The Annals of Statistics
2011, Vol. 39, No. 6, 3003–3031
DOI: 10.1214/11-AOS930

POSTERIOR CONSISTENCY OF NONPARAMETRIC CONDITIONAL MOMENT RESTRICTED MODELS

BY YUAN LIAO AND WENXIN JIANG

Princeton University and Northwestern University

This paper addresses the estimation of the nonparametric conditional moment restricted model that involves an infinite-dimensional parameter g_0. We estimate it in a *quasi-Bayesian* way, based on the limited information likelihood, and investigate the impact of three types of priors on the posterior consistency: (i) truncated prior (priors supported on a bounded set), (ii) thin-tail prior (a prior that has very thin tail outside a growing bounded set) and (iii) normal prior with nonshrinking variance. In addition, g_0 is allowed to be only partially identified in the frequentist sense, and the parameter space does not need to be compact. The posterior is regularized using a slowly growing sieve dimension, and it is shown that the posterior converges to any small neighborhood of the identified region. We then apply our results to the nonparametric instrumental regression model. Finally, the posterior consistency using a random sieve dimension parameter is studied.

1. Introduction.

We consider a conditional moment restricted model

$$E(\rho(Z, g_0)|W, g_0) = 0, \tag{1.1}$$

where (Z^T, W^T) is a vector of observable random variables, and W may or may not be included in Z. Here ρ is a one-dimensional residual function known up to g_0. The conditional expectation is taken with respect to the conditional distribution of Z given W and g_0, assumed unknown. The parameter of interest is g_0, which is infinite dimensional. Moreover, suppose we observe independent and identically distributed data $\{(Z_i^T, W_i^T)\}_{i=1}^n$ of (Z^T, W^T).

Model (1.1) is a very general setting, which encompasses many important classes of nonparametric and semiparametric models.

EXAMPLE 1.1 (Regular nonparametric regression). Consider the model

$$Y = g_0(W) + \varepsilon$$

assuming $E(\varepsilon|W) = 0$. Let $Z = (Y, W)$, then it can be written as the conditional moment restricted model with $\rho(Z, g_0) = Y - g_0(W)$.

Received March 2010; revised September 2011.
MSC2010 subject classifications. Primary 62F15, 62G08, 62G20; secondary 62P20.
Key words and phrases. Identified region, limited information likelihood, sieve approximation, nonparametric instrumental variable, ill-posed problem, partial identification, Bayesian inference, shrinkage prior, regularization.

EXAMPLE 1.2 (Single index model). Consider the single index model

$$Y = h_0(W^T\theta_0) + \varepsilon,$$

where $E(\varepsilon|W) = 0$. The parameter of interest is (h_0, θ_0), with h_0 being nonparametric. This type of model is studied by Ichimura (1993) and Antoniadis, Grégoire and McKeague (2004). By defining $Z = (Y, W)$, $g_0 = (h_0, \theta_0)$ and $\rho(Z, g_0) = Y - h_0(W^T\theta_0)$, we can write $E(\rho(Z, g_0)|W, g_0) = 0$.

EXAMPLE 1.3 (Nonparametric IV regression). Consider the nonparametric model

$$Y = g_0(X) + \varepsilon,$$

where X is an endogenous regressor, meaning that $E(\varepsilon|X)$ does not vanish. However, suppose we have observed an instrumental variable W for which $E(\varepsilon|W) = 0$; then it becomes a nonparametric regression model with instrumental variables (NPIV), studied by Newey and Powell (2003) and Hall and Horowitz (2005). Define $\rho(Z, g_0) = Y - g_0(X)$, with $Z = (Y, X)$. Then we have the conditional moment restriction.

EXAMPLE 1.4 (Nonparametric quantile IV regression). The nonparametric quantile IV regression was previously studied by Chernozhukov and Hansen (2005), Chernozhukov, Imbens and Newey (2007) and Horowitz and Lee (2007). The model is

$$y = g_0(X) + \varepsilon, \qquad P(\varepsilon \le 0|W) = \gamma,$$

where g_0 is the unknown function of interest, and $\gamma \in (0, 1)$ is known and fixed. Assume X is a continuous random variable. Then the conditional moment restriction is given by

$$E(\rho(Z, g_0)|W, g_0) = 0, \qquad \rho(Z, g_0) = I_{(y \le g_0(X))} - \gamma.$$

If we define $G(g) = E_W[E(\rho(Z, g)|W, g_0)]^2$, an equivalent way of writing model (1.1) is then $G(g_0) = 0$. When the unknown function g_0 depends on certain endogenous variable as in Examples 1.3 and 1.4, the identification and consistent estimation of g_0 is challenging. On one hand, there can be multiple functions in the parameter space that satisfy the moment restriction (1.1). On the other hand, even if g_0 is identified, [in which case the functional $G(g)$ is uniquely minimized at $g = g_0$, as is typically assumed in the literature], reducing $G(g)$ toward $G(g_0)$ does not guarantee that $\|g - g_0\|_s$ will also be close to zero, for a certain norm $\|\cdot\|_s$ of interest. Therefore, minimizing a consistent estimator of $G(g)$ does not lead to a consistent estimator of g_0 under $\|\cdot\|_s$. This phenomenon is usually known as the "ill-posed inverse problem" in the literature.

The general form of (1.1) was first studied by Ai and Chen (2003) and Newey and Powell (2003), where the authors considered sieve approximation of g_0 and estimated it in a compact parameter space. Recently, Chen and Pouzo (2009a) relaxed the compactness assumption and achieved the consistency and convergence rate using the penalized sieve minimum distance estimation. In recent years there has also been extensive literature on the NPIV model (Example 1.3) itself. In these papers, the authors introduce a Tikhonov tuning parameter to play a role of "regularization" in order to overcome the ill-posed inverse problem; see, for example, Hall and Horowitz (2005) and Darolles et al. (2011). Other related works on the nonparametric instrumental variables can be found in Chernozhukov, Gagliardini and Scaillet (2008), Johannes, Van Bellegem and Vanhems (2010), Horowitz (2007, 2011), among others.

Compared to the growing literature from the frequentist perspective, there is very little understanding of the consistent estimation using either a Bayesian or a quasi-Bayesian approach. This paper proposes a quasi-Bayesian procedure and studies the impact of various priors of g_0 on the posterior consistency. Our setup is built on a sieve approximation technique similar to Chen and Pouzo (2009a), which assumes that g_0 can be approximated arbitrarily well on a finite-dimensional sieve space. In order to keep our procedure robust to the distribution specification and convenient for practical implementation, without specifying a known distribution on the data generating process, we employ a limited information likelihood [Kim (2002) and Liao and Jiang (2010)], a moment-condition-based Gaussian approximated likelihood. The use of such a likelihood is more straightforward for models characterized by either moment conditions or estimating equations than the common methods based on Dirichlet process priors in the nonparametric Bayesian literature. With priors placed directly on the sieve coefficients, we show that the proposed posterior is consistent. Due to the difficulty of identifying g_0 in practice, we do not assume g_0 to be necessarily identified. As a result the posterior consistency here means that, asymptotically, the posterior converges into arbitrarily small neighborhood of the region where g_0 is partially identified. Therefore, we also extend model (1.1) to the partial identification setup [Chernozhukov, Hong and Tamer (2007) and Santos (2012)]. We will consider three types of priors: (i) priors supported on a bounded set (truncated prior), (ii) priors with tails decaying fast outside a bounded set (thin-tail prior) and (iii) Gaussian priors with nonshrinking variance.

Recently, Florens and Simoni (2009a) proposed a quasi-Bayesian approach for the NPIV model. They assumed that the error term follows a normal distribution and achieved consistency by regularizing an operator that defines the posterior mean. Our approach differs from theirs essentially in the way of overcoming the ill-posed inverse problem. While Florens and Simoni (2009a) put a Gaussian prior on an infinite-dimensional function space, they require the variance of the prior to shrink to zero. In contrast, we place the prior directly on the sieve coefficients in a finite-dimensional vector space and require the sieve dimension to grow slowly with the sample size. Our approach then corresponds to Chen and Pouzo's (2009a) sieve minimum distance procedure using slowly growing sieves. As a result, it is the finite-dimensional sieve that plays the role of regularization instead of a shrinking prior. In addition, our approach allows nonnormal priors.

Models based on moment conditions as (1.1) have been proved to be essential in many statistical applications, such as financial asset pricing [Gallant and Tauchen (1989), Chen and Ludvigson (2009)], consumer behavior in economics [Blundell, Chen and Kristensen (2007), Santos (2012)] and return to college education [Horowitz (2011)]. Therefore, this paper develops a quite convenient and straightforward quasi-Bayesian approach for these applied problems.

The remainder of this paper is organized as follows: Section 2 introduces general theorems on two types of posterior consistency, which provide sufficient conditions under which a posterior constructed on a sieve space is consistent. Section 3 specifies the priors and shows the consistency results by verifying the sufficient conditions given in Section 2. Section 4 studies in detail the NPIV model as a specific example. Section 5 discusses the case of the random sieve dimension. Finally, Section 6 concludes with further discussions. Proofs are given in the supplementary material.

Throughout the paper, for any two positive deterministic sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, write $a_n \succ b_n$ and $b_n \prec a_n$ if $b_n = o(a_n)$. In addition, $a_n \sim b_n$ if there exist c_1 and $c_2 > 0$ such that $c_1 b_n \leq a_n \leq c_2 b_n$ for all large enough n.

2. General posterior consistency theorems.

2.1. *Sieve approximation*. Suppose we are interested in a nonparametric regression function $g_0 \in (\mathcal{H}, \|\cdot\|_s)$. which is assumed to be inside an infinite-dimensional Banach space $\mathcal{H}$ endowed with norm $\|\cdot\|_s$. Examples of the space $(\mathcal{H}, \|\cdot\|_s)$ include: space of bounded continuous functions with norm $\|g\|_s = \sup_x |g(x)|$, the space of square integrable functions $\{g : E[g(X)^2] < \infty\}$ with $\|g\|_s = \sqrt{E[g(X)^2]}$, etc. In addition, suppose there exists a set of basis functions $\{\phi_1, \phi_2, \ldots\} \subset \mathcal{H}$ such that $g_0 \in \mathcal{H}$ can be approximated by a truncated sum $g_b = \sum_{i=1}^{q_n} b_i \phi_i$ for a vector of coefficients $(b_1, \ldots, b_{q_n})^T$, where q_n is a predetermined constant that grows to infinity. Then g_b lies in an approximating space $\mathcal{H}_n$ spanned by $\{\phi_1, \ldots, \phi_{q_n}\}$. Here $\mathcal{H}_n$ grows to be dense in $\mathcal{H}$, called a sieve approximating space.

There is extensive literature on the posterior consistency using sieve approximation. Shen and Wasserman (2001) applied an orthogonal basis expansion to the nonparametric regression problem. Walker (2003) and Choi and Schervish (2007) provided general results for a class of Bayesian regression models when the data have a normal distribution. Other results on nonparametric regression problems can be found, for example, in Huang (2004), Ghosal and van der Vaart (2007), etc.

Suppose we are given n independent identically distributed observations $X^n = (X_1, X_2, \ldots, X_n)$. In this paper we do not assume any specific distribution of

$X^n|g_0$, but propose a quasi-Bayesian approach, which is based on a pseudo-likelihood,

$$L(g_b) = \exp\left(-\frac{n}{2}\bar{G}(g_b)\right),$$

where $\bar{G}: \mathcal{H}_n \to [0, \infty)$ is a stochastic functional, which we call the *sample risk functional*. Suppose there exists a nonnegative functional G, such that for a bounded set $\mathcal{F}_n \subset \mathcal{H}_n$,

$$\sup_{g_b \in \mathcal{F}_n} |\bar{G}(g_b) - G(g_b)| = o_p(1).$$

We call G the *objective functional* or *risk functional* throughout the paper.

In the literature, it is often assumed that the true regression function g_0 is point identified (as opposed to "partially identified" in the following) as the unique minimizer of G on $\mathcal{H}$, that is,

$$\{g_0\} = \arg\min_{g \in \mathcal{H}} G(g).$$

Then quasi-Bayesian approaches usually construct $\bar{G}$ as the sample analog of G [see Chernozhukov and Hong (2003)]. In many applications of the model considered in this paper, however, it is more natural to assume that G has multiple global minimizers on $\mathcal{H}$; see detailed discussions in Section 3. In this case, we say g_0 is *partially identified* (in the frequentist sense) on

$$\Theta_I = \arg\min_{g \in \mathcal{H}} G(g),$$

and Θ_I is called the *identified region*. Therefore Θ_I is the main object of interest in this paper.

For any $b = (b_1, \dots, b_{q_n})^T \in \mathbb{R}^{q_n}$, let $g_b = \sum_{i=1}^{q_n} b_i \phi_i$. Similarly to the standard treatments in Smith and Kohn (1996) and Antoniadis, Grégoire and McKeague (2004), we put prior $\pi(b)$ on the sieve coefficients $b = (b_1, b_2, \dots, b_{q_n})$, and obtain a posterior distribution,

$$P(g_b|X^n) \propto \pi(b) L(g_b).$$

For any $g_1 \in \mathcal{H}$, define

$$d(g_1, \Theta_I) = \inf_{g \in \Theta_I} \|g_1 - g\|_s,$$

and the ε*-expansion* as a neighborhood of the identified region

$$\Theta_I^{\varepsilon} = \{g \in \mathcal{H} : d(g, \Theta_I) < \varepsilon\}.$$

Then the posterior consistency in this paper refers to the following: for any $\varepsilon > 0$,

$$P(g \in \Theta_I^{\varepsilon}|X^n) \to^p 1.$$

2.2. *Posterior consistency theorems*. We first present two theorems of general posterior consistency using the sieve approximation, which involve conditions on the tail probability of π as well as the performance of $\bar{G}$. They are based on the following variant of an inequality from Jiang and Tanner (2008), Proposition 6. These inequalities will be proved in the supplementary material [Liao and Jiang (2011a)]:

LEMMA 2.1. *Suppose the support of the prior π can be partitioned as $\mathcal{F}_n \cup \mathcal{F}_n^c$. Then for any deterministic sequence $\delta_n > 0$,*

$$\begin{aligned}
&E\Big\{P\Big(G(g_b) - \inf_{g\in\mathcal{H}} G(g) > 5\delta_n | X^n\Big)\Big\} \\
&\quad \le P\Big(\sup_{g\in\mathcal{F}_n} |\bar{G}(g) - G(g)| \ge \delta_n\Big) \\
&\qquad + \frac{e^{-2n\delta_n}}{\pi(G(g_b) - \inf_{g\in\mathcal{H}} G(g) < \delta_n \cap g_b \in \mathcal{F}_n)} \\
&\qquad + EP(g_b \in \mathcal{F}_n^c | X^n).
\end{aligned} \tag{2.1}$$

In addition,

$$\begin{aligned}
EP(g_b \in \mathcal{F}_n^c | X^n) &\le P\Big(\sup_{g\in\mathcal{F}_n} |\bar{G}(g) - G(g)| \ge \delta_n\Big) \\
&\qquad + \frac{\pi(\mathcal{F}_n^c) e^{2n\delta_n}}{\pi(G(g_b) - \inf_{g\in\mathcal{H}} G(g) < \delta_n \cap g_b \in \mathcal{F}_n)}.
\end{aligned}$$

These inequalities imply the following result on the risk consistency:

THEOREM 2.1 (Risk consistency). *Suppose the following conditions hold with respect to a deterministic positive sequence δ_n*:

(i) *Tail condition*: *as q_n and $n \to \infty$, either $EP(g_b \in \mathcal{F}_n^c | X^n) = o(1)$ or $\pi(\mathcal{F}_n^c) = O(e^{-4n\delta_n})$.*

(ii) *Approximation condition*: $\pi(G(g_b) - \inf_{g\in\mathcal{H}} G(g) < \delta_n, g_b \in \mathcal{F}_n) \succ e^{-2n\delta_n}$.

(iii) *Uniform convergence*: $P[\sup_{g\in\mathcal{F}_n} |\bar{G}(g) - G(g)| \ge \delta_n] = o(1)$.

Then we have the risk consistency result at rate δ_n

$$P\Big(G(g_b) - \inf_{g\in\mathcal{H}} G(g) < \delta_n | X^n\Big) = 1 - o_p(1).$$

The naming of these conditions is obvious, except for (ii). There, the approximation refers to the ability of the functions in $\mathcal{F}_n$ (proposed by the prior π) to approximately minimize the risk G over $\mathcal{H}$ with not-too-small prior probability.

When the following condition is added, the risk consistency leads to the estimation consistency.

THEOREM 2.2 (Estimation consistency). *Suppose there exists a sequence* δ_n *such that the following conditions hold*:

(i), (ii), (iii) *in the previous theorem*;
(iv) (*distinguishing ability*) *for any* $\varepsilon > 0$,

$$\inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g) - \inf_{g\in\mathcal{H}} G(g) \succ \delta_n.$$

Then for any $\varepsilon > 0$, *we have*

$$P(g_b \in \Theta_I^\varepsilon | X^n) \to^p 1. \tag{2.2}$$

PROOF. Theorem 2.1 is implied by Lemma 2.1. Now we prove Theorem 2.2. For any $\varepsilon > 0$, by Theorem 2.1,

$$\begin{aligned}
&P(g_b \notin \Theta_I^\varepsilon | X^n) \\
&\quad \le P\Big(g_b \notin \Theta_I^\varepsilon, G(g_b) - \inf_{g\in\mathcal{H}} G(g) < \delta_n | X^n\Big) + o_p(1) \\
&\quad \le P\Big(g_b \notin \Theta_I^\varepsilon, G(g_b) \ge \inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g), G(g_b) - \inf_{g\in\mathcal{H}} G(g) < \delta_n | X^n\Big) \\
&\qquad + o_p(1) \\
&\quad \le P\Big(g_b \notin \Theta_I^\varepsilon, \delta_n < G(g_b) - \inf_{g\in\mathcal{H}} G(g) < \delta_n | X^n\Big) + o_p(1) \\
&\quad = o_p(1),
\end{aligned}$$

where the third inequality is implied by condition (iv) for all large n. □

As a special case of these results, note that when g_0 is point identified as the unique minimizer of $G(g)$ on $\mathcal{H}$, that is, $\Theta_I = \{g_0\}$, (2.2) then becomes

$$P(\|g_b - g_0\|_s < \varepsilon | X^n) \to^p 1,$$

the regular posterior consistency result.

In the subsequent sections, we will construct a so-called *limited information likelihood* $\bar{G}(g)$ and apply the previous two theorems to the conditional moment restricted model (1.1), by verifying conditions (i)–(iv).

3. Conditional moment-restricted model.

3.1. *Limited information likelihood.* Consider a conditional moment condition

$$E[\rho(Z, g_0)|W, g_0] = 0, \tag{3.1}$$

where $g_0 \in \mathcal{H}$ is the true nonparametric structural function. Here W is d-dimensional, with fixed d. For simplicity, throughout the paper, let us assume

W is supported on $[0,1]^d$, as one can always apply the transformation on each component of W, $W_i \to \Phi(W_i)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. We focus on the case when ρ is a one-dimensional function.

Following the setting of Ai and Chen (2003) and Chen and Pouzo (2009a), we approximate $\mathcal{H}$ by a sieve space $\mathcal{H}_n$ that grows to be dense in $\mathcal{H}$. Here $\mathcal{H}_n$ is a finite-dimensional space spanned by sieve basis functions$\{\phi_1, \dots, \phi_{q_n}\}$ such as splines, power series, wavelets and Fourier series.

As the first step, we transform the conditional moment restriction into unconditional moment restrictions (but still conditional on g_0). Let $\{[(i-1)/k_n, i/k_n]\}_{i=1}^{k_n}$ be a partition of $[0,1]$, for some $k_n \in \mathbb{N}$. We then obtain a partition of the support of $W: [0,1]^d = \bigcup_{j=1}^{k_n^d} R_j^n$, where for each $j = 1, \dots, k_n^d$,

$$R_j^n = \prod_{l=1}^{d} \left[\frac{i_l - 1}{k_n}, \frac{i_l}{k_n}\right] \qquad \text{for some } i_l \in \{1, \dots, k_n\}. \tag{3.2}$$

We require $k_n \to \infty$ as $n \to \infty$. Let $X = (Z, W)$. For each j, define

$$m_{nj}(g, X) = \rho(Z, g) I_{(W \in R_j^n)},$$

where $I_{(\cdot)}$ is the indicator function. Let $m_n(g, X) = (m_{n1}(g, X), \dots, m_{nk_n^d}(g, X))^T$, which is a $k_n^d \times 1$ vector. Equation (3.1) then implies

$$E m_n(g_0, X) = 0, \tag{3.3}$$

where the expectation is taken with respect to the joint distribution of $X = (Z, W)$ conditional on g_0. Throughout the paper, the expectation is always taken conditionally on g_0. When $k_n > q_n$ there are more moment conditions than the parameters, and hence (3.3) is a problem of many moment conditions with increasing number of moments studied by Han and Phillips (2006).

It is straightforward to verify that

$$V_0 \equiv \operatorname{Var}(m_n(g_0, X)) = \operatorname{diag}\{E(\rho(Z, g_0)^2 I_{(W \in R_1^n)}), \dots, E(\rho(Z, g_0)^2 I_{(W \in R_{k_n^d}^n)})\}.$$

For each $g \in \mathcal{H}$, and $j = 1, \dots, k_n^d$, write $\bar{m}_{nj}(g) = \frac{1}{n}\sum_{i=1}^n m_{nj}(g, X_i)$ and $\bar{m}_n(g) = (\bar{m}_{n1}(g), \dots, \bar{m}_{nk_n^d}(g))^T$. Instead of g_0, we construct the posterior for its approximating function inside $\mathcal{H}_n$. Under some regularity conditions, for each fixed k, $\bar{m}_n(g_0)$ would satisfy the central limit theorem: for any $\alpha \in \mathbb{R}^k$, as n goes to infinity,

$$\left| P(\sqrt{n} V_0^{-1/2} \bar{m}_n(g_0) \leq \alpha) - \prod_{i=1}^{k} \Phi(\alpha_i) \right| \to 0. \tag{3.4}$$

This motivates a likelihood function on the sieve space $\mathcal{H}_n$,

$$\operatorname{LIL}(g_b) \propto \exp\left(-\frac{n}{2} \bar{m}_n(g_b)^T V_0^{-1} \bar{m}_n(g_b)\right).$$

According to Kim (2002), the function $\mathrm{LIL}(g_b)$ can be more appropriately interpreted as the best approximation to the true likelihood function under the conditional moment restriction by minimizing the Kullback–Leibler divergence, which is known as the *limited information likelihood* (LIL). Note that $\mathrm{LIL}(g_b)$ is not feasible, as V_0 depends on the unknown function g_0; therefore Kim (2002) suggested replacing V_0 with a constant matrix (not dependent on g_0), while maintaining the order of each element. For each element on the diagonal, suppose we have the integration mean value theorem: for some $w^* \in R_j^n$,

$$E\big(\rho(Z, g_0)^2 I_{(W \in R_j^n)}\big) = E\big(\rho(Z, g_0)^2 | W = w^*\big) P(W \in R_j^n) = O\big(P(W \in R_j^n)\big)$$

provided that $\sup_{w \in [0,1]^d} E[\rho(Z, g_0)^2 | w] < \infty$. Hence each diagonal element of V_0 is of the same order as $P(W \in R_j^n)$. We replace V_0 by

$$\hat{V} = \mathrm{diag}\{\hat{v}_1, \dots, \hat{v}_{k_n^d}\} \qquad \text{where } \hat{v}_j = \frac{1}{n} \sum_{i=1}^{n} I_{(W_i \in R_j^n)}.$$

Each $\hat{v}_j$ is a consistent estimate of $P(W \in R_j^n)$. We thus obtain the feasible LIL to be used as the likelihood function throughout this paper,

$$L(g_b) = \exp\biggl(-\frac{n}{2} \bar{m}_n(g_b)^T \hat{V}^{-1} \bar{m}_n(g_b)\biggr). \tag{3.5}$$

The feasible likelihood puts more weights on the moment conditions with smaller variance, having the same spirit of the optimal weight matrix in *generalized method of moments* [Hansen (1982)]. A more refined approach can be based on a second-stage estimation of V_0, where a consistent first-stage estimator of g_0 is used if g_0 is assumed to be point identified. However, it turns out that V_0 does not have to be estimated very precisely in order to achieve the posterior consistency for the inference on g. We will show that our simple estimator $\hat{V}$ is already good enough for proving posterior consistency in the development to be described below and is simple for practical computations.

For the approximated Gaussian likelihood function (3.5), the sample risk functional defined in Section 2 is given by

$$\bar{G}(g_b) \equiv \bar{m}_n(g_b)^T \hat{V}^{-1} \bar{m}_n(g_b). \tag{3.6}$$

Let

$$\mathcal{F}_n = \left\{ \sum_{i=1}^{q_n} b_i \phi_i(x) : \max_{i \le q_n} |b_i| \le B_n \right\}$$

for some sequence $B_n \to \infty$; then we partition the sieve space into $\mathcal{H}_n = \mathcal{F}_n \cup \mathcal{F}_n^c$. Under some regularity conditions, it can be shown that[1] $\bar{G}$ converges in probability

[1] We will verify this for the nonparametric IV regression model in Section 4.

to the risk functional

$$G(g) = E_W\{[E(\rho(Z,g)|W)]^2\} = \int_{[0,1]^d} [E(\rho(Z,g)|W=w)]^2\, dF_W(w) \tag{3.7}$$

uniformly on $\mathcal{F}_n$.

3.2. *Identification and ill-posedness.* The identification of g_0 is characterized by minimizing G. To be specific, define the identified region for g_0,

$$\Theta_I = \{g \in \mathcal{H} : E(\rho(Z,g)|W=w) = 0 \text{ for almost all } w \in [0,1]^d\},$$

which is assumed to be nonempty, then

$$\Theta_I = \arg\min_{g \in \mathcal{H}} G(g) = \{g \in \mathcal{H} : G(g) = 0\}.$$

If Θ_I is a singleton, then $\Theta_I = \{g_0\}$. Otherwise g_0 is partially identified on Θ_I; see, for example, Santos (2012).

In the conditional moment restriction literature, the problem of identification and estimation of g_0 is well known to be *ill posed*. The ill-posed problem was postulated in detail by Kress [(1999), Chapter 15], which occurs, in our context, if one of the following three properties does not hold: (1) there exist solutions to $G(g) = 0$, and here we assume $g_0 \in \Theta_I$; (2) the solution is unique, that is, Θ_I is a singleton; (3) the solution is continuously dependent on the data; that is, roughly speaking, when $G(g)$ is close to zero, g should be close to Θ_I. However, when g_0 depends on the endogenous variable X, the third property may fail because for any $\varepsilon > 0$, there are sequences $\{g_n\}_{n=1}^{\infty} \subset \mathcal{H}$ such that

$$\liminf_{n\to\infty} \inf_{g_n \notin \Theta_I^{\varepsilon}} G(g_n) = 0.$$

Throughout this paper, we call such a problem as the *type-III ill-posed inverse problem*. In order to achieve the posterior consistency, we need certain regularization scheme to make the metric $d(g, \Theta_I)$ be continuous with respect to the risk functional $G(g)$.

While the literature puts a primary interest on dealing with the type-III ill-posedness [Hall and Horowitz (2005), etc.], there are relatively fewer results that deal with the second type of ill-posedness: Θ_I is not necessarily a singleton. In this paper, we also allow g_0 to be only partially identified[2] by the conditional moment restriction (3.1). Such a treatment arises for two reasons. First, when the conditional moment restriction is given by the nonparametric instrumental variable regression (Example 1.3), the identification of g_0 depends on the completeness of the conditional distribution of $X|W$ [Newey and Powell (2003)]; however, the completeness assumption is hard to verify if the conditional distribution

[2] In this paper, the *partial identification* is meant in the frequentist sense, as opposed to the Bayesian identification. See a recent work by Florens and Simoni (2011) for a discussion of these concepts.

of $X|W$ does not belong to the exponential family. Severini and Tripathi (2006) explored identification issues with these models and noted that the point identification can easily fail; see Example 3.2 of Severini and Tripathi (2006). For another reason, sometimes instead of g_0 itself, we are only interested in a particular characteristic of it, for example, its linear functional $h(g_0)$. For example, in the nonparametric IV regression, if $g_0(x)$ represents the inverse demand function, then its consumer surplus at some level x^* can be written as a functional $h(g_0) = \int_0^{x^*} g_0(x)\,dx - g_0(x^*)x^*$. In this case, the identification of g_0 might not be necessary; as Severini and Tripathi (2006) showed, even if g_0 is not identified, it is still possible to point identify its functional $h(g_0)$.

3.3. *Prior specification.* We will apply Theorems 2.1 and 2.2 to three types of priors: (i) truncated prior, (ii) thin-tail prior and (iii) normal prior. In this section we will focus on the first two types of priors, with which more generally consistent results can be derived.[3]

Truncated prior. The prior is supported only on $\mathcal{F}_n$. In particular, we consider the uniform and truncated normal priors, respectively,

$$\text{uniform prior } \pi(b) = \prod_{i=1}^{q_n} I(|b_i| \le B_n);$$

$$\text{truncated normal } \pi(b) = \prod_{i=1}^{q_n} \frac{f(b_i)I(|b_i| \le B_n)}{P(|Z_i| \le B_n)},$$

where $\{Z_i\}_{i=1}^{q_n}$ are i.i.d. random variables from $N(0, \sigma^2)$ for some $\sigma^2 > 0$, and $f(\cdot)$ is the probability density function of Z_i. The tail probability

$$\pi(g_b \in \mathcal{F}_n^c) = 0.$$

Thin-tail prior. The prior π on $b \in \mathbb{R}^{q_n}$ is defined such that the density is symmetric in all directions, and $\|b\|^r$ follows an exponential distribution with mean β^{-r} (for some $\beta > 0$, $r > 0$). Here $\|b\|$ denotes a Euclidean norm,

$$\pi(\|b\|^r > u^r) = e^{-\beta^r u^r},$$

which, together with the spherical symmetry, is enough to derive the density function,

$$\pi(b) = \frac{r\|b\|^{r-q_n}\beta^r e^{-\beta^r\|b\|^r}}{S_{q_n}}, \tag{3.8}$$

[3] We will describe the normal prior in a later section (Section 4.4) since the technique used is somewhat different, which handles mainly the situation of the NPIV model in an identifiable situation.

where S_{q_n} is the area of the $q_n - 1$-dimensional unit sphere in Euclidean norm. For this prior, the parameter $1/\beta$ is roughly the radius of most of the prior mass, and r denotes the thinness of the tails outside. The bigger the r is, the thinner the tail.

This prior is very similar to the class of distributions defined in Azzalini (1986). Both allow any positive power of the distance to the origin to be placed on the exponent. Our density is slightly different and does not, in general, include the normal density exactly. However, it is derived in a way so that the tail probability has an exact expression. Hence it is convenient to impose a regularity condition on the tail probability.

Florens and Simoni (2009a, 2009b) placed a Gaussian prior whose variance decreases to zero with the sample size. Our priors specified here are similar to theirs in the sense that the prior tail probability is small: when the truncated prior is used, $\pi(g_b \in \mathcal{F}_n^c) = 0$; when the thin-tail prior is used, $\pi(g_b \in \mathcal{F}_n^c)$ decreases exponentially fast in n. Both types of priors ensure that

$$P\big(G(g_b) \geq \delta_n | X^n\big) = o_p(1)$$

for some decaying sequence $\delta_n > 0$ that depends on the convergence rate of $\sup_{\mathcal{F}_n} |\bar{G}(g) - G(g)|$. The technique of using a prior that decays exponentially fast outside a bounded sieve set is commonly used in the nonparametric posterior consistency literature; see, for example, Ghosh and Ramamoorthi (2003), Ghosal and Roy (2006), Choi and Schervish (2007), Walker (2003) and many references therein.

However, there is an important difference between Florens and Simoni's prior settings (2009a) and our own. While Florens and Simoni (2009a) put their prior on an infinite-dimensional function space, they require the variance of the Gaussian prior to shrink to zero as a regularization scheme in order to achieve the posterior consistency. In contrast, our prior is placed directly on the sieve coefficients $(b_1, \dots, b_{q_n})$ in a finite-dimensional vector space, and neither the truncated prior nor the thin-tail prior shrinks to a point mass. When q_n grows slowly with n, it can be shown that[4] for any $\varepsilon > 0$,

$$\inf_{g_b \in \mathcal{H}_n, d(g_b, \Theta_I) \geq \varepsilon} G(g_b) \succ \delta_n;$$

hence the distinguishing ability condition in Theorem 2.2 is satisfied. As a result, in our procedure it is the fact that q_n grows slowly that plays the role of regularization instead of a shrinking prior. Later in Section 4.4, we will also verify that with a suitably chosen q_n, a nonshrinking normal prior can be used to achieve the posterior consistency in the identified NPIV model.

[4]We will verify this for the nonparametric IV regression model.

3.4. *Posterior consistency.* The following assumptions are imposed.

ASSUMPTION 3.1. The data $X^n = (X_1, \dots, X_n)$ are independent and identically distributed.

ASSUMPTION 3.2. There exists a positive sequence $\lambda_n \to 0$ such that

$$\sup_{g \in \mathcal{F}_n} |\bar{G}(g) - G(g)| = O_p(\lambda_n).$$

Since $\mathcal{F}_n$ is compact in $\mathcal{H}_n$, as long as the radius of $\mathcal{F}_n$ grows slowly, the uniform convergence condition in Assumption 3.2 can be shown using similar techniques to those in Han and Phillips (2006). We will verify it for the nonparametric IV regression example in Section 4.

ASSUMPTION 3.3. (i) $\{\phi_1, \phi_2, \dots, \phi_{q_n}\}$ forms an orthonormal basis of $\mathcal{H}_n$ such that $E(\phi_i(X)\phi_j(X)) = \delta_{ij}$, the Kronecker δ.

(ii) There exist $g_0 \in \Theta_I$, and $g^*_{q_n} = \sum_{i=1}^{q_n} b^*_i \phi_i \in \mathcal{H}_n$ such that $\|g^*_{q_n} - g_0\|_s = o(1)$ as $q_n \to \infty$.

The existence of $g^*_{q_n}$ is simply implied by the definition of a sieve space. It is satisfied by the spaces that are spanned by commonly used sieve basis functions such as splines, power series, wavelets and Fourier series. For example, if the parameter space is a Sobolev space $\mathcal{W}^2_p[0,1]^{d_x}$, where $d_x = \dim(X)$, and $\|\cdot\|_s$ is the Sobolev norm, then $\|g^*_{q_n} - g_0\|_s = O(q_n^{-p/d_x})$ for some $p > 0$; see, for example, Kress [(1999), Chapter 8] and Chen (2007); see also Schumaker (1981) and Meyer (1990) for splines and orthogonal wavelets in other function spaces.

ASSUMPTION 3.4. There exists $C > 0$ such that $\forall g_1, g_2 \in \mathcal{H}$,

$$E|\rho(Z, g_1) - \rho(Z, g_2)| \le CE|g_1(X) - g_2(X)|.$$

This assumption is trivially satisfied by the nonparametric IV regression in Example 1.3. Here we give another example that satisfies this assumption.

EXAMPLE 3.1 (Nonparametric quantile IV regression). Consider the model in Example 1.4, in which the conditional moment restriction is given by

$$E(\rho(Z, g_0)|W, g_0) = 0, \qquad \rho(Z, g_0) = I_{(y \le g_0(X))} - \gamma.$$

It is straightforward to verify that for any g_1, g_2,

$$\begin{aligned} E|\rho(Z, g_1) - \rho(Z, g_2)| &= E\big|I_{(g_1(X) \le y \le g_2(X))} + I_{(g_2(X) \le y \le g_1(X))}\big| \\ &= E\big[P\big(g_1(X) \le y \le g_2(X)|X\big)\big] \\ &\quad + E\big[P\big(g_2(X) \le y \le g_1(X)|X\big)\big]. \end{aligned}$$

Suppose there exists a constant $C > 0$ such that $F_{y|X}(\cdot)$, the conditional c.d.f. of $y|X$, satisfies

$$|F_{y|x}(y_1) - F_{y|x}(y_2)| \le C|y_1 - y_2|$$

for any $y_1, y_2 \in \mathbb{R}$ and x in the support of X. Then the first term on the right-hand side is bounded by

$$\begin{aligned} E\big[P\big(g_1(X) \le y \le g_2(X)|X\big)\big] &\le E|F_{y|X}(g_2(X)) - F_{y|X}(g_1(X))| \\ &\le CE|g_2(X) - g_1(X)|. \end{aligned}$$

Likewise, $E[P(g_2(X) \le y \le g_1(X)|X)] \le CE|g_2(X) - g_1(X)|$. Therefore Assumption 3.4 is satisfied.

Define

$$\gamma_n = \sup_{g \in \mathcal{F}_n, w \in [0,1]^d} \big|E\big(\rho(Z, g)|W = w\big)\big| + 1. \tag{3.9}$$

We are able to verify the conditions in Theorem 2.1 with the previous assumptions, and establish the following theorem:

THEOREM 3.1 (Risk consistency: truncated prior). *Suppose $q_n = o(n)$ and $B_n = o(n)$. Assume $\delta_n = O(1)$ is such that there exists $g_0 \in \Theta_I$ whose sieve approximation $g^*_{q_n}$ satisfies*

$$\max\left\{G(g^*_{q_n}), \lambda_n, \frac{q_n}{n}\log(\gamma_n n)\right\} = o(\delta_n).$$

Then when either the uniform prior or the truncated normal prior is used, under Assumptions 3.1–3.4,

$$P\big(G(g_b) < \delta_n|X^n\big) \to^p 1.$$

In the following theorem, write $\lambda(B_n) = \lambda_n$ and $\gamma(B_n) = \gamma_n$ to indicate the dependence of λ_n and γ_n on B_n, defined in Assumption 3.2 and (3.9), respectively.

THEOREM 3.2 (Risk consistency: thin-tail prior). *Suppose there exists $g_0 \in \Theta_I$ with $g^*_{q_n}$ being its sieve approximation in $\mathcal{H}_n$, and a sequence $B^*_n \to \infty$ such that $\max\{G(g^*_{q_n}), \lambda(B^*_n), \gamma(B^*_n)e^{-n\lambda(B^*_n)/q_n}\} = o(B^{*r}_n/n)$. In addition, suppose $\delta_n = O(1)$ is such that*

$$\max\big\{G(g^*_{q_n}), \lambda(B^*_n), \gamma(B^*_n)e^{-n\lambda(B^*_n)/q_n}\big\} = o(\delta_n).$$

Then under Assumptions 3.1–3.4,

$$P\big(G(g_b) < \delta_n|X^n\big) \to^p 1.$$

REMARK 3.1. (1) We will show in the next section that in the nonparametric IV regression model, $\gamma_n = O(q_n B_n)$. For the nonparametric quantile IV regression in Example 3.1, γ_n is a constant that is bounded away from zero.

(2) Under the conditions of Theorems 3.1 and 3.2, δ_n can be fixed as a constant. Namely, $\forall\delta > 0$,

$$P\big(G(g_b) > \delta | X^n\big) = o_p(1).$$

Roughly speaking, the posterior distribution is asymptotically supported on the set where G is minimized. This result has many important applications. For example, in the binary treatment effect study, let $Y \in \{0, 1\}$ indicate whether a treatment is successful, which is associated with a covariate X. Suppose we model the success probability $P(Y = 1|X = x)$ by a nonparametric function $g(x)$. In this model,

$$G(g) = E_X\{[E(Y|X) - g(X)]^2\} = \|P(Y = 1|X) - g(X)\|_s^2,$$

where $\|g\|_s^2 = E(g(X)^2)$. By Theorems 3.1, 3.2, for any $\varepsilon > 0$, the posterior

$$P\big(\|P(Y = 1|X) - g_b(X)\|_s^2 < \varepsilon|\text{Data}\big) \to^p 1,$$

which implies that the posterior of g_b can recover the success probability arbitrarily well with high probability.

(3) In data mining, this type of result is sometimes called the "risk consistency." For example, if G was the classification risk, the risk consistency result would show that the posterior would effectively minimize the misclassification error. The current definition of G, however, is not the classification risk. In nonparametric regression and in the NPIV example, the risk G becomes, respectively, $E_W\{[E(Y|W) - g(W)]^2\}$ and $E_W\{[E(Y|W) - E(g(X)|W)]^2\}$, which is related to how much $E(Y|W)$ would be missed if it was estimated by (something derived from) g.

The following two theorems establish the posterior consistency without assuming the compactness of the parameter space $\mathcal{H}$.

THEOREM 3.3 (Posterior consistency: truncated prior). *Suppose there exists $g_0 \in \Theta_I$ whose sieve approximation $g^*_{q_n}$ satisfies* $\forall\varepsilon > 0$

$$\max\left\{G(g^*_{q_n}), \lambda_n, \frac{q_n}{n}\log(\gamma_n n)\right\} = o\Big(\inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g)\Big). \tag{3.10}$$

Then under Assumptions 3.1–3.4, for any $\varepsilon > 0$,

$$P\big(d(g_b, \Theta_I) < \varepsilon|X^n\big) \to^p 1.$$

THEOREM 3.4 (Posterior consistency: thin-tail prior). *Suppose there exists $g_0 \in \Theta_I$ with $g^*_{q_n}$ being its sieve approximation in $\mathcal{H}_n$, and a sequence $B^*_n \to \infty$*

such that $\max\{G(g^*_{q_n}), \lambda(B^*_n), \gamma(B^*_n)e^{-n\lambda(B^*_n)/q_n}\} = o(B^{*r}_n/n)$. *In addition, suppose* $\forall\varepsilon > 0$,

$$\max\{G(g^*_{q_n}), \lambda(B^*_n), \gamma(B^*_n)e^{-n\lambda(B^*_n)/q_n}\} = o\Big(\inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g)\Big). \tag{3.11}$$

Then under Assumptions 3.1–3.4, *for any* $\varepsilon > 0$,

$$P\big(d(g_b, \Theta_I) < \varepsilon|X^n\big) \to^p 1.$$

REMARK 3.2. (1) The restriction $\lambda(B^*_n) = o(B^{*r}_n/n)$ in both Theorems 3.2 and 3.4 requires that r, the thin-tail prior parameter, should not be too small; otherwise, no such B^*_n exists. In the NPIV model which will be illustrated in the next section, we need $r > 6d + 4$, where $d = \dim(W)$.

(2) Conditions (3.10) and (3.11) are similar to Chen and Pouzo's [(2009a), condition (3.1)], where they require that q_n grow slowly enough so that $\inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g)$ does not decrease too fast for any fixed $\varepsilon > 0$. This will also be illustrated in Section 4.

Let $h(g_0)$ be a linear functional of g_0, whose practical meaning may be of direct interest. For example, if $h(g_0) = E[g_0(X)\omega(X)]$ for some weight function ω, then with proper choices of ω, h can be used to test some special properties of g_0, such as the monotonicity, the convexity, etc. Santos (2011). On the other hand, h itself may have interesting meanings. For example, when g_0 denotes the inverse demand function in nonparametric regression, $h(g_0)$ can be the consumer surplus [Santos (2012)]. Severini and Tripathi (2006) have provided conditions to point identify $h(g_0)$ even if g_0 itself is not identified.

EXAMPLE 3.2. Suppose we want to test whether the unknown function g_0 is weakly increasing. Note that any weakly increasing function $g(x)$ must satisfy $\int_{-\pi}^{\pi} \sin(x)g(x)\,dx \ge 0$; hence the functional of interest here is $h(g_0) = \int_{-\pi}^{\pi} \sin(x)g_0(x)\,dx$. Suppose the joint distribution of (X, W) has density function $f_{XW}(x, w)$. By Severini and Tripathi (2006), $h(g_0)$ is point identified, if there exists $p(w)$ such that $E[p(W)^2] < \infty$ and $E(p(W)|X) = \sin(X)/f_X(X)$ almost surely.

Theorems 3.3 and 3.4 imply a flexible way to consistently estimate h without identifying g_0. In the following assumption, condition (i) assumes the point identification of $h(g_0)$. Condition (ii) requires the uniform continuity of h, which is satisfied when $h(g) = E[g(X)\omega(X)]$ if $\sup_x |w(x)| < \infty$ and $E|g_1 - g_2| \le C\|g_1(X) - g_2(X)\|_s$ for any $g_1, g_2 \in \mathcal{H}$.

ASSUMPTION 3.5. (i) $\{h(g) : g \in \Theta_I\} = \{h(g_0)\}$; (ii) $h : (\mathcal{H}, \|\cdot\|_s) \to \mathbb{R}$ is uniformly continuous.

COROLLARY 3.1. *Suppose the assumptions of Theorem 3.3 (if the truncated priors are used) and Theorem 3.4 (if the thin-tail prior is used) are satisfied. In addition, suppose Assumption 3.5 holds. When g_0 is not necessarily point identified,* $\forall\delta > 0$,

$$P\big(|h(g_b) - h(g_0)| < \delta|X^n\big) \to^p 1.$$

4. Nonparametric instrumental variable regression.

4.1. *The model.* The nonparametric instrumental variable regression (NPIV) model is given by

$$Y = g_0(X) + \varepsilon,$$

where X is endogenous, which is correlated with ε. We consider the following parameter space and the norm $\|\cdot\|_s$:

$$\mathcal{H} = L^2(X) = \{g : E[g(X)^2] < \infty\}, \qquad \|g\|_s^2 = E[g(X)^2].$$

In addition, suppose we observe an instrumental variable $W \in [0,1]^d$ such that $E(\varepsilon|W) = 0$. Applications of instrumental variables can be found in many standard econometrics texts, for example, Hansen (2002). Let $Z = (Y, X)$; the NPIV model is then essentially a conditional moment restricted model with $\rho(Z, g) = Y - g(X)$.

Let $\{\phi_1, \phi_2, \ldots\}$ be a set of orthonormal basis functions of $L^2(X)$. We consider the sieve space $\mathcal{H}_n = \{g \in L^2(X) : g = \sum_{i=1}^{q_n} b_i\phi_i\}$, which can be partitioned into $\mathcal{H}_n = \mathcal{F}_n \cup \mathcal{F}_n^c$, where $\mathcal{F}_n = \{\sum_{i=1}^{q_n} b_i\phi_i \in \mathcal{H}_n, \max_{i\le q_n}|b_i| \le B_n\}$ as in Section 3.

We apply the feasible LIL (3.5) to construct the posterior. The log-likelihood involves the sample risk functional

$$\bar{G}(g) = \sum_{j=1}^{k_n^d}\left(\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - g(X_i)\big)I_{(W_i\in R_j^n)}\right)^2 \hat{v}_j^{-1},$$

which later will be shown to uniformly converge to

$$G(g) = E_W\big\{\big[E\big(Y - g(X)|W\big)\big]^2\big\}$$

over $\mathcal{F}_n$. The identified region Θ_I is defined as a subset of $L^2(X)$ on which $G(g) = 0$.

4.2. *Risk consistency.* Under mild conditions, we can derive the convergence rate of $\sup_{g\in\mathcal{F}_n}|\bar{G}(g) - G(g)|$. The following assumptions are imposed.

ASSUMPTION 4.1. (i) $k_n^{-d} = O(\min_{j\le k_n^d} P(W \in R_j^n))$;
(ii) $\max_{j\le k_n^d} P(W \in R_j^n) = O(k_n^{-d})$.

This assumption is satisfied, for example, when W has a continuous density function on $[0, 1]^d$ that is bounded away from both zero and infinity.

ASSUMPTION 4.2. There exists $C > 0$ such that for all $i = 1, \dots, q_n$:

(i) $\sup_w E(Y^2|W = w) < C$, $\sup_w E(\phi_i(X)^2|W = w) < C$;
(ii) $E(Y|W = w)$ is Lipschitz continuous with respect to w on $[0, 1]^d$;
(iii) for any $w_1, w_2 \in [0, 1]^d$,

$$\left|E\left(\phi_i(X)|W = w_1\right) - E\left(\phi_i(X)|W = w_2\right)\right| \le C\|w_1 - w_2\|.$$

Condition (iii) requires that the family $\{E(\phi_i(X)|W = w) : i \le q_n\}$ is Lipschitz equicontinuous on $[0, 1]^d$, which is satisfied, for example, when X has a density function that is bounded away from zero on the support of X; in addition, $X|W$ has a conditional density function $f_{X|W}$ such that for some $C > 0$,

$$\left|f_{X|W}(x|w_1) - f_{X|W}(x|w_2)\right| \le C\|w_1 - w_2\|$$

for all x and $w_1, w_2 \in [0, 1]^d$.[5]

ASSUMPTION 4.3. There exist $g_0 \in \Theta_I$, $g^*_{q_n} = \sum_{i=1}^{q_n} b^*_i \phi_i$ with $\sum_{i=1}^{\infty} b^{*2}_i < \infty$, and a positive sequence $\{\eta_j\}_{j=1}^{\infty}$ that strictly decreases to zero as $j \to \infty$ such that $\|g^*_{q_n} - g_0\|_s = O(\eta_{q_n})$ as $q_n \to \infty$. (We will choose $g^*_{q_n}$ to be the projection of g_0 onto $\mathcal{H}_n$, unless otherwise noted.)

Examples of the rate η_{q_n} are discussed earlier behind Assumption 3.3.

THEOREM 4.1. *Assume $q_n^2 B_n^2 = o(\min\{\sqrt{n}/k_n^{3d/2}, k_n\})$. Then under Assumptions 3.1, 4.1, 4.2,*

$$\sup_{g \in \mathcal{F}_n} |\bar{G}(g) - G(g)| = O_p\left(\frac{q_n^2 B_n^2 k_n^{3d/2}}{\sqrt{n}} + \frac{q_n^2 B_n^2}{k_n}\right).$$

Define a semi-norm $\|\cdot\|_w$, which is weaker than $\|\cdot\|_s$, as

$$\|g\|_w^2 = E_W\{(E(g(X)|W))^2\}. \tag{4.1}$$

It can be easily verified that $\|\cdot\|_w$ satisfies the triangular inequality, but $\|g\|_w = 0$ does not necessarily imply $g = 0$ if the conditional distribution $X|W$ is not complete. Note that $G(g) = \|g_0 - g\|_w^2$; hence this semi-norm induces an equivalence class characterized by the identified region $\Theta_I = \{g \in L^2(X) : E(Y - g(X)|W) =$

[5]This is simple to show: for any w_1, w_2, $|E(\phi_i(X)|W = w_1) - E(\phi_i(X)|W = w_2)| \le (\inf f_X(x))^{-1} \int |\phi_i(x) f_X(x)||f_{X|W}(x|w_1) - f_{X|W}(x|w_2)|\,dx \le C\|w_1 - w_2\| E|\phi_i(X)| \le C'\|w_1 - w_2\|$, where the fact that $E|\phi_i(X)|$ is bounded away from infinity is guaranteed by condition (i).

0, a.s.}, such that $\|g - g_0\|_w = 0$ if and only if $g \in \Theta_I$. In other words, we can say that g_0 *is weakly identified* under $\|\cdot\|_w$, since for any $g \in \Theta_I$, g and g_0 are equivalent under $\|\cdot\|_w$.

The following theorem is a straightforward application of Theorems 3.1 and 3.2:

THEOREM 4.2 (Risk-consistency). *Under Assumptions 3.1, 4.1–4.3, suppose* $\delta_n = O(1)$ *is such that*:

(i) *for the truncated priors assuming* $q_n^2 B_n^2 = o(n^{1/(3d+2)})$,

$$\max\left\{\eta_{q_n}^2, q_n^2 B_n^2\left(\frac{k_n^{3d/2}}{\sqrt{n}} + \frac{1}{k_n}\right)\right\} = o(\delta_n),$$

(ii) *for the thin-tail prior with* $r > 6d + 4$, *assuming* $q_n = o(n^{1/(6d+4)-1/r})$,

$$\max\left\{\eta_{q_n}^2, n^{2/(r-2)} q_n^{2r/(r-2)}\left(\frac{k_n^{3d/2}}{\sqrt{n}} + \frac{1}{k_n}\right)^{r/(r-2)}\right\} = o(\delta_n),$$

then

$$P(\|g_b - g_0\|_w > \delta_n | X^n) = o_p(1).$$

4.3. *Ill-posedness and posterior consistency.* Define

$$T : L^2(X) \to \{\zeta : E[\zeta(W)^2] < \infty\}, \qquad T(g) = E(g(X)|W)$$

and write $E(Y|W = w) \equiv \zeta(w)$. Then the NPIV model can be equivalently written as

$$T g_0 = \zeta. \tag{4.2}$$

Under Assumption 4.4, T is a compact linear operator [see Carrasco, Florens and Renault (2007)], and therefore is continuous. Equation (4.2) is usually called the *Fredholm integral equation of the first kind*.

ASSUMPTION 4.4. The joint distribution (Y, X, W) is absolutely continuous with respect to the Lebesgue measure. In addition, suppose $f_{XW}(x, w)$, $f_X(x)$, $f_W(w)$ denote the density functions of (X, W), X and W, respectively, then

$$\iint \left(\frac{f_{XW}(x, w)}{f_X(x) f_W(w)}\right)^2 f_X(x) f_W(w)\, dx\, dw < \infty.$$

As described before, the problem of inference about g_0 is ill-posed in two aspects. The first ill-posedness comes from the identification, which depends on the invertibility of T. If T is nonsingular, in which case its null space is {0}, g_0 can be point identified by $g_0 = T^{-1}\zeta$, but not otherwise. See Severini and Tripathi (2006) and D'Haultfoeuille (2011) for detailed descriptions of the identification issues.

Even when g_0 is identified, in which case T^{-1} exists, as pointed out by Florens (2003) and Hall and Horowitz (2005), since $L^2(X)$ is of infinite dimension, and T is compact, T^{-1} is not bounded (therefore is not continuous). As a result, small inaccuracy in the estimation of ζ can lead to large inaccuracy in the estimation of g_0, which is known as the type-III ill-posed inverse problem described in Section 3.2. When g_0 is partially identified, this problem is still present when

$$\liminf_{n\to\infty} \inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g) = \liminf_{n\to\infty} \inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} E\{[T(g-g_0)]^2\} = 0.$$

By Theorems 3.3, 3.4 and 4.2, in order to achieve the posterior consistency, it suffices to verify

$$\delta_n^* = o\Big(\inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g)\Big), \tag{4.3}$$

where

$$\text{for truncated prior } \delta_n^* = \max\left\{\eta_{q_n}^2, q_n^2 B_n^2\left(\frac{k_n^{3d/2}}{\sqrt{n}} + \frac{1}{k_n}\right)\right\},$$

$$\text{for thin-tail prior } \delta_n^* = \max\left\{\eta_{q_n}^2, n^{2/(r-2)} q_n^{2r/(r-2)}\left(\frac{k_n^{3d/2}}{\sqrt{n}} + \frac{1}{k_n}\right)^{r/(r-2)}\right\}.$$

Hence it requires us to derive a lower bound of $\inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g)$ first, and, in addition, this lower bound should decay at a rate slower than δ_n^*.

When g_0 is point identified and a slowly growing finite-dimensional sieve is used, Chen and Pouzo (2009a) showed the existence of such a lower bound using the singular value decomposition of T. Their approach is briefly illustrated in the following example.

EXAMPLE 4.1. Let $\langle g_1, g_2\rangle_X = E[g_1(X)g_2(X)]$ denote the inner product of two elements in $L^2(X)$, and $\{\nu_j, \phi_{1j}, \phi_{2j}\}_{j=1}^\infty$ be the ordered singular value system of T such that

$$T\phi_{1j} = \nu_j\phi_{2j}, \qquad \nu_1^2 \ge \nu_2^2 \ge \cdots.$$

Suppose T is nonsingular, then $\{\phi_{1j}\}_{j=1}^\infty$ forms an orthonormal basis of $L^2(X)$. Chen and Pouzo (2009a) showed that when $\{\phi_{1j}\}_{j=1}^{q_n}$ is used as the basis in the sieve approximation space, $\forall\varepsilon > 0$, $\nu_{q_n}^2 = O(\inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g))$. Therefore, condition (4.3) is satisfied if we assume $\delta_n^* = o(\nu_{q_n}^2)$. In addition, suppose $\{\nu_j^2\}_{j=1}^\infty$ decays at a polynomial rate $j^{-\alpha}$ for some $\alpha > 0$; then we require $q_n = o(\delta_n^{*-1/\alpha})$, a slowly growing sieve dimension.

We impose the following assumption to derive a lower bound for $\inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g)$ and verify (4.3), which, in the identified case, uses more general basis functions for the sieve space. Therefore we allow the sieve basis to be different from the eigenfunctions of T. A similar approach was used by Chen and Reiss [(2011), Section 6.1], who used the wavelets as the sieve basis functions while the eigenfunctions of T form a Fourier basis.

ASSUMPTION 4.5. There is a continuous and increasing function $\varphi(\cdot) > 0$ satisfying $\lim_{t \to 0^+} \varphi(t) = 0$ such that, for $\{g_0, g^*_{q_n}, \{\eta_j\}_{j=1}^\infty\}$ as defined in Assumption 4.3 and some constants $C_1, C_2 > 0$:

(i) $\|g - g_0\|_w^2 \geq C_1 \sum_{j=1}^\infty \varphi(\eta_j^2) |\langle g - g_0, \phi_j \rangle_X|^2$ for all $g \in L^2(X)$;
(ii) $\|g^*_{q_n} - g_0\|_w^2 \leq C_2 \sum_j \varphi(\eta_j^2) |\langle g_0 - g^*_{q_n}, \phi_j \rangle_X|^2$.

REMARK 4.1. (1) This assumption implies a generalization of the relation $\nu_{q_n}^2 = O(\inf_{g \in \mathcal{H}_n, g \notin \Theta_I^\varepsilon} G(g))$ in Example 4.1. In this assumption, $\{\phi_j\}_{j=1}^\infty$ are the basis functions whose first q_n terms span the sieve approximation space. In the identified case, $\{\phi_j\}_{j=1}^\infty$ can be a general set of basis functions that is different from the eigenfunctions of T. Chen and Pouzo [(2009a), Section 5.3] identified the singular value ν_j^2 of Example 4.1 as a special case of the general $\varphi(\eta_j^2)$, in which case Assumption 4.5 is satisfied. In its general form, Assumption 4.5 is standard in the literature for the linear ill-posed inverse problem when the convergence rate of the estimator is studied; see, for example, Nair, Pereverzev and Tautenhahn (2005), Chen and Pouzo [(2009a), Assumption 5.2], Chen and Reiss [(2011), Section 2.1], etc. As described above, however, this assumption is also needed in order to verify (4.3) and show consistency when general basis functions are used. Blundell, Chen and Kristensen (2007) provided sufficient conditions of Assumption 4.5 for the NPIV model setting.

(2) In the partially identified case when Θ_I is not a singleton, Assumption 4.5 is still satisfied, if we take $\{\phi_j\}_{j=1}^\infty$ to be the eigenfunctions of T^*T that correspond to its nonzero eigenvalues, where T is the conditional expectation operator, and T^* is its adjoint. The spectral theory of compact operators [Kress (1999)] implies that $\|T(g - g_0)\|_s^2 = \sum_{j=1}^\infty \nu_j^2 |\langle g - g_0, \phi_j \rangle_X|^2$ for all $g \in L^2(X)$, where $\{\nu_j^2\}$ represent all the (nonzero) eigenvalues of T^*T, and $\{\phi_j\}$ are the corresponding eigenfunctions (the zero eigenvalues of T^*T do not contribute to the right-hand side of the spectral decomposition). Therefore, Assumption 4.5 remains valid with $\varphi(\eta_j^2) = \nu_j^2$, with $\{\nu_j^2\}$ denoting the sequence of decreasing nonzero eigenvalues. This idea of using the spectral representation of T^*T is related to the commonly used "general source condition" in the literature [Tautenhahn (1998) and Darolles et al. (2011)], where, for example, Darolles et al. (2011) used this condition to derive the convergence rate of their kernel-based Tikhonov regularized estimator in NPIV regression.

(3) When a more general sieve basis $\{\phi_j\}_{j=1}^\infty$ is used in the partially identified case, condition (i) of Assumption 4.5 is not generally satisfied. For example, suppose there exists $g \in \Theta_I$, but $g \neq g_0$. By the definition of $\|\cdot\|_w$, $\|g - g_0\|_w^2 = 0$,

but the right-hand side of the displayed inequality in condition (i) is strictly positive unless $\{\phi_j\}_{j=1}^{\infty}$ are the eigenfunctions of T^*T. To allow for more general sieve basis in this case, a possible approach is to assume the true g_0 in the data generating process to lie in a compact set Θ, for example., a Sobolev ball [Chen and Reiss (2011)]. It is then not hard to show that $\inf_{g\in\Theta,g\notin\Theta_I^{\varepsilon}} G(g)$ is bounded away from zero. Restricting g_0 inside a compact set is actually a quite common approach in nonparametric IV regression, and the literature is found in Newey and Powell (2003), Blundell, Chen and Kristensen (2007), Chen and Reiss (2011), etc. Recently, Santos (2012) extended this approach to the partially identified case, with the compactness restriction. We do not pursue this approach here, since our other results on posterior consistency allow a noncompact parameter space.

As in Chen and Pouzo (2009a), generally the degree of ill-posedness has two types:

(1) *mild ill-posedness*: $\varphi(\eta)=\eta^{\alpha}$ for some $\alpha>0$.
(2) *severe ill-posedness*: $\varphi(\eta)=\exp(-\eta^{-\alpha})$ for some $\alpha>0$.

Under Assumption 4.5, it can be shown that $\varphi(\eta_{q_n}^2)=O(\inf_{g\in\mathcal{H}_n,g\notin\Theta_I^{\varepsilon}} G(g))$ for any $\varepsilon>0$; see Lemma C.5 of the supplementary material. Intuitively speaking, $\varphi(\cdot)$ is associated with the singular values of T and is related to how severe the type-III ill-posed inverse problem is. When the nonzero singular values decay at a polynomial rate, φ corresponds to the mildly ill-posed case; when the singular values decay at an exponential rate, it corresponds to the severely ill-posed case.

Before formally presenting our posterior consistency result, we briefly comment on the role of condition (ii) of Assumption 4.5. Assumption 5.2(ii) is the so-called "stability condition" in Chen and Pouzo (2009a) that is required to hold only in terms of the sieve approximation error on one element in Θ_I. By Theorems 3.3 and 3.4, we require $G(g_{q_n}^*)=o(\inf_{g\in\mathcal{H}_n,g\notin\Theta_I^{\varepsilon}} G(g))$. It can be easily shown that $G(g_{q_n}^*)=O(\eta_{q_n}^2)$, and hence $G(g_{q_n}^*)$ was replaced with $\eta_{q_n}^2$ in the condition of Theorem 4.2. In addition, condition (i) of Assumption 4.5 implies that $\varphi(\eta_{q_n}^2)=O(\inf_{g\in\mathcal{H}_n,g\notin\Theta_I^{\varepsilon}} G(g))$. With condition (ii) of Assumption 4.5, it can be further shown that $G(g_{q_n}^*)=O(\eta_{q_n}^2\varphi(\eta_{q_n}^2))$ (see Lemma C.6 in the supplementary material). Since $\eta_{q_n}^2=o(1)$, $G(g_{q_n}^*)=o(\varphi(\eta_{q_n}^2))=o(\inf_{g\in\mathcal{H}_n,g\notin\Theta_I^{\varepsilon}} G(g))$ is verified.

Under this framework, we have the posterior consistency under $\|\cdot\|_s$:

THEOREM 4.3 (Posterior consistency). *Under Assumptions* 3.1, 4.1–4.5, *suppose*:

(i) *for the truncated priors assuming* $q_n^2B_n^2=o(n^{1/(3d+2)})$,

$$q_n^2B_n^2\left(\frac{k_n^{3d/2}}{\sqrt{n}}+\frac{1}{k_n}\right)=o(\varphi(\eta_{q_n}^2)); \tag{4.4}$$

(ii) *for the thin-tail prior with* $r > 6d + 4$, *assuming* $q_n = o(n^{1/(6d+4)-1/r})$,

$$n^{2/(r-2)} q_n^{2r/(r-2)} \left(\frac{k_n^{3d/2}}{\sqrt{n}} + \frac{1}{k_n} \right)^{r/(r-2)} = o(\varphi(\eta_{q_n}^2)). \tag{4.5}$$

Then for any $\varepsilon > 0$,

$$P\big(d(g_b, \Theta_I) > \varepsilon | X^n\big) = o_p(1).$$

4.4. *Normal prior.* When g_0 is point identified, we can also establish the posterior consistency using normal priors

$$\pi(b) = \prod_{i=1}^{q_n} \pi_i(b_i), \qquad \pi_i(b_i) \sim N(0, \sigma^2), \tag{4.6}$$

for some constant $\sigma^2 > 0$. As discussed previously, by restricting q_n to grow slowly as $n \to \infty$, we do not need a shrinking prior to function as a penalty term attached to the log-likelihood for the regularization purpose.[6] Therefore σ^2 is treated to be a fixed constant that does not depend on n.

With the assumptions imposed in Sections 4.2 and 4.3, we can verify all the conditions in Theorem 2.2, which then leads to the following theorem:

THEOREM 4.4 (Posterior consistency using Gaussian prior). *Assume* g_0 *is point identified. Under Assumptions* 3.1, 4.1–4.5, *suppose the normal prior* (4.6) *is used, and*

$$q_n \left(\frac{k_n^{3d/2}}{\sqrt{n}} + \frac{1}{k_n} \right)^{1/3} = o(\varphi(\eta_{q_n}^2)), \tag{4.7}$$

then for any $\varepsilon > 0$,

$$P(\|g_b - g_0\|_s > \varepsilon | X^n) = o_p(1).$$

4.5. *Choice of tuning parameters.* To choose (k_n, q_n, B_n) that satisfy (4.4) (4.5) and (4.7) for each specified prior, consider the case where η_{q_n} is decreasing as some power of q_n [see, e.g., Schumaker (1981) and Meyer (1990)], and k_n grows at a polynomial rate of n, that is,

$$\begin{aligned} &\eta_{q_n} \sim q_n^{-\upsilon} && \text{for some } \upsilon > 0, \\ &\frac{k_n^{3d/2}}{\sqrt{n}} + \frac{1}{k_n} \sim n^{-p}, && 0 < p \leq \frac{1}{3d+2}. \end{aligned} \tag{4.8}$$

We then have the following corollaries:

[6] We thank a referee for pointing this out.

COROLLARY 4.1 (Truncated prior). *Suppose the truncated prior* (*either uniform or truncated normal*) *is used*; *then the following choice of* (q_n, B_n) *achieves the posterior consistency*, *for* $b < p$:

(i) *in the mildly ill-posed case*,

$$B_n^2 \sim n^b, \qquad q_n = o\big(n^{(p-b)/(2+2\alpha v)}\big);$$

(ii) *in the severely ill-posed case*,

$$B_n^2 \sim n^b, \qquad q_n = o\big((\log n)^{1/(2\alpha v)}\big).$$

COROLLARY 4.2 (Thin-tail prior). *Suppose the thin-tail prior is used*; *then the following choice of* q_n *achieves the posterior consistency*, *for* $pr > 2$:

(i) *in the mildly ill-posed case*,

$$q_n = o\big(n^{(pr-2)/(2r+2\alpha v(r-2))}\big);$$

(ii) *in the severely ill-posed case*,

$$q_n = o\big((\log n)^{1/(2\alpha v)}\big).$$

COROLLARY 4.3 (Normal prior). *Suppose the normal prior is used*, *and* g_0 *is point identified*, *the following choice of* q_n *achieves the posterior consistency*:

(i) *in the mildly ill-posed case*,

$$q_n = o\big(n^{p/(3(1+2\alpha v))}\big);$$

(ii) *in the severely ill-posed case*,

$$q_n = o\big((\log n)^{1/(2\alpha v)}\big).$$

In the conditions of these consistency results, the choice of tuning parameters (q_n, B_n, r) depend on some parameters that one either knows or chooses (d, p), as well as some parameters related to the true model (α, v). The latter, although undesirable, cannot be totally avoided when we study the frequentist convergence properties under ill-posedness. [Conditions depending on the true model are also used, e.g., by Chen and Pouzo (2009a), directly in their Corollary 5.1, and indirectly at the end of their Section 3.1.]

On the other hand, these results can still have meaningful implications that do not explicitly depend on the indexes α and p (which are probably unknown in practice). For example, we note that in the mildly ill-posed situations, the condition on q_n would be satisfied if it grows as any finite power of $\log n$. Likewise, in the severely ill-posed situations, the condition on q_n would be satisfied if it grows as any finite power of $\log\log n$.

In addition, we will indicate in the next section that the current Bayesian-flavored treatment can even allow a data-driven choice of the sieve dimension q_n, using a posterior distribution derived from a mixed prior.

5. Random sieve dimension. As the sieve dimension q_n plays an important role not only in dealing with the ill-posed inverse problem, but also in many applied sieve estimation methods, in this section we briefly discuss the possibility of choosing it based on a posterior distribution. This will require specifying a prior distribution on the sieve dimension first. Since the conditions of a deterministic q_n for consistency only restricts the growth rate, as a result, Mq_n would also lead to consistency for a positive constant $M > 1$, if q_n ensures consistency.

We denote the sieve dimension by q, let it be random and place a discrete uniform prior

$$\pi(q) = \mathrm{Unif}\{1, \dots, Mq_n\} \tag{5.1}$$

for some deterministic sequence $q_n \to \infty$ and constant $M > 1$. Then the prior on the sieve coefficients b becomes a mixture prior

$$\pi(b) = \sum_{q=1}^{Mq_n} \pi(q)\pi(b|q) = \sum_{q=1}^{Mq_n} (Mq_n)^{-1}\pi(b|q), \tag{5.2}$$

where $\pi(b|q)$ follows a prior as specified before for a given sieve dimension q. The feasible limited information likelihood is, as before, denoted by $L_n(b, q)$. We have the joint posterior

$$p(g_b, q|X^n) \propto \pi(b|q)L_n(b, q).$$

It can be shown that the uniform mixture prior can also lead to the posterior consistency.

THEOREM 5.1 (RANDOM q). *For each theorem in Sections 3 and 4, suppose the corresponding conditions are satisfied for the deterministic sieve dimension Mq_n instead of q_n, for some $M > 1$. Then all the posterior consistency results stated in Sections 3 and 4 (on risk consistency and on estimation consistency) remain valid for the mixed prior (5.2) with random q following prior (5.1), with no extra conditions, with the following two exceptions:*

(1) *We will additionally assume that $(\log q_n)/n = o(\delta_n)$ holds for the statement of Theorem 3.2 to hold.*

(2) *We will additionally assume that $(\log q_n)/n = o(\inf_{g\in\mathcal{H}_n, g\notin\Theta_I^\varepsilon} G(g))$ for the statement of Theorem 3.2 to hold.*

Note that the uniform prior is used for q, which gives zero prior probability on very large choice beyond Mq_n. However, from a technical point of view, the result can be extended to the case with tails of prior on q extending to infinity, as long as the tail is thin enough so that $\pi(q > Mq_n)$ is dominated by a small enough upper bound.

The marginal posterior of q is given by

$$p(q|X^n) \propto \int \pi(b|q)L_n(b, q)\,db. \tag{5.3}$$

Practically, we can choose q from $p(q|X^n)$.

6. Conclusion and discussion. We studied the nonparametric conditional moment restricted model in a quasi-Bayesian approach, with a special focus on the large sample frequentist properties of the posterior distribution. There was no distribution assumed on the data generating process. Instead, we derived the posterior using the *limited information likelihood* (*LIL*), allowing the proposed procedure to be simpler than the traditional nonparametric Bayesian approach which would model the data distribution nonparametrically. There are several alternative moment-condition-based likelihood functions. The empirical likelihood [Owen (1990))] and the generalized empirical likelihood [Imbens, Spady and Johnson (1998), Newey and Smith (2004) and Kitamura (2006)] are typical examples. It is still possible to establish the posterior consistency if these alternative nonparametric likelihoods are used, which is left as a future research direction.

The parameter space $\mathcal{H}$ does not need to be compact. We approximate $\mathcal{H}$ using a finite-dimensional sieve space $\mathcal{H}_n$, and the regularization is carried out by a slowly growing sieve dimension q_n. We then studied in detail the NPIV model and verified all the sufficient conditions proposed in Section 3 in order for the posterior to be consistent.

It is also possible to achieve the posterior consistency using a larger sieve dimension q_n. In this case, the regularization is carried out by a truncated normal prior with shrinking variance, and the log-prior is then a regularization penalty attached to the log-likelihood. Conditions (3.10), (3.11) and Assumption 4.5 can be relaxed. We describe this procedure in the Technical Report [Liao and Jiang (2011b)].

An interesting research direction is to derive the convergence rate. With all the tools given in this paper, it is possible to obtain the rate of convergence of our procedure. However, the rate would be sub-optimal, possibly due to the technical bound (2.1) used in this paper. It would be interesting to develop a method based on a bound tighter than (2.1), in order to prove the nonparametric minimax optimal rate of convergence as in Chen and Pouzo (2009b).

In applications, our method requires a priori choices of (k_n, q_n), and B_n for the truncated prior. We conjecture that the finite sample behavior of the posterior is robust to the choice of (k_n, B_n). However, it should be sensitive to q_n, as a large value of q_n may lead to over-fitting. Therefore, we proposed an approach to allow for a random sieve dimension by putting a discrete uniform prior on it and selecting it from its posterior. With the upper bound of the uniform prior Mq_n growing under the same rate restriction as before, the posterior consistency is also achieved. This feature, however, requires specifying Mq_n. In practice, one may start with a moderate level Mq_n that is less than ten. In the NPIV setting, Horowitz (2010) recently introduced an empirical approach for selecting q_n. Moreover, developing methods of selecting (k_n, B_n) in a Bayesian (or quasi-Bayesian) approach is another important research topic.

Acknowledgments. This paper develops from a chapter of the first author's Ph.D. dissertation at Northwestern University. We are grateful to Joel Horowitz, Elie Tamer, Hidehiko Ichimura, Jia-Young Fu, Tom Severini, Xiaohong Chen, Anna Simoni, an Associate Editor and two referees for many helpful comments and suggestions on this paper. We also thank the discussions with seminar participants at the 2010 Summer CEMMAP conference on "Recent developments in nonparametric instrumental variable methods" in London. The first author appreciates the constant encouragements from his Ph.D. committee members at Northwestern University.

SUPPLEMENTARY MATERIAL

Technical proofs (DOI: 10.1214/11-AOS930SUPP; .pdf). This supplementary material contains the proofs of all the results developed in the main paper.

REFERENCES

AI, C. and CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71** 1795–1843. MR2015420

ANTONIADIS, A., GRÉGOIRE, G. and MCKEAGUE, I. W. (2004). Bayesian estimation in single-index models. *Statist. Sinica* **14** 1147–1164. MR2126345

AZZALINI, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica* (*Bologna*) **46** 199–208. MR0877720

BLUNDELL, R., CHEN, X. and KRISTENSEN, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75** 1613–1669. MR2351452

CARRASCO, M., FLORENS, J. and RENAULT, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In *Handbook of Econometrics* (J. J. Heckman and E. E. Leamer, eds.) **VI** Chapter 77. North-Holland, Amsterdam.

CHEN, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics* (J. J. Heckman and E. E. Leamer, eds.) **VI** Chapter 76. North-Holland, Amsterdam.

CHEN, X. and LUDVIGSON, S. C. (2009). Land of addicts? An empirical investigation of habit-based asset pricing models. *J. Appl. Econometrics* **24** 1057–1093. MR2750279

CHEN, X. and POUZO, D. (2009a). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*. To appear. Cowles Foundation Discussion Paper 1650R, Yale Univ.

CHEN, X. and POUZO, D. (2009b). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *J. Econometrics* **152** 46–60. MR2562763

CHEN, X. and REISS, M. (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory* **27** 497–521. MR2806258

CHERNOZHUKOV, V., GAGLIARDINI, P. and SCAILLET, O. (2008). Nonparametric instrumental variable estimation of quantile structural effects. Unpublished manuscript. MIT, Cambridge, MA.

CHERNOZHUKOV, V. and HANSEN, C. (2005). An IV model of quantile treatment effects. *Econometrica* **73** 245–261. MR2115636

CHERNOZHUKOV, V. and HONG, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346.

CHERNOZHUKOV, V., HONG, H. and TAMER, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica* **75** 1243–1284. MR2347346

CHERNOZHUKOV, V., IMBENS, G. W. and NEWEY, W. K. (2007). Instrumental variable estimation of nonseparable models. *J. Econometrics* **139** 4–14. MR2380756

CHOI, T. and SCHERVISH, M. J. (2007). On posterior consistency in nonparametric regression problems. *J. Multivariate Anal*. **98** 1969–1987. MR2396949

DAROLLES, S., FAN, Y., FLORENS, J. P. and RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica* **79** 1541–1565.

D'HAULTFOEUILLE, X. (2011). On the completeness condition in nonparametric instrumental problems. *Econometric Theory* **27** 460–471. MR2806256

FLORENS, J. (2003). Inverse problems and structural econometrics: The example of instrumental variables. In *Advances in Economics End Econometrics*: *Theory and Applications* (M. Dewatripont, L. P. Hansen and S. J. Turnovsky, eds.). *Invited Lectures to the World Congress of the Econometric Society*, *Seattle* 2000 **II** 284–311. Cambridge Univ. Press, Cambridge.

FLORENS, J. and SIMONI, A. (2009a). Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior. Unpublished manuscript. Toulouse School of Economics, Toulouse, France.

FLORENS, J. and SIMONI, A. (2009b). Regularizing priors for linear inverse problems. Unpublished manuscript. Toulouse School of Economics, Toulouse, France.

FLORENS, J. and SIMONI, A. (2011). Bayesian identification and partial identification. Unpublished manuscript. Toulouse School of Economics, Toulouse, France.

GALLANT, A. R. and TAUCHEN, G. (1989). Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications. *Econometrica* **57** 1091–1120. MR1014542

GHOSAL, S. and ROY, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *Ann. Statist*. **34** 2413–2429. MR2291505

GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist*. **35** 192–223. MR2332274

GHOSH, J. K. and RAMAMOORTHI, R. V. (2003). *Bayesian Nonparametrics*. Springer, New York. MR1992245

HALL, P. and HOROWITZ, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist*. **33** 2904–2929. MR2253107

HAN, C. and PHILLIPS, P. C. B. (2006). GMM with many moment conditions. *Econometrica* **74** 147–192. MR2194322

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123

HANSEN, B. (2002). Econometrics. Unpublished manuscript. Univ. Wisconsin, Madison.

HOROWITZ, J. L. (2007). Asymptotic normality of a nonparametric instrumental variables estimator. *Internat. Econom. Rev.* **48** 1329–1349. MR2375627

HOROWITZ, J. (2010). Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. Unpublished manuscript. Northwestern Univ.

HOROWITZ, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica* **79** 347–394. MR2809374

HOROWITZ, J. L. and LEE, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica* **75** 1191–1208. MR2333498

HUANG, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist*. **32** 1556–1593. MR2089134

ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. MR1230981

IMBENS, G. W., SPADY, R. H. and JOHNSON, P. (1998). Information-theoretic approaches to inference in moment condition models. *Econometrica* **66** 333–357. MR1612246

JIANG, W. and TANNER, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist*. **36** 2207–2231. MR2458185

JOHANNES, J., VAN BELLEGEM, S. and VANHEMS, A. (2010). Iterative regularization in nonparametric instrumental regression. Unpublished manuscript. Toulouse School of Economics, Toulouse, France.

KIM, J.-Y. (2002). Limited information likelihood and Bayesian analysis. *J. Econometrics* **107** 175–193. MR1889958

KITAMURA, Y. (2006). Empirical likelihood methods in econometrics: Theory and practice. Unpublished manuscript. Yale Univ.

KRESS, R. (1999). *Linear Integral Equations*, 2nd ed. *Applied Mathematical Sciences* **82**. Springer, New York. MR1723850

LIAO, Y. and JIANG, W. (2010). Bayesian analysis in moment inequality models. *Ann. Statist.* **38** 275–316. MR2589323

LIAO, Y. and JIANG, W. (2011a). Supplement to "Posterior consistency of nonparametric conditional moment restricted models." DOI:10.1214/11-AOS930SUPP.

LIAO, Y. and JIANG, W. (2011b). Posterior consistency of nonparametric conditional moment restricted models using a shrinking prior. Technical report, Northwestern Univ. Available at http://newton.stats.northwestern.edu/~jiang/cmrm/suppG.pdf.

MEYER, Y. (1990). *Ondelettes et Opérateurs. I*. Hermann, Paris. MR1085487

NAIR, M. T., PEREVERZEV, S. V. and TAUTENHAHN, U. (2005). Regularization in Hilbert scales under general smoothing conditions. *Inverse Problems* **21** 1851–1869. MR2183654

NEWEY, W. K. and POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71** 1565–1578. MR2000257

NEWEY, W. K. and SMITH, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72** 219–255. MR2031017

OWEN, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120. MR1041387

SANTOS, A. (2011). Instrumental variable methods for recovering continuous linear functionals. *J. Econometrics* **161** 129–146. MR2774933

SANTOS, A. (2012). Inference in nonparametric instrumental variables with partial identification. *Econometrica* **80** 213–275.

SCHUMAKER, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York. MR0606200

SEVERINI, T. A. and TRIPATHI, G. (2006). Some identification issues in nonparametric linear models with endogenous regressors. *Econometric Theory* **22** 258–278. MR2230389

SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

TAUTENHAHN, U. (1998). Optimality for ill-posed problems under general source conditions. *Numer. Funct. Anal. Optim.* **19** 377–398. MR1624930

WALKER, S. (2003). Bayesian consistency for a class of regression problems. *South African Statist. J.* **37** 149–167. MR2042627

DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
SHERRERD HALL
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: yuanliao@princeton.edu

DEPARTMENT OF STATISTICS
NORTHWESTERN UNIVERSITY
2006 SHERIDAN RD
EVANSTON, ILLINOIS 60208
USA
E-MAIL: wjiang@northwestern.edu