GMM is Inadmissible Under Weak Identification

By Isaiah Andrews[1] and Anna Mikusheva[2]

Abstract

We consider estimation in moment condition models and show that under squared error loss and bounds on identification strength, asymptotically admissible (i.e. undominated) estimators must be Lipschitz functions of the sample moments. GMM estimators are in general discontinuous in the sample moment function, and are thus inadmissible under weak identification. We show, by contrast, that quasi-Bayes posterior means and bagged, or bootstrap aggregated, GMM estimators have superior continuity properties, while results in the literaure imply that they are equivalent to GMM when identification is strong. In simulations calibrated to published instrumental variables specifications, we find that these alternatives often outperform GMM. Hence, quasi-Bayes and bagged GMM present attractive alternatives to GMM.

Keywords: Limit Experiment, Weak Identification, Nonlinear GMM

JEL Codes: C11, C12, C20

1 Introduction

Generalized method of moments (GMM) estimators are ubiquitous in empirical economics, and many popular estimation methods including linear and nonlinear instrumental variables, moment-matching, and many cases of maximum likelihood, can be cast as special cases. Appropriately constructed GMM estimators are known to be efficient in large samples, in the sense of minimizing mean squared error over a large class of estimators, provided model parameters are strongly identified (i.e. the data are sufficiently informative) and other regularity conditions hold (see Hansen 1982, Chamberlain 1987).

Unfortunately, however, in many contexts of economic interest the data provide only limited information about model parameters (Mavroeidis et al. 2014, Armstrong 2016, Andrews et al. 2019). In such cases, asymptotic results assuming strong identification

[1]Harvard Department of Economics, Littauer Center M18, Cambridge, MA 02138. Email: iandrews@fas.harvard.edu. Support from the National Science Foundation under grant number 1654234 is gratefully acknowledged.

[2]Department of Economics, M.I.T., 50 Memorial Drive, E52-526, Cambridge, MA, 02142. Email: amikushe@mit.edu. We thank Andrew Wang and Bas Sanders for research assistance.

can be unreliable, and weak-identification approximations, which model informativeness of the data as limited even in large samples, often provide a better description of finite-sample behavior (Staiger and Stock 1997, D. Andrews and Cheng 2012, Andrews and Mikusheva 2022). Standard arguments for the efficiency of GMM no longer apply under weak identification, raising the question of whether GMM estimators should be used in such settings and, if not, what alternatives we should prefer.

There generally exists no single best estimator under weak identification, since optimizing performance over different parts of the parameter space leads to different estimators. A minimal requirement is that an estimator be admissible, meaning that there exists no alternative estimator which performs at least as well for all parameter values and strictly better for some. Our main result shows that GMM estimators are inadmissible under bounds on the strength of identification.

Our proof for inadmissibility is non-constructive, in the sense that it does not deliver a dominating estimator, but nonetheless suggests directions for improvement. Specifically, we show that admissible estimators must be Lipschitz in the sample moments. To prove this result, we first note that by a complete class theorem, any admissible estimator under squared-error loss must be equal to the limit of a sequence of Bayes posterior means for some sequence of priors. Under bounds on identification strength, however, Bayes posterior means are Lipschitz in the sample moments, so small changes in these moments lead only to small changes in the posterior mean. Moreover, the Lipschitz property is preserved under limits. GMM estimators, however, change discontinuously in the sample moments when the minimizer of the sample GMM objective function is non-unique, and so fail to satisfy this necessary condition for admissibility.

Motivated by the necessity of Lipschitz continuity for admissibility, we next explore more continuous alternatives to GMM. We discuss two such estimators: first a quasi-Bayes posterior mean, and second a bagged (or bootstrap aggregated) GMM estimator. While we do not claim these estimators are admissible, we show that they have better continuity properties than GMM under weak identification, while existing results imply that both are asymptotically equivalent to GMM under strong identification and standard regularity conditions. Hence, in large samples there is no first-order loss from using these estimators if identification is strong.

Quasi-Bayes puts a prior on the structural parameters, treats the GMM objective as a negative log-likelihood, and combines the two to compute a quasi-posterior distribution. This approach was initially proposed by Chernozhukov and Hong (2003) for settings where minimization is computationally intractable, and they showed that quasi-Bayes is asymptotically equivalent to GMM under strong identification. More recently, in Andrews and Mikusheva (2022) we showed that quasi-Bayes arises as the limit of a sequence of Bayes decision rules under weak identification. In the present paper, we show that quasi-Bayes posterior means are Lipschitz in the GMM objective function. While quasi-Bayes is not in general Lipschitz in the moments, we show that it is Lipschitz in the special case where (i) the structural parameter takes only a finite number of possible values and (ii) the J-statistic for testing over-identifying restrictions is bounded.

The second alternative estimator we discuss, bagged GMM, corresponds to the average of the GMM estimator across bootstrap realizations. Bagging smooths any discontinuities in the estimator, and we show that bagged GMM is Lipschitz in many cases. Moreover, bagged GMM has a Bayesian interpretation, corresponding to the posterior mean of the GMM estimand under an uninformative prior that does not impose correct specification of the GMM model. Moreover, standard results on bootstrap bias correction (see e.g. Horowitz, 2001, Chen and Hall 2003) imply that bagged GMM is asymptotically equivalent to GMM in the strongly-identified case.[3]

We compare these estimators in simulation designs (from Andrews et al., 2019) calibrated to 124 linear instrumental variables specifications published in the American Economic Review. We find that the performance of quasi-Bayes depends strongly on the prior. Specifically, quasi-Bayes estimators with a flat prior perform the worst of all estimators considered, while quasi-Bayes estimators with a novel invariant prior (motivated by parameterization invariance in the spirit of Jeffreys, 1946) perform the best. Bagged GMM estimators typically have smaller mean squared error than their conventional counterparts, consistent with poor performance for GMM under weak identification.

Section 2 describes the estimation problem we consider, the limit experiment (based on Andrews and Mikusheva 2022) in which we conduct our analysis, and defines a the-

[3]As Chen and Hall (2003) show for estimating equation models, however, bagging is essentially the opposite of standard bias-correction and will increase higher-order bias in the well-identified case.

oretical measure of identification strength. Section 3 shows that admissible estimators under bounds on identification strength must be Lipschitz in the moments, and shows that GMM fails to satisfy this condition. Section 4 turns to alternative estimators, discussing quasi-Bayes in Section 4.1 and bagged GMM in Section 4.2. Section 5 compares the performance of these estimators in simulation.

2 Setting

Consider a researcher who observes a sample of independent and identically distributed observations $X^n = \{X_i, i = 1, ..., n\}$ with $X_i \in \mathcal{X}$, and who wants to estimate some bounded function $r(\theta^*) \in \mathbb{R}^p$ of a structural parameter $\theta^* \in \Theta$. The researcher might, for instance, be interested in the full parameter vector, $r(\theta^*) = \theta^*$, or in a lower-dimensional function such as a counterfactual or average causal effect. We assume Θ is compact and that the true structural parameter value θ^* satisfies a moment condition $\mathbb{E}\left[\phi(X, \theta^*)\right] = 0$ for $\phi(\cdot, \cdot)$ a known $\mathbb{R}^k$-valued function of the data and parameters. The researcher selects an estimate $a \in \mathcal{A} \subset \mathbb{R}^p$, where we assume that $\mathcal{A}$ is compact and contains the convex hull of $\{r(\theta) : \theta \in \Theta\}$. For a given choice of a the researcher incurs squared error loss

$$L\left(a, \theta^*\right) = \left(r\left(\theta^*\right) - a\right)' \Xi \left(r\left(\theta^*\right) - a\right). \tag{1}$$

The researcher's goal is to select an estimator $\delta_n : \mathcal{X}^n \to \mathcal{A}$ that yields low risk, or expected loss, $\mathbb{E}[L(\delta_n(X^n), \theta^*)]$, where θ^* and the distribution of X are both unknown.

GMM estimators are popular in this setting. GMM estimates θ^* by minimizing some distance between the scaled sample moments $g_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(X_i, \theta)$ and zero,

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} g_n(\theta)' W_n(\theta) g_n(\theta),$$

for a potentially data- and parameter-dependent weighting matrix $W_n(\theta)$. GMM then estimates $r(\theta)$ using the plug-in method, $\delta_n^{GMM}(X^n) = r(\hat{\theta}_n)$. Well-known asymptotic arguments (see Hansen 1982) provide conditions under which $r(\hat{\theta}_n)$ is consistent for $r(\theta^*)$ and asymptotically normal as $n \to \infty$. These results further establish that if $W_n(\theta^*)$ is proportional to the inverse of the variance of $g_n(\theta^*)$ then the GMM estimator is asymptotically efficient, in the sense that δ_n^{GMM} minimizes the asymptotic risk $\lim_{n\to\infty} n \cdot \mathbb{E}[L(\delta_n(X^n), \theta^*)]$

over a large class of estimators.[4]

Standard asymptotic results for GMM require, among other assumptions, that θ^* is point identified and strongly identified. Specifically, the moment condition $\mathbb{E}\left[\phi(X,\theta)\right] = 0$ should be uniquely solved at θ^*, and the sample moment function $g_n(\theta)$ should be well-separated from zero, asymptotically, outside infinitesimal neighborhoods of θ^*. These point- and strong-identification assumptions are a poor fit for many economic applications, so in Andrews and Mikusheva (2022) we derived an alternative asymptotic efficiency theory for moment condition models with weak and partial identification. There, we showed that under mild conditions the problem of inference on θ^* under weak identification reduces, asymptotically, to observing a single realization of a Gaussian process

$$g(\cdot) \sim \mathcal{GP}(m, \Sigma) \tag{2}$$

with an unknown mean function m satisfying $m(\theta^*) = 0$, and a known covariance function Σ. We assume that $\Sigma(\theta,\theta)$ has full rank for all θ and that $\Sigma(\theta,\tilde{\theta})$ is continuous on $\Theta \times \Theta$. In this limit experiment, as in the finite-sample problem, the goal is to choose an estimator δ, which now maps realizations of $g(\cdot)$ to estimates $\delta(g,\Sigma) \in \mathcal{A}$, in a way which yields a low risk $\mathbb{E}_m[L(\delta(g,\Sigma),\theta^*)]$, where $\mathbb{E}_m[\cdot]$ denotes the expectation taken under (2). Andrews and Mikusheva (2022) shows that the risk in the limit experiment lower-bounds the (appropriately scaled) asymptotic risk in the original problem.

In addition to deriving lower bounds, we can use the limit experiment to construct asymptotically optimal estimators. Intuitively, $g_n(\cdot) \Rightarrow g(\cdot) \sim \mathcal{GP}(m,\Sigma)$ in large samples, where $\Sigma(\theta,\tilde{\theta}) = Cov(\phi(X_i,\theta),\phi(X_i,\tilde{\theta}))$ and Σ is consistently estimated by the sample covariance $\widehat{\Sigma}$. Hence, for finite-sample estimators of the form $\delta_n(X^n) = \delta\left(g_n,\widehat{\Sigma}\right)$, we have $\delta\left(g_n,\widehat{\Sigma}\right) \Rightarrow \delta(g,\Sigma)$ under mild conditions, and the asymptotic performance of $\delta_n(X^n)$ coincides with the performance of $\delta(g,\Sigma)$. Thus, if $\delta(g,\Sigma)$ is optimal in the limit experiment, the plug-in estimator $\delta_n(X^n) = \delta\left(g_n,\widehat{\Sigma}\right)$ is asymptotically optimal. Moreover, we can evaluate the large-sample performance of GMM by studying the behavior of $\delta^{GMM}(g,\Sigma) = r(\hat{\theta})$ in the limit experiment, where $\hat{\theta} \in \arg\min_{\theta\in\Theta} g(\theta)'W(\theta)g(\theta)$ for $W(\theta)$ the probability limit of $W_n(\theta)$.

[4]Uniform integrability conditions are needed to ensure that $\lim_{n\to\infty} n \cdot \mathbb{E}[L(\delta_n(X^n),\theta^*)]$ is well-behaved. Absent such conditions, analogous results hold for trimmed losses.

Motivated by the results of Andrews and Mikusheva (2022), the following sections focus on properties for the limit experiment (2). First, however, we introduce two special cases and characterize the parameter space for the limit experiment.

Special Case: Linear IV For our first special case we consider the linear IV model. Suppose $X_i = (Y_i, D_i, Z_i')$ for $Y_i \in \mathbb{R}$ an outcome of interest, $D_i \in \mathbb{R}$ an endogenous regressor, and $Z_i \in \mathbb{R}^k$ a vector of instruments. The familiar linear IV estimators correspond to GMM with moment condition $\phi(X_i, \theta) = (Y_i - D_i\theta)Z_i$ and different choices of weighting matrix, for instance $W_n = (\frac{1}{n}\sum Z_i Z_i')^{-1}$ for two-stage least squares.

Weak-identification asymptotics in this case correspond to weak IV asymptotics as in Staiger and Stock (1997), and model the first stage parameter as shrinking with the sample size to ensure that it cannot be distinguished from zero with certainty, with $\mathbb{E}[D_i Z_i] = \frac{1}{\sqrt{n}}\pi^*$ for a fixed vector π^*. The $\mathbb{R}^k$-valued Gaussian process $g(\cdot)$ is linear in θ, and so is fully characterized by its intercept $g(0) = \xi_0$ and slope $\frac{\partial}{\partial\theta}g(\theta) = -\xi_1$, where

$$(\xi_0', \xi_1')' \sim N((\pi^{*\prime}\theta, \pi^{*\prime})', \Omega), \quad \Omega = Var((Z_i'Y_i, Z_i'D_i)'). \tag{3}$$

Intuitively, ξ_0 corresponds (up to a linear tranformation) to the reduced-form coefficient from regressing Y_i on Z_i, while ξ_1 corresponds to the first-stage regression of D_i on Z_i. For $\Theta = [\theta_L, \theta_U]$ an interval, the two-stage least squares estimator for θ is

$$\hat{\theta} = \arg\min_{\theta\in\Theta}(\xi_0 - \xi_1\theta)'W(\xi_0 - \xi_1\theta) = \min\left\{\theta_U, \max\left\{\frac{\xi_1'W\xi_0}{\xi_1'W\xi_1}, \theta_L\right\}\right\}, \tag{4}$$

for $W = \mathbb{E}[Z_i Z_i']^{-1}$, and the corresponding GMM estimator is $\delta(g, \Sigma) = r(\hat{\theta})$. □

Special Case: Finite Θ For our second special case we consider a potentially nonlinear moment condition $\phi(X_i, \theta)$ but restrict the structural parameter space to contain only a finite number of points, $\Theta = \{\theta_1, ..., \theta_r\}$. While theoretical models in economics are typically written using continuous parameterizations, computational implementation is limited by machine precision, so the case with a finite parameter space Θ is arguably a better description of empirical practice.

Weak-identification asymptotics in this setting correspond to the weak-GMM asymptotics of Stock and Wright (2000), and imply that the mean of the moments is of the same order as sampling uncertainty, $\mathbb{E}[\phi(X_i, \theta)] = \frac{1}{\sqrt{n}}m(\theta)$, so $\mathbb{E}[g_n(\theta)] = m(\theta)$ for all n.

The limit experiment thus corresponds to observing the rk-dimensional normal vector $g = (g(\theta_1)', ..., g(\theta_r)')' \sim N(m, \Sigma)$ for $m \in \mathbb{R}^{rk}$ and Σ an $(rk) \times (rk)$ matrix. The GMM estimator $\hat{\theta}$ for θ solves

$$g(\hat{\theta})'W(\hat{\theta})g(\hat{\theta}) = \min\{g(\theta_1)'W(\theta_1)g(\theta_1), ..., g(\theta_r)'W(\theta_r)g(\theta_r)\},$$

and the GMM estimator for $r(\theta)$ is $\delta^{GMM}(g, \Sigma) = r(\hat{\theta})$. □

2.1 Parameter Space for the Limit Experiment

To complete our description of the limit experiment (2) we need to specify the parameter space. As in the finite sample problem, we take the parameter space for the structural parameter θ^* to be Θ.[5] Andrews and Mikusheva (2022) show that the parameter space for the functional parameter m in the limit experiment is related to the reproducing kernel Hilbert space (RKHS) associated with Σ, which we denote by $\mathcal{H}$.[6]

Intuitively, $\mathcal{H}$ is the set of mean functions such that for any $m \in \mathcal{H}$, we cannot tell with certainty whether a given draw g was generated by $\mathcal{GP}(m, \Sigma)$ or $\mathcal{GP}(0, \Sigma)$. Since $m = 0$ corresponds to the case of complete non-identification of θ^*, $\mathcal{H}$ is thus the largest parameter space for m such that the data never rule out complete identification failure. Imposing the identifying restriction that $m(\theta^*) = 0$, the resulting joint parameter space for (θ^*, m) is

$$\Gamma = \{(\theta^*, m) : \theta^* \in \Theta, m \in \mathcal{H}, m(\theta^*) = 0\}. \tag{5}$$

We treat the structural parameter θ^* as a well-defined economic quantity that may or may not be point-identified by the moment conditions. Hence, it is meaningful to discuss the "true" value of θ^* even when m has more than one zero so θ^* is set-identified.

For the purposes of the present paper, it is helpful to work with another representation of the parameter space. Consider a mean-zero Gaussian process $G \sim \mathcal{GP}(0, \Sigma)$. Compactness of Θ and continuity of Σ together imply that G can be realized as a process

[5]The results of Andrews and Mikusheva (2022) allow a potentially smaller limiting parameter space $\Theta_0 \subseteq \Theta$. This distinction is unimportant for the results of the present paper, so we take $\Theta_0 = \Theta$.

[6]For finite sets of vectors $\{a_i\}_{i=1}^s \subset \mathbb{R}^k$ and $\{\theta_i\}_{i=1}^s \subset \Theta$, consider functions of the form $\sum_{i=1}^s \Sigma(\cdot, \theta_i) a_i$, with scalar product $\left\langle \sum_{i=1}^s \Sigma(\cdot, \theta_i) a_i, \sum_{j=1}^{s^*} \Sigma(\cdot, \theta_j^*) b_j \right\rangle_{\mathcal{H}} = \sum_{i=1}^s \sum_{j=1}^{s^*} a_i' \Sigma(\theta_i, \theta_j^*) b_j$. The RKHS $\mathcal{H}$ is the completion of $\left\{\sum_{i=1}^s \Sigma(\cdot, \theta_i) a_i : a_i \in \mathbb{R}^k, \theta_i \in \Theta, s < \infty\right\}$ under $\langle \cdot, \cdot \rangle_{\mathcal{H}}$.

with almost surely continuous paths. Denote by $\mathcal{C}^k$ the space of $\mathbb{R}^k$-valued continuous functions on Θ with norm $\|f\|_\infty = \max_{j=1,\ldots,k} \sup_{\theta\in\Theta} |f_j(\theta)|$ for $f \in \mathcal{C}^k$. Let $\mathbb{H}$ be the space of continuous linear functionals on $\mathcal{C}^k$ with the norm $\|\eta\|^* = \sup_{f\in\mathcal{C}^k, \|f\|_\infty \leq 1} |\eta(f)|$. For each $\eta \in \mathbb{H}$ we define the Pettis integral of η as $m_\eta(\cdot) \equiv \mathbb{E}[G(\cdot)\eta(G)]$. The RKHS can be represented as the image of $\mathbb{H}$ under the Pettis integral.

Lemma 1 *The image of $\mathbb{H}$ under the Pettis integral transformation coincides with the RKHS: $\mathcal{H} = \{m_\eta : \eta \in \mathbb{H}\}$. Furthermore, the transformation is continuous with $\|m_\eta\|_\infty \leq \sigma^2(G)\|\eta\|^*$, where $\sigma^2(G) = \sup_{\|\eta\|^*\leq 1} \mathbb{E}[\eta(G)^2]$ is finite.*

Hence, we may equivalently parameterize the limit experiment by $\mathbb{H}$,

$$\Gamma = \{(\theta^*, m_\eta) : \theta^* \in \Theta, \eta \in \mathbb{H}, m_\eta(\theta^*) = 0\}.$$

Bounding Identification Strength Our main result concerns parameter spaces that bound the norm of η, which we interpret as a measure of identification strength. To understand this interpretation, consider a restricted parameter space with $\|\eta\|^*$ bounded by a positive constant W,

$$\Gamma_W = \{(\theta^*, m_\eta) : \theta^* \in \Theta, \eta \in \mathbb{H}, \|\eta\|^* \leq W, m_\eta(\theta^*) = 0\}.$$

At one extreme, if $W = 0$, $\Gamma_0 = \Theta \times \{0\}$ implies that $m(\theta) = 0$ for all θ, so θ^* is completely unidentified. At the other extreme $\Gamma_\infty = \cup_W \Gamma_W = \Gamma$, so for unrestricted W we recover the original parameter space Γ. Between these two extremes, Lemma 1 shows that for any $(\theta^*, m) \in \Gamma_W$, $\|m\|_\infty \leq \sigma^2(G)W$. Since we observe only a noisy measure of m, $g(\cdot) \sim \mathcal{GP}(m, \Sigma)$, bounds on $\|m\|_\infty$ limit the ease with which we can distinguish $m(\theta)$ from 0 for any θ value and so, in that sense, limit how informative the data can be about θ^*. Thus, we can interpret Γ_W as a parameter space which imposes a uniform upper bound on the strength of identification.

Finite-Dimensional Limit Experiments In many cases of empirical interest the limit experiment is finite-dimensional, in the sense that $g(\cdot)$ can be written as a function of a finite-dimensional normal random vector or, equivalently, that the covariance function Σ has a finite number of nonzero eigenvalues.

Definition 1 *The limit experiment is finite-dimensional if the covariance function Σ has finitely many nonzero eigenvalues.*

Most of our results apply to both finite- and infinite-dimensional limit experiments, but the interpretation of some conditions is simpler in the finite-dimensional case. Finite-dimensional limit experiments can arise in many ways, for instance because the support $\mathcal{X}$ of the data is finite, because the moments are additively or multiplicatively separable in the data, $\phi(X,\theta) = \phi_1(X) - \phi_2(X)\phi_3(\theta)$, or because the parameter space is finite. Whatever the source of finite dimension, our bounds on identification strength are particularly easy to interpret in this case. Specifically, since all norms are equivalent on finite-dimensional spaces, there exists a (Σ-dependent) constant λ such that $\lambda^{-1}\|m_\eta\|_\infty \leq \|\eta\|^* \leq \lambda\|m_\eta\|_\infty$, so bounds on $\|\eta\|^*$ not only imply, but are also implied by, bounds on $\|m\|_\infty$. In infinite-dimensional settings, by contrast, bounds on $\|\eta\|^*$ imply upper, but not in general lower, bounds on $\|m\|_\infty$, so weak identification neighborhoods defined using $\|\eta\|^*$ imply that identification is "weaker" than neighborhoods defined using $\|m\|_\infty$.

Special Case: Linear IV (continued) Recall that in the linear IV model, $g(\theta) = \xi_0 - \xi_1\theta$, for (ξ_0, ξ_1) a Gaussian vector in $\mathbb{R}^{2k}$. This is therefore a finite-dimensional setting. The mean function is $m(\theta) = \pi^*(\theta^* - \theta)$, so bounding $\|m\|_\infty$ is equivalent to bounding the first stage π^*. Consequently, for π^*_η the first stage implied by η and $\|\pi^*_\eta\|$ its Euclidean norm, there exists a constant λ^* such that $\lambda^{*-1}\|\pi^*_\eta\| \leq \|\eta\|^* \leq \lambda^*\|\pi^*_\eta\|$, and bounding $\|\eta\|^*$ is equivalent to bounding the first stage coefficient π^*. □

Special Case: Finite Θ (continued) In this example the process g reduces to a Gaussian vector in $\mathbb{R}^{rk}$, so this is again a finite-dimensional case. Thus, there exists a constant λ^* for which $\lambda^{*-1}\|m_\eta\|_\infty \leq \|\eta\|^* \leq \lambda^*\|m_\eta\|_\infty$, and bounding identification strength in terms of $\|\eta\|^*$ is equivalent to bounding $\|m\|_\infty$, the maximal deviation of the moments from zero. □

3 Admissibility

Recall that the limit experiment corresponds to observing $g \sim \mathcal{GP}(m, \Sigma)$, where $(\theta^*, m) \in \Gamma$. The researcher aims to choose an estimator δ that yields a low risk $\mathbb{E}_m[L(\delta(g, \Sigma), \theta^*)]$ for the loss function L defined in (1), where since Σ is known in the limit experiment we abbreviate $\delta(g, \Sigma) = \delta(g)$ going forward. Unfortunately there is not in general a uniformly best estimator in this setting, as minimizing risk at different parameter values $(\theta^*, m), (\theta^{*\prime}, m') \in \Gamma$ usually leads to distinct estimators δ and δ'. It is without loss of performance to limit attention to the set of admissible, (i.e. undominated), estimators.

Definition 2 *An estimator δ is dominated on $\widetilde{\Gamma} \subseteq \Gamma$ if there exists another estimator δ' such that $\mathbb{E}_m[L(\delta'(g), \theta^*)] \leq \mathbb{E}_m[L(\delta(g), \theta^*)]$ for all $(\theta^*, m) \in \widetilde{\Gamma}$, with a strict inequality for some $(\theta^*, m) \in \widetilde{\Gamma}$. The estimator δ is admissible on $\widetilde{\Gamma}$ if it is not dominated on $\widetilde{\Gamma}$.*

An estimator is admissible if its performance, measured in terms of risk, cannot be uniformly improved. Since no admissible estimator dominates any other, selecting from among sets of admissible estimators requires taking a stand on how we value performance over different regions of the parameter space, for instance by specifying a prior and considering Bayes estimators as in Andrews and Mikusheva (2022). In the present paper we set a more modest goal, and aim to provide necessary conditions for admissibility under bounds on identification strength. Our main technical contribution is to establish a close connection between the set of admissible estimators under bounded identification strength and the set of estimators that are Lipschitz in g.

Definition 3 *An estimator δ is almost-surely Lipschitz with Lipschitz constant K if there exists another estimator δ^* such that $\delta(g) = \delta^*(g)$ for almost every g and $\|\delta^*(g) - \delta^*(g')\| \leq K\|g - g'\|_\infty$ for all g, g' in the support of the process $\mathcal{GP}(0, \Sigma)$.*

Theorem 1 *Assume that an estimator δ is admissible on $\widetilde{\Gamma}$, where $\widetilde{\Gamma} \subseteq \Gamma_W$ for $W < \infty$. Then δ is almost-surely Lipschitz with Lipschitz constant $K = \bar{r}\sqrt{p}W$, where $\bar{r} = \sup_\theta \|r(\theta)\|$.*

The proof of Theorem 1 builds on Theorem 2 of Andrews and Mikusheva (2022), which is itself a minor extension of a result from Brown (1986). This result, reproduced

in the appendix for completeness, shows that for convex loss functions admissible estimators must be the (almost everywhere) pointwise limit of Bayes decision rules for finitely-supported priors. We then show that under bounded identification strength, small changes in the moments g lead to only small changes in the posterior probability of different θ values. Since Bayes decision rules under squared error loss are posterior means, this implies that Bayes decision rules with finitely-supported priors are Lipschitz. Finally, we note that the Lipschitz property is preserved under pointwise convergence, from which the conclusions of the theorem follow.

It is important to emphasize that the set of admissible estimators depends on the set of parameter values $\widetilde{\Gamma}$ over which the performance is evaluated, and that the set of admissible estimators is in general not monotone in $\widetilde{\Gamma}$. That is, if we enlarge $\widetilde{\Gamma}$ the set of admissible estimators may lose some estimators but gain others. Motivated by this fact, Theorem 1 considers the set of estimators which are admissible for any set $\widetilde{\Gamma}$ that obeys a numerical bound W on identification strength.

While Theorem 1 translates numerical bounds on identification strength to numerical bounds on the Lipschitz constant, selecting a value of W for a given application seems challenging. We next provide a necessary condition for admissibility under *any* bound on identification strength.

Corollary 1 *If δ is not almost-surely Lipschitz, then it is inadmissible on $\widetilde{\Gamma}$ for all $\widetilde{\Gamma}$ with bounded identification strength (that is, $\widetilde{\Gamma} \subseteq \Gamma_W$ for some $W < \infty$).*

Corollary 1 states that under any bound on the strength of identification, no matter how large, admissible estimators are Lipschitz in the moment process g, so small changes in the realized sample moments (measured in the supremum norm $\|\cdot\|_\infty$) can induce only small changes in the estimate. While this may seem a minimal requirement, we show in the next section that GMM estimators do not have this property.

We show in Appendix B that bounded identification strength is crucial for the Lipschitz property. There, we provide an example with an unrestricted parameter space Γ where the limit of Bayes posterior means is discontinuous, and thus not Lipschitz.

3.1 Inadmissbility of GMM

We next show that GMM estimators are not generally Lipschitz, and so are inadmissible under any bound on the strength of identification. GMM estimators take the form

$$\delta^{GMM}(g, \Sigma) = r(\hat{\theta}), \quad \hat{\theta} \in \arg\min_{\theta \in \Theta} g(\theta)' W(\theta) g(\theta), \tag{6}$$

where $W(\theta)$ is a deterministic weight function. If there are multiple points where the minimum is achieved, we assume that $\hat{\theta}$ applies some selection rule.

GMM estimators are invariant to the scale of g.[7]

Definition 4 *An estimator δ is scale-invariant if $\delta(c \cdot g, \Sigma) = \delta(g, \Sigma)$ for all g and all $c > 0$.*

This scale invariance is important for our purposes, since scale-invariant estimators are Lipschitz if and only if they are constant.

Lemma 2 *Let δ be a scale-invariant estimator. If δ is almost-surely Lipschitz, then there exists $a^* \in \mathcal{A}$ such that $\delta(g, \Sigma) = a^*$ almost surely.*

GMM estimators δ^{GMM} are scale-invariant and non-constant, so Lemma 2 implies that δ^{GMM} is not Lipschitz. Hence, by Corollary 1, δ^{GMM} is inadmissible under bounded identification strength. The source of this inadmissibility is intuitive, namely that small changes in data can cause the GMM estimator to jump discontinuously.

Special Case: Linear IV (continued) In this example, δ is Lipschitz in $g(\cdot)$ if and only if it is Lipschitz in (ξ_0, ξ_1), and the two-stage least squared estimator (4) is discontinuous when the first stage estimate is zero, $\xi_1 = 0$, and hence is not Lipschitz. This is consistent with the intuition that instrumental variables estimation is badly behaved when the instrument is irrelevant.

Interestingly, if we use other instrumental variables estimators (with multiple instruments, $k > 1$) we may encounter additional points of discontinuity. For instance the limited information maximum likelihood estimator corresponds to GMM with weighting matrix $W(\theta) = (\sigma_Y^2 - 2\sigma_{DY}\theta + \sigma_D^2\theta^2)^{-1}\mathbb{E}[Z_i Z_i']^{-1}$ for σ_Y^2, σ_D^2, and σ_{DY} the residual

[7] To be precise, GMM estimators are scale-invariant so long as the rule for selecting from a non-unique argmin is likewise invariant.

variances and covariance from regressing (Y, D) on Z. The resulting estimator $\hat{\theta}$ is discontinuous at $\xi_1 = 0$, but also at (ξ_0, ξ_1) where (i) the OLS and two stage least squares estimates coincide and (ii) the reduced-form R^2 coefficient exceeds the first stage R^2, which may be interpreted as a sign of model misspecification – see Andrews (2018). □

4 Alternative Estimators

In the last section we showed that GMM estimators are inadmissible under bounds on identification strength. In particular, small changes in the moments may lead to large changes in GMM estimators. In settings with bounded identification strength, however, small changes in the moments lead only to small changes in posterior beliefs for Bayesian decision makers. Hence, the discontinuous behavior of GMM is unreasonable from a Bayesian perspective. Since a complete class theorem applies, it follows that GMM is unreasonable (specifically, inadmissible) from a frequentist perspective as well.

Unfortunately, our proof that GMM is inadmissible is non-constructive, and yields no characterization for a dominating estimator. Even were a dominating estimator known, it could depend on the bound W imposed on identification strength, which as previously discussed seems difficult to choose. The *reasons* for GMM's inadmissibility are nonetheless instructive, and suggest a route to more reasonable estimators.

The source of GMM's inadmissibility is that it depends only on the minimizer of the GMM objective. This results in the scale-invariance discussed in the last section, but also implies, for instance, that the GMM estimator is unaffected by existence of alternative, but distant, parameter values which deliver nearly the same level of the GMM objective as the minimizer. As a summary for the information contained in the data this seems an undesirable property. In this section we present two estimators which depend on the moments in a more continuous way, the first based on quasi-Bayes and the second based on bagging or bootstrap aggregation. While the admissibility of these estimators is an open question, both are continuous in the GMM moments, and Lipschitz under additional conditions.[8]

[8]To guarantee admissibility under bounded identification strength one may also report Bayes posterior means based on full-support priors on Γ_W. In the infinite-dimensional case, however, it is not obvious

4.1 Quasi-Bayes

The first alternative estimator is quasi-Bayes. For a prior π on Θ, the quasi-Bayes posterior mean of $r(\theta)$ in the limit experiment is

$$\delta_{\pi}^{QB}(g) = \frac{\int r(\theta) \exp\left(-\frac{1}{2}Q(\theta|g)\right) d\pi(\theta)}{\int \exp\left(-\frac{1}{2}Q(\theta|g)\right) d\pi(\theta)}, \tag{7}$$

where $Q(\theta|g) = g(\theta)' \Sigma(\theta,\theta)^{-1} g(\theta)$ is the GMM objective function with weight equal to the inverse of the variance.

This estimator corresponds to the posterior mean after updating $\pi(\theta)$ based on "log-likelihood" $-\frac{1}{2}Q(\theta|g)$, and was initially suggested by Chernozhukov and Hong (2003) as an estimator for settings where minimizing the GMM objective $Q(\theta|g)$ is computationally intractable (for instance due to the presence of many local minima), since $\delta_{\pi}^{QB}(g)$ can instead be computed using Bayesian numerical methods such as Markov chain Monte Carlo. Since $Q(\theta|g)$ is not in general the likelihood of the researcher's model, however, the interpretation of $\delta_{\pi}^{QB}(g)$ from a strict Bayesian perspective may not be obvious. Andrews and Mikusheva (2022) show that this estimator arises as the limit of a sequence of Bayes posterior means with proper priors. These priors imply independence between the structural parameter θ and an infinite-dimensional nuisance parameter governing the mean function m. The priors on the infinite-dimensional component are motivated from particular invariance properties, and quasi-Bayes arises from taking the diffuse (i.e. infinite-variance) limit within the resulting class. See Chernozhukov and Hong (2003) and Andrews and Mikusheva (2022) for further motivation and discussion of the quasi-Bayes approach, as well as asymptotic results under both strong and weak identification.

A key feature of the quasi-Bayes approach for our purposes is that it takes a weighted average of $r(\theta)$ over the parameter space Θ, weighting each value of θ proportionally to $\exp(-\frac{1}{2}Q(\theta))$. This directly incorporates the level of the GMM objective function, so the minimizer receives only slightly more weight than near-minimizers. Hence, if the objective function is close to flat over some region of the parameter space and high outside of this region, the quasi-Bayes estimator will correspond to a weighted average over the flat region, and will not be very sensitive to the precise location of the global minimum.

to us how to construct such priors or compute the resulting posteriors, and there further remains the question of how to choose W.

If instead the GMM objective has a well-separated minimum, the quasi-Bayes estimator will be close to the argmin. This structure implies that the quasi-Bayes estimator is continuous, and indeed Lipschitz, in the GMM objective function.

Lemma 3 *Quasi-Bayes is Lipschitz in the GMM objective function Q:*

$$\left\|\delta^{QB}(g) - \delta^{QB}(g')\right\| \leq K \left\|Q(\cdot|g) - Q(\cdot|g')\right\|_\infty,$$

where $K = \frac{1}{2}\bar{r}\sqrt{p}$.

Unfortunately, the GMM objective $Q(\cdot|g)$ is continuous but not Lipschitz in the moments g, as can be seen from the fact that $Q(\theta|g)$ is quadratic in $g(\theta)$. Consequently, the Lipschitz continuity required by Corollary 1 does not follow from Lemma 3. Indeed, while quasi-Bayes is continuous in g, the following example shows that it is not in general Lipschitz.

Special Case: Finite Θ (continued) Suppose that the parameter space consists of just two points, $\Theta = \{0, 1\}$, that we have a one-dimensional moment condition ($k = 1$), and that $\Sigma = I_2$. Consider the quasi-Bayes estimator using a prior π that puts weight $\frac{1}{2}$ on each parameter value. The quasi-Bayes estimator of θ is

$$\delta_\pi^{QB}(g) = \frac{\exp\left(-\frac{1}{2}g(1)^2\right)}{\exp\left(-\frac{1}{2}g(0)^2\right) + \exp\left(-\frac{1}{2}g(1)^2\right)} = \frac{1}{1 + \exp\left(\frac{1}{2}g(1)^2 - \frac{1}{2}g(0)^2\right)}.$$

While this estimator is differentiable in $(g(0), g(1))$, it is not Lipschitz. Indeed,

$$\left.\frac{\partial \delta_\pi^{QB}(g)}{\partial g(0)}\right|_{g(0)=g(1)} = \frac{g(0)}{4},$$

which exceeds any finite constant for large values of both $g(0)$ and $g(1)$. Intuitively, when both $g(0)$ and $g(1)$ are large, $\delta_\pi^{QB}(g)$ behaves like $\delta(g) = \arg\min_{\theta\in\{0,1\}} g(\theta)^2$. □

An interesting feature of this example is that the non-Lipschitz behavior of the quasi-Bayes estimator appears for realizations of g which suggest misspecification of the model. Specifically, the GMM model with parameter space $\Theta = \{0, 1\}$ requires that either $m(0) = 0$ or $m(1) = 0$. Hence, under the model the distribution of $\min_{\theta\in\{0,1\}} g(\theta)^2$ is bounded by a χ_1^2, and data realizations with both $g(0)$ and $g(1)$ large are highly unlikely. This suggests that if we limit attention to data realizations which appear consistent with the model the quasi-Bayes estimator may be Lipschitz. The following result shows that this is the case provided Θ is finite and π has full support.

Proposition 1 *Assume that the parameter space is finite, $|\Theta| < \infty$. For $C > 0$ define*

$$\mathcal{G}_C = \left\{ g : \inf_{\theta \in \Theta} Q\left(\theta | g\right) \leq C \right\}.$$

If π has support Θ, then the quasi-Bayes estimator $\delta_\pi^{QB}(g)$ is Lipschitz in g on $\mathcal{G}_C$.

The minimized GMM objective $Q\left(\theta|g\right)$ is often termed a J-statistic, and researchers commonly reject correct specification of the model when this statistic exceeds a threshold. Under the assumption of correct specification we have $\lim_{C\to\infty} \inf_{\gamma\in\Gamma} \mathbb{P}_\gamma \left\{ g \in \mathcal{G}_C \right\} = 1$, so moment realizations $g \notin \mathcal{G}_C$ have low probability under all data generating processes consistent with the GMM model. Hence, for finite Θ, quasi-Bayes is Lipschitz over data realizations such that the GMM model is not rejected.

Outside of the finite Θ case, bounding the J-statistic is insufficient to ensure that quasi-Bayes is Lipschitz, but one may modify quasi-Bayes to make it Lipschitz.[9] For $q > 0$, if we define the (Huber 1964-type) function $f_q(\cdot)$ by

$$f_q\left(x\right) = \begin{cases} x & \text{if } x < q \\ \sqrt{x}\sqrt{q} & \text{if } x \geq q \end{cases},$$

then $f_q(Q(\cdot|g))$ is Lipschitz in g. Hence, if we define a modified quasi-Bayes estimator $\delta_{\pi,q}^{QB}(g)$ which replaces $Q(\theta|g)$ in (7) by $f_q(Q(\cdot|g))$, the same argument used to prove Lemma 1 implies that $\delta_{\pi,q}^{QB}(g)$ is Lipschitz in g. The decision-theoretic motivation for $\delta_{\pi,q}^{QB}(g)$ is unclear, however, so in finite-dimensional settings where a fully Lipschitz estimator is desired we prefer the bagging approach discussed next.

4.2 Bagged GMM

Quasi-Bayes gives a "smoothed" analog of GMM by replacing minimization with integration. Another option is to directly smooth the GMM estimator. One way to do this is to average over bootstrap draws, which leads to the bagged, or bootstrap aggregated, GMM estimator.

[9] Since Andrews and Mikusheva (2022) show that quasi-Bayes emerges as the limit of a sequence of Bayesian estimators, it may be surprising that it does not satisfy the necessary condition for admissibility under bounded identification strength. The diffuse priors underlying quasi-Bayes imply, however, that $\|m\|_\infty \to_p \infty$, and so correspond to the case of unbounded identification strength.

We again consider the limit experiment where we observe a single draw of the moment process $g \sim \mathcal{GP}(m, \Sigma)$. For an estimator $\delta(g)$, let us draw Gaussian noise $\zeta \sim \mathcal{GP}(0, \Sigma)$ and define the bagged version of δ as the average of $\delta(g + \zeta)$ over noise realizations,

$$\delta^B(g) = \mathbb{E}\left[\delta(g + \zeta)|g\right].$$

We refer to $\delta^B(g)$ as a bagged estimator because the distribution of $g^* = g + \zeta$ given g is exactly the large-sample distribution of the moments across bootstrap replications, conditional on the initial data delivering moments g (see e.g. Section 3.6 and Van der Vaart and Wellner 1994). Hence, $\delta^B(g)$ corresponds to the (asymptotic analog of the) average of $\delta(\cdot)$ across bootstrap draws.

Bagging is a well-known smoothing approach. Bühlmann and Yu (2002) show that bagging can reduce both bias and variance when estimators are unstable, in the sense of being sensitive to small changes in the data. The discontinuity of GMM under weak identification shows that these estimators are also unstable, and suggests that their performance may be improved by bagging.

We next formalize this smoothing intuition for our setting by showing that, provided the limit experiment is finite-dimensional, any bagged estimator is Lipschitz.

Proposition 2 *If the limit experiment is finite-dimensional, for any estimator $\delta(g)$ with range contained in $\mathcal{A}$ the bagged estimator $\delta^B(g)$ is Lipschitz.*

Proposition 2 implies, in particular, that for $\delta^{GMM}(g)$ the GMM estimator as defined in (6), the bagged GMM estimator $\delta^{BGMM}(g) = \mathbb{E}[\delta^{GMM}(g + \zeta)|g]$ satisfies the global Lipschitz property required by Corollary 1. On an intuitive level, this estimator "averages out" the discontinuities of the GMM estimator, resulting in a Lipschitz (and in fact differentiable) estimator.[10]

A practical limitation of the bagged GMM estimator is that it requires repeatedly minimizing the GMM objective function to compute $\delta^{GMM}(g + \zeta)$. In settings where minimization is difficult this can make computing the bagged estimator much slower than computing e.g. the quasi-Bayes estimator. At the same time, bagged GMM is

[10]The same is true for the whole family of estimators $\delta^{BGMM}_\tau(g) = \mathbb{E}\left[\delta(g + \tau \cdot \zeta)|g\right]$ for $\tau > 0$. However, values $\tau \neq 1$ complicate the bootstrap interpretation, as well as the Bayesian interpretation discussed below, so we focus on the case with $\tau = 1$.

globally Lipschitz (in the finite-dimensional case), while quasi-Bayes is only Lipschitz under stronger conditions.

The bagged GMM estimator also has a Bayesian interpretation. In the finite-dimensional case the mean function m is simply a finite-dimensional vector. For a flat prior on m, the posterior distribution on m after observing g corresponds to a $\mathcal{GP}(g, \Sigma)$, which is precisely the distribution of $g + \zeta$. Unlike the priors underlying the quasi-Bayes approach, however, the flat prior on m allows the possibility that $m(\theta) \neq 0$ for all θ and so does not impose correct specification of the GMM model. This raises the question of how to define the object of interest for such values of m. One natural approach is to consider the GMM estimand $\theta^*(m) = \arg\min_{\theta \in \Theta} m(\theta)' W(\theta) m(\theta)$ which minimizes the population analog of the GMM objective. The bagged GMM estimator then corresponds to the posterior mean of $r(\theta^*(m))$ under the flat prior. Hence, like quasi-Bayes, bagged GMM has a Bayesian interpretation, where the difference between the two methods is that the priors underlying quasi-Bayes impose correct specification of the GMM model while the priors underlying bagged GMM do not.

5 Linear IV Simulations

While our theoretical results show that GMM estimators are dominated under bounds on identification strength, they do not imply that GMM is dominated by either quasi-Bayes or bagged GMM. Relative performance of these estimators in applications is thus an open question. In this section we explore this comparison in the context of linear IV, using simulation designs from Andrews et al. (2019).

Andrews et al. (2019) calibrate simulations based on all instrumental variables specifications published in the American Economic Review from 2014 to 2018 for which sufficient information is available to estimate the variance matrix Ω in (3), yielding 124 specifications. Of these, 34 specifictions are just-identified ($k = 1$), 20 have $k = 2$, 30 have $k = 3$, and the remaining 40 have $k \geq 4$. For each of these 124 specifications, we simulate data from the normal model (3) with π^* equal to the first stage estimate in the Andrews et al. (2019) data and θ^* equal to the two-stage least squares estimate. See Andrews et al. (2019) for further details on the data and simulation design.

For consistency with our theoretical results we restrict attention to bounded parameter spaces, taking the parameter space in specification s equal to $\Theta_s = [\theta_s^* \pm 10\sigma_s^*]$ for σ_s^* the standard error in the Andrews et al. (2019) data. We consider six different estimators. The first two are GMM estimators, specifically two stage least squares, which as noted above corresponds to GMM with weighting matrix $W(\theta) = E[Z_i Z_i']^{-1}$, and continuously updating GMM, which corresponds to GMM with weighting matrix $W(\theta) = \Sigma(\theta)^{-1}$. The second two are quasi-Bayes posterior means, including quasi-Bayes with a flat prior $\pi(\theta) \propto 1$, and quasi-Bayes with a novel invariant prior. This invariant prior is motivated by the idea, similar to arguments for the Jeffreys (1946) prior for parametric models, that a non-informative prior should be invariant across different representations of the same problem.[11] Our final two estimators are bagged versions of the two GMM estimators, where we approximate the expectation by averages over 400 draws of the noise ζ. We report results based on 10,000 simulation draws for all estimators.

Table 1 reports our findings. For each specification s we consider the root mean squared error for that estimator, normalized by the standard error σ_s^* to account for differences in units, $\sqrt{\mathbb{E}_s\left[(\delta(g) - \theta_s^*)^2\right]}/\sigma_s^*$, where $\mathbb{E}_s[\cdot]$ denotes the expectation in specification s. We report the average of this ratio for each estimator across four different categories based on the effective first stage F statistic of Montiel-Olea and Pflueger (2013). The effective F statistic, which we denote by F, is a measure of instrument strength and in the just-identified case is equal to the squared t-statistic for testing $\pi^* = 0$, $\xi_1^2/Var(\xi_1)$. See Montiel-Olea and Pflueger (2013) for details and motivation for this statistic. To complement these results, Figure 1 plots the root mean squared error for each alternative estimator, relative to its GMM counterpart, against the average effective F statistic $\mathbb{E}_s[F]$, limiting attention to specifications where $\mathbb{E}_s[F] \leq 50$ for visibility.

A number of patterns emerge in Table 1 and Figure 1. First, two stage least squares outperforms continuously updating GMM in almost all cases. Second, quasi-Bayes with a flat prior underperforms all the other estimators considered, while quasi-Bayes with

[11] For $\vartheta : \Theta \to \Psi$ an invertible function and $B(\psi)$ everywhere full-rank, $h(\cdot) = B(\cdot)g(\vartheta^{-1}(\cdot))$ is a one-to-one transformation of g. Hence, it seems reasonable to require that a non-informative prior be "the same" in both cases, but the flat prior generally will not be. Our invariant prior, by contrast, guarantees this property when θ is scalar and $\vartheta(\cdot)$, $B(\cdot)$ are differentiable – see Appendix C for details.

	$\mathbb{E}_s[F] \leq 10$	$10 < \mathbb{E}_s[F] \leq 20$	$20 < \mathbb{E}_s[F] \leq 50$	$50 < \mathbb{E}_s[F]$
Two Stage Least Squares	1.21	1.17	1.02	1.00
Continuously Updating GMM	1.38	1.22	1.04	0.99
Quasi-Bayes, Flat Prior	1.42	1.59	1.15	1.00
Quasi-Bayes, Invariant Prior	1.00	1.09	1.01	0.98
Bagged Two Stage Least Squares	1.03	1.15	1.03	0.99
Bagged Continuously Updating GMM	1.04	1.18	1.06	0.99
Number of Specifications	56	30	20	18

Table 1: Estimator performance in Andrews et al. (2019) specifications. Entries correspond the root mean squared error normalized by the standard error in the Andrews et al. (2019) data, $\sqrt{\mathbb{E}_s[(\delta(g) - \theta_s^*)^2]}/\sigma_s^*$, averaged across specifications. Columns correspond to ranges of values for the average effective first-stage F statistic of Montiel-Olea and Pflueger (2013).

the invariant prior outperforms in all cases. The performance gap between the two quasi-Bayes approaches demonstrates the influence of the prior, and that the greater smoothness of quasi-Bayes as a function of the moment functions does not guarantee improved performance for all priors. By contrast, the bagged GMM estimators, which do not require specification of a prior, each outperform their standard GMM analogs in most cases. Specifically, these estimators show substantial improvements in the category where identification is weakest ($\mathbb{E}_s[F] \leq 10$), a smaller improvement in the next-weakest category ($10 < \mathbb{E}_s[F] \leq 20$), and a small deterioration in performance in the second-strongest category ($20 < \mathbb{E}_s[F] \leq 50$). All estimators considered have essentially indistinguishable performance in the strongest category ($50 < \mathbb{E}_s[F]$).

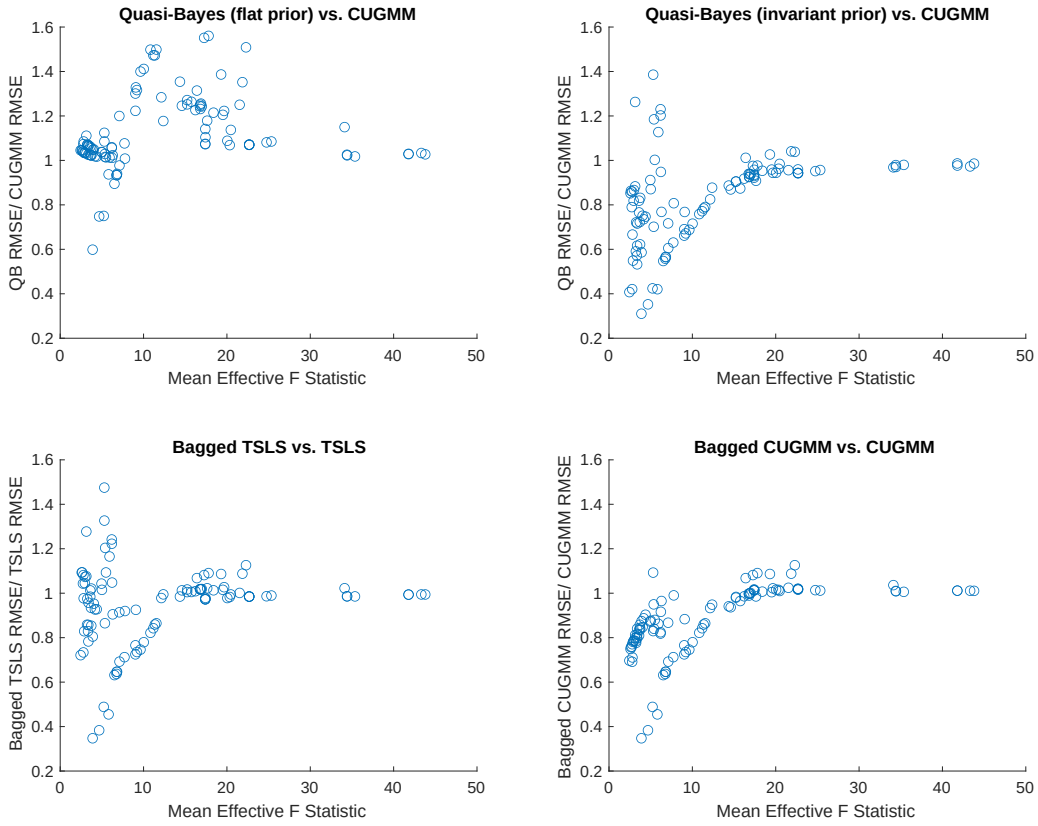

Figure 1: Estimator performance in Andrews et al. (2019) specifications. Each point corresponds to one of the Andrews et al. (2019) specifications. The vertical axis measures the ratio of root mean squared error for the alternative estimator compared to GMM, $\sqrt{\mathbb{E}_s[(\delta(g) - \theta_s^*)^2]/\mathbb{E}_s[(\delta^{GMM}(g) - \theta_s^*)^2]}$. So, for instance, a value of 0.8 means the RMSE for the alternative estimator is 20% lower. The horizontal axis shows the average effective first-stage F statistic of Montiel-Olea and Pflueger (2013), $\mathbb{E}_s[F]$. We limit attention to specifications with $\mathbb{E}_s[F] \leq 50$ for visibility.

6 References

Adler, R.J., and J.E. Taylor (2007): *Random Fields and Geometry*, New York : Springer

Andrews, D.W.K. and X. Cheng (2012): "Estimation and Inference with Weak, Semi-strong and Strong Identification," *Econometrica*, 80(5), 2153-2211.

Andrews, I. (2019): "On the Structure of IV Estimands," *Journal of Econometrics*, 211(1), 294-307.

Andrews, I. and A. Mikusheva (2022): "Optimal Decision Rules for Weak GMM," *Econometrica*, 90(2), 715-748.

Andrews, I., J. Stock, and L. Sun, (2019): "Weak Instruments in IV Regression: Theory and Practice." *Annual Review of Economics*, 11, 727-753.

Armstrong, T. (2016): "Large Market Asymptotics for Differentiated Product Demand Estimators With Economic Models of Supply," *Econometrica*, 84(5), 1961-1980.

Bühlmann, P. and B. Yu (2002): "Analyzing bagging," *The Annals of Statistics*, 30(4), 927-961.

Brown, L.D. (1986): *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, Hayward, CA: Institute of Mathematical Statistics

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions," *Journal of Econometrics*, 34(3), 305-334.

Chen, S.X. and P. Hall (2003): "Effects of Bagging and Bias Correction on Estimators Defined by Estimating Equations," *Statistica Sinica*, 13(1), 97-109

Chernozhukov, V. and H. Hong (2003): "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115(2), 293-346

Hansen, L.P. (1982) : "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029-1054.

Horowitz, J.L. (2001): "The Bootstrap," Handbook of Econometrics, Volume 5, Eds. James J. Heckman, Edward Leamer, 3159-3228.

Huber, P.J. (1964): "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35(1), 73-101.

Jeffreys, H. (1946): "An Invariant Form of the Prior in Estimation Problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*,

186(1007), 453-461.

Mavroeidis, S., M. Plagborg-Møller, and J.H. Stock (2014): "Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve." *Journal of Economic Literature* 52(1): 124-188.

Montiel-Olea J. and C. Pflueger C (2013): "A Robust Test for Weak Instruments." *Journal of Business and Economic Statistics* 31(3): 358-369.

Staiger, D. and J.H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557-586.

Stock, J.H. and J. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055-96.

Van der Vaart, A.W. and J.A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer

Van der Vaart, A.W. and H. Van Zanten (2008): "Reproducing Kernel Hilbert Spaces of Gaussian Processes," in *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, Bertrand Clarke and Subhashis Ghosal, eds., (Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2008), 200-222

A Proofs

Proof of Lemma 1 If $k = 1$, the result is immediate from Theorem 2.1 of van der Vaart and van Zanten (2008). We are left to prove it for $k > 1$. Define an augmented parameter space $\Theta^* = \Theta \times V$, where $V = \{v \in \mathbb{R}^k : \|v\|_1 = 1\}$ and consider a Gaussian process $g^*(\cdot)$ defined on Θ^* as $g^*(\theta, v) = v'g(\theta)$. Note that the process g^*, its mean $m^*(\theta, v)$ and its covariance function $\Sigma^*(\theta, v, \tilde{\theta}, \tilde{v})$ are one-to-one transformations of g, m, and Σ. For $\mathcal{H}^*$ the RKHS associated with Σ^*, $\mathcal{H}^*$ is isometric to $\mathcal{H}$. Indeed, for $m^* \in \mathcal{H}^*$:

$$m^*(\theta, v) = \sum \alpha_i \Sigma^*(\theta_i, v_i, \theta, v) = \left(\sum \alpha_i v_i' \Sigma(\theta_i, \theta)\right) v = m(\theta)'v, \tag{8}$$

where $m \in \mathcal{H}$ and $\|m\|_{\mathcal{H}} = \|m^*\|_{\mathcal{H}^*}$.

Since Θ^* is compact and Σ^* is continuous, Lemma 1.3.1 of Adler and Taylor (2007) implies that process $G^*(\cdot) \sim \mathcal{GP}(0, \Sigma^*)$ can be realized as a process with almost surely continuous sample paths. Let $\mathcal{C}^*$ be the space of $\mathbb{R}$-valued continuous functions on Θ^*

with the property that any $f^* \in \mathcal{C}^*$ can be represented as $f^*(\theta, v) = f(\theta)'v$ for $f \in \mathcal{C}^k$ and $v \in V$. Due to the structure of Σ^*, realizations of the process G^* almost surely belong to $\mathcal{C}^*$ and the process can be represented as $G^*(\theta, v) = v'G(\theta)$, where $G \sim \mathcal{GP}(0, \Sigma)$. Take any linear functional defined on the space of continuous functions with index set Θ^* and denote by η^* its restriction to $\mathcal{C}^*$. Since the relation between $f^* \in \mathcal{C}^*$ and $f \in \mathcal{C}^k$ is one-to-one, we can define a linear functional on $\mathcal{C}^k$ as $\eta(f) = \eta^*(f^*)$. This creates a one-to-one correspondence between linear functionals on $\mathcal{C}^*$ and linear functionals on $\mathcal{C}^k$. Note that the definition of the Pettis integral for process G^* depends on η^* only and all functionals that are the same once restricted to $\mathcal{C}^*$ lead to the same Pettis integral:

$$m^*_{\eta^*}(\theta, v) = \mathbb{E}[G^*(\theta, v)\eta^*(G^*)] = \mathbb{E}[v'G(\theta)\eta^*(G^*)] = v'\mathbb{E}[G(\theta)\eta(G)] = v'm_\eta(\theta). \tag{9}$$

Due to Theorem 2.1 of van der Vaart and van Zanten (2008), $\mathcal{H}^*$ coincides with the image of the space of linear functionals defined on $\mathcal{C}^*$ under the Pettis integral transformation and $\|m_{\eta^*}\|_\infty \leq \sigma^2(G^*)\|\eta^*\|^*$. Comparing equation (9) to (8), we see that the first statement of Lemma 1 holds. We further notice that all norms of starred objects coincide with the norms of the corresponding objects without stars. For example,

$$\|m\|_\infty = \sup_{j,\theta\in\Theta} |m_j(\theta)| = \sup_{v\in V,\theta\in\Theta} |v'm(\theta)| = \sup_{(\theta,v)\in\Theta^*} |m^*(\theta, v)| = \|m^*\|_\infty.$$

Note further that η and η^* have the same total variation norm.

$$\|\eta^*\|^* = \sup_{f^*\in\mathcal{C}^*, \|f^*\|_\infty\leq 1} \eta^*(f^*) = \sup_{f\in\mathcal{C}^k, \|f\|_\infty\leq 1} \eta(f) = \|\eta\|^*.$$

Finally,

$$\sigma^2(G^*) = \sup_{\|\eta^*\|^*\leq 1} \mathbb{E}[\eta^*(G^*)^2] = \sup_{\|\eta\|^*\leq 1} \mathbb{E}[\eta(G)^2] = \sigma^2(G).$$

This completes the proof. □

Theorem 2 *(Brown 1986, Andrews and Mikusheva 2022): For any parameter space $\widetilde{\Gamma} \subseteq \Gamma$, any loss $L(a, \theta)$ which is convex in a for all θ, and any decision rule δ that is admissible on $\widetilde{\Gamma}$, there exists a sequence of finitely supported priors π_r on $\widetilde{\Gamma}$ and corresponding Bayes decision rules δ_{π_r},*

$$\int \mathbb{E}_m[L(\delta_{\pi_r}(g), \theta^*)]d\pi_r(\theta^*, m) = \min_{\tilde{\delta}} \int \mathbb{E}_m[L(\tilde{\delta}(g), \theta^*)]d\pi_r(\theta^*, m),$$

such that $\delta_{\pi_r}(g) \to \delta(g)$ as $r \to \infty$ for almost every g.

Proof of Theorem 1 First consider $\widetilde{\Gamma} \subseteq \Gamma_W$. We show that for any finitely-supported prior π on $\widetilde{\Gamma}$, $\mathbb{E}_\pi[r(\theta)|g]$ is Lipschitz in g:

$$\|\mathbb{E}_\pi[r(\theta)|g=w] - \mathbb{E}_\pi[r(\theta)|g=w']\| \leq KW\|w-w'\|_\infty.$$

Let $\{(\theta_1, m_1), ..., (\theta_J, m_J)\}$ be the support of π. For each m_j we know from Lemma 1 that there exists $\eta_{m,j} \in \mathbb{H}$ with $\|\eta\|^* \leq W$ and $m_j(\cdot) = E[G(\cdot)\eta_{m,j}(G)]$. Further note that by e.g. Lemma 3.1 of van der Vaart and van Zanten (2008), for each $m \in \mathcal{H}$ the likelihood ratio relative to $m' = 0$ takes the form

$$\frac{dQ_m}{dQ_0}(g) = \exp\left(\eta_m(g) - \frac{1}{2}\|m\|^2_{\mathcal{H}}\right).$$

Define $\tilde{w} = w' - w$, and let $w_t = w + t \cdot \tilde{w}$. Note that

$$\mathbb{E}_\pi[r(\theta)|g=w_t] = \frac{\sum_j r(\theta_j)\exp\left\{\eta_{m,j}(w_t) - \frac{1}{2}\|m_j\|^2_{\mathcal{H}}\right\}\pi(\theta_j, m_j)}{\sum_j \exp\left\{\eta_{m,j}(w_t) - \frac{1}{2}\|m_j\|^2_{\mathcal{H}}\right\}\pi(\theta_j, m_j)}. \tag{10}$$

Linearity implies that $\eta_{m,j}(w_t) = \eta_{m,j}(w) + t\eta_{m,j}(\tilde{w})$, and thus

$$\frac{\partial}{\partial t}\exp\left\{\eta_{m,j}(w_t) - \frac{1}{2}\|m_j\|^2_{\mathcal{H}}\right\} = \eta_{m,j}(\tilde{w})\exp\left\{\eta_{m,j}(w_t) - \frac{1}{2}\|m_j\|^2_{\mathcal{H}}\right\}.$$

By differentiating (10) we get

$$\frac{\partial}{\partial t}\mathbb{E}_\pi[r(\theta)|g=w_t] = Cov_\pi(r(\theta), \eta_m(\tilde{w})|g=w_t),$$

where the only posterior uncertainty about $\eta_m(\tilde{w})$ comes from the unknown parameter m, while $\tilde{w}$ is fixed. Cauchy-Schwarz implies that

$$\|Cov_\pi(r(\theta), \eta_m(\tilde{w})|g=w_t)\| \leq \bar{r}\sqrt{p}\sqrt{Var(\eta_m(\tilde{w})|g=w_t)}.$$

For $(\theta_j, m_j) \in \Gamma_W$ we have $\|\eta_{m,j}\|^* \leq W$, thus

$$|\eta_{m,j}(\tilde{w})| \leq W\|\tilde{w}\|_\infty = W\|w-w'\|_\infty,$$

which implies that $\sqrt{Var(\eta_m(\tilde{w})|g=w_t)} \leq W\|w-w'\|_\infty$. Hence,

$$\|\mathbb{E}_\pi[r(\theta)|g=w] - \mathbb{E}_\pi[r(\theta)|g=w']\| =$$

$$\left\|\int_0^1 \frac{\partial}{\partial t}\mathbb{E}_\pi[r(\theta)|g=w_t]\,dt\right\| \leq \bar{r}\sqrt{p}W\|w-w'\|_\infty.$$

We next show that if a sequence of Bayes estimators δ_{π_s} converges almost-everywhere pointwise to δ, then δ must be almost-everywhere Lipschitz. Indeed, for almost every pair w and w' we have

$$\left(\delta_{\pi_s}(w), \delta_{\pi_s}(w')\right) \to \left(\delta^*(w), \delta^*(w')\right).$$

Hence, $\delta_{\pi_s}(w) - \delta_{\pi_s}(w') \to \delta^*(w) - \delta^*(w')$. Since

$$\left\|\delta_{\pi_s}(w) - \delta_{\pi_s}(w')\right\| \leq \bar{r}\sqrt{p}W \left\|w - w'\right\|_\infty$$

for all s, $\left\|\delta^*(w) - \delta^*(w')\right\| \leq \bar{r}\sqrt{p}W \left\|w - w'\right\|_\infty$ as well.

Further, note that for any w in the support of g and any $\varepsilon > 0$, there exists a $\tilde{w}$ with $\left\|w - \tilde{w}\right\|_\infty < \varepsilon$ and $\delta_{\pi_s}(\tilde{w}) \to \delta(\tilde{w})$. As we proved, $\limsup_{s\to\infty} \left\|\delta_{\pi_s}(w) - \delta(\tilde{w})\right\| \leq \bar{r}\sqrt{p}W\varepsilon$. Since we can repeat this argument for all ε, we see that $\delta_{\pi_s}(w)$ has a limit. Define $\delta^*(\cdot)$ as the pointwise limit of $\delta_{\pi_s}(\cdot)$, and note that the same argument as used above shows that δ^* is everywhere Lipschitz with Lipschitz constant $\bar{r}\sqrt{p}W$. By construction, $\delta^*(w) = \delta(w)$ for almost every w. □

Proof of Corollary 1 Immediate from Theorem 1. □

Proof of Lemma 2 Suppose that δ is both almost-surely Lipschitz and scale-invariant. Consider two independent draws g and g', and note that by scale-invariance we have

$$\delta(g, \Sigma) - \delta(g', \Sigma) = \delta(c \cdot g, \Sigma) - \delta(c \cdot g', \Sigma)$$

for all $c > 0$. However, the Lipschitz property implies there exists a constant K such that for almost every (g, g') and any fixed c,

$$\left\|\delta(c \cdot g, \Sigma) - \delta(c \cdot g', \Sigma)\right\| \leq cK \cdot \left\|g - g'\right\|_\infty$$

with probability one. Hence, $\mathbb{E}[\|\delta(g, \Sigma) - \delta(g', \Sigma)\|] \leq cK \cdot \mathbb{E}[\|g - g'\|_\infty]$. Since $\mathbb{E}[\|g - g'\|_\infty] = \sqrt{2}\mathbb{E}[\|G\|_\infty]$ is finite when $G \sim \mathcal{GP}(0, \Sigma)$, it follows that $\mathbb{E}[\|\delta(g, \Sigma) - \delta(g', \Sigma)\|] = 0$, and we may take $a^* = \mathbb{E}[\delta(g, \Sigma)]$ to complete the proof □

Proof of Lemma 3 Let $Q(\theta) = Q(\theta|g)$ and $Q'(\theta) = Q(\theta|g')$, and consider

$$Q_t(\theta) = Q(\theta) + t(Q'(\theta) - Q(\theta)) = Q(\theta) + t \cdot \Delta(\theta)$$

for $\Delta(\theta) = Q'(\theta) - Q(\theta)$. Let us write $\mathbb{E}_\pi^{QB}[\cdot|Q]$ for the expectation under the quasi-Bayes posterior distribution, which draws θ from the distribution with density $\frac{\exp(-\frac{1}{2}Q(\theta))}{\int \exp(-\frac{1}{2}Q(\theta))d\pi(\theta)}$ relative to π, and define $Cov_\pi^{QB}(\cdot,\cdot|Q)$ analogously. Note that $\delta_\pi^{QB}(g) = \mathbb{E}_\pi^{QB}[r(\theta)|Q(\cdot|g)]$, and that

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_\pi^{QB}[r(\theta)|Q_t] &= \frac{\partial}{\partial t}\left[\frac{\int r(\theta)\exp\left(-\frac{1}{2}Q_t(\theta)\right)d\pi(\theta)}{\int \exp\left(-\frac{1}{2}Q_t(\theta)\right)d\pi(\theta)}\right] \\ &= -\frac{1}{2}\left(\mathbb{E}_\pi^{QB}[r(\theta)\Delta(\theta)|Q_t] - \mathbb{E}_\pi^{QB}[r(\theta)|Q_t]\mathbb{E}_\pi^{QB}[\Delta(\theta)|Q_t]\right) \\ &= -\frac{1}{2}Cov_\pi^{QB}\left(r(\theta), \Delta(\theta)|Q_t\right)\end{aligned}$$

By the Cauchy-Schwarz inequality, however,

$$\|Cov_\pi^{QB}\left(r(\theta), \Delta(\theta)|Q_t\right)\| \leq \bar{r}\sqrt{p}\sup_\theta |\Delta(\theta)|,$$

which completes the proof. □

Proof of Proposition 1 Similar to the proof of Lemma 3, let $\mathbb{E}_\pi^{QB}[\cdot|g]$ be the expectation under the quasi-Bayes posterior distribution, which draws θ from the distribution with density $\frac{\exp(-\frac{1}{2}Q(\theta|g))}{\int \exp(-\frac{1}{2}Q(\theta|g))d\pi(\theta)}$ relative to π, and define $Cov_\pi^{QB}(\cdot,\cdot|g)$, $Var_\pi^{QB}(\cdot|g)$ analogously. For $g_t = g + t \cdot \tilde{g}$, note that by Cauchy-Schwarz

$$\left\|\frac{\partial}{\partial t}\delta_\pi^{QB}(g_t)\right\| = \left\|\frac{\partial}{\partial t}\mathbb{E}_\pi^{QB}\left[r(\theta)|g_t\right]\right\| = \frac{1}{2}\left\|Cov_\pi^{QB}\left(r(\theta), \tilde{g}(\theta)'\Sigma(\theta)^{-1}g_t(\theta)|g_t\right)\right\| \leq$$

$$\frac{1}{2}\bar{r}\sqrt{p}\sqrt{\mathbb{E}_\pi^{QB}\left[\left(\tilde{g}(\theta)'\Sigma(\theta)^{-1}g_t(\theta)\right)^2|g_t\right]}.$$

By another application of Cauchy-Schwarz,

$$\left(\tilde{g}(\theta)'\Sigma(\theta)^{-1}g_t(\theta)\right)^2 \leq \tilde{g}(\theta)'\Sigma(\theta)^{-1}\tilde{g}(\theta) \cdot Q(\theta|g_t),$$

so

$$\mathbb{E}_\pi^{QB}\left[\left(\tilde{g}(\theta)'\Sigma(\theta)^{-1}g_t(\theta)\right)^2|g_t\right] \leq \|\tilde{g}\|_{\Sigma,\infty}^2\,\mathbb{E}_\pi^{QB}\left[Q(\theta|g_t)|g_t\right]$$

for

$$\|\tilde{g}\|_{\Sigma,\infty} = \sup_{\theta\in\Theta}\sqrt{\tilde{g}(\theta)'\Sigma(\theta)^{-1}\tilde{g}(\theta)} = \sup_{\theta\in\Theta}\sqrt{Q(\theta|\tilde{g})}.$$

Altogether, we obtain that

$$\left\| \frac{\partial}{\partial t} \mathbb{E}_{\pi}^{QB} \left[r\left(\theta\right) |g_t \right] \Big|_{t=0} \right\| \leq \frac{1}{2} \bar{r} \sqrt{p} \left\| \tilde{g} \right\|_{\Sigma,\infty} \sqrt{\mathbb{E}_{\pi}^{QB} \left[Q(\theta|g) |g \right]}.$$

Note, next, that

$$\mathbb{E}_{\pi}^{QB} \left[Q(\theta|g) |g \right] = \frac{\int Q(\theta|g) \exp\left(-\frac{1}{2} Q(\theta|g)\right) d\pi\left(\theta\right)}{\int \exp\left(-\frac{1}{2} Q(\theta|g)\right) d\pi\left(\theta\right)}.$$

Since the function $h\left(x\right) = x \exp\left(-\frac{1}{2}x\right)$ is maximized at $x = 2$, if

$$\int \exp\left(-\frac{1}{2} Q(\theta|g)\right) d\pi\left(\theta\right) \geq \varepsilon,$$

then

$$E_{\pi} \left[Q(\theta|g) |g \right] \leq 2 \exp\left(-1\right) \varepsilon^{-1}.$$

Note, however, that for $|\Theta|$ finite and $\underline{\pi} = \min_{\theta \in \Theta} \pi(\theta)$,

$$\int \exp\left(-\frac{1}{2} Q(\theta|g)\right) d\pi\left(\theta\right) \geq \exp\left(-\frac{1}{2} \min_{\theta \in \Theta} Q(\theta|g)\right) \underline{\pi}.$$

Hence, for $g \in \mathcal{G}_C$ as defined in the proposition,

$$\left\| \frac{\partial}{\partial t} \mathbb{E}_{\pi}^{QB} \left[r\left(\theta\right) |g_t \right] \Big|_{t=0} \right\| \leq \bar{r} \sqrt{p} \left\| \tilde{g} \right\|_{\Sigma,\infty} \exp\left(\frac{1}{2} C - 1\right) \underline{\pi}^{-1},$$

which completes the proof. □

Proof of Proposition 2 If the covariance function Σ has a finite number of nonzero eigenvalues, it follows that for $G \sim \mathcal{GP}(0, \Sigma)$ the process $G\left(\cdot\right)$ is a transformation of a finite-dimensional normal random vector, so we can write $G\left(\theta\right) = A\left(\theta\right) Y$ for $Y \in \mathbb{R}^q$ a standard normal random vector and $A\left(\cdot\right)$ a matrix-valued function which depends on Σ. Correspondingly, the RKHS $\mathcal{H}$ can be written as $\left\{ A\left(\cdot\right) x : x \in \mathbb{R}^q \right\}$.

Combining these observations, we can write $g\left(\cdot\right) = A\left(\cdot\right) y$ for $y \sim N\left(x, I\right)$, and any estimator $\delta\left(g\right)$ can be equivalently expressed as $\gamma\left(y\right) = \delta\left(A\left(\cdot\right) y\right)$. For υ a standard normal random vector and ζ as defined in the main text, we likewise have the equality

$$\delta^B \equiv \mathbb{E}\left[\delta\left(g + \zeta\right) |g\right] = \mathbb{E}\left[\gamma\left(y + \upsilon\right) |y\right] \equiv \gamma^B\left(y\right)$$

for the bagged estimators. Since $\Sigma\left(\theta, \tilde{\theta}\right) = A\left(\theta\right) A\left(\tilde{\theta}\right)'$ while Σ is continuous and Θ is compact, the largest singular value of $A\left(\theta\right)$, $\sigma_{\max}\left(A\left(\theta\right)\right)$, is uniformly bounded. For $\bar{\sigma} = \sup_{\theta \in \Theta} \sigma_{\max}\left(A\left(\theta\right)\right)$,

$$\sup_{\theta} \left\| g\left(\theta\right) - \tilde{g}\left(\theta\right) \right\| \leq \sup_{\theta} \sigma_{\max}\left(A\left(\theta\right)\right) \left\| y - \tilde{y} \right\| \leq \bar{\sigma} \left\| y - \tilde{y} \right\|.$$

Hence, it suffices to show that $\gamma^B(y)$ is Lipschitz in y.

Note that for φ the standard (multivariate) normal density,

$$\gamma^B(y) = \int \gamma(y+\upsilon)\,\varphi(\upsilon)\,d\upsilon = \int \gamma(\upsilon)\,\varphi(\upsilon-y)\,d\upsilon.$$

Hence, if we let $y_t = y + t\cdot\tilde{y}$, we have

$$\frac{\partial}{\partial t}\,\gamma^B(y_t)\big|_{t=0} = \int \gamma(\upsilon)\,\frac{\partial}{\partial t}\,\varphi(\upsilon-y_t)\big|_{t=0}\,d\upsilon$$

$$= \int \gamma(\upsilon)\,(\upsilon-y)'\,\tilde{y}\varphi(\upsilon-y)\,d\upsilon = Cov_{N(y,I)}\left(\gamma(\upsilon),\upsilon'\tilde{y}\right),$$

where $Cov_{N(y,I)}\left(\gamma(\upsilon),\upsilon'\right)$ denotes the covariance of $\gamma(\upsilon)$ and υ when $\upsilon\sim N(y,I)$. Note, however, that since the range of $\gamma(\upsilon)$ is contained in a, Cauchy-Schwarz implies that for $\bar{a} = \sup_{a\in\mathcal{A}}\|a\|$,

$$\left\|\frac{\partial}{\partial t}\,\gamma^B(y_t)\big|_{t=0}\right\| = \left\|Cov_{N(y,I)}\left(\gamma(\upsilon),\upsilon'\tilde{y}\right)\right\| \le \bar{a}\sqrt{p}\,\|\tilde{y}\|,$$

which completes the proof. □

B Example: Discontinuous Limit-of-Bayes

This appendix provides an example to demonstrate that without bounds on identification strength, the pointwise limit of Bayes posterior means can be discontinuous.

Let us continue the finite Θ special case discussed in main text, and further suppose that the parameter space consists of just two points, $\Theta = \{\theta_1,\theta_2\}$, that we have a one-dimensional moment condition ($k=1$), and that $\Sigma = I_2$. The function m is thus described by two numbers – the values at θ_1 and θ_2. Consider prior π_C, supported on just two values of (θ,m), which assigns probability $\frac{1}{2}$ to each of θ_1 and θ_2 and, conditional on $\theta=\theta_j$, implies that $(m(\theta_j),m(\theta_{-j})) = (0,C)$ with probability one, where θ_{-j} denotes the element of Θ other than θ_j. Suppose we are interested in estimating $r(\theta)=\theta$ and note that the Bayes decision corresponds to the posterior mean

$$\delta_{\pi_C}(g,\Sigma) = \mathbb{E}_{\pi_C}[\theta|g] = \frac{\sum_{j=1}^2 \theta_j \exp(-\frac{1}{2}g(\theta_j)^2 - \frac{1}{2}(g(\theta_{-j})-C)^2)}{\sum_{j'=1}^2 \exp(-\frac{1}{2}g(\theta_{j'})^2 - \frac{1}{2}(g(\theta_{-j'})-C)^2)}.$$

For a given g, $\mathbb{E}_{\pi_C}[\theta|g] \to \arg\min_{\theta\in\{\theta_1,\theta_2\}} g(\theta)$ as $C\to\infty$. Note that the limiting estimator $\delta(g) = \arg\min_{\theta\in\{\theta_1,\theta_2\}} g(\theta)$ is discontinuous for realizations where $g(\theta_1)=g(\theta_2)$ and is therefore not Lipschitz for any Lipschitz constant.

C Invariant Prior

From a subjective Bayesian perspective the prior π on the GMM parameter θ should reflect the researcher's beliefs about the structural parameters in a given application. In practice, however, subjective priors may be difficult to specify or controversial, and it may be helpful to have default options.

One common default is to use a flat prior, with π proportional to Lebesgue measure. As has been noted in many settings, however, "flatness" of a prior is specific to a given parameterization, so the default of "use a flat prior" can lead to different conclusions depending on the parameterization used. In likelihood settings, the desire for a parameterization-invariant default prior has led to the use of the Jeffreys (1946) prior, which ensures such invariance in parametric models.

Our goal in this section is to develop a prior with analogous invariance properties for our setting. Specifically, since we treat the covariance function Σ as known, we seek a default prior $\pi(\cdot; \Sigma)$ on Θ which is invariant to certain transformations of the problem. To describe the invariance we seek, we assume θ is scalar and that $\Sigma\left(\theta, \tilde{\theta}\right)$ is continuously differentiable in both arguments. Let Ψ be a compact set, and let $\vartheta : \Theta \to \Psi$ be a diffeomorphism between Θ and Ψ, corresponding to a reparameterization of the model. Further, let $\mathcal{B}$ denote the set of full-rank $k \times k$ matrices, and let $B : \Psi \to \mathcal{B}$ be a differentiable function from Ψ to $\mathcal{B}$. For each such (ϑ, B) pair we can define a new moment process

$$h(\cdot) = B(\cdot)\, g\left(\vartheta^{-1}(\cdot)\right)$$

with domain Ψ, where by construction $h(\cdot) \sim \mathcal{GP}(m_h, \Sigma_h)$ for

$$m_h(\psi) = B(\psi)\, m\left(\vartheta^{-1}(\psi)\right), \;\; \Sigma_h(\psi_1, \psi_2) = B(\psi_1)\, \Sigma\left(\vartheta^{-1}(\psi_1), \vartheta^{-1}(\psi_2)\right) B(\psi_2)'.$$

Since $B(\psi)$ has full rank by definition, $(g(\cdot), \Sigma(\cdot,\cdot))$ and $(h(\cdot), \Sigma_h(\cdot,\cdot))$ are one-to-one transformations of each other. Thus, observing $g(\cdot)$ with Σ known is in a strong sense equivalent to observing h with Σ_h known. We require that the default priors for these two problems also be equivalent, in the sense that the pushforward for $\pi(\cdot; \Sigma)$ under $\vartheta(\cdot)$ must equal $\pi(\cdot; \Sigma_h)$ for all $(\vartheta(\cdot), B(\cdot))$ pairs.

We verify below that the prior

$$\pi(\theta;\Sigma) = |\Sigma(\theta,\theta)|^{-\frac{1}{2}} \left| \frac{\partial^2}{\partial\theta\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right) - \frac{\partial}{\partial\theta}\Sigma\left(\theta,\tilde{\theta}\right)\Sigma(\theta,\theta)^{-1}\frac{\partial}{\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right)\right|_{\tilde{\theta}=\theta}^{\frac{1}{2}},$$

for $|A|$ the absolute value of the determinant of the matrix A, satisfies the desired invariance. In the linear IV example, this prior simplifies to

$$\pi(\theta;\Sigma) = \begin{array}{c} |\Omega_{00} - (\Omega_{10}+\Omega_{01})\theta + \Omega_{11}\theta^2|^{-\frac{1}{2}} \times \\ \left|\Omega_{11} - (\Omega_{01}-\Omega_{11}\theta)(\Omega_{00} - (\Omega_{10}+\Omega_{01})\theta + \Omega_{11}\theta^2)^{-1}(\Omega_{10}-\Omega_{11}\theta)\right|^{\frac{1}{2}}. \end{array}$$

This is equal to the square root of the determinant of the conditional variance for the reduced form given the moment evaluated at θ, over the square root of the determinant of the variance,

$$\pi(\theta;\Sigma) = \frac{|Var(\xi_1|g(\theta))|^{\frac{1}{2}}}{|Var(g(\theta))|^{\frac{1}{2}}}.$$

Special Case: Homoskedastic Errors In the special case of linear IV with homoskedastic errors the asymptotic variance matrix has Kronecker product structure, with

$$\begin{pmatrix} \Omega_{00} & \Omega_{01} \\ \Omega_{10} & \Omega_{11} \end{pmatrix} = \begin{pmatrix} \omega_0^2 & \omega_{10} \\ \omega_{10} & \omega_1^2 \end{pmatrix} \otimes \mathbb{E}\left[Z_i Z_i'\right]$$

for ω_0^2 and ω_1^2 the variance of the reduced form and first stage residuals, respectively, and ω_{10} their covariance. In this case, one can show that both $|Var(\xi_1|g(\theta))|^{\frac{1}{2}}$ and $|Var(g(\theta))|^{-\frac{1}{2}}$ are maximized at $\theta = \frac{\omega_{10}}{\omega_1^2}$, which corresponds to the probability limit of least squares under weak instrument asymptotics. Hence, in this case we can interpret the invariant prior as shrinking towards (the probability limit of) the least squares estimate.

Verifying Invariance: For the desired invariance to hold, the default prior $\pi(\cdot;\cdot)$ must satisfy the change-of-variables formula

$$\pi(\psi;\Sigma_h) = \pi\left(\vartheta^{-1}(\psi),\Sigma\right)\left|\frac{\partial}{\partial\psi}\vartheta^{-1}(\psi)\right|.$$

For the proposed prior, however, we have

$$\pi(\psi;\Sigma_h) = |\Sigma_h(\psi,\psi)|^{-\frac{1}{2}} \left| \frac{\partial^2}{\partial\psi\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right) - \frac{\partial}{\partial\psi}\Sigma_h\left(\psi,\tilde{\psi}\right)\Sigma_h\left(\psi,\tilde{\psi}\right)^{-1}\frac{\partial}{\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right)\right|_{\tilde{\psi}=\psi}^{\frac{1}{2}}.$$

Letting $\theta = \vartheta^{-1}(\psi)$ and $\tilde{\theta} = \vartheta^{-1}\left(\tilde{\psi}\right)$, note that

$$\Sigma_h\left(\psi,\tilde{\psi}\right)^{-1} = B\left(\tilde{\psi}\right)'^{-1}\Sigma\left(\theta,\tilde{\theta}\right)^{-1}B(\psi)^{-1}$$

$$\frac{\partial}{\partial\psi}\Sigma_h\left(\psi,\tilde{\psi}\right) = \frac{\partial}{\partial\psi}B(\psi)\,\Sigma\left(\theta,\tilde{\theta}\right)B\left(\tilde{\psi}\right)' + \frac{\partial}{\partial\psi}\vartheta^{-1}(\psi)\,B(\psi)\,\frac{\partial}{\partial\theta}\Sigma\left(\theta,\tilde{\theta}\right)B\left(\tilde{\psi}\right)'$$

$$\frac{\partial^2}{\partial\psi\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right) = \frac{\partial}{\partial\psi}B(\psi)\,\Sigma\left(\theta,\tilde{\theta}\right)\frac{\partial}{\partial\psi}B\left(\tilde{\psi}\right)' + \frac{\partial}{\partial\psi}\vartheta^{-1}(\psi)\,B(\psi)\,\frac{\partial}{\partial\theta}\Sigma\left(\theta,\tilde{\theta}\right)\frac{\partial}{\partial\psi}B\left(\tilde{\psi}\right)'$$
$$+\frac{\partial}{\partial\psi}\vartheta^{-1}\left(\tilde{\psi}\right)\frac{\partial}{\partial\psi}B(\psi)\,\frac{\partial}{\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right)B\left(\tilde{\psi}\right) + \frac{\partial}{\partial\psi}\vartheta^{-1}(\psi)\,\frac{\partial}{\partial\psi}\vartheta^{-1}\left(\tilde{\psi}\right)B(\psi)\,\frac{\partial^2}{\partial\theta\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right)B\left(\tilde{\psi}\right)'.$$

Hence,

$$\frac{\partial}{\partial\psi}\Sigma_h\left(\psi,\tilde{\psi}\right)\Sigma_h\left(\psi,\tilde{\psi}\right)^{-1}\frac{\partial}{\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right) =$$

$$\frac{\partial}{\partial\psi}B(\psi)\,\Sigma\left(\theta,\tilde{\theta}\right)\frac{\partial}{\partial\psi}B\left(\tilde{\psi}\right)'$$

$$+\frac{\partial}{\partial\psi}\vartheta^{-1}\left(\tilde{\psi}\right)B(\psi)\,\frac{\partial}{\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right)\frac{\partial}{\partial\psi}B\left(\tilde{\psi}'\right) + \frac{\partial}{\partial\psi}\vartheta^{-1}(\psi)\,\frac{\partial}{\partial\psi}B(\psi)\,\frac{\partial}{\partial\theta}\Sigma\left(\theta,\tilde{\theta}\right)B\left(\tilde{\psi}\right)'$$
$$+\frac{\partial}{\partial\psi}\vartheta^{-1}(\psi)\,\frac{\partial}{\partial\psi}\vartheta^{-1}\left(\tilde{\psi}\right)B(\psi)\,\frac{\partial}{\partial\theta}\Sigma\left(\theta,\tilde{\theta}\right)\Sigma\left(\theta,\tilde{\theta}\right)^{-1}\frac{\partial}{\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right)B\left(\tilde{\psi}\right)',$$

and

$$\frac{\partial^2}{\partial\psi\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right) - \frac{\partial}{\partial\psi}\Sigma_h\left(\psi,\tilde{\psi}\right)\Sigma_h\left(\psi,\tilde{\psi}\right)^{-1}\frac{\partial}{\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right) =$$

$$\frac{\partial}{\partial\psi}\vartheta^{-1}(\psi)\,\frac{\partial}{\partial\psi}\vartheta^{-1}\left(\tilde{\psi}\right)B(\psi)\left(\frac{\partial^2}{\partial\theta\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right) - \frac{\partial}{\partial\theta}\Sigma\left(\theta,\tilde{\theta}\right)\Sigma\left(\theta,\tilde{\theta}\right)^{-1}\frac{\partial}{\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right)\right)B\left(\tilde{\psi}\right)'.$$

It follows that

$$\left|\frac{\partial^2}{\partial\psi\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right) - \frac{\partial}{\partial\psi}\Sigma_h\left(\psi,\tilde{\psi}\right)\Sigma_h\left(\psi,\tilde{\psi}\right)^{-1}\frac{\partial}{\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right)\right|^{\frac{1}{2}}_{\tilde{\psi}=\psi} =$$

$$\left|\frac{\partial}{\partial\psi}\vartheta^{-1}(\psi)\right|\left|B(\psi)\right|\left|\frac{\partial^2}{\partial\theta\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right) - \frac{\partial}{\partial\theta}\Sigma\left(\theta,\tilde{\theta}\right)\Sigma\left(\theta,\tilde{\theta}\right)^{-1}\frac{\partial}{\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right)\right|^{\frac{1}{2}}_{\tilde{\theta}=\theta}.$$

However, $\left|\Sigma_h(\psi,\psi)\right|^{-\frac{1}{2}} = \left|B(\psi)\right|^{-1}\left|\Sigma\left(\theta,\tilde{\theta}\right)\right|^{-\frac{1}{2}}$, so

$$\left|\Sigma_h(\psi,\psi)\right|^{-\frac{1}{2}}\left|\frac{\partial^2}{\partial\psi\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right) - \frac{\partial}{\partial\psi}\Sigma_h\left(\psi,\tilde{\psi}\right)\Sigma_h\left(\psi,\tilde{\psi}\right)^{-1}\frac{\partial}{\partial\tilde{\psi}}\Sigma_h\left(\psi,\tilde{\psi}\right)\right|^{\frac{1}{2}}_{\tilde{\psi}=\psi} =$$

$$\left|\frac{\partial}{\partial\psi}\vartheta^{-1}(\psi)\right|\left|\Sigma\left(\theta,\tilde{\theta}\right)\right|^{-\frac{1}{2}}\left|\frac{\partial^2}{\partial\theta\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right) - \frac{\partial}{\partial\theta}\Sigma\left(\theta,\tilde{\theta}\right)\Sigma\left(\theta,\tilde{\theta}\right)^{-1}\frac{\partial}{\partial\tilde{\theta}}\Sigma\left(\theta,\tilde{\theta}\right)\right|^{\frac{1}{2}}_{\tilde{\theta}=\theta}$$

as desired. □