An MCMC Approach to Classical Estimation with Overidentifying Restrictions

Luis E. Quintero [*]

Johns Hopkins University

June 1, 2021

Abstract

I extend the Laplace estimators approach proposed by Chernozhukov and Hong (2003) for an overidentified system by decomposing the m moments into the identifying space and the overidentifying space, and using both to construct a transformed criterion function for a new just-identified system. Parameters and test statistics are estimated simultaneously using the entire equation domain, not only the global minimum. As in Chernozhukov and Hong (2003), Markov-Chain Monte Carlo (MCMC) avoids the curse of dimensionality in this method. It is also applicable to non-smooth criterion functions. Incorporating the ORs in the objective function amounts to using economic theory as criterion for estimate selection when facing multiple local solutions. The proposed estimators outperform counterparts in simulation of an asset-pricing model in Hall and Horowitz (1996).

KEYWORDS: Overidentifying restrictions, Markov Chain Monte Carlo, Generalized Method of Moments, Laplace, Multiple Minima. JEL:C10,C11,C13,C15.

[*] Carey Business School, Johns Hopkins University. Phone: (412)-304-4389. leq@jhu.edu.

I am grateful to Fallaw Sowell for valuable help, comments, and suggestions. Helpful comments were provided by Benjamin Holcblat, Jason Imbrogno, Christian Julliard, Jay Kadane, Surya Tapas Tokdar, and seminar participants at Carnegie Mellon University.

1 Introduction

Bayesian techniques are often used under a classical estimation framework to solve computational challenges of extremum estimation in econometrics. Chernozhukov and Hong (2003) uses Laplace-type estimators (LTEs) to provide consistent parameter estimation. LTEs, transform the Bayesian approach by using general criterion functions instead of parametric likelihood functions. The LTEs avoid the curse of dimensionality by using simulation techniques, such as Markov-Chain Monte Carlo (MCMC), that approximate integral transformations of the criterion function, thus estimating parameters without the need of an exponentially growing number of function evaluations. Chernozhukov and Hong (2003) develop asymptotic theory for these estimators and apply them to econometric models. These estimators perform well when compared to standard methods, while circumventing the curse of dimensionality.

This paper proposes an extension of these estimators that takes into account the overidentifying restrictions in a generalized method of moments (GMM) framework. In GMM estimation, k linear combinations of the moment conditions are set equal to zero, where k is the number of parameters (Hansen, 1982). For an over-identified system with $m > k$ moment conditions, the remaining $m - k$ linearly independent combinations implied by the model are the overidentifying restrictions (OR). These OR will also be close to zero in expectation when the model is well specified. Sowell (2009) discusses problems that may arise if parameter values are estimated independently from these overidentifying restrictions, emphasizing the role the overidentifying restrictions may play in selecting between different local minima of the GMM objective function. I incorporate OR estimation in the computation of the LTEs for overidentified systems by decomposing the m moments into the identifying space and the overidentifying space, and using both to construct a transformed criterion function for a new just-identified system[1] .

This method inherits the computational advantages of LTEs, allowing us to use MCMC methods that avoid the curse of dimensionality. Also, this approach allows us to apply them to non-smooth criterion functions.[2] It simultaneously estimates the parameters and the test statistic for overidentifying restrictions. Additionally, I use this information in alternative estimators that condition on

[1] Decompositions into an identifying and overidentifying space is shown in Sowell (1996), and used in Sowell (2009) to derive an empirical saddlepoint estimator for GMM.

[2] A common feature of structural estimation criterion functions is to be nonsmooth and highly nonconvex with numerous local optima (Chernozhukov and Hong, 2003), making extremum estimation difficult.

the over identifying restrictions being satisfied. In the presence of multiple minima, this approach amounts to using economic theory to choose between local solutions.

The proposed approach uses information in the overidentifying restrictions that other estimation methods neglect in estimation and that is always available from the moment conditions. In a simulation, the proposed estimators, in particular those that condition on the overidentifying restrictions being satisfied, perform better than other LTE estimates. In the presence of multiple minima, some of which occur far from the true population-parameter values, the new proposed conditional estimators select the one closest to the true population-parameter value more often. As a result, they show lower root-mean-square errors (RMSEs) and variances. They also sometimes outperform two-step GMM estimation, a method that often presents the aforementioned computational challenges.

Section 2 reviews the importance of the overidentifying restrictions and shows how to transform the GMM criterion function to include both the identifying and overidentifying spaces. Section 3 adapts the transformed objective function to LTEs. MCMC sample data generation and kernel-density estimation are motivated for the new transformed criterion function. Corollaries proving the consistency of the proposed estimators are presented. Section 4 shows simulation results and compares the proposed approach with others that do not account for the overidentifying restrictions.

2 Incorporating overidentifying restrictions in estimation

In an overidentified system of m moments, $m - k$ linear combinations are not set to zero by the estimator obtained from the first order conditions (FOCs) with respect to the k-dimensional set of parameters θ.

Assumption 1. *θ is a k-dimensional set of parameters that belongs to Θ. Θ is a k-dimensional compact subset. A population parameter value $\theta_0 \in int(\Theta)$ exists.*

However, under the null hypothesis that the model is valid, these $m - k$ dimensions, the ORs, are restricted to be zero on average as well. We can test the hypothesis that these ORs are zero to asses the validity of our model at some parameter values.

Consider the m dimensional set of moment conditions $g(x_i, \theta) = g_i(\theta)$.

Assumption 2. *The moment condition functions satisfy regularity conditions.*

- $g(x_i, \theta)$ *is measurable for each* $\theta \in \Theta$*, and is continuously differentiable with respect to* θ *in a neighborhood of* θ_0*; only the true parameter value solves* $Eg_i(\theta_0) = 0; E\left[\sup_{\theta\in\Theta} \|g(x, \theta)\|\right] < \infty$.
- $M(\theta) \equiv E\left[\frac{\partial g(x_i,\theta)}{\partial \theta'}\right]$ *is full rank* $\forall \theta \in O_\epsilon(\theta_0)$. $\frac{\partial g(x_i,\theta)}{\partial \theta'}$ *is measurable for each* $\theta \in \Theta; E\left[\sup_{\theta\in\Theta} \|\frac{\partial g(x_i,\theta)}{\partial \theta'}\|\right] < \infty$.
- $\Sigma(\theta) \equiv E\left[g(x_i, \theta)' g(x_i, \theta)\right]$ *is positive definite* $\forall \theta \in O_\epsilon(\theta_0)$*;*

  $E\left[g(x, \theta_0)' g(x, \theta_0)\right] < \infty$.

The sample analog of the moment condition, $G_N(\theta) = \frac{1}{N}\sum_{i=1}^N g_i(\theta)$, is used for estimation with a finite sample. An estimate is obtained by $\hat{\theta} = \arg\min_{\theta\in\Theta} Q_N(\theta)$, where

$$Q_N(\theta) = \frac{1}{2} G_N(\theta)' W_N G_N(\theta), \tag{2.1}$$

Assumption 3. G_N *satisfies a central limit theorem at* θ_0*, such that* $\sqrt{N} G_N \sim N(0, \Sigma_g)$.

W_N is the optimal weighting matrix, a consistent estimate of Σ_g^{-1} usually obtained using parameters from a first-round estimation. The estimate $\hat{\theta}$ satisfies

$$M_N(\hat{\theta})' W_N G_N(\hat{\theta}) = \underset{k\times 1}{0}, \tag{2.2}$$

where $M_N(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^N \frac{\partial g(x_i,\hat{\theta})}{\partial \theta'}$. These *identifying restrictions* set k moments to zero. For a given value of θ, we call the space spanned by the FOCs the identifying space. Defining $\overline{m}_N \equiv W_N^{\frac{1}{2}} M_N$, I can rewrite (2.2):

$$\overline{m}_N(\hat{\theta})' W_N^{\frac{1}{2}} G_N(\hat{\theta}) = \underset{k\times 1}{0}, \tag{2.3}$$

The columns of $\overline{m}_N$ define a basis for the subspace of dimension k spanned by the FOCs. The projection matrix[3] $P_{\overline{m}_N}(\theta) \equiv \overline{m}_N(\overline{m}_N' \overline{m}_N)^{-1} \overline{m}_N'$ can be written as the spectral decomposition[4]:

[3] Explicit dependence of $\overline{m}_N$ on θ is dropped to simplify notation

[4] The spectral decomposition is not unique, raising a potential concern. However, Sowell (2009) proves that inference is invariant with respect to alternative spectral decompositions

$$P_{\overline{m}_N}(\theta) = C_N(\theta)\Lambda C_N(\theta)'. \tag{2.4}$$

C_N columns determine an orthonormal basis for $\overline{m}_N$. Let C_{1N} be the column vectors in C_N that have the same column span as $\overline{m}_N$ (and therefore span the space that is defined by the FOCs), and let C_{2N} be the column vectors that span the orthogonal complement.[5] I use this spectral decomposition to locally parameterize the space spanned by the sample moments G_N, decomposing it into the k-dimensional identifying space and $(m-k)$-overidentifying space:

$$\underset{m\times 1}{\Psi_N}(\alpha) = \begin{bmatrix} \left(C_1(\theta)'W_N^{\frac{1}{2}}G_N(\theta)\right)_{k\times 1} \\ \left(\lambda - C_2(\theta)'W_N^{\frac{1}{2}}G_N(\theta)\right)_{(m-k)\times 1} \end{bmatrix}, \tag{2.5}$$

with $\alpha \equiv (\theta'\lambda')'$. For each value of θ, λ is a set of $m-k$ parameters that span the space orthogonal to the space spanned by θ. Under the null hypothesis that the moments are satisfied at θ_0, the population parameter values for λ is $\lambda_0 = 0$. The ORs test is then equivalent to testing $\lambda = 0$. In fact, $N\hat{\lambda}\hat{\lambda}$ is equivalent to the J-test statistic in Hansen (1982).

I use

$$L_N(\alpha) = \frac{1}{2}\Psi_N(\alpha)'\Psi_N(\alpha) \tag{2.6}$$

as the new objective function to obtain parameter estimates. $\Psi_N(\alpha)$ defines a just-identified system of moments, so $L_N(\alpha)$ does not require a weighting matrix to give efficient estimates. (2.5) defines valid moments because they are zero in expectation at population parameter values. The minimization of (2.6) results in θ estimates equivalent to those from GMM because the first k equations are equivalent to (2.3). Independently, for every θ, the remaining $m-k$ equations in (2.6) give the estimate $\hat{\lambda} = C_{2N}(\hat{\theta})'W_N^{\frac{1}{2}}G_n(\hat{\theta})$. Thus, the estimate $\hat{\alpha} = (\hat{\theta}'\hat{\lambda}')'$ sets all of the m equations in (2.5) exactly to zero.

Introducing the ORs in the criterion function allows us to use Bayesian methods to create a joint distribution for both the parameter θ and the ORs test statistic λ, and estimate them

[5] Equation (2.4) is idempotent. C_{1N} has the same column span as $\overline{m}_N$ since $P_{\overline{m}_N}(\theta) = \left[C_{1N\,m\times k}\; C_{2N\,m\times(m-k)}\right]\begin{bmatrix} I_k & 0 \\ 0 & 0_{m-k}\end{bmatrix}\begin{bmatrix} C'_{1N\,k\times m} \\ C'_{2N\,(m-k)\times m}\end{bmatrix} = C_{1N}C'_{1N}$. In a similar fashion, I have for the projection matrix onto the orthogonal space $P^{\perp}_{\overline{m}_N}(\theta) = C_{2N}C'_{2N}$.

simultaneously.

Additionally, it allows for alternative estimators that can use the overidentifying restrictions as a criteria to select between multiple local minima in finite samples. Dominguez and Lobato (2003) stress the existence of multiple local optima with finite samples in GMM estimation even under asymptotic identification. Under multiple local minima, the global minimum changes as observations are added. Given a finite sample, nothing guarantees the global minimum observed will be equal to the minimum that is closest to the population value (among all local minima). This sample multiple local minima issue is something GMM as a method ignores, focusing only on global minima.

Multiple minima is a recurrent phenomenon in GMM empirical work, as noted in Eichenbaum (1989) , Stock et al. (2002) , Dominguez and Lobato (2003) , and Imbens et al. (1998) . Phillips (1989) points out that most empirical work is performed under conditions of apparent identification. Estimation and testing is carried out as if convergence to a non-random function with a unique minimum has already been attained. The same is not true for small sample sizes often used in empirical work.

Results in section 4 show that introducing information embedded in the overidentifying restrictions space in the estimation procedure improves the probability of selecting the true population-parameter value.

3 Laplace Type Estimation in a Classical Framework

LTEs[6] proposed in Chernozhukov and Hong (2003) use Bayesian techniques in a classical framework. The framework is classical in the sense that it does not use likelihood fucntion, but rather other criterion functions that are transformed into proper distributions over the relevant parameters. Moments of these distributions are defined as point estimates, and quantiles are taken as confidence intervals. I follow their approach but modify the way the criterion functions are constructed to incorporate the overidentifying restrictions estimate. As a result, I obtain a joint distribution of θ and λ and can calculate alternative point estimates.

[6] Equivalent to Quasi-Bayesian estimators. Like Chernozhukov and Hong (2003) , I use the term Laplace all along to avoid confusion with other definitions of Quasi-Bayesian.

3.1 Computation of Laplace estimators

I use the criterion function $L_N(\alpha)$ in (2.6) to construct a proper posterior distribution, $p_N(\alpha \mid Y)$, from which estimates will be calculated:

$$p_N(\alpha \mid Y = y) = \frac{\pi(\alpha)e^{L_N(\alpha)}}{\int_A \pi(\widetilde{\alpha})e^{L_N(\widetilde{\alpha})}\, d\widetilde{\alpha}}, \tag{3.1}$$

where $\pi(\alpha)$ is a prior probability distribution, Y is the data observed, and $\widetilde{\alpha}$ is an integration dummy. Given the continuity of the moment conditions $g_i(\theta)$ and the way λ is constructed, assumption 1 implies assumption 4.

Assumption 4. *α is an m-dimensional set of parameters that belongs to A. A is an m-dimensional compact subset. A population parameter value $\alpha_0 \in$ int(A) exists.*

Assumption 5. *$\rho_N(\alpha, \bar{\alpha}) : \Re^m \to \Re$ is a loss function associated with selecting $\bar{\alpha}$, when the value of the parameter is α.*

1. *$\rho_N(u)$ is convex and bounded by $1 + |u|^p$ for some $p \geq 1$.*

2. *$\rho_{_N}(\alpha, \tilde{\alpha}) = \rho(\sqrt{N}(\alpha - \widetilde{\alpha}))$, where $\rho_{_N}(\alpha, \widetilde{\alpha}) \geq 0$ and $\rho_{_N}(\alpha, \widetilde{\alpha}) = 0$ iff $\alpha = \widetilde{\alpha}$.*

3. *Identification: $\phi(\xi) = \int_{\mathbb{R}^m} \rho(u - \xi)e^{u'Ju}du$ is minimized uniquely at some $\xi^* \in \mathbb{R}^m$ for a finite $J > 0$. Here, u is the error incurred when selecting $\widetilde{\alpha}$.*

The assumptions for the loss function are similar to those in Chernozhukov and Hong (2003) , with the difference being that I will only consider symmetric loss functions. In that case, LTEs are equivalent to extremum estimators. The Laplace estimator $\hat{\alpha}_{LP}$ minimizes the expected loss for different forms of the loss function,

$$\hat{\alpha}_{LP} = \arg \inf_{\alpha \in A} R_N(\alpha) \tag{3.2}$$

with the risk function,

$$R_{_N}(\tilde{\alpha}) = E[\rho_{_N}(\alpha, \bar{\alpha}) \mid Y] = \int_{\Theta} \rho_{_N}(\alpha, \bar{\alpha}) p_{_N}(\alpha \mid Y = y)\, d\alpha \tag{3.3}$$

where $p_{_N}(\alpha \mid Y = y)$ is the posterior probability in (3.1) and $\bar{\alpha}$ is the selected value. Different loss functions change the objective function such that the estimators bear different interpretations. For instance, if the minimum squared loss function is used, the estimator corresponds to the quasi-posterior-mean,

$$\hat{\alpha}_{\text{qpm}} = \int_A \alpha p_N(\alpha)\, d\alpha \tag{3.4}$$

Other familiar forms obtained for different loss functions are modes, medians, and quantiles.

3.1.1 MCMC and Sample Generation

A large number of moments m, or a high-dimensional parameter set k, causes the curse of dimensionality in extremum estimation. This method circumvents the curse of dimensionality, which is usually found in extremum estimation (Jacquier et al., 2001), and is computationally efficient (Jacquier et al., 1994). Under the same conditions, the computation of integrals in Laplace estimators suffers as well, which justifies the use of the MCMC procedure.

MCMC methods generate samples that form an ergodic Markov chain with a target stationary distribution (Gelman et al., 2004). In our case, the sample generated consists of parameter values α with a stationary distribution equivalent to the posterior $p_N(\alpha)$. When constructing the posterior as in (3.1), this method allows for nonsmooth criterion functions $L_N(\alpha)$ and multiple minima.

The evaluation of integrals can be approximated using the generated samples $\alpha^T = \{\alpha^t\}_{t=1}^T$. Asymptotically, as $T \to \infty$,

$$\int_\Theta \alpha p_N(\alpha)\, d\alpha \approx \frac{1}{T}\sum_{t=1}^{T} f(\alpha^{(t)}), \tag{3.5}$$

This method has been shown to quickly converge to the integral evaluation in the ergodicity result of Proposition 1 in Chib and Greenberg (2003) , which I reproduce here without proof.

Proposition 1. *Let $P(\alpha^*, \alpha)$ be a Markov transition kernel for a target distribution $p(\alpha)$. If $P(\alpha^*, \alpha)$ is irreducible and has invariant distribution $p(\alpha)$, then $\mid P^m(\alpha^*, |\alpha) - p(\alpha) \mid\Rightarrow\;\; 0$ as $m \Rightarrow \infty$.* [7]

As the number of iterations m increases, a sample drawn from the distribution P will have

[7] Irreducibility refers to distributions in which no values under the space of possible values is ruled out after reaching any particular point. For more rigorous definitions, refer to Mengersen and Tweedie (1996).

a distribution equal to the probability kernel p. MCMC algorithms use this result to generate transition kernels that satisfy Proposition 1. One of the most common algorithms in MCMC is Metropolis-Hastings.

3.1.2 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a general algorithm of which other algorithms used are special cases (Chib and Greenberg, 1995). The Metropolis-Hastings algorithm was first developed by Metropolis et al. (1953) and was generalized for asymmetric proposal distributions by Hastings (1970). Chernozhukov and Hong (2003) provide an excellent introduction to this algorithm. I include its steps here for completeness.

Let $J_t(\alpha^t \mid \alpha^{t-1})$ be a proposal distribution at time t, and $p(\alpha \mid y)$ be the target distribution. For computation purposes, the proposal distribution is one for which a random variable generator exists. Metropolis-Hasting canonical usage, Gelman et al. (2004) , suggests proposal distributions of the form $J_t(\alpha^* \mid \alpha^{t-1}) = \phi(|\alpha^* - \alpha^{t-1}|)$, where ϕ is a Gaussian distribution density symmetric around 0.

The algorithm proceeds as follows:

Step-1 Draw a starting value for the parameter α^0 for which $p(\alpha^0 \mid y) > 0$.

Step-2 For t=1,2...

-a) Sample a proposal α^* from a proposal distribution $J_t(\alpha^* \mid \alpha^{t-1})$ for iteration t.

-b) Calculate the acceptance criteria r using the ratio of the densities weighted by the proposal distribution to correct for asymmetries (see equation (3.6)). If a symmetric distribution is used, then $J_t(\alpha^* \mid \alpha^{t-1}) = J_t(\alpha^{t-1} \mid \alpha^*)$, the weights cancel out, and the algorithm is simply called Metropolis algorithm.

$$r = \inf\left(1, \frac{p(\alpha^* \mid y)/J_t(\alpha^* \mid \alpha^{t-1})}{p(\alpha^{t-1} \mid y)/J_t(\alpha^{t-1} \mid \alpha^*)}\right). \tag{3.6}$$

-c) To choose the next iteration value α^t assign

$$\alpha^t = \left\{ \begin{array}{c} \alpha^{*}\text{with probability}r \\ \alpha^{t-1} \text{ with probability } 1-r \end{array} \right\}. \tag{3.7}$$

Step-3 Finish after a criterion of convergence has been attained.

A very convenient characteristic of this algorithm is that the target distribution posterior density in (3.1) needs to be known only up to a constant, so I can use $q(\alpha|Y)$ as in

$$p_N(\alpha|Y=y) \propto q(\alpha|Y=y) = \pi(\alpha)f_N(Y|\alpha). \tag{3.8}$$

It is sufficient that $J(\alpha^t \,|\, \alpha^{t-1})$ and $p(\alpha)$ be positive and continuous $\forall \alpha \in A$ for the distribution $p_{MH}(\alpha)$ of the sample generated $\{\alpha^t\}_{t=1}^{T}$ to converge to an invariant distribution equal to the target distribution $p(\alpha \,|\, y)$ (Proposition 1). Chib and Greenberg (2003); Tierney (1994); Chib and Greenberg (1995); Smith and Roberts (1993)) further discuss the convergence of this algorithm, while Gelman et al. (2004) discusses ways to improve convergence speed.

r becomes

$$r = \inf\left(1, \frac{\pi(\alpha^*)e^{L_N(\alpha^*)}/J_t(\alpha^* \,|\, \alpha^{t-1})}{\pi(\alpha^{t-1})e^{L_N(\alpha^{t-1})}/J_t(\alpha^{t-1} \mid \alpha^*)}\right) \tag{3.9}$$

after using the posterior in (3.1) and cancelling the denominator,

The algorithm generates both a sample for parameter values θ and for the estimator of the overidentifying restriction test λ. To generate both samples alternating conditional sampling [8] is suggested (Rossi et al., 2005).

3.1.3 Kernel density estimation for alternative conditional estimators.

LTEs are typically moments, the mode, or quantiles of a quasi-posterior distribution. The probability measure for any specific value generated by MCMC is 0, and thus, calculation of LTEs and hypothesis testing require smooth approximations to the population distribution. Kernel-density estimation achieves that requirement approximations through nonparametric techniques. Given a

[8] Namely, the algorithm would alternate the parameter for which a proposal value is generated and tested. In iteration t, I follow steps 1-3, generating a proposal θ^*, and determining the value for θ^t, for a given λ^t. For iteration $t+1$, I generate a proposal λ^*, and determine λ^{t+1}, fixing θ at θ^t for this iteration.

sample $\alpha^1, \alpha^2, ..., \alpha^B \sim p(\alpha)$, the kernel density approximation is given by

$$\widehat{p}_h(\alpha) = \frac{1}{Bh} \sum_{i=1}^{B} K\left(\frac{\alpha - \alpha_i}{h}\right), \tag{3.10}$$

where K is a probability kernel, and h is a bandwidth or smoothing parameter. The canonical implementation uses a standard normal density probability kernel, (Silverman, 1986; Jones and Henderson, 2007). The optimal bandwidth depends on the sample size.

I additionally implement an alternative estimate that condition explicitly on $\lambda = 0$. I use multivariate kernel-density estimation to calculate the joint distribution of θ and λ, univariate kernel density estimation for the marginal density for λ and the probability $p(\lambda = 0)$, together with Bayes theorem to obtain the conditional density $p(\theta \mid \lambda = 0)$. In contrast to previously used methods, this alternative estimator is feasible only as a result of the proposed parametrization in equation (2.5). As explained in section 2, this alternative estimator allows for the use of the ORs as a criterion to select between local minima. The estimates that are based on densities conditional on $\lambda = 0$ on average select parameter values that more often satisfy economic theory (sample moments are zero).

3.2 Consistency of LTEs

Chernozhukov and Hong (2003) show LTEs are consistent and asymptotically normal for general criterion functions $L_N(\theta)$. Here, I explore a similar approach applied to the transformed objective function $\Psi(\theta, \lambda)$ in (2.6). I establish corollaries of the theorems in Chernozhukov and Hong (2003) that apply to the specific case proposed here. First, I state any additional assumptions required. Then I show the quasi-posterior density concentrates around the true population value α_0, converging at the rate$\frac{1}{\sqrt{N}}$. Finally, I show the estimator is consistent and asymptotically normal.

3.2.1 Assumptions

Assumption 6. ***Compactness.*** *The extended parameter space A is a compact subset of the Euclidian space $\Re^m$, where m is the number of moments.*

Assumption 7. ***Identification.***

(i) Identification for every N: *For every $\delta > 0$, $\exists\, \epsilon > 0$, such that*

$$\lim_{N\to\infty} \inf P\left\{ \sup_{|\alpha-\alpha_0|\geq\delta} \frac{1}{N}(L_N(\alpha) - L_N(\alpha_0)) \leq -\epsilon \right\} = 1,$$

,

(ii) Asymptotic Identification*: $\exists$ a nonstochastic function $M_N(\alpha)$ that is continuous on A with a unique sup at α_0, such that*

(a) for any $\delta > 0$, $\epsilon > 0$

$$\lim_{N\to\infty} \sup P\left\{ \sup_{|\alpha-\alpha_0|\geq\delta} M_N(\alpha) - M_N(\alpha_0) > \epsilon \right\} = 0,$$

(b) $L_N(\alpha)/N \to_p M_N(\alpha)$, and

(c) $L_N(\alpha)$ is continuously differentiable. $\triangledown_{\alpha\alpha'} L_N(\alpha_0) = \mathcal{O}(1)$, and for some $d > 0$ and each $\epsilon > 0$

$$\lim_{N\to\infty} \sup P\left\{ \sup_{|\alpha-\alpha_0|\geq\delta} \left| \frac{\triangledown_{\alpha\alpha'} L_N}{N} - \triangledown_{\alpha\alpha'} M_N(\alpha_0) \right| > \epsilon \right\} = 0.$$

Assumption 8. ***Convergence of expansion****. For α in an open neighborhood of α_0,*

(i) Taylor expansion:

$$L_N(\alpha) - L_N(\alpha_0) = (\alpha - \alpha_0)' \triangledown_\alpha L_N(\alpha_0) - \frac{1}{2}(\alpha - \alpha_0)' \triangledown_{\alpha\alpha'} L_N(\alpha - \alpha_0) + R_N(\alpha)$$

(ii) Asymptotic normality of the FOCs*: $\exists$ a positive definite constant matrix $\Omega_N(\alpha_0)$ such that $\forall N$ such that $\Omega_N^{-\frac{1}{2}}(\alpha_0) \triangledown_\alpha L_N(\alpha_0)/\sqrt{N} \to_d \mathcal{N}(0, \mathcal{I}$.*

(iii) Convergence of Taylor expansion: *for every $\epsilon > 0$, $\exists\ \delta > 0$ and $S > 0$ such that*

(a)

$$\lim_{N\to\infty} \inf P\left\{ \sup_{|\alpha-\alpha_0|\leq\frac{S}{\sqrt{N}}} |R_N(\alpha)| > \epsilon \right\} = 0,$$

(b)

$$\lim_{N\to\infty} \inf P\left\{ \sup_{\frac{S}{\sqrt{N}}<|\alpha-\alpha_0|\leq\delta} \frac{|R_N(\alpha)|}{N\,|\alpha - \alpha_0|^2} > \epsilon \right\} = 0.$$

3.2.2 Corollaries

Results in this section are corollaries of the theorems in Chernozhukov and Hong (2003). Discussion is provided here and proofs are included in the appendix. Corollary 1 shows the quasiposterior density converges around α_0 at a rate of $\frac{1}{\sqrt{N}}$, which suggests the consistency of estimators based on the quasiposterior density. The norm used to show convergence is the total variation of moments norm. Let h be the normalized deviation from α_0, displaced by the normalized *score function*

$$h_N \equiv \sqrt{N}(\alpha - \alpha_0) + \sqrt{N}(\triangledown_{\alpha\alpha'} L_N(\alpha_0))^{-1} \triangledown_\alpha L_N(\alpha_0). \tag{3.11}$$

The quasiposterior density of α can be written as

$$p_N(\alpha) = p_N\left(h_N/\sqrt{N} + \alpha_0 - (\triangledown_{\alpha\alpha'} L_N(\alpha_0))^{-1} \triangledown_\alpha L_N(\alpha_0)\right) \tag{3.12}$$

The localized quasiposterior density for h_N is defined as $p_N^*(h_N) = \frac{1}{\sqrt{N}} p_N(\alpha)$. The total variation of moments norm for a function p on A is

$$\|p\|_{TVM} \equiv \int_A (1 + |h_N|^\beta) \, |p(h_N)| \, dh_N. \tag{3.13}$$

This norm is equivalent to the total variation norm when $\beta = 0$.

Corollary 1. *(Convergence of the quasiposterior density) Under assumptions 1-4, and for any finite nonnegative β,*

$$\|p_N^*(h_N) - p_\infty^*(h_N)\|_{TVM} \equiv \int_{H_N} (1 + |h_N|^\beta) \, |p_N^*(h_N) - p_\infty^*(h_N)| \, dh_N \to_p 0, \tag{3.14}$$

where $H_N = \{h_N : \alpha \in A\}$ and

$$p_\infty^*(h_N) \equiv \sqrt{\frac{\det J_N(\alpha_0)}{(2\pi)^m}} \exp\left(-\frac{1}{2} h_N' \frac{\triangledown_{\alpha\alpha'} L_N(\alpha_0)}{N} h_N\right). \tag{3.15}$$

For a large N, the density $p_N(\alpha)$ is random normal, with mean $\alpha_0 - \sqrt{N}(\triangledown_{\alpha\alpha'} L_N(\alpha_0))^{-1} \triangledown_\alpha$

$L_N(\alpha_0)$, and variance $\sqrt{N}(\bigtriangledown_{\alpha\alpha'} L_N(\alpha_0))^{-1}$, which implies, in the context of the LTEs proposed, the quasiposterior density has the same asymptotic distribution and convergence properties as the GMM estimator for the set of moments $\Psi(\alpha)$ in (2.5).

To establish $\sqrt{N}$-consistency, recall that extremum estimators $\breve{\alpha}$ are defined as the $\arg\sup_{\alpha\in A} L_N(\alpha)$, and regularly satisfy a set of FOCs ($\bigtriangledown_\alpha L_N(\breve{\alpha}) = 0$). The Taylor expansion of the FOCs gives

$$\bigtriangledown_\alpha L_N(\alpha) = \bigtriangledown_\alpha L_N(\alpha_0) + (\alpha_0 - \alpha) \bigtriangledown_{\alpha\alpha'} L_N(\bar{\alpha}) + \mathcal{O}(\sqrt{N}). \tag{3.16}$$

Evaluating at $\breve{\alpha}$ sets the LHS to zero. Rearranging, I get that the extremum estimator $\sqrt{N}(\breve{\alpha} - \alpha_0)$ is first-order equivalent to

$$U_N = \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0)^{-1} \bigtriangledown_\alpha L_N(\alpha_0).$$

From (3.2) and $p_N^* \to p_\infty^*$, $\sqrt{N}(\alpha_{\hat{LTE}} - \alpha)$ can be estimated asymptotically by

$$z_n = \arg\inf_{\alpha\in A} \left\{ \int_A \rho(z-u) p_\infty^*(u - U_N) du \right\} \tag{3.17}$$

Notice the similarity between (3.17) and the equation that defines LTEs estimator (3.2). z_N and U_N are related by the following

$$z_N - U_N = \epsilon_N \tag{3.18}$$

where

$$\epsilon_N \equiv \arg\inf_{\alpha\in A} \int_A \rho(z-u) p_\infty^*(u) du \tag{3.19}$$

which is similar to (3.17) with a shifted probability function.

Corollary 2. *($\sqrt{N}$-consistency and asymptotic normality) Under assumptions 1-4, I get the following relationships:*

(i) $L_N(\alpha) - L_N(\alpha_0) = (\alpha - \alpha_0)' \bigtriangledown_\alpha L_N(\alpha_0) - \frac{1}{2}(\alpha - \alpha_0)' \bigtriangledown_{\alpha\alpha'} L_N(\alpha - \alpha_0) + O_N(\alpha)$,

(ii) $\sqrt{N}(\hat{\alpha} - \alpha_0) = \epsilon_N + U_N + o_p(1)$, *and*

(iii) $\Omega_N^{-\frac{1}{2}} \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0) U_N \rightarrow_d \mathcal{N}(0, I)$.

Hence,

(iv) $\Omega_N^{-\frac{1}{2}}(\alpha_0) \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0)(\sqrt{N}(\hat{\alpha} - \alpha_0) - \epsilon_N) \rightarrow_d \mathcal{N}(0, I)$

and $\epsilon_N = 0$ *for symmetric loss functions.*

Valid large-sample confidence intervals from the quasi-posterior distribution require that an information equality property, $\Omega_N(\alpha_0) \sim \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0)$, be satisfied. It is easy to show this property holds for GMM, and thus holds for the transformed objective function (2.6). Asymptotically, the FOCs of L_N are distributed:

$$\sqrt{N} G_N(\alpha) \Sigma^{-1} \bigtriangledown_\alpha G_N(\alpha) \sim \mathcal{N}(0, \bigtriangledown_\alpha G_N(\alpha)' \Sigma^{-1} \bigtriangledown_\alpha G_N(\alpha)) \tag{3.20}$$

implying that asymptotically $\bigtriangledown_{\alpha\alpha'} G_N(\alpha) = \bigtriangledown_\alpha G_N(\alpha)' \Sigma^{-1} \bigtriangledown_\alpha G_N(\alpha)$. Combining this result with assumption 8(ii) I get the information equality property. In GMM, the use of an optimal weighting matrix ensures information equality.

The usual asymptotic intervals for a function $w(\alpha)$, where α is an estimator that satisfies (3.11) are

$$\begin{aligned} &[w(\hat{\alpha}) + t_{\frac{d}{2}} \sqrt{\bigtriangledown_\alpha w(\hat{\alpha})' \bigtriangledown_{\alpha\alpha'} G_N(\alpha)^{-1} \bigtriangledown_\alpha w(\hat{\alpha})}, \\ &w(\hat{\alpha}) + t_{1-\frac{d}{2}} \sqrt{\bigtriangledown_\alpha w(\hat{\alpha})' \bigtriangledown_{\alpha\alpha'} G_N(\alpha)^{-1} \bigtriangledown_\alpha w(\hat{\alpha})}] \end{aligned} \tag{3.21}$$

Alternatively, I can calculate intervals from the quantiles of the sequence $\{w(\alpha^{(1)}), ..., w(\alpha^{(B)})\}$, constructed using the α-sample generated through MCMC. Finally, Corollary 3 shows that confidence intervals generated from the quasi-posterior distribution are asymptotically equivalent to the asymptotic confidence intervals.

Corollary 3. *Under assumptions 1-4, information equality,* $\forall d \in (0, 1)$*, and for any continuously differentiable function* $w(\alpha)$*,*

1. $c_{g,N}(d) = g(\hat{\alpha}) + t_d \sqrt{\bigtriangledown_\alpha f_N(\alpha_0)'(-\bigtriangledown_{\alpha\alpha'} L_N(\alpha_0))^{-1} \bigtriangledown_\alpha f_N(\alpha_0)} + o_p(1/\sqrt{N})$ *and*

2. $\lim_{N\rightarrow\infty} P\{c_{g,N}(d/2) \leq g(\alpha_0) \leq c_{g,N}(1 - d/2)\} = 1 - d$.

To feasibly estimate $\bigtriangledown_{\alpha\alpha'} L_N(\alpha_0)^{-1}$, we can use the variance-covariance matrix $\{w(\alpha^{(1)}), ..., w(\alpha^{(B)})\}$.

4 Simulation

I simulate a model with one overidentifying restriction. I generate a sample for θ and λ through MCMC. I estimate quasi-posterior means as in Chernozhukov and Hong (2003) , and compare them to quasi-posterior means using the transformed equations from section 2. I calculate joint densities and the corresponding quasi-posterior modes. Finally, I calculate an alternative estimator where I impose that $\lambda = 0$. For this estimator, I calculate the conditional density for $\lambda = 0$, and the corresponding conditional means and modes.

4.1 Model

The model used follows Hall and Horowitz (1996). This model has one parameter ($k = 1$), two moment conditions ($m = 2$), and hence one overidentifying restriction ($m - k = 1$). The sample moments are

$$\underset{2\times 1}{G_N(\theta)} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} \exp(\mu - \theta(X_i + Z_i) + 3Z_i) - 1 \\ \frac{1}{N}\sum_{i=1}^{N} (\exp(\mu - \theta(X_i + Z_i) + 3Z_i) - 1)Z_i \;. \end{bmatrix} \tag{4.1}$$

This model has been widely used in recent literature. For examples, see Sowell (2009) , Imbens et al. (1998) , Kitamura (2001) , and Schennach (2007) . Gregory et al. (2002) present an economic interpretation of the model. The population-parameter value is $\theta_0 = 3$. X_i and Z_i are scalar observations distributed iid $\sim N(0, s^2)$. μ is a normalization constant set to $\mu = \dfrac{\theta_0^2 s^2}{2}$ to make moment expectation zero at θ_0. I use rather small sample sizes to capture the occurrence of multiple minima that typically occur in estimation before they wash away asymptotically.

This model often presents two modes or local solutions. The first moment condition (upper row of (4.1)) is set to zero at both $\theta = 0$ and $\theta = \theta_0$. The second moment condition is zero only at $\theta = \theta_0$. Appendix C presents an analytical solution of the system and shows why multiple modes happen despite identification under a unique population value. The common occurrence of multiple modes in GMM estimation,[9]make this a convenient model to evaluate estimator performance.

In finite samples, the parameter θ and the overidentifying restriction statistic λ are not independent (Sowell, 2009). If we condition on $\lambda = 0$, we can estimate the θ value for which the null hypothesis of moments being zero (the theory being true) is not rejected. I report the parameter estimates conditional on $\lambda = 0$, and compare them with estimations that do not explicitly condition on the overidentifying restrictions.

Burn-in iterations precede simulations to remove dependence from initial values. Posterior densities are produced using the methods described in section 3, and using the criterion function $L_N(\alpha)$ in (2.6). To construct the function $\Psi_N(\alpha)$ in (2.5), the moments in (4.1) are used.

For MCMC, I use a normal symmetric distribution, and a flat or diffuse prior. Figure 4.1 shows examples of the distribution of the sample generated with MCMC for a representative random draw. Figures A.1 and A.2 in Appendix A show a plot of the posterior densities analytically calculated for the same draws. The distribution of the sample corresponds very closely to the true distributions, and both have their modes around the local minima of the objective function. For small sample sizes, we often observe multiple minima as depicted for this draw. From the analytical solution of the model (see Appendix C) we know that only the unique population parameter value $\theta_0 = 3$ makes the expectation of all moments equal to zero, satisfying the overidentifying restrictions. Figure 4.1-b shows how when I condition on the null hypothesis $\lambda = 0$, the probability mass for the *incorrect* local minima (around $\theta = 0$) almost disappears and most of it is now located around the true population value θ_0. On average, by conditioning on $\lambda = 0$, I rule out the local minima for which the theory is not satisfied. Figures 4.1c, 4.1d, and A.2 show the case of a representative draw for $N = 50$. In this case, most of the probability mass is located around the true population value θ_0, both for conditional and unconditional estimators. Figure A.2 shows the GMM objective function is converging and presents presents a unique minimum.[10]

Figures 4.1-c and d show that in the case of a unique minimum, conditioning on $\lambda = 0$ barely affects the quasiposterior distribution and the means and modes, suggesting the improvement of this method is found mainly in small samples.

[9] See discussion in section (2).

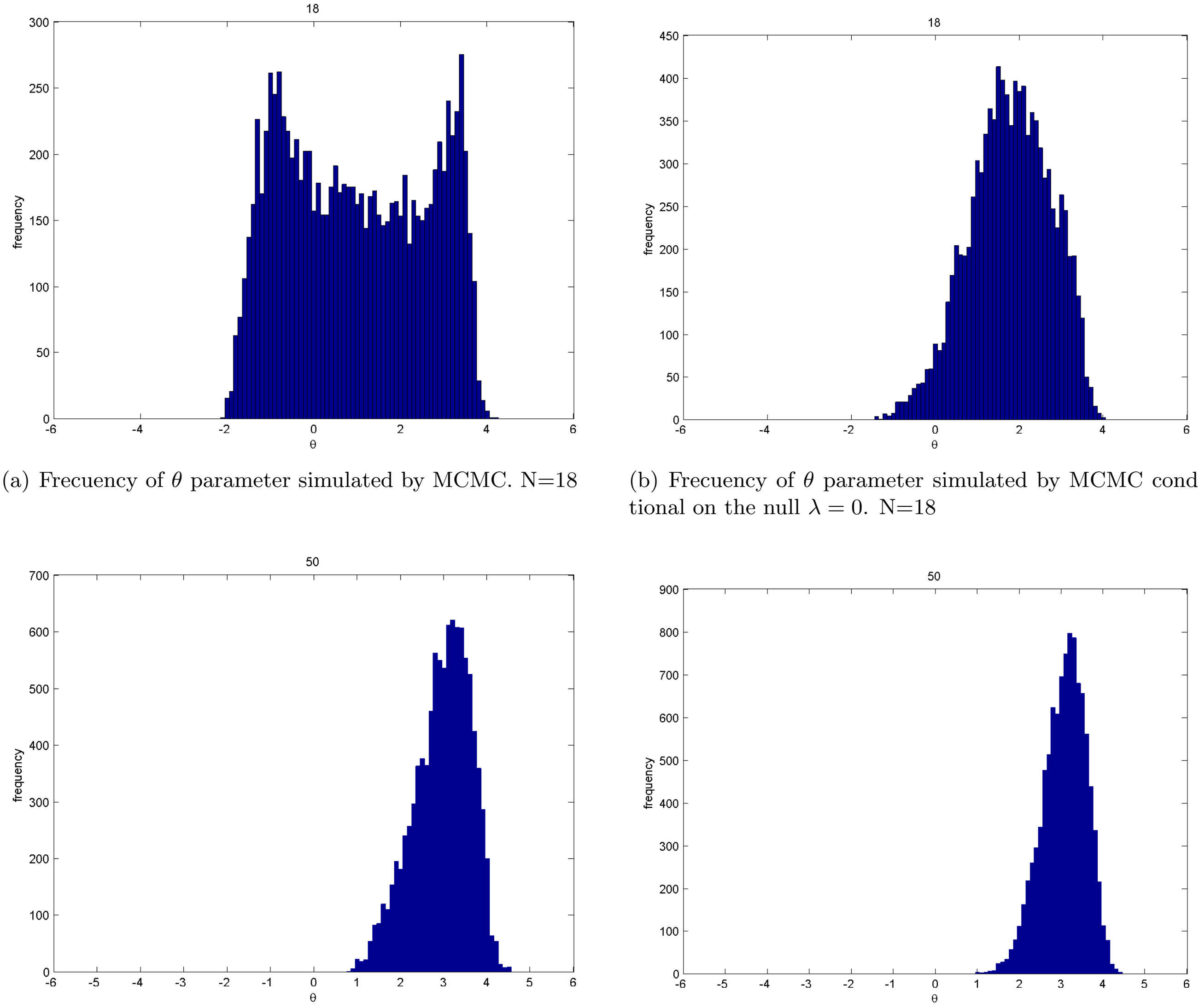
[10] Plots show a representative draw. The probability of multiple minima remains positive for any finite sample size; therefore, introducing the overidentifying restrictions remains useful. In particular, for simulations with $N \geq 50$, I see that multiple minima still occur although significantly less often than for smaller sample sizes.

Table 1: Simulation for the Hall and Horowitz (1996) model using MCMC. 1000 repetitions.

Sample	Estimator	RMSE		BIAS	
Size		s=0.4	s=0.6	s=0.4	s=0.6
18	Q-posterior mean	1.38	1.20	0.88	0.88
	Q-posterior mean,$\lambda = 0$	1.09	1.12	0.37	0.72
	Q-posterior mode	2.10	1.65	4.03	2.79
	Q-posterior mode, $\lambda = 0$	1.04	1.50	1.09	2.29
	GMM	1.46	1.51	0.17	0.90
	C-H	1.31	1.38	0.40	0.40
37	Q-posterior mean	0.75	4.84	0.40	1.39
	Q-posterior mean,$\lambda = 0$	0.73	4.75	0.35	1.29
	Q-posterior mode	1.80	6.03	1.06	1.27
	Q-posterior mode,$\lambda = 0$	1.10	5.82	0.27	0.93
	GMM	0.86	1.28	0.17	0.51
	C-H	1.90	5.59	0.54	1.47
50	Q-posterior mean	0.60	1.42	0.38	0.80
	Q-posterior mean,$\lambda = 0$	0.53	0.89	0.30	0.51
	Q-posterior mode	0.96	1.26	0.10	0.29
	Q-posterior mode,$\lambda = 0$	0.57	1.11	0.10	0.29
	GMM	0.59	1.06	0.10	0.22
	C-H	0.75	1.45	0.10	0.91

Table 2: Simulation for the Hall and Horowitz (1996) model using MCMC. 10,000 repetitions.

Sample	Estimator	RMSE		BIAS	
Size		s=0.4	s=0.6	s=0.4	s=0.6
18	Q-posterior mean	4.83	4.27	0.81	0.95
	Q-posterior mean,$\lambda=0$	4.29	4.02	0.34	0.73
	Q-posterior mode	6.03	4.68	0.63	1.14
	Q-posterior mode, $\lambda=0$	5.15	4.15	1.10	0.99
	GMM	1.46	1.50	0.17	1.73
	C-H	4.88	3.00	0.71	0.58
37	Q-posterior mean	0.75	4.83	0.48	0.62
	Q-posterior mean,$\lambda=0$	0.64	3.99	0.35	0.57
	Q-posterior mode	0.83	4.00	0.27	0.40
	Q-posterior mode,$\lambda=0$	0.76	4.10	0.17	0.30
	GMM	0.78	1.15	0.16	0.45
	C-H	3.67	6.39	0.35	0.89
50	Q-posterior mean	3.40	6.11	0.60	1.20
	Q-posterior mean,$\lambda=0$	3.31	5.10	0.47	0.76
	Q-posterior mode	5.80	7.71	0.27	0.99
	Q-posterior mode,$\lambda=0$	3.17	6.15	0.25	0.75
	GMM	0.78	2.30	0.05	0.35
	C-H	5.24	7.61	0.57	1.21



(a) Frecuency of θ parameter simulated by MCMC. N=18

(b) Frecuency of θ parameter simulated by MCMC conditional on the null $\lambda = 0$. N=18

(c) Frecuency of θ parameter simulated by MCMC. N=50

(d) [Frecuency of θ parameter simulated by MCMC conditional on the null $\lambda = 0$. N=50

Figure 4.1: Hall and Horowitz GMM model. MCMC Sample Generation. 10000 iterations after burn-in.

In the tables, C-H refers to the estimator used in Chernozhukov and Hong (2003) , namely the quasiposterior mean calculated using the criterion function that does not consider the overidentifying restrictions, $Q_N(\theta)$ in (2.1), using the moments in (4.1). The other Q-posterior estimates report the transformed Laplace estimators, those that use the criterion function function $L_N(\theta)$ in (2.6) constructed using the same moments. GMM refers to the results of the typical GMM method.

Root mean square error (RMSE) indicates accuracy of the estimator including both a measure of bias and of variability. On average, all Laplace estimators perform better than the GMM estimator. In all cases, I should keep in mind that GMM is an extremum estimation procedure and suffers from the curse of dimensionality for high dimensional models, whereas the other LTEs estimators do not. Thus, I prefer to focus our attention on the relative performances of the LTEs here. The transformed LTEs, in most cases perform better than the C-H estimator, although results are mixed. Quasi-posterior modes and means conditional on $\lambda = 0$ perform better on average than the rest of the LTEs. In particular, their RMSEs are usually lower. This is in agreement with the intuition discussed when motivating the method, in which I claimed that these estimators, by conditioning on the null hypothesis $\lambda = 0$ being satisfied, use the theory as a useful criterion to select between multiple minima.

5 Conclusion

I use Bayesian techniques in a classical estimation framework to calculate estimators that are computationally efficient, overcome the curse of dimensionality, and work for non-smooth criterion functions, following Chernozhukov and Hong (2003). Estimators are extended by incorporating the ORs in our estimation equations to simultaneously estimate the parameters and the overidentifying restrictions test statistic, no longer assuming both are independent. Corollaries show consistency, asymptotic normality, and the validity of the confidence intervals calculated for the transformed LTEs.

Introducing the information in the $m - k$ dimensions spanned by the overidentifying restrictions improves the estimation of the k parameters. This new information allows the creation of multivariate densities that include both indentifying and over-identifying dimensions, allowing, for example, the calculation of alternative conditional estimators that condition on the overidentifying conditions being satisfied. In the presence of multiple minima of the objective function, the conditional estimators select more often the local minimum that is closer to satisfying the overidentifying restrictions, getting closer, on average, to the true population parameter value.

A simulation study for a nonlinear asset-pricing model in Hall and Horowitz (1996) that often presents multiple local minima illustrates the properties of the extended quasiposterior means and modes. Overall, among the extended estimators, those that condition on the overidentifying restrictions being satisfied ($\lambda = 0$) perform better than their unconditional counterparts. Moreover, the transformed LTEs often perform better than the quasiposterior mean proposed in Chernozhukov and Hong (2003), which does not simultaneously estimate the model parameters and the ORs.

Introducing the ORs in parameter estimation uses information from economic theory embedded in the moment conditions. This information was previously ignored during the process of parameter estimation in econometrics with Bayesian methods. Results justify the use of these estimators as an alternative to extremum estimators for general criterion functions.

Further work should extend simulation to more complex models. Specifically, simulations should be carried out for a large number of moments, parameters, and ORs, such that the curse of dimensionality would actually affect GMM estimation, making the benefits of LTEs stand out. More complicated models usually present multiple modes (multiple local minima in the sample objective functions) more frequently, which might increase the performance improvement observed in the proposed extended conditional estimators.

References

Andrews, D. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric Theory 11*(3).

Andrews, D. (2010). The large sample correspondence between classical hypothesis tests and bayesian posterior odds tests. *Econometrica 62*(5).

Angers, J. and P. Kim (2005). Multivariate bayesian function estimation. *Annals of Statistics 33*(6).

Chernozhukov, V. and H. Hong (2003). An mcmc approach to classical estimation. journal of econometrics. *Journal of Econometrics 115.*

Chib, S. and E. Greenberg (1995). Understanding the metropolis-hastings algorithm. *The American Statistician 49*(4).

Chib, S. and E. Greenberg (2003). Markov chain monte carlo simulation methods in econometrics. *Econometric Theory 12*(3).

Dominguez, M. and I. Lobato (2003). Consistent estimation of models defined by conditional moment restrictions. *Journal of Econometrics 115.*

Eichenbaum, M. (1989). Some empirical evidence on the production level and production cost smoothing models of inventory investment. *Econometrica 79*(4).

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian Data Analysis.* Chapman and Hall/CRC.

Gregory, A., J. Lamarche, and G. Smith (2002). Information-theoretic estimation of preference parameters: macroeconomic applications and simulation evidence. *Journal of Econometrics 107.*

Hall, P. and J. L. Horowitz (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica: Journal of the Econometric Society*, 891–916.

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

Harrison, P. and C. Stevens (1976). Bayesian forecasting. *Journal of the Royal Statistical Society 38*(3).

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrica 57*.

Imbens, G., R. Spady, and P. Jhonson (1998). Information theoretic approaches to inference in moment condition models. *Econometrica 66*(2).

Jacquier, E., M. Johanne, and N. Polson (2001). Mcmc maximum likelihood for latent state models. *Journal of Econometrics 69*(6).

Jacquier, E., N. Polson, and P. Rossi (1994). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics 12*(4).

Jones, M. and D. Henderson (2007). Kernel-type density estimation on the unit interval. *Biometrika*, 977–984.

Kitamura, Y. (2001). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica 69*(6).

Mengersen, K. and R. Tweedie (1996). Rates of convergence of the hastings and metropolis algorithms. *Annals of Statistics 24*(1).

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*.

Phillips, P. (1989). Partially identified econometric models. *Econometric Theory 5*.

Rossi, P. E., G. M. Allenby, and R. E. McCulloch (2005). *Bayesian statistics and marketing.* Wiley New York.

Schennach, S. (2007). Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica 75*.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall/CRC.

Smith, A. and G. Roberts (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society Series B 55*(1).

Sowell, F. (1996). Optimal tests for parameter instability in the generalized method of moments framework. *Econometrica 64*(5), 1085–1107.

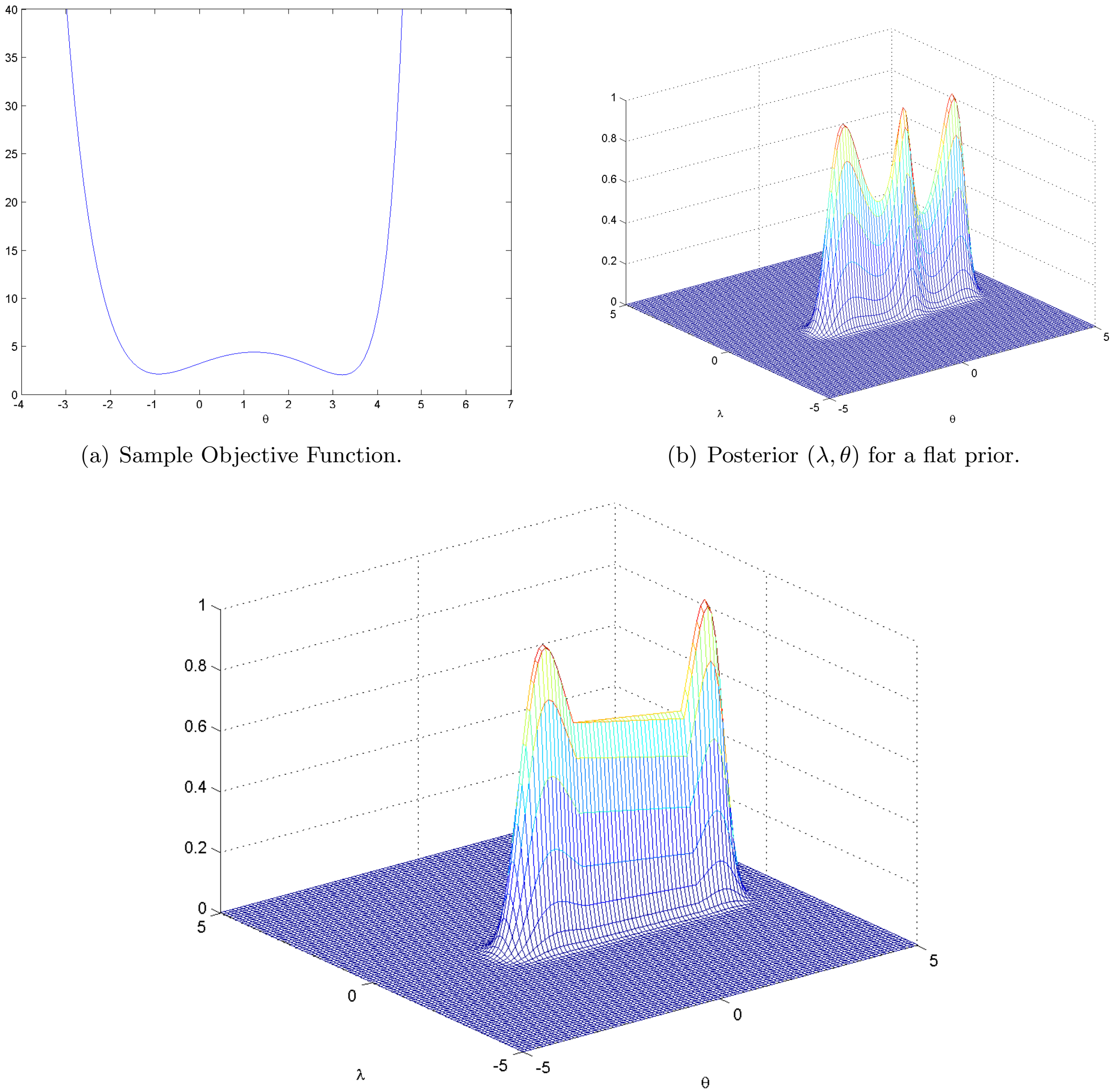
Sowell, F. (2009). The empirical saddlepoint likelihood estimator applied to two-step gmm.

Stock, J., J. Wright, and M. Y. M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economics Statistics 20*.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics 22*(4).

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* Wiley.

A Simulation details

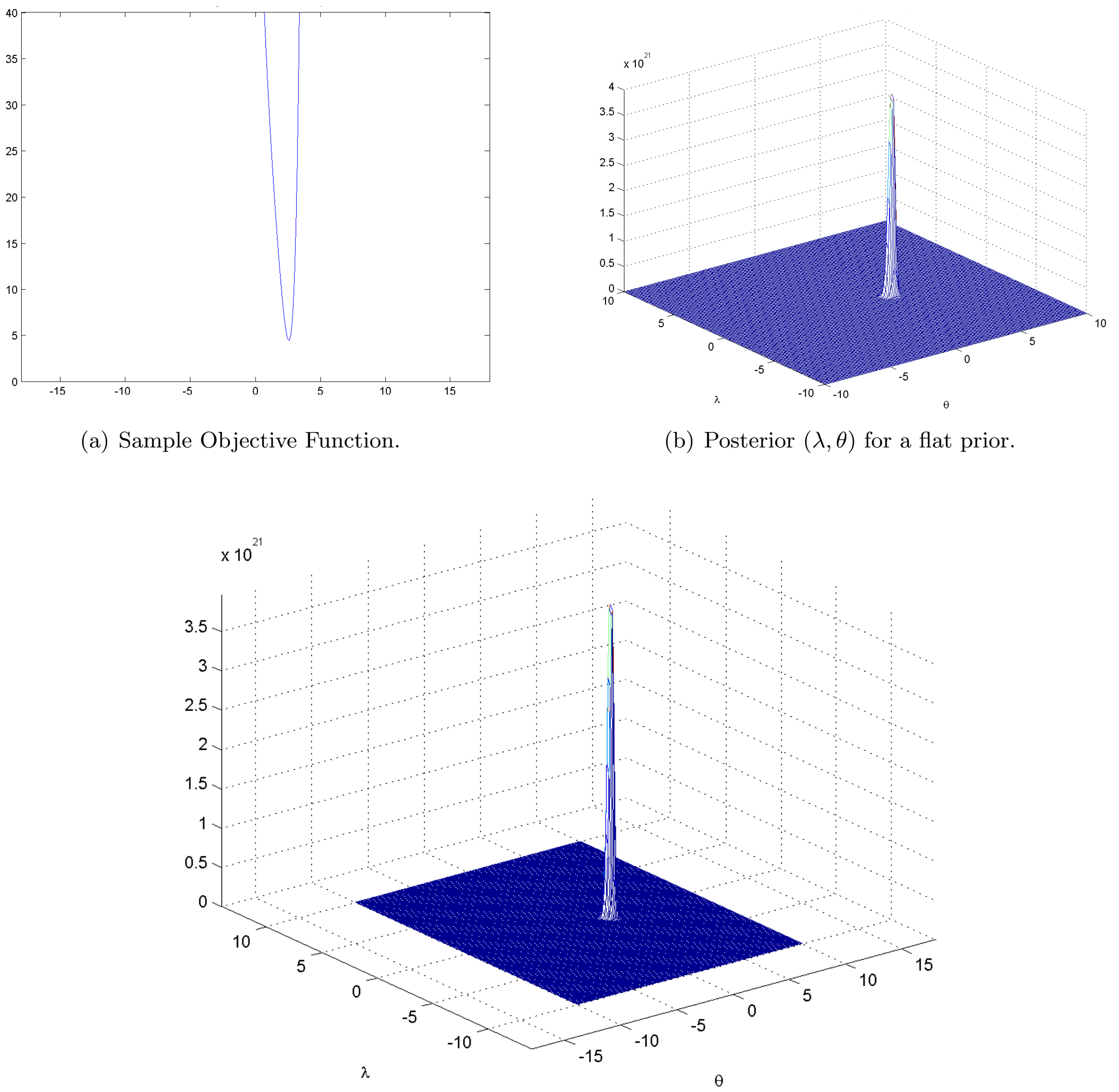


(a) Sample Objective Function.

(b) Posterior (λ, θ) for a flat prior.

(c) Transformed posterior (λ, θ) using linear interpolation

Figure A.1: Hall and Horowitz GMM model. Sample size $N = 18$.



(a) Sample Objective Function.

(b) Posterior (λ, θ) for a flat prior.

(c) Transformed posterior (λ, θ) using linear interpolation

Figure A.2: Hall and Horowitz GMM model. Sample size $N = 50$.

B Proofs.

B.1 Proof of Corollary 1

It is enough to show that

$$\int (|h_N|^{\beta}) \left| p_N^*(h_N) - p_\infty^*(h_N) \right| dh_N \rightarrow_p 0 \tag{B.1}$$

for a general β. Rewrite the localized quasiposterior density for h_N in equation (3.11) using the modified quasiposterior density in equation (3.1),

$$\begin{aligned} p_N^*(h_N) &\equiv \frac{1}{\sqrt{N}} p_N(\alpha) \\ &= \frac{\pi(\alpha) e^{L_N(\alpha)}}{\int_A \pi(\widetilde{\alpha}) e^{L_N(\widetilde{\alpha})}\, d\widetilde{\alpha}} \\ & \frac{= \pi(\alpha) e^{L_N(\alpha)}}{K_N} \end{aligned} \tag{B.2}$$

For the rest of the proof, recall the dependence of α with respect to h_N, which will be dropped to reduce notation.

Define

$$A_N \equiv \int |h_N|^{\beta} \left| e^{L_N(\alpha)} \pi(\alpha) - (2\pi)^{-d/2} \left| \det \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0) \right|^{1/2} e^{\frac{-1}{2} h' \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0) h} \cdot K_N \right| dh_N$$

Now, I can rewrite equation (B.1) as $A_N K_N^{-1}$, using the definition of $p_\infty^*(h_N)$ and (B.2).

$$\begin{aligned} A_N K_N^{-1} \quad &= \int |h_N|^{\beta} \, |e^{L_N(\alpha)} \pi(\alpha) K_N^{-1} - (2\pi)^{-d/2} \left|\det \triangledown_{\alpha\alpha'} L_N(\alpha_0)\right|^{1/2} \\ &\qquad e^{\frac{-1}{2} h' \triangledown_{\alpha\alpha'} L_N(\alpha_0) h} | dh_N \\ &= \int |h_N|^{\beta} \, | \frac{e^{L_N(\alpha)} \pi(\alpha)}{\int_A \pi(\widetilde{\alpha}) e^{L_N(\widetilde{\alpha})} \, d\widetilde{\alpha}} - (2\pi)^{-d/2} \left|\det \triangledown_{\alpha\alpha'} L_N(\alpha_0)\right|^{1/2} \\ &\qquad e^{\frac{-1}{2} h' \triangledown_{\alpha\alpha'} L_N(\alpha_0) h} | dh_N \\ &= \int (|h_N|^{\beta}) \, |p_N^*(h_N) - p_\infty^*(h_N)| \, dh_N \end{aligned}$$

To show that $K_N = \mathcal{O}_p(1)$, define

$$A_{1N} \equiv \int |h|^{\beta} \left| e^{L_N(\alpha_h)} \pi(\alpha_h) - \exp^{-\frac{1}{2} h' \triangledown_{\alpha\alpha'} L_N(\alpha_0) h} \pi(\alpha_0) \right| dh$$

Chernozhukov and Hong (2003) shows that $A_{1N} \to^p 0$. Thus, for $\beta = 0$,

$$\begin{aligned} \int \left| e^{L_N(\alpha_h)} \pi(\alpha_h) \right| dh \to^p & \int \left| \exp^{-\frac{1}{2} h' \triangledown_{\alpha\alpha'} L_N(\alpha_0) h} \pi(\alpha_0) \right| dh \\ K_N \to^p & \int \left| \exp^{-\frac{1}{2} h' \triangledown_{\alpha\alpha'} L_N(\alpha_0) h} \pi(\alpha_0) \right| dh \\ &= \pi(\alpha_0) (2\pi)^{\frac{m}{2}} \left|\det \triangledown_{\alpha\alpha'} L_N(\alpha_0)\right|^{\frac{-1}{2}} \end{aligned} \tag{B.3}$$

where the last equality follows from the fact that $p_\infty^*(h_N)$ is a well defined probability, and therefore

$$\int (2\pi)^{-m/2} \left|\det \triangledown_{\alpha\alpha'} L_N(\alpha_0)\right|^{1/2} \;\; e^{\frac{-1}{2} h' \triangledown_{\alpha\alpha'} L_N(\alpha_0) h} | dh = 1$$

rearranging

$$\begin{aligned} \int e^{\frac{-1}{2} h' \triangledown_{\alpha\alpha'} L_N(\alpha_0) h} \, dh \qquad &= (2\pi)^{\frac{m}{2}} \left|\det \triangledown_{\alpha\alpha'} L_N(\alpha_0)\right|^{-1/2} \\ K_N \quad &= \pi(\alpha_0) (2\pi)^{\frac{m}{2}} \left|\det \triangledown_{\alpha\alpha'} L_N(\alpha_0)\right|^{-1/2} \end{aligned}$$

Thus, it is enough to show that $A_N \to_p 0$. To do so I write $A_N \to^p A_{1N} + A_{2N}$ where

$$A_{2N} \equiv \int |h|^{\beta} \left| K_N (2\pi)^{-m/2} \left| \det \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0)^{1/2} \right| e^{-\frac{1}{2}h' \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0) h} \right.$$
$$\left. - \pi(\alpha_0) e^{-\frac{1}{2}h' \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0) h} \right| dh$$

Rearranging,

$$A_{2N} = \left| C_N (2\pi)^{-\frac{m}{2}} \left| \det \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0) \right|^{1/2} - \pi(\alpha_0) \right| \int |h^{\beta}| e^{-\frac{1}{2}h' \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0) h} dhb \rightarrow^{p} 0$$

since

$$K_n \rightarrow^{p} \pi(\alpha_0)(2\pi)^{\frac{m}{2}} \left| \det \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0) \right|^{-1/2}$$

B.2 Proof of Corollary 2

To show part (i), I need to show that

$$z_n = \sqrt{N}(\alpha - \alpha_0) + o_p(1) \tag{B.4}$$

Recall $z_N = \epsilon_N + U_N$. $U_N = \mathcal{O}_p(1)$, by assumption 4(ii) and CLT, and $\epsilon_N = \mathcal{O}_p(1)$ (by definition of $p^*_{\infty}(\alpha)$), I get that $z_N = O_p(1)$. citetjureckova-77 shows that this result, plus the uniform convergence and convexity properties in the assumptions, imply $z_n - \sqrt{N}(\hat{\alpha} - \alpha_0) \rightarrow_p 0$, which is equivalent to the desired result.

To show part (ii), use assumption 4(ii). Multiplying and dividing by $\bigtriangledown_{\alpha\alpha'} L_N(\alpha_0)$, and substituting for U_N, I get

$$\Omega^{-\frac{1}{2}} \bigtriangledown_{\alpha\alpha'} L_N(\alpha_0) U_N \sim N(0, \mathcal{I}) \tag{B.5}$$

the desired result.

Solve for U_N in part(i) and substitute into B.5 to get the final part of the corollary.

B.3 Proof of Corollary 3

Define

$$c_{w,N}(d) \equiv \inf\{y : F_{w,N}(y) \geq d\} \quad \text{and} \quad F_{w,N}(\bar{w}) \equiv \int_{\alpha \in A: w(\alpha) \leq \bar{w}} p_N(\alpha) d\alpha$$

The d-quantile of the sequence $(w(\alpha^{(1)}), ..., w(\alpha^{(B)}))$ derived using MCMC, can be written as $[c_{w,N}(d/2), c_{w,N}(1-d/2)]$.

Define $\hat{F}_{w,N}$ as the asymptotic cumulative distribution function and make a change of variables to write α in terms of h_N, as defined in (3.11).

$$\hat{F}_{w,N}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) \equiv \int_{\alpha \in A: w(\alpha) \leq w(\alpha_0) + \frac{s}{\sqrt{N}}} p^*_\infty(h_N) d\alpha$$

Define a similar cumulative function over a modified integral area

$$F_{w,\infty}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) \equiv \int_{\alpha \in A: \bigtriangledown_\alpha w(\alpha - \alpha_0)' \leq \frac{s}{\sqrt{N}}} p^*_\infty(h_N) d\alpha$$

By Corollary 1,

$$\sup_s \left| F_{w,N}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) - \hat{F}_{w,N}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right)\right| \to^p 0 \tag{B.6}$$

Because I are dealing with a normal density $p^*_\infty(\alpha)$, the continuity of the integral of the normal density gives

$$\sup_s \left| \hat{F}_{w,N}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) - F_{w,\infty}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right)\right| \to^p 0$$

which implies with (B.6)

$$\sup_s \left| F_{w,N}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) - F_{w,\infty}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right)\right| \to^p 0 \tag{B.7}$$

Because both distributions converge, their quantiles will converge as well.

$$\left| F^{-1}_{w,N}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) - F^{-1}_{w,\infty}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right)\right| \to^p 0$$

Since a variable α with a density function $p^*_\infty(\alpha)$ is distributed

$$\alpha \sim \mathcal{N}(\alpha_0 - \sqrt{N} \bigtriangledown_{\alpha'\alpha} L_N(\alpha_0)^{-1} \bigtriangledown_\alpha L_N(\alpha_0), -N \bigtriangledown_{\alpha'\alpha} L_N(\alpha_0)^{-1})$$

Thus,

$$F_{w,\infty}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) = P\{\bigtriangledown_\alpha w(\alpha_0)'\mathcal{N}(S_N, N \bigtriangledown_{\alpha'\alpha} L_N(\alpha_0)^{-1} < s|S_N\} \tag{B.8}$$

where $S_N \equiv -\sqrt{N} \bigtriangledown_{\alpha'\alpha} L_N(\alpha_0)^{-1} \bigtriangledown_\alpha L_N(\alpha_0)$. Defining $\bar{F}_{w,N}(s) \equiv F_{w,N}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right)$, and inverting the cumulative function in (B.8) I get the quantile

$$\bar{F}^{-1}_{w,\infty}(d) = \bigtriangledown w(\alpha_0)'S_N + t_d\sqrt{\bigtriangledown_\alpha w(\alpha_0) N \bigtriangledown_{\alpha'\alpha} L_N(\alpha_0)^{-1} \bigtriangledown_\alpha w(\alpha_0)} \tag{B.9}$$

where the d-quantile from a standard normal distribution is used for t_d. Recall $c_{w,\infty}(d) = \bar{F}^{-1}_{w,\infty}(d)$. Then,

$$\begin{aligned} \bar{F}^{-1}_{w,N}(d) = & \sqrt{N}(F^{-1}_{w,N}(d) - w(\alpha_0)) \\ = & \sqrt{N}(c_{w,N}(d) - w(\alpha_0)) \end{aligned}$$

Using (B.7),(B.10) and (B.9),

$$\sqrt{N}(c_{q,N}(d) - w(\alpha_0)) = \bigtriangledown w(\alpha_0)'S_N + t_d\sqrt{\bigtriangledown_\alpha w(\alpha_0) N \bigtriangledown_{\alpha'\alpha} L_N(\alpha_0)^{-1} \bigtriangledown_\alpha w(\alpha_0)} + o_p(1) \tag{B.10}$$

Rearranging,

$$c_{q,N}(d) - w(\alpha_0) - \frac{\bigtriangledown w(\alpha_0)'S_N}{\sqrt{N}} - t_d\frac{\sqrt{\bigtriangledown_\alpha w(\alpha_0) N \bigtriangledown_{\alpha'\alpha} L_N(\alpha_0)^{-1} \bigtriangledown_\alpha w(\alpha_0)}}{\sqrt{N}} = o_p(1) \tag{B.11}$$

Recall the Taylor approximation gives $\sqrt{N}(\hat{\alpha} - \alpha_0) = S_N + \mathcal{O}(\frac{1}{\sqrt{N}})$.

A similar approximation for $w(\alpha)$ gives

$$\begin{aligned} w(\hat{\alpha}) = & \quad w(\alpha_0) + \bigtriangledown w(\alpha)(\hat{\alpha} - \alpha_0) + \mathcal{O}\left(\tfrac{1}{\sqrt{N}}\right) \\ = & \quad w(\alpha_0) + \bigtriangledown w(\alpha)\tfrac{S_N}{\sqrt{N}} + \mathcal{O}\left(\tfrac{1}{\sqrt{N}}\right) \end{aligned}$$

Using (B.12) to substitute in for the second and third term of (B.11), I get

$$c_{q,N}(d) - w(\hat{\alpha}) - t_d \frac{\sqrt{\nabla_\alpha w(\alpha_0) N \nabla_{\alpha'\alpha} L_N(\alpha_0)^{-1} \nabla_\alpha w(\alpha_0)}}{\sqrt{N}} = o_p(1)$$

which is the desired result. Part b follows from the fact that both quantiles are equivalent asymptotically, and thus $c_{q,N}(d)$ gives valid asymptotic probabilities for $w(\alpha_0)$.

C Solution to the Hall and Horowitz (1996) Model

Section 4.1 presents the model used in simulations as presented in Hall and Horowitz (1996). One of the reasons supporting the selection of this model for our simulation is that it usually presents multiple modes in small samples. This feature is common in models used in empirical work. Here I show analytically why these multiple modes occur for this particular model, even when the true population-parameter value is unique and therefore no identification issues exist asymptotically.

Recall X_i and Z_i are scalar observations distributed iid $\sim N(0, s^2)$. μ is a normalization constant set to $\mu \equiv -\frac{\theta_0^2 s^2}{2}$ to make the moment conditions zero at the population-parameter values.

The moment conditions are

$$E[G(\theta)] = \begin{bmatrix} \exp(\mu + \theta(X+Z) - \theta_0 Z) - 1 \\ Z(\exp(\mu + \theta(X+Z) - \theta_0 Z) - 1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{C.1}$$

In small samples, multiple minima occur because the moment condition in the first row does not uniquely identify $\theta_0 = 3$, but is also satisfied by $\theta = 0$. The second moment condition is only satisfied by $\theta = \theta_0$. When the optimal weighting matrix gives enough weight to the first moment condition in the estimation, then the sample GMM objective function Q_N (2.1), the sample transformed objective function L_N in (2.6), and the quasiposterior density (3.1) can present a local minimum (local mode) at $\theta = 0$.

Consider the analytical solution to the first moment condition:

$$\begin{aligned} E[g(\theta)] = \quad E[\exp(\mu + \theta(X+Z) - \theta_0 Z) - 1] &= \quad 0 \\ E[\exp(\mu + \theta X + (\theta - \theta_0) Z)] &= \quad 1. \end{aligned}$$

Given the data distribution, $\theta X + Z(\theta - \theta_0) \equiv \gamma \sim \mathcal{N}(0, \theta s^2 + (\theta - \theta_0)s^2)$. For a variable $w \sim \mathcal{N}(0, I)$, and a constant s,

$$\begin{aligned} E(e^{sw}) = &\quad \textstyle\int_{-\infty}^{\infty} \tfrac{1}{2\pi} e^{sw} e^{-\frac{w^2}{2}} dw = &\quad \displaystyle\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{\frac{1}{2}(2sw - w^2 + s^2 - s^2)} dw \\ = &\quad \textstyle\int_{-\infty}^{\infty} \tfrac{1}{2\pi} e^{\frac{1}{2}(-(w-s)^2 + s^2)} dw = &\quad \displaystyle\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{1}{2}(w-s)^2} e^{\frac{1}{2}s^2} dw \\ = &\quad \exp\left(\tfrac{s^2}{2}\right). \end{aligned}$$

For $y = w\sigma + n \sim \mathcal{N}(n, \sigma^2)$, expectation becomes

$$E(e^{sy}) = E(e^{\sigma w + n}) = e^n E(e^{\sigma w}) = e^n e^{-\frac{\sigma^2}{2}}. \tag{C.2}$$

Using this notation, the moment condition is satisfied when $E(e^{\mu} e^{\gamma}) = 1$. Thus,

$$e^{\mu} e^{\frac{\theta s^2 + s^2(\theta - \theta_0)^2}{2}} = 1$$

for $s = 1, n = 0$ and $\sigma^2 \theta^2 s^2 + s^2(\theta - \theta_0)^2$, which implies

$$\begin{aligned} \mu = &\quad \tfrac{-\theta^2 s^2 - s^2(\theta - \theta_0)^2}{2} \\ -\frac{\theta_0 s^2}{2} = &\quad \tfrac{-\theta^2 s^2 - s^2(\theta - \theta_0)^2}{2} \\ \theta_0^2 = &\quad -\theta^2 - \theta^2 + 2\theta\theta_0 - \theta_0^2, \end{aligned}$$

which is solved by $\theta = 0$, and $\theta = \theta_0$ as discussed.

Consider the analytical solution to the second moment condition

$$\begin{aligned} E[g(\theta)] &= E[Z(\exp(\mu + \theta(X + Z) - \theta_0 Z) - 1)] &= 0 \\ E[Z\exp(\mu + \theta X - (\theta_0 - \theta)Z)] & &= E[Z] \\ E[Z\exp((\theta - \theta_0)Z)]E[\exp(-\tfrac{\theta_0^2 s^2}{2} + \theta X)] & &= 0 \end{aligned}$$

Because $-\frac{\theta_0^2 s^2}{2} + \theta X \sim \mathcal{N}(-\frac{\theta_0^2 s^2}{2}, \theta^2 s^2)$, then $E[e^{-\frac{\theta_0^2 s^2}{2} + \theta X}] = e^{-\frac{\theta_0^2 s^2}{2}} e^{-\frac{\theta^2 s^2}{2}}$.

Thus,

$$e^{-\frac{\theta_0^2 s^2}{2}} e^{-\frac{\theta^2 s^2}{2}} E[Z \exp((\theta - \theta_0)Z)] = e^{-\frac{\theta_0^2 s^2}{2}} e^{-\frac{\theta^2 s^2}{2}} \int_{-\infty}^{\infty} Z e^{(\theta-\theta_0)Z} \frac{1}{\sqrt{2\pi}} e^{-\frac{(Z)^2}{2s^2}} dz = 0$$

$$\begin{aligned} e^{-\frac{(\theta_0^2 s^2+\theta^2 s^2)}{2}} \int_{-\infty}^{\infty} \frac{Z}{\sqrt{2\pi}} e^{\frac{2s^2(\theta-\theta_0)Z-Z^2}{2s^2}} dz \quad &= e^{-\frac{(\theta_0^2 s^2+\theta^2 s^2)}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} Z e^{\frac{-(Z-s^2(\theta_0-\theta))^2-s^4(\theta_0-\theta)^2}{2s^2}} dz \\ &= e^{-\frac{(\theta_0^2 s^2+\theta^2 s^2)}{2}+\frac{s^4(\theta_0-\theta)^2}{2s^2}} \end{aligned}$$

which is only solved by $\theta = \theta_0$.