Approximate Maximum Likelihood for Complex Structural Models

Veronika Czellar, David T. Frazier and Eric Renault

March 2021 No: 1337

Warwick Economics Research Papers

ISSN 2059-4283 (online)

ISSN 0083-7350 (print)

Approximate Maximum Likelihood for Complex Structural Models

Veronika Czellar[*], David T. Frazier[†]and Eric Renault[‡][§]

March 1, 2021

Abstract

Indirect Inference (I-I) is a popular technique for estimating complex parametric models whose likelihood function is intractable, however, the statistical efficiency of I-I estimation is questionable. While the efficient method of moments, Gallant and Tauchen (1996), promises efficiency, the price to pay for this efficiency is a loss of parsimony and thereby a potential lack of robustness to model misspecification. This stands in contrast to simpler I-I estimation strategies, which are known to display less sensitivity to model misspecification due in large part to their focus on specific elements of the underlying structural model. In this research, we propose a new simulation-based approach that maintains the parsimony of I-I estimation, which is often critical in empirical applications, but can also deliver estimators that are nearly as efficient as maximum likelihood. This new approach is based on using a constrained approximation to the structural model, which ensures identification and can deliver estimators that are consistent and nearly efficient. We demonstrate this approach through several examples, and show that this approach can deliver estimators that are nearly as efficient as maximum likelihood, when feasible, but can be employed in many situations where maximum likelihood is infeasible.

Keywords: Equality Restrictions; Constrained Inference; Indirect Inference; Generalized Tobit; Markov-Switching Multifractal Models.

1 Introduction

Indirect inference (hereafter, I-I), as proposed by Smith (1993) and Gourieroux, et al. (1993), is a simulation-based estimation method often used when the underlying likelihood for the model

[*] Department of Data Science, Economics and Finance, EDHEC Business School, France

[†] Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia. Corresponding author: `david.frazier@monash.edu`

[‡] Department of Economics, University of Warwick and Department of Econometrics and Business Statistics, Monash University.

[§] Renault was supported by CY Initiative of Excellence (grant "Investissements dÁvenir" ANR-16-IDEX-0008), Project "EcoDep" PSI-AAP2020-0000000013. Frazier was supported by the Australian Research Council's Discovery Early Career Researcher Award funding scheme (DE200101070) and the Australian Center for Excellence in Mathematics and Statistics (ACEMS).

of interest is computationally challenging, or intractable. The key idea underpinning I-I is that, regardless how complicated the structural model, it is often feasible to simulate artificial data from this fully parametric model. As a result, statistics based on the observed data and data simulated from the model can be compared, with the resulting difference minimized in a given norm to produce an estimator of the structural parameters.

The implementation of I-I is most often carried out using an auxiliary model that represents an incorrect, but tractable version of the structural model under analysis. User-friendly estimators for the parameters of this auxiliary model provide the statistics, based on the observed and simulated data, respectively, that are used to conduct inference on the underlying structural parameters. However, by definition the information encapsulated in the auxiliary parameter estimates is less than the information carried in the likelihood for the structural parameters. As such, in any implementation of I-I there is a fundamental trade-off between the statistical efficiency of the resulting estimators and their computational feasibility.

The main contribution of this paper is to propose an alternative to I-I that produces structural parameter estimates that, albeit also simulation-based, are arguably closer to reaching the Cramer-Rao efficiency bound for the parametric structural model. The new method proposed herein, dubbed "Approximate Maximum Likelihood" (hereafter, AML), maintains the standard philosophy of I-I that one can resort to a possibly biased approximation of the structural model, insofar as matching statistics calculated from this approximation using both simulated and observed data will allow us to erase the misspecification bias. In contrast to standard I-I, instead of matching estimators of auxiliary parameters, we directly match a proxy/approximation to the score vector of the intractable log-likelihood. These proxies are indexed by the vector of structural parameters, for which a preliminary plug-in estimator (based on observed data) must be used.

However, as we later demonstrate, the dependence of this approach on the preliminary plug-in estimator differs from standard I-I estimation: as far as the asymptotic distribution of our AML estimator is concerned, the asymptotic distribution of the preliminary estimator is immaterial, and only its probability limit (a pseudo-true value possibly different from the true unknown value) will impact the information conveyed by the approximate score. This is in stark contrast to I-I estimation, where the key feature in determining the asymptotic efficiency of I-I is the efficiency of the auxiliary parameter estimates. As such, since it is only the probability limits of the plug-in estimators that matters, our new AML approach can not be directly placed in the standard I-I framework.

While this new approach is based on matching types of scores, it should not be confused with the score-based version of I-I proposed by Gallant and Tauchen (1996). As shown by Gourieroux, Monfort and Renault (1993) (see "The Third Version of the Indirect Estimator" in their Appendix 1), as far as asymptotic properties are concerned, Gallant and Tauchen's (1996) estimator is actually tantamount to match estimators of auxiliary parameters. In particular, when fishing for efficiency, Gallant and Tauchen (1996) (see the proof of their Theorem 2) ultimately import the efficiency for the estimator of auxiliary parameters to reach the Cramer-Rao efficiency bound for the structural parameters, with this efficiency claim ultimately requiring that the auxiliary model "smoothly embeds" the structural model.

In short, "efficient method of moments", Gallant and Tauchen (1996), must in general resort to a semi-nonparametric score generator as an auxiliary model. Thanks to its steadily increasing dimension, the score of this auxiliary model may asymptotically span the score of the structural model, and thereby deliver efficient estimators of the resulting structural parameters. However, the price to pay for this efficiency is a highly-parametrized auxiliary model that may be ill-behaved (due to the non-parsimonious nature of the auxiliary model) when there are deviations from the assumed model, i.e., when the structural model may be misspecified (for further examples and details see Dridi and Renault, 2000, Section 4.2). This is in contrast to the original intuition of I-I estimation, which has been shown to be somewhat robust to deviations from the underlying modelling assumptions (Dridi et al., 2007), precisely because it is based on calibrating a limited number of structural parameters. Our new method remains true to this parsimony principle since we match proxies for the actual score vector, whose dimension is the same as the structural parameters.

In our AML approach, (approximate) efficiency of structural parameter estimates does not rest upon high-dimensional inference or the near-efficiency of auxiliary parameter estimates, but on the conjunction of two properties.

- First, the efficiency gap between our estimates and the MLE is tightly related to the difference between the asymptotic value of our plug-in estimator for the structural parameters (i.e., the pseudo-true value that will asymptotically feature in our proxy/ approximation for the true limiting score function) and the true unknown value of the structural parameters.

- Second, the fact that the Cramer-Rao efficiency bound can be (nearly) reached if the information identity is (nearly) maintained. More precisely, the question is to assess the difference between the curvature of the log-likelihood at the true value of the structural parameters (as measured by the slope of the expected score vector as a function of the structural parameters) and the slope of the score vector when the structural parameters enter the score through data simulated at a specific parameter value. Satisfaction of the information identity in this context requires a type of multiplicative separability of the score vector, which we later demonstrate is satisfied for exponential models.

The motivation for our AML approach is the observation that there are many cases of interest where the intractability of the assumed model, and its likelihood, is entirely due to a sub-vector of structural parameters. Examples include, for instance, dynamic discrete choice models with ARMA errors (Robinson, 1982, Gourieroux et al., 1985, Poirier and Ruud, 1988), spatial discrete choice models (see, e.g., Pinske and Slade, 1998), and many dynamic equilibrium models. In such models, a few well-chosen restrictions would allow us to alleviate the intractability of the likelihood caused by the presence of certain latent variables.

More generally, many complex economic models are such that imposing a (potentially false) constraint on the structural model yields a simpler auxiliary model with a computationally tractable likelihood. This is precisely the reason why score/LM tests are popular in econometrics: estimation and testing “under the null”is feasible even in very complicated models. Unfortunately, imposition of this constraint, and subsequent optimization of the constrained log-likelihood, will not deliver consistent estimates of the structural parameters if the constraint is not satisfied at the truth.

As recently pointed out by Calvet and Czellar (2015), imposing potentially false equality constraints on a given structural model can be an attractive method for obtaining simple and rich auxiliary models for the purposes of I-I. For instance, in the context of a long-run risk model (Bansal and Yaron, 2004), Calvet and Czellar (2015) demonstrate that imposing specific equality constraints on certain parameters produces a simple auxiliary model for use in I-I (with a computationally tractable likelihood function) that closely resemble the structural model. The fact that this resulting auxiliary model may not deliver consistent estimates of the true structural parameters is immaterial insofar as matching a simulation-based approximation against the observation-based version will allow us to erase the misspecification bias. The benefits of such an approach are two-fold: one, by using constraints to define the auxiliary model, we sketch a systematic strategy for the choice of an auxiliary model; two, this auxiliary model closely matches the structural model and so for issues of robustness and efficiency this auxiliary model is very useful.

However, while highly-useful, the suggestion of Calvet and Czellar (2015) is incomplete, and does not allow for consistent estimation of the structural parameters on its own. That is, since the auxiliary model imposes a number of constraints on the structural model, by definition the auxiliary model can not consistently estimate all the structural parameters, except in the unlikely case where the constraints are satisfied at the true value of the structural parameters. To circumvent this issue, Calvet and Czellar (2015) propose to add to the statistics obtained from the auxiliary model additional statistics so that, when considered jointly, this new vector can jointly identify the structural parameters when estimated by I-I.

Motivated by the above ideas and the approach to handling constraints within I-I proposed in Calzolari et al. (2004) and Frazier and Renault (2019), we propose a novel inference approach based on constraining the structural model parameters to create a simple, but highly informative, proxy for the score vector that can be used to estimate the structural parameters. However, unlike the strategy put forward by Calvet and Czellar (2015), our approach provides an automatic, and nearly-efficient, method to identify the structural parameters.

In addition, we demonstrate that this AML strategy can be based on a proxy for the score vector that entails additional layers of approximation beyond simply plugging in a (wrongly) constrained estimation of the structural parameters. For example, in the context of stable probability distributions, the likelihood function is known in closed-form only at certain specific values of the parameters; as an example, a unit shape parameter ($a = 1$) and a zero value of the asymmetry parameter ($b = 0$) yield a Cauchy likelihood, however, even then the partial derivatives of the likelihood function, with respect to a and b, are not available in closed-form.[1] In such settings, our AML strategy can be implemented by invoking an additional layer of approximation and replacing the directions of the score vector that can not be obtained in closed-form by a finite-difference approximation. Approximating certain directions of the score vector by finite-differences is obviously even more useful when some structural parameters are only defined on the integers. We demonstrate our methodology in such cases using the example of Markov-Switching multifractal (MSM) volatility processes, Calvet and Fisher (2004, 2008), which are especially well-suited to capture volatility dynamics through an unknown, but finite, number of multiplicative components. We document that there is some hope to reach near-efficiency with AML (irrespective of the number of latent multiplicative components) while MLE is computationally intractable with more than a small number of multiplicative components. Moreover, as documented by Calvet and Fisher (2004), goodness of fit for the MSM model with financial data generally requires many multiplicative components.

While we apply our AML methodology within the confines of a MSM volatility model, we note here that the use of MSM models are not exclusive to the analysis of volatility. Indeed, Chen, Diebold and Schorfheide (2013) propose a novel Markov-switching multifractal duration (MSMD) model to analyze inter-trade duration data in financial markets, and demonstrate its

[1] We refer the reader to Appendix C for details.

superiority over competing duration models. While we exemplify the AML procedure within a MSM volatility model, we note here that AML can be equivalently applied to the MSMD model of Chen et al. (2013) using precisely the same approach detailed in this paper.

The remainder of the paper is organized as follows. In Section 2, we give the general setup, and consider examples where equality constraints on the structural model yield a tractable score vector that can be used for inference through score matching, and discuss our AML estimation strategy. In Section 3, we provide the asymptotic theory of AML. Further, we demonstrate that, in the case of an exponential model, a sufficient (but not necessary) condition for AML estimators to achieve the Cramer-Rao efficiency bound is that the pseudo-true value used in AML coincides with the true one. Section 4 provides Monte Carlo evidence on the finite-sample performance of AML in two examples: one based on false equality constraints, and one where we are required to define some of the pseudo-score components using a finite-difference approximation, with the later example containing an empirical application to financial returns data using a multifractal stochastic volatility model. Section 5 concludes with suggestions for future research on extensions of I-I where not only the two vectors to match both depend on the observed data, as in this paper, but even the simulator itself may depend on the observed data. Mathematical details for the proofs of main results and developments of theoretical examples are provided in Appendices A, and B. Appendix C provides additional Monte Carlo evidence in the context of stable distributions, while an application of AML to discrete choice models with serial correlated errors is sketched in Appendix D.

2 Approximate Maximum Likelihood vs Indirect Inference

2.1 Model Setup: Nonlinear State Space Models

Following Gourieroux, et al. (1993) (hereafter, GMR), our goal is inference on the unknown parameters of a dynamic structural model that has a nonlinear state space representation. The structural model is specified through a transition, or state, equation and a measurement equation. The transition equation is of the following form:

$$u_t = \varphi\left(u_{t-1}, \varepsilon_t, \theta\right); \theta \in \Theta \subset \mathbb{R}^p,$$

where $\varphi(\cdot)$ is a known function, $(u_t, \varepsilon_t)_{t=1}^T$ are latent processes and ε_t is a strong white noise process with a known distribution; and the measurement equation satisfies

$$y_t = r\left(y_{t-1}, x_t, u_t, \varepsilon_t, \theta\right); \theta \in \Theta \subset \mathbb{R}^p,$$

where $r(\cdot)$ is a known function and $(x_t, y_t)_{t=1}^T$ are observed processes. In the two equations, known functions $\varphi(\cdot)$ and $r(\cdot)$ are indexed by a p-dimensional vector of unknown parameters $\theta \in \Theta$. We assume that $(x_t)_{t\leq T}$ is a homogenous Markov process of order 1, and is independent of the process $(\varepsilon_t)_{t\leq T}$ (and $(u_t)_{t\leq T}$). Then the process $(x_t)_{t\leq T}$ is exogenous and the process $(x_t, y_t)_{t\leq T}$ is stationary. It is worth recalling that, by standard arguments, the fact that the Markov process is of order 1 and the probability distribution of the white noise ε_t is known are not restrictive assumptions.

Under the above conditions, assuming absolute continuity with respect to some dominating measure, for a given initial condition $z_0 = (y_0, u_0)$, it should be possible to write down the joint conditional probability density function

$$l^* \left\{ (y_t)_{1\leq t\leq T}, (u_t)_{1\leq t\leq T} \left| (x_t)_{1\leq t\leq T}, z_0; \theta \right. \right\}. \tag{1}$$

The density of the observed sequence $(y_t)_{t\leq T}$, conditional on $(x_t)_{t\leq T}$, is obtained by integrating out the latent variables $(u_t)_{1\leq t\leq T}$ from the density (1) and can generally be stated as

$$l \left\{ (y_t)_{1\leq t\leq T} \left| (x_t)_{1\leq t\leq T}; \theta \right. \right\} = \prod_{1\leq t\leq T} l\{ y_t \left| (y_\tau)_{1\leq \tau\leq t-1}, x_t, z_0; \theta \right. \}, \tag{2}$$

where the last equality comes from the Markovianity and exogeneity of the process $(x_t)_{t\leq T}$. This density function allows us to construct the log-likelihood function

$$L_T(\theta) = \frac{1}{T} \sum_{1\leq t\leq T} \log \left(l\{ y_t \left| (y_\tau)_{1\leq \tau\leq t-1}, x_t, z_0; \theta \right. \} \right). \tag{3}$$

A maintained assumption in this paper will be that the log-likelihood asymptotically identifies some true unknown value, θ^0, of the unknown parameters, θ, and is the unique maximizer of the population criterion:

$$\theta^0 = \arg\max_{\theta\in\Theta} L_\infty(\theta), \text{ where } L_\infty(\theta) = \operatorname*{plim}_{T\to\infty} L_T(\theta).$$

It is important to realize that more often than not, this assumption is neither testable nor associated to a feasible estimator of θ^0. The likelihood function in equation (2) does not have an analytically tractable form: it is constructed via the latent likelihood in (1) through an integration step that is infeasible to carry out, integration with respect to the T variables $(u_t)_{t\leq T}$, with T going to infinity.[2]

Even though direct inference on θ^0 associated with $L_T(\theta)$ may be infeasible, it is well-known that inference can be carried out using simulation-based filtering and inference approaches. Under the assumed model, it is possible to simulate values of $y_1, ..., y_T$, for a given initial condition $z_0 = (y_0, u_0)$ and a given value θ of the parameters, conditionally on the observed path of the exogenous variables $x_1, ..., x_T$. This is done by independently drawing simulated values $\tilde{\varepsilon}_1, ..., \tilde{\varepsilon}_T$ from the assumed distribution of the strong white noise $(\varepsilon_t)_{t\leq T}$ (the simulated values are also independent of the realized values $\varepsilon_1, ..., \varepsilon_T$ that underpin the observations) and by computing

$$\tilde{y}_t(\theta, z_0), \text{ for } t = 0, 1, \ldots, T,$$

with $\tilde{y}_0(\theta, z_0) = y_0$ and where

$$\begin{aligned} \tilde{y}_t(\theta, z_0) &= r\left[\tilde{y}_{t-1}(\theta, z_0), x_t, \tilde{u}_t(\theta, u_0), \tilde{\varepsilon}_t, \theta\right] \\ \tilde{u}_t(\theta, u_0) &= \varphi\left[\tilde{u}_{t-1}(\theta, u_0), \tilde{\varepsilon}_t, \theta\right]. \end{aligned}$$

[2] Clearly, such examples are exclusive of cases where the integration, or filtering, can be performed analytically, such as cases where the Kalman filter is valid, as in linear Gaussian state space models, or as in certain qualitative Markov switching models. The focus of this paper is nonlinear state space models, where the above simplifications are not generally applicable.

While simulation is the most prevalent mechanism for inference in such settings, we note that in many cases inference could be based directly on $L_T(\theta)$ if we were to instead consider sub-models defined by restricting the parameters θ to lie in a given set $\Theta_0 \subset \Theta$. Indeed, it will often be that case that the sub-models could be chosen by imposing $\theta \in \Theta_0$ so that we obtain a convenient factorization of the probability density function, which ensures that integration of the T latent variables, $(u_t)_{t\leq T}$, no longer requires solving a T-dimensional integral, and consequently inference (over the sub-models) could be based directly on the log-likelihood function (3). However, in general the sub-models specified by this constraint will not be correctly specified and the resulting estimates will be asymptotically biased for the parameter of interest θ^0. However, as we will later see, following the intuition of I-I, this misspecification bias can be erased by matching these estimators against a simulated counterpart.

2.2 Illustrative Examples

In this section (and Appendices C and D), we demonstrate that there are many interesting cases where restricting the parameters θ to lie in some set $\Theta_0 \subset \Theta$ results in tractable log-likelihood functions.

2.2.1 Example 1: *Generalized Tobit Model*

Amemiya (1985) defines the generalized Tobit model of Type 2 by the following observation scheme for the outcome variable y_i :

$$y_i = \begin{cases} y^*_{1i} & \text{if } y^*_{2i} \geq 0 \\ \text{missing} & \text{if } y^*_{2i} < 0 \end{cases}, \tag{4}$$

with

$$y^*_{1i} = x_i'\theta_1 + \sigma\varepsilon_i, \tag{5}$$

where x_i is a vector of exogenous explanatory variables, $(\theta_1', \sigma)'$ a vector of unknown parameters and ε_i is a standardized Gaussian error $\varepsilon_i \sim \aleph(0,1)$. A complete specification for the likelihood function requires specifying the conditional probability of missingness in the data:

$$\Pr[y^*_{2i} < 0 \,| y^*_{1i}, z_i, \theta_2, \theta_3],$$

where z_i is a vector of exogenous explanatory variables and $(\theta_2', \theta_3')'$ is a vector of unknown parameters. The parameter θ_2 governs the relationship between z_i and the missingness mechanism, and the parameter θ_3 characterizes the dependence between the two latent endogenous variables y^*_{1i} and y^*_{2i}. Then, if I_1 (resp. I_0) stands for the subset of indices for which $y^*_{2i} \geq 0$ (resp. $y^*_{2i} < 0$), the likelihood function can be written as

$$l\left\{(y_i)_{1\leq i\leq T}\,|(x_i, z_i)_{1\leq i\leq T}\,; \theta\right\} = \prod_{i\in I_1} \frac{1}{\sigma}\varphi\left(\frac{y_i - x_i'\theta_1}{\sigma}\right) \Pr[y^*_{2i} \geq 0\,| y_i, z_i, \theta_2, \theta_3] \prod_{i\in I_0} \Pr[y^*_{2i} < 0\,| z_i, \theta],$$

with

$$\Pr[y^*_{2i} < 0\,| z_i, \theta] = \int \Pr[y^*_{2i} < 0\,| y^*_{1i}, z_i, \theta_2, \theta_3]\, \frac{1}{\sigma}\varphi\left(\frac{y^*_{1i} - x_i'\theta_1}{\sigma}\right) dy^*_{1i},$$

where $\varphi(\cdot)$ denotes the probability density function of the standard normal distribution and

$$\theta = (\theta_1', \theta_2', \theta_3', \sigma)' \text{ where } \theta_1 \in \mathbb{R}^{p_1},\ \theta_2 \in \mathbb{R}^{p_2},\ \theta_3 \in \mathbb{R},\ \sigma > 0$$

denotes the vector of unknown structural parameters. We note that estimation of θ may be challenging because the likelihood function involves an integral that may be necessary to compute numerically.

In general, Amemiya (1985) considers that the joint conditional distribution of $(y_{1i}^*, y_{2i}^*)'$ given (x_i, z_i) is Gaussian and θ_3 stands for the correlation coefficient between y_{1i}^* and y_{2i}^*. An alternative, and computationally more convenient choice, is to assume that the conditional probability distribution of y_{2i}^* given (y_{1i}^*, x_i, z_i) is logistic, which yields

$$\Pr[y_{2i}^* \geq 0\,|y_{1i}^*, z_i, x_i, \theta_2, \theta_3] = [1 + \exp(-z_i'\theta_2 - \theta_3 y_{1i}^*)]^{-1}. \tag{6}$$

While the likelihood may be difficult to compute in the general case, imposing the (possibly false) equality constraint $\theta_3 = 0$ implies that y_{1i}^* and y_{2i}^* are conditionally independent, given z_i, and the resulting likelihood function (in either case) is almost as simple as a standard Tobit likelihood. Under the logistic specification, imposing the false equality constraint $\theta_3 = 0$ yields the log-likelihood function

$$\begin{aligned} & L_T\left[(\theta_1', \theta_2', 0, \sigma)\right] \\ = \ & \frac{1}{T}\sum_{i\in I_1}\left\{-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(y_i - x_i'\theta_1\right)^2 - \log\left(1 + e^{-z_i'\theta_2}\right)\right\} - \frac{1}{T}\sum_{i\in I_0}\log\left(1 + e^{z_i'\theta_2}\right). \end{aligned}$$

However, we note that, with reference to the empirical study of Dudley and Montmarquette (1976), Amemiya (1985) argues that the assumption of independence between y_{1i}^* and y_{2i}^* is likely unrealistic. Therefore, there is no reason to expect that $\theta_3 = 0$ should be fulfilled in practice.

2.2.2 Example 2: *Markov-Switching Multifractal (MSM) Model*

Consider that observed asset returns evolve according to

$$r_t = \sigma_t u_t,\ \ E[u_t\,|I_{t-1}] = 0,\ E[u_t^2\,|I_{t-1}] = 1$$

with σ_t denoting the volatility process. Our goal is to analyze the volatility process and we use the Binomial MSM model proposed in Calvet and Fisher (2001, 2004, 2008), and consider that the volatility process is defined as the product of several volatility components

$$\sigma_t^2 = \sigma^2 \prod_{k=1}^{\overline{k}} M_{k,t}.$$

The components $M_{k,t}$ are unobservable (i.e., latent) variables that are often referred to as multipliers or volatility components, and the overall number of components, $\overline{k}$, is unknown.

We will assume that the standardized return u_t is i.i.d with a probability density function $f_u(\cdot)$. The latent state variables $M_{k,t}, k = 1, ..., \overline{k}$, are assumed to be stationary Markov processes with common marginal distribution, denoted by M. Given a value $M_{k,t}$ for the k^{th} component at time t, the next-period multiplier is assumed to evolve according to

$$M_{k,t+1} = \begin{cases} \sim M & \text{with probability } \gamma_k \\ M_{k,t} & \text{with probability } (1-\gamma_k) \end{cases}$$

where the notation $(\sim M)$ stands for "drawn in the distribution M" and M_0 is generated from the stationary distribution π_0, where

$$\pi_0^j = \Pr[M_0 = m^j] = 1/d, \ \forall j = 1, ..., d,$$

and where $d = 2^{\overline{k}}$.

The switching events (with transition probabilities $\gamma_k, k = 1, ..., \overline{k}$) and new draws from M are assumed to be independent across k and t. To ensure a non-negative and stationary volatility process $(E\left(\sigma_t^2\right) = \sigma^2)$, we assume

$$E(M) = 1, \ M \geq 0$$

For sake of parsimony, we introduce an unknown parameter $m_0 \in (1, 2)$ such that:

$$\Pr\left[M = m_0\right] = \Pr\left[M = 2 - m_0\right] = \frac{1}{2}.$$

Then the state vector $M_t = \left(M_{1,t}, ..., M_{\overline{k},t}\right)'$ can take d possible values m^j, $j = 1, ..., d$, so that at each date the squared volatility process takes d possible values

$$\sigma^2 g\left(m^j\right), \text{ where } g\left[\left(M_{1,t}, ..., M_{\overline{k},t}\right)\right] = \prod_{k=1}^{\overline{k}} M_{k,t}.$$

Furthermore, we parametrize the transition probabilities $\gamma_k, k = 1, ..., \overline{k}$, such that the first components (small k) are the most persistent

$$\gamma_k = \bar{\gamma} b^{k-\overline{k}}, \bar{\gamma} \in (0, 1], b > 1, k = 1, ..., \overline{k},$$

and where a possibly higher "volatility of volatility" can be accommodated by increasing $\overline{k}$.

For this model, the structural parameter vector is

$$\theta = \left(m_0, \bar{\gamma}, b, \sigma, \overline{k}\right)'$$

and the log-likelihood associated with observed returns $(r_t)_{t \leq T}$ is given by:

$$L_T\left(\theta\right) = \frac{1}{T} \sum_{t=1}^{T} \log\left(\sum_{j=1}^{d} \frac{1}{\sigma\sqrt{g\left(m^j\right)}} f_u\left(\frac{r_t}{\sigma\sqrt{g\left(m^j\right)}}\right) \Pr[M_t = m^j \left| r_\tau, \tau < t\right]\right) \tag{7}$$

where the conditional probabilities $\pi_t^j = \Pr[M_t = m^j \left| r_\tau, \tau < t\right]$ are computed recursively. By Bayes' rule, the probability π_t^j can be expressed as a function of the previous probabilities $\pi_{t-1} = \left(\pi_{t-1}^1, ..., \pi_{t-1}^d\right)$:

$$\begin{aligned} \pi_t^j &\propto \sum_{i=1}^{d} \frac{1}{\sigma\sqrt{g\left(m^i\right)}} f_u\left(\frac{r_{t-1}}{\sigma\sqrt{g\left(m^i\right)}}\right) \pi_{t-1}^i a_{i,j} \\ a_{i,j} &= \Pr[M_t = j \left| M_{t-1} = i\right] = \prod_{k=1}^{\overline{k}} \left[(1-\gamma_k) 1_{\left[m_k^i = m_k^j\right]} + \frac{\gamma_k}{2}\right]. \end{aligned}$$

Hence, unlike continuous stochastic volatility models, the Markov-switching multifractal model has a closed-form likelihood, precisely because the filtering techniques a la Hamilton can be applied. However, the price to pay for a volatility process with a discrete state space is that, for sake of goodness of fit, it often takes a state space with many elements, which implies a large number of multipliers $\overline{k}$. Calvet and Fisher (2004) documents that for exchange rate data, the multifractal model "works better for larger values of $\overline{k}$" and choose to set the focus on the case $\overline{k} = 10$ for all currencies.

While the log-likelihood is available in closed-form, a single evaluation requires $O\left(2^{2\overline{k}}T\right)$ computations, where $O\left(\cdot\right)$ denotes the order of the evaluation. Therefore, if the upper bound on the parameter space for $\overline{k}$ is too large, estimation via maximum likelihood becomes prohibitively expensive.

Given the potentially prohibitive computational requirements associated with a large value of $\overline{k}$, it is worth revisiting the likelihood function with the false equality constraint $\overline{k} = 2$, which is the smallest possible value of $\overline{k}$ allowing to identify all the other parameters. Under the constraint $\overline{k} = 2$, a single likelihood evaluation requires only $16 \cdot T$, i.e., $2^4 T$, computations. Therefore, such a constraint could easily be imposed, and the resulting estimation procedure implemented, to alleviate the computational burden associated with searching over the entire parameter space for $\overline{k}$.

2.3 Pseudo-Score Vector

In the above examples, for all values of θ in a subset $\Theta_0 \subset \Theta$, obtained by imposing some (possibly false) equality constraints, the log-likelihood function $L_T(\theta)$ in (3) is available in closed-form. For all $\theta \in \Theta_0$, the closed-form log-likelihood function will often allow us to compute a closed-form pseudo-score vector, denoted by

$$M_T(\theta);\ \theta \in \Theta_0, \tag{8}$$

which could be used as the basis for inference on the unknown θ^0. In the case where derivatives of the log-likelihood function $L_T(\theta)$ exist for all $\theta \in \Theta_0$, we set

$$M_T(\theta) = \partial L_T(\theta)/\partial\theta.$$

The notation $M_T(\theta)$ in (8) is used instead of the more explicit $\partial L_T(\theta)/\partial\theta$ notation since there exist situations where exact partial derivatives do not exist, which will then require us to use approximate derivatives. For example, an approximation will be required when some components of θ are integers, such as $\overline{k}$ in the multifractal case (Example 2).

Importantly, we note that the pseudo-score vector in (8) is of the same dimension as the unknown parameters, i.e., it is a p-dimensional vector. That is, the partial derivatives for the pseudo-score are computed with respect to all components of θ, including those dimensions whose values are fixed when $\theta \in \Theta_0$. In the following, we demonstrate that, in the examples considered above, constraining $\theta \in \Theta_0$ allows us to compute the pseudo-score in closed-form, at least up to the evaluation of univariate integrals.

Example 1: (*Generalized Tobit Model*) The generalized Tobit model is a striking example of the fact that even though the complete likelihood function can only be stated as the product of T univariate integrals, which have no closed-form and must be numerically calculated, the sub-model defined by $\theta_3 = 0$ is as simple as the usual Tobit likelihood. Moreover, under this constraint the partial derivatives of the likelihood function are also available in closed-form.

Recall that the log-likelihood for the generalized Tobit model is given by

$$\begin{aligned} L_T(\theta) &= \frac{1}{T}\sum_{i\in I_1}\log\left[\frac{1}{\sigma}\varphi\left(\frac{y_i - x_i'\theta_1}{\sigma}\right)\Pr[y_{2i}^* \geq 0\,|y_i, z_i, \theta_2, \theta_3]\right] + \frac{1}{T}\sum_{i\in I_0}\log\left[\Pr[y_{2i}^* < 0\,|z_i, \theta]\right] \\ &= L_{1,T}\left(\theta\right) + L_{2,T}\left(\theta\right), \end{aligned}$$

where

$$\begin{aligned} \Pr[y_{2i}^* &< 0\,|z_i, \theta] = \int \Pr[y_{2i}^* < 0\,|y_{1i}^*, z_i, \theta_2, \theta_3]\,\frac{1}{\sigma}\varphi\left(\frac{y_{1i}^* - x_i'\theta_1}{\sigma}\right)dy_{1i}^*, \\ \Pr[y_{2i}^* &< 0\,|y_{1i}^*, z_i, \theta_2, \theta_3] = \left[1 + \exp\left(z_i'\theta_2 + \theta_3 y_{1i}^*\right)\right]^{-1}. \end{aligned}$$

As was noted previously, under the restrictions $\theta_3 = 0$, the above log-likelihood has a simple closed form.

The score of this likelihood under the restriction $\theta_3 = 0$ can also be obtained in closed-form. First, we can compute

$$\begin{aligned} \frac{\partial L_{1,T}\left(\theta_1, \theta_2, 0, \sigma\right)}{\partial\theta_1} &= -\frac{1}{T}\sum_{i\in I_1} x_i\left[\frac{y_i - x_i'\theta_1}{\sigma^2}\right], \quad \frac{\partial L_{1,T}\left(\theta_1, \theta_2, 0, \sigma\right)}{\partial\theta_2} = \frac{1}{T}\sum_{i\in I_1} z_i\left[1 + e^{z_i'\theta_2}\right]^{-1} \\ \frac{\partial L_{1,T}\left(\theta_1, \theta_2, 0, \sigma\right)}{\partial\theta_3} &= \frac{1}{T}\sum_{i\in I_1} y_i\left[1 + e^{z_i'\theta_2}\right]^{-1}, \quad \frac{\partial L_{1,T}\left(\theta_1, \theta_2, 0, \sigma\right)}{\partial\sigma} = \frac{1}{T}\sum_{i\in I_1}\left[-\frac{1}{\sigma} + \frac{(y_i - x_i'\theta_1)^2}{\sigma^3}\right] \end{aligned}$$

While we can also check that

$$\begin{aligned} \frac{\partial L_{2,T}\left(\theta_1, \theta_2, 0, \sigma\right)}{\partial\theta_1} &= 0, \quad \frac{\partial L_{2,T}\left(\theta_1, \theta_2, 0, \sigma\right)}{\partial\sigma} = 0, \\ \frac{\partial L_{2,T}\left(\theta_1, \theta_2, 0, \sigma\right)}{\partial\theta_2} &= -\frac{1}{T}\sum_{i\in I_0} z_i\left[1 + e^{-z_i'\theta_2}\right]^{-1}, \\ \frac{\partial L_{2,T}\left(\theta_1, \theta_2, 0, \sigma\right)}{\partial\theta_3} &= -\frac{1}{T}\sum_{i\in I_0} x_i'\theta_1\left[1 + e^{-z_i'\theta_2}\right]^{-1}. \end{aligned}$$

The pseudo-score can then be taken as the above derivatives, computed under the restriction $\theta_3 = 0$, i.e.,

$$M_T(\theta) = \frac{\partial L_{1,T}(\theta_1, \theta_2, 0, \sigma)}{\partial\theta} + \frac{\partial L_{2,T}(\theta_1, \theta_2, 0, \sigma)}{\partial\theta}.$$

Example 2: (*Markov-Switching Multifractal (MSM) Model*) For this model, the structural parameter vector is given by:

$$\theta = \left(\zeta', \overline{k}\right)', \zeta = \left(m_0, \bar{\gamma}, b, \sigma\right)'.$$

As already announced, if we consider this model under the false equality constraint

$$\overline{k} = 2,$$

the log-likelihood associated with observed data $\{r_t\}_{t=1}^{T}$ is given by

$$L_T(\zeta, 2) = \frac{1}{T}\sum_{t=1}^{T}\log\left(\sum_{j=1}^{4}\frac{1}{\sigma\sqrt{g(m^j)}}f_u\left(\frac{r_t}{\sigma\sqrt{g(m^j)}}\right)\Pr[M_t = m^j\,|r_\tau, \tau < t]\right).$$

We can then define a pseudo-score vector by

$$M_T(\zeta, 2) = \left(\frac{\partial L_T(\zeta, 2)}{\partial \zeta'}, L_T(\zeta, 3) - L_T(\zeta, 2)\right)'.$$

Note that filtered $\Pr[M_t = m^j\,|r_\tau, \tau < t]$ probabilities depend on all structural parameters as explained above through in particular two transition probabilities:

$$\gamma_1 = \frac{\bar{\gamma}}{b}, \gamma_2 = \bar{\gamma}.$$

2.4 Pseudo-Score Matching and AML Estimation

The previous examples have exemplified the computation of pseudo-score vectors $M_T(\theta)$ as either partial derivatives of the log-likelihood function $L_T(\theta)$ calculated under $\theta \in \Theta_0$, i.e.,

$$M_T(\theta) := \partial L_T(\theta)/\partial\theta = \frac{1}{T}\sum_{t=1}^{T}\partial\log\left(l\{y_t\,|(y_\tau)_{1\le\tau\le t-1}, x_t, z_0; \theta\}\right)/\partial\theta,$$

or as containing a mix of partial derivatives, and other approximate derivatives of $L_T(\theta)$. In addition, even when we compute genuine partial derivatives, they are most likely computed under some simplifying false equality constraints. For these two reasons, while $M_T(\theta)$ can be feasibly computed, it is important to realize that minimizing $\|M_T(\theta)\|$ is unlikely to deliver a consistent estimator of θ_0.

In general, the restrictions used to produce the pseudo-score will not be satisfied at θ^0, so that $M_T(\theta)$, computed under $\theta \in \Theta_0$, will not characterize the argument maximum of $L_T(\theta)$ computed under $\theta \in \Theta$. To make this explicit, we maintain the following assumption throughout the remainder.

Assumption A1(*False Equality Constraints*): The parameter space Θ is compact and can be partitioned as

$$\begin{aligned}\Theta &= \Theta^1 \times \Theta^2, \quad \Theta^1 \subset \mathbb{R}^{p_1}, \Theta^2 \subset \mathbb{R}^{p_2}, \quad p = p_1 + p_2 \\ \Theta_0 &= \Theta^1 \times \left\{\left(\beta_j^0\right)_{p_1 < j \le p}\right\} = \Theta^1 \times \left\{\beta^{2,0}\right\}\end{aligned}$$

and the map

$$\beta^1 = (\theta_j)_{1\le j\le p_1} \mapsto M_T[(\beta^{1'}, \beta^{2,0'})']$$

is continuously differentiable on the interior of Θ^1.

Assumption A1 clarifies precisely how the false equality constraints artificially restrict certain directions of the parameter space. Consequently, under **Assumption A1** the pseudo-score $M_T(\theta)$ is actually targetting a vector of parameters that differs from the generic elements $\theta \in \Theta$, and for which we denote throughout the remainder as $\beta \in \Theta_0$. In general, minimizing

$\|M_T(\beta)\|$ over $\beta \in \Theta_0$ will only deliver an estimator $\hat{\beta}_T$ ($\hat{\beta}_T \in \Theta_0$) that is consistent for some pseudo-true value β^0. We emphasize that $\hat{\beta}_T$ is an estimator of β^0, and not θ^0, since the constraints are chosen for computational convenience and thus are unlikely to be satisfied at $\theta = \theta^0$.

However, given $M_T(\beta)$ and $\hat{\beta}_T$ it is feasible to consistently estimate θ^0 using simulation from the structural model. Consider the log-likelihood function computed for a simulated path $\{\tilde{y}_t^{(h)}(\theta, z_0)\}_{t=1}^T$ (for $h = 1, \ldots, H$) and at a value $\beta \in \Theta_0$ of the structural parameters:[3]

$$L_T^{(h)}(\theta, \beta) = \frac{1}{T}\sum_{t=1}^{T} \log\left(l\left\{\tilde{y}_t^{(h)}(\theta) \,\middle|\, \left(\tilde{y}_\tau^{(h)}(\theta)\right)_{1\le\tau\le t-1}, x_t; \beta\right\}\right). \tag{9}$$

Associated to $L_T^{(h)}(\theta, \beta)$ is a simulated pseudo-score vector

$$M_T^{(h)}(\theta, \beta); \beta \in \Theta_0.$$

In accordance with $M_T(\beta)$, in the case where partial derivatives of $L_T^{(h)}(\theta, \beta)$ exist, and for $\beta \in \Theta_0$, we take

$$M_T^{(h)}(\theta, \beta) = \left\{\frac{\partial}{\partial \xi} L_T^{(h)}(\gamma, \xi)\right\}_{\gamma=\theta, \xi=\beta} = \frac{1}{T}\sum_{t=1}^{T}\left\{\partial \log\left(l\left\{\tilde{y}_t^{(h)}(\gamma) \,\middle|\, \left(\tilde{y}_\tau^{(h)}(\gamma)\right)_{1\le\tau\le t-1}, x_t; \xi\right\}\right)/\partial\xi\right\}_{\gamma=\theta,\xi=\beta},$$

and with $M_T^{(h)}(\theta, \beta)$ comprising a mix of partial and approximate derivatives in situations where not all partial derivatives exist. We note that the derivatives (approximate or exact) defining $M_T^{(h)}(\theta, \beta)$ are not computed with respect to the value $\gamma \in \Theta$ that is used to generate simulated data but with respect to the vector of structural parameters θ that enters as the variable of the likelihood function (given a simulated sample path). This variable is eventually only considered when it fulfils the possibly false equality constraints, so that it belongs to Θ_0 and is the reason why we eventually denote it by β.

We can now define our pseudo-score matching estimator of θ^0 as follows.

Definition 1: The Approximate Maximum Likelihood (AML) estimator $\hat{\theta}_{T,H}$ of θ^0 is defined as the exact solution to

$$M_T(\hat{\beta}_T) - \frac{1}{H}\sum_{h=1}^{H} M_T^{(h)}(\hat{\theta}_{T,H}, \hat{\beta}_T) = 0. \tag{10}$$

The AML estimator in (10) is the solution of p nonlinear equations, in p unknown parameters, so that we may expect existence of such a solution. However, in practice it will be safer to minimize a squared norm of a difference between the two terms on the left hand side of equation (10). The fact that the system in (10) is just identified tells us that asymptotically, the behavior of the minimum should not depend on the weighting matrix used in the squared norm, insofar as (10) asymptotically defines a unique solution, which, hopefully coincides with the true unknown value θ^0. This will be the purpose of the main identification assumption (stated in Section 3).

It is also useful to point out that since the estimated auxiliary parameters, $\hat{\beta}_T$, show up on both sides of equation (10), their asymptotic distribution will ultimately not impact the asymptotic distribution of AML estimator. As proven in Section 3, only the pseudo-true value of $\hat{\beta}_T$ will impact the asymptotic distribution of the AML estimator.

[3] For the sake of notational simplicity, we have not made explicit the dependence of the likelihood function on the initial value z_0 of the simulated data. Since we are confining ourselves to standard settings, the dependence of $L_T^{(h)}$ on z_0 will be immaterial asymptotically.

2.5 Comparison with I-I Approaches

2.5.1 Score Matching a la Gallant and Tauchen (1996)

The pseudo-score that is considered by Gallant and Tauchen (1996) (GT hereafter) is not, in general, a proxy of the structural score where the parameter vector β is of the same dimension as the structural parameter vector θ. On the contrary, GT consider an auxiliary model with likelihood function

$$Q_T(\beta) = \frac{1}{T}\sum_{1\leq t\leq T} \log\left(q\{y_t \left|(y_\tau)_{1\leq\tau\leq t-1}\right., x_t, z_0; \beta\}\right),\ \beta \in B \subset \mathbb{R}^b,$$

where $q\{y_t \left|(y_\tau)_{1\leq\tau\leq t-1}\right., x_t, z_0; .\}$ is not, in general, the true transition density of the process $\{y_t\}_{t=1}^T$, and is a pseudo-likelihood in the sense of Gourieroux, et al. (1984); which is precisely the reason for using the notations $q\{.\,|.\}$ and $Q_T(\cdot)$ instead of $l\{.\,|.\}$ and $L_T(\cdot)$. The pseudo maximum likelihood estimator $\hat{\beta}_T$ then satisfies

$$\frac{\partial}{\partial\beta}Q_T(\hat{\beta}_T) = 0.$$

Using $\hat{\beta}_T$, GT define an I-I estimator $\hat{\theta}_{T,H}$ of θ^0 as the solution to

$$\min_{\theta\in\Theta}\left\|\frac{1}{H}\sum_{h=1}^H \frac{\partial}{\partial\beta}Q_T^{(h)}(\theta, \hat{\beta}_T)\right\|_{W_T}^2, \tag{11}$$

for W_T a positive-definite matrix, and where $\|x\|_{W_T}^2 = x'W_T x$. While GT only consider the case $H = \infty$, the above definition is indeed the extension of GT proposed by GMR. In GMR, the authors demonstrate that the estimator $\hat{\theta}_{T,H}$ described above is asymptotically equivalent to the standard I-I estimator based on matching estimators of β, and which implicitly requires

$$\dim(\beta) = b \geq p = \dim(\theta).$$

The GT estimator $\hat{\theta}_{T,H}$ can be equivalently viewed as the solution of

$$\min_{\theta}\left\|\frac{\partial}{\partial\beta}Q_T(\hat{\beta}_T) - \frac{1}{H}\sum_{h=1}^H \frac{\partial}{\partial\beta}Q_T^{(h)}\left(\theta, \hat{\beta}_T\right)\right\|_{W_T}^2.$$

In a simplistic setting where $\hat{\beta}_T$ is not subject to false equality constraints, so that the auxiliary model is taken to be the unconstrained structural model, the pseudo-likelihood $Q_T(\cdot)$ coincides with the true likelihood $L_T(\cdot)$ and the GT I-I estimator coincides with our AML estimator. However, it is worth keeping in mind that our philosophy for AML is precisely the opposite: we are explicitly concerned with cases where, by the nature of the constraints we employ,

$$\frac{\partial}{\partial\beta}Q_T(\hat{\beta}_T) \neq 0.$$

A consequence of this difference in estimation philosophy is that GT underpin the accuracy of the I-I estimator $\hat{\theta}_{T,H}$ by the asymptotic distribution of the auxiliary estimator $\hat{\beta}_T$. This point of view can be seen via a Taylor expansion of the first-order conditions

$$\left[\frac{1}{H}\sum_{h=1}^H \frac{\partial^2}{\partial\beta\partial\theta'}Q_T^{(h)}(\hat{\theta}_{T,H}, \hat{\beta}_T)\right]' W_T \frac{\sqrt{T}}{H}\sum_{h=1}^H \frac{\partial}{\partial\beta}Q_T^{(h)}(\hat{\theta}_{T,H}, \hat{\beta}_T) = 0.$$

Defining

$$J^0(\theta^0,\beta^0) = \underset{T\to\infty}{\text{plim}}\frac{\partial^2}{\partial\beta\partial\theta'}Q_T^{(h)}(\theta^0,\beta^0) \text{ and } K^0(\beta^0) = \underset{T\to\infty}{\text{plim}}\frac{\partial^2}{\partial\beta\partial\beta'}Q_T(\beta^0),$$

(and with abuse of notation as if $L_T = Q_T$), we see that

$$\begin{aligned} o_P(1) &= J^0\left(\theta^0,\beta^0\right)' W_T \frac{\sqrt{T}}{H}\sum_{h=1}^{H}\frac{\partial}{\partial\beta}Q_T^{(h)}\left(\theta^0,\beta^0\right) \\ &\quad + J^0\left(\theta^0,\beta^0\right)' W_T K^0\left(\beta^0\right)\sqrt{T}\left(\hat{\beta}_T-\beta^0\right) + J^0\left(\theta^0,\beta^0\right)' W_T J^0\left(\theta^0,\beta^0\right)\sqrt{T}\left(\hat{\theta}_{T,H}-\theta^0\right) \end{aligned}$$

GMR show that the above Taylor expansion allows us to view $\sqrt{T}(\hat{\theta}_{T,H}-\theta^0)$ as an asymptotically linear function of the difference between $\hat{\beta}_T$ and a similar estimator computed on simulated data (see the part of their Appendix 1 entitled "The Third Version of the Indirect Estimator"). For this reason, and for H large the asymptotic distribution of $\sqrt{T}(\hat{\theta}_{T,H}-\theta^0)$ is entirely determined by the asymptotic distribution of $\sqrt{T}(\hat{\beta}_T-\beta^0)$.

As the above expansion clarifies, the fact that $\sqrt{T}(\hat{\theta}_{T,H}-\theta^0)$ depends on the distribution of $\sqrt{T}(\hat{\beta}_T-\beta^0)$ ultimately impacts the estimators efficiency. In particular, if we consider the case where $H=\infty$, it is precisely this dependence on $\sqrt{T}(\hat{\beta}_T-\beta^0)$ that allows one to achieve efficiency in the EMM estimation framework: viewing $\sqrt{T}(\hat{\beta}_T-\beta^0)$ as an asymptotically linear functional of some score generator, and if we allow the dimension of the score generator (i.e., the auxiliary parameters) to increase as T increases, then eventually it may span the score vector of the structural model. It is therefore feasible to suggest that one may eventually achieve fully efficient inferences for θ^0.

However, the drive for efficient inference in EMM ultimately entails estimating a high-dimensional vector of auxiliary parameters, β, and this use of a high-dimensional auxiliary model may lead to over-fitting and a lack of robustness when the structural model is partly mis-specified; see Dridi and Renault (2000) for further discussion on the potential lack of robustness of EMM to misspecification of the structural model. In contrast to EMM and score matching I-I estimators, later we will demonstrate that the distribution of the AML estimator does not depend on the asymptotic distribution of the estimated auxiliary parameters (see **Proposition 2** for details).

2.5.2 Score Matching a la Calzolari, Fiorentini and Sentana (2004)

Consider that the false equality constraints under which AML is implemented can be written in the implicit form

$$g(\theta) = 0,$$

for some given function $g:\Theta\to\mathbb{R}^{d_g}$, with $d_g < p$. Recall that the log-likelihood function $L_T(\theta)$ is assumed to be tractable for the set of parameters satisfying this constraint. It is then possible to estimate the parameters from the Lagrangian function

$$\mathcal{L}_T(\beta,\lambda) = L_T(\beta) + g(\beta)'\lambda,$$

where $\lambda \in \mathbb{R}^{d_g}$ is the vector of Lagrange multipliers. The estimator $\hat{\zeta}_T = (\hat{\beta}_T', \hat{\lambda}_T')'$ can then be defined from the first-order conditions

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}_T(\hat{\beta}_T, \hat{\lambda}_T)}{\partial \beta} = \frac{\partial}{\partial \beta} L_T(\hat{\beta}_T) + \frac{\partial g(\hat{\beta}_T)'}{\partial \beta} \hat{\lambda}_T, \\ 0 &= g(\hat{\beta}_T). \end{aligned}$$

From these conditions, the philosophy of I-I estimation put forth by Calzolari et al. (2004) is to argue that score matching should be corrected by the information contained in the Lagrange multipliers. In our context, this philosophy would lead us to estimation θ^0 by $\hat{\theta}_{T,H}$, the solution to,

$$\frac{1}{H} \sum_{h=1}^{H} \frac{\partial}{\partial \beta} L_T^{(h)} \left(\hat{\theta}_{T,H}, \hat{\beta}_T \right) + \frac{\partial g(\hat{\beta}_T)'}{\partial \beta} \hat{\lambda}_T = 0, \tag{12}$$

which is equivalent to solving

$$\frac{1}{H} \sum_{h=1}^{H} \frac{\partial}{\partial \beta} L_T^{(h)} \left(\hat{\theta}_{T,H}, \hat{\beta}_T \right) - \frac{\partial}{\partial \beta} L_T \left(\hat{\beta}_T \right) = 0,$$

and coincides with our AML estimator.[4]

Our claim is that, even when we have no such thing as Lagrange multipliers $\hat{\lambda}_T$ to encapsulate the information about the violation of constraints (information that should be added to the information brought by the constrained estimators $\hat{\beta}_T$), it still makes sense to imagine that the full score vector accounts for this missing information. This will be confirmed by our general analysis in the next subsections.

In addition, it is worth noting that even though our AML approach is similar to the I-I estimators proposed in Calzolari et al. (2004), it stems from a completely different point of view. We have defined an auxiliary model with parameter vector β as a version of the structural model that has been simplified. In contrast to Calzolari et al. (2004), we never contemplate simplifying the auxiliary model, which in their case has already chosen to be a simple approximation to the structural model.

2.5.3 Indirect Inference a la Calvet and Czellar

The examples in Section 2.2 demonstrate that there are important cases where imposing a simplifying constraint of the form $\theta = h(\gamma), \gamma \in \mathbb{R}^d, d < p$, results in an auxiliary model that is a computationally feasible version of the structural model of interest. As explained in Calvet and Czellar (2015): "Since [under the constraints] the auxiliary and structural models are then

[4] It is worth knowing that Calzolari et al. (2004) also contemplate the I-I estimator defined by (12) in the case of inequality constraints on the auxiliary parameters, so that $\hat{\lambda}_T$ is a vector of Kuhn-Tucker multipliers. In this case, the argument to consider the recentered score vector (12) instead of a score vector (11) a la Gallant and Tauchen (1996) is not any more to correct for a misspecification bias but to hedge against possible non asymptotic normality of estimators constrained by inequality restrictions. Then, it can be shown (see also Frazier and Renault (2019) for a detailed asymptotic theory in case of parameters near the boundary of the parameter space) that making the difference of the two score vectors as in (10) will restore asymptotic normality even though each of them is not asymptotically normal, due to the fact that the inequality constrained estimator $\hat{\beta}_T$ is not asymptotically normal.

closely related, the resulting indirect inference estimator is expected to have good accuracy properties."

Calvet and Czellar (2015) propose to use estimators of the auxiliary parameters based on the observed data, say $\hat{\gamma}_T$, and the simulated data, say $\tilde{\gamma}_T(\theta)$, to estimate the structural parameters. However, while $\hat{\gamma}_T$ and $\tilde{\gamma}_T(\theta)$ can often be obtained relatively easily, it is important to realize that these auxiliary parameters can not generally identify the structural parameters θ, except in the unlikely case that the constraints $\{\exists\gamma \in \Gamma, \theta = h(\gamma)\}$ are satisfied at θ^0 (the true value of the structural parameters).

To circumvent this identification issue, Calvet and Czellar (2015) propose to add additional auxiliary statistics, with dimension at least as large as $p - d$, within the I-I procedure. Denote these statistics based on observed data by $\hat{\eta}_T$ and simulated data by $\tilde{\eta}_T(\theta)$, then Calvet and Czellar (2015) propose to estimate θ from the following program: for $\hat{\beta}_T := (\hat{\gamma}'_T, \hat{\eta}'_T)'$, $\tilde{\beta}_T(\theta) := (\tilde{\gamma}_T(\theta)', \tilde{\eta}_T(\theta)')'$, an estimator of θ^0 can be obtained by

$$\min_{\theta\in\Theta} \left(\hat{\beta}_T - \tilde{\beta}_T(\theta)\right)' W \left(\hat{\beta}_T - \tilde{\beta}_T(\theta)\right), \tag{13}$$

where W is a positive-definite weighting matrix of conformable dimension.

In a sense, the approach of Calvet and Czellar (2015) follows the idea of estimation under the null that is commonly encountered in testing situations in econometrics; namely, we estimate a simpler version of the model that is formed as a constrained version of the model we assume has actually generated the data, and then we construct statistics about this simpler model to determine whether or not the simpler model is appropriate to model the observed data. Several remarks are in order.

First, it is important to keep in mind that for the minimization program (13), the simulated data are obtained from the unconstrained structural model, meaning by considering possibly any $\theta \in \Theta$ and not only $\theta \in \Theta^0 = \{\theta \in \Theta; \exists\gamma \in \Gamma, \theta = h(\gamma)\}$.

Second, since the Calvet and Czellar (2015) approach directly imposes the constraints in explicit form within the structural model, they obtain what they consider as an "unconstrained" auxiliary model. The result is that this approach will generate simple auxiliary estimators of β. However, the downside is that since we have disregarded the impact of the constraints the approach can not identify the entire vector of structural parameters without resorting to ad-hoc statistics. While the addition of $\hat{\eta}_T$ to the auxiliary estimators may result in a vector of statistics that can identify θ^0, the precise choice of $\hat{\eta}_T$ in any given example is somewhat arbitrary and likely sub-optimal.

Third, for sake of efficient inference, one should realize that, by definition, the estimator of the simplified structural model (indexed by a lower dimensional parameter), while convenient, overlooks relevant information. In the following section, we demonstrate that AML can, in a sense, account for this information loss, and, thus, get close to the efficiency of maximum likelihood estimation without giving up the convenient simplification of our structural model.

3 Asymptotic Behavior of AML Estimators

In this section, we describe the asymptotic distribution of the AML estimator $\hat{\theta}_{T,H}$ given in

Definition 1, i.e., the solution, in θ, to

$$M_T(\hat{\beta}_T) = \frac{1}{H}\sum_{h=1}^{H} M_T^{(h)}(\theta, \hat{\beta}_T). \tag{14}$$

The asymptotic theory of this estimator is not completely standard since, for each $h = 1, ..., H$, $M_T^{(h)}(\theta, \hat{\beta}_T)$ is a sample mean of T terms, each of them depending on $\hat{\beta}_T$, hence it is a double array.

Since $M_T(\beta^0)$ is a pseudo-score, and may include components that can not be represented as partial derivatives of $L_T(\cdot)$, we follow van der Vaart (1998) (Chapter 5) and refer to $\hat{\theta}_{T,H}$ as a Z-estimator of θ^0. Moreover, it is worth recalling that we do not accommodate here the case where one component of the structural parameter vector is an integer. The discussion of this case could be achieved by extending the range of the integer parameter to the complete set of non-negative real numbers, which is feasible by a piecewise linear extension.

3.1 Consistency

Consistency of the AML estimator $\hat{\theta}_{T,H}$, defined in (14), can be obtained under the hypothesis that $\hat{\beta}_T$ is consistent for some pseudo-true value β^0. However, to deduce an asymptotic distribution for $\hat{\theta}_{T,H}$ we will ultimately require that $\sqrt{T}(\hat{\beta}_T - \beta^0) = O_P(1)$. Consistency of $\hat{\theta}_{T,H}$, for θ^0, can be seen to follow by applying Theorem 5.9 in van der Vaart (1998), under the following regularity condition.

Assumption B1: (i) There exists some $\beta^0 \in \Theta_0$ such that $\sqrt{T}(\hat{\beta}_T - \beta^0) = O_p(1)$; (ii) There exists a vector function $M : \Theta \times \Theta_0 \to \mathbb{R}^p$ such that, for any $h = 1, ..., H$, and any $\gamma > 0$,

$$\sup_{\theta \in \Theta} \sup_{\|\beta - \beta^0\| \leq \gamma/\sqrt{T}} \left\| M_T^{(h)}(\theta, \beta) - M\left(\theta, \beta^0\right) \right\| = o_P(1);$$

(iii) For every $\varepsilon > 0$,

$$\inf_{\theta \in \Theta : \|\theta - \theta^0\| \geq \varepsilon} \left\| M\left(\theta, \beta^0\right) - M\left(\theta^0, \beta^0\right) \right\| > 0.$$

From the i.i.d. nature of the simulation, and the definition of the simulated log-likelihood $L_T^{(h)}(\theta, \beta^0)$ in (9), it is not restrictive to assume that $M(\theta, \beta^0)$ does not depend on h. Furthermore, this condition similarly implies that $M_T(\beta^0)$ converges towards $M(\theta^0, \beta^0)$. Under **Assumption B1**, we can state the following result.

Proposition 1: Under **Assumptions A1 and B1**, the AML estimator $\hat{\theta}_{T,H}$ is a consistent estimator of the true unknown value θ^0: $\text{plim}_{T \to \infty} \hat{\theta}_{T,H} = \theta^0$. □

We now illustrate the identification condition in **Assumption B1**(iii) in two examples, and demonstrate that this condition is similar to the identification condition required by maximum likelihood. For the purpose of these illustrations, we only consider that **Assumption B1** enforces

$$M\left(\theta, \beta^0\right) - M\left(\theta^0, \beta^0\right) \neq 0, \forall \theta \neq \theta^0.$$

That is, we temporarily overlook the fact that the well-separated minimum of $\|M(\theta, \beta^0) - M(\theta^0, \beta^0)\|$ generally requires additional regularity, e.g., continuity of the function $M(., \beta^0)$ and compactness of Θ.

Example: Well-specified Models

Assume that $M_T^{(h)}(\theta,\beta)$ is the score vector of a well-specified parametric model for which $\beta^0=\theta^0$ is the true unknown value of the parameters, i.e.,

$$M_T^{(h)}(\theta,\beta)=\frac{1}{T}\sum_{t=1}^{T}\frac{\partial\log\left[l\{\tilde{y}_t^{(h)}(\theta)|\{\tilde{y}_\tau^{(h)}(\theta)\}_{1\leq\tau\leq t-1},x_t;\beta\}\right]}{\partial\beta}.$$

Under standard regularity conditions

$$M(\theta,\beta)=E_\theta\left\{\frac{\partial\log\left[l\{y_t|\{y_\tau\}_{1\leq\tau\leq t-1},x_t;\beta\}\right]}{\partial\beta}\right\},$$

where E_θ denotes expectation computed under the probability distribution of the process $\{y_t\}_{t=1}^T$ at the parameter value θ. The standard identification condition for maximum likelihood is then

$$M(\theta,\beta)=0\Longleftrightarrow\theta=\beta.$$

In particular,

$$M\left(\theta,\beta^0\right)-M\left(\theta^0,\beta^0\right)\neq0,\forall\theta\neq\theta^0=\beta^0.$$

In other words, the identification condition in **Assumption B1** is tantamount to the identification condition for maximum likelihood. □

Example: Exponential Models

Assume that conditionally on $\{x_t\}_{t=1}^T$, the variables y_t are independent, for $t=1,...,T$, and the conditional distribution of y_t only depends on the exogenous variable x_t with the same index. Further, assume that this distribution has a density $l\{y_t\,|x_t;\theta\}$ that is of the exponential form

$$l\{y_t\,|x_t;\theta\}=\exp\left[c\left(x_t,\theta\right)+h(y_t,x_t)+a(x_t,\theta)'T(y_t)\right],$$

where $c(.,.)$ and $h(.,.)$ are given functions and $a(x_t,\theta)$ and $T(y_t)$ are r-dimensional random vectors, all known up to the unknown θ^0. The extension to dynamic models, in which conditioning values would also include lagged values of the process y_t, can also be considered at the cost of additional notations. From

$$\frac{\partial\log\left[l\{y_t\,|x_t;\theta\}\right]}{\partial\theta}=\frac{\partial c\left(x_t,\theta\right)}{\partial\theta}+\frac{\partial a\left(x_t,\theta\right)'}{\partial\theta}T(y_t),$$

since the conditional score vector has, by definition, a zero conditional expectation, we deduce that

$$\frac{\partial L_T\left(\theta\right)}{\partial\theta}=\frac{1}{T}\sum_{t=1}^{T}\frac{\partial a'\left(x_t,\theta\right)}{\partial\theta}\left\{T(y_t)-E_\theta[T(y_t)\,|x_t]\right\}.$$

Defining,

$$m\left(x_t,\theta\right)=E_\theta[T(y_t)\,|x_t]\,,\ \ \Omega\left(x_t,\theta\right),=\mathrm{Var}_\theta[T(y_t)\,|x_t]$$

we have that

$$\frac{\partial a\left(x_t,\theta\right)'}{\partial\theta}=\frac{\partial m'\left(x_t,\theta\right)}{\partial\theta}\Omega^{-1}\left(x_t,\theta\right).$$

Therefore, the maximum likelihood estimator $\hat{\theta}_T$ can be defined as the solution to

$$\frac{\partial L_T(\theta)}{\partial \theta} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial m'(x_t, \theta)}{\partial \theta} \Omega^{-1}(x_t, \theta)\left\{T(y_t) - m(x_t, \theta)\right\} = 0. \tag{15}$$

The first-order conditions (15) show that maximum likelihood is a GMM estimator with optimal instruments for the conditional moment restrictions

$$E_\theta[T(y_t) - m(x_t, \theta) \,|x_t] = 0.$$

Under the assumptions for standard asymptotic theory of efficient GMM (Hansen, 1982), i.e., for all $\theta \in \Theta$, the conditional variance $\Omega(x_t, \theta)$ of the moment conditions is non-singular and the Jacobian matrix $E[\partial m'(x_t, \theta)/\partial \theta \,|x_t]$ is full row rank, the identification condition for consistency of maximum likelihood is that

$$E\left\{\frac{\partial m'(x_t, \theta)}{\partial \theta}\Omega^{-1}(x_t, \theta)\left\{m(x_t, \theta^0) - m(x_t, \theta)\right\}\right\} = 0 \Longrightarrow \theta = \theta^0,$$

or equivalently, by the Law of Iterated Expectation and exogeneity of x_t,

$$E\left\{\frac{\partial m'(x_t, \theta^0)}{\partial \theta}\Omega^{-1}\left(x_t, \theta^0\right)\left\{T(y_t) - m(x_t, \theta)\right\}\right\} = 0 \Longrightarrow \theta = \theta^0.$$

In Appendix B, we demonstrate that the identification condition required by AML amounts to the condition that

$$E\left\{\frac{\partial m'(x_t, \beta^0)}{\partial \theta}\Omega^{-1}\left(x_t, \beta^0\right)\left\{m(x_t, \theta) - m\left(x_t, \theta^0\right)\right\}\right\} \Longrightarrow \theta = \theta^0,$$

which allows us to demonstrate that the AML identification condition is tantamount to the ML identification condition in at least two primary cases of interest: regression models, and unconditional models.

□

While the above example demonstrates the (near) equivalence of the identification conditions for MLE and AML in the context of exponential models, this example gives strong intuition that the two identification conditions are likely to be similar in many different models. After all, many complex models can be seen as exponential models that have been rendered intractable by the introduction of latent variables, and are such that if the latent variables were observed, then the model would be of the exponential form (or nearly so); indeed, the generalized Tobit example is one particular example of this phenomena.

Further to the above point, we stress that the similarity between the identification conditions needed for MLE and AML, respectively, is warranted not only for exponential models but also for a broad class of models that can be seen as particular transformation of an exponential model. We describe in Appendix B2 the case of "latent exponential models" where we only observe $y = g(y^*, x)$ for some known function $g(\cdot, \cdot)$, and where the conditional model of y^* given x is exponential. Following Gourieroux et al. (1987), the concept of generalized residuals allows us to see how the identification analogy between MLE and AML is maintained from (latent) exponential models to a much broader class of models that are not exponential, such as Probit, Tobit (in particular our example of generalized Tobit), Gompit, and Disequilibrium models.

3.2 Asymptotic Normality and Efficiency

To deduce the asymptotic behavior of the AML estimator, we require the following standard regularity conditions.

Assumption A2: Uniformly on the interior of Θ^1, for some $(p \times p_1)$-dimensional matrix K^0,

$$\underset{T\to\infty}{\text{plim}} \frac{\partial M_T\left[(\beta^{1'}, \beta^{2,0'})'\right]}{\partial \beta^{1'}} = -K^0\left[(\beta^{1'}, \beta^{2,0'})'\right],$$

and where $-K^0\left[(\beta^{1'}, \beta^{2,0'})'\right]$ has full column-rank.

Assumption B2: (i) For all $\beta \in \Theta_0$, the map $\theta \mapsto M_T^{(h)}(\theta, \beta)$ is continuously differentiable on the interior of Θ; (ii) There exists a $(p \times p)$-dimensional matrix, $J^0(\theta, \beta)$, with $J^0(\theta^0, \beta^0)$ non-singular, and such that, for any $h = 1, ..., H$ and any $\gamma > 0$,

$$\sup_{\theta\in\Theta} \sup_{\|\beta-\beta^0\|\le\gamma/\sqrt{T}} \left\|\frac{\partial M_T^{(h)}(\theta, \beta)}{\partial\theta'} + J^0(\theta, \beta)\right\| = o_P(1).$$

We are now ready to state the asymptotic distribution results for the AML estimator. Before doing so, we recall that $M_T(\beta)$ denotes the pseudo-score and $M_T^{(h)}(\theta, \beta)$ its simulated counterpart.

Lemma 1: Under **Assumptions A1-A2, and B1-B2**, the AML estimator satisfies

$$\left\|\sqrt{T}M_T\left(\beta^0\right) - \frac{1}{H}\sum_{h=1}^{H}\sqrt{T}M_T^{(h)}(\hat{\theta}_{T,H}, \beta^0)\right\| = o_P(1).$$

□

The definition of the AML estimator in equation (14) together with **Lemma 1** imply that the AML estimator is asymptotically equivalent to an unfeasible AML (UAML) estimator of θ^0 that solves

$$M_T\left(\beta^0\right) = \frac{1}{H}\sum_{h=1}^{H} M_T^{(h)}\left(\theta, \beta^0\right).$$

The equivalence between the UAML and AML estimators suggests that the asymptotic distribution of $\sqrt{T}(\hat{\beta}_T - \beta^0)$, has no impact on the asymptotic distribution of the AML estimator. Stated another way, equivalence between the UAML and AML estimators implies that estimation of $\hat{\beta}_T$ in the first stage has no impact on the (second stage) standard errors of the AML estimator, other than through the pseudo-true value β^0. This is a significant departure from standard I-I estimation, where the distribution of the estimators intimately depends on the asymptotic distribution of $\sqrt{T}(\hat{\beta}_T - \beta^0)$. The following result clarifies the AML estimators dependence on the pseudo-true value β^0.

Proposition 2: Under **Assumptions A1-A2, and B1-B2**, and if

$$\sqrt{T}M_T(\beta^0) - \frac{1}{H}\sum_{h=1}^{H}\sqrt{T}M_T^{(h)}\left(\theta^0, \beta^0\right) \to_d \aleph\left[0, I^0_{(H)}(\theta^0, \beta^0)\right]$$

with

$$I^0_{(H)}(\theta^0,\beta^0) = (1+1/H)\lim_{T\to\infty}\mathrm{Var}\left\{\sqrt{T}M_T\left(\beta^0\right) - E\left[\sqrt{T}M_T\left(\beta^0\right)\mid\{x_t\}_{t=1}^T\right]\right\},$$

then

$$\sqrt{T}(\hat{\theta}_{T,H}-\theta^0)\to_d \aleph\left(0,\Omega_{(H)}\right),\ \ \Omega_{(H)} = \left[J^0\left(\theta^0,\beta^0\right)\right]^{-1}\left[I^0_{(H)}(\theta^0,\beta^0)\right]\left[J^0\left(\theta^0,\beta^0\right)\right]^{-1\prime}.$$

□

Proposition 2 demonstrates that the (first stage) estimator $\hat{\beta}_T$ only impacts the the statistical efficiency of the AML estimator $\hat{\theta}_{T,H}$ through the pseudo-true value, β^0, that it converges to. We recall that this feature of AML is in contrast to the standard score-based I-I approaches put forward by GT and GMR: the asymptotic distribution of all I-I estimators can be characterized as an asymptotically linear function of $\sqrt{T}(\hat{\beta}_T-\beta^0)$ (see Section 2.5.1 for details).

A natural question to ask is how close is the asymptotic variance matrix $\Omega = \lim_{H\to\infty}\Omega_{(H)}$ to the Cramer-Rao efficiency bound. It is important to realize that efficiency loss can only occur if $\beta^0\neq\theta^0$ or if the pseudo score vector $M_T\left(\theta^0\right)$ is not the true score vector.

Proposition 3: Under the assumptions of **Proposition 2**, if for all $\theta\in\Theta_0$

$$M_T\left(\theta\right) = \frac{1}{T}\sum_{t=1}^{T}\partial\log\left(l\{y_t\left|(y_\tau)_{1\leq\tau\leq t-1}\right., x_t, z_0, \theta\}\right)/\partial\theta,$$

where $l\{y_t\mid(y_\tau)_{1\leq\tau\leq t-1}, x_t, z_0, \theta\}$ denotes the likelihood in equation (2), and if $\beta^0=\theta^0$ and $H\to\infty$, then the AML estimator achieves the Cramer-Rao efficiency bound. □

It is also important to note that even if $M_T\left(\beta^0\right)$ is accurately computed at the pseudo-true value β^0, the matrix

$$I^0\left(\theta^0,\beta^0\right) = \lim_{T\to\infty}\mathrm{Var}\left\{\sqrt{T}M_T\left(\beta^0\right) - E\left[\sqrt{T}M_T\left(\beta^0\right)\mid\{x_t\}_{t=1}^T\right]\right\}$$

will coincide with the Fisher Information Matrix only if

$$\lim_{T\to\infty}\mathrm{Var}\left\{E\left[\sqrt{T}M_T\left(\beta^0\right)\mid\{x_t\}_{t=1}^T\right]\right\} = 0.$$

This property is unlikely to be fulfilled in the case of a conditional model when $\beta^0\neq\theta^0$. However, it is automatically fulfilled in a model that is not conditional. Moreover, it is possible to analytically calculate the proximity between the asymptotic variances of AML and genuine maximum likelihood in the, previously considered, case of exponential models. We refer the interested reader to Appendix B for full details.

4 Examples

In this section, we apply AML to two of the examples considered in Section 2.2. First, we analyze the repeated sampling behavior of AML in the confines of the generalized Tobit model, with a pseudo-score computed under the false inequality constraint discussed in Section 2.2.1. Next, we evaluate the performance of AML relative to ML in the MSM model, described in Section 2.2.2, and use AML to estimate the MSM model on daily S&P500 returns. The empirical results suggest a large value of $\overline{k}$ for this data, which ensures ML can not be feasibly implemented.

4.1 Example 1: Generalized Tobit Model

We first illustrate the performance of AML in the generalized Tobit-type model via a Monte Carlo study, and compare the standard errors of the AML estimator calculated via Monte Carlo to their asymptotic counterparts (given in Proposition 2).

We generate 10,000 replications from the structural model in equations (4)-(5), jointly with the logistic distribution specification for y_{2i}^*, as in equation (6), for three different samples sizes $T = 100$, $T = 1,000$ and $T = 10,000$. We fix the true parameter values at $\theta_1 = \theta_2 = 0.2$, and $\theta_3 = 1$, and the scale parameter for the model is set to $\sigma = 0.5$. The explanatory variables x_i and z_i are generated i.i.d. from the uniform distribution on $[0, 1]$. For AML, we take $H = 10$ simulated samples.

For each Monte Carlo replication, we calculate the constrained auxiliary estimates (using the false equality constraint $\theta_3 = 0$ as described in subsection 2.2.1) and the associated AML estimates. We compare the resulting estimates graphically in Figure 1. For each of the parameters, the left figure represents the auxiliary estimator over the replications, and the right figure displays the AML estimator. True parameter values are reported as horizontal lines.

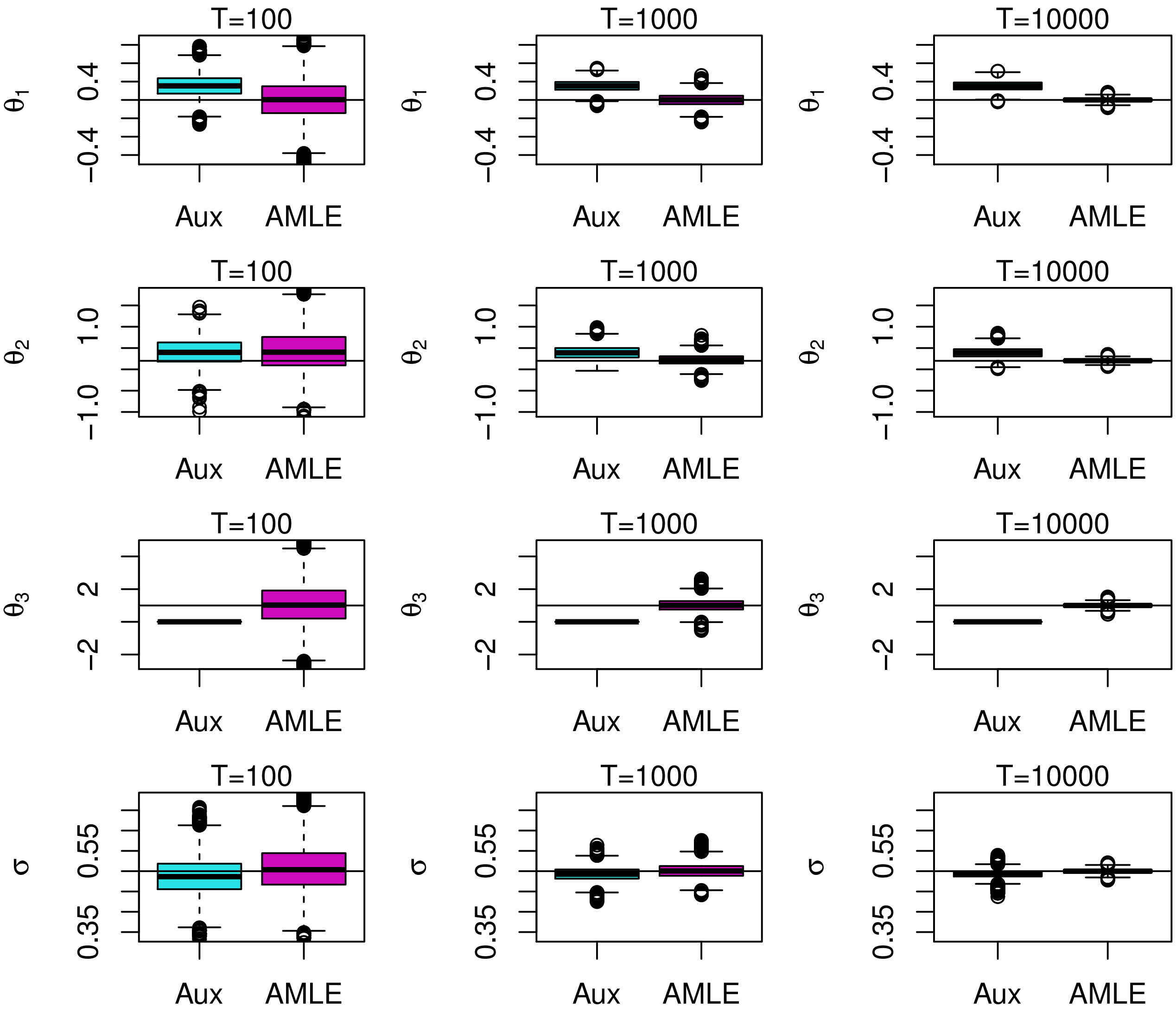

Figure 1: Each boxplot reports the auxiliary (left boxplots) and AML (right boxplots) parameter estimates for the generalized Tobit model with $T = 100$ (left figures), $T = 1,000$ (middle figures) and $T = 10,000$ (right figures) across the Monte Carlo replications. The true parameter values are $\theta_1 = \theta_2 = 0.2$, $\theta_3 = 1$, $\sigma = 0.5$ and are reported as horizontal lines.

The results demonstrate that while the restricted model is easy to estimate, it ultimately provides biased estimators of the resulting parameters for θ_1, θ_2 and σ (as well as θ_3, which is fixed at a value of zero). In contrast, AML delivers point estimators that are well-centred over the true values.

Table 1 compares the AML and auxiliary estimators across the three samples sizes in terms of bias (Bias), mean squared error (MSE), and Monte Carlo coverage (COV).[5] The results demonstrate that AML delivers estimators with relatively small biases, and good (Monte Carlo)

[5] Monte Carlo coverage is calculated as the average number of times, across the Monte Carlo trials, that θ_j^0, i.e., the true value of the j-th parameter, is contained in the univariate confidence interval $\hat{\theta}_j^i \pm \hat{\sigma}_j 1.96$, where $\hat{\sigma}_j$ is the standard deviation for the j-th parameter over the Monte Carlo replications and $\hat{\theta}_j^i$ is the estimator of the j-th parameter in the i-th Monte Carlo trial.

coverage.

		θ_1	θ_2	θ_3	σ
$T = 100$					
Auxiliary	Bias	0.15258	0.21231	-	-0.01281
	MSE	0.03853	0.13049	-	0.00231
	COV	0.76460	0.86360	-	0.94070
AML	Bias	0.00064	0.29606	0.14294	0.00550
	MSE	0.04860	0.58569	2.63985	0.00662
	COV	0.94750	0.95100	0.96290	0.98320
$T = 1,000$					
Auxiliary	Bias	0.15366	0.19205	-	-0.00722
	MSE	0.02707	0.06178	-	0.00034
	COV	0.27070	0.76910	-	0.92980
AML	Bias	-0.00045	0.02416	0.01519	0.00104
	MSE	0.00467	0.01626	0.14967	0.00033
	COV	0.94630	0.94550	0.95160	0.94650
$T = 10,000$					
Auxiliary	Bias	0.15251	0.18972	-	-0.00688
	MSE	0.02560	0.05204	-	0.00015
	COV	0.12030	0.68070	-	0.90290
AML	Bias	0.00034	0.00253	-0.00028	0.00005
	MSE	0.00047	0.00143	0.01474	0.00003
	COV	0.94960	0.94910	0.94890	0.98320

Table 1: Accuracy measures for auxiliary and AML parameter estimates of the generalized Tobit model, across the sample sizes $T = 100$, $T = 1,000$ and $T = 10,000$, and across the 10,000 Monte Carlo replications. The true parameter values are $\theta_1 = \theta_2 = 0.2$, $\theta_3 = 1$, $\sigma = 0.5$.

We now compare the standard errors of AML estimates obtained using the Monte Carlo simulations to their asymptotic counterpart. Concretely, we compare the sample standard deviation (multiplied by $\sqrt{T}$) of the 10,000 simulated AML estimates against the asymptotic standard errors calculated using the formula in Proposition 2.[6]To approximate the expectations in Proposition 2, we use a long simulated series with $T = 10^9$ observations. The asymptotic standard errors are reported in the last row of Table 2.[7] The upper rows of Table 2 provide the Monte

[6] We note that the asymptotic standard errors are infeasible in practice, and in general only their estimated value can be computed. While infeasible in practice, as they are calculated at the true value θ^0, in the context of AML estimation and for a long enough simulated sample size, the only difference between the estimated standard errors and their infeasible counterpart (i.e., the asymptotic standard errors) is the replacement of θ^0 by the estimated value $\hat{\theta}_{H,T}$.

[7] Note that the asymptotic variance formula is infeasible to use directly in empirical examples since calculation of $I^0_{(H)}(\theta^0, \beta^0)$ and $J^0(\theta^0, \beta^0)$ necessitate simulating $\{x_i\}$ and $\{z_i\}$ under their true distributions, which typically would not be known in practice.

Carlo standard errors (multiplied by $\sqrt{T}$) calculated using the 10,000 AML estimates of the parameters θ_1, θ_2, θ_3 and σ across the sample sizes $T = 100$, $1,000$ and $10,000$ and using $H = 10$. Table 2 shows that the asymptotic variance formula in Proposition 2 is accurate in the case of the Generalized Tobit model for sample sizes as small as $T = 1,000$.

Table 2: Finite-sample (multiplied by $\sqrt{T}$) and asymptotic standard errors of AML estimators in the Generalized Tobit model.

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\sigma}$
$\sqrt{T}\times$ MC SE, $T = 100$	2.20461	7.05757	16.18541	0.59598
$\sqrt{T}\times$ MC SE, $T = 1000$	2.16101	3.95946	12.22533	0.57007
$\sqrt{T}\times$ MC SE, $T = 10000$	2.17106	3.77924	12.14220	0.57296
Asympt. SE	2.17448	3.72847	12.23216	0.57327

4.2 Example 4: Markov-Switching Multifractal Model

In this sub-section, we explore the behavior of AML and, when feasible, compare AML and ML. As discussed in Section 2.2.2, the structural parameters in the MSM model are $\theta = (\zeta', \overline{k})'$, where the parameter $\zeta = (m_0, \bar{\gamma}, b, \sigma)'$ govern the behavior of the individual volatility processes, and where $\overline{k}$ denotes the (unknown) number of volatility components. The likelihood of the MSM model, $L_T(\zeta, \overline{k})$, is given in equation (7), and can be optimized so long as small values of $\overline{k}$ are considered. Indeed, for fixed ζ, computation of the likelihood is only feasible for values of $\overline{k}$ that are not too large: a single evaluation of the log-likelihood for a sample of size T requires $O(2^{2\overline{k}}T)$ computations, and ML estimation becomes infeasible if the true value of $\overline{k}$ is large.

However, under the constraint $\overline{k} = 2$, the likelihood $L_T(\zeta, \overline{k})$ requires only $O(2^4 T)$ computations. This suggest the following constrained estimator for the purpose of AML:[8]

$$\hat{\beta}_T = \underset{\beta\in\Theta}{\arg\max}\, L_T(\zeta, \overline{k}), \text{ s.t } \overline{k} = 2. \tag{16}$$

The likelihood $L_T(\zeta, \overline{k})$ is not differentiable in $\overline{k}$, since $\overline{k} \in \{1, 2, \dots, \}$, and so for the $\overline{k}$ component of the AML pseudo-score we use the difference approximation $L_T(\zeta, 3) - L_T(\zeta, 2)$, which yields

$$M_T(\zeta, 2) = \left(\frac{\partial L_T(\zeta, 2)}{\partial \zeta'}, L_T(\zeta, 3) - L_T(\zeta, 2)\right)'. \tag{17}$$

where we note that $\partial L_T(\zeta, 2)/\partial\zeta'$ can be reliably obtained using numerical differentiation.

To implement AML in this example, we consider H i.i.d. simulated samples, from the MSM model. From these simulated samples, the AML estimator is obtained by minimizing, in the Euclidean norm, the difference between the average simulated pseudo-score $\sum_{h=1}^{H} M_T^{(h)}(\theta, \hat{\beta}_T)/H$ and $M_T(\hat{\beta}_T)$.

[8] The more computationally convenient constraint $\overline{k} = 1$ can not be readily used as the parameter b vanishes from the log-likelihood function when $\overline{k} = 1$.

Monte Carlo

We first consider data generated from the MSM model with a relatively small value of $\overline{k}$ so that ML is computationally feasible. This allows us to compare AML and ML, and directly assess the efficiency loss of AML relative to ML. To this end, we generate 1,000 synthetic data sets from the MSM model in Section 2.2.2 with $T = 5,000$ observations, and where the parameter values are set as follows: $m_0 = 1.5$, $\overline{\gamma} = 0.2$, $b = 4$, $\sigma = 0.01$ and $\overline{k} = 4$.

Numerical implementation of AML and ML require optimization over the integer parameter space for $\overline{k}$, while optimization for the ζ components can proceed via standard approaches. For both approaches, optimization over the ζ components is carried out using a quasi-Newton approach, with finite-differences used to estimate the derivatives. For the $\overline{k}$ components, the likelihood is optimized across the grid $\{1, \dots, 7\}$, while AML considers a much larger grid of values.[9]

The ability of AML to consider large values for $\overline{k}$ is possible because the computational cost required to evaluate the AML criterion function *does not* increase with $\overline{k}$, and requires $O(HT)$ computations for any value of $\overline{k}$. In this Monte Carlo exercise, AML is implemented using $H = 100$ pseudo-samples, as the large value of H smooths the criterion function and increases the accuracy of numerical differentiation methods.[10]

Figure 2 displays the results of this Monte Carlo experiment. For each sub-figure, the left plot contains the ML estimator and the right plot contains the associated AML estimator. The true parameter values are reported as horizontal lines. AML provides estimators that are well-centred over the true value of the structural parameters with, as expected, a larger variance than the ML estimator in some cases.

Table 3 compares the bias (Bias), mean squared error (MSE) and Monte Carlo coverage (COV) of the estimators. In addition, for each replication we calculate the efficiency loss of AML with respect to ML via the average relative standard error, denoted by SE(ML)/SE(AML) in Table 3. Using this measure, numbers below unity suggest that, on average, the ML estimator is more efficient than the AML estimator. The results in Table 3 suggest that the two estimators are comparable in terms of bias and MSE for m_0, $\bar{\gamma}$ and b, with ML yielding more accurate estimators for $\overline{k}$ and σ. Analyzing the efficiency of the two estimators, we see that, according to the SE(ML)/SE(AML) measure, AML is nearly as efficient as ML for m_0, $\bar{\gamma}$ and b, but less so for σ and $\overline{k}$. The later is not entirely unexpected as imposing the invalid restriction $\overline{k} = 2$ within the pseudo-score should lead to some efficiency loss (with respect to ML). However, this example also demonstrates that imposing this restriction only leads to a minor loss in accuracy for estimating m_0, $\bar{\gamma}$ and b.

[9] Technically, we implement AML by extending the grid of values over which $\overline{k}$ is optimized to the entire real line. This is done by considering a piecewise linear extension of the pseudo-score for the $\overline{k}$ component, and by taking the closest integer to the resulting optimized value.

[10] An alternative to the finite-differences considered herein would be to use the simulation-based differentiation approach in Frazier et al. (2019).

		m_0	$\overline{\gamma}$	b	σ	$\overline{k}$
ML	Bias	-0.0014244	0.0134517	0.1367587	0.0000088	-0.0120000
	MSE	0.0004834	0.0121796	1.0688123	0.0000003	0.0900000
	COV	0.9380000	0.9520000	0.9560000	0.9490000	0.9130000
AML	Bias	-0.0036913	0.0280103	0.0653309	0.0002228	-0.0878051
	MSE	0.0005691	0.0142423	0.9924541	0.0000009	0.1727888
	COV	0.9510000	0.9430000	0.9440000	0.9310000	0.9150000
	SE(ML)/SE(AML)	0.9309007	0.9442337	1.0308558	0.5860551	0.7377811

Table 3: Accuracy measures for ML and AML parameter estimates of the MSM for $T = 5,000$, and across the 1,000 Monte Carlo replications. The true parameter values are $m_0 = 1.5$, $\overline{\gamma} = 0.4$, $b = 5$, $\sigma = 0.01$ and $\overline{k} = 4$. In ML estimation, $\overline{k}$ only takes values in $\{1, \dots, 7\}$.

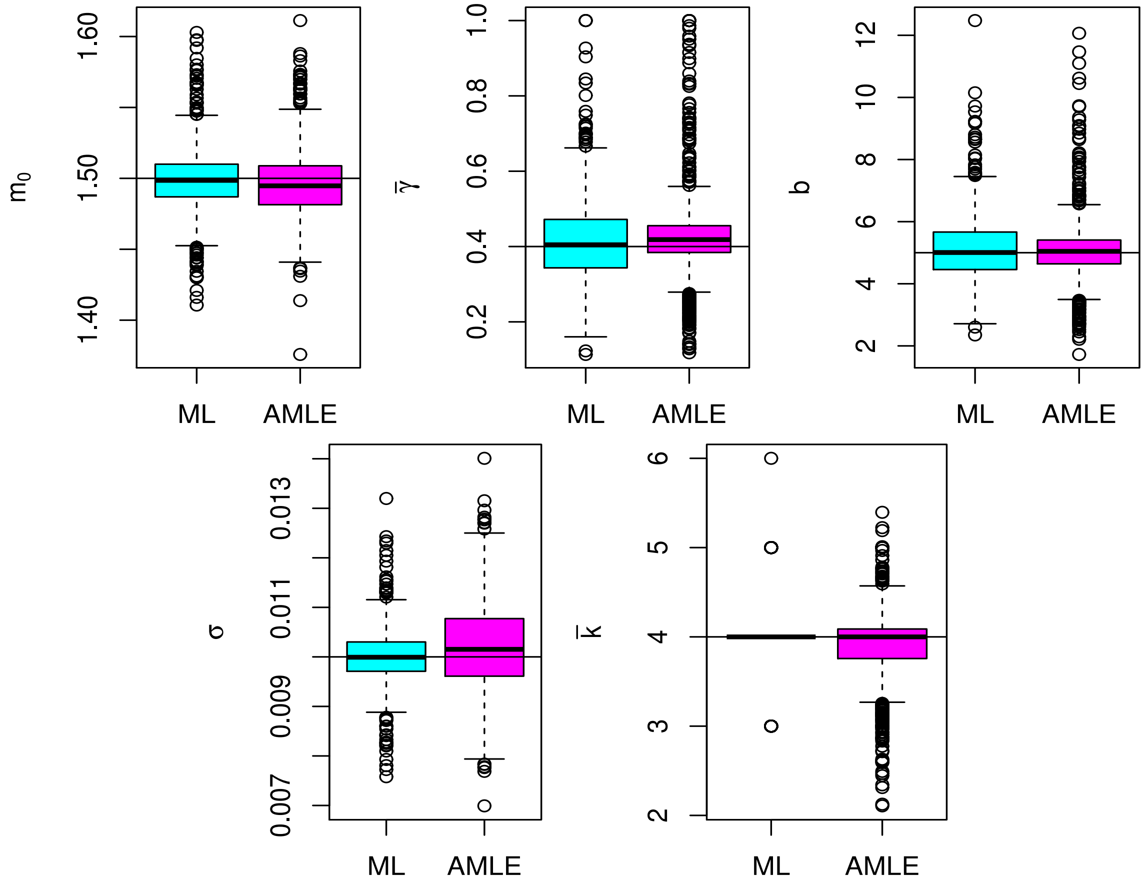

Figure 2: Each boxplot reports the ML (left boxplots) and AML (right boxplots) parameter estimates for the MSM model with sample size $T = 5,000$ across the Monte Carlo replications. The true parameter values are $m_0 = 1.5$, $\overline{\gamma} = 0.4$, $b = 5$, $\sigma = 0.01$ and $\overline{k} = 4$ and are reported as horizontal lines.

While ML has an edge in terms of accuracy, due to computational cost, ML is infeasible if the true value of $\overline{k}$ is large. To illustrate this point, we compare the time, in $\log_{10}$ seconds, required to evaluate the log-likelihood function and the AML criterion function for various values of $\overline{k}$ and for a sample size of $T = 5,000$. Programs were implemented in C and computation was performed on an Intel(R) Xeon(R) CPU E7-4830 v3 @ 2.10GHz. For each $\overline{k} = 6, 7, \ldots, 21$, we evaluate twenty Monte Carlo replications and report the mean computation time for the AML criterion function based on $H = 100$ simulated samples. We repeat the same exercises for the log-likelihood function and for $\overline{k} = 6, 7, \ldots, 14$, with linear extrapolation used for values of $\overline{k} \geq 15$. Figure 3 compares the mean computation times. For $\overline{k}$ small, evaluation of the likelihood is faster than the AML criterion, given the large number of simulated paths used in the AML criterion. However, when $\overline{k}$ becomes even moderately large, AML is clearly superior in terms of computational cost. For values of $\overline{k} > 9$, AML is particularly attractive in terms of computation time. At a value of $\overline{k} = 21$, a single evaluation of the log-likelihood would require 5459.2 days (approximately 15 years), whereas an evaluation of the AML criterion only requires 1.45 seconds.

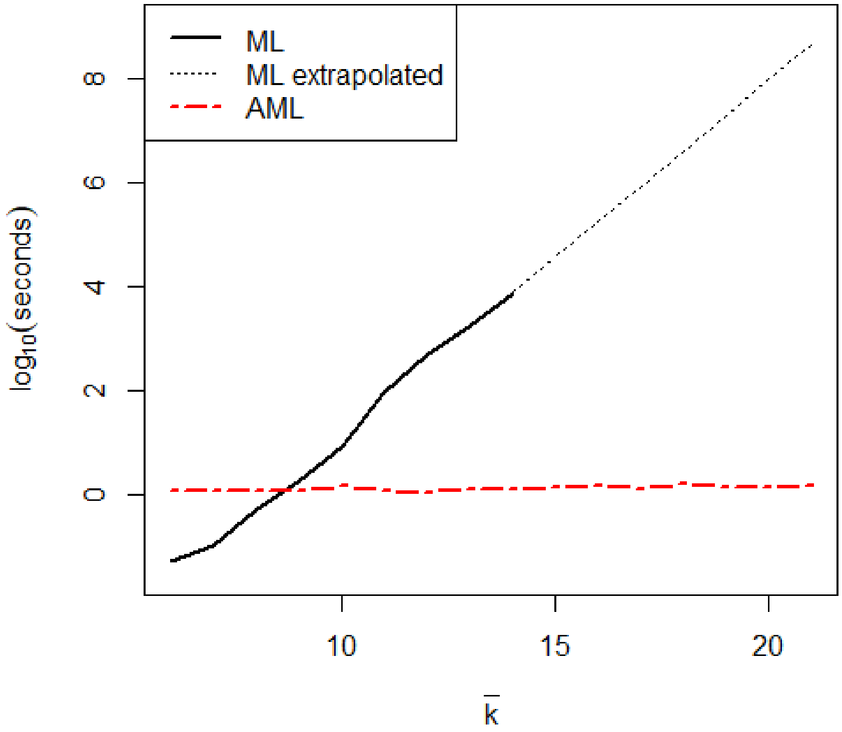

Figure 3: Computation times, in $\log_{10}$ seconds, of the likelihood function (continuous line) and AML criterion function (dash-dotted line) using $H = 100$. The averages presented are taken over twenty data sets simulated from the MSM model with $T = 5,000$, $m_0 = 1.5$, $\overline{\gamma} = 0.2$, $b = 4$, $\sigma = 0.01$ and $\overline{k} = 6, 7, \ldots, 21$. Small dotted line indicates extrapolated computation time for ML estimation for $\overline{k} \geq 15$.

We now assess the performance of AML for a large value of $\overline{k}$. We choose $\overline{k} = 18$ and other parameter values that resulted from the empirical example conducted later (see Table 5 in the following subsection). Figure 4 displays the estimation results over 1,000 Monte Carlo replications from the DGP associated with $T = 23,202$ (as in the empirical dataset in the following subsection), and where the parameter values are $m_0 = 1.2708$, $\overline{\gamma} = 0.1215$, $b = 1.5663$, $\sigma = 0.0149$ and $\overline{k} = 18$. For each sample, we calculate the constrained estimator and AML estimator using $H = 100$ pseudo-samples. For each sub-figure, the left plot contains the constrained auxiliary estimates and the right plot contains the associated AML estimator. The true parameter values are reported with horizontal lines. While the restricted model is easy to estimate, it provides estimators that are significantly biased for all parameters except σ. AML corrects the resulting bias for all structural parameters and delivers estimators that are, on average, centred over the true values. Analyzing the other accuracy measures given in Table 4, we see that AML generally yields estimators with low bias and Monte Carlo coverage close to the nominal level.

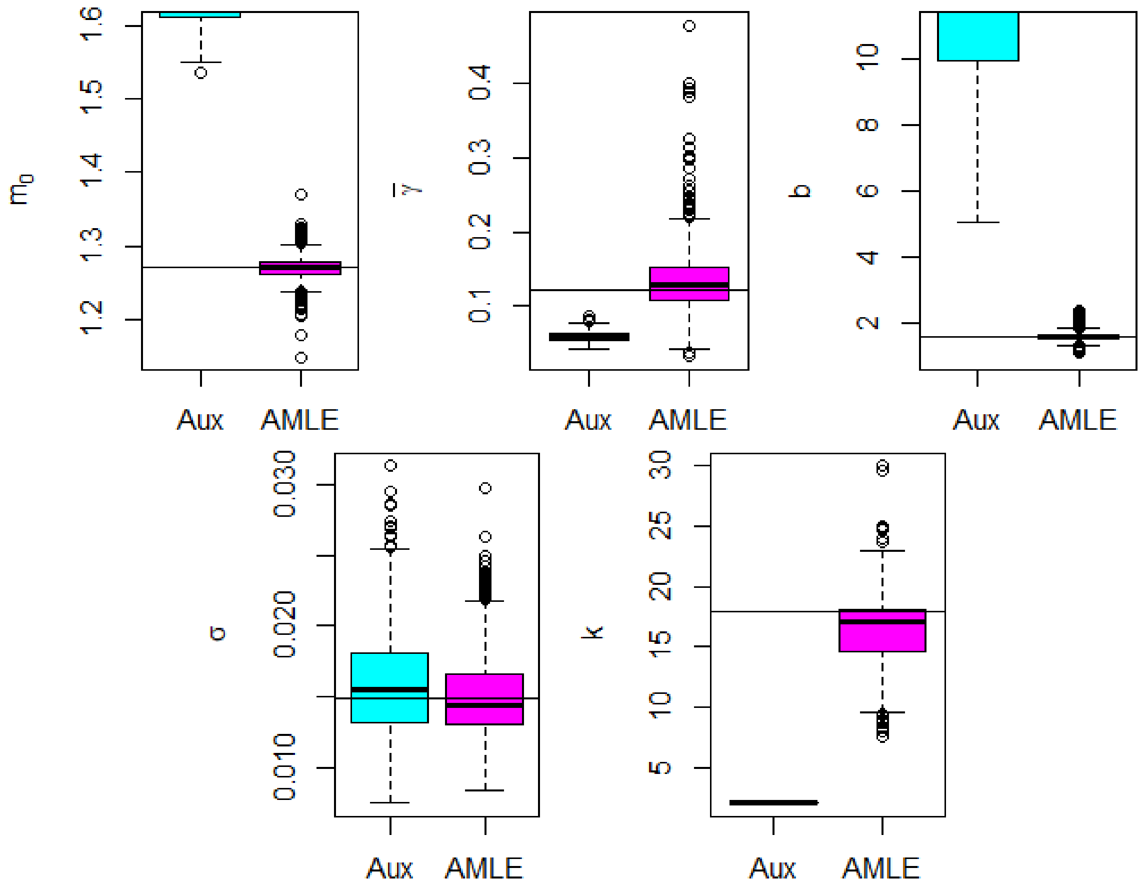

Figure 4: Each boxplot reports the auxiliary (left boxplots) and AML (right boxplots) parameter estimates for the MSM model with sample size $T = 23,202$ across the Monte Carlo replications. The true parameter values are $m_0 = 1.2708$, $\overline{\gamma} = 0.1215$, $b = 1.5663$, $\sigma = 0.0149$ and $\overline{k} = 18$ and reported with horizontal lines.

		m_0	$\overline{\gamma}$	b	σ	$\overline{k}$
Auxiliary	Bias	0.363348	-0.061777	12.002244	0.000943	-
	MSE	0.133257	0.003867	174.209123	0.000014	-
	COV	0.000000	0.000000	0.480000	0.939000	-
AML	Bias	-0.001502	0.012439	0.025719	0.000033	-1.558178
	MSE	0.000303	0.002176	0.022416	0.000009	11.391885
	COV	0.936000	0.955000	0.937000	0.945000	0.897000

Table 4: Accuracy measures for auxiliary and AML estimator parameter estimates of the MSM model with $T = 23,202$, and across the 1,000 Monte Carlo replications. True parameter values are $m_0 = 1.2708$, $\overline{\gamma} = 0.1215$, $b = 1.5663$, $\sigma = 0.0149$ and $\overline{k} = 18$.

Application: S&P500 Returns

We now estimate the Binomial MSM model on demeaned daily S&P500 (simple) returns between January 3, 1928 and May 15, 2020[11]. The sample size is $T = 23,202$. The data are plotted in Figure 5. Using this data set, Table 5 compares the AML estimators with those obtained from maximum likelihood for fixed values of $\overline{k}$ ranging from $\overline{k} = 1$ up to $\overline{k} = 10$. The estimated value of $\overline{k}$ obtained by AML is far larger than the feasible value associated with ML. Moreover, except for m_0, the remaining estimated parameters are also significantly different, with the estimated values of $\bar{\gamma}$ and b being markedly different across the two approaches. The standard errors for ML are calculated using the asymptotic formula, while those for AML are calculated using a parametric bootstrap-based and 1,000 simulated data sets from the assumed DGP.[12]

In order to compare the goodness-of-fit of the eleven models enumerated in Table 5, for each model we provide one-day-ahead forecasts at each in-sample date $t = 1, \dots, T$ using a particle filter of size $N = 10^6$. For a given model, at each date t, the particle filter provides N simulated values from the approximate distribution of $r_t | \{r_1, \dots, r_{t-1}\}$:

$$r_t^{(1)}, \dots, r_t^{(N)} .$$

At each date $t = 1, \dots, T$, we calculate the $\alpha = 1\%$ and $\alpha = 5\%$ value-at-risk forecasts defined by

$$\text{VaR}_{\alpha,t} = -q_\alpha(r_t^{(1)}, \dots, r_t^{(N)}) ,$$

where $q_\alpha(\cdot)$ indicates the α-th sample quantile, and report the failure rate of $\text{VaR}_{\alpha,t}$:

$$p_\alpha = \frac{1}{T} \sum_{t=1}^{T} 1_{r_t < (-\text{VaR}_{\alpha,t})} .$$

The closer p_α is to α, the better the forecasts. The left panel of Table 7 reports p_α for $\alpha = 0.01$ and $\alpha = 0.05$ for each model specification along with asymptotic standard errors in parentheses.

[11] Downloaded from finance.yahoo.com on May 15, 2020.

[12] The reasoning behind using the parametric bootstrap instead of estimated asymptotic standard errors is two-fold: first, the parametric bootstrap can provide high-order asymptotic refinements (Andrews, 2005); second, estimating the asymptotic standard errors for AML requires numerically computing second-order derivatives with respect to the parameters governing the simulated data, which can result in numerical instabilities within the estimated derivatives, and ultimately in poor estimates of asymptotic standard errors (see Frazier et al., 2019 for further discussion of such issue in the context of I-I estimation).

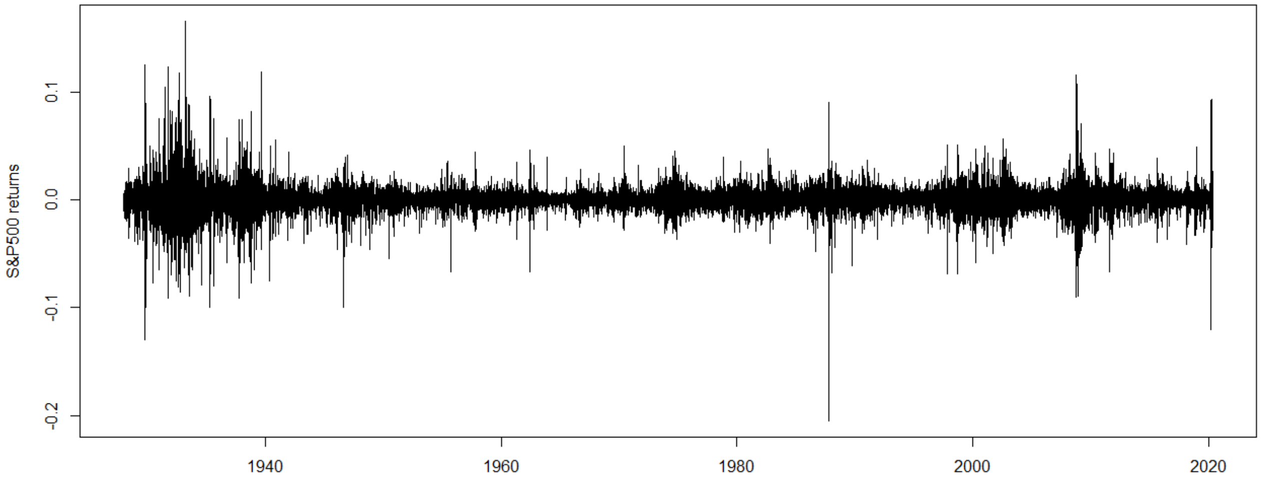

Figure 5: Daily S&P500 returns between January 3, 1928 and May 15, 2020.

AML provides the only model specification for which both failure rates are not significantly different from their nominal levels. In addition, we also assess the accuracy of the $\alpha = 5\%$ expected shortfall forecasts:

$$ES_{\alpha,t} = \sum_{i=1}^{N} r_t^{(i)} 1_{r_t^{(i)} < (-\mathrm{VaR}_\alpha^t)} \Big/ \sum_{i=1}^{N} 1_{r_t^{(i)} < (-\mathrm{VaR}_\alpha^t)} \, .$$

To this end, we collect the empirical returns satisfying $r_t | r_t < -\mathrm{Var}_{0.05,t}^{(\overline{k}=10)}$, under the model with $\overline{k} = 10$, and for each value of $\overline{k}$ in Table 5, we regress these returns on $ES_{\alpha,t}$, calculated under the corresponding value of $\overline{k}$ in Table 5. Regression intercepts, slopes, R^2 values and p-values of the Wald test associated with the joint hypothesis $(\text{intercept}, \text{slope})' = (0, 1)$ are reported in the right panel of Table 7. The $\overline{k} = 18$ specification provides the best expected shortfall forecasts, as measured by the magnitude of the corresponding p-values.

	$\overline{k}$	m_0	$\overline{\gamma}$	b	σ	Log-like.
ML	1	1.8168 (0.0040)	0.0269 (0.0062)	-	0.0164 (0.0002)	75476.07
	2	1.6654 (0.0040)	0.0593 (0.0062)	14.6239 (1.1063)	0.0157 (0.0002)	76409.45
	3	1.5890 (0.0040)	0.0922 (0.0062)	9.0988 (1.1063)	0.0161 (0.0002)	76779.51
	4	1.5199 (0.0052)	0.1149 (0.0861)	5.3760 (0.3414)	0.0151 (0.0003)	76874.11
	5	1.4745 (0.0052)	0.1461 (0.0861)	4.6768 (0.3414)	0.0161 (0.0003)	76940.79
	6	1.4517 (0.0052)	0.9441 (0.0861)	6.5357 (0.3414)	0.0152 (0.0003)	76978.39
	7	1.4291 (0.0055)	0.9999 (0.0929)	5.5954 (0.2772)	0.0132 (0.0002)	76994.82
	8	1.3882 (0.0060)	1.0000 (0.1093)	3.9099 (0.1854)	0.0128 (0.0003)	77001.80
	9	1.3568 (0.0062)	1.0000 (0.1224)	3.1657 (0.1427)	0.0137 (0.0005)	77006.94
	10	1.3383 (0.0067)	1.0000 (0.1305)	2.8090 (0.1328)	0.0130 (0.0006)	77009.94
AML	18 (2.9955)	1.2708 (0.0173)	0.1215 (0.0450)	1.5663 (0.1476)	0.0149 (0.0030)	-

Table 5: The table reports the ML estimator (ML) and AML estimator (AML) of the demeaned empirical S&P500 returns (left panel). Asymptotic standard errors for the ML estimator are reported in parentheses below each value. The AML standard errors are obtained using a parametric bootstrap based on 1,000 simulated samples (of length $T = 23,202$) generated from the MSM model at the AML point estimates.

Table 6: Goodness-of-fit comparisons of AML and ML with various $\overline{k}$.

		VaR failure rates		$\text{ES}_{0.05}$ regressions			
	$\overline{k}$	$p_{0.05}$	$p_{0.01}$	Intercept	Slope	R^2	Wald
ML	1	0.0427 (0.0013)	0.0081 (0.0006)	0.0007 (0.0008)	0.9112 (0.0277)	0.4771	$3 \cdot 10^{-19}$
	2	0.0463 (0.0014)	0.0082 (0.0006)	0.0024 (0.0007)	1.0322 (0.0244)	0.6015	$5 \cdot 10^{-7}$
	3	0.0463 (0.0014)	0.0082 (0.0006)	0.0009 (0.0006)	0.9947 (0.0220)	0.6331	0.0013
	4	0.0486 (0.0014)	0.0082 (0.0006)	0.0005 (0.0006)	0.9975 (0.0216)	0.6420	0.1256
	5	0.0479 (0.0014)	0.0085 (0.0006)	−0.0003 (0.0006)	0.9622 (0.0209)	0.6420	0.0151
	6	0.0461 (0.0014)	0.0075 (0.0006)	0.0004 (0.0006)	0.9666 (0.0222)	0.6149	$7 \cdot 10^{-5}$
	7	0.0477 (0.0014)	0.0078 (0.0006)	0.0006 (0.0006)	0.9873 (0.0228)	0.6127	0.0115
	8	0.0489 (0.0014)	0.0080 (0.0006)	0.0008 (0.0006)	1.0071 (0.0228)	0.6215	0.0741
	9	0.0486 (0.0014)	0.0081 (0.0006)	0.0005 (0.0006)	0.9944 (0.0224)	0.6236	0.0725
	10	0.0488 (0.0014)	0.0082 (0.0006)	0.0006 (0.0006)	1.0054 (0.0225)	0.6271	0.1981
AML	18	0.0522 (0.0015)	0.0106 (0.0007)	−0.0005 (0.0006)	1.0010 (0.0217)	0.6412	0.2022

Table 7: The table reports accuracies of the 1% and 5% value-at-risk (left panel) and 5% expected shortfall forecasts (right panel) using a particle filter with 10^6 particles. In the left panel, failure rates of the 1% and 5% value-at-risk are reported with asymptotic standard errors in parentheses. In the right panel, for each $\overline{k}$, the empirical returns satisfying $\{r_t | r_t < -\text{VaR}_{0.05,t}^{(\overline{k}=10)}\}$ are regressed on $\{\text{ES}_{0.05,t}^{\overline{k}} | r_t < -\text{VaR}_{0.05,t}^{(\overline{k})}\}$, where $\text{VaR}_{0.05,t}^{(\overline{k})}$ corresponds to the 5% value-at-risk at date t forecasted with $\overline{k}$ and $\text{ES}_{0.05,t}^{(\overline{k})}$ corresponds to the 5% expected shortfall at date t forecasted with $\overline{k}$. For each regression, the intercepts and slopes are reported with standard errors in parentheses along with the R^2 values and the p-values of the Wald test H_0 : $(\text{intercept}, \text{slope}) = (0, 1)$.

5 Conclusion

In this paper, we provide an alternative to indirect inference (hereafter, I-I) estimation that simultaneously allows us to circumvent the intractability of maximum likelihood estimation (as with standard I-I), but which, in contrast to naive I-I, respects the goal of obtaining asymptotically efficient inference in the context of a fully parametric model. Although close in spirit to I-I, the AML method developed in this paper does *not* belong to the realm of I-I for two reasons: First, the asymptotic distribution of the AML estimator only depends on the probability limit of the estimated auxiliary parameters and not on its asymptotic distribution. Second, while the AML estimator is obtained by matching two sample moments, one computed on observed data, and one computed on simulated data, both sample moments depend on the observed data through the value of the preliminary estimator of the auxiliary parameters. Interestingly, the sampling uncertainty carried by this preliminary estimator has no impact on the asymptotic distribution of the AML estimator because it is erased through the matching procedure.

The message of our paper is threefold. First, we demonstrate that the idea of matching proxies of the score for the structural model seems productive to reach near efficiency for inference on the structural parameters. We show theoretically that, at least for exponential models or transformation of them, the efficiency loss should be manageable since it is mainly due to the effect of a misspecification bias created by our simplification of the structural model.

Second, there are many non-linear time series models, which are popular in financial econometrics and dynamic/nonlinear microeconometrics, where a natural simplification of the structural model yields a convenient proxy for the score of the structural model. Since the misspecification bias created by this simplification is only due to imposing some possible false equality constraints, or to numerical approximations for certain elements of the gradient vector, one may reasonably hope that the resulting efficiency loss is minimal. While our general results (and theoretical examples) suggest that this finding is valid in many examples, including dynamic discrete choice and stochastic volatility models, we provide numerical evidence in three specific examples: generalized Tobit, Markov-switching multifractal models and stable distributions. The numerical results largely confirm our intuitions. Our method can alleviate the computational cost of maximum likelihood associated with complex models, at the cost of a limited loss in efficiency. Moreover, we confirm that even in finite-samples, the Wald confidence intervals associated to AML estimators display excellent coverage, since, thanks to matching the misspecification bias, the preliminary estimators have no impact on the central tendency of the AML estimator.

A third and even more general message is that the matching principle put forward by I-I estimation can be extended to situations where the two empirical moments to match, one based on observed data, one based on simulated data may both depend on the observed data through a convenient summary of them. While we have used this idea to aim for (nearly) efficient inference, Gospodinov, et al., (2017) employ a similar approach to hedge against misspecification bias due to the use of a misspecified simulator. Even though they have not derived the asymptotic distribution theory in their case, the two methods are essentially similar and could be nested within a general asymptotic theory where both the moments to match and the simulator depend on observed data.

References

[1] Andrews, Donald WK. “Higher-order improvements of the parametric bootstrap for Markov processes.” Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg (2005): 171-215.

[2] Amemiya, Takeshi. Advanced econometrics. Harvard university press, 1985.

[3] Bansal, Ravi, and Amir Yaron. “Risks for the long run: A potential resolution of asset pricing puzzles.” The Journal of Finance 59, no. 4 (2004): 1481-1509.

[4] Behrens, S. and Melissinos, A.C., Univ. of Rochester Preprint UR-776 (1981).

[5] Calvet, Laurent E., and Veronika Czellar. “Through the looking glass: Indirect inference via simple equilibria.” Journal of Econometrics 185, no. 2 (2015): 343-358.

[6] Calvet, Laurent, and Adlai Fisher. “Forecasting multifractal volatility.” Journal of Econometrics 105, no. 1 (2001): 27-58.

[7] Calvet, Laurent E., and Adlai J. Fisher. “How to forecast long-run volatility: Regime switching and the estimation of multifractal processes.” Journal of Financial Econometrics 2, no. 1 (2004): 49-83.

[8] Calvet, Laurent E., and Adlai Fisher. Multifractal volatility: theory, forecasting, and pricing. Academic Press (2008).

[9] Calzolari, Giorgio, Gabriele Fiorentini, and Enrique Sentana. “Constrained indirect estimation.” The Review of Economic Studies 71, no. 4 (2004): 945-973.

[10] Chambers, John M., Colin L. Mallows, and B. W. Stuck. “A method for simulating stable random variables.” Journal of the American Statistical Association 71, no. 354 (1976): 340-344.

[11] Chen, Fei, Francis X. Diebold, and Frank Schorfheide. “A Markov-switching multifractal inter-trade duration model, with application to US equities.” Journal of Econometrics 177, no. 2 (2013): 320-342.

[12] Dridi, Ramdan, and Eric Renault. “Semi-parametric indirect inference.” LSE STICERD Research Paper No. EM392 (2000).

[13] Dridi, Ramdan, Alain Guay, and Eric Renault.“Indirect inference and calibration of dynamic stochastic general equilibrium models.” Journal of Econometrics 136, no. 2 (2007): 397-430.

[14] Dudley, Leonard, and Claude Montmarquette. “A model of the supply of bilateral foreign aid.” The American Economic Review 66, no. 1 (1976): 132-142.

[15] Franses, Philip Hans, Marco Van Der Leij, and Richard Paap. “A simple test for GARCH against a stochastic volatility model.” Journal of Financial Econometrics 6, no. 3 (2008): 291-306.

[16] Frazier, David T., Tatsushi Oka, and Dan Zhu. “Indirect inference with a non-smooth criterion function.” Journal of Econometrics 212, no. 2 (2019): 623-645.

[17] Frazier, David T., and Eric Renault. “Indirect inference with(out) constraints.” Quantitative Economics, 11 (2020): 113-159.

[18] Gallant, A. Ronald. and George Tauchen. “Which moments to match.” Econometric Theory 12, (1996): 657-681.

[19] Gospodinov, Nikolay, Ivana Komunjer, and Serena Ng. “Simulated minimum distance estimation of dynamic models with errors-in-variables.” Journal of Econometrics 200, no. 2 (2017): 181-193.

[20] Gourieroux, Christian, Alain Monfort, and Alain Trognon. “Pseudo maximum likelihood methods: Theory.” Econometrica: Journal of the Econometric Society (1984): 681-700.

[21] Gourieroux, Christian, Alain Monfort, and Alain Trognon. "A general approach to serial correlation." Econometric Theory 1, no. 3 (1985): 315-340.

[22] Gourieroux, Christian, Alain Monfort, Eric Renault, and Alain Trognon. "Generalised residuals." Journal of econometrics 34, no. 1-2 (1987): 5-32.

[23] Gourieroux, Christian and Alain Monfort. Simulation-based Econometric Methods, OUP, (1996).

[24] Gourieroux, Christian, Alain Monfort and Eric Renault. "Indirect inference." Journal of Applied Econometrics 85, (1993): S85–S118.

[25] Gourieroux, Christian, Alain Monfort, Eric Renault, and Alain Trognon. "Generalised residuals." Journal of Econometrics 34, no. 1 (1987): 5-32.

[26] Hansen, Lars Peter. "Large sample properties of generalized method of moments estimators." Econometrica: Journal of the Econometric Society (1982): 1029-1054.

[27] Koutrouvelis, Ioannis A. "An iterative procedure for the estimation of the parameters of stable laws: An iterative procedure for the estimation." Communications in Statistics-Simulation and Computation 10, no. 1 (1981): 17-28.

[28] Louis, Thomas A. "Finding the observed information matrix when using the EM algorithm." Journal of the Royal Statistical Society: Series B (Methodological) 44, no. 2 (1982): 226-233.

[29] McCulloch, J. Huston. "Simple consistent estimators of stable distribution parameters." Communications in Statistics-Simulation and Computation 15, no. 4 (1986): 1109-1136.

[30] Meddahi, Nour, and Eric Renault. "Temporal aggregation of volatility models." Journal of Econometrics 119, no. 2 (2004): 355-379.

[31] Pinkse, Joris, and Margaret E. Slade. "Contracting in space: An application of spatial statistics to discrete-choice models." Journal of Econometrics 85, no. 1 (1998): 125-154.

[32] Poirier, Dale J., and Paul A. Ruud. "Probit with dependent observations." The Review of Economic Studies 55, no. 4 (1988): 593-614.

[33] Robinson, Peter M. "On the asymptotic properties of estimators of models containing limited dependent variables." Econometrica, (1982): 27-41.

[34] Smith, Anthony A. "Estimating nonlinear time series models using simulated vector autoregressions." Journal of Applied Econometrics 8, no. S1 (1993): S63-S84.

[35] van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press, 1998.

A Proofs of Main Results

A.1 Proof of Lemma 1

From the differentiability in **Assumption A2**, with a standard abuse of notation, a Taylor expansion gives:

$$\begin{aligned}\sqrt{T}M_T\left(\hat{\beta}_T\right) &= \sqrt{T}M_T\left(\beta^0\right) - K^0\left[\tilde{\beta}_T\right]\sqrt{T}\left[\hat{\beta}_T-\beta^0\right]\\ \frac{\sqrt{T}}{H}\sum_{h=1}^{H}M_T^{(h)}\left(\theta,\hat{\beta}_T\right) &= \frac{1}{H}\sum_{h=1}^{H}\sqrt{T}M_T^{(h)}\left(\theta,\beta^0\right) - \left\{\frac{1}{H}\sum_{h=1}^{H}K^0\left(\tilde{\beta}_T^{(h)}(\theta)\right)\right\}\sqrt{T}\left[\hat{\beta}_T-\beta^0\right]\end{aligned}$$

where $\tilde{\beta}_T$ and $\tilde{\beta}_T^{(h)}(\theta),\ h=1,...,H$, are all in the interval $\left[\beta^0,\hat{\beta}_T\right]$. Hence,

$$\sqrt{T}M_T\left(\beta^0\right)-\frac{1}{H}\sum_{h=1}^{H}\sqrt{T}M_T^{(h)}\left(\theta,\beta^0\right)=\left\{K^0\left[\tilde{\beta}_T\right]-\frac{1}{H}\sum_{h=1}^{H}K^0\left(\tilde{\beta}_T^{(h)}(\theta)\right)\right\}\sqrt{T}\left[\hat{\beta}_T-\beta^0\right].$$

From **Assumptions A1 and A2, and Assumption B1(i)**, in particular, $\sqrt{T}(\hat{\beta}_T-\beta^0)=O_P(1)$, the AML estimator satisfies

$$\sqrt{T}M_T\left(\beta^0\right)-\frac{1}{H}\sum_{h=1}^{H}\sqrt{T}M_T^{(h)}\left(\hat{\theta}_{T,H},\beta^0\right)=o_P(1).$$

A.2 Proof of Proposition 2

Under **Assumption B2**, an additional Taylor expansion of $M_T^{(h)}\left(\hat{\theta}_{T,H},\beta^0\right)$ gives

$$\sqrt{T}M_T\left(\beta^0\right)-\frac{1}{H}\sum_{h=1}^{H}\sqrt{T}M_T^{(h)}\left(\theta^0,\beta^0\right)+o_P(1)=-\left(\frac{1}{H}\sum_{h=1}^{H}J^0\left(\tilde{\theta}_T^{(h)},\beta^0\right)\right)\sqrt{T}\left(\hat{\theta}_{T,H}-\theta^0\right),$$

where $\tilde{\theta}_T^{(h)}, h=1,...,H$ are all in the interval $\left[\theta^0,\hat{\theta}_{T,H}\right]$. Invertibility of $J^0(\theta^0,\beta^0)$, **Assumption B2**, then allows us to write

$$\sqrt{T}\left(\hat{\theta}_{T,H}-\theta^0\right)=\left[J^0\left(\theta^0,\beta^0\right)\right]^{-1}\left\{\sqrt{T}M_T\left(\beta^0\right)-\frac{1}{H}\sum_{h=1}^{H}\sqrt{T}M_T^{(h)}\left(\theta^0,\beta^0\right)\right\}+o_P(1).$$

From the hypothesis of the proposition,

$$\left\{\sqrt{T}M_T\left(\beta^0\right)-\frac{1}{H}\sum_{h=1}^{H}\sqrt{T}M_T^{(h)}\left(\theta^0,\beta^0\right)\right\}\to_d \aleph\left[0,I_{(H)}^0\left(\theta^0,\beta^0\right)\right],$$

and the result follows. □

A.3 Proof of Proposition 3

By virtue of the first part of **Proposition 2**, we only need to prove that the asymptotic variance $\Omega_{(H)}$ of the AML estimator $\hat{\theta}_{T,H}$ coincides with the Cramer-Rao efficiency bound when $H \to \infty$ and in the specific case where $\hat{\beta}_T = \beta^0 = \theta^0$. When $H \to \infty$, this estimator, denoted $\breve{\theta}_T$, can be seen as the solution in θ of the system of equations:

$$M_T\left(\theta^0\right) = E_\theta[M_T\left(\theta^0\right) \mid \{x_t\}_{t=1}^T].$$

If we define

$$g_T\left(\beta, \theta\right) = M_T\left(\beta\right) - E_\theta[M_T\left(\beta\right) \big| \{x_t\}_{t=1}^T],$$

we have, by definition,

$$\begin{aligned} 0 &= \sqrt{T} g_T\left(\theta^0, \breve{\theta}_T\right) \\ &= \sqrt{T} g_T\left(\theta^0, \theta^0\right) + \frac{\partial g_T\left(\theta^0, \theta^0\right)}{\partial \theta'} \sqrt{T}\left(\breve{\theta}_T - \theta^0\right) + o_P(1). \end{aligned}$$

Recall the definition of $M_T\left(\theta^0\right)$,

$$\begin{aligned} M_T\left(\theta^0\right) &= \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \log\left(l\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\right\}\right)}{\partial \theta} \qquad (18) \\ &= \frac{1}{T}\sum_{t=1}^{T} S\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\right\}, \end{aligned}$$

and note that, by virtue of (18),

$$\sqrt{T} g_T\left(\theta^0, \theta^0\right) = \sqrt{T} M_T\left(\theta^0\right) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial \log\left(l\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\right\}\right)}{\partial \theta}$$

converges in distribution to a $\aleph\left(0, I^0\right)$ random variable, where $I^0 = I^0\left(\theta^0, \theta^0\right)$ is the Fisher information matrix.

Moreover,

$$\operatorname*{plim}_{T\to\infty} \frac{\partial g_T\left(\theta^0, \theta^0\right)}{\partial \theta'} = \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\partial}{\partial \theta'} E_\theta \left\{ \frac{\partial \log\left(l\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\right\}\right)}{\partial \theta} \right\}_{\theta=\theta^0},$$

and we have

$$\begin{aligned} & \frac{\partial}{\partial \theta'} E_\theta \left\{ \frac{\partial \log\left(l\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\right\}\right)}{\partial \theta} \right\} \\ = & \frac{\partial}{\partial \theta'} \int \frac{\partial \log\left(l\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\right\}\right)}{\partial \theta} l\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta\right\} d\nu \left(y_t \middle| \{y_\tau\}_{\tau=1}^{t-1}, x_t\right), \end{aligned}$$

where ν denotes some dominating measure. Thus,

$$\begin{aligned} & \frac{\partial}{\partial \theta'} E_\theta \left\{ \frac{\partial \log\left(l\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\right\}\right)}{\partial \theta} \right\} \\ & = \int S\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\right\} S\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta\right\}' l\{y_t \left| \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta\right\} d\nu \left(y_t \middle| \{y_\tau\}_{\tau=1}^{t-1}, x_t\right). \end{aligned}$$

Therefore,

$$\operatorname*{plim}_{T\to\infty} \frac{\partial g_T\left(\theta^0,\theta^0\right)}{\partial\theta'} = E\left[S\{y_t \mid \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\} S\{y_t \mid \{y_\tau\}_{\tau=1}^{t-1}, x_t, z_0, \theta^0\}' \mid \{x_\tau\}_{\tau=1}^{t}\right]$$

is the Fisher information matrix I^0. Consequently,

$$\sqrt{T}\left(\breve{\theta}_T - \theta^0\right) = -\left(I^0\right)^{-1}\sqrt{T}g_T\left(\theta^0,\theta^0\right) + o_P(1) \longrightarrow_d \aleph\left(0, \left(I^0\right)^{-1}\right).$$

B Details for Examples in Section 3

In this section, we give the details regarding the identification and efficiency of AML in the exponential models example considered in Section 3. In addition, we also extend this example to consider latent exponential models.

B.1 Example: Exponential Models

For the sake of exposition, we assume that conditionally on $\{x_t\}_{t=1}^T$, the variables $y_t, t = 1, ..., T$ are independent and the conditional distribution of y_t only depends on the exogenous variable x_t with the same index. This distribution has a density $l\{y_t \,|x_t;\theta\}$ that is assumed to be exponential:

$$l\{y_t \,|x_t;\theta\} = \exp\left[c\left(x_t,\theta\right) + h(y_t,x_t) + a'(x_t,\theta)T(y_t)\right]$$

where $c(.,.)$ and $h(.,.)$ are given numerical functions and $a(x_t,\theta)$ and $T(y_t)$ are r-dimensional random vectors. Note that the extension to dynamic models in which conditioning values would also include some lagged values of the process y_t would be easy to devise. From:

$$\frac{\partial\log\left[l\{y_t \,|x_t;\theta\}\right]}{\partial\theta} = \frac{\partial c\left(x_t,\theta\right)}{\partial\theta} + \frac{\partial a'\left(x_t,\theta\right)}{\partial\theta}T(y_t),$$

we deduce, since the conditional score vector has by definition a zero conditional expectation, that:

$$\frac{\partial L_T\left(\theta\right)}{\partial\theta} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial a'\left(x_t,\theta\right)}{\partial\theta}\left\{T(y_t) - E_\theta[T(y_t)\,|x_t]\right\}.$$

Following Theorem 1 in Gourieroux et al. (1987),

$$\begin{aligned} E_\theta[T(y_t)\,|x_t] &= m\left(x_t,\theta\right), Var_\theta[T(y_t)\,|x_t] = \Omega\left(x_t,\theta\right) \\ &\Longrightarrow \frac{\partial a'\left(x_t,\theta\right)}{\partial\theta} = \frac{\partial m'\left(x_t,\theta\right)}{\partial\theta}\Omega^{-1}\left(x_t,\theta\right). \end{aligned}$$

Therefore, the maximum likelihood estimator $\hat{\theta}_T$ is defined as solution of:

$$\frac{\partial L_T\left(\theta\right)}{\partial\theta} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial m'\left(x_t,\theta\right)}{\partial\theta}\Omega^{-1}\left(x_t,\theta\right)\left\{T(y_t) - m\left(x_t,\theta\right)\right\} = 0. \tag{19}$$

We actually generalize the remark of van der Vaart (1998), Section 4.2, noting that "the maximum likelihood estimators are moment estimators" based on the (conditional) expectation of the sufficient statistic $T(y)$. The first-order conditions (19) show that maximum likelihood is the GMM estimator with optimal instruments for the conditional moment restrictions:

$$E_\theta[T(y_t) - m\left(x_t, \theta\right)|x_t] = 0.$$

Note that we implicitly maintain the assumptions for standard asymptotic theory of efficient GMM (Hansen, 1982): for all $\theta \in \Theta$, the conditional variance $\Omega\left(x_t, \theta\right)$ of the moment conditions is non-singular and the Jacobian matrix $E[\partial m'\left(x_t, \theta\right)/\partial\theta\,|x_t]$ is full row rank.

The identification condition for consistency of maximum likelihood is then that:

$$E\left\{\frac{\partial m'\left(x_t, \theta\right)}{\partial\theta}\Omega^{-1}\left(x_t, \theta\right)\left\{T(y_t) - m\left(x_t, \theta\right)\right\}\right\} = 0 \Longrightarrow \theta = \theta^0.$$

In terms of GMM, it means that optimal instruments are assumed to identify the true unknown value θ^0 of the parameter vector θ, by contrast with cases put forward by Dominguez and Lobato (2004). By the Law of Iterated Expectations, this can be rewritten:

$$E\left\{\frac{\partial m'\left(x_t, \theta\right)}{\partial\theta}\Omega^{-1}\left(x_t, \theta\right)\left\{m\left(x_t, \theta^0\right) - m\left(x_t, \theta\right)\right\}\right\} = 0 \Longrightarrow \theta = \theta^0$$

or equivalently (by symmetry):

$$E\left\{\frac{\partial m'\left(x_t, \theta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t, \theta^0\right)\left\{m\left(x_t, \theta\right) - m\left(x_t, \theta^0\right)\right\}\right\} = 0 \Longrightarrow \theta = \theta^0. \tag{20}$$

By extension of (19), we have:

$$M_T^{(h)}\left(\theta, \beta\right) = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial m'\left(x_t, \beta\right)}{\partial\theta}\Omega^{-1}\left(x_t, \beta\right)\left\{T\left[\tilde{y}_t^{(h)}\left(\theta\right)\right] - m\left(x_t, \beta\right)\right\} \tag{21}$$

so that:

$$M\left(\theta, \beta^0\right) = E\left\{\frac{\partial m'\left(x_t, \beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t, \beta^0\right)\left\{T\left[\tilde{y}_t^{(h)}\left(\theta\right)\right] - m\left(x_t, \beta^0\right)\right\}\right\}.$$

Hence:

$$M\left(\theta, \beta^0\right) - M\left(\theta^0, \beta^0\right) = E\left\{\frac{\partial m'\left(x_t, \beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t, \beta^0\right)\left\{T\left[\tilde{y}_t^{(h)}\left(\theta\right)\right] - T\left[\tilde{y}_t^{(h)}\left(\theta^0\right)\right]\right\}\right\}.$$

By the Law of Iterated Expectations:

$$M\left(\theta, \beta^0\right) - M\left(\theta^0, \beta^0\right) = E\left\{\frac{\partial m'\left(x_t, \beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t, \beta^0\right)\left\{m\left(x_t, \theta\right) - m\left(x_t, \theta^0\right)\right\}\right\},$$

so that the identification Assumption B1 amounts to:

$$E\left\{\frac{\partial m'\left(x_t, \beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t, \beta^0\right)\left\{m\left(x_t, \theta\right) - m\left(x_t, \theta^0\right)\right\}\right\} \Longrightarrow \theta = \theta^0. \tag{22}$$

When $\beta^0 = \theta^0$, we are back to the well-specified example and (22) is obviously identical to the identification condition (20) for consistency of maximum likelihood. Moreover, the identification assumption (22) for consistency of the AML estimator $\breve{\theta}_{T,H}(\beta^0)$ is clearly likely implied by the standard condition (20) for consistency of maximum likelihood, at least in two particular cases.

Result 1 ***Case 1:*** *The model is a linear regression. For some known multivariate function $\kappa(x_t)$ of x_t,*

$$m\left(x_t,\theta\right)=\kappa\left(x_t\right)'\theta.$$

The identification condition (22) is then equivalent to

$$E\left[\kappa(x_t)\Omega^{-1}\left(x_t,\beta^0\right)\kappa(x_t)'\right]\left(\theta-\theta^0\right)=0\Longrightarrow\theta=\theta^0.$$

Moreover, if $E\left[\kappa(x_t)\Omega^{-1}\left(x_t,\beta^0\right)\kappa(x_t)'\right]$ is full rank at $\beta^0=\theta^0$, it is full rank for any $\beta^0\in\Theta_0$.

Case 2: *The model is unconditional. In this case, a necessary identification condition is given by*

$$E_\theta[T(y_1)]=E_{\theta^0}[T(y_1)]\iff\theta=\theta^0.$$

In this case, the AML identification condition (22) can be equivalently stated as

$$\frac{\partial m'\left(\beta^0\right)}{\partial\theta}\Omega^{-1}\left(\beta^0\right)\left\{E_\theta\left[T(y_1)\right]-E_{\theta^0}\left[T(y_1)\right]\right\}\Longrightarrow\theta=\theta^0.$$

The matrix $\partial m\left(\beta^0\right)'/\partial\theta$ is full row rank, irrespective of the value of β^0, so that if $\Omega(\beta^0)$ is non-singular for any $\beta^0\in\Theta_0$, the above identification condition is implied by the identification condition $E_\theta[T(y_1)]=E_{\theta^0}[T(y_1)]\iff\theta=\theta^0$.

Proof: The model is a linear regression model w.r.t. some known multivariate function $\kappa(x_t)$ of x_t:

$$m\left(x_t,\theta\right)=\kappa'\left(x_t\right)\theta.$$

In this case, the identification condition (22) is akin to:

$$E\left[\kappa(x_t)\Omega^{-1}\left(x_t,\beta^0\right)\kappa'(x_t)\right]\left(\theta-\theta^0\right)=0\Longrightarrow\theta=\theta^0.$$

Obviously, when the matrix:

$$E\left[\kappa(x_t)\Omega^{-1}\left(x_t,\beta^0\right)\kappa'(x_t)\right]$$

is positive definite for $\beta^0=\theta^0$, it is positive definite for any possible value of β^0. □

Result 2 *The model is unconditional. In this case, a necessary identification condition is given by*

$$E_\theta[T(y_1)]=E_{\theta^0}[T(y_1)]\iff\theta=\theta^0.$$

In this case, the AML identification condition (22) can be equivalently stated as

$$\frac{\partial m'\left(\beta^0\right)}{\partial\theta}\Omega^{-1}\left(\beta^0\right)\left\{E_\theta\left[T(y_1)\right]-E_{\theta^0}\left[T(y_1)\right]\right\}\Longrightarrow\theta=\theta^0.$$

The matrix $\partial m\left(\beta^0\right)'/\partial\theta$ is full row rank, irrespective of the value of β^0, so that if $\Omega(\beta^0)$ is non-singular for any $\beta^0\in\Theta_0$, the above identification condition is implied by the identification condition $E_\theta[T(y_1)]=E_{\theta^0}[T(y_1)]\iff\theta=\theta^0$.

Proof: The model is not conditional. In this case, a necessary condition for identification condition is:

$$E_\theta \left\{T(y_1)\right\} = E_{\theta^0} \left\{T(y_1)\right\} \Longleftrightarrow \theta = \theta^0. \tag{23}$$

This is basically the case considered by van der Vaart (1998) when noting that "the maximum likelihood estimators are moment estimators" based on the expectation of the sufficient statistic $T(y)$. This identification condition should be maintained when picking p linear independent equations out of possibly overidentified equations (23). More precisely, the identification condition for AML, written as:

$$\frac{\partial m'\left(\beta^0\right)}{\partial \theta}\Omega^{-1}\left(\beta^0\right)\left\{E_\theta\left\{T(y_1)\right\} - E_{\theta^0}\left\{T(y_1)\right\}\right\} \Longrightarrow \theta = \theta^0$$

should generically be implied by (23), since, irrespective of the value of β^0, the matrix $\partial m'\left(\beta^0\right)/\partial\theta$ is full row rank.

More generally, one may expect that the identification condition (22), when fulfilled for $\beta^0 = \theta^0$, should be more often than not fulfilled for any value of β^0. □

To compare the efficiency of AML and ML, recall that, from the first-order conditions (15), the simulated pseudo-score can be stated as

$$M_T^{(h)}\left(\theta,\beta\right) = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial m'\left(x_t,\beta\right)}{\partial\theta}\Omega^{-1}\left(x_t,\beta\right)\left\{T\left[\tilde{y}_t^{(h)}\left(\theta\right)\right] - m\left(x_t,\beta\right)\right\}.$$

Recalling that the AML estimator only depends on the pseudo-true value of the estimate $\hat{\beta}_T$, the AML estimator is asymptotically equivalent to an unfeasible AML (UAML) estimator that solves

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial m'\left(x_t,\beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t,\beta^0\right)\left\{T(y_t) - m\left(x_t,\theta\right)\right\} = 0,$$

where we recall that $E_\theta[T(y_t)|x_t] = m(x_t,\theta) = \lim_{H\to\infty}\sum_{h=1}^{H} T[\tilde{y}_t^{(h)}(\theta)]/H$.

Comparing the above equation with (15), the only reason why AML may be less efficient than ML is that the evaluation of the "optimal instruments" is carried out at a pseudo-true value of the structural parameters (i.e., $\beta^0 \neq \theta^0$). It is worth revisiting the implications of this in the two cases discussed above.

Case 1: The model is a linear regression. For some known multivariate function $\kappa(x_t)$ of x_t,

$$m\left(x_t,\theta\right) = \kappa\left(x_t\right)'\theta.$$

In this case, the equation defining the UAML estimator is then

$$\frac{1}{T}\sum_{t=1}^{T}\kappa\left(x_t\right)\Omega^{-1}\left(x_t,\beta^0\right)\left\{T(y_t) - \kappa\left(x_t\right)'\theta\right\} = 0.$$

From the above, we see that the presence of conditional heteroskedasticity or cross-correlation, of a parametric nature, can result in a loss of efficiency for UAML. However, if $\Omega\left(x_t,\beta^0\right) = \sigma^2\mathrm{Id}$, UAML is asymptotically equivalent to maximum likelihood.

Case 2: The model is unconditional. The equation defining the UAML estimator is then given by

$$\frac{\partial m'\left(\beta^0\right)}{\partial\theta}\Omega^{-1}\left(\beta^0\right)\frac{1}{T}\sum_{t=1}^{T}\left\{T(y_t)-m\left(\theta\right)\right\}=0.$$

In this case, the only possible loss of efficiency will occur if the moment conditions that identify θ are overidentified, i.e., when $r=\dim(T)\geq p$, so that the selection matrix $\frac{\partial m'\left(\beta^0\right)}{\partial\theta}\Omega^{-1}\left(\beta^0\right)$ is optimal only at $\beta^0=\theta^0$. An efficiency loss will then occur if, when evaluated at $\beta^0\neq\theta^0$, the vector space spanned by the rows of the selection matrix do not coincide with the space spanned by the rows when $\beta^0=\theta^0$.

B.2 Example: Latent Exponential Model

We now extend the exponential model example to incorporate a sequence of latent variables $\{y_t^*\}_{t=1}^{T}$, such that, conditionally on $\{x_t\}_{t=1}^{T}$, the variables y_t^* are independent, for all $t=1,\dots,T$, and the conditional distribution of y_t^* only depends on the exogenous variable x_t with the same index. This distribution has a density $l\{y_t^*\,|x_t;\theta\}$, with respect to the dominating measure $\nu(dy_t^*)$, that is assumed to be exponential:

$$l\{y_t^*\,|x_t;\theta\}=\exp\left[c\left(x_t,\theta\right)+h(y_t^*,x_t)+a'(x_t,\theta)T(y_t^*)\right].$$

Let g be a known vector function that defines the observed endogenous variable y_t as:

$$y_t=g\left(y_t^*,x_t\right).$$

Then, conditionally on $\{x_t\}_{t=1}^{T}$, the variables $y_t,t=1,...,T$ are independent and the conditional distribution of y_t only depends on the exogenous variables x_t with the same index. This conditional distribution has a density $l\{y_t\,|x_t;\theta\}$, with respect to the measure $\nu^g(dy)$, which is the transformation of the original measure $\nu(dy_t^*)$ by g, and where we recall that $\nu(dy_t^*)$ was the dominating measure used to define the latent density $l\{y_t^*\,|x_t;\theta\}$. The observable log-likelihood can then be stated as

$$L_T(\theta)=\frac{1}{T}\sum_{t=1}^{T}\log\left[l\{y_t\,|x_t;\theta\}\right].$$

In general, the observable density is not of an exponential form, see Gourieroux et al. (1987) for the particular case where $y_t=g\left(y_t^*\right)$ and for examples of Probit, bivariate Probit, Tobit, generalized Tobit, disequilibrium and Gompit models. As already mentioned in Section 2.3, Gourieroux et al. (1987), extending a result of Louis (1982), give a method to compute the observable score as a conditional expectation of the latent score

$$\frac{\partial L_T\left(\theta\right)}{\partial\theta}=\frac{1}{T}\sum_{t=1}^{T}E_\theta\left[\frac{\partial\log\left[l\{y_t^*|x_t;\theta\}\right]}{\partial\theta}\bigg|y_t,x_t\right].$$

Then, by applying (19) we get

$$\frac{\partial L_T\left(\theta\right)}{\partial\theta}=\frac{1}{T}\sum_{t=1}^{T}\frac{\partial m'\left(x_t,\theta\right)}{\partial\theta}\Omega^{-1}\left(x_t,\theta\right)\left\{E_\theta[T(y_t^*)\,|y_t,x_t]-m\left(x_t,\theta\right)\right\}. \tag{24}$$

As exemplified by Gourieroux et al. (1987) for many limited dependent variable models, we can define and compute a generalized error as:

$$\begin{aligned} u\left(y_t, x_t, \theta\right) &= \tilde{T}\left(y_t, x_t, \theta\right) - m\left(x_t, \theta\right) \\ \tilde{T}\left(y_t, x_t, \theta\right) &= E_\theta[T(y_t^*)\left|y_t, x_t\right]. \end{aligned}$$

Then, the maximum likelihood estimator $\hat{\theta}_T$ is defined as solution of

$$\frac{\partial L_T\left(\theta\right)}{\partial\theta} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial m'\left(x_t, \theta\right)}{\partial\theta}\Omega^{-1}\left(x_t, \theta\right)u\left(y_t, x_t, \theta\right) = 0. \tag{25}$$

Hence, the identification condition for consistency of maximum likelihood can be written:

$$E\left[\frac{\partial m'\left(x_t, \theta\right)}{\partial\theta}\Omega^{-1}\left(x_t, \theta\right)u\left(y_t, x_t, \theta\right)\right] = 0 \Longleftrightarrow \theta = \theta^0. \tag{26}$$

We also note that MLE is not any more a moment estimator with optimal instruments (confirming that the model is not exponential any more) since:

$$Var[u\left(y_t, x_t, \theta^0\right)|x_t] = Var\left[E_{\theta^0}[T(y_t^*)\left|y_t, x_t\right]|x_t]\right] \neq \Omega\left(x_t, \theta^0\right) = Var[T(y_t^*)\left|x_t\right].$$

More generally, by extension of (25) we have:

$$M_T^{(h)}\left(\theta, \beta\right) = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial m'\left(x_t, \beta\right)}{\partial\theta}\Omega^{-1}\left(x_t, \beta\right)u\left[\tilde{y}_t^{(h)}\left(\theta\right), x_t, \beta\right].$$

Hence,

$$M\left(\theta, \beta^0\right) = E\left\{\frac{\partial m'\left(x_t, \beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t, \beta^0\right)u\left[\tilde{y}_t^{(h)}\left(\theta\right), x_t, \beta^0\right]\right\}.$$

so that

$$\begin{aligned} &M\left(\theta, \beta^0\right) - M\left(\theta^0, \beta^0\right) \\ = \; &E\left\{\frac{\partial m'\left(x_t, \beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t, \beta^0\right)\left[u\left[\tilde{y}_t^{(h)}\left(\theta\right), x_t, \beta^0\right] - u\left[\tilde{y}_t^{(h)}\left(\theta^0\right), x_t, \beta^0\right]\right]\right\}. \end{aligned}$$

When $\beta^0 = \theta^0$, we are back to the well-specified example and we note that by definition:

$$\begin{aligned} E\left\{u\left[\tilde{y}_t^{(h)}\left(\theta^0\right), x_t, \theta^0\right]|x_t\right\} &= 0 \Longrightarrow \forall h \\ E\left\{h(x_t)u\left[\tilde{y}_t^{(h)}\left(\theta^0\right), x_t, \theta^0\right]\right\} &= 0 \Longrightarrow \\ M\left(\theta, \beta^0\right) - M\left(\theta^0, \beta^0\right) &= E\left\{\frac{\partial m'\left(x_t, \theta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t, \theta^0\right)u\left[\tilde{y}_t^{(h)}\left(\theta\right), x_t, \theta^0\right]\right\} = 0. \end{aligned}$$

so that the identification condition

$$M\left(\theta, \beta^0\right) - M\left(\theta^0, \beta^0\right) \Longleftrightarrow \theta = \theta^0,$$

can be written

$$E\left\{\frac{\partial m'\left(x_t,\theta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t,\theta^0\right)u\left[\tilde{y}_t^{(h)}\left(\theta\right),x_t,\theta^0\right]\right\}=0\Longleftrightarrow\theta=\theta^0. \tag{27}$$

By commuting the roles of θ and θ^0, this is clearly tantamount to the identification condition (26) for maximum likelihood. In the general case, the identification condition B1(β^0) for UAML can be written:

$$E\left\{\frac{\partial m'\left(x_t,\beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t,\beta^0\right)\left[u\left[\tilde{y}_t^{(h)}\left(\theta\right),x_t,\beta^0\right]-u\left[\tilde{y}_t^{(h)}\left(\theta^0\right),x_t,\beta^0\right]\right]\right\}=0\Longleftrightarrow\theta=\theta^0.$$

Note that by the Law of Iterated Expectations, this can be written:

$$E\left\{\frac{\partial m'\left(x_t,\beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t,\beta^0\right)\left[\tilde{m}(x_t,\theta,\beta^0)-\tilde{m}(x_t,\theta^0,\beta^0)\right]\right\}=0\Longleftrightarrow\theta=\theta^0,$$

where

$$\tilde{m}(x_t,\theta,\beta^0)=E\left[u\left(\tilde{y}_t^{(h)}\left(\theta\right),x_t,\beta^0\right)|x_t\right].$$

By comparison with (27), we see that while both generalized errors $u\left[\tilde{y}_t^{(h)}\left(\theta\right),x_t,\beta^0\right]$ and $u\left[\tilde{y}_t^{(h)}\left(\theta^0\right),x_t,\beta^0\right]$ will in general have a non-zero conditional expectation given x_t (when $\beta^0\notin\{\theta,\theta^0\}$), identification means that when $\theta\neq\theta^0$, their difference cannot be orthogonal to the p specific functions of x_t that define the rows of the selection matrix:

$$\frac{\partial m'\left(x_t,\beta^0\right)}{\partial\theta}\Omega^{-1}\left(x_t,\beta^0\right).$$

This condition is similar to the condition (22) of identification for UAML in the exponential model example, except that, due to the transformation $y_t=g\left(y_t^*,x_t\right)$, the conditional expectation given x_t along simulated paths still depend on β^0. In the particular case of a latent model defined by a univariate linear and homoskedastic regression equation:

$$m\left(x_t,\theta\right)=x_t'\theta,\Omega\left(x_t,\theta\right)=\sigma^2,$$

the identification condition in **Assumption B1** for UAML becomes:

$$E\left\{x_t\left[\tilde{m}(x_t,\theta,\beta^0)-\tilde{m}(x_t,\theta^0,\beta^0)\right]\right\}=0\Longleftrightarrow\theta=\theta^0.$$

For instance, in the case of a Probit model ($\sigma^2=1$):

$$E\left\{x_t\frac{\varphi(x_t'\beta^0)}{\Phi(x_t'\beta^0)\left[1-\Phi(x_t'\beta^0)\right]}\left[\Phi(x_t'\theta)-\Phi(x_t'\theta^0)\right]\right\}=0\Longleftrightarrow\theta=\theta^0,$$

which we can compare to the standard identification condition for a Probit model

$$E\left\{x_t\frac{\varphi(x_t'\theta)}{\Phi(x_t'\theta)\left[1-\Phi(x_t'\theta)\right]}\left[\Phi(x_t'\theta)-\Phi(x_t'\theta^0)\right]\right\}=0\Longleftrightarrow\theta=\theta^0.$$

These conditions appear to be quite reasonable.

C Additional Example: Stable Distribution

Consider i.i.d. observations $y_1, \ldots, y_T$ generated from a stable distribution with stability parameter $a \in (0, 2]$, skewness parameter $b \in [-1, 1]$, scale parameter $c > 0$ and location parameter $\mu \in \mathbb{R}$. The structural parameter vector is given by:

$$\theta = (a, b, c, \mu)' \tag{28}$$

The practical problem for maximum likelihood inference in this context does not come from a non-linear state space where the likelihood function would involve integrals over the state variables. However, it is known that the log-likelihood function $L_T(\theta)$ is not available in general, except for some specific values of the parameters a and b. As such, maximum likelihood inference can only be implemented by the time-consuming task of numerical inverting the characteristic function, which is known in closed-form, to obtain the resulting (numerical approximation to) the stable density.

However, for $a = 1$ and $b = 0$, the stable distribution coincides with the Cauchy distribution which has a closed-form log-likelihood function $L_T(1, 0, c, \mu)$. Moreover, the stable model also allows to simulate sample paths, for instance with the method of Chambers, Mallows and Stuck (1976). This will pave the way again for an AML strategy.

Consider i.i.d. observations $y_1, \ldots, y_T$ generated from a stable distribution with stability parameter $a \in (0, 2]$, skewness parameter $b \in [-1, 1]$, scale parameter $c > 0$ and location parameter $\mu \in \mathbb{R}$. The structural parameter vector is given by

$$\theta = (a, b, \zeta')', \zeta = (c, \mu)'.$$

We consider this model under the false equality constraint:

$$(a, b)' = (1, 0)'$$

corresponding to a Cauchy distribution with location μ and scale c, which gives the log-likelihood:

$$L_T(1, 0, \zeta) = -\log[\pi c] - \frac{1}{T}\sum_{t=1}^{T} \log\left[1 + \left(\frac{y_t - \mu}{c}\right)^2\right]$$

We can define the pseudo-score vector as:

$$\Delta_\theta L_T(1, 0, \zeta) = \left(\frac{\partial L_T(1, 0, \zeta)}{\partial \zeta'}, L_T(2, 0, \zeta) - L_T(1, 0, \zeta), \tilde{L}_T(1, 1, \zeta) - L_T(1, 0, \zeta)\right)'.$$

Note that, the finite difference $[L_T(2, 0, \zeta) - L_T(1, 0, \zeta)]$ is a convenient approximation of the partial derivative $\partial L_T(1, 0, \zeta)/\partial a$ since the log-likelihood function $L_T(2, 0, \zeta)$ is computed as the likelihood for i.i.d. draws in a Normal distribution with mean μ and variance $2c^2$. Second, the finite difference $[L_T(1, 1, \zeta) - L_T(1, 0, \zeta)]$ is a convenient approximation of the partial derivative $\partial L_T(1, 0, \zeta)/\partial b$ since the log-likelihood function $L_T(1, 1, \zeta)$ could be computed as the likelihood for i.i.d. draws in a Landau distribution with location parameter μ and scale parameter c

$$L_T(1, 1, \zeta) = \sum_{t=1}^{T} \log(f(y_t)), \text{ where } f(y) = \frac{1}{\pi c}\int_0^\infty e^{-x} \cos\left[x\left(\frac{y - \mu}{e}\right) + \frac{2x}{\pi}\log\left(\frac{x}{c}\right)\right] dx.$$

To speed up the computation, we use the following approximation to $f(y)$ given by Behrens and Melissinos (1981).[13]

$$f(y) \approxeq \frac{1}{\sqrt{2\pi c}} \exp\left\{-(y-\mu)/(2c) - \exp\left[-\left(\{y-\mu\}/c\right)\right]/2\right\}.$$

C.1 Monte Carlo

We now compare the behavior of AML using the above pseudo-score, and $H = 10$ simulations, against two alternative approaches: one based on sample quantiles, due to McCullough (1986), and one based on an auxiliary regression model, due to Koutrouvelis (1981). To this end, we generate 1,000 synthetic datasets from the alpha stable models, each with $T = 10,000$ observations, and under $\theta = (1.8, -0.1, 1, 0)'$.

We display the resulting estimators across the replications in Figure 6.[14] Analyzing the results, we see that the three procedures perform similarly for σ, but display different behavior for α, β, δ, although all estimators seem quite reliable, and are well-centred over the true values.

Table 8 records the Monte Carlo bias (Bias), root mean squared error (RMSE), and Monte Carlo coverage (COV), based on individual 95% Wald interval, across the replications. The results demonstrate that the methods all yield accurate estimators of the corresponding true values. However, we note that the simpler methods do outperform AML in terms of bias and RMSE, but display worse coverage than AML in almost all cases.

Table 8: Summary accuracy measures for stable example. Acronyms are as described in Figure 6, while Aux refers to the auxiliary estimator estimated under the restriction $(a, b) = (1, 0)$. To aid readability of the table, the reported bias has been multiplied by 1000, and reported RMSE has been multiplied by 100.

	a					b			
	AML	Aux	Kout	McC		AML	Aux	Kout	McC
Mean	1.8190	1.0000	1.7994	1.8031	Mean	-0.0948	0.0000	-0.0966	-0.1039
Bias	19.0315	-800.0000	-0.6072	3.1120	Bias	5.1846	100.0000	3.4381	-3.8520
RMSE	9.6607	80.0000	1.4561	2.9785	RMSE	13.6959	10.0000	6.5433	7.9542
COV	0.9600	0.0000	0.9410	0.9540	COV	0.9650	0.0000	0.9540	0.9440

	c					μ			
	AML	Aux	Kout	McC		AML	Aux	Kout	McC
Mean	0.1002	0.0881	0.1000	0.1000	Mean	0.0007	0.0025	0.0032	0.0031
Bias	0.1636	-11.8524	0.0016	-0.0074	Bias	0.6945	2.4720	3.2351	3.1469
RMSE	0.1488	1.1884	0.0978	0.1251	RMSE	0.6313	0.2986	0.3737	0.3781
COV	0.9480	0.0000	0.9480	0.9510	COV	0.9810	0.6650	0.6030	0.6640

[13] Similar results were obtained whether or not the approximation was employed. Given the similarity of the results, and the drastic speed difference, the approximation approach is more reasonable to apply in practice.

[14] We remark that while ML estimation is feasible in the α-stable model for small numbers of observations, given the sample size considered herein, obtaining the MLE proved to be computationally infeasible.

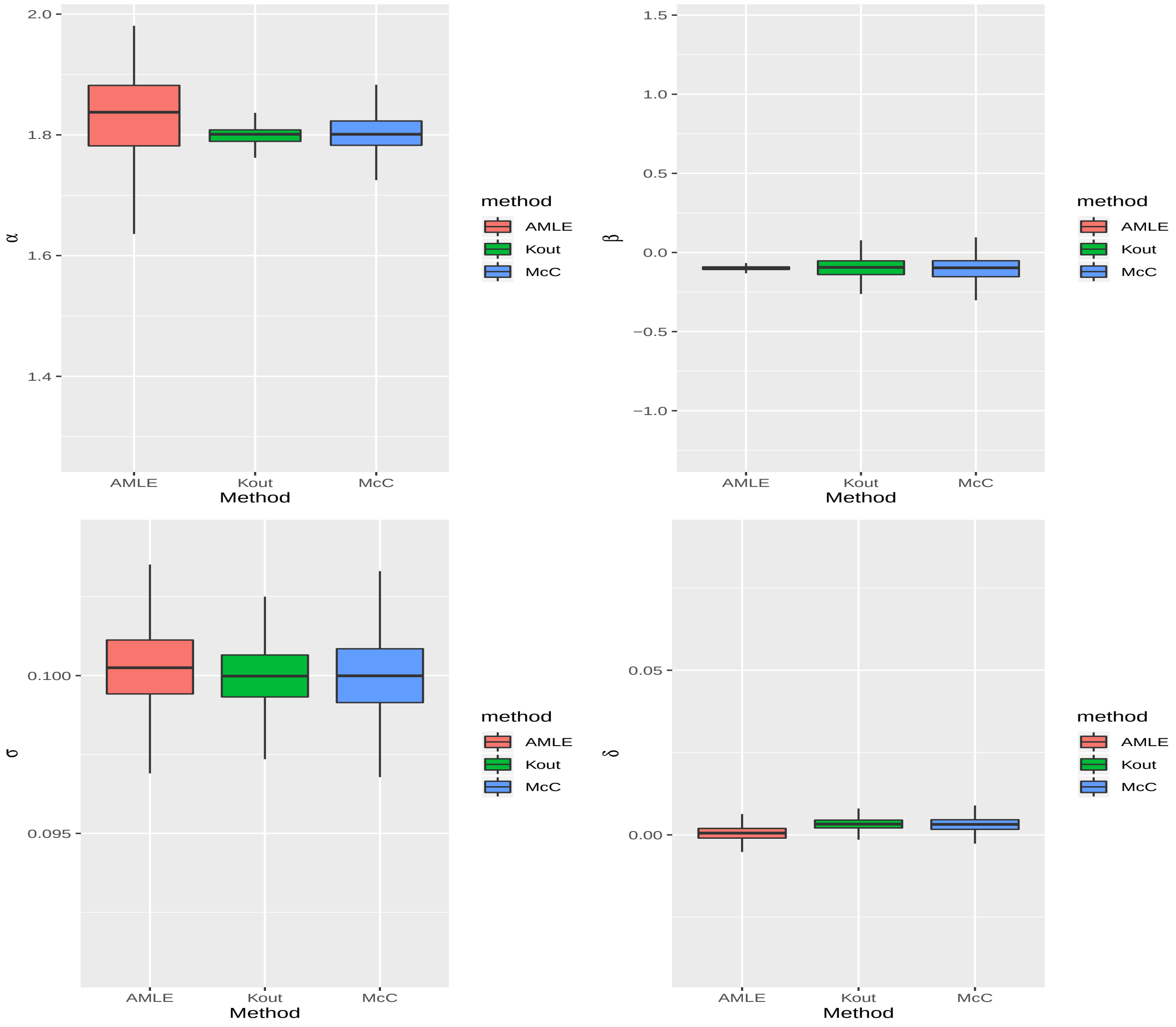

Figure 6: Boxplots of estimators across 1000 Monte Carlo replications from the stable distribution. The true values used to generate the data are $\theta = (a, b, c, \mu) = (1.8, -0.1, 0.1, 0)'$. AML- approximate maximum likelihood estimator, Kout- Koutrouvelis (1981) regression approach, McC- McCullough (1986) quantile approach.

D Additional Example: Autoregressive Discrete Choice Models

We observe the sample $\{y_t, x_t\}_{t=1}^{T}$ generated from

$$y_t \;=\; \begin{cases} 1 & \text{if } y_t^* > 0 \\ 0 & \text{if } y_t^* \leq 0 \end{cases}, \quad y_t^* = x_t'\theta_1 + u_t, \quad u_t = \theta_2 u_{t-1} + \nu_t,$$

where x_t is a vector of explanatory variables, ν_t is a Gaussian white noise and the $AR(1)$ process $(u_t)_{t \leq T}$ is stationary ($-1 < \theta_2 < 1$), $\theta = (\theta_1', \theta_2)'$. Following the standard normalization practice for a Probit error term, we set $\nu_t \sim \aleph(0, 1)$. In what follows, panel data can easily be accommodated at the cost of more involved notations, and so we omit this extension for simplicity.

Unlike the standard Probit model, the autoregressive nature of u_t means that the data density can only be stated as the T-dimensional integral: Let $A_t = [0, +\infty)$ if $y_t = 1$ and $A_t = (-\infty, 0)$ if $y_t = 0$,

$$l\left\{(y_t)_{t\leq T}\left|(x_t)_{t\leq T}\right.;\theta\right\} = \int_{A_1}\cdots\int_{A_T} l^*\left\{(y_t^*)_{t\leq T}\left|(x_t)_{t\leq T}\right., z_0;\theta\right\} dy_1^*\cdots dy_T^*,$$

$$l^*\left\{(y_t^*)_{t\leq T}\left|(x_t)_{t\leq T}\right., z_0;\theta\right\} = (2\pi)^{-T/2}R(\theta_2)^{-1/2}\exp\left(-\frac{1}{2R(\theta_2)}u_1^2(\theta_1)\right)\prod_{t=2}^{T}\exp\left(-\frac{\left[u_t(\theta_1)-\theta_2 u_{t-1}(\theta_1)\right]^2}{2}\right)$$

where $R(\theta_2) = 1/(1-\theta_2^2)$ and $u_t(\theta_1) = y_t^* - x_t'\theta_1$. However, note that if one were to impose the constraint $\theta_2 = 0$ in $l^*\left\{(y_t^*)_{t\leq T}\left|(x_t)_{t\leq T}\right., z_0;\theta\right\}$, the integral that defines this density can be factorized into a product of T integrals, ultimately yields the usual Probit likelihood function. As such, a convenient parametric sub-model is given by

$$l\left\{(y_t)_{t\leq T}\left|(x_t)_{t\leq T}\right.;\theta\right\};\theta\in\Theta_0 = \left\{\theta\in\Theta, \theta = (\theta_1', 0)'\right\}$$

A similar finding to the above can also be applied, albeit with different notations, to spatially correlated Probit models, instead of the autoregressive Probit model.

The dynamic Probit model is a striking example of the fact that, while the complete likelihood function $l\left\{(y_t)_{t\leq T}\left|(x_t)_{t\leq T}\right.;\theta\right\}$ can only be stated as a T-dimensional integral, the sub-model defined by $\theta_2 = 0$ is much simpler, since it coincides with the usual Probit likelihood. Not only does the (possibly false) equality constraint $\theta_2 = 0$ lead to a closed-form likelihood, but the results of Gourieroux, et al. (1985) demonstrate that the partial derivatives of the likelihood function are also available in closed-form.

Under the restriction $\theta_2 = 0$, for

$$\tilde{u}_t(\theta_1, 0) = \frac{\varphi\left(x_t'\theta_1\right)}{\Phi\left(x_t'\theta_1\right)\left[1-\Phi\left(x_t'\theta_1\right)\right]}\left[y_t - \Phi\left(x_t'\theta_1\right)\right],$$

where φ (resp. Φ) denotes the probability density function (resp. the cumulative distribution function) of the standard normal, the computations in Gourieroux et al. (1985) yield

$$\frac{\partial L_T(\theta_1, 0)}{\partial\theta_1} = \frac{1}{T}\sum_{t=1}^{T} x_t\tilde{u}_t(\theta_1, 0), \quad \left.\frac{\partial L_T(\theta_1, \theta_2)}{\partial\theta_2}\right|_{\theta_2=0} = \frac{1}{T}\sum_{t=2}^{T}\tilde{u}_{t-1}(\theta_1, 0)\tilde{u}_t(\theta_1, 0)$$

The term $\tilde{u}_t(\theta_1, 0)$ is the generalized residual under the restriction $\theta_2 = 0$. Gourieroux et al. (1987) show that $\tilde{u}_t(\theta_1, 0)$ can be interpreted as the conditional expectation of the error term u_t given y_t when the true value of θ is $(\theta_1', 0)'$.