Optimal Decision Rules for Weak GMM

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

Abstract

This paper derives the limit experiment for nonlinear GMM models with weak and partial identification. We propose a theoretically-motivated class of default priors on a nonparametric nuisance parameter. These priors imply computationally tractable Bayes decision rules in the limit problem, while leaving the prior on the structural parameter free to be selected by the researcher. We further obtain quasi-Bayes decision rules as the limit of sequences in this class, and derive weighted average power-optimal identification-robust frequentist tests. Finally, we prove a Bernstein-von Mises-type result for the quasi-Bayes posterior under weak and partial identification.

Keywords: Limit Experiment, Quasi Bayes, Weak Identification, Nonlinear GMM

JEL Codes: C11, C12, C20

This draft: July 2020.

1 Introduction

Weak and partial identification arise in a wide range of empirical settings. The problem of weak identification in linear IV is well-studied, but much less is known about weak identification in nonlinear models. In particular, while there is clear evidence of identification problems in some nonlinear applications, with objective functions that have multiple minima or are close to zero over non-trivial regions of the parameter space, there are not yet commonly-accepted methods for detecting weak identification. Even less is known about optimality: while there are some results on optimal tests for parameters in weak IV settings (e.g. D. Andrews, Moreira, and Stock 2006, Moreira and Moreira 2019,

[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, and from the Sloan Research Fellowship is gratefully acknowledged.

[2]Department of Economics, M.I.T., 50 Memorial Drive, E52-526, Cambridge, MA, 02142. Email: amikushe@mit.edu. Financial support from the Castle-Krob Career Development Chair and the Sloan Research Fellowship is gratefully acknowledged.

Moreira and Ridder 2019) we are unaware of similar testing results for nonlinear GMM, much less results on optimal decision rules for general decision problems.[3]

This paper develops a theory of optimality under weak and partial identification in nonlinear GMM. We first derive the limit experiment for weakly identified GMM. We then study Bayes decision rules in the limit problem and propose a theoretically-motivated class of priors that implies computationally tractable decision rules. This class yields the quasi-Bayes decision rules studied by Kim (2002) and Chernozhukov and Hong (2003) as their diffuse-prior limit. We further prove a Bernstein-von Mises-type result establishing the asymptotic properties of quasi-Bayes under weak and partial identification.

Our results show that the quasi-Bayes approach has a number of appealing properties regardless of the identification status of the model. Kim (2002) suggested the quasi-Bayes approach based on maximum entropy arguments, while Chernozhukov and Hong (2003) discussed it as a computational device for strongly-identified, point-identified settings, where they showed that quasi-Bayes procedures are asymptotically equivalent to optimally-weighted GMM, and so are efficient in the usual sense. We show that quasi-Bayes is the limit of a sequence of Bayes decision rules for theoretically motivated priors even under weak identification. In addition to quasi-Bayes decision rules, we also derive new weighted average power-optimal, identification-robust frequentist tests.

There are three main results in this paper. The first derives the limit experiment for weakly and partially identified models, laying the foundation for our analysis of optimality. The observation in the limit experiment is a Gaussian process corresponding to the normalized sample average of the GMM moments. Consistent with the semiparametric nature of GMM, the parameter space is infinite dimensional. The parameter consists of the structural parameter (i.e. the parameter that enters the GMM moments) and the non-parametric mean function of the moments, which lies in a reproducing kernel Hilbert space (RKHS). For convex loss functions, a complete class theorem from Brown (1986) implies that all admissible decision rules in the limit experiment are the pointwise limits of Bayes decision rules, so we focus our attention on Bayesian approaches.

[3]Kaji (2020) studies weakly identified parameters in semiparametric models, and introduces a notion of weak efficiency for estimators. Weak efficiency is necessary, but not in general sufficient, for decision-theoretic optimality (e.g. admissibility) in many contexts.

For the second main result, we motivate and derive a class of default priors in the limit experiment. We first reparameterize the limit experiment to better separate the structural parameter from an infinite-dimensional nuisance parameter governing the mean of the moments. For computational tractability we put independent priors on the two components, with Gaussian process priors on the nuisance parameter. Since researchers may know something about the structural parameter, we leave this prior free to be specified on a case-specific basis. By contrast, researchers seem unlikely to know much about the infinite-dimensional nuisance parameter, so we seek default priors there. Default priors should deliver reasonable decision rules when combined with a wide variety of researcher-selected priors on the structural parameter. This suggests a particular invariance restriction, which we show dramatically reduces the class of candidate priors, generically down to a one-parameter family that we term proportional priors. Taking the diffuse (prior variance to infinity) limit in this family yields the quasi-posterior of Kim (2002) and Chernozhukov and Hong (2003).

The third main result in the paper is an analog of the Bernstein von Mises theorem for quasi-Bayes in weakly and partially identified models. We show two results: that the quasi-posterior concentrates on the subset of the parameter space where identification is problematic, and that it is asymptotically equivalent to quasi-Bayes with priors supported on the set of problematic parameters and a modified moment condition. Our results complement prior work by Chen, Christensen, and Tamer (2018) who show that quasi-Bayes highest posterior density sets have correct coverage of the identified set in partially (but not in general weakly) identified settings.

Though our main focus is on Bayesian procedures, similar derivations are applicable to the problem of optimal frequentist hypothesis testing. We thus derive weighted average power optimal similar tests, which can be inverted to form robust confidence sets.

We illustrate our results with an application to quantile IV using data from Graddy (1995) on the demand for fish. The GMM objective function in this example is highly non-quadratic, suggesting weak identification. We compute the quasi-posterior distribution, contrast the posterior mean with GMM, and compare the highest posterior density set to the frequentist identification-robust confidence set. We find that the confidence set is slightly smaller than the credible set, but has a similar shape.

The next section introduces our setting and derives the limit experiment. Section 3 motivates and derives the class of proportional priors, shows that quasi-Bayes procedures arise as the diffuse limit of proportional priors, and constructs optimal frequentist tests. Section 4 discusses feasible quasi-Bayes decision rules and characterizes their large-sample behavior. Finally, Section 5 provides a numerical illustration, considering quantile IV applied to data from Graddy (1995).

2 Limit Experiment for Weakly Identified GMM

2.1 Weak Identification in Nonlinear GMM

Suppose we observe a sample of independent and identically distributed observations $\{X_i, i = 1, ..., n\}$ from an unknown distribution P^*. The true structural parameter value $\theta^* \in \Theta$ satisfies the moment equality $\mathbb{E}_{P^*}[\phi(X, \theta^*)] = 0$ for $\phi(\cdot, \cdot)$ a known function of the data and parameters with $\phi(x, \theta) \in \mathbb{R}^k$. We aim to choose an action $a \in \mathcal{A}$, and will incur a loss $L(a, \theta^*)$ that depends only on a and the structural parameter θ^*.

We are interested in settings where identification is weak, in the sense that the mean of the moment function $\mathbb{E}_{P^*}[\phi(X, \theta)]$ is close to zero relative to sampling uncertainty, or exactly zero, over a non-trivial part of the parameter space Θ. To obtain asymptotic approximations that reflect this, we adopt a nonparametric version of weak identification asymptotics and model the data generating process as local to identification failure. Specifically, we assume that the true distribution P^* is close to some (unknown) distribution P where the identified set for the structural parameter,

$$\Theta_0 = \{\theta \in \Theta : \mathbb{E}_P[\phi(X, \theta)] = 0\}$$

contains at least two distinct elements, and further assume that $\theta^* \in \Theta_0$.[4] To derive results that reflect proximity to identification failure, we embed P^* in a sequence of distributions $P_{n,f}$ converging to P in the sense that

$$\int \left[\sqrt{n}(dP_{n,f}^{1/2} - dP^{1/2}) - \frac{1}{2} f dP^{1/2} \right]^2 \to 0 \tag{1}$$

[4] The more general assumption that θ^* is local to Θ_0 yields a limit experiment similar to that derived below, at the cost of heavier notation. Hence, we focus on the case with $\theta^* \in \Theta_0$.

as $n \to \infty$, where $P^* = P_{n,f}$ for the observed sample size n.[5] A measurable function f in equation (1) is called score, and (1) implies that $\mathbb{E}_P[f(X)] = 0$, and that $\mathbb{E}_P[f^2(X)]$ is finite (see Van der Vaart and Wellner, Lemma 3.10.10). Denote the space of score functions by $T(P)$, and note that this is a linear subspace of $L_2(P)$.

While θ^* is the structural parameter of interest, it does not fully describe the distribution of the data, even in large samples. Instead, asymptotic behavior under $P_{n,f}$ is governed by the score f. Identifying information about θ^* then comes from the fact that not all elements of $T(P)$ are consistent with a given θ^*. Specifically, one can show that the scaled sample average of the moments has (asymptotic) mean zero at θ^* under $P_{n,f}$ if and only if $\mathbb{E}_P[f(X)\phi(X,\theta^*)] = 0$. Correspondingly, define the sub-space of scores consistent with θ^* as

$$T_{\theta^*}(P) = \{f \in T(P) : \mathbb{E}_P[f(X)\phi(X,\theta^*)] = 0\}.$$

We are now equipped to define the finite sample statistical experiment.

Definition 1 *The finite sample experiment for sample size n, $\mathcal{E}^*_{n,P}$, corresponds to observing an i.i.d. sample of random variables $X_i, i = 1, ..., n$ distributed according to $P_{n,f}$, with parameter space $\{(\theta^*, f) : \theta^* \in \Theta_0, f \in T_{\theta^*}(P)\}$.*

Note that the parameter space for this experiment is infinite-dimensional, consistent with the semi-parametric nature of the GMM model.

We next introduce two running examples, based on linear and quantile IV respectively. While our focus is on nonlinear models, we include the linear IV example to illustrate the implications of our approach in a more familiar setting.

Example 1. Linear IV. Assume that the observed data $X_i = (Y_i, W_i, Z_i')$ consists of an outcome variable Y, a scalar regressor W, and a k-dimensional instrument Z. The moment function is $\phi(X,\theta) = Z(Y - \theta W)$. Let P be a distribution such that $\mathbb{E}_P[ZY] = \mathbb{E}_P[ZW] = 0$, so the mean of the moments is identically zero under P and the structural parameter is unidentified. We model the true distribution of the data as part

[5] Prior work by Kaji (2020) also analyzes weak identification using paths of the form (1).

of a sequence $P_{n,f}$ local to P in the sense of (1). This implies that there exists a $k \times 1$ first stage vector δ (which depends on score f) such that

$$\left(\zeta_{1,n}', \zeta_{2,n}'\right)' = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i' Y_i, \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i' W_i\right)' \Rightarrow (\zeta_1', \zeta_2')' \sim N\left((\theta^*\delta', \delta')', \boldsymbol{\Sigma}\right).$$

Here $\boldsymbol{\Sigma}$ is $(2k) \times (2k)$ reduced-form covariance matrix that is consistently estimable and unaffected by f. We assume for simplicity that $\boldsymbol{\Sigma}$ is full rank, but impose no other restrictions and so allow heteroskedastic errors. Hence, in this setting our approach nests the weak-instrument asymptotics introduced by Staiger and Stock (1997). $\square$

Example 2. Quantile IV. Consider the moment condition

$$\phi\left(X, \alpha, \beta\right) = \left(\mathbb{I}\{Y - \alpha - W'\beta \leq 0\} - 0.5\right) Z, \tag{2}$$

introduced by Chernozhukov and Hansen (2005). The observed data $X_i = (Y_i, W_i, Z_i)$ consist of an outcome Y, a $(p-1)$-dimensional vector of potentially endogenous regressors W, and k-dimensional vector of instruments Z. The structural parameters $\theta = (\alpha, \beta)$ lie in a set $\Theta \subset \mathbb{R}^p$. A variety of different distributions P give rise to non-trivial identified sets Θ_0 in this model. Correspondingly, there are many ways weak identification may arise. For this example, suppose that the first element of Z is a constant, while the remaining elements of Z can be written as the element-wise product $U \cdot Z^*$, for U a k-dimensional mean-zero random vector independent of (Y, W, Z^*) and Z^* a potentially informative, but unobserved, instrument. In this setting, the last $k-1$ elements of $E_P[\phi\left(X, \alpha, \beta\right)]$ are identically zero on Θ, while the first element of $E_P[\phi\left(X, \alpha, \beta\right)]$ is zero if and only if α is equal to the median of $Y - W'\beta$. Hence the identified set under P is $\Theta_0 = \{\theta = (\alpha, \beta) \in \Theta : \alpha = median_P(Y - W'\beta)\}$.

2.2 Asymptotic Representation Theorem

This section shows that in order to construct asymptotically optimal decision rules for weakly identified GMM, it suffices to derive optimal decision rules in a limit experiment. This limit experiment corresponds to observing a Gaussian process $g(\cdot)$ with unknown mean function $m(\cdot)$ and known covariance function $\Sigma(\cdot, \cdot)$, where θ^* satisfies $m(\theta^*) = 0$.

Intuitively, $g(\cdot)$ corresponds to the scaled sample average of the moments, since as we discuss below, under mild regularity conditions

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(X_i,\cdot)\Rightarrow g(\cdot)\sim\mathcal{GP}(m,\Sigma), \tag{3}$$

on Θ_0 under $P_{n,f}$, where $m(\cdot)=\mathbb{E}_P[f(X)\phi(X,\cdot)]$, $\Sigma(\theta_1,\theta_2)=\mathbb{E}_P[\phi(X,\theta_1)\phi(X,\theta_2)']$, and Σ is consistently estimable.

To derive the limit experiment, we first discuss the parameter space for the mean function $m(\cdot)$, and its connection to the space of scores f. We next discuss a standard non-parametric limit experiment, which we then use to derive our Gaussian process limit experiment. At each stage, we follow the usual limits-of-experiments approach and relate the experiments studied in terms of the attainable risk functions.

Functional parameter space. Consider the set of $k\times 1$ vector-valued functions $\sum_{j=1}^{s}\Sigma(\cdot,\theta_j)b_j:\Theta_0\to\mathbb{R}^k$, defined for any finite set of constant vectors $\{b_j\}\subset\mathbb{R}^k$, parameters $\{\theta_j\}\subset\Theta_0$ and a covariance function $\Sigma(\cdot,\cdot)$. Define a scalar product on this set by $\langle\sum_{j=1}^{s}\Sigma(\cdot,\theta_j)b_j,\sum_{l=1}^{\tilde{s}}\Sigma(\cdot,\tilde{\theta}_l)c_l\rangle_{\mathcal{H}}=\sum_{j=1}^{s}\sum_{l=1}^{\tilde{s}}b_j'\Sigma(\theta_j,\tilde{\theta}_l)c_l$.

Definition 2 *The Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ associated with Σ is the completion of the space spanned by functions of the form $\sum_{j=1}^{s}\Sigma(\cdot,\theta_j)b_j$ with respect to the scalar product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.*

Let $\mathcal{H}^*$ be the completion of the space spanned by scores of the form $f(X)=\sum_{j=1}^{s}\phi(X,\theta_j)'b_j$ in $L_2(P)$. This is a linear subspace of $T(P)$. Let $(\mathcal{H}^*)^\perp$ be the orthogonal complement to $\mathcal{H}^*$. For any score $f\in T(P)$, denote by f^* and $f^\perp$ its projections onto $\mathcal{H}^*$ and $(\mathcal{H}^*)^\perp$ respectively.

Lemma 1 *Define a linear transformation mapping scores $f\in T(P)$ to functions $m(\cdot)$:*

$$m(\cdot)=\mathbb{E}_P[f(X)\phi(X,\cdot)]. \tag{4}$$

The image of $T(P)$ under this transformation is $\mathcal{H}$. The null space of this transformation is $(\mathcal{H}^*)^\perp$. The transformation (4) restricted to $\mathcal{H}^*$ establishes an isomorphism between $\mathcal{H}$ and $\mathcal{H}^*$. In particular, for any two $f_1,f_2\in\mathcal{H}^*$ and the corresponding $m_1(\cdot),m_2(\cdot)$, we have $\langle m_1,m_2\rangle_{\mathcal{H}}=\mathbb{E}_P\left[f_1(X)f_2(X)\right]$.

Lemma 1 states that all mean functions implied by $f \in T(P)$ lie in $\mathcal{H}$, and all mean functions in $\mathcal{H}$ correspond to some $f \in T(P)$. The correspondence between scores and mean functions is many-to-one, however, as all scores with the same projection f^* onto $\mathcal{H}^*$ imply the same mean function.

Definition of limit experiments. We next introduce two limit experiments. The first, $\mathcal{E}^*_\infty$, is a variant of a Gaussian sequence experiment discussed in Van der Vaart (1991), adapted to incorporate the moment restriction. The second, $\mathcal{E}^*_{GP}$, is our final goal. Let $\{\varphi_j\}$ be a complete orthonormal basis in $T(P)$, and let $\mathcal{H}_{\theta^*} = \{m \in \mathcal{H} : m(\theta^*) = 0\}$ denote the subset of $\mathcal{H}$ with a zero at θ^*.

Definition 3 *The limit experiment $\mathcal{E}^*_\infty$ corresponds to observing the (infinite) sequence of independent random variables $W_j \sim N(\mathbb{E}_P[f(X)\varphi_j(X)], 1)$, with parameter space $\{(\theta^*, f) : \theta^* \in \Theta_0, f \in T_{\theta^*}(P)\}$.*

Definition 4 *The Gaussian process experiment $\mathcal{E}^*_{GP}$ corresponds to observing a Gaussian process $g(\cdot) \sim \mathcal{GP}(m(\cdot), \Sigma)$ with known covariance function $\Sigma(\cdot, \cdot)$, unknown mean m, and parameter space $\{(\theta^*, m) : \theta^* \in \Theta_0, m \in \mathcal{H}_{\theta^*}\}$.*

The parameter space in $\mathcal{E}^*_\infty$ is the same as in the finite-sample experiment, while by Lemma 1 the parameter space in $\mathcal{E}^*_{GP}$ is smaller. In all experiments the true value θ^* corresponds to a zero of the moment function, in the sense that $\mathbb{E}_P[f(X)\phi(X, \theta^*)] = 0$ or $m(\theta^*) = 0$.

Attainable risk functions. Following the literature on limits of experiments (c.f. Le Cam 1986) we will compare the experiments described above in terms of attainable risk functions. We begin with an asymptotic representation theorem.

Lemma 2 *(Theorem 3.1 in Van der Vaart (1991)) Consider a sequence of statistics S_n which has a limit distribution under $\mathcal{E}^*_{n,P}$, in the sense that under any $P_{n,f}$ for $f \in T(P)$, $S_n(X_1, ..., X_n) \Rightarrow S_f$ as $n \to \infty$. Assume there exists a complete separable set $\mathbb{S}_0$ such that $S_f(\mathbb{S}_0) = 1$ for all $f \in T(P)$. Then in the experiment $\mathcal{E}^*_\infty$ there exists a (possibly randomized) statistic $S^* = s^*(\{W_j\}, U)$ for a random variable $U \sim U[0,1]$ independent of W_j such that $S^* \sim S_f$ under f for all $f \in T(P)$.*

Lemma 2 implies that for well-behaved loss functions, the set of risk functions in the limit experiment $\mathcal{E}_\infty^*$ nests that in the finite sample experiment $\mathcal{E}_{n,P}^*$ asymptotically.

Corollary 1 *If $L(a,\theta^*)$ is bounded and continuous in a for all $\theta^* \in \Theta_0$, and the sequence of decision rules S_n satisfies the conditions of Lemma 2, then there exists a statistic S^* in the limit experiment $\mathcal{E}_\infty^*$ such that*

$$\lim_{n\to\infty} \mathbb{E}_{P_{n,f}}[L(S_n,\theta^*)] = \mathbb{E}_f[L(S^*,\theta^*)] \text{ for all } \{(\theta^*, f) : \theta^* \in \Theta_0, f \in T_{\theta^*}(P)\}.$$

We next compare the experiments $\mathcal{E}_\infty^*$ and $\mathcal{E}_{GP}^*$. By Lemma 1 any score can be written as $f = f^* + f^\perp$, where $f^\perp \in (\mathcal{H}^*)^\perp$ has no effect on the mean function. We can re-write the parameter space of the limit experiment $\mathcal{E}_\infty^*$ as a Cartesian product

$$\{\theta^* \in \Theta_0, f = (f^*, f^\perp) \in T_{\theta^*}(P)\} = \{\theta^* \in \Theta_0, f^* \in \mathcal{H}_{\theta^*}^*\} \times \{f^\perp \in (\mathcal{H}^*)^\perp\}$$

for $\mathcal{H}_{\theta^*}^* = \{f \in \mathcal{H}^* : \mathbb{E}_P[f(X)\phi(X,\theta^*)] = 0\}$. The parameter $f^\perp$ is unrelated to the structural parameter θ^*, and the restriction of the experiment $\mathcal{E}_\infty^*$ that fixes this parameter is equivalent to the experiment $\mathcal{E}_{GP}^*$.

Theorem 1 *Fix any $f^\perp \in (\mathcal{H}^*)^\perp$. For any statistic S^* in $\mathcal{E}_\infty^*$, there exists a (possibly randomized) statistic S in $\mathcal{E}_{GP}^*$ such that for all $f^* \in \mathcal{H}^*$, the distribution of S^* under $(f^*, f^\perp)$ is the same as that of S under $m(\cdot) = \mathbb{E}_P[f^*(X)\phi(X,\cdot)]$. Identifying f^* and m, the set of risk functions $\left\{\mathbb{E}_{(f^*,f^\perp)}\left[L(S^*,\theta^*)\right] : \theta^* \in \Theta_0, f \in \mathcal{H}_{\theta^*}^*\right\}$ in $\mathcal{E}_\infty^*$ is equal to the set of risk functions $\{\mathbb{E}_m\left[L(S,\theta^*)\right] : \theta^* \in \Theta_0, m \in \mathcal{H}_{\theta^*}\}$ in $\mathcal{E}_{GP}^*$.*

The idea of holding $f^\perp$ fixed is similar to the "slicing" argument of Hirano and Porter (2009). Specifically, $f^\perp$ is a nuisance parameter that neither interacts with the parameter of interest θ^*, nor enters the loss function. Thus, to derive optimal procedures it suffices to study optimality holding $f^\perp$ fixed, which in turn implies equivalence with the simpler Gaussian process experiment $\mathcal{E}_{GP}^*$. Hence, the performance of optimal decision rules in $\mathcal{E}_{GP}^*$ bounds asymptotic performance in $\mathcal{E}_{n,P}^*$. Thus, if a sequence of decision rules S_n has risk converging to an optimal risk function in $\mathcal{E}_{GP}^*$, it must be asymptotically optimal.

Theorem 1 gives a criterion which may be checked to verify asymptotic optimality. In many cases, a plug-in approach further suggests the form of an asymptotically optimal rule. Suppose we know an optimal decision rule $S = s(g(\cdot), \Sigma, U)$ in the Gaussian process experiment $\mathcal{E}^*_{GP}$, where we now make dependence on the covariance function explicit. If a uniform central limit theorem holds under P and we have a consistent estimator $\widehat{\Sigma}$ for Σ (e.g. the sample covariance function $\widehat{\Sigma}(\theta,\tilde{\theta}) = \widehat{Cov}(\phi(X_i;\theta),\phi(X_i;\tilde{\theta})))$, then Le Cam's third lemma implies that the weak convergence (3) holds and $\widehat{\Sigma}$ remains consistent under $P_{n,f}$. Hence, provided $s(g(\cdot),\Sigma,U)$ is almost-everywhere continuous in $(g(\cdot),\Sigma,U)$, the Continuous Mapping Theorem implies that

$$S_n = s\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(X_i,\cdot),\widehat{\Sigma},U\right) \Rightarrow s(g(\cdot),\Sigma,U)$$

under $P_{n,f}$, so the sequence of rules S_n is asymptotically optimal so long as the loss satisfies the conditions of Corollary 1.

The idea of solving the limit problem in order to derive asymptotically optimal decision rules is of course not new (see e.g. Le Cam, 1986). More recently, Mueller (2011) proposed an alternative approach to derive asymptotically optimal tests based on weak convergence conditions like (3). Relative to the approach of Mueller (2011) applied to our setting, the benefits of Theorem 1 are (i) to show that there is, in a sense, no asymptotic information loss from limiting attention to the sample average of the moments and (ii) the ability to consider general decision problems in addition to tests. The weak convergence (3) was the starting point of Andrews and Mikusheva (2016), where we proposed a general approach to constructing identification-robust, but not necessarily optimal, tests.

Example 1. Linear IV (continued). The Gaussian process experiment corresponds to observing the linear-in-θ process $g(\theta) = \zeta_1 - \theta\zeta_2$. For Σ_{ij} $k \times k$ sub-blocks of the $(2k)\times(2k)$ covariance matrix $\mathbf{\Sigma}$, this process has covariance function

$$\Sigma(\theta_1,\theta_2) = \mathbb{E}_P\left[(Y-\theta_1 W)ZZ'(Y-\theta_2 W)\right] = \Sigma_{11} - \theta_1\Sigma_{12} - \theta_2\Sigma_{21} + \theta_1\theta_2\Sigma_{22}.$$

The corresponding RKHS consists of $\mathbb{R}^k$-valued linear functions of θ. Theorem 1 implies that the performance of optimal decision rules in the limit experiment bounds the attainable asymptotic performance. Moreover, given an optimal rule in the limit experiment we can construct an asymptotically optimal rule by plugging in $\zeta_{1,n} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_iY_i$ $\zeta_{2,n} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_iW_i$, and a covariance estimator $\widehat{\mathbf{\Sigma}}$, in place of ζ_1, ζ_2, and $\mathbf{\Sigma}$. $\square$

Example 2. Quantile IV (continued). Our analysis in this section treats the identified set Θ_0 under P as known, and limits attention to $\theta \in \Theta_0$. Under $P_{n,f}$ the scaled sample average of the moments converges as a process on Θ_0. Specifically,

$$g_n(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbb{I}\{Y_i - \alpha(\beta) - W_i'\beta \leq 0\} - \frac{1}{2} \right) Z_i \Rightarrow g(\beta), \quad g(\cdot) \sim \mathcal{GP}(m, \Sigma)$$

where $\alpha(\beta) = median_P(Y - W'\beta)$. According to Theorem 1, to derive asymptotically optimal decision rules in this setting, it suffices to derive optimal decision rules based on observing $g(\cdot)$. We cannot calculate $g_n(\beta)$ in practice, since P, and thus $\alpha(\beta)$ and Θ_0, are unknown. Section 4 discusses feasible procedures, showing that they implicitly estimate Θ_0 based on a subset of the moments and behave like their Gaussian process experiment analogs based on the remaining moments. $\square$

2.3 A Complete Class Theorem

Theorem 1 allows us to characterize the class of asymptotically admissible decision rules. A decision rule $s(g)$ in the experiment $\mathcal{E}_{GP}^*$ is admissible if there exists no rule $s'(g)$ with weakly lower risk for all parameter values, and strictly lower risk for some. The infinite-dimensional parameter space for $\mathcal{E}_{GP}^*$ puts it beyond the scope of many complete class theorems (theorems characterizing the set of admissible rules), but for convex loss functions a result from Brown (1986) applies.

Theorem 2 *(Brown, 1986) Suppose that $\mathcal{A}$ is closed, with $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ for some d_a, that $L(a, \theta)$ is continuous and strictly convex in a for every θ, and that either $\mathcal{A}$ is bounded or $\lim_{\|a\| \to \infty} L(a, \theta) = \infty$. Then for every admissible decision rule s in $\mathcal{E}_{GP}^*$ there exists a sequence of priors π_r and corresponding Bayes decision rules s_{π_r},*

$$\int E_{\theta^*,m}[L(s_{\pi_r}(g), \theta)] d\pi_r(\theta^*, m) = \min_{\tilde{s}} \int E_{\theta^*,m}[L(\tilde{s}(g), \theta)] d\pi_r(\theta^*, m),$$

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

Hence, for convex loss functions all admissible decision rules in the limit experiment are pointwise limits of Bayes decision rules.

3 Priors for GMM Limit Problem

The previous section shows that we can reduce the search for asymptotically optimal decision rules to a search for optimal rules based on the Gaussian process $g(\cdot) \sim \mathcal{GP}(m, \Sigma)$, with a known covariance function $\Sigma(\cdot, \cdot)$ and an unknown mean function m such that $m(\theta^*) = 0$. Motivated by the complete class result in Theorem 2, we concentrate our attention on Bayes decision rules.

The parameter in the limit experiment consists of the finite-dimensional parameter of interest θ^*, and the infinite-dimensional nuisance parameter m that determines the identification status of θ^*. Researchers may have prior information about θ^*, but it seems impractical to elicit priors about the infinite-dimensional parameter m. We thus aim to propose a class of default priors on the infinite-dimensional component.

We proceed in three steps. First, we re-parameterize the limit experiment to further separate the structural parameter θ^* from the infinite-dimensional nuisance parameter. Second, we leave the choice of prior on θ^* free, and consider a class of Gaussian process priors on the infinite-dimensional parameter which lead to tractable decision rules. Third, we argue that it is natural to impose a particular invariance property for default priors, and find that this restriction dramatically reduces the class of candidate priors. This leads to our suggested default priors.

We assume from this point on that Θ_0 is compact and Σ is continuous. By Corollary 4 of Berlinet and Thomas-Agnan (2004), this implies that $\mathcal{H}$ is a separable space of continuous functions. Lemma 1.3.1 of Adler and Taylor (2007) implies we can take g to be everywhere continuous almost surely.

3.1 Linear Reparameterization

The parameter space $\{(\theta^*, m) : \theta^* \in \Theta_0, m \in \mathcal{H}_{\theta^*}\}$ requires that for fixed θ^*, the mean function m must lie in the linear subspace $\mathcal{H}_{\theta^*}$. Hence, the marginal parameter space for m, leaving θ^* unrestricted, is the subset of $\mathcal{H}$ with at least one zero on Θ_0. This set is highly non-convex, as it is easy to find pairs of functions, each of which has a zero, such that the average has no zeros.

To simplify the construction of the prior, we re-parameterize the model to disentangle

θ^* and the infinite-dimensional nuisance parameter. Our reparameterization is based on what we term an anchor functional. Denote by $\mathcal{C}$ the space of continuous functions from Θ_0 to $\mathbb{R}^k$, and let A be a linear functional from $\mathcal{C}$ to $\mathbb{R}^k$. Let $G(\cdot) = g(\cdot) - m(\cdot)$ denote the mean-zero Gaussian process with covariance function Σ. The regression of the process G on the anchor $A(G)$ defines a Pettis integral

$$\psi(\cdot) = [\psi_1(\cdot), ..., \psi_k(\cdot)] = \mathbb{E}\left[G(\cdot)A(G)'\right] \left(\mathbb{E}\left[A(G)A(G)'\right]\right)^{-1} \in \mathcal{H}^k,$$

where each column is again a function in $\mathcal{H}$ (see Van der Vaart and Van Zanten, 2008, for discussion). Since $\psi(\cdot)$ depends only on Σ and A, it is known in the limit experiment. An example of an anchor functional is the point-evaluation functional at a point $\theta_0 \in \Theta_0$, $A(G) = G(\theta_0)$. For this anchor $\psi(\cdot) = \Sigma(\cdot, \theta_0)\Sigma(\theta_0, \theta_0)^{-1}$.

Let $\mathcal{H}_\mu$ be the linear sub-space of $\mathcal{H}$ orthogonal to $\{\psi_1(\cdot), ..., \psi_k(\cdot)\}$. Assume that the $k \times k$ matrix $\psi(\theta)$ has full rank for all $\theta \in \Theta_0$. For each $m(\cdot) \in \mathcal{H}$, define $\mu(\cdot)$ to be the projection of m on the linear sub-space $\mathcal{H}_\mu$. The properties of Pettis integrals imply that $\langle \psi, m \rangle_{\mathcal{H}} = A(m)$ and $\langle \psi, \psi \rangle_{\mathcal{H}} = 1$, which yields the orthogonal decomposition

$$m(\cdot) = \mu(\cdot) + \psi(\cdot)A(m).$$

For any θ^* and $m \in \mathcal{H}_{\theta^*}$, $m(\theta^*) = 0$, so $A(m) = -[\psi(\theta^*)]^{-1}\mu(\theta^*)$. We can consequently re-write $m(\cdot)$ as a function of (θ^*, μ),

$$m(\cdot) = \mu(\cdot) - \psi(\cdot)[\psi(\theta^*)]^{-1}\mu(\theta^*).$$

This establishes a one-to-one correspondence between $\{(\theta^*, m) : \theta^* \in \Theta_0, m \in \mathcal{H}_{\theta^*}\}$ and $(\theta^*, \mu) \in \Theta_0 \times \mathcal{H}_\mu$. Hence, the transformation from (θ^*, m) to (θ^*, μ) is a reparameterization of the model. The parameter space in the reparameterized model is a Cartesian product, $\Theta_0 \times \mathcal{H}_\mu$. Moreover, $\mathcal{H}_\mu$ is the RKHS generated by the covariance function

$$\widetilde{\Sigma}(\theta_1, \theta_2) = \Sigma(\theta_1, \theta_2) - \psi(\theta_1)\mathbb{E}\left[A(G)A(G)'\right]\psi(\theta_2)',$$

and thus is a linear space. The combination of Cartesian product structure and linearity for the infinite-dimensional component greatly simplifies the task of constructing priors.

There is a stochastic decomposition associated with this re-parametrization. Define the random vector and stochastic process ξ and h, respectively, by

$$\xi = A(g) \quad \text{and} \quad h(\cdot) = g(\cdot) - \psi(\cdot)\xi. \tag{5}$$

By construction $\xi \sim N(A(m), \Sigma_\xi)$ for $\Sigma_\xi = E[A(G)A(G)']$, while $h(\cdot) \sim \mathcal{GP}(\mu, \widetilde{\Sigma})$. Moreover, ξ and h are jointly normal and uncorrelated, and therefore independent. Note that the distribution of $h(\cdot)$ does not depend on θ^*. In Andrews and Mikusheva (2016) we showed that when A is the point evaluation functional at θ_0, $h(\cdot)$ is a sufficient statistic for the nuisance parameter in the problem of testing $H_0 : \theta^* = \theta_0$.

Example 1. Linear IV (continued). The linear process $g(\theta) = \zeta_1 - \theta\zeta_2$ is a one-to-one transformation of the $k \times 1$ Gaussian vectors ζ_1 and ζ_2, where the mean vectors of ζ_1 and ζ_2 are proportional with constant of proportionality θ^*. For any anchor, the corresponding process h is also linear, and can be fully described by a single $k \times 1$ Gaussian vector, equal to a full rank linear transformation of ζ_1 and ζ_2. The mean μ of h is correspondingly described by a $k \times 1$-vector $\tilde{\mu}$, equal to the same linear transformation applied to the means of ζ_1 and ζ_2. See Section S1 of the Supplementary Appendix for details. Different choices of anchor yield different re-parameterizations $(\theta^*, \tilde{\mu})$, where one natural option is to select the anchor that implies $\tilde{\mu}$ equal to the first stage coefficient.

Example 2. Quantile IV (continued). Many different anchor functionals may be used in this example. For instance, in many econometric applications $\beta = 0$ is a point of particular interest. Correspondingly, one could take the anchor to equal the point-evaluation functional at $\beta = 0$, $A(g) = g(0)$. $\square$

3.2 Structure of the Prior

We next derive a class of default priors on the infinite-dimensional nuisance parameters μ, leaving the prior on θ^* free to be specified based on context-specific knowledge. We seek priors on μ that (i) yield analytically and computationally tractable Bayes decision rules and (ii) behave reasonably when combined with many different priors on θ^*.

Our specification of the prior is guided in part by the structure of the likelihood. The last section decomposed the observed process $g(\cdot)$ as $g(\cdot) = h(\cdot) + \psi(\cdot)\xi$, for $\xi = A(g)$. By construction ξ and $h(\cdot)$ are independent. Thus, the likelihood function $\ell(\mu, \theta^*; g)$ based on the observed data $g(\cdot)$ factors as[6]

$$\ell\left(\mu,\theta^{*};g\right)=\ell\left(\mu,\theta^{*};\xi\right)\ell\left(\mu;h\right),$$

where $\ell\left(\mu,\theta^{*};\xi\right)$ and $\ell\left(\mu;h\right)$ are the likelihood functions based on ξ and h, with the latter depending only on μ but not on θ^{*}. Since the loss function depends only on θ^{*}, to derive Bayes decision rules it suffices to construct the marginal posterior distribution for θ^{*}.

For analytical tractability we consider only independent priors $\pi(\theta^{*})\pi(\mu)$ on θ^{*} and μ. Under such priors the marginal posterior for θ^{*} is

$$\pi(\theta^{*}|\xi,h)=\frac{\pi(\theta^{*})\int\ell\left(\mu,\theta^{*};\xi\right)\ell\left(\mu;h\right)d\pi(\mu)}{\int\int\ell\left(\mu,\theta;\xi\right)\ell\left(\mu;h\right)d\pi(\mu)d\pi(\theta)}=\frac{\pi(\theta^{*})\int\ell\left(\mu,\theta^{*};\xi\right)d\pi(\mu|h)f(h)}{\int\int\ell\left(\mu,\theta;\xi\right)d\pi(\mu|h)d\pi(\theta)f(h)}.$$

Here $f(h)=\int\ell\left(\mu;h\right)d\pi(\mu)$ denotes the marginal density of h, and $\pi(\mu|h)$ the posterior for μ given h. Prior independence of θ^{*} and μ ensures that $f(h)$ does not depend on θ^{*}, and so drops out. Thus, the posterior simplifies to

$$\pi(\theta^{*}|g)=\frac{\pi(\theta^{*})\ell^{*}\left(\theta^{*}\right)}{\int\pi(\theta)\ell^{*}\left(\theta\right)d\theta},\quad\text{for}\quad\ell^{*}\left(\theta\right)=\int\ell\left(\mu,\theta;\xi\right)d\pi\left(\mu|h\right). \tag{6}$$

Cartesian product structure of the parameter space $\Theta_0\times\mathcal{H}_\mu$ is necessary for independent priors and so plays a crucial role in this result. The restriction to independent priors is made less stringent than it might appear by the freedom to choose the anchor functional A, which in turn determines the content of independence.

We further restrict attention to Gaussian process priors $\mu\sim\mathcal{GP}(0,\Omega)$, where $\Omega(\cdot,\cdot)$ is a continuous covariance function. This allows us to exploit conjugacy results, greatly simplifying the form of the posterior. Specifically, $\ell^{*}\left(\theta^{*}\right)$ is based on the conditional distribution of $\xi\sim N(-[\psi(\theta^{*})]^{-1}\mu(\theta^{*}),\Sigma_\xi)$ conditional on the realization of $h=\mu+\mathcal{GP}(0,\Sigma)$, where $\mu\sim\mathcal{GP}(0,\Omega)$. For a Gaussian process prior on μ, $\ell^{*}\left(\theta^{*}\right)$ corresponds to a Gaussian likelihood for observation ξ with mean given by the best linear predictor of ξ based on h. The solution to this linear prediction problem is obtained in Parzen (1962), and details appear in the proof of Theorem 3 stated below. See Berlinet and Thomas-Agnan (2004) Ch. 2.4 for a textbook treatment.

[6] All Gaussian processes with covariance function Σ and mean functions in $\mathcal{H}$ are mutually absolutely continuous, so we can define the likelihood with respect to any base measure in this class.

3.3 Invariance Restriction

The posterior (6) depends on the researcher-specified prior $\pi(\theta^*)$ and the Gaussian likelihood $\ell^*(\theta^*)$. The latter in turn depends on ξ, along with the best linear predictor for the vector ξ based on the process h. While the best linear predictor is mathematically well-defined, direct calculation involves infinite-dimensional objects and will often be practically unappealing. In most cases one would need to approximate the best linear predictor numerically, for instance using discretization or eigenvector expansions. See e.g. Parzen (1962) for discussion. A further challenge is that the form of the best linear predictor depends on the precise specification of the prior covariance function Ω. The space of such covariance functions is enormous, and it seems challenging to directly evaluate whether a given covariance function is reasonable or not.

To derive default priors, we thus take a different approach, and ask what choices of prior covariance Ω lead to decision rules with desirable properties. Since the prior on θ^* may be specified based on application-specific knowledge, it is particularly important that a default prior on μ produce reasonable results when combined with many different choices of $\pi(\theta^*)$. To this end, we require that if a researcher rules out some parameter values ex-ante, limiting the support of $\pi(\theta^*)$, the implied Bayes decision rules should not depend on the behavior of the moments at the excluded parameter values.

Formally, we require that for priors $\pi(\theta^*)$ with restricted support $\widetilde{\Theta} \subset \Theta_0$, Bayes decision rules based on the prior $\pi(\theta^*)\pi(\mu)$ should depend on the data only through ξ and the restriction of g to $\widetilde{\Theta}$. For this invariance property to hold for all possible priors $\pi(\theta^*)$ and all loss functions $L(a, \theta^*)$, however, it must be that $\ell^*(\theta)$ depends on the data only through $(\xi, g(\theta))$ for all $\theta \in \Theta_0$.[7] This restriction dramatically narrows the class of candidate covariance functions Ω.

Theorem 3 *Consider the setting described above with a Gaussian process prior on* μ, *where the covariance function* Ω *is continuous. For all* $\theta^* \in \Theta_0$ *such that* $\widetilde{\Sigma}(\theta^*, \theta^*)$ *and* $\Omega(\theta^*, \theta^*)$ *have full rank, the integrated likelihood* $\ell^*(\theta^*)$ *depends on the data only through* $(\xi, g(\theta^*))$, *or equivalently through* $(\xi, h(\theta^*))$, *if and only if*

$$\Omega(\theta^*, \theta^*)^{-1}\Omega(\theta^*, \theta) = \widetilde{\Sigma}(\theta^*, \theta^*)^{-1}\widetilde{\Sigma}(\theta^*, \theta) \text{ for all } \theta \in \Theta_0. \tag{7}$$

[7] We provide a formal invariance argument in Section S2 of the Supplementary Appendix.

One implication of (7) is that $\mathcal{H}_\Omega$, the RKHS generated by Ω, coincides with $\mathcal{H}_\mu$, the parameter space for μ. It is natural to require that $\mathcal{H}_\Omega \subseteq \mathcal{H}_\mu$. On the other hand, covariance functions that imply $\mathcal{H}_\Omega \subset \mathcal{H}_\mu$ rule out parts of the parameter space a-priori, and, as discussed in Florens and Simoni (2012), this can be understood as a smoothing assumption. Specifically, if $\mathcal{H}_\Omega$ is strictly contained in $\mathcal{H}_\mu$, then $\mathcal{H}_\Omega$ can be expressed as the image of $\mathcal{H}_\mu$ under an integral, or smoothing, operator related to the covariance functions Ω and $\widetilde{\Sigma}$. In such cases, it is intuitive that the best linear predictor of $\mu(\theta^*)$ will smooth the realized process h, using values of the process at points other than θ^*. Thus, it makes sense that the requirement to use only $h(\theta^*)$ forces $\mathcal{H}_\Omega = \mathcal{H}_\mu$.

The next lemma shows that invariance generically reduces the class of candidate priors to a one dimensional family, namely covariance functions Ω proportional to $\widetilde{\Sigma}$. We first recall a definition. A linear subspace $V \subseteq \mathbb{R}^k$ is invariant for a linear operator L if for any $v \in V$ we have $Lv \in V$. Invariant sub-spaces for a symmetric matrix L are the sub-spaces spanned by subsets of its eigenvectors.

Lemma 3 *Fix some $\theta_0 \in \Theta_0$ such that $\widetilde{\Sigma}(\theta_0, \theta_0)$ is full rank. Assume there does not exist a non-trivial (non-empty, but strictly smaller than $\mathbb{R}^k$) linear subspace $V \subseteq \mathbb{R}^k$ that is invariant for the whole family of symmetric operators*

$$\mathcal{D} = \left\{ D(\theta) = R(\theta_0, \theta) R(\theta_0, \theta)', \theta \in \Theta_0 : \det(\widetilde{\Sigma}(\theta, \theta)) > 0 \right\},$$

where $R(\theta_0, \theta) = \widetilde{\Sigma}(\theta_0, \theta_0)^{-1/2} \widetilde{\Sigma}(\theta_0, \theta) \widetilde{\Sigma}(\theta, \theta)^{-1/2}$ is a correlation function. Then condition (7) is equivalent to $\Omega(\cdot, \cdot) = \lambda \widetilde{\Sigma}(\cdot, \cdot)$ for some $\lambda > 0$.

Two positive-definite matrices share a common subspace if and only if several eigenvectors of one matrix span the same sub-space as several eigenvectors of the other. The set of matrix pairs that share a non-trivial invariant subspace is of lower dimension than the set of positive-definite matrix pairs, so generically (that is, everywhere but on a nowhere-dense subset) two positive-definite matrices share no non-trivial invariant subspace. For the condition of Lemma 3 to fail requires something still stronger, namely that the same subspace be invariant for a whole family of matrices indexed by θ. This often entails special structure on the moment conditions. Such structure arises naturally in some cases. For instance, suppose a researcher forms moments based on two independent datasets, where one dataset is used to form the first block of moments, while the other is used for the rest. In this case Σ will be block-diagonal, and will imply two orthogonal invariant sub-spaces that are common across all θ. If these are the only nontrivial invariant subspaces, the family of Ω satisfying condition (7) is two-dimensional, allowing a researcher to put different coefficients of proportionality on two invariant sub-spaces.

Example 1. Linear IV (continued). Our analysis allows a researcher-selected marginal prior θ^* and an independent Gaussian prior $N(0,\Omega)$ on $\tilde{\mu}$. Interestingly, in this setting our invariance condition imposes no restrictions, and any $k \times k$ covariance matrix Ω is allowed. This reflects the parametric and low-dimensional nature of the model: $(\xi, h(\theta^*))$ is an invertible linear transformation of (ζ_1, ζ_2), so $\ell^*(\theta^*)$ necessarily depends on the data only through $(\xi, h(\theta^*))$.

Choosing the anchor such that $\tilde{\mu}$ is the first stage coefficient leads to independent priors on the structural and first-stage parameters (θ^*, δ). The class of priors we consider therefore nests some independent priors discussed in the literature, including the MM1 prior of Moreira and Moreira (2019) and (taking the diffuse limit for Ω) the relatively invariant prior of Moreira and Ridder (2019). Our class of priors is wider than this, however, and also allows some form of dependence between (θ^*, δ). See Section S1 of the Supplementary Appendix for details. $\square$

Example 2. Quantile IV (continued). Assume without loss of generality that $Var(U) = I_k$ and denote $\tilde{Z}' = (1\ Z^{*\prime})$. Absent further restrictions on the joint distribution of (Y, W, Z^*), in general, the covariance function

$$\Sigma(\beta,\tilde{\beta}) = \mathbb{E}_P\left[\left(\mathbb{I}\{Y-\alpha(\beta)-W\beta\leq 0\}-\frac{1}{2}\right)\left(\mathbb{I}\{Y-\alpha(\tilde{\beta})-W\tilde{\beta}\leq 0\}-\frac{1}{2}\right)\tilde{Z}\tilde{Z}'\right].$$

does not have non-trivial invariant subspace and satisfies assumptions of Lemma 3, so only the proportional prior yields invariant decision rules. However, if P imposes independence of Z^* from (Y, W), then $\Sigma(\beta, \tilde{\beta})$ is the product of a scalar function of $\beta, \tilde{\beta}$ with the $k \times k$ matrix $\mathbb{E}_P\left[\tilde{Z}\tilde{Z}'\right]$. In this special case, invariance only determines Ω up to a $k \times k$ positive definite matrix. $\square$

3.4 Bayes Decision Rules for Proportional Priors

Motivated by Theorem 3 and Lemma 3, we focus on proportional prior covariance functions, $\Omega(\cdot,\cdot) = \lambda\widetilde{\Sigma}(\cdot,\cdot)$. Bayes decision rules minimize the posterior risk,

$$s(g) = \arg\min_{a\in\mathcal{A}} \int_{\Theta_0} L(a,\theta^*)\pi(\theta^*|g)d\theta^* \tag{8}$$

for almost every realization of the data (see e.g. Chapter 3 of Lehmann and Casella (1998)), for $\pi(\theta^*|g)$ as defined in (6). Under proportional priors,

$$\ell^*(\theta) = \ell(\theta; g, \Sigma, \lambda) = |\Lambda(\theta)|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}u(\theta)'\Lambda(\theta)^{-1}u(\theta)\right),$$

where $u(\theta) = \frac{\lambda}{1+\lambda}\psi(\theta)^{-1} g(\theta) + \frac{1}{1+\lambda}\xi$, $\Lambda(\theta) = \frac{\lambda}{1+\lambda}\left[\psi(\theta)^{-1}\right]\Sigma(\theta,\theta)\left[\psi(\theta)^{-1}\right]' + \frac{1}{1+\lambda}Var(\xi)$, and $\xi = A(g)$ is the value of the anchor functional applied to the process g. Hence, for proportional priors and a given choice of λ, the posterior distribution takes a simple form. Standard numerical techniques like Markov chain Monte Carlo can be used to sample from the posterior and implement Bayes decision rules using (8).

Intuitively, the constant of proportionality λ controls the strength of identification under the prior. When $\lambda = 0$ the prior implies that the mean function m is zero with probability one, so nothing can be learned from g and the posterior on θ^* is equal to the prior. Under the diffuse limit, $\lambda \to \infty$, by contrast, the prior implies that m diverges everywhere it is nonzero, and dominates sampling uncertainty. Note that

$$\lim_{\lambda\to\infty} \ell(\theta; g, \Sigma, \lambda) = \lambda(\theta; g, \Sigma, \infty) = |\psi(\theta)| \cdot |\Sigma(\theta,\theta)|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}g(\theta)'\Sigma(\theta,\theta)^{-1}g(\theta)\right). \tag{9}$$

Hence, as $\lambda \to \infty$, $\ell^*(\theta)$ converges to a transformation of the continuously updating GMM objective function, multiplied by factors that do not depend on g and so may be absorbed into the prior. Chernozhukov and Hong (2003) advocate the quasi-likelihood (9) (without the first two terms) as a computational device for point-identified, strongly-identified settings where Bayesian techniques are more tractable than minimization, and show that the resulting estimators are asymptotically efficient. We obtain the same quasi-likelihood as a diffuse-prior limit in a setting that allows for weak and partial identification. Since Bayes decision rules with respect to full-support priors are admissible under mild conditions, quasi-Bayes decision rules based on (9) are the limit of a sequence of admissible decision rules. See Section S3 of the Supplementary Appendix. Indeed, since the limit (9) arises for any choice of anchor A and the term $|\psi(\theta)|$ may be absorbed into the prior, a given quasi-Bayes decision rule corresponds to the limit of many different sequences of priors. Given these desirable properties for quasi-Bayes rules, together with the asymptotic results discussed in Section 4 below, we recommend choosing $\lambda = \infty$.

3.5 Bayesian Approaches to GMM

Several other papers have justified quasi-Bayes decision rules based on (9) from a Bayesian perspective. Closest to our approach, Florens and Simoni (2019) consider Bayesian inference based on an asymptotic normal approximation to a transformation of the data, and obtain the quasi-likelihood (9) as a diffuse-prior limit. Unlike our analysis, however, they specify a Gaussian process prior on the finite-sample density of the data X, rather than on the mean function in the limit experiment. Earlier work by Kim (2002) obtained the same quasi-likelihood via maximum entropy arguments, while Gallant (2016) obtains it as a Bayesian likelihood based on a coarsened sigma-algebra. Unlike our analysis, none of these papers speak to questions of optimality.

Other authors have considered alternative Bayesian approaches for moment condition models that do not run through the quasi-likelihood (9). Chamberlain and Imbens (2003) consider inference for just-identified moment condition models with discrete data, while Bornn et al. (2019) consider discrete data and potentially over-identified moment conditions. Both procedures have a finite-sample Bayesian justification, unlike our approach. Kitamura and Otsu (2011) and Shin (2015) consider Bayesian approaches based on Dirichlet process priors and exponential tilting arguments, while Schennach (2005) shows that a particular generalized empirical likelihood-type objective arises in the limit for a family of nonparametric priors.

3.6 Optimal Tests

While we focus on Bayes decision rules, the calculations needed to derive weighted average power optimal tests are nearly the same. Such tests provide a natural complement to Bayesian approaches to uncertainty quantification such as credible sets. Hence, we briefly describe how our results may be used to construct optimal tests.

Consider the problem of testing $H_0 : \theta^* = \theta_0$ against the composite alternative $H_1 : \theta^* \neq \theta_0$. There generally exists no uniformly most powerful test in this setting, so let us instead maximize average power with respect to weights (i.e. prior) $\pi(\theta, \mu)$ over the alternative. The problem is further complicated by the presence of the infinite-dimensional nuisance parameter μ under the null. Andrews and Mikusheva (2016) show that the process h based on anchor $A(g) = g(\theta_0)$ is sufficient for μ under the null. Building on this result, if we limit attention to tests that are similar, with rejection probability equal to α under the null for all values of μ, optimal test takes a simple form.

Theorem 4 *Consider anchor $A(g) = g(\theta_0)$ and h defined in (5). Let $\pi(\theta, \mu) = \pi(\theta)\pi(\mu)$ be the weight function over $\Theta_0 \times \mathcal{H}_\mu$. Define the test $\varphi^*(\theta_0) = \mathbb{I}\left\{\frac{\int \ell^*(\theta)\pi(\theta)d\theta}{\ell^*(\theta_0)} > c_\alpha(h)\right\}$, for testing the null hypothesis $H_0 : \theta^* = \theta_0$, where $\ell^*(\cdot)$ is defined in (6) and $c_\alpha(h)$ is the $1-\alpha$ quantile of the random variable $\frac{\int \ell^*(\theta)\pi(\theta)d\theta}{\ell^*(\theta_0)}$ conditional on h under the null, provided the last distribution is almost surely continuous. Then $\varphi^*(\theta_0)$ is a similar test, with $\mathbb{E}_{\theta_0,\mu}[\varphi^*(\theta_0)] = \alpha$ for all $\mu \in \mathcal{H}_\mu$. Moreover, $\varphi^*(\theta_0)$ maximizes $\pi(\theta, \mu)$-weighted average power over the class of similar tests, in the sense that for any other test φ with $\mathbb{E}_{\theta_0,\mu}[\varphi] = \alpha$ for all $\mu \in \mathcal{H}_\mu$, $\int \mathbb{E}_{\theta,\mu}[\varphi^*(\theta_0) - \varphi]\, d\pi(\theta, \mu) \geq 0$.*

For further discussion of similarity and the construction of the conditional critical value $c_\alpha(h)$, see Andrews and Mikusheva (2016). Note that while the test $\varphi^*(\theta_0)$ depends on the weight function π, it controls the rejection probability for all parameter values consistent with the null. Hence, the confidence set $CS = \{\theta \in \Theta : \varphi^*(\theta) = 0\}$ formed by inverting this family of tests will have coverage $1-\alpha$ no matter the choice of π.

4 Feasible Procedures

The limit experiment studied in the previous sections treats Θ_0 as known. In practice, however, the structure of weak identification, and thus the set Θ_0, is often unknown. Feasible quasi-Bayes procedures use the normalized sample moment $g_n(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(X_i, \cdot)$ and estimated covariance $\widehat{\Sigma}_n$ in place of the limit process g and known covariance Σ. The researcher specifies a prior over the whole parameter space Θ, and for $Q_n(\theta) =$

$g_n(\theta)' \widehat{\Sigma}_n^{-1}(\theta) g_n(\theta)$ uses the decision rule

$$s_n(g_n) = \arg\min_{a \in \mathcal{A}} \frac{\int_\Theta L(a,\theta)\pi(\theta)\exp\left\{-\frac{1}{2}Q_n(\theta)\right\}d\theta}{\int_\Theta \pi(\theta)\exp\left\{-\frac{1}{2}Q_n(\theta)\right\}d\theta}. \tag{10}$$

This section shows that feasible decision rules (10) are asymptotically equivalent to infeasible rules based on knowledge of Θ_0. In particular, the feasible quasi-posterior $\pi(\widetilde{\Theta}|g_n) = \frac{\int_{\widetilde{\Theta}} \pi(\theta)\exp\left\{-\frac{1}{2}Q_n(\theta)\right\}d\theta}{\int_\Theta \pi(\theta)\exp\left\{-\frac{1}{2}Q_n(\theta)\right\}d\theta}$ concentrates on neighborhoods of Θ_0.[8] Moreover, there exist infeasible decision rules that are asymptotically equivalent to (10) for a large class of loss functions. These rules correspond to a prior π^0 supported on Θ_0 and a transformation of the moments.[9] Specifically, some of the original moments are used to estimate Θ_0, while the remainder are used to form the posterior on Θ_0.

Assumption 1 *The distribution P is such that $\Phi(\theta) = \mathbb{E}_P[\phi(X_i;\theta)]$ and $\Sigma(\cdot,\cdot)$ are continuous, and the determinant of $\Sigma(\theta) = \Sigma(\theta,\theta)$ is nonzero. Further, under P*

$$G_n(\theta) = g_n(\theta) - \sqrt{n}\Phi(\theta) \Rightarrow G \sim \mathcal{GP}(0,\Sigma),$$

and the covariance estimator $\widehat{\Sigma}_n$ is uniformly consistent, $\sup_{\theta\in\Theta}\left\|\widehat{\Sigma}_n(\theta) - \Sigma(\theta)\right\| \to_p 0$.

Assumption 2 *There exists a continuously differentiable function $\vartheta(\beta,\gamma) : \Xi \to \Theta^* \subseteq \Theta$ where $\Theta_0 \subseteq \Theta^*$, $\Xi = \{(\beta,\gamma) : \beta \in B, \gamma \in \Gamma(\beta) \subseteq \mathbb{R}^{p_\gamma}\}$ is compact, $\vartheta(\beta,\gamma) \in \Theta_0$ if and only if $\gamma = 0$, and 0 lies in the interior of $\Gamma(\beta)$ for all $\beta \in B$. There exist a positive measure $\pi(\beta,\gamma)$ on Ξ such that $\pi(\theta)$ on Θ^* is the pushforward of $\pi(\beta,\gamma)$ under ϑ. The conditional prior on γ given β has uniformly bounded density $\pi_\gamma(\gamma|\beta)$ that is uniformly continuous and positive at $\gamma = 0$, and $\int_B d\pi(\beta) > 0$.*

We adopt the shorthand $\Phi(\beta,\gamma) = \Phi(\vartheta(\beta,\gamma))$ and $\Phi(\beta) = \Phi(\beta,0)$ for all functions.

Assumption 3 *The function $\Phi(\beta,\gamma)$ is uniformly (over $\beta \in B$) differentiable in γ at $\gamma = 0$. Further, for $\nabla(\beta) = \frac{\partial}{\partial\gamma}\Phi(\beta)$, $J(\beta) = \frac{1}{2}\nabla(\beta)'\Sigma(\beta)^{-1}\nabla(\beta)$ is everywhere positive definite.*

[8] Liao and Jiang (2010) establish a similar consistency result for the case where the weighting matrix does not vary with θ. Chen, Christensen, and Tamer (2018) characterize the behavior of the quasi-posterior when the GMM objective depends on a finite-dimensional reduced-form parameter.

[9] We focus on unconstrained decision rules for brevity, but a similar analysis applies for tests.

Assumption 1 standard for asymptotic analysis. Assumption 2 imposes that on a neighborhood of Θ_0 there exists some (unknown to the researcher) re-parameterization of the model in terms of β and γ, where β indexes the weakly or partially identified parameter, while γ can be called strongly identified. The set Θ_0 corresponds to $\gamma = 0$, and is parameterized by $\beta \in B$. Han and McCloskey (2019) provide sufficient conditions for such a reparameterization to exist. The mapping from (β, γ) to θ can be many-to-one, and we impose very little structure on the set B, which may, for example, be a collection of points or intervals. We also note that $\pi(\beta, \gamma)$ need not integrate to one, since Θ^* may be a strict subset of Θ. Assumption 3 requires that γ be strongly identified, in the sense that the Jacobian of the moments with respect to γ has full rank at $\gamma = 0$.

Theorem 5 *Assume that P satisfies Assumptions 1, 2, and 3. If the prior $\pi(\theta)$ has bounded density on the set Θ, then for any sequence $c_n \to \infty$, under sequences $P_{n,f}$ local to P in the sense of (1) we have*

$$\pi\left(\left\{\theta \in \Theta : \Phi(\theta)\Sigma(\theta)^{-1}\Phi(\theta) \geq \frac{c_n}{n}\right\} | g_n\right) = o_p(1). \tag{11}$$

Moreover, for any bounded function $c(\theta)$ uniformly continuous at Θ_0

$$\int_\Theta c(\theta) d\pi(\theta|g_n) - \frac{\int_B c(\vartheta(\beta)) \exp\left\{-\frac{1}{2}Q_n^\beta(\beta)\right\} d\pi^0(\beta)}{\int_B \exp\left\{-\frac{1}{2}Q_n^\beta(\beta)\right\} d\pi^0(\beta)} \to^p 0, \tag{12}$$

where $d\pi^0(\beta) = \pi_\gamma(0|\beta)|J(\beta)|^{-\frac{1}{2}} d\pi(\beta)$, $Q_n^\beta(\beta) = g_n(\beta)' M(\beta) g_n(\beta)$, and

$$M(\beta) = \Sigma(\beta)^{-1} - \Sigma(\beta)^{-1} \nabla(\beta) J(\beta)^{-1} \nabla(\beta)' \Sigma(\beta)^{-1}.$$

Theorem 5 is a version of the Bernstein-von Mises theorem for weakly and partially identified quasi-Bayesian settings. The GMM objective function $Q_n(\theta)$ is bounded on Θ_0 but diverges away from Θ_0. As (11) highlights, this forces the posterior to concentrate on infinitesimal neighborhoods of Θ_0, corresponding to consistent estimation of the strongly identified parameter γ. The rank $k - p_\gamma$ matrix $M(\beta)$ then selects the linear combination of moments orthogonal to those used to estimate γ, and this combination is used to form the posterior on β. Unlike in the classical Bernstein-von Mises theorem, the prior on Θ_0 (i.e. on B) matters asymptotically, and is adjusted based on the precision of the estimate for γ as measured by $J(\beta)$. Overall, we obtain that feasible quasi-Bayes posteriors are asymptotically equivalent to infeasible posteriors based on a transformation of the prior and moment conditions. This likewise implies asymptotic equivalence of feasible and infeasible decision rules.

Corollary 2 *Let the assumptions of Theorem 5 hold. Assume that the loss function $L(a,\theta)$ is Lipschitz in a and continuous in θ over Θ^*, and that $\mathcal{A}$ is compact. Assume further that for almost all realization of process $G(\beta) \sim \mathcal{GP}(0,\Sigma)$, the process $L(a) = \int_B L(a,\vartheta(\beta)) \exp\left\{-\frac{1}{2}G(\beta)' M(\beta) G(\beta)\right\} d\pi^0(\beta)$ has a unique minimizer over $\mathcal{A}$. Then*

$$s_n^0(g_n) = \arg\min_{a\in\mathcal{A}} \frac{\int_B L(a,\vartheta(\beta)) \exp\left\{-\frac{1}{2}Q_n^\beta(\beta)\right\} d\pi^0(\beta)}{\int_B \exp\left\{-\frac{1}{2}Q_n^\beta(\beta)\right\} d\pi^0(\beta)} \to^p s_n(g_n).$$

Uniqueness of the minimizer $L(a)$ is guaranteed to hold if the loss function is convex in a. Sufficient conditions for uniqueness in non-convex cases are discussed in Cox (2020).

Overall, we obtain that feasible quasi-Bayes decision rules, computed without knowledge of Θ_0, converge to infeasible quasi-Bayes rules based on knowledge of Θ_0 and a transformation of the moments and prior. These rules are, in turn, the limit of sequences of proper-prior Bayes decision rules in the limit problem by our previous results.

5 Empirical Illustration

Following Kim (2002) and Cherzhukov and Hong (2003), quasi-Bayes procedures have been used in a range of applications. Here, we briefly illustrate our results with an application of quantile IV to data from Graddy (1995) on the demand for fish at the Fulton fish market. Following Chernozhukov et al. (2009), who discuss finite-sample frequentist inference in this setting, we consider the quantile IV moment conditions stated in equation (2) with Y the log quantity of fish purchased, W the log price, and Z a vector of instruments consisting of a constant, a dummy for whether the weather offshore was mixed (with wave height above 3.8 feet and windspeed over 13 knots), and a dummy for whether the weather offshore was stormy (with wave height above 4.5 feet and windspeed over 18 knots). For further details on the data and setting, see Graddy (1995) and Chernozhukov et al. (2009).

The results of Chernozhukov et al. (2009) suggest that the data are not very informative when we consider inference on the 0.25 or 0.75 quantiles, consistent with weak

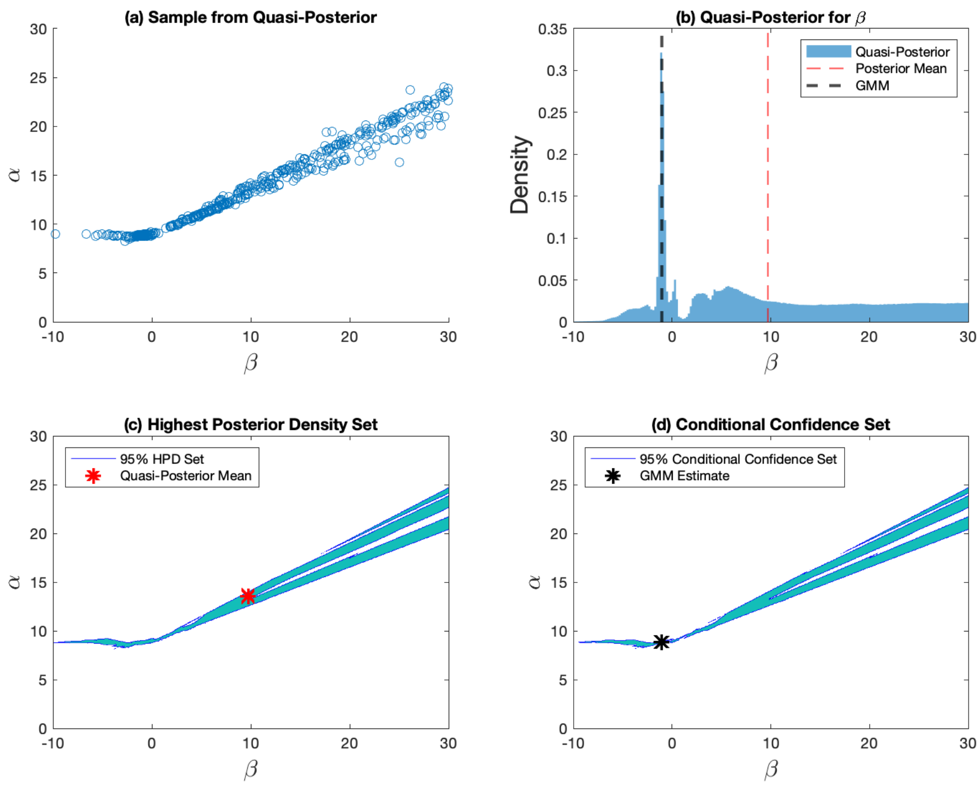

Figure 1: Results for Graddy (1995) data

identification. Here we discuss results for the 0.75 quantile. Following Chernozhukov et al. (2009), we restrict attention to $\alpha \in [0, 30]$ and $\beta \in [-10, 30]$, and consider a flat prior $\pi(\theta^*)$ on $\theta = (\alpha, \beta)$.[10] We calculate the quasi-Bayes posterior distribution discussed in Section 4 using random-walk Metropolis Hastings with ten million draws. We plot 500 draws from the quasi-Bayes posterior in panel (a) of Figure 1. Panel (b) plots the marginal quasi-Bayes posterior distribution for the price coefficient β, with vertical lines marking the (continuously-updating) GMM estimate and the quasi-posterior mean. Panel (c) plots the 95% highest posterior density set, while panel (d) plots the conditional frequentist confidence set discussed in Section 3.6, along with the GMM estimate.

A few aspects of these results warrant discussion. Given the flat prior, the quasi-posterior is a monotonic transformation of the GMM objective function. Figure 1 thus highlights that the GMM objective is far from quadratic in this case, so conventional

[10] These choices are discussed in the working paper version, Chernozhukov et al. (2006).

asymptotic results based on quadratic approximations may be unreliable. By contrast, the results in Figure 1 are entirely consistent with weak and partial identification. Panel (a) shows that the quasi-posterior is concentrated around a lower-dimensional set, in line with the asymptotic results of Theorem 5, and the form of Θ_0 in Example 2. Panel (b) shows that the GMM estimate is quite different from the quasi-posterior mean and that the quasi-posterior is highly non-normal, again contrary to what we would expect in the point-identified, strongly-identified case (see Chernozhukov and Hong, 2003). The quasi-posterior mean seems a more reasonable summary than the GMM estimate, as the latter ignores the large region of uncertainty to the right.

We also report credible and confidence sets. Panel (c) reports the quasi-Bayesian highest posterior density credible set, which has no frequentist coverage guarantees in the current setting.[11] Panel (d) reports the frequentist confidence set obtained through inversion of the weighted average power optimal conditional tests discussed in Section 3.6, using the quasi-Bayes objective. The quasi-Bayesian credible set and the frequentist confidence set have a quite similar shape, but the frequentist confidence set is slightly smaller, covering 4.74% of the parameter space, as compared to 4.82% for the highest posterior density set.

6 References

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

Andrews, D.W.K., M. Moreira, and J. Stock (2006), "Optimal Two-Sided Invariant Similar Tests of Instrumental Variables Regression," *Econometrica*, 74(3), 715-752

Andrews, I. and A. Mikusheva (2016), "Conditional Inference with a Functional Nuisance Parameter," *Econometrica*, 84(4), 1571-1612

Berlinet, A., and C. Thomas-Agnan (2004) *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers

[11] The results of Chen et al (2018) show that quasi-Bayesian highest posterior density sets have correct asymptotic coverage of the identified set in certain partially identified settings. This result does not extend to general weakly identified models, however, since highest posterior density sets are necessarily a strict subset of the parameter space and so under-cover the identified set when $\Theta_0 = \Theta$.

Bornn, L., N. Shephard, and R. Solgi (2019) "Moment Conditions and Bayesian Nonparametrics," *Journal of the Royal Statistical Society, Series B*, 81: 5-43

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

Chamberlain, G. and G. Imbens (2003): "Nonparametric Applications of Bayesian Inference," *Journal of Business and Economic Statistics*, 21, 12–18

Chen, X., T. Christensen, and E. Tamer (2018). "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86(6), 1965-2018

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245-261

Chernozhukov, V., C. Hansen, and M. Jansson (2006), "Finite Sample Inference for Quantile Regression Models," Unpublished Manuscript

Chernozhukov, V., C. Hansen, and M. Jansson (2009), "Finite Sample Inference for Quantile Regression Models," *Journal of Econometrics*, 152(2), 93-103

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

Cox, G. (2020), "Almost Sure Uniqueness of a Global Minimum Without Convexity," *Annals of Statistics*, 48(1), 584-606

Florens, J.P. and A. Simoni (2012): "Nonparametric Estimation of an Instrumental Regression: A Quasi-Bayesian Approach Based on Regularized Posterior," *Journal of Econometrics*, 170, 458-475

Florens, J.P. and A. Simoni (2019): "Gaussian Processes and Bayesian Moment Estimation," *Journal of Business and Economic Statistics*, Forthcoming.

Gallant, R. (2016), "Reflections on the Probability Space Induced by Moment Conditions with Implications for Bayesian Inference," *Journal of Financial Econometrics*, 14(2), 284–294

Graddy, K. (1995), "Testing for Imperfect Competition at the Fulton Fish Market," *Rand Journal of Economics,* 26(1), 75-92

Han, S. and A. McCloskey (2019), "Estimation and Inference with a (Nearly) Singular Jacobian," *Quantitative Economics*, 10(3), 1019-1068

Hirano, K. and J. Porter (2009), "Asymptotics for Statistical Treatment Rules," *Econo-*

metrica, 77(5), 1683-1701

Kaji, T. (2020): "Theory of Weak Identification in Semiparametric Models," Unpublished Manuscript

Kim, J.Y. (2002): "Limited Information Likelihood and Bayesian Analysis," *Journal of Econometrics*, 107, 175-193

Kitamura, Y. and T. Otsu (2011), "Bayesian Analysis of Moment Restriction Models Using Nonparametric Priors," Unpublished Manuscript

Le Cam, L. (1986), *Asymptotic Theory of Statistical Inference,* John Wiley & Sons

Lehman, E.L. and G. Casella (1998), *Theory of Point Estimation,* Springer Texts in Statistics

Liao, Y. and W. Jiang (2010), "Bayesian Analysis in Moment Inequality Models," *Annals of Statistics*, 38(1), 275-316

Moreira, H. and M. Moreira (2019), "Optimal Two-Sided Tests for Instrumental Variables Regression with Heteroskedastic and Autocorrelated Errors," *Journal of Econometrics*, 213(2), 398-433

Moreira, M. and G. Ridder (2019), "Optimal Invariant Tests in an Instrumental Variables Regression With Heteroskedastic and Autocorrelated Errors," Unpublished Manuscript

Mueller, U.K. (2011), "Efficient Tests Under a Weak Convergence Assumption," *Econometrica*, 79(2), 395-435

Neveu J . (1968) *Processus AIeatoires Gaussiens.* Seminaire Math. Sup., Les presses de I'Universite de Montreal

Parzen, E. (1962), "Extraction and Detection Problems and Reproducing Kernel Hilbert Spaces," *J. SIAM Control*, Ser. A., 1(1) , 35–62

Schennach, S. M. (2005), "Bayesian Exponentially Tilted Empirical Likelihood," *Biometrika* 92, 31–46

Shin, M. (2015), "Bayesian GMM," Dissertation Chapter, University of Pennsylvania

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

Van der Vaart, A.W. (1991) "An Asymptotic Representation Theorem," *International Statistical Review*, Vol. 59, No. 1, pp. 97-121

Van der Vaart , A.W. and J.A. Wellner (1996), *Weak Convergence and Empirical Pro-*

cesses, 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

7 Appendix

Proof of Lemma 1. A score $f(X) = \sum_{i=1}^{s} \phi(X, \theta_i)' a_i$ corresponds to the function

$$m(\cdot) = \mathbb{E}_P[f(X)\phi(X, \cdot)] = \sum_{i=1}^{s} \mathbb{E}_P[\phi(X, \cdot)\phi(X, \theta_i)' a_i] = \sum_{i=1}^{s} \Sigma(\cdot, \theta_i) a_i.$$

For two scores $f_1(X) = \sum_{i=1}^{s} \phi(X, \theta_i)' a_i$ and $f_2(X) = \sum_{j=1}^{s^*} \phi(X, \theta_j^*)' b_j$ and corresponding mean functions $m_1(\cdot) = \sum_{i=1}^{s} \Sigma(\cdot, \theta_i) a_i$ and $m_2(\cdot) = \sum_{j=1}^{s^*} \Sigma(\cdot, \theta_j^*) b_j$ we have

$$\mathbb{E}_P\left[f_1(X) f_2(X)\right] = \sum_{i=1}^{s} \sum_{j=1}^{s^*} a_i' \mathbb{E}_P\left[\phi(X, \theta_i)\phi(X, \theta_j^*)'\right] b_j = \sum_{i=1}^{s} \sum_{j=1}^{s^*} a_i' \Sigma(\theta_i, \theta_j^*) b_j = \langle m_1, m_2 \rangle_{\mathcal{H}}.$$

This implies that there is an isomorphism between $\mathcal{H}$ and $\mathcal{H}^*$.

It remains to show is that for any $f^\perp \in T(P)$ that is orthogonal (in $L_2(P)$ sense) to $\mathcal{H}^*$ we have $\mathbb{E}_P[f^\perp(X)\phi(X, \cdot)] \equiv 0_k \in \mathbb{R}^k$. Indeed, $f^\perp$ is orthogonal to $a'\phi(X, \theta) \in \mathcal{H}^*$ for any vector $a \in \mathbb{R}^k$ and any $\theta \in \Theta_0$. Thus $\mathbb{E}_P[f^\perp(X) a' \phi(X, \theta)] = a' m(\theta) = 0$.

Proof of Theorem 1. Define the orthonormal basis $\{\phi_j(X)\}$ of $T(P)$ to consist of the union of an orthonormal basis $\{\phi_j^*(X)\}$ of $\mathcal{H}^*$ and an orthonormal basis $\{\phi_j^\perp(X)\}$ of $(\mathcal{H}^*)^\perp$. The limit experiment $\mathcal{E}_\infty^*$ corresponds to observing the union of two sets of mutually independent random variables:

$$W_j^* \sim N\left(\mathbb{E}_P[f(X)\varphi_j^*(X)], 1\right) \text{ and } W_j^\perp \sim N\left(\mathbb{E}_P[f(X)\varphi_j^\perp(X)], 1\right).$$

Due to Lemma 1 $\mathbb{E}_P[f(X)\varphi_j^*(X)] = \langle m, \varphi_j^* \rangle_{\mathcal{H}}$. The experiment of observing only $W_j^* \sim N\left(\langle m, \varphi_j^* \rangle_{\mathcal{H}}, 1\right)$ is equivalent to the Gaussian Process experiment $\mathcal{E}_{GP}^*$ by the Karhunen-Loeve theorem.

By independence $dP_f(W^*, W^\perp) = dP_{f^*}(W^*) \times dP_{f^\perp}(W^\perp)$. The loss function depends only on θ^*, and the parameter space for $(\theta^*, f^*, f^\perp)$ is the Cartesian product $\{\theta^* \in$

$\Theta_0, f^* \in \mathcal{H}^*_{\theta^*}\} \times \{f^\perp \in (\mathcal{H}^*)^\perp\}$. The risk of a decision rule δ is

$$\widetilde{R}(\theta^*, f) = \widetilde{R}(\theta^*, f^*, f^\perp) = \mathbb{E}_f \left[L(\delta(W^*, W^\perp), \theta^*) \right].$$

We claim that for any fixed value $f^\perp$ there exists a decision rule in the experiment $\mathcal{E}^*_{GP}$ with risk $R(\theta^*, m) = \widetilde{R}(\theta^*, f^*, f^\perp)$ for all $(\theta^*, f^*) \in \{\theta^* \in \Theta_0, f^* \in \mathcal{H}^*_{\theta^*}\}$, where m corresponds to f^* as in Lemma 1. Indeed, since experiment $\mathcal{E}^*_{GP}$ is equivalent to observing only the W^*_j variables, it is enough for each realization $W^* = w$ to draw a random variable $(W^\perp)$ from distribution $dP_{f^\perp}$ (which is fixed) and produce a randomized decision as $\widetilde{\delta}(w) = \delta(w, W^\perp)$.

Proof of Theorem 2. The distribution of g for any $m \in \mathcal{H}$ is dominated by the distribution under $m = 0$. Moreover, the form of the likelihood ratio for Gaussian processes (see e.g. Theorem 54 in Berlinet and Thomas-Agnan) implies that condition (1) in Section 4A.1 of Brown (1986) holds. Our assumptions likewise imply condition (2) of Brown (1986). This result is thus immediate from Theorem 4A.12 of Brown (1986).

Proof of Theorem 3. We first find the conditional mean of $\xi \sim N(-[\psi(\theta^*)]^{-1}\mu(\theta^*), \Sigma_\xi)$ given the realization of $h = \mu + GP(0, \widetilde{\Sigma})$, assuming $\mu \sim GP(0, \Omega)$. Neveu (1968) proves that for the Gaussian family the conditional mean coincides with the best linear predictor. Note next that

$$\mathbb{E}[\xi h(\cdot)] = -[\psi(\theta^*)]^{-1}\Omega(\theta^*, \cdot) = \rho(\cdot), \text{ and } \mathbb{E}[h(\theta_1)h(\theta_2)] = \Omega(\theta_1, \theta_2) + \widetilde{\Sigma}(\theta_1, \theta_2).$$

Denote by $\mathcal{K}$ the RKHS corresponding to the covariance function $\Omega + \widetilde{\Sigma}$, and by $L(h)$ the subspace of $L_2(P)$ random variables obtained as the closure of linear combinations of h.

Define ξ^* as the projection of ξ on to $L(h)$. By definition it is the best linear predictor of ξ given h, and $\mathbb{E}[\xi h(\cdot)] = \mathbb{E}[\xi^* h(\cdot)] = \rho(\cdot)$. Lemma 13 in Berlinet and Thomas-Agnan (2004) implies that $\rho(\cdot) \in \mathcal{K}$. Denote by Ψ the canonical congruence between $L(h)$ and $\mathcal{K}$, defined by

$$\Psi\left(\sum_j a_j h(\theta_j)\right) = \sum_j a_j (\Omega(\theta_j, \cdot) + \widetilde{\Sigma}(\theta_j, \cdot)) \in \mathcal{K},$$

and extended by continuity. Then $\xi^* = \Psi^{-1}(\rho(\cdot))$. See Section 3 of Berlinet and Thomas-Agnan (2004) for further discussion.

We next fix θ^*, assume that condition (7) holds, and show that the best linear predictor depends on $(\xi, g(\theta^*))$ only. Condition (7) implies that

$$\Omega(\theta^*, \cdot) + \widetilde{\Sigma}(\theta^*, \cdot) = \left(I_k + \widetilde{\Sigma}(\theta^*, \theta^*)\Omega(\theta^*, \theta^*)^{-1}\right)\Omega(\theta^*, \cdot).$$

Thus,

$$\rho(\cdot) = -[\psi(\theta^*)]^{-1}\left(I_k + \widetilde{\Sigma}(\theta^*, \theta^*)\Omega(\theta^*, \theta^*)^{-1}\right)^{-1}\left[\Omega(\theta^*, \cdot) + \widetilde{\Sigma}(\theta^*, \cdot)\right],$$

and the canonical congruence has the form

$$\xi^* = \Psi^{-1}(\rho(\cdot)) = -[\psi(\theta^*)]^{-1}\left(I_k + \widetilde{\Sigma}(\theta^*, \theta^*)\Omega(\theta^*, \theta^*)^{-1}\right)^{-1} h(\theta^*),$$

which depends on the data only through $h(\theta^*) = g(\theta^*) - \psi(\theta^*)\xi$.

Finally, we prove the converse, assuming that for each θ^* the likelihood depends only on $(\xi, g(\theta^*))$ and proving that (7) holds. Since the conditional distribution of ξ given θ^* and ξ^* is $N(\xi^*, \Sigma_\xi)$, ξ^* must depend only on $(\xi, g(\theta^*))$ or, equivalently, on $(\xi, h(\theta^*))$. Linearity of ξ^* in h then implies that there exists a non-random $k \times k$ matrix $B(\theta^*)$ such that

$$\Psi^{-1}(\rho(\cdot)) = B(\theta^*)h(\theta^*).$$

By the definition of the canonical congruence this implies $\rho(\cdot) = B(\theta^*)\left[\Omega(\theta^*, \cdot) + \widetilde{\Sigma}(\theta^*, \cdot)\right]$. Since $\rho(\cdot) = -[\psi(\theta^*)]^{-1}\Omega(\theta^*, \cdot)$, however, $[\psi(\theta^*)]^{-1}\Omega(\theta^*, \cdot) = B(\theta^*)\left[\Omega(\theta^*, \cdot) + \widetilde{\Sigma}(\theta^*, \cdot)\right]$. Since $\psi(\theta^*)$ has full rank, both sides are invertible when evaluated at θ^*, and some rearrangement yields (7).

Proof of Lemma 3. Let $B = \Omega(\theta_0, \theta_0)$. Condition (7) implies that

$$\Omega(\theta_0, \theta) = B\widetilde{\Sigma}(\theta_0, \theta_0)^{-1}\widetilde{\Sigma}(\theta_0, \theta) \text{ and } \Omega(\theta, \theta)^{-1}\Omega(\theta, \theta_0) = \widetilde{\Sigma}(\theta, \theta)^{-1}\widetilde{\Sigma}(\theta, \theta_0).$$

The transposed equations are

$$\Omega(\theta, \theta_0) = \widetilde{\Sigma}(\theta, \theta_0)\widetilde{\Sigma}(\theta_0, \theta_0)^{-1}B \text{ and } \Omega(\theta_0, \theta)\Omega(\theta, \theta)^{-1} = \widetilde{\Sigma}(\theta_0, \theta)\widetilde{\Sigma}(\theta, \theta)^{-1}.$$

We can calculate $\Omega(\theta_0, \theta)\Omega(\theta, \theta)^{-1}\Omega(\theta, \theta_0)$ in two ways, so

$$\widetilde{\Sigma}(\theta_0, \theta)\widetilde{\Sigma}(\theta, \theta)^{-1}\widetilde{\Sigma}(\theta, \theta_0)\widetilde{\Sigma}(\theta_0, \theta_0)^{-1}B = B\widetilde{\Sigma}(\theta_0, \theta_0)^{-1}\widetilde{\Sigma}(\theta_0, \theta)\widetilde{\Sigma}(\theta, \theta)^{-1}\widetilde{\Sigma}(\theta, \theta_0).$$

Pre and post multiply the last equation with $\widetilde{\Sigma}(\theta_0,\theta_0)^{-1/2}$ and let

$$\widetilde{B}=\widetilde{\Sigma}(\theta_0,\theta_0)^{-1/2}B\widetilde{\Sigma}(\theta_0,\theta_0)^{-1/2}.$$

We obtain that $\widetilde{B}$ commutes with a whole family of symmetric matrices : $D(\theta)\widetilde{B}=\widetilde{B}D(\theta)$. Assume $\widetilde{B}$ has r distinct eigenvalues. Since $\widetilde{B}$ is symmetric, all eigenvectors corresponding distinct eigenvalues are orthogonal. Let V_1,..., V_r be the orthogonal subspaces spanned by eigenvectors of the $\widetilde{B}$ corresponding to eigenvalues $\lambda_1,...,\lambda_r$, respectively. Consider a symmetric matrix $D(\theta)\in\mathcal{D}$ that commutes with $\widetilde{B}$. Take any $v_i\in V_i$ and $v_j\in V_j$:

$$v_i'D(\theta)\widetilde{B}v_j=\lambda_jv_i'D(\theta)v_j=v_i'\widetilde{B}D(\theta)v_j=\lambda_iv_i'D(\theta)v_j.$$

This implies $v_i'D(\theta)v_j=0$ for any $i\neq j$. Thus $D(\theta)v_j\in V_j$, and V_1,..., V_r are invariant subspaces for $D(\theta)$. Thus, we proved that V_1,...,V_r are invariant spaces for the whole family of operators $\mathcal{D}$. Under the conditions of the lemma this implies that $\widetilde{B}$ has single eigenvalue $\lambda>0$, and thus $\Omega(\cdot,\cdot)=\lambda\widetilde{\Sigma}(\cdot,\cdot)$.

Proof of Theorem 4. Similarity of $\varphi^*(\theta_0)$ follows from Lemma 1 in Andrews and Mikusheva (2016). Theorem S2.1 in the supplement to Andrews and Mikusheva (2016) implies that any similar test in this setting must be conditionally similar given h, with $\mathbb{E}_{\theta_0,\mu}[\varphi|h]=\alpha$ almost surely. For any $\mu\in\mathcal{H}_\mu$ and any test φ,

$$\int\int\mathbb{E}_{\theta,\mu}[\varphi]\,d\pi(\mu)\,d\pi(\theta)=\mathbb{E}_\pi[\varphi]=\mathbb{E}_{\theta_0,\mu}\left[\frac{\int\ell^*(\theta)\,d\pi(\theta)f(h)}{\ell(\mu,\theta_0;\xi)\,\ell(\mu;h)}\varphi\right].$$

Since $\xi=g(\theta_0)$, $\ell(\mu,\theta_0;\xi)$ does not depend on μ, and is equal to $\ell^*(\theta_0)$. Lemma 2 of Andrews and Mikusheva (2016) implies that $\tilde{\varphi}(\theta_0)=1\left\{\frac{\int\ell^*(\theta)d\pi(\theta)f(h)}{\ell^*(\theta_0)\ell(\mu;h)}>\tilde{c}_\alpha(h)\right\}$ maximizes $\mathbb{E}_\pi[\varphi]$ over the class of size-α similar tests, where $\tilde{c}_\alpha(h)$ is the $1-\alpha$ quantile of $\frac{\int\ell^*(\theta)d\pi(\theta)f(h)}{\ell^*(\theta_0)\ell(\mu;h)}$ conditional on h under the null. The test statistic in $\tilde{\varphi}(\theta_0)$ differs from that in $\varphi^*(\theta_0)$ only through terms depending on h. These can be absorbed into the critical value, so $\tilde{\varphi}(\theta_0)=\varphi^*(\theta_0)$.

Proof of Theorem 5. By contiguity it is enough to prove the statement under P. Denote $Q(\theta)=\Phi(\theta)'\Sigma(\theta)^{-1}\Phi(\theta)$. Due to Assumption 1, $\widehat{\Sigma}_n(\theta)^{-1}$ is uniformly bounded in probability and $G_n(\cdot)\Rightarrow\mathcal{GP}(0,\Sigma)$, thus $\max_{\theta\in\Theta}G_n(\theta)'\widehat{\Sigma}_n(\theta)^{-1}G_n(\theta)=O_p(1)$. Since

$g_n(\theta) = \sqrt{n}\Phi(\theta) + G_n(\theta)$ we have

$$\frac{1}{2}Q_n(\theta) \leq nQ(\theta)(1+o_p(1)) + \max_{\theta\in\Theta} G_n(\theta)'\widehat{\Sigma}_n(\theta)^{-1}G_n(\theta);$$
$$\frac{1}{2}Q_n(\theta) \geq \frac{n}{2}Q(\theta)(1+o_p(1)) - \max_{\theta\in\Theta} G_n(\theta)'\widehat{\Sigma}_n(\theta)^{-1}G_n(\theta).$$

Define a set $\Theta_{\delta,n} = \left\{\theta\in\Theta : Q(\theta) \leq \frac{\delta}{n}\right\}$ for some $\delta > 0$ and $\Theta_{c_n}^c = \{\theta\in\Theta : Q(\theta) \geq \frac{c_n}{n}\}$.

$$\pi\left(\Theta_{c_n}^c|g_n\right) \leq \frac{\int_{\Theta_{c_n}^c}\pi(\theta)\exp\left\{-\frac{1}{2}Q_n(\theta)\right\}d\theta}{\int_{\Theta_{\delta,n}}\pi(\theta)\exp\left\{-\frac{1}{2}Q_n(\theta)\right\}} \leq O_p(1)\cdot\frac{\int_{\Theta_{c_n}^c}\pi(\theta)\exp\{-\frac{n}{2}Q(\theta)\}d\theta}{\int_{\Theta_{\delta,n}}\pi(\theta)d\theta}.$$

Due to uniform differentiability of $\Phi(\beta,\gamma)$ there exist positive constants C_1, C_2 and small enough $\varepsilon > 0$ such that for all $\theta\in\Theta_\varepsilon^* = \{\theta = \vartheta(\beta,\gamma) : \|\gamma\| < \varepsilon\}$ we have $C_1\|\gamma\|^2 \leq Q(\beta,\gamma) \leq C_2\|\gamma\|^2$. For large enough n we have $\Theta_{\delta,n}\subseteq\Theta_\varepsilon^*$. Thus,

$$\int_{\Theta_{\delta,n}}\pi(\theta)d\theta = \int_B\int_{Q(\beta,\gamma)\leq\frac{\delta}{n}}\pi_\gamma(\gamma|\beta)d\gamma d\pi(\beta) \geq C\int_{\|\gamma\|^2\leq\frac{\delta}{C_2 n}}d\gamma \geq Cn^{-\frac{p_\gamma}{2}}.$$

Divide the integral over $\Theta_{c_n}^c$ into integrals over $\Theta_{c_n}^c\cap\Theta_\varepsilon^*$ and over $\Theta_{c_n}^c\cap(\Theta_\varepsilon^*)^c$, where $(\Theta_\varepsilon^*)^c = (\Theta\setminus\Theta_\varepsilon^*)$. We have $\Theta_{c_n}^c\cap\Theta_\varepsilon^*\subseteq\{\theta=\vartheta(\beta,\gamma) : C_2\|\gamma\|^2 \geq \frac{c_n}{n}\}$. Denote by $\bar{Q}$ the non-zero minimum of $Q(\theta)$ over $(\Theta_\varepsilon^*)^c$. Thus,

$$\frac{\int_{\Theta_{c_n}^c}\pi(\theta)\exp\left\{-\frac{n}{2}Q(\theta)\right\}d\theta}{\int_{\Theta_{\delta,n}}\pi(\theta)d\theta} \leq Cn^{\frac{p_\gamma}{2}}\left(\exp\left\{-\frac{n}{2}\bar{Q}\right\} + \int_{C_2\|\gamma\|^2\geq\frac{c_n}{n}}\exp\{-nC_1\|\gamma\|^2\}d\gamma\right) \leq$$
$$\leq o(1) + \int_{C_2\|y\|^2\geq c_n}\exp\{-C_1\|y\|^2\}dy \to 0.$$

In the last line we used the change of variables $y = \sqrt{n}\gamma$ and integrability of $\exp\{-\|y\|^2\}$. This proves (11), and implies that for $\pi_{\Theta_{c_n}}$ the prior restricted to Θ_{c_n}, the posterior $\pi_{\Theta_{c_n}}(\Upsilon|g_n) = \frac{\pi(\Upsilon\cap\Theta_{c_n}|g_n)}{\pi(\Theta_{c_n}|g_n)}$ defined on sets $\Upsilon\subseteq\Theta$ is asymptotically the same as $\pi(\Upsilon|g_n)$. If $\frac{c_n}{n}\to 0$, then due to compactness of Θ^*, for large enough n we have $\Theta_{c_n}\subseteq\Theta^*$. Thus, we can treat the parameterization described in Assumption 2 as applying to the whole parameter space Θ.

The above implies that for any $\varepsilon > 0$ there exists δ large enough such that

$$P\left\{\sup_\beta n^{\frac{p_\gamma}{2}}\int_{\frac{\delta}{n}\leq\|\gamma\|^2}\exp\left\{-\frac{1}{2}Q_n(\beta,\gamma)\right\} \geq \varepsilon\right\} \leq \varepsilon, \tag{13}$$

and also $\sup_\beta P\left\{\|N(0,J^{-1}(\beta))\| \geq \delta\right\} < \varepsilon$. Define $g_n^0(\beta,\gamma) = g_n(\beta) + \sqrt{n}\nabla(\beta)\gamma$ and $R_n(\beta,\gamma) = g_n(\beta,\gamma) - g_n^0(\beta,\gamma)$. Let us show that

$$\sup_\beta\sup_{\|\gamma\|^2\leq\frac{\delta}{n}}\|R_n(\beta,\gamma)\| \to^p 0. \tag{14}$$

Indeed, $\|R_n(\beta,\gamma)\| \leq \sqrt{n}\|\Phi(\beta,\gamma) - \nabla(\beta)\gamma\| + \|G_n(\beta,\gamma) - G_n(\beta)\|$. We have that $\sup_\beta \sup_{\|\gamma\|^2\leq\frac{\delta}{n}} \|G_n(\beta,\gamma) - G_n(\beta)\| \to^p 0$ due to stochastic equicontinuity. Uniform differentiability implies $\sup_\beta \sup_{\|\gamma\|^2\leq\frac{\delta}{n}} \sqrt{n}\|\Phi(\beta,\gamma) - \nabla(\beta)\gamma\| \to 0$.

Denote $Q_n^0(\beta,\gamma) = g_n^0(\beta,\gamma)\Sigma^{-1}(\beta)g_n^0(\beta,\gamma)$. Equation (14) implies that

$$\sup_\beta \sup_{\|\gamma\|^2\leq\frac{\delta}{n}} \left|1 - \exp\left\{-\frac{1}{2}(Q_n(\beta,\gamma) - Q_n^0(\beta,\gamma))\right\}\right| \to^p 0. \tag{15}$$

Indeed, the left-hand side is bounded above by

$$\begin{aligned}&\sup_\beta \sup_{\|\gamma\|^2\leq\frac{\delta}{n}} |Q_n(\beta,\gamma) - Q_n^0(\beta,\gamma)| \leq \\ \leq &\sup_\beta \sup_{\|\gamma\|^2\leq\frac{\delta}{n}} \left\{|(g_n + g_n^0)\widehat{\Sigma}_n^{-1}R_n| + |g_n^0(\widehat{\Sigma}_n^{-1}(\beta,\gamma) - \Sigma^{-1}(\beta))g_n^0|\right\} \to^p 0.\end{aligned}$$

The last convergence follows from continuity of covariance function Σ, equation (14), and boundedness in probability of g_n, g_n^0 and $\widehat{\Sigma}^{-1}$ over $\{\|\gamma\|^2 \leq \frac{\delta}{n}\}$.

Denote $Q_n^\beta(\beta) = g_n(\beta)' M(\beta) g_n(\beta)$. Let us define a projection operator $P(\beta) = \Sigma^{-\frac{1}{2}}(\beta)\nabla(\beta)J(\beta)^{-1}\nabla(\beta)'\Sigma^{-\frac{1}{2}}(\beta)$. Notice that $M(\beta) = \Sigma^{-\frac{1}{2}}(\beta)(I_k - P(\beta))\Sigma^{-\frac{1}{2}}(\beta)$.

$$\begin{aligned}Q_n^0(\beta,\gamma) =& g_n^0(\beta,\gamma)'M(\beta)g_n^0(\beta,\gamma) + g_n^0(\beta,\gamma)'\Sigma^{-\frac{1}{2}}(\beta)P(\beta)\Sigma^{-\frac{1}{2}}(\beta)g_n^0(\beta,\gamma) = \\ =& Q_n^\beta(\beta) + \left(G^*(\beta) + \sqrt{n}\gamma\right)' J(\beta)\left(G^*(\beta) + \sqrt{n}\gamma\right),\end{aligned}$$

where $G^*(\beta) = J(\beta)^{-1}\nabla(\beta)'\Sigma^{-1}(\beta)G_n(\beta)$. Integration of the Gaussian pdf gives

$$\begin{aligned}&n^{\frac{p_\gamma}{2}} \int_{\|\gamma\|^2\leq\frac{\delta}{n}} \exp\left\{-\frac{1}{2}\left(G^*(\beta) + \sqrt{n}\gamma\right)' J(\beta)\left(G^*(\beta) + \sqrt{n}\gamma\right)\right\} d\gamma = \\ =& |J(\beta)|^{-\frac{1}{2}} P\left\{\left\|N\left(\frac{1}{\sqrt{n}}G^*(\beta), J^{-1}(\beta)\right)\right\| \leq \delta\right\} \to^p |J(\beta)|^{-\frac{1}{2}} P\left\{\|N\left(0, J^{-1}(\beta)\right)\| \leq \delta\right\},\end{aligned}$$

where δ was chosen large enough that $q(\delta) = P\{\|N(0, J^{-1}(\beta))\| \leq \delta\} \geq 1 - \varepsilon$. Thus,

$$\sup_\beta \left| n^{\frac{p_\gamma}{2}} \int_{\|\gamma\|^2\leq\frac{\delta}{n}} \exp\left\{-\frac{1}{2}Q_n^0(\beta,\gamma)\right\} d\gamma - |J(\beta)|^{-\frac{1}{2}} q(\delta) \exp\left\{-\frac{1}{2}Q_n^\beta(\beta)\right\}\right| \to^p 0.$$

Joining together last statement with equations (13) and (15) we get

$$\sup_\beta \left| n^{\frac{p_\gamma}{2}} \int_{\Gamma(\beta)} \exp\left\{-\frac{1}{2}Q_n(\beta,\gamma)\right\} d\gamma - |J(\beta)|^{-\frac{1}{2}} \exp\{-\frac{1}{2}Q_n^\beta(\beta)\}\right| \to^p 0.$$

Given statement (13), for $c(\beta,\gamma)$ satisfying assumptions of Theorem 5 we have

$$\sup_\beta \left| n^{\frac{p_\gamma}{2}} \int_{\Gamma(\beta)} c(\beta,\gamma) \exp\left\{-\frac{1}{2}Q_n(\beta,\gamma)\right\} d\gamma - c(\beta,0)|J(\beta)|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Q_n^\beta(\beta)\right\}\right| \to^p 0.$$

Assumption 2 implies $\int_B \pi_\gamma(0|\beta)|J(\beta)|^{-\frac{1}{2}} \exp\{-\frac{1}{2}Q_n^\beta(\beta)\}d\pi(\beta)$ is stochastically bounded aways from zero. Thus, (12) holds. $\square$

Proof of Corollary 2. For each $a \in \mathcal{A}$ we can apply (12) to $c(\theta) = L(a, \theta)$. Since $L(a, \theta)$ is Lipschitz in a and $\mathcal{A}$ is compact, this implies

$$\sup_{a \in \mathcal{A}} \left| \int_{\Theta} L(a, \theta) d\pi(\theta | g_n) - \frac{\int_B L(a, \vartheta(\beta, 0)) \exp\left\{-\frac{1}{2} Q_n^{\beta}(\beta)\right\} d\pi^0(\beta)}{\int_B \exp\left\{-\frac{1}{2} Q_n^{\beta}(\beta)\right\} d\pi^0(\beta)} \right| \to^p 0.$$

We also have weak convergence of the process

$$\frac{\int_B L(\cdot, \vartheta(\beta, 0)) \exp\left\{-\frac{1}{2} Q_n^{\beta}(\beta)\right\} d\pi^0(\beta)}{\int_B \exp\left\{-\frac{1}{2} Q_n^{\beta}(\beta)\right\} d\pi^0(\beta)} \Rightarrow L(\cdot)$$

on $\mathcal{A}$. This implies $\int_{\Theta} L(\cdot, \theta) d\pi(\theta | g_n) \Rightarrow L(\cdot)$. Due to Theorem 3.2.2 in Van der Vaart and Wellner (1996), $(s_n(g_n), s_n^0(g_n)) \Rightarrow (\text{argmin}_{a \in \mathcal{A}} L(a), \text{argmin}_{a \in \mathcal{A}} L(a))$. Thus, $s_n(g_n) - s_n^0(g_n) \to^p 0$. $\square$