Evaluating the normal approximation in econometric models

Andrew Wang

Presented to the Department of Applied Mathematics
in partial fulfillment of the honors requirement
for a Bachelor of Arts degree

Harvard College
Cambridge, Massachusetts

March 25, 2022

Dedicated to my late father, Yong Wang, who first showed me the joy of learning and discovery

Evaluating the normal approximation in econometric models

Andrew Wang

Abstract

Many models in the economics literature are estimated using generalized method of moments (GMM) and other extremum estimators that attempt to minimize an objective function with few assumptions about the underlying data. These estimates have been proven to be consistent and asymptotically normal under a range of regularity assumptions, and researchers often take these assumptions for granted when calculating standard errors and conducting inference. Using a quasi-Bayesian alternative to the standard GMM estimation procedure, we test whether asymptotic normality is actually a reasonable assumption in three well-cited papers. we use Markov chain Monte Carlo (MCMC) algorithms to sample from the quasi-posterior and find deviations from normality in all cases, though to varying extents.

Acknowledgements

I am extremely grateful to Isaiah Andrews for his advising on this thesis. Our meetings always left me with new insights and a greater appreciation for the research process. He is an inspiring teacher, a brilliant researcher, and one of the kindest people I have met at Harvard.

Thank you to the Applied Math department, especially my academic advisors Margo Levine and Michael Brenner, for providing a wonderful environment where I have been able to freely explore my intellectual interests (and for department lunches, which always make my day).

Shout-out to past research mentors and teachers who have shaped my academic path: Rediet Abebe, Guiping Wang, Xiaowei Zhuang, Igor Sokolov, Karin Knudson, Caroline Odden, Denis Barkats, and John Kovac. Thank you for taking a chance and giving me the opportunity to learn and contribute to your work.

To the many friends who I have gotten to know at Harvard: it has been an honor, and I hope our paths cross again. Thanks in particular to Nancy Hu for proofreading a draft of this thesis and for her support throughout the process.

And to my family: Mom, Dad, Allison, and now Archie, thank you for everything.

1 Introduction

The normal (or Gaussian) distribution is familiar to most people who have taken a class in statistics or econometrics. In this thesis, we examine whether generalized method of moments (GMM), a popular estimation procedure used in economics and other research fields, produces estimates that are approximately normally distributed. In a broad sense, GMM finds parameters for economic models that produce outcomes most closely aligned with empirical observations. The fact that GMM estimates converge to a normal distribution at large sample sizes and under certain regularity conditions is a standard result in econometrics.[1] Due to its generality and minimal requirements on the underlying distribution, GMM has been used in a wide range of applications.

Why does normality matter? Researchers often conduct hypothesis testing and construct confidence intervals based on estimated standard errors together with an assumption that parameter estimates are approximately normal at the sample sizes of their data. If the true distribution of a parameter is heavily skewed or bimodal or displays other non-normal properties, standard inference procedures may not be valid, and we may need to use other methods more suited for each case.

The quasi-Bayesian procedure introduced by Chernozhukov and Hong (2003) provides an alternative to GMM that allows us to better understand the behavior of GMM.[2] Similar to a purely Bayesian approach to estimating parameter values, quasi-Bayes involves generating a posterior distribution over the parameter space from a prior and the data, and then calculating estimates based on this distribution. The difference is that we substitute a transformation of the objective function in place of the typical Bayesian likelihood function, which is in general not known. Chernozhukov and Hong (2003) showed that under the standard conditions required for GMM to be asymptotically normal, the quasi-Bayes posterior (or quasi-posterior) converges to a normal distribution centered around the GMM estimate. We demonstrate that the quasi-posterior is non-normal and thus that the normal approximation for GMM is not necessarily valid in three well-cited empirical studies spanning several fields. The first example uses a rational expectations model from Hansen and Singleton (1982) with financial market data from Campbell and Shiller (1987). The second example uses a demand model from Berry et al. (1995) with US automobile market data. The third uses a behavioral model from DellaVigna et al. (2012) with data from the authors' own field experiments. In all three cases, the quasi-posterior distributions of parameter values are non-normal in a variety of different ways, suggesting that inference based on the standard asymptotic formulas may not be valid.

The paper is organized as follows. Section 2 provides an overview of GMM, while Section 3 provides an overview of quasi-Bayes and estimation via Markov chain Monte Carlo (MCMC) methods. Sections 4, 5, and 6 each cover one empirical example, and Section 7 concludes.

[1] See Section 2 for the specific result that is relevant to GMM's asymptotic normality and Newey and McFadden (1994) for an overview of large-sample estimation in econometrics.

[2] This method was originally introduced as "Laplace-type estimation" by the authors.

1.1 Related Literature

Violations of asymptotic normality have been studied in a variety of other settings. The instrumental variables model, which is a specific case of GMM, is known to break down when the instruments used are not related to the endogenous variables being instrumented (so-called "weak instruments"). This is well-documented in the economics literature, such as in Staiger and Stock (1997), Stock and Wright (2000), Stock et al. (2002) and Andrews et al. (2019). Example 1 is a case of this problem, as pointed out by Stock and Wright (2000) and Andrews (2018).

Another violation occurs when parameters are close to the boundary of the relevant parameter space, which we will see in Example 2. The *parameter on the boundary* problem was studied in depth by Donald Andrews (1999, 2000, 2001, 2002).

This thesis is also related to the broader literature on estimation and inference using quasi-Bayes, which includes Chernozhukov and Hong (2003), Chernozhukov et al. (2007), Tian et al. (2007), Belloni and Chernozhukov (2009), and Andrews and Mikusheva (2022).

2 Generalized Method of Moments

Generalized method of moments (GMM) is an estimation technique frequently used in the economics literature. We observe some data $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ drawn from an unknown distribution F, and we are interested in parameters θ in parameter space Θ. The true value for θ is denoted by θ_0, and we hope to estimate some reasonable approximation to θ_0 from the data.[3]

Since the specific distribution is unknown, we cannot use maximum likelihood estimation. Instead, the researcher specifies some *moment conditions*. These are a vector-valued function of moments $g(X, \theta)$ that each have expected value of 0 when $\theta = \theta_0$ is at its true value (so that $E[g(X, \theta_0)] = 0$). In order for θ_0 to be *point-identified*, we also need $E[g(X, \theta)] \neq 0$ for $\theta \neq \theta_0$ (e.g. if we observed the whole distribution, we would know that θ_0 was the true value because it uniquely zeros $E[g(X, \theta)]$). Cases of weak identification occur when $E[g(X, \theta)] \approx 0$, making it difficult to estimate θ_0.

Define the GMM population objective function as a weighted squared norm of the moments:

$$Q(\theta; X) = \frac{1}{2} E[g(X, \theta)]' W E[g(X, \theta)]$$

for a specified, positive semi-definite weighting matrix W.[4] Note that $Q(\theta_0; X) = 0 < Q(\theta; X)$ for $\theta \neq \theta_0$, i.e. θ_0 uniquely minimizes this objective function so long as θ_0 is point-identified. GMM provides an estimate $\hat{\theta}$ by minimizing the sample analog of this objective function:

$$Q_n(\theta; X) = \frac{1}{2} g_n(X, \theta)' \hat{W} g_n(X, \theta) \tag{1}$$

[3] Note that θ does not have to fully describe the distribution F, though it can.

[4] The $\frac{1}{2}$ helps to simplify notation but does not affect estimation or inference.

for $g_n(X,\theta) = \frac{1}{n}\sum g(\vec{x}_i,\theta)$ and $\hat{W}$ an estimate of W. Then

$$\hat{\theta} = \arg\min_{\theta\in\Theta} Q_n(\theta; X)$$

Example 2.1. *OLS and IV are cases of GMM. The moment conditions are* $E[X(Y - X\beta)] = 0$ *and* $E[Z(Y - X\beta)] = 0$ *respectively, where* X *are regressors,* Y *is outcome, and* Z *are instruments.*[5]

We will see several other examples of moment conditions in the empirical sections. GMM is a specific case of *extremum estimation*, in which a parameter is estimated by minimizing some objective function dependent on the data. It turns out that under certain regularity conditions, GMM is consistent, asymptotically normal, and even asymptotically efficient if the right weighting matrix is chosen. We focus in this thesis on asymptotic normality, and reproduce the relevant result below:

Theorem 2.1 (Asymptotic normality of GMM; Newey and McFadden (1994), Theorem 3.4)**.** *If* $\hat{W} \xrightarrow{p} W$*, the data are i.i.d., and:*

1. *Consistency holds (e.g.* $\hat{\theta} \xrightarrow{p} \theta_0$*)*
2. θ_0 *is in the interior of* Θ
3. $g(X,\theta)$ *is continuously differentiable in a neighborhood* $\mathcal{N}$ *of* θ_0 *with probability approaching one*
4. $E[g(X,\theta_0)] = 0$ *and* $E[||g(X,\theta_0)||]^2$ *is finite*
5. $E[\sup_{\theta\in\mathcal{N}} ||\nabla_\theta g(X,\theta)|| < \infty$
6. $G'WG$ *is nonsingular for* $G = E[\nabla_\theta g(X,\theta_0)]$

then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,\Sigma)$$

for asymptotic covariance $\Sigma = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$ *and* $\Omega = E[g(X,\theta_0)g(X,\theta_0)'] = Cov(g(X,\theta_0))$.[6]

A complete proof can be found in Newey and McFadden (1994). Sufficient conditions for consistency are identification of θ_0, a compact parameter space Θ, continuity of $Q(\theta; X)$, and uniform convergence of $Q_n(\theta; X)$ to $Q(\theta; X)$ [Newey and McFadden (1994), Theorem 2.6].

A sample analog estimate of the variance Σ can be constructed by using $\hat{G} = \frac{1}{n}\sum \nabla_\theta g(X,\hat{\theta})$ and $\hat{\Omega} = \hat{Var}(g(X,\hat{\theta}))$, which are consistent for G and Ω under mild conditions. This yields the following estimate for the asymptotic variance of θ:

$$\hat{\Sigma} = (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\Omega}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1} \tag{2}$$

[5] GMM, as the name suggests, is in some sense a generalization of the method of moments in which parameters are selected to match theoretical moments with empirical moments of the data. However, given that we don't know the underlying distribution, we cannot calculate the theoretical moments. If instead some function of the parameters $f(\theta)$ could be captured by the data as $h(X)$, then we could set our moment as $g(\theta; X) = f(\theta) - h(X)$. We will see this in the DLM example.

[6] This is asymptotically efficient when $W = \Omega^{-1}$.

If we are not confident that the asymptotic normality assumptions hold up, however, we may want to conduct some heuristic tests before relying on the asymptotic covariance formula above. One way to do this is by bootstrapping values of $\hat{\theta}$. Bootstrapping is conceptually simple; we repeatedly construct new samples from our existing data and get new estimates. However, we face the problem of having to minimize $Q_n(\theta; X^*)$ among $\theta \in \Theta$ each time we draw a bootstrap sample X^*. Moment functions $Q_n(\theta) = \frac{1}{2} g_n(\theta)' \hat{W} g_n(\theta)$ can often have many local minima, and especially in higher-dimensional parameter spaces, finding a global minimum repeatedly can be computationally intensive or even infeasible.

Quasi-Bayes provides an elegant alternative, as we will see in the next section.

3 Quasi-Bayes

The fundamental idea of quasi-Bayes is to use a Bayes-like procedure to learn about our parameters from the data, but substituting the objective function in place of the log-likelihood. Specifically, given some prior weighting $\pi(\theta)$ on the parameters $\theta \in \Theta$, the quasi-posterior is

$$\pi^Q(\theta; X) = \frac{\exp(-n \cdot Q_n(\theta))\pi(\theta)}{\int_{\theta' \in \Theta} \exp(-n \cdot Q_n(\theta')\pi(\theta')d\theta'}$$

for the objective function Q_n given in Equation 1.

This quasi-posterior integrates to 1 and is well-defined on Θ. Moreover, given a flat prior, the mode of this distribution occurs exactly at $\theta = \hat{\theta}$, when the objective function is minimized. The intuition here is that the objective function provides information about the parameter, and as more samples accumulate (i.e. n grows larger), the data provides stronger evidence.

We can derive parameter estimates and credible sets from the quasi-posterior, e.g. by the mean, median, or quantiles. For example, an estimate based on the quasi-posterior mean would equal

$$\bar{\theta} = \int_{\Theta} \theta \pi^Q(\theta; X) d\theta$$

Chernozhukov and Hong (2003) show that the estimates, confidence sets, and tests based on the quasi-posterior have nice asymptotic properties given some regularity conditions. Here we present a more approachable version of their results:

Theorem 3.1 (Equivalence of Quasi-Bayes and GMM and Asymptotic Normality; Informal Version of Chernozhukov and Hong (2003), Theorem 1)**.** *Given the assumptions underlying asymptotic normality of GMM and some basic requirements on the prior π, the quasi-posterior mean $\bar{\theta}$ converges in probability to the GMM estimate $\hat{\theta}$, and the quasi-posterior distribution of $\sqrt{n}(\theta - \hat{\theta})$ converges in distribution to $N(0, \Sigma)$.*[7]

This implies that with large enough sample sizes, if the GMM asymptotic normality assumptions hold and we use a flat prior, the quasi-posterior distribution should look roughly normal and our point estimate should

[7] See Chernozhukov and Hong (2003) for a formal statement and conditions.

be close to the one derived via minimization of Q_n. One benefit of a quasi-Bayesian approach is that we avoid the problem of having to minimize a potentially complicated objective function with discontinuities and local minima.[8] However, exact formulas for the quasi-posterior $\pi^Q(\theta)$ are rare in applications, so we cannot analytically calculate the quasi-posterior distribution or our estimates. We need some way to approximate them.

3.1 MCMC

One effective way to approximate the distribution is through MCMC methods, in which we draw a chain of vector-values in the parameter space:

$$R = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(J)})$$

For large enough J, the distribution of the chain approaches that of π^Q. After an burn-in period of B draws, the chain should look roughly stationary. Our parameter estimate is then

$$\hat{\theta}_{\text{approx QP mean}} = \frac{1}{J-B} \sum_{i=B+1}^{J} \theta^{(i)}$$

The sample distributions derived via MCMC asymptotically converge to the target distribution under weak regularity conditions, so running the chain for a large number of iterations should get us a close approximation to the desired quasi-posterior. Furthermore, these methods focus on areas with higher density, making them more efficient than sampling the distribution across a grid.

I will describe two algorithms below that we use for our exercise. In practice, we do not know the normalizing constant in the denominator of π^Q and only know that

$$\pi^Q(\theta; X) \propto \exp(-n \cdot Q_n(\theta; X))\pi(\theta) \tag{3}$$

but both of these methods are still valid. For generality, we denote the quasi-likelihood $\exp(-n \cdot Q_n(\theta; X))$ by $f(\theta; X)$. All of our empirical results are based on a flat prior $\pi(\theta) \propto 1$.

3.2 Metropolis-Hastings

Metropolis-Hastings (M-H) creates a chain of parameter values by proposing a move from the current value θ_j in the chain to a new proposal value ζ drawn from a researcher-specified proposal density $p(\zeta|\theta_j)$. Starting from initial value $\theta_0 = \theta_0$, we iterate to generate the chain $(\theta_0, \theta_1, \ldots, \theta_J)$:

M-H Algorithm For $j = 0, \ldots, J-1$, draw θ_{j+1} as follows:

1. Draw from the proposal density: $\zeta \sim p(\zeta|\theta_j)$
2. Calculate the acceptance ratio:

$$\rho_j = \frac{f(X|\zeta)\pi(\zeta)}{f(d|\theta^j)\pi(\theta^j)} \cdot \frac{p(\theta^j|\zeta)}{p(\zeta|\theta^j)}$$

[8] The quasi-posterior also provides estimation procedures that are more robust under weak identification, as shown in Andrews and Mikusheva (2022), but that is beyond the scope of this thesis.

3. Draw U uniformly from $[0, 1]$

4. Take $\theta_{j+1} = \zeta$ if $U \leq \rho^j$ and $\theta_{j+1} = \theta_j$ otherwise.

In our case, we specify the proposal density to be normal and centered around the current value, such that $\zeta = \theta_j + \varepsilon_j$ for $\varepsilon_j \sim N(0, \Xi)$. The proposal density Ξ is customized for each application.

M-H is arguably the most well-known MCMC algorithm, but it requires the researcher to specify a suitable proposal density. If the step sizes ε_j are small, then the chain can take too long to explore the parameter space and may get easily stuck in a local optimum. If the step sizes are too large, the proposals are rarely accepted, causing a slow convergence as well.[9]

3.3 Slice Sampling

The central idea of slice sampling, a method introduced by Neal (2003), is that a random variable θ can be sampled by uniformly sampling from the area under its density function. This is most easily seen in the one-dimensional case but generalizes to higher dimensions.

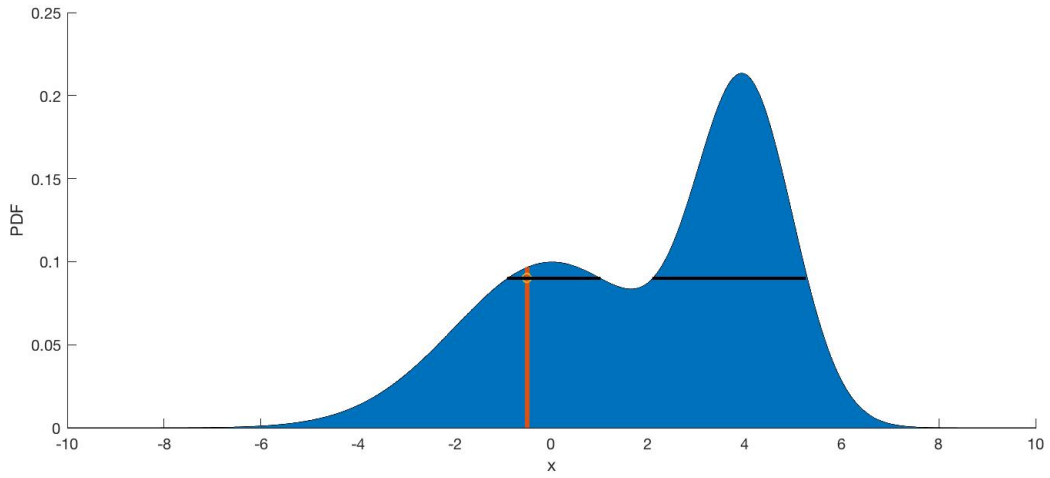


Figure 1: One-dimensional example of a slice sampler step. Given a current value x, a point y is chosen from the red line $[0, f(x)]$, and another x-value is sampled uniformly from the horizontal slice at height y (shown in black).

In the one-dimensional case, suppose we want to sample a random variable X with density $f(x)$. Starting from some value x_0 with $f(x_0) > 0$, the key idea for generating the chain is to take horizontal slices in the density plot at a randomly selected height between 0 and $f(x_0)$.

1-D Slice Sampling Algorithm For $j = 0, \ldots, J-1$, draw x_{j+1} as follows:

1. Sample y uniformly from $[0, f(x_j)]$.

[9] See Chib (2001) for more details on M-H.

2. Draw a horizontal line at height y that crosses all relevant parts of the density function.

3. Randomly select a value x_{j+1} from these horizontal slices at height y.

To implement steps 2 and 3, we can move our endpoints until they are outside of the shaded area and then select a x_{j+1} value via rejection sampling, as described by Neal. With a complicated distribution, the slice from step 2 may not reach all relevant x values, so there is some level of serial dependence in the chain and the possibility of getting stuck in a local neighborhood. Steps 2 and 3 can also be more computationally intensive than simple M-H steps.

The multi-dimensional case is similar and can be done by updating each variable in turn or by using hyper-rectangular slices instead of line segments in step 2. Details can be found in Neal (2003).

3.4 Evaluating convergence

Under mild conditions, both M-H and slice sampling converge asymptotically to the target distribution.[10] However, reaching convergence may require a very large value of J, which is computationally infeasible. We use two primary methods of evaluating whether our MCMC chains have reached a sensible result.

First, we can plot the chains for each component of the parameter vector in *trace plots*. An MCMC chain that looks like noise has likely converged to a local minimum. It is possible, however, that the chain has missed a part of the parameter space. Example 1 is a case where the trace plot looks like it has converged and has found two separate regions with high density. If the proposal densities were too small, the chain would only show one region and would rarely propose moves to the other one.

Our second approach is to initiate chains at different points, which helps to address the concern about getting stuck in a local neighborhood. If these chains converge roughly to the same distribution after a burn-in period, this makes us more confident in our results. We use this as a robustness check in Examples 2 and 3.

We show results based on M-H in the main text, and results based on slice sampling are shown for robustness measures in the Appendix.

4 Empirical Example 1: Hansen and Singleton (1982)

Hansen and Singleton (or HS) was one of the first papers to introduce GMM to the economics literature. They estimate several behavioral parameters in a non-linear rational expectations model for an economic agent.

The moment condition arises from the agent's desire to maximize aggregate expected utility U_0 in a multi-period setting:

$$U_0 = E_0 \left[\sum_{t=0}^{\infty} \delta^t U(C_t) \right]$$

[10] See Chib (2001) and Neal (2003) for formal statements in M-H and slice sampling, respectively.

for consumption C_t in time period t, *discount factor* $\delta \in (0,1)$, and concave utility function $U(\cdot)$. The agent is also subject to some budget constraints. In the case of constant relative risk aversion (CRRA), we have $U(C_t) = \frac{C_t^\eta}{\eta}$ for *coefficient of relative risk aversion* $\eta < 1$. The Euler equation found from taking the first-order condition of the constrained utility is:

$$E_{t-1}\left[\delta\left(\frac{C_t}{C_{t-1}}\right)^{-\eta} R_t - 1\right] = 0$$

where R_t is an aggregate stock return from $t-1$ to t. We can multiply the expectation by any function of observables at time $t-1$, which allows us to incorporate instruments Z_{t-1}. This suggests the following empirical moments as a function of $\theta = (\delta, \eta)$:

$$g_n(\theta, X) = \frac{1}{n}\sum_{t=1}^{n}\left(\delta\left(\frac{C_t}{C_{t-1}}\right)^{-\eta} R_t - 1\right) Z_{t-1}$$

Following Andrews (2018) and Stock and Wright (2000), we use annual US data of stock returns and consumption between 1891-1991 (inclusive) constructed by Campbell and Shiller (1987), so that $n = 101$.[11]

Stock and Wright (2000) report GMM estimates for several different utility specifications and instrument sets; we will follow the main text of Andrews (2018) instead and focus on the CRRA model with a constant and 1-year lagged versions of consumption growth and stock returns as our instruments. The variance-covariance matrix $\hat{\Omega}$ of the moments is estimated as in Andrews (2018) to allow for serial dependence, and we use its inverse as our weighting matrix ($\hat{W} = \hat{\Omega}^{-1}$).[12]

We estimate $\theta = (\delta, \eta)$, e.g. the discount factor and coefficient of relative risk aversion, via our quasi-Bayesian approach described above.

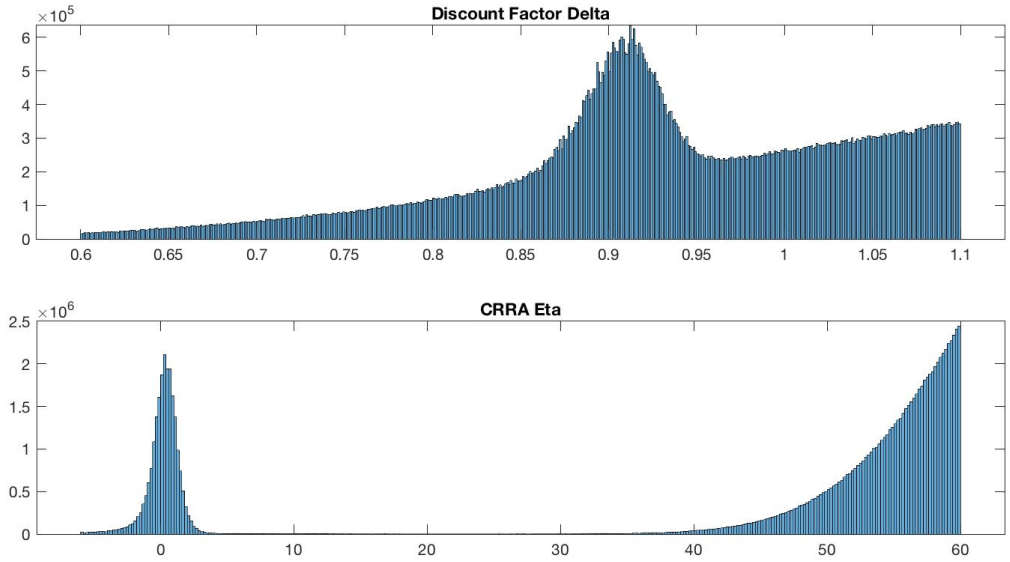
Param	Q-P mean	Pub. estimate	Q-P std	Pub. SE	Q-P median	Q-P quantiles
δ	0.932	0.903	0.106	0.022	0.928	[0.729, 1.085]
η	42.828	0.132	22.995	1.037	54.365	[-0.364, 59.587]

Table 1: Quasi-Bayesian estimates and central 90% credible set based on the quasi-posterior in Figure 2, along with the published estimates and standard errors from Stock and Wright (our own GMM estimates based on the standard GMM formula are similar to theirs and are not shown). Point estimates for η are significantly different between the two methods, and the quasi-posterior standard deviations are much larger than published SE's. Note that the published values are based on more years of data and a different weighting matrix.

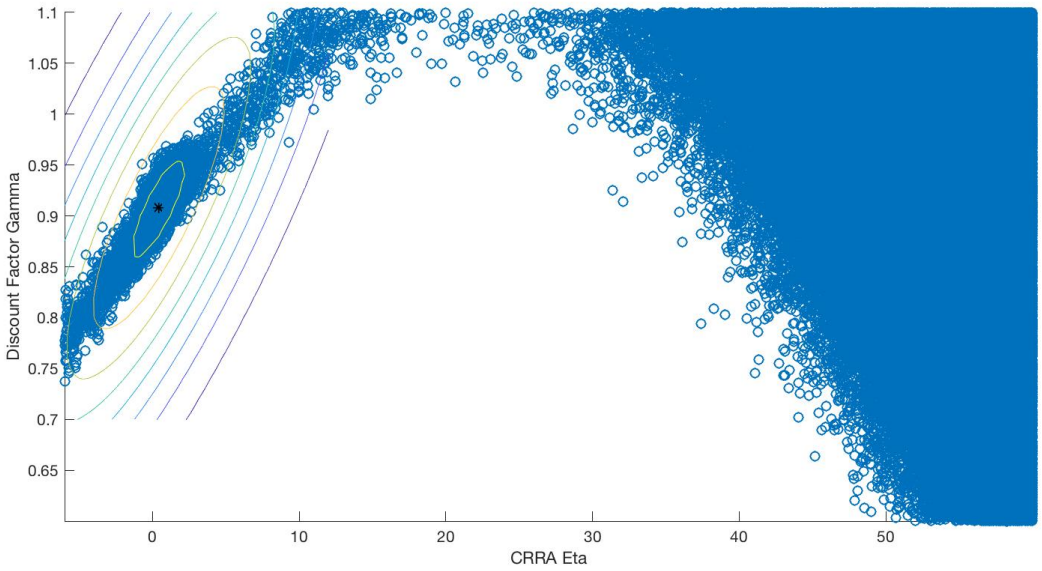
The marginal and joint distributions in Figure 2 make it very clear that the quasi-posterior is not normal. The quasi-posterior seems to have several regions of high density. One is centered around the GMM estimate

[11] Stock and Wright (2000) uses a longer time series from the same dataset, while Hansen and Singleton (1982) uses a different dataset entirely. The underlying model and parameters are the same in all cases.

[12] This is the optimal GMM weighting matrix. Chernozhukov and Hong (2003) also show that in this case, a generalized information equality holds and the quasi-posterior standard deviations can be directly compared with GMM standard errors.



(a) Quasi-posterior marginal distributions.



(b) Quasi-posterior joint distribution. Contours are centered at the GMM extremum estimate and correspond to a normal density with covariance from the asymptotic normality formula.

Figure 2: Discount factor and CRRA in the HS model with data from Andrews (2018). Sampled using M-H for $J = 10,000,000$ runs in 10 chains, for 100M total draws.

and is roughly normal, but the other is much wider.

Given these results, it is no surprise that our Q-P estimates of the mean and standard deviation are quite different than the published GMM estimates. Specifically, our standard deviations are 5-20x larger, as seen

in Table 1. Similar non-normality results have been found by Andrews (2018) and Stock and Wright (2000).

5 Empirical Example 2: Berry, Levinsohn, and Pakes (1995)

Our second empirical example is based on a seminal paper in the industrial organization literature, henceforth referred to as BLP. The paper introduces a random-coefficients logit model to estimate demand and cost parameters in the American automobile market. The details of the model can be found in BLP or Nevo (2000). Here, I give a high-level overview of the data, moments, parameters, and estimation details.

BLP observes annual sales and product characteristics in the automobile market for 20 years between 1971-1990. These include supply-side and demand-side characteristics, such as horsepower/weight ($hpwt$), air conditioning standards (air), miles/dollar (mpd), and space. Through a random-coefficients discrete-choice model, they generate a moment vector with components $g_l(\theta; X) = \frac{1}{n_m} \sum_m g_{l,m}(\theta)$ for instruments l. As a weighting matrix W, we use the inverse of the covariance matrix for $g(\hat{\theta}_{published})$, which differs somewhat from the authors' procedure.

The six components in the parameter vector θ include: (i) the standard deviation of the utility distribution for each automobile characteristic (σ_i for $i \in \{const, hpwt, air, mpd, space\}$) and (ii) the utility from cash (α, or the term on $\ln(y - p)$ in the utility function). Other parameters in the model can be calculated directly from $\theta = (\sigma_i, \alpha)$, as described by BLP and the replication notes for Andrews et al. (2017).

Our quasi-posterior from M-H sampling is shown in Figure 3. The most striking result is that although most of the parameters look roughly normal, values of σ_{air} in the chain are bunched up close to 0. The estimated Q-P mean is less than the standard deviation for that parameter, as seen in Table 2. Although our estimation procedure is somewhat different from BLP, the broad results match up. The published estimate is barely larger than the published standard error, and since σ_{air} is constrained to be greater than 0, it cannot be close to normally distributed even with BLP's procedure. (The problem occurs to a lesser extent for σ_{hpwt} and σ_{mpd}, whose distributions both seem to be cut off sharply at 0.)

This is an example of the *parameter on the boundary* problem, which arises when the true parameter θ_0 is close to the boundary of parameter space Θ. It is clear that inference dependent on the normality of σ_{air} is unreliable. These results hold true across multiple M-H chains and slice sampling as well (see Table 4, Table 5, Figure 7, and Figure 8 in the Appendix).[13]

6 Empirical Example 3: DellaVigna, List, and Malmendier (2012)

Our third empirical example (which we refer to as DLM) comes from a set of field experiments and surveys designed to estimate behavioral parameters related to altruism and giving. In these experiments, the research team went door-to-door in towns around Chicago asking households to donate to one of two charities or

[13]Two authors of BLP also published results about the asymptotic behavior of their estimation procedure. However, their results do not seem to address this problem, at least at the current sample size.

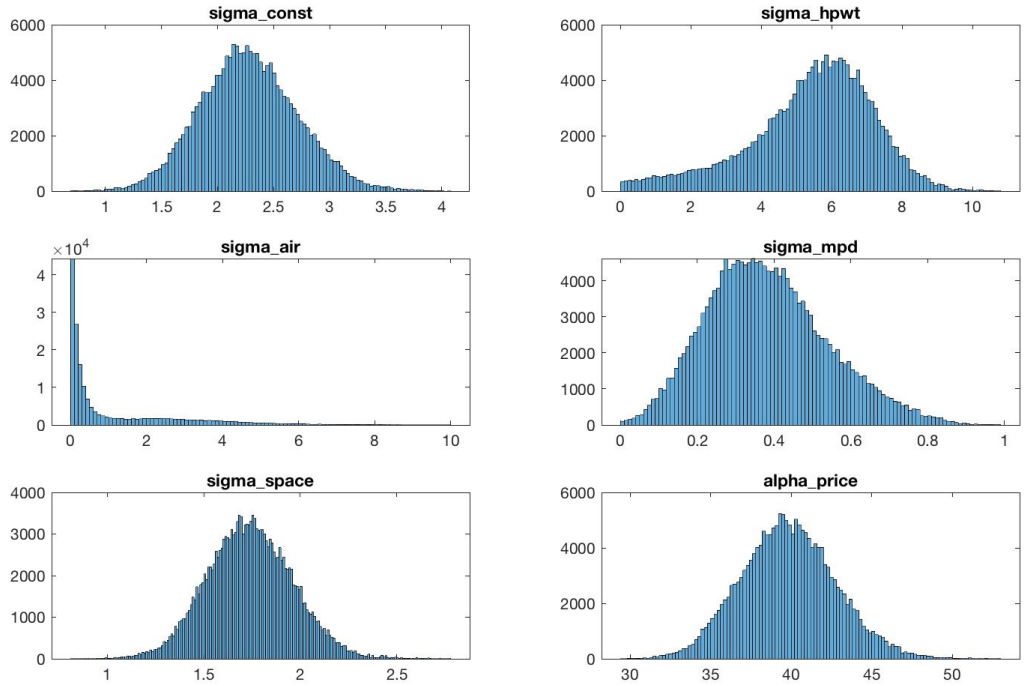


Figure 3: Quasi-posterior marginal distributions for parameters in the BLP model. Sampled using M-H for $J = 50,000$, $B = 5,000$, and 4 chains for a total of 180,000 draws in each histogram.

Param	Q-P mean	Pub. est.	Q-P std	Pub. SE	Q-P median	Q-P quantiles
σ_{const}	2.280	3.612	0.434	1.485	2.266	[1.593, 3.014]
σ_{hpwt}	5.413	4.628	1.783	1.885	5.662	[1.836, 7.913]
σ_{air}	1.186	1.818	1.631	1.695	0.323	[0.017, 4.705]
σ_{mpd}	0.383	1.050	0.158	0.272	0.369	[0.146, 0.669]
σ_{space}	1.739	2.056	0.222	0.585	1.736	[1.385, 2.108]
α_{price}	39.821	43.501	2.938	6.427	39.759	[35.087, 44.759]

Table 2: BLP parameter estimates from M-H sampling, as shown in Figure 3. Published estimate and standard errors are not directly comparable with the Q-P values due to differences in weighting matrices.

complete a survey. Some households were randomly given a flyer to inform them before the visit, while other households were also given the option to opt out.

The behavioral model is a two-period model in which the household's utility is affected by several competing forces: altruism (positive utility from supporting the charity), warm glow (positive utility from the act of giving), and social pressure (negative utility from having to respond to the solicitor). The solicitor randomly chooses whether to provide a flyer and/or the choice of opting-out to each household in the first

period, and the household chooses an action in the second period by rationally anticipating their future utility from various options (e.g. opening the door, donating certain amounts, completing the survey, etc).

The two charities are the East Carolina Hazard Center (ECU) and the La Rabida Children's Hospital (LAR). Only the latter is local to the Chicago area where the survey is conducted, and parameters in the model can differ between the two charities to reflect that altruism towards a local, child-oriented charity may be stronger.

The fifteen components in parameter vector ζ include: (i) the probability of opening the door without seeing a flyer in each year of the experiment ($h_{2008} = h_0$, h_{2009}), (ii) the elasticity of home presence (η), (iii) the probability of observing and remembering the flyer (r), (iv) the mean and standard deviation of the utility of doing a 10-minute survey (μ_s, σ_s), (v) the value of an hour doing the survey ($timeval$), (vi) the social pressure cost associated with saying no to the survey request (S^s), (vii) the mean and standard deviation of utility from altruism (μ_{ch}, σ_{ch} for $ch \in \{LAR, ECU\}$), (viii) the curvature of the altruism function (Gi), and (ix) the social pressure cost from giving zero in response to a donation request (S_{ch}).

There are 70 moments $m(\zeta)$ predicted by this model, which include the probabilities of various actions the household can take in the second period. We use a minimum-distance (MD) estimator to match these theoretical moments with empirical moments $\hat{m}$ to find $\hat{\zeta}$:

$$\hat{\zeta} = \arg\min(m(\zeta) - \hat{m})' W (m(\zeta) - \hat{m})$$

where weighting matrix W is chosen to be the diagonal of the inverse of the variance-covariance matrix of the moments (using the full inverse matrix is computationally challenging). A standard MD estimate is found by the authors via numerical optimization, and they estimate the variance using the standard formula.

This is just a special case of GMM, so we can use Quasi-Bayes to check whether the standard conditions underlying asymptotic normality hold. With a 15-dimensional sample space and a much more complicated set of moments, the MCMC chains take longer to converge. However, we can see in Table 6, Table 7, Figure 9, and Figure 11 of the Appendix that different M-H and slice sampling chains are close to convergence and provide roughly similar outcomes across all parameters.

The quasi-posterior in Figure 4 looks closer to normal than the previous two examples, but it suffers from skewness in several parameters, violating the normal approximation. In particular, η, μ_{lar}, μ_{ecu}, σ_s, S_{svy}, and $timeval$ skew right, while μ_s skews strongly left.[14] Although the MCMC chains have not completely converged, the skew of these parameters holds up across chains and algorithms (see Figure 10 for the case of μ_{lar}).

[14] The magnitude of the skewness values for each of these marginal distributions is greater than 0.6 in both the M-H and slice sampling distributions. μ_{lar}, μ_s, σ_s, and S_{svy} in particular have skewness magnitudes greater than 1 with both algorithms.

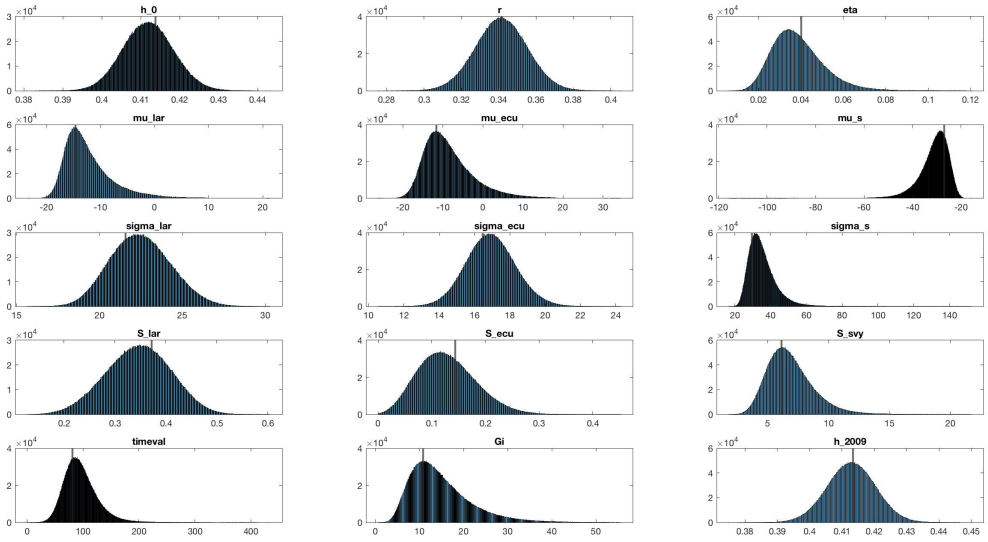


Figure 4: Quasi-posterior marginal distributions for parameters in the DLM model. Sampled using M-H for $J = 500,000$, $B = 50,000$, and 10 chains, for a total of 4.5M draws after burn-in. Vertical black lines denote GMM estimates.

7 Conclusion

These three examples illustrate that the assumptions underlying GMM asymptotic normality are often suspect and provide a possible alternative for GMM inference based on quasi-Bayes. We have seen a variety of violations: multiple local optima in HS, a parameter on the boundary in BLP, and excessive skewness in DLM. Note that it is possible that there are further violations that our MCMC chains did not detect, especially in the latter two cases with higher dimensions.

Given these violations, what should we do for inference? One option for GMM is to use the S-set proposed by Stock and Wright (2000), which is a generalization of Anderson-Rubin (1949) robust confidence sets. This involves collecting all values $\theta \in \Theta$ such that $n \cdot Q_n(\theta) \leq \frac{1}{2}\chi^2_{dim(g),1-\alpha}$ to construct a level $1-\alpha$ confidence set.[15] Another option that was previously mentioned is to bootstrap parameter values and estimate the distribution. This still faces the computational problem of having to repeatedly minimize the objective function and also does not ensure validity of the confidence sets, so we prefer to avoid this approach.

The quasi-posteriors simulated here may provide one of the best alternatives for doing inference in these settings. As Andrews and Mikusheva (2022) show, quasi-Bayes is more robust than GMM in cases of weak identification. Chernozhukov and Hong showed in their original paper that with an optimal weighting matrix, the Q-P standard deviation can be interpreted as a frequentist standard error, and Q-P credible sets can be used for frequentist inference. These methods can also be applied to extremum estimators beyond GMM.

[15] This is based on the asymptotic result that $n \cdot \hat{Q}_n(\theta_0) \to \frac{1}{2}\chi^2_{dim(g)}$ when $W = \hat{Var}(g(X,\theta))^{-1}$.

Param	Q-P mean	Pub. est	Q-P std	Pub. SE	Q-P median	Q-P quantiles
h_0	0.412	0.414	0.006	0.004	0.412	[0.401, 0.423]
r	0.341	0.341	0.014	0.012	0.341	[0.319, 0.364]
η	0.039	0.040	0.012	0.011	0.037	[0.022, 0.060]
μ_{lar}	-12.221	-14.495	4.627	1.444	-13.298	[-17.524, -3.134]
μ_{ecu}	-8.516	-11.522	6.287	1.485	-9.716	[-16.404, 3.588]
μ_s	-31.242	-26.956	6.291	4.204	-30.139	[-42.584, -23.482]
σ_{lar}	22.473	21.584	1.799	1.038	22.419	[19.611, 25.529]
σ_{ecu}	16.917	16.517	1.395	1.103	16.893	[14.669, 19.258]
σ_s	35.286	29.697	7.902	5.129	33.845	[25.693, 49.553]
S_{lar}	0.346	0.372	0.063	0.058	0.347	[0.240, 0.447]
S_{ecu}	0.128	0.142	0.053	0.078	0.124	[0.048, 0.223]
S_{svy}	6.844	6.197	1.898	1.492	6.588	[4.251, 10.281]
$timeval$	95.496	80.656	30.005	22.762	91.711	[54.508, 148.743]
Gi	14.659	10.606	6.778	4.466	13.313	[6.234, 27.739]
h_{2009}	0.413	0.414	0.007	0.007	0.413	[0.401, 0.425]

Table 3: DLM parameter estimates with the distributions from Figure 4. Note that we report a value of the parameter vector that produces a slightly lower objective function than the published values, though the difference is negligible. Additionally, the QP standard deviation and published SE are not directly comparable because the QP does not use the optimal weighting matrix.

References

D. W. Andrews. Estimation when a parameter is on a boundary. *Econometrica*, 67(6):1341–1383, 1999.

D. W. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000.

D. W. Andrews. Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69(3):683–734, 2001.

D. W. K. Andrews. Generalized method of moments estimation when a parameter is on a boundary. *Journal of Business & Economic Statistics*, 20(4):530–544, 2002.

I. Andrews. Valid two-step identification-robust confidence sets for gmm. *Review of Economics and Statistics*, 100(2):337–348, 2018.

I. Andrews and A. Mikusheva. Optimal decision rules for weak gmm. *Econometrica*, 90(2):715–748, 2022.

I. Andrews, M. Gentzkow, and J. M. Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, 132(4):1553–1592, 2017.

I. Andrews, J. H. Stock, and L. Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753, 2019.

A. Belloni and V. Chernozhukov. On the computational complexity of mcmc-based estimators in large samples. *The Annals of Statistics*, 37(4):2011–2055, 2009.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

J. Y. Campbell and R. J. Shiller. Cointegration and tests of present value models. *Journal of Political Economy*, 95(5):1062–1088, 1987.

V. Chernozhukov and H. Hong. An mcmc approach to classical estimation. *Journal of Econometrics*, 115 (2):293–346, 2003.

V. Chernozhukov, H. Hong, and E. Tamer. Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, 75(5):1243–1284, 2007.

S. Chib. Markov chain monte carlo methods: computation and inference. *Handbook of econometrics*, 5: 3569–3649, 2001.

S. DellaVigna, J. A. List, and U. Malmendier. Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, 127(1):1–56, 2012.

L. P. Hansen and K. J. Singleton. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5):1269–1286, 1982.

R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

A. Nevo. A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy*, 9(4):513–548, 2000.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3): 557–586, 1997.

J. H. Stock and J. H. Wright. Gmm with weak identification. *Econometrica*, 68(5):1055–1096, 2000.

J. H. Stock, J. H. Wright, and M. Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529, 2002.

L. Tian, J. S. Liu, and L. Wei. Implementation of estimating function-based inference procedures with markov chain monte carlo samplers. *Journal of the American Statistical Association*, 102(479):881–888, 2007.

A Appendix

A.1 Computational Notes

For HS and DLM, we can run separate MCMC chains on parallel cores. For BLP, the function to calculate $g(X, \theta)$ uses multiple cores itself, so we can only run one chain at a time. We do 10 cores for H-S and DLM, and then run 4 cores for BLP. All code is run on MATLAB 2018, and parallelization is done using Harvard's Odyssey cluster.

A.2 HS

We build off of replication code for Andrews (2018) and set our parameter bounds as $\delta \in [0.6, 1.1]$ and $\eta \in [-6, 60]$. For our M-H proposal density covariance Ξ, we use $16\hat{\Sigma}$, where $\hat{\Sigma}$ is the asymptotic covariance estimate. It takes less than 15 minutes to run an M-H chain of $J = 10^7$ values and a slice sampling chain of $J = 5 \cdot 10^5$ values. We initiate both algorithms at a GMM estimate calculated prior to generating the chains, and we do not test different starting points because the whole parameter space is explored anyways.

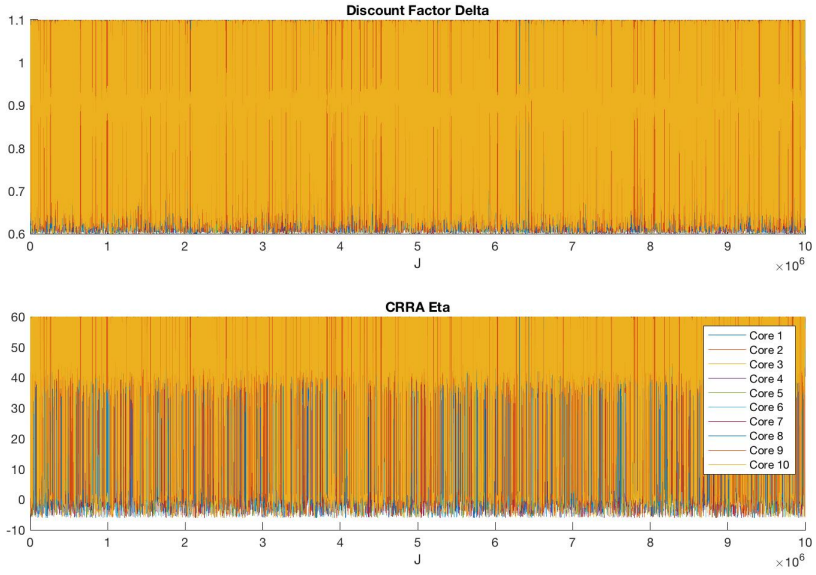


Figure 5: M-H trace plots for HS parameters with $J = 10,000,000$.

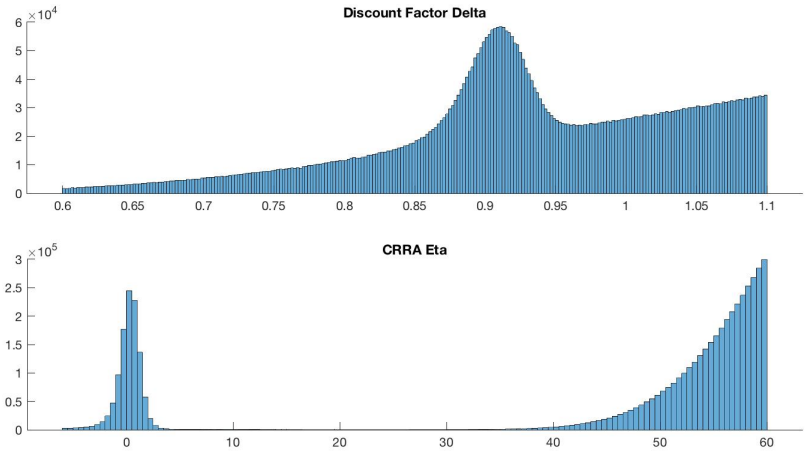
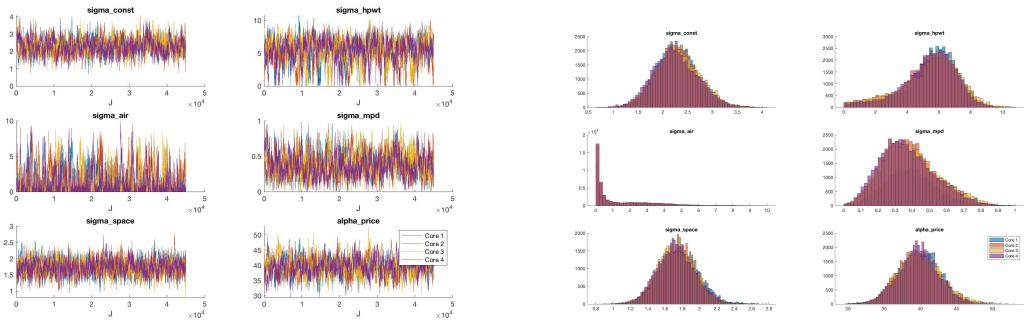


Figure 6: Histogram of parameter values combined from 10 slice sampling chains, each of length $J = 500,000$. This matches the M-H result.

A.3 BLP

We build off of replication code for BLP from Andrews, Gentzkow, and Shapiro (2017) and set our parameter bounds such that $\sigma_i > 0$. For our M-H proposal density covariance Ξ, we use the scaled covariance of the slice sampling chain $4 \cdot Cov(\hat{\theta}_{slice})$. It takes around 0.5s to run one M-H iteration and 1.25s to run one slice sampling iteration, so our chain lengths are shorter due to running time constraints.

For our initial values in each chain j, we scale the i-th component of the published GMM estimate $\hat{\zeta}$ randomly by $s_{ij} \sim U[0.99, 1.01]$ for M-H. We initiate our slice sampling chain at the published estimate. We use $n = 999$, the number of automobile models in the dataset, because the moments are averaged across these models.



(a) Trace plots.

(b) Individual histograms across cores.

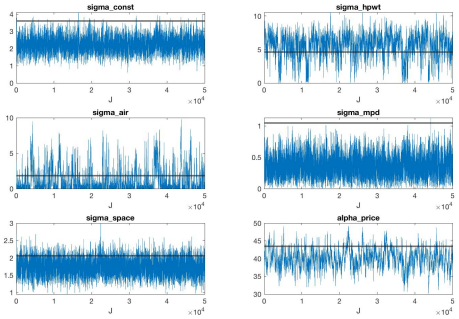
Figure 7: Trace plots and histograms for 4 M-H chains with $J = 50,000$ and $B = 5,000$ (for 180,000 total draws after burn-in).

A.3.1 Heuristic Convergence Checks

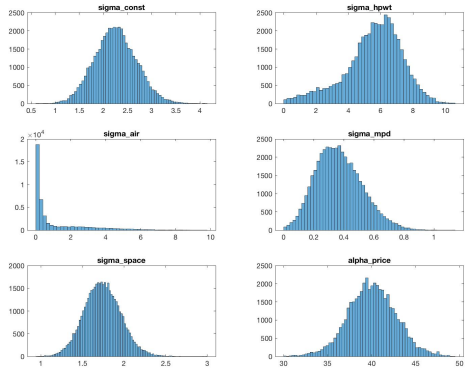
Here, we compare the parameter estimates derived from each independent M-H chain. Note that this is purely heuristic and should not be interpreted as a formal significance test.

Param	Max Q-P mean	Min Q-P mean	Difference	Q-P standard deviation
σ_{const}	2.304	2.252	0.0526	0.434
σ_{hpwt}	5.533	5.269	0.2641	1.783
σ_{air}	1.321	0.997	0.3235	1.631
σ_{mpd}	0.388	0.372	0.0160	0.158
σ_{space}	1.755	1.729	0.0261	0.222
α_{price}	39.956	39.707	0.2490	2.938

Table 4: Maximum and minimum Q-P means across 4 M-H chains after burn-in, initiated at different points in the parameter space. Standard deviation calculated from the aggregate of data from all 4 chains.



(a) Trace plots. Black lines denote published values.



(b) Histograms.

Figure 8: Trace plots and histograms for parameters σ^s and α from slice sampling with $J = 50,000$ and $B = 5,000$ in 1 chain.

Param	Q-P mean	Pub. est.	Q-P std	Pub. SE	Q-P median	Q-P quantiles
σ_{const}	2.251	3.612	0.434	1.485	2.243	[1.548, 2.972]
σ_{hpwt}	5.553	4.628	1.776	1.885	5.766	[2.016, 8.103]
σ_{air}	1.110	1.818	1.600	1.695	0.288	[0.015, 4.710]
σ_{mpd}	0.368	1.050	0.154	0.272	0.356	[0.136, 0.639]
σ_{space}	1.740	2.056	0.219	0.585	1.735	[1.386, 2.105]
α_{price}	40.068	43.501	2.736	6.427	40.044	[35.644, 44.528]

Table 5: BLP parameter estimates from slice sampling, as described in Figure 8. These match closely with the estimates from M-H in Table 2.

A.4 DLM

We build off of replication code from the authors and use the parameter bounds from their paper. For our M-H proposal density covariance Ξ, we use $\hat{\Sigma}/10$. For our initial values in each chain j, we scale the i-th component of the GMM estimate $\hat{\zeta}$ randomly by $s_{ij} \sim U[0.9, 1.1]$ for M-H and $s_{ij} \sim U[0.99, 1.01]$ for slice sampling. It takes 0.005s for one M-H iteration and 0.09s for one slice sampling iteration, though it get significantly slower at large chain lengths (possibly due to memory constraints).

We set $n = 1$ because the moments are calculated from separate field experiments and the weighting matrix builds in the sample sizes. The moments $\hat{m}$ and covariance matrix $\hat{\Omega}$ of the moments are calculated by the authors from a first-stage regression of an outcome (e.g. responding to survey) on 70 indicator variables from all four empirical treatments, with a set of control variables and fixed effects. Other values of n based on the experiment sample sizes give standard deviations a factor of $\sqrt{n}$ off, providing further support for the use of $n = 1$.[16]

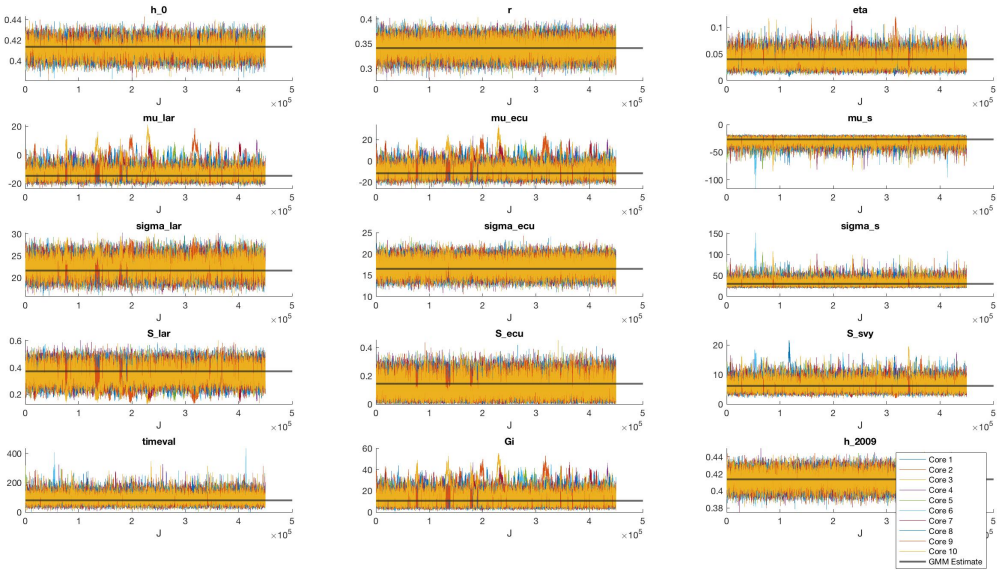
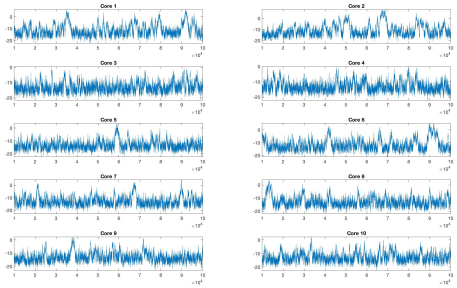


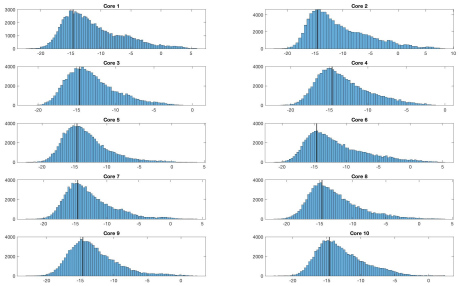
Figure 9: M-H trace plots for DLM parameters with $J = 500,000$, $B = 50,000$, and 10 chains as in Figure 4. Horizontal black lines denote parameter values of the GMM estimate.

A.4.1 Heuristic Convergence Checks

[16] There were a total of 19,658 people across the different field components.



(a) Some autocorrelation in the chains.

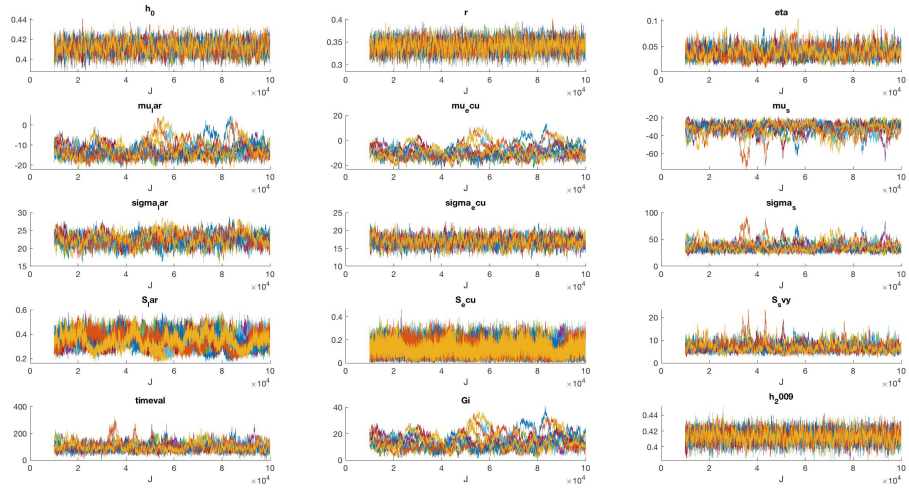


(b) Marginals all skew in the same direction. Vertical lines denote GMM estimate.

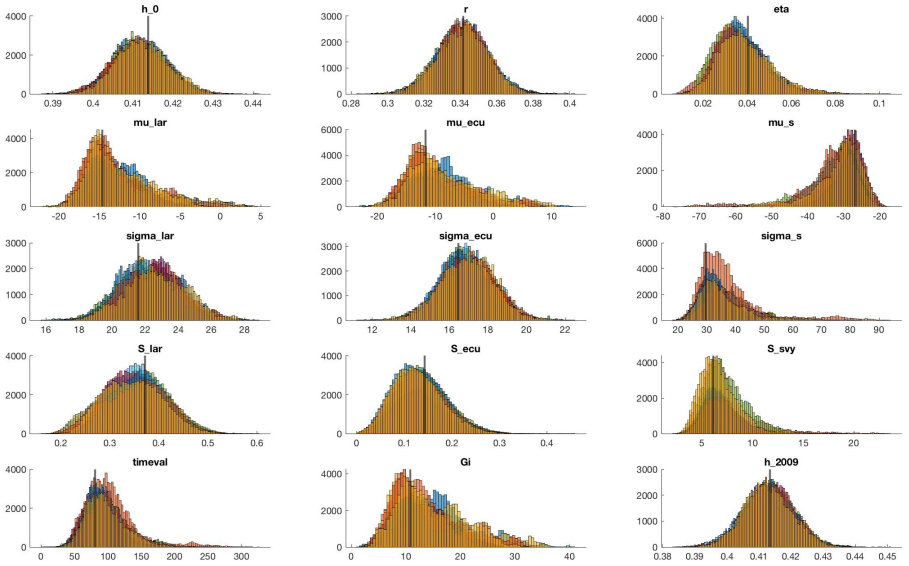
Figure 10: Trace plots and histograms for parameter μ_{lar}, the mean utility derived from altruism towards La Rabida Children's Hospital, across 10 M-H chains.

Param	Max Q-P mean	Min Q-P mean	Difference	Q-P standard deviation
h_0	0.412	0.412	0.000	0.006
r	0.342	0.341	0.001	0.014
η	0.039	0.038	0.001	0.012
μ_{lar}	-11.641	-12.727	1.086	4.627
μ_{ecu}	-7.901	-9.194	1.293	6.287
μ_s	-30.881	-31.397	0.516	6.291
σ_{lar}	22.587	22.350	0.237	1.799
σ_{ecu}	16.957	16.888	0.068	1.395
σ_s	35.490	34.834	0.656	7.902
S_{lar}	0.350	0.341	0.010	0.063
S_{ecu}	0.131	0.126	0.005	0.053
S_{svy}	6.935	6.745	0.189	1.898
$timeval$	96.365	94.255	2.111	30.005
Gi	15.296	13.935	1.361	6.778
h_{2009}	0.413	0.413	0.000	0.007

Table 6: Maximum and minimum Q-P means across 10 M-H chains after burn-in, initiated at different points in the parameter space. Standard deviation calculated from the aggregate of data from all 10 chains.



(a) Trace plots.



(b) Individual histograms across cores.

Figure 11: Trace plots and histograms for 10 slice sampling chains in DLM with $J = 100,000$ and $B = 10,000$. Vertical lines denote GMM estimate.

Param	Q-P mean	Pub. est.	Q-P std	Pub. SE	Q-P median	Q-P quantiles
h_0	0.412	0.414	0.007	0.004	0.412	[0.401, 0.423]
r	0.341	0.341	0.014	0.012	0.341	[0.319, 0.364]
η	0.038	0.040	0.011	0.011	0.036	[0.021, 0.058]
μ_{lar}	-12.946	-14.495	3.674	1.444	-13.721	[-17.544, -5.740]
μ_{ecu}	-9.541	-11.522	5.217	1.485	-10.307	[-16.629, 0.350]
μ_s	-31.451	-26.956	6.398	4.204	-30.266	[-43.152, -23.522]
σ_{lar}	22.292	21.584	1.735	1.038	22.250	[19.518, 25.208]
σ_{ecu}	16.913	16.517	1.368	1.103	16.902	[14.662, 19.164]
σ_s	35.466	29.697	7.985	5.129	33.956	[25.698, 50.324]
S_{lar}	0.354	0.372	0.059	0.581	0.354	[0.258, 0.451]
S_{ecu}	0.133	0.142	0.053	0.784	0.129	[0.052, 0.226]
S_{svy}	6.978	6.197	1.946	1.492	6.692	[4.407, 10.482]
$timeval$	96.363	80.656	30.096	22.762	92.799	[54.758, 150.800]
Gi	13.598	10.606	5.639	4.466	12.649	[6.069, 24.420]
h_{2009}	0.413	0.414	0.008	0.007	0.413	[0.401, 0.425]

Table 7: DLM parameter estimates via slice sampling instead of M-H. Based on the data from Figure 11, with a total of 900M points after burn-in.