Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA
IN ECONOMIA

Ciclo XX

Settore scientifico disciplinare di afferenza: SECS-P/05 - ECONOMETRIA.

Bayesian Analysis of Linear Inverse Problems with Applications in Economics and Finance

Presentata da: Anna SIMONI

Coordinatore Dottorato:
Andrea ICHINO

Relatore:
Sergio PASTORELLO

Esame Finale Anno 2009

Contents

List of Figures

Chapter 1

Introduction

Recent econometric theory has developed an increasing interest in nonparametric estimation of structural parameters of economic models. The (possibly functional) parameter of interest describes the economic agent's behavior or the equilibrium market's characteristics. In other words, structural econometrics has not for purpose the estimation of objects directly linked to the data's distribution F, like the density or the hazard function, or the estimation of some characteristics of the data's conditional distribution, such as the conditional expectation. The structural parameters have not a statistical interpretation and, in general, they are not simple transformations of the sampling distribution of the data.

The relationship between structural parameters and the sampling distribution F is, in most of the cases, only implicitly defined through a functional equation. Hence, the structural parameter is characterized as solution of this functional equation. When the dimension of the structural parameter is finite, the functional equation reduces to a matrix equation.

Several authors, see, among others, Florens (2003), Hall and Horowitz (2005), Linton and Mammen (2005), Carrasco, Florens and Renault (2007), have developed a general framework for structural functional inference in connection with the inverse problem literature. A complete list of references can be found in Carrasco, Florens and Renault (2007).

Economic theory can provide information about the shape (like convexity, concavity) or the differentiability of the parameter of interest, but never provides a parametric form for it. Hence, it is really suitable to not restrict inference to parametric classes. Nonparametric estimation reduces the risk of mispecification but, at the same time, rises problems of continuity, uniqueness and existence of the solution of the corresponding functional equation, so that some care must be taken in solving it. Such a problem is known as the *ill-posedness* of the inverse problem we want to solve. The lack of continuity of the solution entails a strong sensibility of the solution to the estimation errors of certain elements of the functional equation. The noises in the functional equation arise because some elements in it can be (and usually are) imperfectly known (like for instance the sampling distribution) and they are replaced by consistent estimators. Hence, small estimation errors can be strongly amplified in the estimated parameter.

Classical econometric literature deals with these problems by proposing different techniques of stabilization of the solution. Classical stabilization techniques consist in replacing the non-continuous estimator with an approximation of it that is continuous and that converges in the sense that, as the noise level in the functional equation tends to zero, the approximated estimator tends to the true one.

My work develops bayesian nonparametric methods to estimate structural economic quantities. It analyzes the role played by the prior distribution in solving these problems of continuity and existence of the solution of the functional equation.

Bayesian analysis considers an inverse problem in a different way with respect to the classical analysis since it restates the functional equation in a larger space of probability distributions. This reformulation of an inverse problem was due to Franklin (1970) [33]. From a Bayesian point of view all the quantities in the structural model are random functions. Hence, the structural parameter of interest having generated the data is the realization of some random process. I substantially look for an estimation of such process and I take the posterior distribution of the parameter as the estimator.

When the dimension of the problem is finite the ill-posedness is principally due to a problem of multicollinearity. In this case, classical and Bayesian approaches are strongly related since the *ridge regression* (that is a classical method proposed for dealing with multicollinearity) has a Bayesian interpretation. Therefore, in finite dimension we can remove the ill-posedness by incorporating the available prior information. On the contrary, I prove that in infinite dimensional problems, a general prior distribution does not get ride of the ill-posedness of the problem since general prior covariance matrices do not have the regularization properties that have in the finite dimensional case. In particular, being covariance matrices impossible to continuously inverse we still need some regularization technique and the bayesian approach only lies in changing the nature of the problem.

Nevertheless, an exciting result of my work is that there exists a class of prior distributions, or more precisely, of covariance operators, that are able to solve for the ill-posedness as in the finite dimensional case.

I believe that a bayesian approach for solving functional equations is suitable for many reasons that I discuss in the following. *(i)* It is very important to incorporate, in the estimation procedure, the information that we may have *a-priori* on the structural parameter that we want to recover, for instance shape information or other constraints given by economic theory. This can be easily done through a Bayesian procedure. We could choose a prior mean function of the form suggested by the economic theory. Otherwise, we can specify a prior covariance operator that restricts the space in which the solution can lie to a subspace of functions satisfying the constraints we want to impose. The importance of incorporating the prior information that is available is not to underestimate, above all in nonparametric estimation where the parameter is often weakly identified by data because the amount of data is small with respect to the dimension of the parameter to estimate (although it is identified from a mathematical point of view). Moreover, in the most applied research, for instance in finance or in consulting studies, people are strongly inclined to exploit all the prior information that they can have, like the opinion of some expert.

(ii). The fact to get a posterior distribution of the structural parameter of interest represents a big advantage with respect to classical estimation procedures. The posterior distribution has good small sample properties and so it can be used for recovering every quantity linked to it (as quantiles and credible sets) and for implementing testing procedures. On the contrary, classical procedures give punctual estimators that have good properties asymptotically and not in small samples. Moreover, the proof of such properties is sometimes very demanding and very complicate.

(iii). In nonparametric estimation there usually are some free parameters (*i.e.* tuning parameters) to choose, like the bandwidth in kernel estimation or the regularization parameter in stabilization techniques. Some methods for choosing them are provided by the existing theory, but bayesian theory could give some further insight, from a practical point of view, for optimally choosing them. In particular, the prior-to-posterior transformation would provide a value for the tuning parameter that incorporates information in both our prior knowledge and data.

(iv). Bayesian nonparametric analysis of structural models broadens the nonparametric

estimation techniques available to bayesian statisticians. In fact, I assume the error in the functional equation is gaussian and this suggests to use a conjugate model where the prior distribution is also a gaussian process. This makes the bayesian nonparametric estimation that I propose different from the usual bayesian nonparametric estimation based on Dirichlet process and its transformations. A gaussian prior distribution has the big advantage, with respect to the Dirichlet process, that it is able to generate trajectories that are continuous. Furthermore, being the model conjugate, the posterior distribution is still gaussian and all the interesting quantities of a gaussian distribution (as quantiles or confidence intervals) are well-known. The gaussian prior measure cannot be substituted by a non-informative prior since we are working in infinite dimensional spaces and it does not exists an invariant non-informative distribution on such spaces.

(v). Developing a Bayesian approach for structural estimation permits to create a link between econometric theory and the *Machine Learning and Computer Science theory* where huge amount of data are available. In this field the object of interest is a functional parameter and it is estimated with bayesian techniques by specifying a gaussian prior on it.

The infinite dimension of the problem we are considering makes the Bayesian analysis interesting *per se* from a theoretical point of view. In fact, properties that are usually satisfied by bayesian estimators in finite dimension do not necessary hold when the dimension of the problem is infinite. The infinite dimension of the objects I am working with demands to verify properties like *admissibility* of the estimator and in particular *posterior consistency*. A Bayesian estimator is consistent in the bayesian sense, *i.e.* with respect to the joint distribution, even in infinite dimensional problems, but it does not need to be consistent in the frequentist, or sampling, sense, *i.e.* with respect to the sampling distribution. Frequentist consistency is also referred to as *posterior consistency* and it will be of peculiar interest in my study. The concept of posterior consistency implies the assumption that a true value of the structural parameter of interest that generates the data exists; it can be interpreted as the realization of a random process that has occurred in a very preliminary step. We say that the posterior distribution is consistent if, as the number of observations increases, it degenerates in a point mass in correspondence of the true value of the parameter.

Nonparametric structural estimation represents a good example of lack of posterior consistency in infinite dimensional problems under very general assumptions. I prove this fact, that is substantially due to the lack of continuity stressed before, and I confirm what other authors have already found in situations that are, for some aspects, similar to the mine, see Diaconis *et al.* (1986) and Mandelbaum (1984).

One of the main contributions of my work consists in proposing a new posterior distribution, for the gaussian model, that is consistent in the sampling sense and that is based on regularization techniques, like Tikhonov regularization, similar to the one used for solving functional equations in a classical way. I call this new posterior distribution *regularized posterior distribution*.

Even if this approach holds for very general gaussian sampling models and priors, it is possible to find a class of gaussian prior distribution for which the posterior distribution is consistent, in the sampling sense, even without regularizing it. I detect such a class and I study the particular conditions that must be verified in order to have this nice result. The idea is to specify a prior distribution that degenerates, at a suitable speed, towards a point mass concentrated on the prior mean as the sample size increases. Moreover, it must exist a certain link between the prior covariance operator and the sampling model. Under such conditions, we do not need to regularize the posterior distribution since the prior-to-posterior transformation has a regularizing effect of the same type as the Tikhonov

regularization in the Hilbert Scale induced by the prior covariance operator. I exploit the regularizing power of the prior distribution as it happens in the finite dimensional case (see, for instance, the ridge regression to correct for multicollinearity). Therefore, when the prior distribution is linked to the sample size and particular conditions are satisfied, the stabilization procedure that in classical nonparametric estimation of structural models is done through regularization schemes, in the bayesian approach is simply obtained by incorporating the prior distribution. In these particular cases the prior distribution is linked to the sample size n through a function $\frac{1}{g}$ that scales the prior variance, with $g = g(n)$ and $g \to \infty$ as $n \to \infty$. This prior distribution is an extension of the g-prior introduced by Zellner (1986) [81]. The parameter g, scaled by the inverse of the sample size, plays the role of a *regularization parameter* and it can be either considered as known or interpreted as an hyperparameter for which a prior must be specified. In this second case, the maximum a posteriori gives an estimator for g that has the optimal speed of convergence.

I compute the rate of convergence, towards the Dirac measure concentrated in the true value of the parameter, of both the regularized posterior distribution and the posterior distribution based on extended g-prior and I make comparison with the classical rates. The rate of convergence obtained when I regularize the posterior distribution through a simple Tikhonov regularization scheme can be considerably improved. In order to improve it, I propose to use a Tikhonov regularization in the Hilbert scale induced by the prior covariance operator. What is noteworthy is that the choice of the Hilbert scale is naturally induced by the prior distribution and I do not need to use some ad-hoc operator. There exists a strong relationship between the choice of the prior distribution and the choice of the Hilbert scale. The convergence of the regularized posterior distribution is faster since the Hilbert Scale regularization allows to exploit all the regularity of the structural function of interest and this impact on a faster speed if the structural function is highly regular.

Finally, the optimal value of the regularization parameter and the optimal rate of convergence are provided. The speed of convergence depends on the regularity conditions, like differentiability, that we are willing to assume about the true value of the structural parameter that we want to estimate. The fact to consider only functions that are quite regular is not astonishing in nonparametric estimation since estimating infinite dimensional parameters with a finite number of observations does not allow to recover too complicate functions. Consequently, assumption of good behavior and smoothness of the function of interest must be introduced.

I start my work by laying the foundations of the bayesian approach to nonparametric structural econometric models. My bayesian theory is stated and developed for infinite dimensional Hilbert spaces In Chapter 2 I state the theory for a general functional equation. This is the classical signal-noise model in a broad gaussian framework. The regularized posterior distribution is defined and is shown to have good asymptotic properties. As example of application of the signal-noise model, I analyze and develop different estimation problems in statistics and econometrics and I rewrite them as functional equations that have the parameter of interest as solution. Examples are the density estimation, regression estimation, deconvolution, hazard rate estimation and instrumental variable estimation. For each of them I compute the exact or asymptotic sampling distribution on Hilbert space.

Furthermore, I analyze extensions and variations of the basic functional equation. In particular, I develop the case in which also the operator in the functional equation is unknown

and estimated. This situation is frequently encountered in econometrics, for instance in the instrumental regression estimation.

In Chapter 3 I analyze the conditions under which the posterior distribution does not need to be regularized in order to be consistent. I show that the prior-to-posterior transformation is equivalent to compute a Tikhonov regularization in Hilbert Scale if we use an extended g-prior distribution, namely a gaussian distribution with a covariance operator strictly linked to the sampling model. Moreover, I am able to introduce the regularization parameter of the Hilbert Scale regularization as a transformation of the g hyperparameter of the prior distribution since it is assumed that g is a function of the sample size.

The second part of this essay focuses on econometric applications of the general theory developed in these preliminary chapters. In Chapter 4 I deal with the estimation of an instrumental regression by exploiting the informative content of instrumental variables and moment restrictions concerning them. I consider a joint prior distribution on the instrumental regression function and on the variance parameter of the reduced form. The marginal posterior distribution of the instrumental regression is more complicated and it is recovered either in a closed form as a Student process or through a Gibbs Sampling implemented in infinite dimensional Hilbert Spaces.

In Chapter 5 I deal with integral equations of type II. These functional equations are different with respect to the type considered in the basic model of Chapters 2 and 3, that is a functional equation of first kind. The proposed bayesian methodology can be applied to estimate every kind of Euler Equations. In particular, I go into the consumption based asset pricing model in the style of Lucas (1978) to recover the equilibrium asset pricing functional. In order to obtain the sampling distribution, I need to transform this model in such a way that, at the end, I get a functional equation of the basic type.

The bayesian nonparametric approach to structural inference that I propose is completely new. In the inverse problems literature some attempts to solve functional equation in a bayesian way are present, see Mandelbaum (1984) or Lehtinen *et al.* (1989). Anyway, the problem of posterior consistency is not considered at all and the lack of continuity is solved in a way that is not applicable in most of empirical problems.

The incorporation of the prior distribution in nonparametric estimation is noteworthy, in particular in economics and finance. This work opens the way for plenty of further development that may be done and I intend to do them in the future.

Chapter 2

Regularized Posterior in linear ill-Posed Inverse Problems[1]

joint with Jean-Pierre Florens.

Abstract

This chapter studies Bayesian solution for a signal-noise problem stated in infinite dimensional Hilbert spaces. The infinite dimensional parameter of interest is characterized as the solution of a functional equation which is ill-posed because of compactness of the operator appearing in it. We restate the problem as a parameter estimation problem where inference is performed in a Bayesian way. The solution of this problem is the posterior distribution of the parameter of interest, but the infinite dimension of the considered spaces causes a problem of non continuity of the posterior mean and a consequent problem of frequentist inconsistency of this solution. We propose to solve this problem through a regularization of the posterior distribution.We adopt a Tikhonov regularization scheme for constructing a new posterior distribution that is continuous and that we call *regularized posterior distribution*. We prove that this regularized posterior distribution is consistent in a "frequentist" sense. Our results agree with previous literature on infinite-dimensional Bayesian experiments, see for instance Diaconis and Freedman (1986).

2.1 Introduction

We consider the functional equation

$$\hat{Y} = Kx + U, \qquad x \in \mathcal{X}, \ \hat{Y} \in \mathcal{Y} \tag{2.1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are infinite dimensional separable Hilbert spaces over $\mathbb{R}$ supposed to be Polish [2] with inner product $< \cdot, \cdot >$ and norm $||\cdot||$. $K : \mathcal{X} \to \mathcal{Y}$ is a known Hilbert-Schmidt (HS, hereafter), then compact, linear operator with infinite dimensional range [3]. K^* will

---

[1]Sections 2.1 to 2.4 of this chapter are adapted from: Florens, J.P., and A., Simoni (2008), *Regularized Posteriors in Linear Ill-Posed Inverse Problems*, mimeo. Sections 2.5 and 2.6 are adapted from: Florens, J.P., and A., Simoni (2008), *Regularized Posteriors in Linear Ill-Posed Inverse Problems: Extensions*, mimeo.

[2]In mathematics, a Polish space is a separable completely metrizable topological space.

[3]Every Hilbert-Schmidt operator is compact. Hence, the Hilbert-Schmidt property is commonly used to show that an operator is compact

denote the adjoint of K, *i.e.* K^* is such that $< K\varphi, \psi > = < \varphi, K^*\psi >$, $\forall \varphi \in \mathcal{X}$ and $\psi \in \mathcal{Y}$. Compactness of operator K and the infinite dimension of the range of K make the inverse K^{-1} not continuous.

The error term U is an Hilbert-valued gaussian random variable with zero mean and covariance operator $\Sigma_n : \mathcal{Y} \to \mathcal{Y}$: $U \sim \mathcal{GP}(0, \Sigma_n)$, where n can be interpreted as the sample size.

The aim of this paper is to recover x from the noisy observation $\hat{Y}$, *i.e* to solve an *ill-posed inverse problem*, through a Bayesian approach. The Hilbert-valued random element x is supposed to induce a gaussian measure on $\mathcal{X}$: $x \sim \mathcal{GP}(x_0, \Omega_0)$, $x_0 \in \mathcal{X}$ and $\Omega_0 : \mathcal{X} \to \mathcal{X}$. From a Bayesian point of view, the solution to an inverse problem is the posterior distribution of the quantity of interest. This reformulation of an inverse problem as a parameter estimation is due to Franklin [33].

As an example of spaces, we could take both $\mathcal{X}$ and $\mathcal{Y}$ to be L^2 spaces. An L^2 space, endowed with a Gaussian measure defined on it, is a Polish space, see [45].

If problem (2.1) was formulated in finite dimension then it would have as solution the classical gaussian posterior distribution of x given $\hat{Y}$: $x|\hat{Y} \sim \mathcal{N}(x_0 + Cov(x, \hat{Y})Var(\hat{Y})^{-1}(\hat{Y} - Kx_0), Var(\hat{Y}) - Cov(x, \hat{Y})Var(\hat{Y})^{-1}Cov(x, \hat{Y}))$, see [49]. In finite dimensional inverse problems it is possible to remove ill-posedness by incorporating the available prior information, but this is no longer true when dimension is infinite since the covariance operator $Var(\hat{Y})$ is no longer continuously invertible. So, covariance operators do not have the regularization properties that have in the finite dimensional case. This prevents the posterior mean from being continuous in $\hat{Y}$ and consequently from being a consistent estimator. Therefore, the posterior distribution suffers of a problem of inconsistency in the sampling sense too.

This problem has been solved in past literature by restricting the space of definition of $\hat{Y} - Kx_0$, see [33], [65], [60] and [53]. However, this solution is not always appropriate since the observed data do not always satisfy this restriction.

Our contribution lies in proposing an alternative method to deal with the lack of continuity of the inverted covariance operator. The idea is to apply a regularization scheme to this inverse. We introduce the concept of *regularization of a distribution*, that consists in regularizing the moments characterizing it in order to obtain new moments with good asymptotic behavior. The posterior distribution that results is slightly modified and we call this new distribution *regularized posterior distribution*. We show that, as the number of observations in the sample grows, our proposed solution degenerates, with respect to the sampling measure, to a probability mass in a neighborhood of the true value of the parameter x having generated the data. This is the concept of posterior consistency, or frequentist consistency (or consistency in the sampling sense) of the posterior distribution, see [16].

We propose to use two alternative regularization schemes for computing the regularized posterior distribution. We consider first a classical Tikhonov regularization scheme where $(Var(\hat{Y}))^{-1}$ is replaced by $(\alpha_n I + Var(\hat{Y}))^{-1}$. Secondly, we use a Tikhonov regularization scheme in the Hilbert scale induced by the prior covariance operator Ω_0 that substitutes $(Var(\hat{Y}))^{-1}$ with $(\alpha_n L^{2s} + Var(\hat{Y}))^{-1}$, with $L = \Omega_0^{-\frac{1}{2}}$, $s \in \mathbb{R}$ and α_n the regularization parameter. Two facts have to be remarked: *(i)* the choice of the Hilbert scale is naturally suggested by the prior; *(ii)* the speed of convergence is considerably improved with the second regularization scheme.

The chapter is developed as follow. Section 4.2 presents the Bayesian experiment associated to (2.1). In Section 2.3 we define the *regularized posterior distribution*, for both regularization schemes; its consistency is proved in Section 2.4.

Extensions of the basic model are developed in Sections 2.5 and 2.6. More specifically, we

propose in Section 2.5 the more general case where the operator K is unknown and we prove that, under some minor assumption, this does not affect the speed of convergence of the regularized estimated solution. In Section 2.6 we consider the case in which the transformation of the parameter of interest x is made by an operator K_i that is different from an observation to another one. Section 5.7 concludes. All the proofs are given in Appendix A, examples of possible applications are given in Appendix B and numerical simulations are provided in Appendix C.

2.2 The Model

2.2.1 Sampling Probability Measure

Quantities $\hat{Y}$, x and U in equation (2.1) have to be meant as Hilbert-random variables. Let $\mathcal{F}$ denote the σ-field of subsets of the sample space $\mathcal{Y}$. We endow the measurable space $(\mathcal{Y}, \mathcal{F})$ with the sampling distribution $\mathbb{P}(\hat{Y}|x)$ of $\hat{Y}$ given x, denoted with P^x and characterized by Assumption 1 below.

Assumption 1 *Let P^x be a conditional probability measure on $(\mathcal{Y}, \mathcal{F})$ given x such that $\mathbb{E}(||\hat{Y}||^2) < \infty$, $\hat{Y} \in \mathcal{Y}$. P^x is a Gaussian measure that defines a mean element $Kx \in \mathcal{Y}$ and a covariance operator $\Sigma_n : \mathcal{Y} \to \mathcal{Y}$.*

For a characterization of Gaussian measures in Hilbert spaces we refer to Baker (1973) [3]. Assumption 1 implies that the covariance operator Σ_n is linear, bounded, nonnegative, selfadjoint and trace-class. A covariance operator need to be trace-class in order the associated measure be able to generate trajectories belonging to the Hilbert space of reference ($\mathcal{Y}$ in this case). The fact that Σ_n is trace-class entails that $\Sigma_n^{\frac{1}{2}}$ is HS. HS operators are compact and compacity of $\Sigma_n^{\frac{1}{2}}$ implies compacity of Σ_n. Compact operators are particularly attractive since they can be approximated by a sequence of finite dimensional operators and this is useful when we do not know such an operator and we need to estimate it.

We also suppose that $\Sigma_n \to 0$ as $n \to \infty$, where n is usually meant as the sample size. Indexing Σ_n to a parameter n is natural since in several applications $\hat{Y}$ is a consistent estimator of the transformed signal Kx, see examples in Appendix B.

Usually, in functional analysis literature the curve $\hat{Y}$ is supposed to be observed only at a finite number of points. A peculiarity of our model is to allow for more general observational schemes. The whole infinite dimensional object $\hat{Y}$ may be observed, for example this is the case with high-frequency financial data. Alternatively, we could observe a sample of discrete objects and the curve $\hat{Y}$ is a mathematical object obtained by transforming these data. Transformations of this kind are frequent in statistic and econometrics: for instance nonparametric estimators like kernel density estimator, empirical characteristic function, empirical cumulative distribution function or estimated integrated hazard function, see examples in Appendix B.

2.2.2 Prior Specification and Identification

Let μ denote the prior measure induced by x on the parameter space $\mathcal{X}$ endowed with the σ-field $\mathcal{E}$. We specify a conjugate prior:

Assumption 2 *Let μ be a probability measure on $(\mathcal{X}, \mathcal{E})$ such that $\mathbb{E}(||x||^2) < \infty$, $\forall x \in \mathcal{X}$* [4]*. μ is a Gaussian measure that defines a mean element $x_0 \in \mathcal{X}$ and a covariance operator $\Omega_0 : \mathcal{X} \to \mathcal{X}$.*

The covariance operator Ω_0 has the same properties discussed for Σ_n, then it is compact. We introduce the *Reproducing Kernel Hilbert Space* associated to the covariance operator Ω_0 (denoted with $\mathcal{H}(\Omega_0)$). Let $\{\lambda_j^{\Omega_0}, \varphi_j^{\Omega_0}\}_j$ be the eigensystem of Ω_0. We define the space $\mathcal{H}(\Omega_0)$ embedded in $\mathcal{X}$ as:

$$\mathcal{H}(\Omega_0) = \{\varphi : \varphi \in \mathcal{X} \quad \text{and} \quad \sum_{j=1}^{\infty} \frac{|<\varphi, \varphi_j^{\Omega_0}>|^2}{\lambda_j^{\Omega_0}} < \infty\} \tag{2.2}$$

and, following Proposition 3.6 in [10], $\mathcal{H}(\Omega_0) = \mathcal{R}(\Omega_0^{\frac{1}{2}})$, where $\mathcal{R}(\cdot)$ denotes the range of the corresponding operator. Let x_* denote the true value of the parameter having generated the data $\hat{Y}$, we assume that

Assumption 3 $(x_* - x_0) \in \mathcal{H}(\Omega_0)$, *i.e. there exists $\delta_* \in \mathcal{X}$ such that $(x_* - x_0) = \Omega_0^{\frac{1}{2}} \delta_*$.*

This assumption is only a regularity condition and it will be exploited for proving asymptotic results.

The support of a centered Gaussian process, taking its values in an Hilbert space $\mathcal{X}$, is the closure in $\mathcal{X}$ of the *Reproducing Kernel Hilbert Space* associated with the covariance operator of this process, see[77]. Then, for the prior distribution, $x - x_0 \in \overline{\mathcal{H}(\Omega_0)}$ with μ-probability 1, but, with μ- probability 1, $x - x_0$ is not in $\mathcal{H}(\Omega_0)$. This implies that the prior distribution is not able to generate a trajectory x that satisfies Assumption (24) or, in other words, the true value x_* having generated $\hat{Y}$ cannot have been drawn from μ. Anyway, if Ω_0 is injective, the space $\mathcal{H}(\Omega_0)$ is dense in $\mathcal{X}$ and therefore μ can generate trajectories as close as possible to the true value x_* even if μ puts zero probability on $\mathcal{H}(\Omega_0)$. This kind of problem is known as *prior inconsistency* and it is due to the fact that, because of the infinite dimensionality of the parameter space, the support of the prior can cover only a very "small" part of it. A similar result is found with the Dirichlet process, in nonparametric probabilities estimation, in the sense that such a process puts zero probability mass on absolutely continuous probability measures but it is able to generate probability functions close to them.

From a Bayesian point of view we say that a model is identified if the posterior distribution completely revises the prior distribution, for what we do not need to introduce strong assumptions, see [27] Section 4.6 for an exhaustive explanation of this concept. Nevertheless, this chapter focuses on consistency in the sampling sense of the posterior distribution, see Section 2.4, and for that we need the following assumption for identification.

Assumption 4 *The operator $K\Omega_0^{\frac{1}{2}} : \mathcal{X} \to \mathcal{Y}$ is one-to-one on $\mathcal{X}$.*

This assumption guarantees continuity of the *regularized posterior mean* defined below. The classical hypothesis for identification of x in model (2.1) requires that K be one-to-one. This is a stronger condition since, if $\Omega_0^{\frac{1}{2}}$ is one-to-one, K one-to-one implies $K\Omega_0^{\frac{1}{2}}$ one-to-one, but the reverse is not true. Therefore, frequentist consistency in a Bayesian model requires a weaker identification condition than a classical model does.

---

[4]We use the same notation $\mathbb{E}(\cdot)$ for denoting the mathematical expectation taken with respect to whatever distribution. Which distribution is meant should result clear.

2.2.3 Construction of the Bayesian Experiment

The relevant probability space associated to (2.1) is the real linear product space $\mathcal{X} \times \mathcal{Y}$ defined as the set

$$\mathcal{X} \times \mathcal{Y} := \{(x, y); x \in \mathcal{X}, y \in \mathcal{Y}\}$$

with addition, scalar multiplication and scalar product defined in the usual way. The product σ-field associated to $\mathcal{X} \times \mathcal{Y}$ is denoted with $\mathcal{E} \otimes \mathcal{F}$ and the probability measure defined on $(\mathcal{X} \times \mathcal{Y}, \mathcal{E} \otimes \mathcal{F})$ is denoted with Π and constructed by recomposing μ and P^x. Let $\Upsilon_{yy} = (\Sigma_n + K\Omega_0 K^*)$ be the covariance operator of the *predictive distribution P* and Υ be the covariance operator associated to Π defined as $\Upsilon(\varphi, \psi) = (\Omega_0\varphi + \Omega_0 K^*\psi, (\Sigma_n + K\Omega_0 K^*)\psi + K\Omega_0\varphi)$, for all $(\varphi, \psi) \in \mathcal{X} \times \mathcal{Y}$.

Lemma 1 *The covariance operators Υ and Υ_{yy} are trace class. In particular, Υ_{yy} trace class is a necessary condition for Υ being trace class.*

Next, we state that the joint and predictive probabilities, Π and P, are gaussian.

Theorem 1

(i). *Under Assumptions 1 and 2, the joint measure Π on $(\mathcal{X} \times \mathcal{Y}, \mathcal{E} \otimes \mathcal{F})$ is Gaussian with mean function $m_{xy} = (x_0, Kx_0) \in \mathcal{X} \times \mathcal{Y}$ and covariance operator Υ.*

(ii). *Let P be a gaussian measure on $(\mathcal{Y}, \mathcal{F})$ with mean function $m_y = Kx_0$ in $\mathcal{Y}$ and covariance operator Υ_{yy}. Then, P is the marginal distribution on $(\mathcal{Y}, \mathcal{F})$ associated to the joint gaussian measure Π defined in (i).*

The *Bayesian Experiment Ξ* associated to inverse problem (2.1) is:

$$\Xi = (\mathcal{X} \times \mathcal{Y}, \mathcal{E} \otimes \mathcal{F}, \Pi = P^x \otimes \mu) \tag{2.3}$$

and it constitutes the object of our study. The aim will be to determine the inverse decomposition of Π into the marginal P and the posterior distribution $\mu^{\mathcal{F}} = \mathbb{P}(x|\hat{Y})$: $\Pi = P \otimes \mu^{\mathcal{F}}$. Existence of this inverse decomposition is ensured if a regular version of the posterior probability exists. The posterior distribution $\mu^{\mathcal{F}}$ will be the solution to the initial inverse problem.

2.3 Solution of the Ill-Posed Inverse Problem

Due to infinite dimension of the Bayesian experiment, application of Bayes theorem is not evident and in computing the posterior distribution three points require a particular attention: *(i)* the existence of a regular version of the conditional probability on $\mathcal{E}$ given $\mathcal{F}$, *(ii)* the fact that it is a gaussian measure and *(iii)* its continuity.

(i) The conditional probability on $\mathcal{E}$ given $\mathcal{F}$ exists and it is unique since it is the projection on a closed convex subset of $L^2(\mathcal{X} \times \mathcal{Y})$, where $L^2(\mathcal{X} \times \mathcal{Y})$ is the Hilbert space of random variables defined on $\mathcal{X} \times \mathcal{Y}$ that are square integrable with respect to Π. A conditional probability is called *regular* if a transition probability characterizing it exists. The existence of such a transition for $\mu^{\mathcal{F}}$ is guaranteed by *Jirina Theorem*, see [62], if the space $(\mathcal{X} \times \mathcal{Y})$ is Polish.

(ii) By slightly modifying the proof given in Section 2.2 of [60] it is trivial to show that $\mu^{\mathcal{F}}$ is gaussian. The associated characteristic function takes the form

$$\mathbb{E}(e^{i<x,h>}|\hat{Y}) = e^{i<A\hat{Y}+b,h>-\frac{1}{2}<(\Omega_0-AK\Omega_0)h,h>}, \qquad h \in \mathcal{X}.$$

Then, $x|\hat{Y}$ has mean: $A\hat{Y} + b$, and variance $V = \Omega_0 - AK\Omega_0$. Since $\mathbb{E}(x) = \mathbb{E}(\mathbb{E}(x|\hat{Y}))$, $b = (I - AK)x_0$ and A is deducible from the following development:

$$
\begin{aligned}
< \Upsilon_{12}\varphi, \psi > &= Cov(< x, \varphi >, < \hat{Y}, \psi >) \\
&= Cov(\mathbb{E}(< x, \varphi > |\hat{Y}), < \hat{Y}, \psi >) \\
&= Cov(< A\hat{Y}, \varphi >, < \hat{Y}, \psi >) \\
&= < (\Sigma_n + K\Omega_0 K^*)A^*\varphi, \psi >, \qquad \forall \varphi \in \mathcal{X}, \psi \in \mathcal{Y} \quad (2.4)
\end{aligned}
$$

where $\Upsilon_{12} = K\Omega_0$ is a component of operator Υ determined in Theorem 1. Hence, $A : \mathcal{Y} \to \mathcal{X}$ is solution of

$$A(\Sigma_n + K\Omega_0 K^*)\psi = \Omega_0 K^*\psi, \qquad \psi \in \mathcal{Y} \tag{2.5}$$

and then $A = \Omega_0 K^*(\Sigma_n + K\Omega_0 K^*)^{-1}$.

(iii) Expression for A is not well-defined since $(\Sigma_n + K\Omega_0 K^*)$ is a compact operator with infinite range and its inverse is not continuous on the whole $\mathcal{Y}$. Therefore, the posterior mean is not continuous in $\hat{Y}$ and we have to deal with a further ill-posed inverse problem. Continuity is crucial for asymptotic properties of the estimator, in particular for posterior consistency. Problems of inconsistency are frequent in nonparametric Bayesian experiments, see [16].

Therefore, Bayesian analysis of inverse problems changes the cause of the ill-posedness of the equation (2.1): inconsistency is no longer due to non-continuity of K^{-1} but to non-continuity of $(\Sigma_n + K\Omega_0 K^*)^{-1}$.

Past literature on Bayesian inverse problems, see [60] and [53], proposed, as strategy to solve this problem of non-continuity, to restraint the space of the observable $\hat{Y}$. It was implicitly assumed that $\hat{Y}$ belongs to $\mathcal{R}(\Sigma_n + K\Omega_0 K^*)$ or to a subspace of it. We do not wish to make this kind of restriction since we admit any trajectory $\hat{Y}$ in $\overline{\mathcal{R}(\Sigma_n + K\Omega_0 K^*)}$. Thus, a different strategy, based on Tikhonov regularization, will be proposed in the next paragraph.

2.3.1 Tikhonov Regularized Posterior distribution

We propose to solve the problem of unboundedness of A by applying a Tikhonov regularization scheme to equation (2.5). We define the regularized operator A_α as:

$$A_\alpha = \Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1} \tag{2.6}$$

where $\alpha_n > 0$ is a regularization parameter appropriately chosen such that $\alpha_n \to 0$ with n.

We could interpret the Tikhonov regularized A_α as the operator that we would obtain if we considered a new Bayesian experiment $\hat{Y} = Kx + U + \eta$, with η a further error term with variance $\alpha_n I$. In this case the sampling measure would define a covariance operator $\alpha_n I + \Sigma_n$. This covariance operator is not trace class so that the trajectories generated

by this distribution would not be in the Hilbert space $\mathcal{Y}$. This interpretation is useful since it provides a new Bayesian method that can be used for selecting the regularization parameter α_n from its posterior distribution once a prior distribution on it has been specified. We do not develop this point here and it will be object of future research. The regularized versions of b and V, with A replaced by A_α are

$$
\begin{aligned}
b_\alpha &= (I - A_\alpha K)x_0, \\
V_\alpha &= \Omega_0 - \Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}K\Omega_0.
\end{aligned}
\tag{2.7}
$$

These regularized objects characterize a new distribution that is gaussian with mean $(A_\alpha \hat{Y} + b_\alpha)$ and covariance operator V_α. This distribution is called *regularized posterior distribution* and is denoted with $\mu_\alpha^{\mathcal{F}}$. It is a new object that we define to be the solution of the signal-noise problem and that we will show in Section 2.4, is consistent. Moreover, we keep as punctual estimator of x the regularized posterior mean

$$
\mathbb{E}_\alpha(x|\hat{Y}) = x_0 + \Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}(\hat{Y} - Kx_0).
\tag{2.8}
$$

2.3.2 Tikhonov regularization in the Prior Variance Hilbert scale

We propose in this subsection an alternative regularization scheme, for recovering A, based on *Tikhonov regularization in the Hilbert scale induced by the inverse of the prior covariance operator*, see [19] for general theory. Let $L = \Omega_0^{-\frac{1}{2}}$ be a densely defined unbounded self-adjoint strictly positive operator in the Hilbert space $\mathcal{X}$. More clearly, if $\mathcal{D}(L)$ denotes the domain of L, L is a closed operator in $\mathcal{X}$ satisfying: $\mathcal{D}(L) = \mathcal{D}(L^*)$ is dense in $\mathcal{X}$, $< Lx, y > = < x, Ly >$ for all $x, y \in \mathcal{D}(L)$, and there exists $\gamma > 0$ such that $< Lx, x > \geq \gamma ||x||^2$ for all $x \in \mathcal{D}(L)$. The norm $|| \cdot ||_s$ is defined as $||x||_s := ||L^s x||$. We define the Hilbert Scale $\mathcal{X}_s$ induced by L as the completion of the domain of L^s, $\mathcal{D}(L^s)$, with respect to the norm $|| \cdot ||_s$; moreover $\mathcal{X}_s \subseteq \mathcal{X}_{s'}$ if $s' \leq s$, $\forall s \in \mathbb{R}$. Usually, when a regularization scheme in Hilbert Scale is adopted, the operator L, and consequently the Hilbert Scale, is created *ad hoc*. In our case the Hilbert Scale is not created ad-hoc but is suggested by the prior information we have and this is a noteworthy fact that represents a considerable advantage with respect to the standard methods. For the theory to work it is necessary the first of the following assumptions to be satisfied.

Assumption 5

(i) $||K\Omega_0^{\frac{1}{2}}x|| \sim ||\Omega_0^{\frac{a}{2}}x||$, $\forall x \in \mathcal{X}$ and for some $a > 0$;

(ii) $(x_* - x_0) \in \mathcal{X}_{\beta+1}$ for some $\beta > 0$, i.e. $\exists \rho_* \in \mathcal{X}$ such that $(x_* - x_0) = \Omega_0^{\frac{\beta+1}{2}}\rho_*$

(iii) $a \leq s \leq \beta + 1 \leq 2s + a$.

Assumption *(i)* is equivalent to say that in specifying the prior distribution we take into account the sampling model, hence the prior variance is linked to the sampling model (2.1) we are studying and, in particular, to operator K. This kind of prior specification is not new in Bayesian literature since it is similar to the Zellner's *g-prior*, see [81] or [1]. Parameter a can be interpreted as a *degree of ill-posedness*. Therefore, the prior is specified not only by taking into account the sampling model but also the degree of ill-posedness of the problem.

Assumption *(ii)* is known as a *source condition* and is formulated in order to reach a certain speed of convergence of the regularized solution. Under Assumption 24, it says that $\delta_* \in \mathcal{R}(\Omega_0^{\frac{\beta}{2}})$, hence $\mathcal{X}_{\beta+1} \equiv \mathcal{R}(\Omega_0^{\frac{\beta+1}{2}}) \equiv \mathcal{D}(L^{\beta+1})$. The meaning of such an assumption is that the prior distribution contains information about the regularity of the true value of x. In fact, parameter β is interpreted as the *regularity parameter*. These two remarks stress the fact that we are not taking whatever Hilbert Scale, but the Hilbert Scale linked to the prior. Either we first choose the Hilbert Scale and then we use the information contained in it to specify the prior distribution or we use the information contained in the prior distribution to specify the Hilbert Scale.

The restriction $\beta + 1 \geq s$ means that the centered true value x_* has to be at least an element of $\mathcal{X}_s$ and it guarantees that the norm $||L^s x||$ exists $\forall x \in \mathcal{X}_{\beta+1}$. The parameter a denotes the degree of ill-posedness in the Bayesian problem.

Under such assumptions the Tikhonov regularized solution in $\mathcal{X}_s$ to equation (2.5) is:

$$A_s = \Omega_0 K^* (\alpha_n L^{2s} + \Sigma_n + K\Omega_0 K^*)^{-1}. \tag{2.9}$$

The regularized posterior distribution is thus defined similarly as in Section 2.3.1 with A_α substituted by A_s and is denoted with $\mu_s^{\mathcal{F}}$. The regularized posterior mean and variance are

$$
\begin{aligned}
\mathbb{E}_s(x|\hat{Y}) &= A_s \hat{Y} + (I - A_s)x_0 \\
V_s &= \Omega_0 - A_s K \Omega_0.
\end{aligned}
\tag{2.10}
$$

This regularization method has the advantage that it permits to better exploit the regularity of the true function x_*. A classical Tikhonov regularization method allows to obtain a rate of convergence to zero of the regularization bias that is at most of order 2; on the contrary with a Tikhonov scheme in an Hilbert Scale the smoother the function x_* is, the faster the rate of convergence to zero of the regularization bias will be. Moreover, we will show in Section 5.4.2 that $\mu_s^{\mathcal{F}}$ reaches a faster speed of convergence toward the true solution.

2.4   Asymptotic Analysis

This section focuses on the study of the consistency of the regularized posterior distribution and of the regularized posterior mean. We start by showing the consistency, and by computing the rate of convergence, of the Tikhonov regularized posterior distribution $\mu_\alpha^{\mathcal{F}}$ defined in paragraph 2.3.1. Consistency of $\mu_s^{\mathcal{F}}$ defined in 2.3.2 will be analyzed in subsection 5.4.2.

The aim of this section is to analyze "frequentist" consistency of the recovered posterior distribution. If P^x denotes the sampling probability, this means that we analyze convergence P^x-*a.s.*, or convergence in probability with respect to the measure P^x, of the regularized version of the posterior distribution that we have defined.

Following Diaconis *et al-* (1986) [16] we give the following definition of *posterior consistency*:

Definition 1 *The pair $(x, \mu^{\mathcal{F}})$ is consistent if $\mu^{\mathcal{F}}$ converges weakly to δ_x as $n \to \infty$ under P^x-probability or P^x-a.s., where δ_x is the Dirac measure in x.*
The posterior probability $\mu^{\mathcal{F}}$ is consistent if $(x, \mu^{\mathcal{F}})$ is consistent for all x.

If $(x, \mu^{\mathcal{F}})$ is consistent in the previous sense, the Bayes estimate for x, (*i.e.* the posterior mean for a quadratic loss function), is consistent too.

The meaning of this definition is that, for any neighborhood U of the true parameter x, the posterior probability of the complement of U converges toward zero when $n \to \infty$: $\mu^{\mathcal{F}}(U^c) \to 0$ in P^x-probability, or P^x-a.s. Therefore, since distribution expresses one's knowledge about the parameter, consistency stands for convergence of knowledge towards the perfect knowledge with increasing amount of data.

In general, in an identified *i.i.d.* model with final dimensional parameter space we have posterior consistency if the true value of the parameter is in the support of the prior distribution. On the contrary, when the parameter space is of infinite dimension, this condition is no more sufficient to guarantee the consistency of the posterior, as it is remarked in [16]. Besides the problem of infinite dimension of the parameter space, we also encounter the difficulty that we are dealing with the regularized posterior distribution, $\mu_{\alpha}^{\mathcal{F}}$. Then, we are going to extend the concept of posterior consistency in order to be applied to the regularized posterior distribution and it makes sense to speak about *regularized posterior consistency*.

To prove posterior consistency in the case of a Gaussian posterior measure, it is sufficient to prove consistency of the posterior mean and convergence to zero of the posterior variance. In fact, let x_* be the true value of the parameter characterizing the DGP of $\hat{Y}$, by using *Chebyshev's Inequality* and for any sequence $M_n \to \infty$

$$
\begin{aligned}
\mu_{\alpha}^{\mathcal{F}}\{x : ||x - x_*|| \geq M_n \varepsilon_n\} &\leq \frac{\mathbb{E}_{\alpha}(||x - x_*||^2|\hat{Y})}{(M_n \varepsilon_n)^2} \\
&= \frac{< V_{\alpha}(x(t)|\hat{Y}), 1 >_{\mathcal{X}} + ||\mathbb{E}_{\alpha}(x|\hat{Y}) - x_*||^2}{(M_n \varepsilon_n)^2} \\
&\leq \frac{||V_{\alpha}(x|\hat{Y})||_{\mathcal{X}} + ||\mathbb{E}_{\alpha}(x|\hat{Y}) - x_*||^2}{(M_n \varepsilon_n)^2}
\end{aligned} \tag{2.11}
$$

with π a measure on $\mathbb{R}$. The RHS of (5.19) goes to 0 if both the terms in the numerator converge to zero. We start by proving consistency of the regularized posterior mean, *i.e.* $||\mathbb{E}_{\alpha}(x|\hat{Y}) - x_*|| \to 0$ P^{x*}-a.s. when $n \to \infty$. For any true value $x_* \in \mathcal{X}$, the Bayes estimation error is

$$
\begin{aligned}
\mathbb{E}_{\alpha}(x|\hat{Y}) - x_* &= \Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1} K(x_* - x_0) \\
&\quad + \Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1} U - (x_* - x_0)
\end{aligned}
$$

and it converges to 0 under conditions given in the theorem below. Let Φ_{β} denote the β-regularity space of the operator $K\Omega_0^{\frac{1}{2}}$, *i.e.* $\Phi_{\beta} \equiv \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$.

Theorem 2 *Under Assumptions 24 and 25 if $\alpha_n \to 0$, $\frac{1}{\alpha_n} tr\Sigma_n \to 0$ and $\frac{1}{\alpha_n^3} ||\Sigma_n||^2 \sim \mathcal{O}_p(1)$, then:*

(i) $\mathbb{E}(x|\hat{Y}) \to^{\mathcal{P}^{x*}} 0$ *in $\mathcal{X}$ norm;*

(ii) *moreover, if $\delta_* \in \Phi_{\beta}$, for some $\beta > 0$, the bias is of order*

$$
||\mathbb{E}_{\alpha}(x|\hat{Y}) - x_*||^2 = \mathcal{O}_p(\alpha_n^{\beta \wedge 2} + \frac{1}{\alpha_n^4} ||\Sigma_n||^2 \alpha_n^{(\beta+1) \wedge 2} + \frac{1}{\alpha_n} tr\Sigma_n).
$$

The larger β is, the smoother the function $\delta_* \in \Phi_\beta$ will be and faster the regularization bias will converge to zero. However, for a Tikhonov regularization scheme, β cannot be grater than 2, this is the reason why we bound it by 2 in α_n^β. With classical Tikhonov regularization scheme it is useless to have a function x_* with a degree of smoothness larger than 2. In the remaining of this section, for simplifying writing, we will not explicitly write $\beta \wedge 2$, but it will be implicit that we are assuming $\beta \le 2$ and if $\beta > 2$ it must be set at 2. Condition $\frac{1}{\alpha_n^3}||\Sigma_n||^2 \sim \mathcal{O}_p(1)$ is sufficient to guarantee that $\frac{1}{\alpha_n^4}||\Sigma_n||^2\alpha_n^{(\beta+1)\wedge 2} \to 0$ since for every β, $(\beta+1) \wedge 2 > 1$ and then $\alpha_n^{(\beta+1)\wedge 2}$ converges to zero even after having been simplified with the α_n in the denominator.

Furthermore, if we assume that $tr\Sigma_n$ is of the same order as $||\Sigma_n||$, for instance $tr\Sigma_n \sim ||\Sigma_n|| \sim \mathcal{O}_p(\frac{1}{n})$, convergence to zero of the second and third rates in the bias require satisfaction of conditions $\alpha_n \to 0$ and $\alpha_n^{\frac{3}{2}}n \to \infty$. Classical conditions for convergence of the solution of stochastic ill-posed problems are $\alpha_n \to 0$ and $\alpha_n^2 n \to \infty$ (see [78]). Therefore, we require weaker conditions to get optimal speed of convergence.

If $tr\Sigma_n$ is of the same order as $||\Sigma_n||$ the fastest global rate of convergence is obtained when $\alpha_n^\beta = \frac{1}{\alpha_n}||\Sigma_n||$, that is, when the optimal regularization parameter α_n^* is proportional to

$$\alpha_n^* \propto ||\Sigma_n||^{\frac{1}{\beta+1}}.$$

With the optimal value α_n^*, the condition $\frac{1}{\alpha_n^3}||\Sigma_n||^2 \sim \mathcal{O}_p(1)$ is ensured if $\beta \ge \frac{1}{2}$. Hence, the speed of convergence of the regularized posterior mean is proportional to $||\Sigma_n||^{\frac{\beta}{\beta+1}}$. Assuming the trace and the norm of the covariance operator be of the same order is not really stringent. For instance, in almost all real examples they are both of order $\frac{1}{n}$.

Let us proceed now to the study of the regularized posterior variance. We want to check that $||V_\alpha \varphi|| \to 0$ for all $\varphi \in \mathcal{X}$.

Theorem 3 *Under Assumption 25, if $\alpha_n \to 0$ and $\frac{1}{\alpha_n^3}||\Sigma_n||^2 \sim \mathcal{O}_p(1)$ then*

(i) $V_\alpha(x|\hat{Y})\varphi \to^{\mathcal{P}^{x*}} 0$ *in $\mathcal{X}$ norm;*

(ii) *moreover, if the posterior variance is applied to $\varphi \in \mathcal{X}$ such that $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, for some $\beta > 0$, it is of order*

$$||V_\alpha(x|\hat{Y})\varphi||^2 = \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^4}||\Sigma_n||^2\alpha_n^{(\beta+1)\wedge 2}).$$

With the optimal α_n^*, under the conditions in the above theorem and if $\beta \ge \frac{1}{2}$, the squared norm of the regularized posterior variance converges to zero at the speed of $||\Sigma_n||^{\frac{\beta}{\beta+1}}$. Its norm is slower and is of order $||\Sigma_n||^{\frac{\beta}{2(\beta+1)}}$.

Finally, from inequality (5.19) it follows that $\mu_\alpha^{\mathcal{F}}$ degenerates to the Dirac measure in x_*. Thus, under the fundamental assumption $(x_* - x_0) \in \mathcal{H}(\Omega_0)$, the regularized posterior probability of the complement of any neighborhood of the true parameter x_*, $\mu_\alpha^{\mathcal{F}}\{x : ||x - x_*||_\mathcal{X} \ge M_n\varepsilon_n\}$, goes to zero and, if $tr\Sigma_n \sim \mathcal{O}_p(||\Sigma_n||)$, it is of optimal order $||\Sigma_n||^{\frac{\beta}{2(\beta+1)}}$. We have in this way proved the posterior consistency of $\mu_\alpha^{\mathcal{F}}$.

Lastly, we wish to compare the speed of convergence that we find with the Bayesian method with the rate founded by applying a classical Tikhonov resolution method to equation (2.1) (that is suggested by the classical literature on inverse problems). In the following, we shall call these two methods *Bayesian method* and *classical method*, respectively; we refer to [19] and [10] for a review of the classical method. For simplifying, we set $x_0 = 0$. To make this comparison possible we have to consider a particular case for the prior covariance operator: $\Omega_0 = c_1 (K^* K)^\gamma$, with c_1 a constant of proportionality. In this particular case the fastest rate of convergence of the regularized posterior distribution is slower than the rate of convergence that would be obtained with the classical method. The regularity condition required in the classical method is $x_* \in \mathcal{R}(K^* K)^{\frac{\gamma}{2}}$ and the optimal speed of convergence is $(tr\Sigma_n)^{\frac{\gamma}{\gamma+1}}$, with $\gamma \leq 2$ or γ set equal to 2 if $\gamma \geq 2$. Therefore, if we choose β in order to have the same regularity condition, *i.e.* $\mathcal{R}(K^* K)^{\frac{(\gamma+1)\beta}{2}} = \mathcal{R}(K^* K)^{\frac{\gamma}{2}}$ and then $\beta = \frac{\gamma}{\gamma+1}$, the fastest rate of convergence in the Bayesian method is proportional to $(tr\Sigma_n)^{\frac{\gamma}{2\gamma+1}}$ that is slower with respect to the classical one. This result is due to the fact that the Bayesian method increases the degree of ill-posedness. However, no comparison can be done outside of this particular form taken by Ω_0. In the following subsection we show that the speed of convergence is improved when we use $\mu_s^{\mathcal{F}}$ and the same speed of convergence as with the classical method is attained.

2.4.1 Speed of convergence with Tikhonov regularization in the Prior Variance Hilbert Scale

We compute in this subsection the speed of convergence for the regularized posterior distribution with Tikhonov regularization in Hilbert scale, under Assumption 5. The speed obtained in this case is faster than that one with a simple Tikhonov regularization scheme and it is the same speed as we would have obtained if we had solved the functional equation directly in an Hilbert scale without applying the bayesian method. We suppose Assumption 5 holds, the attainable speed of convergence is given in the following theorem, the proof of which is provided in Appendix 5.8.

Theorem 4 *Let $\mathbb{E}_s(x|\hat{Y})$ and V_s be as in (5.18). Under Assumptions 24, 25 and 5*

$$||\mathbb{E}_s(x|\hat{Y}) - x_*||^2 \sim \mathcal{O}_p\Big(\alpha_n^{\frac{\beta+1}{a+s}} + \alpha_n^{\frac{1-a}{a+s}} tr\Sigma_n + \frac{1}{\alpha_n^2}||\Sigma_n||^2 \alpha_n^{\frac{\beta-a}{a+s}} + \frac{1}{\alpha_n^2}||\Sigma_n||^2 tr\Sigma_n \alpha_n^{\frac{1-a}{a+s}}\Big).$$

Moreover, if the covariance operator V_s is applied to elements $\varphi \in \mathcal{X}$ such that $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{\beta}{2}})$, then

$$||V_s\varphi||^2 \sim \mathcal{O}_p\Big(\alpha_n^{\frac{\beta+1}{a+s}} + \frac{1}{\alpha_n^2}||\Sigma_n||^2 \alpha_n^{\frac{\beta-a}{a+s}}\Big).$$

The optimal α_n is obtained by equating the first two rates of convergence of the posterior mean, that gives:

$$\alpha_n^* \propto (tr\Sigma_n)^{\frac{a+s}{a+\beta}}$$

and the corresponding optimal speed is proportional to $(tr\Sigma_n)^{\frac{\beta+1}{a+\beta}}$. With this choice of the regularization parameter the remaining rates goes to zero if $\beta > \frac{a+2s}{3}$. This constraint is binding with respect to the constraint in Assumption 5 *(iii)*, *i.e.* $\frac{a+2s}{3} \geq s - 1$, if the

ill-posedness parameter satisfies $a \geq s - 3$. It should be noted that parameter s characterizing the norm in the Hilbert scale does not play any role in the speed of convergence. An advantage of the Tikhonov regularization in Hilbert Scale is that we can even obtain a rate of convergence for other norms, namely $|| \cdot ||_r$ for $-a \leq r \leq \beta + 1 \leq a + 2s$. Actually, the speed of convergence of these norms gives the speed of convergence of the estimate of the r-th derivative of the parameter of interest x.

If we directly solved functional equation (2.1) with a Tikhonov regularization in an Hilbert scale we would obtain a solution $x_s = (\alpha_n L^{2s} + K^* K)^{-1} K^* \hat{Y}$ and a speed of convergence of order $(tr\Sigma_n)^{\frac{u}{\bar{a}+u}}$, under the hypothesis $||Kx|| \sim ||L^{-\bar{a}}x||$ and $x \in \mathcal{X}_u$, with $\bar{a}$ the degree of ill-posedness in the classical problem. By comparing these assumptions to the bayesian ones it results that $||K\Omega_0^{\frac{1}{2}}x|| \sim ||L^{-\bar{a}}\Omega_0^{\frac{1}{2}}x||$ and, substituting to L the operator $\Omega_0^{-\frac{1}{2}}$, this norm is equivalent to $||\Omega_0^{\frac{\bar{a}+1}{2}}x||$, that implies that the degree of ill-posedness in the Bayesian problem is greater than the degree of ill-posedness in the classical problem: $a = \bar{a} + 1$. Moreover, if we take the same regularity condition in the two problems, *i.e.* $\beta + 1 = u$, the rate of convergence of the regularized posterior and of the Tikhonov regularized solution in Hilbert scale would be the same.

This confirms the improvement, in terms of speed of convergence, of the Tikhonov regularization in Hilbert scale with respect to the classical Tikhonov regularization. Take for instance the particular case with $\Omega_0 = (K^* K)$ and impose the same regularity condition in $\mathcal{X}$ and in the Hilbert scale $\mathcal{X}_s$. The regularity condition in Theorem 2 requires that $\delta_* \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\gamma}{2}} \equiv \mathcal{R}((K^* K)^{\gamma})$ for a certain $\gamma > 0$ [5], that implies $(x_* - x_0) \in \mathcal{R}((K^* K)^{\gamma + \frac{1}{2}})$. The regularity condition for the Hilbert scale regularization is $(x_* - x_0) \in \mathcal{R}(\Omega_0^{\frac{\beta+1}{2}}) \equiv \mathcal{R}((K^* K)^{\frac{\beta+1}{2}})$; henceforth the conditions are equal if $2\gamma = \beta$. Taking this value for β, the rate of convergence in the Hilbert scale $\mathcal{X}_s$ is proportional to $(tr\Sigma_n)^{\frac{2\gamma+1}{2\gamma+2}}$ that is faster than the speed of convergence in $\mathcal{X}$ (that is proportional to $(tr\Sigma_n)^{\frac{\gamma}{\gamma+1}}$).

Even without restricting to this particular form for Ω_0 it is possible to show the improvement in term of speed of convergence obtained with an Hilbert scale. To this end, it is sufficient that Assumption 5 *(i)* holds since it implies the equivalence $||(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\gamma}{2}} v|| \sim ||\Omega_0^{\frac{a\gamma}{2}} v||$, for some $v \in \mathcal{X}$. Then, $||\Omega_0^{\frac{\beta}{2}} v|| \sim ||\Omega_0^{\frac{a\gamma}{2}} v||$ if and only if $\beta = a\gamma$ (or $\beta = (\bar{a}+1)\gamma$). The optimal bayesian speed of convergence with an Hilbert scale is $(tr\Sigma_n)^{\frac{a\gamma+1}{a+a\gamma}}$ that is fastest than the bayesian speed of convergence with a classical Tikhonov: $(tr\Sigma_n)^{\frac{\gamma}{\gamma+1}}$, $\forall \gamma > 0$.

2.5 The case with unknown operator K

In several econometric and statistical applications of inverse problem (2.1) both Y and K are unknown. In such situations we are faced with the so-called *stochastic ill-posed problem* described in Vapnik (1998) [78]. Examples are the nonparametric instrumental regression model, see Darolles *et al.* (2006) [15] and Florens and Simoni (2008a) [29], the conditional density estimation and the regression function estimation. We study in this section the signal-noise problem when operator K is unknown and we analyze as this affects the rate of convergence of the regularized posterior distribution.

When operator K is unknown, the measurement error U can be defined in two different

---

[5]Note that for diversify with respect to the regularity parameter in the Hilbert scale we use letter γ, instead of β as used in Theorem 2 for the regularity on $\mathcal{X}$.

ways. Either *(i)* we can interpret U as the estimation error in Y: $U = \hat{Y} - Kx$ like in (2.1), or *(ii)* U denotes the difference between the two estimated quantities $\hat{Y}$ and $\hat{K}$: $U = \hat{Y} - \hat{K}x$. The second way to define the error is motivated by the fact that in real applications it is not always possible to recover the distribution of $U = \hat{Y} - Kx$ (even asymptotically), while the distribution of $U = \hat{Y} - \hat{K}x$ results easier to obtain. It should be noted that Y and Kx are equal but they are not the same object. Consequently, the equality is no more valid for their estimators.

We focus on the first definition of the measurement error: $\hat{Y} = Kx + U$, that is our general model (2.1). Nevertheless, when K is unknown the *regularized posterior distribution $\mu_\alpha^\mathcal{F}$* previously defined is no longer computable. Thus, we replace the infeasible distribution $\mu_\alpha^\mathcal{F}$ with the *estimated regularized posterior distribution $\hat{\mu}_\alpha^\mathcal{F}$* that is feasible and that is obtained by replacing K with a consistent estimator of it. For instance, if K is the conditional expectation its kernel can be estimated through a kernel smoothing. We denote with $\hat{K}$ the consistent estimator of K and with $\widehat{K^*}$ the consistent estimator of the adjoint K^*. In general, the adjoint of the estimator is different than the estimator of the adjoint: $(\hat{K})^* \neq \widehat{K^*}$.

The estimated $\hat{\mu}_\alpha^\mathcal{F}$ is a gaussian measure on $\mathcal{X}$ characterized by the *estimated regularized mean function* and *covariance operator*:

$$
\begin{aligned}
\hat{\mathbb{E}}_\alpha(x|\hat{Y}) &= \Omega_0 \widehat{K^*}(\alpha_n I + \Sigma_n + \hat{K}\Omega_0\widehat{K^*})^{-1}(\hat{Y} - \hat{K}x_0) + x_0 \\
\hat{V}_\alpha &= \Omega_0 - \Omega_0 \widehat{K^*}(\alpha_n I + \Sigma_n + \hat{K}\Omega_0\widehat{K^*})^{-1}\hat{K}\Omega_0.
\end{aligned}
$$

The posterior $\hat{\mu}_\alpha^\mathcal{F}$ is the solution to the ill-posed inverse problem (2.1) and it has been computed by applying a classical Tikhonov regularization scheme to the inverse of $(\Sigma_n + \hat{K}\Omega_0\widehat{K^*})$.

Alternatively, we could use a Tikhonov regularization scheme in the prior variance Hilbert Scale, as it has been proposed in section 2.3.2. We develop in this essay general computations and asymptotic analysis only for the classical Tikhonov case. Extension to Tikhonov regularization in Hilbert Scale are trivial and would require only minor modifications.

2.5.1 Asymptotic Analysis

We proceed to analyze consistency of the estimated regularized posterior distribution $\hat{\mu}_\alpha^\mathcal{F}$ and we adopt the frequentist notion of consistency given in definition 1.

We start by analyzing convergence of the mean function and we decompose the estimation error as

$$\hat{x}_\alpha - x_* = (\hat{x}_\alpha - \tilde{x}_\alpha) + (\tilde{x}_\alpha - x_\alpha) + (x_\alpha - x_*), \tag{2.12}$$

where $\hat{x}_\alpha = \hat{\mathbb{E}}_\alpha(x|\hat{Y})$ and $x_\alpha = \mathbb{E}_\alpha(x|\hat{Y})$ denote the *regularized posterior mean* with unknown and with known operator K, respectively, and

$$\tilde{x}_\alpha = \Omega_0\widehat{K^*}(\alpha_n I + \Sigma_n + \hat{K}\Omega_0\widehat{K^*})^{-1}(\hat{K}x_* + U - \hat{K}x_0) + x_0$$

denotes the *regularized posterior mean* with unknown K conditioned to the observation $\tilde{Y} = \hat{K}x + U$. We have decomposed the estimation error in three terms: the first one takes into account the estimation error about $Y - Kx_*$, the second one is the estimation error about the operator K and the third term is the approximation error due to have approximated x by using the regularized posterior mean. We analyze convergence in $\mathcal{X}$-norm:

$$||\hat{x}_\alpha - x_*||^2 \leq ||\hat{x}_\alpha - \tilde{x}_\alpha||^2 + ||\tilde{x}_\alpha - x_\alpha||^2 + ||x_\alpha - x_*||^2.$$

The asymptotic behavior of the third component of the estimation error was provided in Theorem 2. In order to have convergence to zero of the other two terms, we require that $\hat{K}$ and $\hat{Y}$ converge towards the true values K and Y, respectively, at a suitable rate.

Assumption 6 *(a)* $||\Omega_0^{\frac{1}{2}} \widehat{K^*}\hat{K} - \Omega_0^{\frac{1}{2}} K^* K||^2 = \mathcal{O}_p(\delta_1)$;

(b) $||\hat{K}\Omega_0\widehat{K^*} - K\Omega_0 K^*||^2 = \mathcal{O}_p(\delta_2)$;

(c) $||\Omega_0^{\frac{1}{2}}(\widehat{K^*} - K^*)||^2 = \mathcal{O}_p(\delta_3)$;

(d) $||\hat{K} - K||^2 = \mathcal{O}_p(\delta_4)$.

The order of convergence δ_1, δ_2, δ_3 and δ_4 of these operators depend on the nature of K and Ω_0 and on the kind of estimators used. Therefore, they need to be determined contextually to every problem. In most of the cases, even if we use a nonparametric estimator for the operator K, the rates δ_1 and δ_3 are faster than the nonparametric rate of convergence. This is caused by the smoothing effect due to application of operator $\Omega_0^{\frac{1}{2}}$ that allows to improve the speed of convergence and sometimes even to reach the parametric one (*e.g.* instrumental variable estimation, see Florens *et al.* (2008a) [29]). The following theorem states consistency of the *estimated regularized posterior mean* $\hat{x}_\alpha$.

Theorem 5 *Under Assumptions 24, 25 and 6, if $\alpha_n \to 0$, $\frac{1}{\alpha_n}tr\Sigma_n \to 0$, $\frac{1}{\alpha_n^3}||\Sigma_n||^2 \sim \mathcal{O}_p(1)$, $\frac{\delta_1}{\alpha^2} \sim \mathcal{O}_p(1)$, $\frac{\delta_2}{\alpha^2} \sim \mathcal{O}_p(1)$, $\frac{\delta_3}{\alpha} \sim \mathcal{O}_p(1)$ and $\frac{\delta_4}{\alpha^2} \sim \mathcal{O}_p(1)$ then:*

$$||\hat{x}_\alpha - x_*||^2 \to 0$$

in P^{x_*}-probability as $n \to \infty$.
Moreover, if $\delta_* \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$

$$\begin{aligned}
||\hat{x}_\alpha - x_*||^2 &= \mathcal{O}_p\Big(\frac{\delta_1}{\alpha_n^2}\alpha^\beta + \frac{\delta_2}{\alpha^3}tr\Sigma_n + \frac{\delta_3}{\alpha^2}tr\Sigma_n + \frac{\delta_4}{\alpha^4}||\Sigma_n||^2 \\
&\quad + \alpha^\beta + \frac{1}{\alpha_n^4}||\Sigma_n||^2\alpha_n^{(\beta+1)\wedge 2} + \frac{1}{\alpha}tr\Sigma_n\Big).
\end{aligned}$$

A remark is in order, the rate $\frac{1}{\alpha_n^3}tr\Sigma_n\delta_2$ can be written in an equivalent way as $\frac{1}{\alpha_n^3}(tr\Sigma_n)^3\frac{\delta_2}{(tr\Sigma_n)^2}$. The first factor of this expression converges to 0 under the hypothesis in Theorem 5. The second factor $\frac{\delta_2}{(tr\Sigma_n)^2}$ is particularly interesting. It is the square of the ratio between the rate of the estimating error of the operator and the rate of decline of the measurement error in the inverse problem (2.1). For this ratio being bounded it is necessary that the combination of estimated operators $\hat{K}\Omega_0\hat{K}^*$ does not converge too slowly with respect to the residuals. In other words, the combination of estimated operators must have at least the same speed of decline as the measurement error U, otherwise the ratio explodes as $n \to \infty$.

If the hypothesis set in the theorem hold, the terms in the rate of convergence that account for the operator estimation error are negligible with respect to the terms due to the approximation errors $||x_\alpha - x_*||^2$. Henceforth, if $tr\Sigma_n \sim ||\Sigma_n||$, for instance $\Sigma_n = \frac{1}{n}\Sigma$, and $\beta \geq \frac{1}{2}$, we find the same optimal regularization parameter as in the case with known

K: $\alpha_n^* \propto ||\Sigma_n||^{\frac{1}{\beta+1}}$ and the optimal speed of convergence is $||\Sigma_n||^{\frac{\beta}{\beta+1}}$.

Analysis of the asymptotic behavior of the estimated regularized posterior variance requires a decomposition similar to that one done for the mean: $\hat{V}_\alpha = (\hat{V}_\alpha - V_\alpha) + V_\alpha$. Here, we only have the error about estimation of K. Then, $||\hat{V}_\alpha||^2 \leq ||\hat{V}_\alpha - V_\alpha||^2 + ||V_\alpha||^2$ and we have the following theorem.

Theorem 6 *Under Assumptions 25 and 6 (a) and (d), if $\alpha_n \to 0$, $\frac{\delta_1}{\alpha_n^2} \sim \mathcal{O}_p(1)$, $\frac{\delta_4}{\alpha_n^2} \sim \mathcal{O}_p(1)$ and $\frac{1}{\alpha_n^3}||\Sigma_n||^2 \sim \mathcal{O}_p(1)$ then $\forall \varphi \in \mathcal{X}$*

$$||\hat{V}_\alpha \varphi||^2 \to 0$$

in P^{x_*}-probability as $n \to \infty$.

Moreover, if $\hat{V}_\alpha$ is applied to $\varphi \in \mathcal{X}$ such that $\Omega_0^{\frac{1}{2}} \varphi \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, for some $\beta > 0$, it is of order

$$||\hat{V}_\alpha \varphi||^2 = \mathcal{O}_p\Big(\frac{\delta_1}{\alpha_n^2}\alpha_n^\beta + \frac{\delta_4}{\alpha_n^4}||\Sigma_n||^2 + \frac{1}{\alpha_n^4}||\Sigma_n||^2\alpha_n^{(\beta+1)\wedge 2} + \alpha_n^\beta\Big).$$

If we set the regularization parameter α_n equal to the optimal one α_n^*, this guarantees convergence to zero of $||\hat{V}_\alpha \varphi||^2$ at the speed $||\Sigma_n||^{\frac{\beta}{\beta+1}}$ if $\beta \geq \frac{1}{2}$. Moreover, under all the hypothesis given in Theorems 5 and 6, the estimated regularized posterior distribution $\hat{\mu}_\alpha^{\mathcal{F}}$ converges at the rate of $||\Sigma_n||^{\frac{\beta}{2(\beta+1)}}$. This is proved by using a Chebyshev's inequality:

$$\hat{\mu}_\alpha^{\mathcal{F}}\{x : ||x - x_*|| \geq M_n \varepsilon_n\} \leq \frac{||\hat{V}_\alpha(x|\hat{Y})|| + ||\hat{\mathbb{E}}_\alpha(x|\hat{Y}) - x_*||^2}{(M_n \varepsilon_n)^2}.$$

By summarizing, when operator K is unknown and substituted by some nonparametric consistent estimator with suitable rate of convergence, the speed of convergence of the regularized posterior distribution is not affected and we are able to achieve the same optimal speed as when the operator is known.

2.6 The case with different operator for each observation

We present in this section a slightly modified version of model (2.1). Suppose to observe an n-sample of Hilbert-valued curves $\hat{Y}_1, \ldots, \hat{Y}_n$, with $\hat{Y}_i \in \mathcal{Y}$, $\forall i$, each of them is a noisy transformation of the parameter of interest x through an observation-specific transformation K_i, namely operator K changes with the index of observation. More clearly,

$$\hat{Y}_i = K_i x + U_i \qquad i = 1, \ldots, n \qquad U_i \sim iid \tag{2.13}$$

where we still assume K_i, $i = 1, \ldots, n$, is a known, linear, non-random and Hilbert-Schmidt operator. This is the classical linear regression model with fixed regressors, where the operators $\{K_i\}$ play the role of explanatory variables.

The observational scheme is the following: either we directly observe n curves or we dispose of n samples of discrete observations and through a mathematical transformation of the observations in each of these samples we obtain n infinite dimensional objects $\hat{Y}_i$. In the first case we observe infinite dimensional elements of the Hilbert space $\mathcal{Y}$, in the second case we observe elements in $\mathbb{R}^l$, for some $l > 0$.

The parameter of interest x belongs to the probability space $(\mathcal{X}, \mathcal{E}, \mu)$ where the prior distribution μ is still assumed to be gaussian with mean function x_0 and covariance operator Ω_0. On the contrary, the sampling distribution is not the same for all the observations and it is specified in Assumption 7 below.

Assumption 7 *Let P_i^x be a probability measure on $(\mathcal{Y}, \mathcal{F})$ conditioned on $\mathcal{E}$ such that $\mathbb{E}(||\hat{Y}_i||^2) < \infty$, where $\mathbb{E}$ denotes the expectation taken with respect to P_i^x. P_i^x is a Gaussian measure that defines a mean element $K_i x \in \mathcal{Y}$ and a covariance operator $\Sigma : \mathcal{Y} \to \mathcal{Y}$.*

Therefore,

$$\hat{Y}_i | x \sim i\mathcal{GP}(K_i x, \Sigma) \qquad i = 1, \dots, n. \tag{2.14}$$

and Σ satisfies the usual properties of covariance operators (*i.e.* it is linear, bounded, nonnegative, self-adjoint and trace-class).

Throughout this section we will adopt the following notation: $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_n)'$ is the $n \times 1$ vector of noisy observations, $U = (U_1, \dots, U_n)'$ is the $n \times 1$ vector of error terms, $K = (K_1, \dots, K_n)' : \mathcal{X} \to \mathcal{Y}^n$ is the $n \times 1$ vector of observation-specific operators, $K^* = (K_1^*, \dots, K_n^*) : \mathcal{Y}^n \to \mathcal{X}$ is the $1 \times n$ adjoint vector of K. Moreover, $(\mathcal{Y}^n, \mathcal{F}^n)$ denotes the product of the measurable spaces $\{(\mathcal{Y}, \mathcal{F}) : i = 1, \dots, n\}$ and the joint sampling measure on it will be denoted with $P^x \equiv P_1^x \otimes \dots \otimes P_n^x$. The corresponding Bayesian Experiment will be defined by the following probability space, denoted with Ξ_D:

$$\Xi_D = (\mathcal{X} \times \mathcal{Y}^n, \mathcal{E} \otimes \mathcal{F}^n, \Pi^n), \tag{2.15}$$

where $\Pi^n = \mu \otimes P_1^x \otimes \dots \otimes P_n^x$.

The following Lemma is only an adaptation of Theorem 1 to the particular case with different operators. For this reason the proof is omitted.

Lemma 2 *Under Assumptions 7 and 2:*

(i) *the joint measure Π^n on $(\mathcal{X} \times \mathcal{Y}^n, \mathcal{E} \otimes \mathcal{F}^n)$ is Gaussian with mean function $m_{xy}^n = (x_0, K x_0) \in \mathcal{X} \times \mathcal{Y}^n$ and covariance operator Υ^n such that $\Upsilon^n(\varphi, \psi) = (\Omega_0 \varphi + \Omega_0 K^* \psi, K \Omega_0 \varphi + (I_n \otimes \Sigma + K \Omega_0 K^*) \psi)$, for all (φ, ψ) in $\mathcal{X} \times \mathcal{Y}^n$.*

(ii) *The marginal distribution $P = P_1 \otimes \dots \otimes P_n$ on $(\mathcal{Y}^n, \mathcal{F}^n)$ is a gaussian measure with mean function $m_y^n = K x_0 \in \mathcal{Y}^n$ and covariance operator $\Upsilon_{yy}^n = (I_n \otimes \Sigma + K \Omega_0 K^*)$.*

For clarifying the writing, we rewrite in matrix form the covariance operator of the marginal distribution of the n-dimensional gaussian process $\hat{Y}$.

$$
\begin{aligned}
Var(\hat{Y}) &= (I_n \otimes \Sigma + K \Omega_0 K^*) \\
&= \begin{pmatrix}
\Sigma + K_1 \Omega_0 K_1^* & K_1 \Omega_0 K_2^* & \dots & K_1 \Omega_0 K_n^* \\
K_2 \Omega_0 K_1^* & \Sigma + K_2 \Omega_0 K_2^* & & \\
\vdots & & \ddots & \vdots \\
K_n \Omega_0 K_1^* & \dots & & \Sigma + K_n \Omega_0 K_n^*
\end{pmatrix},
\end{aligned}
$$

where I_n is the n-dimensional diagonal matrix with the non-null elements equal to identity operators.

2.6.1 Marginalization of the Bayesian experiment

In order to simplify long computations caused by large amount of statistical data, we can reduce Bayesian experiment (2.15) through a marginalization of it. We consider a marginalization on the sample space, namely, if $\mathcal{T} \subset \mathcal{F}$ is the sub-σ-field generated by

a sufficient statistic t defined on the sample space $(\mathcal{Y}^n, \mathcal{F}^n)$, we are considering the restriction of Π^n on $\mathcal{E} \otimes \mathcal{T}$, denoted with $\Pi^n_{\mathcal{E} \otimes \mathcal{T}}$ and defined as the *trace* of Π^n on $\mathcal{E} \otimes \mathcal{T}$, i.e. $\Pi^n_{\mathcal{E} \otimes \mathcal{T}}(A) = \Pi^n(A)$, $\forall A \in \mathcal{E} \otimes \mathcal{T}$. In the following, we work with the statistic $t = \sum_{i=1}^n K_i^* \hat{Y}_i \ (\equiv K^* \hat{Y})$ and we show in Appendix 2.8 that it is sufficient for the Bayesian experiment Ξ_D. To prove sufficiency is easy in finite dimension, but in infinite dimension it is more troublesome and requires more caution. By exploiting the sufficiency property and Assumption 7 we get that t, conditioned on x, induces a gaussian measure on $\mathcal{X}$: $K^* \hat{Y} | x \sim \mathcal{GP}(K^* K x, K^* \Sigma K)$, where $K^* \Sigma K = \sum_i K_i^* \Sigma K_i$.

The statistic t is not well-defined for n large, so that we should divide it by n. We denote $t_n = \frac{1}{n} K^* \hat{Y}$ the scaled sufficient statistic. Actually, the factor scaling $\frac{1}{n}$ is appropriate only if the data are *i.i.d.*, if, for instance, we had time series data we could necessitate to divide by n^2 if there is a trend. Also the asymptotic behavior of $K^* \Sigma K$ and $K^* K = \sum_i K_i^* K_i$ are noteworthy. If these operators are deterministic, then we should divide this sum by n to guarantee summability of the series. On the contrary, K_i could be a transformation of some sample of random variables w_i, therefore the rate for which we have to divide t depends on properties of w_i. For keeping thinks simple, we assume in this analysis that the data are such that the scaling factor $\frac{1}{n}$ is appropriate for all the sums.

Let $\mathcal{T}_T$ be the σ-field generated by t_n, i.e. $\mathcal{T}_T = \sigma(\frac{1}{n} K^* \hat{Y})$. The sampling probability of t_n restricted to $\mathcal{T}_T$, $P^x_{\mathcal{T}_T}$, is gaussian with mean function $\frac{1}{n} K^* K x$ and covariance operator $\frac{1}{n^2} K^* \Sigma K$. In the analysis of asymptotic properties, we shall assume that $\frac{1}{n} K^* \Sigma K$ converges towards a well-defined operator, so that $\frac{1}{n^2} K^* \Sigma K$ converges to 0 with n. The joint measure restricted to $\mathcal{E} \otimes \mathcal{T}_T$ is gaussian with mean $(x_0, \frac{1}{n} K^* K x_0)$ and covariance operator

$$
\begin{bmatrix}
\Omega_0 & \Omega_0 \frac{K^* K}{n} \\
\frac{K^* K}{n} \Omega_0 & \left(\frac{1}{n} \left(\frac{K^* \Sigma K}{n} \right) + \frac{K^* K}{n} \Omega_0 \frac{K^* K}{n} \right)
\end{bmatrix}
$$

and the marginal distribution $P_{\mathcal{T}_T}$ restricted to $\mathcal{T}_T$ is also Gaussian.

The solution to the ill-posed problem (2.13) is the regularized conditional distribution $\mu^{\mathcal{T}_T}_\alpha$ of x given the observed t_n, that we denote with $\mu^{\mathcal{T}_T}_\alpha$ for analogy with previous notation. The regularization is obtained with a Tikhonov scheme and the regularized quantities defining the posterior distribution $\mu^{\mathcal{T}_T}_\alpha$ are

$$
\begin{aligned}
A_{D,\alpha} &= \Omega_0 \frac{K^* K}{n} \left(\alpha_n I + \frac{1}{n} \left(\frac{K^* \Sigma K}{n} \right) + \frac{K^* K}{n} \Omega_0 \frac{K^* K}{n} \right)^{-1}, \\
b_{D,\alpha} &= (I - A_{D,\alpha} \frac{K^* K}{n}) x_0, \\
V_{D,\alpha} &= \Omega_0 - A_{D,\alpha} \frac{K^* K}{n} \Omega_0 \\
&= \Omega_0 - \Omega_0 \frac{K^* K}{n} \left(\alpha_n I + \frac{1}{n} \left(\frac{K^* \Sigma K}{n} \right) + \frac{K^* K}{n} \Omega_0 \frac{K^* K}{n} \right)^{-1} \frac{K^* K}{n} \Omega_0 \quad (2.16)
\end{aligned}
$$

Therefore, the regularized posterior mean is

$$
\begin{aligned}
\mathbb{E}_{D,\alpha}(x|t) &= \Omega_0 \frac{K^* K}{n} \left(\alpha_n I + \frac{1}{n} \left(\frac{K^* \Sigma K}{n} \right) + \frac{K^* K}{n} \Omega_0 \frac{K^* K}{n} \right)^{-1} t + \\
&\quad (I - \Omega_0 \frac{K^* K}{n} \left(\alpha_n I + \frac{1}{n} \left(\frac{K^* \Sigma K}{n} \right) + \frac{K^* K}{n} \Omega_0 \frac{K^* K}{n} \right)^{-1} \frac{K^* K}{n}) x_0 \quad (2.17)
\end{aligned}
$$

Note that also in this extension of the basic model we could use a Tikhonov regularization in the prior variance Hilbert Scale. We do not develop this here since, as we have already stressed in Section 2.5, computations and asymptotic analysis can be easily recovered by minor modifications.

2.6.2 Asymptotic Analysis

We analyze asymptotic properties and we check that posterior consistency is verified. The arguments for deriving convergence and speed of convergence are essentially the same as in previous sections, thus the details given here and in the proofs of theorems will be minimal. To have well-defined sums of operators for $n \to \infty$ we shall assume that $\frac{1}{n} \sum_i K_i^* \Sigma K_i \to Q_1$ and $= \frac{1}{n} \sum_i K_i^* K_i \to Q_2$, for Q_1 and Q_2 two bounded operators. Posterior consistency means convergence of the regularized posterior distribution toward the Dirac mass in x_* and convergence will be in $P_{\mathcal{T}_T}^{x_*}$-probability. Consistency is obtained under a slightly modified identification condition: Assumption 25 is substituted by

Assumption 8 *The operator $K^* K \Omega_0^{\frac{1}{2}} : \mathcal{X} \to \mathcal{X}$ is one-to-one on $\mathcal{X}$.*

The following Theorem formalizes convergence to zero of the squared norm of the bias of the regularized posterior mean. In the following of this section we assume that $\beta \leq 2$ and if $\beta > 2$ then it must be set to 2 (comments on this fact is given in previous sections).

Theorem 7 *Consider inverse problem (2.13) and the regularized posterior distribution $\mu_\alpha^{\mathcal{T}_T}$ with mean and variance defined in (2.17) and in (2.16), respectively. Under Assumptions 24 and 8, if $\alpha_n \to 0$, $\frac{1}{n} \sum_i K_i^* \Sigma K_i \to Q_1$, $\frac{1}{n} \sum_i K_i^* K_i \to Q_2$ with Q_1 and Q_2 two bounded operators, $\frac{1}{\alpha_n} tr(\frac{Q_1}{n}) \to 0$ and $\frac{1}{\alpha_n^3} ||\frac{Q_1}{n}||^2 \sim \mathcal{O}_p(1)$ then:*

$$\mathbb{E}_\alpha(x|t) \to^{P_{\mathcal{T}_T}^{x_*}} x_*$$

in $\mathcal{X}$ norm. Moreover, if $\delta_* \in \mathcal{R}(\Omega_0^{\frac{1}{2}} \frac{K^* K}{n} \frac{K^* K}{n} \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ the bias is of order:

$$||\mathbb{E}_{D,\alpha}(x|t) - x_*||^2 = \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n} tr\left(\frac{Q_1}{n}\right) + \frac{1}{\alpha_n^4} \left|\left|\frac{Q_1}{n}\right|\right|^2 \alpha_n^{(\beta+1)\wedge 2}).$$

It should be noted that the third rate is negligible with respect to the other two if $\frac{1}{\alpha_n^3} \left|\left|\frac{Q_1}{n}\right|\right|^2$ is bounded. In general we can suppose $tr(\frac{Q_1}{n})$ is of the same order as $||\frac{Q_1}{n}||$, in particular, this is satisfied when $tr(\frac{Q_1}{n}) \sim ||\frac{Q_1}{n}|| \sim \mathcal{O}_p(\frac{1}{n})$ that is very frequent for U_i being an estimation error. We deduce that the optimal α_n is determined by equating the first two rates: $\alpha_n^\beta = \frac{1}{\alpha_n} tr(\frac{Q_1}{n^2})$. Thus,

$$\alpha_n^* \propto \left|\left|\frac{Q_1}{n}\right|\right|^{\frac{1}{\beta+1}} \quad \text{and} \quad ||\mathbb{E}_{D,\alpha}(x|t) - x_*||^2 = \mathcal{O}_p\left(\left|\left|\frac{Q_1}{n}\right|\right|^{\frac{\beta}{\beta+1}}\right).$$

For having consistency of the posterior distribution we also need to prove convergence to zero of the regularized posterior variance. This is shown in the next theorem:

Theorem 8 *Under Assumption 8, if $\alpha_n \to 0$, $\frac{1}{n} \sum_i K_i^* \Sigma K_i \to Q_1$, $\frac{1}{n} \sum_i K_i^* K_i \to Q_2$, with Q_1 and Q_2 two bounded operators, and $\frac{1}{\alpha_n^3} ||\frac{Q_1}{n}||^2 \sim \mathcal{O}_p(1)$ then for all $\varphi \in \mathcal{X}$:*

$$V_{D,\alpha}(x|t)\varphi \to^{P_{\mathcal{T}_T}^{x_*}} 0$$

in $\mathcal{X}$ norm.
Moreover, if $V_{D,\alpha}$ is applied to $\varphi \in \mathcal{X}$ such that $\Omega_0^{\frac{1}{2}} \varphi \in \mathcal{R}(\Omega_0^{\frac{1}{2}} (\frac{K^* K}{n})^2 \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, the squared norm of $V_{D,\alpha}$ is of order:

$$||V_{D,\alpha}\varphi||^2 = \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^4} \left|\left|\frac{Q_1}{n}\right|\right|^2 \alpha_n^{(\beta+1)\wedge 2}).$$

The optimal speed of convergence of the norm of the regularized posterior variance $V_{D,\alpha}$ and of the regularized posterior distribution $\mu_\alpha^{\mathcal{T}_T}$ is of order $||\frac{Q_1}{n}||^{\frac{\beta}{2(\beta+1)}}$.

2.7 Conclusions

This chapter analyzes posterior distribution of the solution of a functional equation in Hilbert Spaces. When the parameter of interest is of infinite dimension its posterior mean is not continuous. What is new in this paper is the construction of a new kind of posterior distribution that we call *Regularized Posterior Distribution* and that has the important property to be continuous in the observed quantity.

We have computed the regularized posterior distribution in two ways: with a classical Tikhonov regularization scheme and with a Tikhonov regularization in the prior variance Hilbert Scale. The Hilbert Scale that we use is naturally suggested by the prior distribution and it is not chosen ad-hoc as usually happens in inverse problems literature.

The regularization parameter α_n is in practice unknown. An estimation method for it is the data-driven method discussed in [19] Ch. 4., and implemented, among others, in [29]. Alternatively, a new method that we have suggested consists in putting a prior distribution on it and obtain an estimator from its posterior distribution.

In this paper we have considered the basic case with both K and Σ_n known. We have extended this basic model in [32] where we consider the cases where K is unknown, where operator K is specific to every observation and the case with partially unknown Σ_n.

2.8 Appendix A: Proofs

Proof of Lemma1

Note that $tr(\Sigma_n + K\Omega_0 K^*) = tr\Sigma_n + tr(K\Omega_0 K^*)$. Since Σ_n is trace class, we only have to prove that $K\Omega_0 K^*$ is trace class, or that $\Omega_0^{\frac{1}{2}} K^*$ is an HS operator.

Let $\Omega_0^{\frac{1}{2}} = \int_{\mathbb{R}} a(z,t)g(t)dt$ and $K^* = \int_{\mathbb{R}} b(s,t)f(s)ds$ with g and f measures on $\mathbb{R}$, then $\Omega_0^{\frac{1}{2}} K^* = \int_{\mathbb{R}\times\mathbb{R}} a(z,t)b(s,t)g(t)f(s)dsdt$ and its HS square norm is

$$
\int_{\mathbb{R}\times\mathbb{R}} \Big| \int_{\mathbb{R}} a(z,t)b(s,t)g(t)dt \Big|^2 f(s)h(z)dsdz
$$

$$
\leq \int_{\mathbb{R}\times\mathbb{R}} \Big(\int_{\mathbb{R}} |a(z,t)b(s,t)|g(t)dt \Big)^2 f(s)h(z)dsdz
$$

$$
\leq \int_{\mathbb{R}\times\mathbb{R}} \Big(\Big(\int_{\mathbb{R}} a^2(z,t)g(t)dt \Big)^{\frac{1}{2}} \Big(\int_{\mathbb{R}} b^2(s,t)g(t)dt \Big)^{\frac{1}{2}} \Big)^2 f(s)h(z)dsdz
$$

$$
= \int_{\mathbb{R}} \int_{\mathbb{R}} a^2(z,t)g(t)h(z)dtdz \int_{\mathbb{R}} \int_{\mathbb{R}} b^2(s,t)g(t)f(s)dsdt
$$

$$
< \infty
$$

since both $\Omega_0^{\frac{1}{2}}$ and K^* are Hilbert Schmidt operators. This prove that $\Omega_0^{\frac{1}{2}} K^*$ is Hilbert Schmidt and then $(\Sigma_n + K\Omega_0 K^*)$ is trace-class.

Let now consider Υ:

$$
\Upsilon = \left[\begin{array}{cc} \Omega_0 & \Omega_0 K^* \\ K\Omega_0 & \Sigma_n + K\Omega_0 K^* \end{array} \right].
$$

Let $e_j = (e_{1j}, e_{2j})$ be a basis in $\mathcal{X} \times \mathcal{Y}$, the trace of Υ is:

$$
\begin{aligned}
tr(\Upsilon) &= \sum_j <\Upsilon e_j, e_j> \\
&= \sum_j (<\Omega_0 e_{1j}, e_{1j}> + <\Omega_0 K^* e_{2j}, e_{1j}> + <K\Omega_0 e_{1j}, e_{2j}> \\
&\quad + <(\Sigma_n + K\Omega_0 K^*)e_{2j}, e_{1j}>).
\end{aligned}
$$

For the above part of this proof and since Ω_0 is trace-class, the infinite sum of the first and last terms are finite. We only have to consider the two terms in the center: $\sum_j (< \Omega_0 K^* e_{2j}, e_{1j} > + < K\Omega_0 e_{1j}, e_{2j} >)$. This term is equal to $2\sum_j < \Omega_0^{\frac{1}{2}} K^* e_{2j}, \Omega_0^{\frac{1}{2}} e_{1j} >$ and

$$
\begin{aligned}
2\sum_j < \Omega_0^{\frac{1}{2}} K^* e_{2j}, \Omega_0^{\frac{1}{2}} e_{1j} > &\leq 2\sum_j ||\Omega_0^{\frac{1}{2}} K^* e_{2j}|| \sup_j ||\Omega_0^{\frac{1}{2}} e_{1j}|| \\
&\leq 2||\Omega_0^{\frac{1}{2}}|| \sum_j ||\Omega_0^{\frac{1}{2}}|| ||K^* e_{2k}||
\end{aligned}
$$

that is finite since $\Omega_0^{\frac{1}{2}}$ is bounded and K^* is HS. The necessity of Υ_{yy} being trace-class to have Υ trace-class is evident and this complete the proof.

Proof of Theorem 1

(i). Let $(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}$. Assumptions 1 implies that $\tilde{y} = \tilde{y}_1 + \tilde{y}_2$, with $\tilde{y}_1 \in \mathcal{R}(K)$ and $\tilde{y}_2 \in \mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.(\Sigma_n)$. Therefore, $\tilde{y}_1$ and $\tilde{y}_2$ are independent and for all $(\varphi, \psi) \in \mathcal{X} \times \mathcal{Y}$

$$
\begin{aligned}
< (\tilde{x}, \tilde{y}), (\varphi, \psi) > &= < \tilde{x}, \varphi > + < \tilde{y}_1 + \tilde{y}_2, \psi > \\
&= < \tilde{x}, \varphi > + < K\tilde{x}, \psi > + < \tilde{y}_2, \psi > \\
&= < \tilde{x}, \varphi + K^* \psi > + < \tilde{y}_2, \psi >
\end{aligned}
$$

and $< \tilde{x}, \varphi + K^* \psi > + < \tilde{y}_2, \psi >$ is distributed as

$$
= \mathcal{N}(< x_0, \varphi + K^* \psi >, < \Omega_0(\varphi + K^* \psi), (\varphi + K^* \psi) > + < \Sigma_n \psi, \psi >).
$$

We have proved that the joint measure Π on $\mathcal{X} \times \mathcal{Y}$ is gaussian. The mean m_{xy} is defined through $< m_{xy}, (\varphi, \psi) >= \mathbb{E}_\Pi < (\tilde{x}, \tilde{y}), (\varphi, \psi) >$ and since $< x_0, \varphi + K^* \psi >=< (x_0, Kx_0), (\varphi, \psi) >$ we get $m_{xy} = (x_0, Kx_0)$. From the definition of Υ, we get

$$
< \Upsilon(\varphi, \psi), (\varphi, \psi) >=< \Omega_0 \varphi, \varphi > + < (\Sigma_n + K\Omega_0 K^*)\psi, \psi >
$$

that concludes the proof.

(ii). Let Q be the projection of Π on $(\mathcal{Y}, \mathcal{F})$ with mean function m_Q and covariance operator R_Q. Since Π is gaussian, the projection must be gaussian. Moreover, $\forall \psi \in \mathcal{Y}$

$$
\begin{aligned}
< m_Q, \psi > &= < m_{xy}, (0, \psi) > \\
&= < (x_0, Kx_0), (0, \psi) > = < Kx_0, \psi >
\end{aligned}
$$

and

$$
\begin{aligned}
< R_Q \psi, \psi > &= < \Upsilon(0, \psi), (0, \psi) > \\
&= < (\Omega_0 0 + \Omega_0 K^* \psi, (\Sigma_n + K\Omega_0 K^*)\psi + K\Omega_0 0), (0, \psi) > \\
&= < (\Sigma_n + K\Omega_0 K^*)\psi, \psi > .
\end{aligned}
$$

Hence, $m_Q = m_y$ and $R_Q = \Upsilon_{yy}$. This implies $Q \equiv P$ since there is an unique correspondence between a gaussian measure and its covariance operator and mean element.

Proof of Theorem 2

Write $(\mathbb{E}_\alpha(x|\hat{Y}) - x_*)$ as:

$$
\overbrace{-[I - \Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K](x_* - x_0)}^{I}
$$
$$
+ \underbrace{[\Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}K - \Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K](x_* - x_0)}_{II}
$$
$$
+ \underbrace{\Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}U}_{III}. \tag{2.18}
$$

The first term looks very similar to the regularization bias of the solution of a functional equation. More properly, to obtain such a kind of object we use Assumption 24:

$$
\begin{aligned}
I &= [I - \Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K]\Omega_0^{\frac{1}{2}}\delta_* \\
&= \Omega_0^{\frac{1}{2}}[I - \Omega_0^{\frac{1}{2}}K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}}]\delta_*,
\end{aligned}
$$

We take the norm in $\mathcal{X}$ of I:

$$
||I||^2 \leq ||\Omega_0^{\frac{1}{2}}||^2 ||(I - \Omega_0^{\frac{1}{2}}K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}})||^2 ||\delta||^2.
$$

Note that $(I - \Omega_0^{\frac{1}{2}}K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}})$ has the same eigenvalues as

$$
[I - (\alpha_n I + \Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{-1}\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}}]. \tag{2.19}
$$

that is the regularization bias associated to the regularized solution of the ill-posed inverse problem $K\Omega_0^{\frac{1}{2}}\delta_* = r$ computed using Tikhonov regularization scheme. It converges to zero when $\alpha_n \to 0$ and then the second norm in $||I||^2$ is bounded. This way to rewrite the above operator justifies the identification condition in Assumption 25. Injectivity of $K\Omega_0^{\frac{1}{2}}$ ensures that the solution of $K\Omega_0^{\frac{1}{2}}\delta = r$ is identified.

The speed of convergence to zero of $||(I - \Omega_0^{\frac{1}{2}}K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}})||^2$ depends on the regularity of δ_*, and consequently of $(x_* - x_0)$. If $\delta_* \in \Phi_\beta$, it is at most of order α_n^β, see [10]. We admit without proof the following lemma. Then $||I||^2 = \mathcal{O}_p(\alpha_n^\beta)$.

Now, let us consider the II and III terms. We have $||II||^2 = ||\Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}(-\Sigma_n)(\alpha_n I + K\Omega_0 K^*)^{-1}K(x_* - x_0)||^2$ and it is less than or equal to

$$
||\Omega_0 K^*||^2 ||(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}||^2 ||\Sigma_n||^2 ||(\alpha_n I + K\Omega_0 K^*)^{-1}K(x_* - x_0)||^2
$$

where the first norm is bounded and the second and the third ones are $\mathcal{O}_p(\frac{1}{\alpha_n^2})$ and $\mathcal{O}_p(||\Sigma_n||^2)$ respectively. The last norm can be written as:

$$
||(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}}\delta_*||^2,
$$

and, by using the hypothesis that $\delta_* \in \Phi_\beta$

$$
||(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}}\delta_*||^2 = \frac{1}{\alpha^2}||\alpha(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}\rho||^2,
$$

for some $\rho \in \mathcal{X}$ and it is at least of order $\frac{1}{\alpha^2}\alpha^{\beta+1}$. As a consequence of the fact that, with a Tikhonov regularization, a degree of smoothness greater than or equal to 2 may be useless, we get $||(\alpha_n I + K\Omega_0 K^*)^{-1}K(x_* - x_0)||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^2}\alpha_n^{(\beta+1)\wedge 2})$.

To find speed of convergence of term III we re-write it as:

$$III = \underbrace{\Omega_0 K^*[(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1} - (\alpha_n I + K\Omega_0 K^*)^{-1}]U}_{A} +$$

$$\underbrace{\Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}U}_{B}.$$

By standard computation and by Kolmogorov theorem, it is trivial to determine that $||A||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^3}||\Sigma_n||^2 tr\Sigma_n)$ and $||B||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n} tr\Sigma_n)$, since $||U||^2$ is bounded in probability if $\mathbb{E}||U||^2 < \infty$. Finally, $\mathbb{E}||U||^2 = tr\Sigma_n$.
The first term of $||III||^2$ is negligible with respect to the other terms in $||II||^2$ and $||III||^2$.

Proof of Theorem 3

By recalling expression (2.7), we can rewrite the regularized posterior variance as

$$V_\alpha = \overbrace{\Omega_0 - \Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0}^{IV} +$$

$$\underbrace{\Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0 - \Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}K\Omega_0}_{V}.$$

Since Ω_0 is a positive definite self-adjoint operator, it can be decomposed as $\Omega_0 = \Omega_0^{\frac{1}{2}}\Omega_0^{\frac{1}{2}}$. For term IV we follow the same reasoning done for term I in (5.21), so that we conclude that, if $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, $||IV\varphi||^2 = \mathcal{O}_p(\alpha_n^\beta)$. Operator V in (2.20) applied to $\varphi \in \mathcal{X}$ is equivalently rewritten as

$$\Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}\Sigma_n(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}}\Omega_0^{\frac{1}{2}}\varphi$$

and by using the same proof as for term II in (5.21), its squared norm is bounded and of order $||V||^2 = \mathcal{O}_p(\frac{1}{\alpha_n^4}||\Sigma_n||^2\alpha_n^{(\beta+1)\wedge 2})$.

Proof of Theorem 4

We admit the following Lemma:

Lemma 3 *Let $\mathcal{X}_s$, $s \in \mathbb{R}$, be a Hilbert scale induced by L and let $T : \mathcal{X} \to \mathcal{Y}$ be a bounded operator satisfying $||x||_{-a} \sim ||Tx||$ on $\mathcal{X}$ for some $a > 0$. Then for $B := TL^{-s}$, $s \geq 0$ and $|\nu| \leq 1$*

$$||x||_{-\nu(a+s)} \sim ||(B^*B)^{\frac{\nu}{2}}x||.$$

Moreover, $\mathcal{R}((B^*B)^{\frac{\nu}{2}}) = \mathcal{X}_{\nu(a+s)}$.

<u>Proof:</u>ee proof of Corollary 8.22 in [19]. ∎

The bias $\mathbb{E}_s(x|\hat{Y}) - x_*$ is rewritten as

$$\overbrace{[I - \Omega_0 K^*(\alpha_n L^{2s} + K\Omega_0 K^*)^{-1}K](x_* - x_0)}^{I}$$
$$+\underbrace{\Omega_0 K^*[(\alpha_n L^{2s} + \Sigma_n + K\Omega_0 K^*)^{-1}K - (\alpha_n L^{2s} + K\Omega_0 K^*)^{-1}K](x_* - x_0)}_{II}$$
$$+\underbrace{\Omega_0 K^*(\alpha_n L^{2s} + \Sigma_n + K\Omega_0 K^*)^{-1}U}_{III}.$$

Let us start by considering term I, note that

$$||I|| \leq ||\Omega_0^{\frac{1}{2}}||^2||[I - (\alpha_n\Omega_0^{-s} + \Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{-1}\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}}]\Omega_0^{\frac{\beta}{2}}\rho_*||$$

if Ω_0 is such that $\Omega_0^{\frac{1}{2}}K^*(\alpha_n L^{2s} + K\Omega_0K^*)^{-1} = (\alpha_n\Omega_0^{-s} + \Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{-1}\Omega_0^{\frac{1}{2}}K^*$, *i.e.* $\Omega_0^{-s+\frac{1}{2}}K^* = \Omega_0^{\frac{1}{2}}K^*L^{2s}$. By using Assumption 5 *(ii)* and the notation $B = K\Omega_0^{\frac{s+1}{2}}$, we rewrite

$$
\begin{aligned}
||I|| &\leq ||\Omega_0^{\frac{1}{2}}||^2||\Omega_0^{\frac{s+1}{2}}(I - (\alpha_n I + B^*B)^{-1}B^*B)\Omega_0^{\frac{\beta-s}{2}}\rho_*|| \\
&\leq ||\Omega_0^{\frac{1}{2}}||^2||\Omega_0^{\frac{s+1}{2}}(I - (\alpha_n I + B^*B)^{-1}B^*B)(B^*B)^{\frac{\beta-s}{2(a+s)}}v|| \\
&\sim ||(B^*B)^{\frac{\beta+1}{2(a+s)}}\alpha_n(\alpha_n I + B^*B)^{-1}v|| \\
&\sim \mathcal{O}_p(\alpha_n^{\frac{\beta+1}{2(a+s)}})
\end{aligned}
$$

where the second line follows from the fact that $\mathcal{R}(\Omega_0^{\frac{\beta-s}{2}}) \equiv \mathcal{X}_{\beta-s} \equiv \mathcal{R}((B^*B)^{\frac{\beta-s}{2(a+s)}})$, then $\Omega_0^{\frac{\beta-s}{2}}\rho_* = (B^*B)^{\frac{\beta-s}{2(a+s)}}v$, for some $v \in \mathcal{X}$. The third equivalence is a consequence of Lemma 3. It follows that $||I||^2 \sim \mathcal{O}_p(\alpha_n^{\frac{\beta+1}{a+s}})$.

We use similar steps for obtaining the convergence of the other terms, so that we omit any redundant comment.

$$||II|| \leq ||\Omega_0 K^*(\alpha_n L^{2s} + \Sigma_n + K\Omega_0 K^*)^{-1}||||\Sigma_n||||(\alpha_n L^{2s} + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}}\delta_*||$$

and the norm in the last term can be developed as

$$
\begin{aligned}
||(\alpha_n L^{2s} + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}}\delta_*|| &= ||K\Omega_0^{\frac{1}{2}}(\alpha_n\Omega_0^{-s} + \Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{-1}\delta_*|| \\
&= ||B(\alpha_n I + B^*B)^{-1}\Omega_0^{\frac{s+\beta}{2}}v|| \\
&\sim ||(B^*B)^{\frac{2s+\beta+a}{2(a+s)}}(\alpha_n I + B^*B)^{-1}v|| \\
&\sim \mathcal{O}_p(\frac{1}{\alpha_n}\alpha_n^{\frac{2s+\beta+a}{2(a+s)}}).
\end{aligned}
$$

Thus, $||II||^2 \sim \mathcal{O}_p\left(\frac{1}{\alpha_n^4}||\Sigma_n||^2\alpha_n^{\frac{2s+\beta+a}{(a+s)}}\right)$.

We proceed with term III that can be decomposed as

$$
\begin{aligned}
III &= \underbrace{\Omega_0 K^*[(\alpha_n L^{2s} + \Sigma_n + K\Omega_0 K^*)^{-1} - (\alpha_n L^{2s} + K\Omega_0 K^*)^{-1}]U}_{IIIA} + \\
&\quad \underbrace{\Omega_0 K^*(\alpha_n L^{2s} + K\Omega_0 K^*)^{-1}U}_{IIIB},
\end{aligned}
$$

where the squared norm $||IIIA||^2$ of the first term is less or equal then

$$
\begin{aligned}
&||\Omega_0 K^*(\alpha_n L^{2s} + K\Omega_0 K^*)^{-1}||^2||\Sigma_n||^2||(\alpha_n L^{2s} + \Sigma_n + K\Omega_0 K^*)^{-1}||^2||U||^2 \\
\leq\ &||\Omega_0^{\frac{s+1}{2}}(\alpha_n I + \Omega_0^{\frac{s+1}{2}}K^*K\Omega_0^{\frac{s+1}{2}})^{-1}\Omega_0^{\frac{s+1}{2}}K^*||^2||\Sigma_n||^2||(\alpha_n L^{2s} + \Sigma_n + K\Omega_0 K^*)^{-1}||^2||U||^2 \\
\sim\ &||(B^*B)^{\frac{a+2s+1}{2(a+s)}}(\alpha_n I + B^*B)^{-1}||^2||\Sigma_n||^2||(\alpha_n L^{2s} + \Sigma_n + K\Omega_0 K^*)^{-1}||^2||U||^2 \\
\sim\ &\mathcal{O}_p\left(\frac{1}{\alpha_n^4}||\Sigma_n||^2 tr\Sigma_n\alpha_n^{\frac{a+2s+1}{a+s}}\right).
\end{aligned}
$$

The norm of the term $IIIB$ is:

$$
\begin{aligned}
||IIIB|| &= ||\Omega_0 K^*(\alpha_n L^{2s} + K\Omega_0 K^*)^{-1} U|| \\
&= ||\Omega_0^{\frac{1}{2}}(\alpha_n \Omega_0^{-s} + \Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{-1} \Omega_0^{\frac{1}{2}} K^* U|| \\
&= ||\Omega_0^{\frac{s+1}{2}}(\alpha_n I + B^* B)^{-1} B^* U|| \\
&\sim ||(B^* B)^{\frac{s+1}{2(a+s)}}(\alpha_n I + B^* B)^{-1} B^* U|| \\
&\leq ||(B^* B)^{\frac{2s+a+1}{2(a+s)}}(\alpha_n I + B^* B)^{-1}|| \, ||U|| \\
&\sim \mathcal{O}_p(\alpha_n^{\frac{1-a}{2(a+s)}} ||U||).
\end{aligned}
$$

Thus $||IIIB||^2 \sim \mathcal{O}_p(\alpha_n^{\frac{1-a}{(a+s)}} tr\Sigma_n)$.

The variance V_s is applied to an element $\varphi \in \mathcal{X}$ such that $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{\beta}{2}})$ and $\Omega_0^{\frac{s+1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{\beta-s}{2}})$. Then the variance can be decomposed as

$$
V_s \varphi = \overbrace{[\Omega_0 - \Omega_0 K^*(\alpha_n L^{2s} + K\Omega_0 K^*)^{-1} K\Omega_0]\varphi}^{IV} \\
+ \underbrace{\Omega_0 K^*[(\alpha_n L^{2s} + K\Omega_0 K^*)^{-1} - (\alpha_n L^{2s} + \Sigma_n + K\Omega_0 K^*)^{-1}]K\Omega_0\varphi}_{V}.
$$

Computation of $||IV||$ is specular to that one for term $||I||$ above and computation of $||V||$ to that one for term $||II||$, therefore we give only the result: $||IV||^2 \sim \mathcal{O}_p(\alpha_n^{\frac{\beta+1}{a+s}})$ and $||V||^2 \sim \mathcal{O}_p\left(\frac{1}{\alpha_n^2}||\Sigma_n||^2 \alpha_n^{\frac{\beta-a}{(a+s)}}\right)$.

The result follows.

Proof of Theorem 5

For brevity, let $H = K\Omega_0^{\frac{1}{2}}$, $\hat{H} = \hat{K}\Omega_0^{\frac{1}{2}}$, $H^* = \Omega_0^{\frac{1}{2}} K^*$ and $\hat{H}^* = \Omega_0^{\frac{1}{2}} \widehat{K}^*$. Moreover, we denote $\tilde{Y} = \hat{K}x_* + U$, $\tilde{x}_\alpha = \mathbb{E}_\alpha(x|\tilde{Y})$ and we use decomposition (2.12) of the regularized estimation bias. We start to analyze the estimation error about $Y - Kx_*$:

$$
\begin{aligned}
\hat{x}_\alpha - \tilde{x}_\alpha &= \Omega_0 \widehat{K}^*(\alpha I + \hat{K}\Omega_0 \widehat{K}^* + \Sigma_n)^{-1}(K - \hat{K})x_* \\
&= \Omega_0^{\frac{1}{2}}[\hat{H}^*(\alpha I + \hat{H}\hat{H}^* + \Sigma_n)^{-1} - \hat{H}^*(\alpha I + \hat{H}\hat{H}^*)^{-1}](K - \hat{K})x_* \\
&\quad + \Omega_0^{\frac{1}{2}} \hat{H}^*(\alpha I + \hat{H}\hat{H}^*)^{-1}(K - \hat{K})x_*.
\end{aligned}
$$

Then, $||\hat{x}_\alpha - \tilde{x}_\alpha||^2$ is less or equal than

$$
\begin{aligned}
&||\Omega_0^{\frac{1}{2}}||^2 ||\hat{H}^*(\alpha_n I + \hat{H}\hat{H}^*)^{-1}||^2 (||\Sigma_n||^2 ||(\alpha I + \hat{H}\hat{H}^* + \Sigma_n)^{-1}||^2 ||(K - \hat{K})x_*||^2 + \\
&\quad ||(K - \hat{K})x_*||^2) \\
&\sim \mathcal{O}_p(\frac{1}{\alpha_n^3}||\Sigma_n||^2 ||\hat{K} - K||^2 + \frac{||\hat{K} - K||^2}{\alpha_n}) \\
&\sim \mathcal{O}_p(\frac{1}{\alpha_n^3}||\Sigma_n||^2 \delta_4 + \frac{\delta_4}{\alpha_n}),
\end{aligned}
$$

The second term $\tilde{x}_\alpha - x_\alpha$ in decomposition (2.12) is equal to

$$
\overbrace{\Omega_0[\widehat{K}^*(\alpha_n I + \hat{K}\Omega_0 \widehat{K}^* + \Sigma_n)^{-1}\hat{K} - K^*(\alpha_n I + K\Omega_0 K^* + \Sigma_n)^{-1}K](x_* - x_0)}^{A} \\
+ \underbrace{\Omega_0[\widehat{K}^*(\alpha_n I + \hat{K}\Omega_0 \widehat{K}^* + \Sigma_n)^{-1} - K^*(\alpha I + K\Omega_0 K^* + \Sigma_n)^{-1}K]}_{B}.
$$

Term A can in turn be decomposed as

$$
A = \overbrace{\Omega_0^{\frac{1}{2}}[\hat{H}^*(\alpha_n I + \hat{H}\hat{H}^* + \Sigma_n)^{-1}\hat{H} - \hat{H}^*(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H}]\delta_*}^{AI}
$$
$$
+ \underbrace{\Omega_0^{\frac{1}{2}}[\hat{H}^*(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H} - H^*(\alpha_n I + HH^*)^{-1}H]\delta_*}_{AII}
$$
$$
- \underbrace{\Omega_0^{\frac{1}{2}}[H^*(\alpha_n I + HH^* + \Sigma_n)^{-1}\hat{H} - H^*(\alpha_n I + HH^*)^{-1}H]\delta_*}_{AIII} .
$$

The squared norm of term AI is

$$
\begin{aligned}
||AI||^2 &\leq ||\Omega_0^{\frac{1}{2}}||^2||\hat{H}^*(\alpha_n I + \hat{H}\hat{H}^* + \Sigma_n)^{-1}||^2||\Sigma_n||^2||(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H}\delta_*||^2 \\
&\sim \mathcal{O}_p(\frac{1}{\alpha_n^2}||\Sigma_n||^2\frac{1}{\alpha_n^2}||\alpha_n(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H}\delta_*||^2)
\end{aligned}
$$

and with the first order Taylor expansion around the true value of the operator

$$
\begin{aligned}
(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H}\delta_* &\simeq (\alpha_n I + HH^*)^{-1}H\delta_* + (\alpha_n I + HH^*)^{-1}(\hat{H} - H)\delta_* \\
&\quad + (\alpha_n I + \hat{H}\hat{H}^*)^{-1}(\hat{H}\hat{H}^* - HH^*)(\alpha_n I + HH^*)^{-1}H\delta_*
\end{aligned}
$$

we get $||\alpha_n(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H}\delta_*||^2 \sim \mathcal{O}_p(\alpha_n^{(\beta+1)\wedge 2}(1 + \frac{\delta_2}{\alpha_n^2}) + \delta_4)$ by using the regularity condition $\delta_* \in \mathcal{R}(H^*H)^{\frac{\beta}{2}}$. Then $||AI||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^4}||\Sigma_n||^2[\alpha_n^{(\beta+1)\wedge 2}(1 + \frac{\delta_2}{\alpha_n^2}) + \delta_4])$ and under the hypothesis that $\frac{\delta_2}{\alpha_n^2} \sim \mathcal{O}_p(1)$ the second term in the squared brackets is negligible with respect to the first one so that $||AI||^2 \sim \mathcal{O}_p\left(\frac{1}{\alpha_n^4}||\Sigma_n||^2\left(\alpha_n^{(\beta+1)\wedge 2} + \delta_4\right)\right)$.

After permutation of operators H and H^* the squared norm $||AII||^2$ of term AII is less or equal than

$$
\begin{aligned}
&||\Omega_0^{\frac{1}{2}}||^2|| - [I - (\alpha_n I + \hat{H}^*\hat{H})^{-1}\hat{H}^*\hat{H}]\delta_* + [I - (\alpha_n I + H^*H)^{-1}H^*H]\delta_*||^2 \\
\leq\ &||\Omega_0^{\frac{1}{2}}||^2||(\alpha_n I + \hat{H}^*\hat{H})^{-1}||^2||\hat{H}^*\hat{H} - H^*H||^2||\alpha_n(\alpha_n I + H^*H)^{-1}||^2 \\
\leq\ &\sim \mathcal{O}_p(\frac{\delta_1}{\alpha_n^2}\alpha_n^\beta).
\end{aligned}
$$

We use the same logic to recover the rate of $||AIII||^2$: $||AIII||^2$ is less or equal than

$$
\begin{aligned}
&||\Omega_0^{\frac{1}{2}}||^2||H^*(\alpha_n I + H^*H + \Sigma_n)^{-1}||^2||\Sigma_n||^2||(\alpha_n I + H^*H)^{-1}H\delta_*||^2 \\
\sim\ &\mathcal{O}_p(\frac{1}{\alpha_n^4}||\Sigma_n||^2\alpha_n^{(\beta+1)\wedge 2}).
\end{aligned}
$$

Therefore, $||A||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^4}||\Sigma_n||^2\alpha_n^{(\beta+1)\wedge 2} + \frac{\delta_1}{\alpha_n^2}\alpha_n^\beta + \frac{\delta_4}{\alpha_n^4}||\Sigma_n||^2)$. Lastly, term B can be decomposed in a way specular to that one used for A and the proof is substantially identical, then omitted. We conclude that $||B||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^3}||\Sigma_n||^2 tr\Sigma_n + \frac{\delta_2}{\alpha_n^3}tr\Sigma_n + \frac{\delta_3}{\alpha_n^2}tr\Sigma_n)$.

The norm of the last term in decomposition (2.12) is given by Theorem 2. By collecting all these results and by deleting the negligible terms we find the rate given in the theorem. The convergence to zero of $||\hat{x}_\alpha - x_*||^2$ is proved thanks to the rate of convergence we have determined and the conditions $\alpha_n \to 0$ and $\frac{1}{\alpha_n}tr\Sigma_n \to 0$.

Proof of Theorem 6

For brevity, let $H = K\Omega_0^{\frac{1}{2}}$, $\hat{H} = \hat{K}\Omega_0^{\frac{1}{2}}$, $H^* = \Omega_0^{\frac{1}{2}}K^*$ and $\hat{H}^* = \Omega_0^{\frac{1}{2}}\hat{K}^*$. First, we rewrite the decomposition of the estimated regularized posterior variance:

$$\begin{aligned}
\hat{V}_\alpha\varphi &= (\hat{V}_\alpha - V_\alpha)\varphi + V_\alpha\varphi \\
||\hat{V}_\alpha\varphi||^2 &\leq ||(\hat{V}_\alpha - V_\alpha)\varphi||^2 + ||V_\alpha\varphi||^2, \qquad \varphi \in \mathcal{X}.
\end{aligned}$$

Convergence and rate of convergence of the second norm in the right hand side is given in Theorem 3. The first term: $(\hat{V}_\alpha - V_\alpha)\varphi$ is equal to

$$\Omega_0[-\widehat{K^*}(\alpha_n I + \hat{K}\Omega_0\widehat{K^*} + \Sigma_n)^{-1}\hat{K} + K^*(\alpha_n I + K\Omega_0 K^* + \Sigma_n)^{-1}K]\Omega_0\varphi$$

that is less or equal than

$$\begin{aligned}
&\Omega_0^{\frac{1}{2}}[-\hat{H}^*(\alpha_n I + \hat{H}\hat{H}^* + \Sigma_n)^{-1}\hat{H} + \hat{H}^*(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H}]\Omega_0^{\frac{1}{2}}\varphi \\
&-\Omega_0^{\frac{1}{2}}[\hat{H}^*(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H} - H^*(\alpha_n I + HH^*)^{-1}H]\Omega_0^{\frac{1}{2}}\varphi \\
&+\Omega_0^{\frac{1}{2}}[H^*(\alpha_n I + HH^* + \Sigma_n)^{-1}\hat{H} - H^*(\alpha_n I + HH^*)^{-1}H]\Omega_0^{\frac{1}{2}}\varphi.
\end{aligned}$$

This expression coincide with term A in the proof 2.8 of Theorem 5 with the only difference that δ_* must be substituted with $\Omega_0^{\frac{1}{2}}\varphi$. Hence, $||(\hat{V}_\alpha - V_\alpha)\varphi||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^4}||\Sigma_n||^2\alpha_n^{(\beta+1)\wedge 2} + \frac{\delta_1}{\alpha_n^2}\alpha_n^\beta + \frac{\delta_4}{\alpha_n^4}||\Sigma_n||^2)$ and it converges to 0. By collecting the rate of convergence of the two terms in the variance decomposition and after neglecting the redundant term we get the result.

Sufficiency of $K^*\hat{Y}$

In Section 2.6.1 the use of $t = K^*\hat{Y}$ to make inference was justified under the condition that it is sufficient. We prove here that it is a sufficient statistics, namely that $\mathcal{F}^n \perp\!\!\!\perp \mathcal{E}|\sigma(K^*\hat{Y})$, where $\sigma(K^*\hat{Y})$ denotes the σ-field generated by statistic t. Actually, sufficiency of statistic t entails sufficiency of every bijective transformation of $\sum_{i=1}^n K_i^*\hat{Y}_i$.

Due to the fact that we are working in infinite dimensional spaces and we have not a likelihood function, we can not use the factorization criterion in order to prove sufficiency. Hence, we propose to consider a sequential model, obtained by projecting the model on an orthonormal bases, and to take into account only a finite number k of projections. The idea is to find a sufficient statistic for the sequential model and to analyze its asymptotic behavior.

Let $\{\lambda_j, \psi_j\}_j$ be the singular system of the covariance operator Σ, the sequence $\{\psi_j\}_j$ represents a base that will be used to project the model. A *sequential Bayesian Experiment* is defined by

$$\Xi_{Ds} = (\mathcal{X} \times \mathcal{Y}^n, \mathcal{E} \otimes \mathcal{F}^n, \Pi^n, \mathcal{E}_k \uparrow \mathcal{E}_\infty, \mathcal{F}_k^n \uparrow \mathcal{F}_\infty^n), \tag{2.20}$$

with $\mathcal{E}_k \subset \mathcal{E}_{k+1} \subset \mathcal{E}$ and $\mathcal{F}_k^n \subset \mathcal{F}_{k+1}^n \subset \mathcal{F}^n$ two filtrations in $(\mathcal{X} \times \mathcal{Y}^n, \mathcal{E} \otimes \mathcal{F}^n)$. The filtration $\mathcal{E}_k$ is generated by the projected true parameter x: $\mathcal{E}_k = \sigma(\{< x, K^*\psi_j >\}_{j=1,\ldots,k})$[6]. The filtration $\mathcal{F}_k^n$ is generated by the n-dimensional vector of projected observed curves $\hat{Y}$: $\mathcal{F}_k^n = \sigma(\{< \hat{Y}, \psi_j >\}_{j=1,\ldots,k})$. The sub-$\sigma$-field $\mathcal{E}_\infty$ and $\mathcal{F}_\infty^n$ are defined to be the σ-field generated by the random functions x and Y respectively: $\mathcal{E}_\infty = \mathcal{E}$ and $\mathcal{F}_\infty^n = \mathcal{F}^n$. The k-dimensional sequential model is written as

$$\underbrace{< \hat{Y}_i, \psi_j >_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}}}_{n\times k\ matrix:\ \{\hat{y}_{ij}\}_{ij}} = \underbrace{< K_i x, \psi_j >_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}}}_{n\times k\ matrix} + \underbrace{< U_i, \psi_j >_{\substack{i=1,\ldots,n \\ j=1,\ldots k}}}_{n\times k\ matrix},$$

with $< U_i, \psi_j > \sim \mathcal{N}(0, \lambda_j)$ and $Cov(< U_i, \psi_j >, < U_i, \psi_{j'} >) = 0$, $\forall j \neq j'$. If we consider the sequential bayesian model with only a finite number k of projections we are able to find a sufficient

---

[6]More clearly, the sub-σ-field $\mathcal{E}_k$ is identified with the sub-σ-field of cylinder sets $\mathcal{E}_k \times F = \{B \times F; B \in \mathcal{E}_k\}$.

statistic by using the factorization criterion. We use the result of the following Lemma that is proved in the next section.

Lemma 4 *Let $\hat{Y}_i(k)$ be the partial sum of the Fourier series of $\hat{Y}_i$, i.e. $\sum_{j=1}^{k} < \hat{Y}_i, \psi_j > \frac{\psi_j}{\lambda_j}$. The statistic $t = \sum_{i=1}^{n} K_i^* \hat{Y}_i(k)$ is sufficient in the sequential Bayesian experiment Ξ_{Ds}.*

Let $\mathcal{T}_k = \sigma(K^* \hat{Y}(k))$ be the sub-σ-field and we define the *tail σ-field* $\mathcal{T}_T$ as

$$\mathcal{T}_T = \bigcap_{k \geq 0} \bigvee_{m \geq k} \mathcal{T}_m, \tag{2.21}$$

that, by definition of $\mathcal{T}_k$, is equal to the smallest σ-field that makes measurable the function $K^* \hat{Y}(k)$ depending on the last coordinate, *i.e.* $\mathcal{T}_T = \sigma(K^* \hat{Y})$, see Florens *et al.* (1990) [27]. Our point is to prove that $\mathcal{T}_T$ is a sufficient statistic for the initial Bayesian experiment Ξ_D, namely: $\mathcal{F}^n \parallel \mathcal{E} | \mathcal{T}_T$. The following theorem, that is a slightly modified version of Theorem 7.2.7 in Florens *et al.* (1990) [27], shows that sufficiency in the sequential Bayesian Experiment implies sufficiency in the limit Bayesian Experiment. The proof is given in Appendix 2.8

Theorem 9 *Let $(E \times F, \mathcal{E} \otimes \mathcal{F})$ be a measurable space. Consider two filtrations in $\mathcal{E} \otimes \mathcal{F}$: $\mathcal{F}_k \uparrow \mathcal{F}_\infty$ and $\mathcal{E}_k \uparrow \mathcal{E}_\infty$ along with a sequence $\mathcal{T}_k$ adapted to $\mathcal{F}_k$, i.e.,*

(o) $\mathcal{T}_k \subset \mathcal{F}_k \ \forall k \in \mathbb{N}$.

If

(i) $\mathcal{F}_k \perp \mathcal{E}_k | \mathcal{T}_k$,

then

(ii) $\mathcal{F} \perp \mathcal{E} | \mathcal{T}_T$.

This theorem applies to Bayesian Experiment defined in (2.15) with $(E \times F, \mathcal{E} \otimes \mathcal{F})$ substituted by $(\mathcal{X} \times \mathcal{Y}^n, \mathcal{E} \otimes \mathcal{F}^n)$, $\mathcal{F}_k$ and $\mathcal{F}$ replaced by $\mathcal{F}_k^n$ and $\mathcal{F}_\infty^n \equiv \mathcal{F}^n$, respectively. Hence, $K^* \hat{Y}$ is a sufficient statistic for parameter x.

Proof of Lemma 4

In the sequel, indices i and j in the sums and products are meant to belong to $\{1, 2, \ldots, n\}$ and $\{1, 2, \ldots, k\}$, respectively. Consider the likelihood function of the sequential experiment (2.20):

$$L(\{< x, K_i^* \psi_j >\}_{ij} | \{\hat{y}_{ij}\}_{ij}) =$$
$$\prod_i [\frac{1}{(2\pi)^{\frac{k}{2}}} \prod_j \lambda_j^{-\frac{1}{2}} \exp\{-\frac{1}{2} \frac{1}{\lambda_j} (\hat{y}_{ij} - < x, K_i^* \psi_j >)^2\}],$$

and the log-likelihood $l(:= \log L)$ is proportional to the following expression:

$$
\begin{aligned}
l(\{< x, K_i^* \psi_j >\}_{ij} | \{\hat{y}_{ij}\}_{ij}) &\propto \sum_{ij} \frac{1}{\lambda_j} (\hat{y}_{ij} - < x, K_i^* \psi_j >)^2 \\
&\propto \sum_{ij} (< \hat{Y}_i, \tilde{\psi}_j >^2 + < x, K_i^* \tilde{\psi}_j >^2 \\
&\quad - 2 < \hat{Y}_i, \tilde{\psi}_j >< x, K_i^* \tilde{\psi}_j >),
\end{aligned}
\tag{2.22}
$$

where $\tilde{\psi}_j = \frac{\psi_j}{\sqrt{\lambda_j}}$ is the scaled singular value. By the factorization principle, $t \ (\equiv S(\{\hat{y}_{ij}\}_{ij}))$ is sufficient if the loglikelihood can be written as $l(\{< x, K_i^* \psi_j >\}_{ij} | \{\hat{y}_{ij}\}_{ij}) \propto f(\{\hat{y}_{ij}\}_{ij}) + g(x) + h(S(\{\hat{y}_{ij}\}_{ij}); x)$, where f, g and h are three real-valued functions. We develop the third term in (2.22) in such a way to obtain a function of x and $K^* \hat{Y}(k)$:

$$\sum_{ij} <\hat{Y}_i, \tilde{\psi}_j><x, K_i^* \tilde{\psi}_j> \quad = \quad <x, \sum_{ij} <\hat{Y}_i, \tilde{\psi}_j> K_i^* \tilde{\psi}_j>$$

$$= \quad <x, \sum_i K_i^* \sum_j <\hat{Y}_i, \tilde{\psi}_j> \tilde{\psi}_j>$$

$$= \quad <x, \sum_i K_i^* \hat{Y}_i(k)>$$

$$= \quad <x, K^* \hat{Y}(k)>,$$

where $\hat{Y}(k)$ is the partial sum of the Fourier series of $\hat{Y}$.

Proof of Theorem 9

Let a be a random variable defined on $\mathcal{E}_{k'}$, $0 \leq k' \leq k$ and belonging to L^1. Since $\mathcal{T}_k \subset \mathcal{F}_k$, *(i)* is equivalent to

(iii) $\mathbb{E}(a|\mathcal{F}_k) = \mathbb{E}(a|\mathcal{T}_k)$,

where *(iii)* is true with probability 1. Moreover, $\mathbb{E}(a|\mathcal{F}_k)$ is an $\mathcal{F}_k$-martingale, therefore by martingale properties

(iv) $\mathbb{E}(a|\mathcal{F}_k) \to^{L^1} \mathbb{E}(a|\mathcal{F}_\infty) \qquad a.s.$

Taking the $\limsup_k$ on both sides of *(iii)* we have, by using *(iv)*

$$\limsup_k \mathbb{E}(a|\mathcal{F}_k) = \limsup_k \mathbb{E}(a|\mathcal{T}_k),$$

therefore

$$\mathbb{E}(a|\mathcal{F}_\infty) = \limsup_k \mathbb{E}(a|\mathcal{T}_k) \qquad a.s.$$

Now, $\limsup_k \mathbb{E}(a|\mathcal{T}_k) = \limsup_k \mathbb{E}(a|\overline{\mathcal{T}}_k) = \mathbb{E}(a|\limsup_k \overline{\mathcal{T}}_k)$ [7] and it is a random variable defined on $\limsup_k \overline{\mathcal{T}}_k = \bigcap_{k \geq 0} \bigvee_{m \geq k} \overline{\mathcal{T}}_m = (\overline{\mathcal{T}})_T = \overline{\mathcal{T}_T}$. Then,

(v) $\mathbb{E}(a|\mathcal{F}_\infty) = \mathbb{E}(a|\overline{\mathcal{T}_T}) = \mathbb{E}(a|\mathcal{T}_T)$.

By definition of the *tail σ-field* $\mathcal{T}_T$ and filtration, we have $\mathcal{T}_T \subset \mathcal{F}_\infty$. It follows that *(v)* is equivalent to $\mathcal{F}_\infty \perp \mathcal{E}_\infty | \mathcal{T}_T$.

Proof of Theorem 7

$$||\mathbb{E}_\alpha(x|t) - x_*||^2 \quad = \quad ||\frac{1}{n} A_{D,\alpha}(K^*(Kx_* + U) - \frac{1}{n} K^* K x_0) + x_0 - x_*||^2$$

that is less or equal than

$$\leq \quad \overbrace{||\Omega_0 \frac{K^* K}{n}[\alpha_n I + \frac{1}{n}\Big(\frac{K^* \Sigma K}{n}\Big) + \frac{K^* K}{n}\Omega_0 \frac{K^* K}{n}]^{-1}\frac{K^* U}{n}||^2}^{I}$$
$$+ \underbrace{||(I - \Omega_0 \frac{K^* K}{n}[\alpha_n I + \frac{1}{n}\Big(\frac{K^* \Sigma K}{n}\Big) + \frac{K^* K}{n}\Omega_0 \frac{K^* K}{n}]^{-1}\frac{K^* K}{n})(x_* - x_0)||^2}_{II}.$$

---

[7] In general, for a σ-field $\mathcal{M}$, we denote with $\overline{\mathcal{M}}$ the completed σ-field.

We consider terms I and II separately by starting with term I:

$$
\begin{aligned}
I &= \Omega_0 \frac{K^*K}{n}\Big[(\alpha_n I + \Big(\frac{1}{n}\frac{K^*\Sigma K}{n}\Big) + \frac{K^*K}{n}\Omega_0\frac{K^*K}{n})^{-1} \\
&\quad -(\alpha_n I + \frac{K^*K}{n}\Omega_0\frac{K^*K}{n})^{-1}\Big]\frac{K^*U}{n} \\
&\quad +\Omega_0\frac{K^*K}{n}(\alpha_n I + \frac{K^*K}{n}\Omega_0\frac{K^*K}{n})^{-1}\frac{K^*U}{n} \\
||I||^2 &\leq ||\Omega_0^{\frac{1}{2}}||^2\Big(||\Omega_0^{\frac{1}{2}}\frac{K^*K}{n}(\alpha_n I + \frac{K^*K}{n}\Omega_0\frac{K^*K}{n})^{-1}||^2||\frac{K^*\Sigma_N K}{n^2}||^2 \\
&\quad ||(\alpha_n I + \frac{K^*Var(Y)K}{n^2})^{-1}||^2||\frac{K^*U}{n}||^2 + \\
&\quad ||\Omega_0^{\frac{1}{2}}\frac{K^*K}{n}(\alpha_n I + \frac{K^*K}{n}\Omega_0\frac{K^*K}{n})^{-1}||^2||\frac{K^*U}{n}||^2\Big)
\end{aligned}
$$

From the distribution of U_i we can infer $K_i^*U_i \sim i\mathcal{GP}(0, K_i^*\Sigma K_i)$ and we can write

$$
\begin{aligned}
||\frac{1}{n}K^*U||^2 &= ||\frac{1}{n}\sum_{i=1}^{n}K_i^*U_i||^2 \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}||K_i^*U_i||^2
\end{aligned}
$$

that by Kolmogorov theorem is bounded in probability if $\mathbb{E}||\frac{1}{n}K^*U||^2 < \infty$. Therefore, asymptotic behavior of $||\frac{1}{n}K^*U||^2$ will be determined by asymptotic behavior of $\mathbb{E}||\frac{1}{n}K^*U||^2$. Let $(\tilde{\lambda}_{ij}, \tilde{\varphi}_{ij})_j$ be the eigensystem of the self-adjoint compact operator $\frac{K_i^*\Sigma K_i}{n}$, then, since $\left|\left|\frac{K_i^*U_i}{n}\right|\right|^2 = \sum_{j=1}^{\infty} < \frac{K_i^*U_i}{n}, \tilde{\varphi}_{ij} >^2$, we have

$$
\begin{aligned}
\mathbb{E}||\frac{1}{n}K^*U||^2 &\leq \mathbb{E}\sum_{i=1}^{\infty}\left|\left|\frac{K_i^*U_i}{n}\right|\right|^2 \\
&\leq \sum_{i=1}^{\infty}\mathbb{E}\left|\left|\frac{K_i^*U_i}{n}\right|\right|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\tilde{\lambda}_{ij} \\
&\leq \frac{1}{n}tr\Big(\frac{K^*\Sigma K}{n}\Big)
\end{aligned}
$$

that goes to zero with n. Moreover, we assume that $\frac{1}{n^2}K^*Var(Y)K \to (\frac{1}{n}Q_1 + Q_2\Omega_0 Q_2)$ with n, with Q_1 and Q_2 are bounded operators. It follows that $||(\alpha_n I + \frac{1}{n^2}K^*Var(Y)K)^{-1}||^2 = \mathcal{O}_p(\alpha_n^{-2})$ and $||\Omega_0^{\frac{1}{2}}\frac{K^*K}{n}(\alpha_n I + \frac{K^*K}{n}\Omega_0\frac{K^*K}{n})^{-1}||^2 = \mathcal{O}_p(\alpha_n^{-1})$. Therefore, we get

$$
I = \mathcal{O}_p\Big(\frac{1}{\alpha_n^3}||\frac{Q_1}{n}||^2\frac{1}{n}tr(Q_1) + \frac{1}{\alpha_n}tr(\frac{Q_1}{n})\Big)
$$

where we have substituted $\frac{K^*\Sigma K}{n}$ with its limit. Then, the trace of this operator converges to zero.

To analyze term II, it is advisable to rewrite it in the following way by using the notation $T_n = \frac{K^*K}{n}$:

$$
\begin{aligned}
II &= ||(I - \Omega_0 T_n(\alpha_n I + T_n\Omega_0 T_n)^{-1}T_n)(x_* - x_0) \\
&\quad -\Omega_0 T_n(\alpha_n I + \frac{1}{n^2}K^* Var(\hat{Y})K)^{-1}T_n(x_* - x_0) \\
&\quad +\Omega_0 T_n(\alpha_n I + T_n\Omega_0 T_n)^{-1}T_n(x_* - x_0)||^2 \\
&\leq \overbrace{||(I - \Omega_0 T_n(\alpha_n I + T_n\Omega_0 T_n)^{-1}T_n)(x_* - x_0)||^2}^{III} \\
&\quad + \underbrace{||\Omega_0 T_n(\alpha_n I + \frac{K^* Var(Y)K}{n^2})^{-1}\frac{1}{n}\Big(\frac{K^*\Sigma K}{n}\Big)(\alpha_n I + T\Omega_0 T)^{-1}T(x_* - x_0)||^2}_{IV}.
\end{aligned}
$$

For term IV we proceed exactly as for term II in the proof of Theorem 3 in [**?**] with the operators opportunely replaced. Thus, we get $IV = \mathcal{O}_p(||\frac{Q_1}{n}||^2\frac{1}{\alpha_n^4}\alpha_n^{(\beta+1)\wedge 2})$ by using Assumption 24, 8 and the regularity condition $\delta_* \in \mathcal{R}(\Omega_0^{\frac{1}{2}}T_n^2\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$.

Lastly, we carry out asymptotic analysis of term III, we rewrite it as

$$
\begin{aligned}
III &= ||(\Omega_0^{\frac{1}{2}} - \Omega_0 T_n(\alpha_n I + T_n\Omega_0 T_n)^{-1}T_n\Omega_0^{\frac{1}{2}})\delta_*||^2 \\
&\leq ||\Omega_0^{\frac{1}{2}}||^2||(I - \Omega_0^{\frac{1}{2}}T_n(\alpha_n I + T_n\Omega_0 T_n)^{-1}T_n\Omega_0^{\frac{1}{2}})\delta_*||^2.
\end{aligned}
$$

The second norm is of the same order as the bias of regularization of the solution of $T_n\Omega_0^{\frac{1}{2}}\delta_* = r$:

$$
||(\alpha_n I + \Omega_0^{\frac{1}{2}}T_n^2\Omega_0^{\frac{1}{2}})^{-1}\Omega_0^{\frac{1}{2}}T_n^2\Omega_0^{\frac{1}{2}}||^2,
$$

then it goes to zero. Under the regularity condition $\delta_* \in \mathcal{R}(\Omega_0^{\frac{1}{2}}T_n^2\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, the bias of regularization is at most of order $\mathcal{O}_p(\alpha_n^\beta)$. By summing the rate of convergence found and neglecting the negligible rate in term I we get the result of the theorem.

Proof of Theorem 8

We outline only the principal steps of this proof since it is substantially similar to Proof 2.8 of Theorem 7 and the proof of Theorem 4 in [**?**]. The basic decomposition that is employed is :

$$
\begin{aligned}
V_{D,\alpha}\varphi &= \overbrace{(\Omega_0 - \Omega_0 T_n(\alpha_n I + T_n\Omega_0 T_n)^{-1}T_n\Omega_0)\varphi}^{A} \\
&\quad \underbrace{\Omega_0 T_n(\alpha_n I + \frac{1}{n^2}K^* Var(Y)K)^{-1}\frac{1}{n}\Big(\frac{1}{n}K^*\Sigma K\Big)(\alpha_n I + T_n\Omega_0 T_n)^{-1}T_n\Omega_0\varphi}_{B}.
\end{aligned}
$$

if $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{1}{2}}T_n^2\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, the norm $||A||^2$ is equivalent to $||I||^2$ and $||B||^2 \sim ||IV||^2$ in Proof 2.8 of Theorem 7.

Thus $||A||^2 \sim \mathcal{O}_p(\alpha_n^\beta)$ and $||B||^2 \sim \mathcal{O}_p(||\frac{Q_1}{n}||^2\frac{1}{\alpha_n^4}\alpha_n^{(\beta+1)\wedge 2})$.

2.9 Appendix B: Examples

Our estimator can be applied to all the classical examples of linear inverse problems, for instance *digital image analysis*, see [11], *tomography*, *cancer therapy*, *time resolved fluorescence* problem. Statistics and econometrics offers several examples of applications, see [78] and [10], and we develop in this section some examples in these fields.

Example 1: Density estimation

We propose a new approach for density estimation that is substantially different from the other Bayesian methods existing in the literature like [47], [21], [28], [66], [24] and [52].

Let $\mathcal{X} = L_\pi^2(\mathbb{R})$ and $\mathcal{Y} = L_\rho^2(\mathbb{R})$, with π and ρ two measures on $\mathbb{R}$ different than the Lebeasgue measure. We consider a real-valued random variable ξ with *c.d.f.* F, $F(\bar{\xi}) = \mathbb{P}(\xi \leq \bar{\xi})$, admitting a density $f(\xi) \in \mathcal{X}$ that is characterized as the solution of an inverse problem.

If an *i.i.d.* sample $\xi_1, \ldots, \xi_n$ from F is available we estimate F by $\hat{F}_n(\bar{\xi}) = \frac{1}{n} \sum_{i=1}^n 1\{\xi_i \leq \bar{\xi}\}$ and the probability density function is obtained by solving

$$\hat{F}_n(\bar{\xi}) = \int_{-\infty}^{\bar{\xi}} f(u)du + U_n,$$

with $K : L_\pi^2(\mathbb{R}) \to L_\rho^2(\mathbb{R})$ the integral operator with kernel $1\{u \leq \bar{\xi}\}\frac{1}{\pi(u)}$ and U_n the estimation error. The adjoint of K, $K^* : L_\rho^2(\mathbb{R}) \to L_\pi^2(\mathbb{R})$, has kernel $\frac{1\{\bar{\xi} \geq u\}}{\rho(\xi)}$. If $1\{u \leq \bar{\xi}\}\frac{1}{\pi(u)}$ is square integrable with respect to the product of measures $\pi(u)\rho(\xi)$, K is an HS operator and then it is compact.

The sampling probability P^f is inferred from asymptotic properties of the empirical distribution function, so that it is asymptotically a Gaussian measure with mean F and covariance operator $\Sigma_n = \frac{1}{n} \int_{\mathbb{R}} F(t_j \wedge t_l) - F(t_j)F(t_l)dt_j$.

Example 2: Regression estimation

Let (ξ, w) be a $\mathbb{R}^{1+p}$-valued random vector with *cdf* F and $L_F^2(w)$ be the space of square integrable functions of w, integrable with respect to F. We define the regression function of ξ given w as a function $m(w) \in L_F^2(w)$ such that $\xi = m(w) + \varepsilon$, $\mathbb{E}(\varepsilon|w) = 0$ and $\mathbb{E}(\varepsilon^2|w) = \sigma^2$. Then, $m(w) = \mathbb{E}(\xi|w)$.

Let $g(w,t) : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ be a known function defining an HS integral operator with respect to w, then $\mathbb{E}(g(w,t)\xi) = \mathbb{E}(g(w,t)m(w))$, with the expectation taken with respect to F, and $m(w)$ is the solution to a linear inverse problem. The fact that K is HS ensures that $Km \in L_\pi^2(\mathbb{R})$, with π a measure on $\mathbb{R}$; moreover, the fact that ξ has finite second moment ensures that $\mathbb{E}(g(w,t)\xi) \in L_\pi^2(\mathbb{R})$. We suppose $F(\xi|w)$ is unknown while $F(\cdot, w)$ is known; this implies that $\mathbb{E}(g(w,t)\xi)$ must be estimated but the operator $K = \int g(w,t)dF(\cdot, w)$ is known. If we dispose of a random sample (ξ_i, w_i) we get the consistent estimator

$$\hat{\mathbb{E}}(g(w,t)\xi) := \frac{1}{n} \sum_{i=1}^n g(w_i, t)\xi_i.$$

The statistical inverse problem with estimated LHS becomes

$$\hat{\mathbb{E}}(g(w,t)\xi) = Km(t) + U_n(t).$$

The empirical process $\sqrt{n}(\hat{\mathbb{E}}(g(w,t)\xi) - \mathbb{E}(g(w,t)\xi))$ weakly converges toward a zero mean gaussian process with covariance operator

$$\Lambda = \int_{\mathbb{R}} (\sigma^2 \int_{\mathbb{R}^p} g(w,t)g(w,s)f(w)dw - \mathbb{E}(g(w,t)\xi)\mathbb{E}(g(w,s)\xi))\pi(s)ds.$$

So, the sampling measure P^m is approximately gaussian with mean $\mathbb{E}(g(w,t)\xi)$ and variance $\frac{1}{n}\Lambda$. In most of the cases the *cdf* F is completely unknown and also operator K must be estimated. However, under some regularity assumption, this does not affect the speed of convergence of our estimator to the true solution, see [32].

Alternative approaches existing in Bayesian literature can be found in [43] or [74].

Example 3: Hazard rate function estimation with Right-Censored Survival data

Let $X_1, \ldots, X_n$ be i.i.d. survival times with absolutely continuous distribution function, characterized by the *cdf* F, hazard rate function $h = \frac{F'}{1-F}$ and integrated hazard function $A(t) = \int_0^t h(u)du$. We consider a sequence of survival times $X_{1n}, X_{2n}, \ldots, X_{nn}$. In reality we do not observe $X_{1n}, \ldots, X_{nn}$ but only the right-censored sample $(\tilde{X}_{in}, D_{in})$, $i = 1, \ldots, n$, where $\tilde{X}_{in} = X_{in} \wedge U_{in}$ and $D_{in} = 1(\tilde{X}_{in} = X_{in})$ for some sequence of censoring times $U_{1n}, \ldots, U_{nn}$ from a distribution function G_{in}. We suppose that the survival times $X_{1n}, \ldots, X_{nn}$ and the censoring times $U_{1n}, \ldots, U_{nn}$ are mutually independent for each n.

The aim is to get an estimate of the hazard rate function h, given an estimate of $A(t)$, by solving the functional equation

$$\hat{A}_n(t) = \int_0^t h(u)du + U_n(t)$$

where $U_n(t)$ is introduced to account for the estimation error. We propose to estimate $A(t)$ with the *Nelson-Aalen estimator*, see [2] and from asymptotic properties of this estimator we can infer an approximate sampling distribution.

This inference method is really new with respect to previous bayesian literature, see [46], [25], [80], [18], [48],[69].

Example 4: Deconvolution.

Let (X, Y, Z) be a random vector in $\mathbb{R}^3$ such that $Y = X + Z$, X be independent of Z and $\varphi(\cdot)$, $f(\cdot)$, $g(\cdot)$ be the marginal density functions of X, Y and Z respectively. The density $f(y)$ is defined to be the convolution of $\varphi(\cdot)$ and $g(\cdot)$

$$f(y) = \varphi * g := \int \varphi(x)g(y - x)dx.$$

We assume that $\varphi(\cdot)$, $f(\cdot)$, $g(\cdot)$ are elements of $L^2_\pi(\mathbb{R})$ where π is a symmetric measure assigning a weight decreasing to zero to points far from the median. We suppose $g(\cdot)$ is known, x is not observable, $f(y)$ is estimated nonparametrically and our interest is to recover the density $\varphi(x)$. The corresponding statistical model is

$$\hat{f}(y) = K\varphi(y) + U,$$

where $K = \int g(y - x)dx$ is known and U is the estimation error. Distribution of process U should be inferred from asymptotic properties of the nonparametric estimator $\hat{f}(y)$. This is not possible for a nonparametric estimation since a nonparametric estimator defines an empirical process with trajectories that are discontinuous and independent at each point.

To solve this problem, we propose to transform the model. Let A be a known operator with the property of smoothing the nonparametric estimate. For instance, it could be an integral operator $A = \int a(y, t)dy$, between Hilbert spaces. The *transformed deconvolution model* becomes:

$$\mathbb{E}_y(a(y, t))(t) = AK\varphi(t),$$

where $\mathbb{E}_y$ denotes the expectation taken with respect to $f(y)$. We substitute $f(y)$ with a kernel estimator and we get the error term V defined as $V = \int a(y, t)\hat{f}(y)dy - AK\varphi$. $\sqrt{n}V$ weakly converges toward a gaussian process with zero mean and covariance operator with kernel $\mathbb{E}(a(y_i, t) - \mathbb{E}(a(y, t)))(a(y_i, \tau) - \mathbb{E}(a(y, \tau)))$, from which we infer the sampling distribution.

Example 5: Instrumental Regression Model.

Let (Y, Z, W) be a random vector in $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ with *cdf* F. Let L^2_F be the space of square integrable functions of (Y, Z, W) and $L^2_F(Z) \subseteq L^2_F$ be the space of square integrable functions depending on Z. The instrumental regression $\varphi(Z) \in L^2_F(Z)$ is defined by

$$Y = \varphi(Z) + \varepsilon, \qquad \mathbb{E}(U|W) = 0, \qquad Var(\varepsilon) = \sigma^2. \tag{2.23}$$

$\varphi(Z)$ is the parameter of interest and is solution of an integral equation of first kind: $\mathbb{E}(Y|W) = \mathbb{E}(\varphi(Z)|W)$. If we want to stay completely nonparametric, the estimator of the LHS gives an empirical process with discontinuous trajectories. We have the same kind of problem as in deconvolution to determine the (asymptotic) distribution of the estimation error. Hence, we need to transform the model by re-projecting it on $L_F^2(Z)$. The instrumental regression is now characterized as the solution of

$$\mathbb{E}(\mathbb{E}(Y|W)|Z) = K\varphi, \qquad K = \mathbb{E}(\mathbb{E}(\cdot|W)|Z).$$

By substituting the LHS with a nonparametric estimator, we get a model like (2.1)

$$\hat{\mathbb{E}}(\hat{\mathbb{E}}(Y|W)|Z) = K\varphi + U.$$

The (approximated) distribution of U is gaussian with zero mean and covariance operator $\frac{1}{n}\sigma^2 K^* K$, where K^* denotes the adjoint of K, see [29].

2.10 Appendix C: Monte Carlo Simulations

In all these simulations we take the regularized posterior mean as punctual estimator for the solution of inverse problem (2.1).

Functional equation with a parabola as solution

We take $\mathcal{X} = L_\pi^2$ and $\mathcal{Y} = L_\rho^2$, with π and ρ two measures taken to be uniform on $[0,1]$. The data generating process is

$$\begin{aligned}
\hat{Y} &= \int_0^1 x(s)(s \wedge t)ds + U, \qquad x_* = -3s^2 + 3s \tag{2.24}\\
U &\sim \mathcal{GP}(0, \Sigma_n), \qquad \Sigma_n = n^{-1}\int_0^1 \exp\{-(s-t)^2\}ds \\
x &\sim \mathcal{GP}(x_0, \Omega_0), \qquad x_0 = -2.8s^2 + 2.8s \\
\Omega_0\varphi(t) &= \omega_0 \int_0^1 \exp(-(s-t)^2)\varphi(s)ds.
\end{aligned}$$

The covariance operators have eigenvalues of order $\mathcal{O}(e^{-j})$, the regularization parameter α has been set to $2.e-03$, $n = 1000$ and the discretization step is 0.01.

We show in Figure 2.1a the true function x_* (continuous line) and the regularized posterior mean estimation (dotted line) for the prior given above with $\omega_0 = 2$. We propose, in Figure 2.1b a comparison between our estimator and the estimator obtained by solving equation (2.1) with a classical Tikhonov regularization method (small dotted line) (with $\alpha = 2.e-04$).

To analyze the role of the prior distribution we have performed the simulation for different priors, see Figures 2.1c and 2.1d. It should be noted that the far the prior mean is from the true parameter the bigger should be the prior covariance operator.

Finally, in Figure 2.1 results of a Monte Carlo experiment with 100 iterations are shown. Panels (2.1e), (2.1g) and (2.1h) are Monte Carlo experiment conducted for the three different priors distribution considered. The dotted line represents the mean of the regularized posterior means obtained for each iteration. Panel (2.1f) shows the Monte Carlo mean of the regularized posterior means for the first specification of the prior distribution (dotted line) and of the classical Tikhonov solutions (small dotted line).

(a) $x_0 = -2.8s^2 + 2.8s$,
$\Omega_0\varphi(t) = 2\int_0^1 \exp(-(s-t)^2)\varphi(s)ds$

(b) comparison between the performance of our estimator and the classical Tikhonov regularized solution.

(c) $x_0 = -2s^2 + 2s$,
$\Omega_0\varphi(t) = 40\int_0^1 ((s \wedge t) - st)\varphi(s)ds$

(d) $x_0 = -2.22s^2 + 2.67s - 0.05$,
$\Omega_0 = 100\int_0^1 (0.9(s-t)^2 - 1.9|s-t| + 1)ds$

(e) $x_0 = -2.8s^2 + 2.8s$,
$\Omega_0\varphi(t) = 2\int_0^1 \exp(-(s-t)^2)\varphi(s)ds$

(f)

(g) $x_0 = -2s^2 + 2s$,
$\Omega_0\varphi(t) = 40\int_0^1 ((s \wedge t) - st)\varphi(s)ds$

(h) $x_0 = -2.22s^2 + 2.67s - 0.05$,
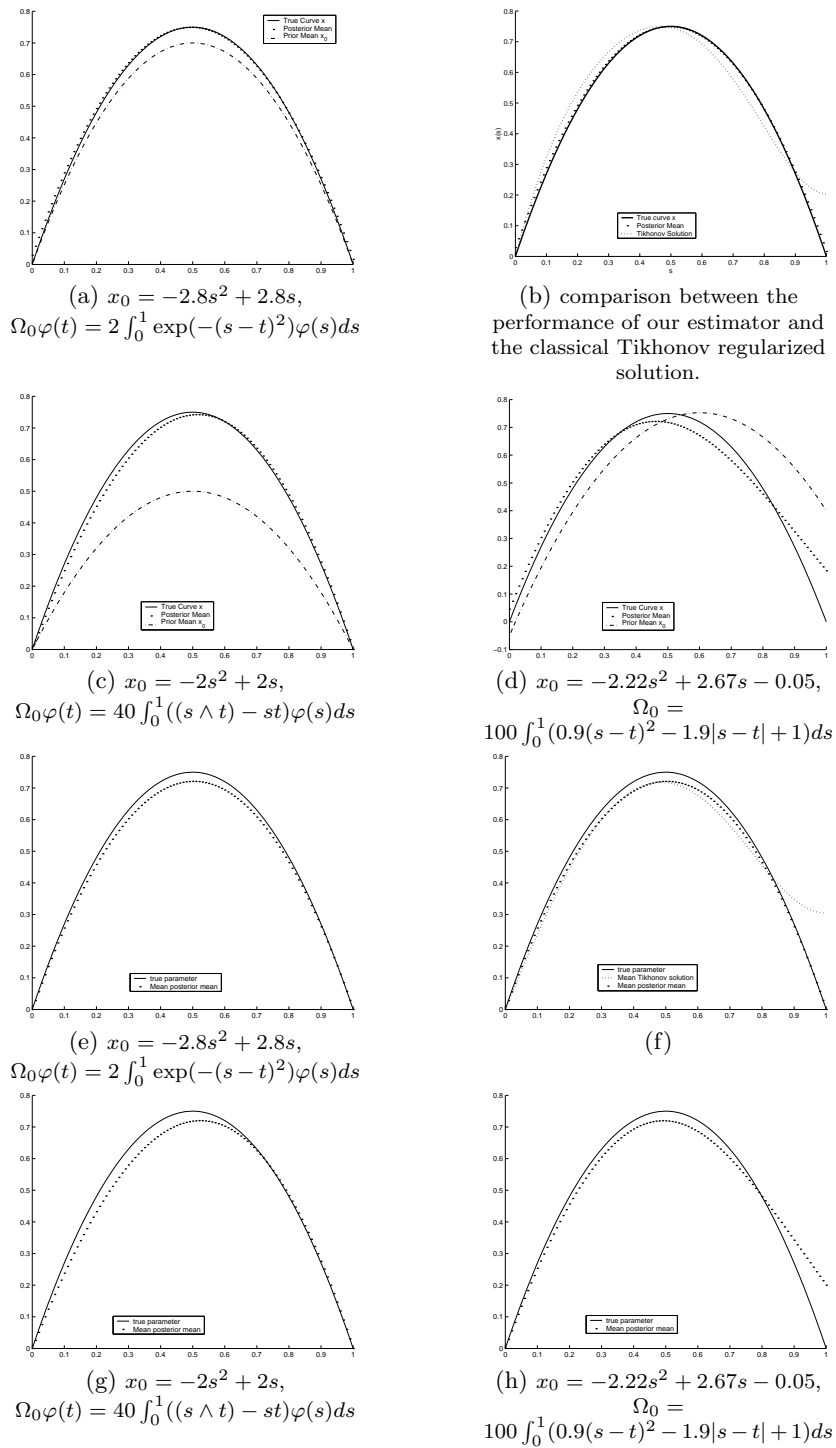$\Omega_0 = 100\int_0^1 (0.9(s-t)^2 - 1.9|s-t| + 1)ds$

Figure 2.1: Figures (2.1a) - (2.1d) represents simulations with only one trial. Figures (2.1e) - (2.1h) represent the Monte Carlo experiment.

Density Estimation

This is a simulation of example 2.9 and the notation will be the same. The true density f_* is the density of a standard gaussian measure on $\mathbb{R}$ and the measures π and ρ, defining the L^2 spaces, are uniform measure on $[-3, 3]$. We use the sample $\xi_1, \ldots, \xi_n$ to estimate F and the sampling variance Σ_n. The operator K is known.

The prior mean is $f_0 = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(\xi - \theta)^2\}$ and the prior variance is $\Omega_0 \varphi(t) = \omega_0 \int_{-3}^{3} \exp(-(s - t)^2)\varphi(s)\frac{1}{6}ds$.

Parameters $(\sigma, \theta, \omega_0)$ have been differently set to see the effect of prior changes on the estimated solution. The regularization parameter α_n has been set equal to 0.05 and the sample size is of $n = 1000$. Figures (2.2a)- (2.2d) shows the regularized posterior mean estimator for different specification of the parameters. In panels (a) and (c) the true density (continuous line), the prior mean (dotted line) and the regularized posterior mean estimator (dashed-dotted line) are drawn; panels (b) and (d) show the comparison between our estimator and the classical Tikhonov solution (dotted line).

Figures 2.2e and 2.2f represent a sample of curves dawn from the prior distribution together with the prior mean (continuous line) and the true density (dotted line). Lastly, in Figures 2.2g and 2.2h, the results of a Monte Carlo experiment are shown. The dashed-dotted line is the mean of the regularized posterior means obtained in each replication, the dashed line is the mean of Tikhonov solutions for each Monte Carlo iteration and the solid line is thetrue density function.

Regression Estimation

This is a simulation of example 2.9; the notation is the same. We consider $w \in \mathbb{R} \sim F = \mathcal{N}(2, 1)$ and a Gaussian white noise $\varepsilon \sim \mathcal{N}(0, 2)$ independently drawn. Function $g(w, t)$ has been alternatively specified as an exponential function, $g(w, t) = exp(-(w - t)^2)$, or as an indicator function, $g(w, t) = 1\{w \leq t\}$, but we only report here the results for the second specification. $g(w, t)$ define an HS operator $K : L_F^2(w) \to L_\pi^2$, with $\pi \sim \mathcal{N}(2, 1)$.

The true regression function is $m_*(w) = cos(w)sin(w)$ and the prior distribution is Gaussian: $m(w) \sim \mathcal{GP}(m_0(w), \Omega_0 \varphi(w))$, with $\Omega_0 \varphi(w_1) = \omega_0 \int \exp(-(w_1 - w_2)^2)\varphi(w_2)f(w_2)dw_2, \forall \varphi \in L_F^2(w)$ and $\omega_0 = 2$ or $\omega_0 = 10$. We have considered three different prior mean specifications: $m_0(w) = m_*(w)$, $m_0(w) = 0.067w - 0.2$, or $m_0 = 0$. After having drawn a sample of (ξ, w) we estimate $\mathbb{E}(g(w, t)\xi)$ for any t by using the sample mean. The regularization parameter α is set equal to 0.05, the sample size is $n = 1000$ for a single estimation and $n = 500$ for Monte Carlo simulations. In Monte Carlo Simulation we have done 50 replications.

Figure 2.3 shows the results. Panels (a), (c) and (e) shows the estimation for only one replication, Panels (b), (d) and (f) shows the estimation for each Monte Carlo replication and the mean over all the replications (dashed-dotted line).
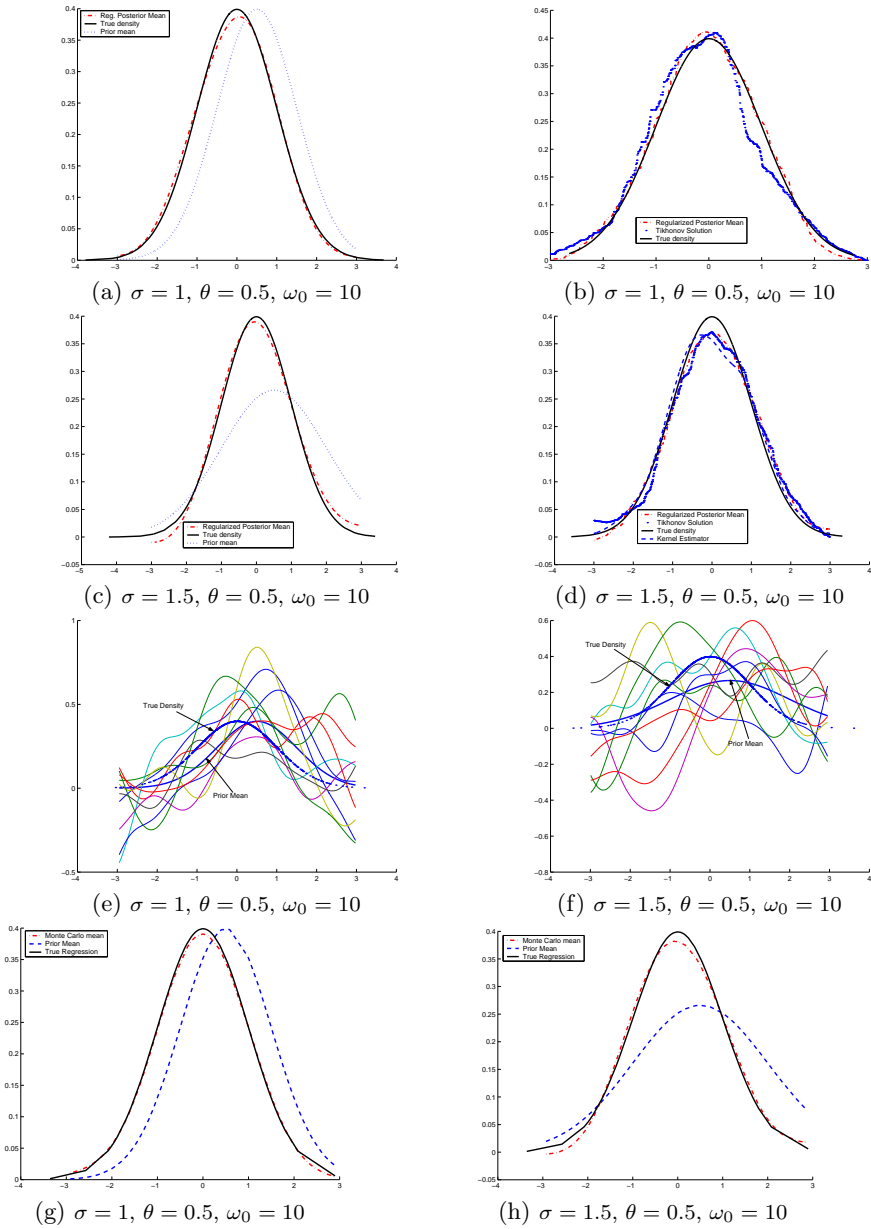
Figure 2.2: Panels (2.2a)- (2.2d): regularized posterior mean and Tikhonov estimators. Panels (2.2e) - (2.2f): Drawn from the prior distribution. Panels: (2.2g) - (2.2h): Monte Carlo simulation.

(a) $m_0(w) = cos(w)sin(w)$, $\Omega_0\varphi(w_1) = 2\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$

(b) $m_0(w) = cos(w)sin(w)$, $\Omega_0\varphi(w_1) = 2\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$

(c) $m_0(w) = 0.067w - 0.2$, $\Omega_0\varphi(w_1) = 10\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$

(d) $m_0(w) = 0.067w - 0.2$, $\Omega_0\varphi(w_1) = 10\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$

(e) $m_0(w) = 0$, $\Omega_0\varphi(w_1) = 10\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$

(f) $m_0(w) = 0$, $\Omega_0\varphi(w_1) = 10\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$

Figure 2.3: Panels (2.3a), (2.3c) and (2.3e): estimation for different prior means. Panels (2.3b), (2.3d) and (2.3f): Monte Carlo Experiment with $N = 100$, $\alpha = 0.05$, 50 iterations.

Chapter 3

On the Regularization Power of the Prior Distribution in Linear ill-Posed Inverse Problems[1]

joint with Jean-Pierre Florens.

Abstract

We consider models described by a functional equation in an Hilbert space of the type $\hat{Y} = Kx + U$. We wish to recover the functional parameter of interest x after having observed $\hat{Y}$. This problem is ill-posed because the operator K is assumed to be compact so that its inverse is not continuous on the whole space of reference and the estimator of x is in general non consistent.

We specify a prior distribution on x of the g-prior type and we detect a class of models for which the prior distribution on x is able to correc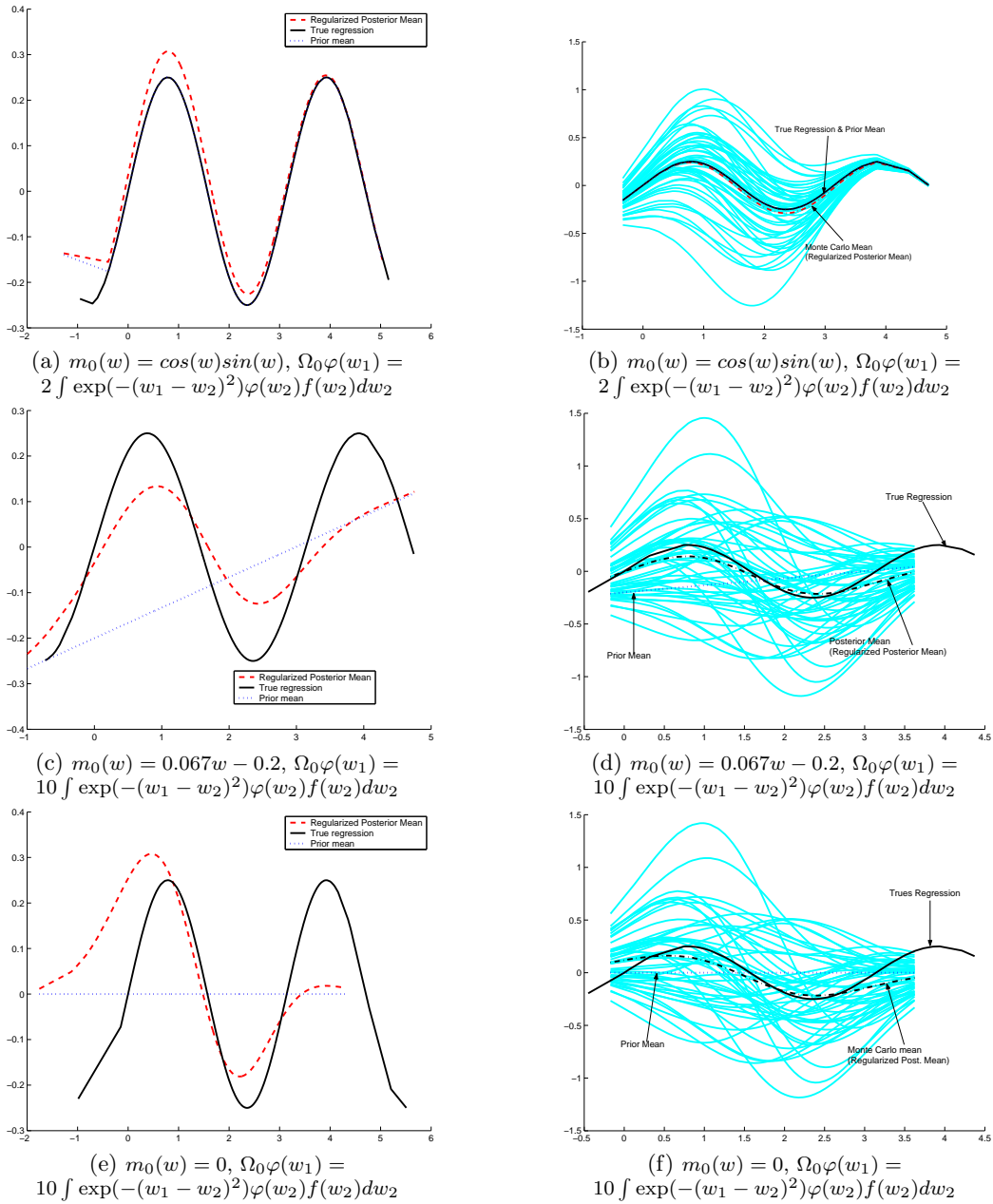t for the ill-posedness also in infinite dimensional problems. The prior distribution depends on the regularization parameter and on the degree of penalization. We prove that, under some conditions, the posterior distribution is consistent in the sampling sense. In particular, the prior-to-posterior transformation can be interpreted as a Tikhonov regularization in the Hilbert scale induced by the prior covariance operator.

Finally, the regularization parameter is treated as an hyperparameter and we propose how to exploit its posterior distribution for optimally selecting it.

3.1 Introduction

Let consider the solution to the noisy functional equation

$$\hat{Y} = Kx + U, \qquad x \in \mathcal{X}, \hat{Y} \in \mathcal{Y} \tag{3.1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are infinite dimensional separable Hilbert spaces over $\mathbb{R}$ supposed to be Polish with inner product $< \cdot, \cdot >$ and norm $|| \cdot ||$. U is a measurement error. $K : \mathcal{X} \to \mathcal{Y}$ is a known Hilbert-Schmidt (HS, hereafter), then compact, linear operator with infinite dimensional range. K^* will denote the adjoint of K, *i.e.* K^* is such that $< K\varphi, \psi > = < \varphi, K^*\psi >$, $\forall \varphi \in \mathcal{X}$ and $\psi \in \mathcal{Y}$. Compactness of operator K and the infinite dimension of the range of K make the inverse K^{-1} not continuous on the whole

---

[1]This chapter is adapted from: Florens, J.P., and A., Simoni (2008), *On the Regularization Power of the Prior Distribution in Linear ill-Posed Inverse Problem*, mimeo.

$\mathcal{Y}$ so that some regularization of this inverse is demanded.

This kind of model is classical in the inverse problem literature and it is encountered in many real applications. Classical techniques of regularization consist in Spectral cut-off regularization, Tikhonov regularization, or Landweber-Fridman regularization, among other, see Kress (1999) [50]. On the other side, Bayesian methodologies propose the posterior distribution of x as solution for 3.1. This posterior distribution is in general non well-defined, in the sense that it is not consistent in a frequentist sense. In the previous Chapter we have proposed to regularize this distribution and we have defined a new object called *Regularized Posterior distribution* that plays the role of the posterior distribution. Lehtinen *et al.* (1989) [53] and Mandelbaum (1984) [60] propose to regularize through a restriction of the space of definition of $\hat{Y}$.

In this chapter we consider a class of models where the regularization is automatically performed by the prior-to-posterior transformation, so that the posterior distribution that we obtain is well-defined and no ad-hoc regularization need to be introduced. In particular, the prior distribution depends on the regularization parameter and the degree of penalization, chosen for measuring the variability of the solution (as, for instance, the higher order of derivatives in a Sobolev penalization).

We assume that U induces a gaussian process ($\mathcal{GP}$ in the following) on $\mathcal{Y}$. Consequently, the sampling distribution of $\hat{Y}$ is gaussian:

$$\hat{Y}|x \sim \mathcal{GP}(Kx, \delta\Sigma) \tag{3.2}$$

with $\delta = \delta(n)$ a function of the sample size n such that $\delta \to 0$ as $n \to \infty$. The covariance operator $\Sigma : \mathcal{Y} \to \mathcal{Y}$ is assumed to be a fixed and given operator. It follows that it is linear, bounded, nonnegative, self-adjoint, compact and trace-class. Let $\mathcal{R}(\cdot)$ denote the range of an operator and $\mathcal{D}(\cdot)$ its domain. We make the following assumption:

Assumption 9

(a) $\mathcal{R}(K) \subset \mathcal{D}(\Sigma^{-\frac{1}{2}})$;

(b) there exists an unbounded densely defined operator L that is self-adjoint and positive such that $||L^{-a}x|| \sim ||\Sigma^{-\frac{1}{2}}Kx||$.

Part *(a)* of Assumption 9 ensures that operator $\Sigma^{-\frac{1}{2}}K$ is well-defined and it is equivalent to say that we are demanding a compatibility between the sampling covariance operator Σ and the operator K in the sampling mechanism. This is very common in practical examples, like estimation of a density, a regression or an instrumental variable regression, where the covariance operator is of the form $\Sigma = (KK^*)^r$, for some $r \geq 1$. We develop this particular case in Section 3.3.

For all $s \in \mathbb{R}$, operator L in Assumption 9 *(b)* induces the Hilbert scale $(\mathcal{X}_s)_{s \in \mathbb{R}}$, where $\mathcal{X}_s$ is an Hilbert space defined as the completion of $\bigcap_{s \in \mathbb{R}} \mathcal{D}(L^s)$ with respect to the norm $||x||_s := ||L^s x||$. Parameter a is the degree of ill-posedness in the bayesian experiment. It is usually different than the degree of ill-posedness in the classical problem $\hat{Y} = Kx$.

We assume that the functional parameter of interest x is characterized by the following gaussian distribution:

$$x|g, s \sim \mathcal{GP}\left(x_0, \frac{1}{g}L^{-2s}\right), \tag{3.3}$$

with $g = g(n)$ a function of n such that $g \to \infty$ with n. The two conditioning parameters g and s are for the moment treated as fixed. In Section 3.4 we partially relax this assumption and treat g as an hyperparameter. The operator L^{-2s} plays the role of the prior covariance

operator, then, following notation in Chapter 2, $\Omega_0 = L^{-2s}$, where $\Omega_0 : \mathcal{X} \to \mathcal{X}$ is a linear operator that is bounded, nonnegative, self-adjoint, compact and trace-class. This choice of the prior covariance is aimed to link the prior distribution with the operator K and the sampling model. Such a link is evident from Assumption 9 *(b)* and it is a natural idea in linear regression models, see for instance Zellner's g-prior (1986) [81]. Our prior is an extension of the Zellner'g-prior.

The predictive distribution, obtained by integrating out x, is $\hat{Y}|g,s \sim \mathcal{GP}(Kx_0, (\delta\Sigma + \frac{1}{g}K\Omega_0 K^*))$.

From a frequentist point of view, there exists a true value of the parameter of interest x having generated the data $\hat{Y}$. We denote this value with x_* and it will be used in the asymptotic analysis since we care for the weak convergence of the posterior distribution of x towards a point mass in x_* as $n \to \infty$. This type of convergence has been widely discussed in Section 3.2 of Chapter 2; it is a convergence with respect to the sampling probability and it is known as *posterior consistency*. We introduce a regularity assumption about the centered true value of the parameter of interest.

Assumption 10 *For some $\beta \geq s$, we assume that $(x_* - x_0) \in \mathcal{X}_\beta$, i.e. there exists a $\rho_* \in \mathcal{X}$ such that $(x_* - x_0) = L^{-\beta}\rho_* \ (\equiv \Omega_0^{\frac{\beta}{2s}}\rho_*)$.*

Because $\beta \geq s$, it follows that $\mathcal{R}(\Omega_0^{\frac{\beta}{2s}}) \subset \mathcal{R}(\Omega_0^{\frac{1}{2}})$ and Assumption 10 implies that there exists a δ_* such that $(x_* - x_0) = \Omega_0^{\frac{1}{2}}\delta_*$ and $\delta_* = \Omega_0^{\frac{\beta-s}{2s}}\rho_*$. Moreover, after Proposition 3.6 in Carrasco *et al.* (2007), we can write $\mathcal{R}(\Omega_0^{\frac{1}{2}}) = \mathcal{H}(\Omega_0)$, where $\mathcal{H}(\Omega_0)$ denotes the *Reproducing Kernel Hilbert Space* associated to Ω_0 and embedded in $\mathcal{X}$, *i.e.*

$$\mathcal{H}(\Omega_0) = \left\{ \varphi : \varphi \in \mathcal{X} \quad \text{and} \quad ||\varphi||_{\Omega_0} := \sum_{j=1}^{\infty} \frac{|<\varphi, \varphi_j^{\Omega_0}>|^2}{\lambda_j^{\Omega_0}} < \infty \right\}.$$

Hence, Assumption 10 implies that $(x_* - x_0) \in \mathcal{H}(\Omega_0)$.

Hereafter we use the notation: $\alpha = \delta g$, $B = \Sigma^{-\frac{1}{2}}K\Omega_0^{\frac{1}{2}}$, $T = \Sigma^{-\frac{1}{2}}K$. Operator T is well defined under Assumption 9 *(a)*. A further assumption needs to be introduced in order that the operator B be well-defined.

Assumption 11

(a) $\mathcal{R}(K\Omega_0^{\frac{1}{2}}) \subset \mathcal{D}(\Sigma^{-1})$;

(b) a, β and s are three real parameters satisfying the inequalities $0 < a \leq s \leq \beta \leq 2s+a$;

(c) there exists a $\gamma \in]0,1]$ such that the operator $(B^*B)^\gamma$ is trace class, i.e. if $\{\lambda_j^2\}$ denotes the eigenvalues of B^*B, then $\sum_j \lambda_j^{2\gamma} < \infty$ must be verified.

Under Assumption 11 *(a)*, $\mathcal{R}(K\Omega_0^{\frac{1}{2}}) \subset \mathcal{D}(\Sigma^{-1})$ and, since $\mathcal{D}(\Sigma^{-1}) \subset \mathcal{D}(\Sigma^{-\frac{1}{2}})$, operator B is well-defined.

The last assumption will be exploited for computing the speed of convergence of the posterior distribution. When $\gamma = 1$, Assumption 11 *(c)* is the classical Hilbert-Schmidt assumption of operator $\Sigma^{-\frac{1}{2}}K\Omega_0^{\frac{1}{2}}$. For $\gamma < 1$ this assumption is more demanding. The parameter $\alpha := \delta g$ will be used as the index for the family of posterior distributions, it plays the role of a regularization parameter and it is linked to the error δ in the observations. It must satisfy the two classical properties required for a regularization parameter: $\alpha \to 0$ and $\alpha^2 n \to \infty$ as $n \to \infty$. If $\delta \propto \frac{1}{n}$, this implies that $\frac{g}{n} \sim o_p(1)$ and $\frac{g}{\sqrt{n}} \to \infty$, or

equivalently $\frac{\sqrt{n}}{g} \sim o_p(1)$, *i.e.* g must increase faster than $\sqrt{n}$ but slower than n.

The solution of (3.1) is the posterior distribution of x, denoted with μ^Y. μ^Y is a conditional probability on $\mathcal{X}$ that exists and is gaussian, see Section 2.3 of the previous Chapter. It has mean function $A(\hat{Y} - Kx_0) + x_0$ and covariance operator $\Omega_0 - AK\Omega_0$, where $A : \mathcal{Y} \to \mathcal{X}$ is an operator such that its adjoint is defined as the solution of the functional equation:

$$\left(\delta\Sigma + \frac{1}{g}K\Omega_0 K^*\right)A^*\varphi = \frac{1}{g}K\Omega_0\varphi, \qquad \forall\varphi \in \mathcal{X}. \tag{3.4}$$

Hence,

$$
\begin{aligned}
(\alpha\Sigma + K\Omega_0 K^*)A^* &= K\Omega_0 \\
\Leftrightarrow \quad \Sigma^{\frac{1}{2}}(\alpha I + \Sigma^{-\frac{1}{2}}K\Omega_0 K^*\Sigma^{-\frac{1}{2}})\Sigma^{\frac{1}{2}}A^* &= K\Omega_0 \\
\Leftrightarrow \quad (\alpha I + BB^*)\Sigma^{\frac{1}{2}}A^* &= B\Omega_0^{\frac{1}{2}} \\
\Leftrightarrow \quad \Sigma^{\frac{1}{2}}A^* &= (\alpha I + BB^*)^{-1}B\Omega_0^{\frac{1}{2}} \\
\Leftrightarrow \quad \Sigma^{\frac{1}{2}}A^* &= B(\alpha I + B^*B)^{-1}\Omega_0^{\frac{1}{2}} \\
\Leftrightarrow \quad A^* &= \Sigma^{-\frac{1}{2}}B(\alpha I + B^*B)^{-1}\Omega_0^{\frac{1}{2}}.
\end{aligned}
$$

that is well-defined under Assumption 11 *(i)* since $\mathcal{R}(K\Omega_0^{\frac{1}{2}}) \subset \mathcal{D}(\Sigma^{-1})$. Such assumption concerns the degree of regularity (*i.e.* the differentiability) of the prior covariance operator with respect to the sampling covariance operator. Then,

$$A = \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^* \tag{3.5}$$

that is continuous and defined everywhere. In general, it is not sure that the inverse of operator B^*B exists, since if it is compact its eigenvalues are countable and they accumulate only at zero, then $(B^*B)^{-1}$ explodes. However, this possible problem is solved by the presence of operator αI that translates the eigenvalues sufficiently far from zero, or equivalently extends the range of B^*B to the whole space $\mathcal{Y}$. In other words, when Assumption 11 holds, the prior-to-posterior transformation is equivalent to apply a Tikhonov regularization scheme to the inverse of B^*B, *i.e.* to regularize the solution of the equation $B\varphi = r$, with $\varphi \in \mathcal{Y}$ and $r \in \mathcal{X}$.

Two comments are noteworthy to be pointed out.
1) The construction of the posterior mean can be interpreted as a regularization in the Hilbert scale induced by L^s. Take for simplicity $x_0 = 0$, then

$$
\begin{aligned}
\mathbb{E}(x|\hat{Y}, g, s) &= A\hat{Y} \\
&= L^{-s}(\alpha I + L^{-s}K^*\Sigma^{-1}KL^{-s})^{-1}L^{-s}K^*\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}\hat{Y} \\
&= (\alpha L^{2s} + T^*T)^{-1}T^*\Sigma^{-\frac{1}{2}}\hat{Y}
\end{aligned}
$$

that results to be the regularization, in the prior variance Hilbert Scale, of the solution of the model

$$\Sigma^{-\frac{1}{2}}\hat{Y} = Tx + \Sigma^{-\frac{1}{2}}U.$$

This model is the transformation of (3.1) through operator $\Sigma^{-\frac{1}{2}}$. We remark that there is no reason why the quantities $\Sigma^{-\frac{1}{2}}\hat{Y}$ and $\Sigma^{-\frac{1}{2}}U$ exist, so that this model is *per se* incorrect, but it is useful in order to interpret the prior-to-posterior transformation as an Hilbert Scale regularization.

2) In the specification of the prior distribution we may wish to stay as general as possible by choosing a prior variance of the form $\Omega_0 = \frac{1}{g}QL^{-2s}Q^*$, for some bounded operator Q not necessarily compact. Then, the previous case is a particular case of this one for $Q = I$. Operator A takes the form

$$A = QL^{-s}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^*,$$

for $B = \Sigma^{-\frac{1}{2}}KQL^{-s}$. Hence, L^s is the Hilbert Scale for $\Sigma^{-\frac{1}{2}}KQ$ and Assumption 9 *(a)* is replaced by $\mathcal{R}(KQ) \subset \mathcal{D}(\Sigma^{-\frac{1}{2}})$ that is weaker. Moreover, operator B is well-defined if $\mathcal{R}(KQL^{-s}) \subset \mathcal{D}(\Sigma^{-1})$ that is also less demanding than Assumption 11 *(a)*.
In order to obtain the same order of convergence of the posterior distribution we also have to replace Assumption 10 with the assumption that there exists an element $\tilde{\delta}_* \in \mathcal{R}(L^{-(\beta-s)})$ such that $(x_* - x_0) = QL^{-s}\tilde{\delta}_*$.

3.2 Asymptotic Analysis

The posterior distribution μ^Y, previously defined, can reveal to be useful also for classical statisticians if, as more and more observations are accumulated, it degenerates towards a Dirac measure in x_*. This is the concept of posterior consistency. In other words, if the posterior distribution is consistent with respect to the sampling distribution, then it can be used as an estimator not only by bayesian statisticians but also by classical statisticians. In this section we study convergence in $\mathcal{X}$-norm with respect to the sample distribution as $n \to \infty$. This reduces to study consistency of the posterior mean and convergence to zero of the posterior variance.
In order to prove posterior consistency we make use of Corollary 8.22 in Engl *et al.* (2000) [19]. We give a simplified version of it:

Corollary 1 *Let $\mathcal{X}_s$, $s \in \mathbb{R}$ be a Hilbert scale induced by L and let $\Sigma^{-\frac{1}{2}}K : \mathcal{X} \to \mathcal{Y}$ be a bounded operator satisfying $||L^{-a}x|| \sim ||\Sigma^{-\frac{1}{2}}Kx||$, $\forall x \in \mathcal{X}$ and for some $a > 0$. Then, for $B = \Sigma^{-\frac{1}{2}}KL^{-s}$, $s \geq 0$ and $|\nu| \leq 1$*

$$||(B^*B)^{\frac{\nu}{2}}x|| \sim ||L^{-\nu(a+s)}x||$$

and $\mathcal{R}((B^*B)^{\frac{\nu}{2}}) = \mathcal{X}_{\nu(a+s)} \equiv \mathcal{D}(L^{\nu(a+s)})$.

We refer to [19] for the proof of it.

Let start by analyzing the posterior bias $\mathbb{E}(x|\hat{Y}) - x_*$ that we re-write as

$$\mathbb{E}(x|\hat{Y}) - x_* = \overbrace{-(I - AK)(x_* - x_0)}^{C} + \overbrace{AU}^{D},$$

with A is as defined in (3.5). Let $v \in \mathcal{X}$ be such that $(x_* - x_0) = L^{-\beta}v$, then

$$
\begin{aligned}
||C||^2 &= ||[I - \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^*K]L^{-\beta}v||^2 \\
&= ||\Omega_0^{\frac{1}{2}}[I - (\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^*K\Omega_0^{\frac{1}{2}}]L^{s-\beta}v||^2 \\
&= ||(B^*B)^{\frac{s}{2(a+s)}}[I - (\alpha I + B^*B)^{-1}B^*B](B^*B)^{\frac{\beta-s}{2(a+s)}}\tilde{v}||^2 \\
&= ||\alpha(\alpha I + B^*B)^{-1}B^*B](B^*B)^{\frac{\beta}{2(a+s)}}\tilde{v}||^2 \\
&\sim \mathcal{O}_p(\alpha^{\frac{\beta}{a+s}}).
\end{aligned}
$$

The third equality is obtained by applying Corollary 1 and $\tilde{v}$ is an element of $\mathcal{X}$ such that $L^{s-\beta}v = (B^*B)^{\frac{\beta-s}{2(a+s)}}\tilde{v}$.

Let consider now term D:

$$
\begin{aligned}
||D||^2 &= ||AU||^2 \\
&\leq tr(AVar(U)A^*).
\end{aligned}
$$

The last inequality is obtained by applying Markov inequality: $\mathbb{P}\{U \in \mathcal{Y}; ||AU||^2 \geq \epsilon\} \leq \frac{1}{\epsilon}\mathbb{E}(||AU||^2)$ and $\mathbb{E}(||AU||^2) = Var(AU)$ since U has zero mean. Application of Corollary 1 implies that $\mathcal{R}(\Omega_0^{\frac{1}{2}}) \equiv \mathcal{D}(L^s)$ is equal to $\mathcal{R}(B^*B)^{\frac{s}{2(a+s)}}$ so that $A = (B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^*$ and then

$$
\begin{aligned}
tr(AVar(U)A^*) &= tr((B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^*\delta\Sigma\Sigma^{-\frac{1}{2}}B(\alpha I + B^*B)^{-1}(B^*B)^{\frac{s}{2(a+s)}}) \\
&= \delta tr((B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}B^*B(\alpha I + B^*B)^{-1}(B^*B)^{\frac{s}{2(a+s)}})
\end{aligned}
$$

after simplification. By denoting with $\{\lambda_j^2\}$ the sequence of eigenvalues associated to BB^*, or equivalently to B^*B, we have

$$
\begin{aligned}
tr(AVar(U)A^*) &\leq \delta \sum_j \frac{\lambda_j^{\frac{2s}{a+s}+2}}{(\alpha + \lambda_j^2)^2} \\
&= \delta \sum_j \frac{\lambda_j^{\frac{2s}{a+s}+2-2\gamma}}{(\alpha + \lambda_j^2)^2}\lambda_j^{2\gamma} \\
&\leq \delta \sup_j \frac{\lambda_j^{\frac{2s}{a+s}+2-2\gamma}}{(\alpha + \lambda_j^2)^2} \sum_j \lambda_j^{2\gamma} \\
&\sim \mathcal{O}_p\left(\delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}}\right),
\end{aligned}
$$

where we have exploited Assumption 11 *(c)*.

In choosing α we find the usual trade-off: while $||C||^2$ is increasing in α, $||D||^2$ is decreasing in α. The optimal α, denoted with α^*, is the value for which $||C||^2$ and $||D||^2$ are of the same order:

$$
\begin{aligned}
\alpha^{\frac{\beta}{a+s}} &\sim \delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}} \\
\Leftrightarrow \quad \alpha^* &= c_1\delta^{\frac{a+s}{\beta+a+\gamma(a+s)}}
\end{aligned}
$$

with c_1 some constant. The fastest speed of convergence of the posterior mean, obtained by substituting the optimal α^*, is of order $\delta^{\frac{\beta}{\beta+a+\gamma(a+s)}}$ that is decreasing in $s\gamma$. We have therefore proved the following theorem.

Theorem 10 *Let consider the probability specification in (3.2) and (3.3). Under Assumptions 9, 10 and 11 the posterior mean of x is consistent in the sense that $||\mathbb{E}(x|\hat{Y},g,s)-x_*||^2$ converges to zero with respect to the sampling probability. It is of order*

$$||\mathbb{E}(x|\hat{Y},g,s) - x_*||^2 \sim \mathcal{O}_p\Big(\alpha^{\frac{\beta}{a+s}} + \delta\alpha^{-\frac{\gamma(a+s)+a}{a+s}}\Big).$$

Moreover, if $\alpha = c_1\delta^{\frac{a+s}{\beta+a+\gamma(a+s)}}$, for some constant c_1,

$$\delta^{-\frac{\beta}{\beta+a+\gamma(a+s)}}||\mathbb{E}(x|\hat{Y}) - x_*||^2 \sim \mathcal{O}_p(1).$$

When g is not treated as an hyperparameter [2], it has to be chosen so that $g \to \infty$ holds. This in turn guarantees that the prior distribution degenerates to a point mass in correspondence of the prior mean, but in order this makes sense, it must degenerate at the good rate that, as we have already stressed, must be faster than $\sqrt{n}$ and slower than n.

Once the optimal α has been determined, the corresponding optimal g can be obtained through the relationship $\alpha = \delta g$:

$$
\begin{aligned}
g^* &\propto \alpha^*\delta^{-1} \\
&= c_2\delta^{-\frac{\beta-s+\gamma(a+s)}{\beta+a(1+\gamma)+s\gamma}},
\end{aligned}
$$

with c_2 some constant. The requirement that g must goes to infinity slower than n is satisfied if $-a < s$, that is always true under Assumption 11 *(b)*. In addition, in order to have that g converges to $+\infty$ faster than $\sqrt{n}$ one demands that $\beta > (2s+a) - \gamma(a+s)$, that makes sense under Assumption 11 *(b)* since $2s + a > \beta > (2s+a) - \gamma(a+s)$.

The asymptotic behavior of the posterior variance is similar to that one of term C previously considered:

$$Var(x|\hat{Y},g,s)\phi = [\Omega_0 - \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^*K\Omega_0]\phi,$$

for any $\phi \in \mathcal{X}$. If $\phi \in \mathcal{X}$ is such that $\Omega_0^{\frac{1}{2}}\phi \in \mathcal{R}(\Omega_0^{\frac{\beta-s}{2s}})$, then

$$
\begin{aligned}
||Var(x|\hat{Y},g,s)\phi||^2 &= ||\Omega_0^{\frac{1}{2}}[\Omega_0^{\frac{1}{2}} - (\alpha I + B^*B)^{-1}(\Sigma^{-\frac{1}{2}}B)^*K\Omega_0]\phi||^2 \\
&= ||(B^*B)^{\frac{s}{2(a+s)}}[I - (\alpha I + B^*B)^{-1}B^*\Sigma^{-\frac{1}{2}}K\Omega_0^{\frac{1}{2}}]\Omega_0^{\frac{1}{2}}\phi||^2 \\
&= ||(B^*B)^{\frac{s}{2(a+s)}}[I - (\alpha I + B^*B)^{-1}B^*B](B^*B)^{\frac{\beta-s}{2(a+s)}}\upsilon||^2 \\
&\sim \mathcal{O}_p\Big(\alpha^{\frac{\beta}{a+s}}\Big)
\end{aligned}
$$

where $\upsilon \in \mathcal{X}$ is such that $\Omega_0^{\frac{1}{2}}\phi = (B^*B)^{\frac{\beta-s}{2(a+s)}}\upsilon$. We summarize this result in the following theorem.

---

[2]We shall consider g as an hyperparmater in Section 3.4.

Theorem 11 *Let consider the probability specification in (3.2) and (3.3). Under Assumptions 9 and 11 the posterior variance of x converges to zero in $\mathcal{X}$-norm with respect to the sampling probability: $||Var(x|\hat{Y}, g, s)\phi|| \to 0$, $\forall \phi \in \mathcal{X}$. If $\phi \in \mathcal{X}$ is such that $\Omega_0^{\frac{1}{2}}\phi \in \mathcal{R}(\Omega_0^{\frac{\beta-s}{2s}})$, it is of order*

$$||Var(x|\hat{Y}, g, s)\phi||^2 \sim \mathcal{O}_p(\alpha^{\frac{\beta}{a+s}}).$$

When the optimal α is used, the posterior variance converges at the optimal speed of $\mathcal{O}_p(\delta^{\frac{\beta}{\beta+a+\gamma(a+s)}})$.

3.3 A particular case

We consider in this section the particular case where L is chosen to be the canonical Hilbert scale $L = (K^*K)^{-\frac{1}{2}}$, *i.e.* L is chosen in according to the sampling model, and where, for some $r, s \in \mathbb{R}_+$

$$\delta = \frac{\sigma^2}{n}, \qquad \Sigma = (KK^*)^r, \qquad \Omega_0 = (K^*K)^s.$$

Then,

$$
\begin{aligned}
\hat{Y}|x &\sim \mathcal{GP}\Big(Kx, \frac{\sigma^2}{n}(KK^*)^r\Big) \\
x|g, s &\sim \mathcal{GP}\Big(x_0, \frac{\sigma^2}{g}(K^*K)^s\Big) \\
\hat{Y}|g, s &\sim \mathcal{GP}(Kx_0, \sigma^2\Big(\frac{1}{n}(KK^*)^r + \frac{1}{g}K(K^*K)^sK^*\Big)).
\end{aligned}
\tag{3.6}
$$

The prior distribution is in the extended Zellner's g-prior form, but when $s = 1$ we exactly have the Zellner's g-prior.

In this case, Assumption 9 *(a)* and *(b)* holds for $r \leq 1$ and $a = 1 - r$, respectively. Assumption 11 *(a)* holds for $s \geq 1$, while Assumption 11 *(c)* is trivially verified for $\gamma = \frac{1}{s+1-r}$ since in this case the eigenvalues of $(B^*B)^\gamma$ are equal to the square of the eigenvalues of K.

Hence, we replace Assumptions 9 and 11 by

Assumption 12

  (a) a, b and s are three real parameters satisfying the inequalities $0 < a \leq s \leq \beta \leq 2s + a$;

  (b) $r \leq 1$ and $s \geq 1$;

  (c) $a = 1 - r$ so that $||(K^*K)^{\frac{a}{2}}x|| = ||(KK^*)^{-\frac{r}{2}}Kx||$;

  (d) there exists a $\gamma \in]0, 1]$ such that the operator $(B^*B)^\gamma$ is trace class, *i.e.* if $\{\lambda_j^2\}$ denotes the eigenvalues of B^*B, then $\sum_j \lambda_j^{2\gamma} < \infty$.

Assumption 10 remains valid.
The expressions obtained for the general case simplify, so that

$$A = (K^*K)^{\frac{s}{2}}(\alpha I + (K^*K)^{\frac{s}{2}}K^*(KK^*)^{-r}K(K^*K)^{\frac{s}{2}})^{-1}((K^*K)^{-r}K(K^*K)^{\frac{s}{2}})^*,$$

with $\alpha = \frac{g}{n}$. We use the same decomposition of the posterior bias in the sum $C + D$ as in the previous section. Hence,

$$
\begin{aligned}
||C||^2 &= ||[I - (K^*K)^{\frac{s}{2}}(\alpha I + (K^*K)^{\frac{s}{2}}K^*(KK^*)^{-r}K(K^*K)^{\frac{s}{2}})^{-1}(K^*K)^{\frac{s}{2}}K^*(KK^*)^{-r}K](x_* - x_0)||^2 \\
&= ||[I - (K^*K)^s K^*(\alpha(KK^*)^r + K(K^*K)^s K^*)^{-1}K](K^*K)^{\frac{\beta}{2}}v||^2
\end{aligned}
$$

where the second equality has been obtained after permutation of the operator $(KK^*)^{-\frac{r}{2}}K(K^*K)^{\frac{s}{2}}$ with its adjoint and under Assumption 10. Let $\{\rho_j^2\}$ be the sequence of eigenvalues associated to operator K^*K (or equivalently to KK^*). The order of the squared norm $||C||^2$ is equal to the square of the maximum eigenvalues of C:

$$
\begin{aligned}
||C||^2 &\sim \left(\sup_j \left[\rho_j^\beta - \frac{\rho_j^{2(s+1)+\beta}}{\alpha \rho_j^{2r} + \rho_j^{2(s+1)}}\right]\right)^2 \\
&\sim \left(\sup_j \left[\rho_j^\beta - \frac{\rho_j^{2(s+1-r)+\beta}}{\alpha + \rho_j^{2(s+1-r)}}\right]\right)^2 \\
&\sim \left(\sup_j \left[\frac{\alpha \rho_j^\beta}{\alpha + \rho_j^{2(s+1-r)}}\right]\right)^2 \\
&\sim \mathcal{O}_p(\alpha^{\frac{\beta}{s+1-r}})
\end{aligned}
$$

that converges to zero if $r < s + 1$. Note in particular that, for the case considered here, $B = (KK^*)^{-\frac{r}{2}}K(K^*K)^{\frac{s}{2}}$ and it is well defined if $\sup_j \rho_j^{s+1-r} < \infty$, that is guaranteed if $s + 1 > r$ since ρ_j accumulates at zero. This condition is satisfied under Assumption 12 (b).

Markov inequality is still used to analyze term D, so that we obtain:

$$
\begin{aligned}
||D||^2 &\leq tr(Var(D)) \\
&= \frac{\sigma^2}{n} tr(A(KK^*)^r A^*) \\
&= \frac{\sigma^2}{n} \sum_j \frac{\rho_j^{2(2s+1-r)}}{(\alpha + \rho_j^{2(s+1-r)})^2} \\
&= \frac{\sigma^2}{n} \sum_j \frac{\rho_j^{2(s+1-r)(1-\gamma)+2s}}{(\alpha + \rho_j^{2(s+1-r)})^2} \rho_j^{2(s+1-r)\gamma} \\
&\leq \frac{\sigma^2}{n} \sup_j \frac{\rho_j^{2(s+1-r)(1-\gamma)+2s}}{(\alpha + \rho_j^{2(s+1-r)})^2} \sum_j \rho_j^{2(s+1-r)\gamma} \\
&\sim \mathcal{O}_p\left(\frac{1}{n}\alpha^{\frac{-\gamma(1-r+s)-1+r}{1-r+s}}\right).
\end{aligned}
$$

By equating the speed of convergence of $||C||^2$ and $||D||^2$ we get the optimal α:

$$
\begin{aligned}
\alpha^* &= c_3\left(\frac{1}{n}\right)^{\frac{s+1-r}{\beta+1-r+\gamma(1-r+s)}} \\
& c_3\left(\frac{1}{n}\right)^{\frac{s+a}{\beta+a+\gamma(a+s)}},
\end{aligned}
$$

for some constant c_3, that is the same rate obtained for the general case if $\delta = \frac{1}{n}$ and under Assumption 12 *(c)*. The fastest speed of convergence of the squared norm of the posterior mean is of order $\left(\frac{1}{n}\right)^{\frac{\beta}{\beta+a+1}}$, where we have used the value for a and γ.

From the optimal α we can find the optimal value of the associated g by using the relation $\alpha \propto \frac{g}{n}$:

$$g^* = c_4 \left(\frac{1}{n}\right)^{-\frac{\beta+\gamma(a+s)-s}{\beta+a+\gamma(a+s)}},$$

for some constant c_4, and it goes to ∞ if $\beta > s - \gamma(a+s)$ that is a condition than that one required for the general case.

The posterior variance has norm

$$
\begin{aligned}
||Var(x|\hat{Y}, g, s)\phi||^2 &= ||(K^*K)^{\frac{s}{2}}[I - (\alpha I + B^*B)^{-1}(K^*K)^{\frac{s}{2}}K^*(KK^*)^{-r}K(K^*K)^{\frac{s}{2}}](K^*K)^{\frac{s}{2}}\phi||^2 \\
&= ||(K^*K)^{\frac{s}{2}}[I - (\alpha I + B^*B)^{-1}(K^*K)^{\frac{s}{2}}K^*(KK^*)^{-r}K(K^*K)^{\frac{s}{2}}](K^*K)^{\frac{\beta-s}{2}}v||^2 \\
&\sim \mathcal{O}_p\left(\alpha^{\frac{\beta}{s+1-r}}\right)
\end{aligned}
$$

for any $\phi \in \mathcal{X}$ such that there exists a $v \in \mathcal{X}$ for which $(K^*K)^{\frac{s}{2}}\phi = (K^*K)^{\frac{\beta-s}{2}}v$.

Thus, we have proved the following Corollary to Theorems 14 and 11,

Corollary 2 *Under the distributional assumptions given in (3.6), under Assumptions 10 and 12 and if $\gamma = \frac{1}{s+1-r}$, then $||\mathbb{E}(x|\hat{Y}, g, s) - x_*||^2$ and $||Var(x|\hat{Y}, g, s)\phi||^2$ converge to zero with respect to the sampling probability. Moreover,*

$$||\mathbb{E}(x|\hat{Y}, g, s) - x_*||^2 \sim \mathcal{O}_p\left(\alpha^{\frac{\beta}{s+a}} + \frac{1}{n}\alpha^{-\frac{\gamma(a+s)+a}{s+a}}\right)$$

and $\forall \phi$ such that $(K^*K)^{\frac{s}{2}}\phi \in \mathcal{R}((K^*K)^{\frac{\beta-s}{2}})$

$$||Var(x|\hat{Y}, g, s)\phi||^2 \sim \mathcal{O}_p(\alpha^{\frac{\beta}{s+a}}).$$

Furthermore, if $\alpha = c_3 \left(\frac{1}{n}\right)^{\frac{s+a}{\beta+a+1}}$, for some constant c_3,

$$
\begin{aligned}
n^{\frac{\beta}{\beta+1+a}}||\mathbb{E}(x|\hat{Y}, g, s) - x_*||^2 &\sim \mathcal{O}_p(1) \\
n^{\frac{\beta}{\beta+1+a}}||Var(x|\hat{Y}, g, s)\phi||^2 &\sim \mathcal{O}_p(1).
\end{aligned}
$$

The definition of α as a regularization parameter demands that it satisfies the two conditions: $\alpha \to 0$ and $\alpha^2 n \to \infty$. Then, since $\alpha = \frac{g}{n}$, the optimal g must go to ∞ faster than $\sqrt{n}$ and slower than n. This is verified under the same conditions as in the general case: $\beta > (2s + a - \gamma(a+s))$ and $-a < s$.

3.4 g as an hyperparameter

In the preceding sections we have treated the parameter g in the prior distribution as a fixed parameter that has to be chosen in order to get the good rate of contraction of the prior distribution. Now, we want to consider g as an hyperparameter and express our degree of ignorance of the prior through a prior distribution on g.

The distributional scheme is the following:

$$\begin{aligned} g &\sim \nu \\ x|g &\sim \mu^g \\ \hat{Y}|x,g &\sim P^x. \end{aligned}$$

The indices g and x mean that the prior and the sampling distributions are conditioned on g and x, respectively. Hence, implicitly we are saying that, conditionally on x, $\hat{Y}$ is independent on g, in symbols $\hat{Y} \parallel g \,|x$. The specification of P^x and μ^g remains as in (3.2) and (3.3), respectively, *i.e.* $P^x \sim \mathcal{GP}(Kx, \delta\Sigma)$ and $\mu^g \sim \mathcal{GP}(x_0, \frac{1}{g}L^{-2s})$.

We use the joint conditional distribution of $(\hat{Y}, x)$, conditioned on g, to integrate out x from the sampling distribution. The resulting predictive distribution P^g is conditional on g. The model that we use to recover a posterior estimator for g is

$$\begin{aligned} g &\sim \nu \\ \hat{Y}|g &\sim P^g, \end{aligned}$$

with $P^g \sim \mathcal{GP}(Kx_0, \delta\Sigma + \frac{1}{g}K\Omega_0 K^*)$. A result of Kuo (1975) [51] shows that it is possible to define a density for P^g with respect to another measure different than the Lebesgue measure. We restate this result applied to our case in the following Theorem.

Theorem 12 *Let P^g be a gaussian measure on $\mathcal{Y}$ with mean Kx_0 and covariance operator $S_2 = (\delta\Sigma + \frac{1}{g}K\Omega_0 K^*)$ and P^∞ another gaussian measure on the same space with same mean and covariance operator $S_1 = \delta\Sigma$. If there exists a positive definite, bounded, invertible operator T such that $S_2 = S_1^{\frac{1}{2}}TS_1^{\frac{1}{2}}$ and $T - I$ is Hilbert-Schmidt, then P^g is equivalent to P^0. Moreover, the Radon-Nikodym derivative is given by*

$$\frac{dP^g}{dP^\infty} = \prod_{j=1}^{\infty} \sqrt{\frac{\alpha}{\lambda_j^2 + \alpha}} e^{\frac{\lambda_j^2}{2(\lambda_j^2 + \alpha)}z_j^2}, \tag{3.7}$$

with $\frac{\lambda_j^2}{\alpha}$ the eigenvalues of $T - I$, $z_j^2 = \frac{<\hat{Y} - Kx_0, \varphi_j>^2}{\delta l_j^2}$ and $\{l_j^2, \varphi_j\}$ the eigensystem associated to Σ.

It is possible to notice that

$$\left(\delta\Sigma + \frac{1}{g}K\Omega_0 K^*\right) = \sqrt{\delta}\Sigma^{\frac{1}{2}}\left[I + \frac{1}{g\sqrt{\delta}}\Sigma^{-\frac{1}{2}}K\Omega_0 K^*\Sigma^{-\frac{1}{2}}\frac{1}{\sqrt{\delta}}\right]\Sigma^{\frac{1}{2}}\sqrt{\delta},$$

so that $T = [I + \frac{1}{g\sqrt{\delta}}\Sigma^{-\frac{1}{2}}K\Omega_0 K^*\Sigma^{-\frac{1}{2}}\frac{1}{\sqrt{\delta}}]$. All the properties of T in the Theorem are trivially satisfied. Assumption 11 *(c)* guarantees that $T - I$ is Hilbert Schmidt, since it guarantees that $\sum_j \lambda_j^2 < \infty$ that implies that $\sum_j \lambda_j^4 < \infty$, where $\{\lambda_j^2\}$ are the eigenvalues of $\Sigma^{-\frac{1}{2}}K\Omega_0 K^*\Sigma^{-\frac{1}{2}}$.

The density in (3.7) has been expressed as function of α instead of g. This is aimed to directly select the regularization parameter $\alpha = \delta g$. We put a non-informative prior distribution on α (or equivalently on g) and we select the regularization parameter that maximizes the posterior distribution of α. Clearly, the posterior distribution of α is proportional to the density in (3.7) so that it is enough to maximize it with respect to α. The nice results that we get is that the value of α maximizing the posterior distribution is of the same order as the optimal one.

Lemma 5 *Under Assumptions 9, 10 and 11*

$$\frac{\partial \log(dP^g/dP^\infty)}{\partial \alpha} \sim \mathcal{O}_p(\alpha^{\frac{a+\beta}{a+s}} + \delta\alpha^{-\gamma}).$$

The Maximum a Posteriori (MAP) estimator for α is of order $\alpha^{MAP} \propto \delta^{\frac{a+s}{a+\beta+\gamma(a+s)}}$.

3.5 Conclusion

In this Chapter we have introduced a new class of prior distributions called *extended g-priors* in honour of Zellner's g-prior. These prior distributions are gaussian measures with a covariance operator that is linked to the sampling mechanism. The difference with respect to the classical g-priors is that the covariance operator does not need to be an exact transformation of operator K, but we admit for a more general relationship between the prior covariance operator and K. Furthermore, we require that, as the sample size increases, the prior distribution degenerates towards the prior mean at a rate faster than $\sqrt{n}$ and slower than n.

We analyze the classical signal-noise problem stated in infinite dimensional Hilbert spaces. We prove that when the prior distribution belongs to the class of extended g-prior, and under a certain compatibility between operators K and Σ in the sampling model, the posterior distribution of the signal is consistent. Thus, it can be used as a well-defined estimator of the solution of the signal-noise problem.

The assumptions that are necessary for having consistency of the posterior distribution, are satisfied by several statistical and econometric estimation problems. In these example the sampling covariance operator assumes a particular structure that simplifies computations and the proof of consistency. We have explicitly treated this particular case in Section 3.3. We show that the prior-to-posterior transformation acts as a regularization scheme and it can be interpreted either as a Tikhonov regularization or as a prior variance Hilbert scale regularization but that is directly introduced by the prior distribution. Therefore, the regularization parameter is part of the prior distribution of the signal x. Finally, we consider the regularization parameter as an hyperparameter and we propose a completely Bayesian method for optimally selecting the regularization parameter.

3.6 Appendix A: proofs

Proof of Lemma 5

We first consider the density $\frac{dP^g}{dP^\infty}$ in (3.7) with the product truncated at $J < \infty$. Its logarithm is proportional to

$$\sum_{j=1}^{J} \log \frac{\alpha}{\alpha + \lambda_j^2} + \sum_{j=1}^{J} \frac{(<K(x_* - x_0), \varphi_j>^2 + <U, \varphi_j>^2 + <K(x_* - x_0), \varphi_j><U, \varphi_j>\lambda_j^2)}{\delta l_j^2(\alpha + \lambda_j^2)}$$

after having replaced $\hat{Y}$ with its expression. Then, we equate to zero the derivative with respect to α and we multiply by $\delta\alpha$:

$$\overbrace{\frac{\delta}{\alpha}\sum_{j=1}^{J}\frac{\alpha\lambda_j^2}{\alpha+\lambda_j^2}}^{I_J} = \overbrace{\alpha\sum_{j=1}^{J}\frac{<K(x_*-x_0),\varphi_j>^2\lambda_j^2}{l_j^2(\alpha+\lambda_j^2)^2}}^{II_J} + \overbrace{\alpha\sum_{j=1}^{J}\frac{<U,\varphi_j>^2\lambda_j^2}{l_j^2(\alpha+\lambda_j^2)^2}}^{III_J}$$

$$+ \overbrace{\alpha\sum_{j=1}^{J}\frac{<K(x_*-x_0),\varphi_j><U,\varphi_j>\lambda_j^2}{l_j^2(\alpha+\lambda_j^2)^2}}^{IV_J}.$$

We take the limit for $J \to \infty$ of each term:

$$\begin{aligned}
\lim_J I_J &= \frac{\delta}{\alpha}\lim_J\sum_{j=1}^{J}\frac{\alpha\lambda_j^{2(1-\gamma)}}{\alpha+\lambda_j^2}\lambda_j^{2\gamma} \\
&\leq \frac{\delta}{\alpha}\Big(\sup_j\frac{\alpha\lambda_j^{2(1-\gamma)}}{\alpha+\lambda_j^2}\Big)\lim_J\sum_{j=1}^{J}\lambda_j^{2\gamma} \\
&\sim \mathcal{O}_p\Big(\frac{\delta}{\alpha^{\gamma}}\lim_J\sum_{j=1}^{J}\lambda_j^{2\gamma}\Big)
\end{aligned}$$

and the limit of the sum is finite under Assumption 11 *(c)*. To analyze term II_j, note that Assumption 9 *(b)* implies that $\Sigma^{-\frac{1}{2}}$ and $K\Omega_0 K^*$ have the same eigenfunctions. Then there exists $\{b_j\}$ such that $K\Omega_0 K^*\varphi_j = b_j\varphi_j$. Moreover, $\{\varphi_j\}$ are also the eigenvalues of BB^* since $BB^*\varphi_j = \Sigma^{-\frac{1}{2}}K\Omega_0 K^*\Sigma^{-\frac{1}{2}}\varphi_j = (b_j/l_j^2)\varphi_j$. Hence,

$$\begin{aligned}
\lim_j II_J &= \alpha\lim_J\sum_{j=1}^{J}\frac{<K\Omega_0^{\frac{1}{2}}\delta_*,\Sigma^{-\frac{1}{2}}\varphi_j>^2\lambda_j^2}{(\alpha+\lambda_j^2)^2} \\
&= \alpha\lim_J\sum_{j=1}^{J}\frac{<\Omega_0^{\frac{\beta-s}{2s}}\rho_*,\psi_j>^2\lambda_j^4}{(\alpha+\lambda_j^2)^2} \\
&= \alpha\lim_J\sum_{j=1}^{J}\frac{<(B^*B)^{\frac{\beta-s}{2(a+s)}}v,\psi_j>^2\lambda_j^4}{(\alpha+\lambda_j^2)^2} \\
&= \alpha\lim_J\sum_{j=1}^{J}\frac{<v,\psi_j>^2\lambda_j^{2(2+\frac{\beta-s}{(a+s)})}}{(\alpha+\lambda_j^2)^2} \\
&\leq \alpha\Big(\sup_j\frac{\lambda_j^{2(2+\frac{\beta-s}{(a+s)})}}{(\alpha+\lambda_j^2)^2}\Big)\lim_J\sum_{j=1}^{J}<v,\psi_j>^2 \\
&\sim \mathcal{O}_p\Big(\alpha^{\frac{1+\beta}{a+s}}||v||^2\Big).
\end{aligned}$$

By using Markov inequality it is possible to show that term III_J is negligible with respect to term I_J and that term IV_J is equal to zero in probability.

Then, the α^{MAP} is such that $\alpha^{\frac{a+\beta}{a+s}} = \frac{\delta}{\alpha^{\gamma}}$ and the result follows.

3.7 Appendix B: Numerical Implementation

We take $\mathcal{X} = L_\pi^2$ and $\mathcal{Y} = L_\pi^2$, with π the uniform distribution on $[0,1]$. Let $K = \int_0^1(s\wedge t)ds$, then K is self-adjoint: $K = K^*$. The data generating process is

$$
\begin{aligned}
\hat{Y} &= \int_0^1 x(s)(s \wedge t)ds + U, \qquad x_* = -3s^2 + 3s \tag{3.8} \\
U &\sim \mathcal{GP}(0, \frac{1}{n}KK), \\
x &\sim \mathcal{GP}(x_0, \frac{1}{\alpha n}\Omega_0), \qquad x_0 = -2.8s^2 + 2.8s \\
\Omega_0 \varphi(t) &= (KK)^s, \quad s = 1, \ \alpha n = g.
\end{aligned}
$$

We first compute estimation of x_* by fixing α to 0.3. In a second step we estimate α by using the technique suggested in Section 3.4. In order to compute it we need to write down the density $\frac{dP^g}{dP^\infty}$ with the product in it truncated at a certain $J < \infty$. We make use of the eigensystem $\{\lambda_j^2, \varphi_j\}$ associated to K. This eigensystem is well known to be $\lambda_j = \frac{4}{\pi^2 j^2}$, $\varphi_j(t) = \sqrt{2}\sin(\frac{\pi j t}{2})$, $j = 1, 3, 5, \ldots$.

In Figures 3.1a and 3.1c we represent the likelihood $\frac{dP^g}{dP^\infty}$ drawn against different values for α. We select the value of α that maximizes this curve and that is shown with an arrow in the figure. Then, we recompute the posterior distribution by taking this value of α. In Figure 3.1b and 3.1d, we show the true curve in black continuous line and the prior mean in dashed line. Then, in each graph, we represent the posterior means obtained for an arbitrarily selected value of α and for the value of α selected with the previous method.

(a) $x_0 = -2.8s^2 + 2.8s$, $N = 1000$, $J = 100$, $s = r = 1$

(b) $x_0 = -2.8s^2 + 2.8s$, $N = 1000$, $J = 100$, $s = r = 1$

(c) $x_0 = -2s^2 + 2s$, $N = 1000$, $J = 100$, $s = r = 1$

(d) $x_0 = -2s^2 + 2s$, $N = 1000$, $J = 100$, $s = r = 1$
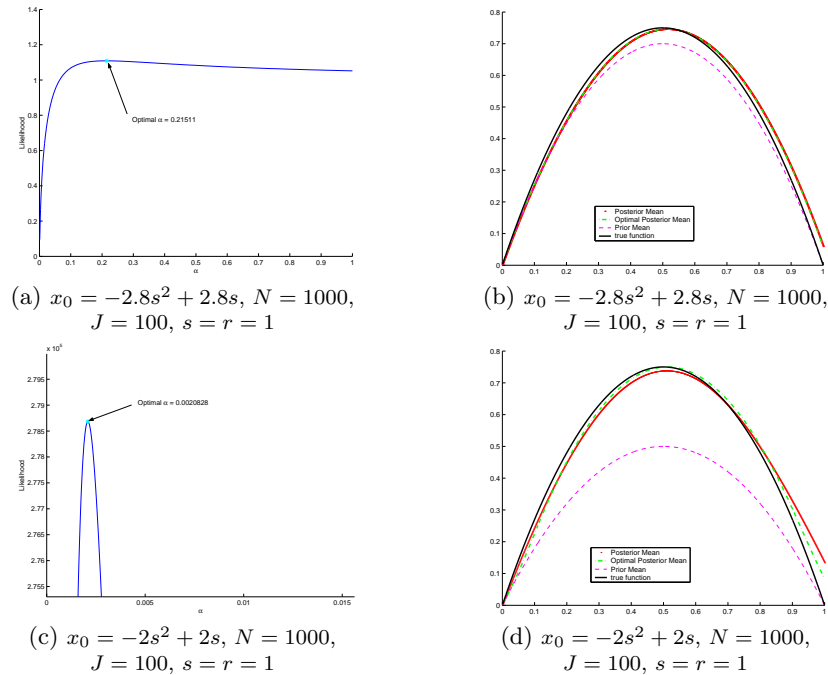
Figure 3.1

Chapter 4

Nonparametric Estimation of an Instrumental Regression: a Bayesian Approach Based on Regularized Posterior[1]

joint with Jean-Pierre Florens.

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 according to 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 models 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. Since this distribution is inconsistent in the sampling sense, 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: in the first step we get either a bayesian parametric estimator or a classical nonparametric estimator of $F(\cdot, Z, W)$ and in the second step we compute the regularized bayesian estimator of φ. We develop asymptotic analysis in a frequentist sense and posterior consistency is proved in all the three cases.

---

[1]This chapter is adapted from: Florens, J.P., and A., Simoni (2008), *Nonparametric Estimation of Instrumental Regression: a Bayesian Approach Based on Regularized Posterior*, mimeo.

4.1 Introduction

Instrumental regression estimation plays 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, with Z 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). Moreover, in order to characterize and define φ, some restriction concerning the disturbances in the model and the instrumental variables W must be satisfied by W.

A third approach proposed in literature for treating endogeneity problems is the *control function approach* proposed by Newey *et al.* (1999) [64]. They consider a triangular nonparametric simultaneous equations model with some restriction on the exogenous variables and on the error terms of the structural and reduced form equations.

In this chapter we adopt the instrumental regression approach. We increase the vector (Y, Z) with the vector W of instruments and we replace the classical hypothesis of exogeneity $\mathbb{E}(U|Z) = 0$ with the hypothesis $\mathbb{E}(U|W) = 0$. Our aim is to obtain a nonparametric Bayesian estimation of φ. As stressed by Newey and Powell (2003) [63], 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*. Hence, estimating φ is the same as solving an inverse problem.

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. 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) [33]. We refer to Chapter 2 for a more complete discussion about this approach.

Application of Bayes theorem in infinite dimensional spaces is perfectly known (see [33] and [60]), the posterior distribution of φ is well defined and the posterior mean is bounded and continuous in Y in finite samples. On the contrary, as the sample size increases, the posterior mean looses property of continuity and it is not consistent in the frequentist sense. This is due to the fact that its expression involves the inverse of a covariance operator that converges towards an unbounded operator. To overcome this problem, we adopt the strategy proposed in Chapter 2 consisting in applying a Tikhonov regularization scheme to the inverse of the covariance operator. The posterior distribution that results is slightly modified and it is called *regularized posterior distribution*.

The idea of estimating the instrumental regression φ by exploiting the theory of inverse problems is primarely due to Florens (2002) [26], Hall and Horowitz (2005) [38] and Darolles, Florens and Renault (2006) [15].

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, our proposed Bayesian estimator of the instrumental regression converges toward this true value. This convergence is known as posterior, or frequentist, consistency and it demands that the regularized posterior distribution degenerates in a Dirac measure in correspondence of the true value of the parameter of interest.

The paper is organized as follows. In Section 4.2 the instrumental variable model is write down. Section 4.3 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 4.4 we develop inference on φ when the joint distribution 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 distribution 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. Numerical simulations are in Section 4.5 and Section 5.7 concludes. All the proofs can be found in the Appendix.

4.2   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 defines the Hilbert space L_F^2 of square integrable functions with respect to F. We denote with $||\cdot||$ the norm in this space. We consider a model of the type

$$Y = \varphi(Z) + U, \qquad \mathbb{E}(U|Z) \neq 0. \tag{4.1}$$

This model is a *structural model* in the sense that it is directly proposed by the economic theory; it is characterized by the fact that the intervening variables Y and Z are both endogenous. The endogeneity of Y and Z can be explained by the fact that they have been simultaneously generated 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, makes 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

$$\mathbb{E}(Y|W) = \mathbb{E}(\varphi(Z)|W) \tag{4.2}$$

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 (4.1) is that it belongs to $L_F^2(Z)$, where $L_F^2(Z) \subset L_F^2$ is the subset of square integrable functions of Z with norm $|| \cdot ||$.

Uniqueness of the solution in (4.2) ensures identifiability, in the classical sense, of the parameter of interest φ by the moment condition (4.2) 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 zero).

Furthermore, a classical solution to equation (4.2) exists if and only if the regression function $\mathbb{E}(Y|W)$ belongs to the range of the conditional expectation operator $\mathbb{E}(\cdot|W) : L_F^2(Z) \to L_F^2(W)$, where $L_F^2(W) \subset L_F^2$ 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 classical solution are properties of the *cdf* F of S.

Anyway, we are not concerned with under and over-identification since our approach is Bayesian and we need a weaker condition for identification than the conditions necessary to guarantee existence and uniqueness of the classical solution. In a Bayesian optics, a model is identified if the prior distribution is completely revised. Though we are moved by a Bayesian philosophy in constructing our estimator, we adopt a classical (frequentist) notion of consistency, *i.e. posterior consistency* and then we need a condition for identification, but this condition is weaker than demanding injectivity of the conditional expectation operator, as it is done in the most of the classical literature about nonparametric instrumental regression estimation. In particular, if Ω_0 denotes the prior covariance operator in $L_F^2(Z)$, we will prove that our estimator will be consistent under the hypothesis that $\mathbb{E}(\Omega_0^{\frac{1}{2}}|W) : L_F^2(Z) \to L_F^2(W)$ is one-to-one on $L_F^2(Z)$. Hence, the identification assumption is

Assumption 13 *The operator* $\mathbb{E}(\Omega_0^{\frac{1}{2}}|W) : L_F^2(Z) \to L_F^2(W)$, *characterized by the true cdf* F, *is one-to-one on* $L_F^2(Z)$.

This assumption is weaker than requiring that $\mathbb{E}(\cdot|W)$ is one-to-one since if $\Omega_0^{\frac{1}{2}}$ and $\mathbb{E}(\Omega_0^{\frac{1}{2}}|W)$ are both one-to-one, this does not imply that $\mathbb{E}(\cdot|W)$ is one-to-one. This is caused by the fact that we are working in spaces of infinite dimension. [2]. On the contrary,

---

[2] If we were working on finite dimensional spaces, and consequently $\Omega_0^{\frac{1}{2}}$ and $\mathbb{E}(\cdot|W)$ would be matrices, $\Omega_0^{\frac{1}{2}}$ one to one and $\mathbb{E}(\Omega_0^{\frac{1}{2}}|W)$ one to one would imply $\mathbb{E}(\cdot|W)$ is one to one

if $\Omega_0^{\frac{1}{2}}$ and $\mathbb{E}(\cdot|W)$ are both one-to-one this do imply $\mathbb{E}(\Omega_0^{\frac{1}{2}}|W)$ is one-to-one.

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 (4.1) is

$$
\begin{aligned}
Y &= \mathbb{E}(Y|W) + \varepsilon, & \mathbb{E}(\varepsilon|W) = 0 \\
&= \mathbb{E}(\varphi(Z)|W) + \varepsilon, & \mathbb{E}(\varepsilon|W) = 0.
\end{aligned} \tag{4.3}
$$

The reduced form will be used as sampling model for inference. We stress that the error term ε in the reduced form is different from the structural error term U in (4.1). It should be noted that model (4.3) 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 variables 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. Note that with a *control function approach* we probably 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. We assume to observe a sample of S:

Assumption 14 $s_i = (y_i, z_i, w_i)$, $i = 1, \ldots, n$ 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 $i = 1, \ldots, n$. After having scaled every term in the reduced form by $\frac{1}{\sqrt{n}}$, we rewrite it in matrix form as

$$
y_{(n)} = K_{(n)}\varphi + \varepsilon_{(n)}, \tag{4.4}
$$

where

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

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

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

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 with $K = \mathbb{E}(\cdot|W)$ the operator from $L_F^2(Z)$ in $L_F^2(W)$ and with K^* its adjoint: $K^* = \mathbb{E}(\cdot|Z) : L_F^2(W) \to L_F^2(Z)$.

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 15 *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 $\varepsilon_{(n)} | \mathbf{w} \sim \mathcal{N}(0, \frac{\sigma^2}{n} I_n)$, where I_n is the identity matrix of order n. We only treat the homoskedastic case.

4.3 Bayesian Analysis

In this section we develop and analyze the Bayesian experiment associated to the reduced form model (4.4) and we consider a sample from it. Elements $y_{(n)}^i$ of vector $y_{(n)}$ represent n independent, but not identically distributed, draws from a sampling probability P^{σ, φ, w_i} conditional on $W = w_i$ [3]. 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 conditional sampling measure on $\mathcal{F}_Y$ associated to the whole vector $y_{(n)}$ and conditioned on the vector of instruments $\mathbf{w}$.

Two parameters characterize the model: the nuisance variance parameter σ^2 and the instrumental regression φ that represents the parameter of interest. We use 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 only been constrained to be square integrable with respect to F, implying that it belongs to $L_F^2(Z)$. We denote with $\mathcal{E}$ the σ-field of measurable subsets of $L_F^2(Z)$ and with μ^σ the prior distribution , conditional on σ^2.

Finally, the product parameter space is $(\mathbb{R}_+ \times L_F^2(Z), \mathcal{B} \otimes \mathcal{E}, \nu \times \mu^\sigma)$ and there exist two possible ways for specifying the probability measure on it. 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 subsection 4.3.3.

The conjugate bayesian experiment associated to model (4.4) 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 $\mathbb{E}(||\phi||^2) < \infty$, $\forall \phi \in L_F^2(Z)$, where $\mathbb{E}(\cdot)$ is the expectation taken with respect to μ^σ. Moreover, Ω_0 results to be a trace-class operator and this guarantees that realizations of this process will be in $L_F^2(Z)$ with probability 1.

The support of the centered prior distribution μ^σ is the closure in $L_F^2(Z)$ of the *Reproducing Kernel Hilbert Space* associated to Ω_0, ($\mathcal{H}(\Omega_0)$ in the following). Let $\{\lambda_j^{\Omega_0}, \varphi_j^{\Omega_0}\}_j$ be the eigensystem of the compact self-adjoint operator Ω_0, see Kress (1999) [50] for a

---

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

definition of eigensystem and singular value decomposition. We define the $\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.(\Omega_0)$ embedded in $L^2_F(Z)$ as:

$$\mathcal{H}(\Omega_0) = \{\phi : \phi \in L^2_F(Z) \quad \text{and} \quad \sum_{j=1}^{\infty} \frac{|<\phi, \phi_j^{\Omega_0}>|^2}{\lambda_j^{\Omega_0}} < \infty\} \tag{4.5}$$

and, following Proposition 3.6 in [10], we have the relation $\mathcal{H}(\Omega_0) = \mathcal{R}(\Omega_0^{\frac{1}{2}})$.

If Ω_0 is injective then $\mathcal{H}(\Omega_0)$ is dense in $L^2_F(Z)$ and the support of μ^σ will be the whole $L^2_F(Z)$.

Under Assumption 15 the sampling probability $P^{\sigma,\varphi,\mathbf{w}}$ is gaussian with mean $K_{(n)}\varphi$ and covariance matrix $\frac{\sigma^2}{n}I_n$. 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.

From a classical point of view, there exists a true value of the regression function that has generated data $y_{(n)}$ through model (4.4). We denote this value with φ_* and we assume that

Assumption 16 $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$, *i.e. there exists* $\delta_* \in L^2_F(Z)$ *such that* $\varphi_* - \varphi_0 = \Omega_0^{\frac{1}{2}}\delta_*$.

This assumption is only a regularity condition and it will be exploited for proving asymptotic results with a convergence analyzed in the sampling sense. In reality, the gaussian prior measure μ^σ is not able to generate trajectories in this space since $\mu^\sigma\{\varphi; \varphi \in \overline{\mathcal{H}(\Omega_0)}\} = 1$, but $\mu^\sigma\{\varphi; \varphi \in \mathcal{H}(\Omega_0)\} = 0$. However, if Ω_0 is injective, $\mathcal{H}(\Omega_0)$ is dense in $L^2_F(Z)$ and μ^σ is able to generate trajectories as close as possible to the true one. The incapability of the prior to generate the true parameter characterizing the data generating process is known in literature as *prior inconsistency*. 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. [34] or [37].

The elements in $K_{(n)}$ and $K^*_{(n)}$ depends on the density $f(Z, W)$ and its marginalizations. For the moment we take these densities as known, but this is not always true in real applications. When they are unknown they can be seen as nuisance parameters affecting both distributions $P^{\sigma,\varphi,\mathbf{w}}$ and $P^{\sigma,\mathbf{w}}$. In Section 4.4 we will analyze the unknown density case and we will index these two probabilities with f: $P^{f,\sigma,\varphi,\mathbf{w}}$ and $P^{f,\sigma,\mathbf{w}}$.

Summarizing, we have

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

so that $(\varphi, y_{(n)})$ is a jointly gaussian process conditionally on σ^2.

The posterior distribution of σ^2 is easily computable and does not rise any relevant problem; we will handle it in the next subsection.

More problems are found concerning the posterior distribution of φ. We start by considering the conditional posterior distribution of φ, conditional on σ^2. The main theoretical question concerns the existence of conditional gaussian processes in Hilbert spaces, namely

the existence of a transition probability characterizing the posterior distribution of φ conditional on σ^2. The existence of a well-defined posterior distribution is guaranteed by *Jirina theorem*, see Neveu (1965) [62] since the spaces we are working in are Polish spaces[4]. We refer to Chapter 2 of this essay, to Franklin (1970) [33] and Mandelbaum (1984) [60] for a complete discussion about this point.

The conditional posterior distribution of φ, given σ^2, is gaussian; this follows from the form assumed by the characteristic function of φ given $(y_{(n)}, \sigma^2)$. We characterize this distribution in the following Theorem. 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 remark again that all the posterior probability have to be meant computed for a given **w**.

Theorem 13 *Let φ and $y_{(n)}$ be two conditionally jointly distributed gaussian random elements, conditional on σ^2, in $L_F^2(Z)$ and $\mathbb{R}^N$, respectively. Then, the conditional distribution of φ given $y_{(n)}$ and σ^2 is gaussian with mean $Ay_{(n)} + b$, where*

$$A = \Omega_0 K_{(n)}^* C_n^{-1}, \qquad b = (I - AK_{(n)})\varphi_0 \qquad (4.6)$$

and covariance given by

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

Proof of this theorem can be found in Mandelbaum (1984) [60]. 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 $\mathbb{E}(\varphi|y_{(n)}, \sigma^2) = \mathbb{E}(\varphi|y_{(n)})$.

For small samples, the posterior distribution $\mu^{\sigma,\mathcal{F}}$ is well defined in the sense the operators in its mean and variance are bounded due to the fact that C_n is an invertible $n \times n$ matrix because its n eigenvalues are all different than zero. On the contrary, as $n \to \infty$, the inverse C_n^{-1} that appears in operator A converges towards a noncontinuous operator and then A converges to a non-continuous linear operator defined in the set of $(y_i)_{i \in \mathcal{N}}$. This prevents the posterior mean from being consistent in a sampling sense even if it is consistent in a Bayesian sense, *i.e.* with respect to the joint distribution of observations and parameters.

For different reasons explained just below, we want that our Bayesian estimator be consistent in the sampling sense, namely with respect to the sampling probability. In the following, terms like *frequentist consistency*, *classical consistency* or *posterior consistency* will be equally used for referring to this convergence. The pair $(\varphi, \mu^{\sigma,\mathcal{F}})$ is consistent,in the classical sense 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 classical sense if $(\varphi, \mu^{\sigma,\mathcal{F}})$ is consistent for all φ. This concept of frequentist consistency is extensively developed in Diaconis and Freedman (1986) [16] among others, where Bayesians are separated into two groups: "classical" and "subjectivist". Classical bayesians believe 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, posterior 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 [16] and [27]).

Furthermore, having a posterior distribution (and hence a bayesian estimator) that is consistent justifies, also from a classical point of view, the estimator obtained with a bayesian

---

[4]A *Polish space* is a separable completely metrizable topological space. Both $L_F^2(Z)$ and $\mathbb{R}$ are Polish spaces, see for instance [45].

approach. On the basis of this argument we are persuaded about the importance to find an estimator that is consistent in the sampling sense. The following lemma states the non consistency of the posterior $\mu^{\sigma,\mathcal{F}}$.

Lemma 6 *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 a point mass δ_{φ_*} in φ_*.*

<u>Proof:</u> See Appendix 4.7. ■

The intuition of this lemma is that, 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 of the limit of the inverse of $K_{(n)}\Omega_0 K_{(n)}^*$. The number of eigenvalues of C_n grows with n up to form a decreasing sequence having 0 as the only accumulating point. Contrarily to finite dimensional cases, where the Ridge regression has a Bayesian interpretation and a regularization effect, the prior specification does not solve the problem of ill-posedness in infinite dimensional problems because of compacity of Ω_0.

For solving the lack of continuity we propose to apply a Tikhonov regularization scheme to the inverse of C_n: $C_{n,\alpha}^{-1} = (\alpha_n I_n + \frac{1}{n}I_n + K_{(n)}\Omega_0 K_{(n)}^*)^{-1}$, where α_n is a regularization parameter. In practice, this consists in translating the eigenvalues of C_n far from 0 by a factor $\alpha_n > 0$. As $n \to \infty$, $\alpha_n \to 0$ at a suitable rate to ensure operator C_n stays well defined asymptotically.

We call *Regularized (conditional) Posterior Distribution*, denoted with $\mu_\alpha^{\sigma,\mathcal{F}}$, the conditional distribution on $\mathcal{E}$, given $(y_{(n)},\sigma^2)$, defined in Theorem 13 in which operator A has been substituted by the Tikhonov regularized operator $A_\alpha = \Omega_0 K_{(n)}^* C_{n,\alpha}^{-1}$. This object has been introduced in Chapter 2 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 φ on a space of smaller dimension. Even if the problem we are considering is substantially different with respect to that one considered in Chapter 2, asymptotic arguments motivates us to adopt the regularized posterior distribution $\mu_\alpha^{\sigma,\mathcal{F}}$ as solution for our inference problem. On the other side, if we wanted to solve (4.4) 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.

Summarizing, the regularized (conditional) posterior distribution $\mu_\alpha^{\sigma,\mathcal{F}}$ is a gaussian measure defining a mean element and a covariance operator

$$
\begin{aligned}
\hat{\varphi}_\alpha &:= \mathbb{E}_\alpha(\varphi|y_n,\sigma^2) = A_\alpha y_{(n)} + b_\alpha \\
\sigma^2 \Omega_{y,\alpha} &:= Var_\alpha(\varphi|y_n,\sigma^2) = \sigma^2(\Omega_0 - A_\alpha K_{(n)}\Omega_0),
\end{aligned}
\tag{4.7}
$$

with

$$
\begin{aligned}
A_\alpha &= \Omega_0 K_n^* \Big(\alpha_n I + \frac{1}{n}I + K_{(n)}\Omega_0 K_{(n)}^*\Big)^{-1} \\
b_\alpha &= (I - A_\alpha K_{(n)})\varphi_0.
\end{aligned}
$$

We will take the regularized posterior mean as punctual estimator for the instrumental regression, as suggested for a quadratic loss function. In section 4.3.2 we will state consistency of this solution.

4.3.1 The Student t Process

We proceed now to computation of the posterior distribution of σ^2. Then, this distribution will be exploited in order to marginalize the regularized conditional posterior distribution $\mu_\alpha^{\sigma,\mathcal{F}}$ of φ, given σ^2, with respect to σ^2.

A conjugate model allows to integrate out φ from the sampling probability $P^{\sigma,\varphi,\mathbf{w}}$ to obtain $P^{\sigma,\mathbf{w}} := \mathbb{P}(y_{(n)}|\sigma^2, \mathbf{w})$ and then to use the two probabilities

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

to make inference on σ^2. The posterior distribution of σ^2 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 $\mathcal{IT}$ distribution [5]. Then

$$\sigma^2|y_{(n)} \sim \mathcal{IT}(\nu_*, s_*^2), \quad \text{with}$$
$$\nu_* = \nu_0 + n$$
$$s_*^2 = s_0^2 + (y_{(n)} - K_{(n)}\varphi_0)'\Big(\frac{1}{n}I_n + K_{(n)}\Omega_0 K_{(n)}^*\Big)^{-1}(y_{(n)} - K_{(n)}\varphi_0).$$

Obviously the posterior distribution of σ^2 does not depend on φ and then it can be used for marginalizing the regularized conditional posterior distribution of φ by directly integrating out σ^2. When the model is conjugate we do not necessitate of a Gibbs sampling, as in the case with independent priors.

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: $\varphi|y_{(n)}$ should be a *Student t Process* in $L_F^2(Z)$. We introduce a new process called *Student t Process*. In the next definition, we define it in a general Hilbert space through the scalar product in this space.

Definition 2 *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

---

[5]There exist different specifications of the *Inverse Gamma* distribution; we use in our study an $\mathcal{IT}(\nu_0, s_0^2)$ 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.

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

with mean and variance

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

At the best of our knowledge, this kind of process has never been encountered in the existing literature.

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

Lemma 7 *Let $\sigma^2 \sim \mathcal{IT}(\nu, s^2)$ and $x|\sigma^2 \sim \mathcal{GP}(x_0, \sigma^2\Omega_0)$, with $\sigma^2 \in \mathbb{R}_+$ and $x \in \mathcal{X}$. Then,*

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

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 the instrumental variable process φ for integrating out σ^2 in the regularized posterior distribution. Hence,

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

with marginal mean $\hat{\varphi}_\alpha$ and marginal variance $\frac{s_*^2}{\nu_*-2}\Omega_{y,\alpha}$. We call this distribution *regularized posterior distribution* and denote it with $\mu_\alpha^\mathcal{F}$.

4.3.2 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.

The regularized posterior distribution $\mu_\alpha^{\sigma,\mathcal{F}}$ is consistent if the probability, taken with respect to this distribution, of any complement of a neighborhood of φ_* converges to zero. Posterior consistency of $\mu_\alpha^{\sigma,\mathcal{F}}$ is stated in the following theorem.

Theorem 14 *Let φ_* be the true value having generated the data 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 $\sigma^2\Omega_{y,\alpha}$ defined in 4.7. Under Assumption 24, if $\alpha_n \to 0$ and $\alpha_n^2 n \to \infty$, then:*

(i) $\mu_\alpha^{\sigma,\mathcal{F}}$ weakly converges to a 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

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

The condition required for the second part of the theorem is only a regularity condition that is necessary for having convergence at a certain speed. The hypothesis that really matters for having posterior consistency is the fact that $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$.

A corollary provides the necessary results for Theorem 14, it concerns consistency of the regularized posterior mean and convergence to zero of the regularized posterior variance.

Corollary 3 *Under Assumption 24, if $\alpha_n \to 0$ and $\alpha_n^2 n \to \infty$,*

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

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

(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 $\beta > 0$,

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

The rates governing the bias are the first and the third one in brackets, 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 and variance (in squared norm) proportional to $n^{-\frac{\beta}{\beta+2}}$ that is the fastest one. The regularized posterior distribution converges at the slower rate $n^{-\frac{\beta}{2(\beta+2)}}$.

Now we concentrate on the posterior consistency of $\nu^{\mathcal{F}}$. We denote with $g(Z, w_i)$ 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 $\Omega_0^{\frac{1}{2}}$, $g(Z, w_i) = \Omega_0^{\frac{1}{2}} \frac{f(s,w_i)}{f(s)f(w_i)} = \int \omega_0(s, Z)\frac{f(s,w_i)}{f(s)f(w_i)}f(s)ds$. 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 15 *Let σ_*^2 be the true value of σ^2 having generated the data and $\nu^{\mathcal{F}}$ the posterior Inverse Gamma distribution on $\mathbb{R}_+$ described in subsection 4.3.1. Under Assumption 24, 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:*

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

It follows that $\nu^{\mathcal{F}}\{\sigma^2 : |\sigma^2 - \sigma_*^2| \geq \epsilon_n\} \to \delta_{\sigma_*^2}$, where $\delta_{\sigma_*^2}$ is the point mass in σ_*^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 8 *Under condition of Theorems 14 and 15, the joint measure*

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

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

4.3.3 Independent Priors

We adopt in this section an alternative specification of the joint prior measure on the parameter space: we assume that the prior for φ does not depend on σ^2. Then, we denote with μ the prior on $L_F^2(Z)$ and the joint prior distribution on $\mathbb{R}_+ \times L_F^2(Z)$ is equal to the product of the two marginal ν and μ. Hence,

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

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 marginal posterior distributions of φ and σ^2 unfeasible, nevertheless it is possible to obtain closed form for the posterior distribution of φ conditional on σ^2, $\mu^{\sigma,\mathcal{F}}$ and for the posterior distribution of σ^2 conditional on φ, denoted with $\nu^{\varphi,\mathcal{F}}$. 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 $\nu_\alpha^{\mathcal{F}}$.[6]
We start by computing the conditional posterior distribution of φ. Conditionally on σ^2 $(y_{(n)}, \varphi)$ are jointly normally distributed with mean and variance

$$
\mathbb{E}\left(\begin{array}{c} y_{(n)} \\ \varphi \end{array} \right) = \left(\begin{array}{c} K_{(n)}\varphi_0 \\ \varphi_0 \end{array} \right), \qquad Var\left(\begin{array}{c} y_{(n)} \\ \varphi \end{array} \middle| \right) = \left(\begin{array}{cc} \Omega_0 & \Omega_0 K_{(n)}^* \\ K_{(n)}\Omega_0 & (\frac{\sigma^2}{n}I + K_{(n)}\Omega_0 K_{(n)}^*) \end{array} \right)
$$

and the parameter σ^2 only affects the variance of $y_{(n)}$. The conditional posterior of φ still suffers of a problem of inconsistency since it demands the inversion of the covariance operator of $y_{(n)}|\sigma^2$ that, as $n \to \infty$, converges towards an operator with non continuous inverse. Hence, we use the Tikhonov regularization scheme already introduced and the regularized conditional posterior distribution of φ, still denoted with $\mu_\alpha^{\sigma,\mathcal{F}}$ is a gaussian measure:

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

It should be remarked the difference between operators A_α^σ and A_α and operators $\Omega_{y,\alpha}^\sigma$ and $\Omega_{y,\alpha}$ defined in the conjugate case.
For computing the posterior distribution of σ^2 given φ, we use the homoskedastic model specified in Assumption 15 for the reduced form error terms: $\varepsilon_{(n)}|\sigma^2$ i.i.d. $\sim \mathcal{N}(0, \frac{\sigma^2}{n}I_n)$, with $\varepsilon_{(n)} = y_{(n)} - K_{(n)}\varphi$. Computation of this posterior distribution demands to know φ and this means, in a Gibbs sampling algorithm, that we have to draw φ from $\mu_\alpha^{\sigma,\mathcal{F}}$ before to use it for constructing $\varepsilon_{(n)}$. This makes clear the fact that the regularization scheme

---

[6]The meaning of the index α will be clarified below.

plays a role also in the conditional posterior distribution for σ^2 through φ, so that we also index the conditional posterior distribution of σ^2 with α: $\nu_\alpha^{\varphi,\mathcal{F}}$. Then, we talk about *regularized error term* for $\varepsilon_{i,\alpha} = y_i - K_{(n)}^i \varphi$, with $\varphi \sim \mu_\alpha^{\sigma,\mathcal{F}}$ and $K_{(n)}^i$ denoting the i-th component of the vector. Trivial computations provide us with the conditional posterior

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

The associated Gibbs sampling algorithm is the following:

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

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

(iii) draw $\sigma_{(i)}^2$ 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 $\sigma_{(0)}^2$ and the number of iterations J necessary to get the stationary distribution. We propose to draw the starting value $\sigma_{(0)}^2$ from an $\mathcal{IT}$ 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 $\sigma^2 \sim \mathcal{IT}(\tilde{\nu}_0, \tilde{s}_0^2)$, $\mathbb{E}(\sigma^2) = \frac{\tilde{s}_0^2}{\tilde{\nu}_0 - 2}$ and then $\tilde{s}_0^2 = \hat{\sigma}^2(\tilde{\nu}_0 - 2)$. Lastly, $\tilde{\nu}_0$ 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 $\sigma_{(0)}^2$ from $\mathcal{IT}(\tilde{\nu}_0, \tilde{s}_0^2)$. In order to determine the number of iterations J we propose a method that is an adaptation of the technique proposed in Gelman and Rubin (1992) [35]. 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 quantities of the target distribution to see how close the simulation process is to convergence.

We simulate M independent sequences, each one with length $2J$, with different starting points drawn from $\mathcal{IT}(\tilde{\nu}_0, \tilde{s}_0^2)$, with $\tilde{\nu}_0$ and $\tilde{s}_0^2$ determined as described above:

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

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.

The first J iterations of each sequence are discarded and the last J are used to compute the following quantities:

$$B = \frac{J}{M-1}\sum_{i=1}^{M}(\overline{\sigma_{i.}^2} - \overline{\sigma_{..}^2})^2, \quad \overline{\sigma_{i.}^2} = \frac{1}{J}\sum_{j=1}^{J}\sigma_{ij}^2, \quad \overline{\sigma_{..}^2} = \frac{1}{M}\sum_{i=1}^{M}\overline{\sigma_{i.}^2}$$

$$WW = \frac{1}{M}\sum_{i=1}^{M}s_i^2, \quad s_i^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\sigma_{ij}^2 - \overline{\sigma_{i.}^2})^2$$

$$\widehat{Var}(\sigma^2) = \frac{J-1}{J}WW + \frac{1}{J}B.$$

B is the between-sequence variance and WW 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 overestimation 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.

$$B^\varphi(\cdot) = \frac{J}{M-1}\sum_{i=1}^{M}(\overline{\varphi_{i.}}(\cdot) - \overline{\varphi_{..}}(\cdot))^2, \quad \overline{\varphi_{i.}}(\cdot) = \frac{1}{J}\sum_{j=1}^{J}\varphi_{ij}(\cdot), \quad \overline{\varphi_{..}}(\cdot) = \frac{1}{M}\sum_{i=1}^{M}\overline{\varphi_{i.}}(\cdot)$$

$$WW^\varphi(\cdot) = \frac{1}{M}\sum_{i=1}^{M}(s_i^\varphi)^2(\cdot), \quad (s_i^\varphi)^2(\cdot) = \frac{1}{J-1}\sum_{j=1}^{J}(\varphi_{ij}(\cdot) - \overline{\varphi_{i.}}(\cdot))^2$$

$$\widehat{Var}(\varphi(\cdot)) = \frac{J-1}{J}WW^\varphi(\cdot) + \frac{1}{J}B^\varphi(\cdot).$$

To monitor convergence of the iterative simulation, it is suggested in Gelman and Rubin (1992) [35] 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 $\hat{R} = \frac{\widehat{Var}(\sigma^2)}{WW}$ 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 [35] is adapted only 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.

4.4 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 considerably simplifies 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 when the joint density $f(Z, W)$ is known up to a parameter θ of finite dimension and then when $f(Z, W)$ is totally unknown. In the latter case, nonparametric estimation methods require to be considered.

4.4.1 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 from $F(Z, W)$, the Bayesian experiment is

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

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. F^θ represents the sampling distribution on $\mathcal{F}_{Z,W}$. 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 17 $s_{1,i} = (y_i, w_i)$, $i = 1, \ldots, n$ and $s_{2,j} = (\tilde{z}_j, \tilde{w}_j)$, $\tilde{i} = 1, \ldots, \tilde{n}$ are two *i.i.d. samples of observations on* $S_1 = (Y, W)$ *and* $S_2 = (Z, W)$, *respectively. Moreover, we assume that the samples* $(w_i)_{i=1,\ldots,n}$ *and* $(\tilde{w}_j)_{j=1,\ldots,\tilde{n}}$ *have been generated by the same distribution* $F(W)$ *characterized by the true value of parameter* θ_*.

To simplify things we suppose the variance parameter σ^2 to be known and we use the notation Ξ_θ, instead of Ξ as before, for the Bayesian experiment for φ. Hence, Bayesian model Ξ_θ results to be

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

The sampling and marginal probabilities in Ξ_θ depends on the realized value of θ, this justifies the 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 4.3 can be seen as a particular case of Ξ_θ, 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 only demands the use of 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 $(p+q) \times (p+q)$ positive definite matrices. Therefore $\theta = (m, V) \in \Theta = \mathbb{R}^{p+q} \times \mathcal{C}_{p+q}$ and

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

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:

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

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)

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

and its decomposition

$$\rho(\Sigma|(s_{2,i})_{i=1,\ldots,\tilde{n}}) \sim \mathcal{W}(\Sigma_*, v_*) \tag{4.8}$$

$$\rho(m|\Sigma; (s_{2,i})_{i=1,\ldots,\tilde{n}}) \sim \mathcal{N}_{(p+q)}(m_*, \frac{1}{u_*}\Sigma^{-1}), \tag{4.9}$$

with

$$
\begin{aligned}
u_* &= \tilde{n} + u_0 \\
m_* &= \frac{1}{u_*}\left(\sum_i s_{2,i} + u_0 m_0\right) \\
v_* &= \tilde{n} + v_0 \\
\Sigma_*^{-1} &= \Sigma_0^{-1} + \sum_i s_{2,i} s'_{2,i} + u_0 m_0 m'_0 - \frac{\tilde{n}^2}{u_*}\bar{s}_2 \bar{s}'_2 - \tilde{n}\frac{u_0}{u_*}(\bar{s}_2 m'_0 + m_0 \bar{s}'_2) - \frac{u_0^2}{u_*} m_0 m'_0 \\
\bar{s}_2 &= \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} s_{2,i}.
\end{aligned}
$$

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:

$$
\mathbb{E}_\alpha(\varphi|y_{(n)}, \mathbf{w}) = \int \mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}})\rho(\theta|y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}})d\theta \tag{4.10}
$$

$$
\begin{aligned}
Var_\alpha(\varphi|y_{(n)}, \mathbf{w}) = &\int Var_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}})\rho(\theta|y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}})d\theta + \\
&Var(\mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}})|y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}})
\end{aligned} \tag{4.11}
$$

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

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

Quantities (4.10) and (4.11) 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 $\rho(\theta|(s_{2,i})_{i=1,\dots,\tilde{n}})$. 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) to draw $\theta^{(j)}$ from the posterior $\rho(m, \Sigma|(s_{2,i})_{i=1,\dots,\tilde{n}})$;

(ii) to 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) to 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 $A_\alpha^{(j)} = \Omega_0 \hat{K}^{*(j)}(\alpha_n I_n + \hat{K}^{(j)}\Omega_0 \hat{K}^{*(j)} + \frac{\sigma^2}{n} I_n)^{-1}$ and $b_\alpha^{(j)} = (I - A_\alpha^{(j)} \hat{K}^{(j)})\varphi_0$;

(iv) to 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) to 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, $\tilde{\Omega}_{y,\alpha} = \frac{1}{J}\sum_j \Omega_{y,\alpha}^{(j)}$, to approximate the first term in the RHS of (4.11). Approximate the second term in the RHS of (4.11) by $\frac{1}{J}\sum_j (\hat{\varphi}_\alpha^{(j)})^2 - (\frac{1}{J}\sum_j \hat{\varphi}_\alpha^{(j)})^2$. Let 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:

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

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.* $\hat{\theta} = \int \theta\rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}})d\theta$. 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 4.

The following theorem shows consistency of the estimated posterior mean under some minor hypothesis.

Theorem 16 *Let $\hat{\varphi}_\alpha = \hat{\mathbb{E}}_\alpha(\varphi|y_{(n)}, \boldsymbol{w})$. Under Assumption 24, if $\alpha_n \to 0$, $\frac{1}{\alpha_n n} \to 0$, $\frac{1}{\alpha_n^3 n^2} \sim \mathcal{O}_p(1)$, 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 $\delta_* \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$, then

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

We implicitly assume in Theorem 16 that all the conditions necessary to guarantee consistency of the posterior mean $\hat{\theta}$ of a finite dimensional parameter are satisfied, see Bernstein (1934) [5], Gosh and Ramamoorthi (2003) [36] or Von Mises (1964) [79] for technical details.

Let study convergence to zero of the regularized posterior variance $\hat{\Omega}_{y,\alpha}$:

$$
\hat{\Omega}_{y,\alpha}\phi = \frac{1}{J}\sum_{j=1}^J [\Omega_{y,\alpha}(\theta^{(j)})\phi](\zeta) + \frac{1}{J}\sum_{j=1}^J (\hat{\varphi}_\alpha^{(j)})^2(\zeta) - \left(\frac{1}{J}\sum_{j=1}^J \hat{\varphi}_\alpha\right)^2(\zeta) \tag{4.12}
$$

with $\phi \in L_F^2(Z)$, $\Omega_{y,\alpha}(\theta^{(j)}) = Var_\alpha(\varphi|\theta^{(j)}, y_{(n)}, \mathbf{w}) = \Omega_{y,\alpha}^{(j)}$.

Theorem 17 *Let $\hat{\Omega}_{y,\alpha} : L_F^2(Z) \to L_F^2(Z)$ be computed as in (4.12). If $\alpha_n \to 0$, $\frac{1}{\alpha_n n} \to 0$, $\frac{1}{\alpha_n^3 n^2} \sim \mathcal{O}_p(1)$, $\frac{\partial \mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \boldsymbol{w})}{\partial\theta} \in L_F^2(Z)$ and $\frac{\partial}{\partial\theta}\Omega_{y,\alpha}(\theta) \in L_F^2(Z)$ for $\theta = \theta_*$ and $\theta = \hat{\theta}$ then*

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

(ii) moreover, $\forall \phi \in L_F^2(Z)$ 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 $\beta > 0$

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

Chebyshev's Inequality allows to show that, under conditions given for point *(i)* of Theorems (16) and (17), posterior consistency is preserved also in the case with unknown θ:

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

Moreover, under conditions given in point *(ii)* of Theorems 16 and 17, with optimal regularization parameter α_*, $\hat{\mu}_\alpha^{\mathcal{F}}$ degenerates towards a point mass in φ_* at the optimal speed of $n^{-\frac{\beta}{2(\beta+2)}}$. This means that the optimal speed of convergence does not change with respect to the better case in which F is completely known.

4.4.2 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 estimation was based on Dirichlet processes (introduced by Ferguson (1973) [23]) that has the drawback of producing discrete random probabilities measures with probability one. In alternative, Polya tree priors, initially considered by Ferguson (1974) [24] and then by Lavine (1992) [52], can be chosen to generate only absolutely continuous distributions. We refer to Choudhuri *et al.* (2005) [13] for a complete review on Bayesian nonparametric methods. The technique that we propose in this paper for dealing with this case is essentially different and it does not appear among 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 the solution of the estimated reduced form equation

$$y_{(n)} = \hat{K}_{(n)}\varphi + \eta_{(n)} + \varepsilon_{(n)}. \tag{4.13}$$

We use the notation $\hat{K}_{(n)}$ and $\hat{K}_{(n)}^*$ for the corresponding operators with the density $f(Z,W)$ substituted by a nonparametric estimator. We have two error terms: $\varepsilon_{(n)}$ is the classical error term of the reduced form and $\eta_{(n)}$ 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 $\eta_{(n)} = (\eta_1, \ldots, \eta_n)'$ [7]. 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). Equation (4.13) defines a new Bayesian experiment that results to be a slightly modification of Ξ in Section 4.3 primarily for the fact that σ^2 is known and then it no more enters the Bayesian experiment, and secondly for the fact that the sampling distribution depends on $\hat{f}(Z,W)$ instead of on $f(Z,W)$ (we will see this below):

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

Nonparametric estimation of $f(Z,W)$ is performed by kernel smoothing; we stress the fact that here, contrarily to the previous case where f was known up to a parameter θ, we use the same sample for getting an estimate of f and the posterior distribution of φ. Let

---

[7]It should be noted that properties of η_i are not affected by the function φ at which operator $\hat{K}_{(n)}$ is applied.

L be a kernel function satisfying the usual properties and ρ be the minimum between the order of L and the order of differentiability of f. We use the notation $L(u)$ for $L(\frac{u}{h})$ where h is the bandwidth used for kernel estimation such that $h \to 0$ as $n \to \infty$ (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 subscripts in the kernel for W and for Z, L_w and L_z respectively. Estimate of $K_{(n)}$ and $K_{(n)}^*$ are obtained by plugging in the estimate $\hat{f}$:

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

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

and

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

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 18 $\eta_{(n)} \sim \mathcal{N}_n(0, \frac{\sigma^2}{n^2 h^q} D_{(n)})$, where $D_{(n)} = diag(\frac{1}{f(w_i)} \int L_w^2(u) du)$, $i = 1, \ldots, n$.

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 for the covariance between $\eta_{(n)}$ and $\varepsilon_{(n)}$; therefore we are content to simply write $Var(y_{(n)} | \varphi) = (\frac{\sigma^2}{n} + o_p(\frac{1}{n})) I_n$ and we denote this matrix with Σ_n. At this point we are able to specify the prior and sampling probabilities μ and $\hat{P}^{\varphi, \mathbf{w}}$:

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

The sampling probability depends on the sample size and, as n becomes large, $\hat{P}^{\varphi, \mathbf{w}}$ weakly converges to $P^{\varphi, \mathbf{w}}$. As in the basic case, the factor $\frac{1}{n}$ in Σ_n dose not stabilize the inverse of the covariance operator: 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 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 $\frac{1}{n^2 h^q}$. Hence, we need to consider a regularized posterior distribution that, since in this case it employs $\hat{K}_{(n)}$ instead of $K_{(n)}$, we will call *estimated regularized posterior distribution*. It will be denoted with $\hat{\mu}_\alpha^{\mathcal{F}}$ while the predictive will be denoted with $\hat{P}^{\mathbf{w}}$ and they are given respectively by:

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

with

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

Asymptotic properties of the posterior distribution for the case with unknown f are very similar to that one shown in Theorem 14 and in Corollary 4. In fact, the estimation error associated to $\hat{K}_{(n)}$ is negligible with respect to the other terms in the bias and variance.

Theorem 18 *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 a 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 $\beta > 0$, $||\hat{K}_{(n)}^* \hat{K}_{(n)} - K^* K||^2 \sim \mathcal{O}_p(\frac{1}{nh^p} + h^{2\rho})$ 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

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

If the bandwidth h is chosen in such a way to guarantee that the last factor rate is negligible with respect to the first two, the optimal speed of convergence is obtained by equating $\alpha_n^\beta = \frac{1}{\alpha_n^2 n}$, 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 last factor is negligible if $\frac{1}{\alpha^2}(\frac{1}{n} + h^{2\rho}) \sim \mathcal{O}_p(1)$, that implies a choice of the bandwidth such that

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

4.5 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 is

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

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

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

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

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. We present in the next section a data-driven method to select α_n.

We have performed simulations for the conjugate model with known $f(Z, W)$ (CASE I) and for the case with completely unknown $f(Z, W)$ and 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:

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

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 $L_F^2(Z)$. 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 4.1. The first three graphs are drawn for $\varphi_0(Z) =$

$0.95Z^2 + 0.25$ 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 (4.1b) - (4.1c) and (4.1e) - (4.1f) represent drawn from the prior and posterior distribution of φ. In Figure 4.2 we show the histogram of a sample of observations drawn from the prior and the posterior distribution of σ^2.
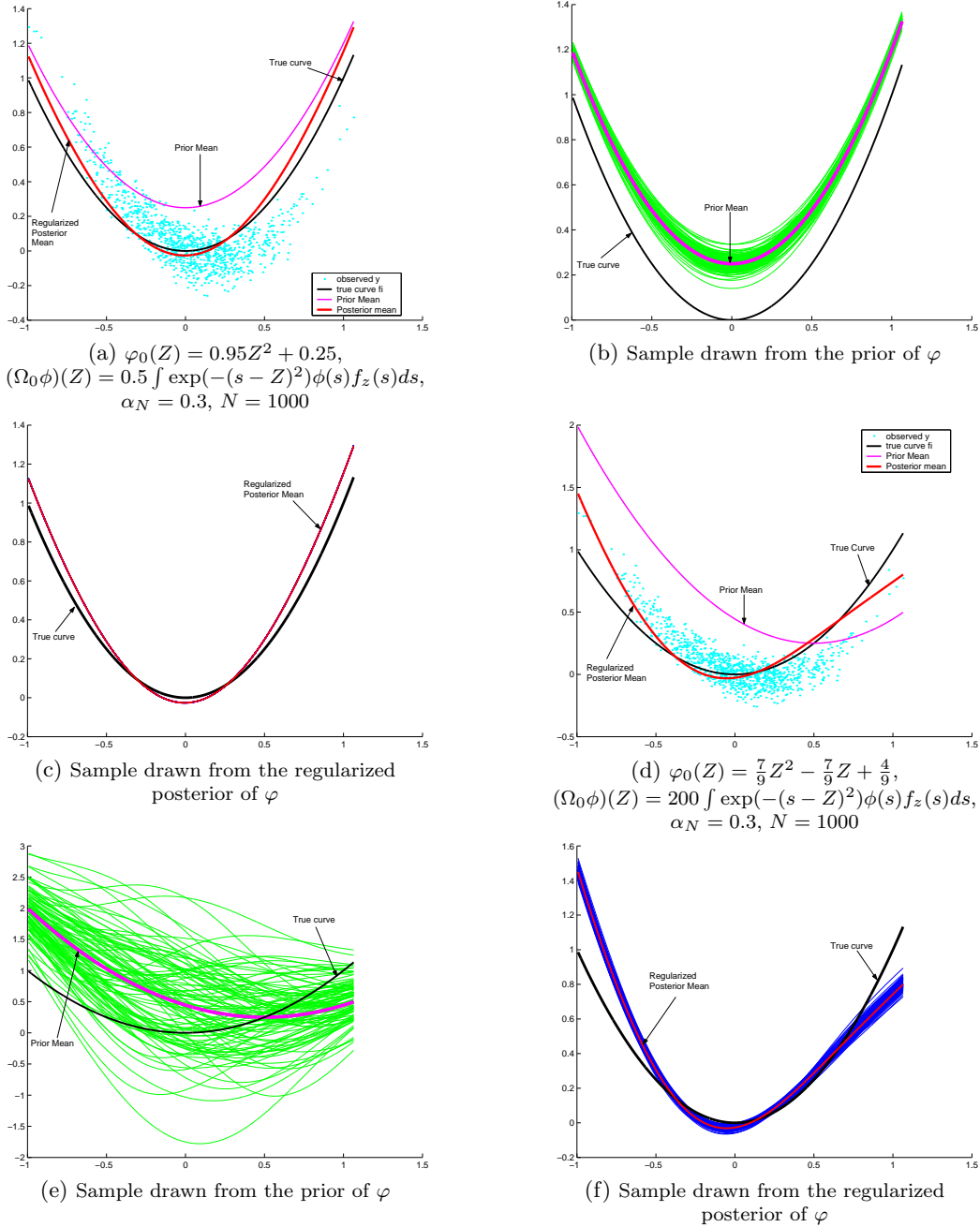

(a) $\varphi_0(Z) = 0.95Z^2 + 0.25$, $(\Omega_0\phi)(Z) = 0.5 \int \exp(-(s - Z)^2)\phi(s)f_z(s)ds$, $\alpha_N = 0.3$, $N = 1000$

(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 4.1: CASE I. *Conjugate Model with $f(Z, W)$ known.* Graphs (4.1a) - (4.1c) are for $\varphi_0(Z) = 0.95Z^2 + 0.25$ and $\sigma_0 = 0.5$; graphs (4.1d) - (4.1f) are for $\frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$ and $\sigma_0 = 200$

CASE II. $f(Z, W)$ *unknown and* σ^2 *known.*
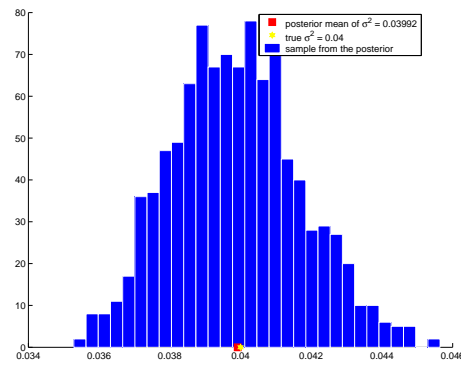In this simulation we have specified a prior only on φ since σ^2 is supposed to be known:
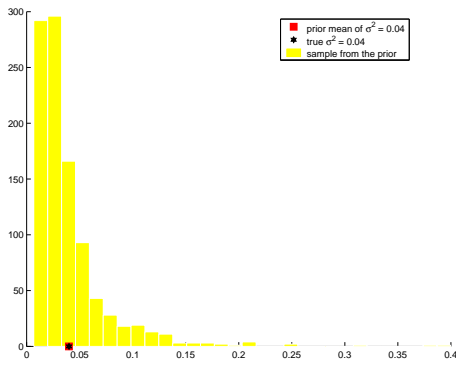
(a) Histogram of a sample drawn from the prior of σ^2
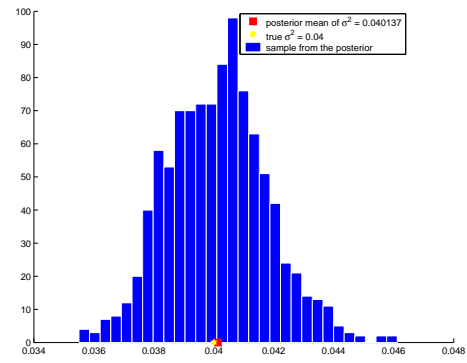
(b) Histogram of a sample drawn from the posterior of σ^2, for $\varphi_0(Z) = 0.95Z^2 + 0.25$ and $\sigma_0 = 0.5$

(c) Histogram of a 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 4.2: CASE I. *Conjugate Model with $f(Z, W)$ known.*

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

with φ_0 and Ω_0 specified as in CASE I. We show in Figure 4.3 only the results for the prior distribution with $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$ and $\sigma_0 = 200$. Panels 4.3a shows the estimated regularized posterior mean, together with the true curve and the prior mean; panel 4.3b reports a sample drawn from the estimated posterior distribution.

(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
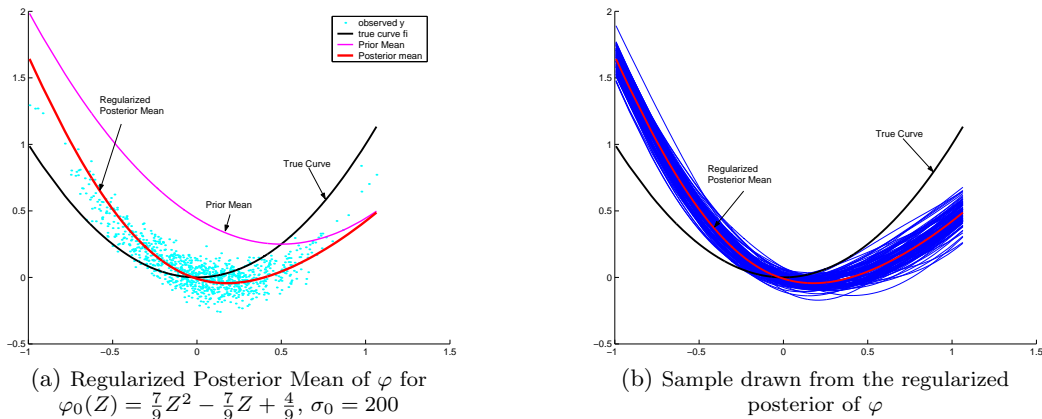
Figure 4.3: CASE II. $f(Z,W)$ *unknown and* σ^2 *known*

4.5.1 Data driven method for choosing α

In regularization of inverse problems theory there exist several a-posteriori parameter choice rules for choosing α_n which depend on the noise level δ in the observed data $y_{(n)}$, with δ such that $||y_{(n)} - K_{(n)}\varphi|| \leq \delta$. In the real world, such noise level information is not always available, therefore it is often advisable to consider alternative parameter choice rules that does not require knowledge of δ. The idea is to select α_n only on the basis of the performance of the regularization method under consideration. This parameter choice technique is widely known and developed in literature and is called *error free*, see for instance Engl *et al.* (2000) [19].

The data-driven method that we apply in this section rests upon a slightly modification of the estimated residuals derived when the regularized posterior mean of φ is used as a punctual estimator of the instrumental regression. This choice for the estimator is dictated by the use of a quadratic loss function. The use of residuals instead of the estimation error $||\hat{\varphi}_\alpha - \varphi_*||$ is justified only if the residuals are adjusted in order to preserve the same speed of convergence as the estimation error. In particular, as it is noted in Engl *et al.* (2000) [19], there exists a relation between the estimation error and the residuals re-scaled by a convenient power of $\frac{1}{\alpha_n}$. Let ν_α denote the residual we are considering, we have to find the value d such that asymptotically

$$\frac{||\nu_\alpha||}{\alpha^d} \sim ||\hat{\varphi}_\alpha - \varphi_*||.$$

Therefore, it seems to make sense to take $\frac{||\nu_\alpha||}{\alpha^d}$ as a measure of the estimation error and to select the optimal α_n as the one which minimizes the ratio:

$$\alpha_n^* = \arg\min \frac{||\nu_\alpha||}{\alpha_n^d}.$$

In the light of this argument, while the classical residual $y_{(n)} - K_{(n)}\hat{\varphi}_\alpha$ would seem the natural choice, it is not acceptable since it does not converge to zero at the good rate. On the contrary, convergence is satisfied by the *projected residuals* defined as

$$\nu_\alpha = \Omega_0^{\frac{1}{2}} K_{(n)}^* y_{(n)} - \Omega_0^{\frac{1}{2}} K_{(n)}^* K_{(n)} \varphi_\alpha$$

that for simplicity we rewrite as $\nu_\alpha = T_{(n)}^* y_{(n)} - T_{(n)}^* K_{(n)} \hat{\varphi}_\alpha$, using notation $T_{(n)}^* = \Omega_0^{\frac{1}{2}} K_{(n)}^*$ and $T_{(n)} = K_{(n)} \Omega_0^{\frac{1}{2}}$.

Nevertheless, even if the ratio is constructed through projected residuals, there is no value of d for which it achieves the same speed of convergence as the estimation error. This is due to the undesirable fact that Tikhonov regularization (that has been used to construct $\hat{\varphi}_\alpha$) can allow to achieve a speed of convergence of at most α^2. Thus, we solve this problem by substituting the Tikhonov regularization scheme with an *iterated Tikhonov* that results in better convergence rate. In our case, it is sufficient to iterate only two times, so that the resulting operator $A_\alpha^{(2)}$, for the conjugate case described in Section 4.3, takes the form:

$$A_\alpha^{(2)} = (\alpha \Omega_0 K_{(n)}^* C_{n,\alpha}^{-1} + \Omega_0 K_{(n)}^*) C_{n,\alpha}^{-1}.$$

We denote with $\hat{\varphi}_\alpha^{(2)}$ the regularized posterior mean obtained by using operator $A_\alpha^{(2)}$ and with $\nu_\alpha^{(2)}$ the corresponding projected residuals.

Lemma 9 *Let $\hat{\varphi}_\alpha^{(2)}$ be the regularized posterior mean obtained through a two-times-iterated Tikhonov scheme in the conjugate case and $\nu_\alpha^{(2)} = T_{(n)}^*(y_{(n)} - K_{(n)} \hat{\varphi}_\alpha^{(2)})$. If $\alpha_n \to 0$, $\alpha_n^2 n \to \infty$, $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$ and $\Omega_0^{-\frac{1}{2}}(\varphi_* - \varphi_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$, then*

$$||\nu_\alpha^{(2)}||^2 \sim \mathcal{O}_p\left(\alpha_n^{\beta+2} + \frac{1}{n}\right).$$

The rate of convergence given in Lemma 9 can be made equivalent, up to negligible terms, to the rate given in Corollary 4 *(i)* by dividing the squared norm of the residual by α_n^2.

Hence, once we have performed estimation for a given sample, we construct the curve $\frac{||\nu_\alpha^{(2)}||^2}{\alpha_n^2}$, as a function of α_n and we select the value for the regularization parameter which minimizes it. The minimization program does not change if we take an increasing transformation of this ratio, for instance we have considered the logarithmic transformation. This simplifies representation of the curve.

Figure 4.4 shows the performance of the data-driven method for the simulation scheme applied to CASE I. In Panels (4.4a) and (4.4c) the log-ratio curve $log\frac{||\nu_\alpha^{(2)}||^2}{\alpha_n^2}$ is plotted against a range of values for α_n in the interval $[0,1]$, for two different choices of the prior specification. For the first specification of the prior it is selected a value $\alpha_n = 0.0285$, while for the second prior a larger value of $\alpha_n = 0.1233$ is selected. In Panels (4.4b) and (4.4d) we show the goodness of our estimation method when the data-driven selected value for α_n is used. We see that the regularized posterior mean is more affected by the data than by the prior mean; this is due to the smaller value selected for α_n with respect to the value we had previously chosen.

A result similar to Lemma 9 can be derived when the density $f(Z, W)$ is unknown and the nonparametric method described in subsection 4.4.2 is applied. In this case we denote $\hat{T}_{(n)} = \hat{K}_{(n)} \Omega_0^{\frac{1}{2}}$ and $\hat{T}_{(n)}^* = \Omega_0^{\frac{1}{2}} \hat{K}_{(n)}^*$ the corresponding estimator for $T_{(n)}$ and $T_{(n)}^*$.
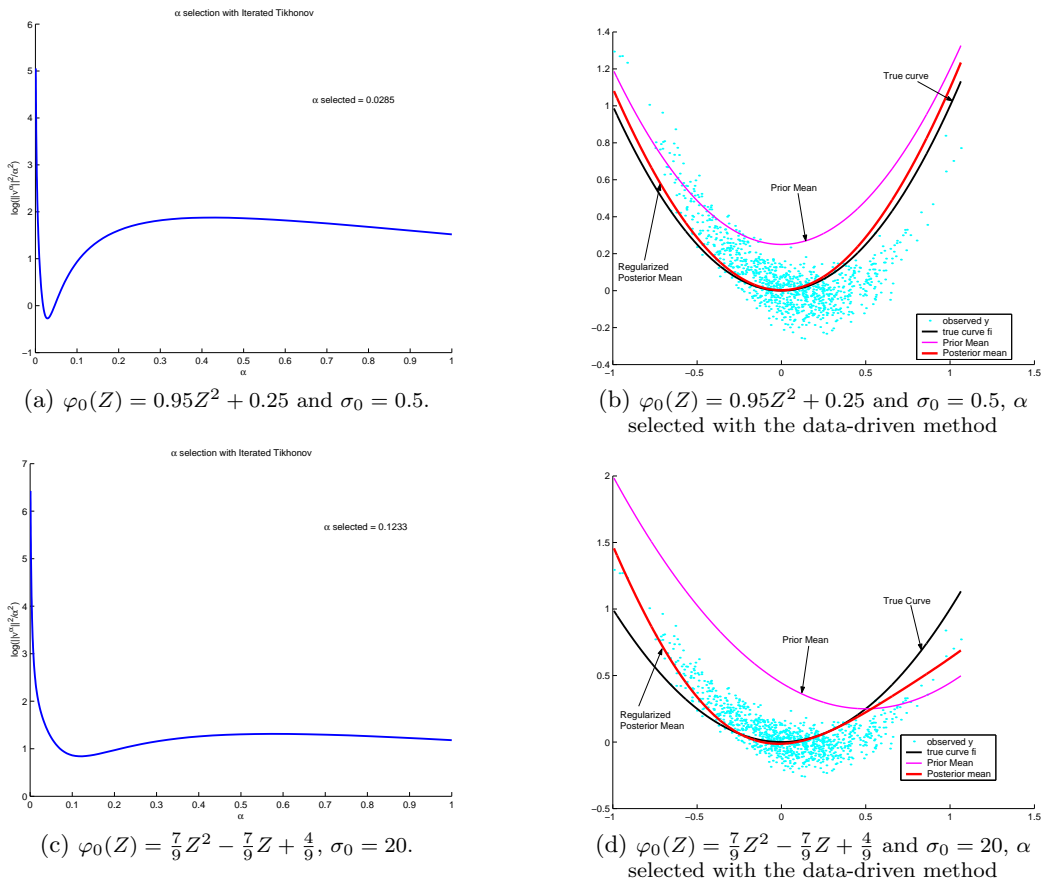
(a) $\varphi_0(Z) = 0.95Z^2 + 0.25$ and $\sigma_0 = 0.5$.

(b) $\varphi_0(Z) = 0.95Z^2 + 0.25$ and $\sigma_0 = 0.5$, α selected with the data-driven method

(c) $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$, $\sigma_0 = 20$.

(d) $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$ and $\sigma_0 = 20$, α selected with the data-driven method

Figure 4.4: Known Operator. Logarithm of the norm of the ratio of the Projected Residuals and α^2: $\log \frac{||\nu_\alpha||^2}{\alpha^2}$ for two different prior specification for φ. In Panel (4.4a) it is selected an $\alpha = 0.0285$; in Panel (4.4c) it is selected an $\alpha = 0.1233$.

We define the *estimated projected residual*, obtained with a two times iterated Tikhonov, as: $\hat{\nu}_\alpha^{(2)} = \hat{T}_{(n)}^*(y_{(n)} - \hat{K}_{(n)}\hat{\mathbb{E}}_\alpha(\varphi|y_{(n)}))$. We obtain the following result.

Lemma 10 *If $\alpha_n \to 0$, $\alpha_n^2 n \to \infty$, $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$, $\Omega_0^{-\frac{1}{2}}(\varphi_* - \varphi_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta \in (0, 2]$, $\|\hat{K}_{(n)}^*\hat{K}_{(n)} - K^*K\|^2 \sim \mathcal{O}_p(\frac{1}{nh^p} + h^{2\rho})$ 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*

$$\|\hat{\nu}_\alpha^{(2)}\|^2 \sim \mathcal{O}_p\left(\alpha_n^{\beta+2} + (\frac{1}{n} + h^{2\rho})(\alpha_n^\beta + \frac{1}{\alpha_n^2}(\frac{1}{n} + h^{2\rho}) + \frac{1}{\alpha_n^2 n}) + \frac{1}{n}\right).$$

It is necessary to re-scale the residual by $\frac{1}{\alpha_n^2}$ to reach the same speed of convergence given in Theorem 18. Figures 4.5a and 4.5c represent the curve $\log\frac{\|\hat{\nu}_\alpha\|^2}{\alpha^2}$ against different values for $\alpha \in [0, 1]$, for two different prior mean specifications. It is indicated the value of α for which the curve reach its minimum and the in Figures 4.5b and 4.5d it is drawn the regularized posterior mean obtained with these selected αs.
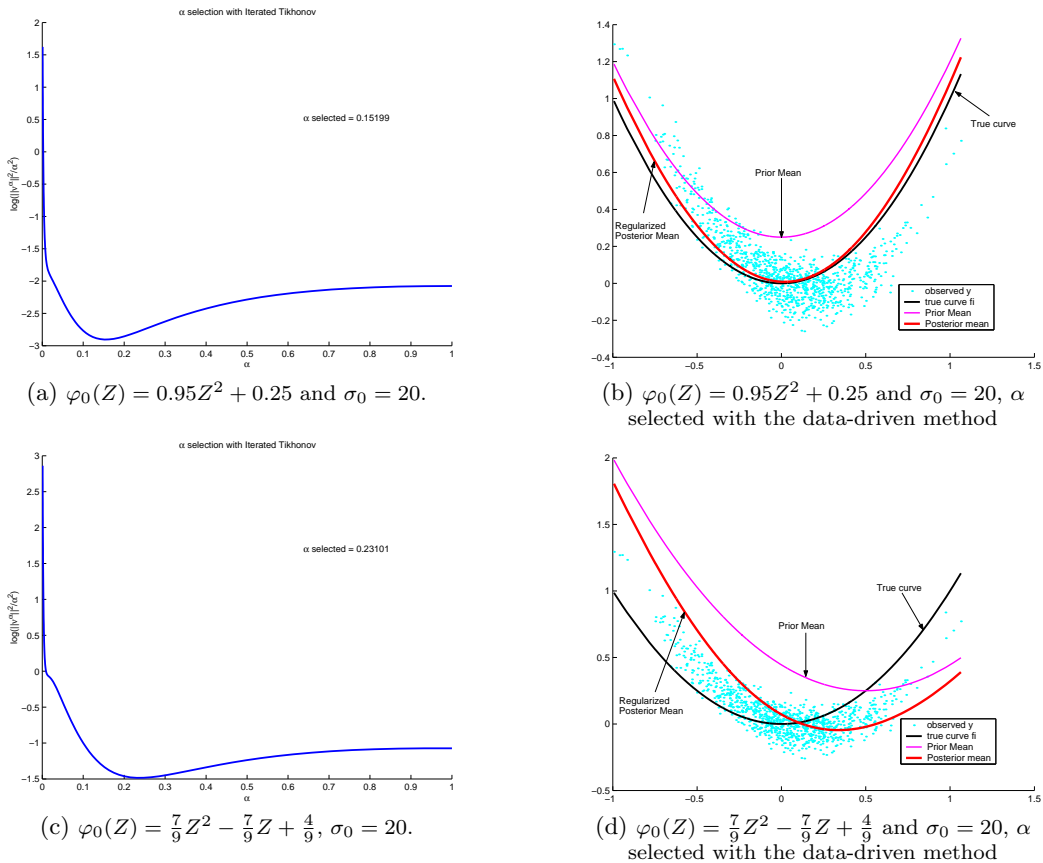

(a) $\varphi_0(Z) = 0.95Z^2 + 0.25$ and $\sigma_0 = 20$.

(b) $\varphi_0(Z) = 0.95Z^2 + 0.25$ and $\sigma_0 = 20$, α selected with the data-driven method

(c) $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$, $\sigma_0 = 20$.

(d) $\varphi_0(Z) = \frac{7}{9}Z^2 - \frac{7}{9}Z + \frac{4}{9}$ and $\sigma_0 = 20$, α selected with the data-driven method

Figure 4.5: Unknown Operator - Kernel estimation. Logarithm of the norm of the ratio of the Projected Residuals and α_n^2: $\log\frac{\|\hat{\nu}_\alpha\|^2}{\alpha^2}$ for two different prior specification for φ. In Panel (4.5a) it is selected an $\alpha = 0.15199$; in Panel (4.5c) it is selected an $\alpha = 0.23101$.

4.6   Conclusions

We have proposed in this chapter 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 available. However, a deeper analysis of the role played by the prior distribution seems to be advisable.

Several possible extensions of our model can be developed. First of all, it would be interesting to consider other regularization methods, different from Tikhonov scheme, and to analyze the way in which the regularized posterior mean is affected. We could also consider Sobolev spaces, instead of general Hilbert spaces, and regularization methods using differential norms.

Lastly, a fully nonparametric Bayesian approach, that uses some kind of Polya tree or Bernstein polynomials prior on density functions, it is noteworthy as an alternative to the classical nonparametric model that we propose.

4.7   Appendix A

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

- (φ_*, σ_*^2) is the true parameter having generated the data;
- $\mathcal{H}(\Omega_0) = \mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}(\Omega_0)$;
- If $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$, we write $(\varphi_* - \varphi) = \Omega_0^{\frac{1}{2}}\psi$, $\psi \in L_F^2(Z)$;
- $T = K\Omega_0^{\frac{1}{2}}$, $T : L_F^2(Z) \to L_F^2(W)$;
- $T_{(n)} = K_{(n)}\Omega_0^{\frac{1}{2}}$, $T_{(n)} : L_F^2(Z) \to \mathbb{R}^n$;
- $\hat{T}_{(n)} = \hat{K}_{(n)}\Omega_0^{-\frac{1}{2}}$, $\hat{T}_{(n)} : L_F^2(Z) \to \mathbb{R}^n$;
- $T^* = \Omega_0^{\frac{1}{2}}K^*$, $T^* : L_F^2(W) \to L_F^2(Z)$;
- $T_{(n)}^* = \Omega_0^{\frac{1}{2}}K_{(n)}^*$, $T_{(n)}^* : \mathbb{R}^n \to L_F^2(Z)$;
- $\hat{T}_{(n)}^* = \Omega_0^{\frac{1}{2}}\hat{K}_{(n)}^*$, $\hat{T}_{(n)}^* : \mathbb{R}^n \to L_F^2(Z)$;
- $\Omega_0^{\frac{1}{2}} = \int \omega_0(s, Z)f(s)ds$;
- $g(Z, w_i) = \int \omega_0(s, Z)\frac{f(s,w_i)}{f(s)f(w_i)}f(s)ds$

Proof of Lemma 6

To clarify the discussion in the following, we will index the posterior distribution with the sample size n, so that $\mu_n^{\sigma,\mathcal{F}}$ will substitute the usual notation $\mu^{\sigma,\mathcal{F}}$. The limits are taken 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

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

for every bounded and continuous functional $a : L_F^2(Z) \to L_F^2(Z)$.

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. Let take, for brevity, null prior mean, $\varphi_0 = 0$, the posterior mean estimator for φ is

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

We are interested in the L_F^2 norm:

$$\|\mathbb{E}(\varphi|y_{(n)}) - \varphi_*\| \leq \overbrace{\|\Omega_0 K_{(n)}^* \Big(\frac{1}{n}I + K_{(n)}\Omega_0 K_{(n)}^*\Big)^{-1} K_{(n)}\varphi_* - \varphi_*\|}^{I}$$
$$+ \underbrace{\|\Omega_0 K_{(n)}^* \Big(\frac{1}{n}I + K_{(n)}\Omega_0 K_{(n)}^*\Big)^{-1} \varepsilon_{(n)}\|}_{II}.$$

If we assume $\varphi_* \in \mathcal{H}(\Omega_0)$ [8], 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

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

obtained by permuting the operators. The term in squared brackets is the regularization bias of the equation $T_{(n)}\gamma = r$ 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 $T_{(n)}^* T_{(n)}$ degenerates towards an infinite rank operator T^*T with unbounded inverse. In particular, by Kolmogorov's Theorem $\|T_{(n)}^* T_{(n)} - T^*T\|^2 \sim \mathcal{O}_p(\delta)$ if $\mathbb{E}(\|T_{(n)}^* T_{(n)} - T^*T\|^2) \sim \mathcal{O}_p(\delta)$, where the expectation is taken with respect to the distribution of w_i. This is the usual MISE and it can be decomposed into the sum of the squared bias and the variance. The bias is zero since $\mathbb{E}(T_{(n)}^* T_{(n)}) - T^*T = 0$, while the variance goes to zero at the speed of $\frac{1}{n}$, so that $\|T_{(n)}^* T_{(n)}\| \sim \mathcal{O}_p(\frac{1}{\sqrt{n}})$. 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.

Proof of Corollary 4

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

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

We start by term I:

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

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

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

that is less than or equal to

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

---

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

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 Carrasco *et al.* (2007) [10]. Therefore, $||IA||^2 \sim \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta)$.

Term $IB = \Omega_0^{\frac{1}{2}}T_{(n)}^*(\alpha_n I + \frac{1}{n}I + T_{(n)}T_{(n)}^*)^{-1}(\frac{1}{n}I)(\alpha_n I + T_{(n)}T_{(n)}^*)^{-1}T_{(n)}\psi$ 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

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

where $T_{(n)}^*\varepsilon_{(n)} = \frac{1}{\sqrt{n}}\left[\frac{1}{\sqrt{n}}\sum_i \varepsilon_i g(Z, w_i)\right]$. By Central Limit Theorem (CLT) the term into squared brackets is bounded because it 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 $||(\alpha_n I + T_{(n)}^*T_{(n)})^{-1}||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n})$ because $T_{(n)}^*T_{(n)}$ 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

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

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):

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

We have to consider the squared norm in L_F^2 of $\Omega_{y,\alpha}\phi$: $||\Omega_{y,\alpha}\phi||^2 \leq |\sigma^2|^2(||A||^2 + ||B||^2)$. 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 15 that $|\sigma^2|^2 \sim \mathcal{O}_p(1)$. Concerning term A we have

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

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, $||B||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^4 n^2}(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta))$. To sum up, $||\Omega_{y,\alpha}||^2 \sim \mathcal{O}_p((1 + \frac{1}{\alpha_n^4 n^2})(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta))$ that, once the fastest terms are neglected, becomes $\mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^2 n}\alpha_n^\beta)$.

Proof of Theorem 14

Both points *(i)* and *(ii)* in the Theorem are consequences of Corollary 4 and Chebishev's Inequality. More clearly, we have

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

and the result follows.

Proof of Theorem 15

The posterior mean $\mathbb{E}(\sigma^2|y_{(n)}) = \frac{s_*^2}{\nu_* - 2}$ is asymptotically equal to

$$
\begin{aligned}
\mathbb{E}(\sigma^2|y_{(n)}) &\approx \underbrace{\frac{1}{n}(y_{(n)} - K_{(n)}\varphi_0)'\Big(\frac{1}{n}I_n + K_{(n)}\Omega_0 K_{(n)}^*\Big)^{-1}(y_{(n)} - K_{(n)}\varphi_0)}_{I} \\
&= \underbrace{\frac{1}{n}(K_{(n)}(\varphi_* - \varphi_0))'\Big(\frac{1}{n}I_n + K_{(n)}\Omega_0 K_{(n)}^*\Big)^{-1}(K_{(n)}(\varphi_* - \varphi_0))}_{} + \\
&\quad \underbrace{\frac{2}{n}(K_{(n)}(\varphi_* - \varphi_0))'\Big(\frac{1}{n}I_n + K_{(n)}\Omega_0 K_{(n)}^*\Big)^{-1}\varepsilon_{(n)}}_{II} + \underbrace{\frac{1}{n}\varepsilon'_{(n)}\Big(\frac{1}{n}I_n + K_{(n)}\Omega_0 K_{(n)}^*\Big)^{-1}(K_{(n)}\varepsilon_{(n)})}_{III}.
\end{aligned}
$$

Under the assumption that $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0) \equiv \mathcal{R}(\Omega_0^{\frac{1}{2}})$, there exists a $\psi \in L_F^2(Z)$ such that $(\varphi_* - \varphi_0) = \Omega_0^{\frac{1}{2}}\psi$, then

$$I = \frac{1}{n} < K_{(n)}\Omega_0^{\frac{1}{2}}\psi, \left(\frac{1}{n}I_n + K_{(n)}\Omega_0 K_{(n)}^*\right)^{-1} K_{(n)}\Omega_0^{\frac{1}{2}}\psi >$$

$$= < \psi, \Omega_0^{\frac{1}{2}} K_{(n)}^* \left(\frac{1}{n}I_n + K_{(n)}\Omega_0 K_{(n)}^*\right)^{-1} K_{(n)}\Omega_0^{\frac{1}{2}}\psi >_{L^2}$$

$$\leq \frac{1}{n}||\psi||_{L^2}||T_{(n)}^* \left(\frac{1}{n}I_n + T_{(n)}T_{(n)}^*\right)^{-1} T_{(n)}||_{L^2}||\psi||_{L^2}$$

$$\sim \mathcal{O}_p\left(\frac{1}{n}\right)$$

since $||T_{(n)}^*(\frac{1}{n}I_n + T_{(n)}T_{(n)}^*)^{-1}T_{(n)}||_{L^2} = ||(\frac{1}{n}I_n + T_{(n)}^*T_{(n)})^{-1}T_{(n)}^*T_{(n)}||$ and it is bounded.

Let notice that $||\varepsilon_{(n)}|| = \sqrt{\frac{1}{n}\sum_i \varepsilon_i^2}$ converges to the true value σ_* and that $||\frac{1}{n}(\frac{1}{n}I_n + T_{(n)}T_{(n)}^*)^{-1}T_{(n)}||_{L^2} = \frac{1}{\sqrt{n}}||\frac{1}{\sqrt{n}}(\frac{1}{n}I_n + T_{(n)}T_{(n)}^*)^{-1}T_{(n)}||_{L^2} = \frac{1}{\sqrt{n}}\mathcal{O}_p(1)$ and then it converges to 0 as $n \to \infty$. Therefore,

$$II = \frac{1}{n} < \varepsilon_{(n)}, \left(\frac{1}{n}I_n + T_{(n)}T_{(n)}^*\right)^{-1} T_{(n)}\psi >$$

$$\leq ||\varepsilon_{(n)}||\left|\left|\frac{1}{n}\left(\frac{1}{n}I_n + T_{(n)}T_{(n)}^*\right)^{-1} T_{(n)}\right|\right|||\psi||$$

$$\sim \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right).$$

Third term requires a little bit more computations. First, we recall that, by Binomial Inverse Theorem, $(\frac{1}{n}I_n + T_{(n)}T_{(n)}^*)^{-1} = nI_n - n^2 T_{(n)}(I_{L^2} + nT_{(n)}^*T_{(n)})^{-1}T_{(n)}^*$, where I_{L^2} denotes the identity operator in L_F^2. Hence,

$$III = \varepsilon_{(n)}'\varepsilon_{(n)} - n\varepsilon_{(n)}'T_{(n)}(I_{L^2} + nT_{(n)}^*T_{(n)})^{-1}T_{(n)}^*\varepsilon_{(n)} \qquad (4.14)$$

Moreover, it is easy to see that

$$\varepsilon_{(n)}'\varepsilon_{(n)} \to \sigma_*^2$$

$$\hat{T}^*\varepsilon_{(n)} = \frac{1}{n}\sum_i \varepsilon_i g(Z, w_i)$$

$$n(I_{L^2} + nT^*T_{(n)})^{-1}T_{(n)}^*\varepsilon_{(n)} = \frac{1}{n}\sum_i \varepsilon_i\left((\frac{1}{n}I_{L^2} + T_{(n)}^*T_{(n)})^{-1}g(Z, w_i)\right).$$

The second term in (4.14) becomes

$$n\varepsilon_{(n)}'T_{(n)}(I_{L^2} + nT_{(n)}^*T_{(n)})^{-1}T_{(n)}'\varepsilon_{(n)} = < T_{(n)}^*\varepsilon_{(n)}, \left(\frac{1}{n}I_{L^2} + T_{(n)}^*T_{(n)}\right)^{-1}T_{(n)}^*\varepsilon_{(n)} >$$

$$\leq ||T_{(n)}^*\varepsilon_{(n)}||\left|\left|\left(\frac{1}{n}I_{L^2} + T_{(n)}^*T_{(n)}\right)^{-1}T_{(n)}^*\varepsilon_{(n)}\right|\right|.$$

The first norm is an $\mathcal{O}_p(\frac{1}{\sqrt{n}})$ since $||T_{(n)}^*\varepsilon_{(n)}|| = \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{n}}\sum_i \varepsilon_i||g(Z, w_i)||_{L^2}\right)$ and the factor in brackets is bounded in probability because it converges to a normal random variable (by the CLT). If $g(Z, w_i) \in \mathcal{R}(T^*T)^{\frac{\gamma}{2}}$, for some $\gamma > 0$, then there exists a function $h(Z, w_i) \in L_F^2$ such that $g = (T^*T)^{\frac{\gamma}{2}}h(Z, w_i)$ and hence

$$||\left(\frac{1}{n}I_{L^2} + T_{(n)}^*T_{(n)}\right)^{-1}T_{(n)}^*\varepsilon_{(n)}|| = ||\frac{1}{n}\sum_i \varepsilon_i\left((\frac{1}{n}I_{L^2} + T^*T)^{-1}(T^*T)^{\frac{\gamma}{2}}h(Z, w_i)\right)|| \qquad (4.15)$$

$$+||\frac{1}{n}\sum_i \varepsilon_i\left([(\frac{1}{n}I_{L^2} + T_{(n)}^*T_{(n)})^{-1} - (\frac{1}{n}I_{L^2} + T^*T)^{-1}]g(Z, w_i)\right)||$$

The first norm in the left hand side is

$$||\frac{1}{n}\sum_i \varepsilon_i\Big((\frac{1}{n}I_{L^2}+T^*T)^{-1}(T^*T)^{\frac{\gamma}{2}}h(Z,w_i)\Big)|| \quad \leq \quad \frac{n}{n}\sum_i |\varepsilon_i|\underbrace{\Big|\Big|\frac{1}{n}(\frac{1}{n}I_{L^2}+T^*T)^{-1}(T^*T)^{\frac{\gamma}{2}}\Big|\Big|}_{\sim \mathcal{O}_p(n^{-\frac{\gamma}{2}})}||h(Z,w_i)||$$

$$= \quad \mathcal{O}_p(n^{1-\frac{\gamma}{2}})\frac{1}{n}\sum_i |\varepsilon_i|||h(Z,w_i)||$$

that is bounded since ε_i is absolutely integrable.

The second norm in (4.15), $||\frac{1}{n}\sum_i \varepsilon_i\Big([(\frac{1}{n}I_{L^2}+T^*_{(n)}T_{(n)})^{-1}-(\frac{1}{n}I_{L^2}+T^*T)^{-1}]g(Z,w_i)\Big)||$, can be developed as

$$||\frac{1}{n}\sum_i \varepsilon_i\Big([(\frac{1}{n}I_{L^2}+T^*_{(n)}T_{(n)})^{-1}(T^*T-T^*_{(n)}T_{(n)})(\frac{1}{n}I_{L^2}+T^*T)^{-1}](T^*T)^{\frac{\gamma}{2}}h(Z,w_i)\Big)||$$

that is an $\mathcal{O}_p(n^{1-\frac{\gamma}{2}})$. Finally, $n\varepsilon'_{(n)}\hat{T}(I_{L^2}+nT^*_{(n)}T_{(n)})^{-1}T^*_{(n)}\varepsilon_{(n)} \sim \mathcal{O}_p(n^{\frac{1-\gamma}{2}})$ and it goes to 0 if $\gamma > 1$. Therefore, by eliminating negligible terms,

$$\mathbb{E}(\sigma^2|y_{(n)}) - \sigma^2_* \sim \mathcal{O}_p(\frac{1}{\sqrt{n}}\Big(\frac{1}{n}\Big)^{\frac{\gamma-1}{2}}).$$

Proof of Theorem 8

Note that

$$
\begin{aligned}
||(\sigma^2,\varphi)-(\sigma^2_*,\varphi_*)||_{\mathbb{R}_+ \times L^2_F(Z)} &= ||(\sigma^2-\sigma^2_*,\varphi-\varphi_*)||_{\mathbb{R}_+ \times L^2_F(Z)} \\
&= \sqrt{<(\sigma^2-\sigma^2_*,\varphi-\varphi_*)>_{\mathbb{R}_+ \times L^2_F(Z)}} \\
&= \sqrt{<(\sigma^2-\sigma^2_*),(\sigma^2-\sigma^2_* >_{\mathbb{R}_+} + <(\varphi-\varphi_*),(\varphi-\varphi_*)>_{L^2_F(Z)}} \\
&= (||\sigma^2-\sigma^2_*||^2_{\mathbb{R}_+}+||\varphi-\varphi_*||^2_{L^2_F(Z)})^{\frac{1}{2}} \\
&\leq ((||\sigma^2-\sigma^2_*||_{\mathbb{R}_+}+||\varphi-\varphi_*||_{L^2_F(Z)})^2)^{\frac{1}{2}} \\
&= ||\sigma^2-\sigma^2_*||_{\mathbb{R}_+}+||\varphi-\varphi_*||_{L^2_F(Z)}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\nu^{\mathcal{F}}\times\mu^{\sigma,\mathcal{F}}\{(\sigma^2,\varphi)\in\mathbb{R}_+\times L^2_F(Z),||(\sigma^2,\varphi)-(\sigma^2_*,\varphi_*)||_{\mathbb{R}_+\times L^2_F(Z)}>\varepsilon\} \\
&\leq \nu^{\mathcal{F}}\times\mu^{\sigma,\mathcal{F}}\{(\sigma^2,\varphi)\in\mathbb{R}_+\times L^2_F(Z),||\sigma^2-\sigma^2_*||_{\mathbb{R}_+}+||\varphi-\varphi_*||_{L^2_F(Z)}>\varepsilon\} \\
&= \mathbb{E}^{\nu^{\mathcal{F}}}(\mu^{\sigma,\mathcal{F}}\{||\varphi-\varphi_*||_{L^2_F(Z)}>\varepsilon-||\sigma^2-\sigma^2_*||_{\mathbb{R}_+}\}),
\end{aligned}
$$

where $\mathbb{E}^{\nu^{\mathcal{F}}}$ denotes the mean taken with respect to the posterior distribution of σ^2. Since $\mu^{\sigma,\mathcal{F}}$ is a bounded and continuous function of σ^2, by definition of weak convergence of a probability measure and by Theorem 15, this expectation converges in $\mathbb{R}$-norm toward

$$\mu^{\sigma_*,\mathcal{F}}\{||\varphi-\varphi_*||_{L^2_F(Z)}>\varepsilon\}$$

that in turn converges to 0 by Theorem 14.

Proof of Theorem 16

We start by decomposing the estimation error in four 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 (4.10) 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 14. We shall show that the other two terms are converging at a faster speed and then are negligible.

We start with the second term. Note that $\mathbb{E}_\alpha(\varphi|y_{(n)}, \mathbf{w}) = \int \hat{\varphi}_\alpha^\theta \rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}}) d\theta$, then

$$
\begin{aligned}
||\mathbb{E}_\alpha(\varphi|y_{(n)}, \mathbf{w}) - \hat{\varphi}_\alpha^{\hat{\theta}}||^2 &= \int \left(\int (\hat{\varphi}_\alpha^\theta(Z) - \hat{\varphi}_\alpha^{\hat{\theta}}(Z)\rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}}) d\theta \right)^2 f(Z, \cdot|\theta_*) dZ \\
&\leq \int \left(\hat{\varphi}_\alpha^\theta(Z) - \hat{\varphi}_\alpha^{\hat{\theta}}(Z) \right)^2 \rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}}) d\theta \, f(Z, \cdot|\theta_*) dZ \\
&\approx tr Var(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}}) \int \left(\frac{\partial \hat{\varphi}_\alpha^{\hat{\theta}}}{\partial \theta} \frac{\partial \hat{\varphi}_\alpha^{\hat{\theta}}}{\partial \theta'} \right)(Z) f(Z, \cdot|\theta_*) dZ \\
&\sim \mathcal{O}_p\left(\frac{1}{\tilde{n}} \right)
\end{aligned}
$$

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. Bernstein (1934) [5], Gosh and Ramamoorthi (2003) [36] or Von Mises (1964) [79]) show that, under some regularity condition that we assume to be satisfied, $\sqrt{N}||\hat{\theta} - \theta_*|| \sim \mathcal{O}_p(1)$. This 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.

Proof of Theorem 17

In order to show convergence to 0 of $\hat{\Omega}_{y,\alpha}$ we decompose it in different terms and study each of them separately. Let $\phi \in L_F^2(Z)$ be 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 $\beta > 0$, then

$$
\begin{aligned}
||\hat{\Omega}_{y,\alpha}\phi||^2 &\leq ||\hat{\Omega}_{y,\alpha}\phi - Var_\alpha(\varphi|y_{(n)}, \mathbf{w})\phi||^2 \\
&\quad + ||[\int Var_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w})\rho(\theta|(s_{2,i})_{i=1,\ldots,n}) - Var_\alpha(\varphi|\hat{\theta}, y_{(n)}, \mathbf{w})]\phi||^2 \\
&\quad + ||Var_\alpha(\varphi|\hat{\theta}, (y_{(n)}, \mathbf{w}))\phi - Var_\alpha(\varphi|\theta_*, (y_{(n)}, \mathbf{w}))\phi||^2 + ||\Omega_{y,\alpha}(\theta_*)\phi||^2 \\
&\quad + ||\frac{1}{J}\sum_{j=1}^J (\hat{\varphi}_\alpha^{(j)})^2 - \left(\frac{1}{J}\sum_{j=1}^J \hat{\varphi}_\alpha \right)^2 (\zeta)||^2 \qquad (4.16)
\end{aligned}
$$

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 it is negligible assuming that we are taking a large number of discretization points drawn from $\rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}})$. For simplicity, we rewrite $Var_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w})$ as $\Omega_{y,\alpha}(\theta)$, thus the second error term becomes:

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

that is equal to

$$\int \left(\int [\Omega_{y,\alpha}(\theta)\phi - \Omega_{y,\alpha}(\hat{\theta})\phi](\zeta)\rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}})d\theta\right)^2 f(\zeta, \cdot|\theta_*)d\zeta$$

$$\leq \int \int [\Omega_{y,\alpha}(\theta)\phi - \Omega_{y,\alpha}(\hat{\theta})\phi]^2(\zeta)\rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}})d\theta\, f(\zeta, \cdot|\theta_*)d\zeta$$

$$\approx tr Var(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}}) \int \frac{\partial \Omega_{y,\alpha}(\hat{\theta})\phi}{\partial \theta} \frac{\partial \Omega_{y,\alpha}(\hat{\theta})\phi}{\partial \theta'}(\zeta)f(\zeta, \cdot|\theta_*)d\zeta$$

$$\sim \mathcal{O}_p\left(\frac{1}{\tilde{n}}\right) \qquad \text{if } \frac{\partial \Omega_{y,\alpha}(\hat{\theta})\phi}{\partial \theta} \in L_F^2(Z).$$

Using the same notation as before the third error term is

$$||\Omega_{y,\alpha}(\hat{\theta})\phi - \Omega_{y,\alpha}(\theta_*)\phi||^2 = \int [(\Omega_{y,\alpha}(\hat{\theta})\phi)(\zeta) - (\Omega_{y,\alpha}(\theta_*)\phi)(\zeta)]^2 f(\zeta, \cdot|\theta_*)d\zeta$$

$$\approx tr(\hat{\theta} - \theta_*)(\hat{\theta} - \theta_*)' \int \frac{\partial(\Omega_{y,\alpha}(\theta_*)\phi)(\zeta)}{\partial \theta} \frac{\partial(\Omega_{y,\alpha}(\theta_*)\phi)(\zeta)}{\partial \theta'} f(\zeta, \cdot|\theta_*)$$

$$\sim \mathcal{O}_p\left(\frac{1}{n}\right) \qquad \text{if } \frac{\partial \Omega_{y,\alpha}(\theta_*)\phi}{\partial \theta} \in L_F^2(Z).$$

It should be noticed that all the approximated equalities in previous terms are obtained thanks to a first order Taylor expansion.

Consider the last norm of (4.16):

$$||\widehat{Var}(\mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}}))||^2 \leq ||\widehat{Var}(\mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}})) - Var(\mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}}))||^2$$
$$||Var(\mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}}))||^2,$$

where the first term of the decomposition is the Monte Carlo approximation error and it is negligible. By using the notation: $\hat{\varphi}_\alpha^\theta := \mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}}) = \mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w})$ and $\hat{\varphi}_\alpha := \int \hat{\varphi}_\alpha^\theta \rho(\theta|\tilde{\mathbf{z}}, \tilde{\mathbf{w}})d\theta$, we can rewrite the last norm as

$$||\int \left(\hat{\varphi}_\alpha^\theta - \hat{\varphi}_\alpha\right)^2 \rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}})d\theta||^2$$

$$= ||\int \left(\hat{\varphi}_\alpha^\theta \pm \hat{\varphi}_\alpha^{\hat{\theta}} - \hat{\varphi}_\alpha\right)^2 \rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}})d\theta||^2$$

$$\leq ||\int \left[(\hat{\varphi}_\alpha^\theta - \hat{\varphi}_\alpha^{\hat{\theta}})^2 + (\hat{\varphi}_\alpha^{\hat{\theta}} - \hat{\varphi}_\alpha)^2\right]\rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}})d\theta||^2$$

since $2(\hat{\varphi}_\alpha^\theta - \hat{\varphi}_\alpha^{\hat{\theta}})(\hat{\varphi}_\alpha^{\hat{\theta}} - \hat{\varphi}_\alpha) = -2(\hat{\varphi}_\alpha^{\hat{\theta}} - \hat{\varphi}_\alpha)$. Therefore,

$$||Var(\mathbb{E}_\alpha(\varphi|\theta, y_{(n)}, \mathbf{w}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}}))||^2 \leq ||\int (\hat{\varphi}_\alpha^\theta - \hat{\varphi}_\alpha^{\hat{\theta}})^2 \rho(\theta|(s_{2,i})_{i=1,\ldots,\tilde{n}})d\theta||^2 + ||(\hat{\varphi}_\alpha - \hat{\varphi}_\alpha^{\hat{\theta}})^2||^2.$$

Proof of Theorem 16 shows that these norms are $\mathcal{O}_p(\frac{1}{\tilde{n}})$.

Therefore, all the error terms in (4.16) 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 14 and this proves the result.

Proof of Theorem 18

The proof is analogous to that one for Theorem 14 with the regularized posterior distribution replaced by its estimated version. Therefore, we limit ourselves to recover the speed of convergence of the estimated regularized posterior mean and variance. First, we use the following decomposition:

$$
\begin{aligned}
\hat{\mathbb{E}}_\alpha(\varphi|y_{(n)}) - \varphi_* = \ &\overbrace{-(I - \Omega_0\hat{K}_{(n)}^*(\alpha_n I + \hat{K}\Omega_0\hat{K}_{(n)}^*)^{-1}\hat{K}_{(n)})(\varphi_* - \varphi_0)}^{I} \\
&+ \underbrace{\Omega_0\hat{K}_{(n)}[(\alpha_n I + \Sigma_n + \hat{K}\Omega_0\hat{K}_{(n)}^*)^{-1} - (\alpha_n I + \hat{K}\Omega_0\hat{K}_{(n)}^*)^{-1}]\hat{K}_{(n)}(\varphi_* - \varphi_0)}_{II} \\
&+ \underbrace{\Omega_0\hat{K}_{(n)}^*(\alpha_n I + \Sigma_n\hat{K}\Omega_0\hat{K}_{(n)}^*)^{-1}(\eta_{(n)} + \varepsilon_{(n)})}_{III}
\end{aligned}
$$

As usual, we assume $(\varphi_* - \varphi_0) \in \mathcal{H}(\Omega_0)$, then $(\varphi_* - \varphi_0) = \Omega_0^{\frac{1}{2}}\psi$. Hence,

$$
\begin{aligned}
||I||^2 &\leq ||\Omega_0^{\frac{1}{2}}||^2||I - \hat{T}_{(n)}^*(\alpha_n I + \hat{T}_{(n)}\hat{T}_{(n)}^*)-1\hat{T}_{(n)}||^2||\psi||^2 \\
&= ||\Omega_0^{\frac{1}{2}}||^2||I - (\alpha_n I + \hat{T}_{(n)}^*\hat{T}_{(n)})^{-1}\hat{T}_{(n)}^*\hat{T}_{(n)}||^2||\psi||^2 \\
&\leq \underbrace{||\Omega_0^{\frac{1}{2}}||^2}_{\sim\mathcal{O}_p(1)}(\underbrace{||\alpha_n(\alpha_n I + T^*T)^{-1}||^2}_{\sim\mathcal{O}_p(\alpha_n^\beta)} + \underbrace{||\alpha_n(\alpha_n I + \hat{T}_{(n)}^*\hat{T}_{(n)})^{-1}(T^*T - \hat{T}_{(n)}^*\hat{T}_{(n)})(\alpha_n I + T^*T)^{-1}||^2}_{\sim\mathcal{O}_p(\alpha_n^{\beta-2}||\hat{T}_{(n)}^*\hat{T}_{(n)} - T^*T||^2)})||\psi||^2
\end{aligned}
$$

where the power β is found under the assumption $\Omega_0^{\frac{1}{2}}(\varphi_* - \varphi_0) \in \mathcal{R}(T^*T)^{\frac{\beta}{2}}$. Let consider term II:

$$
\begin{aligned}
||II||^2 &\leq ||\Omega_0^{\frac{1}{2}}||^2||\hat{T}_{(n)}^*(\alpha_n I + \frac{\sigma^2}{n}I + \hat{T}_{(n)}\hat{T}_{(n)}^*)^{-1}(-\frac{\sigma^2}{n}I)(\alpha_n I + \hat{T}_{(n)}\hat{T}_{(n)}^*)^{-1}\hat{T}_{(n)}||^2||\psi||^2 \\
&\leq ||\Omega_0^{\frac{1}{2}}||^2||(\alpha_n I + \frac{\sigma^2}{n}I + \hat{T}_{(n)}^*\hat{T}_{(n)})^{-1}||^2||\frac{\sigma^2}{n}I||^2||(\alpha_n I + \hat{T}_{(n)}\hat{T}_{(n)}^*)^{-1}\hat{T}_{(n)}^*\hat{T}_{(n)}||^2||\psi||^2 \\
&\simeq ||\Omega_0^{\frac{1}{2}}||^2||(\alpha_n I + \frac{\sigma^2}{n}I + T^*T)^{-1} + (\alpha_n I + \frac{\sigma^2}{n}I + T^*T)^{-2}(\hat{T}_{(n)}^*\hat{T}_{(n)} - T^*T)||^2||\frac{\sigma^2}{n}I||^2 \\
&\quad ||(\alpha_n I + T^*T)^{-1}T^*T + [(\alpha_n I + T^*T)^{-1} + (\alpha_n I + T^*T)^{-2}T^*T](\hat{T}_{(n)}^*\hat{T}_{(n)} - T^*T)||^2||\psi||^2 \\
&\sim \mathcal{O}_p\Big((\frac{1}{\alpha_n^2 n^2} + \frac{1}{\alpha_n^4 n^2}||\hat{T}_{(n)}^*\hat{T}_{(n)} - T^*T||^2)(1 + \frac{1}{\alpha_n^2}||\hat{T}_{(n)}^*\hat{T}_{(n)} - T^*T||^2)\Big),
\end{aligned}
$$

where the third approximated equality follows from a first order Taylor expansion around the true value of operator T^*T. Lastly, term III can be rewritten as

$$
\begin{aligned}
||III||^2 &\leq ||\Omega_0^{\frac{1}{2}}||^2\Big(\overbrace{||\hat{T}_{(n)}^*(\alpha_n I + \hat{T}_{(n)}\hat{T}_{(n)}^*)^{-1}(\eta_{(n)} + \varepsilon_{(n)})||^2}^{IIIA} + \\
&\quad \underbrace{||\hat{T}_{(n)}^*(\alpha_n I + \frac{\sigma^2}{n}I + \hat{T}_{(n)}\hat{T}_{(n)}^*)^{-1}(-\frac{\sigma^2}{n}I)(\alpha_n I + \hat{T}_{(n)}\hat{T}_{(n)}^*)^{-1}(\eta_{(n)} + \varepsilon_{(n)})||^2}_{IIIB}\Big);
\end{aligned}
$$

$$
\begin{aligned}
||IIIA||^2 &\simeq ||(\alpha_n I + T^*T)^{-1} - (\alpha_n I + T^*T)^{-2}(\hat{T}_{(n)}^*\hat{T}_{(n)} - T^*T)||^2||\Omega_0^{\frac{1}{2}}||^2||\hat{K}_{(n)}^*(\eta_{(n)} + \varepsilon_{(n)})||^2 \\
&\sim \mathcal{O}_p(\frac{1}{\alpha_n^2 n} + \frac{1}{\alpha_n^4 n}||\hat{T}_{(n)}^*\hat{T}_{(n)} - T^*T||^2)
\end{aligned}
$$

$$
||IIIB||^2 \sim \mathcal{O}_p\Big((\frac{1}{\alpha_n^2 n^2} + \frac{1}{\alpha_n^4 n^2}||\hat{T}_{(n)}^*\hat{T}_{(n)} - T^*T||^2)(\frac{1}{\alpha_n^2 n} + \frac{1}{\alpha_n^4 n}||\hat{T}_{(n)}^*\hat{T}_{(n)} - T^*T||^2)\Big).
$$

The rates of $IIIA$ and $IIIB$ have been obtained through a first order Taylor expansion. The last thing we need to prove is that $||\hat{T}^*_{(n)}\hat{T}_{(n)} - T^*T||^2 \sim \mathcal{O}_p(\frac{1}{n} + h^{2\rho})$. This is an easy task if we note that $\hat{K}^*_{(n)}\hat{K}_{(n)}\varphi$ has the same asymptotic behavior of $\int\int \varphi(z)\hat{f}(z|w_i)dz\frac{\hat{f}(z,w_i)}{\hat{f}(z)}dw_i$ that is the operator $T^*_{\hat{F}}T_{\hat{F}}$ defined in Darolles *et al.* (2006) [15]. We use their result (that they prove in the Appendix B): $||T^*_{\hat{F}}T_{\hat{F}} - T^*_FT_F||^2 = \mathcal{O}_p(\frac{1}{nh^p} + h^{2\rho})$ and it follows that $\hat{K}^*_{(n)}\hat{K}_{(n)}$ is of the same order. Then, we smooth by applying the integral operator $\Omega_0^{\frac{1}{2}}$ to the first order Taylor expansion $T^*_{\hat{F}}T_{\hat{F}} - T^*_FT_F = (T^*_{\hat{F}} - T^*)T + T^*(T_{\hat{F}} - T)$. We compute the squared bias and the variance of the last two terms as described in Darolles *et al.* (2006) [15]; the smoothing effect acts only on the variance (that is now of order $\frac{1}{n}$) and not on the squared bias (that remains equal to $h^{2\rho}$). Henceforth, after having deleting the negligible terms we get: $||I||^2 \sim \mathcal{O}_p(\alpha_n^\beta + \alpha_n^{\beta-2}(\frac{1}{n} + h^{2\rho}))$, $||II||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^2 n^2} + \frac{1}{\alpha_n^4 n^2}(\frac{1}{n} + h^{2\rho}))$ and $||III||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n^2 n} + \frac{1}{\alpha_n^4 n}(\frac{1}{n} + h^{2\rho}))$ since term $IIIB$ is negligible with respect to $IIIA$. Finally, simplification of the redundant and negligible terms in $||I||^2$ establishes the result.

Proof of Lemma 9

Let $R^\alpha = (\alpha I_n + T_{(n)}T^*_{(n)})^{-1}$ and $R^\alpha = (\alpha I_n + \frac{1}{n}I + T_{(n)}T^*_{(n)})^{-1}$. We decompose the residual as

$$
\begin{aligned}
\nu_{\alpha^{(2)}} &= \overbrace{T^*_{(n)}[I - (\alpha K_{(n)}\Omega_0 K^*_{(n)}R^\alpha + K_{(n)}\Omega_0 K^*_{(n)})R^\alpha]K_{(n)}(\varphi_* - \varphi_0)}^{I} \\
&+ \overbrace{T^*_{(n)}[(\alpha K_{(n)}\Omega_0 K^*_{(n)}R^\alpha + K_{(n)}\Omega_0 K^*_{(n)})R^\alpha - (\alpha K_{(n)}\Omega_0 K^*_{(n)}R^\alpha_{(n)} + K_{(n)}\Omega_0 K^*_{(n)})R^\alpha_{(n)}]K_{(n)}(\varphi_* - \varphi_0)}^{II} \\
&+ \overbrace{T^*_{(n)}[I - (\alpha K_{(n)}\Omega_0 K^*_{(n)}R^\alpha + K_{(n)}\Omega_0 K^*_{(n)})R^\alpha]\varepsilon_{(n)}}^{III} \\
&+ \underbrace{T^*_{(n)}[(\alpha K_{(n)}\Omega_0 K^*_{(n)}R^\alpha + K_{(n)}\Omega_0 K^*_{(n)})R^\alpha - (\alpha K_{(n)}\Omega_0 K^*_{(n)}R^\alpha_{(n)} + K_{(n)}\Omega_0 K^*_{(n)})R^\alpha_{(n)}]\varepsilon_{(n)}}_{IV} \, .
\end{aligned}
$$

Standard computations similar to those one used in previous proof allows to show that: $||I||^2 \sim \mathcal{O}_p(\alpha^{\beta+2} + \frac{1}{n})$, $||II||^2 \sim \mathcal{O}_p(\frac{1}{n^2} + \frac{1}{\alpha^2 n^2} + \frac{\alpha^2}{n})$, $||III||^2 \sim \mathcal{O}_p(\frac{1}{n} + \frac{1}{\alpha^2 n^2})$, $||IV||^2 \sim \mathcal{O}_p(\frac{1}{\alpha^2 n^3} + \frac{1}{\alpha^4 n^3})$.

Proof of Lemma 10

The same as the Proof of Lemma 9 with $T_{(n)}$, $T^*_{(n)}$, $K_{(n)}$ and $K^*_{(n)}$ replaced by $\hat{T}_{(n)}$, $\hat{T}^*_{(n)}$, $\hat{K}_{(n)}$ and $\hat{K}^*_{(n)}$. Then we have the same decomposition and we get: $||I||^2 \sim \mathcal{O}_p(\alpha^{\beta+2} + (\frac{1}{n} + h^{2\rho}))$, $||II||^2 \sim \mathcal{O}_p(\alpha^{\beta+2} + (\frac{1}{n} + h^{2\rho})\alpha^\beta)$, $||III||^2 \sim \mathcal{O}_p(\frac{1}{\alpha^2 n}(\frac{1}{n} + h^{2\rho}))$, $||IV||^2 \sim \mathcal{O}_p(\frac{1}{n} + \frac{1}{\alpha^2 n}(\frac{1}{n} + h^{2\rho}))$.

Chapter 5

Bayesian Nonparametric Estimation of Asset Pricing Functionals[1]

Abstract

We recover the posterior distribution of the equilibrium asset pricing functional p in a completely nonparametric way. We consider rational expectation models for assets pricing as in Lucas (1978), where the pricing functional p is a function of a vector of n state variables and is characterized as the solution of an integral equation of second kind, stated in an Hilbert space. We adopt a Bayesian procedure since it allows to incorporate all the prior information we have and this is particular useful in nonparametric estimation. Moreover, a Bayesian estimation mimics the Bayesian learning process of economic agents that leads to form rational expectations.

Integral equations of second kind are well posed inverse problems, but the use of a Bayesian approach for solving them introduces ill-posedness. Therefore, the posterior distribution of p is inconsistent, due to non-continuity of its posterior mean. The contribution of this paper is to propose two consistent estimators for the pricing functionals. The first one is a regularized posterior distribution and the second estimator is the posterior distribution obtained through a prior distribution of the g-prior type, like in Zellner (1986), that we show is able to get rid of the ill-posedness in the posterior distribution.

These two estimators allow to benefit from the advantages of being Bayesian without suffering of the drawbacks that we had with the usual posterior distribution.

5.1 Introduction

In this chapter we propose a new nonparametric Bayesian estimator for the solution of an Euler equation. In particular, we focus on the Euler equation defined in consumption-based asset pricing model.

We link two ingredients. The first one is the bayesian nonparametric approach we have proposed in Chapters 2 and 3 to solve integral equations of first kind, stated in infinite dimensional Hilbert spaces. In this paper we develop a similar bayesian procedure for solving integral equations of second kind, whose Euler Equations are a well-known example in economics. The second ingredient is the consumption-based asset pricing model in the style of the Lucas'(1978) tree model.

We have introduced the nonparametric bayesian approach in a general setting where the

---

[1]This chapter is adapted from: Simoni, A. (2008), *Bayesian Nonparametric Estimation of Asset Pricing Functionals*, mimeo.

object of interest was the solution of an integral equation of first kind. Several estimation problems in econometrics can be restated as problems of recovering the solution of a functional equation (*i.e.* as an inverse problem) and there exist numerous techniques to solve them, see Carrasco *et al.* (2007). Our contribution is the development of a Bayesian approach that is new both as solution technique of inverse problem and as bayesian nonparametric estimation method. The main Bayesian solution of a functional equation, that we propose, is the *regularized posterior distribution* of the parameter of interest. It is a regularized version of the "classical" posterior distribution where the regularization is performed through alternative techniques, like Tikhonov scheme or Hilbert Scale regularization, and it is necessary in order to guarantee posterior consistency.

The application of these bayesian techniques to dynamic rational expectation models is a first attempt to illustrate the usefulness in economics and econometrics of our new Bayesian approach.

Dynamic rational expectation models have been extensively studied in economic and econometric theory. In these models economic agents are supposed to face an intertemporal choice problem in which they have to determine their consumption and investment plans through a maximization of an infinite horizon expected utility function under budget and positivity constraints. The result is a model for general equilibrium assets pricing where the assumption of *rational expectations* is fundamental. In fact, it is assumed that the market clearing price, implied by consumer behavior, is the same as the price on which consumer decisions are based.

This paper exploits the equilibrium characterization provided by such kind of models in order to analyze the performance of the Bayesian nonparametric approach for estimating the equilibrium asset pricing functional. In dynamic rational expectation models, such a functional is characterized as the solution of a functional equation. The aim of this paper is to recover the stochastic character of the price process $\{p_t\}$ of a financial asset. Consumption-based asset pricing models assume that at each time t, the price of a financial asset is equal to a fixed function of the state of the economy Y_t, namely $\forall t$, $p_t = p(Y_t)$. Our idea is to estimate both $p(\cdot)$ and the dynamic of the state of the economy in a nonparametric way and to combine them for obtaining $\{p_t\}$.

Having a nonparametric estimation of $\{p_t\}$ is useful for many reasons. First, it allow to test parametric specifications on the price process. If we take as the state of the economy the aggregate consumption, the price series that we obtain can be interpreted as a measure of the market portfolio and this is very useful since usually we observe it only through proxies. Moreover, $\{p_t\}$ can be used in order to empirically study the implications of the consumption-based asset pricing model for explaining observed data on asset returns and dividends, that is for trying to explain the equity premium puzzle. Lastly, we can use it for analyzing if a financial asset is over- or under-priced.

The Bayesian approach is appropriate to analyze rational expectation models since the way in which economic agents form rational expectations is driven by a Bayesian learning process. The theory of rational expectations was introduced by Muth (1961) and applied to the economy as a whole by Lucas during the 1970s, see Lucas (1976) and Lucas (1978). This theory revolutionized macroeconomics and economic thinking. It is based on the belief that economic agents make their economic choices by taking into account their previous experiences and their rational expectations of the result of those choices. So, as Lucas (1978) points out, the hypothesis of rational expectation "is not *behavioral*: it does not describe the way agents think about their environment, how they learn...It is rather a properly likely to be (approximately) possessed by the outcome of this unspecified process of learning and adapting".

Furthermore, a bayesian analysis is interesting, from an econometric point of view, for many other reasons. *(i)*, in computing the estimator of the pricing functional, it allows to exploit the prior information we could have. This is very important for nonparametric estimation since it is difficult to estimate infinite dimensional objects with a finite number of data and parameters that are identified from a mathematical point of view are usually partially identified by the data. Hence, any kind of prior information can helps in restoring identification. In financial markets it is usual to possess this kind of information and it is efficient to use it for improving forecasting. *(ii)*, the Bayesian method that we propose for recovering solution of integral equations broadens the nonparametric estimation techniques in the background of the bayesian statisticians. In fact, we consider a prior different than the Dirichlet process, or its transformations, that is the usual prior for nonparametric estimation. In this paper we are able to stay completely nonparametric by using a gaussian process prior. *(iii)*, the fact that we get the whole posterior distribution of the pricing functional represents a big advantage with respect to the classical estimation procedure that provides only a punctual estimator. The posterior distribution has good small sample properties and so it can be used for recovering every quantity linked to it (as quantiles and confidence intervals) and for implementing testing procedures.

The econometric analysis of dynamic rational expectation models is widely developed. Lucas (1976) and Hansen *et al.* (1980) observed that, instead of estimating the parameter of agents' decision rules, we should estimate the parameters of agents' objective functions and the random process they face as decision makers. This is enough for enabling the econometricians to predict how agents' decision rules change over time across alterations in their stochastic environment.

Dynamic rational expectation models have been exploited by econometricians in order to pursue two different aims. The first scope, that has motivated the literature on GMM, has been to estimate parameters of economic agents' preferences. The dynamic optimization problem of economic agents provides a set of stochastic Euler equations that must be satisfied in equilibrium. These Euler equations, in turn, imply a set of population orthogonality conditions that can be exploited to estimate the parameters of interest. Several authors have proposed to use Euler equations to estimate parameters, see Hayashi (1980) [44], Fair and Taylor (1980) [22], Hansen and Sargent (1980) [39], Hansen and Sargent (1981) [40], Sargent (1981) [71], Hansen and Singleton (1982) [41].

An other branch of econometric literature concerning dynamic rational expectation models, is interested in directly recovering the equilibrium asset pricing functional by solving the Euler equation that characterize it. Our paper gets into this literature. In the simple Lucas' tree model (1978) [59], characterized by a one-good, pure exchange economy with identical consumers, the equilibrium asset vector price is described as a functional $p(\cdot)$, of the Markov state of the economy, solution of an integral equation of second kind. The functional equation is of the form $(I - K)p = d$, where I and K are two operators (the identity and an integral operator, respectively) onto an infinite dimensional Hilbert space and d is a known element of this Hilbert space [2]. Such characterization is particularly useful since it allows to recover equilibrium asset prices without imposing any parametric restriction on them and by using the theory of inverse problems. Only regularity and smoothness conditions will be imposed.

---

[2]An integral equation of second kind is a particular type of inverse problem and it can be ill-posed or well-posed according to the fact that the integral operator K in it has an eigenvalue equal to one or not. Methods for treating integral equations of second kind are extensively treated in Kress (1999) and Carrasco *et al.* (2007).

Literature dealing with determination of *equilibrium asset pricing functionals* by solving Euler equations can be split in three veins. *(i)* The literature that proposes an exact solution to the Euler equation. This requires strong parametric assumptions on the stochastic discount factor and on the dynamic of the state of the economy. Not only a parametric specification is necessary, but the methodology works uniquely for a specific parametric form, see for instance Burnside (1998) [8], Tsionas (2003) [76], Bidarkota and McCulloch (2003) [6] and Calin *et al.* (2005) [9]. *(ii)* The literature proposing numerical solution methods for the Euler equation, see Hussey and Tauchen (1991) [75] and Rust *et al.* (2002) [70]. These methodologies require to specify a parametric form for both the stochastic discount factor and the dynamic of the state of the economy, but they works for whatever parametric specification. Hussey and Tauchen (1991) compute a discrete state space solution method for the pricing functional based on numerical quadrature approximation of the integral operator K. Rust *et al.* (2002) use the observation that operator $K + r$ is a *quasi linear contraction* and compute a pointwise ε-approximation of its fixed point. This approximation is shown to converge at a rate close to T^{-1}. *(iii)* The econometric methods proposing to estimate the solution of the Euler equation, see Carrasco *et al.* (2007) [10]. In this literature whatever parametric specification for the stochastic discount factor is required but the dynamic of the state of the economy does not need to be specified and it is estimated nonparametrically. Henceforth, more flexibility is admitted.

Our methodology belongs to this third vein and it will be compared with the methodology proposed by Carrasco *et al.* (2007). In the following, we refer to this methodology as the *classical approach* since they propose a classical method for estimating the asset price in Lucas' model based on an estimation of d and K and on the simple inversion of operator $(I - K)$. The inverse problem is well-posed so that no regularization technique is demanded for solving it.

A particular feature of the method that we propose in this chapter is that we stay nonparametric also in the dynamic of the state of the economy. This choice is motivated by the fact that we want to stay as general as possible and, in particular, by the result of Bansal and Yaron (2004) that it is empirically "difficult to distinguish an *i.i.d.* consumption growth model and a long-run risk model.

The new approach that we propose to estimate the asset pricing functional is different from the previous ones first of all because it is bayesian. Our approach restates the integral equation in a larger space of probability distributions so that each quantity in it (p and d in our case) are re-interpreted as random functions. Hence, from a Bayesian point of view, the solution to an Euler equation is the *posterior distribution* of the quantity of interest p.

Some element of the integral equation defining the asset pricing functional is unknown and requires to be estimated, so that finally we obtain an Euler equation that is only approximately true: $\hat{d} \approx (I - \hat{K})p$. In particular, what is unknown is the transition density of the Markov state of the economy and it is estimated nonparametrically. The asymptotic properties of such estimator define the sampling probability associated to this functional equation. In fact, the exact sampling distribution is not computable. Moreover, in order to derive a suitable asymptotic distribution, the original model must be transformed as $\hat{K}^*\hat{d} = \hat{K}^*(I - \hat{K})p$, where $\hat{K}^*$ denotes the estimation of the adjoint of K. We end up with an integral equation of first kind that is solvable through the technique we have proposed in Chapter 2. Hence, even if both the classical and the bayesian approaches start with the same functional equation, they finally solve two substantially different, though linked, functional equations.

The infinite dimension of the pricing functional inverse problem makes the posterior dis-

tribution not well defined due to lack of continuity of its mean function. Hence, the posterior mean, and consequently the posterior distribution, is prevented from being consistent in the frequentist sense. This is an interesting example of frequentist inconsistency in Bayesian nonparametric estimation, see Diaconis and Freedman (1986) [16]. If p_* denotes the true value of the pricing functional having generated the data, the posterior distribution is said to be consistent in the frequentist sense if it degenerates, with respect to the sampling distribution, towards a point mass in p_* as more and more observations are collected.

The strategy that we use consists in getting rid of the lack of continuity by applying a regularization scheme in the computation of the posterior distribution. We propose two alternative regularization schemes: a classical Tikhonov scheme and a Tikhonov regularization in the Hilbert scale induced by the prior covariance operator. The posterior distribution that we get is slightly modified and it is called *Regularized Posterior distribution* to highlight the role played by the regularization scheme. We take as punctual Bayesian estimator the mean of this distribution. Under some regularity condition on the true pricing functional p_*, our bayesian estimator converges towards p_* faster, in L^2-norm and in the sampling probability, than the classical estimator proposed in Carrasco *et al.* (2007).

Finally, we study a particular prior distribution that is able by itself to introduce the regularization scheme necessary for making the posterior distribution consistent. This prior is an extended version of the g-prior proposed by Zellner (1986).

The chapter is organized as follows. In Section 5.2 we briefly remind the rational expectation general equilibrium model of Lucas (1978) and we explicit the functional equation in the equilibrium asset price as an integral equation of second kind. We properly define the Hilbert space we are working in and the integral operator K. The Bayesian approach will be explained and adapted to this particular inverse problem in Section 5.3. In this section we compute the regularized posterior distribution by using the two alternative regularization schemes. In Section 5.4, posterior consistency of the regularized posterior distribution of the asset price p is proved. Section 5.5 presents the particular g-prior distribution for the pricing functional that is able to regularize. We develop an extension of our model in Section 5.6 where the variance parameter in the sampling covariance operator is unknown. Section 5.7 concludes. All the proofs and some numerical simulation can be found in the Appendix.

5.2 Rational Expectations Asset Pricing Model

Our Bayesian estimator does not require any particular assumption about preferences to be satisfied in the asset pricing model. It is general and it can be applied to every asset pricing model that characterizes the asset pricing functional as solution of the Euler Equation. In order to stay as general as possible in this paper we take the asset pricing model of Lucas (1978) since it represents the basis for all the subsequent models. Every extension to more specific models with, for instance, non-separable utility functions, habit preferences or Epstein and Zin (1991) utility function is possible with only minor modifications.

5.2.1 Lucas' (1978) Model

Lucas (1978) [59] constructed the equilibrium in an exchange economy under the assumption of *rational expectations*. The first-order conditions for attaining the optimum define a functional equation in the vector of equilibrium prices of financial assets which is solved for price as a function of the physical state of the economy.

We consider a one-good pure exchange economy with a single consumer interpreted as representative of a large number of identical consumers. The consumer faces the intertemporal choice problem between consumption and trading in financial assets and she/he maximizes the expectation of a time-separable utility function:

$$\mathbb{E}_t\Big[\sum_{j=0}^{\infty}\beta^j U(C_{t+j})\Big] \tag{5.1}$$

where $\mathbb{E}_t$ denotes the conditional expectation operator conditional on the information set $\mathcal{F}_t$ available in t, $\beta \in (0,1)$ is the time discount factor, $U(\cdot)$ is a current period strictly concave utility function and C_{t+j} is a stochastic process representing the consumption of a single good at time $t+j$. Since expectations are supposed to be formed rationally, $\mathbb{E}_t$ denotes both the mathematical conditional expectation and the agents' subjective expectations at time t.

In this economy there exist n distinct productive units (denoted with $i = 1,\ldots,n$) each one producing a quantity Y_{it} of the consumption good in period t. The production $Y_t = (Y_{1t},\ldots,Y_{nt})$ is assumed to be entirely *exogenous* and to follow a Markov process defined by its transition distribution function $F(y_{t+1}|y_t) = \mathbb{P}\{Y_{t+1} \leq y_{t+1}|Y_t = y_t\}$. Moreover, since the produced output is perishable, feasible consumption levels are those which satisfy $0 \leq C_t \leq \sum_{i=1}^{n} Y_{it}$. Each productive unit has outstanding one perfectly divisible equity share held by the representative consumer and traded at a competitively determined price vector $p_t = (p_{1t},\ldots,p_{nt})$. We denote with $z_t = (z_{1t},\ldots,z_{nt})$ the consumer's share holding at the beginning of period t, *i.e.* z_{it} is the period t share holding in the i-th productive unit.

Definition of the equilibrium of this economy requires to determine the equilibrium *quantities* of consumption and asset holdings and the equilibrium *price vector p*. As Lucas stresses, the equilibrium quantities of consumption and asset holdings are easily determined since all output will be consumed and all shares will be held, then

$$C_t = \sum_{i=1}^{n} Y_{it}, \qquad z_t = (1,\ldots,1), \quad \forall t. \tag{5.2}$$

The feasible equilibrium consumption and investment plans must satisfy, at each period t, the budget constraint

$$C_{t+1} + p_t z_{t+1} \leq Y_t z_t + p_t z_t, \quad C_t \geq 0 \ z_t \geq 0. \tag{5.3}$$

The important economic variable whose equilibrium value remains to be determined is the asset price. Equilibrium prices are set by the asset market by solving a problem of the same form each period, so that it seems natural to express them as some fixed function $p(\cdot)$ of the state of the economy: $p_t = p(Y_t)$, where the i-th coordinate $p_i(Y_t)$ is the price of a share of unit i when the economy is in the state Y_t.

The first order conditions for maximizing (5.1) subject to (5.3), once equilibrium conditions (5.2) have been incorporated, gives a functional equation in the equilibrium price vector, or equivalently, n functional equations:

$$p_i(Y_t) = \beta \int \frac{U'(\sum_i Y_{i,t+1})}{U'(\sum_i Y_{i,t})}\,(Y_{i,t+1} + p_i(Y_{t+1}))\,d\,F(Y_{t+1}|Y_t), \tag{5.4}$$

for $i = 1,\ldots,n$, where the conditional expectation $\mathbb{E}_t$ in (5.1) has been explicited. This equilibrium asset-pricing relation is the classical Euler equation that equates current price of the i-th security to its expected discounted future payoff, discounted using the *stochastic*

discount factor $M_{t+1}(Y_t, Y_{t+1}) = \beta \frac{U'(\sum_i Y_{i,t+1})}{U'(\sum_i Y_{i,t})}$. The *stochastic discount factor* is expressed as a function of the vectorial Markov state $\{Y_t\}$ instead of consumption process $\{C_t\}$. In the following of the paper, sometimes we shall denote it, at time $t+1$, simply by M_{t+1}, by neglecting its arguments.

Two remarks are in order. First, we choose to use the Lucas' model and a separable utility function because this represents the most general setting and it allows to explain in a clear way our bayesian estimation approach. In any case, our bayesian procedure does not require them and it perfectly works with every other specification of the utility function (*e.g.* non-separable utility function over time and goods, habit utility function, Epstein-Zin utility function, etc...) or with a model in continuous time as Cox, Ingersoll and Ross (1985). A different kind of utility function only affects the stochastic discount factor M_{t+1}, but it does not change the characterization of the asset pricing functional p as the solution of an integral equation.

Second, it is possible to note that the validity of equation (5.4) implies the validity of the projected model

$$\mathbb{E}[p_i(Y_t)|\tilde{Y}_{t+1}] = \mathbb{E}\left[M_{t+1}(Y_t, \tilde{Y}_{t+1})\mathbb{E}\left(M_{t+1}(Y_t, Y_{t+1})(Y_{i,t+1} + p_i(Y_{t+1}))\Big|Y_t\right)\Big|\tilde{Y}_{t+1}\right] \quad (5.5)$$

for $i = 1, \ldots, n$, where we re-project the Euler equation through a conditional expectation operator conditioned on the future state of the economy. This more complicated integral equation will be required in order to compute the sampling distribution in the Bayesian experiment. This is the price to pay for being bayesian.

The object of interest of this paper will be the determination of the vector of pricing functionals $p(\cdot)$. Since equilibrium prices are a fixed function of the state of the economy, once the transition function $F(y_{t+1}|y_t)$ is known or estimated, this will be sufficient to determine the stochastic process of prices p_t.

5.2.2 Martingale Property

The equilibrium asset-pricing relation (5.4) says that $p_i(Y_t) = \mathbb{E}[M_{t+1}(Y_{i,t+1} + p_{i,t+1})|Y_t]$. Therefore, we can write:

$$M_{t+1}(Y_{i,t+1} + p_{i,t+1}) = p_i(Y_t) + \varepsilon_{t+1}. \quad (5.6)$$

The variable ε_{t+1} is a noise satisfying the following assumption that will turn out useful in determining the covariance operator of the sampling distribution in the Bayesian experiment.

Assumption 19 $\{\varepsilon_{t+1}\}$ *is a weak white noise with variance σ^2 that is constant for each time t.*

The fact that error terms are serially uncorrelated prevents problems of endogeneity of the regressors.

Lucas (1978) [59] stresses that *"asset prices themselves do not possess the Martingale property"*, but that asset prices properly corrected for dividends and for the stochastic discount factor β possess this property, how can be seen from equation (5.4). This observation confirms the finding of Leroy (1973) [56] that the martingale property is neither a necessary nor sufficient condition for rationally determined asset prices. However, it is possible to show that there exists a probability, known as *risk-neutral probability* (or

equivalent martingale measure - EMM) under which the discounted price process corrected for dividends is a martingale. To show this, note that relation (5.4), divided by the value of the function $p_i(Y_t)$, gives for a risk-free security

$$1 = (1 + r_f)\mathbb{E}_F(M_{t+1}|Y_t),$$

where r_f denotes the risk-free rate compounded once in period $[t, t+1]$. We make the following assumption concerning the transition distribution function of the Markov state

Assumption 20 *The transition distribution function $F(y_{t+1}|y_t)$ is absolutely continuous with respect to the Lebesgue measure and there exists a positive function f such that $\frac{dF(y_{t+1}|y_t)}{dy_{t+1}} = f(y_{t+1}|y_t)$.*

Hence, under this hypothesis, we have $\forall i = 1, \ldots, n$

$$
\begin{aligned}
p_i(Y_t) &= \int \frac{Y_{i,t+1} + p_i(Y_{i,t+1})}{1 + r_f} \frac{M_{t+1}(Y_t, Y_{t+1})}{\mathbb{E}(M_{t+1}|Y_t)} f(Y_{t+1}|Y_t) dY_{t+1} \\
&= \int \frac{Y_{i,t+1} + p_i(Y_{i,t+1})}{1 + r_f} f^*(Y_{t+1}|Y_t) dY_{t+1},
\end{aligned}
$$

where $f^*(Y_{t+1}|Y_t) = \frac{M_{t+1}}{\mathbb{E}(M_{t+1}|Y_t)} f(Y_{t+1}|Y_t)$ is the equivalent martingale measure. In the following we denote with $\mathbb{E}^*$ the expectation taken with respect to this probability.

5.2.3 Integral Equations of Second Kind and Characterization of the Operator

In this subsection, we study mathematical properties of functional equations (5.4) and (5.5), meant as a functional equations in $p_i(\cdot)$, and we properly characterize all the elements appearing in it. If Assumption 20 holds, we can restate equation (5.4) in a more general form:

$$p_i(Y_t) - \int M_{t+1}(Y_t, Y_{t+1}) p_i(Y_{t+1}) f(Y_{t+1}|Y_t) dY_{t+1} = \int M_{t+1}(Y_t, Y_{t+1}) b_i(Y_{t+1}) f(Y_{t+1}|Y_t) dY_{t+1},$$
$$(5.7)$$

for $i = 1, \ldots, n$. Function b_i is the coordinate function associating vector Y_{t+1} to its i-th component. $\{Y_t\}$ is an n-dimensional stationary stochastic process that satisfies Markov property with stationary distribution Π, *i.e.* Π is the unique solution to

$$\Pi(Y_{t+1}) = \int F(Y_{t+1}|Y_t) d\Pi(Y_t).$$

We denote with π the density function associated to Π.

Let $\mathcal{X}$ be the space of square integrable functions of one realization of $\{Y_t\}$ with respect to the stationary distribution Π endowed with the scalar product $< \cdot, \cdot >$ inducing the norm $|| \cdot ||$, *i.e.* $\mathcal{X} = L^2_\pi(Y)$. We assume that $p \in \mathcal{X}$ [3] and we define an operator K acting on this space as:

$$\forall \phi \in \mathcal{X}, \qquad K\phi(Y_t) = \mathbb{E}_F(M_{t+1}(Y_t, Y_{t+1})\phi(Y_{t+1})|Y_t),$$

---

[3]This assumption is simply an assumption on the distribution of the state of the economy Y_t.

where the conditional expectation is taken with respect to the transition distribution $F(Y_{t+1}|Y_t)$. Operator K is a contraction operator with norm strictly less then 1. The contraction property can be easily proved by using Theorem 5 in Blackwell (1965) [7] or directly through the definition of contraction operator. In particular, $||K|| := \sup_{\phi:||\phi||\leq 1} ||K\phi|| \leq \frac{1}{1+r_f} \sup_{\phi:||\phi||\leq 1} ||\mathbb{E}^*(\phi|Y_t)|| < 1$ since the conditional operator has norm equal to 1 and $\frac{1}{1+r_f} < 1$.

The adjoint K^* of this operator is defined through the equality $< K\phi, \psi >=< \phi, K^*\psi >$, $\forall \phi, \psi \in \mathcal{X}$, so that $K^*\psi = \mathbb{E}_F(M_{t+1}(Y_t, Y_{t+1})\psi(Y_t)|Y_{t+1}) = \int \beta \frac{U'(Y_{t+1})}{U'(y_t)}\psi(y_t)f(y_t|y_{t+1})dy_t$ and it is the operator characterizing the projected model (5.5). Although $F(Y_t|Y_{t+1}) = F(Y_{t+1}|Y_t)$, the two operators K and K^* are substantially different due to the fact that M_{t+1} is not symmetric in its arguments. Thus, $K\phi$ coincides, up to a constant, with the conditional expectation of the product of ϕ and the marginal utility function whereas $K^*\phi$ is proportional to the conditional expectation of the ratio $\frac{\phi}{U'}$.

We call $d_i(Y_t)$, or simply d_i, the right hand side of equation (5.7), so that we rewrite the equilibrium model as

$$
\begin{aligned}
d_i(Y_t) &= (I - K)p_i(Y_t), \qquad i = 1, \dots, n \\
d_i(Y_t) &:= \mathbb{E}_F(M_{t+1}(Y_t, Y_{t+1})b_i(Y_{t+1})|Y_t), \qquad i = 1, \dots, n
\end{aligned}
\tag{5.8}
$$

where I is the identity operator onto $\mathcal{X}$. In the following we eliminate the subscript i in the price, b_i and d_i functions and it will be implied that the functional equation $(I - K)p = d$ refers to a single security.

We will now introduce an assumption, that is only a regularity assumption but that is useful to guarantee compacity of operator K.

Assumption 21 *The Equivalent Martingale Measure $f^*(Y_{t+1}|Y_t)$ is dominated by the marginal distribution of Y_{t+1} and its density is square integrable with respect to the product of margins of Y_{t+1} and Y_t.*

Exploiting this assumption it is possible to show that K is an Hilbert-Schmidt operator. Let $k(Y_t, Y_{t+1}) = M_{t+1}\frac{f(Y_{t+1}|Y_t)}{\pi(Y_{t+1})}$ be the kernel characterizing operator K. K is an Hilbert-Schmidt operator if the Hilbert-Schmidt norm $|| \cdot ||_{HS}$ is finite:

$$
\begin{aligned}
||K||^2_{HS} &= \int |k(Y_t, Y_{t+1})|^2 \pi(Y_t)\pi(Y_{t+1})dY_t dY_{t+1} \\
&\leq (1 + r_f)^2 \int (M_{t+1}\frac{f(Y_{t+1}|Y_t)}{\pi(Y_{t+1})})^2 \pi(Y_t)\pi(Y_{t+1})dY_t dY_{t+1} \\
&= \int (\frac{M_{t+1}}{\mathbb{E}(M_{t+1}|Y_t)}\frac{f(Y_{t+1}|Y_t)}{\pi(Y_{t+1})})^2 \pi(Y_t)\pi(Y_{t+1})dY_t dY_{t+1} \\
&= \int (g^*(Y_{t+1}|Y_t))^2 \pi(Y_t)\pi(Y_{t+1})dY_t dY_{t+1} < \infty
\end{aligned}
$$

where the second line follows from the fact that $(1 + r_f)^2 \geq 1$ and g^* is the density of the EMM f^* with respect to $\pi(Y_{t+1})$, i.e. $\frac{dF^*(Y_{t+1}|Y_t)}{d\Pi(Y_{t+1})} = g^*(Y_{t+1|Y_t})$.

Hilbert-Schmidt operators are compact; this is a very attractive property since every compact operator is the limit of a sequence of operators with finite dimensional range.

Hence, when operator K has to be estimated it can be approached by a sequence of finite dimensional operators. Furthermore, a compact operator has peculiar spectral properties. The eigenvectors of a self-adjoint compact operator can be orthonormalized, the set of its eigenvalues $\{\lambda_j^2\}$ is at most countable and if there are infinitely many eigenvalues they accumulate only at 0. For a compact operator that is non self-adjoint, like K, we consider its *singular values* that are defined to be the square roots of the eigenvalues of the nonnegative self-adjoint compact operator K^*K. Then, there exist orthonormal sequences $\{\varphi_j\}$ and $\{\psi_j\}$ of $\mathcal{X}$ such that

$$K\varphi_j = \lambda_j \psi_j, \quad K^*\psi_j = \lambda_j \varphi_j.$$

Assumption 21 also implies that $r(Y_t) \in \mathcal{X}$, $\mathcal{R}(K) \subseteq \mathcal{X}$ and $\mathcal{R}(K^*) \subseteq \mathcal{X}$, then $K : \mathcal{X} \to \mathcal{X}$ and $K^* : \mathcal{X} \to \mathcal{X}$.

Functional equation (5.7) is an integral equation of second kind and its properties are well known in the literature (see Kress (1999) [50]). While K is compact, $(I - K)$ is not compact. Moreover, 1 is not an eigenvalue of K so that $(I - K)$ is one-to-one and its inverse is bounded. Therefore, the inverse problem defined by (5.7) is well-posed in the sense that it satisfies Hadamard's conditions, see Engl *et al.* [19]. Unfortunately, when we consider the projected model (5.5) we loose the well-posed character of the inverse problem. The projection operation transforms a well-posed inverse problem in an ill-posed one since operator $K^*(I - K)$ is compact and its inverse is not continuous on $\mathcal{X}$, so that the recovered pricing functional p is very sensitive to small measurement errors in r.

5.3 Bayesian Econometric Analysis

The aim moving our econometric analysis is the characterization and estimation of the price process $\{p_t\}$. The price process can be expressed at each period t as a fixed function $p(\cdot)$ of the state of the economy: $p_t = p(Y_t)$. Therefore, once function $p(\cdot)$ is known, knowledge of the transition function $F(y_{t+1}|y_t)$ is enough to determine the stochastic character of the price process. While the transition function will be approximated in a classical nonparametric way (*e.g.* with a kernel method) the whole pricing function $p(\cdot)$ will be the object of a Bayesian analysis.

The rationalization for our estimation choice is that prices are economic variables that economic agents have to take into consideration when they make their economic decisions and on which they performs a Bayesian learning through a continuous updating of the prior distribution. Hence, it seems natural to consider a similar learning process for the econometrician. On the contrary, the transition probability of the state of the economy is exogenous to the learning process of the economic agents and so it does not seem suitable to treat it in a Bayesian way. Roughly speaking, we could consider $F(y_{t+1}|y_t)$ as a nuisance parameter. This approach has nothing of strange since it is the same as in the classical linear model, where the parameters are estimated in a bayesian way while the second moment of covariates and the second cross moment are estimated with a classical procedure, see Zellner (1986) [82].

The stochastic discount factor M_t will be considered as known. In the case in which it is unknown we can calibrate it.

5.3.1 Nonparametric Estimation of the Transition Density

The transition density function $f(Y_{t+1}|Y_t)$ is usually unknown. In this subsection, it will be briefly reviewed the construction and properties of the kernel density estimation considered in Roussas (1967) [68].

With abuse of notation, we use f to denote both the transition density and the two-dimensional joint density of the Markov process $\{Y_t\}$ with respect to Lebeasgue measure. It is assumed that π is strictly positive on $\mathbb{R}_+$. Then, the transition density of the process is written as $\frac{f(Y_t, Y_{t+1})}{\pi(Y_t)}$. We state the following assumption where small letters denote realizations of the random variable Y_t.

Assumption 22 *We dispose of a $(T+1)$ sample $(y_1, \ldots, y_{T+1})$ from the weakly stationary Markov process $\{Y_t\}$.*

As already stated we want to stay as general as possible, hence we follow the original setup of Lucas (1978) [59] which assumes stationarity of dividends levels, so we take Y_t as the aggregate consumption process.

In some case, data may not confirm the hypothesis of stationarity of the consumption process. When this is the case, it is sufficient to rewrite the basic asset pricing equation (5.4) to express it in terms of consumption growth rates, which is shown to be stationary and Markov by empirical evidence. Then, Y_t will denote either the consumption growth rate process or a stationary state variable whose the consumption growth rate is a transformation, see Chen *et al.* (2008) [12]. The slightly modified asset pricing equation can be rewritten as

$$v_i(Y_t) = \mathbb{E}(m(Y_{t+1}, Y_t)[1 + v_i(Y_{t+1})]\frac{Y_{t+1}}{Y_t}|Y_t) \tag{5.9}$$

where v_i denotes the i-th asset's price-dividend ratio, $m(Y_{t+1}, Y_t) = \beta\frac{U'(C_{t+1})}{U'(C_t)}$, under the hypothesis of homogeneous utility function, and $\frac{Y_{t+1}}{Y_t}$ is the dividend growth variable.

In the following, this specification is not used and for clarity and simplicity of exposition we consider the basic Lucas setting. All the results in the following can be trivially adapted to the functional equation (5.9) with only minor modifications.

Let $L : (\mathbb{R}^n) \to \mathbb{R}$ be a measurable function satisfying properties:

$$|L(u)| \le M_1(< \infty), \quad u \in \mathbb{R}^n; \qquad \int |K(u)|du < \infty,$$

$$||u||^m|K(u)| \to 0, \text{ as } ||u|| \to \infty; \qquad \int K(u)du = 1,$$

$h = h(T)$ be a function of T such that $h \to 0$ as $T \to \infty$ and $L_h(u)$ stands for $L(\frac{u}{h_T})$. Then, the kernel transition density estimation is obtained as the ratio of the kernel density estimation of the joint f and of π, $\hat{f}(Y_{t+1}|Y_t) = \frac{\hat{f}(Y_t, Y_{t+1})}{\hat{\pi}(Y_t)}$:

$$\hat{f}(Y_{t+1}|Y_t) = \frac{\frac{1}{Th^{2n}}\sum_{j=1}^{T} L_h(Y_t - y_j)L_h(Y_{t+1} - y_{j+1})}{\frac{1}{Th^n}\sum_{l=1}^{T} L_h(Y_t - y_l)}.$$

We plug this estimator in the operator K and in d:

$$\begin{aligned}
\hat{K}p(Y_t) &= \hat{\mathbb{E}}(M_{t+1}(Y_t, Y_{t+1})p(Y_{t+1})|Y_t) \\
&= \int M_{t+1}(Y_t, Y_{t+1})p(Y_{t+1})\hat{f}(Y_{t+1}|Y_t)dY_{t+1} \\
&= \frac{1}{Th^{2n}}\sum_{j=1}^{T}\frac{L_h(Y_t - y_j)}{\frac{1}{Th^n}\sum_{l=1}^{T}L_h(Y_t - y_l)}\int M_{t+1}(Y_t, Y_{t+1})p(Y_{t+1})L_h(Y_{t+1} - y_{j+1})dY_{t+1} \\
\hat{d}(Y_t) &= \hat{\mathbb{E}}(M_{t+1}(Y_t, Y_{t+1})b(Y_{t+1})|Y_t) \\
&= \int M_{t+1}(Y_t, Y_{t+1})b(Y_{t+1})\hat{f}(Y_{t+1}|Y_t)dY_{t+1} \\
&= \frac{1}{Th^{2n}}\sum_{j=1}^{T}\frac{L_h(Y_t - y_j)}{\frac{1}{Th^n}\sum_{l=1}^{T}L_h(Y_t - y_l)}\int M_{t+1}(Y_t, Y_{t+1})b(Y_{t+1})L_h(Y_{t+1} - y_{j+1})dY_{t+1}.
\end{aligned}$$

The expression for $\hat{K}^*$ can be easily deduced from that one for $\hat{K}$. We assume that $\hat{K}$ and $\hat{K}^*$ define operators from $\mathcal{X}$ into $\mathcal{X}$ and $\hat{d}$ is an element of $\mathcal{X}$. These assumptions are actually integrability assumptions on the kernel function L. Hence, both $\hat{K}$ and $\hat{K}^*$ are degenerate operators with range of dimension T, they are compact and have at most T nonzero eigenvalues $\hat{\lambda}_j$ that implies they have not continuous inverses.

For numerical simulations and asymptotic properties it is useful to approximate $\hat{K}$ and $\hat{d}$ through a change of variable $\frac{Y_{t+1} - y_{j+1}}{h} = u$ and a Taylor expansion at the first order:

$$\begin{aligned}
\hat{K}p &= \frac{\frac{1}{Th}\sum_{j=1}^{T}M_{t+1}(Y_t, y_{j+1})p(y_{j+1})L_h(Y_t - y_j)}{\frac{1}{Th}\sum_{l=1}^{T}L_h(Y_t - y_l)} \\
\hat{d} &= \frac{\frac{1}{Th}\sum_{j=1}^{T}M_{t+1}(Y_t, y_{j+1})b(y_{j+1})L_h(Y_t - y_j)}{\frac{1}{Th}\sum_{l=1}^{T}L_h(Y_t - y_l)}.
\end{aligned}$$

Asymptotic properties of this kernel estimator will affect the asymptotic properties of the Bayesian estimator for p. Note that the use of these estimated quantities implies that the Euler Equation defining the pricing functional is now only approximately true: $\hat{d} \approx (I - \hat{K})p$.

5.3.2 Construction of the Bayesian experiment

We concentrate in this paragraph on the characterization of the Bayesian experiment associated to (5.8). Given the reasons discussed at the beginning of Section 5.3, preference parameters and β are assumed as known and the transition density is substituted with the kernel estimator previously described.

Prior Distribution

The first step in order to well define the Bayesian experiment is the characterization of a *prior probability* μ induced by the pricing functional p on the parameter space $\mathcal{X}$ [4]. We endow the parameter space with the σ-field $\mathcal{E}$ and we assume that μ is a gaussian measure.

---

[4]Note that the distribution μ has nothing to do with the stochastic character of p_t. The latter only depends on the state of the economy once a pricing functional has been drawn from μ

Assumption 23 *Let μ be a probability measure on $(\mathcal{X}, \mathcal{E})$ such that $\mathbb{E}(||p||^2) < \infty$, with $\mathbb{E}$ the expectation taken with respect to μ. μ is a Gaussian measure that defines a mean element $p_0 \in \mathcal{X}$ and a covariance operator $\Omega_0 : \mathcal{X} \to \mathcal{X}$.*

μ is gaussian if the probability distribution on the Borel sets of $\mathbb{R}$ induced from μ by every bounded linear functional on $\mathcal{X}$ is gaussian. More clearly, μ gaussian means that $\forall B \in \mathcal{B}(\mathbb{R})$

$$\mathbb{P}(B) = \mu\{p; <p, \varphi> \in B\}$$

is gaussian for all $\varphi \in \mathcal{X}$, see Baker (1973) [3]. The mean element p_0 in $\mathcal{X}$ is defined by

$$<p_0, \varphi> = \int_{\mathcal{X}} <p, \varphi> d\mu(p)$$

and the operator Ω_0 by

$$<\Omega_0\varphi_1, \varphi_2> = \int_{\mathcal{X}} <p - p_0, \varphi_1><p - p_0, \varphi_2> d\mu(p)$$

for every $\varphi_1, \varphi_2 \in \mathcal{X}$. Let $\mathcal{S}(\mathcal{X})$ denote the set of all linear, bounded, self-adjoint, positive semi-definite and trace-class operators onto $\mathcal{X}$. In particular, $\mathcal{S}(\mathcal{X})$ is the set of all covariance operators of Gaussian measure on $\mathcal{X}$. On the basis of Assumption 23, Ω_0 is correctly specified as a covariance operator in the sense that it belongs to $\mathcal{S}(\mathcal{X})$. A covariance operator needs to be trace-class in order the associated measure be able to generate trajectories in the well suited space. Indeed, by Kolmogorov's inequality a realization of the random function p is in $\mathcal{X}$ if $\mathbb{E}(||p||^2)$ is finite[5]. Since $\mathbb{E}(||p||^2) = \sum_j \lambda_j^{\Omega_0}$, this is guaranteed if Ω_0 is trace-class, that is if $\sum_j \lambda_j^{\Omega_0} < \infty$, with $\{\lambda_j^{\Omega_0}\}$ the eigenvalues associated to Ω_0 and $\mathbb{E}(\cdot)$ the expectation taken with respect to μ.

Since the eigenvalues of $\Omega_0^{\frac{1}{2}}$ are the square roots of the eigenvalues of Ω_0 the fact to be trace-class entails that $\Omega_0^{\frac{1}{2}}$ is Hilbert-Schmidt. Hilbert-Schmidt operators are compact and the adjoint is still Hilbert-Schmidt. Compacity of $\Omega_0^{\frac{1}{2}}$ implies compacity of Ω_0.

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, ($\overline{\mathcal{H}(\Omega_0)}$ in the following), that is dense in $\mathcal{X}$ if Ω_0 is one to one. Let $\{\lambda_j^{\Omega_0}, \varphi_j^{\Omega_0}\}$ be the eigensystem of Ω_0. We define the space $\mathcal{H}(\Omega_0)$ embedded in $\mathcal{X}$ as

$$\mathcal{H}(\Omega_0) = \{\varphi : \varphi \in \mathcal{X} \quad \text{and} \quad \sum_{j=1}^{\infty} \frac{|<\varphi, \varphi_j^{\Omega_0}>|^2}{\lambda_j^{\Omega_0}} < \infty\} \tag{5.10}$$

and, following Proposition 3.6 in Carrasco *et al.* (2007), we have the relation $\mathcal{H}(\Omega_0) = \mathcal{R}(\Omega_0^{\frac{1}{2}})$. It results evident how the choice of the covariance operator can modify the support of a gaussian measure. In particular, if Ω_0 is injective then the support of μ is the whole space $\mathcal{X}$, otherwise, the support is any subset of $\mathcal{X}$; henceforth, a particular choice of the covariance operator allows to incorporate in the prior distribution constraints on the parameter of interest.

An other way to incorporate constraints on the functional form of p consists in specifying a prior mean satisfying them. The trajectories drawn from the corresponding prior distribution will almost surely satisfy the constraints. Let p_* denote the true value of the pricing functional having generated the data $\hat{d}$, we assume that

---

[5]Namely, following Kolmogorov's inequality $\mathbb{P}(||p|| > \epsilon_n) \sim \mathcal{O}_p(1)$ if and only if $\mathbb{E}(||p||^2)$ is finite.

Assumption 24 $(p_* - p_0) \in \mathcal{H}(\Omega_0)$, *i.e. there exists* $\delta_* \in \mathcal{X}$ *such that* $(p_* - p_0) = \Omega_0^{\frac{1}{2}} \delta_*$.

In other words, we are supposing there exists a function $\delta_* \in \mathcal{X}$ such that the centered true value of the pricing functional is the image of it through operator $\Omega_0^{\frac{1}{2}}$. This assumption is only a regularity condition on p_* and will be exploited for proving asymptotic results.

Sampling Distribution

In our model, both the parameter and the sample space coincide with $\mathcal{X}$. We denote with Q^p the sampling probability on $\mathcal{X}$, namely the conditional probability of the observations given p, and it can be inferred from the conditional distribution of the measurement error process $\hat{d} - (I - \hat{K})p$ given p. An exact conditional distribution of this process is impossible, or at least too complicate, to compute due to nonparametric estimation. Hence, we need to compute its asymptotic distribution. However, the nonparametric estimator used for obtaining $\hat{K}$ and $\hat{d}$ prevents us to find convergence of $\hat{d} - (I - \hat{K})p$ to a well defined process with continuous trajectories,like a gaussian process. In fact, it converges towards a process with trajectories that are discontinuous. In order to obtain weak convergence of this process it is necessary to smooth its trajectories. For this, we consider the projected model (5.5) instead of the original one (5.4) and we redefine p as the solution of the estimated integral equation of type one

$$\hat{K}^* \hat{d} = \hat{K}^* (I - \hat{K})p + U \tag{5.11}$$

that is the estimated counterpart of (5.5). We introduce the notation $\hat{R}$ for denoting $\hat{K}^* \hat{d}$ and $\hat{H}$ for denoting $\hat{K}^*(I - \hat{K})$ so that

$$\hat{R} = \hat{H}p + U \tag{5.12}$$

and $\hat{H}$ is the estimator of $H = K^*(I - K)$ that is a compact operator onto $\mathcal{X}$. Hereinafter we denote with H^* the adjoint of H and $H^* = (I - K^*)K$. In this new model the estimated operator $\hat{H}$ becomes the true operator defining the functional equation for p and p is now solution of an integral equation of first kind. The compacity of H makes this inverse problem ill-posed.

The error term process can be rewritten as $U = \hat{K}^*((\hat{d} + \hat{K}p) - (d + Kp))$ and the following theorem shows that it is asymptotically gaussian.

Theorem 19 *Under Assumption 22, there exists a random element $\vartheta \in \mathcal{X}$ such that* $\sqrt{T}\hat{K}^*((\hat{d} + \hat{K}p) - (d + Kp))$ *is asymptotically equivalent to*

$$\frac{\sqrt{T}}{T} \sum_j M_{t+1}(y_j, Y_{t+1})[M_{t+1}(y_j, y_{j+1})(b(y_{j+1}) + p(y_{j+1})) - p(y_j)]\frac{f(y_j, Y_{t+1})}{\pi(y_j)\pi(Y_{t+1})} + h^\rho \vartheta.$$

Moreover, $\sqrt{T}\hat{K}^*((\hat{d} + \hat{K}p) - (d + Kp)) \Rightarrow \mathcal{GP}(0, \sigma^2 K^* K)$ *(weak convergence in $\mathcal{X}$) and* $K^* K$ *is a trace-class operator.*

It will be proved in the Appendix that the first term of the above equality and ϑ weakly converge to a gaussian element in $\mathcal{X}$, but that the second term becomes negligible after having been scaled by $h \to 0$.

Assumption 22, concerning the weakly stationarity of the sample, is necessary only for having a speed of convergence of $\sqrt{T}$, but it does not matter for having weakly convergence

towards a gaussian process. Our guess is that without the weakly stationarity assumption we would get a slower speed of convergence equal to $\delta(T)$, for some function $\delta(\cdot)$.

The sampling distribution Q^p of $\hat{R}$ given p is characterized by the transition probability $\mathbb{P}(\cdot|p)$ that associates to each p a probability measure on $(\mathcal{X}, \mathcal{F})$: $Q^p = \mathbb{P}(\hat{R} \in B|p)$, for all $B \in \mathcal{F}$, where $\mathcal{F}$ is the σ-field associated to the sample space. This probability is deduced from the above theorem, thus Q^p is approximately gaussian with mean $\hat{H}p$ and covariance operator $\Sigma_T = \frac{\sigma^2}{T}K^*K$. Because K is unknown, operator Σ_T is replaced by the estimator $\hat{\Sigma}_T = \frac{\sigma^2}{T}\hat{K}^*\hat{K}$ when we want to compute the posterior distribution (under the assumption that σ^2 is known, the case with σ^2 unknown will be considered in Section 5.6).

Some remarks are in order. First, the fact that the sampling probability is only asymptotically gaussian does not affect properties of our estimator. Indeed, we need normality only to construct the estimator of p and it is not used at all to prove consistency (that is the argument that justifies the proposed estimator).

Second, in order to recover the sampling probability, we have considered the estimated projected model (that is an ill-posed inverse problem) instead of the more natural one $\hat{d} = (I - \hat{K})p + U$ (that is a well-posed inverse problem). This is because such error term does not weakly converge to any well-defined stochastic process since kernel estimation produces an empirical process converging to a process with discontinuous trajectories. Projecting the model through a further application of operator K^* allows to smooth trajectories and to increase the speed of convergence. We loose the well-posedness of the initial inverse problem (5.4), but this is the price to pay in order to be bayesian.

Third, $\Sigma_T \in \mathcal{S}(\mathcal{X})$, thus it possesses all the properties that characterize a covariance operator.

Fourth, the sampling model (5.12) is different than standard econometric models since the sample is represented by only one variable of infinite dimension, that plays the role of the observation, instead of by several finite dimensional observations as usual. The variable $\hat{R}$, playing the role of the sample, is a mathematical object obtained through a transformation of a sample of finite dimensional observations. Therefore, its distribution (in particular its covariance operator) depends on the way the data are generated.

Identification

In our estimation, we are interested in frequentist consistency of the posterior distribution, *i.e.* convergence with respect to the sampling distribution. We will give in Section 5.4 the definition of *frequentist consistency*, also called *posterior consistency* or consistency in the sampling sense. In order this type of consistency be verified we need the following assumption for identification.

Assumption 25 *The operator $H\Omega_0^{\frac{1}{2}} := K^*(I - K)\Omega_0^{\frac{1}{2}} : \mathcal{X} \to \mathcal{X}$ is one-to-one on $\mathcal{X}$.*

This assumption guarantees continuity of the *regularized posterior mean* that we shall define below, so that posterior consistency is satisfied.

Some comments about this hypothesis are in order. If we use the classical model $d = (I - K)p$ and a classical (non bayesian) procedure to recover p then no further identification condition would be required since operator $(I-K)$ is one-to-one (due to the fact that 1 is not an eigenvalue of K). In reality, we are using the projected model $K^*d = K^*(I-K)p$, so that, if a classical resolution method is used, the identification of p would require injectivity of $K^*(I - K)$ that is not guaranteed by injectivity of $(I - K)$. If we compare Assumption 25 to this last one, we see that it is weaker in the sense that if $\Omega_0^{\frac{1}{2}}$ is one-to-one

then $K^*(I-K)\Omega_0^{\frac{1}{2}}$ injective does not imply $K^*(I-K)$ injective while the reverse is true.

Joint Probability Distribution

With relevant space we refer to the product of the sample and parameter space, associated to model (5.11), endowed with the associated σ-field $\mathcal{E} \otimes \mathcal{F}$ and with the joint measure determined by recomposing the prior and sampling distributions. We define the product space $\mathcal{X} \times \mathcal{X}$ as the set

$$\mathcal{X} \times \mathcal{X} := \{(\phi, \psi); \phi, \psi \in \mathcal{X}\}$$

with addition and scalar multiplication defined by $(\phi_1, \psi_1) + (\phi_2, \psi_2) = (\phi_1 + \phi_2, \psi_1 + \psi_2)$ and $h(\phi_1, \psi_1) = (h\phi_1, h\psi_1)$, $\forall h \in \mathbb{R}$. $\mathcal{X} \times \mathcal{X}$ is a separable Hilbert space under the norm induced by the scalar product defined as

$$< (\phi_1, \psi_1), (\phi_2, \psi_2) > := < \phi_1, \phi_2 > + < (\psi_1, \psi_2) >, \quad \forall (\phi_i, \psi_i) \in \mathcal{X} \times \mathcal{X}, \, i = 1, 2.$$

The joint probability measure on $\mathcal{X} \times \mathcal{X}$, denoted with Λ, is constructed by recomposing the prior μ and the sampling distribution Q^p in the following way:

$$\Lambda(A \times B) = \int_A Q^p(B)\mu(dp), \qquad A, B \in \mathcal{X}.$$

After that, function Λ is extended to $\mathcal{E} \otimes \mathcal{F}$. Following discussion in Chapter 2, it is trivial to prove that $(\hat{R}, p)$ are (asymptotically) jointly distributed as a gaussian process:

$$\begin{pmatrix} \hat{R} \\ p \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} \hat{H}p_0 \\ p_0 \end{pmatrix}, \begin{pmatrix} \Sigma_T + \hat{H}\Omega_0\hat{H}^* & \hat{H}\Omega_0 \\ \Omega_0\hat{H}^* & \Omega_0 \end{pmatrix} \right) \tag{5.13}$$

The marginal distribution induced by $\hat{R}$ on $\mathcal{X}$, denoted with Q, is gaussian with mean $\hat{H}p_0$ and covariance $C_T := \Sigma_T + \hat{H}\Omega_0\hat{H}^*$ that is trace class. We shall denote with $\hat{C}_T = \hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*$ the estimated marginal covariance operator. It should be noted that $\hat{H}$ and H are compact operators since they are the product of a bounded and a compact operator, see Theorem 2.16 in Kress [50]. While $\hat{H}$ has a finite number of non-zero singular values, H has a countable number of singular values only accumulating at 0.
Summarizing, the bayesian experiment associated to model (5.5) can be written as

$$\Xi = (\mathcal{X} \times \mathcal{X}, \mathcal{E} \otimes \mathcal{F}, \Lambda = \mu \otimes Q^p).$$

Bayesian inference consists in finding the inverse decomposition of Λ in the product of the posterior distribution, denoted with $\mu^{\mathcal{F}}$, and the predictive measure Q.

5.3.3 Analysis of the Posterior Distribution

The infinite dimension of the Bayesian experiment makes application of Bayes theorem not evident, so that in defining and computing the posterior distribution we should care about three points: *(i)* existence of a *regular* version of the conditional probability on $\mathcal{E}$ given $\mathcal{F}$, *(ii)* the fact that it is a gaussian measure and *(iii)* its continuity. The conditional probability $\mu^{\mathcal{F}}$, given $\hat{R}$, is said *regular* if a transition probability characterizing it exists, *i.e.* there exists a probability $\mathbb{P}(\cdot|\mathcal{F})$ such that $\mathbb{P}(A|\mathcal{F}) = \mu^{\mathcal{F}}(A)$, $\forall A \in \mathcal{E}$. The next theorem answers to the first two questions:

Theorem 20

(i) Let $(\mathcal{X} \times \mathcal{X}, \mathcal{E} \otimes \mathcal{F}, \Lambda)$ be a probability space that is Polish [6], then there exists at least one regular conditional probability $\mathbb{P}(\cdot|\mathcal{F})$ such that $\mathbb{P}(A|\mathcal{F}) = \mu^{\mathcal{F}}(A)$, $\forall A \in \mathcal{E}$.

(ii) The probability $\mu^{\mathcal{F}}$ is characterized by the characteristic function

$$\mathbb{E}(e^{i<p,h>}|\hat{Y}) = e^{i<A\hat{R}+b,h> - \frac{1}{2}<(\Omega_0 - A\hat{H}\Omega_0)h,h>}, \qquad h \in \mathcal{X},$$

where i is the imaginary unit, $A : \mathcal{X} \to \mathcal{X}$ and $b \in \mathcal{X}$. Then $\mu^{\mathcal{F}}$ is gaussian with mean $A\hat{R} + b$ and covariance operator $(\Omega_0 - A\hat{H}\Omega_0)$.

The first point of the theorem is an application of *Jirina theorem*, see Neveu (1965). We find that the space $\mathcal{X}$ we are considering, defined as the space $L^2_\pi(Y)$ of square integrable functions with respect to π, is Polish, see Hiroshi *et al.* (1975). Concerning the second part of the theorem, a proof of this part can be found in Mandelbaum [60]. The characteristic function takes the form of the characteristic function of a gaussian random variable. The posterior mean is $A\hat{R} + b$ and the posterior variance is $\Omega_0 - A\hat{H}\Omega_0$. The deterministic function b has the following form: $b = (I - A\hat{H})p_0$ and operator A is determined through the equality between the two expressions for the covariance operator:

$$\forall \phi, \psi \in \mathcal{X}, \quad \begin{aligned} Cov(<p,\phi>,<\hat{R},\psi>) &= Cov(<\mathbb{E}(p|\hat{R}),\phi>,<\hat{R},\psi>) \\ &= Cov(<A\hat{R},\phi>,<\hat{R},\psi>) \\ &= Cov(<\hat{R},A^*\phi>,<\hat{R},\psi>) \\ &= <(\Sigma_T + \hat{H}\Omega_0\hat{H}^*)A^*\phi,\psi>, \end{aligned}$$

where A^* denotes the adjoint of A, and from (5.13)

$$Cov(<p,\phi>,<\hat{R},\psi>) = <\hat{H}\Omega_0\phi,\psi>.$$

Therefore, by equating these two terms, A is defined as the solution of the functional equation:

$$(\Sigma_T + \hat{H}\Omega_0\hat{H}^*)A^*\phi = \hat{H}\Omega_0\phi \qquad \forall \phi \in \mathcal{X}. \tag{5.14}$$

In reality, Σ_T is unknown and replaced by its estimated version. Therefore, it is more appropriate to define A as the solution of

$$(\hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*)A^*\phi = \hat{H}\Omega_0\phi \qquad \forall \phi \in \mathcal{X}. \tag{5.15}$$

With the transition distribution F replaced by the estimator $\hat{F}$, which is of finite rank, the null set of operators $\hat{H}$, $\hat{H}^*$ and $\hat{\Sigma}_T$ is not reduced to zero. Furthermore, $\hat{\Sigma}_T$, $\hat{H}$ and $\hat{H}^*$ are operators from $\mathcal{X}$ in $\mathcal{X}$, so that they have an infinite number of eigenvalues equal to zero. Hence, $\hat{C}_T$ has not an inverse continuously defined on $\mathcal{X}$ and A^* is unbounded. This causes A to be unbounded and the posterior mean to be not continuous in $\hat{R}$. This is a huge problem because it entails that small measurement errors in $\hat{R}$ will have a severe impact on the posterior mean of p that consequently will be prevented from being a consistent estimator (in the sampling sense). Then, the posterior distribution is not consistent in the sampling sense when we are considering the whole space $\mathcal{X}$. Nevertheless, the posterior mean remain a consistent estimator in the Bayesian sense, *i.e.* with respect to the joint

---

[6]A *Polish space* is a separable completely metrizable topological space.

distribution Λ.

In practice, the computation of the posterior distribution in infinite dimensional spaces requires to solve the further inverse problem (5.15) that is ill-posed. Henceforth, the degree of ill-posedness of the Bayesian problem is different than the degree of ill-posedness of the classical problem. In the following two subsections we propose two solutions to deal with this lack of consistency. These solutions are based on two different regularization techniques of the inverse of operator $(\hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*)$ in (5.15); the first one uses a classical Tikhonov regularization scheme and the second one uses a Tikhonov regularization in the Hilbert scale induced by the inverse of the prior covariance operator.

5.3.4 Tikhonov Regularized Posterior Distribution

We solve the problem of unboundedness of operator A in the posterior mean function by applying a Tikhonov regularization scheme, see Kress (1999), to the inverse of operator $(\hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*)$. We define the regularized operator A_α as:

$$A_\alpha\phi = \Omega_0\hat{H}^*(\alpha I + \hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*)^{-1}\phi \tag{5.16}$$

where $\alpha > 0$ is a regularization parameter that is function of the sample size T, $\alpha = \alpha(T)$, and it is such that $\alpha \to 0$ as $T \to \infty$. This parameter must be chosen in order to balance the trade-off between the bias due to the regularization and the variance due to the instability of the inversion. Operator $(\alpha I + \hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*)$ is surjective and then injective and it has a bounded inverse.

The regularized operator A_α is used to construct a new posterior distribution that we denote with $\mu_\alpha^{\mathcal{F}}$ and that we guess is the solution of the projected Euler equation (5.12). Asymptotic arguments will justify this choice as far as it is proved, in Section 5.4, that $\mu_\alpha^{\mathcal{F}}$ weakly converges, with respect to the sampling probability, to the Dirac measure concentrated in p_*, where p_* is the true value of the pricing functional.

The *regularized posterior distribution* $\mu_\alpha^{\mathcal{F}}$ is a conditional gaussian measure on the σ-field $\mathcal{E}$ given $\mathcal{F}$, with mean and variance

$$
\begin{aligned}
\mathbb{E}_\alpha(p|\hat{R}) &= A_\alpha(\hat{R} - \hat{H}p_0) + p_0 \\
\Omega_{\alpha,R} &= \Omega_0 - A_\alpha\hat{H}\Omega_0.
\end{aligned}
$$

This probability measure is characterized by the estimated operator $\hat{K}$, therefore it must be meant as an estimation of the corresponding regularized posterior distribution with true K. We select as punctual estimator of the equilibrium price function the regularized posterior mean $\mathbb{E}_\alpha(p|\hat{R})$, as it is suggested by a quadratic loss function. This estimator is a continuous function of $\hat{R}$ and then it is consistent.

Tikhonov regularization is a stabilization procedure and it is the equivalent, in inverse problem theory, of shrinkage estimators in statistics and econometrics. These estimators are defined through the addition of a bias in order to stabilize the inversion. One example of shrinkage estimator is the well-known ridge regression. In particular, in finite dimensional Bayesian inverse problem, for particular choices of the prior and sampling variance, the posterior mean and the Tikhonov regularized solution coincides.

Tikhonov regularization is easy to implement but in certain situations the rate of convergence of the regularized solution, toward the true value p_*, is not optimal. More properly, when the true pricing functional p_* is highly regular, Tikhonov regularization does not permit to exploit all its regularity to reach a faster rate of convergence. This is what is called *saturation* or *qualification* effect.

5.3.5 Tikhonov regularization in the Prior Variance Hilbert scale

Different methods for better exploiting the regularity of function p_* have been proposed in literature. Among these, we find the iterative methods, as the *iterated Tikhonov regularization*, and the Tikhonov regularization in Hilbert Scale, see Engl *et al.*(2000) for general theory of regularization in Hilbert scale.

In this subsection, we recover A by applying a *Tikhonov regularization in the Hilbert scale induced by the inverse of the prior covariance operator*. Let $L = \Omega_0^{-\frac{1}{2}}$ be a densely defined, unbounded, self-adjoint, strictly positive operator in the Hilbert space $\mathcal{X}$ [7]. The norm $||\cdot||_s$ is defined as $||x||_s := ||L^s x||$. We define the Hilbert Scale $\mathcal{X}_s$ induced by L as the completion of the domain of L^s, $\mathcal{D}(L^s)$, with respect to the norm $||\cdot||_s$ previously defined; moreover $\mathcal{X}_s \subseteq \mathcal{X}_{s'}$ if $s' \leq s$, $\forall s \in \mathbb{R}$. Usually, when a regularization scheme in Hilbert Scale is adopted, the operator L, and consequently the Hilbert Scale, is created *ad hoc*. The operator L is in general a differential operator. In the Bayesian case this regularization scheme results to be very interesting since the Hilbert Scale is not created ad-hoc but is suggested by the prior information we have and this represents a big difference and advantage with respect to the standard methods. Hence, the regularization scheme is strictly linked to the prior distribution. The following assumption is necessary in order the theory of regularization in Hilbert scale works and gives suitable rates of convergence.

Assumption 26 *(i)* $||H\Omega_0^{\frac{1}{2}}x|| \sim ||\Omega_0^{\frac{a}{2}}x||$, $\forall x \in \mathcal{X}$;

(ii) $(p_* - p_0) \in \mathcal{X}_{\beta+1}$, *i.e.* $\exists \rho_* \in \mathcal{X}$ *such that* $(p_* - p_0) = \Omega_0^{\frac{\beta+1}{2}} \rho_*$

(iii) $a, s, \beta \in \mathbb{R}_+$ *and* $a \leq s \leq \beta + 1 \leq 2s + a$.

This Assumption is the analogous of Assumption 5 in Chapter 2. Therefore, we refer to section 2.3.2 for comments on this assumption.

Under such the regularized solution in $\mathcal{X}_s$ to equation (5.15) is:

$$A_s = \Omega_0 \hat{H}^* (\alpha L^{2s} + \hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*)^{-1}. \tag{5.17}$$

The regularized posterior distribution is thus defined similarly as in Section 5.3.4 with A_α substituted by A_s and is denoted with $\mu_s^{\mathcal{F}}$. The regularized posterior mean and variance are

$$\begin{aligned} \mathbb{E}_s(p|\hat{R}) &= A_s\hat{R} + (I - A_s\hat{H})p_0 \\ \Omega_{s,R} &= \Omega_0 - A_s\hat{H}\Omega_0. \end{aligned} \tag{5.18}$$

A classical Tikhonov regularization method allows to obtain a rate of convergence to zero of the regularization bias that is at most of order 2; on the contrary with a Tikhonov scheme in an Hilbert Scale the smoother the function p_* is, the faster the rate of convergence to zero of the regularization bias will be.

5.4 Asymptotic Analysis

A very important result, due to Doob (1949), see Doob (1949) and Florens *et al.* (1990), states that for any prior, the posterior distribution is consistent in the sense that it converges to a point mass at the unknown parameter that is outside a set of prior mass zero.

---

[7]More clearly, $L = \Omega_0^{-\frac{1}{2}}$ is a closed operator in $\mathcal{X}$ satisfying: $\mathcal{D}(L) = \mathcal{D}(L^*)$ is dense in $\mathcal{X}$, $< Lx, y > = < x, Ly >$ for all $x, y \in \mathcal{D}(L)$, and there exists $\gamma > 0$ such that $< Lx, x > \geq \gamma ||x||^2$ for all $x \in \mathcal{D}(L)$.

Actually, no one can be so certain about the prior, above all when the parameter is of infinite dimension, and values of the parameter for which consistency is not verified may be obtained. To move around this problem it is customary to use a frequentist notion of consistency. The idea of this consistency lies in thinking the data as generated from a distribution characterized by the true value of the parameter and in checking the accumulation of the posterior distribution in a neighborhood of this true value.

This is the so-called "classical bayesian" point of view and, in according to it, we assume there exists a true value of the pricing functional, already denoted with p_*, and we check that the regularized posterior distribution becomes more and more accurate and precise, around p_*, as the number of observed data increases indefinitely. Thus, it is a convergence in the sampling probability sense and it is known as *consistency of the posterior distribution*.

Following Diaconis *et al.* (1986) we give the following definition of *posterior consistency* (or consistency in the sampling sense):

Definition 3 *The pair $(p, \mu^{\mathcal{F}})$ is consistent if $\mu^{\mathcal{F}}$ converges weakly to δ_p as $T \to \infty$ under Q^p-probability or Q^p-a.s., where δ_p is the Dirac measure in p.*
The posterior probability $\mu^{\mathcal{F}}$ is consistent if $(p, \mu^{\mathcal{F}})$ is consistent for all p.

If $(p, \mu^{\mathcal{F}})$ is consistent in the previous sense, the Bayes estimate for p, for instance the posterior mean for a quadratic loss function, is consistent too.

The meaning of this definition is that, for any neighborhood $\mathcal{U}$ of the true parameter p_*, the posterior probability of the complement of $\mathcal{U}$ converges toward zero when $T \to \infty$: $\mu^{\mathcal{F}}(\mathcal{U}^c) \to 0$ in Q^p-probability, or Q^p-a.s. Therefore, since distribution expresses one's knowledge about the parameter, consistency stands for convergence of knowledge towards the perfect knowledge with increasing amount of data.

We refer to Section 2.4 of Chapter 2 for a discussion on this definition. We are persuaded about the importance of studying posterior consistency and in this section we study this concept of consistency for the regularized posterior distribution. By *Chebyshev's Inequality* in L^2 spaces we have, for any sequence $M_n \to \infty$:

$$
\begin{aligned}
\mu_\alpha^{\mathcal{F}}\{p : ||p - p_*|| \geq M_n \varepsilon_n\} &\leq \frac{\mathbb{E}_\alpha(||p - p_*||^2|\hat{R})}{(M_n \varepsilon_n)^2} \\
&= \frac{1}{(M_n \varepsilon_n)^2}[< \Omega_{\alpha, R} 1, 1 > + ||\mathbb{E}_\alpha(p|\hat{R}) - p_*||^2] \\
&\leq \frac{||\Omega_{\alpha, R}|| + ||\mathbb{E}_\alpha(p|\hat{R}) - p_*||^2}{(M_n \varepsilon_n)^2}.
\end{aligned}
\tag{5.19}
$$

The same inequality is valid for $\mu_s^{\mathcal{F}}$.

5.4.1   Speed of convergence with classical Tikhonov regularization

We begins by checking posterior consistency of the regularized posterior $\mu_\alpha^{\mathcal{F}}$ computed with the classical Tikhonov, namely we check accumulation of $\mu_\alpha^{\mathcal{F}}$ to the point mass δ_{p_*}. The main results are contained in the following theorem.

Theorem 21 *Let p_* be the true value of the asset pricing functional and $\mu_\alpha^{\mathcal{F}}$ a gaussian measure on $\mathcal{X}$ with mean $A_\alpha(\hat{R} - \hat{H}p_0) + p_0$ and covariance operator $\Omega_{\alpha, R}$. Under Assumptions 24 and 25, and if $\alpha \to 0$, $\alpha^2 T \to \infty$,*

(i) $\mu_\alpha^{\mathcal{F}}$ weakly converges towards a point mass δ_{p_*} in p_*;

(ii) if moreover $\delta_* \in \mathcal{R}(\Omega_0^{\frac{1}{2}} H^* H \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$, then for $\rho \geq 2$

$$\mu_\alpha^{\mathcal{F}}\{p : ||p - p_*|| \geq \varepsilon_T\} \sim \mathcal{O}_p(\alpha^{\frac{\beta}{2}} + \frac{1}{\alpha_T T} + \frac{1}{\alpha}\Big(\frac{1}{T} + h^{2\rho}\Big)^{\frac{1}{2}}\alpha^{\frac{\beta}{2}} + \frac{1}{\alpha^2 T}\frac{1}{\alpha}\Big(\frac{1}{Th^n} + h^{2\rho}\Big)^{\frac{1}{2}}$$
$$+ \frac{1}{\alpha^2 T}\alpha^{\frac{(\beta+1)}{2} \wedge 1}).$$

The parameter ρ is the minimum between the order of the kernel and the order of differentiability of the density function f.

It should be noted that the condition for the second part of the theorem is only a regularity condition that is necessary for having 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, *i.e.* $(p_* - p_0) \in \mathcal{H}(\Omega_0)$.

The support of a centered gaussian process, taking its value in an Hilbert space $\mathcal{X}$, is the closure in $\mathcal{X}$ of the *Reproducing Kernel Hilbert Space* associated with the covariance operator of this process, see VanDerVaart *et al.* (2000). Then, for p drawn from the prior distribution μ, $(p - p_0) \in \overline{\mathcal{H}(\Omega_0)}$ with μ-probability 1, but with μ-probability 1, $(p - p_0)$ is not in $\mathcal{H}(\Omega_0)$. Hence, the prior distribution is not able to generate trajectories that satisfy Assumption 24 or, in other words, the true value of the price functional p_* cannot be generated by the prior μ specified in Assumption 23. This concept is known in literature as *prior inconsistency* and it refers to a prior that is unable to generate the true parameter having characterized the data generating process. This problem is present only for infinite dimensional parameter sets and it is due to the fact that it is difficult to be sure about a prior on an infinite dimensional parameter space so that it can happen that the true value of the parameter is not in the support of the prior, see e.g. Freedman (1965) [34] or Ghoshal (1998) [37].

Anyway, if Ω_0 is one-to-one, $\mathcal{H}(\Omega_0)$ is dense in $\mathcal{X}$ 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.

The next corollary states consistency of the regularized posterior mean and convergence to zero of the regularized posterior variance; it provides the necessary results for having Theorem 21.

Corollary 4 *Under Assumptions 24 and 25, and if $\alpha \to 0$, $\alpha^2 T \to \infty$, $\rho \geq 2$ then:*

(i) $||\hat{\mathbb{E}}_\alpha(p|\hat{R}) - p_*|| \to 0$ in Q^{p_*}-probability and if $\Omega_0^{-\frac{1}{2}}(p_* - p_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}} H^* H \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$,

$$||\hat{\mathbb{E}}_\alpha(p|\hat{R}) - p_*||^2 \sim \mathcal{O}_p(\alpha^\beta + \frac{1}{(\alpha^2 T)^2}\alpha^{(\beta+1)\wedge 2} + \frac{1}{\alpha T} +$$
$$\frac{1}{\alpha^2}\Big(\frac{1}{T} + h^{2\rho}\Big)\alpha^\beta + \frac{1}{\alpha^2 T}\frac{1}{T}\frac{1}{\alpha^2}\Big(\frac{1}{Th^n} + h^{2\rho}\Big)).$$

(ii) $||\Omega_{\alpha,R}|| \to 0$ in P^{p_*}-probability and $\forall \phi \in \mathcal{X}$ such that $\Omega_0^{\frac{1}{2}}\phi \in \mathcal{R}(\Omega_0^{\frac{1}{2}} H^* H \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ for some $\beta > 0$,

$$||\Omega_{\alpha,R}\phi||^2 \sim \mathcal{O}_p\Big(\alpha^\beta + \frac{1}{\alpha^2}\Big(\frac{1}{T} + h^{2\rho}\Big)\alpha^\beta + \frac{1}{(\alpha^2 T)^2}\frac{1}{\alpha^2}\Big(\frac{1}{Th^n} + h^{2\rho}\Big) + \frac{1}{(\alpha^2 T)^2}\alpha^{(\beta+1)\wedge 2}\Big).$$

The parameter β denotes the regularity of the true p_* and, in the previous rate of convergence, it must be meant as $\beta \wedge 2$ since 2 is the qualification for Tikhonov regularization.

Then, the rate of convergence cannot exceed α^2.

The rate of convergence to zero of the posterior variance is negligible with respect to the rate in the bias, so that the optimal parameter of regularization will be chosen by taking into account the rate of the squared norm of the bias. Concerning this rate, only the first and third terms matter, being the other three terms negligible for particular choices of β and of the bandwidth h. While the first rate α^β requires a regularization parameter α going to zero as fast as possible, the third one requires an α going to zero as slow as possible. In choosing the regularization parameter we should take into account this trade-off, so that the optimal regularization parameter α_* will be obtained when the two rates are made equal: $\alpha^\beta = \frac{1}{\alpha T}$. This implies

$$\alpha_* \propto T^{-\frac{1}{\beta+1}}.$$

The optimal rate of convergence of the squared norm of the regularized posterior mean and variance is $T^{-\frac{\beta}{\beta+1}}$, while the optimal rate of the regularized posterior distribution is $T^{-\frac{\beta}{2(\beta+1)}}$ since, when the optimal α is used, $\alpha^{\frac{\beta}{2}}$ dominates all the other rates.

Let us analyze conditions on β and h to guarantee convergence to zero of the other rates in the bias. A sufficient condition for $\frac{1}{(\alpha^2 T)^2}\alpha^{(\beta+1)\wedge 2}$ converging to zero is that $\frac{1}{(\alpha^2 T)} \sim \mathcal{O}_p(1)$, i.e. $\alpha^2 \sim \mathcal{O}_p\left(\frac{1}{T}\right)$. With α replaced by its optimal value, this condition is met for $\beta \geq 1$. For $\frac{1}{\alpha_T^2}\left(\frac{1}{T}+h^{2\rho}\right)\alpha_T^\beta$ being negligible we have to choose h in such a way that $h^{2\rho} \sim \mathcal{O}_p(\frac{1}{T})$, i.e.

$$h \propto \left(\frac{1}{T}\right)^{\frac{1}{2\rho}}$$

To guarantee that the last rate $\frac{1}{\alpha^4 T^2}(\frac{1}{Th^n}+h^{2\rho})$ converges to zero we simply have to check that $\frac{1}{\alpha^2 Th^n} \sim \mathcal{O}_p(1)$ since the second term is $o_p(1)$ due to the choice of h and to the fact that $\frac{1}{(\alpha^2 T)^2} \sim o_p(1)$. Then, $\frac{1}{\alpha^2}\frac{1}{Th^n} = (\frac{1}{T})^{-\frac{2}{\beta+1}+1-\frac{n}{2\rho}}$ and it goes to zero if $\beta > \frac{2\rho+n}{2\rho-n}$ when $2\rho - n > 0$ and if $\beta < \frac{2\rho+n}{2\rho-n}$ when $2\rho - n < 0$. This constraint is binding with respect to the constraint $\beta \geq 1$, previously introduced, when $2\rho - n > 0$. Summarizing, if $2\rho - n > 0$ the only constraint is $\beta > \frac{2\rho+n}{2\rho-n}$; otherwise, we have two constraints: $1 \leq \beta < \frac{2\rho+n}{2\rho-n}$.

Lastly, it should be noticed that the second, third and fourth rates of the squared norm of the regularized variance operator goes to zero if conditions for ensuring convergence to zero of the terms in the bias are satisfied.

5.4.2 Speed of convergence with Tikhonov regularization in the Prior Variance Hilbert Scale

We compute in this subsection the speed of convergence for $\mu_s^{\mathcal{F}}$. The speed obtained in this case is faster than that one obtained with a simple Tikhonov regularization scheme. In this section we suppose Assumption 26 holds, the attainable speed of convergence is given in the following theorem, the proof of which can be found in Appendix 5.8.

Theorem 22 *Let $\mathbb{E}_s(x|\hat{Y})$ and V_s be as in (5.18). Under Assumptions 24, 25 and 26*

$$||\mathbb{E}_s(p|\hat{R}) - p_*||^2 \quad \sim \quad \mathcal{O}_p\Big(\alpha^{\frac{\beta+1}{a+s}} + \alpha^{\frac{1-a}{a+s}}\frac{1}{T} + \frac{1}{\alpha^4}\frac{1}{T^2}\alpha^{\frac{a+\beta+2s}{a+s}} + \alpha^{\frac{\beta+1}{a+s}}\frac{1}{\alpha^2}\Big(\frac{1}{T}+h^{2\rho}\Big) + \frac{1}{\alpha^3}\frac{1}{T^2}\Big).$$

Moreover, if the covariance operator $\Omega_{s,R}$ is applied to any element $\varphi \in \mathcal{X}$ such that $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{\beta}{2}})$, then

$$||\Omega_{s,R}\varphi||^2 \sim \mathcal{O}_p\Big(\alpha^{\frac{\beta+1}{a+s}} + \frac{1}{\alpha^4 T^2}\alpha^{\frac{2s+a+\beta}{a+s}} + \alpha^{\frac{\beta+1}{a+s}}\frac{1}{\alpha^2}\Big(\frac{1}{T} + h^{2\rho}\Big) + \frac{1}{\alpha^3 T^2}\Big).$$

The optimal α is obtained by equating the first two rates of convergence of the posterior mean: $\alpha^{\frac{\beta+1}{a+s}} = \alpha^{\frac{1-a}{a+s}}\frac{1}{T}$ and is proportional to

$$\alpha_* \propto \Big(\frac{1}{T}\Big)^{\frac{a+s}{a+\beta}}.$$

The optimal bandwidth is determined in the same way as before, hence $h = c_1(\frac{1}{T})^{\frac{1}{2\rho}}$, for some given constant c_1. With this optimal choice of the regularization parameter, in order to guarantee the other rates in the bias and variance are of order $o_p(1)$, we have to restrict the values of β. In particular, if $2a+s > 1$ then the regularity parameter must satisfy $\frac{2s+a-1}{2} < \beta < 2s+a-1$; otherwise $\frac{s-a}{2} < \beta < 2s+a-1$. The corresponding optimal speed of the squared bias and variance is proportional to $(\frac{1}{T})^{\frac{\beta+1}{a+\beta}}$, while the regularized posterior distribution $\mu_s^{\mathcal{F}}$ is of order $\mathcal{O}_p((\frac{1}{T})^{\frac{\beta+1}{2(a+\beta)}})$. It should be remarked that parameter s characterizing the norm in the Hilbert scale does not play any role on the speed of convergence.

An advantage of the Tikhonov regularization in Hilbert Scale is that we can even obtain a rate of convergence for other norms, namely $||\cdot||_r$ for $-a \le r \le \beta+1 \le a+2s$. Actually, the speed of convergence of these norms gives the speed of convergence of the estimate of the r-th derivative of the parameter of interest p.

Tikhonov regularization in Hilbert scale improves the speed of convergence of the regularized posterior distribution with respect to the classical Tikhonov regularization. Let us call γ, instead of β, the regularity parameter of function $(p_* - p_0)$ used in the source condition of subsection 5.4.1, namely $\delta_* \in \mathcal{R}(\Omega_0^{\frac{1}{2}} H^* H \Omega_0^{\frac{1}{2}})^{\frac{\gamma}{2}}$. This is for differentiating with respect to the regularity parameter in the Hilbert scale regularization that will continue to be denoted with β. If Assumption 26 *(i)* holds, it implies the equivalence $||(\Omega_0^{\frac{1}{2}} H^* H \Omega_0^{\frac{1}{2}})^{\frac{\gamma}{2}} v|| \sim ||\Omega_0^{\frac{a\gamma}{2}} v||$, for some $v \in \mathcal{X}$. Then, equivalence of the source conditions in the two regularized solutions implies $||\Omega_0^{\frac{\beta}{2}} v|| \sim ||\Omega_0^{\frac{a\gamma}{2}} v||$ that is verified if $\beta = a\gamma$. In terms of γ, the optimal bayesian speed of convergence with an Hilbert scale regularization is $\Big(\frac{1}{T}\Big)^{\frac{a\gamma+1}{a(1+\gamma)}}$ that is fastest than the bayesian speed of convergence with a classical Tikhonov: $\Big(\frac{1}{T}\Big)^{\frac{\gamma}{\gamma+1}}$, $\forall \gamma > 0$.

5.4.3 Comparison with the classical estimation of the pricing functional

We develop in this paragraph a comparison between the bayesian method we have proposed in this paper for recovering the asset pricing functional and the classical solution to the integral equation (5.7) computed in Carrasco *et al.* (2007) [10]. The classical solution does not require the use of any regularization scheme since the operator $(I - K)$ is continuously invertible. Since K is unknown it is substituted by $\hat{K}$ as defined in subsection 5.3.1, the estimated pricing functional $\hat{p}$ is

$$\hat{p} = (I - \hat{K})^{-1}\hat{d},$$

with $\hat{d}$ defined in subsection 5.3.1. By applying Theorem 7.2 in Carrasco *et al.* [10], the squared norm of the asymptotic bias is of order

$$||\hat{p} - p_*||^2 \sim \mathcal{O}_p\Big(\frac{1}{Th^n} + h^{2\rho}\Big).$$

The optimal speed of convergence is obtained when $\frac{1}{Th^n} = h^{2\rho}$, that is when $h = c_1(\frac{1}{T})^{\frac{1}{2\rho+n}}$. With this optimal choice of bandwidth the classical estimator $\hat{p}$ converges at the rate of $(\frac{1}{T})^{\frac{2\rho}{2\rho+n}}$: $||\hat{p} - p_*||^2 \sim \mathcal{O}_p((\frac{1}{T})^{\frac{2\rho}{2\rho+n}})$.

We compare this rate of convergence with the rate of the estimated regularized posterior mean obtained when a classical Tikhonov scheme and the optimal α are used: $||\hat{\mathbb{E}}_\alpha(p|\hat{R}) - p_*||^2 \sim \mathcal{O}_p((\frac{1}{T})^{\frac{\beta}{\beta+1}})$. The comparison will be possible only in the subset $\Phi_\beta \in \mathcal{X}$ of the pricing functionals p such that $\Omega_0^{-\frac{1}{2}}(p - p_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}} H^* H \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, since we are able to compute the Bayesian speed of convergence for true value p_* belonging to this set. In this subspace, our solution converges faster if $\beta > \frac{2\rho}{n}$. This condition is more likely to be satisfied when the parameter ρ (that is a measure of regularity of the transition density function) is small or equivalently, for a given value of ρ, when the dimension of Y_t, *i.e.* the number of conditioning variables in the transition probability, increases.

Anyway, with Tikhonov regularization the qualification matters, so that we can only exploit a regularity β of the function p that is less or equal than 2. Therefore, in order condition $\beta > \frac{2\rho}{n}$ is satisfied, it must be $\frac{2\rho}{n} \le 2$, that holds when $\rho \le n$.

Let us consider the regularized posterior mean obtained through a Tikhonov scheme in Hilbert scale. In this case the comparison will be possible only on the subspace $\mathcal{X}_{\beta+1}$. With the optimal regularization parameter α_* the rate of convergence is $||\mathbb{E}_s(p|\hat{R}) - p_*||^2 \sim \mathcal{O}_p((\frac{1}{T})^{\frac{\beta+1}{a+\beta}})$ and it is faster than the rate of convergence with classical solution if $\beta > \frac{2\rho(a-1)}{n} - 1$. When $a > 2$ and $\rho < \frac{n}{2(a-2)}$, this condition is less stringent than condition $\beta > \frac{2\rho}{n}$, demanded for Tikhonov regularized posterior mean converging faster than the classical estimator $\hat{p}$. When the degree of ill-posedness a is less than 2, then the condition $\beta > \frac{2\rho(a-1)}{n} - 1$ is less stringent than condition $\beta > \frac{2\rho}{n}$ if $\rho > \frac{n}{2(a-2)}$.

Summarizing, under some condition on the regularity of the function p_*, in particular if the price function is highly smooth, or if n is high or ρ is small, our Bayesian estimator converges faster than the classical one. The price to pay for having this fastest speed of convergence is to impose a regularity assumption on the price functional that we do not impose with the classical resolution method.

5.5 A g-prior with Regularizing Power

We have shown in preceding sections that, in general, the prior distribution does not regularize and we need to artificially introduce a regularization scheme in order to obtain consistency of the posterior distribution.

Nevertheless, there exists a particular specification of the prior distribution that has a regularizing power in the sense that the prior-to-posterior transformation has the same effect as the application of a regularization scheme so that the recovered posterior mean is consistent. This type of prior distribution is suggested by the Zellner' (1986) g-prior but it extends the latter because it is linked to a slightly modified sampling mechanism. More precisely, it is linked to the sampling mechanism of the non-projected model $\hat{d} = (I - \hat{K})p + error$. This *extended g-prior* was introduced in Chapter 3 where its regularizing power was shown.

Let suppose that the prior measure specified in 5.3.2 is replaced by the *extended g-prior* with a covariance operator related to operator K in the sampling mechanism:

$$p \sim \mathcal{GP}\Big(p_0, \frac{\sigma^2}{g}(K^*K)^s\Big), \qquad \text{for some } s > 0 \tag{5.20}$$

with $g = g(T)$ a function of the sample size T such that $g \to \infty$ with T. We use the notation $\Omega_0 = (K^*K)^s$. Let $\alpha = \frac{1}{T}g$ be the parameter playing the role of regularization parameter. For that, it must go to zero with T and it must be such that $\alpha^2 T \to \infty$. These conditions imply that g must go to infinity faster than $\sqrt{T}$ and slower than T.

Equation (5.14) implies an operator $A = (K^*K)^s \hat{H}^*(\alpha(K^*K) + \hat{H}(K^*K)^s \hat{H}^*)^{-1}$ that, as $T \to \infty$, is well-defined if it is applied to $(\hat{R} - \hat{H}p_0)$. The fact that (K^*K) multiplying α can be factorized out allows to directly obtain a regularization of the inverse of the limit of $(K^*K)^{-\frac{1}{2}}\hat{H}(K^*K)^s \hat{H}^*(K^*K)^{\frac{1}{2}}$. Using equation (5.15) for defining A we have

$$
\begin{aligned}
A &= \frac{\sigma^2}{g}(K^*K)^s \hat{H}^*(\hat{\Sigma}_T + \frac{\sigma^2}{g}\hat{H}(K^*K)^s \hat{H}^*)^{-1} \\
&= ((\hat{K}^*\hat{K})^{-\frac{1}{2}}\hat{H}(K^*K)^s)^*(\alpha I + (\hat{K}^*\hat{K})^{-\frac{1}{2}}\hat{H}(K^*K)^s \hat{H}^*(\hat{K}^*\hat{K})^{-\frac{1}{2}})^{-1}(\hat{K}^*\hat{K})^{-\frac{1}{2}}
\end{aligned}
$$

that is a continuous operator. This is due to the fact that $\mathcal{R}(K^*K) \subset \mathcal{R}(K) = \mathcal{D}(K^{-1}) \subset \mathcal{D}((K^*K)^{-\frac{1}{2}})$, so that $(K^*K)^{-\frac{1}{2}}H$ is well defined. The posterior mean and variance are $\mathbb{E}^g(p|\hat{R}) = A(\hat{R} - \hat{H}p_0) + p_0$ and $Var^g(p|\hat{R}) = (K^*K)^s - A\hat{H}(K^*K)^s$. Because operators K and K^* are unknown, it follows that they must be substituted by their consistent estimators in the prior covariance. We denote with $\hat{\mathbb{E}}^g(p|\hat{R})$ and $\widehat{Var}^g(p|\hat{R})$ the corresponding estimated mean and variance.

Study of asymptotic behavior of the posterior distribution is based on the decompositions:

$$
\begin{aligned}
\hat{\mathbb{E}}^g(p|\hat{R}) - p_* &= [\hat{\mathbb{E}}^g(p|\hat{R}) - \tilde{\mathbb{E}}^g(p|\hat{R})] + [\tilde{\mathbb{E}}^g(p|\hat{R}) - \mathbb{E}^g(p|\hat{R})] + [\mathbb{E}^g(p|\hat{R}) - p_*] \\
\widehat{Var}^g(p|\hat{R}) &= [\widehat{Var}^g(p|\hat{R}) - \widetilde{Var}^g(p|\hat{R})] + [\widetilde{Var}^g(p|\hat{R}) - Var^g(p|\hat{R})] + Var^g(p|\hat{R}).
\end{aligned}
$$

The only difference between $\hat{\mathbb{E}}^g(p|\hat{R})$ and $\tilde{\mathbb{E}}^g(p|\hat{R})$ is that in the first one the prior covariance operator is estimated while in the latter it is known. The same difference characterizes $\widehat{Var}^g(p|\hat{R})$ and $\widetilde{Var}^g(p|\hat{R})$. Hence, the first square brackets term of both the two decompositions above is due to estimation of Ω_0, the second error is due to estimation of all the other operators and the last one is the bias and the variance, respectively, for known operators.

We show in the following theorems that the posterior distribution corresponding to the g-prior is consistent. This is guaranteed by convergence to zero of the bias and the posterior variance.

Theorem 23 *Let (5.20) be the prior distribution for the functional p in the sampling equation (5.12). If, for some $\gamma > 0$, $(K^*K)^{s\gamma}$ is trace class and if $(p_* - p_0) \in \mathcal{R}(\Omega_0^{\frac{\beta}{2s}})$ then $||\mathbb{E}^g(p|\hat{R}) - p_*||^2$ converges to zero with respect to the sampling probability at the speed*

$$
\begin{aligned}
||\hat{\mathbb{E}}^g(p|\hat{R}) - p_*||^2 \sim\ & \mathcal{O}_p\Big(\alpha^{\frac{\beta}{s}} + \frac{1}{T}\alpha^{-\gamma} + \frac{1}{\alpha^2}\Big(\frac{1}{Th^n} + h^{2\rho}\Big)(\alpha^{\frac{3s-\beta}{\beta+s}} + \frac{1}{T}\alpha^{-\gamma}) \\
& + \frac{1}{\alpha^2}\Big(\frac{1}{T} + h^{2\rho}\Big)\frac{1}{T}\alpha^{1-\gamma}\Big).
\end{aligned}
$$

Furthermore, if $\alpha = c_1(\frac{1}{T})^{\frac{s}{(\beta+\gamma s)}}$, $h = c_2(\frac{1}{T})^{\frac{1}{2\rho}}$ for some constants c_1 and c_2,

$$T^{\frac{\beta}{\beta+\gamma s}}||\mathbb{E}(p|\hat{R}) - p_*||^2 \sim \mathcal{O}_p(1)$$

if $s \geq 2$, $\frac{n}{2\rho} \leq \frac{\beta+\gamma s-2s}{\beta+\gamma s}$, $(2-\gamma)s \leq \beta \leq 3s$.

It should be remarked that the condition $(p_* - p_0) \in \mathcal{R}(\Omega_0^{\frac{\beta}{2s}})$ in the theorem implies Assumption 24 if $\beta \geq 1$.

The fastest speed of convergence of the posterior mean is of order $T^{-\frac{\beta}{\beta+\gamma s}}$. It is faster than the rate in the classical resolution method (illustrated in subsection 5.4.3) if $\beta > \frac{2\rho}{n}\gamma s$.

Theorem 24 *Let (5.20) be the prior distribution for the functional p in the sampling equation (5.12). If $s \geq 2$ then $||\widehat{Var}^g(p|\hat{R})||^2$ converges to zero with respect to the sampling probability. Moreover, $\forall \phi \in \mathcal{X}$ such that $\Omega_0^{\frac{1}{2}}\phi \in \mathcal{R}(\Omega_0^{\frac{\beta-s}{2s}})$, the posterior variance converges at the speed*

$$||\widehat{Var}^g(p|\hat{R})||^2 \quad \sim \quad \mathcal{O}_p\Big(\alpha^{\frac{\beta}{s}} + \frac{1}{\alpha^2}\Big(\frac{1}{Th^n} + h^{2\rho}\Big)\alpha^{\frac{\beta}{s}}\Big).$$

When α is set equal to the optimal one, *i.e.* $\alpha = c_1(\frac{1}{T})^{\frac{s}{\beta+\gamma s}}$, the posterior variance converges to zero if $\frac{n}{2\rho} \leq \frac{\beta+\gamma s-2s}{\beta+\gamma s}$.

The value of g corresponding to the optimal α is: $g = (\frac{1}{T})^{-\frac{\beta+\gamma s-s}{\beta+\gamma s}}$. It converges at infinite faster than $\sqrt{T}$ and slower than T if $\beta > (2-\gamma)s$. In particular, convergence at a slower rate than T is always guaranteed.

5.6 Prior on the Variance Parameter

Until now we have considered the variance parameter σ^2 in the covariance operator of the sampling measure as known. This parameter is the variance of the white noise in the regression model (5.6) defined by the Lucas' equilibrium model. In reality this parameter is often unknown and needs to be estimated. In this section, we redefine the Bayesian experiment in order to incorporate the parameter space of definition of the variance parameter σ^2: $(\mathbb{R}_+, \mathcal{B}, \nu)$, with $\mathcal{B}$ the Borel σ-field and ν a measure on it.

There exist two possibilities to specify the probability measure on the parameter space. 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 $\mathcal{X}$. 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 paper we only consider the traditional approach since in this case it is possible to define a closed form for the marginal posterior distribution of both the parameters without demanding the implementation of some MCMC procedure as a Gibbs sampling.

5.6.1 Conjugate model

The modified Bayesian experiment is

$$\Xi_\sigma = (\mathbb{R}_+ \times \mathcal{X} \times \mathcal{X}, \ \mathcal{B} \otimes \mathcal{E} \otimes \mathcal{F}, \ \Pi = \nu \times \mu^\sigma \times Q^{\sigma,p}).$$

μ^σ represents the conditional prior distribution for p conditioned on σ^2: $\mu^\sigma \sim \mathcal{GP}(p_0, \sigma^2\Omega_0)$. $Q^{\sigma,p}$ denotes the sampling distribution conditional on both the parameters and it is characterized by the covariance operator $\frac{\sigma^2}{T}\hat{K}^*\hat{K}$.

We take, as prior distribution for the variance parameter σ^2, an *Inverse Gamma* distribution: $\sigma^2 \sim \Gamma^{-1}(v_0, s_0^2)$, with v_0 and s_0^2 two known parameters.

A conjugate model allows to easily integrate out p from the sampling distribution by using the prior μ^σ so that we obtain a sampling measure Q^σ depending only on σ^2:

$$
\begin{aligned}
\sigma^2 &\sim \Gamma^{-1}(v_0, s_0^2) \\
\hat{R}|\sigma^2 &\sim \mathcal{GP}(\hat{H}p_0, \sigma^2(\frac{1}{T}\hat{K}^*\hat{K} + \hat{H}\Omega_0\hat{H}^*)).
\end{aligned}
$$

Anyway, computation of the posterior of σ^2 is not trivial due to the fact that, because $\hat{R}$ is finite dimensional, we do not have a likelihood function. We make up for this lack by using the projected observations $\hat{R}$ projected by using the eigenfunctions associated to the covariance operator $(\frac{1}{T}\hat{K}^*\hat{K} + \hat{H}\Omega_0\hat{H}^*)$. Let $\{\hat{\lambda}_j, \hat{\varphi}_j\}_{j=1}^J$ be the eigensystem associated to this operator; this eigensystem is actually an estimation of the eigensystem associated to the true covariance operator $(\frac{1}{T}K^*K + H\Omega_0 H^*)$ that we would have if K was known. Moreover, the convergence $||(\frac{1}{T}\hat{K}^*\hat{K} + \hat{H}\Omega_0\hat{H}^*) - (\frac{1}{T}K^*K + H\Omega_0 H^*)|| \to 0$ implies that the eigensystem $\{\hat{\lambda}_j, \hat{\varphi}_j\}$ converges uniformly to the $\{\lambda_j, \varphi_j\}$. Thus, when the sample size is finite, we only have a finite number of eigenvalues $\hat{\lambda}_j$ different than 0. The projected observation $<\hat{R}, \hat{\varphi}_j>$ is normally distributed with mean and variance

$$
\begin{aligned}
\mathbb{E}(<\hat{R}, \hat{\varphi}_j>|\sigma^2) &= <\mathbb{E}(\hat{R}|\sigma^2), \hat{\varphi}_j> \\
&= <\hat{H}p_0, \hat{\varphi}_j> \\
Var(<\hat{R}, \hat{\varphi}_j>|\sigma^2) &= <Var(\hat{R}|\sigma^2), \hat{\varphi}_j> \\
&= \sigma^2 <(\frac{1}{T}\hat{K}^*\hat{K} + \hat{H}\Omega_0\hat{H}^*)\hat{\varphi}_j, \hat{\varphi}_j> \\
&= \sigma^2\hat{\lambda}_j,
\end{aligned}
$$

and $<\hat{R}, \hat{\varphi}_j>$ is independent of $<\hat{R}, \hat{\varphi}_i>, \forall j \neq i$ due to orthogonality between eigenfunctions. It should be noted that if operator K was known we would know all its eigensystem and then we would know the variance parameter σ^2, in fact $\frac{<\hat{R}-Hp_0, \varphi_j>^2}{\lambda_j}|\sigma^2 \sim \sigma^2\chi_1^2$ with mean equal to σ^2. Then, $\frac{1}{J}\sum_{j=1}^J \frac{<\hat{R}-Hp_0, \varphi_j>^2}{\lambda_j} \to \sigma^2$ and we know the limit since we know all the eigenvalues.

From classical computations we obtain the posterior distribution $\nu^\mathcal{F}$ of σ^2 given the sample $<\hat{R}, \hat{\varphi}_1>, \ldots, <\hat{R}, \hat{\varphi}_J>$:

$$
\nu(\sigma^2|\{<\hat{R}, \hat{\varphi}_j>\}_{j=1}^J) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{v_0+J}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}[s_0^2 + \sum_{j=1}^J \frac{1}{\hat{\lambda}_j}(<\hat{R}-\hat{H}p_0, \hat{\varphi}_j>)^2]\right\}
$$

then

$$
\begin{aligned}
\sigma^2|\{<\hat{R}, \hat{\varphi}_j>\}_{j=1}^J &\sim \Gamma^{-1}(v_*, s_*^2), \\
v_* = v_0 + J, \qquad & s_*^2 = s_0^2 + \sum_{j=1}^J \frac{1}{\hat{\lambda}_j}(<\hat{R}-\hat{H}p_0, \hat{\varphi}_j>)^2
\end{aligned}
$$

$$
\mathbb{E}(\sigma^2|\{<\hat{R}, \hat{\varphi}_j>\}_{j=1}^J) = \frac{s_*^2}{v_0+J-2}, \qquad Var(\sigma^2|\{<\hat{R}, \hat{\varphi}_j>\}_{j=1}^J) = \frac{\frac{s_*^4}{4}}{(\frac{v_*}{2}-1)^2(\frac{v_*}{2}-2)}.
$$

In order to compute the posterior distribution for p we first need to compute the conditional posterior distribution of p given σ^2, denoted with $\mu^{\mathcal{F},\sigma}$ and then to integrate out

σ^2 by using its posterior distribution.

Also in this case, problems of continuity of $\mu^{\mathcal{F},\sigma}$ require some technique of regularization. For simplicity, we consider only a classical Tikhonov regularization scheme. Extension to other regularization schemes is immediate. The regularized conditional posterior distribution, denoted with $\mu_\alpha^{\mathcal{F},\sigma}$ is a gaussian process with mean function and covariance operator given by:

$$
\begin{aligned}
\mathbb{E}_\alpha(p|\hat{R},\sigma^2) &= \Omega_0\hat{H}^*(\alpha I + \frac{1}{T}\hat{K}^*\hat{K} + \hat{H}\Omega_0\hat{H}^*)^{-1}(\hat{R} - \hat{H}p_0) + p_0 \\
Var_\alpha(p|\hat{R},\sigma^2) &= \sigma^2[\Omega_0 - \Omega_0\hat{H}^*(\alpha I + \frac{1}{T}\hat{K}^*\hat{K} + \hat{H}\Omega_0\hat{H}^*)^{-1}\hat{H}\Omega_0],
\end{aligned}
$$

where α still denotes the regularization parameter. While the regularized conditional posterior mean does not depend on σ^2, so that $\mathbb{E}_\alpha(p|\hat{R},\sigma^2) = \mathbb{E}_\alpha(p|\hat{R})$, the regularized conditional posterior variance does and then we need to integrate out σ^2 with respect to $\nu^{\mathcal{F}}$. With analogy to the finite dimensional case, this integration transform the posterior of p in a *Student process*. We refer to Chapter 4 for a definition of this process. Thus the marginal regularized posterior distribution $\mu_\alpha^{\mathcal{F}}$ for p is Student with parameters v_*, $\mathbb{E}_\alpha(p|\hat{R})$ and $\frac{s_*^2}{v_0+J}[\Omega_0 - A_\alpha\hat{H}\Omega_0]$:

$$
\begin{aligned}
p|\hat{R} &\sim \mathcal{S}t\mathcal{P}(\mathbb{E}_\alpha(p|\hat{R}), \frac{s_*^2}{v_*}[\Omega_0 - A_\alpha\hat{H}\Omega_0], v_*) \\
\mathbb{E}_\alpha(p|\hat{R}) &= \Omega_0\hat{H}^*(\alpha I + \frac{1}{T}\hat{K}^*\hat{K} + \hat{H}\Omega_0\hat{H}^*)^{-1}(\hat{R} - \hat{H}p_0) + p_0 \\
Var_\alpha(p|\hat{R}) &= \frac{s_*^2}{v_* - 2}[\Omega_0 - \Omega_0\hat{H}^*(\alpha I + \frac{1}{T}\hat{K}^*\hat{K} + \hat{H}\Omega_0\hat{H}^*)^{-1}\hat{H}\Omega_0].
\end{aligned}
$$

Analysis of posterior consistency of the regularized posterior distribution for p is equal to analysis performed in Section 5.4.1 and Corollary 4 holds with $\Omega_{\alpha,R}$ replaced by $Var_\alpha(p|\hat{R},\sigma^2)$.

Concerning the posterior distribution of σ^2, its posterior mean $\mathbb{E}(\sigma^2|\{< \hat{R}, \hat{\varphi}_j >\}_{j=1}^J)$ is asymptotically equivalent to $\frac{1}{J}\sum_{j=1}^J \frac{1}{\hat{\lambda}_j}(< \hat{R} - \hat{H}p_0, \hat{\varphi}_j >)^2$ and its posterior variance is asymptotically equivalent to $\frac{1}{J}(\frac{s_*^2}{J})^2$. As $T \to \infty$, $\hat{K} \to K$ and the number J of eigenfunctions becomes large. Then, $Var(\sigma^2|\{< \hat{R}, \hat{\varphi}_j >\}_{j=1}^J)$ converges to 0 and $\frac{1}{J}\sum_{j=1}^J \frac{1}{\hat{\lambda}_j}(< \hat{R} - \hat{H}p_0, \hat{\varphi}_j >)^2 \to \mathbb{E}(\frac{1}{\hat{\lambda}_j}(< \hat{R} - \hat{H}p_0, \hat{\varphi}_j >)^2) = \sigma^2$ at the parametric rate. Chebyshev's inequality implies consistency of $\nu^{\mathcal{F}}$.

Computation of eigenvalues and eigenfunction is not an easy task but it can be considerably simplified by noting that for computing posterior distribution we need to know the quantities $< \hat{R}, \hat{\varphi}_j >, j = 1, \ldots, J$ instead of the eigenfunctions $\{\hat{\varphi}_j\}$. Kernel estimation provide us with the following approximations:

$$
\begin{aligned}
\hat{R} &\approx \sum_i \sum_j M(y_i, Y_{t+1})M(y_i, y_{j+1})y_{j+1}\frac{L_h(y_i - y_j)L_h(Y_{t+1} - y_{i+1})}{\sum_l L_h(y_i - y_l)\sum_l L_h(Y_{t+1} - y_{l+1})} \\
\hat{H}p_0 &\approx \sum_i M(y_i, Y_{t+1})p_0(y_i)\frac{L_h(Y_{t+1} - y_{i+1})}{\sum_l L_h(Y_{t+1} - y_{l+1})} \\
&\quad - \sum_i \sum_j M(y_i, Y_{t+1})M(y_i, y_{j+1})p_0(y_{j+1})\frac{L_h(y_i - y_j)L_h(Y_{t+1} - y_{i+1})}{\sum_l L_h(y_i - y_l)\sum_l L_h(Y_{t+1} - y_{l+1})},
\end{aligned}
$$

where, for simplicity, we have eliminated the index $t+1$ in function M. Then,

$$
\begin{aligned}
< \hat{R} - \hat{H}p_0, \hat{\varphi}_j > \ &= \ \int (\hat{R} - \hat{H}p_0)(Y_{t+1})\hat{\varphi}_j(Y_{t+1})\pi(Y_{t+1})dY_{t+1} \\
&\approx \ \sum_i \sum_j [M(y_i, y_{j+1})(y_{j+1} + p_0(y_{j+1}))\frac{L_h(y_i - y_j)}{\sum_l L_h(y_i - y_l)} - p_0(y_i)] \\
&\qquad \int M(y_i, Y_{t+1})\frac{L_h(Y_{t+1} - y_{i+1})}{\sum_l L_h(Y_{t+1} - y_{l+1})}\hat{\varphi}_j(Y_{t+1})\pi(Y_{t+1})dY_{t+1} \\
&= \ \sum_i \sum_j \phi_j(y_i, y_{i+1})[M(y_i, y_{j+1})(y_{j+1} + p_0(y_{j+1}))\frac{L_h(y_i - y_j)}{\sum_l L_h(y_i - y_l)} - p_0(y_i)]
\end{aligned}
$$

with $\phi_j(y_i, y_{i+1}) = \int M(y_i, Y_{t+1})\frac{L_h(Y_{t+1} - y_{i+1})}{\sum_l L_h(Y_{t+1} - y_{l+1})}\hat{\varphi}_j(Y_{t+1})\pi(Y_{t+1})dY_{t+1}$. Finally, by expliciting the stochastic discount function we get

$$
\begin{aligned}
\phi_j(y_i, y_{i+1}) \ &= \ \beta\frac{1}{U'(y_i)}\bar{\phi}_j(y_{i+1}), \\
\text{with} \quad \bar{\phi}_j(y_{i+1}) \ &= \ \int U'(Y_{t+1})\frac{L_h(Y_{t+1} - y_{i+1})}{\sum_l L_h(Y_{t+1} - y_{l+1})}\hat{\varphi}_j(Y_{t+1})\pi(Y_{t+1})dY_{t+1}.
\end{aligned}
$$

Henceforth, we only need to compute $(\lambda_j, \bar{\phi}_j)$, $j = 1, \ldots, J$ that is an easier task. $\bar{\phi}_j$ is a T dimensional vector and it is the jth eigenvector of the $T \times T$ matrix $\mathcal{A}$ with (k,t)-th element

$$
\begin{aligned}
\mathcal{A}(k,t) \ &= \ \sum_i \frac{\beta}{U'(y_i)}\Big[T\int M(y_i, Y)g(Y, y_{k+1})L(x_i, x_t, Y, y_{i+1})\pi(Y)dY + \\
&\quad \sum_{i'}(\int \bar{b}(y_{i'}, Y, y_i)g(Y, y_{k+1})L(y_i, y_t, Y, y_{i'+1})\pi(Y)dY) + \\
&\quad \sum_l \sum_{l'} \int c(y_{l'}, y_{l+1}, Y)\bar{L}(y_{l'}, y_t, Y, y_{l'+1})g(Y, y_{k+1})\pi(Y)dYW(y_i, y_t, y_{i+1}, y_{l+1}) - \\
&\quad \sum_m \sum_{m'} \bar{b}(y_{m'}, y_{m+1}, y_i)\frac{L_h(y_i - y_t)}{\sum_m L_h(y_i - y_m)}\int M(y_{m'}, Y)g(Y, y_{k+1})\bar{L}(y_{m'}, y_t, Y, y_{m'+1})\pi(Y)dY - \\
&\quad T\sum_{k'} W(y_i, y_t, y_{i+1}, y_{k'+1})\int M(y_{k'+1}, Y)g(Y, y_{k+1})\frac{L_h(Y - y_{k'+1})}{\sum_l L_h(Y - y_{l+1})}\pi(Y)dY\Big],
\end{aligned}
$$

with $\bar{b}(y_{i'}, Y, y_i) = M(y_{i'}, Y)\omega(Y, y_i)$, $\omega(\cdot, \cdot)$ is the kernel of the prior covariance operator Ω_0, $c(y_{l'}, y_{l+1}, Y) = M(y_{l'}, y_{l+1})M(y_{l'}, Y)$, $g(Y, y_l) = U'(Y)\frac{L_h(Y - y_l)}{\sum_t L_h(Y - y_{t+1})}$, $\bar{L}(y_i, y_t, Y, y_{i+1}) = \frac{L_h(y_i - y_t)L_h(Y - y_{i+1})}{\sum_t L_h(y_i - y_t)\sum_t L_h(Y - y_{t+1})}$ and

$$
W(y_i, y_t, y_{i+1}, y_{l+1}) = \int \bar{b}(y_i, Y, y_{l+1})\bar{L}(y_i, y_t, Y, y_{i+1})\pi(Y)dY.
$$

Proof for obtaining this matrix are provided in the Appendix.

5.7 Conclusions

In this paper we have proposed a new bayesian nonparametric approach for estimating the solution of Euler equations. In particular, we consider the consumption-based asset pricing

model in the style of the Lucas'(1978) tree model. The aim was to estimate the equilibrium asset pricing functional and the dynamic of the state of the economy. Then, by combining these estimations, it is possible to infer the stochastic character of the equilibrium price process of a financial asset. The bayesian procedure is suitable since it offers a tractable way to introduce structural economic constraints and prior information on the estimation procedure by staying at the same time nonparametric. Moreover, it provides us with the whole posterior distribution of the pricing function. This distribution has good finite sample properties and then it can be used to construct whatever quantity, like quantiles, confidence intervals and tests.

An asset pricing model provides a characterization of the pricing functional as the solution of an integral equation of second kind that is well-posed. The bayesian approach allows to exploit the prior information on the price that we have and allows to obtain faster speed of convergence. The price to pay is the increasing of the degree of ill-posedness and the necessity of applying a regularization scheme. Substantially, the bayesian technique transforms a problem that is well-posed in a new one that is ill-posed. This is due to the compacity of the prior covariance operator.

Nevertheless, we have shown that there exists a class of prior distribution, in particular, a class of prior covariance operators, that preserves the well-posedness of the problem. In this case no further regularization technique is required and the speed of convergence of the posterior distribution towards the true value p_* is faster if p_* is highly smooth.

In order to be as general as possible, our study is based on the Lucas'(1978) model, but it can be extended to other dynamic rational expectation models with some minor modifications. Indeed, our bayesian methodology can easily treat every type of preferences as Epstein-Zin or habit preferences.

5.8   Appendix A: Proofs

Proof of Theorem 19

Let $T(\hat{F})$ denote the functional in the estimated transition distribution function $F(y_{t+1}|y_t)$ of the Markov process $\{Y_t\}$:

$$T(\hat{F}) = \int M_{t+1}(y_t, Y_{t+1})[M_{t+1}(y_t, y_{t+1})(b(y_{t+1}) + p(y_{t+1})) - p(y_t)]d\hat{F}(y_{t+1}|y_t)d\hat{F}(y_t|Y_{t+1}).$$

Note that $T(\hat{F})$ coincides with the error term U since $r + Kp = p$ and that $T(F) = 0$. We make a first order Taylor expansion of $T(\hat{F})$ around the true value F: $T(\hat{F}) - T(F) = d_1 T(F; \hat{F} - F) + R_{1T}$, where d_1 denotes the Gâteaux differential of T at F in the direction of $\hat{F}$ and R_{1T} is the rest. Let λ be a scalar and $\xi(y_t, y_{t+1}, Y_{t+1}) = M_{t+1}(y_t, Y_{t+1})[M_{t+1}(y_t, y_{t+1})(b(y_{t+1}) + p(y_{t+1})) - p(y_t)]$, then

$$
\begin{aligned}
d_1 T(F; \hat{F} - F) &= \left. \frac{d}{d\lambda} T(F + \lambda(\hat{F} - F)) \right|_{\lambda=0} \\
&= \int \xi(y_t, y_{t+1}, Y_{t+1})\hat{F}(dY_{t+1}|Y_t)F(dY_t|y_{t+1}) + \\
&\quad \int \xi(y_t, y_{t+1}, Y_{t+1})F(dY_{t+1}|Y_t)\hat{F}(dY_t|y_{t+1}) \\
&\quad -2 \int \xi(y_t, y_{t+1}, Y_{t+1})F(dY_{t+1}|Y_t)F(dY_t|y_{t+1}).
\end{aligned}
$$

Since the last two terms are null and $T(F) = 0$, we obtain that $T(\hat{F})$, and then U, is asymptotically equivalent to $\int M_{t+1}(y_t, Y_{t+1}) \int [M_{t+1}(y_t, y_{t+1})(b(y_{t+1}) + p(y_{t+1})) - p(y_t)]\hat{f}(y_{t+1}|y_t)dy_{t+1}f(y_t|Y_{t+1})dy_t$. The central integral can be approximated through a first order Taylor expansion around the true

value of F as: $\frac{1}{\pi(y_t)}[\int M_{t+1}(y_t, y_{t+1})(b(y_{t+1}) + p(y_{t+1}))\hat{f}(y_{t+1}, y_t)dy_{t+1} - p(y_t)\int \hat{\pi}(y_t)dy_t]$. Then, by substituting $\hat{f}$ and $\hat{\pi}$ with the expression for their kernel estimations we obtain:

$$U \approx \int M_{t+1}(y_t, Y_{t+1})\frac{1}{Th_t}\sum_{j=1}^{T}[M_{t+1}(y_t, y_{j+1})(b(y_{j+1}) + p(y_{j+1})) - p(y_t)]L_h(y_t - y_j)\frac{f(y_t|Y_{t+1})}{\pi(y_t)}dy_t$$

$$\approx \frac{1}{T}\sum_{j=1}^{T}M_{t+1}(y_j, Y_{t+1})[M_{t+1}(y_j, y_{j+1})(b(y_{j+1}) + p(y_{j+1})) - p(y_j)]\frac{f(y_j|Y_{t+1})}{\pi(y_j)} +$$

$$\frac{1}{T}\sum_{j=1}^{T}\int\sum_{i=1}^{\rho}\Big[\frac{\partial^i}{\partial Y_t^i}M_{t+1}(Y_t, Y_{t+1})M_{t+1}(Y_t, y_{j+1})\frac{f(Y_t|Y_{t+1})}{\pi(Y_t)}\Big|_{Y_t=y_j}(b(y_{j+1}) + p(y_{j+1}))$$

$$-\frac{\partial^i}{\partial Y_t^i}M_{t+1}(Y_t, Y_{t+1})p(Y_t)\frac{f(Y_y|Y_{t+1})}{\pi(Y_t)}\Big|_{Y_t=y_j}\Big]h^i u.$$

The second equality is obtained by making the change of variable $\frac{y_t - y_j}{h_T} = u$ and a Taylor expansion at order ρ around y_t, where ρ is the minimum among the order of the kernel, the order of differentiability of the utility function, of the transition and of the stationary density. By denoting with ϑ the second term in the previous expression, we get

$$\sqrt{T}U(Y_{t+1}) \approx \frac{\sqrt{T}}{T}\sum_{j=1}^{T}M_{t+1}(y_j, Y_{t+1})[M_{t+1}(y_j, y_{j+1})(b(y_{j+1}) + p(y_{j+1})) - p(y_j)]\frac{f(y_j|Y_{t+1})}{\pi(y_j)} + h_T^\rho\vartheta,$$

that is the expression in the theorem. Note that all the terms corresponding to h^i, with $i < \rho$ are null since they integrate to 0. When $T \to \infty$, $h \to 0$ then we can neglect the second term in $\sqrt{T}U$ and rewrite the scaled error term as $\sqrt{T}U = T^{-\frac{1}{2}}\sum_{j=1}^{T}\theta_j(Y_{t+1})$, with

$$\theta_j(Y_{t+1}) = M_{t+1}(y_j, Y_{t+1})[M_{t+1}(y_j, y_{j+1})(b(y_{j+1}) + p(y_{j+1})) - p(y_j)]\frac{f(y_j|Y_{t+1})}{\pi(y_j)}.$$

where $\theta_j(Y_{t+1})$ is a sequence of stationary Hilbert random element such that $||\theta_j(Y_{t+1})||$ is bounded with probability 1 since

$$\mathbb{E}||\theta_j(Y_{t+1})|| = \sigma^2\int M_{t+1}^2(y_j, Y_{t+1})\frac{f^2(Y_{t+1}|y_j)}{\pi^2(Y_{t+1})}\pi(Y_{t+1})\pi(y_j)dY_{t+1}dy_j < \infty.$$

This guarantees that $\sqrt{T}U$ weakly converges toward a Gaussian process, see Theorem 2.46 in Carrasco *et al.* (2007) [10]. Its expectation is equal to 0 since

$$\sqrt{T}\mathbb{E}(U(Y_{t+1})) = \int M_{t+1}(y_j, Y_{t+1})[M_{t+1}(y_j, y_{j+1})(b(y_{j+1}) + p(y_{j+1})) - p(y_j)]\frac{f(y_j|Y_{t+1})}{\pi(y_j)}f(y_j, y_{j+1})dy_jdy_{j+1}$$

$$= \int M_{t+1}(y_j, Y_{t+1})\mathbb{E}[M_{t+1}(y_j, y_{j+1})(b(y_{j+1}) + p(y_{j+1})) - p(y_j)|y_j]\frac{f(y_j|Y_{t+1})}{d}y_j$$

$$= 0$$

and the kernel $\varpi(Y_{t+1}, \tilde{Y}_{t+1})$ of its covariance operator is computed as

$$\varpi(Y_{t+1}, \tilde{Y}_{t+1}) = \frac{1}{T}Cov(\sum_{j=1}^{T}\theta_j(Y_{t+1}), \sum_{j=1}^{T}\theta_j(\tilde{Y}_{t+1}))$$

$$= Cov(\theta_j(Y_{t+1}), \theta_j(\tilde{Y}_{t+1})) + \frac{2}{T}\sum_{l>j}Cov(\theta_j(Y_{t+1}), \theta_l(\tilde{Y}_{t+1})).$$

By exploiting equality (5.7), the second term is null. Then,

$$
\begin{aligned}
\varpi(Y_{t+1}, \tilde{Y}_{t+1}) &= \int M_{t+1}(y_j, Y_{t+1}) M_{t+1}(y_j, \tilde{Y}_{t+1}) [M_{t+1}(y_j, y_{j+1})(b(y_{j+1}) + p(y_{j+1})) - p(y_j)]^2 \\
&\quad \frac{f(y_j|Y_{t+1}) f(y_j|\tilde{Y}_{t+1})}{\pi^2(y_j)} f(y_j, y_{j+1}) dy_j dy_{j+1} \\
&= \int M_{t+1}(y_j, Y_{t+1}) M_{t+1}(y_j, \tilde{Y}_{t+1}) Var[M_{t+1}(y_j, y_{j+1})(b(y_{j+1}) + p(y_{j+1})) - p(y_j)|y_j] \\
&\quad \frac{f(y_j|Y_{t+1}) f(y_j|\tilde{Y}_{t+1})}{\pi(y_j)} f(y_j, y_{j+1}) dy_j \\
&= \sigma^2 \int M_{t+1}(y_j, Y_{t+1}) M_{t+1}(y_j, \tilde{Y}_{t+1}) \frac{f(y_j|Y_{t+1}) f(y_j|\tilde{Y}_{t+1})}{\pi(y_j)} f(y_j, y_{j+1}) dy_j.
\end{aligned}
$$

The factor scaled by σ^2 is the kernel of the operator K^*K. Then, the asymptotic covariance operator associated to $\sqrt{T}U$ is asymptotically equal to $\sigma^2 K^*K$. Then, $\sqrt{T}U \Rightarrow \mathcal{GP}(0, \sigma^2 K^*K)$.

Proof of Corollary 4

The bias associated to $\mu_\alpha^{\mathcal{F}}$ can be decomposed in two terms:

$$
\hat{\mathbb{E}}_\alpha(p|\hat{R}) - p_* = (\hat{\mathbb{E}}_\alpha(p|\hat{R}) - \mathbb{E}_\alpha(p|\tilde{R})) + (\mathbb{E}_\alpha(p|\tilde{R}) - p_*),
$$

where $\mathbb{E}_\alpha(p|\tilde{R}) = \Omega_0 H^*(\alpha_T I + \Sigma_T + H\Omega_0 H^*)^{-1}(\tilde{R} - Hp_0) + p_0$ and $\tilde{R} = Hp_* + U$. The first term represent the estimation error of the operators and the second one stands for the error due to approximate the true value p_* of the asset price with the regularized posterior mean. We begin the analysis from the second term that we rewrite as:

$$
\mathbb{E}_\alpha(p|\tilde{R}) - p_* = -\overbrace{[I - \Omega_0 H^*(\alpha_T I + \Sigma_T + H\Omega_0 H^*)^{-1}H](p_* - p_0)}^{I} \\
+ \underbrace{\Omega_0 H^*(\alpha_T I + \Sigma_T + H\Omega_0 H^*)^{-1}U}_{II}.
$$

The first term can still be decomposed into two terms, in order to isolate the effect of the covariance operator Σ_T:

$$
I = \overbrace{[I - \Omega_0 H^*(\alpha I + H\Omega_0 H^*)^{-1}H](p_* - p_0)}^{IA} \\
+ \underbrace{[\Omega_0 H^*(\alpha I + \Sigma_T + H\Omega_0 H^*)^{-1}H - \Omega_0 H^*(\alpha I + H\Omega_0 H^*)^{-1}H](p_* - p_0)}_{IB}
$$

and term IA looks very similar to the regularization bias of the solution of a functional equation. More properly, to obtain such a kind of object we use the assumption that $(p_* - p_0) \in \mathcal{H}(\Omega_0)$, *i.e.* there exists a δ_* belonging to the domain of $\Omega_0^{\frac{1}{2}}$ such that we can write $(p_* - p_0) = \Omega_0^{\frac{1}{2}}\delta_*$. Therefore,

$$
\begin{aligned}
IA &= [I - \Omega_0 H^*(\alpha I + H\Omega_0 H^*)^{-1}H]\Omega_0^{\frac{1}{2}}\delta_* \\
&= [\Omega_0^{\frac{1}{2}} - \Omega_0 H^*(\alpha I + H\Omega_0 H^*)^{-1}H\Omega_0^{\frac{1}{2}}]\delta_* \\
&= \Omega_0^{\frac{1}{2}}[I - \Omega_0^{\frac{1}{2}} H^*(\alpha I + H\Omega_0 H^*)^{-1}H\Omega_0^{\frac{1}{2}}]\delta_*,
\end{aligned}
$$

where in the last equality we have used the fact that, since Ω_0 is positive definite and self-adjoint, it can be rewritten as $\Omega_0 = \Omega_0^{\frac{1}{2}}\Omega_0^{\frac{1}{2}}$. Let $B = H\Omega_0^{\frac{1}{2}}$ we take the norm in $\mathcal{X}$ of IA and after commutation of operators:

$$||IA||^2 \leq ||\Omega_0^{\frac{1}{2}}||^2||(I - (\alpha I + B^*B)^{-1}B^*B)\delta_*||^2.$$

The second norm in the right hand side of the previous expression is equal to $||\alpha(\alpha I + B^*B)^{-1}\delta_*||^2$ and it appears as the regularization bias associated to the regularized solution of the ill-posed inverse problem $B\delta_* = v$ computed using Tikhonov regularization scheme. It converges to zero when the regularization parameter α goes to zero and therefore also $||IA||^2$ converges to zero. This way to rewrite the above operator justifies the identification condition. Injectivity of $H\Omega_0^{\frac{1}{2}}$ ensures that the solution of $B\delta_* = v$ is identified and therefore, if $\Omega_0^{\frac{1}{2}}$ is injective, that $(p_* - p_0)$ is identified and that the convergence of the regularized posterior mean is towards the right true value.

The speed of convergence to zero of $||(I - (\alpha I + B^*B)^{-1}B^*B)||^2$ depends on the regularity of δ_*, and consequently of $(p_* - p_0)$. If the true solution δ_* lies in the β-regularity space Φ_β of the operator B, i.e. $\delta_* \in \mathcal{R}(\Omega_0^{\frac{1}{2}}H^*H\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, the squared regularization bias is at most of order α^β and then $||IA||^2 = \mathcal{O}_p(\alpha^\beta)$. We refer to Carrasco *et al.* (2007) [10] and Kress (1999) [50] for a proof of it.

The larger β is, the smoother the function $\delta_* \in \Phi_\beta$ will be and the faster the regularization bias will converge to zero. However, since for Tikhonov regularization scheme, β cannot be grater than 2 we implicitly assume that $\delta_* \in \Phi_\beta$ for $\beta \leq 2$.

Now, let us consider term IB:

$$\begin{aligned}
||IB||^2 &\leq ||\Omega_0 H^*||^2(\alpha I + \Sigma_T + H\Omega_0 H^*)^{-1}||^2||\Sigma_T||^2||(\alpha I + H\Omega_0 H^*)^{-1}H(p_* - p_0)||^2 \\
&\sim \mathcal{O}_p\Big(\frac{1}{\alpha^2}||\Sigma_T||^2||(\alpha I + H\Omega_0 H^*)^{-1}H(p_* - p_0)||^2\Big).
\end{aligned}$$

Since $\Sigma_T = \frac{\sigma^2}{T}K^*K$, its squared norm is $||\Sigma_T||^2 \sim \mathcal{O}_p(\frac{1}{T^2})$. Moreover, by using the regularity condition $\delta_* \in \mathcal{R}((\Omega_0^{\frac{1}{2}}H^*H\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}) \equiv \mathcal{R}((B^*B)^{\frac{\beta}{2}})$

$$\begin{aligned}
||(\alpha I + H\Omega_0 H^*)^{-1}H(p_* - p_0)||^2 &\sim ||(\alpha I + B^*B)^{-1}B\delta_*||^2 \\
&\sim ||(\alpha I + B^*B)^{-1}(B^*B)^{\frac{\beta+1}{2}}\rho_*||^2 \\
&\sim \frac{1}{\alpha^2}||\alpha(\alpha I + B^*B)^{-1}(B^*B)^{\frac{\beta+1}{2}}\rho_*||^2 \\
&\sim \mathcal{O}_p\Big(\frac{1}{\alpha^2}\alpha^{(\beta+1)\wedge 2}\Big),
\end{aligned}$$

since $||B|| = ||(B^*B)^{\frac{1}{2}}||$. Thus $||IB||^2 \sim \mathcal{O}_p\Big(\frac{1}{\alpha^4 T^2}\alpha^{(\beta+1)\wedge 2}\Big)$.

To find speed of convergence of term II we decompose it in the following equivalent way:

$$\begin{aligned}
II &= \overbrace{\Omega_0^{\frac{1}{2}}B^*(\alpha I + BB^*)^{-1}U}^{IIA} \\
&\quad + \underbrace{\Omega_0 H^*[(\alpha I + \Sigma_T + H\Omega_0 H^*)^{-1} - (\alpha I + H\Omega_0 H^*)^{-1}]U}_{IIB} \\
||IIA||^2 &\leq ||\Omega_0^{\frac{1}{2}}||^2||(\alpha I + B^*B)^{-1}B^*||^2||U||^2 \\
||IIB||^2 &\leq ||\Omega_0^{\frac{1}{2}}||^2||B^*(\alpha I + BB^*)^{-1}||^2||\Sigma_T||^2||(\alpha I + \Sigma_T + BB^*)^{-1}||^2||U||^2.
\end{aligned}$$

By Kolmogorov theorem, $||U||^2$ is bounded in probability if $\mathbb{E}||U||^2 < \infty$ and $\mathbb{E}||U||^2 = tr\Sigma_T$. Then, $||IIA||^2 \sim \mathcal{O}_p(\frac{1}{\alpha}tr\Sigma_T)$ and $||IIB||^2 \sim \mathcal{O}_p(\frac{1}{\alpha^3}||\Sigma_T||^2 tr\Sigma_T)$. Since $tr\Sigma_T \sim \mathcal{O}_p(\frac{1}{T})$ and $||\Sigma_T||^2 \sim \mathcal{O}_p(\frac{1}{T^2})$ we conclude that $||II||^2 \sim \mathcal{O}_p(\frac{1}{\alpha T} + \frac{1}{\alpha^3 T^3}) \sim \mathcal{O}_p(\frac{1}{\alpha T})$ because the second rate is negligible with respect to the first one.

Let consider now the term $(\hat{\mathbb{E}}_\alpha(p|\hat{R}) - \mathbb{E}_\alpha(p|\tilde{R}))$ due to the estimation error. We make a decomposition similar to that done before:

$$\hat{\mathbb{E}}_\alpha(p|\hat{R}) - \mathbb{E}_\alpha(p|\tilde{R}) = \overbrace{\Omega_0[\hat{H}^*(\alpha I + \Sigma_T + \hat{H}\Omega_0\hat{H}^*)^{-1}\hat{H} - H^*(\alpha I + \Sigma_T + H\Omega_0 H^*)^{-1}H](p_* - p_0)}^{A}$$
$$+ \underbrace{\Omega_0[\hat{H}^*(\alpha I + \Sigma_T + \hat{H}\Omega_0\hat{H}^*)^{-1} - H^*(\alpha I + \Sigma_T + H\Omega_0 H^*)^{-1}]U}_{B},$$

$$A = \overbrace{\Omega_0^{\frac{1}{2}}[\hat{B}^*(\alpha I + \hat{B}\hat{B}^*)^{-1}\hat{B} - B^*(\alpha I + BB^*)^{-1}B]\delta_*}^{A1}$$
$$+ \underbrace{\Omega_0^{\frac{1}{2}}[\hat{B}^*(\alpha I + \Sigma_T + \hat{B}\hat{B}^*)^{-1}\hat{B} - [\hat{B}^*(\alpha I + \hat{B}\hat{B}^*)^{-1}\hat{B}]\delta_*}_{A2}$$
$$- \underbrace{\Omega_0^{\frac{1}{2}}[B^*(\alpha I + \Sigma_T + BB^*)^{-1}B - B^*(\alpha I + BB^*)^{-1}B]\delta_*}_{A3},$$

$$B = \Omega_0^{\frac{1}{2}}[\hat{B}^*(\alpha I + \hat{B}\hat{B}^*)^{-1} - B^*(\alpha I + BB^*)^{-1}]U$$
$$+ \Omega_0^{\frac{1}{2}}[\hat{B}^*(\alpha I + \Sigma_T + \hat{B}\hat{B}^*)^{-1} - [\hat{B}^*(\alpha I + \hat{B}\hat{B}^*)^{-1}]U$$
$$- \Omega_0^{\frac{1}{2}}[B^*(\alpha I + \Sigma_T + BB^*)^{-1} - B^*(\alpha I + BB^*)^{-1}]U.$$

The norm $||A3||^2$ is equal to $||IB||^2$. Note that $||\hat{B}^*\hat{B} - B^*B||^2 \sim \mathcal{O}_p(\frac{1}{T} + h^{2\rho})$ and $||\hat{B}\hat{B}^* - BB^*||^2 \sim \mathcal{O}_p(\frac{1}{Th^n} + h^{2\rho})$, see Darolles *et al.* (2007) [15]. By using methods similar to those one used before and a Taylor expansion of $(\alpha I + \hat{B}^*\hat{B})$ around the true operator B, we get

$$||A1||^2 \sim \mathcal{O}_p\left(\left(\frac{1}{\alpha^2} + \frac{1}{\alpha^4}\left(\frac{1}{T} + h^{2\rho}\right)\right)\left(\frac{1}{T} + h^{2\rho}\right)\alpha^\beta\right)$$
$$||A2||^2 \sim \mathcal{O}_p\left(\frac{1}{T^2\alpha^4}(1 + \frac{1}{\alpha^2\left(\left(\frac{1}{Th^n} + h^{2\rho}\right)\right)})(\alpha^{(\beta+1)\wedge 2} + \left(\frac{1}{Th^n} + h^{2\rho}\right))(1 + \frac{1}{\alpha^2}\left(\frac{1}{Th^n} + h^{2\rho}\right))\right).$$

In a similar way we obtain

$$||B||^2 \sim \mathcal{O}_p(\frac{1}{\alpha^4 T^3}(1 + \frac{1}{\alpha^2}\left(\frac{1}{Th^n} + h^{2\rho}\right))(1 + \left(\frac{1}{Th^n} + h^{2\rho}\right)) + \frac{1}{\alpha T}\left(\frac{1}{Th^n} + h^{2\rho}\right)(\frac{1}{\alpha} + \frac{1}{\alpha^3}\left(\frac{1}{Th^n} + h^{2\rho}\right)) + \frac{1}{\alpha^3 T^3}).$$

Elimination of the negligible terms allows to conclude.
The procedure to obtain the rate of convergence of $\Omega_{\alpha,R}$ is equivalent, hence in this proof we only show the fundamental decomposition that we have to perform:

$$\Omega_{\alpha,R} = -\Omega_0^{\frac{1}{2}}[\hat{B}^*(\alpha I + \hat{\Sigma}_T + \hat{B}\hat{B}^*)^{-1}\hat{B} - B^*(\alpha I + \Sigma_T + BB^*)^{-1}B]\Omega_0^{\frac{1}{2}}$$
$$-\Omega_0^{\frac{1}{2}}B^*(\alpha I + \Sigma_T + BB^*)^{-1}B]\Omega_0^{\frac{1}{2}}.$$

Proof of Theorem 21

Point *(i)* follows from Chebyshev's Inequality (5.19) and results in Corollary 4.
Point *(ii)* can be obtained by Chebishev's Inequality (5.19) and by keeping the non negligible rates in $||\hat{\mathbb{E}}_\alpha(p|\hat{R}) - p_*||^2$ and in $||\Omega_{\alpha,R}||$.

Proof of Theorem 22

Write the bias $(\mathbb{E}_s(p|\hat{R}) - p_*)$ as

$$
\begin{aligned}
\mathbb{E}_s(p|\hat{R}) - p_* &= (\mathbb{E}_s(p|\hat{R}) - \mathbb{E}_s(p|\tilde{R})) + (\mathbb{E}_s(p|\tilde{R}) - p_*), \\
\mathbb{E}_s(p|\hat{R}) - \mathbb{E}_s(p|\tilde{R}) &= [\Omega_0 \hat{H}^*(\alpha L^{2s} + \hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*)^{-1}\hat{H} - \Omega_0 H^*(\alpha L^{2s} + \Sigma + H\Omega_0 H^*)^{-1}H](p_* - p_0) \\
&\quad + [\Omega_0\hat{H}^*(\alpha L^{2s} + \hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*)^{-1} - \Omega_0 H^*(\alpha L^{2s} + \Sigma + H\Omega_0 H^*)^{-1}]U, \\
\mathbb{E}_s(p|\tilde{R}) - p_* &= -[I - \Omega_0 H^*(\alpha L^{2s} + \Sigma + H\Omega_0 H^*)^{-1}H](p_* - p_0) \\
&\quad + \Omega_0 H^*(\alpha L^{2s} + \Sigma + H\Omega_0 H^*)^{-1}U.
\end{aligned}
$$

We omit computation of the rate of convergence of $(\mathbb{E}_s(p|\tilde{R}) - p_*)$ since it is given in the proof of Theorem 4 in Chapter 2. The obtained rate is:

$$
||\mathbb{E}_s(p|\tilde{R}) - p_*||^2 \sim \mathcal{O}_p(\alpha^{\frac{\beta+1}{a+s}} + \alpha^{\frac{1-a}{a+s}} tr\Sigma_T + \frac{1}{\alpha^4}||\Sigma_T||^2 \alpha^{\frac{a+\beta+2s}{a+s}} + \frac{1}{\alpha^2}||\Sigma||^2 \alpha^{\frac{1-a}{a+s}} tr\Sigma_T).
$$

Consider the estimation error $(\mathbb{E}_s(p|\hat{R}) - \mathbb{E}_s(p|\tilde{R}))$, denote $T = H\Omega_0^{\frac{1}{2}}$, the first term in it can be rewritten as:

$$
\begin{aligned}
&\Omega_0^{\frac{1}{2}}\Big(\overbrace{[\hat{T}^*(\alpha\Omega_0^{-s} + \hat{T}\hat{T}^*)^{-1}\hat{T} - T^*(\alpha\Omega_0^{-s} + TT^*)^{-1}T]\delta_*}^{A1} \\
&+ \underbrace{[\hat{T}^*(\alpha\Omega_0^{-s} + \hat{\Sigma}_T + \hat{T}\hat{T}^*)^{-1}\hat{T} - \hat{T}^*(\alpha\Omega_0^{-s} + \hat{T}\hat{T}^*)^{-1}\hat{T}]\delta_*}_{A2} \\
&- \underbrace{[T^*(\alpha\Omega_0^{-s} + \Sigma_T + TT^*)^{-1}T - T^*(\alpha\Omega_0^{-s} + TT^*)^{-1}T]\delta_*}_{A3} \Big).
\end{aligned}
$$

Let $B = T\Omega_0^{\frac{s}{2}} = H\Omega_0^{\frac{s+1}{2}}$ By commuting operators and factorizing $\Omega_0^{\frac{s}{2}}$ we get

$$
\begin{aligned}
||A1|| &= ||\Omega_0^{\frac{s+1}{2}}[(\alpha I + \hat{B}^*\hat{B})^{-1}\hat{B}^*\hat{B} - (\alpha I + B^*B)^{-1}B^*B]\Omega_0^{\frac{\beta-s}{2}}\rho_*|| \\
&= ||\Omega_0^{\frac{s+1}{2}}\Big(-[I - (\alpha I + \hat{B}^*\hat{B})^{-1}\hat{B}^*\hat{B}] + [I - (\alpha I + B^*B)^{-1}B^*B]\Big)\Omega_0^{\frac{\beta-s}{2}}\rho_*|| \\
&= ||\Omega_0^{\frac{s+1}{2}}\Big(-\alpha(\alpha I + \hat{B}^*\hat{B})^{-1} + \alpha(\alpha I + B^*B)^{-1}\Big)\Omega_0^{\frac{\beta-s}{2}}\rho_*|| \\
&= ||\Omega_0^{\frac{s+1}{2}}\alpha(\alpha I + \hat{B}^*\hat{B})^{-1}(\hat{B}^*\hat{B} - B^*B)(\alpha I + B^*B)^{-1}\Omega_0^{\frac{\beta-s}{2}}\rho_*|| \\
&\leq ||(\alpha I + \hat{B}^*\hat{B})^{-1}||_{-(s+1)}||\hat{B}^*\hat{B} - B^*B)||||(\alpha I + B^*B)^{-1}\Omega_0^{\frac{\beta-s}{2}}\rho_*||.
\end{aligned}
$$

The last norm is an $\mathcal{O}_p(\alpha^{\frac{\beta-s}{2(a+s)}})$; moreover $(\alpha I + \hat{B}^*\hat{B})^{-1} = (\alpha I + B^*B)^{-1} - (\alpha I + B^*B)^{-1}(\hat{B}^*\hat{B} - B^*B)(\alpha I + \hat{B}^*\hat{B})^{-1}$. Then, by using the Corollary 8.22 in Engl *et al.* (2000) [19]

$$
\begin{aligned}
||(\alpha I + \hat{B}^*\hat{B})^{-1}||_{-(s+1)} &\leq ||(B^*B)^{\frac{s+1}{2(a+s)}}(\alpha I + B^*B)^{-1}|| + \\
&\quad ||(B^*B)^{\frac{s+1}{2(a+s)}}(\alpha I + B^*B)^{-1}(\hat{B}^*\hat{B} - B^*B)(\alpha I + \hat{B}^*\hat{B})^{-1}|| \\
&\sim \mathcal{O}_p(\alpha^{\frac{1-2a-s}{2(a+s)}}).
\end{aligned}
$$

since the second norm is negligible once multiplied by the remaining terms of $||A1||$. It follows that $||A1||^2 \sim \mathcal{O}_p(\alpha^{\frac{\beta+1}{a+s}}\frac{1}{\alpha^2}||\hat{B}^*\hat{B} - B^*B||^2)$. Following the same logic, term $A2$ is rewritten as

$$
\Omega_0^{\frac{1}{2}}\hat{B}^*(\alpha I + \Omega_0^{\frac{s}{2}}(\hat{\Sigma}_T + \hat{T}\hat{T}^*)\Omega_0^{\frac{s}{2}})^{-1}\Sigma_T(\alpha I + \hat{B}\hat{B}^*)^{-1}\hat{B}]\delta_*
$$

that has norm of order $\mathcal{O}_p(\frac{1}{\alpha^3}||\Sigma_T||^2)$. Lastly,

$$\|A3\| \leq \|\Omega_0^{\frac{1}{2}}B^*(\alpha I + \Omega_0^{\frac{s}{2}}(\Sigma_T + TT^*)\Omega_0^{\frac{s}{2}})^{-1}\Omega_0^{\frac{s}{2}}\|\|\Sigma_T\|\|(\alpha\Omega_0^{-s} + TT^*)^{-1}T\delta_*\|,$$

$$
\begin{aligned}
\|(\alpha\Omega_0^{-s} + TT^*)^{-1}T\delta_*\| &= \|T(\alpha\Omega_0^{-s} + T^*T)^{-1}\Omega_0^{\frac{\beta}{2}}\rho_*\| \\
&= \|T\Omega_0^{\frac{s}{2}}(\alpha I + \Omega_0^{\frac{s}{2}}T^*T\Omega_0^{\frac{s}{2}})^{-1}\Omega_0^{\frac{\beta+s}{2}}\rho_*\| \\
&= \|(B^*B)^{\frac{1}{2}}(\alpha I + B^*B)^{-1}\Omega_0^{\frac{\beta+s}{2}}\rho_*\| \\
&= \|(B^*B)^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(B^*B)^{\frac{\beta+s}{2(a+s)}}v\| \\
&\sim \mathcal{O}_p(\alpha^{\frac{\beta-a}{2(a+s)}}),
\end{aligned}
$$

for some v such that $\Omega_0^{\frac{\beta+s}{2}}\rho_* = (B^*B)^{\frac{\beta+s}{2(a+s)}}v$. Such v exists since, under Assumption 26, $\mathcal{R}(\Omega_0^{a+s}) = \mathcal{R}(B^*B)$. Then, $\|A3\|^2 \sim \mathcal{O}_p(\frac{1}{\alpha^4}\|\Sigma_T\|^2\alpha^{\frac{a+\beta+2s}{a+s}})$.
The second term of $(\mathbb{E}_s(p|\hat{R}) - \mathbb{E}_s(p|\tilde{R}))$ is rewritten

$$
\begin{aligned}
&\overbrace{\Omega_0^{\frac{1}{2}}\Big([\hat{T}^*(\alpha\Omega_0^{-s} + \hat{T}\hat{T}^*)^{-1} - T^*(\alpha\Omega_0^{-s} + TT^*)^{-1}]U}^{A4} \\
+\ &\underbrace{[\hat{T}^*(\alpha\Omega_0^{-s} + \hat{\Sigma}_T + \hat{T}\hat{T}^*)^{-1} - \hat{T}^*(\alpha\Omega_0^{-s} + \hat{T}\hat{T}^*)^{-1}]U}_{A5} \\
-\ &\underbrace{[T^*(\alpha\Omega_0^{-s} + \Sigma_T + TT^*)^{-1} - T^*(\alpha\Omega_0^{-s} + TT^*)^{-1}]U}_{A6}\Big).
\end{aligned}
$$

Then,

$$
\begin{aligned}
\|A4\|^2 &= \|\Omega_0^{\frac{1}{2}}(\alpha\Omega_0^{-s} + \hat{T}^*\hat{T})^{-1}\hat{T}^* - (\alpha\Omega_0^{-s} + T^*T)^{-1}T^*]U\|^2 \\
&\leq \|\Omega_0^{\frac{s+1}{2}}(\alpha I + B^*B)^{-1}\|^2\Big(\|\hat{B}^*\hat{B} - B^*B\|^2\|(\alpha I + \hat{B}^*\hat{B})^{-1}\hat{B}^*\|^2 + \|\hat{B}^* - B^*\|^2\Big)\|U\|^2 \\
&\sim \mathcal{O}_p(\alpha^{\frac{1-2a-s}{2(a+s)}}\|\hat{B}^*\hat{B} - B^*B\|^2\frac{1}{\alpha}tr\Sigma_T + \alpha^{\frac{1-2a-s}{2(a+s)}}\|\hat{B}^* - B^*\|^2 tr\Sigma_T) \\
\|A5\|^2 &\leq \|\Omega_0^{\frac{1}{2}}\|^2\|(\alpha I + \hat{B}^*\hat{B})^{-1}\hat{B}^*\Omega_0^{\frac{s}{2}}\|^2\|\hat{\Sigma}_T\|^2\|\Omega_0^{\frac{s}{2}}(\alpha I + \Omega_0^{\frac{s}{2}}(\hat{\Sigma}_T + \hat{T}\hat{T}^*)\Omega_0^{\frac{s}{2}})^{-1}\|^2\|U\|^2 \\
&\sim (\frac{1}{\alpha^3}\|\hat{\Sigma}_T\|^2 tr\Sigma_T) \\
\|A6\|^2 &\leq \|\Omega_0^{\frac{1}{2}}T^*(\alpha\Omega_0^{-s} + TT^*)\|^2\|\Sigma_T\|^2\|(\alpha\Omega_0^{-s} + \Sigma_T + TT^*)\|\|U\|^2 \\
&\sim \mathcal{O}_p(\frac{1}{\alpha^2}\|\Sigma_T\|^2 tr\Sigma_T\alpha^{\frac{1-a}{a+s}}).
\end{aligned}
$$

Elimination of negligible terms allows to get the result.
The rate of convergence of $\|\Omega_{s,R}\|^2$ is based on specular methods and on the decomposition

$$
\begin{aligned}
\Omega_{s,R} &= -\Omega_0[\hat{H}^*(\alpha L^{2s} + \hat{\Sigma}_T + \hat{H}\Omega_0\hat{H}^*)^{-1}\hat{H} - H^*(\alpha I + \Sigma_T + H\Omega_0 H^*)^{-1}H]\Omega_0 \\
&\quad -\Omega_0 H^*(\alpha I + \Sigma_T + H\Omega_0 H^*)^{-1}H]\Omega_0.
\end{aligned}
$$

Proof of Theorem 23

Consider the decomposition

$$
\hat{\mathbb{E}}^g(p|\hat{R}) - p_* = \overbrace{[\hat{\mathbb{E}}^g(p|\hat{R}) - \tilde{\mathbb{E}}^g(p|\hat{R})]}^{I} + \overbrace{[\tilde{\mathbb{E}}^g(p|\hat{R}) - \mathbb{E}^g(p|\hat{R})]}^{II} + \overbrace{[\mathbb{E}^g(p|\hat{R}) - p_*]}^{III}.
$$

Let $W = (\hat{K}^*\hat{K})^{-\frac{1}{2}}\hat{H}(K^*K)^{\frac{s}{2}}$ and $\hat{W} = (\hat{K}^*\hat{K})^{-\frac{1}{2}}\hat{H}(\hat{K}^*\hat{K})^{\frac{s}{2}}$. Then,

$$\|I\|^2 \leq \overbrace{\|[(\hat{K}^*\hat{K})^{\frac{s}{2}}\hat{W}^*(\alpha I + \hat{W}\hat{W}^*)^{-1}(\hat{K}^*\hat{K})^{-\frac{1}{2}} - (K^*K)^{\frac{s}{2}}W^*(\alpha I + WW^*)^{-1}W(\hat{K}^*\hat{K})^{-\frac{1}{2}}]\hat{H}(p_* - p_0)\|^2}^{IA}$$
$$+ \underbrace{\|[(\hat{K}^*\hat{K})^{\frac{s}{2}}\hat{W}^*(\alpha I + \hat{W}\hat{W}^*)^{-1}(\hat{K}^*\hat{K})^{-\frac{1}{2}} - (K^*K)^{\frac{s}{2}}W^*(\alpha I + WW^*)^{-1}W(\hat{K}^*\hat{K})^{-\frac{1}{2}}]U}_{IB}$$

$$\|IA\|^2 \leq \|(\hat{K}^*\hat{K})^{\frac{s}{2}}(\alpha I + \hat{W}\hat{W}^*)^{-1}\|^2 \Big(\|\alpha(\hat{W}^* - W^*)\|^2 + \|\hat{W}*\|2\|\hat{W} - W\|^2\|W^*\|^2\Big)$$
$$\|W(\alpha I + W^*W)^{-1}(K^*K)^{\frac{\beta-s}{2}}\rho_*\|^2 + \|(\hat{K}^*\hat{K})^{\frac{s}{2}} - (K^*K)^{\frac{s}{2}}\|^2\|W^*(\alpha I + WW^*)^{-1}W(K^*K)^{\frac{\beta-s}{2}}\rho_*\|^2$$
$$\sim \mathcal{O}_p\Big((\frac{1}{Th^n} + h^{2\rho})(\frac{1}{\alpha}\alpha^{\frac{2(s-\beta)}{\beta+s}} + \frac{1}{\alpha}\alpha^{\frac{4s}{\beta}} + \alpha^{\frac{2(s-\beta)}{\beta+s}})\Big)$$
$$\sim \mathcal{O}_p\Big(\frac{1}{\alpha^2}(\frac{1}{Th^n} + h^{2\rho})\alpha^{\frac{3s-\beta}{\beta+s}}\Big),$$

since the second and third rates are negligible with respect to the first one. To get this result we have used the assumption $(p_* - p_0) \in \mathcal{R}(\Omega_0^{\frac{\beta}{2s}})$, i.e. $\exists \rho_* \in \mathcal{X}$ such that $(p_* - p_0) = (K^*K)^{\frac{\beta}{2}}\rho_*$.

$$\|IB\|^2 \leq \|(\hat{K}^*\hat{K})^{\frac{s}{2}}[\hat{W}^*(\alpha I + \hat{W}\hat{W}^*)^{-1} - W^*(\alpha I + WW^*)^{-1}](K^*K)^{-\frac{1}{2}}U\|^2$$
$$+ \|(\hat{K}^*\hat{K})^{\frac{s}{2}} - (K^*K)^{\frac{s}{2}}\|^2\|W^*(\alpha I + WW^*)^{-1}(\hat{K}^*\hat{K})^{-\frac{1}{2}}U\|^2$$
$$\sim \mathcal{O}_p\Big(\frac{1}{\alpha^2}(\frac{1}{Th^n} + h^{2\rho})\frac{1}{T}\alpha^{-\gamma}\Big).$$

Hence,

$$\|[\hat{\mathbb{E}}^g(p|\hat{R}) - \tilde{\mathbb{E}}^g(p|\hat{R})]\|^2 \sim \mathcal{O}_p\Big(\frac{1}{\alpha^2}(\frac{1}{Th^n} + h^{2\rho})(\alpha^{\frac{3s-\beta}{\beta+s}}) + \frac{1}{T}\alpha^{-\gamma}\Big).$$

Let $B = (K^*K)^{-\frac{1}{2}}H(K^*K)^{\frac{s}{2}}$ and $\hat{B} = (\hat{K}^*\hat{K})^{-\frac{1}{2}}H(K^*K)^{\frac{s}{2}}$, the second error is rewritten as:

$$\|II\|^2 \leq \overbrace{\|(K^*K)^{\frac{s}{2}}[\hat{B}^*(\alpha I + \hat{B}\hat{B}^*)^{-1}\hat{B} - B^*(\alpha I + BB^*)^{-1}B](K^*K)^{\frac{\beta-s}{2}}\rho_*\|^2}^{IIA}$$
$$+ \underbrace{\|(K^*K)^{\frac{s}{2}}[\hat{B}^*(\alpha I + \hat{B}\hat{B}^*)^{-1} - B^*(\alpha I + BB^*)^{-1}](K^*K)^{-\frac{1}{2}}U\|^2}_{IIB}$$

$$\|IIA\|^2 = \|(K^*K)^{\frac{s}{2}}(\alpha I + \hat{B}^*\hat{B})^{-1}(\hat{B}\hat{B}^* - BB^*)\alpha(\alpha I + B^*B)^{-1}(K^*K)^{\frac{\beta-s}{2}}\rho_*\|^2$$
$$\sim \mathcal{O}_p(\frac{1}{\alpha}(\frac{1}{T} + h^{2\rho})\alpha^{\frac{2s}{\beta}}).$$

$$\|IIB\|^2 \leq \|(K^*K)^{\frac{s}{2}}[\hat{B}^*(\alpha I + \hat{B}\hat{B}^*)^{-1} - B^*(\alpha I + BB^*)^{-1}](\hat{K}^*\hat{K})^{-\frac{1}{2}}U\|^2$$
$$+ \|(K^*K)^{\frac{s}{2}}\hat{B}^*(\alpha I + \hat{B}\hat{B}^*)^{-1}[(K^*K)^{-\frac{1}{2}} - (\hat{K}^*\hat{K})^{-\frac{1}{2}}]U\|^2$$
$$\sim \mathcal{O}_p\Big(\frac{1}{\alpha}(\frac{1}{T} + h^{2\rho})\frac{1}{T}\alpha^{-\gamma} + \frac{1}{T}\alpha^{-\gamma} + \frac{1}{T}\Big).$$

Then, $\|\tilde{\mathbb{E}}^g(p|\hat{R}) - \mathbb{E}^g(p|\hat{R})\|^2 \sim \mathcal{O}_p(\frac{1}{\alpha^2}(\frac{1}{T} + h^{2\rho})(\alpha^{\frac{2s+\beta}{\beta}} + \frac{1}{T}\alpha^{1-\gamma}))$ that is of the same order as $\mathcal{O}_p(\frac{1}{\alpha^2}(\frac{1}{T} + h^{2\rho})\frac{1}{T}\alpha^{1-\gamma})$. Lastly,

$$
\begin{aligned}
||\mathbb{E}^g(p|\hat{R}) - p_*||^2 \;\leq\; &\overbrace{||[I - (K^*K)^s H^*(\alpha(K^*K) + H(K^*K)^s H^*)^{-1} H(p_* - p_0)]||^2}^{IIIA} \\
&+ \underbrace{||(K^*K)^s H^*(\alpha(K^*K) + H(K^*K)^s H^*)^{-1} U||^2}_{IIIB}
\end{aligned}
$$

$$
\begin{aligned}
||IIIA||^2 \;&=\; \Big(\sup_j(\lambda_j - \frac{\lambda_j^{2s+\beta}(1-\lambda_j)^2}{\alpha + (1-\lambda_j)(\lambda_j^{2s} - \lambda_j^{2s+1})})\Big)^2 \\
&\leq\; \Big(\sup_j \frac{\alpha\lambda_j^{\beta}}{\alpha + \lambda_j^{2s}}\Big)^2 \\
&\sim\; \mathcal{O}_p\Big(\alpha^{\frac{\beta}{s}}\Big) \\
||IIIB||^2 \;&\leq\; tr(Var((K^*K)^s H^*(\alpha(K^*K) + H(K^*K)^s H^*)^{-1} U)) \\
&=\; \frac{\sigma^2}{T}\sum_j \frac{\lambda_j^{4s}(1-\lambda_j)^2}{(\alpha + (1-\lambda_j)^2\lambda_j^{2s})^2} \\
&=\; \frac{\sigma^2}{T}\sum_j \frac{\lambda_j^{2(2s-\gamma s)}}{(\alpha + (1-\lambda_j)^2\lambda_j^{2s})^2}\lambda_j^{2\gamma s} \\
&\leq\; \frac{\sigma^2}{T}\Big(\sup_j \frac{\lambda_j^{(2s-\gamma s)}}{\alpha + \lambda_j^{2s}}\Big)^2\sum_j \lambda_j^{2\gamma s} \\
&\sim\; \mathcal{O}_p\Big(\frac{1}{T}\alpha^{-\gamma}\Big).
\end{aligned}
$$

The optimal α is obtaining by equating the two rates of $||\mathbb{E}^g(p|\hat{R}) - p_*||^2$. Then, $\alpha^{\frac{\beta}{s}} = \frac{1}{T}\alpha^{-\gamma}$ if $\alpha \propto (\frac{1}{T})^{\frac{s}{(\beta+\gamma s)}}$. The corresponding optimal speed of convergence is proportional to $(\frac{1}{T})^{\frac{\beta}{\beta+\gamma s}}$. When α is set equal to the optimal one, the terms I and II go to zero if $\beta < 3s$, $\beta \geq (2-\gamma)s$ and $\frac{n}{2\rho} \leq \frac{\beta+\gamma s-2s}{\beta+\gamma s}$.
Moreover, $||(\hat{K}^*\hat{K})^{\frac{s}{2}} - (K^*K)^{\frac{s}{2}}||^2 \sim \mathcal{O}_p(\frac{1}{Th^n+h^{2\rho}})$ if $s \geq 2$.

Proof of Theorem 24

We consider the posterior variance applied to an element $\phi \in \mathcal{X}$ and its decomposition

$$
\widehat{Var}^g(p|\hat{R})\phi = \overbrace{[\widehat{Var}^g(p|\hat{R}) - \widetilde{Var}^g(p|\hat{R})]\phi}^{I} + \overbrace{[\widetilde{Var}^g(p|\hat{R}) - Var^g(p|\hat{R})]\phi}^{II} + \overbrace{Var^g(p|\hat{R})\phi}^{III}.
$$

Let $W = (\hat{K}^*\hat{K})^{-\frac{1}{2}}\hat{H}(K^*K)^{\frac{s}{2}}$ and $\hat{W} = (\hat{K}^*\hat{K})^{-\frac{1}{2}}\hat{H}(\hat{K}^*\hat{K})^{\frac{s}{2}}$. Then, for any $v \in \mathcal{X}$ such that $(K^*K)^{\frac{s}{2}}\phi = (K^*K)^{\frac{\beta-s}{2}}v$

$$
\begin{aligned}
||I||^2 \;&=\; ||(\hat{K}^*\hat{K})^s\phi - (\hat{K}^*\hat{K})^{\frac{s}{2}}\hat{W}^*(\alpha I + \hat{W}\hat{W}^*)^{-1}\hat{W}(\hat{K}^*\hat{K})^{\frac{s}{2}}\phi \\
&\quad -(K^*K)^s\phi + (K^*K)^{\frac{s}{2}}W^*(\alpha i + WW^*)^{-1}W(K^*K)^{\frac{\beta-s}{2}}v||^2 \\
&=\; ||(\hat{K}^*\hat{K})^{\frac{s}{2}}[I - \hat{W}^*(\alpha I + \hat{W}\hat{W}^*)^{-1}\hat{W}](\hat{K}^*\hat{K})^{\frac{s}{2}}\phi \\
&\quad -(K^*K)^{\frac{s}{2}}[I - W^*(\alpha I + WW^*)^{-1}W](K^*K)^{\frac{s}{2}}\phi||^2 \\
&\leq\; ||(\hat{K}^*\hat{K})^{\frac{s}{2}}\alpha(\alpha I + \hat{W}\hat{W}^*)^{-1}[(\hat{K}^*\hat{K})^{\frac{s}{2}} - (K^*K)^{\frac{s}{2}}]\phi||^2 \\
&\quad +||(\hat{K}^*\hat{K})^{\frac{s}{2}}[\alpha(\alpha I + \hat{W}^*\hat{W})^{-1} - \alpha(\alpha I + W^*W)^{-1}](K^*K)^{\frac{\beta-s}{2}}v||^2 \\
&\quad +||[(\hat{K}^*\hat{K})^{\frac{s}{2}} - (K^*K)^{\frac{s}{2}}]\alpha(\alpha I + W^*W)^{-1}(K^*K)^{\frac{\beta-s}{2}}v||^2 \\
&\sim\; \mathcal{O}_p\Big((\frac{1}{Th^n} + h^{2\rho})(1 + \alpha^{\frac{\beta-s}{2s}})\Big) \\
&\sim\; \mathcal{O}_p\Big((\frac{1}{Th^*} + h^{2\rho})(\alpha^{\frac{2s}{2s+1}} + \alpha^{\frac{2(\beta-2s-1)}{2s+1}})\Big).
\end{aligned}
$$

Let $B = (K^*K)^{-\frac{1}{2}} H (K^*K)^{\frac{s}{2}}$ and $\hat{B} = (\hat{K}^*\hat{K})^{-\frac{1}{2}} H (K^*K)^{\frac{s}{2}}$, term II is:

$$
\begin{aligned}
||II||^2 &\leq ||(K^*K)^s \phi - (K^*K)^{\frac{s}{2}} \hat{B}(\alpha I + \hat{B}\hat{B}^*)^{-1}\hat{B}(K^*K)^{\frac{s}{2}}\phi||^2 \\
&\quad + ||(K^*K)^s \phi - (K^*K)^{\frac{s}{2}} B(\alpha I + BB^*)^{-1}B(K^*K)^{\frac{s}{2}}\phi||^2 \\
&\sim \mathcal{O}_p\Big(\frac{1}{\alpha}\Big(\frac{1}{T} + h^{2\rho}\Big)\alpha^{\frac{\beta-s}{s}}\Big).
\end{aligned}
$$

Lastly, $||III||^2 = ||(K^*K)^{\frac{s}{2}}\alpha(\alpha I + B^*B)^{-1}(K^*K)^{\frac{\beta-s}{2}}v||^2$ that is of order $\mathcal{O}_p(\alpha^{\frac{\beta}{s}})$.

Computation of the Eigensystem for Section 5.6

In this appendix we prove that the eigensystem $\{\lambda_j, \bar{\phi}_j\}$, necessary for obtaining the posterior distribution in Section 5.6, can be computed as the eigensystem associated to matrix $\mathcal{A}$. We start by explicitating the estimated elements of $(\frac{1}{T}\hat{K}\hat{K}^* + \hat{H}\Omega_0\hat{H}^*)$. Note that $\hat{K}\hat{\varphi}_j \approx \int M(y_i, Y)\hat{\varphi}_j \frac{\hat{f}(y_i, Y)}{\pi(\hat{y}_i),\pi(\hat{Y})}\pi(Y)dY$. By remembering the definition of ϕ_j, we have:

$$
\begin{aligned}
\hat{K}\hat{\varphi}_j &= T\sum_t \phi_j(y_i, y_{t+1})\frac{L_h(y_i - y_t)}{\sum_t L_h(y_i - y_t)} \\
\hat{K}^*\hat{K}\hat{\varphi}_j &= T\sum_t \sum_i M(y_i, Y)\phi_j(y_i, y_{t+1})\bar{L}(y_i, y_t, Y, y_{i+1}) \\
\hat{H}\Omega_0\hat{H}^* &= \hat{K}^*\Omega_0\hat{K} + \hat{K}^*\hat{K}\Omega_0\hat{K}^*\hat{K} - \hat{K}^*\hat{K}\Omega_0\hat{K} - \hat{K}^*\Omega_0\hat{K}^*\hat{K} \\
\hat{K}^*\Omega_0\hat{K}\hat{\varphi}_j &= T\sum_t \sum_i \sum_{i'} M(y_{i'}, Y)\omega(y_i, Y)\phi_j(y_i, y_{t+1})\bar{L}(y_i, y_t, Y, y_{i'+1})\pi(y)dy \\
\hat{K}^*\hat{K}\Omega_0\hat{K}^*\hat{K} &= \sum_t \sum_i \sum_l \sum_{l'} M(y_{l'}, y_{l+1})M(y_{l'}, Y)\bar{L}(y_{l'}, y_t, Y, y_{l'+1})\phi_j(y_i, y_{t+1}) \\
&\quad \int M(y_i, y)\omega(y, y_{l+1})\bar{L}(y_i, y_t, y, y_{i+1})\pi(y)dy \\
\hat{K}^*\hat{K}\Omega_0\hat{K} &= \sum_t \sum_i \sum_m \sum_{m'} M(y_{m'}, y_{m+1})\omega(y_{m+1}, y_i)\frac{L_h(y_i - y_t)}{\sum_m L_h(y_i - y_m)}M(y_{m'}, Y) \\
&\quad \bar{L}(y_{m'}, y_t, Y, y_{m'+1})\phi_j(y_i, y_{t+1}) \\
\hat{K}^*\Omega_0\hat{K}^*\hat{K} &= \sum_t \sum_i \sum_{k'} M(y_{k'+1}, Y)\frac{L_h(Y - y_{k'+1})}{\sum_l L_h(Y - y_{l+1})}\int M(y_i, y)\omega(y, y_{k'+1}) \\
&\quad \bar{L}(y_i, y_t, y, y_{i+1})\pi(y)dy\phi_j(y_i, y_{t+1}).
\end{aligned}
$$

Then, $(\frac{1}{T}\hat{K}\hat{K}^* + \hat{H}\Omega_0\hat{H}^*)\hat{\varphi}_j = \hat{\lambda}_j\hat{\varphi}_j$. By taking the integral $\int U'(Y)\frac{L_h(Y-y_{k+1})}{\sum_k L_h(Y-y_{k+1})}\pi(Y)dY$ on both sides of this equality, and developing $\phi_j(y_i, y_{y+1}) = \beta\frac{1}{U'(y_i)}\bar{\phi}_j(y_{t+1})$, we get $\mathcal{A}_k\varphi_j = \hat{\lambda}_j\bar{\phi}_j(y_{k+1})$, where $\mathcal{A}_k$ denotes the $(k+1)$-th row of $\mathcal{A}$, for $k = 0, \ldots, T-1$.

5.9   Appendix B: Numerical Implementation

We present in this subsection a numerical simulation able to show the good properties of our estimator. For simplicity, we take $n = 1$, so that only 1 consumption good is present in the economy. The law of motion for the relevant state variable Y_t is

$$\ln Y_t = a + b\ln Y_{t-1} + \epsilon,$$

where ϵ is a normal random variable with variance 0.01. The agent's per-period utility function is of CRR type: $U(Y_t) = \frac{Y_t^{(1-\gamma)}}{1-\gamma}$, with $\gamma = 0.30$. We chose the agent's subjective discount factor $\beta = 0.97$.
The true value of the pricing functional is taken as the function satisfying equation (5.8) and it is obtained through the classical method described in subsection 5.4.3. This choice is motivated by

the small dimension n. In this situation the classical solution is likely to converge faster than the bayesian solution.

The transition density of the state variable is estimated through a kernel smoothing with a gaussian kernel function and a bandwidth $h = 0.1$. The prior distribution is specified as a gaussian measure with mean set alternatively equal to $p_0 = 525Y_t^2 - 857.5Y_t + 373$ or to $p_0 = 160Y_t - 108$. The prior covariance operator is $\Omega_0 = \int \exp\{-|\tilde{Y} - \wedge Y_t|\}\pi(\tilde{Y})d\tilde{Y}$. We show the results of the simulation in Figure 5.1 for two values of the regularization parameter α: $\alpha = 0.3$ and $\alpha = 1$. The magenta curve is the prior mean. The blue curve is the classical solution $\hat{p} = (I - \hat{K})\hat{r}$, the red one is the regularized posterior mean, regularized through the classical Tikhonov scheme. The difference between this two curves gives a measure of how the bayesian method fit well.

(a) $\alpha = 0.3$, $N = 1000$ (b) $\alpha = 1$, $N = 1000$

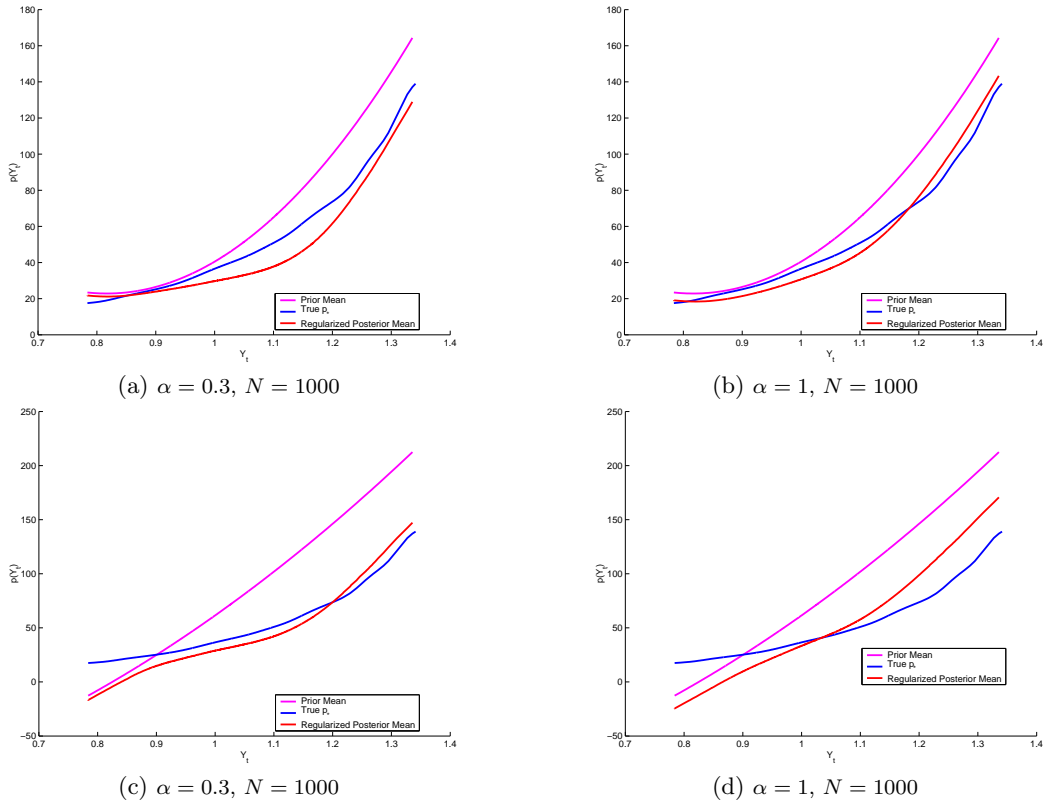(c) $\alpha = 0.3$, $N = 1000$ (d) $\alpha = 1$, $N = 1000$

Figure 5.1: Asset Pricing functional estimation.

In Figure 5.2 we have used the *extended g-prior* distribution with $g = T\alpha$, $T = 1000$ and different values of α are alternatively specified. The covariance operator is $\Omega_0 = (K^*K)^s$, with $s = 1$.

(a) $\alpha = 0.3$, $N = 1000$,
$p_0 = 525Y_t^2 - 857.5Y_t + 373$

(b) $\alpha = 0.2$, $N = 1000$,
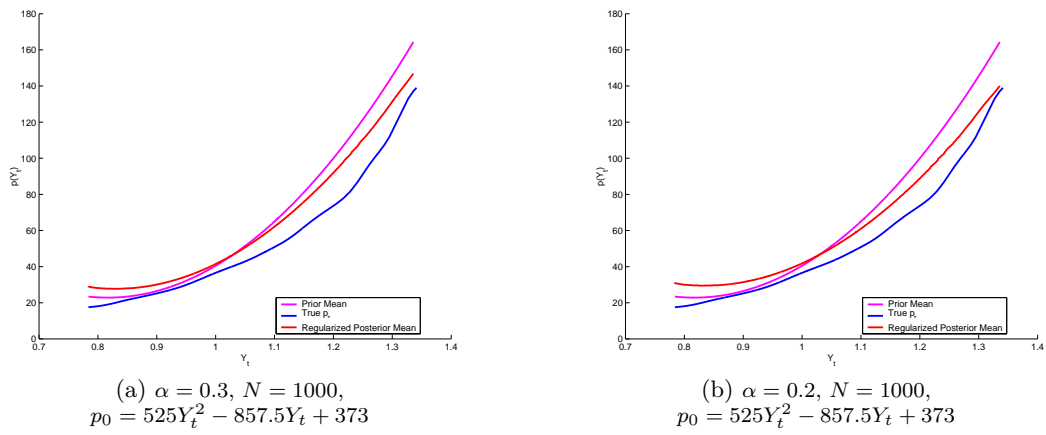$p_0 = 525Y_t^2 - 857.5Y_t + 373$

Figure 5.2: Asset Pricing functional estimation with an *extended g-prior* specification.

References

[1] Agliari, A. and C.C. Parisetti (1988), *A g-reference Informative Prior: a Note on Zellner's g-prior*, The Statistician, Vol.37, 3, 271 - 275.

[2] Andersen, P.K., Borgan, O., Gill, R.D. and N. Keiding (1993), *Statistical Models Based on Counting Processes*, Springer-Verlag

[3] Baker, C.R. (1973), *Joint Measures and Cross-Covariance Operators*, Transactions of the American Mathematical Society, 186, 273-289.

[4] Bansal, R. and A., Yaron (2004), *Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles*, The Journal of Finance, Vol.59, 4, 1481 - 1509.

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

[6] Bidarkota, P.V. and H., McCulloch (2003), *Consumption Asset Pricing with Stable Shocks - Exploring a Solution and its Implications for Mean Asset Returns*, Journal of Economic Dynamics and Control, 27, 399 - 421.

[7] Blackwell, D. (1965), *Discounted Dynamic Programming*, The Annals of Mathematical Statistics, Vol.36, No. 1, 226-235.

[8] Burnside, C. (1998), *Solving Asset Pricing Models with Gaussian Shocks*, Journal of Economic Dynamics and Control, 22, 329 - 340.

[9] Calin, O.L., Chen, Y. and T.F., Cosimano (2995), *Solving Asset Pricing Models when the Price-Dividend Function is Analytic*, Econometrica, Vol.73, 3, 961 - 982.

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

[11] Chalmond, B. (2003), *Modeling and Inverse Problems in Image Analysis*, Springer.

[12] Chen, X., Favilukis, J. and S.C., Ludvigson (2008), *An Estimation of Economic Models with Recursive Preferences*, working paper.

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

[14] Cox, J.C., Ingersoll, J.E. and S.A., Ross (1985), *An Intertemporal General Equilibrium Model of Asset Prices*, Econometrica, Vol. 53, 2, 363 - 384.

[15] Darolles, S., Florens, J.P., and E., Renault (2006), *Nonparametric Instrumental Regression*, under revision for Econometrica.

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

[17] Doob, J.L. (1949), *Application of the Theory of Martingales*, in Le Calcul des Probabilités et ses Applications, Colloques Internationaux du Centre National de la Recherche Scientifique, 13, 23-27.

[18] Dykstra, R.L. and P.W., Laud (1981), *A Bayesian Nonparametric Approach to Reliability*, The Annals of Statistics, 9, 356-367.

[19] Engl, H.W., Hanke, M. and A., Neubauer (2000), *Regularization of Inverse Problems*, Kluwer Academic, Dordrecht.

[20] Epstein, L.G. and S.E., Zin (1991), *Substitution, Risk Aversion, and the Temporal Behavior of Consumption Growth and Asset Returns II: An Empirical Analysis*, The Journal of Political Economy, 99, 2, 263 - 286.

[21] Escobar, M.D. and M., West (1995), *Bayesian Density Estimation and Inference Using Mixtures*, Journal of the American Statistical Association, Vol. 90, 430, 577-588.

[22] Fair, R.C. and J.B., Taylor (1980), *Solution and Maximum Likelihood Estimation of Dynamic Nonlinear Rational Expectation Models*, Cowles Foundation Discussion Paper No. 564.

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

[24] Ferguson, T.S. (1974), *Prior Distributions on Spaces of Probability Measures*, The Annals of Statistics, Vol.2, 4, 615-629.

[25] Ferguson, T.S. and E.G. Phadia (1979), *Bayesian Nonparametric Estimation Based on Censored Data*, Annals of Statistics, 7, 163-186.

[26] Florens, J.P. (2003), *Inverse Problems and Structural Econometrics: The Example of Instrumental Variables*, in *Advances in Economics and Econometrics*, Vol.2, eds. by M. Dewatripont, L.P. Hansen and S.J. Turnovsky, Cambridge University Press, 284 - 311.

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

[28] Florens, J.P., Mouchart, M. and J.M., Rolin (1992) *Bayesian Analysis of Mixture: Some Results on Exact Estimability and Identification*, in Bayesian Statistics IV, edited by J.Bernardo, J.Burger, D. Lindley and A.Smith, North Holland, 127-145.

[29] Florens, J.P., and A., Simoni (2008), *Nonparametric Estimation of Instrumental Regression: a Bayesian Approach Based on Regularized Posterior*, mimeo.

[30] Florens, J.P., and A., Simoni (2008), *Regularized Posteriors in Linear Ill-Posed Inverse Problems*, mimeo.

[31] Florens, J.P., and A., Simoni (2008), *On the Regularization Power of the Prior Distribution in Linear ill-Posed Inverse Problem*, mimeo.

[32] Florens, J.P., and A., Simoni (2008), *Regularized Posteriors in Linear Ill-Posed Inverse Problems: Extensions*, mimeo.

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

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

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

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

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

[38] Hall, P. and J.L., Horowitz (2005), *Nonparametric Methods for Inference in the Presence of Instrumental Variables*, Annals of Statistics, Vol.33, n.6, 2904-2929.

[39] Hansen, L.P., and T.J., Sargent (1980), *Formulating and Estimating Dynamuc Linear Rational Expectations Models*, Journal of Economic Dynamics and Control, 2, 7 - 46.

[40] Hansen, L.P., and T.J., Sargent (1981), *Linear Rational Expectations Models for Dynamically Interrelated Variables*, in Rational Expectation and Economic Practice, ed. by R.E. Lucas, Jr. and T.J. Sargent. Minneapolis: University of Minnesota Press, 1981.

[41] Hansen, L.P., and K.J., Singleton (1982), *Generalized Instrumental Variables Estimation on Nonlinear Rational Expectations Models*, Econometrica, vol. 50, 5, 1269 - 1286.

[42] Hansen, L.P., and K.J., Singleton (1983), *Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns*, the Journal of Political Economy, vol. 91, 2, 249 - 265.

[43] Hanson, T. and W.O., Johnson (2002), *Modeling Regression Error With a Mixture of Polya Trees*, Journal of the American Statistical Association, 97, 1020-1033.

[44] Hayashi, F. (1980), *Estimation Macroeconomic Models Under Rational Expectations: A Survey*, Manuscript, Northwestern University, 1980.

[45] Hiroshi, S. and O., Yoshiaki (1975), *Separabilities of a Gaussian Measure*, Annales de l'I.H.P., section B, tome 11, 3, 287 - 298.

[46] Hjort, N.L. (1990), *Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data*, The Annals of Statistics, Vol.18, 3, 1259-1294.

[47] Hjort, N.L. (1996), *Bayesian Approaches to Non- and Semiparametric Density Estimation*, Bayesian Statistics 5 (J.M. Bernardo *et al.*, eds.), 223-253.

[48] Ishwaran, H. and L., James (2004), *Computational Methods for Multiplicative Intensity Models Using Weighted Gamma Processes: Proportional Hazards, Marked Point Processes, and Panel Count Data*, Journal of the American Statistical Association, Vol.99, 465, 175-190.

[49] Kaipio, J., and E., Somersalo (2004), *Statistical and Computational Inverse Problems*, Applied Mathematical Series, Vol.160, Springer, Berlin.

[50] Kress, R. (1999), *Linear Integral Equation*, Springer.

[51] Kuo, H.H. (1975), *Gaussian Measures in Banach Spaces*, Springer.

[52] Lvine, M. (1992) *Some Aspects of Polya Tree Distributions for Statistical Modelling*, The Annals of Statistics, Vol.20, 3, 1222-1235.

[53] Lehtinen, M.S., Päivärinta, L. and E., Somersalo (1989), *Linear Inverse Problems for Generalised Random Variables*, Inverse Problems, 5, 599-612.

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

[55] Lenk, P.J. (1988), *The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities*, Journal of the American Statistical Association, Vol.83, 402, 509-516.

[56] LeRoy, S.F. (1973), *Risk Aversion and the Martingale Property of Stock Prices*, International Economic Review, Vol.14, No.2, 436-446.

[57] Linton, O.B. and E., Mammen (2005), *Estimating Semiparametric ARCH(∞) Models by Kernel Smoothing Methods*, Econometrica, Vol.73, 771 - 836.

[58] Lucas, R.E. (1976), *Econometric policy evaluation: A critique*, in: K. Brunner and A.H. Meltzer, eds., The Phillips curve and labor markets, Carnegie-Rochester conferences on public policy 1 (North Holland, Amsterdam).

[59] Lucas, R.E. (1978), *Asset Prices in An Exchange Economy*, Econometrica, Vol.46, No.6, 1429-1445.

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

[61] Muth, J.F. (1961), *Rational Expectations and the Theory of Price Movements*, Econometrica, 29, 41 - 49.

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

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

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

[65] Prenter, P.M. and C.R., Vogel (1985), *Stochastic Inversion of Linear First Kind Integral Equations. I. Continuous Theory and the Stochastic Generalized Inverse*, Journal of Mathematical Analysis and Applications, 106, 202 - 212.

[66] Petrone S. (1999) *Bayesian Density Estimation Using Bernstein Polynomials*, The Canadian Journal of Statistics, Vol.27, 1, 15-126.

[67] Rasmussen, C.E. and C.K.I., Williams (2006), *Gaussian Processes for Machine Learning*, The MIT Press.

[68] Roussas, G.G. (1967), *Nonparametric Estimation in Markov Processes*, Ann. Inst. Stat. Math., 21, 73 - 87.

[69] Ruggiero, M. (1994), *Bayesian Semiparametric Estimation of Proportional Hazards Models*, Journal of Econometrics, 62, 277 - 300.

[70] Rust, J.B., Traub, J.F. and H., Wozniakowski (2002), *Is There a Curse of Dimensionality for Contraction Fixed Points in the Worst Case?*, Econometrica, 70, No.1, 285 - 329.

[71] Sargent, T.J. (1981), *Interpreting Economic Time Series*, Journal of Political Economy, 89, 213 - 248.

[72] Simoni, A. (2008), *A Note on Admissibility of the Regularized Posterior Mean Estimator for an Hilbert-valued Parameter*, mimeo.

[73] Simoni, A. (2008), *Bayesian Nonparametric Estimation of Asset Pricing Functionals*, mimeo.

[74] Smith, M. and R., Kohn (1996), *Nonparametric Regression Using Bayesian Variable Selection*, Journal of Econometrics, 75, 317-343.

[75] Tauchen, G. and R., Hussey (1991), *Quadrature Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models*, Econometrica, 59, 371 - 396.

[76] Tsionas, E.G. (2003), *Exact Solution of Asset Pricing Models with Arbitrary Shock Distributions*, Journal of Economic Dynamics and Control, 27, 843 - 851.

[77] Van der Vaart, A.W., and J.H., Van Zanten (2000), *Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors*, Working paper.

[78] Vapnik, V.N. (1998), *Statistical Learning Theory*, John Wiley & Sons, Inc.

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

[80] Walker, S. and P., Muliere (1997), *Beta-Stacy Processes and a Generalization of the Polya-Urn Scheme*, The Annals of Statistics, Vol. 25, 4, 1762-1780.

[81] Zellner, A. (1986), *On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distribution*, in: Goel, P.K. and Zellner, A. (Eds) Bayesian Inference and Decision Techniques: essays in honour of Bruno de Finetti, pp. 233-243 (Amsterdam, North Holland).

[82] Zellner, A. (1986), *An Introduction to Bayesian Inference in Econometrics*, Wiley Classics, 1996.