Regularized Posteriors in Linear Ill-posed Inverse Problems

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

Anna Simoni
Toulouse School of Economics
(GREMAQ)

Preliminary and Incomplete Version

Abstract

The aim of the paper is to obtain a solution for a signal-noise problem, namely we want to make inference on an unobserved infinite dimensional parameter through a noisy indirect observation of this quantity. The 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 place us in a Bayesian framework so that the parameter of interest is a stochastic process and the solution to the inference problem is the posterior distribution of such a parameter. We characterize a regular version of the posterior distribution in functional spaces, but due to the infinite dimension of our problem it is only possible to compute a regularized version of it. We call it regularized posterior distribution and we guess it is the solution of the inverse problem.

We study asymptotic properties of this solution and we prove that, if the true value of the parameter of interest satisfies a regularity condition, it is consistent in a "frequentist" sense. However, the prior is inconsistent since the prior distribution we have specified is not able to generate the true value of the parameter for which posterior consistency is verified. This result perfectly agrees with previous literature and confirms once again the possible prior inconsistency in infinite-dimensional Bayesian experiments already stressed by Diaconis and Freedman (1986).

Furthermore, we study asymptotic properties for the case in which the operator in the functional equation is unknown.

Lastly, we consider a simple extension of the basic model that account for the case in which the operator is specific to each observation. To solve this case, we first determine a sufficient statistic for the infinite dimensional parameter and then we marginalize the Bayesian experiment on the sample space.

Monte Carlo simulations of several examples confirm good properties of the proposed estimator.

JEL codes: C11, C14.
Keywords: Gaussian Process, Conditional Probability, Tikhonov regularization, Sufficient Statistics, Posterior Consistency.

1 Introduction

In this paper we deal with solving a functional equation describing an implicit relation between the object of interest x and the observed one Y:

$$Y = Kx, \qquad x \in \mathcal{X}, Y \in \mathcal{Y} \tag{1}$$

where K is an operator that uniquely maps elements of the set $\mathcal{X}$ onto elements of $\mathcal{Y}$. Such a kind of problem is known in literature as an *inverse problem* and it can be encountered in very different fields. In physics, for instance, in *time resolved fluorescence* problem x represents the intensity of the light-pulse and Y is the observed fluorescence intensity. In signal and image processing, the deblurring image problem amounts to recover the true image x through the blurred image Y. Numerous examples of inverse problems can be also found in statistics, for example the classical example of density estimation where x stands for the density and Y for the cumulative distribution function, see Vapnik 1998 [32]. We will describe this and other applications more deeply in the following of the paper. Further applications of inverse problems in economics are provided by recent literature, see Carrasco, Florens and Renault (2007) [4], that proposes to interpret many problems in structural econometric as inverse problems where the implicit relation is between some functional parameter x of interest and the *cumulative distribution function* F determining the data generating process (DGP).
In most of applications, $\mathcal{X}$ and $\mathcal{Y}$ are infinite dimensional spaces and K is a compact integral operator so that equation (1) is known as *Fredholm Integral Equation of type I*. Otherwise, if $K = I - A$, where I and A are the identity and the integral operator, respectively, equation (1) is known as *Fredholm Integral Equation of type II*. Given the huge number of problems that are linear in x, we only deal with linear inverse problems in this paper. However, it is important to point out that there exists a lot of inverse problems that are nonlinear. The resolution technique that we propose cannot be extended to nonlinear models since the property of linearity is directly exploited in order to compute the posterior distribution of X.
Following the traditions, the analysis in this paper is formulated in terms of Hilbert spaces, so that $\mathcal{X}$ and $\mathcal{Y}$ are two separable real or complex Hilbert spaces and K is supposed to be an Hilbert-Schmidt operator. Equation (1) must be solved in x. If an unique solution exists and depends continuously on data, namely *Hadamard's conditions* are satisfied, the problem is said a *well-posed inverse problem*. On the contrary, the inverse problem is ill-posed if at least one of the Hadamard's conditions is not fulfilled, see Engl, Hanke and Neubauer(2000) [8]. Actually, these are conditions on the operator K so that well-posedness is obtained if the operator is bijective and has a continuous inverse. In finite dimension, linear operators are continuous, but this property no longer holds in infinite dimension, this is for instance the case of compact integral operators. Moreover, il K is continuous and one to one, the range of K, $\mathcal{R}(K)$, can be equal to $\mathcal{Y}$ only if $\mathcal{Y}$ is finite dimensional.
Ill-posedness of an inverse problem is also related to which data are considered admissible. In general we do not observe the exact transformation Y of x but a noisy transformation $\hat{Y}$. This implies, first of all, that we cannot be sure that $\hat{Y} \in \mathcal{R}(K)$, where $\mathcal{R}(\cdot)$ denotes the range of an operator, so the assumption of surjectivity is violated. Moreover, if K^{-1} is not continuous then small observation errors in Y cause large deviations in the solution $\hat{x}$. This can be understood by considering the inequality

$$||\hat{x} - x|| \leq ||K^{-1}||||\hat{Y} - Y||,$$

in which $||K^{-1}||$ denotes the norm of K^{-1}. It is clear that, if $||K^{-1}||$ is very large, errors may be strongly amplified by the action of K^{-1}. From a mathematical point of view we can always make an inverse problem well-posed relative to new spaces and new topology. However, this way to cope with problem (1) is artificial and the most of times it is not appropriate for the concrete problem we are analyzing, so that more correct techniques have to be used.
The general idea, in order to solve an ill-posed inverse problem, is to restate the problem in such a way that the *Hadamard's conditions* be satisfied. Restating the problem always implies reducing one's ambition: it is not possible to restore all the information that an ideal solution would carry. The important thing is then to find the right trade-off between the quantity of information to be

retrieved and the level of accuracy. Two different approaches to face an ill-posed inverse problem have been proposed in literature: the classical approach and the Bayesian one. The classical approach solves the lack of bijectivity of K by looking for the best approximate solution. In other words, equation (1) is transformed in a generalized inverse problem: $\hat{x} = \arg\min_{x \in \mathcal{X}} ||Y - Kx||^2$ and we keep the solution with minimal norm. This is equivalent to compute the Moore-Penrose generalized inverse of K. However, the generalized inverse is generally not continuous and it has to be regularized. Regularization techniques consist in replacing the sequence of explosive inverse singular values $\frac{1}{\mu_j}$ associated to operator K by a sequence of bounded inverse singular values $\frac{q(\alpha,\mu_j)}{\mu_j}$ that are asymptotically unbiased (*i.e.* $\lim_{\alpha \to 0} q(\alpha, \mu_j) = 1$, $\forall j$), see Kress (1999, Theorem 15.21) [22]. The most common regularization schemes are the *Spectral cut-off*, the *Landweber-Fridman* scheme, the *Tikhonov regularization* and the *iterated Tikhonov regulariztion*.
A completely different approach is the Bayesian approach that interprets every quantity in the problem as a random variable or a random function, so that both the parameter of interest x and the observed quantity Y induce some measure on the Hilbert spaces to which they belong. In Bayesian analysis, a functional equation of the type of problem (1) is called a *Bayesian Statistical Inverse Problem*. From a Bayesian Statistical Inversion Theory point of view, the solution to an inverse problem is the posterior distribution of the quantity of interest. In other words, Bayesian analysis considers an inverse problem in a different way with respect to the classical analysis since it restates functional equation (1) in a larger space of probability distributions. The object of interest is also changed: we are no more interested in a punctual estimation of x, but in a distribution that incorporates both the information a priori and in the sample. After which, this distribution can be exploited to obtain a punctual estimation of x.
We follow in this paper a Bayesian approach and we propose a solution to (1) for the Gaussian case by directly working on functional spaces. Before performing the measurement of Y, we suppose to have some information about x that we summarize by a *prior distribution* on the parameter space, and for a given realization of x we know the *sampling probability* of Y, namely the conditional distribution of Y conditioned to the σ-field of subsets of the parameter space. More precisely, to well describe noisy data $\hat{Y}$, an error term U has to be explicitly incorporated in (1) and the sampling probability is determined by the measure induced by this observation error. As usual in Bayesian analysis, we incorporate the error term additively, so that analysis can be simplified by using a conjugate prior distribution.
Previous works in bayesian ill-posed problems literature have considered equations of type (1) in finite dimensional spaces, see for instance examples given in Kaipio and Somersalo (2004) [21]. In finite dimension, ill-posedness is principally due to a problem of multicollinearity and the most commonly used method to solve this problem is Tikhonov regularization, also known as *ridge regression*. In this framework, classical and Bayesian approach are strongly related since Tikhonov regularization method can be justified from a Bayesian point of view. We can assume that both the prior and the sampling measures are gaussian with spherical variances. Under these assumptions, the Tikhonov regularized solution coincides with the posterior mean for a regularization parameter equal to the ratio of the noise and prior variance. Therefore, in finite dimension we can remove ill-posedness by incorporating the available prior information.
On the contrary, in infinite dimension, Bayesian approach does not solve the ill-posedness of the problem since covariance matrices do not have the regularization properties that have in the finite dimensional case. In particular, being impossible to continuously inverse the covariance matrices we still need some regularization technique and the bayesian approach only lies in changing the nature of the problem. In most of the literature functional equations in infinite dimensional spaces are addressed by projecting the model. Instead of observing the whole curve Y, only projections of it on an orthonormal basis of the functional space to which Y belongs are observed. Zhao (2000) studies nonparametric regression and white noise model and considers asymptotic properties of Bayes estimators of Fourier coefficients of the function of interest. She proves that, with independent normal conjugate priors, Bayes estimator is consistent and attains the optimal minimax rate. What is unappealing is that these priors are not supported on the assumed Sobolev parameter space. It is proved that the problem is solved by considering priors that are mixture of Gaussian priors. The same result is found in Belitser and Ghosal (2003). They also consider a hierarchical prior because of the dependence of the prior on a unknown "smoothness parameter".

A peculiarity of our study is to consider infinite dimensional spaces and parameters and to

make inference directly on them without projecting on any basis. We develop a model that can be applied to the case in which we observe the whole curve, (i.e. we have continuous observations) or to the case in which we have discrete observations but a way to transform them in an infinite dimensional object is available. The first question that arise when we apply Bayes theorem in infinite dimensional spaces is the existence of a regular version of the posterior distribution. Thank to *Jirina Theorem* we can state its existence, since we assume to work in Polish spaces and then conditions entailing the existence of a regular conditional probability are satisfied.
However, when we try to compute the posterior distribution we are faced with some problem due to infinite dimension of the space. In particular, the operator that characterizes the posterior mean results to be the solution of a further inverse problem that again results to be ill-posed. This justifies the application of a regularization scheme to this functional equation and the characterization of a new object that we call Regularized Posterior Distribution. We guess this regularized version of the posterior measure is the solution of our inverse problem. The regularized scheme that we use is a Tikhonov regularization scheme.

We also analyze consistency of the regularized posterior distribution in a "frequentist" sense. This means that we are supposing that there exists a true value of the parameter, as in classical statistics, and we check whether the regularized posterior distribution degenerates to a probability mass in a neighborhood of the true value as the number of observations grows indefinitely. Under the assumption that the true value of the parameter of interest belongs to the Reproducing Kernel Hilbert Space associated to the prior covariance operator we prove posterior consistency. Anyway, the supposed prior distribution is not able to generate a trajectory satisfying this regularity condition. This result is in line with the previous literature and is due to the infinite dimension of the parameter of interest and to the impossibility for the support of the prior to cover all the domain of the true parameter. The support of the prior distribution is the closure of the Reproducing Kernel Hilbert Space associated to its covariance operator and so it is possible for the prior distribution to generate a trajectory as close as we want to the true value of the parameter.

The paper is developed as follow. In Section 2 we present the model and the associated Bayesian experiment. Some example of possible applications is given, in particular we apply bayesian inversion analysis to density and regression function estimation. In Section 3 we study the existence, continuity and computability of the posterior distribution. To overcome some problems of continuity and computability, a regularized version of the posterior distribution is defined and it is assumed to be the solution of the inverse problem. Frequency asymptotic properties of this solution are studied in Section 4. After enouncing conditions under which we would have posterior consistency, we prove the inconsistency of the prior distribution, namely the impossibility of the prior distribution to have generated the data.
The basic model analyzed in the first part of the paper is, for some aspects, restrictive in that operator K is supposed to be known and equal for all the observations. In Section 5 we consider the more general case in which operator K is unknown and we study how this affect the speed of convergence of the regularized estimated solution. In Section 6 we consider an extension of the basic model for the case in which the transformation of the parameter of interest is different from an observation to another one. Finally, results of simulations of the basic model and of some example are provided in Sections 7.

2 The Model

We consider a simple linear inverse problem. Let $\mathcal{X}, \mathcal{Y}$ be infinite-dimensional Hilbert spaces over $\mathbb{R}$ that are supposed to be Polish with inner product $< \cdot, \cdot >_{\mathcal{X}}$ (resp., $< \cdot, \cdot >_{\mathcal{Y}}$). Let $|| \cdot ||_{\mathcal{X}}$ denote the norm in $\mathcal{X}$ (resp., $|| \cdot ||_{\mathcal{Y}}$). In the following, the notation will be lightened by eliminating the indices in the inner product and in the norm, the sense being clear from the contest. Our purpose is to recover the infinite dimensional parameter x from the functional equation:

$$Y = Kx \tag{2}$$

where $Y \in \mathcal{Y}$, $x \in \mathcal{X}$ and $K : \mathcal{X} \to \mathcal{Y}$ is an injective, Hilbert-Schmidt operator that is supposed to be known. K^* will denote the adjoint of K.

As an example of spaces, we could take $\mathcal{X} = L^2_\pi$ and $\mathcal{Y} = L^2_\rho$, where π and ρ are two measures on $\mathbb{R}$,

$$L^2_\pi = \{x : \mathbb{R} \to \mathbb{R};\ \int x^2(t)\pi(t)dt < \infty\}$$

and

$$L^2_\rho = \{y : \mathbb{R} \to \mathbb{R};\ \int y^2(t)\rho(t)dt < \infty\},$$

endowed with the L^2 norm. Such spaces are used in the examples given below.
Compactness of operator K and the infinite dimension of the range of K (*i.e.* $dim\mathcal{R}(K) = \infty$), make equation (2) ill-posed. To put it better, instability of the solution is due to the spectrum of K, $\sigma(K)$, consisting of a countable set of eigenvalues that accumulate only at 0. To recover x we exploit information contained in the whole curve $\{Y(t)\}$ that represents the observed quantity. What renders instability a relevant problem is the fact that typically $Y(t)$ is measured with error, so that only the approximate equation $Y \approx Kx$ holds and, due to the ill-posed nature of the problem, these small errors may cause errors of arbitrary size in the recovered x. Following tradition in Bayesian literature, we will denote the noisy observation of Y with $\hat{Y}$ and the measurement error with U, $U \in \mathcal{Y}$. Therefore, the corresponding statistical model is

$$\hat{Y} = Kx + U. \tag{3}$$

A particular interpretation of the observation noise is found in Statistical Inverse Theory that substitutes the true Y with some estimation obtained from some sample and hence U is the estimation error.
Quantities $\hat{Y}$, x and U in equation (3) have to be meant as hilbert-random variables, namely as measurable maps from some probability space in an Hilbert space endowed with its Borel σ-field. This is the principal departure from classical approach to solve inverse problems. Such a functional equation is exploited to characterize the conditional distribution $\mathbb{P}(\hat{Y}|x)$ of $\hat{Y}$ given x. 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 this conditional distribution. In order this conditional probability be properly defined as a measure (in the sense that it represents a regular version of the conditional probability), it is assumed that a transition probability exists that associates to each x a probability measure P^x on $(\mathcal{Y}, \mathcal{F})$ such that

$$\mathbb{P}(A|x) = P^x(A)$$

for every $A \in \mathcal{F}$. We call P^x the *sampling measure* and we suppose that it is gaussian whose mean function and covariance operator are determined by (3). Assumption 1 below characterizes in a rigorous way the measure P^x induced by $\hat{Y}$.

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

P^x is gaussian if the probability distribution on the Borel sets of $\mathbb{R}$ induced from P^x by every bounded linear functional on $\mathcal{Y}$ is Gaussian. More clearly, P^x gaussian means that $\forall B \in \mathcal{B}(\mathbb{R})$

$$\mathbb{P}(B) = P^x\{\hat{Y};\ < \hat{Y}, \psi > \in B\}$$

is Gaussian for all $\psi \in \mathcal{Y}$, see Baker (1973) [2]. The mean element Kx in $\mathcal{Y}$ is defined by

$$< Kx, \psi_1 > \int_{\mathcal{Y}} < \hat{Y}, \psi_1 > dP^x(\hat{Y})$$

and the operator Σ by

$$< \Sigma\psi_1, \psi_2 > = \int_{\mathcal{Y}} < \hat{Y} - Kx, \psi_1 >< \hat{Y} - Kx, \psi_2 > dP^x(\hat{Y})$$

for every $\psi_1, \psi_2 \in \mathcal{Y}$. On the basis of this definition, Σ is correctly specified as a covariance operator in the sense that it 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 in the well suited space. Indeed, by Kolmogorov's inequality a realization of a random function $\hat{Y}$ is in $\mathcal{Y}$ if $\mathbb{E}(||\hat{Y}||^2_{\mathcal{Y}}|x)$ is finite[1]. Since $\mathbb{E}(||\hat{Y}||^2|x) = \sum_j \lambda_j + ||Kx||^2$ and $Kx \in \mathcal{Y}$, this is guaranteed if Σ is trace-class, that is if $\sum_j \lambda_j < \infty$, with $\{\lambda_j\}$ the eigenvalues associated to Σ and $\mathbb{E}(\cdot|x)$ the expectation taken with respect to P^x.
Since the eigenvalues of $\Sigma^{\frac{1}{2}}$ are the square roots of the eigenvalues of Σ the fact to be trace-class entails that $\Sigma^{\frac{1}{2}}$ is Hilbert-Schmidt. Hilbert-Schmidt operators are compact and the adjoint is still Hilbert-Schmidt. Compaticity of $\Sigma^{\frac{1}{2}}$ implies compaticity of Σ. Compact operators are particularly attractive since they can be approximated by a sequence of finite dimensional operators and this is really useful when we do not known such an operator and we need to estimate it.

In general, we can link the covariance operator Σ to some parameter n in such a way that this operator goes to 0 with n: $\Sigma_n \to 0$. For instance, this will be the case if Y is a consistent estimation of the transformed signal Kx and consequently U is an estimation error that disappears as the sample size increases. Otherwise, in order to make inference we may want to consider the mean of the first n observations of an *i.i.d.* sample $\hat{Y}_1, \hat{Y}_2, \ldots$, each of them being a realization of a conditional gaussian stochastic process, conditioned to x, with conditional mean Kx and conditional variance Σ. Thus, $\hat{Y} = \frac{1}{n}\sum_{i=1}^n \hat{Y}_i$ and model (3) can be written as

$$\begin{aligned} \hat{Y} &= Kx + \frac{1}{n}\sum_{i=1}^n U_i \qquad U_i \sim i.i.d.\mathcal{GP}(0,\Sigma) \\ &= Kx + U \qquad U \sim \mathcal{GP}(0, \frac{1}{n}\Sigma). \end{aligned} \tag{4}$$

Since the sample mean is a sufficient statistics, this formulation and $Y_i = Kx + U_i$, $i = 1, \ldots, n$ are statistically equivalent. For a proof of sufficiency of the sample mean, see Appendix 8.1. Henceforth, from now on we will denote the covariance operator with Σ_n, where the index n will be meant, in the most intuitive way, the sample size.
It should result clear, after this remark, that we can have observational schemes of different type. In the most immediate situation, we are concerned with a sample of n curves $\hat{Y}_i$, $i = 1, \ldots, n$, namely we observe an object of infinite dimension. This is different with the usual observational scheme in functional analysis literature where it is supposed to observe the curve only at certain points. In our setting it is supposed to really observe the whole curve, *i.e.* at every point, that could appear not very realistic since it seems difficult to observe this kind of object. However, we can easily obtain high-frequency data, like financial data, that can be seen as an approximation to this curve. Otherwise, the observations could be founded to be discrete and the curve $\hat{Y}$ is a function obtained by transforming these data, like for instance a nonparametric estimator. This is the classical situation in Statistical Inverse Problems that we are going to consider: a sample of observations of a discrete object is available and from it we can obtain an estimate of $\hat{Y}$ in the problem, for instance the empirical characteristic function, the empirical cumulative distribution function or again the estimated integrated hazard function, see Examples 2 and 3 below. With this more realistic observational scheme, by starting from discrete observations we get objects of infinite dimension.
The last step in order to well define the Bayesian inverse problem is to define a probability measure μ on the parameter space $\mathcal{X}$, this measure in induced by x and is called *prior probability*. We specify a conjugate prior for x that means we are considering a Gaussian measure on the Hilbert space $\mathcal{X}$ endowed with the σ-field $\mathcal{E}$, and we assume x is independent of U.

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

Definition of the mean function x_0 and covariance operator Ω_0 are specular to above. Then, $\Omega_0\phi := \mathbb{E}(< \phi, (x - x_0) > (x - x_0))$ is a linear, bounded, self-adjoint, positive semi-definite and

[1] Namely, following Kolmogorov's inequality $\mathbb{P}(||\hat{Y}||^2_{\mathcal{Y}} > \epsilon_n) \sim \mathcal{O}_p(1)$ if and only if $\mathbb{E}(||\hat{Y}||^2_{\mathcal{Y}})$ is finite.

trace-class operator. Moreover, $\Omega_0^{\frac{1}{2}}$ is Hilbert-Schmidt and, exploiting same arguments as for Σ_n, Ω_0 is compact.

Assumption 1 and functional form (3) implies that the error term U is a Gaussian process with zero mean and with variance Σ_n. Moreover, we assume that U is independent of x. Error term U is usually interpreted in Statistical Inverse Theory as a measurement error. In most of real situations the hypothesis of normality of an estimation error is justified only asymptotically, hence our model would seem not to rely over a valid foundation. Actually, this is not the case since hypothesis of normality only matters for determining our estimator, but it is not at all used in study of asymptotic properties and our estimator will be consistent even if the normality hypothesis is not verified. In other words, the justification of our estimator is given from posterior consistency the proof of which does not rely on the fact that the sampling and prior measures are assumed to be gaussian.

To have an identified model, we add the assumption of injectivity of the covariance operators:

Assumption 3 *Both Σ_n and Ω_0 are injective operators, i.e. $\mathcal{N}(\Sigma_n) = \{0\}$ and $\mathcal{N}(\Omega_0) = \{0\}$.*

2.1 Construction of the Bayesian Experiment

We construct in this section the Bayesian experiment associated to inverse problem (2) and based on the prior and sampling distributions specified in preceding section. This is accomplished by defining the relevant probability space, that is the probability space associated to (2), with the joint measure determined by recomposing the prior and sampling distributions as joint probability measure. The relevant space we are working in is the real linear product space $\mathcal{X} \times \mathcal{Y}$ that is defined as the set

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

with addition and scalar multiplication defined by $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$ and $h(x_1, y_1) = (hx_1, hy_1)$, $h \in \mathbb{R}$. $\mathcal{X} \times \mathcal{Y}$ is a separable Hilbert space under the norm induced by the scalar product defined as

$$< (x_1, y_1), (x_2, y_2) >_{\mathcal{X}\times\mathcal{Y}} := < x_1, x_2 >_{\mathcal{X}} + < (y_1, y_2) >_{\mathcal{Y}}, \quad \forall (x_i, y_i) \in \mathcal{X} \times \mathcal{Y},\ i = 1, 2.$$

As already stressed, in the following we eliminate the indices in the scalar product and in the norm. We have already introduced $\mathcal{E}$ and $\mathcal{F}$ for the σ-fields of subsets of the parameter space $\mathcal{X}$ and the sample space $\mathcal{Y}$, respectively. We introduce now the product σ-field of $\mathcal{E}$ and $\mathcal{F}$, denoted with $\mathcal{E} \otimes \mathcal{F}$ and defined as the σ-field generated by the algebra of measurable rectangles $R_1 \times R_2$, with $R_1 \in \mathcal{E}$, $R_2 \in \mathcal{F}$.
The probability measure Π on the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{E} \otimes \mathcal{F})$ can be constructed by endowing the parameter space with a probability measure μ and the sampling space with a sampling measure P^x. This has been done in Assumptions 1 and 2 where a Gaussian Measure for both the prior and the sampling distribution has been specified. Then, Π is obtained by recomposing μ and P^x in the following way:

$$\Pi(B \times C) = \int_B P^x(C)\mu(dx), \qquad B \in \mathcal{E},\ C \in \mathcal{F} \tag{5}$$

and by extending to $\mathcal{E} \otimes \mathcal{F}$ this function Π. The following theorem states that this measure is gaussian.

Theorem 1 *Let $\mathcal{X} \times \mathcal{Y}$ be a separable Hilbert space endowed with the σ-field $\mathcal{E} \otimes \mathcal{F}$. 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 such that $\Upsilon(\varphi, \psi) = (\Omega_0\varphi + \Omega_0 K^*\psi, (\Sigma_n + K\Omega_0 K^*)\psi + K\Omega_0\varphi)$, for all (φ, ψ) in $\mathcal{X} \times \mathcal{Y}$.*

The proof of this theorem is given in Appendix 8.2. A similar argument will be used to prove that the marginal distribution P of $\hat{Y}$ on $(\mathcal{Y}, \mathcal{F})$ is gaussian.

Theorem 2 *Let P be a gaussian measure on $(\mathcal{Y}, \mathcal{F})$ with mean function $m_y = Kx_0$ in $\mathcal{Y}$ and covariance operator $\Upsilon_{yy} = \Sigma_n + K\Omega_0 K^*$. Then, P is the projection on the Hilbert measurable space $(\mathcal{Y}, \mathcal{F})$ of the joint gaussian measure Π defined in Theorem 1.*

Distribution P is called *predictive probability*. We have already pointed out that a covariance operator need to be trace-class in order to have an Hilbert space-valued random function. The following Lemma shows that Υ and Υ_{yy} are trace-class.

Theorem 3 *The covariance operators* Υ *and* $(\Sigma_n + K\Omega_0 K^*)$ *are trace class. In particular,* $(\Sigma_n + K\Omega_0 K^*)$ *trace class is a necessary condition for* Υ *being trace class.*

Summarizing, we have started from the inverse problem (2) and we have characterized the Bayesian Inverse Problem associated to it by making two distributional assumptions. In the following of the paper we will denote the *Bayesian Experiment* as:

$$\Xi = (\mathcal{X} \times \mathcal{Y}, \mathcal{E} \otimes \mathcal{F}, \Pi = P^x \otimes \mu). \tag{6}$$

This Bayesian Experiment constitutes the object of our study and it is now evident how the object analyzed in Bayesian approach is changed with respect to that one analyzed in classical approach. The aim will be to determine the inverse decomposition of Π into the marginal P and posterior distribution $\mu^{\mathcal{F}} = \mathbb{P}(x|\hat{Y})$. Determination of this distribution raises some problems, in particular, we have to check for the regularity of the Bayesian Experiment, namely for the existence of a transition characterizing the posterior distribution.

2.2 Examples

We present in this subsection some example of application of our model. These are examples encountered in statistic and econometric field.

Example 1: Density estimation

We consider the classical problem in statistic of density estimation. A review of bayesian nonparametric and semiparametric density estimates is provided by Hjort (1996) [19]. Bayesian density estimation based on mixture of Dirichlet processes has been proposed for instance by Escobar and West (1995) [9] and by Florens, Mouchart and Rolin (1992) [14]. Density estimation using Bernstein polynomials as priors was proposed by Petrone (1999) [27]. Alternatively, Polya tree prior has been used for instance by Ferguson (1974) [11] and Lavine (1992) [23].
We consider a different approach. Let $\mathcal{X} = L^2_\pi(\mathbb{R})$ and $\mathcal{Y} = L^2_\rho(\mathbb{R})$, with π and ρ two measures on $\mathbb{R}$ different than the Lebeasgue measure. We consider a real-valued random variable ξ with a distribution characterized by the *c.d.f.* F, $F(\bar{\xi}) = \mathbb{P}(\xi \leq \bar{\xi})$. If this distribution admits a density, then it exists a nonnegative function $f(\xi)$ known as the probability density function of ξ such that $f(\xi) = \frac{dF}{d\xi}$. Function $f(\xi)$ belongs to $\mathcal{X}$ and it is characterized as the solution of an inverse problem.
To obtain an estimate of the probability density we should solve an integral equation of I kind under the fact that F is unknown and an *i.i.d.* sample $\xi_1, \ldots, \xi_n$ from F is available. By using the random sample we would approximate F for instance with the *empirical distribution function*

$$\hat{F}_n(\bar{\xi}) = \frac{1}{n} \sum_{i=1}^{n} 1\{\xi_i \leq \bar{\xi}\}.$$

The inverse problem we have to solve becomes

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

with $K : L^2_\pi(\mathbb{R}) \to L^2_\rho(\mathbb{R})$ an integral operator with kernel $1\{u \leq \bar{\xi}\}\frac{1}{\pi(u)}$ and U_n the estimation error. The choice of the spaces of reference entails K is a compact operator. Solution of this inverse problem can be computed also in the classical way by applying for instance a Tikhonov regularization scheme. The regularized solution would be $f_{\alpha_n} = (\alpha_n I + K^*K)^{-1}K^*\hat{F}_n$, where α_n denotes a regularization parameter that declines to zero with n and $K^* : L^2_\rho(\mathbb{R}) \to L^2_\pi(\mathbb{R})$ is the

adjoint of K with kernel $\frac{1\{\bar{\xi}\geq u\}}{\pi(u)}$. The solution f_{α_n} is continuous in $\hat{F}_n$ then, by Glivenko-Cantelli theorem, it converges toward the true density function, see Vapnik (1998) Theorem 7.1 [32].

To get a Bayesian estimation of f exploiting our approach, we suppose that $f(\xi)$ has been generated from a gaussian measure μ. The sampling probability P^f is inferred from asymptotic properties of the empirical distribution function. It is a well known result that $\sqrt{n}(\hat{F}_n - F)$ weakly converges toward a zero mean gaussian process G_F with covariance kernel $\mathbb{E}(G_F(t_j), G_F(t_l)) = F(t_j \wedge t_l) - F(t_j)F(t_l)$ and such that $G_F(\pm\infty) = 0$ and G_F is tight. Therefore, P^f is asymptotically a Gaussian measure with mean F and covariance operator $\Sigma_n = \frac{1}{n}\int_{\mathbb{R}} \mathbb{E}(G_F(t_j), G_F(t_l))dt_j$.

Example 2: Regression estimation

Another example of inverse problem is given by the regression function estimation. This problem is of particular interest in statistic, econometrics and machine learning, see Rasmussen and Williams (2006) [28]. Different approaches can be found in Bayesian literature, see for instance Hanson and Johnson (2002) [16] that uses a mixture of Polya tree or Smith and Kohn (1996) [30] that uses spline models.
Let (ξ, w) be a $\mathbb{R}^{1+p}$-valued random vector with distribution characterized by a *cdf* F and $L^2_F(w)$ be the space of square integrable functions of w, we define the regression function of ξ given w as a function $m(w) \in L^2_F(w)$ such that

$$\xi = m(w) + \varepsilon, \qquad \mathbb{E}(\varepsilon|w) = 0 \qquad \mathbb{E}(\varepsilon^2|w) = \sigma^2.$$

In other words, $m(w) = \mathbb{E}(\xi|w)$ and estimate the regression function is equivalent to estimate the conditional density $f(\xi|w)$ as proposed in Vapnik (1998) [32], Section 1.9 and to plug it in the integral defining $m(w)$. We follow here another approach and we characterize $m(w)$ directly as a solution to an inverse problem. Let $g(w,t)$ be a known function on $\mathbb{R}^p \times \mathbb{R}$ in $\mathbb{R}$ defining an Hilbert-Schmidt integral operator with respect to w, then

$$\mathbb{E}(g(w,t)\xi) = \mathbb{E}(g(w,t)m(w)),$$

where the expectation is taken with respect to F. The fact that K is Hilbert-Schmidt ensures that $Km \in L^2_\pi(\mathbb{R})$, with π a measure on $\mathbb{R}$; moreover, if ξ has finite second moment, it also ensures that $\mathbb{E}(g(w,t)\xi) \in L^2_\pi(\mathbb{R})$. We suppose $F(\xi|w)$ is unknown while $F(\cdot, w)$ is known; this implies that the LHS of the functional equation 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 can replace the LHS of the functional equation with 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

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

with the operator $K : L^2_F(w) \to L^2_\pi(\mathbb{R})$ the integral operator with kernel $g(w,t)$ and π a measure on $\mathbb{R}$. 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 $\sigma^2\Lambda = \int_{\mathbb{R}}\int_{\mathbb{R}^p} g(w,t)g(w,s)\sigma^2 f(w)dw\pi(s)ds$. So, the sampling measure P^m is approximately gaussian with mean $\mathbb{E}(g(w,t)\xi)$ and variance $\frac{\sigma^2}{n}\Lambda$.
In most of cases the *cdf* F is completely unknown and also operator K must be estimated to compute the posterior distribution. However, under some regularity assumption this does not affect the speed of convergence of our estimator to the true solution. We refer to section 5 for a more deep analysis of this extension.

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 $\alpha = \frac{F'}{1-F}$ and integrated hazard function $A(t) =$

$\int_0^t \alpha(u)du$. We consider a sequence of survival times $X_{1n}, X_{2n}, \ldots, X_{nn}$. The observational scheme is particular in the sense that 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 α, given an estimate of $A(t)$, by solving the functional equation

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

where $U_n(t)$ is introduced to account for the estimation error. We propose to estimate $A(t)$ with the *Nelson-Aalen estimator* that takes the form

$$\hat{A}_n(t) = \sum_{i:\tilde{X}_{in} \leq t} \frac{D_{in}}{Y_n(\tilde{X}_{in})},$$

with $Y_n(t) = \sum_{i=1}^n 1(\tilde{X}_{in} \geq t)$, see Andersen, Borgan, Gill and Keiding (1993) [1]. An approximate of the sampling distribution can be inferred from the asymptotic properties of the Nelson-Aalen estimator. $\hat{A}_n$ is uniformly consistent on a compact interval and

$$\sqrt{n}(\hat{A}_n - A) \to^d W \quad \text{as } n \to \infty.$$

Here, W is a Gaussian martingale with $W(0) = 0$ and $Cov(W(s_1), W(s_2)) = \sigma^2(s_1 \wedge s_2)$, where $\sigma^2(s) = \int_0^s \{\alpha(u)/y(u)\}du$ with $y(s) = (1 - F(s))(1 - G(s-))$.
The structure of the bayesian inverse problem (3) has been restated, with the operator $K = \int_0^t \cdot\, du$, and as sampling distribution a zero mean gaussian measure with covariance operator $\Sigma_n\psi = \sigma^2 \int (s_1 \wedge s_2)\psi(s_1)\rho(s_1)ds_1$, with $\psi \in \mathcal{Y}$ and ρ a measure on the domain of the functions in $\mathcal{Y}$.
The method that we just proposed to deal with inferential problems in survival analysis with right-censored data is really new with respect to previous bayesian literature. Alternative approaches can be found in Hjort (1990) [18] that proposes to use a beta process as prior for the cumulative hazard process, in Ferguson and Phadia (1979) [12] that use processes neutral to the right as priors for nonparametric estimation of the *cdf* F or in Walker and Muliere (1997) [33] where the *cdf* F is supposed to be drawn from a Beta-Stacy process. Bayesian models for hazard rates has been proposed by Dykstra and Laud (1981) [10] and by Ishwaran and James (2004) [20]. A semiparametric estimation of proportional hazards models, where the parameter of interest is that one involved in the relation between a duration and explanatory variables and the data distribution is treated as a nuisance parameter has been proposed by Ruggiero (1994) [29].

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. Our interest is to recover density $\varphi(x)$. We suppose $g(\cdot)$ is known and x is not observable but a sample of realizations of the random variable Y is available. Density $f(y)$ is estimated nonparametrically, for instance with a kernel smoothing. The corresponding statistical model is

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

where $K = \int g(y - x)dx$ is the known operator of our model in a certain L^2 space 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, characterized by the kernel function $a(y,t)$. 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 y. If we substitute $f(y)$ with the kernel estimator we get the error term V defined as $V = \int a(y,t)\hat{f}(y)dy - AK\varphi = \frac{1}{n}\sum(a(y_i,t) - \mathbb{E}(a(y,t)))$. It weakly converges toward a gaussian process with zero mean and covariance operator with kernel $\mathbb{E}(a(y_i,t) - \mathbb{E}(a(y,t)))(a(y_i,\tau) - \mathbb{E}(a(y,\tau)))$.

Example 5: Instrumental Regression Model

Let (Y,Z,W) be a random vector whose components $Y \in \mathbb{R}$, $Z \in \mathbb{R}^p$ and $W \in \mathbb{R}^q$. Let F denote its cumulative distribution function and L^2_F be the Hilbert space of square integrable functions of (Y,Z,W). We consider the econometric model

$$Y = \varphi(Z) + \varepsilon, \qquad \mathbb{E}(U|W) = 0 \tag{7}$$

that defines the instrumental regression function $\varphi(Z) \in L^2_F(Z)$, where $L^2_F(Z) \subseteq L^2_F$ is the space of square integrable functions depending on Z. ε is an homoskedastic error term with variance σ^2. $\varphi(Z)$ is the parameter of interest and, by exploiting condition in (7) and under some regularity assumption, can be seen as the solution of an integral equation of I 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 $L^2_F(Z)$. The instrumental regression is now characterized as the solution of

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

with $K = \mathbb{E}(\mathbb{E}(\cdot|W)|Z)$. By substituting the LHS with a nonparametric estimator, we get a model like (3)

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

We can make a different assumption concerning the degree of knowledge of F. If F is partially known, namely $F(Z,W)$ is known, then only the conditional expectation $\mathbb{E}(Y|W)$ is to be estimated. In the alternative case with F completely unknown, both the whole LHS and the operator need to be estimated. The extension of the basic model to the case with unknown operator is treated in Section 5 of this paper.
In both the cases, the (approximated) distribution of the estimation error will be a centered gaussian with covariance operator $\frac{1}{N}\sigma^2 K^*K$, where K^* denotes the adjoint of K. We refer you to Florens and Simoni (2007) [15] for a complete treatment of this model.

3 Solution of the Ill-Posed Inverse Problem

In this section we compute the solution of Statistical Inverse Problem (3). We reminde that the solution to inverse problem (2) restated in a larger space of probability measures is the posterior distribution of the quantity of interest $x(t)$. We will use $\mu^{\mathcal{F}}$ to denote this posterior distribution. Due to infinite dimension of the Bayesian experiment, application of Bayes theorem is not evident and we have to be careful in defining and computing the posterior distribution. Three points require to be discussed: the existence of a regular version of the conditional probability on $\mathcal{E}$ given $\mathcal{F}$, the fact that it is a gaussian measure and its computability.

3.1 Regularity of Bayesian Experiment: existence of a regular version of the posterior probability

In constructing the Bayesian experiment we have defined Π as the recomposition of μ and P^x. To make this experiment operational, it is compulsory the existence of an inverse decomposition of Π: $\Pi = P \otimes \mu^{\mathcal{F}}$, that is ensured if a regular version of the posterior probability exists.
Consider the two probability spaces $(\mathcal{X}, \mathcal{E}, \mu)$ and $(\mathcal{Y}, \mathcal{F}, P)$ defining the Bayesian experiment. First of all, we prove that the conditional probability on $\mathcal{E}$ given $\mathcal{F}$ exists. Then, we find that it can be characterized by a transition so that $\mathbb{P}(x|\hat{Y})$ is well defined and, if recomposed with P, gives the joint measure Π. Let $L^2(\mathcal{X} \times \mathcal{Y})$ be the Hilbert space of square integrable random variables defined on $\mathcal{X} \times \mathcal{Y}$ with the inner product

$$< \varphi(x, \hat{Y}), \psi(x, \hat{Y}) >_{L^2} = \mathbb{E}(\varphi(x, \hat{Y})\psi(x, \hat{Y})), \qquad \forall \varphi, \psi \in L^2(\mathcal{X} \times \mathcal{Y}),$$

where the expectation is taken with respect to Π. Consider the subset $L^2(\mathcal{Y}) \subset L^2(\mathcal{X} \times \mathcal{Y})$ of square integrable functions defined on $\mathcal{Y}$. The conditional probability of a measurable set $A \in \mathcal{E}$ given $\mathcal{F}$, $\mu^{\mathcal{F}}(A)$ is the projection of $1(x \in A)$ on $L^2(\mathcal{Y})$. $L^2(\mathcal{Y})$ is a closed convex subset of $L^2(\mathcal{X} \times \mathcal{Y})$ (closed with respect to the scalar product on $L^2(\mathcal{X} \times \mathcal{Y})$), that implies that $\forall \varphi \in L^2(\mathcal{X} \times \mathcal{Y})$ the conditional expectation $\mathbb{E}(\varphi|\mathcal{F})$ exists and is unique and so the conditional probability exists.
A conditional probability is called *regular* if a transition probability characterizing it exists. Anyway, conditional probability does not need to be a transition since relations characterizing a probability hold only outside of a negligible set. In particular, countable additivity is difficult to satisfy. Nevertheless, the existence of such a transition inducing conditional probability is guaranteed by an important theorem, known as *Jirina Theorem* (see Neveu (1965) [25]):

Proposition 1 *Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{E} \otimes \mathcal{F}, \Pi)$ be a probability space that is Polish. Then for every σ-subalgebras $\mathcal{G}_1 \subset \mathcal{E} \otimes \mathcal{F}$ and $\mathcal{G}_2 \subset \mathcal{E} \otimes \mathcal{F}$, there exists at least one "regular conditional probability" on $\mathcal{G}_2$ given $\mathcal{G}_1$.*

This result implies that if we pick as σ-subalgebras $\mathcal{E}$ and $\mathcal{F}$, $\mathbb{P}(A|\mathcal{F}) \equiv \mu^{\mathcal{F}}(A)$, $\forall A \in \mathcal{E}$ is a regular conditional probability which entails the existence of the posterior distribution and of the inverse decomposition $\Pi = P \otimes \mu^{\mathcal{F}}$. The crucial hypothesis to having a regular Bayesian experiment is that the space we are working in is Polish (complete separable metric space). This is not an unrealistic assumption since a lot of spaces satisfy this property. For instance, the frequently used L^2 space, with a Gaussian measure defined on it, is a Polish space, see Hiroshi and Yoshiaki (1975) [17].
A last remark need to be stressed. Given the existence of the solution of Bayesian Statistical Inverse Problem (3), we are interested in checking whether it is continuous in the sample $\hat{Y}$. Continuity is crucial for asymptotic properties of the estimator. In particular we are interested in the posterior consistency, whose definition will be given in Section 4. Problems of inconsistency are frequent in nonparametric Bayesian experiments, see Diaconis and Freedman (1986) [7]. For circumventing that, we have to guarantee that the posterior mean, characterizing the measure $\mu^{\mathcal{F}}$, be continuous in $\hat{Y}$.

3.2 Gaussian Posterior Probability

In this section, we briefly recover results that are well known in literature, see for instance Mandelbaum (1984) [24], but that are very important for our study. Given two jointly distributed stochastic processes x and $\hat{Y}$, the conditional expectation $\mathbb{E}(x|\hat{Y})$, determined by the relation

$$\mathbb{E}(< x, h > |\hat{Y}) = < h, \mathbb{E}(x|\hat{Y}) > \quad \text{a.s.}, \qquad \forall h \in \mathcal{X}, \tag{8}$$

exists and it is an affine transformation of $\hat{Y}$. Existence is shown in Neveu (1975) [26] together with the convergence, as $k \to \infty$,

$$\mathbb{E}(x| < \hat{Y}, \psi_1 >, \ldots, < \hat{Y}, \psi_k >) \to \mathbb{E}(x|\hat{Y}) \quad \text{a.s.}, \tag{9}$$

true for any orthonormal basis $\{\psi_j\}$. Let $\{\lambda_j, \psi_j\}_j$ be the eigensystem of Υ_{yy}, it follows that $< \hat{Y}, \psi_j >$, $j = 1, 2, \ldots$ are *i.i.d.* $\mathcal{N}(< Kx_0, \psi_j >, \lambda_j)$ and $\forall h \in \mathcal{X}$

$$\begin{aligned}\mathbb{E}(< x,h > | < \hat{Y},\psi_1 >,\ldots,< \hat{Y},\psi_k >) &= < x_0,h > + \sum_{j=1}^{k} \frac{< \Omega_0 K^*\psi_j, h >}{\lambda_j}(< \hat{Y},\psi_j > - < Kx_0,\psi_j >) \\ &= < h, A_k\hat{Y} + b_k > .\end{aligned}$$

with $A_k\hat{Y} = \sum_{j=1}^{k} \Omega_0 K^* \Upsilon_{yy}\psi_j < \hat{Y},\psi_j >$ and $b_k = x_0 - \sum_{j=1}^{k} \Omega_0 K^*\Upsilon_{yy}\psi_j < Kx_0,\psi_j >$. Then, by (8), $\mathbb{E}(x| < \hat{Y},\psi_1 >,\ldots,< \hat{Y},\psi_k >) = A_k\hat{Y}+b_k$ and by (9), $\mathbb{E}(x|\hat{Y}) = A\hat{Y}+b$, where $A = \lim A_k$ and $b = \lim b_k$.
Furthermore, it is trivial to show that the conditional distribution of x given $\hat{Y}$, $\mu^{\mathcal{F}}$, is gaussian. Let $\varepsilon = x - \mathbb{E}(x|\hat{Y})$, then, by definition of conditional expectation, ε is independent of $\hat{Y}$ and $< \varepsilon, h > \sim \mathcal{N}(0, < (\Omega_0 - AK\Omega_0)h, h >)$, for each $h \in \mathcal{X}$. Consider the characteristic function of $< x, h >$:

$$\begin{aligned}\mathbb{E}(e^{it<x,h>}|\hat{Y}) &= \mathbb{E}(e^{it<\varepsilon,h>}e^{it<A\hat{Y}+b,h>}|\hat{Y}) \\ &= \mathbb{E}(e^{it<\varepsilon,h>})e^{it<A\hat{Y}+b,h>} \\ &= e^{it<A\hat{Y}+b,h> - \frac{1}{2}t^2<(\Omega_0 - AK\Omega_0)h,h>}\end{aligned}$$

that it is the characteristic function of a gaussian random variable. Moreover, we will show in the next section that $(\Omega_0 - AK\Omega_0) = Var(x|\hat{Y})$. Given the one-to-one correspondence between distribution and characteristic function, we can say that $< x,h > |\hat{Y}$ is gaussian for every $h \in \mathcal{X}$ and, from definition of Gaussian process, we conclude that $\mu^{\mathcal{F}}$ is gaussian with mean $A\hat{Y} + b$.

3.3 Computation of the Posterior Probability

We have proved the posterior distribution of the parameter x exists, is a transition probability and is a Gaussian measure: $x|\hat{Y} \sim \mathcal{GP}(A\hat{Y} + b, V)$, with $A : \mathcal{Y} \to \mathcal{X}$, $V : \mathcal{X} \to \mathcal{X}$ and $b \in \mathcal{E}$ two operators and a measurable function, respectively, to be determined. Computation of these quantities is trivial in finite dimensional space, but it may rise problems of continuity when spaces are infinite dimensional causing operator A to not be well-defined. Consider the covariance operator of the stochastic process $(x, \hat{Y}) \in \mathcal{X} \times \mathcal{Y}$ and in particular the covariance between the two components of this process:

$$\begin{aligned}Cov(< x,\varphi >, < \hat{Y},\psi >) &= Cov(\mathbb{E}(< x,\varphi > |\hat{Y}), < \hat{Y},\psi >) \\ &= Cov(< \mathbb{E}(x|\hat{Y}),\varphi >, < \hat{Y},\psi >) \\ &= Cov(< A\hat{Y} + b,\varphi >, < \hat{Y},\psi >) \\ &= Cov(< A\hat{Y},\varphi >, < \hat{Y},\psi >) \\ &= Cov(< \hat{Y}, A^*\varphi >, < \hat{Y},\psi >) \\ &= < (\Sigma_n + K\Omega_0 K^*)A^*\varphi,\psi > \qquad (10)\end{aligned}$$

for any $\varphi \in \mathcal{X}$ and $\psi \in \mathcal{Y}$. By definition of covariance operator, $Cov(< x,\varphi >, < \hat{Y},\psi >) =< \Upsilon_{12}\varphi,\psi >$, where Υ_{12} is a component of operator Υ determined in Theorem 1. Exploitation of this equality allows to obtain a functional equation defining operator A^*

$$(\Sigma_n + K\Omega_0 K^*)A^*\varphi = K\Omega_0\varphi, \qquad (11)$$

for any $\varphi \in \mathcal{X}$, or equivalently,

$$A(\Sigma_n + K\Omega_0 K^*)\psi = \Omega_0 K^*\psi$$

where $\psi \in \mathcal{Y}$.
Without paying too much attention to equation (11), we could define A, in an erroneous way, as $A = \Omega_0 K^*(\Sigma_n + K\Omega_0 K^*)^{-1}$. This solution is clearly not well-defined since $(\Sigma_n + K\Omega_0 K^*)$

is compact[2] and its inverse is not continuous. Moreover, $dim\mathcal{R}(\Sigma_n + K\Omega_0 K^*) = \infty$, then the eigensystem associated to operator $(\Sigma_n + K\Omega_0 K^*)$ is such that there is a countable set of eigenvalues that has (only) zero as accumulation point.
Therefore, we are dealing with an ill-posed inverse problem. In other words, restating the inverse problem in a larger space of probability distributions does not remove the ill-posedness since it is not possible to compute the posterior distribution of x and even if it was possible, the posterior mean would not be continuous in $\hat{Y}$ causing problems of posterior inconsistency.

3.4 Regularized Posterior distribution of Gaussian Processes

The unboundedness of A and consequently of the posterior mean, for an infinite dimension model, has been circumvented in different ways in inverse problem literature, for instance Mandelbaum (1984) [24] restricts the analysis to a class of measurable transformations A which are linear on a subspace of measure one. In a different way, we solve this problem by applying Tikhonov regularization scheme to equation (11) and we define the regularized operator A_α as:

$$A_\alpha = \Omega_0 K^* (\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1} \tag{12}$$

where $\alpha_n > 0$ is a regularization parameter appropriately chosen such that $\alpha_n \to 0$ with n. In the following, we recover expressions for b and V; since they are dependent on A we can compute them only if we replace A with its regularized version A_α. To identify function b of the posterior mean we use an iterated expectation law argument:

$$\begin{aligned} \mathbb{E}(x) &= \mathbb{E}(\mathbb{E}(x|\hat{Y})) \\ x_0 &= A\mathbb{E}(\hat{Y}) + b \\ &= AKx_0 + b. \end{aligned}$$

Then, $b = (I - AK)x_0$ and the corresponding regularized version of b is obtained by substituting A with A_α:

$$b_\alpha = (I - A_\alpha K)x_0. \tag{13}$$

To identify operator $V = Var(x|\hat{Y})$ we make the assumption that V is homoskedastic and we use the relation for the unconditional variance:

$$\begin{aligned} Var(x) &= \mathbb{E}(Var(x|\hat{Y})) + Var(\mathbb{E}(x|\hat{Y})) \\ \Omega_0 &= V + A(\Sigma_n + K\Omega_0 K^*)A^* \\ &= V + AK\Omega_0, \end{aligned}$$

where the last equality has been obtained by using (11) and gives the same expression for V recovered in 3.2. By using regularized version of A we get regularized V:

$$V_\alpha = \Omega_0 - \Omega_0 K^* (\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1} K\Omega_0 \tag{14}$$

These regularized objects characterize a new distribution that is normal with mean $(A_\alpha \hat{Y} + b_\alpha)$ and covariance operator V_α. This distribution is called *regularized posterior distribution* and is denoted with $\mu_\alpha^{\mathcal{F}}$. Note that $\mu_\alpha^{\mathcal{F}}$ differs from the posterior distribution $\mu^{\mathcal{F}}$. Actually, it is a new object that we define to be the solution of the signal-noise problem and that converges toward the true posterior distribution. Moreover, we keep as punctual estimator of x the regularized posterior mean

$$\begin{aligned} \mathbb{E}_\alpha(x|\hat{Y}) = \; & \Omega_0 K^* (\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1} \hat{Y} \\ & + (I - \Omega_0 K^* (\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1} K) x_0. \end{aligned} \tag{15}$$

[2] As we have shown in Section 2, operators Σ_n and Ω_0 are compact. Operator K is compact by assumption. It follows that $K\Omega_0 K^*$ is compact and $(\Sigma_n + K\Omega_0 K^*)$ is compact since it is a linear combination of compact operators (Kress, 1999, [22] Theorems 2.15, 2.16).

From this expression we clearly see that, without a regularization technique, the posterior mean would not be continuous in $\hat{Y}$. Therefore, when the measurement error goes to zero, $\hat{Y}$ would go to the exact transformation Y but the solution of the inverse problem (*i.e.* the posterior mean) would not converge toward the true solution.

3.4.1 Alternative regularized solutions

Definition of operator A has called to solve an ill-posed inverse problem: $(\Sigma_n + K\Omega_0 K^*)A^*\varphi = K\Omega_0\varphi$. Actually, there exists two different way to obtain a regularized solution of this functional equation and consequently to define A. The first option concerns regularization of the inverse of operator $(\Sigma_n + K\Omega_0 K^*)$. This is what we have done in the previous section.
Otherwise, we could regularize the Moore-Penrose generalized inverse of $(\Sigma_n + K\Omega_0 K^*)$ that is defined as the unique solution of minimal norm of the normal equation $(\Sigma_n + K\Omega_0 K^*)^2 A^*\varphi = (\Sigma_n + K\Omega_0 K^*)K\Omega_0\varphi$. In this case the regularized operator A^* is

$$A_\alpha^* = (\alpha_n I + (\Sigma_n + K\Omega_0 K^*)^2)^{-1}(\Sigma_n + K\Omega_0 K^*)K\Omega_0. \tag{16}$$

Consideration of asymptotic properties leads us to prefer the first way of regularization. In fact, to have the same speed of convergence of the regularized posterior mean, that we will determine in the next section, regularization of the Moore-Penrose inverse requires the stronger condition $\Omega_0^{-\frac{1}{2}}(x - x_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}} K^* K \Omega_0^{\frac{1}{2}})^\beta$ be satisfied.

4 Asymptotic Analysis

One of the most interesting topic in statistics is the consistency of the estimator. Once posterior distribution has been obtained, it is very important to know whether it becomes more and more accurate and precise as the number of observed data increases indefinitely. This fact is known as the consistency of the posterior distribution and it is from the point of view of a Bayesian who believes in a certain prior distribution.
A very important result, due to Doob (1949), 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. Actually, no one can be so certain about the prior 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 which consists 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 neighborhoods of this true value.

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 and Freedman (1986) 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, for a quadratic loss function (*i.e.* the posterior mean), 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.
Under suitable assumptions on the true value of the parameter it is not too difficult to have consistency of the posterior distribution. On the contrary, it is more difficult obtaining consistency of the prior distribution. We say that (x, μ) is consistent if and only if the prior μ assigns a positive

probability to every open interval around the true value x. A necessary condition to guarantee this kind of consistency is the finite dimensionality of the parameter space. Diaconis and Freedman (1986) have shown that if the underlying probability mechanism has only a finite number of possible outcomes and the support of the prior probability contains the true parameter value, Bayes estimates are consistent in the classic sense. On the contrary, if the underlying mechanism allows an infinite number of possible outcomes, Bayes estimates can be inconsistent.
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 make sense to speak about *regularized posterior consistency*. The principal result that we find is the consistency of the posterior distribution, under the assumption that the true parameter satisfies some regularity conditions, but the true value of the parameter does not belong to the support of the prior $\mu-$*almost surely*. In other words, the prior is not able to generate a trajectory of x that satisfies the necessary condition for consistency. This finding is in line with previous literature about Bayesian nonparametrics.
In this section we will enounce conditions under which regularized posterior consistency is verified; in the next sub-section we will show that these conditions are not satisfied μ-*a.s.* from a parameter generated by the prior, namely prior consistency is not verified.

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 $M_n \to \infty$

$$\begin{aligned}
\mu_\alpha^{\mathcal{F}}\{x : ||x - x_*||_{\mathcal{X}} \geq M_n\varepsilon_n\} &\leq \frac{\mathbb{E}_\alpha(||x - x_*||_{\mathcal{X}}^2|\hat{Y})}{(M_n\varepsilon_n)^2} \\
&= \frac{1}{(M_n\varepsilon_n)^2}\int V_\alpha(x(t)|\hat{Y}) + (\mathbb{E}_\alpha(x(t) - x_*(t)|\hat{Y}))^2\pi(t)dt \\
&\leq \frac{||V_\alpha(x(t)|\hat{Y})||_{\mathcal{X}}^2 + ||\mathbb{E}_\alpha(x(t)|\hat{Y}) - x_*(t)||_{\mathcal{X}}^2}{(M_n\varepsilon_n)^2} \qquad (17)
\end{aligned}$$

with π a measure on $\mathbb{R}$. The RHS of (17) goes to 0 if both the terms in the numerator converge to zero. Firstly, we consider consistency of the regularized posterior mean: $||\mathbb{E}_\alpha(x|\hat{Y}) - x_*||_{\mathcal{X}} \to 0$ P^{x_*}-*a.s.* when $n \to \infty$. For any true value $x_* \in \mathcal{X}$, the Bayes estimation error

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

can be decomposed in the following way:

$$\begin{aligned}
\mathbb{E}_\alpha(x|\hat{Y}) - x_* = {} & -\overbrace{[I - \Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K](x_* - x_0)}^{I} \\
& + \underbrace{[\Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}K - \Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K](x_* - x_0)}_{II} \\
& + \underbrace{\Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}U}_{III}. \qquad (18)
\end{aligned}$$

We start by analyzing the II and III terms. We use the following majorization of the II term:

$$\begin{aligned}
||II||^2 &= ||\Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}(-\Sigma_n)(\alpha_n I + K\Omega_0 K^*)^{-1}K(x_* - x_0)||^2 \\
&\leq ||\Omega_0 K^*||^2||(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}||^2||\Sigma_n||^2||(\alpha_n I + K\Omega_0 K^*)^{-1}K(x_* - x_0)||^2
\end{aligned}$$

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:

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

that is well-defined if $||\Omega_0^{-\frac{1}{2}}(x_* - x_0)||$ is bounded, *i.e.* if we assume that $(x_* - x_0)$ belongs to the *Reproducing Kernel Hilbert Space* associated to the covariance operator Ω_0 ($\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.(\Omega_0)$ or $\mathcal{H}(\Omega_0)$ throughout the paper). Let $\{\lambda_j^{\Omega_0}, \varphi_j^{\Omega_0}\}_j$ be the eigensystem of the compact self-adjoint operator Ω_0. We define the $\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.(\Omega_0)$ embedded in $\mathcal{X}$ as:

$$\mathcal{H}(\Omega_0) = \{\varphi : \varphi \in \mathcal{X} \quad \text{and} \quad \sum_{j=1}^{\infty} \frac{| < \varphi, \varphi_j^{\Omega_0} > |^2}{\lambda_j^{\Omega_0}} < \infty\}. \tag{19}$$

Therefore, since

$$\begin{aligned} ||\Omega_0^{-\frac{1}{2}}(x_* - x_0)||^2 &= \sum_{j=1}^{\infty} | < \Omega_0^{-\frac{1}{2}}(x_* - x_0), \varphi_j^{\Omega_0} > |^2 \\ &= \sum_{j=1}^{\infty} | < (x_* - x_0), \Omega_0^{-\frac{1}{2}} \varphi_j^{\Omega_0} > |^2 \\ &= \sum_{j=1}^{\infty} \frac{| < (x_* - x_0), \varphi_j^{\Omega_0} > |^2}{\lambda_j^{\Omega_0}} \\ &= ||x_* - x_0||^2_{\mathcal{H}(\Omega_0)}, \end{aligned} \tag{20}$$

we get that $||\Omega_0^{-\frac{1}{2}}(x_* - x_0)||^2 < \infty$ if and only if $(x_* - x_0) \in \mathcal{H}(\Omega_0)$.
As a consequence the last norm of term II is an $\mathcal{O}_p(\frac{1}{\alpha_n})$.
Now, we consider term III and we write it in the following equivalent way:

$$\begin{aligned} III &= \underbrace{\Omega_0 K^*[(\alpha_n I + \Sigma_N + K\Omega_0 K^*)^{-1} - (\alpha_n I + K\Omega_0 K^*)^{-1}]U}_{A} + \\ &\quad \underbrace{\Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1} U}_{B}. \end{aligned}$$

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 $||B||^2 \sim \mathcal{O}_p(\frac{1}{\alpha_n} tr\Sigma_n)$. Kolmogorov Theorem entails that if $\mathbb{E}||U||^2 < \infty$ then $||U||^2$ is bounded in probability; moreover, $\mathbb{E}||U||^2 = tr\Sigma_n$.
We summarize our results of convergence in $\mathcal{X}$ in the following lemma:

Lemma 1 *(i)* $||II||^2 = \mathcal{O}_p(\frac{1}{\alpha_n^2}||\Sigma_n||^2 \frac{1}{\alpha_n})$;

(ii) $||III||^2 = \mathcal{O}_p(\frac{1}{\alpha_n^3}||\Sigma_n||^2 tr\Sigma_n + \frac{1}{\alpha_n} tr\Sigma_n)$.

It should be noted that, if we assume that $tr\Sigma_n \sim \mathcal{O}_p(\frac{1}{n})$ and $||\Sigma_n|| \sim \mathcal{O}_p(\frac{1}{n})$, in order terms II and III converge to zero, conditions $\alpha_n \to 0$ and $\alpha^{\frac{3}{2}} n \to \infty$ should be satisfied by the regularization parameter. 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 Vapnik (1998)[32]). Therefore, we require weaker conditions to get optimal speed of convergence.

Now, let us consider the first term of decomposition (18). Since Ω_0 is positive definite and self-adjoint, it can be rewritten as $\Omega_0 = \Omega_0^{\frac{1}{2}} \Omega_0^{\frac{1}{2}}$ and term I becomes:

$$\begin{aligned} I &= \Omega_0^{\frac{1}{2}}[\Omega_0^{-\frac{1}{2}} - \Omega_0^{\frac{1}{2}} K^*(\alpha_n I + K\Omega_0 K^*)^{-1} K\Omega_0^{\frac{1}{2}} \Omega_0^{-\frac{1}{2}}](x_* - x_0) \\ &= \Omega_0^{\frac{1}{2}}[I - \Omega_0^{\frac{1}{2}} K^*(\alpha_n I + K\Omega_0 K^*)^{-1} K\Omega_0^{\frac{1}{2}}]\Omega_0^{-\frac{1}{2}}(x_* - x_0). \end{aligned} \tag{21}$$

It should be noted that we can always write $I = \Omega_0^{\frac{1}{2}}\Omega_0^{-\frac{1}{2}}$ and that it is well defined if it is not applied to noisy data. On the other hand, when we consider the norm of term I the inverse operator $\Omega_0^{-\frac{1}{2}}$ can pose some problem due to its unboundedness and thus we must assume some regularity condition. More specifically,

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

The last norm on the right hand side is bounded if $(x_* - x_0) \in \mathcal{H}(\Omega_0)$ where $\mathcal{H}(\Omega_0)$ denotes the *Reproducing Kernel Hilbert Space* associated to the covariance operator Ω_0. In order to see that the second norm in the right hand side is bounded, we have just to note that the operator $(I - \Omega_0^{\frac{1}{2}}K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0^{\frac{1}{2}})$ has the same eigenvalues as

$$[I - (\alpha_n I + \Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{-1}\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}}]. \tag{22}$$

Expression (22) appears as the regularization bias associated to the regularized solution of the ill-posed inverse problem $K\Omega_0^{\frac{1}{2}}\lambda = r$ computed using Tikhonov regularization scheme. Therefore, it converges to zero when the regularization parameter α_n goes to zero. Moreover, if the true solution λ lies in the β-regularity space Φ_β of the operator $K\Omega_0^{\frac{1}{2}}$, *i.e.* $\lambda \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, the squared regularization bias is at most of order α_n^β. With reference to expression (21) $\lambda = \Omega_0^{-\frac{1}{2}}(x_* - x_0)$. We admit without proof the following lemma. Sketch of the proof can be found in Carrasco, Florens and Renault (2005) and in Kress (1999).

Lemma 2 *If* $(x_* - x_0) \in \mathcal{H}(\Omega_0)$ *and if* $\Omega_0^{-\frac{1}{2}}(x_* - x_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ *then*

$$||I||^2 = \mathcal{O}_p(\alpha_n^\beta).$$

The larger β is, the smoother the function $\lambda \in \Phi_\beta$ will be and faster the regularization bias will converge to zero. However, for Tikhonov regularization scheme, β cannot be grater than 2. The following theorem summarizes the results of convergence that we have obtained.

Theorem 4 *(i) If* $(x_* - x_0) \in \mathcal{R.K.H.S.}(\Omega_0)$ *and if* $\Omega_0^{-\frac{1}{2}}(x_* - x_0) \in \Phi_\beta$ *the bias is of order*

$$||\mathbb{E}_\alpha(x|\hat{Y}) - x_*||^2 = \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^3}||\Sigma_n||^2 + \frac{1}{\alpha_n}tr\Sigma_n).$$

(ii) If $\alpha_n \to 0$, $\frac{1}{\alpha_n}tr\Sigma_n \to 0$ *and* $\frac{1}{\alpha_n^3}||\Sigma_n||^2 \to 0$, *then* $\mathbb{E}(x|\hat{Y}) \to^{\mathcal{P}^{x_*}} 0$ *in* $\mathcal{X}$ *norm.*

If $tr\Sigma_n$ is of the same order as $||\Sigma_n||$ the fastest global rate of convergence is obtained when $\alpha_n^\beta = \frac{1}{\alpha_n^3}||\Sigma_n||^2$, that is, when the regularization parameter is proportional to

$$\alpha_n \propto ||\Sigma_n||^{\frac{2}{\beta+3}}.$$

The speed of convergence of the regularized posterior mean is equal to $||\Sigma_n||^{\frac{2\beta}{\beta+3}}$. Assuming the trace and the norm of the covariance operator be of the same order is not really stringent. For instance, we can assume, as above, they are both of order $\frac{1}{n}$ and this is true in almost all real examples.

Let us proceed now to the study of the regularized posterior variance. We want to check that $||V_\alpha\varphi|| \to 0$ for all $\varphi \in \mathcal{X}$. By recalling expression (14), we can rewrite the regularized posterior variance as

$$\begin{aligned} V_\alpha &= \overbrace{\Omega_0 - \Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0}^{IV} + \\ &\quad \underbrace{\Omega_0 K^*(\alpha_n I + K\Omega_0 K^*)^{-1}K\Omega_0 - \Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}K\Omega_0}_{V}. \end{aligned} \tag{23}$$

Since Ω_0 is a positive definite self-adjoint operator, it can be decomposed as $\Omega_0 = \Omega_0^{\frac{1}{2}}\Omega_0^{\frac{1}{2}}$ and the operator IV in (23) applied to an element φ in $\mathcal{X}$ can be rewritten as

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

Here, we do not have to put any condition on φ for this operator being bounded. Following the same reasoning done for term I in (18), we conclude that, if $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ then $||IV\varphi||^2 = \mathcal{O}_p(\alpha_n^\beta)$. Operator V in (23) applied to $\varphi \in \mathcal{X}$ is equivalently rewritten as

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

and its squared norm is bounded and of order of $\frac{1}{\alpha_n^3}||\Sigma_n||^2$:

$$||V||^2 = \mathcal{O}_p(\frac{1}{\alpha_n^2}||\Sigma_n||^2\frac{1}{\alpha_n}).$$

We can conclude with the following theorem that states the convergence to zero of the posterior regularized variance:

Theorem 5 *(i) if* $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ *then*

$$||V_\alpha(x|\hat{Y})\varphi||^2 = \mathcal{O}_p(\alpha_n^\beta + \frac{1}{\alpha_n^3}||\Sigma_n||^2).$$

(ii) Moreover, if $\frac{1}{\alpha_n^3}||\Sigma_n||^2 \to 0$ *and* $\alpha_n \to 0$ *then* $V_\alpha(x|\hat{Y})\varphi \to^{\mathcal{P}^{x_*}} 0$ *in* $\mathcal{X}$ *norm.*

Finally, from inequality (17) it follows that when the regularized posterior variance converges to zero in P^{x_*}-probability and the posterior mean is consistent for the true value x_*, the regularized posterior distribution of x degenerates to the Dirac measure in x_*. Thus, if $(x_* - x_0) \in \mathcal{H}(\Omega_0)$, the regularized posterior probability of the complement of any neighborhood of the true parameter x_*, $\mu_\alpha^{\mathcal{F}}\{x : ||x - x_*||_{L^2_\pi} \geq M_n\varepsilon_n\}$, goes to zero and, if $tr\Sigma_n \sim \mathcal{O}_p(||\Sigma_n||)$, it is of order $\alpha_n^\beta + \frac{1}{\alpha_n}tr\Sigma_n + \frac{1}{\alpha_n^3}||\Sigma_n||^2$.

4.1 A result of prior inconsistency

We have expressed our prior beliefs as a Gaussian measure on the parameter space. Now, it is well-known that the support of a centered Gaussian process, taking its values in an L^p space, is equal to the closure in L^p of the *Reproducing Kernel Hilbert Space* associated with the covariance operator of this process. Then, we might think that $(x - x_0)$ can be seen as a distribution over $\mathcal{H}(\Omega_0)$, but this is not correct. In fact, for any version of the process $(x - x_0)$ generated by the prior distribution μ we have $(x - x_0) \in \overline{\mathcal{H}(\Omega_0)}$, but with $\mu-$ probability 1 its trajectories are not in $\mathcal{H}(\Omega_0)$. In the previous section we have provided an important theorem that states posterior consistency if certain hypothesis hold. Now, we are ready to show that the prior distribution is not able to generate a trajectory x that satisfies the necessary regularity condition, *i.e.* $(x - x_0) \in \mathcal{H}(\Omega_0)$ is never verified with $\mu-$probability 1. The following Lemma and its proof state this concept in a more detailed way.

Lemma 3 *Let* $(x - x_0)$ *a zero-mean Gaussian process with covariance operator* Ω_0*. Let* $\mathcal{H}(\Omega_0)$ *be the* $\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.$ *associated to* Ω_0*, i.e.*

$$\mathcal{H}(\Omega_0) = \{x : x \in L^2_\pi \quad and \quad ||x||^2_{\Omega_0} = \sum_{j=1}^{\infty}\frac{|< x, \varphi_j^{\Omega_0} >|^2}{\lambda_j^{\Omega_0}} < \infty\},$$

where $(\lambda_j^{\Omega_0}, \varphi_j^{\Omega_0})$ *is the spectral decomposition of* Ω_0 *and* $||\cdot||_{\Omega_0}$ *denotes the* $\mathcal{H}(\Omega_0)$*-norm. Then,* $(x - x_0) \notin \mathcal{H}(\Omega_0)$ μ*-a.s.*

Proof: Let $(\lambda_j^{\Omega_0}, \varphi_j^{\Omega_0})$ be the eigensystem of Ω_0. We represent the zero-mean Gaussian process $(x - x_0)$ with Karhunen-Loeve expansion:

$$(x - x_0) = \sum_{j=1}^{\infty} < x - x_0, \varphi_j^{\Omega_0} > \varphi_j^{\Omega_0}.$$

$(x - x_0) \in \mathcal{H}(\Omega_0)$ if $||x - x_0||^2_{\Omega_0} = \sum_{j=1}^{\infty} \frac{|<x-x_0, \varphi_j^{\Omega_0}>|^2}{\lambda_j^{\Omega_0}} < \infty$.

$$\begin{aligned} \mathbb{E}||x - x_0||^2_{\Omega_0} &= \sum_{j=1}^{\infty} \frac{1}{\lambda_j^{\Omega 0}} \mathbb{E}(< x - x_0, \varphi_j^{\Omega_0} >)^2 \\ &= \sum_{j=1}^{\infty} \frac{1}{\lambda_j^{\Omega_0}} < \Omega_0 \varphi_j^{\Omega_0}, \varphi_j^{\Omega_0} > \\ &= \sum_{j=1}^{\infty} 1 = \infty. \end{aligned}$$

Hence, by Kolmogorov's Theorem $||x - x_0||^2_{\Omega_0} = \infty$ with nonzero probability. ∎

There is not way to have prior consistency, or more precisely it is not possible for the specified prior distribution to generate a true value x_* such that $(x_* - x_0) \in \mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.(\Omega_0)$.

To summarize, the regularized posterior distribution, and consequently the regularized posterior mean, are consistent, but the prior distribution is not able to generate the true value of the parameter of interest for which consistency is satisfied. This problem 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. However, since the $\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.$ associated to Ω_0 is dense in L^2_π, the prior distribution is able to generate a trajectory that is very closed to the true one. Van der Vaart and Van Zanten (2000) stress that, since the posterior distribution puts all its mass on the support of the prior, consistency is possible only if the true parameter belongs to this support. Anyway, the prior becomes less and less important as $n \to \infty$ because the data swamp the prior. This is exactly what happens here: the posterior distribution converges to the true value of the parameter even if the true value is not in the support of the prior.

5 The case with unknown operator K

Until now we have analyzed a functional equation $Y = Kx$ where the LHS is observed with error or estimated and the operator K is perfectly known. This formulation is correct in most common inverse problems, but in a lot of econometric or statistical inverse problems both Y and K are unknown and estimated. Otherwise stated, in such situations we are faced with the *stochastic ill-posed problem* described in Vapnik (1998) [32]. Examples are the nonparametric instrumental regression model described in Darolles, Florens and Renault (2006) [6], the conditional density estimation, the regression function estimation, etc... In this section, we propose to study such a situation where K is an unknown operator.
The statistical specification of the model remain unchanged as in (3)

$$\hat{Y} = Kx + U,$$

with Assumptions 1, 2 and 3 still valid. Actually, when operator K is unknown we can define the measurement error U in two different ways and this is due to the fact that Y and Kx are equal but are not the same object. More specifically, we can consider the estimation error of Y, $U = \hat{Y} - Kx$, or the difference between the two estimated quantities: $U = \hat{Y} - \hat{K}x$. In this paper we concentrate on the first definition of U and we develop the second approach in Florens and Simoni (2007) [15] for the specific case of instrumental variables where it is possible to recover an asymptotic distribution for $U = \hat{Y} - \hat{K}x$.

The solution of (2) is still defined as the *regularized posterior distribution* with mean and covariance operator given in (14) and (15). However, these quantities are functions of K and then, since operator K is unknown, we are not able to compute them. The simple idea is to not consider K in a bayesian way but to replace this operator with a classical consistent estimator of it. We denote with $\hat{K}$ the consistent estimator of K and with $(\hat{K^*})$ the consistent estimator of the adjoint K^*. In general, $(\hat{K})^* \neq (\hat{K^*})$. The solution to ill-posed inverse problem (2) is the *estimated regularized posterior distribution.*

Definition 2 *Let α_n be a parameter that converges to zero with the sample size n. We define the* estimated regularized posterior distribution *as a Gaussian measure on $\mathcal{X}$ characterized by the* estimated regularized mean function

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

and the estimated regularized variance operator

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

In the following of this section we will denote the estimated regularized posterior distribution with $\hat{\mu}_\alpha$.

5.1 Asymptotic Analysis

We proceed to analyze consistency of the estimated regularized posterior distribution and we adopt a frequentist notion of consistency as described in Section 4. We want to prove that $\hat{\mu}_\alpha$ accumulates in a neighborhood of the true value of x, denoted with x_*, in P^{x_*}-probability as the sample size n becomes large. For that, we have already shown it is sufficient to prove convergence, in a suitable norm, of the estimated regularized mean function to the true value x and of the estimated regularized variance operator to zero.
We start by analyzing convergence of the mean function and we consider the estimation error

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

where $x_\alpha = \mathbb{E}_\alpha(x|\hat{Y})$ denotes the *regularized Bayesian solution* of the inverse problem with known operator K. We have decomposed the estimation error in two error terms: the first one takes into account the estimation error of the operator K and the second term is the error due to having estimated x by using the regularized posterior mean. We analyze convergence in L^2_π

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

where the norm $||\cdot||$ mean norm in L^2_π. The second component of the estimation error has already been analyzed in depth in Section 4 and it has been proved that it converges to zero at the optimal speed $n^{-\frac{\beta}{\beta+1}}$. Therefore, consistent estimation of function x requires that $(\hat{x}_\alpha - x_\alpha)$ converges towards zero.

Assumption 4 *(a)* $||\Omega_0^{\frac{1}{2}}\hat{K}^*\hat{K}\Omega_0^{\frac{1}{2}} - \Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}}||^2 = \mathcal{O}_p(\delta_1)$;
(b) $||\hat{K}\Omega_0\hat{K}^* - K\Omega_0K^*||^2 = \mathcal{O}_p(\delta_2)$;
(c) $||\Omega_0^{\frac{1}{2}}\hat{K}^*\hat{K} - \Omega_0^{\frac{1}{2}}\hat{K}^*K||^2 = \mathcal{O}_p(\delta_3)$;
(d) $||\hat{K}^*||^2 = \mathcal{O}_p(1)$;
(e) $||\hat{K}^* - K^*||^2 = \mathcal{O}_p(1)$;
(f) $||\hat{K} - K||^2 = \mathcal{O}_p(1)$.

The rates of convergence δ_1, δ_2 and δ_3 depend on the type of operator and of estimator used for it. Therefore, they will be determined contextually to every problem. However, in most cases, even if we use a nonparametric estimator for the operator K, the rates δ_1, δ_2 and δ_3 are faster than the nonparametric speed of convergence since the smoothing step by application of $\Omega_0^{\frac{1}{2}}\hat{K}^*$ or $\hat{K}\Omega_0$ allows to improve the speed of convergence, sometimes even to reach the parametric one (*e.g.* instrumental variable estimation, see Florens and Simoni (2007) [15]). The following theorem states consistency of the *estimated regularized Bayesian solution* $\hat{x}_\alpha$, see Appendix 8.5 for the proof.

Theorem 6 *Under Assumptions 4 (a)-(f) we get:*

$$||\hat{x}_\alpha - x_*||^2 = \mathcal{O}_p(\delta_1\alpha^{\beta-2} + \frac{1}{\alpha^3}||\Sigma_n||^2 + \frac{1}{\alpha^2}tr\Sigma_n(1 + \frac{1}{\alpha}(\delta_2 + ||\Sigma_n||^2)) + \frac{1}{\alpha^2}\delta_3 + \frac{1}{\alpha}tr\Sigma_n + \alpha^\beta)$$

if $(x_* - x_0) \in \mathcal{R.K.H.S.}(\Omega_0)$ *and* $\Omega_0^{-\frac{1}{2}}(x_* - x_0) \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})$.

Moreover, if $\frac{\delta_1}{\alpha^2} \sim \mathcal{O}_p(1)$, $\frac{\delta_2}{\alpha} \sim \mathcal{O}_p(1)$, $\frac{\delta_3}{\alpha^2} \sim o_p(1)$, $tr\Sigma_n \sim \mathcal{O}_p(\frac{1}{n})$, $||\Sigma_n|| \sim \mathcal{O}_p(\frac{1}{n})$, $\alpha_n^2 n \to \infty$ *and* $\alpha_n \to 0$, *then*

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

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

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 $\frac{1}{\alpha_n^3}(tr\Sigma_n)^{\frac{3}{2}}\frac{\delta_2}{(tr\Sigma_n)^{\frac{1}{2}}}$. The first factor of this expression converges to 0 under the hypothesis in Theorem 6, the second factor is particularly interesting. In fact, it is the ratio between the estimating error of the operator and the measurement error in our statistical inverse problem. For this ratio being bounded it is necessary that the combination of estimated operators does not converge too fast with respect to the residuals. In other words, the combination of estimated operators must have at least the same speed as the measurement error, otherwise the ratio explodes as $n \to \infty$.
We consider now the estimated regularized posterior variance, we state in the next theorem its convergence to zero and we prove this result in Appendix

Theorem 7 *Under Assumptions 4 (a), (d),* $\forall\varphi \in L_\pi^2$

$$||\hat{V}_\alpha\varphi||^2 = \mathcal{O}_p(\delta_1\alpha^{\beta-2} + \frac{1}{\alpha_n^3}||\Sigma_n||^2 + \alpha_n^\beta)$$

if $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\Omega_0^{\frac{1}{2}}K^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$.

Moreover, if $\frac{\delta_1}{\alpha^2} \sim \mathcal{O}_p(1)$, $||\Sigma_n|| \sim \mathcal{O}_p(\frac{1}{n})$, $\alpha_n^2 n \to \infty$ *and* $\alpha_n \to 0$, *then* $\forall\varphi \in L_\pi^2$

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

in P^x*-probability as* $n \to \infty$.

5.2 Determination of the optimal α

6 The case with different operator for each observation

We are now presenting a slightly modified version of model (3). Suppose to observe an n-sample of curves $\hat{Y}_1, \ldots, \hat{Y}_n$, $\hat{Y}_i \in \mathcal{Y}\ \forall\, i$, each of them is a noisy transformation of the parameter of interest x through an observation specific transformation, namely operator K changes with the index of observation. More clearly,

$$\hat{Y}_i = K_i x + U_i \qquad i = 1, \ldots, n \qquad U_i \sim iid \tag{24}$$

where we still assume K_i, $i = 1, \ldots, n$, is known, non-random, Hilbert-Schmidt and injective operator. This is the classical linear regression model with fixed regressors, where the operators play the role of explanatory variables. In the easiest situation we could directly observe n infinite dimensional objects or curves. However, more realistically, we could suppose to dispose of n samples each one composed of n_i, $i = 1, \ldots, n$ discrete observations. $\hat{Y}_i$ is then obtained by transforming these discrete observations in an infinite dimensional object (*i.e.* a curve). To stay as general as possible we denote with N the whole number of observed object and with n the number of curves. Therefore, for the first observational scheme $N \equiv n$ and for the second one $N = n_1 + \ldots + n_n$ and n is the number of samples we are considering. The asymptotic is considered for $N \to \infty$ with the assumption for the second observational scheme that $N \to \infty$ if and only if the number of observations in all the samples goes to infinity, *i.e.* $n_i \to \infty \, \forall i = 1, \ldots n$. The parameter of interest x belongs to the probability space $(\mathcal{X}, \mathcal{E}, \mu)$ where the prior distribution μ is characterized by Assumption 2. On the contrary, Assumption 1 is replaced by Assumption 5 below since the sampling distribution is no more the same for all the observations but it changes from an observation to another one.

Assumption 5 *Let P_i^x be a probability measure on $(\mathcal{Y}, \mathcal{F})$ conditioned on $\mathcal{E}$ such that $\mathbb{E}(||\hat{Y}_i||^2) < \infty$. P_i^x is a Gaussian measure that defines a mean element $K_i x \in \mathcal{Y}$ and a covariance operator $\Sigma_N : \mathcal{Y} \to \mathcal{Y}$ such that $\Sigma_N \to \infty$ when $N \to \infty$.*

Therefore,

$$\hat{Y}_i | x \sim i\mathcal{GP}(K_i x, \Sigma_N) \qquad i = 1, \ldots, n. \tag{25}$$

and the covariance operators satisfy Assumption 3.

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

$$\Xi_D = (\mathcal{X} \times \mathcal{Y}^n, \mathcal{E} \otimes \mathcal{F}^n, \Pi^n), \tag{26}$$

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

The following Lemma is only an adaptation of Theorems 1 and 2 to the particular setting with different operators. For this reason the proof is skep.

Lemma 4 *Under Assumptions 1 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 $m_{xy}^n = (x_0, Kx_0) \in \mathcal{X} \times \mathcal{Y}^n$ and covariance operator Υ^n such that $\Upsilon^n(\varphi, \psi) = (\Omega_0 \varphi + \Omega_0 K^* \psi, K\Omega_0 \varphi + (I_n \otimes \Sigma_N + K\Omega_0 K^*)\psi)$, for all (φ, ψ) in $\mathcal{X} \times \mathcal{Y}^n$.

(ii) The marginal distribution $P = P_1 \otimes \ldots \otimes P_n$ on $(\mathcal{Y}^n, \mathcal{F}^n)$ is a gaussian measure with mean function $m_y^n = Kx_0 \in \mathcal{Y}^n$ and covariance operator $\Upsilon_{yy}^n = (I_n \otimes \Sigma_N + K\Omega_0 K^*)$.

We rewrite in matrix form the covariance operator of the n-dimensional gaussian process $\hat{Y}$.

$$Var(\hat{Y}) = \begin{pmatrix} \Sigma_N + K_1\Omega_0 K_1^* & K_1\Omega_0 K_2^* & \ldots & K_1\Omega_0 K_n^* \\ K_2\Omega_0 K_1^* & \Sigma_N + K_2\Omega_0 K_2^* & & \\ \vdots & & \ddots & \vdots \\ K_n\Omega_0 K_1^* & \ldots & & \Sigma_N + K_n\Omega_0 K_n^* \end{pmatrix},$$

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

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 (26) 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 some 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.* $\Pi^n_{\mathcal{E}\otimes\mathcal{T}}(A) = \Pi^n(A), \quad \forall A \in \mathcal{E} \otimes \mathcal{T}$. In particular, we are considering the sub-σ-field generated by a sufficient statistic. In this way we are sure to not loose any information by considering the reduced experiment instead of the unreduced one. We consider statistic $t = \sum_{i=1}^n K_i^* \hat{Y}_i$ and we are going to prove that it is sufficient for our Bayesian experiment, that means $\mathcal{F}^n \underline{\parallel} \mathcal{E} | \sigma(K^*\hat{Y})$, where $\sigma(K^*\hat{Y})$ denotes the σ-field generated by statistic t. Actually, sufficiency of statistic t entails sufficiency of every transformation of $\sum_{i=1}^n K_i^* \hat{Y}_i$.
Due to the fact that we are working in infinite dimension and we have not a likelihood function, we can not use the factorization criterion in order to prove sufficiency. Hence, we consider a sequential model by projecting the model on an orthonormal bases and we 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_N, where we have skip the index N for simplicity. Sequential Bayesian Experiment is defined by

$$\Xi_{Ds} = (\mathcal{X} \times \mathcal{Y}^n, \mathcal{E} \otimes \mathcal{F}^n, \Pi^n, \mathcal{E}_k \uparrow \mathcal{E}_\infty, \mathcal{F}_k^n \uparrow \mathcal{F}_\infty^n), \tag{27}$$

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: $\mathcal{E}_k = \sigma(\{< x, K^*\psi_j >\}_{j=1,\dots,k})$[3]. The filtration $\mathcal{F}_k^n$ is generated by the n-dimensional vector of projected observed curves $\hat{Y}$: $\mathcal{F}_k^n = \sigma(\{< \hat{Y}, \psi_j > \}_{j=1,\dots,k})$. 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

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

with $< U_i, \psi_j > \sim \mathcal{N}(0, \lambda_j)$ and $Cov(< U_i, \psi_j >, < U_i, \psi_{j'} >) = 0$, $\forall j \neq j'$.

Lemma 5 *The statistic $t = \sum_{i=1}^n K_i^* \hat{Y}_i(k)$ is sufficient in the sequential Bayesian experiment Ξ_{Ds}.*

Proof: Consider the likelihood function of the sequential experiment (27):

$$L(\{< x, K_i^*\psi_j >\}_{ij} | \{\hat{y}_{ij}\}_{ij}) = \prod_i [\frac{1}{(2\pi)^{\frac{k}{2}}} \prod_j \lambda_j^{-\frac{1}{2}} \exp\{-\frac{1}{2}\frac{1}{\lambda_j}(\hat{y}_{ij} - < x, K_i^*\psi_j >)^2\}], \tag{28}$$

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

$$\begin{aligned} l(\{< x, K_i^*\psi_j >\}_{ij} | \{\hat{y}_{ij}\}_{ij}) &\propto \sum_{ij} \frac{1}{\lambda_j}(\hat{y}_{ij} - < x, K_i^*\psi_j >)^2 \\ &\propto \sum_{ij} (< \hat{Y}_i, \tilde{\psi}_j >^2 + < x, K_i^*\tilde{\psi}_j >^2 \\ &\quad -2 < \hat{Y}_i, \tilde{\psi}_j >< x, K_i^*\tilde{\psi}_j >), \end{aligned} \tag{29}$$

where $\tilde{\psi}_j = \frac{\psi_j}{\sqrt{\lambda_j}}$ is the normalized singular value. By the factorization principle, t 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 (29) in such a way to obtain a function of x and $K^*\hat{Y}(k)$:

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

$$\begin{aligned}\sum_{ij} < \hat{Y}_i, \tilde{\psi}_j >< x, K_i^* \tilde{\psi}_j > &= < x, \sum_{ij} < \hat{Y}_i, \tilde{\psi}_j > K_i^* \tilde{\psi}_j > \\ &= < x, \sum_i K_i^* \sum_j < \hat{Y}_i, \tilde{\psi}_j > \tilde{\psi}_j > \\ &= < x, \sum_i K_i^* \hat{Y}_i(k) > \\ &= < x, K^* \hat{Y}(k) >,\end{aligned}$$

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

Let $\mathcal{T}_k = \sigma(K^* \hat{Y}(k))$ be the sub-σ-field generated by $K^* \hat{Y}(k)$. We rewrite sufficiency in a more technical way: $\mathcal{T}_k$ is a sufficient statistic for the Bayesian projected experiment Ξ_{Ds} if and only if $\mathcal{F}_k^n \underline{\parallel} \mathcal{E}_k | \mathcal{T}_k$. Moreover, $\mathcal{E}_\infty = \bigvee_{k\geq 0} \mathcal{E}_k$ and $\mathcal{F}_\infty^n = \bigvee_{k\geq 0} \mathcal{F}_k^n$, and we define the *tail σ-field* $\mathcal{T}_T$ as

$$\mathcal{T}_T = \bigcap_{k\geq 0} \bigvee_{m\geq k} \mathcal{T}_m, \tag{30}$$

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.* $\mathcal{T}_T = \sigma(K^* \hat{Y})$. Our point is to prove that $\mathcal{T}_T$ is a sufficient statistic for the initial Bayesian experiment Ξ_D, namely: $\mathcal{F}^n \underline{\parallel} \mathcal{E} | \mathcal{T}_T$. The following theorem, that is a slightly modification of Theorem 7.2.7 in Florens, Mouchart and Rolin (1990), shows that sufficiency in the sequential Bayesian Experiment implies sufficiency in the limit Bayesian Experiment.

Theorem 8 *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) $\mathcal{F} \perp \mathcal{E} | \mathcal{T}_T$.

Proof: 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) $\mathbb{E}(a|\mathcal{F}_k) \rightarrow^{L^1} \mathbb{E}(a|\mathcal{F}_\infty) \qquad a.s.$
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, $\limsup_k \mathbb{E}(a|\mathcal{T}_k) = \limsup_k \mathbb{E}(a|\overline{\mathcal{T}}_k) = \mathbb{E}(a|\limsup_k \overline{\mathcal{T}}_k)$ [4] 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) $\mathbb{E}(a|\mathcal{F}_\infty) = \mathbb{E}(a|\overline{\mathcal{T}_T}) = \mathbb{E}(a|\mathcal{T}_T)$.

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

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 $\mathcal{F}_\infty \perp \mathcal{E}_\infty | \mathcal{T}_T$ ∎

This theorem applies to Bayesian Experiment defined in (26) 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 $\mathcal{F}_k^n$ replaced by $\mathcal{F}_\infty$ and $\mathcal{F}_\infty^n$, respectively. Hence, $K^*\hat{Y}$ is a sufficient statistic for parameter x [5] and inference on x will be done only by using information contained in $\mathcal{T}_T$. However, being interested in asymptotic properties, $t = K^*\hat{Y}$ must be divided by n since otherwise it is not well defined for n big, then t will be used to denote $t = \frac{1}{n}K^*\hat{Y}$.
The sampling probability restricted to $\mathcal{T}_T$, $P^x_{\mathcal{T}_T}$, is a gaussian measure with mean function $\frac{1}{n}K^*Kx$ and covariance operator $\frac{1}{n^2}K^*(I_n \otimes \Sigma_N)K$. The joint measure restricted to $\mathcal{E} \otimes \mathcal{T}_T$ is gaussian with mean $(x_0, \frac{1}{n}K^*Kx_0)$ and covariance operator

$$\begin{bmatrix} \Omega_0 & \frac{1}{n}\Omega_0 K^*K \\ \frac{1}{n}K^*K\Omega_0 & \frac{1}{n^2}(K^*\Sigma_N K + K^*K\Omega_0 K^*K) \end{bmatrix}$$

and this entails that also marginal distribution $P_{\mathcal{T}_T}$ restricted to $\mathcal{T}_T$ is Gaussian.
The solution to ill-posed problem (24) restated in a larger space of probability distributions is the conditional distribution $\mu^{\mathcal{T}_T}$ of x given the observed t. $\mu^{\mathcal{T}_T}$ is a gaussian process with mean function $\mathbb{E}(x|t) = \frac{1}{n}BK^*\hat{Y} + c$ and covariance operator W. As for the general case, this distribution, and in particular the posterior mean, is not continuous in t, causing a problem of posterior inconsistency. Indeed, $\forall \varphi_1, \varphi_2 \in \mathcal{X}$,

$$\begin{aligned} Cov(<x,\varphi_1>,<t,\varphi_2>) &= Cov(<\mathbb{E}(x|t),\varphi_1>,<t,\varphi_2>) \\ &= Cov(<t,B^*\varphi_1>,<t,\varphi_2>) \\ &= <Var(t)B^*\varphi_1,\varphi_2>, \end{aligned}$$

and operator B is identified by the equality with $Cov(<x,\varphi_1>,<t,\varphi_2>) = <\frac{1}{n}K^*K\Omega_0\varphi_1,\varphi_2>$:

$$\frac{1}{n^2}(K^*\Sigma_N K + K^*K\Omega_0 K^*K)B^*\varphi_1 = \frac{1}{n}K^*K\Omega_0\varphi_1$$

that can not be continuously solved. Therefore, we apply a Tikohonv scheme for rgularize it and we define, also for the case with different operators, a new object called *regularized posterior distribution* that we guess it is the solution of (24). The regularized quantities defining the posterior distribution are

$$\begin{aligned} B_\alpha &= \Omega_0\frac{1}{n}K^*K(\alpha_n I + \frac{1}{n^2}K^*\Sigma_N K + \frac{1}{n^2}K^*K\Omega_0 K^*K)^{-1}, \\ c_\alpha &= (I - B_\alpha\frac{1}{n}K^*K)x_0, \\ W_\alpha &= \Omega_0 - B_\alpha\frac{1}{n}K^*K\Omega_0 \\ &= \Omega_0 - \Omega_0\frac{1}{n}K^*K(\alpha_n I + \frac{1}{n^2}K^*\Sigma_N K + \frac{1}{n^2}K^*K\Omega_0 K^*K)^{-1}\frac{1}{n}K^*K\Omega_0 \end{aligned} \quad (31)$$

Remember that function c_α is identified through the relation $\mathbb{E}(x) = \mathbb{E}(\mathbb{E}(x|\frac{1}{n}K^*Y))$ and W_α through the formula $\Omega_0 = \mathbb{E}(W) + Var(\mathbb{E}(x|\frac{1}{n}K^*Y))$. Therefore, the regularized posterior mean is

$$\mathbb{E}_\alpha(x|t) = \Omega_0\frac{1}{n}K^*K(\alpha_n I + \frac{1}{n^2}K^*\Sigma_N K + \frac{1}{n^2}K^*K\Omega_0 K^*K)^{-1}(t - \frac{1}{n}K^*Kx_0) + x_0. \quad (32)$$

[5] We would like to point out that sufficiency of $K^*\hat{Y}$ entails sufficiency of the OLS estimator. Actually, in functional spaces the OLS estimator, solution of the minimization problem

$$min_x \sum_{i=1}^{n}(\hat{Y}_i - K_i x)^2.$$

must be regularized: $x_\alpha^{OLS} = (\alpha_n I + K^*K)^{-1}K^*\hat{Y}$, for some parameter $\alpha_n \to 0$.

6.2 Asymptotic Analysis

As we have already discussed in Section 4, we analyze asymptotic properties of the posterior distribution in a sampling (or frequentist) sense, namely convergence in P^x-probability. We refer to this section for a discussion on justification and for comments concerning this concept of consistency. The reasoning to derive convergence and speed of convergence is essentially the same, thus the details given here will be minimal.
Let $x_* \in \mathcal{X}$ be the true value that characterizes the data generating process. Posterior consistency means convergence of the regularized posterior distribution toward the Dirac mass in x_* and convergence will be in $P^{x_*}_{\mathcal{T}_T}$-probability. Because of existence of an unique correspondence between a gaussian measure and its mean and covariance we will limit us to prove consistency of $\mathbb{E}_\alpha(x|t)$ and convergence to zero of $W_\alpha\varphi$ for all $\varphi \in \mathcal{X}$. The following Theorem formalizes convergence to zero of the bias of the regularized posterior mean.

Theorem 9 *Consider inverse problem (24) and the regularized posterior distribution with mean 32 and variance 31. Then:*

(i) if $\frac{1}{n^2}K^*Var(Y)K \to Q$ *with* Q *a bounded operator,* $(x_* - x_0) \in \mathcal{R.K.H.S.}(\Omega_0)$ *and if* $\Omega_0^{-\frac{1}{2}}(x_* - x_0) \in \mathcal{R}(\frac{1}{n^2}\Omega_0^{\frac{1}{2}}K^*KK^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ *the bias is of order:*

$$||\mathbb{E}_\alpha(x|t) - x_*||^2 = \mathcal{O}_p(\alpha_N^\beta) + \mathcal{O}_p(\frac{1}{\alpha_N}tr(\frac{K^*\Sigma_N K}{n^2})) + \mathcal{O}_p(\frac{1}{\alpha_N^3}\Big|\Big|\frac{K^*\Sigma_N K}{n^2}\Big|\Big|^2).$$

(ii) Moreover, if $\frac{1}{\alpha_N}tr(\frac{K^*\Sigma_N K}{n^2}) \to 0$, $\frac{1}{\alpha_N^3}||\frac{K^*\Sigma_N K}{n^2}||^2 \to 0$ *and* $\alpha_n \to 0$ *then* $\mathbb{E}_\alpha(x|t) \to^{P^x_{\mathcal{T}_T}} x_*$ *in* $\mathcal{X}$ *norm.*

It should be noted that in general we can suppose $tr(\frac{K^*\Sigma_N K}{n^2})$ is of the same order as $||\frac{K^*\Sigma_N K}{n^2}||$, in particular, this is satisfied when $tr(\frac{K^*\Sigma_N K}{n^2}) \sim ||\frac{K^*\Sigma_N K}{n^2}|| \sim \mathcal{O}_p(\frac{1}{N})$ that is very frequent for U_i being an estimation error. In this case $\frac{1}{\alpha_N}tr(\frac{K^*\Sigma_N K}{n^2})$ is negligible with respect to $\frac{1}{\alpha_N^3}||\frac{K^*\Sigma_N K}{n^2}||^2$ and the optimal α_N will be

$$\alpha_N^* \propto \Big|\Big|\frac{K^*\Sigma_N K}{n^2}\Big|\Big|^{\frac{2}{\beta+3}}.$$

In order to have posterior consistency it is only necessary to have convergence to zero of the regularized posterior variance as it is shown in the next theorem:

Theorem 10 *(i) if,* $\forall\varphi \in \mathcal{X}$, $\Omega_0^{\frac{1}{2}}\varphi \in \mathcal{R}(\frac{1}{n^2}\Omega_0^{\frac{1}{2}}K^*KK^*K\Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$ *and* $\frac{1}{n^2}K^*Var(Y)K \to Q$ *with* Q *a bounded operator then*

$$||W_\alpha\varphi||^2 = \mathcal{O}_p(\alpha_N^\beta + \frac{1}{\alpha_N^3}\Big|\Big|\frac{K^*\Sigma_N K}{n^2}\Big|\Big|^2).$$

(ii) Moreover, if $\frac{1}{\alpha_N}||\frac{K^*\Sigma_N K}{n^2}|| \to 0$ *and* $\alpha_N \to 0$, *then* $W_\alpha\varphi \to 0$ *in* $\mathcal{X}$ *norm.*

We omit proof of this theorem since it is based on standard arguments already used in previous computations. We only show the basic decomposition that has been employed:

$$\begin{aligned} W_\alpha\varphi &= (\Omega_0 - \Omega_0\frac{1}{n}K^*K(\alpha_N I + \frac{1}{n^2}K^*K\Omega_0K^*K)^{-1}\frac{1}{n}K^*K\Omega_0)\varphi \\ &\quad -\Omega_0\frac{1}{n}K^*K(\alpha_N I + \frac{1}{n^2}K^*Var(Y)K)^{-1}(-\frac{1}{n}K^*\Sigma_N K)(\alpha_n I + \frac{1}{n^2}K^*K\Omega_0K^*K)^{-1}\frac{1}{n}K^*K\Omega_0\varphi, \end{aligned}$$

where the first two terms contribute for the first speed α_N^β and the last one for the second speed of convergence. Then, the optimal speed of convergence of the regularized posterior mean, obtained for the optimal α_N^*, is of order $||\frac{K^*\Sigma_N K}{n^2}||^{\frac{2\beta}{\beta+3}}$.

As conclusion, the regularized posterior distribution is consistent, so that it degenerates to a mass in x_*. However, we have the same result of prior inconsistency found for the general case: due to the fact that every trajectory $(x - x_0)$ drawn from the prior distribution does not belong to $\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.(\Omega_0)$ with μ-probability 1, the prior distribution is not able to generate the true value of parameter for which consistency is verified. Consequently, the prior distribution is not well-specified.

7 Numerical Implementation

We will show in this section the performance of the regularized posterior mean as estimator for the solution of ill-posed inverse problem (2) by developing some numerical simulation. First, we will consider numerical implementation of an arbitrary data generating process and a regular parameter of interest. Then we will interest in the estimation of a density function and lastly, of a regression function.

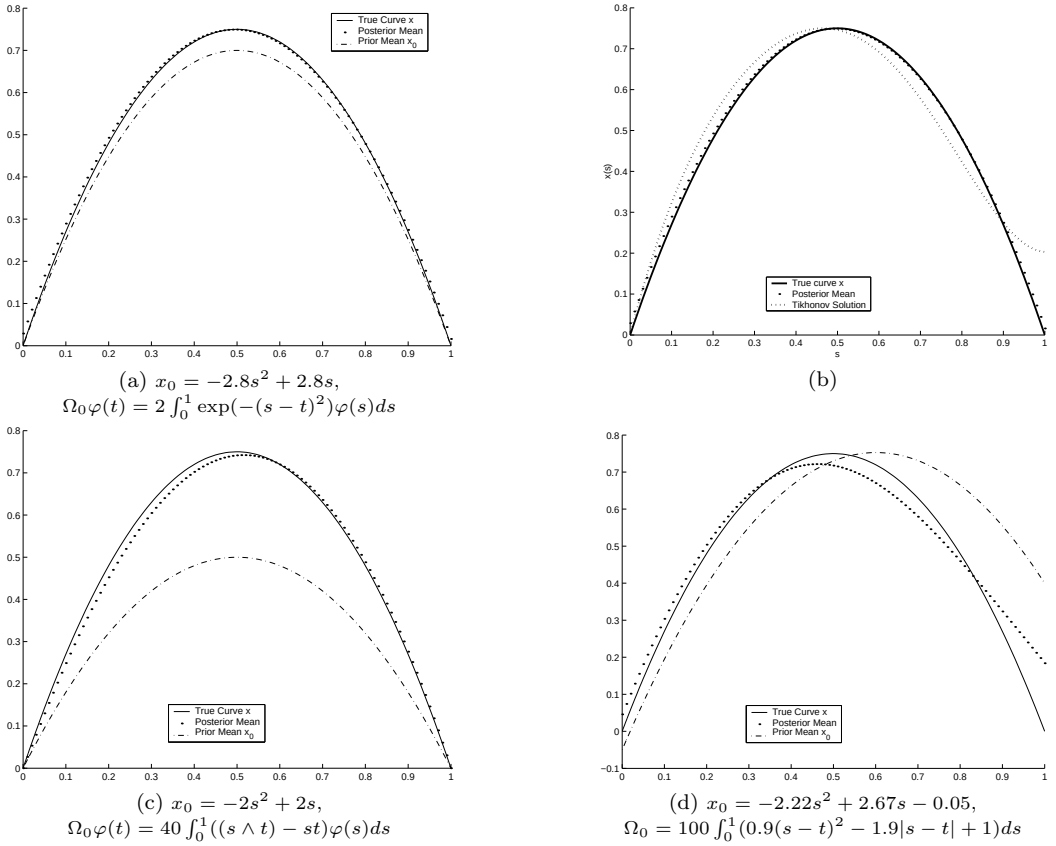
7.1 Functional equation with a parabola as solution

We take L^2 spaces as spaces of reference; more appropriately, $\mathcal{X} = L^2_\pi$ and $\mathcal{Y} = L^2_\rho$, with measures π and ρ taken equal to the uniform measure on $[0, 1]$. The data generating process we have considered is

$$\begin{aligned} \hat{Y} &= \int_0^1 x(s)(s \wedge t)ds + U, \qquad U \sim \mathcal{GP}(0, \Sigma_n) \\ \Sigma_n &= n^{-1} \int_0^1 \exp\{-(s-t)^2\}ds \\ x &\sim \mathcal{GP}(x_0, \Omega_0) \\ x_0 &= -2.8s^2 + 2.8s \\ \Omega_0\varphi(t) &= \omega_0 \int_0^1 \exp(-(s-t)^2)\varphi(s)ds. \end{aligned} \tag{33}$$

The true value of the parameter of interest x has been taken to be a parabola: $x = -3s^2 + 3s$; this choice is not completely arbitrary but it is consistent with the specification of the prior mean and variance. The covariance operators are correctly chosen in the sense that they are self-adjoint, positive semi-definite and nuclear. In particular, their eigenvalues are $\mathcal{O}(e^{-j})$. The regularization parameter α has been set to $2.e-03$, the discretization step is of 0.01 and the sample size is $n = 1000$.
We show in Figure 1a the true function x and the regularized posterior mean estimation for the specified prior with $\omega_0 = 2$. Moreover, we propose, in Figure 1b a comparison between our estimator and the estimator obtained by solving equation (2) with a classical Tikhonov regularization method. To get good results for the Tikhonov estimator we have chosen $\alpha = 2.e-04$.
The choice of the prior distribution is deeply affecting the estimation. To analyze its role we have replicated the same simulation experiment for different specifications of prior distribution. It should be noted that the far the prior mean is from the true parameter the bigger should be the prior covariance operator. The way in which the prior covariance is specified allows to easily increase or decrease it by only changing parameter ω_0. In the first variation of the prior, we consider a prior mean $x_0 = -2s^2 + 2s$, a scale covariance parameter $\omega_0 = 40$ and a covariance kernel equal to the brownian bridge covariance: $\Omega_0\varphi(t) = 40 \int_0^1 ((s \wedge t) - st)\varphi(s)ds$. This covariance operator has eigenvalues of order $\mathcal{O}(j^{-2})$, *i.e.* $\lambda_j = 1/(\pi^2 j^2)$, $j \in \mathbb{N}$. The estimated curve, pictured in Figure 1c, shows that, despite a prior mean far from the true curve, we still get a pretty good estimation. The second variation of the prior specification is the following: the prior mean is $x_0 = -2.22s^2 + 2.67s - 0.05$ and the kernel of covariance operator is now a parabola with scale parameter $\omega_0 = 100$: $\Omega_0 = 100 \int_0^1 (0.9(s-t)^2 - 1.9|s-t| + 1)ds$. The results are shown in Figure 1d. Here we need a large variance to compensate the bad prior information contained in the prior mean.

The choice of the regularization parameter α is particular troublesome. We have adopted here an ad-hoc choice of it, the object of the simulation being to verify the good performance of our estimator. Determination of this parameter is a problem concerning all the methods proposed to solve inverse problems. Therefore the aim of this section is not the optimal choice of α but only analysis of the fit of our estimator for a given α. The evolution of the regularized posterior mean estimator for different values of the regularization parameter α is shown in Figure 2a. Figure 2b is a particular of the previous one.
Finally, in Figure 3 results of a Monte Carlo experiment with 100 iterations are shown. Panels (3a), (3c) and (3d) 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 (3b) 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 (dashed 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)

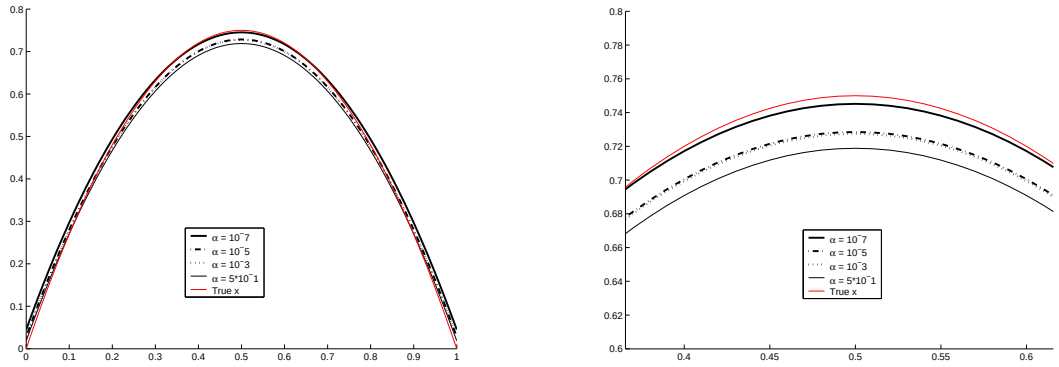
(c) $x_0 = -2s^2 + 2s$,
$\Omega_0\varphi(t) = 40\int_0^1((s \wedge t) - st)\varphi(s)ds$

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

Figure 1: Panels (1a), (1c) and (1d) show the regularized posterior mean estimation and the true solution. Panel (1b) show the comparison between the performance of our estimator and the classical Tikhonov regularized solution.

7.2 Density Estimation

Let $\xi_1, \ldots, \xi_n$ be an observed random sample generated from an unknown distribution F that we assume to be absolutely continuous with respect to the Lebeasgue measure. The aim is to obtain an estimation of the associated density function $f(\xi)$ by solving the functional equation given in Example 1 of subsection 2.2. The notation and all the technicalities are the same as in Example 1. The true parameter of interest f is chosen to be the density of a standard gaussian measure on $\mathbb{R}$ and the measures π and ρ, defining the L^2 spaces, are set equal to an uniform measure on $[-3, 3]$ (*i.e.* $\mathcal{U}[-3, 3]$). We use the sample $\xi_1, \ldots, \xi_n$ to estimate F and the sampling variance Σ_n. The operator K is known and does not need to be estimated.
The prior distribution is specified as follows:



(a) (b)

Figure 2: Prior specification: $x_0 = -2.8s^2 + 2.8s$, $\Omega_0\varphi(t) = 2\int_0^1 \exp(-(s-t)^2)\varphi(s)ds$. Panel (2a) represents the regularized posterior mean estimator for different values of α. Panel (2b) is a zoom of the previous panel.

$$
\begin{aligned}
f_0 &= \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(\xi-\theta)^2\} \\
\Omega_0\varphi(t) &= \omega_0\int_{-3}^{3}\exp(-(s-t)^2)\varphi(s)\frac{1}{6}ds.
\end{aligned}
$$

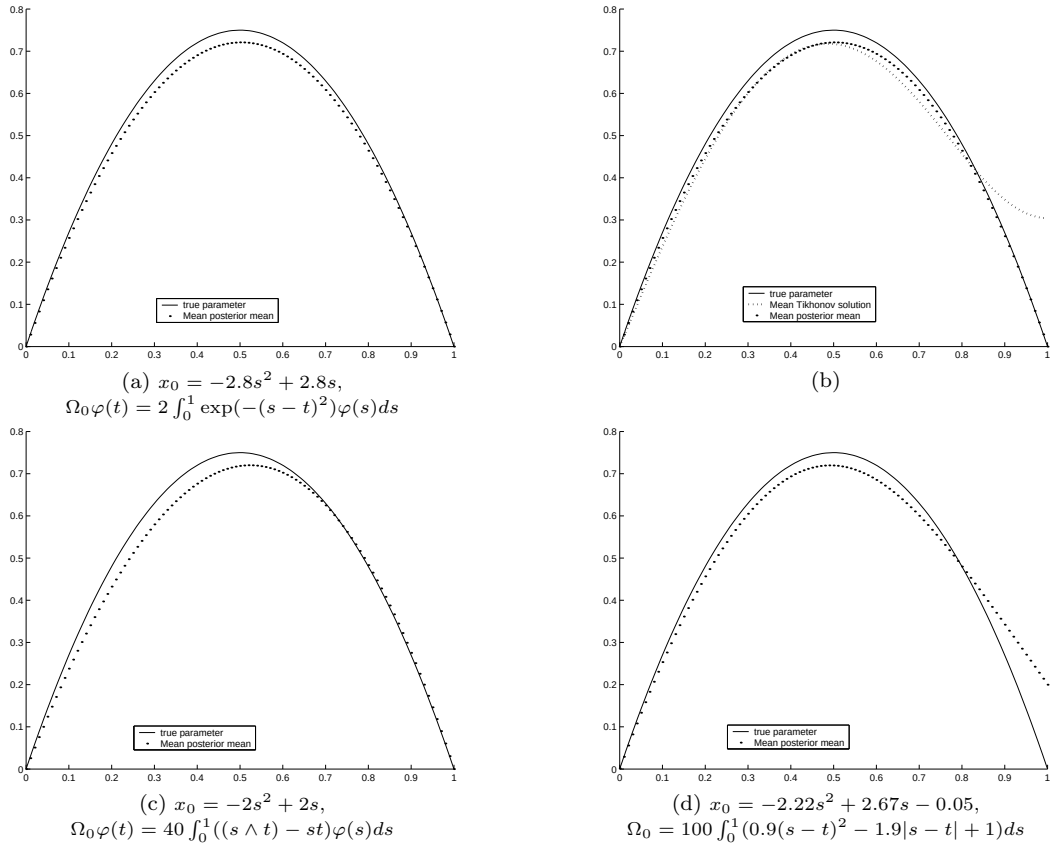
Parameters $(\sigma, \theta, \omega_0)$ have been differently set to see the effect of prior changes on the estimated solution. Figure 4 shows the regularized posterior mean estimator for different specification of the parameters. In panels (a) and (c) the true density, the prior mean and the regularized posterior mean estimator are drawn; panels (b) and (d) show the comparison between our estimator and the classical Tikhonov solution.
We can have an idea about the behavior of the prior distribution by drawing a sample from it. Figure 5 represents a sample of curves dawn from the prior distribution together with the prior mean (in blue) and the true density (dotted line). Lastly, in Figure 6, 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 the true density function.

7.3 Regression Estimation

Consider a real random variable $w \in \mathbb{R}$ with a known normal distribution F with mean 2 and variance 1 and a Gaussian white noise $\varepsilon \sim \mathcal{N}(0,2)$ independently drawn. The regression function $m(w)$ is defined as an element of $L^2_F(w)$ such that $\xi = m(w) + \varepsilon$ and $\mathbb{E}(\varepsilon|w) = 0$. We follow the method outlined in Exemple 2 of Section 2.2 so that, for a given function $g(w,t)$, we can equivalently define $m(w)$ as the solution of the functional equation

$$\mathbb{E}(g(w,t)\xi) = \mathbb{E}(g(w,t)m(w)).$$

We specify a function $g(w,t)$ defining an Hilbert Schmidt operator $K : L^2_F(w) \to L^2_\pi$, with $\pi \sim \mathcal{N}(2,1)$. It is alternatively specified as an exponential function, $g(w,t) = exp(-(w-t)^2)$, or an indicator function, $g(w,t) = 1\{w \le t\}$.
The true regression function is $m_*(w) = cos(w)sin(w)$ and we specify the prior distribution as a Gaussian process: $m(w) \sim \mathcal{GP}(m_0(w), \Omega_0\varphi(w))$, for any $\varphi \in L^2_F(w)$. After having drawn a sample of (ξ, w) we can 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.
CASE I: $g(w,t) = 1\{w \le t\}$. The prior covariance operator is defined through an exponential kernel, $\Omega_0\varphi(w_1) = \omega_0 \int \exp(-(w_1 - w_2)^2)\varphi(w_2)f(w_2)dw_2$, with $\omega_0 = 2$ or $\omega_0 = 10$, and we have



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

(b)

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

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

Figure 3: Monte Carlo experiment.

considered three different prior mean specifications: $m_0(w) = cos(w)sin(w)$, $m_0(w) = 0.067w - 0.2$, or $m_0 = 0$. Figure 7 shows the results for these prior specifications. Despite to the very bad prior mean, we are able to find noteworthy estimation. 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.

CASE II: $g(w,t) = \exp(-(w-t)^2)$. We have conducted single estimations and Monte Carlo experiments for the same prior mean specifications as in Case I. The prior covariance kernel is alternatively specified as an exponential function or as an hyperbola and the parameter ω_0 takes the values of $\omega_0 = 2$, $\omega_0 = 15$ or $\omega_0 = 20$. In Figure 8 is illustrated the results that we have obtained.

8 Appendix

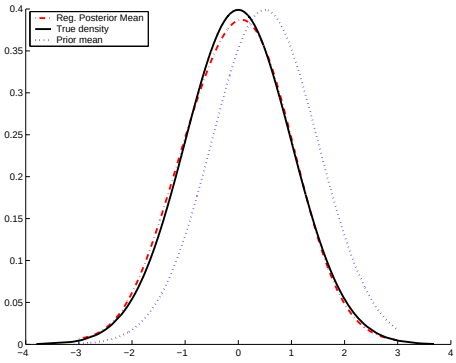
8.1 Sufficiency of the sample mean

Suppose to observe n trajectories of a Gaussian Process $\hat{Y}_i$, $i = 1, \ldots, n$, satisfying statistical model (3):

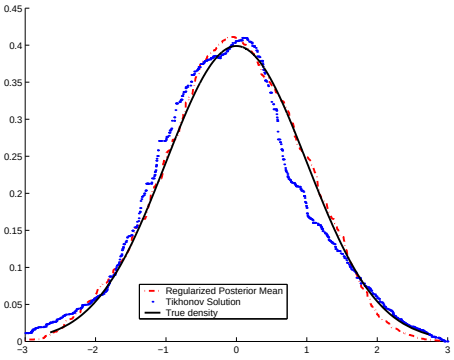
$$\hat{Y} = Kx + U,$$

with all the usual assumptions of Section 2. Let $\mathcal{E} = \sigma(x)$ and $\mathcal{F} = \sigma(\{\hat{Y}_i\}_{i=1,\ldots,n})$ be the σ-fields of the parameter space and of the sample space respectively and $\mathcal{T} = \sigma(\frac{1}{n}\sum_{i=1}^n \hat{Y}_i)$ be the σ-field generated by the sample mean. We prove sufficiency of the sample mean, *i.e.* $\mathcal{E}\underline{||}\mathcal{F}|\mathcal{T}$, by considering the projected model.

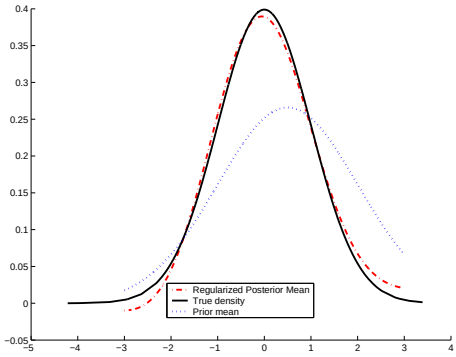
Let $\{\lambda_j; \psi_j\}_j$ be the singular system of the covariance operator Σ of the error term, the projected model is therefore



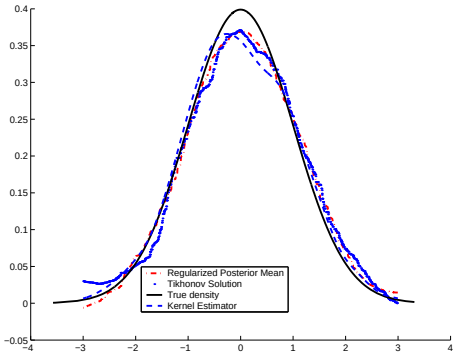
(a) $\sigma = 1$, $\theta = 0.5$, $\omega_0 = 10$



(b) $\sigma = 1$, $\theta = 0.5$, $\omega_0 = 10$



(c) $\sigma = 1.5$, $\theta = 0.5$, $\omega_0 = 10$



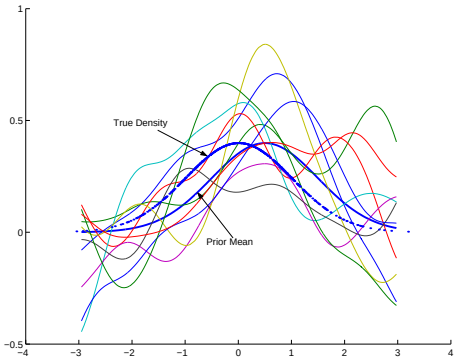
(d) $\sigma = 1.5$, $\theta = 0.5$, $\omega_0 = 10$

Figure 4: Regularized posterior mean and Tikhonov estimators of a density function.

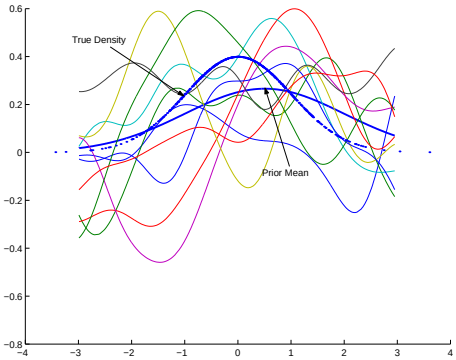
$$\underbrace{< \hat{Y}_i, \psi_j >}_{\equiv y_{ij}} = \underbrace{< Kx, \psi_j >}_{\equiv x_j} + < U_i, \psi_j >, \quad i = 1, \ldots, n \quad j = 1, 2, \ldots$$

with $< U_i, \psi_j > \sim \mathcal{N}(0, \lambda_j)$, $\forall i, j$ and $Cov(< U_i, \psi_j >, < u_i, \psi_{j'} >) = 0$, $\forall j \neq j'$. We only consider a finite number k of projections and, from normality of the error term, I get the loglikelihood:

$$\begin{aligned} \ln L(\{y_{ij}\}_{\substack{i \leq n \\ j \leq k}} | \{x_j\}_{j \leq k}) &\propto \sum_{ij} \frac{1}{\lambda_j}(y_{ij}^2 + x_j^2 - 2 y_{ij} x_j) \\ &\propto \sum_{ij} \frac{1}{\lambda_j} y_{ij}^2 + n \sum_j \frac{1}{\lambda_j} (< x, K^* \psi_j >)^2 \end{aligned}$$

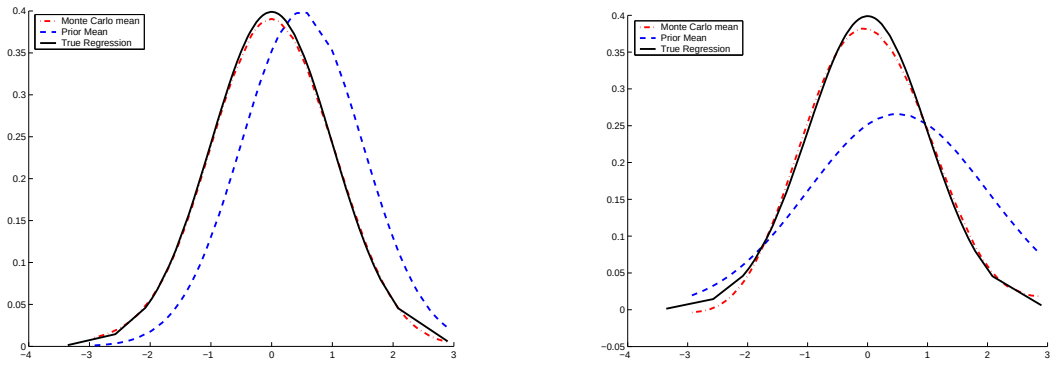


(a) $\sigma = 1$, $\theta = 0.5$, $\omega_0 = 10$



(b) $\sigma = 1.5$, $\theta = 0.5$, $\omega_0 = 10$

Figure 5: Drawn from the prior distribution.



(a) $\sigma = 1$, $\theta = 0.5$, $\omega_0 = 10$ (b) $\sigma = 1.5$, $\theta = 0.5$, $\omega_0 = 10$

Figure 6: Monte Carlo simulation.

$$\begin{aligned} & -2\sum_j \frac{1}{\lambda_j} < \sum_i \hat{Y}_i, \psi_j >< x, K^*\psi_j > \\ \propto \quad & \sum_{ij} \frac{1}{\lambda_j} y_{ij}^2 + n\sum_j \frac{1}{\lambda_j} x_j^2 - 2n\sum_j \frac{1}{\lambda_j} < \bar{Y}, \psi_j > x_j, \end{aligned}$$

where $\bar{Y} = \frac{1}{n}\sum_i \hat{Y}_i$ denotes the sample mean. We have gotten $\ln L(\{y_{ij}\}_{ij}|\{x_j\}_{j\leq k}) = \ln L(\{y_{ij}\}_{ij}|\{< \bar{Y}, \psi_j >\}_{j\leq k}) + \ln L(\{< \bar{Y}, \psi_j >\}_{j\leq k}|\{x_j\}_{j\leq k})$ that, on the basis of factorization principle, shows $\mathcal{E}_k \perp \mathcal{F}_k|\mathcal{T}_k$, where $\mathcal{E}_k = \sigma(\{< x, K^*\psi_j >\}_{j\leq k})$, $\mathcal{F}_k = \sigma(\{< \hat{Y}_i, \psi_j >\}_{\substack{i=1,\dots,n \\ j\leq k}})$ and $\mathcal{T}_k = \sigma(\{< \bar{Y}, \psi_j >\}_{j\leq k})$ are three filtrations. Theorem 8, with the tail σ-field $(\mathcal{T}_k)_T$ given by $\mathcal{T}_\infty = \sigma(\bar{Y})$, allows to conclude.

8.2 Proof of Theorem 1

Let $(\tilde{x}, \tilde{y})$ be an element in $\mathcal{X} \times \mathcal{Y}$. Assumptions 1 and 2 define a conditional probability on $\mathcal{X}$ given $\mathcal{E}$ and a marginal probability on $\mathcal{Y}$ and imply that $\tilde{y}$ can be decomposed in $\tilde{y}_1 + \tilde{y}_2$, with $\tilde{y}_1 \in \mathcal{R}(K)$ and $\tilde{y}_2 \in \overline{\mathcal{R}.\mathcal{K}.\mathcal{H}.\mathcal{S}.(\Sigma_n)}$. More precisely, $\tilde{y}_1$ is a linear transformation, through operator K, of a realization $\tilde{x}$ of μ. In a specular way, $\tilde{y}_2$ is a realization of a centered gaussian measure with covariance operator Σ_n. Therefore, $\tilde{y}_1$ and $\tilde{y}_2$ are independent and for all $(\varphi, \psi) \in \mathcal{X} \times \mathcal{Y}$

$$\begin{aligned} < (\tilde{x}, \tilde{y}), (\varphi, \psi) > &= < \tilde{x}, \varphi > + < \tilde{y}_1 + \tilde{y}_2, \psi > \\ &= < \tilde{x}, \varphi > + < K\tilde{x}, \psi > + < \tilde{y}_2, \psi > \\ &= < \tilde{x}, \varphi + K^*\psi > + < \tilde{y}_2, \psi > \\ &= \mathcal{N}(< x_0, \varphi + K^*\psi >, < \Omega_0(\varphi + K^*\psi), (\varphi + K^*\psi) >) + \mathcal{N}(0, < \Sigma_n\psi, \psi >) \\ &= \mathcal{N}(< x_0, \varphi + K^*\psi >, < \Omega_0(\varphi + K^*\psi), (\varphi + K^*\psi) > + < \Sigma_n\psi, \psi >). \end{aligned}$$

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 $m_{xy} = (x_0, Kx_0)$. In the same way, the covariance operator Υ is defined by

$$\begin{aligned} < \Upsilon(\varphi, \psi), (\varphi, \psi) > &= Var(< (\tilde{x}, \tilde{y}), (\varphi, \psi) >) \\ &= < \Omega_0\varphi, \varphi > + < (\Sigma_n + K\Omega_0K^*)\psi, \psi > + < K\Omega_0\varphi, \psi > + < \Omega_0K^*\psi, \varphi > \\ &= < (\Omega_0\varphi + \Omega_0K^*\psi, (\Sigma_n + K\Omega_0K^*)\psi + K\Omega_0\varphi), (\varphi, \psi) > \end{aligned}$$

that concludes the proof.

8.3 Proof of Theorem 2

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.

8.4 Proof of Theorem 3

We begin by proving that $(\Sigma_n + K\Omega_0 K^*)$ is trace-class. Note that $tr(\Sigma_n + K\Omega_0 K^*) = tr\Sigma_n + tr(K\Omega_0 K^*)$ by linearity of the trace operator. 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 Hilbert-Schmidt operator[6].
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 $g(\cdot)$ and $f(\cdot)$ two measures on $\mathbb{R}$, then

$$
\Omega_0^{\frac{1}{2}} K^* = \int_{\mathbb{R}\times\mathbb{R}} a(z,t)b(s,t)g(t)f(s)dsdt. \tag{34}
$$

We prove that $\Omega_0^{\frac{1}{2}} K^*$ is Hilbert-Schmidt:

$$
\begin{aligned}
\int_{\mathbb{R}\times\mathbb{R}} \Big| \int_{\mathbb{R}} a(z,t)b(s,t)g(t)dt \Big|^2 f(s)h(z)dsdz &\leq \int_{\mathbb{R}\times\mathbb{R}} \Big(\int_{\mathbb{R}} |a(z,t)b(s,t)|g(t)dt \Big)^2 f(s)h(z)dsdz \\
&\leq \int_{\mathbb{R}\times\mathbb{R}} \Big(\Big(\int_{\mathbb{R}} a^2(z,t)g(t)dt \Big)^{\frac{1}{2}} \Big(\int_{\mathbb{R}} b^2(s,t)g(t)dt \Big)^{\frac{1}{2}} \Big)^2 f(s)h(z)dsdz \\
&= \int_{\mathbb{R}\times\mathbb{R}} \int_{\mathbb{R}} a^2(z,t)g(t)dt \int_{\mathbb{R}} b^2(s,t)g(t)dt f(s)h(z)dsdz \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} a^2(z,t)g(t)h(z)dtdz \int_{\mathbb{R}} \int_{\mathbb{R}} b^2(s,t)g(t)f(s)dsdt \\
&< \infty
\end{aligned}
$$

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 the result follows.
Let now consider Υ:

$$
\Upsilon = \begin{bmatrix} \Omega_0 & \Omega_0 K^* \\ K\Omega_0 & \Sigma_n + K\Omega_0 K^* \end{bmatrix}.
$$

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

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

[6]The equivalence is due to the fact that

$$
tr(K\Omega_0 K^*) =< \Omega_0^{\frac{1}{2}} K^*, \Omega_0^{\frac{1}{2}} K^* >_{HS} = ||\Omega_0^{\frac{1}{2}} K^*||^2_{HS}.
$$

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:

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

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

8.5 Proof of Theorem 6

$$\begin{aligned}||\hat{x}_\alpha - x_\alpha||^2 &= \Omega_0\hat{K}^*(\alpha_n I + \Sigma_n + \hat{K}\Omega_0\hat{K}^*)^{-1}(\hat{K}x + (K-\hat{K})x + U) \\ &\quad -\Omega_0\hat{K}^*(\alpha_n I + \Sigma_n + \hat{K}\Omega_0\hat{K}^*)^{-1}\hat{K}x_0 \\ &\quad \Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}(Kx+U) + \Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}Kx_0 \\ &= \underbrace{-\Omega_0(K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}K - \hat{K}^*(\alpha_n I + \Sigma_n + \hat{K}\Omega_0\hat{K}^*)^{-1}\hat{K})(x-x_0)}_{I} \\ &\quad \underbrace{-\Omega_0 K^*(\alpha_n I + \Sigma_n + K\Omega_0 K^*)^{-1}U + \Omega_0\hat{K}^*(\alpha_n I + \Sigma_n + \hat{K}\Omega_0\hat{K}^*)^{-1}U}_{II} \\ &\quad + \underbrace{\Omega_0\hat{K}^*(\alpha_n I + \Sigma_n + \hat{K}\Omega_0\hat{K}^*)^{-1}(K-\hat{K})x}_{III}.\end{aligned}$$

Let $\eta = \Omega^{-\frac{1}{2}}(x-x_0)$, $H = K\Omega_{\frac{1}{2}}$ and $H^* = \Omega_{\frac{1}{2}}K^*$, then we can decompose term I as

$$\begin{aligned}I &= \Omega_{\frac{1}{2}}[\hat{H}^*(\alpha_n I + \Sigma_n + \hat{H}\hat{H}^*)^{-1}\hat{H}\eta - H^*(\alpha_n I + \Sigma_n + HH^*)^{-1}H\eta] \\ &= \Omega_{\frac{1}{2}}[\hat{H}^*(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H}\eta - H^*(\alpha_n I + HH^*)^{-1}H\eta] \\ &\quad +\hat{H}^*(\alpha_n I + \Sigma_n + \hat{H}\hat{H}^*)^{-1}\hat{H}\eta - \hat{H}^*(\alpha_n I + \hat{H}\hat{H}^*)^{-1}\hat{H}\eta] \\ &\quad]\end{aligned}$$

8.6 Proof of Theorem 9

(i) Consider Bayes estimation error:

$$\begin{aligned}||\mathbb{E}_\alpha(x|t) - x_*||^2 &= ||\frac{1}{n}B_\alpha(K^*(Kx_* + U)) + c_\alpha - x_*||^2 \\ &= ||\Omega_0\frac{K^*K}{n}[\alpha_n I + \frac{1}{n^2}K^*(I_n\otimes\Sigma_N + K\Omega_0K^*)K]^{-1}\frac{K^*U}{n} - \\ &\quad (I - \Omega_0\frac{K^*K}{n}[\alpha_n I + \frac{1}{n^2}K^*(I_n\otimes\Sigma_N + K\Omega_0K^*)K]^{-1}\frac{K^*K}{n})(x_* - x_0)||^2 \\ &\leq \overbrace{||\Omega_0\frac{K^*K}{n}[\alpha_n I + \frac{1}{n^2}K^*(I_n\otimes\Sigma_N + K\Omega_0K^*)K]^{-1}\frac{K^*U}{n}||^2}^{I} \\ &\quad + \underbrace{||(I - \Omega_0\frac{K^*K}{n}[\alpha_n I + \frac{1}{n^2}K^*(I_n\otimes\Sigma_N + K\Omega_0K^*)K]^{-1}\frac{K^*K}{n})(x_* - x_0)||^2}_{II}.\end{aligned}$$

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

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

From the distribution of U_i we can infer $K_i^*U_i \sim i\,\mathcal{GP}(0, K_i^*\Sigma_N K_i)$ 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 $K_i^*\Sigma_N K_i$, then, since $||K_i^*U_i||^2 = \sum_{j=1}^\infty < K_i^*U_i, \tilde{\varphi}_{ij} >^2$, we have

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

that goes to zero with N. Moreover, we assume that $\frac{1}{n^2}K^*Var(Y)K \to Q$ with n, where Q is a bounded operator. It follows that $||(\alpha_N I + \frac{1}{n^2}K^*Var(Y)K)^{-1}||^2 = \mathcal{O}_p(\alpha_N^{-2})$ and $||\Omega_0^{\frac{1}{2}}\frac{K^*K}{n}(\alpha_N I + \frac{K^*K\Omega_0 K^*K}{n^2})^{-1}||^2 = \mathcal{O}_p(\alpha_N^{-1})$. Therefore, we get

$$\begin{aligned} I &= \mathcal{O}_p\Big(\frac{1}{\alpha_N^3}||\frac{K^*\Sigma_N K}{n^2}||^2\frac{1}{n^2}tr(K^*\Sigma_N K)\Big) + \mathcal{O}_p\Big(\frac{1}{\alpha_N n^2}tr(K^*\Sigma_N K)\Big) \\ &= \mathcal{O}_p\Big(\frac{1}{\alpha_N^3 n^3}\Big) + \mathcal{O}_p\Big(\frac{1}{\alpha_N n}\Big), \end{aligned}$$

if $\frac{1}{n^2}tr(K^*\Sigma_n K) \sim \mathcal{O}_p(\frac{1}{N})$ and $||\frac{K^*\Sigma_N K}{n^2}|| \sim \mathcal{O}_p(\frac{1}{N})$.

To analyze term II, it is advisable to rewrite it in the following way:

$$II = ||(I - \Omega_0\frac{1}{n}K^*K(\alpha_N I + \frac{1}{n^2}K^*K\Omega_0 K^*K)^{-1}\frac{1}{n}K^*K)(x_* - x_0)$$

$$
\begin{aligned}
& -\Omega_0 \frac{1}{n} K^* K (\alpha_N I + \frac{1}{n^2} K^* Var(Y) K)^{-1} \frac{1}{n} K^* K (x_* - x_0) \\
& + \Omega_0 \frac{1}{n} K^* K (\alpha_N I + \frac{1}{n^2} K^* K \Omega_0 K^* K)^{-1} \frac{1}{n} K^* K)(x_* - x_0)||^2 \\
\leq & \overbrace{||(I - \Omega_0 \frac{1}{n} K^* K (\alpha_N I + \frac{1}{n^2} K^* K \Omega_0 K^* K)^{-1} \frac{1}{n} K^* K)(x_* - x_0)||^2}^{III} \\
& + \underbrace{||\Omega_0 \frac{K^* K}{n} (\alpha_N I + \frac{K^* Var(Y) K}{n^2})^{-1} (\frac{K^* \Sigma_N K}{n^2}) (\alpha_N I + \frac{K^* K \Omega_0 K^* K}{n^2})^{-1} \frac{K^* K}{n} (x_* - x_0)||^2}_{IV}.
\end{aligned}
$$

For term IV we proceed exactly as for term II in (18) and we get $IV = \mathcal{O}_p(||\frac{K^* \Sigma_N K}{n^2}||^2 \frac{1}{\alpha_N^3})$ under the assumption that $(x_* - x_0) \in \mathcal{H}(\Omega_0)$. Moreover, if we assume that $||\frac{K^* \Sigma_N K}{n^2}||$ is of order $\frac{1}{N}$, we get that $IV = \mathcal{O}_p(\frac{1}{\alpha_N^3 N^2})$.
In order to carry out asymptotic analysis of term III, we rewrite it as

$$
\begin{aligned}
III &= ||(\Omega_0^{\frac{1}{2}} - \Omega_0 \frac{1}{n} K^* K (\alpha_N I + \frac{1}{n^2} K^* K \Omega_0 K^* K)^{-1} \frac{1}{n} K^* K \Omega_0^{\frac{1}{2}}) \Omega_0^{-\frac{1}{2}} (x_* - x_0)||^2 \\
&= ||\Omega_0^{\frac{1}{2}} (I - \Omega_0^{\frac{1}{2}} \frac{1}{n} K^* K (\alpha_N I + \frac{1}{n^2} K^* K \Omega_0 K^* K)^{-1} \frac{1}{n} K^* K \Omega_0^{\frac{1}{2}}) \Omega_0^{-\frac{1}{2}} (x_* - x_0)||^2,
\end{aligned}
$$

that is well defined if $(x_* - x_0) \in \mathcal{R.K.H.S.}(\Omega_0)$. Let $\lambda = \Omega_0^{-\frac{1}{2}} (x_* - x_0)$,

$$
(I - \Omega_0^{\frac{1}{2}} \frac{1}{n} K^* K (\alpha_N I + \frac{1}{n^2} K^* K \Omega_0 K^* K)^{-1} \frac{1}{n} K^* K \Omega_0^{\frac{1}{2}}) \lambda
$$

is of the same order as the bias of regularization of the solution of $\frac{1}{n} K^* K \Omega_0^{\frac{1}{2}} \lambda = r$ since

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

has the same eigenvalues of

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

If we add the assumption that $\lambda \in \mathcal{R}(\frac{1}{n^2} \Omega_0^{\frac{1}{2}} K^* K K^* K \Omega_0^{\frac{1}{2}})^{\frac{\beta}{2}}$, term III converges to zero at the speed of α_N^β, where β denotes the regularity of function λ.

(ii) Trivial.

References

[1] Andersen, P.K., Borgan, O., Gill, R.D. and N. Keiding (1993), *Statistical Models Baed on Counting Processes*, Springer-Verlag

[2] Baker, C.R. (1973), *Joint Measures and Cross-Covariance Operators*, Transactions of the American Mathematical Society, **186**, 273-289.

[3] Belitser, E., and S., Ghosal (2003), *Adaptive Bayesian Inference on the Mean of an Infinite Dimensional Normal Distribution*, Annals of Statistics, **31**, 536-559.

[4] Carrasco, M., Florens, J.P., and E., Renault (2005), *Estimation based on Spectral Decomposition and Regularization*, forthcoming in Handbook of Econometrics, J.J. Heckman and E. Leamer, eds., **6**, Elsevier, North Holland.

[5] Cavalier, L., Golubev, G.K., Picard, D., and A.B., Tsybakov (2002), *Oracle Inequalities for Inverse Problems*, Annals of Statistics, **30**, 843-874.

[6] Darolles, S., Florens, J.P., and E., Renault (2006), *Nonparametric Instrumental Regression*, Working paper.

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

[8] Engl, H.W., Hanke, M. and A., Neubauer (2000), *Regularization of Inverse Problems*, Kluwer Academic, Dordrecht.

[9] 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.

[10] Dykstra, R.L. and P.W., Laud (1981), *A Bayesian Nonparametric Approach to Reliability*, The Annals of Statistics, **9**, 356-367.

[11] Ferguson, T.S. (1974), *Prior Distributions on Spaces of Probability Measures*, The Annals of Statistics, Vol.2, **4**, 615-629.

[12] Ferguson, T.S. and E.G. Phadia (1979), *Bayesian Nonparametric Estimation Based on Censored Data*, Annals of Statistics, **7**, 163-186.

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

[14] 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.

[15] Florens, J.P., and A., Simoni (2007), *Nonparametric Estimation of Instrumental Regression: a Bayesian Approach Based on Regularized Posterior*, mimeo.

[16] 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.

[17] Hiroshi, S. and O., Yoshiaki (1975), *Separabilities of a Gaussian Measure*, Annales de l'I.H.P., section B, tome 11, **3**, 287 - 298.

[18] 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.

[19] Hjort, N.L. (1996), *Bayesian Aprroaches to Non- and Semiparametric Density Estimation*, Bayesian Statistics 5 (J.M. Bernardo et *al.*, eds.), 223-253.

[20] 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.

[21] Kaipio, J., and E., Somersalo (2004), *Statistical and Computational Inverse Problems*, Applied Mathematical Series, vol.160, Springer, Berlin.

[22] Kress, R. (1999), *Linear Integral Equation*, Springer.

[23] Lvine, M. (1992) *Some Aspects of Polya Tree Distributions for Statistical Modelling*, The Annals of Statistics, Vol.20, **3**, 1222-1235.

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

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

[26] Neveu, J. (1975), *Discrete Parameter Martingales*, Amsterdam: North-Holland.

[27] Petrone S. (1999) *Bayesian Density Estimation Using Bernstein Polynomials*, The Canadian Journal of Statistics, Vol.27, **1**, 15-126.

[28] Rasmussen, C.E. and C.K.I., Williams (2006), *Gaussian Processes for Machine Learning*, The MIT Press.

[29] Ruggiero, M. (1994), *Bayesian Semiparametric Estimation of Proportional Hazards Models*, Journal of Econometrics, **62**, 277 - 300.

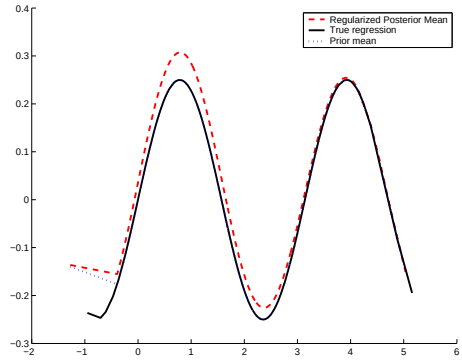
[30] Smith, M. and R., Kohn (1996), *Nonparametric Regression Using Bayesian Variable Selection*, Journal of Econometrics, **75**, 317-343.

[31] Van der Vaart, A.W., and J.H., Van Zanten (2000), *Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors*, Working paper.

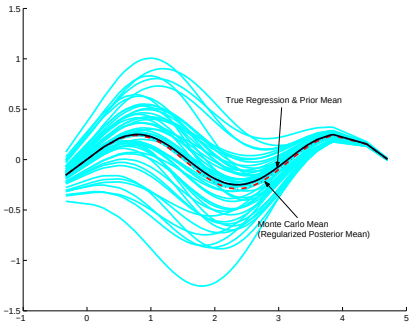
[32] Vapnik, V.N. (1998), *Statistical Learning Theory*, John Wiley & Sons, Inc.

[33] 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.

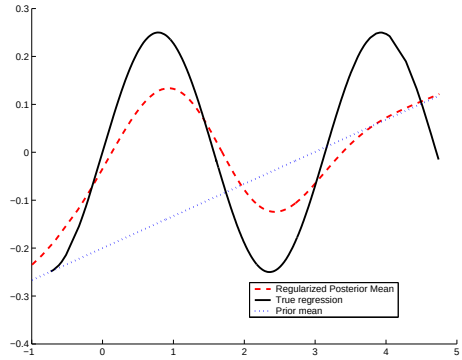
[34] Zhao, L.H. (2000), *Bayesian Aspects of Some Nonparametric Problems*, Annals of Statistics, **28**, 532-552.



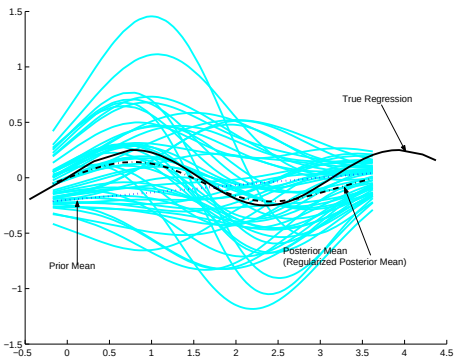
(a) $m_0(w) = cos(w)sin(w)$,
$\Omega_0\varphi(w_1) = 2\int \exp(-(w_1 - w_2)^2)\varphi(w_2)f(w_2)dw_2$



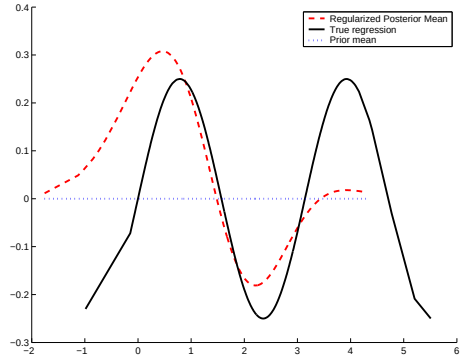
(b) $m_0(w) = cos(w)sin(w)$,
$\Omega_0\varphi(w_1) = 2\int \exp(-(w_1 - w_2)^2)\varphi(w_2)f(w_2)dw_2$



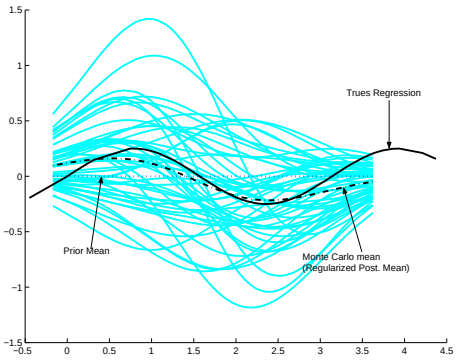
(c) $m_0(w) = 0.067w - 0.2$,
$\Omega_0\varphi(w_1) = 10\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$



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

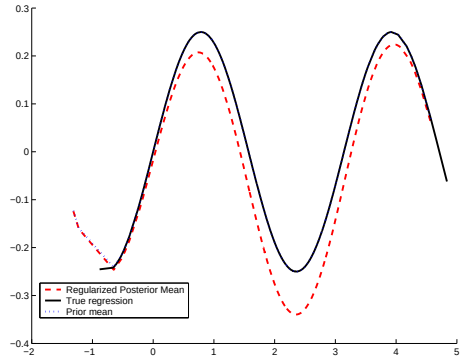


(e) $m_0(w) = 0$,
$\Omega_0\varphi(w_1) = 10\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$

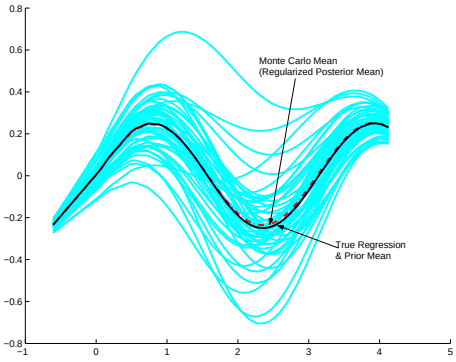


(f) $m_0(w) = 0$,
$\Omega_0\varphi(w_1) = 10\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$

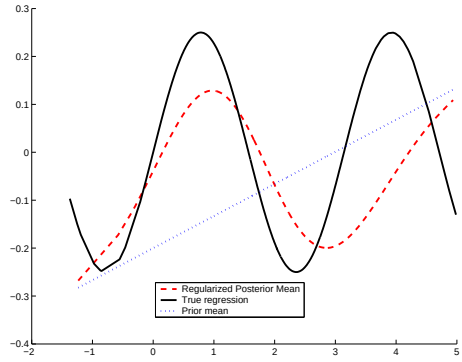
Figure 7: Panels (7a), (7c) and (7e): estimation for different prior means. Panels (7b), (7d) and (7f): Monte Carlo Experiment with $N = 100$, $\alpha = 0.05$, 50 iterations.



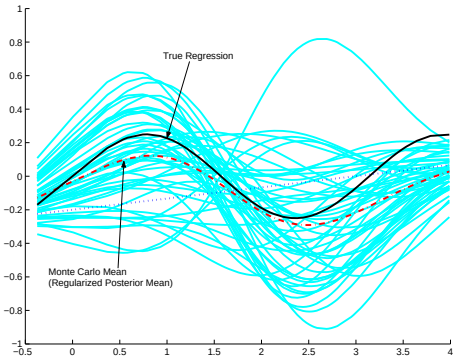
(a) $m_0(w) = cos(w)sin(w)$,
$\Omega_0\varphi(w_1) = 2\int \exp(-(w_1 - w_2)^2)\varphi(w_2)f(w_2)dw_2$



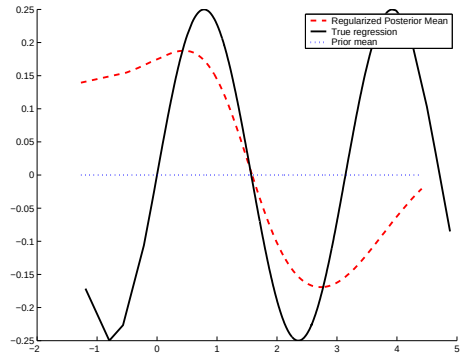
(b) $m_0(w) = cos(w)sin(w)$,
$\Omega_0\varphi(w_1) = 2\int \exp(-(w_1 - w_2)^2)\varphi(w_2)f(w_2)dw_2$



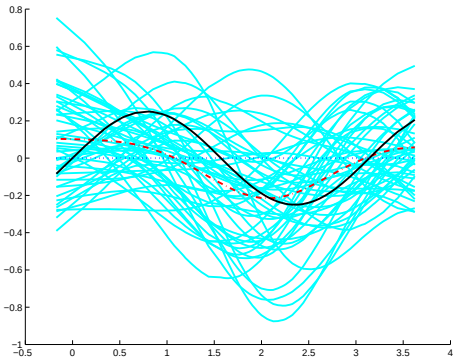
(c) $m_0(w) = 0.067w - 0.2$,
$\Omega_0\varphi(w_1) = 20\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$



(d) $m_0(w) = 0.067w - 0.2$,
$\Omega_0\varphi(w_1) = 20\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$



(e) $m_0(w) = 0$,
$\Omega_0\varphi(w_1) = 20\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$



(f) $m_0(w) = 0$,
$\Omega_0\varphi(w_1) = 20\int \exp(-(w_1-w_2)^2)\varphi(w_2)f(w_2)dw_2$

Figure 8: Panels (8a), (8c) and (8e): estimation for different prior means. Panels (8b), (8d) and (8f): Monte Carlo Experiment with $N = 100$, $\alpha = 0.05$, 50 iterations.