any $i \in \mathcal{M}^0$.

Consider the estimators $\overline{D}_{n,i} := \frac{1}{n}\sum_{j=1}^n d(y, z_j(\theta_i))$ and $\widehat{\sigma}_i^2 := \frac{1}{n}\sum_{j=1}^n d^2(y, z_j(\theta_i)) - \overline{D}_{n,i}^2$. Suppose that we want to test that all $\overline{D}_i$'s are equal. We write $\overline{\mathbf{D}} = (\overline{D}_1, \ldots, \overline{D}_m)'$ and

$$\mathbf{\Sigma} := \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m^2 \end{bmatrix}.$$

Let $\mathbf{\Sigma}_{-1}$ be the $(m-1, m-1)$-matrix obtained from $\mathbf{\Sigma}$ removing the first line and column. In the following, an estimator is indicated adding a hat to the same symbol used for the corresponding parameter. Consider the matrix $\mathbf{A}$ defined by:

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & -1 \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{m-1} & -\mathbf{I}_{m-1} \end{bmatrix}.$$

The null hypothesis $\mathsf{H}_{0,\mathcal{M}}$ is that $\mathbf{A}\overline{\mathbf{D}} = \mathbf{0}_{m-1,1}$. The test statistic is:

$$W_{\mathcal{M}} = n\left(\mathbf{A}\overline{\mathbf{D}}_n\right)'\left[\widehat{\sigma}_1^2\mathbf{U}_{m-1} + \widehat{\mathbf{\Sigma}}_{-1}\right]^{-1}\left(\mathbf{A}\overline{\mathbf{D}}_n\right).$$

As an elimination procedure $e_{\mathcal{M}}$, we choose the index $j \in \mathcal{M}$ with the largest value $\overline{D}_{n,j}$, i.e. we identify $e_{\mathcal{M}} := \arg\max_{j\in\mathcal{M}} \overline{D}_{n,j}$.

Theorem 2. *Let the test procedure $\delta_{\mathcal{M}}$ be based on the test $W_{\mathcal{M}}$, with asymptotic distribution $W_{\mathcal{M}} \xrightarrow{\mathcal{D}} \chi^2_{m-1}$, and let the elimination procedure $e_{\mathcal{M}}$ be based on the elimination from $\mathcal{M}$ of the index $j \in \mathcal{M}$ with the largest value $\overline{D}_{n,j}$. Then, under A1 and A5, we have:*

- $\lim_{n\to\infty} \mathbb{P}\{\mathcal{M}^\star \subset \widehat{\mathcal{M}}^\star_{1-\alpha}\} \geq 1-\alpha$,
- $\lim_{n\to\infty} \mathbb{P}\{i \in \widehat{\mathcal{M}}^\star_{1-\alpha}\} = 0 \quad i \notin \mathcal{M}^\star$.

Remark 2. The confidence interval obtained in this way will likely be conservative. Indeed, in Corollary 1 in Hansen et al. (2011, p. 460) it is shown that, when $\mathcal{M}^\star$ is a singleton, $\lim_{n\to\infty} \mathbb{P}\{\mathcal{M}^\star = \widehat{\mathcal{M}}^\star_{1-\alpha}\} = 1$.

There is an alternative way to see the procedure. We set $\mathcal{M}_1 := \mathcal{M}^0$. Then, let us define a sequence of subsets of $\mathcal{M}^0$ through the elimination rule as:

$$\mathcal{M}_{i+1} = \mathcal{M}_i\backslash e_{\mathcal{M}_i} \qquad i = 1, \ldots, m_0 - 1$$

or:

$$\mathcal{M}_i = \left\{e_{\mathcal{M}_i}, e_{\mathcal{M}_{i+1}}, \ldots, e_{\mathcal{M}_{m_0}}\right\}.$$

In our case, this amounts to ordering the elements $\mathcal{M}^0$ according to the value of $\overline{D}_{n,j}$, from the largest to the smallest.

To each element $e_{\mathcal{M}_i}$, we can associate the p-value of the test procedure $\delta_{\mathcal{M}_i}$ to test the null hypothesis $\mathrm{H}_{0,\mathcal{M}_i}$. We call this p-value $p_{\mathrm{H}_{0,\mathcal{M}_i}}$, with the convention that $p_{\mathrm{H}_{0,\mathcal{M}_{m_0}}} \equiv 1$. These p-values are not necessarily decreasing in i. However, it is possible to define an MCS p-value as:

$$\widehat{p}_{e_{\mathcal{M}_j}} := \max_{i \le j} p_{\mathrm{H}_{0,\mathcal{M}_i}}.$$

The interest of the MCS p-values $\widehat{p}_{e_{\mathcal{M}_j}}$, for $j = 1, \ldots, m_0$, is that $i \in \widehat{\mathcal{M}}^{\star}_{1-\alpha}$ if and only if $\widehat{p}_i \ge \alpha$. This allows us to compute the MCS over a range of values α. In this case, the MCS can be used to assess the stability of the optimal solution.

5. Estimation of rate functions

The asymptotic behavior of the probability is dictated by the infimum of the rate function $\Lambda^{\star}(\cdot)$ over the polytope $\mathcal{P}_j$. This infimum can be approximated as in Duffield et al. (1995), Duffy and Metcalfe (2005a; 2005b): in the following we will provide a method of approximation.

5.1. Approximating the rate function

Consider the empirical moment generating function defined by $\widehat{M}(\mathbf{u}) := \prod_{\ell=1}^{m_0} \widehat{M}_\ell(u_\ell)$, where $\widehat{M}_\ell(u_\ell) := \frac{1}{n}\sum_{k=1}^{n} \exp\{u_\ell D_{k,\ell}\}$. Let $\widehat{\Lambda}_\ell(u_\ell) := \ln \widehat{M}_\ell(u_\ell)$ and $\widehat{\Lambda}(\mathbf{u}) := \sum_{\ell=1}^{m_0} \widehat{\Lambda}_\ell(u_\ell)$. We define:

$$\widehat{\Lambda}^{\star}(\mathbf{y}) := \sup_{\mathbf{u}\in\mathbb{R}^{m_0}} \left[\mathbf{u}'\mathbf{y} - \widehat{\Lambda}(\mathbf{u})\right] = \sum_{\ell=1}^{m_0} \widehat{\Lambda}^{\star}_\ell(y_\ell)$$

where $\widehat{\Lambda}^{\star}_\ell(y_\ell) := \sup_{u_\ell\in\mathbb{R}}\{y_\ell u_\ell - \widehat{\Lambda}_\ell(u_\ell)\}$.

We will need the following assumption.

A6 Both L_i and U_i are finite for $i \in \mathcal{M}^0$.

The assumption that the support of the law of $D_{k,i}$ is bounded can be replaced by an assumption of equi-coercivity on $\widehat{\Lambda}^{\star}_i$ (see Dal Maso, 1993, Chapter 7). As equi-coercivity is rather specialized and difficult to check, we prefer to use the simpler and more manageable assumption of boundedness, that is customary in this literature (Duffy and Metcalfe, 2005a; 2005b).

Theorem 3. *Suppose that $j \notin \mathcal{M}^{\star}$. Under A1-A4 and A6:*

$$\lim_{n\to\infty} \inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^{\star}(\mathbf{y}) = \inf_{\mathbf{y}\in\mathcal{P}_j} \Lambda^{\star}(\mathbf{y})$$

and, for n large enough:

$$\inf_{\mathbf{y}\in\overline{\mathcal{P}_j}} \widehat{\Lambda}^{\star}(\mathbf{y}) = \widehat{\mathbf{u}}'\widehat{\mathbf{y}} - \widehat{\Lambda}(\widehat{\mathbf{u}})$$

where:

- $\widehat{\mathbf{y}} \in \partial\mathcal{P}_j$;
- *the equation* $(\widehat{\Lambda}')^{-1}(\widehat{\mathbf{y}}) = \widehat{\mathbf{u}}$ *has a unique solution* $\widehat{\mathbf{u}}$;
- $\mathcal{P}_j \subset H^{+}(\widehat{\mathbf{u}}, \widehat{\mathbf{y}})$.

This approximation can be studied from two further points of view. First of all, computational issues arise in the identification of the rate function and of the dominating point. Second, one may be interested in its statistical properties and, in particular, its rate of convergence. We will investigate these points in the following sections.

5.2. Computation of the rate function

In this paper, we will estimate $\widehat{\Lambda}^{\star}_\ell$ through one of the algorithms in Lucet (1997) and Lucet (2009), in particular the LLT algorithm of Lucet (1997). The name indicates that the algorithmic complexity of the algorithm, as measured by the number of operations, is a linear function of the number of function evaluations. In particular, if a grid of q points is used to approximate $\widehat{\Lambda}_\ell$ and if one needs the values of $\widehat{\Lambda}^{\star}_\ell$ on a grid of p points, the worst-case time complexity of the LLT algorithm is $O(q + p)$. A graphical representation of the algorithm is in Fig. 3 (see below for details). The figure represents the case in which $\Lambda^{\star}$ is estimated using $n = 10$ observations extracted from an exponential distribution with parameter 1. Note that, when referring to the figure, we remove the index ℓ from functions and scalars.

Let us start with $\widehat{\Lambda}_\ell(u) = \ln \widehat{M}_\ell(u) = \ln \frac{1}{n}\sum_{k=1}^{n} \exp\{uD_{k,\ell}\}$ (here and in the following we drop the index ℓ from the arguments u_ℓ and y_ℓ). It is clear that:

$$\ln \frac{1}{n} \exp\left\{u \max_k D_{k,\ell}\right\} \le \widehat{\Lambda}_\ell(u) \le \ln \frac{1}{n} \sum_{k=1}^{n} \exp\left\{u \max_k D_{k,\ell}\right\}$$

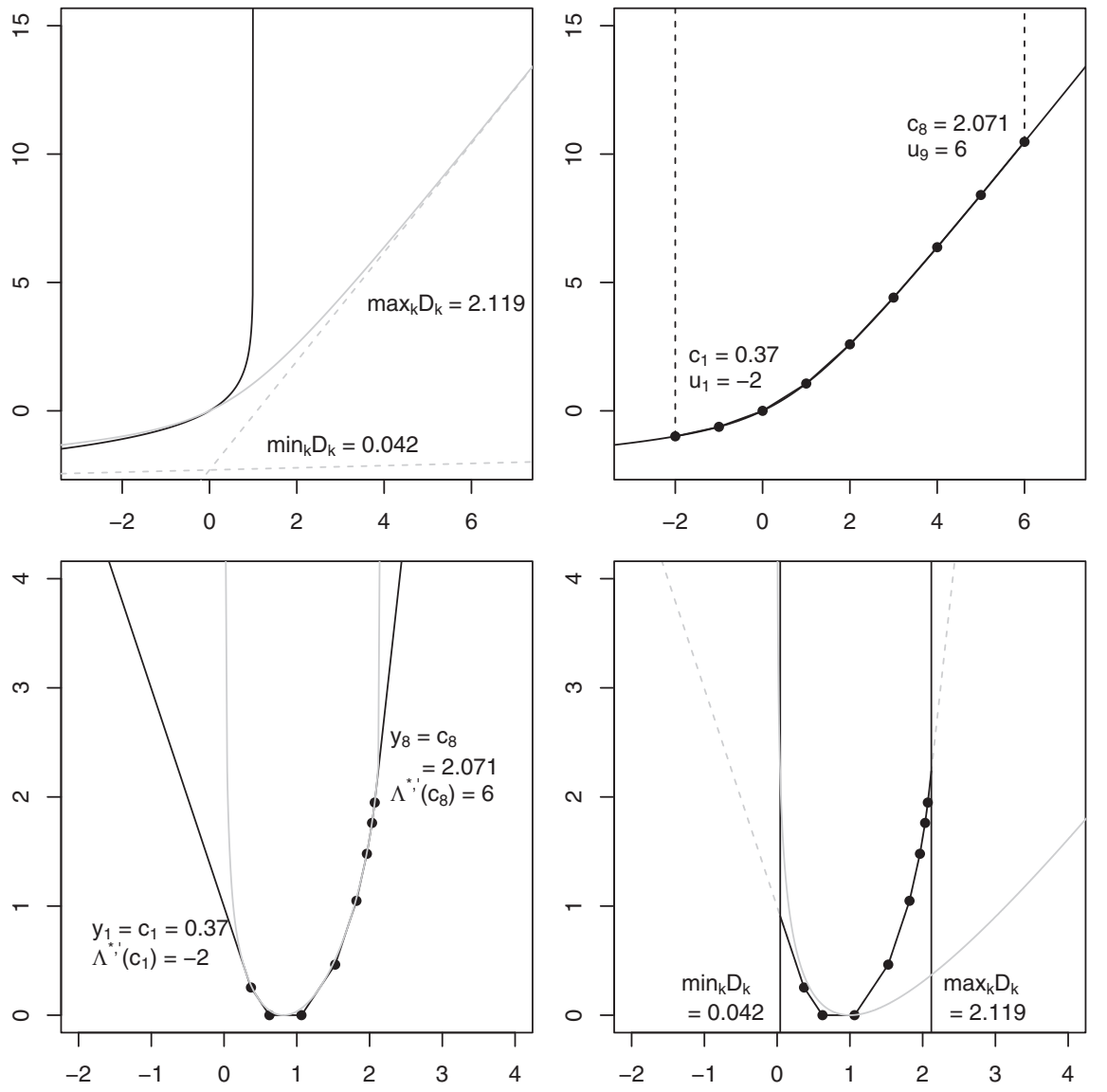

Fig. 3. Illustration of the LLT algorithm with $q = 9$ for $n = 10$ observations extracted from an exponential distribution with parameter 1.
Upper-left quadrant: CGF $\Lambda(u) = -\ln(1-u)$ (in black), estimated CGF $\widehat{\Lambda}(u)$ (in grey), two asymptotes with slope $\max_k D_k$ and $\min_k D_k$.
Upper-right quadrant: points $\{(u_i, \widehat{\Lambda}(u_i)), i = 1, \ldots, 9\}$, piecewise linear approximation $\widetilde{\widehat{\Lambda}}$ to the empirical CGF $\widehat{\Lambda}$, four values c_1, u_1, c_8 and u_9.
Lower-left quadrant: points $\{(c_i, \widetilde{\widehat{\Lambda}}^{\star}(c_i)), i = 1, \ldots, 9\}$, piecewise linear approximation $\widetilde{\widehat{\Lambda}}^{\star}$ (in black), empirical Legendre transform $\widehat{\Lambda}^{\star}$ (in grey), four values $y_1 := c_1$, $u_1 = \widetilde{\widehat{\Lambda}}^{\star,\prime}(c_1)$, $y_8 := c_8$ and $u_9 = \widetilde{\widehat{\Lambda}}^{\star,\prime}(c_8)$.
Lower-right quadrant: modified version of $\widetilde{\widehat{\Lambda}}^{\star}$ with effective domain equal to $[\min_k D_k, \max_k D_k]$ (in black), true rate function $\Lambda^{\star}(y) = y - 1 - \ln y$ for the exponential case (in grey).

and

$$-\ln n + u \max_k D_{k,\ell} \leq \widehat{\Lambda}_\ell(u) \leq u \max_k D_{k,\ell},$$

so that, for $u \to +\infty$:

$$\widehat{\Lambda}_\ell(u) \sim u \max_k D_{k,\ell}.$$

In the same way, for $u \to -\infty$:

$$\widehat{\Lambda}_\ell(u) \sim u \min_k D_{k,\ell}.$$

It can also be shown that:

$$\lim_{u\to+\infty} \widehat{\Lambda}'_\ell(u) = \max_k D_{k,\ell}$$

and

$$\lim_{u\to-\infty} \widehat{\Lambda}'_\ell(u) = \min_k D_{k,\ell}.$$

Indeed, $\widehat{\Lambda}'_\ell(u) = \frac{\sum_{k=1}^n D_{k,\ell}\exp\{uD_{k,\ell}\}}{\sum_{k=1}^n \exp\{uD_{k,\ell}\}}$ and:

$$\min_k D_{k,\ell} \le \widehat{\Lambda}'_\ell(u) = \frac{\sum_{k=1}^n D_{k,\ell}\exp\{uD_{k,\ell}\}}{\sum_{k=1}^n \exp\{uD_{k,\ell}\}} \le \max_k D_{k,\ell}. \tag{1}$$

We will use this fact later. The upper-left quadrant of Fig. 3 represents, in black, the cumulant generating function in the exponential case $\Lambda(u) = -\ln(1-u)$, and, in grey, the estimated CGF $\widehat{\Lambda}(u)$ and the two asymptotes with slope $\max_k D_k$ and $\min_k D_k$.

Then we discuss the approximation of $\widehat{\Lambda}^\star_\ell(y)$. We recall that:

$$\widehat{\Lambda}^\star_\ell(y) = \sup_{u\in\mathbb{R}} \left\{yu - \widehat{\Lambda}_\ell(u)\right\}. \tag{2}$$

The main problem is that the functions $\widehat{\Lambda}^\star_\ell$ involve an optimization step that can turn out to be computationally intensive, especially when it has to be repeated for several values of y.

Most algorithms use therefore a property of the functions to avoid the optimization step. Indeed, convex conjugacy implies that $\widehat{\Lambda}'_\ell(u) = y$ is equivalent to $u = \widehat{\Lambda}^{\star\prime}_\ell(y)$. This means that, if for a fixed u we are able to find $y := \widehat{\Lambda}'_\ell(u)$, we can identify:

$$\widehat{\Lambda}^\star_\ell(y) = u\widehat{\Lambda}'_\ell(u) - \widehat{\Lambda}_\ell(u).$$

The LLT belongs to a whole class of methods based on the *discrete Legendre transform* (DLT), i.e. the replacement in (2) of the maximum over $\mathbb{R}$ with the maximum over a finite set $\{u_1,\dots,u_q\}$ (see Step 0 in Lucet, 1997, p. 176):

$$\widehat{\Lambda}^\star_\ell(y) \simeq \sup_{u\in\{u_1,\dots,u_q\}} \left\{yu - \widehat{\Lambda}_\ell(u)\right\}.$$

In the following, we suppose that $u_i < u_{i+1}$ for all i.

Step 1 in Lucet (1997, p. 176), i.e. the convexification of the function $\widehat{\Lambda}_\ell$ through the construction of the convex hull of the point-set $\{(u_i, \widehat{\Lambda}_\ell(u_i)), i = 1,\dots,q\}$, can be skipped as the function $\widehat{\Lambda}_\ell$ is already convex. Therefore, the function $\widehat{\Lambda}_\ell$ can be approximated on the interval $[u_1, u_q]$ by a piecewise linear function $\widetilde{\widehat{\Lambda}}_\ell$ passing through the points $\{(u_i, \widehat{\Lambda}_\ell(u_i)), i = 1,\dots,q\}$. As our aim is to identify $y := \widehat{\Lambda}'_\ell(u)$, we introduce the slopes of the approximated function $\widetilde{\widehat{\Lambda}}_\ell$ between u_i and u_{i+1}:

$$c_i := \frac{\widehat{\Lambda}_\ell(u_{i+1}) - \widehat{\Lambda}_\ell(u_i)}{u_{i+1} - u_i}, \quad i = 1,\dots,q-1.$$

At the points u_i, for $i = 1,\dots,q$, the function $\widetilde{\widehat{\Lambda}}_\ell$ is not differentiable. The upper-right quadrant of Fig. 3 represents the points $\{(u_i, \widehat{\Lambda}(u_i)), i = 1,\dots,9\}$, the piecewise linear approximation $\widetilde{\widehat{\Lambda}}$ to the empirical CGF $\widehat{\Lambda}$, and the four values c_1, u_1, c_8 and u_9.

Now (see Step 2 in Lucet, 1997, p. 176), for any $c_{i-1} < y < c_i$:

$$\widetilde{\widehat{\Lambda}}^\star_\ell(y) = yu_i - \widehat{\Lambda}_\ell(u_i)$$

so that the function $\widehat{\Lambda}^\star_\ell(y)$ is approximated by a piecewise linear function with intercept $-\widehat{\Lambda}_\ell(u_i)$ and slope u_i over (c_{i-1}, c_i). When $y = c_i$, using the definition of c_i:

$$\widetilde{\widehat{\Lambda}}^\star_\ell(y) = yu_i - \widehat{\Lambda}_\ell(u_i) = yu_{i+1} - \widehat{\Lambda}_\ell(u_{i+1})$$

so that it is immaterial whether we use u_i or u_{i+1} when computing the function. The lower-left quadrant of Fig. 3 represents the points $\{(c_i, \widetilde{\widehat{\Lambda}}^\star(c_i)), i = 1,\dots,9\}$, the piecewise linear approximation $\widetilde{\widehat{\Lambda}}^\star$ (in black), the true empirical Legendre transform $\widehat{\Lambda}^\star$ (in grey) and the four values $y_1 := c_1$, $u_1 = \widetilde{\widehat{\Lambda}}^{\star,\prime}_\ell(c_1)$, $y_8 := c_8$ and $u_9 = \widetilde{\widehat{\Lambda}}^{\star,\prime}_\ell(c_8)$.

Now we see what happens at the boundary of the function $\widetilde{\widehat{\Lambda}}^\star_\ell$ (this topic does not seem to be covered in detail in Lucet, 1997).

The original algorithm does not directly approximate the function $\widehat{\Lambda}_\ell$ but rather the function $\widehat{\Lambda}_\ell + \chi_{[u_1,u_q]}$. Thus, $\widehat{\Lambda}_\ell + \chi_{[u_1,u_q]}$ is a version of $\widehat{\Lambda}_\ell$ taking the value $+\infty$ outside the interval $[u_1, u_q]$. The slopes of $\widehat{\Lambda}_\ell$ at u_1 and u_q are respectively $\widehat{\Lambda}'_\ell(u_1)$ and $\widehat{\Lambda}'_\ell(u_q)$, but are approximated by the values c_1 and c_{q-1}. As the function $\widehat{\Lambda}_\ell$ is convex, $\widehat{\Lambda}'_\ell(u_q) \ge c_{q-1}$ and $\widehat{\Lambda}'_\ell(u_1) \le c_1$. The property of convex conjugacy implies that the function approximating $\widehat{\Lambda}^\star_\ell(y)$ on the basis of $\widehat{\Lambda}_\ell + \chi_{[u_1,u_q]}$ has slope u_1 over $(-\infty, c_1)$ and slope u_q over $(c_{q-1}, +\infty)$. As $u_1 \to -\infty$ and $u_q \to +\infty$, the slopes diverge but this can

take place quite slowly. This means that the effective domain of $\widetilde{\widehat{\Lambda}}_\ell^\star$ will be $\mathbb{R}$. However, we know (see Azencott, 1980, Proposition 9.7) that the effective domain of $\widehat{\Lambda}_\ell^\star$ is $\mathrm{co}\{D_{k,\ell}, k=1,\dots,n\} = [\min_k D_{k,\ell}, \max_k D_{k,\ell}]$, the convex hull of the points $\{D_{k,\ell}, k=1,\dots,n\}$.

A solution is to suppose that, outside $[u_1, u_q]$, the function $\widetilde{\widehat{\Lambda}}_\ell$ has slopes $\min_k D_{k,\ell}$, over $(-\infty, u_1)$, and $\max_k D_{k,\ell}$, over $(u_q, +\infty)$. Note that, from (1), this does not alter the convexity of $\widetilde{\widehat{\Lambda}}_\ell$, as it will remain true that $\min_k D_{k,\ell} \le c_1 \le \cdots \le c_{q-1} \le \max_k D_{k,\ell}$. This corresponds to setting the effective domain of $\widetilde{\widehat{\Lambda}}_\ell^\star$ to $[\min_k D_{k,\ell}, \max_k D_{k,\ell}]$, so that both $\widetilde{\widehat{\Lambda}}_\ell^\star$ and $\widehat{\Lambda}_\ell^\star$ will have the same effective domain. The lower-right quadrant of Fig. 3 represents, in black, the modified version of $\widetilde{\widehat{\Lambda}}^\star$ as well as, in grey, the true rate function $\Lambda^\star(y) = y - 1 - \ln y$ for the exponential case.

This can also be used as a check for the choice of u_1 and u_q. Indeed, if u_1 and u_q are large enough, $c_1 - \min_k D_{k,\ell}$ and $\max_k D_{k,\ell} - c_{q-1}$ should be small. If this is not the case, one should move u_1 and u_q to achieve smaller values of these two quantities.

5.3. Computation of the constrained minimum

Once each $\widehat{\Lambda}_\ell^\star$ has been computed, we need to obtain $\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$, i.e. to optimize it under a constraint.

In order to avoid the inequality constraints, we use the technique of Box (1966, pp. 72-73). The replacements:

$$\begin{cases} y_j \mapsto x_j \\ y_\ell \mapsto x_j + x_\ell^2 \end{cases}$$

transform the computation of $\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ into an unconstrained optimization problem:

$$\inf_{\mathbf{x}\in\mathbb{R}^{m_0}} \sum_{1\le\ell\le m_0, \ell\ne j} \widehat{\Lambda}_\ell^\star\left(x_j + x_\ell^2\right) + \widehat{\Lambda}_j^\star\left(x_j\right).$$

Note that the objective of this technique is not to identify on which face of $\mathcal{P}_j$ the infimum is achieved, but to get a reasonably accurate value of $\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ without using inequality-constrained algorithms.

Example 8. We continue Example 7. The replacement above provides:

$$\Lambda^\star\left(\begin{bmatrix} x_2 + x_1^2 \\ x_2 \end{bmatrix}\right) = 2\left(x_2 + x_1^2\right) - \ln\left(x_2 + x_1^2\right) + x_2 - \ln x_2 - \ln 2e^2.$$

Note that the non-negativity constraint $x_2 > 0$ does not pose any problem, as the function is not defined for $x_2 \le 0$. The level curves of the function are represented in Fig. 4. The point $\widetilde{\mathbf{x}} = (0, 3/2)'$ corresponds to $\widetilde{\mathbf{y}} = (3/2, 3/2)'$ in the new coordinates.

5.4. Asymptotic error of the rate function

In this paper we do not provide a full statistical analysis of this method. However, it is not difficult to obtain some hints about the behavior of the solution. The minimum $\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ is reached in a point $\widehat{\mathbf{y}}$ converging to the point $\widetilde{\mathbf{y}}$ at which $\inf_{\mathbf{y}\in\overline{\mathcal{P}}_j} \Lambda^\star(\mathbf{y})$ is reached. Therefore:

$$\left|\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y}) - \inf_{\mathbf{y}\in\mathcal{P}_j} \Lambda^\star(\mathbf{y})\right| \le \sup_{\mathbf{y}\in\mathcal{N}(\widetilde{\mathbf{y}})} \left|\widehat{\Lambda}^\star(\mathbf{y}) - \Lambda^\star(\mathbf{y})\right|$$

where $\mathcal{N}(\widetilde{\mathbf{y}})$ is a neighborhood of $\widetilde{\mathbf{y}}$ contained in the interior of the effective domain of $\Lambda^\star$. Using the reasoning in Ramm and Zaslavsky (1993, Section 4.3), $\sup_{\mathbf{y}\in\mathcal{N}(\widetilde{\mathbf{y}})} |\widehat{\Lambda}^\star(\mathbf{y}) - \Lambda^\star(\mathbf{y})|$ is of the same order as:

$$\sup_{\mathbf{u}\in\mathcal{N}(\widetilde{\mathbf{u}})} \left|\widehat{\Lambda}(\mathbf{u}) - \Lambda(\mathbf{u})\right| = \sup_{\mathbf{u}\in\mathcal{N}(\widetilde{\mathbf{u}})} \left|\sum_{\ell=1}^{m_0} \widehat{\Lambda}_\ell(u_\ell) - \sum_{\ell=1}^{m_0} \Lambda_\ell(u_\ell)\right| \le \sup_{\mathbf{u}\in\mathcal{N}(\widetilde{\mathbf{u}})} \sum_{\ell=1}^{m_0} \left|\widehat{\Lambda}_\ell(u_\ell) - \Lambda_\ell(u_\ell)\right|.$$

The behavior of $\widehat{\Lambda}_\ell - \Lambda_\ell$ has been considered by Feuerverger (1989) (see also Ben Arous et al., 2005). We first note that, provided $\widehat{M}_\ell$ is a consistent estimator of M_ℓ, the behavior of $\widehat{\Lambda}_\ell - \Lambda_\ell$ is similar to $\widehat{M}_\ell - M_\ell$:

$$\widehat{\Lambda}_\ell - \Lambda_\ell = \ln\left(1 + \frac{\widehat{M}_\ell - M_\ell}{M_\ell}\right) = \frac{\widehat{M}_\ell - M_\ell}{M_\ell} + O\left(\left(\widehat{M}_\ell - M_\ell\right)^2\right).$$

Then, we note that $\mathbb{V}(\widehat{M}_\ell(u_\ell)) = M_\ell(2u_\ell) - M_\ell^2(u_\ell)$ so that, if $M_\ell(2u_\ell)$ is infinite, $\mathbb{V}(\widehat{M}_\ell(u_\ell))$ will be infinite too. This justifies what we write below.

Let I_ℓ be the effective domain of M_ℓ. Theorem 2.3 in Feuerverger (1989) shows that, for any $u_\ell \in \frac{1}{2}\cdot I_\ell$, $\sqrt{n}\cdot(\widehat{M}_\ell - M_\ell)$ converges weakly to a Gaussian process. This means that $\widehat{M}_\ell - M_\ell = O_{\mathbb{P}}(n^{-\frac{1}{2}})$. The convergence is uniform over compact

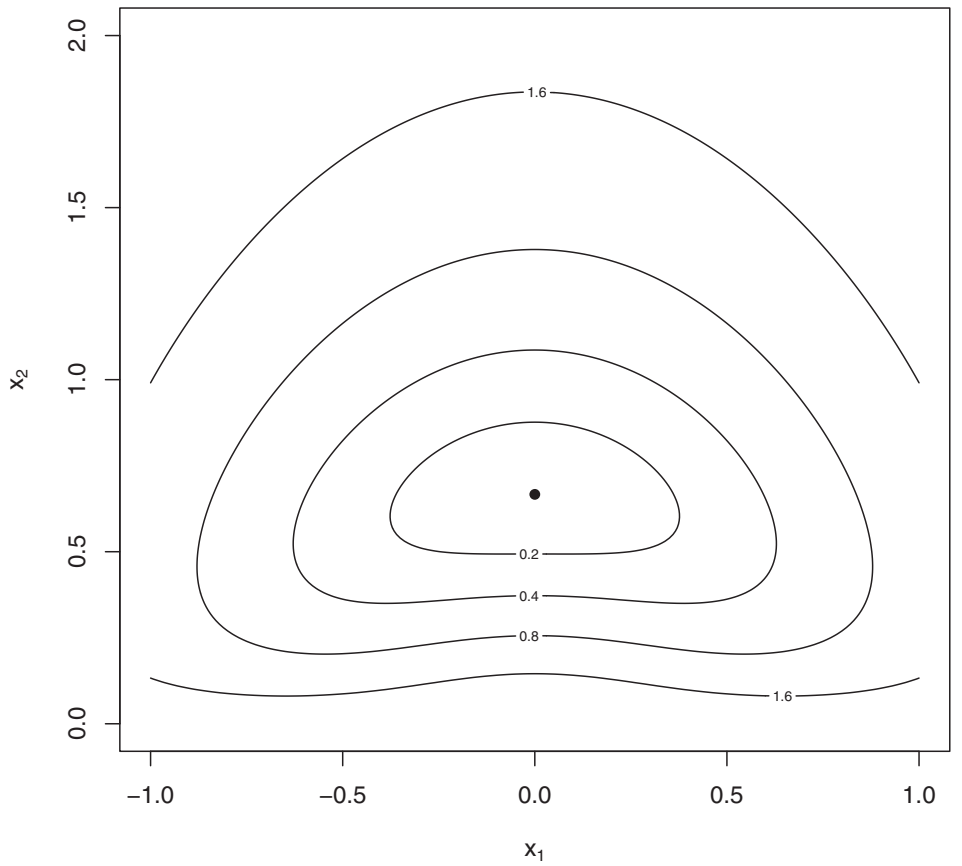

Fig. 4. Example 8: level curves of the function $\Lambda^\star$ in the new coordinate system (in thin solid lines), point $\widetilde{\mathbf{x}} = (0, 3/2)'$.

subsets of $\frac{1}{2} \cdot I_\ell$ and a similar result applies to derivatives. Theorem 2.4 in Feuerverger (1989) extends the result to $\widehat{\Lambda}_\ell - \Lambda_\ell$, that is $\widehat{\Lambda}_\ell - \Lambda_\ell = O_{\mathbb{P}}(n^{-\frac{1}{2}})$. In Feuerverger (1989, p. 460), the interval $\frac{1}{2} \cdot I_\ell$ is called zone of normal convergence.

Consider now any $u_\ell \in I_\ell \backslash (\frac{1}{2} \cdot I_\ell)$. Let $\alpha := C_\ell / u_\ell$ and C_ℓ is the extreme of I_ℓ having the same sign as u_ℓ, i.e. the (left or right) abscissa of convergence of the Laplace-Stieltjes transform. Now, under mild assumptions (see Theorem 1 in Schmidli, 1994), $e^{u_\ell D_{k,\ell}}$ has a tail that is regularly varying with index $-\alpha$. This means that $e^{u_\ell D_{k,\ell}}$ is the domain of attraction of a stable law with exponent α and $n^\delta(\widehat{M}_\ell - M_\ell) \to 0$ and $n^\delta(\widehat{\Lambda}_\ell - \Lambda_\ell) \to 0$ almost surely for any $\delta < \delta_0 = \alpha^{-1} - 1 = -\frac{C_\ell - u_\ell}{C_\ell}$, while the same quantities almost surely diverge for $\delta > \delta_0$ (see Feuerverger, 1989, Theorem 2.5). The result extends to derivatives and holds uniformly over intervals in I_ℓ. Therefore, $\widehat{M}_\ell - M_\ell = o(n^{-\delta})$ and $\widehat{\Lambda}_\ell - \Lambda_\ell = o(n^{-\delta})$ for any $\delta < \delta_0$, where δ_0 is interpreted as the minimum value of $\max\{-\frac{C_\ell - u_\ell}{C_\ell}, -\frac{1}{2}\}$ over the interval.

Now, A6 above implies that the effective domain of the CGF (and of the MGF) is the whole real line. Therefore, also the zone of normal convergence is $\mathbb{R}$, and $\widehat{\Lambda}_\ell - \Lambda_\ell = O_{\mathbb{P}}(n^{-\frac{1}{2}})$ uniformly over compact subsets of $\mathbb{R}$. This means that the mean squared error of $\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$, i.e. the quantity:

$$\mathrm{MSE}\left(\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y}) \right) = \mathbb{E}\left(\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y}) - \inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star(\mathbf{y}) \right)^2$$

is expected to decrease as $O(n^{-1})$. The following example shows that this may take place also when A6 does not hold but the value y corresponds to a value of u that is in the zone of normal convergence.

Example 9. In order to prove that this is the case, we have run a small simulation study. Consider the behavior of the mean of n exponential random variables X_i, $i = 1, \ldots, n$, with parameter 1. We want to study the behavior of the probability $\mathbb{P}\{\frac{1}{n}\sum_{i=1}^n X_i \le \frac{1}{2}\}$ for large n. In this case an exact characterization is possible. Indeed, $\sum_{i=1}^n X_i$ is Gamma distributed with shape n and scale 1. Therefore, the CDF of $\sum_{i=1}^n X_i$ is $\gamma(n, x)/\Gamma(n)$ and:

$$\mathbb{P}\left\{ \frac{1}{n}\sum_{i=1}^n X_i \le \frac{1}{2} \right\} = \frac{\gamma\left(n, \frac{n}{2}\right)}{\Gamma(n)}.$$

Table 1
Example 9: bias, variance and MSE of the estimator $\widehat{\Lambda}^\star(\frac{1}{2})$ of $\Lambda^\star(\frac{1}{2}) = \ln 2 - \frac{1}{2} \doteq 0.1931471806$, i.e. the rate function associated with exponential random variables evaluated at $\frac{1}{2}$, on the basis of 10,000 replications.

n	bias	variance	MSE
10	0.08326409	0.09852722	0.1054601
20	0.03118596	0.02266419	0.02363675
40	0.0135553	0.009384405	0.009568151
80	0.007289302	0.004495908	0.004549042
160	0.003312765	0.002183584	0.002194558
320	0.001757503	0.001064703	0.001067792
640	0.0008518621	0.0005252574	0.000525983

From 8.11(iii) in Olver et al. (2010, p. 180), $\gamma(n, \frac{n}{2}) \sim (\frac{n}{2})^{n-1} e^{-\frac{n}{2}}$ and, from 5.11.3 in Olver et al. (2010, p. 140), $\Gamma(n) \sim e^{-n} n^n (\frac{2\pi}{n})^{\frac{1}{2}}$. At last:

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \leq \frac{1}{2}\right\} \sim \left(\frac{2}{\pi n}\right)^{\frac{1}{2}} e^{-n\left(\ln 2 - \frac{1}{2}\right)}$$

where $\ln 2 - \frac{1}{2} \doteq 0.1931471806$. Large deviations principles give the same solution, but in logarithmic form. Indeed:

$$\lim \frac{1}{n} \ln \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \leq \frac{1}{2}\right\} = \inf_{y \in \left(-\infty, \frac{1}{2}\right]} \Lambda^\star(y) = \Lambda^\star\left(\frac{1}{2}\right)$$

where $\Lambda^\star(y) = y - 1 - \ln y$. Now, $\Lambda^\star(\frac{1}{2}) = \ln 2 - \frac{1}{2} \doteq 0.1931471806$. In Table 1, on the basis of 10,000 replications, we compute the bias, variance and MSE of $\widehat{\Lambda}^\star(\frac{1}{2})$ when $\widehat{\Lambda}^\star$ is based on a sample of n exponential random variables. Apart from an initial transient, each doubling of the sample size leads roughly to a halving of the MSE (and of the variance), thus suggesting that the rate of convergence is indeed $O_{\mathbb{P}}(n^{-\frac{1}{2}})$. Indeed, the value $y = \frac{1}{2}$ corresponds to $\Lambda^{\star\prime}(\frac{1}{2}) = -1$ and this means that $u = -1$ too, that is in the zone of normal convergence $(-\infty, \frac{1}{2})$.

This explains why it is hopeless to use $\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ to reconstruct the probability $\mathbb{P}\{\widehat{i}_n = j\} = \mathbb{P}\{\widehat{\theta} = \theta_j\}$. Indeed, Theorem 1 can be restated as:

$$\frac{1}{n} \ln \mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} = -\inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star(\mathbf{y}) + o(1)$$

but more precise estimates can be obtained (see Ney, 1983; 1984; Iltis, 1995; Choirat and Seri, 2012) as:

$$\frac{c}{n^{\frac{m_0}{2}}} e^{-n \inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star(\mathbf{y})} \leq \mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} \leq \frac{C}{n^{\frac{1}{2}}} e^{-n \inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star(\mathbf{y})}.$$

Note that this is compatible with:

$$\frac{1}{n} \ln \mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} = -\inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star(\mathbf{y}) + O\left(\frac{\ln n}{n}\right).$$

Replacing $\inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star(\mathbf{y})$ with $\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y}) + O_{\mathbb{P}}(n^{-\frac{1}{2}})$ yields:

$$\frac{c}{n^{\frac{m_0}{2}}} e^{-n \inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y}) + O_{\mathbb{P}}\left(n^{\frac{1}{2}}\right)} \leq \mathbb{P}\left\{\widehat{\theta} = \theta_j\right\} \leq \frac{C}{n^{\frac{1}{2}}} e^{-n \inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y}) + O_{\mathbb{P}}\left(n^{\frac{1}{2}}\right)}$$

that is not accurate enough. Nevertheless, the formulas can be used to approximate some useful quantities, such as ratios of logarithms of probabilities that measure their relative rates of decrease. For $i, j \notin \mathcal{M}^\star$:

$$\frac{\ln \mathbb{P}\left\{\widehat{\theta} = \theta_j\right\}}{\ln \mathbb{P}\left\{\widehat{\theta} = \theta_i\right\}} = \frac{\inf_{\mathbf{y} \in \mathcal{P}_j} \Lambda^\star(\mathbf{y})}{\inf_{\mathbf{y} \in \mathcal{P}_i} \Lambda^\star(\mathbf{y})} \left(1 + O\left(\frac{\ln n}{n}\right)\right) = \frac{\inf_{\mathbf{y} \in \mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})}{\inf_{\mathbf{y} \in \mathcal{P}_i} \widehat{\Lambda}^\star(\mathbf{y})} \left(1 + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\right).$$

6. Application

In this section, we consider a published model (see Bardone and Secchi, 2017) available on the online repository OpenABM at https://www.comses.net/codebases/4749/releases/1.0.0/. This is a model of "inquisitiveness in ad hoc teams" and, as already mentioned in the introduction, it has been selected because of (a) theoretical relevance, (b) applicability to economics at large, and (c) computational simulation robustness.

Table 2
Parameter notations and values.

Parameter	Values	Description
problems, $N_{P,0}$	100,300	The number of problems P at time $t = 0$, at the start of the simulation.
problem spin-off, *pso*	4	The maximum rate at which problems can multiply—i.e. spin-off simpler problems.
difficulty, d	$\sim \mathcal{N}(3, 1)$	The level of difficulty each problem is associated with.
decision makers, $N_{dm,0}$	100,300	Number of agents *dm* at time $t = 0$ in the organization.
competence, c	$\sim \mathcal{N}(1, 1.5)$	The level of knowledge that each *dm* carries and that can be applied to any P in order to solve it.
docility, *sodm*	$\sim \mathcal{N}(0, 1)$	Socially-oriented decision making (or, simply, docility) that is associated to each *dm*.
enquiry, e	$\sim \mathcal{N}(0, 1)$	Another characteristic of the *dm* agent, that indicates the extent to which curiosity lead it to explore knowledge of team members other than those of its own team.
inquisitiveness	true/false	A binomial parameter that enables or disables the use of enquiry in *dm*.
range	6,9	The extent to which *dm* explore the environment around them to seek problems P and/or other *dm* to cooperate with.

6.1. Short theoretical background

According to the authors of the model, inquisitiveness is a development of "docility", a concept introduced by Herbert A. Simon (1990, 1993) to indicate individuals that lean on information, advice, recommendations, and suggestions from others to make decisions. As Secchi and Bardone (2009) and Secchi (2016) show, this attitude requires some conditions in place before it is activated. One such conditions is the presence of a *community* of reference—i.e. the idea that individuals must feel like they belong to a group of likeminded others. This could be, for example, the community of mathematical sociologists, or that of a small organization. Or, again, among the many examples, this "community" could also be the work team one belongs to. "Docile" individuals would fit in the team in a way such that there is exchange of information and cognition is very much distributed in an ecological, enacted, embedded sense (see Cowley and Vallée-Tourangeau, 2017). The idea of inquisitiveness comes up when Bardone and Secchi (2017) attempt to break free from the limits of docility. While it can be a good proxy in explaining the dynamics of effective teams, docility could also point at the potentially closed loops in which some may find themselves trapped. This is due to the lack of trust in data coming from outside of the team. When a team member connects and establishes a dialogue with people in other teams then s/he is starting an enquiry. The idea is that such individuals focus on the problem at hand rather than on what can be achieved within the team per se; hence they are driven by a quest for competences, skills, and help that may not be available in just one team. As stated in Bardone and Secchi (2017, p. 68), "[w]e use the word 'inquisitiveness' to refer to an agent who mostly relies on *learning by inquiry* and *open explorations* of his or her own environment, including social channels" (italics in the original text).

6.2. ABM characteristics

The model is based on the theory of docility and attempts to expand it by considering docile (i.e. team-bound), non-docile (i.e. lone wolves), and inquisitive (i.e. docile *sans frontières*) individuals. The aim of the simulation is to understand under which circumstances the inquisitive team member adds something useful to the team. To study this aspect, the model presents problems to individuals and teams and calculates the efficiency with which they are solved. A full description of the model can be found in the paper by Bardone and Secchi (2017) and in the supplementary materials file, available online at the link indicated above. The inquisitiveness ABM can be classified as a highly stochastic simulation. In the following, we have decided to succinctly recall only those features that are relevant to our calculations. The purpose of this exercise is to demonstrate how the Model Confidence Set technique works, not to fully introduce and discuss assumptions and results of a computational simulation model.

Table 2 presents key parameters and describes them shortly. The organizational space—i.e. the simulation environment—features `problems` P that can take any number, but that have been set at {100,300} for this simulation. Each problem is attributed with a `difficulty` level d, distributed normally at random as indicated in Table 2. Very difficult problems (with d larger than 95% the maximum level of d) could, based on a random algorithm, `spin-off` up to 4 other (smaller) problems at each simulation step. Also, a random number of up to 3 very difficult problems increase their `difficulty` level over time at a 2% rate.

The model also features agents, called `decision makers` *dm*; they take two values, {100,300}. Each *dm* has a level of `competence`—distributed normally at random (see Table 2)—that uses to solve the problems that it is set to deal with. When a problem is solved, there is an increase of competence of up to 0.30, when the problem is abandoned (not solved), then there is a decrease in competence of 0.05. In addition to this, `decision makers` are attributed a level of `docility` *sodm* that determines their propensity towards working with others, and a level of `enquiry` that is used to cooperate with coworkers outside of one's team (i.e. an extension of docility, as described above). This latter characteristic is enabled by a binomial parameter `inquisitiveness`, when set to 'on' (or 'true').

Table 3
Definition of the different configurations of parameters.

cop	inquisitiveness	$N_{P,0}$	$N_{dm,0}$	range
1	false	100	100	9
2	false	100	100	6
3	false	100	300	9
4	false	100	300	6
5	false	300	100	9
6	false	300	100	6
7	false	300	300	9
8	false	300	300	6
9	true	100	100	9
10	true	100	100	6
11	true	100	300	9
12	true	100	300	6
13	true	300	100	9
14	true	300	100	6
15	true	300	300	9
16	true	300	300	6

Note: cop: configuration of parameters.

6.3. Rules of interaction

The simulation works with agents moving around in the environment as they attempt to reach the target problem (selected at random). Once agents connect to a problem, they also start a number of interactions with neighbouring agents, forming ad hoc teams. All agents appear in random positions on the environment every time the simulation is ready to start. This allows to generate a random allocation of `decision makers` on a so-called organizational "problem solving space" where `problems` also appear and are found at random—i.e. not as a function of agent's `competence`. When the simulation is performed the appropriate number of times (see below), this stochasticity guarantees a variety of combinations of d, on the one hand, and $sodm$ and c on the other.

While `problems` do not move from their position, `decision makers` do so in a way that sets them to look for problems and move towards them. If they find other problems in their way—i.e. problems that are not their main "goal"—they stop and attempt to solve them as well. They will resume the movement after a solution or abandonment, if that problem is still there. If not, they will look for another to deal with.

As a decision maker deals with a problem, it establishes a link to it. While this happens, it can also establish cooperation links with other decision makers in the radius indicated in Table 2 under parameter `range`. The amount of `competence` that these agents share is proportional to their level of `docility`, such that this latter parameter gives an indication of how much knowledge is actually shared. It is fair to assume that one team member does not transfer its knowledge completely to the rest of the team. Some is transferred while some is kept, for various reasons, tacit and inaccessible to others. Only agents with $sodm > \mu_{sodm} + 0.75 \cdot \sigma_{sodm}$, where μ_{sodm} is the mean docility in the system, and σ_{sodm} is its standard deviation, are sharing significant amounts of their `competence` c. Contrary to the average docile, the way in which competence is used by highly docile agents is incremental and not simply additive. That is, there is an upgrade of the knowledge gained thanks to one's own competence. Agents can make full use of this incremental add-on when the switch `inquisitiveness` is set to true, and for agents mating `docility` with high levels of `enquiry` (see Table 2).

A problem is solved and disappears from the space, when the combination of competences in the team is more than the problem's difficulty. When a problem is too difficult for a team, then each team member evaluates its own contribution and may leave the problem after the efforts have been infused for 20 steps of the simulation.

6.4. Running the simulation model

In Table 3 we describe each one of the 16 configurations of parameters.

The simulation model runs for 300 steps—a "step" could be thought of as an opportunity each agent has to interact with another agent and/or with a problem—and the configurations of parameters (as per Table 3) are $2 \times 2 \times 2 \times 2 = 16$. This factorial design derives from the results shown by the proponents of this model, and has been selected to increase variability in the outcome variable as well as introduce some novelty in the understanding of how this inquisitiveness ABM works. To determine how many times an ABM with a highly stochastic component should be performed, we followed Secchi and Seri (2017) and Seri and Secchi (2017), and calculated power analysis for ANOVA for $\alpha = 0.01, 1 - \beta = 0.95$ and effect size of 0.1, consistent with what found in Bardone and Secchi (2017). As a result, the simulation was performed 200 times per each configuration of parameters, for a total of 3200 runs.

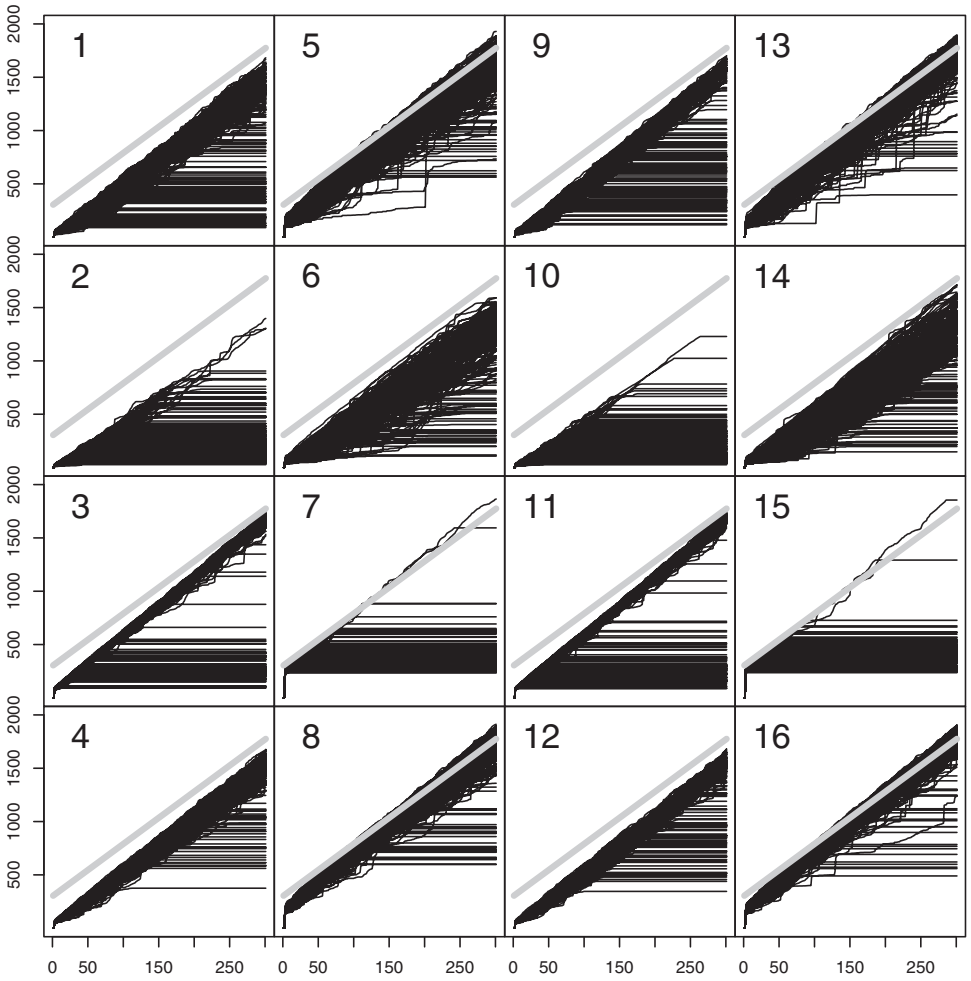

Fig. 5. Trajectories $\mathbf{z}_j(\theta_i)$ for any $j = 1, \ldots, 200$ and $i = 1, \ldots, 16$ (in solid lines) and trajectory $\mathbf{y} = (y_1, \ldots, y_p)$ when $y_h = 300 + 4.9 \cdot h$ (in solid grey line).

6.5. Analyzing the data

In the following, we illustrate the technique outlined above. Since our aim is purely expository, we will consider what happens when the h-th element y_h of $\mathbf{y}$ is $y_h = 300 + 4.9 \cdot h$. This is motivated by the aim to establish an ideal benchmark for the output variable. In so doing, the equation represents an optimal solution threshold that is set at 90% of all problems solved at time 300. Fig. 5 represents the trajectories $\mathbf{z}_j(\theta_i)$ for any $j = 1, \ldots, 200$ and $i = 1, \ldots, 16$, as well as the trajectory $\mathbf{y} = (y_1, \ldots, y_p)$. As a distance, we choose the square of the Euclidean distance, normalized dividing it by $301 \cdot 1{,}000{,}000$, where 301 is the length of the series and 1,000,000 is a normalizing factor. Therefore:

$$d\big(\mathbf{y}, \mathbf{z}_j(\theta_i)\big) = \frac{\sum_{h=1}^{301} \big|y_h - z_{jh}(\theta_i)\big|^2}{301 \cdot 1{,}000{,}000}.$$

These choices respect all the assumptions. A preliminary consideration, that will be used in the following, is that the distances in our example are bounded: indeed, the number of problems to be solved in a finite horizon is bounded and so is the distance between the benchmark and the simulated data. This automatically implies that A2, A3, A5 and A6 are verified. A1 is respected by construction. A4 is verified as shown by an analysis of the data (see also Fig. 7).

In this case we have $\widehat{i}_n = 16$. The construction of the MCS is illustrated in Table 4. The MCS p-values are represented graphically in Fig. 6. The Model Confidence Sets at 95% and at 99% are {8, 13, 16}.

Now we consider the large deviations rate functions associated with each one of the configurations of parameters (apart from $\widehat{i}_n$). We have approximated each $\widehat{\Lambda}_\ell$ on a grid of mesh 0.0005 from -200 to 400. This means that $q = 200{,}001$. The largest value taken by $c_1 - \min_k D_{k,\ell}$ and $\max_k D_{k,\ell} - c_{q-1}$ over $\ell = 1, \ldots, 16$ is 0.005270962, that seems to be small enough.

Table 4
Order of elimination of the different configurations of parameters with means and p-values for $y_h = 300 + 4.9 \cdot h$.

k	$e_{\mathcal{M}_k}$	mean of $e_{\mathcal{M}_k}$	p-value of $\delta_{\mathcal{M}_k}$ ($p_{\mathrm{H}_{0,\mathcal{M}_k}}$)	MCS p-value ($\widehat{p}_{e_{\mathcal{M}_k}}$)
1	10	0.87555	0.00000	0.00000
2	2	0.84660	0.00000	0.00000
3	7	0.63282	0.00000	0.00000
4	15	0.62762	0.00000	0.00000
5	11	0.45591	0.00000	0.00000
6	3	0.37997	0.00000	0.00000
7	9	0.31235	0.00000	0.00000
8	1	0.30822	0.00000	0.00000
9	6	0.26487	0.00000	0.00000
10	14	0.25420	0.00000	0.00000
11	12	0.14426	0.00000	0.00000
12	4	0.10884	0.00000	0.00000
13	5	0.05956	0.00099	0.00099
14	8	0.04722	0.05552	0.05552
15	13	0.04353	0.06277	0.06277
16	16	0.02995	1.00000	1.00000

Note: cop: configuration of parameters.

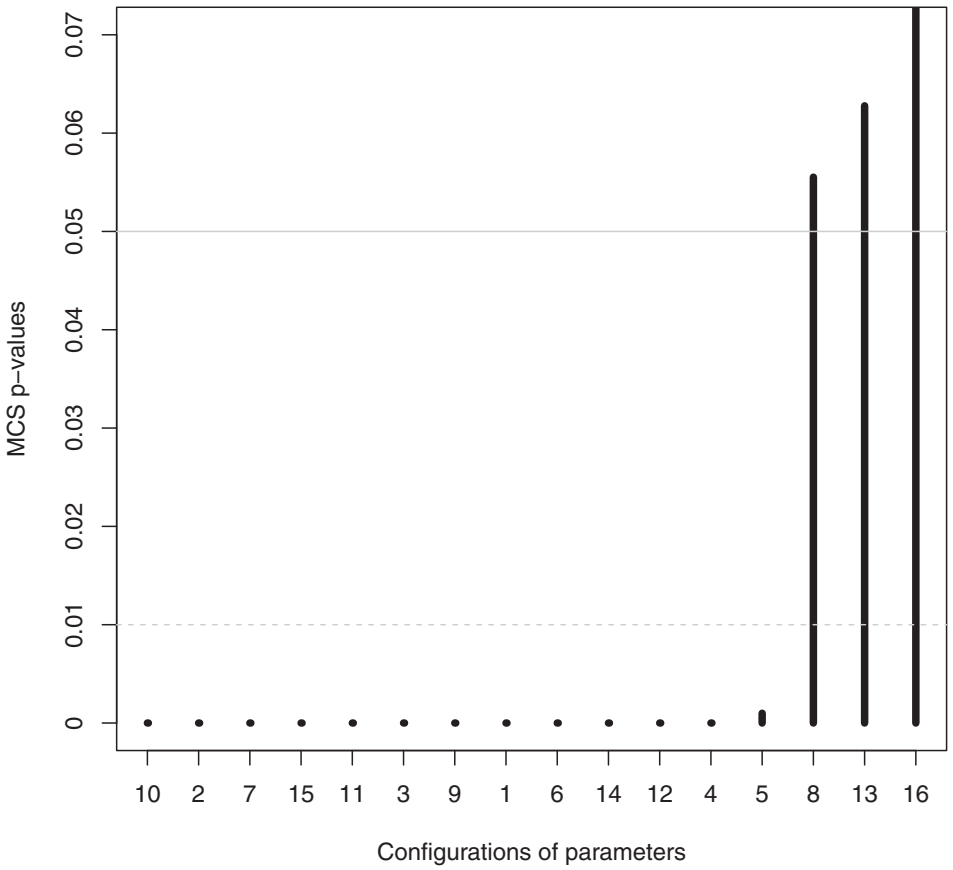

Fig. 6. Graphical representation of the MCS p-values $\widehat{p}_{e_{\mathcal{M}_k}}$ as vertical black lines and of the thresholds as horizontal grey lines.

The functions $\widehat{\Lambda}^{\star}_{\ell}$ are approximated each one on a grid of length 10,000 going from $\min_k D_{k,\ell}$ to $\max_k D_{k,\ell}$. The final results are illustrated in Fig. 7. Note that the $\arg\min$ of each $\widehat{\Lambda}^{\star}_{\ell}$ coincides with $\overline{D}_{n,\ell}$.

Table 5 reports the values of $\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^{\star}(\mathbf{y})$, i.e. the estimated rate of decrease of the probabilities. It can be seen that the accordance with the MCS p-values is extremely good.

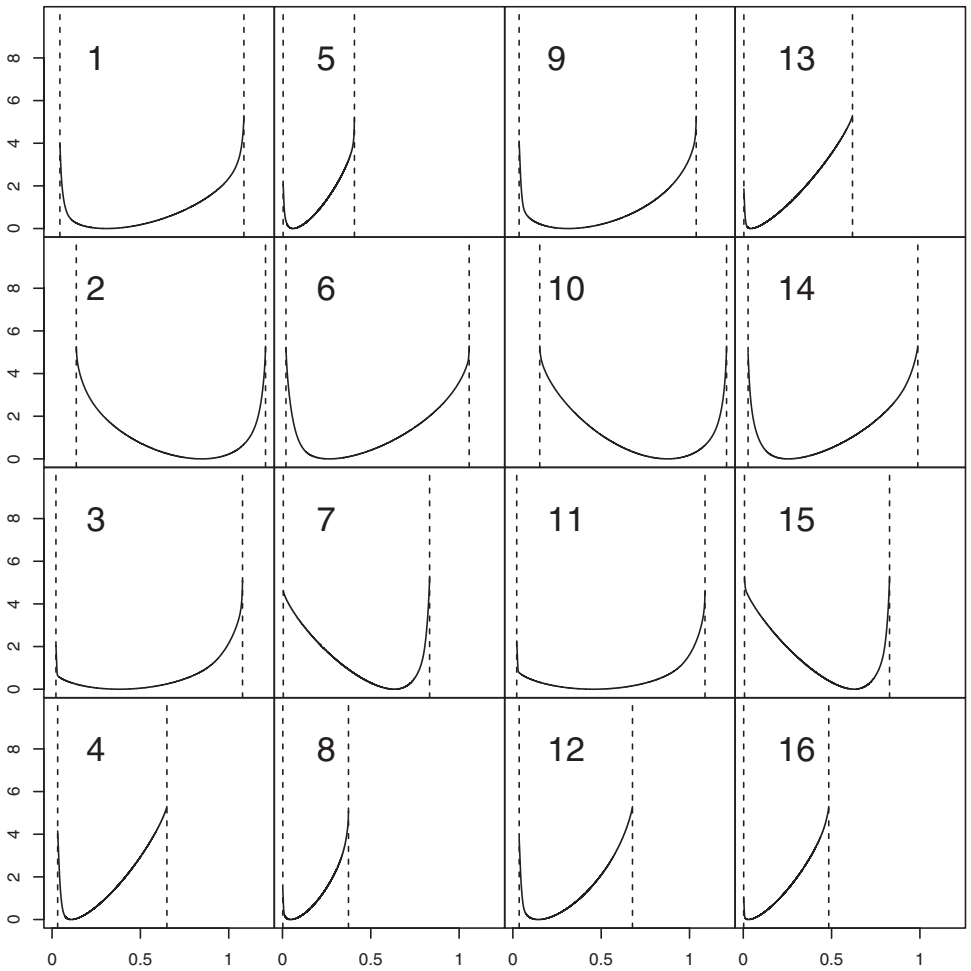

Fig. 7. Rate functions (in solid line) $\widehat{\Lambda}_i^\star$ for all $j = 1, \ldots, 16$; boundaries of the effective domain of $\widehat{\Lambda}_i^\star$ (in dashed vertical lines).

Table 5
Value of the rate function $\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$ for all $j \neq \widehat{i}_n$ ordered from largest to smallest.

j	$\inf_{\mathbf{y}\in\mathcal{P}_j} \widehat{\Lambda}^\star(\mathbf{y})$
10	6.93468540
2	6.14584947
15	3.96367376
7	3.70826307
14	1.35071156
6	1.35043742
1	1.10375307
9	0.91656526
11	0.71091121
3	0.54869452
12	0.49722953
4	0.36842910
5	0.03924210
8	0.01211586
13	0.00860211

7. Conclusions

The standard approach in calibration is to select a unique vector of parameters based on previously available estimations or on the basis of the *wisdom-of-the-crowd*. Assessing the validity of such calibration exercises is daunting as no robustness or sensitivity checks are performed. In this paper, we consider one of these possible checks, based on the comparison of several finite combinations of the model parameters. We propose to use a measure of distance to rank models from the

most plausible to the least plausible. Model Confidence Sets can be employed to further restrict the number of plausible alternatives and provide a sensitivity check for the preferred specification. We also discuss a complementary analysis based on rate functions. The estimation of the latter allows the researcher to assign to all models, except the best one, an estimated rate of decrease of the probability of being the correct model. This approach has comparable interpretation to the Model Confidence Sets, and can be equivalently used to capture the distance between the chosen model and its alternatives, as our empirical application shows.

Declaration of competing interest

The authors of the Manuscript "Calibration and Validation via Confidence Sets" certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgements

The authors are grateful to the editor, the associate editor, and three reviewers, for insightful comments which improved the original manuscript. Moreover, they would like to thank the participants of the CFE'18, EURAM 2019 and CFE'19 conferences, the ABMO4 workshop and the first meeting of the NetCIEx (Network on Counterfactual Impact Evaluation) for helpful discussions. The third and fourth authors gratefully acknowledge financial and logistic support from the DiECO, Università degli Studi dell'Insubria, during a visiting period.

Appendix

Proof of Theorem 1. Let $\mathcal{S}$ be the closure of the convex hull of the support of the law of $\mathbf{D}_k$.

Cramér's Theorem in $\mathbb{R}^d$ (see Corollary 6.1.6 in Dembo and Zeitouni, 2010, p. 253) allows us to derive the first two statements. In particular, the lower bound holds with no supplementary assumption. The upper bound requires the Cramér condition $\mathbf{0} \in \text{int}\mathcal{D}(\Lambda)$ that is verified under A2.

For the third statement, we apply part (ii) in Lemma on page 903 of Ney (1984). Let us start from (I). Under A2, $\text{int}\mathcal{D}(\Lambda)$ is non-empty. On $\text{int}\mathcal{D}(\Lambda)$ the function Λ is differentiable (see Azencott, 1980, Proposition 9.7, p. 49). Under A3, Λ is steep (see Dembo and Zeitouni, 2010, p. 44 for a definition). Therefore, Λ is *essentially smooth* and (I) is verified. According to Corollary 6.1.6 in Dembo and Zeitouni (2010, p. 253), condition (II) is verified if $\mathbf{0} \in \text{int}\mathcal{D}(\Lambda)$, that holds true under A2. As far as (III) is concerned, we first take $B \equiv \text{int}\mathcal{P}_j$. From Azencott (1980, Proposition 9.7, p. 49), $\text{int}\mathcal{S} = \text{int}\mathcal{D}(\Lambda^\star) \subset \mathcal{D}(\Lambda^\star) \subset \overline{\mathcal{S}} = \mathcal{S}$. Therefore, the condition $\text{int}(\text{int}\mathcal{P}_j \cap \mathcal{D}(\Lambda^\star)) \neq \emptyset$ is equivalent to:

$$\begin{aligned}\text{int}\big(\text{int}\mathcal{P}_j \cap \mathcal{D}(\Lambda^\star)\big) &= \text{int}\mathcal{P}_j \cap \text{int}\mathcal{D}(\Lambda^\star)\\ &= \text{int}\mathcal{P}_j \cap \text{int}\mathcal{S} = \text{int}\big(\mathcal{P}_j \cap \mathcal{S}\big) \neq \emptyset.\end{aligned}$$

Under A4, the closure of the convex hull of the support of the law of $\mathbf{D}_k$ is $\prod_{1\le\ell\le m_0}[L_\ell, U_\ell]$. Therefore, $\text{int}\mathcal{S} = \prod_{1\le\ell\le m_0}(L_\ell, U_\ell)$ and A4 implies $\text{int}(\mathcal{P}_j \cap \mathcal{S}) \neq \emptyset$. This implies that $\inf_{\mathbf{y}\in\text{int}\mathcal{P}_j} \Lambda^\star(\mathbf{y})$ is achieved at a unique point belonging to $\overline{\mathcal{P}_j} \cap \text{int}\mathcal{D}(\Lambda^\star) = \overline{\mathcal{P}_j} \cap \text{int}\mathcal{S}$ (see Azencott, 1980, Proposition 9.7, p. 49). The same is true if we take $B \equiv \overline{\mathcal{P}_j}$, thus implying that $\inf_{\mathbf{y}\in\text{int}\mathcal{P}_j} \Lambda^\star(\mathbf{y}) = \inf_{\mathbf{y}\in\overline{\mathcal{P}_j}} \Lambda^\star(\mathbf{y})$.

For the fourth statement, we use the theorem in Ney (1984, p. 904). The conditions are easily verified. Under A2, $\mathcal{D}(\Lambda)$ contains a neighborhood of the origin. The proof that Λ is essentially smooth is provided above. The set $B = \mathcal{P}_j$ is clearly convex. The condition $\text{int}(\mathcal{P}_j \cap \mathcal{S}) \neq \emptyset$ is verified above. At last, it is also clear that $\mathbb{E}\mathbf{D}_k \notin \mathcal{P}_j$. Therefore, there is a dominating point respecting the conditions in the statement (we use here the definition in Ney, 1983, instead of the one in Ney, 1984). □

Proof of Theorem 2. We briefly recall the framework of Hansen et al. (2011). For any $\mathcal{M} \subset \mathcal{M}^0$, we need the following assumptions:

B1 $\limsup_{n\to\infty} \mathbb{P}\{\delta_{\mathcal{M}} = 1 | \mathsf{H}_{0,\mathcal{M}}.\} \le \alpha$.
B2 $\lim_{n\to\infty} \mathbb{P}\{\delta_{\mathcal{M}} = 1 | \mathsf{H}_{A,\mathcal{M}}.\} = 1$.
B3 $\lim_{n\to\infty} \mathbb{P}\{e_{\mathcal{M}} \in \mathcal{M}^\star | \mathsf{H}_{A,\mathcal{M}}.\} = 0$.

Under these conditions, the following results hold (see Theorem 1 in Hansen et al., 2011, p. 459):

- $\lim_{n\to\infty} \mathbb{P}\{\mathcal{M}^\star \subset \widehat{\mathcal{M}}^\star_{1-\alpha}\} \ge 1-\alpha$,
- $\lim_{n\to\infty} \mathbb{P}\{i \in \widehat{\mathcal{M}}^\star_{1-\alpha}\} = 0 \quad i \notin \mathcal{M}^\star$.

Now we turn to the proof in our case. Under A1 and A2 it is trivial to verify that:

$$\sqrt{n}(\overline{\mathbf{D}}_n - \overline{\mathbf{D}}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}_{m,1}, \mathbf{\Sigma}).$$

Under $\mathsf{H}_{0,\mathcal{M}} : \mathbf{A}\overline{\mathbf{D}} = \mathbf{0}_{m-1,1}$, we have:

$$\sqrt{n}\left(\mathbf{A}\overline{\mathbf{D}}_n - \mathbf{A}\overline{\mathbf{D}}\right) = \sqrt{n}\mathbf{A}\overline{\mathbf{D}}_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}_{m-1,1}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}'\right)$$

where:

$$\begin{aligned}\mathbf{A}\mathbf{\Sigma}\mathbf{A}' &= \begin{bmatrix}\mathbf{u}_{m-1} & -\mathbf{I}_{m-1}\end{bmatrix}\begin{bmatrix}\sigma_1^2 & \mathbf{0}_{1,m-1} \\ \mathbf{0}_{m-1,1} & \mathbf{\Sigma}_{-1}\end{bmatrix}\begin{bmatrix}\mathbf{u}'_{m-1} \\ -\mathbf{I}_{m-1}\end{bmatrix} \\ &= \sigma_1^2\mathbf{U}_{m-1} + \mathbf{\Sigma}_{-1}.\end{aligned}$$

Therefore, under the null hypothesis $\mathsf{H}_{0,\mathcal{M}}$, the asymptotic distribution of this test is $W_{\mathcal{M}} \xrightarrow{\mathcal{D}} \chi^2_{m-1}$. This does not change if the σ_i^2's are replaced by estimators. Therefore, Assumption B1 is verified.

Under the alternative $\mathsf{H}_{A,\mathcal{M}}$, i.e. when $\mathbf{A}\overline{\mathbf{D}} \neq \mathbf{0}_{m-1,1}$, the quantity $W_{\mathcal{M}}/n$ converges in probability to $(\mathbf{A}\overline{\mathbf{D}})'[\sigma_1^2\mathbf{U}_{m-1} + \mathbf{\Sigma}_{-1}]^{-1}(\mathbf{A}\overline{\mathbf{D}}) > 0$ and $W_{\mathcal{M}}$ diverges to infinity. Therefore, Assumption B2 is verified.

Under $\mathsf{H}_{A,\mathcal{M}}$, $\mathcal{M} \not\subseteq \mathcal{M}^\star$ and at least an element of $\mathcal{M}$ does not belong to $\mathcal{M}^\star$. It is clear that this element will have an average distance that is larger than the one observed in $\mathcal{M}^\star$. Therefore, selecting the index $j \in \mathcal{M}$ with the largest value $\widehat{\mu}_j$ will eventually lead to eliminate an element in $\mathcal{M}\backslash\mathcal{M}^\star$, as requested by Assumption B3. □

Proof of Theorem 3. As seen above, the moment generating function $M(\mathbf{u})$ can be approximated through $\widehat{M}(\mathbf{u})$. We will define $\widehat{M}_\ell(u_\ell) := \frac{1}{n}\sum_{k=1}^n \exp\{u_\ell D_{k,\ell}\}$. We will introduce the notation $\widehat{\mathbb{P}}_n(\cdot) = \frac{1}{n}\sum_{k=1}^n \delta_{\mathbf{D}_k}(\cdot)$, where δ_x is the Dirac measure in x, and use $\widehat{\mathbb{E}}_n$ for the expectation under $\widehat{\mathbb{P}}_n$.

Using Hess et al. (2010, Section 3), it is simple to see that $\widehat{M}_\ell(u_\ell)$ converges almost surely pointwise to $M_\ell(u_\ell)$. As $\widehat{M}_\ell$ and M_ℓ have in general different effective domains, it is clear that the convergence cannot be uniform. For this reason, we will use epigrahical convergence or, for short, epi-convergence (see, e.g., Dal Maso, 1993; Rockafellar and Wets, 1998) that is especially suitable to analyse the convergence of infima and minimizers of sequences of functions, especially when they have different effective domains (see, e.g., Pennanen and Koivu, 2005; Seri and Choirat, 2013; Hess et al., 2014). Using Theorem 3.1 in Hess and Seri (2019) (or Theorem 2.3 in Choirat et al., 2003), it is trivial to verify that the approximation is not only pointwise but also epigraphically convergent, i.e. $\operatorname{epi}-\lim_n \widehat{M}_\ell(\cdot) = M_\ell(\cdot)$. Now, the very definition of epi-convergence and the strict positivity of $\widehat{M}_\ell(\cdot)$ imply that $\operatorname{epi}-\lim_n \widehat{\Lambda}_\ell(\cdot) = \Lambda_\ell(\cdot)$ where $\widehat{\Lambda}_\ell(\cdot) := \ln \widehat{M}_\ell(\cdot)$. It is in general false that sums of epi-convergent functions are epi-convergent, but exploiting Proposition 6.25 in Dal Maso (1993, p. 64) it is easy to see that $\operatorname{epi}-\lim_n \widehat{\Lambda}(\mathbf{u}) = \Lambda(\mathbf{u})$ where $\Lambda(\mathbf{u}) := \sum_{\ell=1}^{m_0}\Lambda_\ell(u_\ell)$ and $\widehat{\Lambda}(\mathbf{u}) := \sum_{\ell=1}^{m_0}\widehat{\Lambda}_\ell(u_\ell)$. Using the continuity of the Fenchel transform with respect to epi-convergence (Wijsman' s theorem, see Theorem 11.34 in Rockafellar and Wets, 1998, p. 500), this shows that also $\Lambda^\star(\cdot)$ has an epigraphically convergent approximation, in the form:

$$\widehat{\Lambda}^\star(\mathbf{y}) := \sup_{\mathbf{u}\in\mathbb{R}^{m_0}}\left[\mathbf{u}'\mathbf{y} - \widehat{\Lambda}(\mathbf{u})\right].$$

This can at last be written as $\widehat{\Lambda}^\star(\mathbf{y}) = \sum_{\ell=1}^{m_0}\widehat{\Lambda}^\star_\ell(y_\ell)$ where $\widehat{\Lambda}^\star_\ell(y_\ell) := \sup_{u_\ell\in\mathbb{R}}\{y_\ell u_\ell - \widehat{\Lambda}_\ell(u_\ell)\}$.

Now, convergence of the minima holds true only under an equi-coercivity condition (see Dal Maso, 1993, Chapter 7). To avoid messier assumptions, we assume that $\mathcal{S} = \prod_{1\le\ell\le m_0}[L_\ell, U_\ell]$ is compact (see A5; this assumption seems to be common in this literature, see Duffield et al., 1995; Duffy and Metcalfe, 2005a). From Azencott (1980, Proposition 9.7) we know that $\mathcal{D}(\Lambda^\star) \subset \mathcal{S}$. We also know, applying the same result to the function $\widehat{\Lambda}^\star$, that $\mathcal{D}(\widehat{\Lambda}^\star) \subset \overline{\text{co}}\{\mathbf{D}_k, 1 \le k \le n\} \subset \mathcal{S}$. Therefore:

$$\inf_{\mathbf{y}\in\mathcal{P}_j\cap\mathcal{S}}\widehat{\Lambda}^\star(\mathbf{y}) = \inf_{\mathbf{y}\in\mathcal{P}_j}\widehat{\Lambda}^\star(\mathbf{y}).$$

Now we characterize $\inf_{\mathbf{y}\in\mathcal{P}_j}\widehat{\Lambda}^\star(\mathbf{y})$ in terms of its dominating point. As $\widehat{\Lambda}$ is everywhere finite for finite n, $\mathcal{D}(\widehat{\Lambda}) = \mathbb{R}^{m_0}$ contains a neighborhood of the origin. For the same reason $\widehat{\Lambda}$ is essentially smooth. The set $B = \mathcal{P}_j$ is convex. The condition $\text{int}(\mathcal{P}_j \cap \overline{\text{co}}\{\mathbf{D}_k, 1 \le k \le n\}) \neq \emptyset$ is justified for large n. At last, it is also clear that $\widehat{\mathbb{E}}_n\mathbf{D}_k \notin \mathcal{P}_j$ for large enough n. □

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2020.01.001

References

Azencott, R., 1980. Grandes deviations et applications. In: Hennequin, P.L. (Ed.), Ecole d'Eté de Probabilités de Saint-Flour VIII-1978, 774. Springer, Berlin and Heidelberg, pp. 1–176. doi:10.1007/BFb0089623.

Barde, S., 2016. Direct comparison of agent-based models of herding in financial markets. Journal of Economic Dynamics and Control 73, 329–353. doi:10.1016/j.jedc.2016.10.005.

Barde, S., 2017. A Practical, Accurate, Information Criterion for Nth Order Markov Processes. Computational Economics 50 (2), 281–324. doi:10.1007/s10614-016-9617-9.

Bardone, E., Secchi, D., 2017. Inquisitiveness: Distributing rational thinking. Team Performance Management: An International Journal 23 (1/2), 66–81. doi:10.1108/TPM-10-2015-0044.
Basseville, M., 2013. Divergence measures for statistical data processing–An annotated bibliography. Signal Processing 93 (4), 621–633. doi:10.1016/j.sigpro.2012.09.003.
Ben Arous, G., Bogachev, L.V., Molchanov, S.A., 2005. Limit theorems for sums of random exponentials. Probability Theory and Related Fields 132 (4), 579–612. doi:10.1007/s00440-004-0406-3.
Berndt, D.J., Clifford, J., 1994. Using Dynamic Time Warping to Find Patterns in Time Series. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. AAAI Press, pp. 359–370.
Bickel, P.J., Doksum, K.A., 2015. Mathematical statistics—basic ideas and selected topics. Vol. 1, Second edition. CRC Press, Boca Raton. doi:10.1201/b18312.
Boero, R., Squazzoni, F., 2005. Does Empirical Embeddedness Matter? Methodological Issues on Agent-Based Models for Analytical Social Science. Journal of Artificial Societies and Social Simulation 8 (4), 6.
Box, M.J., 1966. A Comparison of Several Current Optimization Methods, and the use of Transformations in Constrained Problems. The Computer Journal 9 (1), 67–77. doi:10.1093/comjnl/9.1.67.
Brenner, T., 2006. Chapter 18. Agent Learning Representation: Advice on Modelling Economic Learning. In: Tesfatsion, L., Judd, K.L. (Eds.), Handbook of Computational Economics, Vol. 2. Elsevier, pp. 895–947. doi:10.1016/S1574-0021(05)02018-6.
Camerer, C.F., Fehr, E., 2006. When does "economic man" dominate social behavior. Science 311 (47), 47–52. doi:10.1126/science.1110600.
Canova, F., Sala, L., 2009. Back to square one: Identification issues in DSGE models. Journal of Monetary Economics 56 (4), 431–449. doi:10.1016/j.jmoneco.2009.03.014.
Chen, S.-H., Chang, C.-L., Du, Y.-R., 2012. Agent-based economic models and econometrics. The Knowledge Engineering Review 27 (2), 187–219. doi:10.1017/S0269888912000136.
Choirat, C., Hess, C., Seri, R., 2003. A functional version of the Birkhoff ergodic theorem for a normal integrand: A variational approach. The Annals of Probability 31 (1), 63–92. doi:10.1214/aop/1046294304.
Choirat, C., Seri, R., 2012. Estimation in Discrete Parameter Models. Statistical Science 27 (2), 278–293. doi:10.1214/11-STS371.
Comman, H., 2009. Differentiability-free conditions on the free-energy function implying large deviations. Confluentes Mathematici 01 (02), 181–196. doi:10.1142/S1793744209000079.
Cooley, T.F., 1997. Calibrated Models. Oxford Review of Economic Policy 13 (3), 55–69. doi:10.1093/oxrep/13.3.55.
Cowley, S.J., Vallée-Tourangeau, F. (Eds.), 2017. Cognition Beyond the Brain: Computation, Interactivity and Human Artifice, Second edition. Springer, Cham. doi:10.1007/978-3-319-49115-8.
Crooks, A., Castle, C., Batty, M., 2008. Key challenges in agent-based modelling for geo-spatial simulation. Computers, Environment and Urban Systems 32 (6), 417–430. doi:10.1016/j.compenvurbsys.2008.09.004.
Dal Maso, G., 1993. An Introduction to Γ-Convergence. Birkhäuser, Basel. doi:10.1007/978-1-4612-0327-8.
De Marco, S., Jacquier, A., Roome, P., 2016. Two examples of non strictly convex large deviations. Electronic Communications in Probability 21 (38). doi:10.1214/16-ECP4088.
Dembo, A., Zeitouni, O., 2010. Large Deviations Techniques and Applications, Second edition. Springer, Berlin and Heidelberg. doi:10.1007/978-3-642-03311-7.
Di Molfetta, G., 2016. On the parameter identifiability problem in Agent Based economical models. arXiv:1602.01271 [q-fin].
Duffield, N.G., Lewis, J.T., O'Connell, N., Russell, R., Toomey, F., 1995. Entropy of ATM traffic streams: A tool for estimating QoS parameters. IEEE Journal on Selected Areas in Communications 13 (6), 981–990. doi:10.1109/49.400654.
Duffy, K.R., Metcalfe, A.P., 2005a. How to Estimate the Rate Function of a Cumulative Process. Journal of Applied Probability 42 (4), 1044–1052. doi:10.1239/jap/1134587815.
Duffy, K.R., Metcalfe, A.P., 2005b. The Large Deviations of Estimating Rate Functions. Journal of Applied Probability 42 (1), 267–274. doi:10.1239/jap/1110381386.
Duffy, K. R., Williamson, B. D., 2015. Estimating large deviation rate functions. arXiv:1511.02295 [math].
Fabretti, A., 2013. On the problem of calibrating an agent based model for financial markets. Journal of Economic Interaction and Coordination 8 (2), 277–293. doi:10.1007/s11403-012-0096-3.
Fagiolo, G., Guerini, M., Lamperti, F., Moneta, A., Roventini, A., 2019. Validation of agent-based models in economics and finance. In: Beisbart, C., Saam, N.J. (Eds.), Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives. Springer, Cham, pp. 763–787. doi:10.1007/978-3-319-70766-2_31.
Fagiolo, G., Moneta, A., Windrum, P., 2007. A Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems. Computational Economics 30 (3), 195–226. doi:10.1007/s10614-007-9104-4.
Feuerverger, A., 1989. On The Empirical Saddlepoint Approximation. Biometrika 76 (3), 457–464. doi:10.2307/2336112.
Fu, T.-C., 2011. A review on time series data mining. Engineering Applications of Artificial Intelligence 24 (1), 164–181. doi:10.1016/j.engappai.2010.09.007.
Gilbert, N., 2007. Computational Social Science: Agent-based social simulation. In: Phan, D., Amblard, F. (Eds.), Agent-Based Modelling and Simulation. Bardwell, Oxford, pp. 115–134.
Gilbert, N., Terna, P., 2000. How to build and use agent-based models in social science. Mind and Society 1, 57–72. doi:10.1007/BF02512229.
Gilli, M., Winker, P., 2002. Indirect Estimation of the Parameters of Agent Based Models of Financial Markets. SSRN Electronic Journal. doi:10.2139/ssrn.300220.
Gilli, M., Winker, P., 2003. A global optimization heuristic for estimating agent based models. Computational Statistics & Data Analysis 42 (3), 299–312. doi:10.1016/S0167-9473(02)00214-1.
Grazzini, J., Richiardi, M., 2015. Estimation of ergodic agent-based models by simulated minimum distance. Journal of Economic Dynamics and Control 51, 148–165. doi:10.1016/j.jedc.2014.10.006.
Grazzini, J., Richiardi, M.G., Tsionas, M., 2017. Bayesian estimation of agent-based models. Journal of Economic Dynamics and Control 77, 26–47. doi:10.1016/j.jedc.2017.01.014.
Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.-H., Weiner, J., Wiegand, T., DeAngelis, D.L., 2005. Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology. Science 310 (5750), 987–991. doi:10.1126/science.1116681.
Gudmundsson, S., Runarsson, T.P., Sigurdsson, S., 2008. Support vector machines and dynamic time warping for time series. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, Hong Kong, pp. 2772–2776. doi:10.1109/IJCNN.2008.4634188.
Guerini, M., Moneta, A., 2017. A method for agent-based models validation. Journal of Economic Dynamics and Control 82, 125–141. doi:10.1016/j.jedc.2017.06.001.
Hansen, L.P., Heckman, J.J., 1996. The Empirical Foundations of Calibration. Journal of Economic Perspectives 10 (1), 87–104. doi:10.1257/jep.10.1.87.
Hansen, P.R., Lunde, A., Nason, J.M., 2011. The Model Confidence Set. Econometrica 79 (2), 453–497. doi:10.3982/ECTA5771.
Hess, C., Seri, R., 2019. Generic Consistency for Approximate Stochastic Programming and Statistical Problems. SIAM Journal on Optimization 29 (1), 290–317. doi:10.1137/17M1156769.
Hess, C., Seri, R., Choirat, C., 2010. Ergodic theorems for extended real-valued random variables. Stochastic Processes and their Applications 120 (10), 1908–1919. doi:10.1016/j.spa.2010.05.008.
Hess, C., Seri, R., Choirat, C., 2014. Essential intersection and approximation results for robust optimization. Journal of Nonlinear and Convex Analysis 15 (5), 979–1002.
Iltis, M., 1995. Sharp asymptotics of large deviations in $\mathbb{R}^d$. Journal of Theoretical Probability 8 (3), 501–522. doi:10.1007/BF02218041.

Kahneman, D., 2003. A perspective of judgement and choice: Mapping bounded rationality. American Psychologist 58 (9), 697–721. doi:10.1037/0003-066X.58.9.697.
Kruskal, J.B., 1983. An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules. SIAM Review 25 (2), 201–237. doi:10.1137/1025045.
Kukacka, J., Barunik, J., 2017. Estimation of financial agent-based models with simulated maximum likelihood. Journal of Economic Dynamics and Control 85, 21–45. doi:10.1016/j.jedc.2017.09.006.
Kydland, F.E., Prescott, E.C., 1982. Time to Build and Aggregate Fluctuations. Econometrica 50 (6), 1345–1370. doi:10.2307/1913386.
Lamperti, F., 2018a. Empirical validation of simulated models through the GSL-div: An illustrative application. Journal of Economic Interaction and Coordination 13 (1), 143–171. doi:10.1007/s11403-017-0206-3.
Lamperti, F., 2018b. An information theoretic criterion for empirical validation of simulation models. Econometrics and Statistics 5, 83–106. doi:10.1016/j.ecosta.2017.01.006.
Lamperti, F., Roventini, A., Sani, A., 2018. Agent-based model calibration using machine learning surrogates. Journal of Economic Dynamics and Control 90, 366–389. doi:10.1016/j.jedc.2018.03.011.
Liao, T.W., 2005. Clustering of time series data—a survey. Pattern Recognition 38 (11), 1857–1874. doi:10.1016/j.patcog.2005.01.025.
Lorscheid, I., Meyer, M., 2016. Divide and conquer: Configuring submodels for valid and efficient analyses of complex simulation models. Ecological Modelling 326, 152–161. doi:10.1016/j.ecolmodel.2015.11.013.
Lucet, Y., 1997. Faster than the Fast Legendre Transform, the Linear-time Legendre Transform. Numerical Algorithms 16 (2), 171–185. doi:10.1023/A:1019191114493.
Lucet, Y., 2009. What Shape Is Your Conjugate? A Survey of Computational Convex Analysis and Its Applications. SIAM Journal on Optimization 20 (1), 216–250. doi:10.1137/080719613.
Marks, R.E., 2013. Validation and model selection: Three similarity measures compared. Complexity Economics 2 (1), 41–61. doi:10.7564/13-COEC10.
Miller, K.D., Lin, S.-J., 2010. Different truths in different worlds. Organization Science 21 (1), 97–114. doi:10.1287/orsc.1080.0409.
Ney, P.E., 1983. Dominating Points and the Asymptotics of Large Deviations for Random Walk on $\mathbb{R}^d$. The Annals of Probability 11 (1), 158–167. doi:10.1214/aop/1176993665.
Ney, P.E., 1984. Convexity and Large Deviations. The Annals of Probability 12 (3), 903–906. doi:10.1214/aop/1176993239.
Ney, P.E., Robinson, S.M., 1995. Polyhedral Approximation of Convex Sets With an Application to Large Deviation Probability Theory. Journal of Convex Analysis 2 (1/2), 229–240.
North, M.J., Macal, C.M., 2007. Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation. Oxford University Press, Oxford and New York. doi:10.1093/acprof:oso/9780195172119.001.0001.
Olver, F.W., Lozier, D.W., Boisvert, R.F., Clark, C.W., 2010. NIST Handbook of Mathematical Functions. Cambridge University Press, New York.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263 (5147), 641–646. doi:10.1126/science.263.5147.641.
Pennanen, T., Koivu, M., 2005. Epi-convergent discretizations of stochastic programs via integration quadratures. Numerische Mathematik 100 (1), 141–163. doi:10.1007/s00211-004-0571-4.
Ramm, A.G., Zaslavsky, A.I., 1993. Reconstructing singularities of a function from its Radon transform. Mathematical and Computer Modelling 18 (1), 109–138. doi:10.1016/0895-7177(93)90083-B.
Recchioni, M.C., Tedeschi, G., Gallegati, M., 2015. A calibration procedure for analyzing stock price dynamics in an agent-based framework. Journal of Economic Dynamics and Control 60, 1–25. doi:10.1016/j.jedc.2015.08.003.
Richiardi, M., Leombruni, R., Saam, N., Sonnessa, M., 2006. A Common Protocol for Agent-Based Social Simulation. Journal of Artificial Societies and Social Simulation 9 (1), 15.
Rockafellar, R.T., Wets, R.J.-B., 1998. Variational Analysis. Grundlehren Der Mathematischen Wissenschaften, 317. Springer, Berlin and Heidelberg. doi:10.1007/978-3-642-02431-3.
Rohwer, C.M., Angeletti, F., Touchette, H., 2015. Convergence of large-deviation estimators. Physical Review E 92 (5), 052104. doi:10.1103/PhysRevE.92.052104.
Romer, D., 2012. Advanced Macroeconomics, Fourth edition. McGraw Hill/Irwin, New York.
Schmidli, H., 1994. Estimation of the abscissa of convergence of the moment generating function. Mimeo.
Secchi, D., 2016. Boundary Conditions for the Emergence of "Docility" in Organizations: Agent-Based Model and Simulation. In: Secchi, D., Neumann, M. (Eds.), Agent-Based Simulation of Organizational Behavior. Springer, Cham, pp. 175–200. doi:10.1007/978-3-319-18153-0_9.
Secchi, D., Bardone, E., 2009. Super-docility in organizations: An evolutionary model. International Journal of Organization Theory & Behavior 12 (3), 339–379. doi:10.1108/IJOTB-12-03-2009-B001.
Secchi, D., Gullekson, N., 2016. Individual and organizational conditions for the emergence and evolution of bandwagons. Computational and Mathematical Organization Theory 22 (1), 88–133. doi:10.1007/s10588-015-9199-4.
Secchi, D., Seri, R., 2017. Controlling for false negatives in agent-based models: A review of power analysis in organizational research. Computational and Mathematical Organization Theory 23 (1), 94–121. doi:10.1007/s10588-016-9218-0.
Seri, R., Choirat, C., 2013. Scenario Approximation of Robust and Chance-Constrained Programs. Journal of Optimization Theory and Applications 158 (2), 590–614. doi:10.1007/s10957-012-0230-3.
Seri, R., Secchi, D., 2017. How Many Times Should One Run a Computational Simulation? In: Edmonds, B., Meyer, R. (Eds.) Simulating Social Complexity: A Handbook. Springer, Cham, pp. 229–251. doi:10.1007/978-3-319-66948-9_11.
Simon, H.A., 1955. A behavioral theory of rational choice. Quarterly Journal of Economics 69 (1), 99–118. doi:10.2307/1884852.
Simon, H.A., 1979. Rational decision making in business organizations. American Economic Review 69 (4), 493–513.
Simon, H.A., 1990. A mechanism for social selection and successful altruism. Science 250 (4988), 1665–1668. doi:10.1126/science.2270480.
Simon, H.A., 1993. Altruism and Economics. American Economic Review 83 (2), 156–161.
Simon, H.A., 1997. Administrative behavior, Fourth edition. The Free Press, New York.
Thaler, R.H., Sunstein, C.R., 2008. Nudge. Improving Decisions About Health, Wealth, and Happiness. Yale University Press, New Haven and London.
Thiele, J.C., Kurth, W., Grimm, V., 2015. Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R. Journal of Artificial Societies and Social Simulation 17 (3), 11.
Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E., 2013. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery 26 (2), 275–309. doi:10.1007/s10618-012-0250-5.
Windrum, P., Fagiolo, G., Moneta, A., 2007. Empirical Validation of Agent-Based Models: Alternatives and Prospects. Journal of Artificial Societies and Social Simulation 10 (2), 8.
Winker, P., Gilli, M., Jeleskovic, V., 2007. An objective function for simulation based inference on exchange rate data. Journal of Economic Interaction and Coordination 2 (2), 125–145. doi:10.1007/s11403-007-0020-4.

Supplement to
"Model Calibration and Validation via Confidence Sets"

Raffaello Seri[a,b], Mario Martinoli[1a], Davide Secchi[b], Samuele Centorrino[c]

[a] *Department of Economics, Università degli Studi dell'Insubria, Varese, Italy*
[b] *Centre for Computational & Organisational Cognition (CORG), University of Southern Denmark, Slagelse, Denmark*
[c] *Economics Department, Stony Brook University, Stony Brook, USA*

The supplemental material contains:

- `Data.csv`: a `csv` file containing the data used for the application in Section 6 of the paper;
- `Plot.R`: an `R` file reading and plotting the data;
- `MCS.R`: an `R` file reading the data and computing the Model Confidence Set;
- `LDP.R`: an `R` file reading the data and computing the rate functions appearing in Large Deviations results.

[1] Corresponding author.