Nonparametric Estimation of an Instrumental Regression: a Bayesian Approach Based on Regularized Posterior

Jean-Pierre Florens
Toulouse School of Economics
(GREMAQ and IDEI)

Anna Simoni
Toulouse School of Economics
(GREMAQ)

May 14, 2008

Preliminary and Incomplete Version

Abstract

In this paper we deal with Bayesian inference about an instrumental regression function φ that is defined through a moment condition involving the random vector $S = (Y, Z, W)$. S is jointly distributed as F; the variables in the subvector (Y, Z) are endogenous while W is a subvector of instruments. Moment restrictions of this kind are very often encountered in structural econometric model and we exploit them to construct a conditional probability measure on the sample space given the parameter φ. The instrumental regression is not constrained to belong to a finite dimensional space, but we only impose some regularity condition and inference is directly performed in the infinite dimensional space L^2.
The solution of this inference problem is the posterior distribution of the unknown random function φ. However, the unboundedness of covariance operators in infinite dimensional spaces causes problems of non continuity of the posterior mean and rises a problem of posterior inconsistency in the frequentist sense. To avoid such kind of problems we adopt a regularized version of the posterior distribution that we compute through a Tikhonov regularization scheme and that we show to satisfy posterior consistency.
We consider three different degrees of knowledge of the joint distribution $F(\cdot, Z, W)$: completely known, known up to a finite dimensional parameter and completely unknown. In the last two cases estimation is performed in two steps, the first one providing a bayesian parametric or classical nonparametric estimator of $F(\cdot, Z, W)$ and the second one the regularized bayesian estimator of φ. We develop asymptotic analysis in a frequentist sense and posterior consistency is proved in all three cases.
Moreover, the model is extended to consider a partially unknown variance in the sampling distribution. We provide the conditional posterior laws and a Gibbs sampling algorithm is implemented.

JEL codes: C11, C14, C30.
Keywords: Instrumental Regression, Nonparametric Estimation, Posterior distribution, Tikhonov Regularization, Posterior Consistency.

1 Introduction

Instrumental regression estimation has always played a central role in econometric theory. Economic analysis provides econometricians with theoretical models, describing a certain phenomenon, that specify relations between economic variables: a response variable, denoted with Y, and a vector of explanatory variables, denoted with Z. The variables in Z can be endogenous or exogenous and the relation is of the form $Y = \varphi(Z) + U$, where $\varphi(\cdot)$ expresses the link we are interested in and, in the most easy case when Z are exogenous, $\varphi(Z) = \mathbb{E}(Y|Z)$. Unfortunately, in several economic models the explanatory variables are endogenous and so the parameter of interest $\varphi(Z)$ is not the conditional expectation function. In this latter case, the structural econometric model we have to deal with can be written in very general terms as

$$Y = \varphi(Z) + U, \qquad \mathbb{E}(U|Z) \neq 0.$$

The hypothesis about the error term plays a crucial role and, if we neglect it and perform a classical estimation by considering Z as exogenous, we get an estimation of the conditional expectation function $\mathbb{E}(Y|Z)$ that is not the structural parameter of interest.
This specification of the model is not enough to estimate the structural parameter of interest φ and some assumption must be added in order to have a further characterization of φ. A first strategy proposed in literature consists in adding hypothesis regarding the joint distribution of U and Z, but this will not be the strategy followed here.

Alternatively, it is possible to add to the vector of observations (Y, Z) a vector of observed variables correlated with Z, that we call W. Since the variables in W are introduced to make inference possible, they are called *instrumental variables* and the vector of observed variables becomes (Y, W, Z). After that, to characterize and define φ, some hypothesis about the relation between W and the disturbances in the model must be introduced.
A third approach proposed in literature for treating endogeneity problems is the *control function approach* proposed by [22]. They consider a triangular nonparametric simultaneous equations model with some restriction on the error terms of the structural and reduced form equations and on the exogenous variables.
In this paper we adopt the instrumental regression approach by adding a vector W of instruments and by replacing the classical hypothesis of exogeneity $\mathbb{E}(U|Z) = 0$ with the hypothesis $\mathbb{E}(U|W) = 0$. As stressed by Newey and Powell (2003) [21], when we are considering a nonparametric estimation the strong condition that the error term is mean independent of the instrument is important for identification while a finite number of zero covariance restriction between the instruments and the disturbances will not suffice to identify an infinite dimension parameter. Therefore, the structural parameter of interest φ is characterized as the solution of

$$\mathbb{E}(Y - \varphi(Z)|W) = 0$$

and it is called *instrumental regression*.

In this paper we are going to exploit this moment restriction in order to make inference about the instrumental regression without imposing any constraint on the functional form of φ. Then, we estimate a parameter of infinite dimension. Anyway, even if we do not limit φ to be in a space of finite dimension, we propose to take into account all the information we have *a priori* on the data generating process of the instrumental regression by incorporating it in a *prior distribution* on the parameter space. We conceive therefore the instrumental regression not as a given parameter but as a realization of a random process and we work in the product space of the sampling and parameter space. This study is primarily aimed by a Bayesian philosophy and we transform an inverse problem in a problem of estimation, as it is natural in the Bayesian approach to inverse problems, see Franklin (1970) [11]. We refer to Florens and Simoni (2008) [10] for a more complete discussion about this approach.
Application of Bayes theorem in infinite dimensional spaces is perfectly known (see [11] and [19]) and the posterior distribution is well defined. However, the posterior mean presents a problem of continuity due to the fact that its expression involves the inverse of a covariance operator that is unbounded if we do not assume that the covariance operator is proportional to the identity operator. Hence, consistency, in the frequentist sense, of the posterior distribution is not verified.

To overcome this problem, we adopt the strategy proposed in [10] consisting in applying a regularization scheme and in replacing the posterior distribution with a *regularized posterior distribution* obtained by using a Tikhonov regularization scheme.
The idea of the instrumental regression as solution of an ill-posed inverse problem is primarely due to Florens (2002) [8], Darolles, Florens and Renault (2005) [5] and Hall and Horowitz (2005) [16].
The Bayesian optics that moves this study is in any case not binding. In particular, if we adopt a classical point of view, where a true value of the parameter of interest that characterizes the distribution having generated the data exists, the proposed Bayesian estimator of the instrumental regression converges toward the true value of the parameter. This is the notion of frequency consistency, more properly we prove that the regularized posterior distribution degenerates to a Dirac measure in correspondence of the true value of the parameter of interest. This is the notion of posterior consistency and is verified under the hypothesis that the true instrumental regression satisfies some minor regularity condition.
The paper is organized as follows. In the following section we characterize the model and introduce some notation. Section 3 presents the formal statement of the Bayesian experiment in the general case with unknown variance parameter and conjugate prior distributions. We characterize the solution of the inference problem as a regularized version of the posterior distribution. Then, we consider the slightly different situation with independent priors. In section 4 we develop inference on φ when the joint density of the explanatory variables and the instruments $f(\cdot, Z, W)$ is unknown. A preliminary step of estimation of this density is required and, in particular, two alternative strategies to accomplish this step are presented. The first one is a Bayesyan parametric method that applies when the joint density is known up to a finite dimensional parameter; the second one consists in a classical nonparametric estimation and applies when the density is completely unknown. Lastly, in section 5 we show some result of numerical simulations of the more relevant cases previously considered and section 6 concludes. All the proofs can be found in the Appendix.

2 The Model

Let $S = (Y, Z, W)$ denote the random vector belonging to $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ with distribution characterized by the cumulative distribution function F. We assume that F is absolutely continuous with respect to Lebeasgue measure with density f and it defines the Hilbert space L^2_F of square integrable functions with respect to F.
We consider a model of the type

$$Y = \varphi(Z) + U, \qquad \mathbb{E}(U|Z) \neq 0. \tag{1}$$

This model is a *structural model* in the sense that it is directly proposed by the economic theory; moreover it is characterized by the fact that the intervening variables Y and Z are endogenous. We could interpret Y and Z as having been simultaneously determined by the relations given in the model. The lack of any further characterization of φ or any constraint on it, except regularity requirements, that will be explicit below, make the model the most general as possible.
In order to be able to estimate the instrumental regression φ, we suppose that a vector of instruments W, such that $\mathbb{E}(U|W) = 0$, is available. This is the *instrumental variables approach* that characterizes the structural model by the relation

$$\mathbb{E}(Y|W) = \mathbb{E}(\varphi(Z)|W) \tag{2}$$

and assumes that there exists an unique element φ_* satisfying this equality. The only requirement we make on the true φ having generated the data according to (1) is that it belongs to $L^2_F(Z)$, where $L^2_F(Z) \subset L^2_F$ is the subset of square integrable functions of Z.
Uniqueness of the solution of (2) ensures identifiability of the parameter of interest φ by the moment condition (2) and, using terminology of functional analysis, it is equivalent to assume that the conditional expectation operator is one-to-one (or equivalently that its kernel is reduced to be zero).
Furthermore, a solution to equation (2) exists if and only if the regression function $\mathbb{E}(Y|W)$ belongs to the range of the conditional expectation operator $\mathbb{E}(\cdot|W) : L^2_F(Z) \to L^2_F(W)$, where

$L^2_F(W) \subset L^2_F$ denotes the space of square integrable functions of W, integrable with respect to F and notation $\mathcal{R}(\cdot)$ will be reserved to denote the range of an operator. Non existence of this solution characterizes the so-called problem of *overidentification*. Henceforth, overidentified solutions come from equations with an operator that is not surjective and non identified solutions, as we have already stressed, come from equations with an operator that is not one-to-one. Indeed, properties ensuring existence and uniqueness of the solution are properties of the *cdf* F of S. Our assumption will be that a unique solution φ to functional equation 2 exists. We summarize in the next assumption both the assumptions allowing for ruling out under- and over-identified solutions.

Assumption 1 *(i) The operator* $\mathbb{E}(\cdot|W) : L^2_F(Z) \to L^2_F(W)$ *characterized by the true cdf* F *is one-to-one;*

(ii) the true cdf F *is such that* $\mathbb{E}(Y|W) \in \mathcal{R}(\mathbb{E}(\cdot|W))$.

In reality, the *cdf* F is unknown, at least partially (for instance, we can know $F(\cdot, Z, W)$, but not $F(Y, \cdot, \cdot)$) and we have to replace it by an estimator $\hat{F}$ of finite rank in $\mathbb{E}(\cdot|W)$. It results that Assumption 1 cannot be satisfied by $\hat{F}$.
However, problems of over and under identification are only marginal in this paper since we are moved by a Bayesian philosophy, in the sense that we are looking for the posterior distribution of φ given the noisy measurement $\hat{\mathbb{E}}(Y|W)$ and not for the exact solution.
A classical procedure in models with endogenous variables consists in transforming the structural model provided by economic theory in a *reduced form* model that is tractable from an estimation point of view. This means that the model is solved for the endogenous variables in function of exogenous variables and random noise. Then, the reduced form corresponding to (1) is

$$Y = \mathbb{E}(Y|W) + \varepsilon, \qquad \mathbb{E}(\varepsilon|W) = 0$$

or equivalently, by using the relation characterizing the instrumental variables:

$$Y = \mathbb{E}(\varphi(Z)|W) + \varepsilon, \qquad \mathbb{E}(\varepsilon|W) = 0. \tag{3}$$

The reduced form will be used as sampling model for inference. It should be noted that it is a conditional model, conditional on W, and that it does not depend on Z. This is a consequence of the fact that the instrumental variable approach specifies a statistical model concerning (Y, W), but not concerning the whole vector (Y, Z, W) since the only information available is that $\mathbb{E}(U|W) = 0$ and nothing is specified about $\mathbb{E}(U|Z)$ except that it is different than 0. An alternative approach to endogeneity is the *control function approach* proposed by [22]; with this approach we could specify a Bayesian experiment concerning the whole vector (Y, Z, W), anyway, we do not consider this approach here.
We will denote with small letters realizations of random variables: $s_i = (y_i, z_i, w_i)$ is the i-th observation on the random vector S. Boldface letters $\mathbf{z}$ and $\mathbf{w}$ will denote the matrix of observations on vectors Z and W, respectively; y will be the vector of observations on Y. We assume to observe a sample of S:

Assumption 2 $s_i = (y_i, z_i, w_i)$, $i = 1, \ldots, n$ *is an i.i.d. sample of observations on* $S = (Y, Z, W)$.

Each observation satisfies the reduced form model: $y_i = \mathbb{E}(\varphi(Z)|w_i) + \varepsilon_i$ with $\mathbb{E}(\varepsilon_i|\mathbf{w}) = 0$, for $i = 1, \ldots, n$. We rewrite it in matrix form:

$$y_{(n)} = K_{(n)}\varphi + \varepsilon_{(n)}, \tag{4}$$

where

$$y_{(n)} = \frac{1}{\sqrt{n}} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \varepsilon_{(n)} = \frac{1}{\sqrt{n}} \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

$$\forall \phi \in L_F^2(Z), \quad K_{(n)}\phi = \frac{1}{\sqrt{n}} \begin{pmatrix} \mathbb{E}(\phi(Z)|W = w_1) \\ \vdots \\ \mathbb{E}(\phi(Z)|W = w_n) \end{pmatrix}, \quad K_{(n)} : L_F^2(Z) \to \mathbb{R}^n$$

$$\text{and} \quad \forall x \in \mathbb{R}^n, \qquad K_{(n)}^*(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \frac{f(Z,w_i)}{f(Z,\cdot)f(\cdot,w_i)}, \qquad K_{(n)}^* : \mathbb{R}^n \to L_F^2(Z).$$

Operator $K_{(n)}^*$ is the adjoint of $K_{(n)}$, as it can be easily checked by solving the equation $< K_{(n)}\phi, x >=< \phi, K_{(n)}^* x > \forall x \in \mathbb{R}$ and $\phi \in L_F^2(Z)$. By analogy with this notation we denote $K = \mathbb{E}(\cdot|W)$ the operator from $L_F^2(Z)$ in $L_F^2(W)$ and with K^* its adjoint: $K^* = \mathbb{E}(\cdot|W) : L_F^2(W) \to L_F^2(Z)$.
It should be noted that $K_{(n)}$ and $K_{(n)}^*$ are finite rank operators, so that they have only n singular values different than zero.
To keep things easy we make a distributional assumption for ε_i,

Assumption 3 *The error terms of the reduced form model are independent and identically distributed gaussian, conditionally on* $(w_1, \ldots, w_n)$*:* $\varepsilon_i|\boldsymbol{w} \sim$ *i.i.d.* $\mathcal{N}(0, \sigma^2)$.

As a consequence $\varepsilon_{(n)}|\mathbf{w} \sim \mathcal{N}(0, \frac{\sigma^2}{n} I)$. We limit to consider the homoskedastic case.

3 Bayesian Analysis

In this section we develop and analyze the Bayesian experiment associated to the reduced form model (3) and we consider a sampling of it. Elements of vector $y_{(n)}$ represent n independent, but not identically distributed, draws from a sampling probability P^{σ,φ,w_i} conditional on $W = w_i$ [1]. The product sample space will be denoted by $\mathcal{Y} = \mathbb{R}^N$ and its associated Borel σ-field by $\mathcal{F}_Y$. We shall denote with $P^{\sigma,\varphi,\mathbf{W}}$ the sampling measure on $\mathcal{F}$ associated to the whole vector $y_{(n)}$ and conditional on the structural parameter φ on the variance parameter σ^2 and on the vector of instruments $\mathbf{w}$. There are two parameters in the model: the nuisance variance parameter σ^2 and the instrumental regression φ that represents the parameter of interest. We use the notation $\mathcal{B}$ for the σ-field associated to $\mathbb{R}_+$ and ν for the prior probability defined on it, then $\sigma^2 \in (\mathbb{R}_+, \mathcal{B}, \nu)$. The parameter of interest $\varphi(Z)$ has been constrained only to be square integrable with respect to F, implying that the parameter space is $L_F^2(Z)$ with the associated σ-field $\mathcal{E}$ and the prior distribution μ^σ, conditional on σ^2.
There exists two possible specifications of the probability measure on the joint parameter space $\mathbb{R}_+ \times L_F^2(Z)$. The traditional approach calls for a conjugate model with a joint distributions on the parameter space that is separable in a marginal on $\mathbb{R}_+$ and a conditional μ^σ, given $\mathcal{B}$, on $L_F^2(Z)$. Otherwise, new developments in Bayesian literature propose more and more models in which the prior distribution on the parameter space is the product of two marginal independent distributions, in this case $\mu^\sigma = \mu$ since it does not depend on the variance parameter. Inference analysis changes in the two cases; we start by treating the conjugate model and we present the independent case in the following section.
The conjugate bayesian experiment associated to model (4) is summarized as

$$\Xi = (\mathbb{R}_+ \times L_F^2(Z) \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{E} \otimes \mathcal{F}, \Pi^{\mathbf{W}} = \nu \times \mu^\sigma \times P^{\sigma,\varphi,\mathbf{W}}),$$

where $\Pi^{\mathbf{W}}$ is the conditional joint measure on the product space conditional on $\mathbf{w}$. Bayesian inference consists in finding the inverse decomposition of $\Pi^{\mathbf{W}}$ in the product of the posterior distribution $\nu^{\mathcal{F},\mathbf{W}} \times \mu^{\sigma,\mathcal{F},\mathbf{W}}$ and the predictive measure $P^{\mathbf{W}}$. In the following, we shall lighten notation by simply writing $\nu^{\mathcal{F}}$ for $\nu^{\mathcal{F},\mathbf{W}}$ and $\mu^{\sigma,\mathcal{F}}$ to denote $\mu^{\sigma,\mathcal{F},\mathbf{W}}$.
We assume that the prior ν is an *Inverse Gamma* distribution with known parameters ν_0 and s_0^2. The distribution μ^σ, conditional on σ^2, is a Gaussian measure on $L_F^2(Z)$ defining a mean element $\varphi_0 \in L_F^2(Z)$ and a covariance operator $\sigma^2\Omega_0 : L_F^2(Z) \to L_F^2(Z)$. μ^σ is such that $\mathbb{E}(||\phi||^2) < \infty$, $\forall \phi \in L_F^2(Z)$. Moreover, Ω_0 results to be a trace-class operator and this guarantees that realizations

[1] Notation P^{σ,φ,w_i} means the conditional probability $\mathbb{P}(\frac{1}{\sqrt{n}} y_i | \sigma^2, \varphi, W = w_i)$.

of this process will be in $L^2_F(Z)$ with probability 1. This specification for the prior measure is suitable in the sense that its support is the closure of the *Reproducing Kernel Hilbert Space* associated to Ω_0, ($\mathcal{H}(\Omega_0)$ in the following), that is dense in $L^2_F(Z)$ if Ω_0 is one to one. So, we assume Ω_0 is injective in such a way the support of μ^σ will be all $L^2_F(Z)$.
The sampling probability $P^{\sigma,\varphi,\mathbf{W}}$ is gaussian with mean $K_{(n)}\varphi$ and covariance operator $\frac{\sigma^2}{n}I_n$, where I_n is the identity matrix of order n. The marginal $P^{\sigma,\mathbf{W}}$, marginalized with respect to μ^σ, is still gaussian with mean $K_{(n)}\varphi_0$ and with covariance matrix $\sigma^2 C_n = \sigma^2(K_{(n)}\Omega_0 K^*_{(n)} + \frac{1}{n}I_n)$ that is positive-definite and of full rank n.
Both measures $P^{\sigma,\varphi,\mathbf{W}}$ and $P^{\sigma,\mathbf{W}}$ depend on the density $f(Z,W)$ through operators $K_{(n)}$ and $K^*_{(n)}$, so that to be precise we should index these two probabilities with f, $P^{f,\sigma,\varphi,\mathbf{W}}$ and $P^{f,\sigma,\mathbf{W}}$. However, we are assuming in this section that the joint density $f(Z,W)$ is known and then this index is only a notational matter since it does not affect in any way the estimate for φ. We omit f for all this section and we shall reintroduce it in Section 4. Summarizing, we have

$$\begin{aligned} \sigma^2 &\sim \mathcal{I}\Gamma(\nu_0, s_0^2) \\ \begin{pmatrix} \varphi \\ y_{(n)} \end{pmatrix}\Big|\sigma^2 &\sim \mathcal{GP}\left(\begin{pmatrix} \varphi_0 \\ K_{(n)}\varphi_0 \end{pmatrix}, \sigma^2 \begin{pmatrix} \Omega_0 & \Omega_0 K^*_{(n)} \\ K^*_{(n)}\Omega_0 & \frac{1}{n}I + K_{(n)}\Omega_0 K^*_{(n)} \end{pmatrix}\right), \end{aligned}$$

so that $(\varphi, y_{(n)})$ are jointly a Gaussian process conditionally on σ^2.
The main theoretical question concerns the existence of conditional gaussian processes in Hilbert spaces, namely we are interested in the existence of a conditional distribution on $L^2_F(Z)$, given information in $\mathcal{F}$. *Jirina theorem*, see [20], allows to confirm the existence of a regular version of the posterior distribution, namely there exists a transition probability characterizing the conditional distribution on $\mathcal{E}$ given $\mathcal{F}$. Condition required by this theorem that the product space $(L^2_F(Z)\times\mathbb{R}^N)$ is a Polish space is verified, see [10]. Anyway, existence of a regular posterior distribution which is a gaussian process, was already stated and proved, for instance, in [10], [11] and [19]. We remark again that all posterior probability have to be meant computed for a given w. Proof that the posterior measure is gaussian follows from the form assumed by the characteristic function of φ given $y_{(n)}$, moreover the conditional expectation of φ, given $(y_{(n)}, \sigma^2)$ exists since $|\varphi|^2$ is integrable and it is an affine transformation of $y_{(n)}$. We characterize in the following theorem the conditional posterior distribution of φ, given σ^2.

Theorem 1 *Let φ and y be two jointly distributed gaussian random elements in $L^2_F(Z)$ and $\mathbb{R}^N$, respectively. Then, the conditional distribution of φ given y and σ^2 is gaussian with mean $Ay_{(n)}+b$, where*

$$A = \Omega_0 K^*_{(n)} C_n^{-1}, \qquad b = (I - AK_{(n)})\varphi_0 \tag{5}$$

and covariance given by

$$\Omega_y = \sigma^2(\Omega_0 - AK_{(n)}\Omega_0).$$

This is a very classical result and we refer to [10] or [19] for a proof of it. Then $\mathbb{E}(\varphi|y_{(n)}, \sigma^2) = \varphi_0 + \Omega_0 K^*_{(n)} C_n^{-1}(y_{(n)} - K_{(n)}\varphi_0)$, if $(y_{(n)} - K_{(n)}\varphi_0) \in \mathcal{R}(C_n)$ that is always satisfied in finite dimension. The variance parameter σ^2 affects the posterior of φ only through the posterior covariance operator, so that $\mathbb{E}(\varphi|y_{(n)}, \sigma^2) = \mathbb{E}(\varphi|y_{(n)})$.
Despite its well definition, the posterior distribution $\mu^{\sigma,\mathcal{F}}$ is not consistent in the classical sense. The pair $(\varphi, \mu^{\sigma,\mathcal{F}})$ is consistent if for $P^{\sigma,\varphi,\mathbf{W}}$-almost all sequences $y_{(n)}$, the posterior $\mu^{\sigma,\mathcal{F}}$ converges weakly to point mass at φ. Moreover, $\mu^{\sigma,\mathcal{F}}$ is consistent in the sampling theory sense (or in the classical sense) if $(\varphi, \mu^{\sigma,\mathcal{F}})$ is consistent for all φ. This concept of *frequentist consistency* is extensively developed in [3], among others, where Bayesians are separated into two groups: "classical" and "subjectivist". Classical bayesians believes there exists a true value of the parameter that has generated the data, therefore they care for, as data set becomes large, the posterior converging to a point mass at the true parameter. In point of fact, consistency is interesting also for subjective Bayesian for different reasons (e.g. "intersubjective agreement" or to check if the posterior is a correct representation of the updated prior, see [3] and [9]).

On the basis of this argument we are persuaded about the importance to have a consistent posterior distribution. The following lemma states the non consistency of the posterior $\mu^{\sigma,\mathcal{F}}$.

Lemma 1 *Let φ_* be the true value of the parameter having characterized the data generating process $P^{\sigma,\varphi_*,\boldsymbol{w}}$. The pair $(\varphi_*, \mu^{\sigma,\mathcal{F}})$ is inconsistent, i.e. $\mu^{\sigma,\mathcal{F}}$ does not weakly converge to point mass δ_{φ_*} in φ_*.*

Proof: See Appendix 7.1. ∎

Problem of inconsistency is due to operator A and to non continuity of the inverse of C_n as $n \to \infty$. In fact, as long as n stays small C_n is an invertible $n \times n$ matrix since its n eigenvalues are all different than zero; however, when n becomes large, even if $(\frac{1}{n} I_n + K_{(n)}\Omega_0 K^*_{(n)})$ looks like a Ridge regularization, $\frac{1}{n}$ goes to 0 too fast to control the ill-posedness. Actually, the number of eigenvalues of C_n grows with n up to form a decreasing sequence having 0 as the only accumulating point. Therefore, $\lim C_n$ is no more bounded. Furthermore, contrarily to finite dimensional cases, where the Ridge regression has a Bayesian interpretation, the prior specification does not solve the problem of ill-posedness here because of compacity of Ω_0.
A natural solution seems to consist in translating the eigenvalues of C_n far from 0 by a factor $\alpha_n > 0$ and such that $\alpha_n \to 0$ with n. Indeed, this is as to say that we apply a Tikhonov regularization scheme to the inverse of C_n with regularization parameter α_n: $C^{-1}_{n,\alpha} = (\alpha_n I_n + K_{(n)}\Omega_0 K^*_{(n)} + \frac{1}{n} I_n)^{-1}$. If $\alpha_n \to 0$ at a suitable rate, this operator stays well defined asymptotically. We call *Regularized Posterior Distribution*, denoted with $\mu^{\sigma,\mathcal{F}}_\alpha$, the posterior distribution in which operator A has been substituted by the regularized operator $A_\alpha = \Omega_0 K^*_{(n)} C^{-1}_{n,\alpha}$. This object has been introduced in [10] and defined as the Bayesian solution to a functional equation in Hilbert spaces. The instrumental variables model we are treating describes an equation in finite dimensional spaces, but the parameter of interest is of infinite dimension, so that the reduced form model can be seen as a projection of it on a space of small dimension. As it has already been stressed, even if the problem we are considering is substantially different with respect to that one considered in [10], asymptotic arguments motivates us to adopt the regularized posterior distribution $\mu^{\sigma,\mathcal{F}}_\alpha$ as solution for our inference problem. In particular, for n big the same problems specific to Bayesian analysis of functional equation reappears in our case. On the other side, if we wanted to solve (4) in a classical way, we would realize that some regularization scheme would be necessary also in the finite sample case since $\hat{\varphi} = (K^*_{(n)}K_{(n)})^{-1}K^*_{(n)}y_{(n)}$, but $K^*_{(n)}K_{(n)}$ is not full rank and than non invertible.
The regularized posterior distribution $\mu^{\sigma,\mathcal{F}}_\alpha$ is a gaussian measure defining a mean element $\hat{\varphi}_\alpha = A_\alpha y_{(n)} + b_\alpha$ and a covariance operator $\Omega_{y,\alpha}$, where b_α and $\Omega_{y,\alpha}$ are obtained by substituting A_α to operator A in b and Ω_y respectively. We will take the regularized posterior mean as punctual estimator for the instrumental regression, as suggested for a quadratic loss function. In section 3.2 we will state consistency of this solution.

3.1 The Student t Process

Let consider the posterior distribution of σ^2. Since we have a conjugate model it is possible to integrate out φ in the sampling probability $P^{\sigma,\varphi,\mathbf{W}}$ of $y_{(n)}$ to obtain $P^{\sigma,\mathbf{W}}$ and then to use the two probabilities

$$\begin{aligned} \sigma^2 &\sim \mathcal{IT}(\nu_0, s_0^2) \\ y_{(n)}|\sigma^2 &\sim \mathcal{N}(K_{(n)}\varphi_0, \sigma^2\Big(\frac{1}{n} I_n + K_{(n)}\Omega_0 K^*_{(n)}\Big)) \end{aligned}$$

to make inference on σ^2. The posterior distribution has the kernel:

$$\nu^{\mathcal{F}} \propto \Big(\frac{1}{\sigma^2}\Big)^{\nu_0/2+n/2+1} \exp\{-\frac{1}{2\sigma^2}[(y_{(n)} - K_{(n)}\varphi_0)'\Big(\frac{1}{n} I_n + K_{(n)}\Omega_0 K^*_{(n)}\Big)^{-1}(y_{(n)} - K_{(n)}\varphi_0) + s_0^2]\}$$

that identifies an $\mathcal{IG}$ distribution [2]. Then

$$\begin{aligned}\sigma^2|y_{(n)} &\sim \mathcal{IG}(\nu_*, s_*^2), \quad \text{with}\\ \nu_* &= \nu + n\\ s_*^2 &= s_0^2 + (y_{(n)} - K_{(n)}\varphi_0)'(\frac{1}{n}I_n + K_{(n)}\Omega_0 K_{(n)}^*)^{-1}(y_{(n)} - K_{(n)}\varphi_0).\end{aligned}$$

It results that the posterior of σ^2 does not depend on φ. This fact is exploited for computing the marginal posterior distribution of φ by directly integrating out σ^2.
Analogy with the finite dimensional case, where integration of a gaussian density with respect to an Inverse Gamma gives a *Student t* distribution, suggests that we should find a similar result in infinite dimension: $\varphi|y_{(n)}$ is a *Student t Process* in $L_F^2(Z)$. We define a *Student t Process* in general Hilbert spaces, through scalar product in it, as:

Definition 1 *Let $\mathcal{X}$ be an Hilbert space with inner product $< \cdot, \cdot >_{\mathcal{X}}$ and $x \in \mathcal{X}$. x is a* Student t Process *with parameters $x_0 \in \mathcal{X}$, $\Omega_0 : \mathcal{X} \to \mathcal{X}$ and $\nu \in \mathbb{R}_+$, denoted $x \sim \mathcal{StP}(x_0, \Omega_0, \nu)$, if and only if $\forall \delta \in \mathcal{X}$,*

$$< x, \delta >_{\mathcal{X}} \sim\ t(< x_0, \delta >_{\mathcal{X}}, < \Omega_0\delta, \delta >_{\mathcal{X}}, \nu),$$

i.e. $< x, \delta >_{\mathcal{X}}$ has a density proportional to

$$\Big[\nu + \frac{(< x, \delta >_{\mathcal{X}} - < x_0, \delta >_{\mathcal{X}})^2}{< \Omega_0\delta, \delta >_{\mathcal{X}}}\Big]^{-\frac{\nu+1}{2}},$$

with mean and variance

$$\begin{aligned}\mathbb{E}(< x, \delta >_{\mathcal{X}}) &= < x_0, \delta >_{\mathcal{X}}, \qquad \text{if } \nu > 1\\ Var(< x, \delta >_{\mathcal{X}}) &= \frac{\nu}{\nu - 2} < \Omega_0\delta, \delta >_{\mathcal{X}}, \qquad \text{if } \nu > 2.\end{aligned}$$

We admit the following Lemma, concerning the marginalization of a Gaussian Process with respect to a variable distributed as an *Inverse Gamma*.

Lemma 2 *Let $\sigma^2 \sim \mathcal{IG}(\nu, s^2)$ and $x|\sigma^2 \sim \mathcal{GP}(x_0, \sigma^2\Omega_0)$. Then,*

$$x\ \sim\ \mathcal{StP}\Big(x_0, \frac{s^2}{\nu}\Omega_0, \nu\Big).$$

Proof of this lemma is trivial and follows immediately if we consider the scalar product $< x, \delta >$, $\forall \delta \in \mathcal{X}$, so that it has a normal distribution on $\mathbb{R}$.
We apply this result to our process φ. Hence,

$$\varphi|y_{(n)} \quad \sim \quad \mathcal{StP}(\nu_*, \hat{\varphi}_\alpha, \frac{s_*^2}{\nu_*}\Omega_{y,\alpha}),$$

with marginal moments $\hat{\varphi}_\alpha$ and $\frac{s_*^2}{\nu_* - 2}\Omega_0$.

[2] There exist different specifications of the *Inverse Gamma* distribution; we use in our study an $\mathcal{IG}(\nu_0, s_0^2)$ with density: $f(\sigma^2) \propto \Big(\frac{1}{\sigma^2}\Big)^{\nu_0/2+1} \exp\Big\{-\frac{1}{2}\frac{s_0^2}{\sigma^2}\Big\}$. The corresponding mean and variance are $\mathbb{E}(\sigma^2) = \frac{s_0^2/2}{\nu_0/2-1} = \frac{s_0^2}{\nu_0-2}$ and $Var(\sigma^2) = \frac{s_0^4/4}{(\nu_0/2-1)^2(\nu_0/2-2)}$, respectively.

3.2 Asymptotic Analysis

We focus, in this section, on asymptotic frequentist properties of the posterior distributions of σ^2 and φ. As it has already been pointed out, our study can be classified among classical bayesian studies in the sense that we believe in the existence of a true value for the parameters having generated the data. This fact gives more generality to our analysis since the bayesian philosophy moving it is less binding.
First of all, we state in the following theorem the *posterior consistency* of the regularized conditional posterior probability $\mu_\alpha^{\sigma,\mathcal{F}}$ of φ to contrast the negative result of inconsistency given in Lemma 1. We shall denote with $\mathcal{H}(\Omega_0)$ the *Reproducing Kernel Hilbert Space* associated to Ω_0.

Theorem 2 *Let φ_* be the true value of the parameter and $\mu_\alpha^{\sigma,\mathcal{F}}$ a gaussian measure on $L_F^2(Z)$ with mean $A_\alpha y_{(n)} + b_\alpha$ and covariance operator $\Omega_{y,\alpha}$. If $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$ and if $\alpha_n \to 0$, $\alpha_n^2 n \to \infty$, then:*

(i) $\mu_\alpha^{\sigma,\mathcal{F}}$ weakly converges to point mass δ_{φ_*} in φ_*;

(ii) if moreover $\Omega_0^{-\frac{1}{2}}(\varphi_* - \varphi_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$, then

$$\mu_\alpha^{\sigma,\mathcal{F}}\{\varphi : ||\varphi - \varphi_*|| \geq \epsilon_n\} \sim \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta + \frac{1}{\alpha_n^2 n}).$$

It should be noted that the condition for the second part of the theorem is only a regularity condition necessary to have convergence at a certain speed. The condition that really matters is the fact that the centered true parameter must belong to the *Reproducing Kernel Hilbert Space* associated to Ω_0, namely $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$. Actually, the gaussian prior mean μ is not able to generate trajectories in this space but, since this space is dense in $L_F^2(Z)$ and since the support of μ is the closure $\overline{\mathcal{H}(\Omega_0)}$, this measure is able to generate trajectories as close as possible to the true one. This concept is known in literature as *prior inconsistency* and it refers to a prior that is unable to generate the true parameter characterizing the data generating process. This problem is present only for infinite dimensional parameter sets since it is difficult to be sure about a prior on an infinite dimensional parameter space and so it can happen that the true value of the parameter is not in the support of the prior, see e.g. [12] or [15].
A corollary provides the necessary results for having Theorem 2, it concerns consistency of the regularized posterior mean and convergence to zero of the regularized posterior variance.

Corollary 1 *Under the same assumptions of Theorem 2,*

(i) $||\hat{\varphi}_\alpha - \varphi_*|| \to 0$ in P^{σ_*,φ_*}-probability and if $\Omega_0^{-\frac{1}{2}}(\varphi_* - \varphi_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$,

$$||\hat{\varphi}_\alpha - \varphi_*||^2 \sim \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta + \frac{1}{\alpha_n^2 n}).$$

(ii) $||\Omega_{y,\alpha}|| \to 0$ in P^{σ_*,φ_*}-probability and $\forall \phi \in L_F^2(Z)$ such that $\Omega_0^{\frac{1}{2}}\phi \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$,

$$||\Omega_{y,\alpha}\phi||^2 \sim \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta).$$

The rates governing the bias are the first and the third one, being the second one the product of the two. While the first rate α_n^β requires a regularization parameter α_n going to zero as fast as possible, the third rate requires an α_n going to zero as slow as possible. Hence, the optimal rate for α_n will be obtained when the two rates are equated: $\alpha_n^\beta = \frac{1}{\alpha_n^2 n}$. This gives an optimal regularization parameter proportional to

$$\alpha_n \propto n^{-\frac{1}{\beta+2}}$$

and a global rate of convergence of the regularized posterior mean proportional to $n^{-\frac{\beta}{\beta+2}}$ that is the fastest one.
Now we concentrate on the posterior consistency of $\nu^{\mathcal{F}}$. We define $g(Z, w_i)$ as the transformation of the kernel of $K_{(n)}^*$ by operator $\Omega_0^{\frac{1}{2}}$, *i.e.* if $\omega_0(s, Z)$ denotes the kernel of $\Omega_0^{\frac{1}{2}}$, $g(Z, w_i) = \int \omega_0(s, Z)\frac{f(s,w_i)}{f(s)f(w_i)}f(s)ds$. In particular, we have $\Omega_0^{\frac{1}{2}} K_{(n)}^* \varepsilon_{(n)} = \frac{1}{n}\sum_i \varepsilon_i g(Z, w_i)$.

Theorem 3 *Let σ_*^2 be the true value of σ^2 having generated the data and ν_* an Inverse Gamma distribution on $\mathbb{R}_+$. If $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$ and if there exists a $\gamma > 1$ such that $\forall\, w\ g(Z,w) \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\gamma}{2}}$, then:*

$$\sqrt{n^{\beta-1}}(\mathbb{E}(\sigma^2|y_{(n)}) - \sigma_*^2) \sim \mathcal{O}_p(1).$$

It follows that $\nu^{\mathcal{F}}\{\sigma^2 : |\sigma^2 - \sigma_*^2| \geq \varepsilon_n\} \to \delta_{\sigma_*^2}$, where $\delta_{\sigma_*^2}$ is the point mass in '$sigma_*^2$.

We conclude this section by giving a result of joint posterior consistency, that is the joint measure $\nu^{\mathcal{F}} \times \mu^{\sigma,\mathcal{F}}$ degenerate towards a Dirac measure in (σ_*^2, φ_*).

Lemma 3 *Under condition of Theorems 2 and 3, the joint measure*

$$\nu^{\mathcal{F}} \times \mu^{\sigma,\mathcal{F}}\{(\sigma^2,\varphi) \in \mathbb{R}_+ \times L_F^2(Z); ||(\sigma^2,\varphi) - (\sigma_*^2,\varphi_*)||_{\mathbb{R}_+ \times L_F^2} \geq \eta_n\}$$

converges to zero in P^{σ_*,φ_*}-probability.

3.3 Independent Prior

We adopt in this section an alternative specification of the joint measure on the parameter space that consists in independent prior distributions. The joint measure on $\mathbb{R}_+ \times L_F^2(Z)$ is equal to the product of the two marginal ν and μ. Actually, only the prior measure of the parameter of interest φ is affected since its variance no more depends on σ^2. We have the following distributions on $\mathbb{R}_+ \times L_F^2(Z) \times \mathcal{Y}$:

$$\begin{aligned} \sigma^2 &\sim \mathcal{IG}(\nu_0, s_0^2) \\ \varphi &\sim \mathcal{GP}(\varphi_0, \Omega_0) \\ y_{(n)}|\varphi,\sigma^2 &\sim \mathcal{N}_n(K_{(n)}\varphi, \frac{\sigma^2}{n} I_n). \end{aligned}$$

In this case it is not allowed to integrate out φ from the sampling distribution of $y_{(n)}$ since we do not have a conditional measure for φ given σ^2. This particular structure of the problem makes computation of the posterior distribution of φ, marginalized with respect to σ^2, not possible, or at least not trivially feasible. Anyway, we are able to compute the posterior distribution of φ conditional on σ^2, $\mu_\alpha^{\sigma,\mathcal{F}}$ and the posterior distribution of σ^2 conditional on φ, $\nu_\alpha^{\varphi,\mathcal{F}}$. Then, a Gibbs sampling algorithm will allow, for a large number of iterations, to get a good approximation of the stationary laws represented by the desired regularized marginal posterior distributions $\mu_\alpha^{\mathcal{F}}$ and $\nu_\alpha^{\mathcal{F}}$. We proceed, in the following of this subsection, in the computation of the two conditional posterior distributions for φ and σ^2. To compute $\nu_\alpha^{\varphi,\mathcal{F}}$ we use the homoskedastic model for error terms in the reduced form (3) $\varepsilon_{(n)}|\sigma^2$ *i.i.d.* $\sim \mathcal{N}(0, \frac{\sigma^2}{n} I_n)$, with $\varepsilon_{(n)} = y_{(n)} - K_{(n)}\varphi$. Trivial computations provide us with the conditional posterior

$$\begin{aligned} \nu_\alpha^{\varphi,\mathcal{F}} &\sim \mathcal{IG}(\nu_*, s_*^2) \\ \nu_* = \nu_0 + n, \qquad & s_*^2 = s_0^2 + \sum_{i=1}^n (y_i - \mathbb{E}(\varphi|w_i)). \end{aligned}$$

The parameter φ enters the conditional posterior distribution $\nu_\alpha^{\varphi,\mathcal{F}}$ through the error term ε_i since the latter is observable only when we know φ: $\varepsilon_i = y_i - \mathbb{E}(\varphi|w_i)$. This justifies the regularized parameter α in the notation $\varepsilon_{i,\alpha}$; it means that after having drawn φ from $\mu_\alpha^{\sigma,\mathcal{F}}$ we have computed the correspondent error term. It is then clear that the regularization parameter plays a role, even if indirect, on $\nu_\alpha^{\varphi,\mathcal{F}}$.
The conditional distribution of $y_{(n)}$ conditional on σ^2 is $\mathcal{N}(K_{(n)}\varphi_0, (\frac{\sigma^2}{n} I_n + K_{(n)}\Omega_0 K_{(n)}^*))$. Hence, we get $\mu_\alpha^{\sigma,\mathcal{F}}$:

$$
\begin{aligned}
\varphi|y_{(n)},\sigma^2 &\sim \mathcal{GP}(A_\alpha^\sigma y + b^\sigma, \Omega_y^\sigma)\\
A_\alpha^\sigma &= \Omega_0 K_{(n)}^*(\alpha_n I_n + K_{(n)}\Omega_0 K_{(n)}^* + \frac{\sigma^2}{n} I_n)^{-1}\\
b^\sigma &= (I_n - A_\alpha^\sigma K_{(n)})\varphi_0\\
\Omega_y^\sigma &= \Omega_0 - \Omega_0 K_{(n)}^*(\alpha_n I_n + K_{(n)}\Omega_0 K_{(n)}^* + \frac{\sigma^2}{n} I_n)^{-1} K_{(n)}\Omega_0).
\end{aligned}
$$

The associated Gibbs sampling algorithm is the following:

(i) fix an initial value for σ^2: $\sigma^2_{(0)}$;

(ii) draw $\varphi^{(i)}$ from $\mu_\alpha(\varphi|\mathcal{F}, \sigma^2_{(i-1)})$;

(iii) draw $\sigma^2_{(i)}$ from $\nu_\alpha(\sigma^2|\mathcal{F}, \varphi^{(i)})$;

(iv) iterate *(ii)* - *(iii)* for $i = 1, \ldots, 2J$;

(v) discard the first J values and use the other ones to estimate the posterior distributions $\mu_\alpha^{\mathcal{F}}$ and $\nu_\alpha^{\mathcal{F}}$.

Implementation of this algorithm requires to determine two elements: the starting value $\sigma^2_{(0)}$ and the number of iterations J necessary to get the stationary distribution. We propose to draw the starting value $\sigma^2_{(0)}$ from a Γ distribution with parameters chosen in such a way that some feature of the sample are reproduced. First, we estimate σ^2 through a nonparametric estimation of ε_i: $\hat{\varepsilon}_i = y_i - \hat{\mathbb{E}}(y|w_i)$. For instance, $\hat{\mathbb{E}}(y|w_i)$ is obtained by using a kernel smoothing estimator. Therefore, $\hat{\sigma}^2 = \widehat{Var}(\hat{\varepsilon}_i)$ and we set the first theoretical moment of σ^2 equal to $\hat{\sigma}^2$. Since $\sigma^2 \sim \mathcal{I}\Gamma(\nu_0, s_0^2)$ $\mathbb{E}(\sigma^2) = \frac{s_0^2}{\nu_0 - 2}$ and then $\tilde{s}_0^2 = \hat{\sigma}^2(\nu_0 - 2)$. Lastly, ν_0 will be fixed such that the degree of freedom associated to the distribution will be smaller than the sample size, *i.e.* $\tilde{\nu}_0 < n$, in order to make the distribution more dispersed. At the end, we draw the starting value $\sigma^{2(0)}$ from $\mathcal{I}\Gamma(\tilde{\nu}_0, \tilde{s}_0^2)$.
In order to determine the number of iterations J we use a method that is an adaptation of the technique proposed in [13]. This strategy consists in using several independent sequences, with starting points sampled from an overdispersed distribution, and in analyzing the multiple sequences by computing estimates of the target distribution to see how close the simulation process is to convergence.
We simulate M independent sequences, each one of length $2J$, with different starting points drawn from $\mathcal{I}\Gamma(\tilde{\nu}_0, \tilde{s}_0^2)$ as described above:

$$
\begin{aligned}
\varphi_{ij} &\sim \mu_\alpha^{\sigma,\mathcal{F}}, \quad i = 1, \ldots, M;\ j = 1, \ldots, 2J\\
\eta_{ij} &\sim \nu_\alpha^{\varphi,\mathcal{F}}, \quad i = 1, \ldots, M;\ j = 1, \ldots, 2J.
\end{aligned}
$$

The target distribution of each parameter can be estimated in two ways. First, a distributional estimate is formed by using between-sequence and within-sequence information; this is more variable than the target distribution, because of the use of overdispersed starting values. Second, a pooled within-sequence estimate is formed and used to monitor the convergence of the simulation process. In principle, when the simulations are far from convergence, the individual sequences will be less variable than the target distribution, but as the individual sequences converge to the target distribution, the variability within each sequence will grow to be as large as the variability of the target distribution.
In practice, the first J iterations of each sequence are discarded and the last J are used to compute the following quantities:

$$
\begin{aligned}
B &= \frac{J}{M-1}\sum_{i=1}^{M}(\overline{\sigma^2_{i.}} - \overline{\sigma^2_{..}})^2, \quad \overline{\sigma^2_{i.}} = \frac{1}{J}\sum_{j=1}^{J}\sigma^2_{ij}, \quad \overline{\sigma^2_{..}} = \frac{1}{M}\sum_{i=1}^{M}\overline{\sigma^2_{i.}}\\
WW &= \frac{1}{M}\sum_{i=1}^{M} s_i^2, \quad s_i^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\sigma^2_{ij} - \overline{\sigma^2_{i.}})^2\\
\widehat{Var}(\sigma^2) &= \frac{J-1}{J}WW + \frac{1}{J}B.
\end{aligned}
$$

B is the between-sequence variance and W is the within-sequence variance of σ^2. $\widehat{Var}(\sigma^2)$ is an estimate of the variance that would be unbiased if the starting points of the simulation were really drawn from the target distribution, and it is an overestimate under the more realistic assumption that the starting values are overdispersed. Meanwhile, for J finite, quantity WW underestimates the variance of σ^2 since the individual sequences have not had time to range over all the support of the target distribution and then will have less variability.
For the parameter φ we compute the same quantities, but due to the fact that the trajectory $\varphi(\cdot)$ is a function on $\mathbb{R}$, all the corresponding quantities will be functions on $\mathbb{R}$. Therefore, we have an uncountable number of these quantities: one for every point in the domain of the realization φ.

$$\begin{aligned} B^{\varphi}(\cdot) &= \frac{J}{M-1}\sum_{i=1}^{M}(\overline{\varphi_{i.}}(\cdot)-\overline{\varphi_{..}}(\cdot))^2, \quad \overline{\varphi_{i.}}(\cdot)=\frac{1}{J}\sum_{j=1}^{J}\varphi_{ij}(\cdot), \quad \overline{\varphi_{..}}(\cdot)=\frac{1}{M}\sum_{i=1}^{M}\overline{\varphi_{i.}}(\cdot) \\ WW^{\varphi}(\cdot) &= \frac{1}{M}\sum_{i=1}^{M}(s_i^{\varphi})^2(\cdot), \quad (s_i^{\varphi})^2(\cdot)=\frac{1}{J-1}\sum_{j=1}^{J}(\varphi_{ij}(\cdot)-\overline{\varphi_{i.}}(\cdot))^2 \\ \widehat{Var}(\varphi(\cdot)) &= \frac{J-1}{J}WW^{\varphi}(\cdot)+\frac{1}{J}B^{\varphi}(\cdot). \end{aligned}$$

To monitor convergence of the iterative simulation, it is suggested in [13] to compute the *potential scale reduction*, denoted with $\hat{R}$ (respectively $\hat{R}^{\varphi}$). This quantity estimates the factor by which the scale of the current distribution for the parameter σ^2 (respectively φ) might be reduced if the iterations were continued in the limit $J\to\infty$. The potential scale reduction for σ^2 is computed as the ratio $\hat{R}=\frac{\widehat{Var}(\sigma^2)}{W}$ and then its square root is taken. The idea is to compare something that overestimates with a quantity that underestimates the variance in the target distribution $(\nu_{\alpha}^{\mathcal{F}})^{-1}$. It will be selected a number of iterations for which the potential scale reduction is near 1 for all parameters of interest. The target distribution will be summarized by using the simulated values from the last halves of the simulated sequences. The strategy described in [13] it is only adapted for scalar parameters. In particular, a problem arises in determining the potential scale reduction for an infinite dimensional parameter. Indeed, we have an uncountable number of $\hat{R}^{\varphi}$ for the parameter φ and check for all of them will be unfeasible. Our suggestion is to consider the uniform norm of this quantity:

$$\sqrt{\hat{R}_{\infty}^{\varphi}}=\sqrt{||\hat{R}^{\varphi}||_{\infty}},$$

where $||\hat{R}^{\varphi}||_{\infty}=\sup_s|\hat{R}^{\varphi}(s)|$ and $\hat{R}^{\varphi}(s)=\frac{\widehat{Var}(\varphi(s))}{W^{\varphi}(s)}$. In practice, with numerical simulations we shall have only a finite number of points s because of discretization of function φ. Therefore, our method can be seen as equivalent to a Gibbs sampling for a large, but finite, number of parameters where we are checking that the potential scale reduction is near 1 for all the parameters.
Alternatively, because of the finite number of discretization points s used in a numerical simulation, instead of computing variance for each fixed point s we suggest to compute the covariance matrix of $\varphi(s)$ for the vector of all discretization points of φ. Then, quantities B^{φ}, WW^{φ}, $\widehat{Var}(\varphi)$ become matrices and we can compute the maximum eigenvalues λ_{max} and λ_{max}^{W} of $\widehat{Var}(\varphi)$ and WW^{φ}, respectively. We propose to estimate the potential scale reduction as the ratio between these two eigenvalues: $\sqrt{\hat{R}}=\sqrt{\frac{\lambda_{max}}{\lambda_{max}^{W}}}$ and again to check that it is near 1.

4 The Unknown Operator Case

In the previous section we have developed Bayesian analysis by supposing that the joint density $f(Z,W)$ was known. Though this hypothesis simplifies considerably inference, it is not always realistic. In most of the cases it is more appropriate to consider that it is partially or completely unknown.
In this section, first we develop inference in a context in which f is known up to a parameter θ of finite dimension and then in a context in which f is totally unknown. In the latter case, nonparametric estimation methods require to be considered.

4.1 Unknown Finite Dimensional Parameter

When $F(Z,W)$ is known up to a finite dimensional parameter θ a further Bayesian experiment, different than Ξ, has to be specified. This is due to the fact that the instrumental variable model that we use to characterize Ξ, and in particular the sampling probability in it, does not specifies any characteristic of the distribution of (Z,W). The parameter space will be denoted with $\Theta \subset \mathbb{R}^l$, $\mathcal{A}$ is the associated σ-field and ρ is the probability measure defined on it.
Let consider an *i.i.d.* sampling of $F(Z,W)$, the Bayesian experiment is

$$\Xi_{Z,W} = (\mathbb{R}^l \times \mathcal{Y}_{Z,W}, \mathcal{A} \otimes \mathcal{F}_{Z,W}, \rho \times F^\theta),$$

with $\mathcal{Y}_{Z,W} = \mathbb{R}^{(p+q)N}$ the sampling space for the sample $(\mathbf{z}, \mathbf{w})$ and $\mathcal{F}_{Z,W}$ its associated σ-field. Remark that we keep the same notation F for the sampling measure on the $\tilde{n}$-product of sample spaces. The instrumental variable approach does not provide any way to rely together the two Bayesian experiments $\Xi_{Z,W}$ and Ξ, actually it only defines Ξ and, when θ is unknown, a Bayesian inference on it is possible only by specifying a new experiment $\Xi_{Z,W}$ and by considering a sample different than that one used to make inference on φ. This means that we have two completely separated model: the first one, $\Xi_{Z,W}$, will be used to estimate θ and the second one, Ξ, will be used to estimate φ given the previously obtained estimate for θ. To make this concept operational we need two samples: one on (Z,W) of size $\tilde{n}$, denoted with $s_2 = (\tilde{\mathbf{z}}, \tilde{\mathbf{w}}) = (s_{2,1}, \ldots, s_{2,\tilde{n}})$ and a different one on (Y,W) of size n, denoted with $s_1 = (y, \mathbf{w}) = (s_{1,1}, \ldots, s_{1,n})$ as specified in the following assumption:

Assumption 4 *$s_{1,i} = (y_i, w_i)$, $i = 1, \ldots, n$ and $s_{2,\tilde{i}} = (\tilde{z}_{\tilde{i}}, \tilde{w}_{\tilde{i}})$, $\tilde{i} = 1, \ldots, \tilde{n}$ are two i.i.d. samples of observations on $S_1 = (Y,W)$ and $S_2 = (Z,W)$, respectively.*

To simplify things we suppose the variance parameter σ^2 to be known, hence Bayesian model Ξ_θ results to be

$$\Xi_\theta = (L^2_F(Z) \times \mathcal{Y}, \mathcal{E} \otimes \mathcal{F}, \Pi^{\mathbf{W}} = \mu \times P^{\theta,\varphi,\mathbf{W}}).$$

The sampling and marginal probabilities in Ξ_θ depends on the realized value of θ, this justifies the use of the notation $P^{\theta,\varphi,\mathbf{W}}$, $P^{\theta,\mathbf{W}}$ for the sampling and marginal distribution and $\mu^{\mathcal{F},\theta}$ for the posterior probability. As already stressed, Bayesian experiment in Section 3 can be seen as a particular case of Ξ_θ that we are considering, in the sense that it is the conditional model in the case in which θ is known. In this case, Θ and $\mathcal{A}$ degenerate in a point θ_* and ρ degenerates into a point mass in θ_*.
Bayesian analysis is separated into two steps. In the first one, the parameter θ is estimated by only using the sample $(\tilde{\mathbf{z}}, \tilde{\mathbf{w}})$. The second step performs posterior analysis of φ conditionally on a θ drawn from the posterior $\rho(\theta|\tilde{\mathbf{z}}, \tilde{\mathbf{w}})$ and it demands the use of only the sample $(y, \mathbf{w})$ and model Ξ_θ.
We assume that the subvector $S_2 = (Z,W)$ induces a gaussian measure on $\mathbb{R}^{p+q}$ with mean vector $m \in \mathbb{R}^{p+q}$ and covariance matrix $V \in \mathcal{C}_{p+q}$, where $\mathcal{C}_{p+q}$ is the cone of the $(p+q) \times (p+q)$ positive definite matrices. Therefore $\theta = (m, V) \in \Theta = \mathbb{R}^{p+q} \times \mathcal{C}_{p+q}$,

$$s_{2,i}|\theta \sim \ i.i.d.\, \mathcal{N}_{p+q}(m, V), \qquad i = 1, \ldots, \tilde{n}$$

and F^θ is the product of $\tilde{n}$ multidimensional normal distributions. In order to simplify simulations, we consider the precision matrix $\Sigma = V^{-1}$ instead of V, hence parameter θ becomes: $\theta = (m, \Sigma)$. We specify a conjugate prior for θ:

$$\begin{aligned} \Sigma &\sim \mathcal{W}(\Sigma_0, v_0), \qquad \Sigma_0 \in \mathcal{C}_{p+q}, \quad v_0 > (p+q)+1 \\ m|\Sigma &\sim \mathcal{N}_{(p+q)}(m_0, \frac{1}{u_0}\Sigma^{-1}), \qquad m_0 \in \mathbb{R}^{p+q},\ u_0 \in \mathbb{R}_+, \end{aligned}$$

where $\mathcal{W}(\Sigma_0, v_0)$ stands for a Wishart distribution with parameters a matrix Σ_0 of conformable dimensions and a scalar v_0. Standard Bayesian computations give the posterior of (m, Σ)

$$\rho(m,\Sigma|(s_{2,i})_{i=1,\dots,\tilde{n}}) \propto |\Sigma|^{\frac{1}{2}+\frac{v_*-(p+q+1)}{2}} \exp\{\frac{1}{2}[u_*(m-m_*)'\Sigma(m-m_*)+tr\Sigma_*^{-1}\Sigma]\}$$

and its decomposition

$$\begin{aligned}
\rho(\Sigma|(s_{2,i})_{i=1,\dots,\tilde{n}}) &\sim \mathcal{W}(\Sigma_*, v_*) &(6)\\
\rho(m|\Sigma;(s_{2,i})_{i=1,\dots,\tilde{n}}) &\sim \mathcal{N}_{(p+q)}(m_*, \frac{1}{u_*}\Sigma^{-1}), &(7)
\end{aligned}$$

with

$$\begin{aligned}
u_* &= \tilde{n}+u_0\\
m_* &= \frac{1}{u_*}(\sum_i s_{2,i}+u_0m_0)\\
v_* &= \tilde{n}+v_0\\
\Sigma_*^{-1} &= \Sigma_0^{-1}+\sum_i s_{2,i}s_{2,i}'+u_0m_0m_0'-\frac{\tilde{n}^2}{u_*}\bar{s}_2\bar{s}_2'-\tilde{n}\frac{u_0}{u_*}(\bar{s}_2m_0'+m_0\bar{s}_2')-\frac{u_0^2}{u_*}m_0m_0'\\
\bar{s}_2 &= \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} s_{2,i}.
\end{aligned}$$

Once the posterior distribution $\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})$ has been obtained, we draw from it a value of θ that will characterize the sampling measure $P^{\theta,\varphi,\mathbf{W}}$ in Ξ_θ and the regularized posterior distribution $\mu_\alpha^{\mathcal{F},\theta}$, conditional on θ, is computed as usual. The dependence of $\mu_\alpha^{\mathcal{F},\theta}$ on the particular value θ extracted from $\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})$ will be eliminated by integrating out θ:

$$\begin{aligned}
\mathbb{E}_\alpha(\varphi|y_{(n)},\mathbf{w}) &= \int \mathbb{E}(\varphi|\theta,y_{(n)},\mathbf{w},\tilde{\mathbf{z}},\tilde{\mathbf{w}})\rho(\theta|y_{(n)},\mathbf{w},\tilde{\mathbf{z}},\tilde{\mathbf{w}})d\theta &(8)\\
Var_\alpha(\varphi|y_{(n)},\mathbf{w}) &= \int Var_\alpha(\varphi|\theta,y_{(n)},\mathbf{w},\tilde{\mathbf{z}},\tilde{\mathbf{w}})\rho(\theta|y_{(n)},\mathbf{w},\tilde{\mathbf{z}},\tilde{\mathbf{w}})d\theta + Var(\mathbb{E}_\alpha(\varphi|\theta,y_{(n)},\mathbf{w},\tilde{\mathbf{z}},\tilde{\mathbf{w}})|\tilde{\mathbf{z}},\tilde{\mathbf{w}}) &(9)
\end{aligned}$$

where the last variance is taken with respect to $\rho(\theta|r_n,\mathbf{w},\tilde{\mathbf{z}},\tilde{\mathbf{w}})$. For statistical coherence, we write all the conditioning variables, but we could simplify things by eliminating the variable with respect to which there is independence:

$$\begin{aligned}
\mathbb{E}(\varphi|\theta,y_{(n)},\mathbf{w},\tilde{\mathbf{z}},\tilde{\mathbf{w}}) &= \mathbb{E}(\varphi|\theta,y_{(n)},\mathbf{w})\\
\rho(\theta|y_{(n)},\mathbf{w},\tilde{\mathbf{z}},\tilde{\mathbf{w}}) &= \rho(\theta|y_{(n)},\mathbf{w})\\
Var_\alpha(\varphi|\theta,y_{(n)},\mathbf{w},\tilde{\mathbf{z}},\tilde{\mathbf{w}}) &= Var_\alpha(\varphi|\theta,y_{(n)},\mathbf{w}).
\end{aligned}$$

Quantities (8) and (9) completely characterize $\mu_\alpha^{\mathcal{F}}$ and integrals in them, with respect to ρ, can be approximated thanks to Monte Carlo integration, after a number J of θ have been drawn from $\rho(\theta|r_n,\mathbf{w})$. In practice, $\mu_\alpha^{\mathcal{F}}$ will be obtained by running the following iterative algorithm. This algorithm assumes $\sigma^2 = Var(\varepsilon_i|W)$ known.

(i) draw $\theta^{(j)}$ from the posterior $\rho(m,\Sigma|(s_{2,i})_{i=1,\dots,\tilde{n}})$;

(ii) compute $f^{(j)}(Z|W)$ and $f^{(j)}(Z)$ in order to compute the kernel of operators $K_{(n)}$ and $K_{(n)}^*$. We will denote the corresponding operators with $\hat{K}^{(j)}$ and $\hat{K}^{*(j)}$, respectively;

(iii) compute the regularized posterior distribution $\mu_\alpha^{\mathcal{F},\theta(j)}$ given $\theta^{(j)}$ characterized by the mean function $\hat{\varphi}_\alpha^{(j)} = A_\alpha^{(j)}y_{(n)} + b_\alpha^{(j)}$ and the covariance operator $\Omega_{y,\alpha}^{(j)} = \Omega_0 - A_\alpha^{(j)}\hat{K}^{(j)}\Omega_0$, with $A_\alpha^{(j)} = \Omega_0\hat{K}^{*(j)}(\alpha_nI_n + \hat{K}^{(j)}\Omega_0\hat{K}^{*(j)} + \frac{\sigma^2}{n}I_n)^{-1}$ and $b_\alpha^{(j)} = (I - A_\alpha^{(j)}\hat{K}^{(j)})\varphi_0$;

(iv) iterate *(i)* - *(iii)* up to obtain a large number J of estimations $\hat{\varphi}_\alpha^{(j)}$ and $\Omega_{y,\alpha}^{(j)}$, $j = 1,\dots,J$;

(v) compute the sample average of the J regularized posterior means: $\hat{\mathbb{E}}_\alpha(\varphi|y_{(n)},\mathbf{w}) = \frac{1}{J}\sum_j \hat{\varphi}_\alpha^{(j)}$ and of the J regularized posterior variances: $\Omega_{y,\alpha} = \frac{1}{J}\sum_j \Omega_{y,\alpha}^{(j)}$ to approximate the first term in the RHS of (9). Approximate the second term in the RHS of (9) by $\int \frac{1}{J}\sum_j(\hat{\varphi}_\alpha^{(j)}(Z)\hat{\varphi}_\alpha^{(j)}(\zeta))f(Z,\cdot|\theta_*)dZ - \int(\frac{1}{J}\sum_j \hat{\varphi}_\alpha^{(j)}(Z))(\frac{1}{J}\sum_j \hat{\varphi}_\alpha^{(j)}(\zeta))f(Z,\cdot|\theta_*)dZ$ with the integral replaced by Monte Carlo integration. Denote $\hat{\varphi}_\alpha = \hat{\mathbb{E}}_\alpha(\varphi|y_{(n)},\mathbf{w})$ and $\hat{\Omega}_{y,\alpha} = \widehat{Var}_\alpha(\varphi|y_{(n)},\mathbf{w})$ the estimated regularized posterior mean and variance characterizing $\mu_\alpha^{\mathcal{F}}$.

The sample counterparts $\hat{\varphi}_\alpha$ $\hat{\Omega}_{y,\alpha}$ of the mean and variance of $\mu_\alpha^{\mathcal{F}}$ characterize the *estimated regularized posterior distribution* $\hat{\mu}_\alpha^{\mathcal{F}}$ that is the solution to the inference problem for φ when $f(Z,W)$ is known up to a parameter. The estimation errors caused by an unknown θ are shown to be negligible with respect to the error due to approximation of φ_* by $\mu_\alpha^{\mathcal{F}}$. More precisely, for the estimated regularized posterior mean we have the decomposition:

$$\begin{aligned}||\hat{\varphi}_\alpha - \varphi_*||^2 \quad \leq \quad & ||\hat{\varphi}_\alpha - \mathbb{E}_\alpha(\varphi|y_{(n)},\mathbf{w})||^2 + ||\mathbb{E}_\alpha(\varphi|y_{(n)},\mathbf{w}) - \mathbb{E}_\alpha(\varphi|\hat{\theta},y_{(n)},\mathbf{w})||^2 \\ & + ||\mathbb{E}_\alpha(\varphi|\hat{\theta},y_{(n)},\mathbf{w}) - \mathbb{E}_\alpha(\varphi|\theta_*,y_{(n)},\mathbf{w})||^2 + ||\mathbb{E}_\alpha(\varphi|\theta_*,y_{(n)},\mathbf{w}) - \varphi_*||^2.\end{aligned}$$

We have denoted with θ_* the true value of θ having generated the data $(\tilde{\mathbf{z}},\tilde{\mathbf{w}})$ and $\hat{\theta}$ the posterior mean of θ, *i.e.* $\hat{\theta} = \int \theta\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta$. The first term is the error due to Monte Carlo integration, then it declines to 0 as fast as more discretization points are considered. Since the second and third error terms are $\mathcal{O}_p(\frac{1}{\tilde{n}})$, they are negligible with respect to the last term which has the speed of convergence given in Theorem 2.
The following theorem shows consistency of the estimated posterior mean under some minor hypothesis.

Theorem 4 *Let $\mathbb{E}_\alpha(\varphi|\theta,y_{(n)},\boldsymbol{w}) \in L_F^2(Z)$ be the regularized posterior mean as defined in Theorem 2. If $(\varphi_*-\varphi_0) \in \mathcal{H}(\Omega_0)$, $\alpha_n \to 0$, $\frac{1}{\alpha_n n} \to 0$, $\frac{1}{\alpha_n^3 n^2} \sim \mathcal{O}_p(1)$, and $\frac{\partial \mathbb{E}_\alpha(\varphi|\theta,y_{(n)},\boldsymbol{w})}{\partial\theta} \in L_F^2(Z)$ for $\theta=\theta_*$ and $\theta=\hat{\theta}$, then*

(i) $||\hat{\varphi}_\alpha - \varphi_*||_{L_F^2}^2 \to 0$ *in* $F^\theta \times P^{\theta,\varphi,\boldsymbol{w}}$;

(ii) if moreover $\Omega_0^{-\frac{1}{2}}(\varphi_*-\varphi_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^2\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ *for some* $\beta>0$, *then*

$$\begin{aligned}||\hat{\varphi}_\alpha - \varphi_*||_{L_F^2}^2 \quad \sim \quad & \mathcal{O}_p(\frac{1}{\alpha_n^4 n^3} + \frac{1}{\alpha_n^2 n^2}\alpha_n^\beta + \frac{1}{\alpha_n^3 n^3} + \frac{1}{\alpha_n^2 n^2} + \\ & \alpha_n^\beta + \frac{1}{\alpha_n^4 n^2}\alpha_n^{(\beta+1)\wedge 2} + \frac{1}{\alpha_n n}).\end{aligned}$$

We implicitly assume in Theorem 4 that all the conditions necessary to guarantee consistency of the posterior mean $\hat{\theta}$ of a finite dimensional parameter are satisfied, see [1], [14] or [24] for technical details.
Let study, at this stage, convergence to zero of the regularized posterior variance $\hat{\Omega}_{y,\alpha}$:

$$\begin{aligned}\hat{\Omega}_{y,\alpha}\phi \quad = \quad & \frac{1}{J}\sum_{j=1}^J[\Omega_{y,\alpha}(\theta^{(j)})\phi](\zeta) + \frac{1}{MJ}\sum_{m=1}^M\sum_{j=1}^J[\mathbb{E}_\alpha(\varphi|\theta^{(j)},y_{(n)},\mathbf{w})\mathbb{E}_\alpha(\varphi(z_m)|\theta^{(j)},y_{(n)},\mathbf{w})\phi(z_m)](\zeta) \\ & -\frac{1}{MJ^2}\sum_{m=1}^M[\sum_{j=1}^J\mathbb{E}_\alpha(\varphi|\theta^{(j)},y_{(n)},\mathbf{w})][\sum_{j=1}^J\mathbb{E}_\alpha(\varphi(z_m)|\theta^{(j)},y_{(n)},\mathbf{w})\phi(z_m)](\zeta), \qquad (10)\end{aligned}$$

with $\phi \in L_F^2(Z)$, $\Omega_{y,\alpha}(\theta^{(j)}) = Var_\alpha(\varphi|\theta,y_{(n)},\mathbf{w}) = \Omega_{y,\alpha}^{(j)}$. The index m denotes the m-th drawn of z from $f(Z,\cdot|\theta_*)$.

Theorem 5 *Let $\hat{\Omega}_{y,\alpha} : L_F^2(Z) \to L_F^2(Z)$ be computed as in (10). If $\alpha_n \to 0$, $\frac{1}{\alpha_n n} \to 0$, $\frac{1}{\alpha_n^3 n^2} \sim \mathcal{O}_p(1)$, $\int \frac{\partial}{\partial\theta}\mathbb{E}_\alpha(\varphi(z)|\theta,y_{(n)},\boldsymbol{w})\mathbb{E}_\alpha(\varphi(\zeta)|\theta,y_{(n)},\boldsymbol{w})f(z|\theta_*)dz$ is an Hilbert-Schmidt operator and $\frac{\partial}{\partial\theta}\Omega_{y,\alpha}(\theta) \in L_F^2(Z)$ for $\theta=\theta_*$ and $\theta=\hat{\theta}$ then*

(i) $||\hat{\Omega}_{y,\alpha}||^2_{L^2_F} \to 0$ *in* $F^\theta \times P^{\theta,\varphi,\boldsymbol{w}}$;

(ii) moreover, $\forall \phi \in L^2_F(Z)$ *such that* $\Omega_0^{\frac{1}{2}}\phi \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ *for some* $\beta > 0$

$$||\hat{\Omega}_{y,\alpha}\phi||^2_{L^2_F} \sim \mathcal{O}_p(\frac{1}{\alpha_n^2 n^2}\alpha_n^\beta + \frac{1}{\alpha_n^4 n^3} + \frac{1}{\alpha_n^4 n^2}\alpha_n^{(\beta+1)\wedge 2} + \alpha_n^\beta).$$

Chebyshev's Inequality allows to show that, under conditions given for point *(i)* of Theorems (4) and (5), posterior consistency is preserved also in the case with unknown θ:

$$\hat{\mu}_\alpha^{\mathcal{F}}\{\varphi : ||\varphi - \varphi_*|| \geq \varepsilon_n\} \leq \frac{1}{\varepsilon_n}(||\hat{\varphi}_\alpha - \varphi_*||^2 + ||\hat{\Omega}_{y,\alpha}||^2).$$

Moreover, under conditions given in point *(ii)* of Theorems 4 and 5, with optimal regularization parameter α_*, $\hat{\mu}_\alpha^{\mathcal{F}}$ degenerates towards a point mass in φ_* at the optimal speed of $n^{-\frac{2\beta}{\beta+1}}$. This means that the optimal speed of convergence does not change with respect to the better case in which F is completely known.

4.2 Unknown Infinite Dimensional Parameter

When the joint density $f(Z,W)$ is totally unknown we have to deal with a nonparametric problem that presents complex difficulties. Pioneer Bayesian nonparametric was based on Dirichlet processes (introduced by [6]) that however has the drawback of producing discrete random probabilities measures with probability one. In alternative, Polya tree priors, initially considered by [7] and then by [17], can be chosen to generate only absolutely continuous distributions. We refer to [2] for a complete review on Bayesian nonparametric methods, being this beyond the scope of this paper.
The technique that we propose in this paper for dealing with this case is essentially different and it is far from Bayesian methods. We propose to substitute the true $f(Z,W)$ in operators $K_{(n)}$ and $K^*_{(n)}$ with a nonparametric classical estimator and to redefine the structural function φ as solution of the estimated reduced form equation

$$y_{(n)} = \hat{K}_{(n)}\varphi + \eta_{(n)} + \varepsilon_{(n)}. \tag{11}$$

We use the notation $\hat{K}_{(n)}$ and $\hat{K}^*_{(n)}$ for the corresponding operators with $f(Z,W)$ substituted by a nonparametric estimator. We have two errors term: $\varepsilon_{(n)}$ that is the classical error term of the reduced form and $\eta_{(n)}$ that accounts for the estimation error of operator $K_{(n)}$, *i.e.* $\eta_i = \frac{1}{\sqrt{n}}(\mathbb{E}(\varphi|w_i) - \hat{\mathbb{E}}(\varphi|w_i))$ and $\eta_{(n)} = (\eta_i, \ldots, \eta_n)'$. The estimated operator $\hat{K}_{(n)}$ is seen as the true operator characterizing a functional equation and it must not be considered as an element of the Bayesian experiment in the sense that we do not specify a probability measure on the space of absolutely continuous probability measure of (Z,W). Indeed, equation (11) defines a new Bayesian experiment that is a slightly modification of Ξ in Section 3 due to the fact that σ^2 is known, and then it no more enters the Bayesian experiment, and to the fact that the sampling distribution is differently specified (we will see it below):

$$\Xi_f = (L^2_F(Z) \times \mathcal{Y}, \mathcal{E} \otimes \mathcal{F}, \Pi^{\mathbf{W}} = \mu \times \hat{P}^{\varphi,\mathbf{W}}).$$

Nonparametric estimation of $f(Z,W)$ is performed by kernel smoothing; we stress the fact that here, contrarily to the previous case with f known up to a parameter θ, we use the same sample for estimating f and for getting the posterior distribution of φ. Let L be a kernel function satisfying the usual properties and ρ the minimum between the order of L and the order of differentiability of f. Moreover, we use the notation $L(u)$ for $L(\frac{u}{h})$ where h is the bandwidth used for kernel estimation such that $h \to 0$ with n (for lightening notation we have eliminated the dependence on n from h). The estimated density function is

$$\hat{f}(W,Z) = \frac{1}{nh^{p+q}}\sum_{i=1}^n L_w(w_i - W)L_z(z_i - Z),$$

where we have used different index in the kernel for W and for Z. Estimate of $K_{(n)}$ and $K^*_{(n)}$ are obtained by plugging in the estimate $\hat{f}$:

$$\hat{K}_{(n)}\varphi = \frac{1}{\sqrt{n}} \begin{pmatrix} \sum_j \varphi(z_j) \frac{L(w_1 - w_j)}{\sum_l L(w_1 - w_l)} \\ \vdots \\ \sum_j \varphi(z_j) \frac{L(w_n - w_j)}{\sum_l L(w_n - w_l)} \end{pmatrix}, \quad \varphi \in L^2_Z$$

$$\hat{K}^*_{(n)}x = \frac{1}{\sqrt{n}} \sum_i x_i \frac{\sum_j L(z - z_j)L(w_1 - w_j)}{\sum_l L(z - z_l)\frac{1}{n}\sum_l L(w_1 - w_l)}, \quad x \in \mathbb{R}^n$$

and

$$\hat{K}^*_{(n)}\hat{K}_{(n)}\varphi = \frac{1}{n}\sum_i \left(\sum_j \varphi(z_j) \frac{L(w_i - w_j)}{\sum_l L(w_i - w_l)} \right) \frac{\sum_j K(Z - z_j)K(w_i - w_j)}{\sum_l L(Z - z_l)\frac{1}{n}\sum_l L(w_i - w_l)}.$$

The element in brackets in the last expression converges to $\mathbb{E}(\varphi|w_i)$, the last ratio converges to $\frac{f(Z,w_i)}{f(Z)f(w_i)}$ and hence by the Law of Large Number $\hat{K}^*_{(n)}\hat{K}_{(n)}\varphi \to \mathbb{E}(\mathbb{E}(\varphi|w_i)|Z)$.
Asymptotic properties for kernel estimation of regression function justifies the following hypothesis:

Assumption 5 *$\eta_{(n)} \sim \mathcal{N}_n(0, \frac{\sigma^2}{n^2h^q}D_{(n)})$, where $D_{(n)} = diag(\frac{1}{f(w_i)}\int L^2(u)du)$, $i = 1, \ldots, n$.*

The fact that the covariance matrix is diagonal follows from the asymptotic independence of kernel estimator of the regression function at different points: $\hat{\mathbb{E}}(\varphi|w_i) \perp\!\!\!\perp \hat{\mathbb{E}}(\varphi|w_j)$, $\forall i \neq j$.
The covariance operator of the sampling measure induced on $\mathbb{R}^n$ by $y_{(n)}$ is determined by the covariance of error term $\eta_{(n)} + \varepsilon_{(n)}$. As in the case with f known, $\varepsilon_{(n)}$ has variance $\frac{\sigma^2}{n}I_n$ so that the variance of $\eta_{(n)}$ is negligible with respect to it, since by definition the bandwidth h satisfies $nh^q \to \infty$. The same can be said concerning the covariance between $\eta_{(n)}$ and $\varepsilon_{(n)}$; therefore we are content to simply write $Var(y_{(n)}) = (\frac{\sigma^2}{n} + o_p(\frac{1}{n}))I_n$ and we denote this matrix with Σ_n. At this point we are able to specify the prior and sampling probabilities $\hat{P}^{\varphi,\mathbf{W}}$:

$$\begin{aligned} \varphi &\sim \mathcal{GP}(\varphi_0, \Omega_0) \\ y_{(n)}|\varphi &\sim \mathcal{N}_n(\hat{K}_{(n)}\varphi, \Sigma_n). \end{aligned}$$

Usual computations, and problems of continuity, give the estimated regularized posterior and predictive probabilities $\hat{\mu}^{\mathcal{F}}_\alpha$ and $\hat{P}^w$, respectively:

$$\begin{aligned} y_{(n)} &\sim \mathcal{N}_n(\hat{K}_{(n)}\varphi_0, \Sigma_n + \hat{K}_{(n)}\Omega_0\hat{K}^*_{(n)}) \\ \varphi|y_{(n)} &\sim \mathcal{GP}(\hat{\mathbb{E}}_\alpha(\varphi|y_{(n)}), \hat{\Omega}_{y,\alpha}) \end{aligned}$$

with

$$\begin{aligned} \hat{\mathbb{E}}_\alpha(\varphi|y_{(n)}) &= \varphi_0 + \overbrace{\Omega_0\hat{K}^*_{(n)}(\alpha_n I_n + \Sigma_n + \hat{K}_{(n)}\Omega_0\hat{K}^*_{(n)})^{-1}}^{\hat{A}_\alpha}(y_{(n)} - \hat{K}_{(n)}\varphi_0) \\ \hat{\Omega}_{y,\alpha} &= \Omega_0 - \Omega_0\hat{K}^*_{(n)}(\alpha_n I_n + \Sigma_n + \hat{K}_{(n)}\Omega_0\hat{K}^*_{(n)})^{-1}\hat{K}_{(n)}\Omega_0. \end{aligned}$$

It should be remarked that regularization is necessary since $\frac{1}{n}$ in Σ_n does not stabilize the inverse. More clearly, it converges to zero too fast to compensate the decline towards 0 of the spectrum of operator $\hat{K}_{(n)}\Omega_0\hat{K}^*_{(n)}$. Therefore, to guarantee continuity and consistency of the posterior distribution it must be introduced a regularization parameter $\alpha_n > 0$ that goes to 0 slower than $\frac{1}{n}$ and $\frac{1}{n^2h^q}$.

Asymptotic properties of the posterior distribution for the case with unknown f are very similar to that one shown in Theorem 2 and in Corollary 1.

Theorem 6 *Let φ_* be the true value of the parameter and $\hat{\mu}_\alpha^{\mathcal{F}}$ a gaussian measure on $L_F^2(Z)$ with mean $\hat{A}_\alpha(y_{(n)} - \hat{K}_{(n)}\varphi_0) + \varphi_0$ and covariance operator $\hat{\Omega}_{y,\alpha}$. If $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$ and if $\alpha_n \to 0$, $\alpha_n^2 n \to \infty$, then*

(i) $\hat{\mu}_\alpha^{\mathcal{F}}$ weakly converges to point mass δ_{φ_*} in φ_*;

(ii) if moreover $\Omega_0^{-\frac{1}{2}}(\varphi_* - \varphi_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ with $\beta \leq 2$, $||\hat{K}_{(n)}^* \hat{K}_{(n)} - K^*K||^2 \sim \mathcal{O}_p(\frac{1}{nh^p} + h^{2\rho})$ and $||\Omega_0^{\frac{1}{2}}(\hat{K}_{(n)}^* \hat{K}_{(n)} - K^*K)\Omega_0^{\frac{1}{2}}||^2 \sim \mathcal{O}_p(\frac{1}{n} + h^{2\rho})$, then

$$\hat{\mu}_\alpha^{\mathcal{F}}\{\varphi : ||\varphi - \varphi_*|| \geq \epsilon_n\} \quad \sim \quad \mathcal{O}_p((\alpha_n^\beta + \frac{1}{\alpha_n^2}(\frac{1}{n} + h^{2\rho})\alpha_n^\beta)(1 + \frac{1}{\alpha_n^4 n^2}\frac{1}{n^2 h^{2q}}) + \frac{1}{\alpha_n^2 n})$$

The optimal speed of convergence will be obtained when $\alpha_n^\beta = \frac{1}{\alpha_n^2 n}$, that provides the optimal regularization parameter $\alpha_n \propto n^{-\frac{1}{\beta+2}}$ and the optimal speed of convergence proportional to $n^{-\frac{\beta}{\beta+2}}$ exactly as for f known. The bandwidth is determined in order to satisfy $\frac{h^{2\rho}}{\alpha_n^2} \sim \mathcal{O}_p(1)$, so that

$$h_n \propto n^{-\frac{1}{2\rho}}.$$

5 Numerical Implementation

In this section we investigate the goodness of fit of the regularized posterior distribution in all the considered cases. A large-sample simulation study of asymptotic properties of the estimator is performed. Only results for two different specifications for the prior distribution of φ are reported here. All the simulations have been performed with Matlab®.

We simulate a model where there is only one covariate that is endogenous and a bivariate vector of instruments is available. Our design uses a simple specification for the true value of the structural function: $\varphi_*(Z) = Z^2$ and the structural model for generating the y_is and z_is.

$$\begin{aligned} y_i &= \varphi_*(z_i) + u_i \\ \varphi_*(z_i) &= z_i^2 \\ u_i &= -0.5 v_i + \xi_i \end{aligned}$$

$$\begin{aligned} \varepsilon_i &\sim \mathcal{N}(0, (0.27)^2) \\ \xi_i &\sim \mathcal{N}(0, (0.05)^2) \\ z_i &= 0.1 w_{i,1} + 0.1 w_{i,2} + v_i \\ w_i &= \begin{pmatrix} w_{1,i} \\ w_{2,i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix} \right). \end{aligned}$$

This mechanism of generation entails w_i, v_i and ξ_i are mutually independent for every i; moreover it entails the joint density f is

We take (Z, W) jointly normal

$$\begin{pmatrix} Z \\ W_1 \\ W_2 \end{pmatrix} \sim \mathcal{N}_3 \quad \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.0989 & 0.13 & 0.13 \\ 0.13 & 1 & 0.3 \\ 0.13 & 0.3 & 1 \end{pmatrix} \right).$$

Endogeneity is caused by correlation between u_i and the error term v_i affecting the covariates. The simulation is made for $n = 1000$ and $\alpha_n = 0.3$. The fixed value for α_n has been determined by letting this parameter vary in a very large range of values and selecting that one producing a

better estimation. Of course, this technique applies only when we work with simulated data and we are working to develop a data-driven method to select α_n when we consider real data.
We have performed simulations for the conjugate model with known $f(Z,W)$ (Case I) and for the case with completely unknown $f(Z,W)$ with known σ^2 (Case II). The most important step in bayesian estimation is a correct specification of the prior distribution. It summarizes our prior knowledge about the parameter we desire to estimate. We chose an *Inverse Gamma - Gaussian distribution* for Case I and a *Gaussian distribution* for the only parameter φ that we have in Case II.

Case I. *Conjugate Model with $f(Z,W)$ known.*
In this simulation we choose a conjugate prior:

$$\begin{aligned} \sigma^2 &\sim \mathcal{IG}(5, 0.12) \\ \varphi &\sim \mathcal{GP}(\varphi_0, \sigma^2\Omega_0) \end{aligned}$$

with covariance operator $(\Omega_0\delta)(Z) = \sigma_0 \int \exp\{-(s-Z)^2\}\delta(s)f(s,\cdot)ds$, where $f(s,\cdot)$ is the marginal density of Z and δ is any function in $L^2_F(Z)$. We have performed simulations for several choices for φ_0 and σ_0 in order to see the impact of different prior distributions on our estimator.
The results are reported in Figure 1, where the first three graphs are drawn $\varphi_0(Z) = 0.95Z^2 + 0.25$ and $\sigma_0 = 0.5$ and the last three for $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$ and $\sigma_0 = 200$. Panels (1b) - (1c) and (1e) - (1f) represent drawn from the prior and posterior distribution of φ. In Figure 2 we show drawn from the prior and the posterior distribution of σ^2.

Case II. *$f(Z,W)$ unknown and σ^2 known.*
In this simulation we have to specify a prior only on φ since σ^2 is supposed to be known:

$$\varphi \sim \mathcal{GP}(\varphi_0, \Omega_0)$$

with φ_0 and Ω_0 specified as in Case I. We show in Figure 3 only the results for the prior distribution specification with $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$ and $\sigma_0 = 200$. Panels 3a shows the estimated regularized posterior mean, together with the true curve and the prior mean; panel 3b reports a sample drawn from the estimated posterior distribution.

6 Conclusions

We have studied in this paper a new method to make bayesian inference on an instrumental regression φ defined through a structural econometric model. The peculiarity of our method is that it does not require any specification of the functional form for φ, though it allows to incorporate all the prior information is available. However, a deeper analysis of the role played by the prior distribution seems to be advisable.
A lot of possible extensions of our model can be developed and we plan to do it in the near future. First of all, we would like to consider other regularization methods, different from the Tikhonov scheme. Moreover, we could consider Sobolev spaces, instead of general Hilbert space, and regularize using differential norms.
Lastly, we intend to develop a data-driven method for choosing the regularization parameter α_n.

7 Appendix A

In all the proofs that follow the notation will be the following:

- $\mathcal{H}(\Omega_0) = \mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}(\Omega_0)$;
- $\psi = \Omega_0^{-\frac{1}{2}}(\varphi_* - \varphi)$, $\phi \in L^2_F(Z)$;
- $T = K\Omega_0^{-\frac{1}{2}}$, $T : L^2_F(Z) \to L^2_F(W)$;
- $\hat{T} = \hat{K}\Omega_0^{-\frac{1}{2}}$, $\hat{T} : L^2_F(Z) \to \mathbb{R}^n$;

- $T^* = \Omega_0^{-\frac{1}{2}} K^*$, $T : L_F^2(W) \to L_F^2(Z)$;
- $\hat{T}^* = \Omega_0^{-\frac{1}{2}} \hat{K}^*$, $\hat{T}^* : \mathbb{R}^n \to L_F^2(Z)$;
- $\Omega_0^{\frac{1}{2}} = \int \omega_0(s, Z) f(s) ds$;
- $g(Z, w_i) = \int \omega_0(s, Z) \frac{f(s, w_i)}{f(s) f(w_i)} f(s) ds$

7.1 Proof of Lemma 1

To clarify the discussion in the following, we will use the notation $\mu_n^{\sigma,\mathcal{F}}$ for the posterior distribution instead of the usual one $\mu^{\sigma,\mathcal{F}}$, but it should result clear that they have the same meaning. The index n denotes the sample size and the limits that we shall consider are for $n \to \infty$.
Definition of *weak convergence* of probability measures says that a sequence of probability measures $\mu_n^{\sigma,\mathcal{F}}$ on an Hilbert space $L_F^2(Z)$, endowed with the Borel σ-field $\mathcal{E}$, converges weakly to a probability measure δ_{φ_*} if

$$|| \int a(\varphi) \mu_n^{\sigma,\mathcal{F}}(d\varphi) - \int a(\varphi) \delta_{\varphi_*}(d\varphi) ||_{L_F^2} \to 0,$$

for every bounded and continuous functional $a : L_F^2(Z) \to L_F^2(Z)$, where $|| \cdot ||_{L_F^2}$ denotes the norm in the space $L_F^2(Z)$.
We prove that this convergence is not satisfied at least for one functional a. We consider the identity functional $a : \phi \mapsto \phi$, $\forall \phi \in L_F^2(Z)$, so that we have to check convergence of the posterior mean. Take, for brevity, null prior mean, $\varphi_0 = 0$, the posterior mean estimator for φ is

$$\mathbb{E}(\varphi | y_{(n)}) = \Omega_0 K_{(n)}^* (\frac{1}{n} I + K_{(n)} \Omega_0 K_{(n)}^*)^{-1} y_{(n)}.$$

We are interested in the L_F^2 norm:

$$\begin{aligned} ||\mathbb{E}(\varphi | y_{(n)}) - \varphi_*|| \quad \leq \quad & \overbrace{||\Omega_0 K_{(n)}^* (\frac{1}{n} I + K_{(n)} \Omega_0 K_{(n)}^*)^{-1} K_{(n)} \varphi_* - \varphi_*||}^{I} \\ & + \underbrace{||\Omega_0 K_{(n)}^* (\frac{1}{n} I + K_{(n)} \Omega_0 K_{(n)}^*)^{-1} \varepsilon||}_{II}. \end{aligned}$$

If we assume $\varphi_* \in \mathcal{H}(\Omega_0)$ [3], term I can be rewritten as

$$||\Omega_0^{\frac{1}{2}} [I - \Omega_0^{\frac{1}{2}} K_{(n)}^* (\frac{1}{n} I + K_{(n)} \Omega_0 K_{(n)}^*)^{-1} K_{(n)} \Omega_0^{\frac{1}{2}}] \gamma||,$$

and it has the same eigenvalues as

$$||\Omega_0^{\frac{1}{2}} [I - (\frac{1}{n} I + \hat{T}^* \hat{T})^{-1} \hat{T}^* \hat{T}] \gamma||$$

obtained by permuting the operator. The term in squared brackets is the regularization bias of the equation $\hat{T}\gamma = r$ with regularization parameter $\frac{1}{n}$. However this regularization scheme does not regularize properly since the regularization parameter goes to 0 at a faster rate than the speed at which $\hat{T}^*\hat{T}$ degenerates towards an infinite rank operator T^*T with unbounded inverse. In particular, by Kolmogorov's Theorem $||\hat{T}^*\hat{T} - T^*T||^2 \sim \mathcal{O}_p(\delta)$ if $\mathbb{E}(||\hat{T}^*\hat{T} - T^*T||^2) \sim \mathcal{O}_p(\delta)$, where the expectation is taken with respect to the distribution of w_i. This is the usual MISE that can be decomposed into the sum of the squared bias and the variance. The bias is zero since $\mathbb{E}(\hat{T}^*\hat{T}) - T^*T = 0$, while the variance goes to zero at the speed of $\frac{1}{n}$, so that $||\hat{T}^*\hat{T}|| \sim \mathcal{O}_p(\frac{1}{\sqrt{n}})$. Therefore this regularization scheme is not well defined and so term I is not convergent.
A similar argument proves that also II term does not go to 0 and this complete the proof.

[3] Note that this condition becomes $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$ in the case with non null prior mean.

7.2 Proof of Corollary 1

To prove the first point we develop the bias in two terms:

$$\begin{aligned}\hat{\varphi}_\alpha - \varphi_* &= -\overbrace{(I - \Omega_0 K_{(n)}^*(\alpha_n I + \frac{1}{n}I + K_{(n)}\Omega_0 K_{(n)}^*)^{-1}K_{(n)})(\varphi_* - \varphi_0)}^{I} \\ &\quad + \underbrace{\Omega_0 K_{(n)}^*(\alpha_n I + \frac{1}{n}I + K_{(n)}\Omega_0 K_{(n)}^*)^{-1}\varepsilon_{(n)}}_{II}.\end{aligned}$$

We start by term I:

$$\begin{aligned}||I||^2 &\leq \overbrace{||(I - \Omega_0 K_{(n)}^*(\alpha_n I + K_{(n)}\Omega_0 K_{(n)}^*)^{-1}K_{(n)})(\varphi_* - \varphi_0)||^2}^{IA} \\ &\quad + \underbrace{||\Omega_0 K_{(n)}^*(\alpha_n I + \frac{1}{n}I + K_{(n)}\Omega_0 K_{(n)}^*)^{-1}\frac{1}{n}I(\alpha_n I + K_{(n)}\Omega_0 K_{(n)}^*)^{-1}K_{(n)})(\varphi_* - \varphi_0)}_{IB}\end{aligned}$$

and by permuting operators, $||IA||^2$ is shown to be equivalent to

$$||\Omega_0^{\frac{1}{2}}[\alpha_n(\alpha_n I + T^*T)^{-1}\psi + (\alpha_n(\alpha_n I + \hat{T}^*\hat{T})^{-1}\psi - \alpha_n(\alpha_n I + T^*T)^{-1}\psi)]||^2$$

that is less than or equal to

$$||\Omega_0^{\frac{1}{2}}||^2\Big(||\alpha_n(\alpha_n I + T^*T)^{-1}\psi||^2 + ||(\alpha_n I + \hat{T}^*\hat{T})^{-1}||^2||\hat{T}^*\hat{T} - T^*T||^2||\alpha_n(\alpha_n I + T^*T)^{-1}\psi||^2\Big).$$

In particular, if $\psi \in \mathcal{R}(T^*T)^{\beta/2}$ then $||\alpha_n(\alpha_n I + T^*T)^{-1}||^2 \sim \mathcal{O}_p(\alpha_n^\beta)$, see [4]. Therefore, $||IA||^2 \sim \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta)$.
Term $IB = \Omega_0^{\frac{1}{2}}\hat{T}^*(\alpha_n I + \frac{1}{n}I + \hat{T}\hat{T}^*)^{-1}(\frac{1}{n}I)(\alpha_n I + \hat{T}\hat{T}^*)^{-1}\hat{T}\psi$ is negligible with respect to IA, in fact, by permuting operators in a similar way as above, we get that $||IB||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^4 n^2}(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta))$ that goes to zero if $||IA||^2 \to 0$.

Let consider now term II. An analogous decomposition as for I gives

$$\begin{aligned}||II||^2 &\leq ||\Omega_0^{\frac{1}{2}}||^2\Big(||\underbrace{\hat{T}^*(\alpha_n I + \hat{T}\hat{T}^*)^{-1}\varepsilon_{(n)}}_{IIA}||^2 + ||\underbrace{\hat{T}^*(\hat{T}\hat{T}^* + \alpha_n I + \frac{1}{n}I)^{-1}(\frac{1}{n}I)(\hat{T}\hat{T}^* + \alpha_n I)^{-1}\varepsilon_{(n)}}_{IIB}||^2\Big) \\ ||IIA||^2 &\leq ||(\alpha_n I + \hat{T}^*\hat{T})^{-1}||^2||\hat{T}^*\varepsilon_{(n)}||^2,\end{aligned}$$

where $\hat{T}^*\varepsilon_{(n)} = \frac{1}{\sqrt{n}}\Big[\frac{1}{\sqrt{n}}\sum_i \varepsilon_i g(Z, w_i)\Big]$. By Central Limit Theorem (CLT) the term into squared brackets is bounded because converges toward a normal random variable; then $||\hat{T}^*\varepsilon_{(n)}||^2 \sim \mathcal{O}_p(\frac{1}{n})$ and $||IIA||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^2 n})$ since $||(\alpha_n I + \hat{T}^*\hat{T})^{-1}||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n})$ because $\hat{T}^*\hat{T}$ converges faster than α_n.
Term IIB accounts for the covariance operator $\frac{1}{n}I$ of the sampling probability and, due to the fact that $\frac{1}{n}$ converges to zero faster than α_n, it is negligible with respect to IIA. Its squared norm is equivalent to

$$||(\hat{T}^*\hat{T} + \alpha_n I + \frac{1}{n}I)^{-1}(\frac{1}{n}I)(\hat{T}^*\hat{T}^* + \alpha_n I)^{-1}\hat{T}^*\varepsilon_{(n)}||^2$$

that goes to zero at the speed of $(\frac{1}{\alpha_n^2 n^2}\frac{1}{\alpha_n^2 n})$.
Summarizing $||\hat{\varphi}_\alpha - \varphi_*||^2 \sim \mathcal{O}_p((\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta)(1 + \frac{1}{\alpha_n^4 n^2}) + \frac{1}{\alpha_n^2 n}(1 + \frac{1}{\alpha_n^2 n^2}))$ that, simplifying the

term that are negligible becomes $\mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta + \frac{1}{\alpha_n^2 n})$.

Derivation of the speed of convergence of the covariance operator $\Omega_{y,\alpha}$ is essentially similar. We apply this operator to an element $\phi \in L_F^2(Z)$ and we decompose it into two terms (one including $\frac{1}{n}I$ and an other one not including it):

$$\begin{aligned}\Omega_{y,\alpha}\phi &= \sigma^2\Big(\overbrace{[\Omega_0 - \Omega_0^{\frac{1}{2}}\hat{T}^*(\alpha_n I + \hat{T}\hat{T}^*)^{-1}\hat{T}\Omega_0^{\frac{1}{2}}]\phi}^{A} \\ &\quad + \underbrace{\Omega_0^{\frac{1}{2}}\hat{T}^*[(\alpha_n I + \hat{T}\hat{T}^*)^{-1} - (\alpha_n I + \frac{1}{n}I + \hat{T}\hat{T}^*)^{-1}]\hat{T}\Omega_0^{\frac{1}{2}}\phi}_{B}\Big)\end{aligned}$$

We have to consider the squared norm in L_F^2 of $\Omega_{y,\alpha}\phi$: $||\Omega_{y,\alpha}\phi||^2 \leq |\sigma^2|^2 \cdot (||A||^2 + ||B||^2)$. By Kolmogorov's theorem $|\sigma^2|^2 \sim \mathcal{O}_p(\delta)$ if and only if $\mathbb{E}[(\sigma^2)^2|y_{(n)}] \sim \mathcal{O}_p(1)$. Since the second moment of σ^2 is $\mathbb{E}[(\sigma^2)^2|y_{(n)}] = Var(\sigma^2|y_{(n)}) + \mathbb{E}^2(\sigma^2|y_{(n)})$, it follows from Theorem 3 that $|\sigma^2|^2 \sim \mathcal{O}_p(1)$. Concerning term A we have

$$\begin{aligned}||A||^2 &\leq ||\Omega_0^{\frac{1}{2}}||^2||(I - \hat{T}^*(\alpha_n I + \hat{T}\hat{T}^*)^{-1}\hat{T})\Omega_0^{\frac{1}{2}}\phi||^2 \\ &\leq ||\Omega_0^{\frac{1}{2}}||^2||(I - (\alpha_n I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T})\Omega_0^{\frac{1}{2}}\phi||^2 \\ &\leq ||\Omega_0^{\frac{1}{2}}||^2||\alpha_n(\alpha_n I + \hat{T}^*\hat{T})^{-1}\Omega_0^{\frac{1}{2}}\phi||^2 \\ &\leq ||\Omega_0^{\frac{1}{2}}||^2\Big(||\alpha_n(\alpha_n I + T^*T)^{-1}\Omega_0^{\frac{1}{2}}\phi||^2 + |||[\alpha_n(\alpha_n I + \hat{T}^*\hat{T})^{-1} - \alpha_n(\alpha_n I + T^*T)^{-1}]\Omega_0^{\frac{1}{2}}\phi||^2\Big) \\ &= ||\Omega_0^{\frac{1}{2}}||^2\Big(||\alpha_n(\alpha_n I + T^*T)^{-1}\Omega_0^{\frac{1}{2}}\phi||^2 + ||(\alpha_n I + \hat{T}^*\hat{T})^{-1}(\hat{T}^*\hat{T} - T^*T)\alpha_n(\alpha_n I + T^*T)^{-1}]\Omega_0^{\frac{1}{2}}\phi||^2\Big)\end{aligned}$$

and $||\alpha_n(\alpha_n I + T^*T)^{-1}\Omega_0^{\frac{1}{2}}\phi||^2 \sim \mathcal{O}_p(\alpha_n^\beta)$ if $\Omega_0^{\frac{1}{2}}\phi \in \mathcal{R}(T^*T)^{\frac{\beta}{2}}$. Moreover, the second term in brackets is an $\mathcal{O}_p(\frac{1}{\alpha_n^2 n}\alpha_n^\beta)$ and $||\Omega_0^{\frac{1}{2}}||^2 \sim \mathcal{O}_p(1)$ since Ω_0 is a compact operator, so we get $||A||^2 \sim \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta)$.
Lastly, term B is equivalent to term IB in the mean decomposition above, except that ψ is substituted by $\Omega_0^{\frac{1}{2}}\phi$, but this does not alter the speed of convergence. Hence, $||B||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^4 n^2}(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta))$. Summarizing, $||\Omega_{y,\alpha}||^2 \sim \mathcal{O}_p((1 + \frac{1}{\alpha_n^4 n^2})(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta))$ that, once neglected the fastest terms becomes $\mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta)$.

7.3 Proof of Theorem 2

Both points *(i)* and *(ii)* in the Theorem are a consequence of Corollary 1 and Chebishev's Inequality. More clearly, we have

$$\begin{aligned}\mu_\alpha^{\sigma,\mathcal{F}}\{\varphi : ||\varphi - \varphi_*|| \geq \epsilon_n\} &\leq \frac{\mathbb{E}_\alpha(||\varphi - \varphi_*||^2|\sigma^2, y_{(n)})}{\epsilon_n^2} \\ &\leq \frac{1}{\epsilon_n^2}(||Var(\varphi|\sigma^2, y_{(n)})||^2 + ||\mathbb{E}_\alpha(\varphi|\sigma^2, y_{(n)}) - \varphi_*||^2)\end{aligned}$$

and the result follows.

7.4 Proof of Theorem 4

We start by decomposing the estimation error in fourth parts:

$$||\hat{\varphi}_\alpha - \varphi_*||^2 \leq ||\hat{\varphi}_\alpha - \mathbb{E}_\alpha(\varphi|y_{(n)}, \mathbf{w})||^2 + ||\mathbb{E}_\alpha(\varphi|y_{(n)}, \mathbf{w}) - \hat{\varphi}_\alpha^{\hat{\theta}}||^2 + ||\hat{\varphi}_\alpha^{\hat{\theta}} - \hat{\varphi}_\alpha^{\theta_*}||^2 + ||\hat{\varphi}_\alpha^{\theta_*} - \varphi_*||^2,$$

where $\hat{\varphi}_\alpha^{\hat{\theta}} = \mathbb{E}_\alpha(\varphi|\hat{\theta}, y_{(n)}, \mathbf{w})$ and $\hat{\varphi}_\alpha^{\theta_*} = \mathbb{E}_\alpha(\varphi|\theta_*, y_{(n)}, \mathbf{w})$. For brevity, we have suppressed the subscript $L_F^2(Z)$ in the norm, being implied that it is the norm in this space. The first term is the

error due to Monte Carlo approximation of (8) and it is negligible as $J \to \infty$. The second error term is due to having integrated out θ instead of to set it equal to the posterior mean. The third one accounts for the estimation error of θ and the last term is the usual regularization bias due to the fact that we approximate parameter φ with a regularized version of the posterior mean and it converges to 0 at the speed given in Theorem 2. We shall show that the other two terms are converging at a faster speed and then are negligible.
We start with the second one. Note that $\mathbb{E}_\alpha(\varphi|y_{(n)}, \mathbf{w}) = \int \hat{\varphi}_\alpha^\theta \rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta$, then

$$
\begin{aligned}
||\mathbb{E}_\alpha(\varphi|y_{(n)}, \mathbf{w}) - \hat{\varphi}_\alpha^{\hat{\theta}}||^2 &= \int \Big(\int (\hat{\varphi}_\alpha^\theta(Z) - \hat{\varphi}_\alpha^{\hat{\theta}}(Z)\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta \Big)^2 f(Z, \cdot|\theta_*)dZ \quad (12)\\
&\leq \int \Big(\hat{\varphi}_\alpha^\theta(Z) - \hat{\varphi}_\alpha^{\hat{\theta}}(Z)\Big)^2 \rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta \; f(Z, \cdot|\theta_*)dZ \quad (13)\\
&\approx trVar(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}}) \int \Big(\frac{\partial \hat{\varphi}_\alpha^{\hat{\theta}}}{\partial \theta}\frac{\partial \hat{\varphi}_\alpha^{\hat{\theta}}}{\partial \theta'}\Big)(Z) f(Z, \cdot|\theta_*)dZ \quad (14)\\
&\sim \mathcal{O}_p(\frac{1}{\tilde{n}}) \quad (15)
\end{aligned}
$$

if $\frac{\partial \hat{\varphi}_\alpha^{\hat{\theta}}}{\partial \theta} \in L_F^2(Z)$. The approximated equality has been obtained through a first order Taylor expansion of $\hat{\varphi}_\alpha^\theta$ around the posterior mean $\hat{\theta}$.
Consider now the third error term. A first order Taylor expansion around the true value θ_* gives:

$$
\hat{\varphi}_\alpha^{\hat{\theta}} \approx \hat{\varphi}_\alpha^{\theta_*} + \frac{\partial \hat{\varphi}_\alpha^{\theta_*}}{\partial \theta}(\hat{\theta} - \theta_*).
$$

Classical results in Bayesian statistic (see e.g. [1], [14] or [24]) show that, under some regularity conditions that we assume to be satisfied, $\sqrt{N}||\hat{\theta} - \theta_*|| \sim \mathcal{O}_p(1)$, that implies

$$
\begin{aligned}
||\hat{\varphi}_\alpha^{\hat{\theta}} - \hat{\varphi}_\alpha^{\theta_*}||^2 &\leq ||\frac{\partial \hat{\varphi}_\alpha(Z, \theta_*)}{\partial \theta}||^2 ||(\hat{\theta} - \theta_*)||^2\\
&\sim \mathcal{O}_p(\frac{1}{\tilde{n}})
\end{aligned}
$$

if $\frac{\partial \hat{\varphi}_\alpha^{\theta_*}}{\partial \theta} \in L_F^2(Z)$. The result follows.

7.5 Proof of Theorem 5

In order to show convergence to 0 of $\hat{\Omega}_{y,\alpha}$ we decompose it in several terms and study each of them separately. First of doing it, note that second term in the RHS of (9) must be interpreted in the following way:

$$
\begin{aligned}
Var(\mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w})|\tilde{\mathbf{z}}, \tilde{\mathbf{w}}) = &\overbrace{\int\int \mathbb{E}_\alpha(\varphi(z)|\theta, y_{(n)}, \mathbf{w})\mathbb{E}_\alpha(\varphi(\zeta)|\theta, y_{(n)}, \mathbf{w})\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta \; g(z|\theta_*)dz}^{C_1}\\
&- \int\int \mathbb{E}_\alpha(\varphi(z)|\theta, y_{(n)}, \mathbf{w})\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta \cdot\\
&\underbrace{\int \mathbb{E}_\alpha(\varphi(\zeta)|\theta, y_{(n)}, \mathbf{w})\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta \; f(z|\theta_*)dz}_{C_2}
\end{aligned}
$$

Let $\phi \in L_F^2(Z)$ be such that $\Omega_0^{\frac{1}{2}}\phi \in \mathcal{R}(\lim \Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$, then

$$
||\hat{\Omega}_{y,\alpha}\phi||^2 \leq ||\hat{\Omega}_{y,\alpha}\phi - Var_\alpha(\varphi|y_{(n)}, \mathbf{w})\phi||^2 + |||[\int Var_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w})\rho(\theta|(s_{2,i})_{i=1,\dots,n}) - Var_\alpha(\varphi|\hat{\theta}, y_{(n)}, \mathbf{w})]\phi||^2
$$

$$+||C_1\phi - \overbrace{\int \mathbb{E}_\alpha(\varphi(z)|\hat{\theta}, y_{(n)}, \mathbf{w})\phi(z)\mathbb{E}_\alpha(\varphi(\zeta)|\hat{\theta}, y_{(n)}, \mathbf{w})\phi(z)f(z|\theta_*)dz}^{E}||^2$$
$$+||C_2\phi - E||^2 + ||Var_\alpha(\varphi|\hat{\theta}, (y_{(n)}, \mathbf{w}))\phi - Var_\alpha(\varphi|\theta_*, (y_{(n)}, \mathbf{w}))\phi||^2$$
$$+||\Omega_{y_{(n)},\alpha}(\theta_*)\phi||^2$$

with $\Omega_{y,\alpha}(\theta_*)$ the covariance operator of the regularized posterior distribution $\mu_\alpha^{\mathcal{F}}$ when F is known. The first term is the error due to Monte Carlo integration, therefore is negligible assuming that we are taking a large number of discretization points drawn from $\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})$ and $f(z|\theta_*)$. For simplicity, we rewrite $Var_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w})$ as $\Omega_{y,\alpha}(\theta)$, then the second error term becomes:

$$|| \int [\Omega_{y,\alpha}(\theta) - \Omega_{y,\alpha}(\hat{\theta})]\phi\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta||^2$$

that is equal to

$$\begin{aligned}
& \int \Big(\int [\Omega_{y,\alpha}(\theta)\phi - \Omega_{y,\alpha}(\hat{\theta})\phi](\zeta)\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta\Big)^2 f(\zeta, \cdot|\theta_*)d\zeta \\
\leq & \int\int [\Omega_{y,\alpha}(\theta)\phi - \Omega_{y,\alpha}(\hat{\theta})\phi]^2(\zeta)\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta\, f(\zeta, \cdot|\theta_*)d\zeta \\
\approx & \; trVar(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}}) \int \frac{\partial\Omega_{y,\alpha}(\hat{\theta})\phi}{\partial\theta}\frac{\partial\Omega_{y,\alpha}(\hat{\theta})\phi}{\partial\theta'}(\zeta)f(\zeta, \cdot|\theta_*)d\zeta \\
\sim & \; \mathcal{O}_p(\frac{1}{\tilde{n}}) \qquad \text{if } \frac{\partial\Omega_{y,\alpha}(\hat{\theta})\phi}{\partial\theta} \in L_F^2(Z).
\end{aligned}$$

Moreover,

$$\begin{aligned}
||(C_1 - E)\phi||^2 \leq & \int\int\int [\mathbb{E}_\alpha(\varphi(z)|\theta, y_{(n)}, \mathbf{w})\mathbb{E}_\alpha(\varphi(\zeta)|\theta, y_{(n)}, \mathbf{w}) \\
& -\mathbb{E}_\alpha(\varphi(z)|\hat{\theta}, y_{(n)}, \mathbf{w})\mathbb{E}_\alpha(\hat{\varphi}(\zeta)|\theta, y_{(n)}, \mathbf{w})]^2\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})d\theta\, \phi^2(z)g(z, \cdot|\theta_*)f(\zeta, \cdot|\theta_*)dzd\zeta \\
\approx & \; trVar(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}}) \int\int \frac{\partial\mathbb{E}_\alpha(\varphi(z)|\hat{\theta}, y_{(n)}, \mathbf{w})\mathbb{E}_\alpha(\varphi(\zeta)|\hat{\theta}, y_{(n)}, \mathbf{w})}{\partial\theta} \cdot \\
& \frac{\partial\mathbb{E}_\alpha(\varphi(z)|\hat{\theta}, y_{(n)}, \mathbf{w})\mathbb{E}_\alpha(\varphi(\zeta)|\hat{\theta}, y_{(n)}, \mathbf{w})}{\partial\theta'}\phi^2(z)f(z, \cdot|\theta_*)f(\zeta, \cdot|\theta_*)dzd\zeta \\
\sim & \; \mathcal{O}_p(\frac{1}{n}) \qquad \text{if}
\end{aligned}$$

$$\begin{aligned}
||(C_2 - E)\phi||^2 \leq & \int\int\int (h(t,\tau) - h(\hat{\theta}, \hat{\theta}))^2(z, \zeta)\rho(t|(s_{2,i})_{i=1,\dots,\tilde{n}})\rho(\tau|(s_{2,i})_{i=1,\dots,\tilde{N}})dtd\tau\phi^2(z)f(z, \cdot|\theta_*)f(\zeta, \cdot|\theta_*)dzd\zeta \\
\approx & \; trVar(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})\Big(\int\int \frac{\partial h(\hat{\theta}, \hat{\theta})}{\partial t}\frac{\partial h(\hat{\theta}, \hat{\theta})}{\partial t'}(z, \zeta)\phi^2(z)f(z, \cdot|\theta_*)f(\zeta, \cdot|\theta_*)dzd\zeta \\
& + \int\int \frac{\partial h(\hat{\theta}, \hat{\theta})}{\partial \tau}\frac{\partial h(\hat{\theta}, \hat{\theta})}{\partial \tau'}(z, \zeta)\phi^2(z)f(z, \cdot|\theta_*)f(\zeta, \cdot|\theta_*)dzd\zeta\Big) \\
\sim & \; \mathcal{O}_p(\frac{1}{n}) \quad \text{if}
\end{aligned}$$

with $h(t, \tau) = \mathbb{E}_\alpha(\varphi|t, y_{(n)}, \mathbf{w})\mathbb{E}_\alpha(\varphi|\tau, y_{(n)}, \mathbf{w})$. Using the same notation as before the fifth error term is

$$\begin{aligned}
||\Omega_{y,\alpha}(\hat{\theta})\phi - \Omega_{y,\alpha}(\theta_*)\phi||^2 = & \int [(\Omega_{y,\alpha}(\hat{\theta})\phi)(\zeta) - (\Omega_{y,\alpha}(\theta_*)\phi)(\zeta)]^2 f(\zeta, \cdot|\theta_*)d\zeta \\
\approx & \; tr(\hat{\theta} - \theta_*)(\hat{\theta} - \theta_*)' \int \frac{\partial(\Omega_{y,\alpha}(\theta_*)\phi)(\zeta)}{\partial\theta}\frac{\partial(\Omega_{y,\alpha}(\theta_*)\phi)(\zeta)}{\partial\theta'}f(\zeta, \cdot|\theta_*) \\
\sim & \; \mathcal{O}_p(\frac{1}{n}) \qquad \text{if } \frac{\partial\Omega_{y,\alpha}(\theta_*)\phi}{\partial\theta} \in L_F^2(Z).
\end{aligned}$$

Note that all the approximated equalities in previous terms are obtained thanks to a first order Taylor expansion.
Therefore, all these error terms are negligible with respect to $||\Omega_{y,\alpha}(\theta_*)||^2$ which is an $\mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^4 n^4}\alpha_N^{(\beta+1)\wedge 2})$ as is shown in Theorem 2 and this proves the result.

References

[1] Bernestein, S. (1934), *Theory of Probability*. Moscow. (Russian).

[2] Choudhuri, N., Ghosal, S., and A., Roy (2005), *Bayesian Methods for Function Estimation*, Handbook of Statistics, Vol. 25, 377-418.

[3] Diaconis, F., and D., Freedman (1986), *On the Consistency of Bayes Estimates*, Annals of Statistics, **14**, 1-26.

[4] Carrasco, M., Florens, J.P., and E., Renault (2007), *Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization*, Hnadbook of Econometrics, **Vol.6**, Part 2, 5633-5751.

[5] Darolles, S., Florens, J.P., and E., Renault (2006), *Nonparametric Instrumental Regression*, Econometrica (forthcoming).

[6] Ferguson, T.S. (1973), *A Bayesian analysis of some nonparametric problems*, Annals of Statistics, **1**, 209-230.

[7] Ferguson, T.S. (1974), *Prior distribution on the spaces of probabilities measures*, Annals of Statistics, **2**, 615-629.

[8] Florens, J.P. (2002), *Inverse Problems and Structural Econometrics: the Example of Instrumental Variables*, Invited Lectures to the World Congress of the Econometric Society, Seattle 2000.

[9] Florens, J.P., Mouchart, M., and J.M., Rolin (1990), *Elements of Bayesian Statistics*, Dekker, New York.

[10] Florens, J.P., and A., Simoni (2007), *Regularizzed Posteriors in Linear Ill-Posed Inverse Problems*, working paper.

[11] Franklin, J.N. (1970), *Well-posed stochastic extension of ill-posed linear problems*, Journal of Mathematical Analysis and Applications, **31**, 682 - 716.

[12] Freedman, D. (1965), *On the Asymptotic Behavior of Bayes Estimates in the Discrete Case II.*, Ann. Math. Statist., **36**, 454-456.

[13] Gelman, A. and D.B., Rubin (1992), *Inference from Iterative Simulation Using Multiple Sequences*, Statistical Science, **7**, 457 - 472.

[14] Ghosh, J.K and R.V. Ramamoorthi (2003), *Bayesian Nonparametrics*. Springer Series in Statistics.

[15] Ghosal, S., *A review of consistency and convergence rates of posterior distribution*, in Proceedings of Varanashi Symposium in Bayesian Inference, Banaras Hindu University.

[16] Hall, P. and J., Horowitz (2005), *Nonparametric Metyhods for Inference in the Presence of Instrumental Variables*, Annals of Statistics, **33**, 2904-2929.

[17] Lavine, M. (1992), *Some aspects of Polya tree distributions for statistical modeling*, Annals of Statistics, **20**, 1222-1235.

[18] Lehmann, E.L. (1997), *Theory of point esimation*. Springer-Verlag, New York. Reprint of the 1983 original.

[19] Mandelbaum, A. (1984), *Linear Estimators and Measurable Linear Transformations on a Hilbert Space*, Z. Wahrcheinlichkeitstheorie, **3**, 385-98.

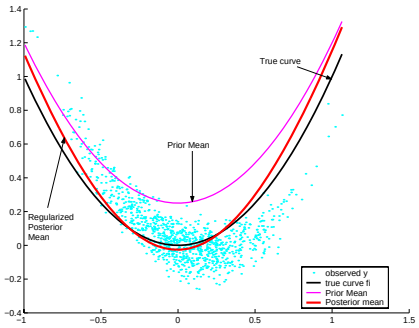
[20] Neveu, J. (1965), *Mathematical Fundations of the Calculus of Probability*, San Francisco: Holden-Day.

[21] Newey, W.K. and J.L., Powell (2003), *Instrumental Variable Estimation of Nonparametric Models*, Econometrica, Vol.71, **5**, 1565-1578.

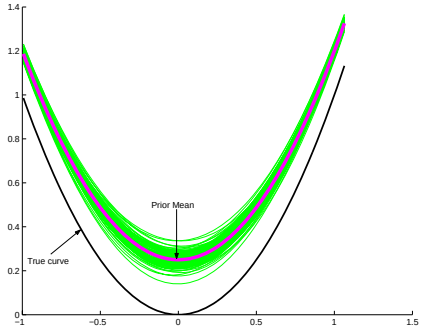
[22] Newey, W.K., Powell, J. and F.Vella (1999), *Nonparametric Estimation of Triangular Simultaneous Equations Models*, Econometrica, **67**, 565 - 604.

[23] Rasmussen, C.E. and C., Williams (2006), *Gaussian Processes for Machine Learning*, the MIT Press.

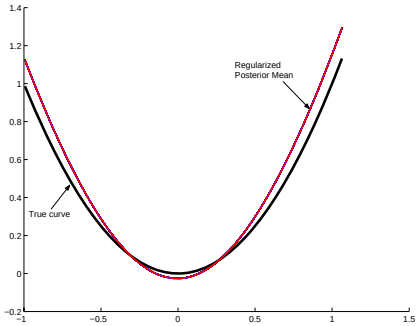
[24] Von Mises, R. (1964), *Mathematical Theory of Probability and Statistics*. H. Geiringer ed. Academic, New York.



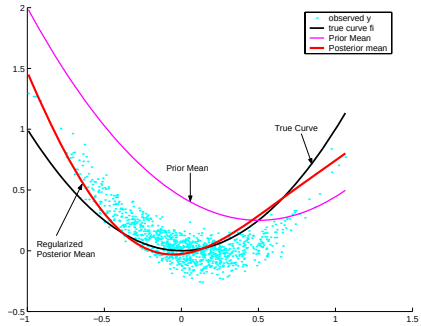
(a) $\varphi_0(Z) = 0.95Z^2 + 0.25$, $(\Omega_0\phi)(Z) = 0.5\int \exp(-(s-Z)^2)\phi(s)f_z(s)ds$, $\alpha_N = 0.3$, $N = 1000$



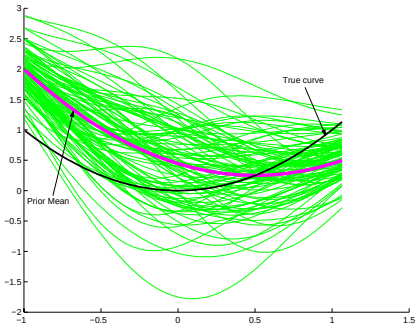
(b) Sample drawn from the prior of φ



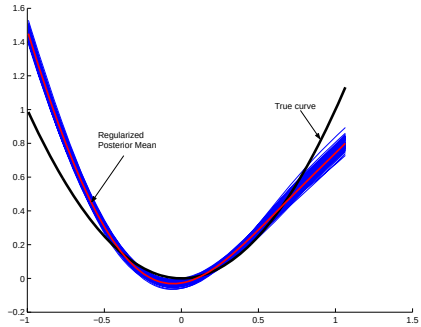
(c) Sample drawn from the regularized posterior of φ



(d) $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$, $(\Omega_0\phi)(Z) = 200\int \exp(-(s-Z)^2)\phi(s)f_z(s)ds$, $\alpha_N = 0.3$, $N = 1000$

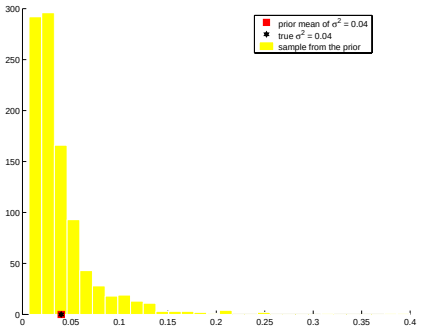


(e) Sample drawn from the prior of φ

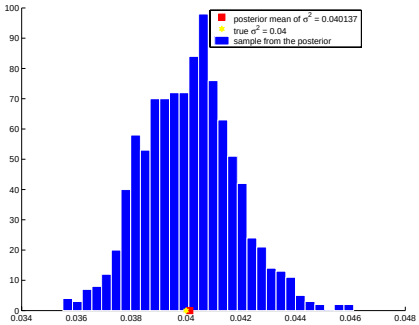


(f) Sample drawn from the regularized posterior of φ

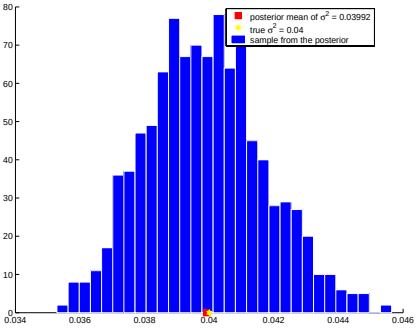
Figure 1: Case I. *Conjugate Model with $f(Z,W)$ known.* Graphs (1a) - (1c) are for $\varphi_0(Z) = 0.95Z^2 + 0.25$ and $\sigma_0 = 0.5$; graphs (1d) - (1f) are for $\frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$ and $\sigma_0 = 200$



(a) Sample drawn from the prior of σ^2

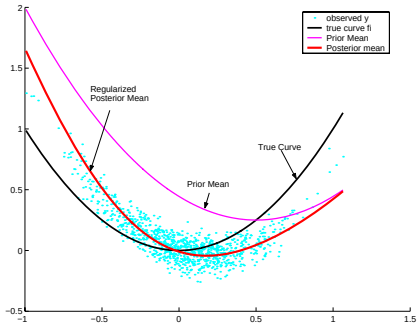


(b) Sample drawn from the posterior of σ^2, for $\varphi_0(Z) = 0.95Z^2 + 0.25$ and $\sigma_0 = 0.5$

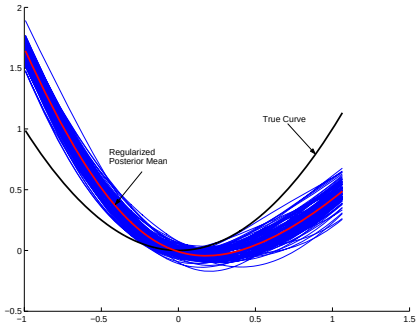


(c) Sample drawn from the posterior of σ^2, for $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$ and $\sigma_0 = 200$

Figure 2: CASE I. *Conjugate Model with $f(Z, W)$ known.*



(a) Regularized Posterior Mean of φ for $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$, $\sigma_0 = 200$



(b) Sample drawn from the regularized posterior of φ

Figure 3: CASE II. *$f(Z, W)$ unknown and σ^2 known*