Econometric Theory, **32**, 2016, 71–121.
doi:10.1017/S0266466614000796

REGULARIZING PRIORS FOR LINEAR INVERSE PROBLEMS

JEAN-PIERRE FLORENS
Toulouse School of Economics

ANNA SIMONI
CNRS – GRECSTA and CREST

This paper proposes a new Bayesian approach for estimating, nonparametrically, functional parameters in econometric models that are characterized as the solution of a linear inverse problem. By using a Gaussian process prior we propose the posterior mean as an estimator and prove frequentist consistency of the posterior distribution. The latter provides the frequentist validation of our Bayesian procedure. We show that the minimax rate of contraction of the posterior distribution can be obtained provided that either the regularity of the prior matches the regularity of the true parameter or the prior is scaled at an appropriate rate. The scaling parameter of the prior distribution plays the role of a regularization parameter. We propose a new data-driven method for optimally selecting in practice this regularization parameter. We also provide sufficient conditions such that the posterior mean, in a conjugate-Gaussian setting, is equal to a Tikhonov-type estimator in a frequentist setting. Under these conditions our data-driven method is valid for selecting the regularization parameter of the Tikhonov estimator as well. Finally, we apply our general methodology to two leading examples in econometrics: instrumental regression and functional regression estimation.

1. INTRODUCTION

In the last decade, econometric theory has shown an increasing interest in the theory of *stochastic inverse problems* (IPs) as a fundamental tool for functional estimation of structural as well as reduced form models. This paper develops an encompassing Bayesian approach for nonparametric estimation of econometric models based on stochastic linear IPs.

A version of this paper circulated under the title *On the Regularization Power of the Prior Distribution in Linear Ill-Posed Inverse Problems* (September 2008) (see also Florens and Simoni, 2010). We thank the co-editor Oliver Linton and three anonymous referees whose comments greatly improved the paper. We also thank Sergios Agapiou, Élise Coudin, Joel Horowitz, Yuan Liao, Enno Mammen, Peter Phillips, Andrew Stuart, and Sébastien Van Bellegem for interesting discussions. We acknowledge helpful comments from participants to workshops and seminars in Banff (2009), Bressanone (2009), Stats in the Château (Paris 2009), Toulouse (2009), CIRM-Marseille (2009), Brown (2010), Mannheim (2010), Northwestern (2010), World Congress of the Econometric Society (Shanghai 2010), Warwick (2011), Cowles Foundation Econometric Summer conference (Yale 2011), and Princeton (2012). We thank financial support from ANR-13-BSH1-0004 (IPANEMA). Anna Simoni gratefully acknowledges financial support from the University of Mannheim through the DFG-SNF Research Group FOR916 and hospitality from University of Mannheim and Boston College. Address correspondence to Anna Simoni, CREST, 15 Boulevard Gabriel Péri, 92240 Malakoff, France; e-mail: simoni.anna@gmail.com.

© Cambridge University Press 2014

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

We construct a Gaussian process prior for the (functional) parameter of interest and establish sufficient conditions for frequentist consistency of the corresponding posterior mean estimator. We prove that under these conditions the posterior mean estimator numerically equals a Tikhonov-type estimator in the frequentist setting. We propose a novel data-driven method for selecting the regularization parameter which enters as hyperparameter of the prior. The value selected by our data-driven method is shown to be optimal in a minimax sense if the prior distribution is sufficiently smooth. Due to the equivalence between Bayes- and Tikhonov-type estimator, our data-driven method has broad applicability and allows to select the regularization parameter necessary for implementing Tikhonov-type estimators.

Stochastic linear IP theory has been used to construct new estimation methods in many subfields of econometrics. Just to mention some of them, it has been shown to be fundamental in nonparametric estimation of an instrumental regression function, see e.g., Florens (2003), Newey and Powell (2003), Hall and Horowitz (2005), Darolles, Fan, Florens, and Renault (2011), and Florens and Simoni (2012a). It has also been used in semiparametric estimation under moment restrictions (see e.g., Carrasco and Florens, 2000; Ai and Chen, 2003; Chen and Pouzo, 2012). In addition, it has been exploited for inference: in econometric models with heterogeneity—e.g., Hoderlein, Nesheim, and Simoni (2013), in stochastic volatility models—e.g., Linton and Mammen (2005), and for frontier estimation for productivity analysis—e.g., Daouia, Florens, and Simar (2009). Finally, IP theory has been used for functional regression estimation (see e.g., Hall and Horowitz, 2007). We refer to Carrasco, Florens, and Renault (2005) and references therein for a general overview of IPs in econometrics.

The general framework considered in this paper—which accommodates many functional estimation problems in econometrics just as mentioned above—is the following. Let $\mathcal{X}$ and $\mathcal{Y}$ be infinite dimensional separable Hilbert spaces over $\mathbb{R}$ and denote by $x \in \mathcal{X}$ the functional parameter that we want to estimate. For instance, x can be an Engel curve or the probability density function of the unobserved heterogeneity. The estimating equation characterizes x as the solution of the functional equation

$$y^\delta = Kx + U^\delta, \qquad x \in \mathcal{X}, \quad y^\delta \in \mathcal{Y}, \quad \delta > 0 \tag{1}$$

where y^δ is an observable function, $K : \mathcal{X} \to \mathcal{Y}$ is a known, bounded, linear operator, and U^δ is a (functional) error term with values in $\mathcal{Y}$ and covariance operator proportional to a positive scalar δ. Model (1) is the standard formulation of a linear IP. It is motivated by many applications across fields, e.g., econometrics, tomography, signal and image processing and has received a great deal of attention in the mathematical and statistical literature. We refer to Engl, Hanke, and Neubauer (2000), Cavalier and Tsybakov (2002), and Kaipio and Somersalo (2004) for an illustration of possible applications. A large class of econometric models writes under the form of equation (1)—like consumption based asset pricing models, density estimation of the heterogeneity parameter in structural models, deconvolution in structural models with measurement errors. In econometrics, y^δ is

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

usually a transformation of an n-sample of finite dimensional objects and the parameter $\delta^{-1} > 0$ represents the "level of information" of the sample, so that $\delta \to 0$ as $n \to \infty$. For instance, equation (1) may arise from a set of moment equations with y^δ an empirical conditional moment and $\delta = \frac{1}{n}$. For $\delta = 0$ (i.e., perfect information) we have $y^0 = Kx$ and $U^0 = 0$. We illustrate two leading examples in econometrics that can be estimated with our Bayesian method.

Example 1 (Instrumental Variable (IV) Regression Estimation)
Let (Y, Z, W) be an observable real random vector and $x(Z)$ be the IV regression defined through the moment condition $\mathbf{E}(Y|W) = \mathbf{E}(x|W)$. Suppose that the distribution of (Z, W) is confined to the unit square $[0,1]^2$, admits a density f_{ZW} with respect to the Lebesgue measure, and $x \in \mathcal{X} \equiv L^2[0,1]$. The moment restriction implies that x is a solution to

$$\mathbf{E}_W\left[\mathbf{E}(Y|w)a(w,v)\right](v) = \int_0^1 \int_0^1 x(z)a(w,v)f_{ZW}(z,w)dwdz,$$

where $a(w,v) \in L^2[0,1]^2$ is a known function, $\mathbf{E}_W$ denotes the expectation with respect to the marginal density f_W of W, and $L^2[0,1]^2$ denotes the space of square integrable functions on $[0,1]^2$. This transformation of the original moment condition is appealing because its empirical counterpart is asymptotically Gaussian (as required by our Bayesian approach). Suppose we have independent and identically distributed (i.i.d.) data (Y_i, Z_i, W_i), $i = 1, \ldots, n$, then the true distribution of (Y, Z, W) can be replaced with a nonparametric estimator and we obtain a problem of the form (1) with $\delta = \frac{1}{n}$,

$$y^\delta = \hat{\mathbf{E}}_W\left[\hat{\mathbf{E}}(Y|w)a(w,v)\right] \quad \text{and} \quad Kx = \int_0^1 \int_0^1 x(z)a(w,v)\hat{f}_{ZW}(z,w)dwdz. \tag{2}$$

Example 2 (Functional Linear Regression Estimation)
The model is the following:

$$\xi = \int_0^1 x(s)Z(s)ds + \varepsilon, \quad \mathbf{E}\big(\varepsilon Z(s)\big) = 0, \quad Z, x \in \mathcal{X} = L^2[0,1], \quad \mathbf{E}\big\langle Z, Z\big\rangle < \infty, \tag{3}$$

and $\varepsilon|Z, \tau \sim \mathcal{N}(0, \tau^2)$, with $\langle \cdot, \cdot \rangle$ the inner product in $L^2[0,1]$. We want to recover the functional regression x. Assuming that Z is a centered random function with covariance operator of trace-class, the most popular approach consists in multiplying both sides of the first equation in (3) by Z and then taking the expectation: $\mathbf{E}(\xi Z(t)) = \int_0^1 x(s)Cov(Z(s), Z(t))ds$, for $t \in [0,1]$. If we dispose of i.i.d. data $(\xi_1, Z_1), \ldots, (\xi_n, Z_n)$ we can estimate the unknown moments in the previous equation. Hence, x is a solution of an equation of the form of equation (1) with $y^\delta = \frac{1}{n}\sum_i \xi_i Z_i(t)$, $U^\delta = \frac{1}{n}\sum_i \varepsilon_i Z_i(t)$, $\delta = \frac{1}{n}$ and $\forall \varphi \in L^2[0,1]$, $K\varphi := \frac{1}{n}\sum_i \langle Z_i, \varphi \rangle Z_i(t)$.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

A Bayesian approach to stochastic IPs combines the prior and sampling information and proposes the posterior distribution as a solution. It allows to deal with two important issues. First, it allows to incorporate in the estimation procedure the prior information about the functional parameter provided by the economic theory or the beliefs of experts. This prior information may be particularly valuable in functional estimation since often the data available are concentrated only in a region of the graph of the functional parameter so that some parts of the function cannot be recovered from the data.

Second, since the quality of the estimation of the solution of the IP (1) relies on the value of a *regularization parameter*, it is particularly important to choose such a parameter in an accurate way. This can be done with our Bayesian approach.

Previous literature on Bayesian IP. The majority of the Bayesian approaches to stochastic IPs proposed so far consider a finite dimensional projection of (1) and recover x only on a finite grid of points (see e.g., Kaipio and Somersalo, 2004, Chap. 5; Hofinger and Pikkarainen, 2007, 2009; Helin, 2009). So, they cannot be used for functional estimation of econometric models that consider functional observations and parameters. Knapik, Van der Vaart, and Van Zanten (2011) study Bayesian IPs with functional observations and parameter. However, their framework does not allow to accommodate the econometric models of interest because of the different definition of the error term U^δ. Indeed, the analysis of Knapik et al. (2011) works under the assumption that U^δ is an isonormal Gaussian process, which implies that U^δ, and by consequence y^δ, is not realizable as a random element in $\mathcal{Y}$. This assumption greatly simplifies the analysis but unfortunately does not hold, in general, in econometric models because real (functional) data cannot be generated by an isonormal Gaussian process. On the contrary, our paper works in the more realistic situation where U^δ, and therefore y^δ, are random elements with realizations in $\mathcal{Y}$, which is the relevant case in econometrics.

For the special case where model (1) results from conditional moment restrictions, Liao and Jiang (2011) propose a quasi-Bayesian procedure. Their approach, which works also in the nonlinear case, is based on limited-information-likelihood and a sieve approximation technique. It is essentially different from our approach since we work with Gaussian process priors and do not use a finite-dimensional approximation.

Working with Gaussian process priors is computationally convenient since in many cases the sampling distribution is (asymptotically) Gaussian and thus the posterior is also Gaussian (conjugate-Gaussian setting). The current paper gives sufficient conditions under which the posterior mean of x, in a conjugate-Gaussian setting, exists in a closed-form and thus can be used to estimate x. Existence of such a closed-form is not verified in general, as we explain in the next paragraph and after Theorem 1. Agapiou, Larsson, and Stuart (2013) propose an alternative approach to deal with this problem. Florens and Simoni (2012b) overcome this problem by constructing a *regularized posterior distribution* which works well in practice but the regularization of the posterior distribution is *ad hoc* and cannot

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

be justified by any prior-to-posterior transformation. In comparison to Florens and Simoni (2012b) and Agapiou et al. (2013), the current paper also provides an adaptive method for choosing the regularization parameter as we explain in the next section.

This paper is also related to the literature on Bayesian adaptation where many important contributions have been made for models different than ours. Some examples of papers on adaptive Bayesian nonparametric estimation include Belitser and Ghosal (2003), Van der Vaart and Van Zanten (2009), Johannes, Schenk, and Simoni (2014) and references therein.

Our contribution. Our estimation procedure is based on a conjugate-Gaussian setting which is suggested by the linearity of problem (1). On one hand, such a setting is appealing because the corresponding posterior distribution can be computed analytically without using any Markov Chain Monte Carlo algorithm which, even if very powerful, slows down the estimate computation. On the other hand, a conjugate-Gaussian Bayesian IP has the drawback that the posterior mean is, in general, not defined as a linear estimator but as a *measurable linear transformation* (*mlt*, hereafter) which is a weaker notion (see Mandelbaum, 1984). In particular, there is no explicit form for the *mlt* estimator and so it is unclear how the posterior mean estimator of x can be computed in practice. Moreover, whether consistency of the *mlt* estimator holds or not is still an open question.

The first contribution of our paper is to provide a sufficient condition under which the posterior mean, in a conjugate-Gaussian setting, is defined as a linear estimator and has a closed-form so that it can be easily computed and implemented. We assume a Gaussian process prior distribution for x, with mean function $x_0 \in \mathcal{X}$ and covariance operator $\Omega_0 : \mathcal{X} \to \mathcal{X}$. In the case where $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional and (x, y^δ) are jointly Gaussian, the posterior mean of x is the linear estimator $[x_0 + \Omega_0 K^* Var(y^\delta)^{-1}(y^\delta - Kx_0)]$ provided $Var(y^\delta)$ is invertible, where $Var(y^\delta)$ denotes the marginal covariance operator of y^δ and K^* denotes the adjoint of K (see Section 3.1). Unfortunately, when the dimension of $\mathcal{X}$ and $\mathcal{Y}$ is infinite, the linear operator $\Omega_0 K^* Var(y^\delta)^{-1}$ is no longer defined on $\mathcal{Y}$ but only on a dense subspace of $\mathcal{Y}$ of measure zero and is typically noncontinuous (i.e., unbounded). This paper: gives a sufficient condition that guarantees that $\Omega_0 K^* Var(y^\delta)^{-1}$ is defined and continuous on the whole $\mathcal{Y}$, shows that this condition is in general satisfied in many econometric models, derives a closed form for the posterior mean of x, and proves that it is a continuous and linear (thus consistent) estimator defined on $\mathcal{Y}$. Under this condition, the prior-to-posterior transformation can be interpreted as a regularization scheme so that no *ad hoc* regularization scheme is required as, e.g., in Florens and Simoni (2012b).

Our second contribution consists in the study of frequentist asymptotic properties of our Bayesian procedure. For that, we admit the existence of a true x, say x_*, that generates the data. We establish that the posterior mean estimator and posterior distribution of the solution x of the IP have good frequentist asymptotic properties for $\delta \to 0$. More precisely, we show that the posterior distribution

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

converges toward a Dirac mass at x_* in probability with respect to the sampling distribution. This is known as *frequentist posterior consistency* (see e.g., Le Cam, 1986; Diaconis and Freedman, 1986) and provides the frequentist validation of our Bayesian procedure.

We also recover the rate of contraction of the risk associated with the posterior mean and of the posterior distribution. This rate depends on the smoothness and scale of the prior as well as on the smoothness of x_*. Depending on the specification of the prior this rate may be *minimax*. In particular, (i) when the regularity of the prior matches the regularity of x_*, the minimax rate of convergence is obtained with a fixed prior covariance; (ii) when the prior is rougher or smoother at any degree than x_*, the minimax rate can still be obtained if the prior is scaled at an appropriate rate depending on the unknown regularity of x_*.

The scaling of the prior is tuned by a hyperparameter α which plays the role of *regularization parameter*. Our third contribution consists in proposing an adaptive data-driven method for optimally selecting α. Our method is based on an Empirical Bayes (EB) approach. Because the posterior mean is, under our assumptions, equal to a Tikhonov-type estimator for problem (1), our EB approach for selecting the regularization parameter is valid, and can be used, also for computing frequentist estimators based on Tikhonov regularization.[1] Finally, the EB-selected regularization parameter is plugged into the prior distribution of x and for the corresponding EB-posterior distribution we prove frequentist posterior consistency.

The paper is organized as follows. In Section 2 we give an overview of our estimator in the framework of the IV regression (Example 1) in order to explain the key idea of our procedure. In Section 3 we formalize our approach by setting the Bayesian experiment associated with (1) and the main assumptions. In Section 4 the posterior distribution of x is computed and its frequentist asymptotic properties are analyzed. Section 5 focuses on the mildly ill-posed case while the severely ill-posed case is treated in the Supplementary Appendix. The EB method is developed in Section 6. Examples 1 and 2 are analyzed in Section 7. Section 8 shows numerical implementations and Section 9 concludes. All the proofs are in the Appendix and in the Supplementary Appendix.

2. IV REGRESSION ESTIMATION: AN OVERVIEW OF OUR ESTIMATOR

In this section we describe our estimator for the IV regression of Example 1. The detailed probabilistic setting is described in Section 3 and the asymptotic frequentist justification of the estimator is provided in Sections 4 and 5. The reader interested in the choice of the regularization parameter can skip Sections 3–5 and go to Section 6. The IV regression x is characterized by the integral equation (1) with y^δ and K defined as in (2), that is

$$\hat{\mathbf{E}}_W\left[\hat{\mathbf{E}}(Y|w)a(w,v)\right] = \int_0^1 \int_0^1 x(z)a(w,v)\hat{f}_{ZW}(z,w)dwdz + U^\delta,$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

where $a(w,v)$ is some known function chosen by the econometrician such that $\int_0^1\int_0^1 a^2(w,v)dwdv < \infty$ and $U^\delta := y^\delta - Kx$. Denote by f_{YW}, f_{ZW}, and f_W the densities with respect to the Lebesgue measure of the distributions of (Y,W), (Z,W), and W. Under mild assumptions on $a(w,v)$, $U^\delta \Rightarrow \mathcal{N}(0,\delta\Sigma)$, where '$\Rightarrow$' means weak convergence in distribution, $\mathcal{N}$ denotes a Gaussian process, and $\Sigma : L^2[0,1] \to L^2[0,1]$ is the operator $\Sigma\varphi = \sigma^2 \int_0^1\int_0^1 a(w,v_1)a(w,v_2)f_W(w)$ $dw\varphi(v_2)f_W(v_2)dv_2$, $\forall\varphi \in L^2[0,1]$ (under the assumption $\mathbf{E}[(Y-x(Z))^2|W] = \sigma^2 < \infty$).

We specify a prior distribution of the type $x \sim \mathcal{N}(0, \frac{1}{n\alpha}\Omega_0)$ where $\alpha > 0$, $\alpha \to 0$ as $n \to \infty$ and Ω_0 is a covariance operator. In practice, examples of how to specify Ω_0 are: (1) $\forall\varphi \in L^2[0,1]$, $(\Omega_0\varphi)(z_2) = b_1\mathbf{E}\left(e^{\{-b_2(Z-z_2)^2\}}\varphi(z_2)\right)$ for some $b_1 > 0$ and $b_2 > 0$ and where the expectation is taken with respect to Z; (2) $\forall\varphi \in L^2[0,1]$, $(\Omega_0\varphi)(z_2) = \sum_{j=1}^\infty l_j\langle\varphi, e_j\rangle e_j(z_2)$ for a decreasing sequence of strictly positive weights $\{l_j\}_{j\geq 1}$ that are summable and for an orthonormal basis $\{e_j\}$ in $L^2[0,1]$ with respect to the distribution of Z.

The estimator for the IV regression $x(\cdot)$ that we propose is the posterior mean

$$\hat{x}_\alpha(\cdot) = \left(\left(\alpha\Omega_0^{-1} + \mathbf{E}\Big(\mathbf{E}(\cdot|W)|Z\Big)\right)^{-1}\mathbf{E}\Big(\hat{\mathbf{E}}(Y|W)|Z\Big)\right)(\cdot) \tag{4}$$

which corresponds to the Bayes estimator under a quadratic loss function. Implementation of $\hat{x}_\alpha$ can be done by: (I) either discretizing the operators in (4), so that Ω_0 and $\mathbf{E}(\mathbf{E}(\cdot|W)|Z)$ become matrices, or (II) computing the singular system $\{\lambda_j, e_{Z,j}, e_{W,j}\}$ of the (estimated) operator $\mathbf{E}(\Omega_0\cdot|W)$. In the second case $\hat{x}_\alpha$ writes:

$$\hat{x}_\alpha(\cdot) = \sum_{j\geq 1}\frac{\lambda_j}{\alpha+\lambda_j^2}\Big\langle\hat{\mathbf{E}}(Y|W), e_{W,j}\Big\rangle\Big(\Omega_0^{1/2}e_{Z,j}\Big)(\cdot), \tag{5}$$

where $\hat{\mathbf{E}}(Y|W)$ is any consistent nonparametric estimator of $\mathbf{E}(Y|W)$ and the infinite sum has to be truncated. We show that this estimator is consistent in a frequentist sense and attains the minimax optimal convergence rate. The parameter α can be data dependent. In practice, it can be "optimally" chosen by using the procedure described in Section 6. A more detailed and involved analysis of the IV regression estimation is presented in Section 7.1.

3. THE MODEL

Let $\mathcal{X}$ and $\mathcal{Y}$ be infinite dimensional separable Hilbert spaces over $\mathbb{R}$ with norm $||\cdot||$ induced by the inner product $\langle\cdot,\cdot\rangle$. Let $\mathfrak{B}(\mathcal{X})$ and $\mathfrak{B}(\mathcal{Y})$ be the Borel σ-fields generated by the open sets of $\mathcal{X}$ and $\mathcal{Y}$, respectively. We consider the IP of estimating the function $x \in \mathcal{X}$ which is linked to the data y^δ through the linear relation

$$y^\delta = Kx + U^\delta, \qquad x \in \mathcal{X}, \quad y^\delta \in \mathcal{Y}, \quad \delta > 0 \tag{6}$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

where y^δ is an observable function and $K : \mathcal{X} \to \mathcal{Y}$ is a known, bounded, linear operator (we refer to Carrasco et al., 2005 for definitions and terminology from functional analysis). The elements y^δ and U^δ are Hilbert space-valued random variables (H-r.v.), that is, for a complete probability space $(S, \mathcal{S}, \mathbb{P})$, U^δ (respectively y^δ) defines a measurable map $U^\delta : (S, \mathcal{S}, \mathbb{P}) \to (\mathcal{Y}, \mathfrak{B}(\mathcal{Y}))$ (see e.g., Kuo, 1975). Typically, in econometrics, realizations of y^δ are functional transformations of the observed data and U^δ is an error term. In the following, we simply write U, instead of U^δ, in order to lighten the notation. The true value of x that generates the data is denoted by x_*.

We assume a mean-zero Gaussian distribution on $\mathfrak{B}(\mathcal{Y})$ for U: $U \sim \mathcal{N}(0, \delta\Sigma)$ where $\delta > 0$ is the noise level and $\Sigma : \mathcal{Y} \to \mathcal{Y}$ is a covariance operator, that is, Σ is such that $\langle \delta\Sigma\phi_1, \phi_2 \rangle = \mathbf{E}(\langle U, \phi_1 \rangle \langle U, \phi_2 \rangle)$ for all $\phi_1, \phi_2 \in \mathcal{Y}$. By definition, Σ is a one-to-one, linear, positive definite, self-adjoint and trace-class operator. Because Σ is one-to-one, the support of U is all $\mathcal{Y}$ (see e.g., Kuo, 1975). A *trace-class* operator is a compact operator with eigenvalues that are summable. This property rules out a covariance Σ proportional to the identity operator I and this is a key difference between our model and the model used in a large part of the statistical IP literature including Cavalier and Tsybakov (2002), Bissantz, Hohage, Munk, and Ruymgaart (2007), and Knapik et al. (2011). The reason why we focus on a covariance Σ different from I is that most of the functional estimation problems in econometrics that writes as equation (6) does not allow for an identity (or proportional to identity) covariance operator (see Examples 1 and 2).

Remark 3.1. We stress that $\Sigma \neq I$ is *not an assumption but a feature of the econometric models we want to treat* and many difficulties arise because $\Sigma \neq I$. One of the main contributions of our paper is to deal with the important case where Σ is trace-class, i.e., $\Sigma \neq I$. In addition, the identity operator cannot be the covariance operator of a well-defined Gaussian process in a Hilbert space because a Gaussian process with an identity covariance operator cannot generate trajectories that are proper elements in infinite dimensional Hilbert spaces and is only defined in a weaker sense as a weak distribution (see e.g., Kuo, 1975). In general, real (functional) data cannot be generated by a Gaussian process with an identity covariance operator. Thus, the model considered by Knapik et al. (2011) is not appropriate for many econometric models of interest like IV and functional linear regression estimation.

Under model (6) with $U \sim \mathcal{N}(0, \delta\Sigma)$, the sampling distribution P^x, i.e., the conditional distribution of y^δ given x, is a Gaussian distribution on $\mathfrak{B}(\mathcal{Y})$:

$$y^\delta | x \sim P^x = \mathcal{N}\left(Kx, \delta\Sigma\right). \tag{7}$$

The parameter $\delta > 0$ is assumed to converge to 0 when the information from the observed data increases. Hereafter, $\mathbf{E}_x(\cdot)$ will denote the expectation taken with respect to P^x.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

Remark 3.2. The assumption of Gaussianity of the error term U in the econometric model (6) is not necessary and only made in order to construct (and give a Bayesian interpretation to) the estimator. The proofs of our results of frequentist consistency do not rely on the normality of U. In particular, asymptotic normality of $y^\delta|x$, which is verified, e.g., in Example 1, is enough for our estimation procedure and also for our EB data-driven method for choosing the regularization parameter.

Remark 3.3. All the results in the paper are given for the general case where K and Σ do not depend on δ. This choice is made in order to keep our presentation as simple as possible. We discuss how our results apply to the case with unknown K and Σ through Examples 1 and 2.

3.1. Notation

We set up some notational convention used in the paper. We simply write U to denote the error term U^δ. For positive quantities M_δ and N_δ depending on a discrete or continuous index δ, we write $M_\delta \asymp N_\delta$ to mean that the ratio M_δ/N_δ is bounded away from zero and infinity. We write $M_\delta = \mathcal{O}(N_\delta)$ if M_δ is at most of the same order as N_δ. For an H-r.v. W we write $W \sim \mathcal{N}$ to denote that W is a Gaussian process. We denote by $\mathcal{R}(\cdot)$ the range of an operator and by $\mathcal{D}(\cdot)$ its domain. For an operator $B : \mathcal{X} \to \mathcal{Y}$, $\overline{\mathcal{R}(B)}$ denotes the closure in $\mathcal{Y}$ of the range of B. For a bounded operator $A : \mathcal{Y} \to \mathcal{X}$, we denote by A^* its adjoint, i.e., $A^* : \mathcal{X} \to \mathcal{Y}$ is such that $\langle A\psi, \varphi\rangle = \langle \psi, A^*\varphi\rangle$, $\forall \varphi \in \mathcal{X}$, $\psi \in \mathcal{Y}$. The operator norm is defined as $||A|| := \sup_{||\phi||=1} ||A\phi|| = \min\{C \geq 0\,;\, ||A\phi|| \leq C||\phi|| \text{ for all } \phi \in \mathcal{Y}\}$. The operator I denotes the identity operator on both spaces $\mathcal{X}$ and $\mathcal{Y}$, i.e., $\forall \psi \in \mathcal{X}$, $\varphi \in \mathcal{Y}$, $I\psi = \psi$, and $I\varphi = \varphi$.

Let $\{\varphi_j\}_j$ denote an orthonormal basis of $\mathcal{Y}$. The trace of a bounded linear operator $A : \mathcal{Y} \to \mathcal{X}$ is defined as $tr(A) := \sum_{j=1}^\infty \langle (A^*A)^{\frac{1}{2}}\varphi_j, \varphi_j\rangle$ independently of the basis $\{\varphi_j\}_j$. If A is compact, then its trace writes $tr(A) = \sum_{j=1}^\infty \lambda_j$, where $\{\lambda_j\}$ are the singular values of A. The Hilbert–Schmidt norm of a bounded linear operator $A : \mathcal{Y} \to \mathcal{X}$ is denoted by $||A||_{HS}$ and defined as $||A||_{HS}^2 = tr(A^*A)$ (see Kato, 1995).

3.2. Prior Distribution and Main Assumptions

We start by constructing the prior distribution of x. Let L be an unbounded, self-adjoint, densely defined operator in the Hilbert space $\mathcal{X}$ for which $\exists\, \eta > 0$ such that $\langle L\psi, \psi\rangle \geq \eta||\psi||^2$, $\forall \psi \in \mathcal{D}(L)$, and L^{-2s} is trace-class for some $s \in \mathbb{R}_+$. We assume a Gaussian *prior distribution* μ on $\mathfrak{B}(\mathcal{X})$:

$$x|\alpha, s \sim \mu = \mathcal{N}\Big(x_0, \frac{\delta}{\alpha}\Omega_0\Big), \qquad x_0 \in \mathcal{X}, \quad \Omega_0 := L^{-2s} \tag{8}$$

with $\alpha > 0$ such that $\alpha \to 0$ as $\delta \to 0$. The parameter α describes a class of prior distributions and it may be viewed as a hyperparameter. Section 6 provides

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

an EB approach for selecting it. By definition of L, the operator $\Omega_0 : \mathcal{X} \to \mathcal{X}$ is linear, bounded, positive-definite, self-adjoint, compact, and trace-class. Thus, Ω_0 cannot be proportional to I.

Next, we introduce two sets of assumptions. (i) The first set is formed by two assumptions (A.1 and A.2) that are new in the literature and guarantee continuity of the posterior mean of x as a linear operator on $\mathcal{Y}$. The detection of these assumptions is an important contribution because, as remarked in Mandelbaum (1984) and Luschgy (1995), in the Gaussian infinite-dimensional model the posterior mean is in general only defined as a *mlt* which is a notion weaker than that of continuous linear operator. Therefore, in general the posterior mean has not an explicit form and may be an inconsistent estimator in the frequentist sense while our Assumptions A.1 and A.2 ensure closed-form (and then easiness of computation) and consistency for the posterior mean. (ii) The assumptions in the second set (Assumptions A.3 and B) will be used for establishing the rate of contraction of the posterior distribution and concern the smoothness of the operator $\Sigma^{-1/2}K$ and of the true value x_*.

Assumption A.1. $\mathcal{R}(K) \subset \mathcal{D}\left(\Sigma^{-1/2}\right)$.

Since K and Σ are integral operators, $\Sigma^{-1/2}$ is a differential operator and Assumption A.1 demands that the functions in $\mathcal{R}(K)$ are at least as smooth as the functions in $\mathcal{R}(\Sigma^{1/2})$. A.1 ensures that $\Sigma^{-1/2}$ is defined on $\mathcal{R}(K)$ so that $\Sigma^{-1/2}K$, which is used in Assumption A.3, exists.

Hereafter, we denote $B = \Sigma^{-1/2}K\Omega_0^{\frac{1}{2}}$. Assumption A.2 guarantees that B and $\Sigma^{-1/2}B$ exist.

Assumption A.2. $\mathcal{R}\left(K\Omega_0^{\frac{1}{2}}\right) \subset \mathcal{D}\left(\Sigma^{-1}\right)$.

Finally, we introduce Assumption A.3 which requires that $\Sigma^{-1/2}K$ regularizes exactly as L^{-a} for some $a > 0$, where L is the operator defined above and used to construct the prior.

Assumption A.3. Let L be the unbounded, self-adjoint, densely defined operator in $\mathcal{X}$ for which $\exists\, \eta > 0$ such that $\langle L\psi, \psi\rangle \geq \eta||\psi||^2$, $\forall \psi \in \mathcal{D}(L)$, that has been used to construct the prior (8). We assume that there exist three real constants $a > 0$ and $0 < \underline{m} \leq \overline{m} < \infty$ such that

$$\underline{m}||L^{-a}x|| \leq ||\Sigma^{-1/2}Kx|| \leq \overline{m}||L^{-a}x|| \tag{9}$$

on $\mathcal{X}$.

Notice that because $\Sigma^{-1/2}K$ must satisfy (9) it is necessarily an injective operator. We refer to the parameter a in (9) as the "degree of ill-posedness" of the estimation problem under study and a is determined by the rate of decreasing of the spectrum of $\Sigma^{-1/2}K$ (and not only by that one of K as it would be in a classical IP framework for (1)) with respect to the prior covariance operator. Since the

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

spectrum of $\Sigma^{-1/2}K$ is decreasing slower than the spectrum of K we have to control for less ill-posedness than if we used the classical approach. Assumption A.3 is used to determine the rate of contraction of the posterior distribution. Indeed, if this assumption is satisfied, the rate may be expressed in terms of the degree of ill-posedness a and the regularity of the prior s. Assumption A.3 makes a link between the smoothness of the prior—which depends on the operator L^{-1} – and the smoothness of the sampling model—which depends on the operator $\Sigma^{-1/2}K$.[2] However, it is important to stress that in order to construct the prior distribution we do not need to know the sampling model and Assumption A.3 is only used in order to derive (minimax) rates of convergence.

For $t > 0$ and $x \in \mathcal{D}(L^t) \subset \mathcal{X}$ define a norm $||x||_t := ||L^t x||$, where L is the operator defined above. Let $\mathcal{X}_t$ be the completion of $\mathcal{D}(L^t)$ under the norm $||\cdot||_t$. Thus, for $t > 0$, $\mathcal{X}_t = \mathcal{D}(L^t)$. The family of Hilbert spaces $(\mathcal{X}_t)_{t>0}$ is called a Hilbert scale and L is called a generating operator of the Hilbert scale (see Definition 8.18 in Engl et al., 2000 for its detailed properties; Krein and Petunin, 1966, p. 142; Chen and Reiss, 2011, p. 5). Typical examples of $\mathcal{X}_t$ are Sobolev spaces of various kinds.

In IP literature it is standard to link the smoothness of the true value x_* to the smoothness of $\Sigma^{-1/2}K$. To this purpose we introduce Assumption B which quantifies the smoothness of x_* in terms of the operator L^{-1} and of the *Hilbert scale* $\mathcal{X}_t$ generated by it.

Assumption B. For some $0 \leq \beta$, $(x_* - x_0) \in \mathcal{X}_\beta$, that is, there exists a $\rho_* \in \mathcal{X}$ such that $(x_* - x_0) = L^{-\beta}\rho_* \left(\equiv \Omega_0^{\frac{\beta}{2s}}\rho_* \right)$.

The parameter β characterizes the "regularity" of the centered true function $(x_* - x_0)$ and is generally unknown. Assumption B is satisfied by regular functions x_*. For instance, when $\mathcal{X}_\beta$ is the scale of Sobolev spaces, Assumption B is equivalent to assume that $(x_* - x_0)$ has at least β square integrable derivatives.[3]

Assumption B is classical in IP literature (see e.g., Chen and Reiss, 2011) and is closely related to the so-called *source condition* which expresses the regularity of x_* according to the spectral representation of the operator K^*K defining the IP (see Engl et al., 2000; Carrasco et al., 2005). In our case, the regularity of $(x_* - x_0)$ is expressed according to the spectral representation of L and is linked to the smoothness of the operator $\Sigma^{-1/2}K$ through Assumption A.3. This way to link the regularity of x_* to the regularity of the operator $\Sigma^{-1/2}K$ that characterizes the IP is natural in IP theory. We refer to Chen and Reiss (2011) for an explanation of the relationship between Hilbert scale and regularity conditions.

Remark 3.4. Assumption A.3 covers not only the mildly ill-posed but also the severely ill-posed case if $(x_* - x_0)$ in Assumption B is infinitely smooth. In the *mildly ill-posed* case the singular values of $\Sigma^{-1/2}K$ decay slowly to zero (typically at a geometric rate) which means that the kernel of $\Sigma^{-1/2}K$ is finitely smooth. In this case the operator L is generally some differential operator so that L^{-1} is finitely smooth. In the *severely ill-posed* case the singular values of

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

$\Sigma^{-1/2}K$ decay very rapidly (typically at an exponential rate). Assumption A.3 covers also this case if $(x_* - x_0)$ is very smooth. This is because when the singular values of $\Sigma^{-1/2}K$ decay exponentially, Assumption A.3 is satisfied if L^{-1} has an exponentially decreasing spectrum too. On the other hand, L^{-1} is used to describe the regularity of $(x_* - x_0)$, so that in the severely ill-posed case, Assumption B can be satisfied only if $(x_* - x_0)$ is infinitely smooth.

To treat the case where the singular values of $\Sigma^{-1/2}K$ decay exponentially fast and $(x_* - x_0)$ is finitely smooth, we should replace (9) in Assumption A.3 by

$$\underline{m}||\exp(-aL)x|| \leq ||\Sigma^{-1/2}Kx|| \leq \overline{m}||\exp(-aL)x||$$

as, e.g., in Cavalier (2011). In this case the minimax rate would be logarithmic. We discuss this case in the Supplementary Appendix.

Remark 3.5. The specification of the prior covariance operator can be generalized as $\frac{\delta}{\alpha}\Omega_0 = \frac{\delta}{\alpha}QL^{-2s}Q^*$, for some bounded operator Q not necessarily compact. Then, our case is a particular case of this one for $Q = I$. In this generalized setting, Assumptions A.1 and A.2 are replaced by the weaker assumptions $\mathcal{R}(KQ) \subset \mathcal{D}\left(\Sigma^{-1/2}\right)$ and $\mathcal{R}\left(KQL^{-s}\right) \subset \mathcal{D}\left(\Sigma^{-1}\right)$, respectively. In Assumption A.3 the operator $\Sigma^{-1/2}K$ must be replaced by $\Sigma^{-1/2}KQ$ and Assumption B becomes: there exists $\tilde{\rho}_* \in \mathcal{X}$ such that $(x_* - x_0) = QL^{-\beta}\tilde{\rho}_*$.

3.2.1. Covariance Operators Proportional to K. A particular situation often encountered in applications is the case where the sampling covariance operator has the form $\Sigma = (KK^*)^r$, for some $r \in \mathbb{R}_+$, and is related to the classical example of g-priors given in Zellner (1986). In this situation it is convenient to choose $L = (K^*K)^{-\frac{1}{2}}$ so that $\Omega_0 = (K^*K)^s$, for $s \in \mathbb{R}_+$. Because $(KK^*)^r$ and $(K^*K)^s$ are proper covariance operators only if they are trace-class then K is necessarily compact. Assumptions A.1 and A.3 hold for $r \leq 1$ and $a = 1 - r$, respectively. Assumption A.2 holds for $s \geq 2r - 1$.

4. MAIN RESULTS

4.1. Posterior Distribution and Point Estimator

The posterior distribution of x, denoted by μ_δ^Y, is the Bayesian solution of the IP (6) and is characterized in the next theorem together with the joint distribution of (x, y^δ). Because a separable Hilbert space is Polish,[4] there exists a regular version of the posterior distribution μ_δ^Y. This means that μ_δ^Y can be characterized by a conditional probability. The notation $\mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{Y})$ means the Borel σ-field generated by the product topology.

THEOREM 1. *Consider two separable infinite dimensional Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $x|\alpha, s$ and $y^\delta|x$ be two Gaussian H-r.v. on $\mathcal{X}$ and $\mathcal{Y}$ as in (8) and (7), respectively. Then,*

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

(i) (x, y^δ) *is a measurable map from* $(S, \mathcal{S}, \mathbb{P})$ *to* $(\mathcal{X} \times \mathcal{Y}, \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{Y}))$ *and has a Gaussian distribution:* $(x, y^\delta)|\alpha, s \sim \mathcal{N}((x_0, Kx_0), \Upsilon)$, *where* Υ *is a trace-class covariance operator defined as* $\Upsilon(\varphi, \psi) = \left(\frac{\delta}{\alpha}\Omega_0\varphi + \frac{\delta}{\alpha}\Omega_0 K^*\psi, \frac{\delta}{\alpha}K\Omega_0\varphi + (\delta\Sigma + \frac{\delta}{\alpha}K\Omega_0K^*)\psi\right)$ *for all* $(\varphi, \psi) \in \mathcal{X} \times \mathcal{Y}$. *The marginal sampling distribution of* $y^\delta|\alpha, s$ *is* $P_\alpha \sim \mathcal{N}(Kx_0, (\delta\Sigma + \frac{\delta}{\alpha}K\Omega_0K^*))$.

Moreover, let $A : \mathcal{Y} \to \mathcal{X}$ *be a* P_α*-measurable linear transformation (*P_α*-mlt), that is,* $\forall\phi \in \mathcal{Y}$, $A\phi$ *is a* P_α*-almost sure limit, in the norm topology, of* $A_k\phi$ *as* $k \to \infty$, *where* $A_k : \mathcal{Y} \to \mathcal{X}$ *is a sequence of continuous linear operators. Then,*

(ii) *the conditional distribution* μ_δ^Y *of* x *given* y^δ *exists, is regular and almost surely unique. It is Gaussian with mean* $\mathbf{E}(x|y^\delta, \alpha, s) = A(y^\delta - Kx_0) + x_0$ *and trace-class covariance operator* $Var(x|y^\delta, \alpha, s) = \frac{\delta}{\alpha}[\Omega_0 - AK\Omega_0] : \mathcal{X} \to \mathcal{X}$. *Furthermore,* $A = \Omega_0K^*(\alpha\Sigma + K\Omega_0K^*)^{-1}$ *on* $\mathcal{R}((\delta\Sigma + \frac{\delta}{\alpha}K\Omega_0K^*)^{\frac{1}{2}})$.

(iii) *Under Assumptions A.1 and A.2, the operator* A *characterizing* μ_δ^Y *is a continuous linear operator on* $\mathcal{Y}$ *and can be written as*

$$A = \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-1/2}B)^* : \mathcal{Y} \to \mathcal{X} \tag{10}$$

with $B = \Sigma^{-1/2}K\Omega_0^{\frac{1}{2}}$.

Point (ii) in the theorem is an immediate application of the results of Mandelbaum (1984) and we refer to this paper for the proof and for a rigorous definition of P_α-mlt. As stated above, the quantity Ay^δ is defined as a P_α-mlt, which is a weaker notion than that of a linear and continuous operator and A is in general not continuous. In fact, since Ay^δ is a P_α-almost sure limit of A_ky^δ, for $k \to \infty$, the null set where this convergence is not satisfied depends on y^δ and we do not have an almost sure convergence of A_k to A. Moreover, in general, A takes the form $A = \Omega_0K^*(\alpha\Sigma + K\Omega_0K^*)^{-1}$ only on a dense subspace of $\mathcal{Y}$ of P_α-probability measure zero. Outside this subspace, A is defined as the unique extension of $\Omega_0K^*(\alpha\Sigma + K\Omega_0K^*)^{-1}$ to $\mathcal{Y}$ for which we do not have an explicit expression. This means that in general it is not possible to construct a feasible estimator for x.

On the contrary, point (iii) of the theorem shows that, under Assumptions A.1 and A.2, A is defined as a continuous linear operator on $\mathcal{Y}$ for which a closed form is provided. Hence, the posterior mean estimator for x can be implemented in practice and our result (iii) makes operational the Bayesian approach for linear statistical IPs in econometrics. When Assumptions A.1 and A.2 do not hold then we can use a quasi-Bayesian approach as proposed in Florens and Simoni (2012a, 2012b).

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

Summarizing, under a quadratic loss function, the Bayes estimator for a functional parameter x characterized by (6) is the posterior mean

$$\hat{x}_\alpha = \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-1/2}B)^*(y^\delta - Kx_0) + x_0, \quad \text{with} \quad B = \Sigma^{-1/2}K\Omega_0^{1/2}. \tag{11}$$

We point out that because the posterior distribution is available in closed-form, then inference may be easily conducted. In particular, Bayesian credible sets for x_* evaluated at a point, as well as for other linear functionals of x_*, are immediately available in closed-form. Bayesian credible sets for nonlinear functionals of x may be constructed by computing the quantiles of the posterior distribution through draws from the posterior.

4.1.1. Bayes and Tikhonov-type Estimators. As stressed before, the distribution μ (respectively P^x) is realizable as a proper random element in $\mathcal{X}$ (respectively $\mathcal{Y}$) if and only if Ω_0 (respectively Σ) is trace-class. Thus, neither Σ nor Ω_0 can be proportional to I and so, in general, in infinite-dimensional IPs, the posterior mean cannot be equal to the Tikhonov regularized estimator $x_\alpha^T := (\alpha I + K^*K)^{-1}K^*y^\delta$.

However, under Assumptions A.1 and A.2, we show the existence of a close relationship between Bayesian and frequentist approach to statistical IPs. In fact, the posterior mean $\hat{x}_\alpha$ is equal to the Tikhonov regularized solution in the Hilbert scale $(\mathcal{X}_t)_{t>0}$ of the equation $\Sigma^{-1/2}y^\delta = \Sigma^{-1/2}Kx + \Sigma^{-1/2}U$. The existence of $\Sigma^{-1/2}K$ is guaranteed by A.1; the stochastic processes $\Sigma^{-1/2}y^\delta$ and $\Sigma^{-1/2}U$ have to be interpreted in the sense of weak distributions (see Kuo, 1975; Bissantz et al., 2007). Since $\mathbf{E}(x|y^\delta, \alpha, s) = A(y^\delta - Kx_0) + x_0$, we have, under A.1 and A.2:

$$\begin{aligned}\mathbf{E}(x|y^\delta, \alpha, s) &= L^{-s}(\alpha I + L^{-s}K^*\Sigma^{-1}KL^{-s})^{-1}L^{-s}K^*\Sigma^{-1/2}\Sigma^{-1/2}(y^\delta - Kx_0) + x_0 \\ &= (\alpha L^{2s} + T^*T)^{-1}T^*(\tilde{y}^\delta - Tx_0) + x_0, \\ &\quad \text{where} \quad T = \Sigma^{-1/2}K, \quad \tilde{y}^\delta = \Sigma^{-1/2}y^\delta\end{aligned}$$

and it is equal to the minimizer, with respect to x, of the Tikhonov functional

$$||\tilde{y}^\delta - Tx||^2 + \alpha||x - x_0||_s^2.$$

4.2. Asymptotic Analysis

In this section we analyze frequentist asymptotic properties of the posterior distribution μ_δ^Y of x. The asymptotic analysis is for $\delta \to 0$. Let P^{x_*} denote the sampling distribution (7) with $x = x_*$. We remind the definition of *posterior consistency* (see Le Cam, 1986; Diaconis and Freedman, 1986; or Ghosh and Ramamoorthi, 2003):

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

DEFINITION 1. *The posterior distribution of x is* consistent *at x_* with respect to P^{x_*} if it weakly converges toward the Dirac measure δ_{x_*} at the point x_*, i.e., if, for every neighborhood $\mathcal{U}$ of x_*, $\mu_\delta^Y(\mathcal{U}|y^\delta, \alpha, s) \to 1$ in P^{x_*}-probability or P^{x_*}-a.s. as $\delta \to 0$.*

Posterior consistency provides the basic frequentist validation of our Bayesian procedure because it ensures that with a sufficiently large amount of data, it is almost possible to recover the truth accurately. Lack of consistency is extremely undesirable, and one should not use a Bayesian procedure if the corresponding posterior distribution is inconsistent.

Our asymptotic analysis is organized as follows. First, we consider the posterior mean $\hat{x}_\alpha$ as an estimator for the solution of (6) and study the rate of convergence of the associated risk. Second, we state posterior consistency and recover the rate of contraction of the posterior distribution. We take the mean integrated square error (MISE) as the risk associated with $\hat{x}_\alpha$ and denote it by $\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2$ where $\mathbf{E}_{x_*}$ denotes the expectation taken with respect to P^{x_*}. The use of a MISE representation of the risk is common in IP literature and in addition, in our framework, it matches the quadratic loss function used to determine the Bayes estimator. We denote by $\mathcal{X}_\beta(\Gamma)$ the ellipsoid of the type

$$\mathcal{X}_\beta(\Gamma) := \{\varphi \in \mathcal{X};\ ||\varphi||_\beta^2 \le \Gamma\}, \qquad 0 < \Gamma < \infty \tag{12}$$

where $||\varphi||_\beta := ||L^\beta \varphi||$. Our asymptotic results will be valid uniformly on $\mathcal{X}_\beta(\Gamma)$. The following theorem gives the asymptotic behavior of $\hat{x}_\alpha$.

THEOREM 2. *Let us consider the observational model (6) with x_* being the true value of x that generates the data and let $\gamma \in [0, 1]$ be such that $tr(L^{-2\gamma(a+s)}) < \infty$. Under Assumptions A.1–A.3 and B, we have*

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} \mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \mathcal{O}\left(\alpha^{\frac{\tilde{\beta}}{a+s}} + \delta\alpha^{-\frac{a+\gamma(a+s)}{a+s}}\right),$$

where $\tilde{\beta} = \min(\beta, a + 2s)$.

The theorem provides a uniform rate over the ellipsoid $\mathcal{X}_\beta(\Gamma)$. It follows from this result that if $(x_* - x_0)$ satisfies Assumption B, then $\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \mathcal{O}\left(\alpha^{\frac{\tilde{\beta}}{a+s}} + \delta\alpha^{-\frac{a+\gamma(a+s)}{a+s}}\right)$.

The value $a + 2s$ in the theorem plays the role of the *qualification* in a classical regularization scheme, that is, $a + 2s$ is the maximum degree of regularity of x_* that can be exploited in order to improve the rate. It is equal to the qualification of a Tikhonov regularization in the Hilbert scale $(\mathcal{X}_s)_{s\in\mathbb{R}}$ (see e.g., Engl et al., 2000, Sect. 8.5).

The value of the prior parameter α that minimizes the rate given in Theorem 2 is: $\alpha_\gamma^{\min} := \kappa\delta^{\frac{a+s}{\tilde{\beta}+a+\gamma(a+s)}}$, for some constant $\kappa > 0$. When α is set equal to $\alpha_\gamma^{\min}$,

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

then $\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} \mathbf{E}_{x_*}||\hat{x}_{\alpha_\gamma^{\min}} - x_*||^2 = \mathcal{O}\left(\delta^{\frac{\tilde{\beta}}{\tilde{\beta}+a+\gamma(a+s)}}\right)$. The minimax rate can be obtained as the infimum of $\delta^{\frac{\tilde{\beta}}{\tilde{\beta}+a+\gamma(a+s)}}$ over the set $\mathcal{G}$ of the possible values for γ and is attained at $\gamma^* := \inf\mathcal{G}$ when $\beta \le a+2s$. The values of γ and γ^* and the lower bound of $\mathcal{G}$ depend on the definition of L^{-1}, the smoothness of (x_*-x_0), and the degree of ill-posedness of the problem. Since β and a depend in turn on the definition of L^{-1} (due to Assumptions A.3 and B), in order to define $\mathcal{G}$ we have to "normalize" L as we now explain. In particular, we consider two cases (see also Remark 3.4):

I. *Mildly ill-posed problem and* (x_*-x_0) *belongs to a Sobolev class.* In this case L^{-1} is finitely smooth so that, typically, its eigenvalues decrease at a geometric rate. Assume that the operator L is such that $tr(L^{-(1+\epsilon)}) < \infty$, $\forall\epsilon > 0$ and $tr(L^{-(1+\epsilon)}) \not< \infty$, $\forall\epsilon \le 0$. Then, $\mathcal{G} := \{\gamma \in \mathbb{R}; \frac{1}{2(a+s)} < \gamma < 1\}$ and $\gamma^* = \frac{1}{2(a+s)}$. The minimum rate of our procedure is $\delta^{\frac{\tilde{\beta}}{\tilde{\beta}+a+\frac{1}{2}}}$ and coincides with the minimax rate for this case if $\beta \le a+2s$.

II. *Severely ill-posed problem and* (x_*-x_0) *belongs to the class of analytic functions.* In this case L^{-1} is infinitely smooth so that, typically, its eigenvalues decrease at an exponential rate. Assume that the operator L is such that $tr(L^{-\epsilon}) < \infty$, $\forall\epsilon > 0$ and $tr(L^{-\epsilon}) \not< \infty$, $\forall\epsilon \le 0$. Then, $\mathcal{G} := \{\gamma \in \mathbb{R}; 0 < \gamma < 1\}$ and $\gamma^* = 0$. The minimum rate of our procedure is $\delta^{\frac{\tilde{\beta}}{\tilde{\beta}+a}}$ and coincides with the minimax rate for this case if $\beta \le a+2s$.

We refer to Cavalier (2011) for a definition of mildly/severely ill posedness and for a review of rates in cases I and II. The value $\alpha = \alpha_{\gamma^*}^{\min}$, which allows to obtain the minimax rate, corresponds to three different behaviors of the prior depending on the values of β and s:

(NS) nonscaling prior if $s = \beta + \frac{1}{2}$ (case I) or $s = \beta$ (case II), i.e., the regularity of the prior matches the regularity of x_*;

(SO) spreading out prior if $s > \beta + \frac{1}{2}$ (case I) or $s > \beta$ (case II);

(S) shrinking prior if $s < \beta + \frac{1}{2}$ (case I) or $s < \beta$ (case II).

This means that when $s \neq \beta + \frac{1}{2}$ (case I) or $s \neq \beta$ (case II), in order to achieve the minimax rate of convergence, the prior must either spread out to become rougher (case *(SO)*) or shrink to become smoother (case *(S)*).

When $\alpha \neq \alpha_{\gamma^*}^{\min}$ or $\gamma \neq \gamma^*$, the rate is slower than the minimax rate but we still have consistency provided that we set $\alpha \asymp \delta^\epsilon$ for $0 < \epsilon < \frac{a+s}{a+\gamma(a+s)}$. In case I, remark that, since $tr(L^{-2s}) < \infty$, $\frac{s}{a+s}$ is a possible value for γ which implies that $\frac{s}{a+s} > \frac{1}{2(a+s)}$ and so 1 is a possible value for ϵ. This implies that consistency is always obtained with a nonscaling prior even if the minimax rate is obtained only in particular cases.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

A similar discussion applies to the rate of contraction of the posterior distribution which is given in the next theorem.

THEOREM 3. *Let the assumptions of Theorem 2 be satisfied. For any sequence* $M_\delta \to \infty$ *the posterior distribution satisfies*

$$\mu_\delta^Y\{x \in \mathcal{X} : ||x - x_*|| > \varepsilon_\delta M_\delta\} \to 0$$

in P^{x_*}*-probability as* $\delta \to 0$, *where* $\varepsilon_\delta = \left(\alpha^{\frac{\tilde\beta}{2(a+s)}} + \delta\alpha^{-\frac{a+\gamma(a+s)}{2(a+s)}}\right)$, $\tilde\beta = \min(\beta, a+2s)$, *and* $\gamma \in [0,1]$ *is such that* $tr(L^{-2\gamma(a+s)}) < \infty$.

We refer to ε_δ as the rate of contraction of the posterior distribution. If α is chosen such that the two terms in ε_δ are balanced, that is, $\alpha \asymp \alpha_\gamma^{\min}$, then $\varepsilon_\delta = \delta^{\frac{\tilde\beta}{2\tilde\beta+2a+2\gamma(a+s)}}$. Then, the minimax rate for cases I and II above is obtained when $\beta \le a+2s$ and we set $\alpha = \alpha_{\gamma*}^{\min}$. In this case, the prior is either fixed or scaling depending whether s equates $\beta + \frac{1}{2}$ (respectively β in case II) or not.

5. OPERATORS WITH GEOMETRIC SPECTRA

In this section we analyze the important case where the IP (6) is *mildly ill-posed* and $(x_* - x_0)$ belongs to a Sobolev class. This case is formalized by Assumptions C and B with L^{-1} constructed as in (13).

Assumption C. Under Assumption A.1, the operator $T := \Sigma^{-1/2}K$ is injective and admits a singular system $(\lambda_{jT}, \varphi_j, \psi_j)_j$, that is, $T\psi_j = \lambda_{jT}\varphi_j$ and $T^*\varphi_j = \lambda_{jT}\psi_j$. Moreover, the eigenvalues of T^*T satisfy

$$\underline{a}\, j^{-2a} \le \lambda_{jT}^2 \le \overline{a}\, j^{-2a}, \qquad j = 1, 2, \ldots$$

for some $a \ge 0$ and $\underline{a}, \overline{a} > 0$.

Assumption C states that the operator T^*T has a polynomially decreasing spectrum which, by using the notation defined in Section 3.1, may be rewritten as $\lambda_{jT}^2 \asymp j^{-2a}$. Imposing such a rate of decay of the spectrum is common in statistical IP literature (see e.g., Assumption A.3 in Hall and Horowitz, 2005 or Assumption B3 in Cavalier and Tsybakov, 2002).

Because Assumption C implies that T^*T is strictly positive definite and $\{\psi_j\}$ is an orthonormal system which spans $\overline{\mathcal{R}(T^*T)}$ we can then construct a self-adjoint bounded operator L^{-1} as

$$\forall \varphi \in \mathcal{X}, \quad L^{-1}\varphi = \sum_{j=1}^{\infty} j^{-1}\langle \varphi, \psi_j\rangle \psi_j. \tag{13}$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

It follows that L is unbounded with domain $\mathcal{D}(L) = \{x \in \mathcal{X}; \sum_{j=1}^{\infty} j^2 \langle x, \psi_j \rangle^2 < \infty\}$. When L^{-1} is constructed as in (13) we may write $\mathcal{X}_\beta(\Gamma)$ as $\mathcal{X}_\beta(\Gamma) := \{\varphi \in \mathcal{X}\,;\, \sum_j j^{2\beta} \langle \varphi, \psi_j \rangle^2 \le \Gamma\}$ and Assumption B implies that $(x_* - x_0)$ belongs to a Sobolev class. The following proposition gives the rate attained by the MISE of $\hat{x}_\alpha$ under Assumption C and (13).

PROPOSITION 1. *Let the operator L^{-1} be defined as in (13) and Assumptions A.1, A.2, B, and C hold with $a > 0$ and $s > \frac{1}{2}$. Then we have*

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} \mathbf{E}_{x_*} \|\hat{x}_\alpha - x_*\|^2 \asymp \alpha^{\frac{\tilde{\beta}}{a+s}} + \delta\alpha^{-\frac{2a+1}{2(a+s)}}, \tag{14}$$

with $\tilde{\beta} = \min(\beta, 2(a+s))$. Moreover, (i) for $\alpha \asymp \delta$ (fixed prior),

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} \mathbf{E}_{x_*} \|\hat{x}_\alpha - x_*\|^2 \asymp \delta^{\frac{\beta \wedge \left(s-\frac{1}{2}\right)}{a+s}};$$

(ii) for $\alpha \asymp \delta^{\frac{a+s}{\tilde{\beta}+a+\frac{1}{2}}}$ (optimal rate of α),

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} \mathbf{E}_{x_*} \|\hat{x}_\alpha - x_*\|^2 \asymp \delta^{\frac{\tilde{\beta}}{\tilde{\beta}+a+\frac{1}{2}}}.$$

The minimax rate of convergence over a Sobolev ellipsoid $\mathcal{X}_\beta(\Gamma)$ (in the mildly ill-posed case) is of the order $\delta^{\frac{\beta}{\beta+a+\frac{1}{2}}}$. Proposition 1 says that the uniform rate of the MISE associated with $\hat{x}_\alpha$ is minimax if: either the parameter s of the prior is chosen such that $s = \beta + \frac{1}{2}$ and the prior is fixed (case (i), i.e., $\alpha \asymp \delta$) or if $\beta \le 2(a+s)$ and the prior is scaling at the optimal rate (case (ii), i.e., $\alpha \asymp \delta^{(a+s)/(\beta+a+\frac{1}{2})}$). In all the other cases the rate is slower than the minimax rate but consistency is still verified provided that $\alpha \asymp \delta^{\epsilon}$ for $0 < \epsilon < \frac{a+s}{a+\frac{1}{2}}$. Remark that since $s > \frac{1}{2}$ then a fixed prior ($\epsilon = 1$) always guarantees consistency (even if the rate is not always minimax).

This result, similar to the one in Theorem 2, means that when the prior is "correctly specified" ("correct" in the sense that the regularity $s - \frac{1}{2}$ of the trajectories generated by the prior is the same as the regularity of x_*) we obtain a minimax rate without scaling the prior covariance. On the other hand, if $s < \beta + \frac{1}{2}$, i.e., the prior is "undersmoothing", the minimax rate can still be achieved as soon as $\beta \le 2(a+s)$ and the prior is shrinking at the optimal rate. When the prior is "oversmoothing", i.e., $s > \beta + \frac{1}{2}$, the minimax rate can be achieved if the prior distribution spreads out at the optimal α (the prior has to be more and more dispersed in order to become rougher).

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

In many cases it is reasonable to assume that the functional parameter x_* has generalized Fourier coefficients (in the basis made of the eigenfunctions $\{\psi_j\}$ of L^{-s}) that are geometrically decreasing (see e.g., Assumption A.3 in Hall and Horowitz, 2005; Theorem 4.1 in Van Rooij and Ruymgaart, 1999). Thus, it is common in statistical IP literature to consider the following assumption instead of Assumption B.

Assumption B'. For some $b_0 > \frac{1}{2}$ and $\{\psi_j\}$ defined in Assumption C, $\langle (x_* - x_0), \psi_j \rangle \asymp j^{-b_0}$.

In the rest of this section we establish asymptotic results under this assumption and Assumption C. Assumption B is more general than Assumption B' since it allows the case of exponentially declining Fourier coefficients $\langle (x_* - x_0), \psi_j \rangle$. We use Assumption B' to show sharp adaptiveness of our EB procedure. If B' holds, then $\exists \Gamma < \infty$ such that Assumption B holds for some $0 \leq \beta < b_0 - \frac{1}{2}$. The following result gives the rate of the MISE when Assumptions B' and C hold.

PROPOSITION 2. *Let the operator L^{-1} be defined as in (13) and Assumptions A.1, A.2, B', and C hold with $a > 0$ and $s > \frac{1}{2}$. Then,*

$$\begin{aligned} \alpha^{\frac{2b_0-1}{2(a+s)}} c_1 + \delta\alpha^{-\frac{2a+1}{2(a+s)}} c_2(\bar{t}) &\leq \mathbf{E}_{x_*} ||\hat{x}_\alpha - x_*||^2 \\ &\leq \alpha^2 I(b_0 \geq (2a+2s)) + \alpha^{\frac{2b_0-1}{2(a+s)}} \tilde{c}_1 \\ &\quad + \delta\alpha^{-\frac{2a+1}{2(a+s)}} \tilde{c}_2(\bar{t}), \end{aligned} \tag{15}$$

where c_1, $\tilde{c}_1$, c_2, $\tilde{c}_2$, and $\bar{t}$ are positive constants and $I(\mathcal{A})$ denotes the indicator function of an event $\mathcal{A}$. Moreover, for $\alpha \asymp \delta^{\frac{a+s}{\tilde{b}_0+a}} \equiv \alpha_*$ and $\tilde{b}_0 = \min(b_0, 2a + 2s + 1/2)$,

$$\mathbf{E}_{x_*} ||\hat{x}_{\alpha_*} - x_*||^2 \asymp \delta^{\frac{2\tilde{b}_0-1}{2(\tilde{b}_0+a)}}. \tag{16}$$

When $b_0 \leq 2a + 2s + 1/2$ the rate of the lower bound $\alpha^{\frac{2b_0-1}{2(a+s)}} + \delta\alpha^{-\frac{2a+1}{2(a+s)}}$ given in (15) provides, up to a constant, a lower and an upper bound for the rate of the estimator $\hat{x}_\alpha$ and so it is optimal. In this case, the minimax-optimal rate $\delta^{\frac{2b_0-1}{2(b_0+a)}}$ is obtained when we set $\alpha \asymp \delta^{\frac{a+s}{b_0+a}}$. We call "optimal" this value of α and denote it by α_*. This value corresponds: to a fixed prior when $s = b_0$, to a shrinking prior when $s < b_0 \leq 2a + 2s + 1/2$, and to a spreading out prior when $s > b_0$. Intuitively, when $s < b_0$ (respectively $s > b_0$) the trajectories generated by the prior are rougher (respectively smoother) than x_* and so the support of the prior must be shrunk (respectively spread out). When $\delta = n^{-1}$ the rate $n^{-\frac{(2b_0-1)}{2(b_0+a)}}$ is shown to be minimax in Hall and Horowitz (2007).

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

Moreover, if we set $\beta = \sup\left\{\breve{\beta} \geq 0;\ (x_* - x_0) \text{ satisfies B' and } \sum_j j^{2\breve{\beta}} \langle (x_* - x_0), \psi_j \rangle^2 < \infty\right\}$, then $\beta = b_0 - \frac{1}{2}$ and the rate $\delta^{\frac{2\tilde{b}_0 - 1}{2(\tilde{b}_0 + a)}}$ is uniform over $\mathcal{X}_\beta(\Gamma)$ and equal to the optimal rate of Proposition 1.

In the following theorem we give the rate of contraction of the posterior distribution under Assumption C.

THEOREM 4. *Let the assumptions of Proposition 2 be satisfied. For any sequence $M_\delta \to \infty$ the posterior distribution satisfies*

$$\mu_\delta^Y \{x \in \mathcal{X} : ||x - x_*|| > \varepsilon_\delta M_\delta\} \to 0$$

in P^{x_*}-probability as $\delta \to 0$, where $\varepsilon_\delta = \left(\alpha^{\frac{2\tilde{b}_0 - 1}{4(a+s)}} + \delta^{\frac{1}{2}} \alpha^{-\frac{2a+1}{4(a+s)}}\right)$ and $\tilde{b}_0 = \min(b_0, 2(a+s) + 1/2)$. Moreover, (i) for $\alpha \asymp \delta$ (fixed prior):

$$\varepsilon_\delta = \delta^{\frac{2(\tilde{b}_0 \wedge s) - 1}{4(a+s)}}$$

and (ii) for $\alpha \asymp \delta^{\frac{a+s}{\tilde{b}_0 + a}}$ (optimal rate of α):

$$\varepsilon_\delta = \delta^{\frac{2\tilde{b}_0 - 1}{4(\tilde{b}_0 + a)}}.$$

The rate of contraction ε_δ is equal to the minimax rate $\delta^{\frac{2b_0 - 1}{4(b_0 + a)}}$ if $b_0 \leq 2(a+s) + 1/2$ and $\alpha \asymp \delta^{\frac{a+s}{b_0 + a}}$. Depending on the relation between s and b_0 the corresponding prior μ is shrinking, spreading out, or fixed (see comments after Proposition 2).

This section has focused on the mildly ill-posed case with $(x_* - x_0)$ in a Sobolev class of functions. The case where the λ_{jT}s in Assumption C are decreasing exponentially fast (*severe ill-posedness*) and $(x_* - x_0)$ is analytic is treated in the Supplementary Appendix.

6. ADAPTIVE SELECTION OF α: AN EMPIRICAL BAYES (EB) APPROACH

As shown by Theorem 1(iii) and Section 4.1.1, the parameter α of the prior plays the role of a *regularization parameter* and $\{\hat{x}_\alpha\}_{\alpha > 0}$ defines a class of estimators for x, which are equal to a Tikhonov-type frequentist estimator. We have shown that for α decreasing at a convenient rate, this estimator converges at the minimax rate. However, this rate and the corresponding value for α are unknown in practice since they depend on the regularity of $\Sigma^{-1/2} K$ and of x_*. Thus, it is crucial for the implementation of the estimation procedure to have an adaptive data-driven method for selecting α.

Adaptive Bayesian estimation can be implemented in two different ways. A first way (Empirical Bayes) consists in constructing the prior for x by using an estimate $\hat{\alpha}$ obtained in a first step through a data-driven method. A second way consists in integrating out α from the conditional prior of x, given α, with respect to

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

the prior of α. The two procedures are adaptive if they lead to a posterior for x which contracts at the optimal rate. Adaptivity relies on the choice of the estimation criterion used to obtain $\hat{\alpha}$ in the first procedure and on the choice of the prior for α in the second procedure. Adaptive Bayesian estimation for models different from ours has been considered, e.g., in Van der Vaart and Van Zanten (2009).

We propose in Sections 6.1 and 6.2 an adaptive data-driven method based on an EB procedure for selecting α. This procedure can be easily implemented for general operators K, Σ, and L satisfying Assumptions A.1 and A.2.

6.1. Characterization of the Likelihood

The marginal distribution of $y^\delta|\alpha, s$, which was given in Theorem 1(i), is

$$y^\delta|\alpha, s \sim P_\alpha, \quad P_\alpha = \mathcal{N}\left(Kx_0, \delta\Sigma + \frac{\delta}{\alpha}K\Omega_0 K^*\right) \tag{17}$$

and is obtained by marginalizing P^x with respect to the prior distribution of x. The following theorem, which is an application of Theorem 3.3 in Kuo (1975), characterizes a probability measure P_0 that is equivalent to P_α for every $\alpha > 0$ and the likelihood of P_α with respect to P_0.

THEOREM 5. *Let P_0 be a Gaussian measure with mean Kx_0 and covariance operator $\delta\Sigma$, that is, $P_0 = \mathcal{N}(Kx_0, \delta\Sigma)$. Under Assumptions A.1 and A.2, the Gaussian measure P_α defined in (17) is equivalent to P_0. Moreover, the Radon–Nikodym derivative is given by*

$$\frac{dP_\alpha}{dP_0}(\{z_j\}) = \prod_{j=1}^{\infty}\sqrt{\frac{\alpha}{\lambda_j^2+\alpha}}\, e^{\frac{\lambda_j^2}{2(\lambda_j^2+\alpha)}z_j^2}, \tag{18}$$

where $\{\varphi_j, \lambda_j^2\}$ are the eigenfunctions and eigenvalues of BB^*, respectively.

In our setting: $z_j = \frac{\langle y^\delta - Kx_0, \Sigma^{-1/2}\varphi_j\rangle}{\sqrt{\delta}}$ and $\Sigma^{-1/2}\varphi_j$ is defined under Assumption A.2.

6.2. Adaptive EB Procedure

Let ν denote a prior distribution for α such that $\frac{d\log\nu(\alpha)}{d\alpha} = \nu_1\alpha^{-1} + \nu_2$ for two constants $\nu_1 > 0$ and $\nu_2 < 0$. Examples of priors for α satisfying this condition include the gamma distribution and the beta distribution on $(0, 1)$. To implement the EB procedure we first have to select a value for α depending on the data y^δ. For that, we define the *marginal maximum a posteriori* (MAP) *estimator* $\hat{\alpha}$ of α to

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

be the maximizer of the marginal log-posterior of $\alpha|(y^\delta, s)$ which is proportional to $\log\left[\frac{dP_\alpha}{dP_0}\nu\right]$:

$$\begin{aligned}\hat{\alpha} &:= \arg\max_{\alpha} \bar{S}(\alpha, y^\delta)\\ \bar{S}(\alpha, y^\delta) &:= \log\left(\frac{dP_\alpha}{dP_0}(\{z_j\})\nu(\alpha)\right)\\ &= \frac{1}{2}\sum_{j=1}^{\infty}\left[\log\left(\frac{\alpha}{\alpha+\lambda_j^2}\right)+\frac{\lambda_j^2}{\alpha+\lambda_j^2}\frac{\langle y^\delta - Kx_0, \Sigma^{-1/2}\varphi_j\rangle^2}{\delta}\right]\\ &\quad +\log\nu(\alpha). \end{aligned} \tag{19}$$

In an equivalent way, $\hat{\alpha}$ is defined as the solution of the first order condition $\frac{\partial}{\partial\alpha}\bar{S}(\alpha, y^\delta) = 0$. We denote $S_{y^\delta}(\alpha) := \frac{\partial}{\partial\alpha}\bar{S}(\alpha, y^\delta)$ where

$$\begin{aligned} S_{y^\delta}(\alpha) &:= \frac{1}{2}\left(\sum_{j=1}^{\infty}\frac{\lambda_j^2}{\alpha(\alpha+\lambda_j^2)} - \sum_{j=1}^{\infty}\frac{\lambda_j^2\langle y^\delta - Kx_0, \Sigma^{-1/2}\varphi_j\rangle^2}{\delta(\alpha+\lambda_j^2)^2}\right)\\ &\quad + \frac{d\log\nu(\alpha)}{d\alpha} \quad \text{and} \quad S_{y^\delta}(\hat{\alpha}) = 0. \end{aligned} \tag{20}$$

Then, our strategy consists in plugging the value $\hat{\alpha}$, found in this way, back into the prior of $x|(\alpha, s)$ and then compute the posterior distribution $\mu^Y_{\delta,\hat{\alpha}}$ of x and the posterior mean estimator $\hat{x}_{\hat{\alpha}}$. We refer to $\hat{x}_{\hat{\alpha}}$ as the EB estimator and to $\mu^Y_{\delta,\hat{\alpha}}$ as the EB posterior of x.

6.3. Posterior Consistency for the EB Posterior (Mildly Ill-Posed Case)

In this section we study existence of $\hat{\alpha}$ and the rate at which $\hat{\alpha}$ decreases to 0 and show posterior consistency of $\mu^Y_{\delta,\hat{\alpha}}$.[5] Theorem 6 shows that when $b_0 - 1/2 \leq a + 2s$ the MAP $\hat{\alpha}$ is of the same order as α_* $(\equiv \delta^{\frac{a+s}{b_0+a}})$ with probability that approaches 1 as $\delta \to 0$. Moreover, it states that the EB-posterior distribution $\mu^Y_{\delta,\hat{\alpha}}$ concentrates around the true x_* in probability. In the next theorem, let $I(\mathcal{A})$ denote the indicator function of an event $\mathcal{A}$

THEOREM 6. *Let the operator L^{-1} be defined as in (13) and Assumptions A.1, A.2, B', and C hold with $a > 0$ and $s > \frac{1}{2}$. Let ν be a prior distribution for α such that $\frac{d\log\nu(\alpha)}{d\alpha} = \nu_1\alpha^{-1} + \nu_2$ for two constants $\nu_1 > 0$, $\nu_2 < 0$. Then, with probability approaching 1, a solution $\hat{\alpha}$ to the equation $S_{y^\delta}(\alpha) = 0$ exists and is of order: $\hat{\alpha} \asymp \delta^{\frac{a+s}{b_0+a+\tilde{\eta}}}$ for $\tilde{\eta} = \eta I(b_0 - a - 2s - 1/2 > 0)$ and any $(b_0 + a) > \eta > \max\{(b_0 - s - 1/2), 0, (b_0 - 2s - a - 1/2)\frac{1}{4}\}$. Moreover, for any sequence $M_\delta \to \infty$ the EB-posterior distribution of x satisfies*

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

$$\mu^{Y}_{\delta,\hat{\alpha}}\{x \in \mathcal{X} : ||x - x_*|| > \varepsilon_\delta M_\delta\} \to 0 \tag{21}$$

in P^{x_*}*-probability as* $\delta \to 0$ *where* $\varepsilon_\delta = \delta^{\frac{\tilde{b}_0 - \frac{1}{2}}{2(\tilde{b}_0 + a + \tilde{\eta})}}$ *and* $\tilde{b}_0 = \min(b_0, 2(a+s) + 1/2)$.

The consistency of the EB estimator $\hat{x}_{\hat{\alpha}}$ follows from the posterior consistency of $\mu^{Y}_{\delta,\hat{\alpha}}$. The theorem says that the posterior contraction rate of the EB-posterior distribution is equal to the minimax rate $\delta^{\frac{2b_0-1}{4(b_0+a)}}$ when $b_0 - 1/2 \leq a + 2s$ which is satisfied, for instance, when the prior is very smooth. In all the other cases the rate is slower.

Remark 6.1. While this theorem is stated and proved for a Gaussian error term U, this result holds also for the more general case where U is only asymptotically Gaussian. In Appendix E we give some hints about how the proof should be modified in this case. Therefore, our EB approach works also in the case where U is only approximately Gaussian. This emphasizes what has already been stressed in Remark 3.2 that Gaussianity of U is not fundamentally needed for our procedure to work and to be consistent.

Remark 6.2. Alternative ways to choose regularization parameters for IPs have been proposed in mathematics, statistics, and econometrics. Some of these methods work under assumptions that are not suitable for econometrics applications. Many of these methods focus on series estimators while very little is known about adaptive procedures in Tikhonov estimation. Our methods contribute to the latter. For adaptive procedures for series estimators in the nonparametric IV framework see e.g., Horowitz (2014) and references therein. Assumptions and settings of these adaptive procedures differ significantly from ours.

7. EXAMPLES

7.1. Example 1: Instrumental Variable (IV) Regression Estimation (Continued)

Recall the framework of Example 1 where the IV regression x is characterized as the solution of the integral equation (6) with $\delta = n^{-1}$ and y^δ and K defined as in (2). Suppose to estimate f_{YW} and f_{ZW} by kernel smoothing, where f_{YW} (respectively f_{ZW}) denotes the density of the distribution of (Y, W) (respectively (Z, W)) with respect to the Lebesgue measure. For simplicity we assume that (Z, W) is a bivariate random vector. Let $K_{Z,h}$ and $K_{W,h}$ denote two univariate kernel functions in $L^2[0,1]$, h be the bandwidth, and $(y_i, w_i, z_i)_{i=1}^n$ be the n-observed random sample. For every $\varphi, \phi \in L^2[0,1]$, denote $\Lambda : L^2[0,1] \to L^2[0,1]$ the operator $\Lambda\varphi = \int a(w,v)\varphi(w)dw$, with $a(w,v)$ a known function chosen by the econometrician, and $\tilde{K}_\delta : L^2[0,1] \to L^2[0,1]$ the operator

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

$\tilde{K}_\delta\phi = \frac{1}{n}\sum_{i=1}^n \frac{K_{W,h}(w_i-w)}{h}\int_0^1 \phi(z)\frac{K_{Z,h}(z_i-z)}{h}dz$. Therefore, $K = \Lambda\tilde{K}_\delta$ so that the quantities in (2) can be rewritten as

$$y^\delta = \Lambda\left[\hat{\mathbf{E}}(Y|W=w)\hat{f}_W\right](v) = \int a(w,v)\frac{1}{nh}\sum_{i=1}^n y_i K_{W,h}(w_i-w)dw$$

$$\text{and} \quad Kx = \int a(w,v)\frac{1}{n}\sum_{i=1}^n \frac{K_{W,h}(w_i-w)}{h}\int x(z)\frac{K_{Z,h}(z_i-z)}{h}dzdw. \qquad \textbf{(22)}$$

Remark that $\lim_{\delta\to 0}\tilde{K}_\delta\phi = f_W(w)\mathbf{E}(\phi|w) = M_f\mathbf{E}(\phi|w)$ where M_f denotes the multiplication operator by f_W. If $a = f_{WZ}$, then $\lim_{\delta\to 0}K$ is the same integral operator as in Hall and Horowitz (2005). Assume that: f_W and f_Z are bounded away from infinity on $[0,1]$, $a(w,v)$ is chosen such that $\Lambda^*\Lambda$ is a trace-class operator, and $\mathbf{E}[(Y-x(Z))^2|W] = \sigma^2 < \infty$.

In this example, $U := y^\delta - Kx$ and the assumption that $U \sim \mathcal{N}(0,\delta\Sigma)$ holds asymptotically. In particular, the transformation of the model through Λ is necessary in order to have weak convergence toward a Gaussian process.[6] Then, $\sqrt{n}\left(y^\delta - Kx\right) \Rightarrow \mathcal{N}(0,\Lambda\tilde{\Sigma}\Lambda^*)$ in $L^2[0,1]$ as $n\to\infty$, where '$\Rightarrow$' denotes weak convergence, and, for every $\phi_2 \in L^2[0,1]$, $\tilde{\Sigma}\phi_2 = \sigma^2[f_W(v)]^2\phi_2(v) = \sigma^2 M_f^2\phi_2(v)$ and $\Lambda^*\phi_2 = M_f^{-1}\int_0^1 a(w,v)\phi_2(v)f_W(v)dv$ is the adjoint operator of Λ. The operator $\Sigma = \Lambda\tilde{\Sigma}\Lambda^*$ is unknown and replaced by $\hat{\Sigma}\cdot = \Lambda\frac{\hat{\sigma}^2}{n}\sum_{i=1}^n \frac{K_{W,h}(w_i-w)}{h}\int_0^1 a(w,v)\varphi(v)\cdot\hat{f}_W(v)dv$. Thus, the sampling distribution P^x writes exactly as (7) with K and Σ two operators of finite rank.[7] The prior covariance operator is specified as $\Omega_0\sigma^2$ so that the Bayes estimator of the IV regression does not depend on the unknown parameter σ^2:

$$\begin{aligned}\hat{x}_\alpha &= \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}\Omega_0^{1/2}\hat{\mathbf{E}}(\hat{\mathbf{E}}(Y-x_0|W)|Z) + x_0\\ &= (\alpha\Omega_0^{-1} + \hat{T}^*\hat{T})^{-1}\hat{T}^*(\hat{\mathbf{E}}(Y-x_0|W)|Z) + x_0,\end{aligned}$$

where $B^*B = \Omega_0^{1/2}\hat{T}^*\hat{T}\Omega_0^{1/2}$, and $\hat{T}\cdot = \hat{\mathbf{E}}(\cdot|W)$ and $\hat{T}^*\cdot = \hat{\mathbf{E}}(\cdot|Z)$ are two operators on $L^2[0,1]$. For $x_0 = 0$ this is, in the framework of Darolles et al. (2011) or Florens, Johannes, and Van Bellegem (2011), the Tikhonov estimator in the Hilbert scale generated by L^{2s}.

Discussion of the Assumptions. In this example, Assumptions A.1 and A.2 need to hold both in finite sample and for $n\to\infty$. On the contrary, A.3 only has to hold for $n\to\infty$. We start by discussing the validity of Assumption A.1. In large sample, the operator K converges to $\Lambda M_f\mathbf{E}(\cdot|w)$ and it is trivial to verify that $\mathcal{D}(\Sigma^{-1/2}) = \mathcal{R}(\Lambda M_f) \supset \mathcal{R}(\Lambda M_f\mathbf{E}(\cdot|w))$. In finite sample, the same holds with M_f and $\mathbf{E}(\cdot|w)$ replaced by their empirical counterparts.

Next, Assumption A.2 is trivially satisfied in finite sample since $\mathcal{R}(\hat{\Sigma})$ $(\equiv \mathcal{D}(\hat{\Sigma}^{-1}))$ and $\mathcal{R}(K\Omega_0^{\frac{1}{2}})$ have finite ranks. In large sample, Assumption A.2

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

is satisfied if and only if $||\Sigma^{-1}\Lambda \lim_{\delta\to 0}\tilde{K}_\delta \Omega_0^{\frac{1}{2}}||^2 < \infty$ (when the conditions for the application of the Dominated Convergence Theorem hold). One way to satisfy this condition is to construct $\Omega_0 = T^*\Lambda^*\Lambda T$ where $T : L^2[0,1] \to L^2[0,1]$ is some trace class integral operator.

Finally, because Assumption A.3 has only to hold for $n \to \infty$, we have to consider the norm of the following operator: $\forall x \in \mathcal{X}$, $||\lim_{\delta\to 0}\Sigma^{-1/2}Kx|| = ||\mathbf{E}(x|W)||$.[8] This shows that, in the IV case, Assumption A.3 is a particular case of Assumptions 2.2 and 4.2 in Chen and Reiss (2011).

Rate of convergence. Next, we explicit the rate of Theorem 2 for the IV regression estimation. We assume that the true IV regression x_* that generates the data satisfies Assumption B. We take two different bandwidths for W and Z: h_w and h_z. In order to determine the rate of the MISE associated with $\hat{x}_\alpha$ the proof of Theorem 2 must be slightly modified to account for the fact that Σ and K depend on δ. Therefore, the rate of the MISE must incorporate the rate of convergence of these operators. The crucial issue in order to establish the rate of the MISE associated with $\hat{x}_\alpha$ is the rate of convergence of B^*B toward $\Omega_0^{\frac{1}{2}}\mathcal{T}^*\mathcal{T}\Omega_0^{\frac{1}{2}}$ where $\mathcal{T}\cdot = \mathbf{E}(\cdot|W)$ and $\mathcal{T}^*\cdot = \mathbf{E}(\cdot|Z)$. This rate is specified by Assumption (HS). Primitive conditions for this assumption and its derivation are provided in Appendix F in the Supplementary Appendix.

Assumption HS. There exists $\rho \geq 1$ such that:

(i) $\mathbf{E}||\Omega_0^{1/2}\hat{\mathcal{T}}^*\hat{f}_W^{-1}\Lambda^{-1}U||^2 = \mathcal{O}\left(n^{-1} + h_z^{2\rho}\right)$;

(ii) $\mathbf{E}||\hat{\mathcal{T}} - \mathcal{T}||^2 = \mathcal{O}\left(\frac{1}{nh_wh_z} + (\max\{h_w, h_z\})^{2\rho}\right)$;

(iii) $\mathbf{E}||\Omega_0^{\frac{1}{2}}(\hat{\mathcal{T}}^* - \mathcal{T}^*)||^2 = \mathcal{O}\left(\frac{1}{nh_w} + h_w^{2\rho}\right)$.

The next corollary to Theorem 2 gives the rate of the MISE of $\hat{x}_\alpha$ under this assumption.

COROLLARY 1. *Let us consider the observational model $y^\delta = Kx_* + U$, with y^δ and K defined as in (22) and x_* being the true value of x. Under Assumptions A.1–A.3, B, and HS:*

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} \mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2$$
$$= \mathcal{O}\left(\left(\alpha^{\frac{\tilde{\beta}}{a+s}} + n^{-1}\alpha^{-\frac{a+\gamma(a+s)}{a+s}}\right) + \alpha^{-2}\left(\frac{1}{nh_w} + h_w^{2\rho}\right)\alpha^{\frac{\tilde{\beta}}{a+s}}\right.$$
$$\left. + \alpha^{-1}\left(\frac{1}{nh_wh_z} + (\max\{h_w, h_z\})^{2\rho}\right)\alpha^{\frac{\tilde{\beta}}{a+s}} + \alpha^{-2}\left(\frac{1}{n} + h_z^{2\rho}\right)\right)$$

with $\tilde{\beta} = \min(\beta, a+2s)$ and $\gamma \in [0,1]$ such that $tr(L^{-2\gamma(a+s)}) < \infty$.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

In this example we have to set three tuning parameters: the bandwidths h_w, h_z, and α. Suppose that we set $h_w = h_z \propto n^{-1/(2\rho)}$ and $\alpha \propto n^{-\frac{a+s}{\tilde{\beta}+2(a+s)}}$. In this case, the rate of convergence is $n^{-\frac{\tilde{\beta}}{\tilde{\beta}+2(a+s)}}$ which is the same rate as in Darolles et al. (2011) if $(a+s)=1$. The leading terms are the first and the last one. The third and fifth terms are negligible if $\tilde{\beta} > (a+s)\max\{2/(4\rho-1), (2-\rho)/(2\rho-1)\}$ while the other terms are always negligible.

As a final remark, we would like to stress that the approach proposed here to estimate an IV regression is an alternative to the approach proposed in Florens and Simoni (2012a). The current approach is more robust because it only relies on asymptotic normality and does not require normality of the reduced-form error term. Moreover, it can be easily implemented by using our adaptive EB procedure for choosing α.

7.2. Example 2: Functional Linear Regression Estimation (Continued)

Let us consider model (3) and the associated integral equation $\mathbf{E}(\xi Z(t)) = \int_0^1 x(s)Cov(Z(s), Z(t))ds$, for $t \in [0,1]$. If we dispose of i.i.d. data $(\xi_1, Z_1), \ldots, (\xi_n, Z_n)$ the unknown moments in this equation can be estimated. Thus, x is solution of (6) with $y^\delta := \frac{1}{n}\sum_i \xi_i Z_i(t)$, $U = \frac{1}{n}\sum_i \varepsilon_i Z_i(t)$, $\delta = \frac{1}{n}$ and $\forall \varphi \in L^2[0,1]$, $K\varphi := \frac{1}{n}\sum_i \langle Z_i, \varphi\rangle Z_i(t)$. The operator $K : L^2[0,1] \to L^2[0,1]$ is self-adjoint, i.e., $K = K^*$. Moreover, conditional on Z, the error term U is exactly a Gaussian process with covariance operator defined as $\forall \varphi \in L^2[0,1]$, $\delta\Sigma\varphi = \delta\tau^2 \int_0^1 \varphi(s)Cov(Z(s), Z(t))ds$, which is trace-class since $\mathbf{E}||Z_i||^2 < \infty$. Without loss of generality we suppose hereinafter that $\tau^2 = 1$.

Suppose that the spectrum of Σ declines at the rate j^{-2a}, with $a \geq 0$, as in Assumption C (for instance, Σ could be the covariance operator of a Brownian motion). Construct the operator L as in (13) with $\{\psi_j\}_{j\geq 1}$ being the eigenfunctions of Σ. Thus, $\Omega_0 = \Sigma^{s/a}$. Because Σ is unknown, it is replaced by K in the posterior mean estimator: $\hat{x}_\alpha = K^{s/(2a)}(\alpha I + K^{s/a+1})^{-1}K^{s/(2a)}(y^\delta - Kx_0) + x_0$, for which the following corollary holds.

COROLLARY 2. *Let $\Sigma : L^2[0,1] \to L^2[0,1]$ have eigenvalues $\{\lambda^2_{j\Sigma}\}$ that satisfy $\underline{a}\, j^{-2a} \leq \lambda^2_{j\Sigma} \leq \overline{a}\, j^{-2a}$, for $\underline{a}, \overline{a} > 0$ and $a > 0$. Assume that $\mathbf{E}||Z_i||^4 < \infty$. Then, under Assumption B' with $b_0 > \max\{a, s\}$, $s > \frac{1}{2}$ and if $\alpha \asymp \alpha_* = n^{-\frac{s+a}{b_0+a}}$, we have*

$$\mathbf{E}_{x_*}||\hat{x}_{\alpha_*} - x_*||^2 = \mathcal{O}\left(n^{-\frac{2\tilde{b}_0-1}{2(\tilde{b}_0+a)}}\right), \quad \text{where} \quad \tilde{b}_0 = \min\left(b_0, 2a+2s+\frac{1}{2}\right).$$

The assumptions and the rate for $\tilde{b}_0 = b_0$ in the corollary are the same as in Hall and Horowitz (2007).

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

8. NUMERICAL IMPLEMENTATION

In this section we first explain the numerical implementation of our estimator $\hat{x}_\alpha$ and then show the results of two simulations. Let M be the number of discretization points and $(s_1, \ldots, s_M)$ (respectively $w_1, \ldots, w_M$) be (equidistant) discretization points for functions in $\mathcal{X}$ (respectively in $\mathcal{Y}$) and $\nu_\mathcal{X}$, $\nu_\mathcal{Y}$ be the measures associated with $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\omega(s,t)$ be the prior covariance kernel, $\sigma(w,v)$ be the covariance kernel that defines Σ, and $k(v,s)$ be the kernel of the operator K. The easiest way to construct $\hat{x}_\alpha$ consists in using the closed-form given in (11) and rewritten as

$$\hat{x}_\alpha = (\alpha\Omega_0^{-1} + K^*\Sigma^{-1}K)^{-1}K^*\Sigma^{-1}(y^\delta - Kx_0) + x_0. \tag{23}$$

The construction of the estimator proceeds with the following steps:

1. Replace the operators in (23) by the corresponding $M \times M$ discretization matrices:
   - Ω_0 is replaced by the matrix $\Omega_{0,d}$ with generic element $\frac{1}{M}(\omega(s_i, s_j)\nu_\mathcal{X}(s_j))_{i,j=1,\ldots,M}$;
   - Σ is replaced by the matrix Σ_d with generic element $\frac{1}{M}(\sigma(w_i, w_j)\nu_\mathcal{Y}(w_j))_{i,j=1,\ldots,M}$;
   - K (respectively K^*) is replaced by the matrix K_d (respectively K_d^*) with generic element

$$\frac{1}{M}\left(k(w_i, s_j)\nu_\mathcal{X}(s_j)\right)_{i,j=1,\ldots,M} \text{ (respectively } \frac{1}{M}\left(k(w_j, s_i)\nu_\mathcal{Y}(w_j)\right)_{i,j=1,\ldots,M}).$$

2. Discretize the prior mean function x_0 to obtain an M-vector: $x_{0,d} = (x_0(s_1), \ldots, x_0(s_M))^T$.
3. Discretize the (functional estimator) y^δ to obtain an M-vector: $y_d^\delta = (y^\delta(w_1), \ldots, y^\delta(w_M))^T$.
4. Compute the estimator $\hat{x}_\alpha$ at the M discretization points $(s_1, \ldots, s_M)$ as:

$$\hat{x}_\alpha = \left(\alpha\Omega_{0,d}^{-1} + K_d^*\Sigma_d^{-1}K_d\right)^{-1}K_d^*\Sigma_d^{-1}(y_d^\delta - K_d x_{0,d}) + x_{0,d}.$$

In order to implement the EB procedure the steps are the following.

5. By using the matrices computed in 1, compute the singular value decomposition of the matrix $\Sigma_d^{-1/2}K_d\Omega_{0,d}^{1/2}$. These singular values and functions correspond to the discretized version of λ_j and φ_j in expression (19) and we denote them by $\lambda_{j,d}$ and $\varphi_{j,d}$, respectively.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

6. Fix J large enough (e.g., $J = 200$) and compute the discretized version of (19) as:

$$\bar{S}_d(\alpha, y^\delta) = \frac{1}{2}\sum_{j=1}^{J}\left[\log\left(\frac{\alpha}{\alpha+\lambda_{j,d}^2}\right) + \frac{\lambda_{j,d}^2}{\alpha+\lambda_{j,d}^2}\frac{[\varphi_{j,d}^T \Sigma_d^{-1/2}(y_d^\delta - K_d x_{0,d})]^2}{\delta}\right] + \log \nu(\alpha) \tag{24}$$

for α taking values over a grid of points in the interval $[0, 1]$.

7. Select the value of α that maximizes (24).

8.1. Instrumental Variable Regression Estimation

This section shows implementation and finite sample properties of our estimator for the IV regression in Example 1. We simulate $n = 1,000$ observations from the following model, which involves only one endogenous covariate Z and two instrumental variables $W = (W_1, W_2)$,

$$W_i = \begin{pmatrix} W_{1,i} \\ W_{2,i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right)$$

$$\begin{aligned} \upsilon_i &\sim \mathcal{N}(0, \sigma_\upsilon^2), \qquad Z_i = 0.1W_{i,1} + 0.1W_{i,2} + \upsilon_i \\ \varepsilon_i &\sim \mathcal{N}(0, (0.4)^2), \qquad \eta_i = -0.5\upsilon_i + \varepsilon_i \\ Y_i &= x_*(Z_i) + \eta_i \end{aligned}$$

for $i = 1, \ldots, n$. Endogeneity is caused by correlation between η_i and the error term υ_i affecting the covariates. The true x_* is the parabola $x_*(Z) = Z^2$. In all the simulations we fix $\sigma_\upsilon = 0.27$. We do not transform the data to the interval $[0, 1]$ and the spaces of reference are $\mathcal{X} = \mathcal{Y} = L^2(Z)$, where $L^2(Z)$ denotes the space of square integrable functions of Z with respect to its marginal distribution. Moreover, the function $a(w, \upsilon)$ is chosen such that $K : L^2(Z) \to L^2(Z)$ is the estimated double conditional expectation operator, that is, $\forall \varphi \in L^2(Z)$, $K\varphi = \hat{\mathbf{E}}(\hat{\mathbf{E}}(\varphi|W)|Z)$, and the functional observation y^δ takes the form $y^\delta(Z) := \hat{\mathbf{E}}(\hat{\mathbf{E}}(Y|W)|Z)$. Here, $\hat{\mathbf{E}}$ denotes the estimated expectation that we compute through a Gaussian kernel smoothing estimator of the joint density of (Y, Z, W). The bandwidth for Z is set equal to $\sqrt{\widehat{Var}(Z)}n^{-1/5}$ and similarly the bandwidths for Y and W.

The sampling covariance operator is $\Sigma\varphi = \sigma^2\mathbf{E}(\mathbf{E}(\varphi|W)|Z)$, $\forall \varphi \in L^2(Z)$ (under the assumption $\mathbf{E}[(Y - x_*(Z))^2|W] = \sigma^2$) and is replaced by its empirical counterpart. We specify the prior covariance operator as $\Omega_0 = \omega_0\sigma^2K^s$ for $s > 0$ and ω_0 a fixed parameter. The advantage of using this specification for Ω_0 is that: (i) it simplifies numerical computations of the estimator $\hat{x}_\alpha$ and (ii) the conditional variance σ^2 in Σ and Ω_0 simplifies and does not need to be estimated.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

We perform simulations for two specifications of x_0, ω_0, and s (where x_0 denotes the prior mean function): Figure 1 refers to $x_0(Z) = 0.95Z^2 + 0.25$, $\omega_0 = 1$, and $s = 2$ while Figure 2 refers to $x_0(Z) = 0$, $\omega_0 = 2$, and $s = 15$. These specifications allow us to analyze the finite sample performance of our method when the prior mean has the correct curvature and when it does not. We first run simulations with a value of α chosen to produce a reasonably good estimation (we fix $\alpha = 0.9$ to obtain Figures 1A, B and 2A, B) and in a second simulation we use the α selected through our EB method. This allows us to compare the sensitivity of $\hat{x}_\alpha$ to the choice of α.

Figures 1A and 2A represent: the n observed Y's (magenta asterisks), the corresponding y^δ (dotted blue line) obtained from the observed sample of (Y, Z, W), the true x_* (black solid line), the nonparametric estimation of the regression function $\mathbf{E}(Y|Z)$ (yellow dashed line), and our posterior mean estimator $\hat{x}_\alpha$ (dotted-dashed red line). We show the estimator of $\mathbf{E}(Y|Z)$ with the purpose of making clear the bias due to endogeneity. Figures 1B and 2B represent: the prior mean function x_0 (magenta dashed line), the observed function y^δ (dotted blue line)

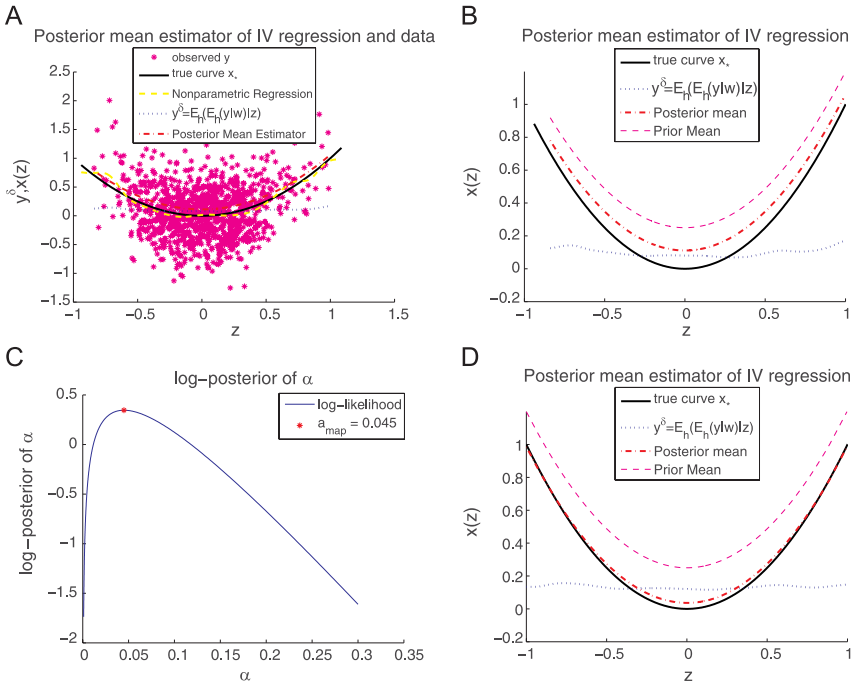


FIGURE 1. Posterior mean estimator for smooth x_*. Graph for: $x_0(Z) = 0.95Z^2 + 0.25$, $\omega_0 = 1$, $s = 2$. (a) Data and posterior mean estimator for $\alpha = 0.9$. (b) Posterior mean estimator for $\alpha = 0.9$. (c) α choice, $\hat{\alpha} = \alpha_{\text{map}}$. (d) Posterior mean estimator for $\alpha = 0.045$.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

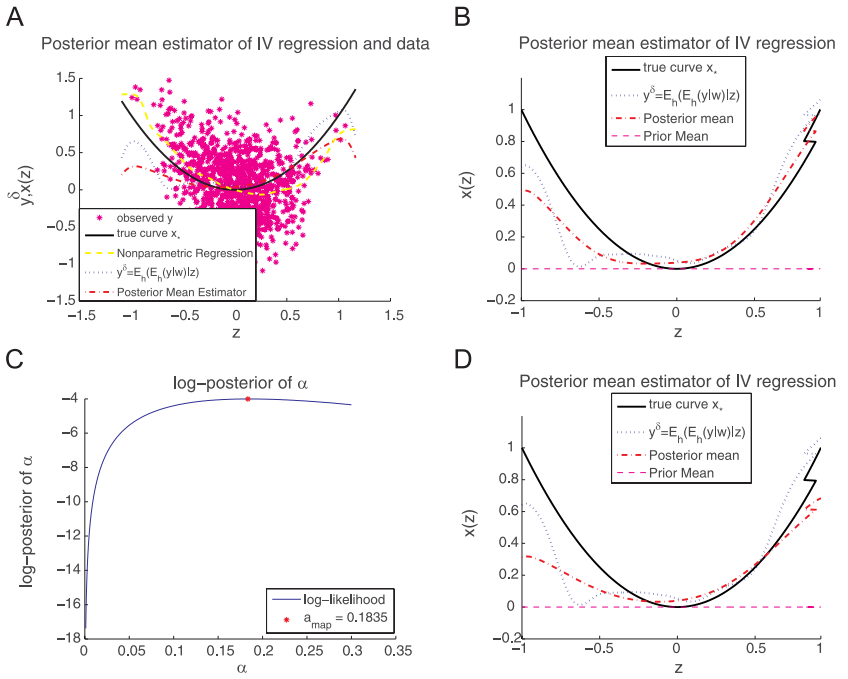


FIGURE 2. Posterior mean estimator for smooth x_*. Graph for: $x_0(Z) = 0$, $\omega_0 = 2$, $s = 15$. (a) Data and posterior mean estimator for $\alpha = 0.9$. (b) Posterior mean estimator for $\alpha = 0.9$. (c) α choice, $\hat{\alpha} = \alpha_{\text{map}}$. (d) Posterior mean estimator for $\alpha = 0.1835$.

obtained from the observed sample of (Y, Z, W), the true x_* (black solid line), and our posterior mean estimator $\hat{x}_\alpha$ (dotted-dashed red line). It is evident from Figure 2B that even with a prior mean x_0 far from x_*, the IV regression is well estimated. Since the sample size is relatively large, this confirms our theoretical results of posterior consistency: the prior is washed away when n is large.

Figures 1C and 2C draw the log-posterior $\log\left[\frac{dP_\alpha}{dP_0}\nu(\alpha)\right]$ against α and show the value of the maximum a posteriori $\hat{\alpha}$. We specify an exponential prior for α: $\nu(\alpha) = 11e^{-11\alpha}$, $\forall \alpha \geq 0$. We point out that the different shape of $\log\left[\frac{dP_\alpha}{dP_0}\nu(\alpha)\right]$ in the two graphs is only due to the different specification of the prior for x (which is used to integrate out x). Finally, Figures 1D and 2D represent our EB-posterior mean estimator $\hat{x}_{\hat{\alpha}}$ (dotted-dashed red line)—obtained by using the $\hat{\alpha}$ selected with the EB procedure—together with the prior mean function x_0 (magenta dashed line), the observed function y^δ (dotted blue line), and the true x_* (black solid line). Without changing the sample size, with respect to the one used to obtain Figures 1A and 2A, the estimation improves greatly when α is selected with the aid of our EB procedure.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

8.2. Geometric Spectrum Case

In this simulation we assume $\mathcal{X} = \mathcal{Y} = L^2(\mathbb{R})$ with respect to the measure $e^{-u^2/2}$ so that the operator K is self-adjoint. We use the Hermite polynomials as common eigenbasis for the operators K, Σ, and L. The Hermite polynomials $\{H_j\}_{j\geq 0}$ form an orthogonal basis of $L^2(\mathbb{R})$ with respect to the measure $e^{-u^2/2}$. The first few Hermite polynomials are $\{1, u, (u^2-1), (u^3-3u), \ldots\}$ and an important property of these polynomials is that they are orthogonal with respect to $e^{-u^2/2}$: $\int_{\mathbb{R}} H_l(u)H_j(u)e^{-u^2/2}du = \sqrt{n}(n!)\delta_{lj}$, where δ_{lj} is equal to 1 if $l = j$ and 0 otherwise. Moreover, they satisfy the recursion $H_{j+1}(u) = uH_j(u) - jH_{j-1}(u)$ which is used in our simulation. We fix: $\delta = 1/n$ and the simulation design is:

$$\Sigma\cdot = \tau\sum_{j=0}^{\infty}\frac{j^{-c_0}}{\sqrt{2\pi}(n!)}\langle H_j,\cdot\rangle H_j, \qquad K\cdot = \sum_{j=0}^{\infty}\frac{j^{-a_0}}{\sqrt{2\pi}(n!)}\langle H_j,\cdot\rangle H_j$$

$$\Omega_0\cdot = \omega_0\sum_{j=0}^{\infty}\frac{j^{-2s}}{\sqrt{2\pi}(n!)}\langle H_j,\cdot\rangle H_j, \qquad y^\delta = Kx_* + U$$

with $a_0 = 1$, $c_0 = 1.2$, $s = 1$, $x_*(u) = u^2$, and $U = \frac{1}{\sqrt{n}}\mathcal{N}(0,\Sigma)$. Moreover, we fix: $\tau = 10$ and $\omega_0 = 20$. The inner product is approximated by discretizing the integral $\int_{\mathbb{R}} H_j(u)\cdot e^{-u^2/2}du$ with 1,000 discretization points uniformly generated between -3 and 3. The infinite sums are truncated at $j = 200$. The simulation results did not change when we increased the value of the truncation.

We first perform simulations for a fixed value of α (we have fixed $\alpha = 0.9$ to obtain Figure 3A) and in a second simulation we use the α selected through our EB method.

Figure 3A represents: the prior mean function x_0 (magenta dashed line), the observed function y^δ (dotted blue line), the true x_* (black solid line), and our posterior mean estimator $\hat{x}_\alpha$ (dotted-dashed red line).

Figure 3B draws the log-posterior $\log\left[\frac{dP_\alpha}{dP_0}\nu(\alpha)\right]$ against α and shows the value of the maximum a posteriori $\hat{\alpha}$. We specify a Gamma prior for α: $\nu(\alpha) \propto \alpha^{11}e^{-10\alpha}$, $\forall\alpha \geq 0$. Finally, Figure 3C represents our EB-posterior mean estimator $\hat{x}_{\hat{\alpha}}$ (dotted-dashed red line)—obtained by using the $\hat{\alpha}$ selected with the EB procedure—together with the prior mean function x_0 (magenta dashed line), the observed function y^δ (dotted blue line), and the true x_* (black solid line). By comparing the estimators in Figure 3C and 3A, it is evident the improvement in the estimation due to the use of our EB method. This shows the importance of having an adaptive data-driven method for choosing α. The tails of the curve x_* are not very well estimated. This is due to the fact that the prior mean function is very far from the true x_*, that is, we have very little prior information. By letting δ decrease, this effect will diminish according to our consistency results.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

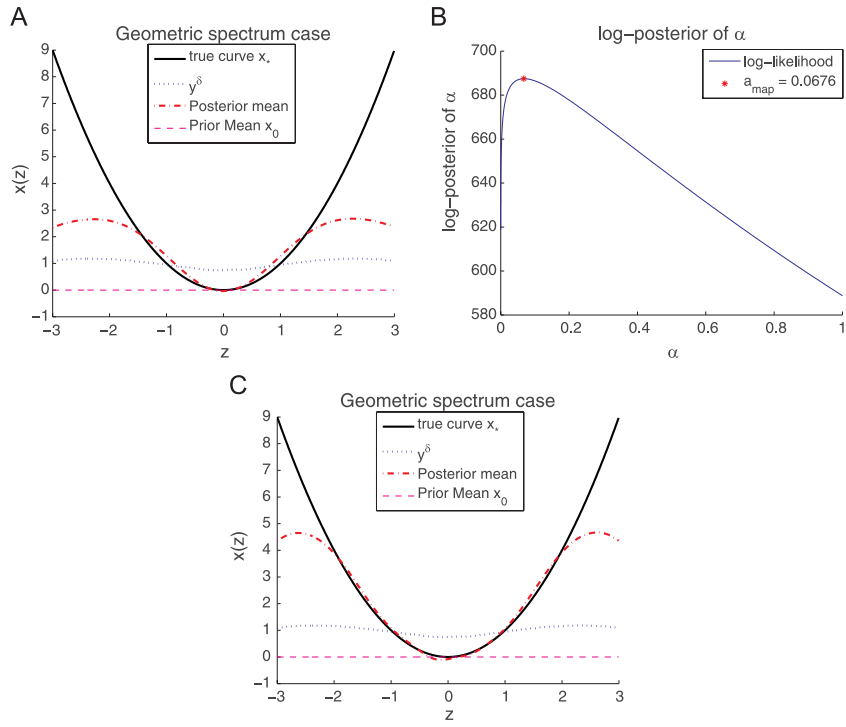


FIGURE 3. Posterior mean estimator for smooth x_*. Graph for: $x_0(Z) = 0$, $\omega_0 = 20$, $s = 1$, and $\tau = 10$. (a) Data and posterior mean estimator for $\alpha = 0.9$. (b) α choice, $\hat{\alpha} = \alpha_{\text{map}}$. (c) Posterior mean estimator for $\alpha = 0.0676$.

9. CONCLUSION

This paper develops a Bayesian approach for nonparametric estimation of functional parameters in econometric models that are characterized as the solution of an IP. We consider a conjugate-Gaussian setting where the "likelihood" is only required to be asymptotically Gaussian. For "likelihood" we mean the sampling distribution of a functional transformation of an n-sample.

We first propose the posterior mean as a point estimator and give sufficient conditions that allow us to show that the posterior mean: (i) has a closed-form, (ii) is easy to implement in practice, (iii) has a pure Bayesian interpretation, and (iv) is consistent in a frequentist sense. Then, we contribute to the literature on IPs by proposing an adaptive data-driven method to select the regularization parameter which enters our estimation procedure. Our data-driven method is based on an Empirical Bayes approach.

We prove that the posterior mean estimator is, under certain conditions, equal to the Tikhonov estimator which has been largely used in frequentist nonparametric

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

estimation based on IPs. Because of this equality, our EB data-driven method for selecting the regularization parameter is valid even for the implementation of the frequentist Tikhonov estimator. Thus, our approach makes an important contribution to the theory of IPs in econometrics in general.

Before concluding, two remarks are in order. First, our procedure is strictly tied to the linearity of the IP. When model (1) is nonlinear in x the posterior mean estimator does not have a closed form anymore. Thus, the techniques used to prove our results should be modified. The second remark concerns the use of a prior for x different from the Gaussian prior. Posterior consistency should still hold when other priors are used (under some conditions) but this type of result would require a completely different strategy of proof. The EB-data-driven method based on maximizing the marginal posterior distribution should also hold when a non-Gaussian prior is used. However, the function to be maximized would no longer have a closed form and this would complicate the implementation. Proving adaptivity in this case would also require a different strategy of proof. Proving all these results with a non-Gaussian prior is far beyond the scope of this paper and we leave it for future research.

NOTES

1. Notice that in general the posterior mean in a conjugate-Gaussian problem stated in infinite-dimensional Hilbert spaces cannot be equal to the Tikhonov solution of (1). This is due to the particular structure of the covariance operator of the error term U^{δ} and it will be detailed in Section 3.

2. A similar idea of linking the prior distribution to the sampling model was proposed by Zellner (1986) for linear regression models for which he constructed a class of priors called g-prior.

3. In principle, this assumption could be satisfied even by irregular x_* if we were able to decompose x_* in the sum of a regular part plus an irregular part and to choose x_0 such that it takes all the irregularity of x_*. This is clearly infeasible in practice as x_* is unknown. On the contrary, we could choose a very smooth function x_0 so that Assumption B would be less demanding about the regularity of x_*.

4. In many applications $\mathcal{X}$ and $\mathcal{Y}$ are L^2 spaces and L^2 spaces are Polish if they are defined on a separable metric space.

5. For simplicity of exposition we limit this analysis to the case where Assumption C holds. It is possible to extend the result of Theorem 6 to the general case at the price of complicate very much the proof and the notation. For this reason we do not treat the general case here.

6. The fact that weak convergence toward a Gaussian process cannot be obtained without transforming the model by Λ can be explained in the following way (see Ruymgaart, 1998 for a detailed explanation). It is possible to show that the covariance operator $\tilde{\Sigma}_h$ of $\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(y_i - \int_0^1 x\,\frac{K_{Z,h}(z_i - z)}{h}dz\right)\frac{K_{W,h}(w_i - w)}{h}$ satisfies

$$\langle \phi_1, \tilde{\Sigma}_h \phi_2 \rangle \quad \longrightarrow \quad \langle \phi_1, \tilde{\Sigma} \phi_2 \rangle, \quad \text{as } h \to 0, \quad \forall \phi_1, \phi_2 \in L^2[0,1]$$

where $\tilde{\Sigma}\phi_2 = \sigma^2 [f_W(v)]^2 \phi_2(v) = \sigma^2 M_f^2 \phi_2(v)$ under the assumption $\mathbf{E}[(Y - x(Z))^2 | W] = \sigma^2 < \infty$. Unfortunately, $\tilde{\Sigma}$ has not finite trace and so it is incompatible with the covariance structure of a Gaussian limiting probability measure. Indeed, there are no scaling factors n^{-r}, for $0 < r < 1$, such that $n^{-r}\sum_{i=1}^{n}\left(y_i - \int_0^1 x\,\frac{K_{Z,h}(z_i - z)}{h}dz\right) \times \frac{K_{W,h}(w_i - w)}{h}$ converges weakly in $L^2[0,1]$ to a Gaussian distribution (unless this distribution is degenerate at the zero function) (see Ruymgaart, 1998). However,

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

if we choose $a(w,v)$ appropriately so that $\Lambda^*\Lambda$ has finite trace, then $\sqrt{n}\left(y^\delta - Kx\right) \Rightarrow \mathcal{N}(0, \Lambda\tilde{\Sigma}\Lambda^*)$ in $L^2[0,1]$ as $n \to \infty$, where '$\Rightarrow$' denotes weak convergence.

7. To be precise we should index the two operators K and Σ by δ and write K_δ and Σ_δ. In fact these two operators change with δ. However, we prefer to keep a uniform notation and do not adopt this notation.

8. This is because the operator $\Sigma^{1/2}$ may be equivalently defined in two ways: (1) as a self-adjoint operator, that is $\Sigma^{1/2} = (\Sigma^{1/2})^* = (\Lambda M_f^2 \Lambda^*)^{1/2}$, so that $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$ or (2) as $\Sigma^{1/2} = \Lambda M_f$ so that $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^*$ where $(\Sigma^{1/2})^* = M_f\Lambda^*$. By using the second definition, we obtain that $\Sigma^{-1/2}K = (\Lambda M_f)^{-1}\Lambda\tilde{K} = f_W^{-1}\Lambda^{-1}\Lambda\tilde{K} = f_W^{-1}\tilde{K}$.

REFERENCES

Agapiou, S., S. Larsson, & A.M. Stuart (2013) Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Processes and Applications* 123, 3828–3860.

Ai, C. & X. Chen (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71, 1795–1843.

Belitser, E. & S. Ghosal (2003) Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Annals of Statistics* 31, 536–559.

Bissantz, N., T. Hohage, A. Munk, & F. Ruymgaart (2007) Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis* 45, 2610–2636.

Carrasco, M. & J.P. Florens (2000) Generalization of GMM to a continuum of moment conditions. *Econometric Theory* 16, 797–834.

Carrasco, M., J.P. Florens, & E. Renault (2005) Estimation based on spectral decomposition and regularization. In J.J. Heckman & E. Leamer (eds.), *Handbook of Econometrics*, vol. 6, pp. 5633–5751. Elsevier.

Cavalier, L. (2011) Inverse problems in statistics. In P. Alquier, E. Gautier, & G. Stoltz (eds.), *Inverse Problems and High-dimensional Estimation*, pp. 3–96. Springer-Verlag.

Cavalier, L. & A.B. Tsybakov (2002) Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields* 123, 323–354.

Chen, X. & D. Pouzo (2012) Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* 80, 277–321.

Chen, X. & M. Reiss (2011) On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory* 27, 497–521.

Daouia, A., J.-P. Florens, & L. Simar (2009) Regularization of nonparametric frontier estimators. *Journal of Econometrics* 168, 285–299.

Darolles, S., Y. Fan, J.P. Florens, & E. Renault (2011) Nonparametric instrumental regression. *Econometrica* 79, 1541–1565.

Diaconis, P.W. & D. Freedman (1986) On the consistency of Bayes estimates. *Annals of Statistics* 14, 1–26.

Engl, H.W., M. Hanke, & A. Neubauer (2000) *Regularization of Inverse Problems*. Kluwer Academic.

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

Florens, J.-P., J. Johannes, & S. Van Bellegem (2011) Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory* 27, 472–496.

Florens, J.P. & A. Simoni (2010) Regularizing Priors for Linear Inverse Problems. IDEI Working paper, n. 621.

Florens, J.P. & A. Simoni (2012a) Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior. *Journal of Econometrics* 170, 458–475.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

Florens, J.P. & A. Simoni (2012b) Regularized posteriors in linear ill-posed inverse problems. *Scandinavian Journal of Statistics* 39, 214–235.

Ghosh, J.K. & R.V. Ramamoorthi (2003) *Bayesian Nonparametrics*. Springer Series in Statistics. Springer.

Hall, P. & J. Horowitz (2005) Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics* 33, 2904–2929.

Hall, P. & J.L. Horowitz (2007) Methodology and convergence rates for functional linear regression. *Annals of Statistics* 35, 70–91.

Helin, T. (2009) On infinite-dimensional hierarchical probability models in statistical inverse problems. *Inverse Problems and Imaging* 3, 567–597.

Hoderlein, S., L. Nesheim, & A. Simoni (2013) Semiparametric Estimation of Random Coefficients in Structural Economic Models. CEMMAP Working paper, CWP09/12.

Hofinger, A. & H.K. Pikkarainen (2007) Convergence rate for the Bayesian approach to linear inverse problems. *Inverse Problems* 23, 2469–2484.

Hofinger, A. & H.K. Pikkarainen (2009) Convergence rate for linear inverse problems in the presence of an additive normal noise. *Stochastic Analysis and Applications* 27, 240–257.

Horowitz, J. (2014) Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularisation parameter. *Journal of Econometrics* 180, 158–173.

Johannes, J., R. Schenk, & A. Simoni (2014) Adaptive Bayesian estimation in Gaussian sequence space models. In E.G. Bongiorno, A. Goia, E. Salinelli, & P. Vieu (eds.), *Contributions in Infinite-Dimensional Statistics and Related Topics*, pp. 167–172. Società Editrice Esculapio.

Kaipio, J. & E. Somersalo (2004) *Statistical and Computational Inverse Problems.* Applied Mathematical Series, 160. Springer.

Kato, T. (1995) *Perturbation Theory for Linear Operators*. Springer.

Knapik, B.T., A.W. Van der Vaart, & J.H. Van Zanten (2011) Bayesian inverse problems. *Annals of Statistics* 39, 2626–2657.

Krein, S.G. & J.I. Petunin (1966) Scales of Banach spaces. *Russian Mathematical Surveys* 21, 85–160.

Kress, R. (1999) *Linear Integral Equation*. Springer-Verlag.

Kuo, H.H. (1975) *Gaussian Measures in Banach Spaces*. Springer-Verlag.

Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag.

Liao, Y. & W. Jiang (2011) Posterior consistency of nonparametric conditional moment restricted models. *Annals of Statistics* 39, 3003–3031.

Linton, O. & E. Mammen (2005) Estimating semiparametric ARCH(∞) models by kernel smoothing methods. *Econometrica* 73, 771–836.

Luschgy, H. (1995) Linear estimators and radonifying operators. *Theory of Probability and Its Applications* 40, 167–175.

Mandelbaum, A. (1984) Linear estimators and measurable linear transformations on a Hilbert space. *Z. Wahrcheinlichkeitstheorie* 3, 385–398.

Natterer, F. (1984) Error bounds for Tikhonov regularization in Hilbert scale. *Applicable Analysis* 18, 29–37.

Newey, W.K. & J.L. Powell (2003) Instrumental variable estimation of nonparametric models. *Econometrica* 71, 1565–1578.

Ruymgaart, F.H. (1998) A note on weak convergence of density in Hilbert spaces. *Statistics* 30, 331–343.

Van der Vaart, A.W. & J.H. Van Zanten (2009) Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Annals of Statistics* 37, 2655–2675.

Van Rooij, A.C.M. & F.H. Ruymgaart (1999) On inverse estimation. In S. Ghosh (ed.), *Asymptotic, Nonparametric and Time Series*, pp. 579–613. Dekker.

Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distribution. In P.K. Goel & A. Zellner (eds.), *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, vol. 6, pp. 233–243. North Holland.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

APPENDIX

In the proofs for Section 4 we use the notation $(\lambda_j, \varphi_j, \psi_j)_j$ to denote the singular value decomposition of $B = \Sigma^{-1/2} K \Omega_0^{1/2}$, that is, $B\psi_j = \lambda_j \varphi_j$ and $B^* \varphi_j = \lambda_j \psi_j$ where φ_j and ψ_j are of norm 1, $\forall j$. In all the proofs we use the notation $I(\mathcal{A})$ or $I_{\mathcal{A}}$ to denote the indicator function of an event $\mathcal{A}$ and the notation $=^d$ to mean "equal in distribution". In order to prove several results we make use of Corollary 8.22 in Engl et al. (2000). We give here a simplified version of this corollary adapted to our framework and we refer to Engl et al. (2000) for the proof of it.

COROLLARY 3. *Let $(\mathcal{X}_t)_{t>0}$ be a Hilbert scale generated by L and let $\Sigma^{-1/2} K : \mathcal{X} \to \mathcal{Y}$ be a bounded operator satisfying Assumption A.3, $\forall x \in \mathcal{X}$ and for some $a > 0$. Then, for $B = \Sigma^{-1/2} K L^{-s}$, $s \geq 0$ and $|\nu| \leq 1$*

$$\underline{c}(\nu) ||L^{-\nu(a+s)} x|| \leq ||(B^*B)^{\frac{\nu}{2}} x|| \leq \overline{c}(\nu) ||L^{-\nu(a+s)} x|| \tag{A.1}$$

holds on $\mathcal{D}((B^*B)^{\frac{\nu}{2}})$ with $\underline{c}(\nu) = \min(\underline{m}^\nu, \overline{m}^\nu)$ and $\overline{c}(\nu) = \max(\underline{m}^\nu, \overline{m}^\nu)$. Moreover, $\mathcal{R}((B^*B)^{\frac{\nu}{2}}) = \mathcal{X}_{\nu(a+s)} \equiv \mathcal{D}(L^{\nu(a+s)})$, where $(B^*B)^{\frac{\nu}{2}}$ has to be replaced by its extension to $\mathcal{X}$ if $\nu < 0$.

A. PROOFS FOR SECTION 4

A.1. Proof of Theorem 1

(i) See the proof of Theorem 1(i) and (ii) in Florens and Simoni (2012b).

(ii) See Theorem 2 and Corollary 2 in Mandelbaum (1984) and their proofs in Sections 3.4 and 3.5, page 392.

(iii) The P_α-mlt A is defined as $A := \Omega_0 K^* (\alpha \Sigma + K \Omega_0 K^*)^{-1}$ on $\mathcal{R}((\delta \Sigma + \frac{\delta}{\alpha} K \Omega_0 K^*)^{\frac{1}{2}})$ (see Luschgy, 1995). Under A.2, the operator $\Omega_0 K^* (\alpha \Sigma + K \Omega_0 K^*)^{-1}$ can equivalently be rewritten as

$$\begin{aligned}
&\Omega_0^{\frac{1}{2}} \Omega_0^{\frac{1}{2}} K^* \Sigma^{-1/2} \left(\alpha I + \Sigma^{-1/2} K \Omega_0 K^* \Sigma^{-1/2} \right)^{-1} \Sigma^{-1/2} \\
&= \Omega_0^{\frac{1}{2}} B^* (\alpha I + B B^*)^{-1} \Sigma^{-1/2} \\
&= \Omega_0^{\frac{1}{2}} (\alpha I + B^* B)^{-1} \left(\Sigma^{-1/2} B \right)^* + \Omega_0^{\frac{1}{2}} [B^* (\alpha I + B B^*)^{-1} \\
&\quad - (\alpha I + B^* B)^{-1} B^*] \Sigma^{-1/2} \\
&= \Omega_0^{\frac{1}{2}} (\alpha I + B^* B)^{-1} \left(\Sigma^{-1/2} B \right)^*
\end{aligned} \tag{A.2}$$

since $[B^*(\alpha I + BB^*)^{-1} - (\alpha I + B^*B)^{-1} B^*]$ is equal to

$$(\alpha I + B^*B)^{-1} [(\alpha I + B^*B) B^* - B^* (\alpha I + BB^*)] (\alpha I + BB^*)^{-1}$$

which is zero. By using expression (A.2) for $\Omega_0 K^* (\alpha \Sigma + K \Omega_0 K^*)^{-1}$ we show that $\Omega_0 K^* (\alpha \Sigma + K \Omega_0 K^*)^{-1}$ is bounded and continuous on $\mathcal{Y}$. By Assumption A.2 the operators B and $\Sigma^{-1/2} B$ exist and are bounded. Under A.1 the operator B is compact because it is the product of a compact and a bounded operator.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

This implies that $B^*B : \mathcal{X} \to \mathcal{X}$ is compact. Because $\langle(\alpha I + B^*B)\phi, \phi\rangle \geq \alpha||\phi||^2$, $\forall \phi \in \mathcal{X}$ we conclude that $(\alpha I + B^*B)$ is injective for $\alpha > 0$. Then, from the Riesz Theorem 3.4 in Kress (1999) it follows that the inverse $(\alpha I + B^*B)^{-1}$ is bounded. Finally, since the product of bounded linear operators is bounded, $\Omega_0 K^*(\alpha\Sigma + K\Omega_0 K^*)^{-1}$ is bounded, i.e., there exists a constant C such that for all $\varphi \in \mathcal{Y}$: $||\Omega_0 K^*(\alpha\Sigma + K\Omega_0 K^*)^{-1}\varphi|| \leq C||\varphi||$. Since $A := \Omega_0 K^*(\alpha\Sigma + K\Omega_0 K^*)^{-1}$ is linear and bounded on $\mathcal{Y}$, it is continuous on $\mathcal{Y}$.

A.2. Proof of Theorem 2

The difference $(\hat{x}_\alpha - x_*)$ is rewritten as

$$\hat{x}_\alpha - x_* = -(I - AK)(x_* - x_0) + AU := \mathcal{C}_1 + \mathcal{C}_2, \tag{A.3}$$

where the expression of A is given in (10). We consider the MISE associated with $\hat{x}_\alpha$: $\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = ||\mathcal{C}_1||^2 + \mathbf{E}_{x_*}||\mathcal{C}_2||^2$ and we start by considering the first term. Our proof follows Natterer (1984). Since $0 < \frac{s}{a+s} < 1$ we can use the left part of inequality (A.1) in Corollary 3 with $\nu = \frac{s}{a+s}$ and $x = (\alpha I + B^*B)^{-1}L^s(x_* - x_0)$ to obtain (A.4):

$$\begin{aligned}
||\mathcal{C}_1||^2 &= ||[I - \Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-1/2}B)^*K](x_* - x_0)||^2 \\
&= ||L^{-s}[I - (\alpha I + B^*B)^{-1}(\Sigma^{-1/2}B)^*KL^{-s}]L^s(x_* - x_0)||^2 \\
&\leq \alpha^2\underline{c}^{-2}\left(\frac{s}{a+s}\right)||(B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}L^s(x_* - x_0)||^2 \qquad \text{(A.4)} \\
&= \alpha^2\underline{c}^{-2}\left(\frac{s}{a+s}\right) \\
&\quad \times ||(B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}(B^*B)^{\frac{\tilde{\beta}-s}{2(a+s)}}(B^*B)^{\frac{s-\tilde{\beta}}{2(a+s)}}L^s(x_* - x_0)||^2 \\
&= \alpha^2\underline{c}^{-2}\left(\frac{s}{a+s}\right)||(B^*B)^{\frac{\tilde{\beta}}{2(a+s)}}(\alpha I + B^*B)^{-1}w||^2, \qquad \text{(A.5)}
\end{aligned}$$

where $\tilde{\beta} = \min(\beta, a + 2s)$ and $w := (B^*B)^{\frac{s-\tilde{\beta}}{2(a+s)}}L^s(x_* - x_0)$. Now, by using the right part of inequality (A.1) in Corollary 3 with $\nu = \frac{s-\tilde{\beta}}{a+s}$ (remark that $|\nu| \leq 1$ is verified since $\tilde{\beta} \leq a + 2s$) and $x = L^s(x_* - x_0)$ we have that

$$\begin{aligned}
\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)}||w|| &\leq \overline{c}\left(\frac{s-\tilde{\beta}}{a+s}\right)\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)}||L^{\tilde{\beta}-s}L^s(x_* - x_0)|| \\
&= \overline{c}\left(\frac{s-\tilde{\beta}}{a+s}\right)\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)}||(x_* - x_0)||_{\tilde{\beta}} \\
&\leq \overline{c}\left(\frac{s-\tilde{\beta}}{a+s}\right)\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)}||(x_* - x_0)||_\beta||L^{\tilde{\beta}-\beta}|| \\
&= \overline{c}\left(\frac{s-\tilde{\beta}}{a+s}\right)\Gamma^{\frac{1}{2}}||L^{\tilde{\beta}-\beta}||, \qquad \text{(A.6)}
\end{aligned}$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

where in the penultimate line the equality holds for $\tilde{\beta} = \beta$. Remark that $||L^{\tilde{\beta}-\beta}|| = 1$ if $\tilde{\beta} = \beta$ and is bounded if $\tilde{\beta} < \beta$. Finally,

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} ||(B^*B)^{\frac{\tilde{\beta}}{2(a+s)}}(\alpha I + B^*B)^{-1}w||^2$$
$$= \sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} \sum_{j=1}^{\infty} \frac{\lambda_j^{\frac{2\tilde{\beta}}{a+s}}}{(\alpha+\lambda_j^2)^2}\langle w, \psi_j\rangle^2 \leq \left(\sup_{\lambda_j\geq 0} \frac{\lambda_j^{\frac{\tilde{\beta}}{a+s}}}{(\alpha+\lambda_j^2)}\right)^2 \sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} ||w||^2 \tag{A.7}$$

and combining (A.5)–(A.7) with the fact that $(\sup_{\lambda\geq 0}\lambda^{2b}(\alpha+\lambda^2)^{-1})^2 = \alpha^{2(b-1)}b^{2b}(1-b)^{2(1-b)}$ for $0 \leq b < 1$ (and in our case $b = \frac{\tilde{\beta}}{2(a+s)}$), we get the result

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} ||\mathcal{C}_1||^2 \leq \alpha^{\frac{\tilde{\beta}}{a+s}} b^{2b}(1-b)^{2(1-b)}\underline{c}^{-2}\left(\frac{s}{a+s}\right)\overline{c}^2\left(\frac{s-\tilde{\beta}}{a+s}\right)||L^{\tilde{\beta}-\beta}||^2\Gamma$$
$$= \mathcal{O}\left(\alpha^{\tilde{\beta}/(a+s)}\right).$$

Next, we address the second term of the MISE. To obtain (A.8), we use the left part of inequality (A.1) in Corollary 3 with $\nu = \frac{s}{a+s}$ and $x = (\alpha I + B^*B)^{-1}(\Sigma^{-1/2}B)^*U$:

$$\mathbf{E}_{x_*}||\mathcal{C}_2||^2 = \mathbf{E}_{x_*}||AU||^2 = \mathbf{E}_{x_*}||\Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}(\Sigma^{-1/2}B)^*U||^2$$
$$\leq \underline{c}^{-2}\left(\frac{s}{a+s}\right)\mathbf{E}_{x_*}||(B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}(\Sigma^{-1/2}B)^*U||^2 \tag{A.8}$$
$$= \delta\underline{c}^{-2}\left(\frac{s}{a+s}\right)$$
$$\times tr\left((B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-1}B^*B(\alpha I + B^*B)^{-1}(B^*B)^{\frac{s}{2(a+s)}}\right)$$
$$\leq \delta\underline{c}^{-2}\left(\frac{s}{a+s}\right)\left(\sup_j \frac{\lambda_j^{\frac{s}{(a+s)}+1-\gamma}}{(\alpha+\lambda_j^2)}\right)^2 \sum_{j=1}^{\infty}\langle (B^*B)^\gamma \psi_j, \psi_j\rangle$$
$$\leq \delta\underline{c}^{-2}\left(\frac{s}{a+s}\right)\alpha^{2(d-1)}d^{2d}(1-d)^{2(1-d)}\overline{c}(\gamma)tr\left(L^{-2\gamma(a+s)}\right), \tag{A.9}$$

where $d = \frac{a+2s-\gamma(a+s)}{2(a+s)}$. The last term is bounded since $tr(L^{-2\gamma(a+s)}) < \infty$ by definition of γ. The last inequality has been obtained by applying the right part of Corollary 3 with $\nu = \gamma$ to $\sum_{j=1}^{\infty}\langle (B^*B)^\gamma \psi_j, \psi_j\rangle$. In fact, $\sum_{j=1}^{\infty}\langle (B^*B)^\gamma \psi_j, \psi_j\rangle = \sum_{j=1}^{\infty}\langle (B^*B)^{\frac{\gamma}{2}}\psi_j, (B^*B)^{\frac{\gamma}{2}}\psi_j\rangle = \sum_{j=1}^{\infty}||(B^*B)^{\frac{\gamma}{2}}\psi_j||^2 \leq \overline{c}(\gamma)\sum_j ||L^{-\gamma(a+s)}\psi_j||^2 = \overline{c}(\gamma)\sum_j\langle L^{-\gamma(a+s)}\psi_j, L^{-\gamma(a+s)}\psi_j\rangle = \overline{c}(\gamma)tr(L^{-2\gamma(a+s)})$.

Therefore, $\mathbf{E}_{x_*}||\mathcal{C}_2||^2 = \mathcal{O}\left(\delta\alpha^{2(d-1)}\right)$ and $\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)}\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \mathcal{O}\left(\alpha^{\frac{\tilde{\beta}}{a+s}} + \delta\alpha^{-\frac{a+\gamma(a+s)}{(a+s)}}\right)$.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

A.3. Proof of Theorem 3

Let $\mathbf{E}_\delta^Y$ be the expectation taken with respect to the posterior μ_δ^Y. By the Chebyshev's inequality, for $\varepsilon_\delta > 0$ small enough and $M_\delta \to \infty$:

$$\mu_\delta^Y\{x \in \mathcal{X} : ||x - x_*|| > \varepsilon_\delta M_\delta\} \leq \frac{1}{\varepsilon_\delta^2 M_\delta^2}\mathbf{E}_\delta^Y||x - x_*||^2$$
$$= \frac{1}{\varepsilon_\delta^2 M_\delta^2}\Big(||\mathbf{E}(x|y^\delta, \alpha, s) - x_*||^2 + trVar(x|y^\delta, \alpha, s)\Big).$$

It remains to determine the rate of $Var(x|y^\delta, \alpha, s)$. Since $Var(x|y^\delta, \alpha, s) = \frac{\delta}{\alpha}[\Omega_0 - AK\Omega_0] = \delta\Omega_0^{\frac{1}{2}}(\alpha I + B^*B)^{-1}\Omega_0^{\frac{1}{2}}$ we have: $trVar(x|y^\delta, \alpha, s) = \delta tr(L^{-s}(\alpha I + B^*B)^{-1}L^{-s}) = \delta tr(RR^*)$ with $R = L^{-s}(\alpha I + B^*B)^{-\frac{1}{2}}$. Let $||\cdot||_{HS}$ denote the Hilbert–Schmidt norm. By using the left part of Corollary 3 with $\nu = \frac{s}{a+s}$ we get

$$trVar(x|y^\delta, \alpha, s) = \delta tr(R^*R) = \delta||R||_{HS}^2 = \delta\sum_{j=1}^\infty ||R\psi_j||^2$$
$$\leq \delta\underline{c}^{-2}\Big(\frac{s}{a+s}\Big)\sum_{j=1}^\infty ||(B^*B)^{\frac{s}{2(a+s)}}(\alpha I + B^*B)^{-\frac{1}{2}}\psi_j||^2$$
$$= \delta\underline{c}^{-2}\Big(\frac{s}{a+s}\Big)\sum_{j=1}^\infty \frac{\lambda_j^{\frac{2s}{a+s}-2\gamma}}{\alpha + \lambda_j^2}\langle(B^*B)^\gamma\psi_j, \psi_j\rangle$$
$$\leq \delta\underline{c}^{-2}\Big(\frac{s}{a+s}\Big)\sup_j \frac{\lambda_j^{\frac{2s}{a+s}-2\gamma}}{\alpha + \lambda_j^2}\sum_{j=1}^\infty\langle(B^*B)^\gamma\psi_j, \psi_j\rangle$$
$$\leq \delta\underline{c}^{-2}\Big(\frac{s}{a+s}\Big)\alpha^{\upsilon-1}\upsilon^\upsilon(1-\upsilon)^{1-\upsilon}\overline{c}(\gamma)tr(L^{-2\gamma(a+s)}), \quad \upsilon := \frac{s}{(a+s)} - \gamma$$

which is finite because $\gamma \in [0, 1]$ is such that $tr(L^{-2\gamma(a+s)}) < \infty$. The last inequality has been obtained by applying the right part of Corollary 3 and by following the steps used to obtain (A.9). Now, by this and the result of Theorem 2:

$$\mu_\delta^Y\{x \in \mathcal{X} : ||x - x_*|| > \varepsilon_\delta M_\delta\} = \mathcal{O}_p\Big(\frac{1}{\varepsilon_\delta^2 M_\delta^2}(\alpha^{\frac{\tilde\beta}{a+s}} + \delta\alpha^{-\frac{a+\gamma(a+s)}{(a+s)}})\Big), \quad \tilde\beta = \min(\beta, a+2s).$$

Hence, for $\varepsilon_\delta = (\alpha^{\frac{\tilde\beta}{2(a+s)}} + \delta^{\frac{1}{2}}\alpha^{-\frac{a+\gamma(a+s)}{2(a+s)}})$ we have $\mu_\delta^Y\{x \in \mathcal{X} : ||x - x_*|| > \varepsilon_\delta M_\delta\} \to 0$ in P^{x_*}-probability.

B. PROOFS FOR SECTION 5

B.1. Proof of Proposition 1

Let us consider the singular system $(\lambda_{jT}, \varphi_j, \psi_j)_j$ associated with $T := \Sigma^{-1/2}K$. Under Assumption C there exist $\underline{a}, \overline{a} > 0$ such that $\underline{a}\, j^{-2a} \leq \lambda_{jT}^2 \leq \overline{a}\, j^{-2a}$. Moreover, because L^{-1} takes the form in (13), $\{\psi_j\}$ and $\{\alpha j^{2s} + j^{-2a}\}$ are the eigenfunctions and (the order

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

of the) eigenvalues, respectively, of the operator $(\alpha L^{2s} + T^*T)$. We can then write: $\forall \phi \in \mathcal{X}_{2s} \cap \mathcal{D}(T^*T)$, $(\alpha L^{2s} + T^*T)\phi = \sum_{j\geq 1}(\alpha j^{2s} + j^{-2a})\langle \phi, \psi_j\rangle \psi_j$ and $\forall \phi \in \mathcal{X}$, $(\alpha L^{2s} + T^*T)^{-1}\phi = \sum_{j\geq 1}(\alpha j^{2s} + j^{-2a})^{-1}\langle \phi, \psi_j\rangle \psi_j$ where, under Assumption A.1, $\mathcal{D}(T^*T) = \mathcal{X}$. Therefore, terms $\mathcal{C}_1$ and $\mathcal{C}_2$ defined in the proof of Theorem 2 can be written in the following way:

$$
\begin{aligned}
||\mathcal{C}_1||^2 &= ||[I - L^{-s}(\alpha I + B^*B)^{-1}(\Sigma^{-1/2}B)^*K](x_* - x_0)||^2 \\
&= ||[I - (\alpha L^{2s} + T^*T)^{-1}T^*T](x_* - x_0)||^2 \\
&= \sum_{j\geq 1}\langle [I - (\alpha L^{2s} + T^*T)^{-1}T^*T](x_* - x_0), \psi_j\rangle^2 \\
&= \sum_{j\geq 1}\frac{\alpha^2 j^{4s}}{(\alpha j^{2s} + j^{-2a})^2}\langle (x_* - x_0), \psi_j\rangle^2 = \sum_{j\geq 1}\frac{\alpha^2}{(\alpha + j^{-2(a+s)})^2}\langle (x_* - x_0), \psi_j\rangle^2
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{E}_{x_*}||\mathcal{C}_2||^2 &= \mathbf{E}_{x_*}||L^{-s}(\alpha I + B^*B)^{-1}(\Sigma^{-1/2}B)^*U||^2 \\
&= \mathbf{E}_{x_*}||(\alpha L^s + B^*BL^s)^{-1}(\Sigma^{-1/2}B)^*U||^2 \\
&= \delta\sum_{j\geq 1}\langle (\alpha L^{2s} + T^*T)^{-1}T^*T(\alpha L^{2s} + T^*T)^{-1}\psi_j, \psi_j\rangle \\
&= \delta\sum_{j\geq 1}\frac{j^{-2a}}{(\alpha j^{2s} + j^{-2a})^2} = \delta\sum_{j\geq 1}\frac{j^{-2(a+s)}\, j^{-2s}}{(\alpha + j^{-2(a+s)})^2}.
\end{aligned}
$$

Hence, by using (A.3), the risk associated with $\hat{x}_\alpha$ writes:

$$
\begin{aligned}
\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 &= \sum_j \frac{\alpha^2}{(\alpha + j^{-2(a+s)})^2}\langle (x_* - x_0), \psi_j\rangle^2 \\
&\quad + \sum_j \frac{\delta j^{-2s-2(a+s)}}{(\alpha + j^{-2(a+s)})^2} =: \mathcal{A}1 + \mathcal{A}2.
\end{aligned}
$$

We have that $\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)}\mathcal{A}1 = \sup_{\sum_j j^{2\beta}\langle (x_*-x_0),\psi_j\rangle^2\leq\Gamma}\mathcal{A}1$ and

$$
\begin{aligned}
&\sup_{\sum_j j^{2\beta}\langle (x_*-x_0),\psi_j\rangle^2\leq\Gamma}\mathcal{A}1 \\
&= \alpha^2 \sup_{\sum_j j^{2\beta}\langle (x_*-x_0),\psi_j\rangle^2\leq\Gamma}\sum_j \frac{1}{\left(\alpha + j^{-2(a+s)}\right)^2}\langle (x_* - x_0), \psi_j\rangle^2 \\
&= \alpha^2 \sup_{\sum_j j^{2\beta}\langle (x_*-x_0),\psi_j\rangle^2\leq\Gamma}\sum_j \frac{j^{-2\beta}}{\left(\alpha + j^{-2(a+s)}\right)^2} j^{2\beta}\langle (x_* - x_0), \psi_j\rangle^2 \\
&= \alpha^2 \sup_j \frac{j^{-2\beta}}{\left(\alpha + j^{-2(a+s)}\right)^2}\Gamma.
\end{aligned}
$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

The supremum is attained at $j = \alpha^{-\frac{1}{2(a+s)}}\left(\frac{\beta}{2(a+s)-\beta}\right)^{-\frac{1}{2(a+s)}}$ as long as $\beta < 2(a+s)$. If $\beta \geq 2(a+s)$, then $\sup_j \frac{j^{-2\beta}}{(\alpha+j^{-2(a+s)})^2} = 1$. Consequently,

$$\sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} \mathcal{A}1 = \alpha^{\frac{\tilde{\beta}}{a+s}} b^{2b}(1-b)^{2(1-b)}\Gamma, \quad b = \frac{\tilde{\beta}}{2(a+s)}, \quad \tilde{\beta} = \min(\beta, 2(a+s))$$

by using the convention $(1-b)^{2(1-b)} = 1$ if $b = 1$. In order to analyze term $\mathcal{A}2$ we first remark that the summand function $f(j) := \frac{j^{-2s-2(a+s)}}{(\alpha+j^{-2(a+s)})^2}$ defined on $\mathbb{R}_+$ is increasing for $j < \bar{j}$ and decreasing for $j > \bar{j}$ where $\bar{j} = \left(\frac{a+2s}{a}\alpha\right)^{-\frac{1}{2(a+s)}}$. Thus,

$$\delta\int_{\bar{j}}^{\infty} \frac{j^{-2s-2(a+s)}}{(\alpha+j^{-2(a+s)})^2}dj \leq \mathcal{A}_2 \leq \delta\sum_{j=1}^{\bar{j}} \frac{j^{-2s-2(a+s)}}{(\alpha+j^{-2(a+s)})^2} + \delta\sum_{j=\bar{j}}^{\infty} \frac{j^{-2s-2(a+s)}}{(\alpha+j^{-2(a+s)})^2}$$

$$\Leftrightarrow \delta\alpha^{-\frac{2a+1}{2(a+s)}}\int_{\bar{t}}^{\infty} \frac{1}{[t^{-a}(t^{2(a+s)}+1)]^2}dt \leq \mathcal{A}_2$$

$$\leq \delta\frac{\bar{j}^{1-2s-2(a+s)}}{(\alpha+\bar{j}^{-2(a+s)})^2} + \delta\alpha^{-\frac{2a+1}{2(a+s)}}\int_{\bar{t}}^{\infty} \frac{1}{[t^{-a}(t^{2(a+s)}+1)]^2}dt,$$

where $\bar{t} = \left(\frac{a+2s}{a}\right)^{-\frac{1}{2(a+s)}}$. Denote $c_2(\bar{t}) = \int_{\bar{t}}^{\infty} \frac{1}{[t^{-a}(t^{2(a+s)}+1)]^2}dt$ and replace $\bar{j}$ by its value to obtain

$$\delta\alpha^{-\frac{2a+1}{2(a+s)}}c_2(\bar{t}) \leq \mathcal{A}_2 \leq \delta\alpha^{-\frac{2a+1}{2(a+s)}}\left(c_2(\bar{t}) + \bar{t}^{(1-2s-2(a+s))}\frac{a^2}{4(a+s)^2}\right) =: \delta\alpha^{-\frac{2a+1}{2(a+s)}}\tilde{c}_2(\bar{t}).$$

Therefore, $\mathcal{A}_2 \asymp \delta\alpha^{-\frac{2a+1}{2(a+s)}}$. Finally, let $G = b^{2b}(1-b)^{2(1-b)}\Gamma$, by putting the two results together we have:

$$\alpha^{\frac{\tilde{\beta}}{(a+s)}}G + \delta\alpha^{-\frac{2a+1}{2(a+s)}}c_2(\bar{t}) \leq \sup_{(x_*-x_0)\in\mathcal{X}_\beta(\Gamma)} \mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 \leq \alpha^{\frac{\tilde{\beta}}{(a+s)}}G + \delta\alpha^{-\frac{2a+1}{2(a+s)}}\tilde{c}_2(\bar{t}),$$

this proves (14). By replacing $\alpha \asymp \delta$ (respectively $\alpha \asymp \delta^{\frac{a+s}{\tilde{\beta}+a+1/2}}$) in this expression we get (i) (respectively (ii)).

B.2. Proof of Proposition 2

The proof proceeds similarly to the proof of Proposition 1, so we just sketch it. The risk $\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2$ associated with $\hat{x}_\alpha$ rewrites as the risk in Section B.1. The only term that we need to analyze is $\mathcal{A}1$ since the analysis of $\mathcal{A}2$ is the same as in B.1. By using Assumption B'

$$\mathcal{A}1 = \alpha^2\sum_j \frac{1}{(\alpha+j^{-2(a+s)})^2}\langle(x_*-x_0),\psi_j\rangle^2 = \alpha^2\sum_j \frac{j^{-2b_0}}{(\alpha+j^{-2(a+s)})^2}$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

and the function $f(j) := \frac{j^{-2b_0}}{(\alpha+j^{-2(a+s)})^2}$ defined on $\mathbb{R}_+$ is decreasing in j if $b_0 \geq 2a+2s$. If $b_0 < 2a+2s$, then $f(j)$ is increasing for $j < \bar{j}$ and decreasing for $j > \bar{j}$ where $\bar{j} = \alpha^{-\frac{1}{2(a+s)}} \left(\frac{b_0}{2a+2s-b_0}\right)^{-\frac{1}{2(a+s)}}$. Therefore, to upper and lower bound $\mathcal{A}_1$ we have to consider these two cases separately. If $b_0 < 2a+2s$

$$\alpha^2 \int_{\bar{j}}^{\infty} \frac{j^{-2b_0}}{(\alpha+j^{-2(a+s)})^2} dj \leq \mathcal{A}_1 \leq \alpha^2 \sum_{j=1}^{\bar{j}} \frac{j^{-2b_0}}{(\alpha+j^{-2(a+s)})^2} + \alpha^2 \sum_{j=\bar{j}}^{\infty} \frac{j^{-2b_0}}{(\alpha+j^{-2(a+s)})^2}$$

$$\Leftrightarrow \alpha^{\frac{2b_0-1}{2(a+s)}} \int_{\underline{u}}^{\infty} \frac{u^{4(a+s)-2b_0}}{(u^{2(a+s)}+1)^2} du \leq \mathcal{A}_1$$

$$\leq \alpha^2 \frac{\bar{j}^{1-2b_0}}{(\alpha+\bar{j}^{-2(a+s)})^2} + \alpha^{\frac{2b_0-1}{2(a+s)}} \int_{\underline{u}}^{\infty} \frac{u^{4(a+s)-2b_0}}{(u^{2(a+s)}+1)^2} du,$$

where $\underline{u} = \left(\frac{b_0}{2a+2s-b_0}\right)^{-\frac{1}{2(a+s)}}$. Denote $c_1(\tilde{u}) = \int_{\tilde{u}}^{\infty} \frac{u^{4(a+s)-2b_0}}{(u^{2(a+s)}+1)^2} du$ for some scalar $\tilde{u}$ and replace $\bar{j}$ by its value to obtain

$$\alpha^{\frac{2b_0-1}{2(a+s)}} c_1(\underline{u}) \leq \mathcal{A}_1 \leq \alpha^{\frac{2b_0-1}{2(a+s)}} \left(c_1(\underline{u}) + \underline{u}^{1-2b_0} \left(\frac{2a+2s-b_0}{2(a+s)}\right)^2 \right) =: \alpha^{\frac{2b_0-1}{2(a+s)}} \tilde{c}_1(\underline{u}).$$

If $b_0 \geq 2a+2s$:

$$\alpha^2 \int_1^{\infty} \frac{j^{-2b_0}}{(\alpha+j^{-2(a+s)})^2} dj \leq \mathcal{A}_1$$

$$\leq \alpha^2 \left. \frac{j^{-2b_0}}{(\alpha+j^{-2(a+s)})^2} \right|_{j=1} + \alpha^2 \int_1^{\infty} \frac{j^{-2b_0}}{(\alpha+j^{-2(a+s)})^2} dj$$

$$\Leftrightarrow \alpha^{\frac{2b_0-1}{2(a+s)}} \int_{\bar{u}}^{\infty} \frac{u^{4(a+s)-2b_0}}{(u^{2(a+s)}+1)^2} du \leq \mathcal{A}_1$$

$$\leq \alpha^2 + \alpha^{\frac{2b_0-1}{2(a+s)}} \int_{\bar{u}}^{\infty} \frac{u^{4(a+s)-2b_0}}{(u^{2(a+s)}+1)^2} du,$$

where $\bar{u} = \alpha^{\frac{1}{2(a+s)}}$. By using the notation defined above we obtain:

$$\alpha^{\frac{2b_0-1}{2(a+s)}} c_1(\bar{u}) \leq \mathcal{A}_1 \leq \alpha^2 + \alpha^{\frac{2b_0-1}{2(a+s)}} c_1(\bar{u}).$$

Since the integral in $c_1(\bar{u})$ is convergent, the upper bound is of the order $\alpha^{\frac{2\tilde{b}_0-1}{2(a+s)}}$ where $\tilde{b}_0 = \min(b_0, 2a+2s+1/2)$. Summarizing the two cases and by defining: $c_1 = c_1(\underline{u}) I(b_0 < 2(a+s)) + c_1(\bar{u}) I(b_0 \geq 2(a+s))$ and $\tilde{c}_1 = \tilde{c}_1(\underline{u}) I(b_0 < 2(a+s)) + c_1(\bar{u}) I(b_0 \geq 2(a+s))$ we have

$$\alpha^{\frac{2b_0-1}{2(a+s)}} c_1 \leq \mathcal{A}_1 \leq \alpha^2 I(b_0 \geq (2a+2s)) + \alpha^{\frac{2b_0-1}{2(a+s)}} \tilde{c}_1.$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

By using the upper and lower bounds for $\mathcal{A}_2$ given in the proof of Proposition 1 we get the expression in (15):

$$\alpha^{\frac{2b_0-1}{2(a+s)}}c_1+\delta\alpha^{-\frac{2a+1}{2(a+s)}}c_2(\bar{t}) \le \mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 \le \alpha^2 I(b_0 \ge (2a+2s)) + \alpha^{\frac{2b_0-1}{2(a+s)}}\tilde{c}_1+\delta\alpha^{-\frac{2a+1}{2(a+s)}}\tilde{c}_2(\bar{t}).$$

By replacing $\alpha \asymp \delta^{\frac{a+s}{b_0+a}}$ we obtain (16).

B.3. Proof of Theorem 4

We use the same strategy used for the proof of Theorem 3. The trace of the posterior covariance operator is:

$$\begin{aligned} tr[Var(x|y^\delta,\alpha,s)] &= \delta\sum_{j=1}^{\infty}\langle L^{-s}(\alpha I + B^*B)^{-1}L^{-s}\psi_j,\psi_j\rangle \\ &= \delta\sum_{j=1}^{\infty}\langle (\alpha L^s + B^*BL^s)^{-1}L^{-s}\psi_j,\psi_j\rangle \\ &= \delta\sum_{j=1}^{\infty}\langle (\alpha L^{2s} + T^*T)^{-1}\psi_j,\psi_j\rangle \\ &= \delta\sum_{j\ge 1}\frac{1}{\alpha j^{2s}+j^{-2a}} = \delta\sum_{j\ge 1}\frac{j^{-2s}}{\alpha + j^{-2(a+s)}}. \end{aligned}$$

Since $f(j) := \frac{j^{-2s}}{\alpha+j^{-2(a+s)}}$ is increasing in j for $j < \bar{j}$ and decreasing for $j > \bar{j}$ where $\bar{j} = (\alpha s/a)^{-\frac{1}{2(a+s)}}$ then we can upper and lower bound the trace of the posterior covariance operator as:

$$\begin{aligned} \delta\sum_{j=\bar{j}}^{\infty} f(j) \le tr[Var(x|y^\delta,\alpha,s)] &\le \delta\sum_{j=1}^{\bar{j}} f(j) + \delta\sum_{j=\bar{j}}^{\infty} f(j) \\ \Leftrightarrow \delta\alpha^{-\frac{2a+1}{2(a+s)}}\int_{\bar{t}}^{\infty}\frac{t^{2a}}{t^{2(a+s)}+1}dt &\le tr[Var(x|y^\delta,\alpha,s)] \\ \le \delta\frac{\bar{j}^{1-2s}}{\alpha+\bar{j}^{-2(a+s)}} + \delta\alpha^{-\frac{2a+1}{2(a+s)}}\int_{\bar{t}}^{\infty}\frac{t^{2a}}{t^{2(a+s)}+1}dt & \\ \Leftrightarrow \delta\alpha^{-\frac{2a+1}{2(a+s)}}\kappa_2(\bar{t}) &\le tr[Var(x|y^\delta,\alpha,s)] \\ \le \delta\alpha^{-\frac{2a+1}{2(a+s)}}\bar{t}^{-\frac{(1-2s)}{2(a+s)}}\frac{a}{a+s} + \delta\alpha^{-\frac{2a+1}{2(a+s)}}\kappa_2(\bar{t}), & \end{aligned}$$

where $\bar{t} = (s/a)^{-\frac{1}{2(a+s)}}$ and $\kappa_2(\bar{t}) = \int_{\bar{t}}^{\infty}\frac{t^{2a}}{t^{2(a+s)}+1}dt$. Thus, $tr[Var(x|y^\delta,\alpha,s)] \asymp \delta\alpha^{-\frac{2a+1}{2(a+s)}}$. By the Chebyshev's inequality and Proposition 2, for $\varepsilon_\delta > 0$ small enough and $M_\delta \to \infty$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

$$\mu_\delta^Y\{x\in\mathcal{X}:||x-x_*||>\varepsilon_\delta M_\delta\}\le\frac{1}{\varepsilon_\delta^2M_\delta^2}\Big(||\mathbf{E}(x|y^\delta,\alpha,s)-x_*||^2+trVar(x|y^\delta,\alpha,s)\Big)$$
$$=\mathcal{O}_p\Big(\frac{1}{\varepsilon_\delta^2M_\delta^2}(\alpha^{\frac{2\tilde b_0-1}{2(a+s)}}+\delta\alpha^{-\frac{2a+1}{2(a+s)}})\Big),$$
$$\tilde b_0=\min(b_0,2(a+s)+1/2)$$

in P^{x_*}-probability. Hence, for $\varepsilon_\delta=(\alpha^{\frac{2\tilde b_0-1}{4(a+s)}}+\delta^{\frac12}\alpha^{-\frac{2a+1}{4(a+s)}})$ we conclude that $\mu_\delta^Y\{x\in\mathcal{X}:||x-x_*||>\varepsilon_\delta M_\delta\}\to0$. For $\alpha\asymp\delta$ we obtain *(i)*. To obtain *(ii)*: $\arg\inf_\alpha\varepsilon_\delta=\delta^{\frac{a+s}{\tilde b_0+a}}$ and $\inf_\alpha\varepsilon_\delta=\delta^{\frac{2\tilde b_0-1}{4(\tilde b_0+a)}}$.

C. PROOFS FOR SECTION 6

C.1. Proof of Theorem 5

To prove Theorem 5 we use Theorem 3.3, p. 123 in Kuo (1975). We first rewrite this theorem and then show that the conditions of this theorem are verified in our case. The proof of Theorem 7 is given in Kuo (1975).

THEOREM 7. *Let P_2 be a Gaussian measure on $\mathcal{Y}$ with mean m and covariance operator S_2 and P_1 be another Gaussian measure on the same space with mean m and covariance operator S_1. If there exists a positive definite, bounded, continuously invertible operator $\mathcal{H}$ such that $S_2=S_1^{\frac12}\mathcal{H}S_1^{\frac12}$ and $\mathcal{H}-I$ is Hilbert–Schmidt, then P_2 is equivalent to P_1. Moreover, the Radon–Nikodym derivative is given by*

$$\frac{dP_2}{dP_1}(\{z_j\})=\prod_{j=1}^\infty\sqrt{\frac{\alpha}{\lambda_j^2+\alpha}}e^{\frac{\lambda_j^2}{2(\lambda_j^2+\alpha)}z_j^2},\qquad\textbf{(C.1)}$$

with $\frac{\lambda_j^2}{\alpha}$ the eigenvalues of $\mathcal{H}-I$ and z_j a sequence of real numbers.

In our case: $P_2=P_\alpha$, $m=Kx_0$, $S_2=\delta\Sigma+\frac{\delta}{\alpha}K\Omega_0K^*$, $P_1=P_0$, and $S_1=\delta\Sigma$. We rewrite S_2 as

$$S_2\equiv\Big(\delta\Sigma+\frac{\delta}{\alpha}K\Omega_0K^*\Big)=\sqrt\delta\Sigma^{\frac12}\Big[I+\frac1\alpha\Sigma^{-1/2}K\Omega_0K^*\Sigma^{-1/2}\Big]\Sigma^{\frac12}\sqrt\delta=S_1^{\frac12}\mathcal{H}S_1^{\frac12}$$

with $\mathcal{H}=(I+\frac1\alpha\Sigma^{-1/2}K\Omega_0K^*\Sigma^{-1/2})=(I+\frac1\alpha BB^*)$. In the following four points we show that $\mathcal{H}$ satisfies all the properties required in Theorem 7.

(1) $\mathcal{H}$ is *positive definite*. In fact, $(I+\frac1\alpha BB^*)$ is self-adjoint, i.e., $(I+\frac1\alpha BB^*)^*=(I+\frac1\alpha BB^*)$ and $\forall\varphi\in\mathcal{Y},\varphi\neq0$

$$\Big\langle\Big(I+\frac1\alpha BB^*\Big)\varphi,\varphi\Big\rangle=\langle\varphi,\varphi\rangle+\frac1\alpha\langle B^*\varphi,B^*\varphi\rangle=||\varphi||^2+\frac1\alpha||B^*\varphi||^2>0.$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

(2) $\mathcal{H}$ is *bounded*. The operators B and B^* are bounded under Assumption A.2, the operator I is bounded by definition, and a linear combination of bounded operators is bounded (see Remark 2.7 in Kress, 1999).
(3) $\mathcal{H}$ is *continuously invertible*. To show this, we first recall that $(I + \frac{1}{\alpha} BB^*)$ is continuously invertible if its inverse is bounded, i.e., there exists a positive number C such that $||(I + \frac{1}{\alpha} BB^*)^{-1}\varphi|| \leq C||\varphi||$, $\forall \varphi \in \mathcal{Y}$. We have $||(I + \frac{1}{\alpha} BB^*)^{-1}\varphi|| \leq (\sup_j \frac{\alpha}{\alpha+\lambda_j^2})||\varphi|| = ||\varphi||, \forall \varphi \in \mathcal{Y}$.
(4) $(\mathcal{H} - I)$ is *Hilbert–Schmidt*. To show this we have to consider the Hilbert–Schmidt norm $||\frac{1}{\alpha} BB^*||_{HS} = \frac{1}{\alpha}\sqrt{tr((BB^*)^2)}$. Now, $tr((BB^*)^2) = tr(\Omega_0 T^* T \Omega_0 T^* T) \leq tr(\Omega_0)||T^* T \Omega_0 T^* T|| < \infty$ since $T := \Sigma^{-1/2} K$ has a bounded norm under Assumption A.1.

C.2. Proof of Theorem 6

We start by showing the first statement of the theorem and then we proceed with the proof of (21). Let $\tilde{c}_3$, c_4, and $\tilde{c}_4$ be the constants defined in Lemma 2. Fix $\epsilon_\delta = \delta^{\frac{1-r}{4(b_0+a)}}$ for $0 < r < 1$, $\alpha_1 = \left(\frac{\tilde{c}_3}{c_4}(1+\epsilon_\delta)\right)^{\frac{a+s}{b_0+a}} \delta^{\tilde{p}}$, $\alpha_2 = \left(\frac{\tilde{c}_3}{\tilde{c}_4}(1-\epsilon_\delta)\right)^{\frac{a+s}{b_0+a}} \delta^{p}$ where: $\tilde{p} = \frac{a+s}{b_0+a+\tilde{\eta}}$, $p = \frac{a+s}{b_0+a-\eta}$, $\tilde{\eta} = \eta I\{b_0 - a - 2s - 1/2 > 0\}$, and $(b_0 + a) > \eta > \max\{(b_0 - s - 1/2), 0, (b_0 - 2s - a - 1/2)\frac{1}{4}\}$. Remark that, since $b_0 - s - 1/2 > (b_0 - 2s - a - 1)$ and $0 < \epsilon_\delta < 1$ (because $0 < \delta < 1$) then the assumptions of Lemmas 4 and 5 are satisfied. Moreover, $\tilde{c}_4 \geq c_4$ (see the proof of Lemma 2). Let $G := \{\exists \hat{\alpha}; S_{y^\delta}(\hat{\alpha}) = 0 \cap \hat{\alpha} \in (\alpha_2, \alpha_1)\}$ and G^c be its relative complement with respect to $\{\exists \hat{\alpha}; S_{y^\delta}(\hat{\alpha}) = 0\}$. In order to prove the first statement of the theorem we make the following decomposition:

$$\begin{aligned} P\{\exists \hat{\alpha}; S_{y^\delta}(\hat{\alpha}) = 0\} &= P\{\exists \hat{\alpha}; S_{y^\delta}(\hat{\alpha}) = 0 \cap \hat{\alpha} \in (\alpha_2, \alpha_1)\} \\ &\quad + P\{\exists \hat{\alpha}; S_{y^\delta}(\hat{\alpha}) = 0 \cap \hat{\alpha} \notin (\alpha_2, \alpha_1)\} \\ &= P\{G\} + P\{G^c\}. \end{aligned}$$

From Lemma 6 we conclude that as $\delta \to 0$: $P(G) \to 1$ and so $\hat{\alpha} \in (\alpha_2, \alpha_1)$ with probability approaching 1. This implies that we can write $\hat{\alpha}$ as a (random) convex combination of α_2 and α_1: $\hat{\alpha} = \rho\alpha_2 + (1-\rho)\alpha_1$ for ρ a random variable with values in $(0, 1)$. Since $\alpha_2 \to 0$ faster than $\alpha_1 \to 0$ then:

$$\begin{aligned} \hat{\alpha} &= o_p\left(\delta^{\frac{a+s}{b_0+a+\tilde{\eta}}}\right) + \delta^{\frac{a+s}{b_0+a+\tilde{\eta}}}(1-\rho)\left[\frac{\tilde{c}_3}{c_4}(1+\epsilon_\delta)\right]^{\frac{a+s}{b_0+a}} \\ &=: o_p\left(\delta^{\frac{a+s}{b_0+a+\tilde{\eta}}}\right) + \delta^{\frac{a+s}{b_0+a+\tilde{\eta}}}(1-\rho)\kappa_1. \end{aligned} \tag{C.2}$$

This establishes the first result of the theorem.

Next, we proceed with the proof of (21). Denote by I_G and I_{G^c} the indicator functions of the events G and G^c, respectively. By the Markov's inequality, to show that $\mu^Y_{\delta,\hat{\alpha}}\{||x - x_*|| > \varepsilon_\delta M_\delta\} \to 0$ in probability it is sufficient to show that its expectation, with respect to P^{x_*}, converges to 0. Then,

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

$$\begin{aligned}
&\mathbf{E}_{x_*}\left(\mu^Y_{\delta,\hat\alpha}\{||x-x_*||>\varepsilon_\delta M_\delta\}\right)\\
&=\mathbf{E}_{x_*}\left(\mu^Y_{\delta,\hat\alpha}\{||x-x_*||>\varepsilon_\delta M_\delta\}I_{G^c}\right)+\mathbf{E}_{x_*}\left(\mu^Y_{\delta,\hat\alpha}\{||x-x_*||>\varepsilon_\delta M_\delta\}I_G\right)\\
&\le \mathbf{E}_{x_*}(I_{G^c})+\mathbf{E}_{x_*}\left(\frac{||\hat x_{\hat\alpha}-x_*||^2}{\varepsilon_\delta^2 M_\delta^2}I_G+\frac{trVar(x|y^\delta,\hat\alpha,s)}{\varepsilon_\delta^2 M_\delta^2}I_G\right), \qquad \textbf{(C.3)}
\end{aligned}$$

where the last inequality follows from the Chebyshev's inequality. By Lemma 6: $\mathbf{E}(I_{G^c}) = P(G^c)\to 0$. We then analyze the second term of (C.3). Let α_* denote the optimum value of α given in Proposition 2, that is, $\alpha_* = \delta^{\frac{a+s}{\tilde b_0+a}}$ with $\tilde b_0 = \min\{b_0, 2(a+s)+1/2\}$. By using the notation $B(\alpha) := (\alpha I + B^*B)^{-1}$ and $\tilde B(\alpha) := (\alpha I + \Omega_0 K^*\Sigma^{-1}K)^{-1}$ and by adding and subtracting $\Omega_0^{\frac12}B(\alpha_*)(\Sigma^{-1/2}B)^*(y^\delta - Kx_0)$ we obtain:

$$\begin{aligned}
\mathbf{E}_{x_*}\left(||\hat x_{\hat\alpha}-x_*||^2 I_G\right) &= \mathbf{E}_{x_*}||\Omega_0^{\frac12}B(\hat\alpha)(\Sigma^{-1/2}B)^*(y^\delta-Kx_0)+x_0-x_*||^2 I_G\\
&\le 2\mathbf{E}_{x_*}||\Omega_0^{\frac12}B(\hat\alpha)(\alpha_*-\hat\alpha)B(\alpha_*)(\Sigma^{-1/2}B)^*(y^\delta-Kx_0)||^2 I_G\\
&\quad +2\mathbf{E}_{x_*}||\Omega_0^{\frac12}B(\alpha_*)(\Sigma^{-1/2}B)^*(y^\delta-Kx_0)+x_0-x_*||^2 I_G\\
&\le 4\mathbf{E}_{x_*}||\tilde B(\hat\alpha)(\alpha_*-\hat\alpha)\left[\Omega_0^{\frac12}B(\alpha_*)(\Sigma^{-1/2}B)^*(y^\delta-Kx_0)+x_0-x_*\right]||^2 I_G\\
&\quad +4\mathbf{E}_{x_*}||\tilde B(\hat\alpha)(\alpha_*-\hat\alpha)(x_0-x_*)||^2 I_G + 2\mathbf{E}_{x_*}||\hat x_{\alpha_*}-x_*||^2\\
&\le 4\mathbf{E}_{x_*}||\tilde B(\hat\alpha)(\alpha_*-\hat\alpha)||^2 I_G \mathbf{E}_{x_*}||\hat x_{\alpha_*}-x_*||^2\\
&\quad +4\mathbf{E}_{x_*}(\alpha_*-\hat\alpha)^2||\tilde B(\hat\alpha)(x_0-x_*)||^2 I_G + 2\mathbf{E}_{x_*}||\hat x_{\alpha_*}-x_*||^2,
\end{aligned}$$

where the last inequality is due to the Cauchy–Schwartz inequality. By the first part of this proof, $\hat\alpha$ can be written as in (C.2) so that there exists a positive constant κ_1 such that:

$$\begin{aligned}
\mathbf{E}_{x_*}||\tilde B(\hat\alpha)(\alpha_*-\hat\alpha)||^2 I_G &= \mathbf{E}_{x_*}||(\hat\alpha L^{2s}+T^*T)^{-1}L^{2s}(\alpha_*-\hat\alpha)||^2 I_G\\
&= \mathbf{E}_{x_*}(\alpha_*-\hat\alpha)^2 \sup_{||\phi||=1}\sum_{j=1}^{\infty}\frac{1}{(\hat\alpha+j^{-2(a+s)})^2}\langle\phi,\psi_j\rangle^2 I_G\\
&\le \mathbf{E}_{x_*}\left(\frac{\alpha_*-\hat\alpha}{\hat\alpha}I_G\right)^2\\
&= \mathbf{E}_{x_*}\left(\frac{\left(\delta^{\tilde p\frac{(b_0-\tilde b_0+\bar\eta)}{(a+\tilde b_0)}}-o_p(1)-(1-\rho)\kappa_1\right)^2}{\left(o_p(1)+(1-\rho)\kappa_1\right)^2}\right)=\mathcal{O}(1). \qquad \textbf{(C.4)}
\end{aligned}$$

Moreover, $||\tilde B(\hat\alpha)(x_0-x_*)||^2 \asymp \sum_j \frac{j^{-2b_0}}{(\hat\alpha+j^{-2(a+s)})^2} = \hat\alpha^{-2}\mathcal{A}1$ where $\mathcal{A}1$ is the term defined in Section B.2 with the only difference that α must be replaced by $\hat\alpha$. With reference to the notation of that section we have: $\hat\alpha^{\frac{2b_0-1}{2(a+s)}-2}c_1 \le ||\tilde B(\hat\alpha)(x_0-x_*)||^2 \le I(b_0\ge 2(a+s)) + \hat\alpha^{\frac{2b_0-1}{2(a+s)}-2}\tilde c_1$. Therefore, by denoting $\tilde\kappa_\rho := [o_p(1)+(1-\rho)\kappa_1]$,

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

$$\delta^{\frac{2b_0-1}{2(b_0+a+\tilde{\eta})}}\mathbf{E}_{x_*}\left((\tilde{\kappa}_\rho)^{\frac{2b_0-1}{2(a+s)}-2}\left[\delta^{\tilde{p}\frac{b_0-\tilde{b}_0+\tilde{\eta}}{(a+\tilde{b}_0)}}-\tilde{\kappa}_\rho\right]^2 c_1\right)$$
$$\leq \mathbf{E}_{x_*}(\alpha_*-\hat{\alpha})^2\,||\tilde{B}(\hat{\alpha})(x_0-x_*)||^2 I_G$$
$$\leq \mathbf{E}_{x_*}\left[\left(\delta^{\frac{2(a+s)}{b_0+a+\tilde{\eta}}}I(b_0\geq 2(a+s))+\delta^{\frac{2b_0-1}{2(b_0+a+\tilde{\eta})}}(\tilde{\kappa}_\rho)^{\frac{2b_0-1}{2(a+s)}-2}\tilde{c}_1\right)\left[\delta^{\tilde{p}\frac{b_0-\tilde{b}_0+\tilde{\eta}}{(a+\tilde{b}_0)}}-\tilde{\kappa}_\rho\right]^2\right],$$

that is, $\mathbf{E}_{x_*}(\alpha_*-\hat{\alpha})^2\,||\tilde{B}(\hat{\alpha})(x_0-x_*)||^2 I_G=\mathcal{O}\left(\delta^{\frac{2\tilde{b}_0-1}{2(b_0+a+\tilde{\eta})}}\right)$ then, we conclude that

$$\mathbf{E}_{x_*}\left(||\hat{x}_{\hat{\alpha}}-x_*||^2 I_G\right)\leq \mathbf{E}_{x_*}||\hat{x}_{\alpha_*}-x_*||^2(2+4\mathcal{O}(1))+4\mathcal{O}\left(\delta^{\frac{2\tilde{b}_0-1}{2(b_0+a+\tilde{\eta})}}\right). \tag{C.5}$$

Let us turn to the analysis of the variance term:

$$\begin{aligned}
trVar(x|y^\delta,\hat{\alpha},s) &= \delta tr\left[\Omega_0(\hat{\alpha}I+B^*B)^{-1}\right]\\
&= \delta tr\Big[\Omega_0(\alpha_*I+B^*B)^{-1}\\
&\qquad +\Omega_0\left[(\hat{\alpha}I+B^*B)^{-1}-(\alpha_*I+B^*B)^{-1}\right]\Big]\\
&= \delta tr\left[L^{-s}(\alpha_*I+B^*B)^{-1}L^{-s}\right]\\
&\quad +\delta tr\left[L^{-s}(\hat{\alpha}I+B^*B)^{-1}(\alpha_*-\hat{\alpha})(\alpha_*I+B^*B)^{-1}L^{-s}\right]\\
&= \delta tr\left[(\alpha_*L^{2s}+T^*T)^{-1}\right]\\
&\quad +\delta tr\left[(\hat{\alpha}L^{2s}+T^*T)^{-1}L^s(\alpha_*-\hat{\alpha})L^s(\alpha_*L^{2s}+T^*T)^{-1}\right]\\
&= \delta\sum_{j=1}^{\infty}\frac{1}{(\alpha_*j^{2s}+j^{-2a})}+\delta\sum_{j=1}^{\infty}\frac{(\alpha_*-\hat{\alpha})j^{2s}}{(\hat{\alpha}j^{2s}+j^{-2a})}\frac{1}{(\alpha_*j^{2s}+j^{-2a})}\\
&= \delta\sum_{j=1}^{\infty}\frac{j^{-2s}}{(\alpha_*+j^{-2(a+s)})}+\delta\sum_{j=1}^{\infty}\frac{(\alpha_*-\hat{\alpha})}{(\hat{\alpha}+j^{-2(a+s)})}\frac{j^{-2s}}{(\alpha_*+j^{-2(a+s)})}\\
&\leq \delta\left(1+\frac{|\alpha_*-\hat{\alpha}|}{\hat{\alpha}}\right)\sum_{j=1}^{\infty}\frac{j^{-2s}}{(\alpha_*+j^{-2(a+s)})}\\
&= \left(1+\frac{|\alpha_*-\hat{\alpha}|}{\hat{\alpha}}\right)tr[Var(x|y^\delta,\alpha_*,s)],
\end{aligned}$$

where the last equality uses the expression of the trace of the variance recovered in the proof of Theorem 4 (see Section B.3). From (C.4) it results that $\mathbf{E}_{x_*}\left(\frac{|\alpha_*-\hat{\alpha}|}{\hat{\alpha}}I_G\right)=\mathcal{O}(1)$. Therefore,

$$\mathbf{E}_{x_*}trVar(x|y^\delta,\hat{\alpha},s)I_G=(1+\mathcal{O}(1))\,tr[Var(x|y^\delta,\alpha_*,s)]$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

and by using the upper and lower bounds for $tr[Var(x|y^\delta,\alpha_*,s)]$ derived in the proof of Theorem 4 we conclude that $\mathbf{E}_{x_*} tr Var(x|y^\delta,\hat\alpha,s) I_G \asymp (1+\mathcal{O}(1))\,\delta\alpha_*^{-\frac{2a+1}{2(a+s)}}$. By replacing these rates in (C.3) we have:

$$\mathbf{E}_{x_*}\left(\mu^Y_{\delta,\hat\alpha}\{||x-x_*||>\varepsilon_\delta M_\delta\}\right)=\mathcal{O}\left(\frac{1}{\varepsilon_\delta^2 M_\delta^2}\left(\delta^{\frac{2\tilde b_0-1}{2(\tilde b_0+a)}}+\delta^{\frac{2\tilde b_0-1}{2(\tilde b_0+a+\tilde\eta)}}\right)\right)$$

which converges to 0 for $\varepsilon_\delta \asymp \delta^{\frac{2\tilde b_0-1}{4(\tilde b_0+a+\tilde\eta)}}$.

D. PROOFS FOR SECTION 7

D.1. Proof of Corollary 1

Let $\widetilde{B^*B}=\Omega_0^{\frac12}\mathcal{T}^*\mathcal{T}\Omega_0^{\frac12}$, $R_\alpha=(\alpha I+\widetilde{B^*B})^{-1}$, $\hat R_\alpha=(\alpha I+B^*B)^{-1}$, $\Sigma^{-1/2}B=(\Lambda^*)^{-1}\frac{1}{\hat f_W}\mathcal{T}\Omega_0^{\frac12}$, $\widetilde{\Sigma^{-1/2}B}=(\Lambda^*)^{-1}\frac{1}{f_W}\mathcal{T}\Omega_0^{\frac12}$. Moreover, we denote $\mathbf{E}\equiv\mathbf{E}_{x_*}$ and define $U_0=U-\mathbf{E}U$, $\tilde y^\delta=\Lambda f_W\mathcal{T}x_*+U_0$, and $\tilde x_\alpha=\Omega_0^{\frac12}R_\alpha(\widetilde{\Sigma^{-1/2}B})^*(\tilde y^\delta-\Lambda f_W\mathcal{T}x_0)$. We decompose $(\hat x_\alpha-x_*)$ as

$$\begin{aligned}(\hat x_\alpha-x_*) &= (\tilde x_\alpha-x_*)+\Omega_0^{\frac12}\hat R_\alpha(\Sigma^{-1/2}B)^*(y^\delta-\Lambda\tilde K x_0)\\ &\quad-\Omega_0^{\frac12}R_\alpha(\widetilde{\Sigma^{-1/2}B})^*(\tilde y^\delta-\Lambda f_W\mathcal{T}x_0)\\ &= (\tilde x_\alpha-x_*)+\Omega_0^{\frac12}\left[\hat R_\alpha(\Sigma^{-1/2}B)^*\Lambda\tilde K-R_\alpha(\widetilde{\Sigma^{-1/2}B})^*\Lambda f_W\mathcal{T}\right](x_*-x_0)\\ &\quad+\Omega_0^{\frac12}\left[\hat R_\alpha(\Sigma^{-1/2}B)^*U-R_\alpha(\widetilde{\Sigma^{-1/2}B})^*U_0\right]\\ &=: (\tilde x_\alpha-x_*)+\mathfrak{A}_1+\mathfrak{A}_2\end{aligned}$$

(I) *Convergence of* $\mathbf{E}||(\tilde x_\alpha-x_*)||^2$. This rate is given in Theorem 2 (since this term depends on operators that do not vary with δ).

(II) *Convergence of* $\mathfrak{A}_1$.

$$\begin{aligned}\mathbf{E}||\mathfrak{A}_1||^2 &= \mathbf{E}||\Omega_0^{\frac12}\left(\hat R_\alpha\left[(\Sigma^{-1/2}B)^*\Lambda\tilde K L^{-s}-(\widetilde{\Sigma^{-1/2}B})^*\Lambda f_W\mathcal{T}L^{-s}\right]\right.\\ &\qquad\left.+(\hat R_\alpha-R_\alpha)(\widetilde{\Sigma^{-1/2}B})^*\Lambda f_W\mathcal{T}L^{-s}\right)L^s(x_*-x_0)||^2\\ &= \mathbf{E}||\Omega_0^{\frac12}\left(\hat R_\alpha\left[B^*B-\widetilde{B^*B}\right]+\hat R_\alpha(\widetilde{B^*B}-B^*B)R_\alpha\widetilde{B^*B}\right)L^s(x_*-x_0)||^2\\ &= \mathbf{E}||\Omega_0^{\frac12}\hat R_\alpha\Omega_0^{\frac12}\left[\hat{\mathcal{T}}^*\hat{\mathcal{T}}-\mathcal{T}^*\mathcal{T}\right]\Omega_0^{\frac12}\alpha R_\alpha L^s(x_*-x_0)||^2\\ &= \mathbf{E}||\Omega_0^{\frac12}\hat R_\alpha\Omega_0^{\frac12}\left(\hat{\mathcal{T}}^*(\hat{\mathcal{T}}-\mathcal{T})+(\hat{\mathcal{T}}^*-\mathcal{T}^*)\mathcal{T}\right)\Omega_0^{\frac12}\alpha R_\alpha L^s(x_*-x_0)||^2\end{aligned}$$

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

$$\leq \left(2\mathbf{E}||\Omega_0^{\frac{1}{2}}\hat{R}_\alpha\Omega_0^{\frac{1}{2}}\hat{T}^*||^2\mathbf{E}||(\hat{T}-T)||^2 + 2\mathbf{E}||\Omega_0^{\frac{1}{2}}\hat{R}_\alpha||^2\mathbf{E}||\Omega_0^{\frac{1}{2}}(\hat{T}^*-T^*)||^2||T||^2\right)\alpha^2||\Omega_0^{\frac{1}{2}}R_\alpha L^s(x_*-x_0)||^2$$
$$= \mathcal{O}\left(\alpha^{-1}\left(\frac{1}{nh_wh_z}+(\max\{h_w,h_z\})^{2\rho}\right)+\alpha^{-2}\left(\frac{1}{nh_w}+h_w^{2\rho}\right)\right)\mathcal{O}\left(\alpha^{\frac{\tilde{\beta}}{a+s}}\right).$$

The last term $||\Omega_0^{\frac{1}{2}}R_\alpha L^s(x_*-x_0)||^2\alpha^2$ is equal to term $\mathcal{C}_1$ in the proof of Theorem 2 and so it is a $\mathcal{O}\left(\alpha^{\frac{\tilde{\beta}}{a+s}}\right)$. Moreover, we have used Assumption HS(ii) and (iii) and the fact that $\mathbf{E}||\Omega_0^{1/2}\hat{R}_\alpha||^2=\mathcal{O}\left(\alpha^{-2}\right)$ and $\mathbf{E}||\Omega_0^{\frac{1}{2}}\hat{R}_\alpha\Omega_0^{\frac{1}{2}}\hat{T}^*||^2=\mathcal{O}\left(\alpha^{-1}\right)$.

(III) *Convergence of* $\mathfrak{A}_2$.

$$\mathbf{E}||\mathfrak{A}_2||^2 = \mathbf{E}||\Omega_0^{\frac{1}{2}}\left(\hat{R}_\alpha(\Sigma^{-1/2}B)^*U - R_\alpha\widetilde{(\Sigma^{-1/2}B)}^*U_0\right)||^2$$
$$\leq 2\mathbf{E}||\Omega_0^{\frac{1}{2}}\hat{R}_\alpha(\Sigma^{-1/2}B)^*U||^2 + 2\mathbf{E}||\Omega_0^{\frac{1}{2}}R_\alpha\widetilde{(\Sigma^{-1/2}B)}^*U_0||^2 \qquad \textbf{(D.1)}$$
$$\leq 2\mathbf{E}||\Omega_0^{\frac{1}{2}}\hat{R}_\alpha\Omega_0^{\frac{1}{2}}\hat{T}^*\hat{f}_W^{-1}\Lambda^{-1}U||^2 + 2\mathbf{E}||\Omega_0^{\frac{1}{2}}R_\alpha\Omega_0^{\frac{1}{2}}T^*\hat{f}_W^{-1}\Lambda^{-1}U_0||^2$$
$$= \mathcal{O}\left(\alpha^{-2}\left(\frac{1}{n}+h_z^{2\rho}\right)+n^{-1}\alpha^{-\frac{a+\gamma(a+s)}{a+s}}\right)$$

by using Assumption HS *(i)* and the fact that the second term in (D.1) is equal to term $\mathcal{C}_2$ in the proof of Theorem 2. By writing $\mathbf{E}||\hat{x}_\alpha-x_*||^2\leq 2\mathbf{E}||\tilde{x}_\alpha-x_*||^2+2\mathbf{E}||\mathfrak{A}_1+\mathfrak{A}_2||^2$ and by putting all these results together we get the result.

D.2. Proof of Corollary 2

Let us denote: $t=s/(2a)$, $\tilde{R}_\alpha=(\alpha I+\Sigma^{2t+1})^{-1}$, $R_\alpha=(\alpha I+K^{2t+1})^{-1}$, $\hat{x}_\alpha = K^tR_\alpha K^t(y^\delta-Kx_0)+x_0$. Then,

$$\mathbf{E}_{x_*}||\hat{x}_\alpha-x_*||^2 = \mathbf{E}_{x_*}||\alpha R_\alpha(x_*-x_0)+K^tR_\alpha K^tU||^2$$
$$= \mathbf{E}_{x_*}||\alpha\tilde{R}_\alpha(x_*-x_0)+\alpha(R_\alpha-\tilde{R}_\alpha)(x_*-x_0)+\tilde{R}_\alpha\Sigma^{2t}U + (R_\alpha K^{2t}-\tilde{R}_\alpha\Sigma^{2t})U||^2$$
$$\leq 2\mathbf{E}_{x_*}||\alpha\tilde{R}_\alpha(x_*-x_0)+\tilde{R}_\alpha\Sigma^{2t}U||^2+4\mathbf{E}_{x_*}||\alpha(R_\alpha-\tilde{R}_\alpha)(x_*-x_0)||^2 + 4\mathbf{E}_{x_*}||(R_\alpha K^{2t}-\tilde{R}_\alpha\Sigma^{2t})U||^2=:2\mathcal{A}_1+4\mathcal{A}_2+4\mathcal{A}_3.$$

The rate of term $\mathcal{A}_1$ is given by Proposition 2 with $\delta=n^{-1}$: $\mathcal{A}_1\asymp\left(\alpha^{\frac{2\tilde{b}_0-1}{2(s+a)}}+n^{-1}\alpha^{-\frac{2a+1}{2(s+a)}}\right)$. By using the result of Lemma 1 and a Taylor expansion of the second order of the function Σ^{2t+1} around K (that is, $\Sigma^{2t+1}=K^{2t+1}+(2t+1)K^{2t}([\Sigma-K])+\mathcal{O}_p([\Sigma-K]^2)$) we obtain

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

$$\begin{aligned}\mathcal{A}_2 &= \mathbf{E}_{x_*}||\alpha(R_\alpha - \tilde{R}_\alpha)(x_* - x_0)||^2 = \mathbf{E}_{x_*}||R_\alpha(\Sigma^{2t+1} - K^{2t+1})\alpha\tilde{R}_\alpha)(x_* - x_0)||^2 \\ &= \mathbf{E}_{x_*}||R_\alpha\left((2t+1)K^{2t}([\Sigma - K]) + \mathcal{O}_p([\Sigma - K]^2)\right)\alpha\tilde{R}_\alpha)(x_* - x_0)||^2 \\ &\leq \mathcal{O}\left(\mathbf{E}_{x_*}||R_\alpha K^{2t}||^2\mathbf{E}_{x_*}||\Sigma - K||^2\mathbf{E}_{x_*}||\alpha\tilde{R}_\alpha(x_* - x_0)||^2\right) \\ &\quad + \mathcal{O}\left(\mathbf{E}_{x_*}||R_\alpha||^2 n^{-2}\mathbf{E}_{x_*}||\alpha\tilde{R}_\alpha(x_* - x_0)||^2\right) \\ &= \mathcal{O}\left(n^{-1}\alpha^{-\frac{2a}{s+a}}\alpha^{\frac{2\tilde{b}_0-1}{2(s+a)}}\right) + \mathcal{O}\left(n^{-2}\alpha^{-2}\alpha^{\frac{2\tilde{b}_0-1}{2(s+a)}}\right),\end{aligned}$$

where we have used the facts that: $\mathbf{E}_{x_*}||R_\alpha K^{2t}||^2 = \mathcal{O}\left(\alpha^{-\frac{2a}{s+a}}\right)$ and, by Lemma 1, $\mathbf{E}_{x_*}||(\Sigma - K)^2||^2 = \mathcal{O}_p\left(n^{-2}\right)$. More precisely, the latter rate is obtained as follows: for some basis $\{\varphi_j\}$

$$\begin{aligned}||(\Sigma - K)^2||^2 &:= \sup_{||\phi||\leq 1}\sum_{j=1}^{\infty}\langle(\Sigma - K)^2\phi, \varphi_j\rangle^2 = \sup_{||\phi||\leq 1}\sum_{j=1}^{\infty}\langle(\Sigma - K)\phi, (\Sigma - K)\varphi_j\rangle^2 \\ &\leq \sup_{||\phi||\leq 1}\sum_{j=1}^{\infty}||(\Sigma - K)\phi||^2||(\Sigma - K)\varphi_j||^2 \\ &= ||(\Sigma - K)||^2\sum_{j=1}^{\infty}||(\Sigma - K)\varphi_j||^2 = ||(\Sigma - K)||^2||(\Sigma - K)||^2_{HS}.\end{aligned}$$

Thus, the proof of Lemma 1 implies: $\mathbf{E}_{x_*}||(\Sigma - K)^2||^2 \leq \left(\mathbf{E}_{x_*}||\Sigma - K||^2_{HS}\right)^2 = \mathcal{O}\left(n^{-2}\right)$. Next,

$$\begin{aligned}\mathcal{A}_3 &= \mathbf{E}_{x_*}||(R_\alpha K^{2t} - \tilde{R}_\alpha\Sigma^{2t})U||^2 = \mathbf{E}_{x_*}||(R_\alpha K^{2t} - \Sigma^{2t}\tilde{R}_\alpha)U||^2 \\ &= \mathbf{E}_{x_*}||R_\alpha\left[\alpha(K^{2t} - \Sigma^{2t}) + K^{2t}(\Sigma - K)\Sigma^{2t}\right]\tilde{R}_\alpha U||^2 \\ &\leq 2\mathbf{E}_{x_*}||R_\alpha\alpha(K^{2t} - \Sigma^{2t})\tilde{R}_\alpha U||^2 + 2\mathbf{E}_{x_*}||R_\alpha K^{2t}(\Sigma - K)\Sigma^{2t}\tilde{R}_\alpha U||^2\end{aligned}$$

and $2\mathbf{E}_{x_*}||R_\alpha K^{2t}(\Sigma - K)\Sigma^{2t}\tilde{R}_\alpha U||^2 = \mathcal{O}\left(\alpha^{-\frac{2a}{s+a}}n^{-2}\alpha^{-\frac{2a+1}{2(s+a)}}\right)$ by Lemma 1 and since $\mathbf{E}_{x_*}||R_\alpha K^{2t}||^2 = \mathcal{O}\left(\alpha^{-\frac{2a}{s+a}}\right)$ and $\mathbf{E}_{x_*}||\Sigma^{2t}\tilde{R}_\alpha U||^2 = \mathcal{O}\left(n^{-1}\alpha^{-\frac{2a+1}{2(s+a)}}\right)$. To obtain the rate for the first term $\mathbf{E}_{x_*}||R_\alpha\alpha(K^{2t} - \Sigma^{2t})\tilde{R}_\alpha U||^2$ we use a Taylor expansion of the second order of the function K^{2t} around Σ (that is, $K^{2t} = \Sigma^{2t} + 2t\Sigma^{2t-1}(K - \Sigma) + \mathcal{O}([K - \Sigma]^2)$ and the $\mathcal{O}_p([K - \Sigma]^2)$ term is negligible as shown above) and obtain:

$$\begin{aligned}\mathbf{E}_{x_*}||R_\alpha\alpha(K^{2t} - \Sigma^{2t})\tilde{R}_\alpha U||^2 &= \mathcal{O}\left(\mathbf{E}_{x_*}||R_\alpha\alpha||\mathbf{E}_{x_*}||K - \Sigma||^2\mathbf{E}_{x_*}||\Sigma^{2t-1}\tilde{R}_\alpha U||^2\right) \\ &= \mathcal{O}\left(n^{-2}\alpha^{-\frac{3a+1/2}{(s+a)}}\right)\end{aligned}$$

by the result of Lemma 1 and since $\mathbf{E}_{x_*}||\alpha R_\alpha||^2 = \mathcal{O}(1)$ and $\mathbf{E}_{x_*}||\Sigma^{2t-1}\tilde{R}_\alpha U||^2 = \mathcal{O}\left(n^{-1}\alpha^{-\frac{3a+1/2}{s+a}}\right)$. Remark that to recover the rate for $\mathbf{E}_{x_*}||\Sigma^{2t-1}\tilde{R}_\alpha U||^2$ a procedure

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press

similar to the one for recovering the rate for term $\mathcal{A}_2$ in Proposition 1 has been used. By putting all these results together we obtain

$$\mathbf{E}_{x_*}||\hat{x}_\alpha - x_*||^2 = \mathcal{O}\Bigg(\alpha^{\frac{2\tilde{b}_0-1}{2(s+a)}} + n^{-1}\alpha^{-\frac{2a+1}{2(s+a)}} + n^{-1}\alpha^{-\frac{2a}{s+a}}\alpha^{\frac{2\tilde{b}_0-1}{2(s+a)}} + n^{-2}\alpha^{-2}\alpha^{\frac{2\tilde{b}_0-1}{2(s+a)}} + n^{-2}\alpha^{-\frac{6a+1}{2(s+a)}} + \alpha^{-\frac{2a}{s+a}}n^{-2}\alpha^{-\frac{2a+1}{2(s+a)}}\Bigg).$$

Finally, by replacing α with $\alpha_* \asymp n^{-\frac{a+s}{\tilde{b}_0+a}}$, the third to sixth terms are negligible with respect to the first and second terms if $\tilde{b}_0 > s$ and $\tilde{b}_0 > a$. This concludes the proof.

https://doi.org/10.1017/S0266466614000796 Published online by Cambridge University Press