MCMC Approach to Classical Estimation with Overidentifying Restrictions

Luis Eduardo Quintero *

April 23, 2009
Preliminary Draft

Abstract

This paper extends the Laplace estimators proposed by Chernozhukov and Hong (2003) to incorporate the statistic that tests the overidentifying restrictions in the GMM. This information was previously ignored during parameter estimation in econometrics with Bayesian methods. The parameters and the test statistic are estimated simultaneously using information in the entire domain of the estimation equations, not at the global minimum only. We avoid the curse of dimensionality by using MCMC, following Chernozhukov and Hong (2003). Multivariate kernel density estimation gives a smooth distribution of the parameter values that are a solution to the optimization in Laplace estimation. The transformed estimators perform better in a simulation exercise than those version that do not use the information in the OR during parameter estimation. Furthermore, the kernel density also allows for the calculation of alternative estimators that condition the estimation on the OR being satisfied. In the presence of multiple solutions of the GMM objective function, conditioning on the OR brings economic theory as a criteria for estimate selection. As a consequence, our estimators perform better than their unconditional counterparts in a simulation exercise. We simulate a model in Hall and Horowitz (1996) that frequently presents multiple local solutions.

KEYWORDS: Overidentifying restrictions; Markov Chain Monte Carlo; Generalized Method of Moments; Laplace multiple minima.

*Contact: leq@cmu.edu
1 Introduction

Bayesian techniques can be used under a classical estimation framework to solve computational challenges of extremum estimation in econometrics. Chernozhukov and Hong (2003) use Laplace type estimators (LTEs) or quasi-Bayesian estimators (QBEs) to provide consistent parameter estimation. These estimators are like Bayesian estimators, cf. Allenby et al. (2005), but they use general criterion functions instead of parametric likelihood functions. They are particularly useful because they avoid the curse of dimensionality to which extremum or m-estimation is generally subject. The LTEs avoid it by using simulation techniques, like Markov Chain Monte Carlo (MCMC). MCMC is used to approximate integral transformations of the criterion function and estimate 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 several econometric models. It concludes that these estimators perform well when compared to standard methods, while circumventing the curse of dimensionality.

In the Generalized Method of Moments (GMM) estimation, economic theory dictates the moments whose expectation is zero for the population parameter values. Hansen (1982) indicates that in the process of optimization in GMM estimation, there are only $k$ linear combinations or dimensions being set to zero, where $k$ is the number of parameters. For an over-identified system ($m > k$, where $m$ is the number of moments), the remaining $m - k$
linearly independent combinations in the objective function will be close to zero in probability when the model or theory is true. These $m - k$ combinations determine the restrictions implied by the model, or the overidentifying restrictions. Sowell (2007) discusses problems that may arise if parameters values are estimated independently from overidentifying restrictions. In particular, it stresses the role that the overidentifying restrictions may play in selecting between different local minima of the GMM objective function.

Overidentifying restrictions can be incorporated in the computation of the LTEs in overidentified systems by decomposing the $m$ dimensions in the moments into the identifying space and the overidentifying space, cf. Sowell (1996), and using both to construct a transformed criterion function for a new just identified system. This method inherits the computational advantages of LTEs, allowing for MCMC methods that avoid the curse of dimensionality, while also allowing for an easy application in non-smooth criterion functions\(^1\). This approach will simultaneously estimate the parameters and the test statistic for overidentifying restrictions. Furthermore, it can be used to estimate alternative estimators that use economic theory to select between estimates in the presence of multiple local minima. Most importantly, we are using information in the overidentifying restrictions that other estimation methods neglect and that is available from in the moment conditions.

\(^1\)In general, it is a common in structural estimation to find criterion functions that are nonsmooth and highly nonconvex and have numerous local optima, making extremum estimation difficult, cf. Chernozhukov and Hong (2003).
means and modes that condition on the overidentifying restrictions being satisfied, perform better than other quasiposterior or Laplace estimates. They also sometimes outperform two-step GMM estimation, a method that occasionally encounters the curse of dimensionality. In the presence of multiple minima, some of which occur around values that do not correspond to the population parameter value, the conditional estimators tend to select the true population parameter value by focusing on solutions in which the overidentifying restrictions are satisfied. Therefore, they show lower root mean square errors (RMSEs) and variances.

Section 2 reviews the importance of the overidentifying restrictions and shows how to transform the criterion function to include both the identifying and overidentifying spaces. Section 3 defines and explains LTEs. MCMC sample data generation and kernel density estimation are explained and 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 Overidentifying restrictions

Economic theory implies $m$ moments whose expectation is zero. When the number of parameters $k$ is less than the number of moments $m$, then $m - k$ linear combinations, or dimensions, are not set to zero by the FOCs (when
minimizing the GMM objective function with respect to the $k$ parameters). However, if our model or theory is true, that is if the expectation of the moments is equal to zero when evaluated at the true parameter value, it restricts these $m - k$ dimensions to also be zero on average. These $m - k$ restrictions are the overidentifying restrictions. If they are indeed zero on average when evaluated at our estimate and the data sample observed, we say our theory cannot be rejected for that estimated parameter value.

Define the $m$ moments $G_N(\theta)$ by

$$G_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} g(x_i, \theta) = \frac{1}{N} \sum_{i=1}^{N} g_i(\theta) \quad (2.1)$$

where $N$ is the sample size, $\theta$ are the $k$ parameters to be estimated, $x_i$ is the data sample, and $g_i(\theta)$ is a moment function given by theory, such that the population parameter value $\theta_0$ solves $E g_i(\theta_0) = 0$. It is assumed that $G_N$ satisfies a CLT at $\theta_0$, such that $\sqrt{N}G_N \sim N(0, \Sigma_g)$. The GMM estimator $\hat{\theta}$ is defined as

$$\hat{\theta} = \arg \min_{\theta \in \Theta} Q_N(\theta) \quad (2.2)$$

where

$$Q_N(\theta) = \frac{1}{2} G_N(\theta)' W_N G_N(\theta) \quad (2.3)$$

$W_N$ is the optimal weighting matrix, obtained through any consistent estimate of $\Sigma_g^{-1}$. In two-step GMM, the covariance matrix between the moments evaluated at a first round estimate $\theta^{(1)}$ is used, as described by Hansen (1982).
The estimated parameter values set the following FOCs to zero with respect to \( \hat{\theta} \)

\[
\tilde{M}_N(\hat{\theta})'W_NG_N(\hat{\theta}) = 0_{k \times 1}
\]  

(2.4)

where \( \tilde{M}_N(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial g(x_i, \hat{\theta})}{\partial \theta} \).

Thus, the FOCs in (2.4) can be written using \( \overline{m}_N \)

\[
\overline{m}_N(\hat{\theta})'W_N^{\frac{1}{2}}G_N(\hat{\theta}) = 0_{k \times 1}
\]  

(2.5)

where \( \overline{m}_N \equiv W_N^{\frac{1}{2}}M_N \) is the derivative of \( G_N \) standardized by the Cholesky decomposition of the weighting matrix. The columns of \( \overline{m}_N \) define a basis for the subspace of dimension \( k \) spanned by the FOCs with respect to the \( \theta \) parameter. The projection matrix

\[
P_{\overline{m}_N}(\theta) \equiv \overline{m}_N(\theta)(\overline{m}_N(\theta)'\overline{m}_N(\theta))^{-1}\overline{m}_N(\theta)'
\]  

(2.6)

is a real symmetric positive semidefinite matrix, and thus can be written as a spectral decomposition

\[
P_{\overline{m}_N}(\theta) = C_N(\theta)\Lambda C_N(\theta)'
\]  

(2.7)

This matrix gives the vector space projection from \( \mathbb{R}^m \) to the \( k \)-dimensional subspace of \( \overline{m}_N \). \( C_N \) columns determine an orthonormal basis for \( \overline{m}_N \), and \( \Lambda \) is a diagonal matrix with the eigenvalues of \( P_{\overline{m}_N}(\theta) \). 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\(^2\). Notice that any linear combination of the FOCs will also be set to zero by the GMM estimate \(\hat{\theta}\). Thus, we can write the identifying restrictions as the linear combinations of the FOCs in (2.5) that use the orthogonal basis vectors in \(C_{1N}\)

\[
C_{1N}(\hat{\theta})'W_\hat{\theta}^{\frac{1}{2}}G_N(\hat{\theta}) = 0
\]  

(2.8)

Alternatively, let \(C_{2N}\) be the column vectors in \(C_N\) whose columns span the space orthogonal to \(\overline{m}_N\) and therefore orthogonal to that spanned by the FOCs. We can create equations that use the orthogonal basis vectors in \(C_{2N}\) to span the \(m - k\) remaining orthogonal dimensions.

\[
C_{2N}(\theta)'W_\theta^{\frac{1}{2}}G_N(\theta)(m-k \times 1)
\]  

(2.9)

These \(m - k\) dimensions or linear combinations form the space of overidentifying restrictions.

Consider new \(m - k\) parameters denoted by \(\lambda\) such that \(C_{2N}(\hat{\theta})'W_\hat{\theta}^{\frac{1}{2}}G_n(\hat{\theta}) = \hat{\lambda}\). The population parameter value that sets the expectation of these equations to zero is \(\lambda_0 = 0\). Also, under the null, \(\sqrt{N}\lambda_0 \sim N(0, I)\). Hence, \(N\hat{\lambda}\hat{\lambda}\) is equivalent to the J-test statistic in Hansen (1982). The overidentifying re-

\(^2\)Since a projection matrix is idempotent, its eigenvalues are either 1 or 0. \(C_{1N}\) contains the vectors that are multiplied by an eigenvalue equal to 1 in (2.7), and not by a eigenvalue equal to 0. Notice that \(P_\theta C_{1N} = C_{1N}C'_{1N}\). In a similar fashion, we have for the projection matrix onto the orthogonal space \(P_{m_N} = I - P_\theta = C_{2N}C'_{2N}\).
restrictions can be tested as the hypothesis \( H_0 : \lambda = 0 \). Defining \( \alpha \equiv (\theta' \lambda)' \),
we can write new moments with expectation 0 that include the identifying
restrictions in (2.8) and the overidentifying restrictions in (2.9) for our \( m \)
original moments \( G_N(\theta) \)

\[
\Psi_N(\alpha) = \begin{bmatrix}
\left( C_1(\theta)'W_N^{\frac{1}{2}}G_N(\theta) \right)_{k \times 1} \\
\left( \lambda - C_2(\theta)'W_N^{\frac{1}{2}}G_N(\theta) \right)_{(m-k) \times 1}
\end{bmatrix}
\] (2.10)

The first \( k \) equations locally vary with the parameter values and span the
space of the FOCs. The next \( m-k \) equations are locally orthogonal to \( \theta \),
i.e. for each value of \( \theta \), \( \lambda \) spans the space that is the orthogonal complement
of the space spanned by \( \theta \). Using the estimate \( \hat{\theta} \) from (2.8) we can use the
latter equations to get the estimate \( \hat{\lambda} \) that sets them to zero.

As in GMM, we want the estimates \( \hat{\alpha} \) to make the objective \( \Psi \) in (2.10)
equal to zero. Intuitively, we want the identifying restrictions (2.8) to be zero
so the \( k \) FOCs are zero when minimizing with respect to the parameters.
Also, under the null, we want the overidentifying restrictions in (2.9) to be
zero because economics theory suggests all \( m \) moments equal zero. However,
so far we have only set \( k \) of those \( m \) dimensions equal to zero through the
FOCs with respect to \( \theta \). The estimation of \( \lambda \) sets the remaining \( m-k \)
combinations so that the moments equal zero on average.

Since \( \Psi_N(\alpha) \) determines a just identified system- \( m \) parameters in \( m \)
moments- whose expectation is zero, we can define a new objective function
$L_N(\alpha)$ that does not require a weighting matrix to give efficient estimates

$$L_N(\alpha) = \frac{1}{2} \Psi_N(\alpha)' \Psi_N(\alpha) \quad (2.11)$$

The estimator we describe in section 3.1 estimates both the parameters $\theta$ and the statistic for the overidentifying restrictions $\lambda$ simultaneously. Also, as an alternative estimator, we propose the overidentifying restrictions as a criteria to select between multiple local minima in finite samples. Sowell (2007) stresses that multiple solutions to the FOCs of GMM objective functions usually exist, and that these can change with the sample size, even if identification for GMM implies there will only be one minimum asymptotically. This is something that GMM as a method ignores, focusing only on global minima. Nevertheless, multiple minima is a recurrent phenomenon in GMM empirical work, cf. Eichenbaum (1989), Stock, Wright, and Yogo (2000), Dominguez and Lobato (2004), Imbens, Spady, and Johnson (1998). Phillips (1989) points out that most of empirical work is performed under conditions of apparent identification. That is, estimation and testing is carried out as if the criterion function had already converged to a non-random function with a unique minimum as asymptotic theory predicts, even when this is evidently not true for small sample sizes often used in empirical work. Introducing the overidentifying restrictions dimensions in the criterion function allows us to use Bayesian methods to create a joint distribution for both the parameter $\theta$ and the overidentifying restriction test statistic $\lambda$, and estimate them si-
multaneously. Introducing the information in the overidentifying restrictions space in the estimation improves the probability of selecting the actual true population parameter values, both when we explicitly use $\lambda$ as a criteria to select between local minima and when we do not.

3 Laplace or Quasi-Bayesian estimation in a Classical Framework

Bayesian statistics interprets population true parameters as a random variable whose distribution is estimated instead of as a constant value. Classical statistics produces probabilities that can be extended to values not observed in the data (i.e. p-values assign probabilities to any value located further in the tail than the critical value, even if it did not occur in the observed realization). In contrast, Bayesian likelihood functions take the data that have already been observed, and estimate parameter values that are considered valid only to fit that observed data. Despite many differences, Bayesian techniques can be used in a classical framework, cf. Chernozhukov and Hong (2003).

3.1 Laplace estimators

Bayesian estimators generally use parametric likelihood functions. Laplace Type Estimators (LTE) replace these functions with general statistical crite-
rion functions\textsuperscript{3}. These criterion functions are used to get a posterior density that determines the probability of selecting different parameter values.

The prior is the belief over the phenomenon of interest, usually stated as values of moments or as a distribution. The posterior probability is an updated version of those beliefs after data is observed. We are interested in the probability of the our estimates taking certain values, so our posterior probability is defined over $\alpha \in A$, the space of parameter values. Given a prior probability distribution $\pi(\alpha)$, the posterior probability distribution can be calculated using Bayes theorem with a likelihood function $f$, and a normalizing constant $P(Y)$,

$$p_N(\alpha \mid Y = y) = \frac{p(\alpha, Y)}{P(Y)} = \frac{\pi(\alpha)f_N(Y \mid \alpha)}{P(Y)} = \frac{\pi(\alpha)f_N(Y \mid \alpha)}{\int_A \pi(\tilde{\alpha})f_N(Y \mid \tilde{\alpha}) d\tilde{\alpha}} \quad (3.1)$$

where $Y$ is the data observed, and $\tilde{\alpha}$ is an integration dummy for $\alpha$. Sometimes the denominator, which is a function of the data, is hard to calculate or just unknown. However, for some methods, including MCMC, we only need to known the posterior density up to a constant, which we call $q(\alpha)$. For this method we are interested in calculating ratios in which this constant cancels out

$$p_N(\alpha \mid Y = y) \propto q(\alpha \mid Y = y) = \pi(\alpha)f_N(Y \mid \alpha). \quad (3.2)$$

\textsuperscript{3}This type of estimator is what we referred previously to as Quasi-Bayesian estimator. Like Chernozhukov and Hong (2003), we use the term Laplace all along to avoid confusion with other definitions of Quasi-Bayesian.
Instead of a likelihood function, our work will use a criterion function $L_N$. Extremum estimators are typically defined as the argmin of a criterion function $L_N$ that is uniquely minimized in the limit by the population value $\alpha_0$. We have discussed that other local minima may exist. Examples of these functions can be (2.3) and (2.11). This approach is generalizable to other objective functions as described in Chernozhukov and Hong (2003). In particular we use the proposed function $L_N(\alpha)$ in (2.11), constructed from the GMM objective function in (2.3). This criterion function includes information on both the identifying and overidentifying restrictions of the GMM original model. Using an exponential transformation and following (3.1), we write the posterior probability as

$$p_N(\alpha \mid Y = y) = \frac{\pi(\alpha)e^{L_N(\alpha)}}{\int_A \pi(\tilde{\alpha})e^{L_N(\tilde{\alpha})} d\tilde{\alpha}}$$ (3.3)

A risk function is the expected loss or error in which the researcher incurs when choosing a certain value for the parameter estimate. Let $\rho_N(\alpha, \tilde{\alpha}) : \mathbb{R}^m \rightarrow \mathbb{R}$ be the loss function associated with selecting $\tilde{\alpha}$, when the value of the parameter is $\alpha$. The loss function can penalize the selection of $\alpha$ asymmetrically, and is a function of the selected value and the rest of the possible values of the parameters in $A$. The risk function takes the form

$$R_N(\tilde{\alpha}) = E[\rho_N(\alpha, \tilde{\alpha}) \mid Y] = \int_{\Theta} \rho_N(\alpha, \tilde{\alpha})p_N(\alpha \mid Y = y) d\alpha$$ (3.4)

where $p_N(\alpha \mid Y = y)$ is the posterior probability in (3.3), $\tilde{\alpha}$ is the selected
value, and \( \alpha \) is all other possible values we are integrating over. The Laplace estimator \( \hat{\alpha}_{LP} \) minimizes the expected loss for different forms of the loss function

\[
\hat{\alpha}_{LP} = \arg \inf_{\alpha \in A} R_N(\alpha) \tag{3.5}
\]

Choosing different loss functions will 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

\[
\hat{\alpha}_{qpm} = \int_A \alpha p_N(\alpha) \, d\alpha \tag{3.6}
\]

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

### 3.2 MCMC and Sample Generation

In practice, when the model has a large number of moments \( m \), a large number of overidentifying restrictions \( m - k \), or a large dimension \( k \) in the \( \theta \) parameter, the curse of dimensionality comes into play, and probability densities become intractable. Therefore the integrals necessary to calculate the quasi-posterior means, median, and quantiles, and the posterior modes become difficult. That is the main reason justifying the use of the MCMC procedure under a classical framework.

Markov Chain Monte Carlo (MCMC) methods can be used to compute Laplace estimators. MCMC methods generate samples from an initial distri-
bution, cf. Gelman et al (2004), and then correct those draws to approximate a target distribution. In our case, the result of the method is a sample of parameter values $\alpha$ that behave as if they had been generated by the posterior $p_N(\alpha)$, the target distribution. The name comes from the generation of random numbers (Monte Carlo) that follow a proposed distribution whose conditional probability for a future value only depends on the current draw (Markov Chain). This method circumvents the curse of dimensionality, which is usually found in extremum estimation, cf. Jaquier et al (2005), and is computationally efficient, cf. Jaquier et al. (1994). Also, when we use a posterior such as (3.3), this method does not require a criterion function $L_N(\alpha)$ that is smooth or that has a unique minimum.

MCMC is generally used to numerically calculate integrals. For instance, if our estimator is the quasi-posterior mean in (3.6), we can approximate it by generating a sample $\alpha^T = \{\alpha^t\}_{t=1}^T$ through MCMC, and calculating the sample mean to get

$$\hat{\alpha}_{qpm} = \int_{\Theta} \alpha p_N(\alpha) d\alpha \approx \frac{1}{T} \sum_{t=1}^T \alpha^t$$

This method has been shown to quickly converge to the integral evaluation in the ergodicity result of Proposition 1 in Chib and Greenberg (1996), which we 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 \( |P^m(\alpha^*, |\alpha) - p(\alpha) |\) \(\Rightarrow\) 0 as \( m \Rightarrow \infty \).

This proposition is a property of Markov chains. It tells us that the probability density of the \( m^{th} \) iterate of the Markov chain converges to its invariant density. If a sample is drawn from the distribution \( P \), then for a large number of iterations \( m \), the distribution of this sample will be equal to the probability kernel \( p \). 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). MCMC Metropolis-Hastings Algorithm uses this result to reproduce a transition kernel that satisfies Proposition 1.

3.2.1 Metropolis-Hastings Algorithm

Different algorithms fall into the classification of MCMC. The Gibbs Sampler and the Metropolis-Hastings Algorithms are widely used examples. The Gibbs algorithm is less general than the Metropolis-Hastings algorithm, and it has in fact been shown to be a special case of it, cf. 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). Let \( J_t(\alpha^t | \alpha^{t-1}) \) be a proposal distribution at time \( t \), and \( p(\alpha | y) \) be the target distribution. The algorithm proceeds as follows:

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

The starting point can be chosen using a crude estimate, i.e the result of a regression.
Step-2 For $t=1,2...$

- a) Sample a proposal $\alpha^*$ 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 (3.8). 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.

\[
r = \min \left(1, \frac{p(\alpha^* \mid y) / J_t(\alpha^* \mid \alpha^{t-1})}{p(\alpha^{t-1} \mid y) / J_t(\alpha^{t-1} \mid \alpha^*)} \right)
\]  
 (3.8)

- c) To choose the next iteration value $\alpha^t$ assign

\[
\alpha^t = \begin{cases} 
\alpha^* & \text{with probability } r \\
\alpha^{t-1} & \text{with probability } 1-r
\end{cases}
\]  
 (3.9)

In this step, we will choose with certainty the proposed $\alpha^*$ if it improves the density of $\alpha$, after weighting for asymmetries in the proposal distribution. If it does not improve it, we might select it with a probability that depends on how much it has worsen it.

Step-3 Finish after some criterion of convergence, i.e. $|\alpha^t - \alpha^{t-1}| < \epsilon$, is satisfied or after a predetermined number of iterations.
A very convenient characteristic of this algorithm is that the target distribution posterior density needs to be known only up to a constant, so we can use \( q(\alpha | Y) \) as described in (3.2).

It is necessary to pick a proposal distribution \( J(\alpha^t | \alpha^{t-1}) \) that can be evaluated for all \( \alpha^* \), and a space of parameter values \( A \) for which the ratio \( r \) can be calculated. For computation purposes, this proposal distribution is usually one for which a random variable generator exists. Chernozhukov and Hong (2003) state that the canonical implementation of the M-H algorithm is to take proposal distributions that are symmetric around 0, such as the Gaussian or the Cauchy density.

Chib and Greenberg (1996) gives conditions for the Metropolis-Hastings Algorithm to satisfy Proposition 1. Let \( p_{MH}(\alpha) \) be the probability distribution of the sample \( \{\alpha^t\}_{t=1}^T \) generated using the Metropolis-Hastings algorithm. If \( J(\alpha^t | \alpha^{t-1}) \) and \( p(\alpha) \) are positive and continuous for all \( \alpha \in A \) then \( p_{MH}(\alpha) \) satisfies Proposition 1. In other words, the sample generated by the Metropolis-Algorithm will be a Markov transition kernel that converges to an invariant distribution equal to the target distribution \( p(\alpha | y) \). Tierney (1994), Chib and Greenberg (1995) and Smith and Roberts (1993) further discuss the convergence of this algorithm. See Gelman (2004) for a discussion about ways to improve convergence speed.

Using the posterior in (3.3) and cancelling the denominator, \( r \) becomes

\[
r = \min \left( 1, \frac{\pi(\alpha^*) e^{LN(\alpha^*)} / J_t(\alpha^* | \alpha^{t-1})}{\pi(\alpha^{t-1}) e^{LN(\alpha^{t-1})} / J_t(\alpha^{t-1} | \alpha^*)} \right)
\]

(3.10)
with $\alpha \equiv (\alpha' \lambda')'$, and $L_N(\alpha)$ as in equation (2.11).

Carrying out the algorithm using functions of $\alpha$ allows us to generate both a sample for parameter values $\theta$ and for the estimator of the overidentifying restriction test $\lambda$. In practice, to generate both samples we can follow an alternating conditional sampling. The only change in the algorithm consists in alternating the parameter for which we generate the proposal value. In iteration $t$, we follow steps 1-3, generating a proposal $\theta^*$, and determining the value for $\theta^t$, for a given $\lambda^t$. For iteration $t + 1$, we generate a proposal $\lambda^*$, and determine $\lambda^{t+1}$, fixing $\theta$ at $\theta^t$ for this iteration. This alternating procedure is used equivalently to generate multiple parameters $\theta_1,\ldots,\theta_k$ when $k > 1$, and the vector $\lambda$ when $m - k > 1$.

### 3.2.2 Kernel density estimation

The calculation of marginal and conditional densities and the evaluation at arbitrary values of the parameters, i.e. $\lambda = 0$, is not available directly through calculations on the sample generated by the MCMC procedure, because the probability measure for any specific value to be found in this sample is 0 (the distribution generated is discrete). Kernel density approximation calculates smooth densities from a sample of data, such as the one generated by MCMC for $\lambda$ and $\theta$. Kernel density estimation is a nonparametric technique that extrapolates the sample data to a density for the whole population. Given a
sample $\alpha^1, \alpha^2, \ldots, \alpha^B \sim p(\alpha)$, the kernel density approximation is given by

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

(3.11)

where $K$ is a probability kernel, and $h$ is a bandwidth or smoothing parameter. The canonical implementation uses a standard normal density\(^4\) probability kernel, cf. Silverman (1986), Jones and Henderson (2007a) and (2007b). The bandwidth depends on the sample size. The estimator does not group observations in bins but places bumps at each observation determined by the kernel function. Thus, the resulting density gives a smoother plot than a histogram.

For the estimates that condition on $\lambda = 0$, we use multivariate kernel density estimation to get the joint distribution, univariate kernel density estimation to get the marginal density for $\lambda$ and the probability $p(\lambda = 0)$, and then Bayes theorem to get the conditional density $p(\theta | \lambda = 0)$. Compared to previous methods, this is possible only because the proposed parametrization in equation (2.10) allows us to generate sample for both $\lambda$ and $\theta$ through MCMC. The estimated densities generate weights to calculate different Laplace estimators, approximating the integrals with sums as in equation (3.7).

As explained in Section 2, the overidentifying restrictions can be a criteria to select between local minima. The estimates that are based on densities

\[K(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\]
conditional on \( \lambda = 0 \) are on average selecting parameter values that satisfy economic theory (for which overidentifying restrictions are satisfied and for which the moments sample averages are zero).

### 3.3 Consistency of LTEs

Chernozhukov and Hong (2003) show LTEs are consistent and asymptotically normal for general criterion functions \( L_N(\theta) \). Here, we explore a similar approach applied to our transformed objective function \( \Psi(\theta, \lambda) \) in (2.11). We establish corollaries of the theorems in Chernozhukov and Hong (2003) that apply to our specific case. First, we state the assumptions. Then we show that the quasi-posterior density concentrates around the true population value \( \alpha_0 \), and that it converges to it at the rate \( \frac{1}{\sqrt{N}} \). Finally, we show that the estimator is consistent and asymptotically normal in Corollary 2.

#### 3.3.1 Assumptions

1. **Compactness.** The extended parameter space \( A \) is a compact subset of the Euclidian space \( \mathbb{R}^m \), where \( m \) is the number of moments.

2. **Loss function properties.** The loss function \( \rho_N(\tilde{\alpha}) \) satisfies:

   (i) Convergence: \( \rho_N(\alpha, \tilde{\alpha}) = \rho(\sqrt{N}(\alpha - \tilde{\alpha})) \), where \( \rho_N(\alpha, \tilde{\alpha}) \geq 0 \) and \( \rho_N(\alpha, \tilde{\alpha}) = 0 \) iff \( \alpha = \tilde{\alpha} \).

   (ii) Boundedness and Convexity: \( \rho \) is convex and \( \rho_N(\alpha, \tilde{\alpha}) \leq 1 + |\alpha - \tilde{\alpha}|^s \) for some \( s \geq 1 \).
(iii) Identification: \( \phi(\xi) = \int_{\mathbb{R}^m} \rho(u - \xi)e^{u'J_n}du \) is minimized uniquely at some \( \xi^* \in \mathbb{R}^m \) for a finite \( J > 0 \). Here, \( u \) is the error incurred in when selecting a \( \tilde{\alpha} \).

(iv) Prior: \( \pi \) is a continuous and uniformly positive density.

3. Identification.

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

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

This assumption is implied by the usual uniform convergence. It says that the expanded objective function \( L_N(\alpha) \) converges uniformly in-probability in \( \alpha \in A \) as \( N \to \infty \) to a function that attains a unique global minimum at \( \alpha_0 \), the population parameter value.

(ii) Asymptotic Convergence 1: \( \exists \) a nonstochastic function \( M_N(\alpha) \) that is continuous on \( A \) with a unique sup on \( A \) at \( \alpha_0 \), 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) \( L_N(\alpha)/N \to_p M_N(\alpha) \).

(c) \( L_N(\alpha) \) is continuously differentiable. \( \nabla_{\alpha'\alpha} L_N(\alpha_0) = \mathcal{O}(1) \), and

\[
\lim_{N \to \infty} \sup P \left\{ \sup_{|\alpha - \alpha_0| \geq \delta} \frac{\nabla_{\alpha'\alpha} L_N}{N} - \nabla_{\alpha'\alpha} M_N(\alpha_0) \right\} > 0 \]

4. Convergence of expansion. For \( \alpha \) in an open neighborhood of \( \alpha_0 \),
Taylor Expansion:

\[ L_N(\alpha) - L_N(\alpha_0) = (\alpha - \alpha_0)\nabla_\alpha L_N(\alpha_0) - \frac{1}{2} (\alpha - \alpha_0)^2 \nabla_{\alpha\alpha} L_N(\alpha - \alpha_0) + R_N(\alpha) \]

Asymptotic normality of the FOCs: \( \exists \) a positive definite constant matrix \( \Omega_N(\alpha_0) \) for all \( N \) such that

\[ \Omega_N^{-\frac{1}{2}} (\alpha_0) \nabla_\alpha L_N(\alpha_0) / \sqrt{N} \to_d \mathcal{N}(0, \mathcal{I}). \]

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)

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

3.3.2 Corollaries

Results in this chapter are corollaries of the theorems in Chernozhukov and Hong (2003). Corollary 1 shows that the quasiposterior density converges around \( \alpha_0 \) at a rate of \( \frac{1}{\sqrt{N}} \). This corollary suggests the consistency of an estimator that is based on a statistic derived from the quasiposterior density. The norm used to show convergence is the total variation of moments norm. Let \( h \) be the normalized deviation from
\( \alpha_0 \), displaced by the normalized score function

\[
h_N \equiv \sqrt{N} (\alpha - \alpha_0) + \sqrt{N} (\nabla_{\alpha\alpha'} L_N(\alpha_0))^{-1} \nabla_{\alpha} L_N(\alpha_0)
\] (3.12)

\( h_N \) is equivalent to the error of the Taylor approximation for the FOCs of the objective function (2.11) around \( \alpha_0 \). Rearranging,

\[
\alpha = h_N/\sqrt{N} + \alpha_0 - (\nabla_{\alpha\alpha'} L_N(\alpha_0))^{-1} \nabla_{\alpha} L_N(\alpha_0)
\]

The quasiposterior density of \( \alpha \) can be written as

\[
p_N(\alpha) = p_N \left( \frac{h_N}{\sqrt{N}} + \alpha_0 - (\nabla_{\alpha\alpha'} L_N(\alpha_0))^{-1} \nabla_{\alpha} L_N(\alpha_0) \right)
\]

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

\[
\|p\|_{TVM} \equiv \int_A (1 + |h_N|^\beta) |p(h_N)| \, dh_N
\] (3.14)

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

**Corollary 1.** (Convergence of the quasiposterior density) Under as-
sumptions 1-4, and for any finite nonnegative $\beta$,

$$
\|p^*_N(h_N) - p^*_\infty(h_N)\|_{TVM} \equiv \int_{H_N} (1 + |h_N|^2) |p^*_N(h_N) - p^*_\infty(h_N)| \, dh_N \to_p 0
$$

(3.15)

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

$$
p^*_\infty(h_N) = \sqrt{\frac{\det J_N(\alpha_0)}{(2\pi)^m}} \exp\left(-\frac{1}{2}h_N' \nabla_{\alpha\alpha'} L_N(\alpha_0) \frac{h_N}{N}\right)
$$

(3.16)

For a large $N$, the density $p_N(\alpha)$ is random normal, with mean $\alpha_0 - \sqrt{N}(\nabla_{\alpha\alpha'} L_N(\alpha_0))^{-1} \nabla_\alpha L_N(\alpha_0)$, and variance $\sqrt{N}(\nabla_{\alpha\alpha'} L_N(\alpha_0))^{-1}$. In the context of the LTEs proposed here, this is equivalent to saying that the quasiposterior density itself has the same asymptotic distribution and convergence properties as a GMM estimator $\hat{\alpha}_{GMM}$ for the set of moments $\Psi(\alpha)$ in (2.10).

To establish $\sqrt{N}$-consistency, recall that the extremum estimators are defined as the $\operatorname{arg\,sup}_{\alpha \in A} L_N(\alpha)$, and that they are usually determined as the solution to a set of FOCs $(\nabla_\alpha L_N(\hat{\alpha}_{ex}) = 0)$. The Taylor expansion of the FOCs gives

$$
\nabla_\alpha L_N(\alpha) = \nabla_\alpha L_N(\alpha_0) + (\alpha_0 - \alpha) \nabla_{\alpha\alpha'} L_N(\bar{\alpha}) + \mathcal{O}(\sqrt{N})
$$

(3.17)

Evaluating at $\hat{\alpha}_{ex}$ sets the LHS to zero. Rearranging and multiplying
by $\sqrt{N}$, we get the expression for $h_N$ in (3.12). Thus, the extremum estimator $\sqrt{N}(\hat{\alpha}_{\text{ex}} - \alpha_0)$ is first-order equivalent to

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

Given that the LTEs are estimated as the solution of the optimization in equation (3.5), and since $p^*_N \rightarrow p^*_\infty$, then we can also estimate $\sqrt{N}(\hat{\alpha}_{\text{LTE}} - \alpha)$ asymptotically by

$$z_n = \arg\inf_{\alpha \in A} \left\{ \int_A \rho(z - u)p^*_\infty(u - U_N)du \right\} \quad (3.18)$$

Notice the similarity between (3.18) and the equation that defines LTEs estimator (3.5). $z_N$ and $U_N$ are related by the following

$$z_N - U_N = \epsilon_N \quad (3.19)$$

where

$$\epsilon_N \equiv \arg\inf_{\alpha \in A} \int_A \rho(z - u)p^*_\infty(u)du \quad (3.20)$$

which is similar to (3.18) with a shifted probability function.

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

(i) $\sqrt{N}(\hat{\alpha} - \alpha_0) = \epsilon_N + U_N + o_p(1)$.

(ii) $\Omega_N^{-\frac{1}{2}} \nabla_{\alpha\alpha'} L_N(\alpha_0)U_N \rightarrow_d \mathcal{N}(0, I)$
Hence,

\[(iii) \Omega_{N}^{-\frac{1}{2}}(\alpha_{0}) \nabla_{\alpha} L_{N}(\alpha_{0}) (\sqrt{N} (\hat{\alpha} - \alpha_{0}) - \epsilon_{N}) \to_{d} N(0, I) \]

and $\epsilon_{N} = 0$ for symmetric loss functions.

It is easy to show that the information equality property holds for GMM, and therefore holds for our transformed objective function (2.11).

The information equality requires that $\Omega_{N}(\alpha_{0}) \sim \nabla_{\alpha} L_{N}(\alpha_{0})$. Recall our assumption on the asymptotic distribution of our objective function

\[\sqrt{N} G_{N}(\alpha) \sim N(0, \Sigma). \]

The FOCs of the $L_{N}$ give

\[\sqrt{N} G_{N}(\alpha) \Sigma^{-1} \nabla_{\alpha} G_{N}(\alpha) \sim N(0, \nabla_{\alpha} G_{N}(\alpha)\Sigma^{-1} \nabla_{\alpha} G_{N}(\alpha)) \] (3.21)

and imply that asymptotically $\nabla_{\alpha} G_{N}(\alpha) = \nabla_{\alpha} G_{N}(\alpha)\Sigma^{-1} \nabla_{\alpha} G_{N}(\alpha)$

Combining this with assumption 4(ii) we get information equality.

The usual asymptotic intervals for a function $w(\alpha)$, where $\alpha$ is an estimator that satisfies (3.12), is

\[w(\hat{\alpha}) + t_{\frac{1}{2}} \sqrt{\nabla_{\alpha} w(\hat{\alpha})\nabla_{\alpha} G_{N}(\alpha)^{-1} \nabla_{\alpha} w(\hat{\alpha})}, \]

\[w(\hat{\alpha}) + t_{1-\frac{1}{2}} \sqrt{\nabla_{\alpha} w(\hat{\alpha})\nabla_{\alpha} G_{N}(\alpha)^{-1} \nabla_{\alpha} w(\hat{\alpha})} \] (3.22)

Alternatively, we can calculate intervals from the quantiles of the sequence \{\(w(\alpha^{(1)}),...,w(\alpha^{(B)})\)\}, constructed using the sample for $\alpha$ generated through MCMC. Finally, Corollary 3 shows confidence inter-
vals 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 \( w(\alpha) \),

\begin{align*}
(a) \quad & c_{g,N}(d) = g(\hat{\alpha}) + t_d \sqrt{\nabla_{\alpha} f_N(\alpha_0)'(- \nabla_{\alpha \alpha'} L_N(\alpha_0))^{-1} \nabla_{\alpha} f_N(\alpha_0)} + o_p(1/\sqrt{N}) \\
(b) \quad & \lim_{N \to \infty} P\{c_{g,N}(d/2) \leq g(\alpha_0) \leq c_{g,N}(1 - d/2)\} = 1 - d
\end{align*}

To feasibly estimate \( \nabla_{\alpha \alpha'} L_N(\alpha_0)^{-1} \) for the confidence interval, we can use the variance-covariance matrix of the generated MCMC parameter values \( \alpha \).

## 4 Simulation

We simulate a model with 1 overidentifying restriction. We generate a sample for \( \theta \) and \( \lambda \) through MCMC. Specifically, we use the Metropolis-Hastings Algorithm. We estimate quasi-posterior means as in Chernozhukov and Hong (2003), and quasi-posterior means using the transformed equations from Section 2. We use kernel density estimation to calculate joint densities and the corresponding quasi-posterior modes. Finally, we 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). The model has one parameter \( k = 1 \), two moment conditions \( m = 2 \), and hence one overidentifying restriction \( m - k = 1 \). The sample moments are

\[
G_N(\theta) = \begin{bmatrix}
\frac{1}{N} \sum_{i=1}^{N} \exp(\mu - \theta(X_i - Z_i) + 3Z_i) - 1 \\
\frac{1}{N} \sum_{i=1}^{N} (\exp(\mu - \theta(X_i - Z_i) + 3Z_i) - 1)Z_i
\end{bmatrix}
\] (4.1)

This model has been widely used in recent literature. For examples, see Sowell (2008), Imbens, Spady and Johnson (1998), Kitamura (2001), and Schennach (2007). Gregory, Lamarche, and Smith (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 \( \sim N(0,s^2) \). \( \mu \) is a normalization constant set to \( \mu = \frac{\theta_0^2s^2}{2} \) to make the moment conditions zero at the population parameter values. \( \lambda \) is a scalar for this case of 1 overidentifying restriction. For the simulations, we will consider different sample sizes that go from 18 to 50. We use rather small sample sizes to capture the occurrence of multiple minima that are washed away asymptotically.

This model usually 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 \( \theta = \theta_0 \). The analytical solution of the system in Appendix B shows why these multiple modes happen even when the true population value is unique. Multiple modes are a common
feature of GMM estimation\textsuperscript{5}, which makes this model a convenient setup to assess the performance of our estimators.

As discussed in Section 2, when we have a finite sample, the parameter $\theta$ and the overidentifying restriction statistic $\lambda$ are not independent. If we condition on $\lambda = 0$, we will estimate the parameter for which the null hypothesis of the theory being true is not rejected, that is of the expectation of all $m$ moments being zero. We report the parameter estimates conditional on $\lambda = 0$, and compare them with estimations that do not explicitly condition on the overidentifying restrictions.

The model is simulated with 1000 and 10000 repetitions for each different sample ($N = 18$, $N = 37$, $N = 50$). Every simulation is preceded by burn-in iterations to achieve independence from initial values. Posterior densities are produced using the methods described in Section 3, and using the criterion function $L_i(\alpha)$ in (2.11). To construct the function $\Psi_N(\alpha)$ in (2.10), the moments in (4.1) are used to calculate the sample moments $G_N(\theta)$, and $\overline{\pi}(\theta)$.

For the MCMC procedure, we use a normal symmetric distribution, and a flat or diffuse prior (equal probability is given to the possible values the parameter can take). Figure 4.1 shows examples of the distribution of the sample generated using the MCMC procedure for a representative random procedure.

\textsuperscript{5}Using the weighting matrix $\hat{W}$ can sometimes assign enough weight to the moment whose minimum is the population value, because of its lower variance. This would wash out the existence of multiple minima in the second round estimation. However, this depends on data randomness, and is not necessarily always the case.
data draw. Figures A.1 and A.2 in Appendix A show a plot of the actual densities for the same draw. These figures display that these true distributions are close to the distributions generated by MCMC. The GMM objective function in Figure A.1 presents local minima for $N = 18$ for this representative case. This is in agreement with the sample generated by MCMC, see Figure 4.1-a, which captures the existence of 2 values for $\theta$ with a higher probability of occurrence. This implies there are multiple values with high probability of being selected as our estimates. We know from the model that only one value, the population parameter value $\theta_0 = 3$, actually makes the expectations of moments equal to zero, and therefore satisfies the overidentifying restrictions. Figure 4.1-b shows that when we condition on the null hypothesis of the overidentifying restriction being satisfied, or $\lambda = 0$, the mass for the second local minima almost disappears and most of it is now located around the true population value $\theta_0$. On average, by conditioning on $\lambda = 0$, we will rule out the other local minima for which the theory is not satisfied. Figures 4.1c, 4.1c, and A.2 show a representative case for $N = 50$. In this case, most of the mass generated is located around the true population value $\theta_0$, both when using conditional and unconditional estimates. For this draw, the GMM objective function is close to converging to a parabola as shown in Figure A.2. Also, the variance is significantly reduced and the tails become thinner. Figures 4.1-c and d show that in the case of a unique

---

Notice the plots discussed in this paragraph are produced using a representative case with only one random draw of the data. For simulation, a different sample of $X$ and $Z$ is generated every time. As sample size increases, asymptotic effects come into play and
minimum, conditioning on $\lambda = 0$ barely affects the quasiposterior means and modes. Tables 1 and 2 give results for different values of the generated data variance $s$.

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

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

the probability of the objective function having multiple minima decreases. However, the probability of multiple minima remains positive for any finite sample size, and therefore introducing the overidentifying restrictions remains useful. In particular, for $N = 50$, we see that multiple minima still occur often in the simulations.
Table 2: Simulation for the Hall and Horowitz (1996) model using MCMC. 10,000 repetitions.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Estimator</th>
<th>RMSE</th>
<th>BIAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>s=0.4</td>
<td>s=0.6</td>
</tr>
<tr>
<td>18</td>
<td>Q-posterior mean</td>
<td>4.83</td>
<td>4.27</td>
</tr>
<tr>
<td></td>
<td>Q-posterior mean, $\lambda=0$</td>
<td>4.29</td>
<td>4.02</td>
</tr>
<tr>
<td></td>
<td>Q-posterior mode</td>
<td>6.03</td>
<td>4.68</td>
</tr>
<tr>
<td></td>
<td>Q-posterior mode, $\lambda=0$</td>
<td>5.15</td>
<td>4.15</td>
</tr>
<tr>
<td></td>
<td>GMM</td>
<td>1.46</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>C-H</td>
<td>4.88</td>
<td>3.00</td>
</tr>
<tr>
<td>37</td>
<td>Q-posterior mean</td>
<td>0.75</td>
<td>4.83</td>
</tr>
<tr>
<td></td>
<td>Q-posterior mean, $\lambda=0$</td>
<td>0.64</td>
<td>3.99</td>
</tr>
<tr>
<td></td>
<td>Q-posterior mode</td>
<td>0.83</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>Q-posterior mode, $\lambda=0$</td>
<td>0.76</td>
<td>4.10</td>
</tr>
<tr>
<td></td>
<td>GMM</td>
<td>0.78</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>C-H</td>
<td>3.67</td>
<td>6.39</td>
</tr>
<tr>
<td>50</td>
<td>Q-posterior mean</td>
<td>3.40</td>
<td>6.11</td>
</tr>
<tr>
<td></td>
<td>Q-posterior mean, $\lambda=0$</td>
<td>3.31</td>
<td>5.10</td>
</tr>
<tr>
<td></td>
<td>Q-posterior mode</td>
<td>5.80</td>
<td>7.71</td>
</tr>
<tr>
<td></td>
<td>Q-posterior mode, $\lambda=0$</td>
<td>3.17</td>
<td>6.15</td>
</tr>
<tr>
<td></td>
<td>GMM</td>
<td>0.78</td>
<td>2.30</td>
</tr>
<tr>
<td></td>
<td>C-H</td>
<td>5.24</td>
<td>7.61</td>
</tr>
</tbody>
</table>

C-H refers to the estimator used in Chernozhukov and Hong (2003). It is the quasiposterior mean using the criterion function $L_N(\theta)$ that does not include the overidentifying restrictions, namely the GMM objective function in equation (2.3) using the moments in (4.1). The other Q-posterior estimates report Laplace estimators that use the criterion function function $L_N(\theta)$ in (2.11) using the moments in (4.1).

The root mean square error (RMSE) is an indicator of the accuracy of the estimate when compared to the true population value. This measure includes
Figure 4.1: Hall and Horowitz GMM model. MCMC Sample Generation. 10000 iterations after burn-in.
both a measure of bias and of variability. We want both to be low. Comparisons between estimators do not change for the two number of repetitions chosen. On average, all Laplace estimators perform better than the GMM estimator. Our proposed LTEs, those that use the transformed objective function, perform better in general than the C-H estimator, although some results are mixed. Quasi-posterior modes and means conditional on $\lambda = 0$ perform better on average than the rest of the LTEs. These estimators also behave better than the GMM estimator in most of the cases. In particular, the RMSEs are usually lower. In any case, we should keep in mind that GMM is an extremum estimation procedure and suffers from the curse for dimensionality for complex systems, while the other LTEs estimators do not. Thus, all LTEs present advantages over GMM, and we rather focus our attention on the performance relative to each other. comparisons between them. The unconditional quasi-posterior mode has a similar performance to the C-H estimator. Different sample sizes and variances change which estimator of the two outperforms the other. Overall, the RMSEs of the estimators that condition on $\lambda = 0$ outperform the RMSEs of the other estimators. This is in agreement with the intuition discussed, in which we claimed that conditioning on the null hypothesis is a useful criteria to select between multiple minima.
5 Conclusion

We use Bayesian techniques in a classical estimation framework to calculate estimators that are computationally efficient following Chernozhukov and Hong (2003). We extend these estimators by incorporating the overidentifying restrictions in our estimation equations to simultaneously estimate the parameters and the overidentifying restrictions statistic, which are no longer forced to be independent. Thus, introducing the information in the $m - k$ dimensions spanned by the overidentifying restrictions improves the estimation of the $k$ parameters as well. These estimators are an extension of the LTEs in Chernozhukov and Hong (2003), and therefore inherit the computational advantages: allowing for the use of MCMC methods that overcome the curse of dimensionality and working for non-smooth criterion functions. This new information allows the creation of multivariate densities that include both dimensions. These can be used to calculate alternative conditional estimators. By selecting the local minimum that is closer to satisfying the overidentifying restrictions when facing multiple minima of the objective function, these alternative estimates gets closer, on average, to the true population parameter value.

A simulation study for a nonlinear model that often presents multiple local minima in the GMM objective function illustrates the properties of the extended quasiposterior means and modes. Overall, among the extended estimators, those that condition on the overidentifying restrictions being satis-
fied (λ = 0) perform better than their unconditional counterparts. Moreover, the extended estimators perform better in most cases than the quasiposterior mean estimator in Chernozhukov and Hong (2003) (these use criterion functions that include the space of identifying restrictions but not the space of overidentifying restrictions.

Corollaries that show consistency and asymptotic normality, and the validity of the confidence intervals calculated from the MCMC generated samples, are presented with their proofs.

Introducing the overidentifying restrictions 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 implement these estimators in a more realistic environment, through a more complex simulation. Specifically, simulations should be carried out for a large number of moments, parameters and overidentifying restrictions. This way, the curse of dimensionality would affect GMM estimation, making the benefits of LTEs stand out. More complicated models usually present multiple modes even more frequently, which might further improve the performance of the extended conditional estimators.
References


A  Simulation details

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

B  Hall and Horowitz (1996) model solution.
Figure A.2: Hall and Horowitz GMM model. Sample size $N = 50$. 
C. Proofs.

C.1 Proof of Corollary 1

It is enough to show that

$$\int (|h_N|^β) \left| p^*_N(h_N) - p^*_\infty(h_N) \right| dh_N \to_p 0 \quad (C.1)$$

for a general $β$. Rewrite the localized quasiposterior density for $h_N$ in equation (3.12) using the modified quasiposterior density in equation (3.3),

$$p^*_N(h_N) \equiv \frac{1}{\sqrt{N}} p_N(\alpha) = \frac{\pi(\alpha)e^{L_N(\alpha)}}{\int_A \pi(\tilde{\alpha})e^{L_N(\tilde{\alpha})} d\tilde{\alpha}} = \frac{\pi(\alpha)e^{L_N(\alpha)}}{K_N} \quad (C.2)$$

For the rest of the proof, recall that $\alpha$ can be written as a function of $h_N$, and this will be taken into account when differentiating or integrating with respect to $h_N$. To reduce notation this dependence will be dropped.

Define

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

Now, we can rewrite equation (C.1) as $A_N K_N^{-1}$, using the definition of $p^*_\infty(h_N)$ and (C.4).
\[ A_N K_N^{-1} = \int |h_N|^{\beta} e^{L_N(\alpha) K_N^{-1} - (2\pi)^{-d/2} \det \nabla_{\alpha} L_N(\alpha_0)} \left\| e^{\frac{1}{2} h' \nabla_{\alpha} L_N(\alpha_0) h} \right\| dh_N \]
\[ = \int |h_N|^{\beta} e^{L_N(\alpha) \pi(\alpha)} \left\| \int A_1 \pi(\alpha) e^{L_N(\alpha)} d\alpha \right\| - (2\pi)^{-d/2} \det \nabla_{\alpha} L_N(\alpha_0)} \left\| e^{\frac{1}{2} h' \nabla_{\alpha} L_N(\alpha_0) h} \right\| dh_N \]
\[ = \int (|h_N|^\beta) p_N^*(h_N) - p^*_\infty(h_N) |dh_N \]

(C.3)

To show that \( K_N = O_p(1) \), define

\[ A_{1N} \equiv \int |h|^{\beta} e^{L_N(\alpha_h) \pi(\alpha_h) - \exp^{-\frac{1}{2} h' \nabla_{\alpha} L_N(\alpha_0)} \pi(\alpha_0)} dh \]

Chernozhukov and Hong (2003) show that \( A_{1N} \to_p 0 \). Thus, for \( \beta = 0 \),

\[ \int e^{L_N(\alpha_h) \pi(\alpha_h)} dh \overset{p}{\to} \int \exp^{-\frac{1}{2} h' \nabla_{\alpha} L_N(\alpha_0)} \pi(\alpha_0) dh \]
\[ K_N \to_p \int \exp^{-\frac{1}{2} h' \nabla_{\alpha} L_N(\alpha_0)} \pi(\alpha_0) dh \]
\[ = \pi(\alpha_0)(2\pi)^{m/2} |\det \nabla_{\alpha} L_N(\alpha_0)|^{-1/2} \]

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

\[ \int (2\pi)^{-m/2} |\det \nabla_{\alpha} L_N(\alpha_0)|^{1/2} e^{\frac{1}{2} h' \nabla_{\alpha} L_N(\alpha_0) h} |dh = 1 \]

rearranging

\[ \int e^{\frac{1}{2} h' \nabla_{\alpha} L_N(\alpha_0) h} dh = (2\pi)^{m/2} |\det \nabla_{\alpha} L_N(\alpha_0)|^{-1/2} \]
\[ K_N = \pi(\alpha_0)(2\pi)^{m/2} |\det \nabla_{\alpha} L_N(\alpha_0)|^{-1/2} \]

Thus, it is enough to show that \( A_N \to_p 0 \). To do so we write \( A_N \to_p A_{1N} + A_{2N} \) where
\[ A_{2N} \equiv \int |h|^\beta \left| K_N(2\pi)^{-m/2} \det \nabla_{\alpha\alpha'} L_N(\alpha_0)^{1/2} \right| e^{-\frac{1}{2} h' \nabla_{\alpha\alpha'} L_N(\alpha_0) h} \] 

\[ - \pi(\alpha_0) e^{-\frac{1}{2} h' \nabla_{\alpha\alpha'} L_N(\alpha_0) h} |dh| \]

Rearranging,

\[ A_{2N} = \left| C_N(2\pi)^{-\frac{m}{2}} \det \nabla_{\alpha\alpha'} L_N(\alpha_0) \right|^{1/2} - \pi(\alpha_0) \left| \int |h|^\beta e^{-\frac{1}{2} h' \nabla_{\alpha\alpha'} L_N(\alpha_0) h} dh \right| \rightarrow_p 0 \]

since

\[ K_n \rightarrow_p \pi(\alpha_0)(2\pi)^{\frac{m}{2}} \det \nabla_{\alpha\alpha'} L_N(\alpha_0)^{-1/2} \]

### C.2 Proof of Corollary 2

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

\[ z_n = \sqrt{N} (\alpha - \alpha_0) + o_p(1) \quad (C.5) \]

Recall \( z_N = \epsilon_N + U_N \). Since \( U_N = O_p(1) \) (by assumption 4(ii) and CLT), and \( \epsilon_N = O_p(1) \) (by definition of \( p_{\infty}^*(\alpha) \)), we get that \( z_N = O_p(1) \). Jureckova (1997) shows that this, plus the uniform convergence and convexity properties in the assumptions, imply \( z_n - \sqrt{N} \rightarrow_p 0 \), which is equivalent to the desired result.

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

\[ \Omega^{-\frac{1}{2}} \nabla_\alpha L_N(\alpha_0) U_N \sim N(0, \mathcal{I}) \quad (C.6) \]

the desired result.

Solve for \( U_N \) in part(i) and substitute into C.6 to get the final part of the corollary.
C.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 \( \alpha \) in terms of \( h_N \), as defined in (3.12).

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

Define a similar cumulative function over a modified integral area

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

By Corollary 1,

\[ \sup_s \left| F_{w,N} \left( w(\alpha_0) + \frac{s}{\sqrt{N}} \right) - \hat{F}_{w,N} \left( w(\alpha_0) + \frac{s}{\sqrt{N}} \right) \right| \rightarrow^p 0 \quad \text{(C.7)} \]

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

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

which implies with (C.7)

\[ \sup_s \left| F_{w,N} \left( w(\alpha_0) + \frac{s}{\sqrt{N}} \right) - F_{w,\infty} \left( w(\alpha_0) + \frac{s}{\sqrt{N}} \right) \right| \rightarrow^p 0 \quad \text{(C.8)} \]

Because both distributions converge, their quantiles will converge as
well.
\[
\left| F_{w,N}^{-1}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) - F_{w,\infty}^{-1}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) \right| \rightarrow^p 0
\]

Since a variable \( \alpha \) with a density function \( p_\infty^* (\alpha) \) is distributed
\[
\alpha \sim \mathcal{N}(\alpha_0 - \sqrt{N} \nabla_{\alpha'} L_N(\alpha_0)^{-1} \nabla_{\alpha} L_N(\alpha_0), -N \nabla_{\alpha'} L_N(\alpha_0)^{-1})
\]

Thus,
\[
F_{w,\infty}\left(w(\alpha_0) + \frac{s}{\sqrt{N}}\right) = P\{\nabla_{\alpha} w(\alpha_0)'N(S_N, N \nabla_{\alpha'} L_N(\alpha_0)^{-1} \nabla_{\alpha} w(\alpha_0) < s|S_N\}
\tag{C.9}
\]

where \( S_N \equiv -\sqrt{N} \nabla_{\alpha'} L_N(\alpha_0)^{-1} \nabla_{\alpha} L_N(\alpha_0) \). 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 (C.9) we get the quantile
\[
\bar{F}_{w,\infty}^{-1}(d) = \nabla w(\alpha_0)'S_N + t_d \sqrt{\nabla_{\alpha} w(\alpha_0)'N \nabla_{\alpha'} L_N(\alpha_0)^{-1} \nabla_{\alpha} w(\alpha_0)}
\tag{C.10}
\]

where the \( d \)-quantile from a standard normal distribution is used for \( t_d \).
Recall \( c_{w,\infty}(d) = \bar{F}_{w,\infty}^{-1}(d) \). Then,
\[
\bar{F}_{w,N}^{-1}(d) = \sqrt{N} (\bar{F}_{w,N}^{-1}(d) - w(\alpha_0)) \tag{C.11}
\]

Using (C.8), (C.11) and (C.10),
\[
\sqrt{N}(c_{q,N}(d) - w(\alpha_0)) = \nabla w(\alpha_0)'S_N + t_d \sqrt{\nabla_{\alpha} w(\alpha_0)'N \nabla_{\alpha'} L_N(\alpha_0)^{-1} \nabla_{\alpha} w(\alpha_0)} + o_p(1)
\tag{C.12}
\]

Rearranging,
\[
c_{q,N}(d) - w(\alpha_0) - \frac{\nabla w(\alpha_0)'S_N}{\sqrt{N}} - t_d \frac{\sqrt{\nabla_{\alpha} w(\alpha_0)'N \nabla_{\alpha'} L_N(\alpha_0)^{-1} \nabla_{\alpha} w(\alpha_0)}}{\sqrt{N}} = o_p(1)
\tag{C.13}
\]

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

$$w(\hat{\alpha}) = w(\alpha_0) + \nabla w(\alpha)(\hat{\alpha} - \alpha_0) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \quad (C.14)$$

$$= w(\alpha_0) + \nabla w(\alpha) \frac{S_N}{\sqrt{N}} + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

Using (C.14) to substitute in for the second and third term of (C.13), we get

$$c_{q,N}(d) - w(\hat{\alpha}) - t_d \frac{\nabla w(\alpha_0) N \nabla \alpha \alpha \mathcal{L}_N(\alpha_0)^{-1} \nabla \alpha w(\alpha_0)}{\sqrt{N}} = o_p(1)$$

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 $w(\alpha_0)$. 

47