I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2014/06

Adaptive Bayesian estimation in Gaussian sequence space models

JOHANNES, J., SCHENK R. AND A. SIMONI

Adaptive Bayesian estimation in Gaussian sequence space models

Jan Johannes, Rudolf Schenk and Anna Simoni

Abstract We consider the inverse problem of recovering a signal θ in a Gaussian sequence space model (GSSM). We adopt a Bayes procedure and study its frequentist properties. We first derive lower and upper bounds for the posterior concentration rate over a family of Gaussian prior distributions indexed by a tuning parameter m. Under a suitable choice of m we derive a concentration rate uniformly over a class of parameters θ and show that this rate coincides with the minimax rate. Then, we construct a hierarchical fully data-driven Bayes procedure and show that it is minimax adaptive.

Introduction

Over the last decade, there has been growing interest in statistical inverse problems (see, *e.g.*, [11], [6], [5] and references therein) due to the fact that they are widely used in many fields of science. Mathematical statistics has paid special attention to minimax estimation and adaptation. Inference for inverse problems in general requires to choose a tuning parameter which is challenging in practice. Minimax estimation is achieved if the tuning parameter is set to an optimal value which relies on knowledge of the smoothness of the unknown parameter of interest. Since this smoothness is unknown, it is necessary to design a feasible procedure to select the tuning parameter which adapts to the unknown regularity of the underlying function and achieves the minimax rate. To obtain such an adaptive procedure it seems natural to adopt a Bayesian point of view where this tuning parameter can be endowed with a prior. As the theory for a general inverse problem – with a possibly unknown or noisy operator – is technically highly involved, we consider as a starting point an indirect Gaussian regression which is well known to be equivalent to an indirect GSSM (in a LeCam sense). Let ℓ_2 be the Hilbert space of square summable real valued sequences endowed with the usual inner product $\langle\cdot,\cdot\rangle_{\ell_2}$ and associated norm $\|\cdot\|_{\ell_2}$. In a GSSM we consider the inverse problem of recovering a signal $\theta = (\theta_j)_{j\geqslant 1} \in \ell_2$ from a version that is blurred by Gaussian white noise. We adopt a Bayesian approach, where the conditional distribution of the observations given the parameter is Gaussian, *i.e.*, $\boldsymbol{Y}_j \,|\, \boldsymbol{\vartheta}_j = \theta_j \sim \mathcal{N}\big(\lambda_j\theta_j, \varepsilon\big)$, independent, for $j \in \mathbb{N}$, with noise level $\varepsilon > 0$. The sequence $(\lambda_j)_{j\geqslant 1}$, λ for short, represents the operator which transforms the

Jan Johannes, Rudolf Schenk
Université catholique de Louvain, Belgium, e-mail: {jan.johannes|rudolf.schenk}@uclouvain.be

Anna Simoni
CNRS and Thema, Université de Cergy-Pontoise, France, e-mail: simoni.anna@gmail.com

signal. We consider a Gaussian prior on $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_j)_{j\geqslant 1}$ and focus on asymptotic frequentist properties of its posterior distribution, that is, we are interested in the rate at which the posterior distribution concentrates towards a point mass on the value of θ that generates the data. For a more detailed discussion see [1, 2, 7] and [8]. Our first contribution is to obtain a lower and upper bound of the concentration rate of the posterior distribution uniformly over a class of parameters. It is interesting to note that [4] derives a similar result in a direct GSSM, where the obtained rate may be up to a logarithmic factor slower than the minimax rate given in, *e.g.*, [9]. However, the rate derived in this paper is shown to coincide with the minimax rate when the hyperparameter of the prior is suitably chosen. Our second contribution consists in introducing a hierarchical structure similar to the compound prior considered by [12] and [13] and showing that the corresponding fully data-driven Bayes procedure achieves minimax adaptive inference for the GSSM. The proofs are given in [10].

1.1 Basic model assumptions

Let us consider a Gaussian prior distribution for the parameter $\boldsymbol{\vartheta}$, that is $\{\boldsymbol{\vartheta}_j\}_{j\geqslant 1}$ are independent, normally distributed with prior means $(\theta_j^{\times})_{j\geqslant 1}$ and prior variances $(\varsigma_j)_{j\geqslant 1}$, *i.e.*, $\boldsymbol{\vartheta}_j \sim \mathcal{N}(\theta_j^{\times}, \varsigma_j)$, independent, for $j \in \mathbb{N}$. Standard calculus shows that the posterior distribution of $\boldsymbol{\vartheta}$ given $\boldsymbol{Y} = (\boldsymbol{Y}_j)_{j\geqslant 1}$ is Gaussian, that is, given $\boldsymbol{Y}$, $\{\boldsymbol{\vartheta}_j\}_{j\geqslant 1}$ are conditionally independent, normally distributed random variables with posterior variance $\sigma_j^2 := \mathbb{V}\mathrm{ar}(\boldsymbol{\vartheta}_j \,|\, \boldsymbol{Y}) = (\lambda_j^2\varepsilon^{-1} + \varsigma_j^{-1})^{-1}$ and posterior mean $\theta_j^{\boldsymbol{Y}} := \mathbb{E}[\boldsymbol{\vartheta}_j \,|\, \boldsymbol{Y}] = \sigma_j^2(\varsigma_j^{-1}\theta_j^{\times} + \lambda_j\varepsilon^{-1}\,\boldsymbol{Y}_j)$, for all $j \in \mathbb{N}$. Taking this as a starting point, we construct a sequence of hierarchical prior distributions. To be more precise, let us denote by δ_x the Dirac measure in the point x. Given $m \in \mathbb{N}$, we consider the independent random variables $\{\boldsymbol{\vartheta}_j^m\}_{j\geqslant 1}$ with marginal distributions $\boldsymbol{\vartheta}_j^m \sim \mathcal{N}(\theta_j^{\times}, \varsigma_j)$, $1 \leqslant j \leqslant m$ and $\boldsymbol{\vartheta}_j^m \sim \delta_{\theta_j^{\times}}$, $m < j$, independent, $j \in \mathbb{N}$, resulting in the degenerate prior distribution $P_{\boldsymbol{\vartheta}^m}$. Here, we use the notation $\boldsymbol{\vartheta}^m = (\boldsymbol{\vartheta}_j^m)_{j\geqslant 1}$. Consequently, $\{\boldsymbol{\vartheta}_j^m\}_{j\geqslant 1}$ are conditionally independent given $\boldsymbol{Y}$ and their posterior distribution $P_{\boldsymbol{\vartheta}^m \,|\, \boldsymbol{Y}}$ is Gaussian with mean $\theta_j^{\boldsymbol{Y}}$ and variance σ_j^2 for $1 \leqslant j \leqslant m$ while being degenerate on $\theta_j^{\times}$ for $j > m$. Let $\mathbb{1}_A$ denote the indicator function which takes the value one if the condition A holds true, and the value zero otherwise. Hence, the common Bayes estimate $\widehat{\theta}^m := \mathbb{E}[\boldsymbol{\vartheta}^m \,|\, \boldsymbol{Y}]$ is given for $j \geqslant 1$ by $\widehat{\theta}_j^m := \theta_j^{\boldsymbol{Y}}\,\mathbb{1}_{\{j\,\leqslant\, m\}} + \theta_j^{\times}\,\mathbb{1}_{\{j\,>\,m\}}$.

From a Bayesian point of view, the thresholding parameter m plays the role of a hyperparameter and hence, we may complete the prior specification by introducing a prior distribution on it. Consider a random thresholding parameter $\boldsymbol{M}$ taking its values in $\{1, \ldots, G_\varepsilon\}$ for some $G_\varepsilon \in \mathbb{N}$ with prior distribution $P_{\boldsymbol{M}}$. Both G_ε and $P_{\boldsymbol{M}}$ will be specified in section **1.3**. Conditionally on $\boldsymbol{M}$, the distributions of the random variables $\{\boldsymbol{Y}_j\}_{j\geqslant 1}$ and $\{\boldsymbol{\vartheta}_j^{\boldsymbol{M}}\}_{j\geqslant 1}$ are determined by $\boldsymbol{Y}_j = \lambda_j\,\boldsymbol{\vartheta}^{\boldsymbol{M}} + \sqrt{\varepsilon}\zeta_j$ and $\boldsymbol{\vartheta}_j^{\boldsymbol{M}} = \theta_j^{\times} + \sqrt{\varsigma_j}\eta_j\,\mathbb{1}_{\{1\,\leqslant\, j\,\leqslant\, \boldsymbol{M}\}}$ where $\{\zeta_j, \eta_j\}_{j\geqslant 1}$ are iid. standard normal random variables independent of $\boldsymbol{M}$. Furthermore the posterior mean $\widehat{\theta} := \mathbb{E}[\boldsymbol{\vartheta}^{\boldsymbol{M}} \,|\, \boldsymbol{Y}]$ is the Bayes estimate which satisfies $\widehat{\theta}_j = \theta_j^{\times}$ for $j > G_\varepsilon$ and for all $1 \leqslant j \leqslant G_\varepsilon$ $\widehat{\theta}_j = \theta_j^{\times}\,P(1 \leqslant \boldsymbol{M} \leqslant j-1|\,\boldsymbol{Y}) + \theta_j^{\boldsymbol{Y}}\,P(j \leqslant \boldsymbol{M} \leqslant G_\varepsilon|\,\boldsymbol{Y})$.

1.2 Optimal concentration rate

Conditional on $\boldsymbol{Y}$, the random variables $\{\boldsymbol{\vartheta}_j^m - \theta_{oj}\}_{j=1}^m$ are independent and normally distributed with conditional mean $\theta_j^{\boldsymbol{Y}} - \theta_{oj}$ and conditional variance σ_j^2. The next assertion presents a version of tail bounds for sums of independent squared Gaussian random variables which is due to [3].

Lemma 1.1. *Let $\{X_j\}_{j\geqslant 1}$ be independent and normally distributed r.v. with mean $\alpha_j \in \mathbb{R}$ and standard deviation $\beta_j \geqslant 0$, $j \in \mathbb{N}$. For $m \in \mathbb{N}$ set $S_m := \sum_{j=1}^m X_j^2$ and consider $v_m \geqslant \sum_{j=1}^m \beta_j^2$, $t_m \geqslant \max_{1\leqslant j\leqslant m} \beta_j^2$ and $r_m \geqslant \sum_{j=1}^m \alpha_j^2$. Then for all $c \geqslant 0$ we have*

$$\sup_{m\geqslant 1} \exp\big(\tfrac{1}{4}c(c\wedge 1)(v_m + 2r_m)t_m^{-1}\big) P\big(S_m - \mathbb{E}S_m \leqslant -c(v_m + 2r_m)\big) \leqslant 1;$$
$$\sup_{m\geqslant 1} \exp\big(\tfrac{1}{4}c(c\wedge 1)(v_m + 2r_m)t_m^{-1}\big) P\big(S_m - \mathbb{E}S_m \geqslant \tfrac{3c}{2}(v_m + 2r_m)\big) \leqslant 1.$$

A major step towards establishing a concentration rate of the posterior distribution consists in finding a finite sample bound for a fixed $m \in \mathbb{N}$. We express these bounds in terms of

$$\mathfrak{b}_m := \textstyle\sum_{j>m}(\theta_{oj} - \theta_j^\times)^2, \quad \mathfrak{v}_m := \sum_{j=1}^m \sigma_j^2 \quad \text{with } \sigma_j^2 = (\lambda_j^2\varepsilon^{-1} + \varsigma_j^{-1})^{-1};$$
$$\mathfrak{t}_m := \max_{1\leqslant j\leqslant m} \sigma_j^2 \quad \text{and} \quad \mathfrak{r}_m := \textstyle\sum_{j=1}^m (\mathbb{E}_{\theta_o}[\theta_j^{\boldsymbol{Y}}] - \theta_{oj})^2 = \sum_{j=1}^m \sigma_j^4(\varsigma_j^{-2}(\theta_j^\times - \theta_{oj})^2).$$

The desired convergence to zero of all the aforementioned sequences necessitates to consider appropriate subsequences in dependence of the noise level ε, notably $(\mathfrak{v}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$, $(\mathfrak{t}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$ and $(\mathfrak{r}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$.

Assumption 1.2. *There exist constants $0 < \varepsilon_o := \varepsilon_o(\theta_o, \lambda, \theta^\times, \varsigma) < 1$ and $0 < K := K(\theta_o, \lambda, \theta^\times, \varsigma) < \infty$ such that the prior distribution satisfies the condition $\sup_{0<\varepsilon<\varepsilon_o}(\mathfrak{r}_{m_\varepsilon} \vee m_\varepsilon\mathfrak{t}_{m_\varepsilon})/(\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}) \leqslant K$.*

Proposition 1.3. *Under Assumption 1.2, for all $0 < \varepsilon < \varepsilon_o$ and $0 < c < 1/(8K)$:*

$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 > (4 + (11/2)K)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}]) \leqslant 2\exp(-\tfrac{m_\varepsilon}{36});$$
$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 < (1 - 8\,c\,K)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}]) \leqslant 2\exp(-c^2 m_\varepsilon).$$

Thereby, if we assume in addition that $\mathfrak{v}_{m_\varepsilon} = o(1)$ and $m_\varepsilon \to \infty$ as $\varepsilon \to 0$ then we obtain by the dominated convergence theorem that also $\mathfrak{b}_{m_\varepsilon} = o(1)$. Hence, $(\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$ converges to zero and is indeed a posterior concentration rate.

Theorem 1.4 (Posterior consistency)**.** *Under Assumption 1.2 if $m_\varepsilon \to \infty$ and $\mathfrak{v}_{m_\varepsilon} = o(1)$ as $\varepsilon \to 0$, then*

$$\lim_{\varepsilon\to 0} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}((1 - 8cK)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}] \leqslant \|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 \leqslant (4 + 11K/2)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}]) = 1.$$

The last assertion shows that $(\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$ is up to a constant a lower and upper bound of the concentration rate.

Proposition 1.5 (Bayes estimate consistency)**.** *Let the assumptions of Theorem 1.4 be satisfied and $\widehat{\theta}^{m_\varepsilon} := \mathbb{E}[\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}]$ then $\mathbb{E}_{\theta_o}\|\widehat{\theta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 \leqslant (3 + K)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}]$ and consequently $\mathbb{E}_{\theta_o}\|\widehat{\theta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 = o(1)$ as $\varepsilon \to 0$.*

The previous results are obtained under Assumption **1.2**. However, it may be difficult to verify whether a prior specification satisfies such an assumption. Therefore, we now introduce an assumption which states a more precise requirement on the prior variance and that can be more easily verified.

Assumption 1.6. *Define* $\Lambda_j := \lambda_j^{-2}$, $j \geqslant 1$, $\Lambda_{(m)} := \max_{1\leqslant j\leqslant m} \Lambda_j$ *and* $\overline{\Lambda}_m := m^{-1}\sum_{j=1}^m \Lambda_j$, $m \geqslant 1$. *There exist constants* $\varepsilon_o \in (0,1)$ *and* $d > 0$ *such that* $\varsigma_j \geqslant d[\varepsilon^{1/2}\Lambda_j^{1/2} \vee \varepsilon\Lambda_j]$ *for all* $1 \leqslant j \leqslant m_\varepsilon$ *and for all* $0 < \varepsilon < \varepsilon_o$.

If there exists in addition to Assumption **1.6** a strictly positive constant $L := L(\theta_o, \lambda, \theta^\times) < \infty$ such that

$$\sup_{0<\varepsilon<\varepsilon_o} \varepsilon\, m_\varepsilon\, \Lambda_{(m_\varepsilon)}\{\mathfrak{b}_{m_\varepsilon} \vee \varepsilon\, m_\varepsilon\, \overline{\Lambda}_{m_\varepsilon}\}^{-1} \leqslant L \tag{1.1}$$

holds true, then Assumption **1.2** is satisfies with $K := (1 \vee d^{-2}\|\theta_o - \theta^\times\|_{\ell_2}^2)L$.

Corollary 1.7. *Let Assumption 1.6 and* (1.1) *be satisfied, then for all* $0 < \varepsilon < \varepsilon_o$ *and* $0 < c < 1/(8K)$ *we have*

$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 > (4 + (11/2)K)[\mathfrak{b}_{m_\varepsilon} \vee \varepsilon m_\varepsilon \overline{\Lambda}_{m_\varepsilon}]) \leqslant 2\exp(-\tfrac{m_\varepsilon}{36});$$
$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 < (1 - 8\,c\,K)(1 + d^{-1})^{-1}[\mathfrak{b}_{m_\varepsilon} \vee \varepsilon m_\varepsilon \overline{\Lambda}_{m_\varepsilon}]) \leqslant 2\exp(-c^2 m_\varepsilon).$$

Under the conditions of the last assertion, the sequence $(\mathfrak{b}_{m_\varepsilon} \vee \varepsilon m_\varepsilon \overline{\Lambda}_{m_\varepsilon})_{m_\varepsilon \geqslant 1}$ provides up to constants a lower and upper bound for the concentration rate. The result implies consistency if $(\mathfrak{b}_{m_\varepsilon} \vee \varepsilon m_\varepsilon \overline{\Lambda}_{m_\varepsilon})_{m_\varepsilon \geqslant 1}$ as $\varepsilon \to 0$ but it does not answer the question of optimality in a satisfactory way. Observe that the rate depends on the parameter of interest θ_o and we could optimize the rate for each θ_o separately, but we are rather interested in a uniform rate over a class of parameters. Given a strictly positive non-decreasing sequence $\mathfrak{a} = \big(\mathfrak{a}_j\big)_{j\geqslant 1}$ with $\mathfrak{a}_1 = 1$ tending to infinity consider for $\theta \in \ell_2$ its weighted norm $\|\theta\|_{\mathfrak{a}}^2 := \sum_{j\geqslant 1} \mathfrak{a}_j \theta_j^2$. We define $\ell_2^{\mathfrak{a}}$ as the completion of ℓ_2 with respect to $\|\cdot\|_{\mathfrak{a}}$. In order to formulate the optimality of the posterior concentration rate let us define

$$m_\varepsilon^\star := m_\varepsilon^\star(\mathfrak{a}, \lambda) := \operatorname*{arg\,min}_{m\geqslant 1}[\mathfrak{a}_m^{-1} \vee \varepsilon\, m\, \overline{\Lambda}_m] \text{ and}$$
$$\mathcal{R}_\varepsilon^\star := \mathcal{R}_\varepsilon^\star\big(\mathfrak{a}, \lambda\big) := [\mathfrak{a}_{m_\varepsilon^\star}^{-1} \vee \varepsilon\, m_\varepsilon^\star\, \overline{\Lambda}_{m_\varepsilon^\star}] \quad \text{for all } \varepsilon > 0.$$

We introduce a further assumption in order to get the next theorem.

Assumption 1.8. *Let* $\mathfrak{a}$ *and* λ *be sequences such that*

$$0 < \kappa := \kappa(\mathfrak{a}, \lambda) := \inf_{0<\varepsilon<\varepsilon_o} \{(\mathcal{R}_\varepsilon^\star)^{-1}[\mathfrak{a}_{m_\varepsilon^\star}^{-1} \wedge \varepsilon\, m_\varepsilon^\star\, \overline{\Lambda}_{m_\varepsilon^\star}]\} \leqslant 1.$$

We illustrate the last assumption for typical choices of the sequences $\mathfrak{a}$ and λ. For two strictly positive sequences $(a_j)_{j\geqslant 1}$ and $(b_j)_{j\geqslant 1}$ we write $a_j \sim b_j$, if $(a_j/b_j)_{j\geqslant 1}$ is bounded away from 0 and infinity.

[P-P] Consider $\mathfrak{a}_j \sim j^{2p}$ and $\lambda_j^2 \sim j^{-2a}$ with $p > 0$ and $a > 0$ then $m_\varepsilon^\star \sim \varepsilon^{-1/(2p+2a+1)}$ and $\mathcal{R}_\varepsilon^\star \sim \varepsilon^{2p/(2a+2p+1)}$.

[E-P] Consider $\mathfrak{a}_j \sim \exp(j^{2p}-1)$ and $\lambda_j^2 \sim j^{-2a}$ with $p > 0$ and $a > 0$ then $m_\varepsilon^\star \sim |\log\varepsilon - \frac{2a+1}{2p}(\log|\log\varepsilon|)|^{1/(2p)}$ and $\mathcal{R}_\varepsilon^\star \sim \varepsilon|\log\varepsilon|^{(2a+1+2s)/(2p)}$.

[P-E] Consider $\mathfrak{a}_j \sim j^{2p}$ and $\lambda_j^2 \sim \exp(-j^{2a}+1)$, with $p > 0$ and $a > 0$ then $m_\varepsilon^\star \sim |\log\varepsilon - \frac{2p+(2a-1)_+}{2a}(\log|\log\varepsilon|)|^{1/(2a)}$ and $\mathcal{R}_\varepsilon^\star \sim |\log\varepsilon|^{-(p-s)/a}$.

In all three cases Assumption **1.8** holds true. We assume in the following that the parameter θ_o belongs to the ellipsoid $\Theta_\mathfrak{a}^r := \{\theta \in \ell_2^\mathfrak{a} : \|\theta - \theta^\times\|_\mathfrak{a}^2 \leqslant r\}$ and therefore, $\mathfrak{b}_m \leqslant \mathfrak{a}_m^{-1} r$. In addition we suppose that

$$\tilde{L} := \tilde{L}(\mathfrak{a}, \lambda) := \sup_{0<\varepsilon<\varepsilon_o} \varepsilon\, m_\varepsilon^\star\, \Lambda_{(m_\varepsilon^\star)} (\mathcal{R}_\varepsilon^\star)^{-1} < \infty. \tag{1.2}$$

We note that under Assumption **1.8** and (1.2) the condition (1.1) is satisfied uniformly for all $\theta_o \in \Theta_\mathfrak{a}^r$ with $L = \tilde{L}/\kappa$.

Theorem 1.9 (Optimal posterior concentration rate)**.** *Suppose that the sequence of prior distributions $(P_{\boldsymbol{\vartheta}^{m_\varepsilon^\star}})_{m_\varepsilon^\star}$ satisfies Assumption **1.6** and let Assumption **1.8** and* (1.2) *be satisfied. Then there exists a constant $K := K(\Theta_\mathfrak{a}^r, \lambda, d, \kappa)$ such that*

$$\lim_{\varepsilon\to 0} \inf_{\theta_o \in \Theta_\mathfrak{a}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon^\star} \mid \boldsymbol{Y}} (\|\boldsymbol{\vartheta}^{m_\varepsilon^\star} - \theta_o\|_{\ell_2}^2 \leqslant K\mathcal{R}_\varepsilon^\star) = 1, \quad \textit{moreover,}$$

if $\Psi_\varepsilon/\mathcal{R}_\varepsilon^\star = o(1)$ as $\varepsilon \to 0$ then $\lim_{\varepsilon\to 0} \sup_{\theta_o \in \Theta_\mathfrak{a}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon^\star} \mid \boldsymbol{Y}} (\|\boldsymbol{\vartheta}^{m_\varepsilon^\star} - \theta_o\|_{\ell_2}^2 \leqslant \Psi_\varepsilon) = 0$.

Remark 1.10. *The rate $\mathcal{R}_\varepsilon^\star = \mathcal{R}_\varepsilon^\star(\Theta_\mathfrak{a}^r, \lambda)$ is optimal in a minimax sense. To be more precise, given an estimator $\widehat{\theta}$ of θ let $\sup_{\theta\in\Theta_\mathfrak{a}^r} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|^2$ denote the maximal mean integrated squared error (MISE) over the class $\Theta_\mathfrak{a}^r$. It has been shown in [9] that $\mathcal{R}_\varepsilon^\star$ provides up to a constant a lower bound for the maximal MISE over the class $\Theta_\mathfrak{a}^r$ and that there exists an estimator attaining this rate.*

Proposition 1.11 (Minimax-optimal Bayes estimate)**.** *Let the assumptions of Theorem **1.12** be satisfied and $\widehat{\theta}^{m_\varepsilon^\star} := \mathbb{E}[\boldsymbol{\vartheta}^{m_\varepsilon^\star} \mid \boldsymbol{Y}]$ then there exists a constant $K := K(\Theta_\mathfrak{a}^r, \lambda, d, \kappa)$ such that $\sup_{\theta_o\in\Theta_\mathfrak{a}^r} \mathbb{E}_{\theta_o} \|\widehat{\theta}^{m_\varepsilon^\star} - \theta_o\|_{\ell_2}^2 \leqslant K\mathcal{R}_\varepsilon^\star$.*

1.3 Adaptive Bayesian estimation

We will derive a concentration rate given the aforementioned hierarchical prior distribution. For this purpose set $G_\varepsilon := \max\{m \in \mathbb{N} : \varepsilon\Lambda_{(m)} \leqslant 1\}$ and

$$p_{\boldsymbol{M}}(m) = \frac{\exp(\frac{-3m}{2\varepsilon}) \prod_{j=1}^m (1 + \lambda_j^2 \varsigma_j \varepsilon^{-1})^{1/2}}{\sum_{m'=1}^{G_\varepsilon} \exp(\frac{-3m'}{2\varepsilon}) \prod_{j=1}^{m'} (1 + \lambda_j^2 \varsigma_j \varepsilon^{-1})^{1/2}} \quad \text{for } 1 \leqslant m \leqslant G_\varepsilon\,.$$

Theorem 1.12 (Optimal posterior concentration rate)**.** *Suppose that the sequence of prior distributions* $(P_{\boldsymbol{\vartheta}^{G_\varepsilon}})_{G_\varepsilon}$ *satisfies Assumption 1.6 and in addition that* $m_\varepsilon^\star$ *satisfies Assumption 1.8 and* (1.2)*. Then there exists a constant* $K := K(\Theta_{\mathfrak{a}}^r, \lambda, d, \kappa)$ *such that*

$$\lim_{\varepsilon\to 0} \inf_{\theta_o\in\Theta_{\mathfrak{a}}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{\boldsymbol{M}} \mid \boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{\boldsymbol{M}} - \theta_o\|_{\ell_2}^2 \leqslant K\mathcal{R}_\varepsilon^\star) = 1, \quad \textit{moreover,}$$

if $\Psi_\varepsilon/\mathcal{R}_\varepsilon^\star = o(1)$ *as* $\varepsilon \to 0$ *then* $\lim_{\varepsilon\to 0} \sup_{\theta_o\in\Theta_{\mathfrak{a}}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{\boldsymbol{M}} \mid \boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{\boldsymbol{M}} - \theta_o\|_{\ell_2}^2 \leqslant \Psi_\varepsilon) = 0.$

We shall emphasize that the concentration rate derived from the hierarchical prior coincides with the minimax optimal rate $\mathcal{R}_\varepsilon^\star = \mathcal{R}_\varepsilon^\star\big(\Theta_{\mathfrak{a}}^r, \lambda\big)$ of the maximal MISE over the class $\Theta_{\mathfrak{a}}^r$. In particular this prior does not involve any knowledge of the class $\Theta_{\mathfrak{a}}^r$, therefore, the corresponding Bayes estimate is fully-data driven. The next assertion establishes its minimax-optimality.

Proposition 1.13 (Minimax-optimal Bayes estimate)**.** *Under the assumptions of Theorem 1.12. Consider the Bayes estimate* $\widehat{\theta} := \mathbb{E}[\boldsymbol{\vartheta}^{\boldsymbol{M}} \mid \boldsymbol{Y}]$ *then there exists a constant* $K := K(\Theta_{\mathfrak{a}}^r, \lambda)$ *such that* $\sup_{\theta_o\in\Theta_{\mathfrak{a}}^r} \mathbb{E}_{\theta_o}\|\widehat{\theta} - \theta_o\|_{\ell_2}^2 \leqslant K\mathcal{R}_\varepsilon^\star$ *for all* $\varepsilon > 0$*.*

Our procedure extends and completes the procedure proposed by [13] in two perspectives. First, it allows a prior variance more general than the polynomially decreasing one. Second, in addition to prove minimax-optimality of the Bayes estimator, we prove concentration at the optimal rate of the posterior distribution.

Conclusions and perspectives. We have presented a hierarchical prior leading to a fully data-driven Bayes estimate that is minimax-optimal in an indirect GSSM. Obviously, the concentration rate based on a hierarchical prior in an indirect GSSM possibly with additional noise in the eigenvalues is only one amongst the many interesting questions for further research and we are currently exploring this topic.

Acknowledgements. This work was supported by the IAP research network no. P7/06 of the Belgian Government (Belgian Science Policy), the contract "Projet d'Actions de Recherche Concertées" No 11/16-039 of the "Communauté française de Belgique" and by the "Fonds Spéciaux de Recherche" from the Université catholique de Louvain.

Bibliography

[1] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

[2] E. Belitser and S. Ghosal. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31:536–559, 2003.

[3] L. Birgé. An alternative point of view on Lepski's method. *State of the art in probability and statistics*, *IMS Lecture Notes*, 36:113–133. 2001.

[4] I. Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Statist.*, 2:1281–1299, 2008.

[5] L. Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24:1–19, 2008.

[6] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30:843–874, 2002.

[7] D. Cox. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, 21:903–923, 1993.

[8] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, pages 500–531, 2000.

[9] J. Johannes and M. Schwarz. Adaptive Gaussian inverse regression with partially unknown operator. *Communications in Statistics - Theory and Methods*, Vol. 42, No. 7, pages 1343-1362, 2013.

[10] J. Johannes, R. Schenk, and A. Simoni. Adaptive Bayesian estimation in Gaussian sequence space models. *Discussion paper at Université catholique de Louvain*, 2014.

[11] A. P. Korostelev, and A. B. Tsybakov. Minimax Theory of Image Reconstruction. *Lecture Notes in Statistics. Springer, New York.*, 82, 1993.

[12] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29:687–714, 2001.

[13] L. H. Zhao. Bayesian aspects of some nonparametric problems. *Ann. Statist.*, 28:532–552, 2000.

Adaptive Bayesian estimation in Gaussian sequence space models

Jan Johannes, Rudolf Schenk and Anna Simoni

Abstract We consider the inverse problem of recovering a signal θ in a Gaussian sequence space model (GSSM). We adopt a Bayes procedure and study its frequentist properties. We first derive lower and upper bounds for the posterior concentration rate over a family of Gaussian prior distributions indexed by a tuning parameter m. Under a suitable choice of m we derive a concentration rate uniformly over a class of parameters θ and show that this rate coincides with the minimax rate. Then, we construct a hierarchical fully data-driven Bayes procedure and show that it is minimax adaptive.

Introduction

Over the last decade, there has been growing interest in statistical inverse problems (see, *e.g.*, [11], [6], [5] and references therein) due to the fact that they are widely used in many fields of science. Mathematical statistics has paid special attention to minimax estimation and adaptation. Inference for inverse problems in general requires to choose a tuning parameter which is challenging in practice. Minimax estimation is achieved if the tuning parameter is set to an optimal value which relies on knowledge of the smoothness of the unknown parameter of interest. Since this smoothness is unknown, it is necessary to design a feasible procedure to select the tuning parameter which adapts to the unknown regularity of the underlying function and achieves the minimax rate. To obtain such an adaptive procedure it seems natural to adopt a Bayesian point of view where this tuning parameter can be endowed with a prior. As the theory for a general inverse problem – with a possibly unknown or noisy operator – is technically highly involved, we consider as a starting point an indirect Gaussian regression which is well known to be equivalent to an indirect GSSM (in a LeCam sense). Let ℓ_2 be the Hilbert space of square summable real valued sequences endowed with the usual inner product $\langle\cdot,\cdot\rangle_{\ell_2}$ and associated norm $\|\cdot\|_{\ell_2}$. In a GSSM we consider the inverse problem of recovering a signal $\theta = (\theta_j)_{j\geqslant 1} \in \ell_2$ from a version that is blurred by Gaussian white noise. We adopt a Bayesian approach, where the conditional distribution of the observations given the parameter is Gaussian, *i.e.*, $\boldsymbol{Y}_j \,|\, \boldsymbol{\vartheta}_j = \theta_j \sim \mathcal{N}\big(\lambda_j\theta_j, \varepsilon\big)$, independent, for $j \in \mathbb{N}$, with noise level $\varepsilon > 0$. The sequence $(\lambda_j)_{j\geqslant 1}$, λ for short, represents the operator which transforms the

Jan Johannes, Rudolf Schenk
Université catholique de Louvain, Belgium, e-mail: {jan.johannes|rudolf.schenk}@uclouvain.be

Anna Simoni
CNRS and Thema, Université de Cergy-Pontoise, France, e-mail: simoni.anna@gmail.com

signal. We consider a Gaussian prior on $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_j)_{j\geqslant 1}$ and focus on asymptotic frequentist properties of its posterior distribution, that is, we are interested in the rate at which the posterior distribution concentrates towards a point mass on the value of θ that generates the data. For a more detailed discussion see [1, 2, 7] and [8]. Our first contribution is to obtain a lower and upper bound of the concentration rate of the posterior distribution uniformly over a class of parameters. It is interesting to note that [4] derives a similar result in a direct GSSM, where the obtained rate may be up to a logarithmic factor slower than the minimax rate given in, *e.g.*, [9]. However, the rate derived in this paper is shown to coincide with the minimax rate when the hyperparameter of the prior is suitably chosen. Our second contribution consists in introducing a hierarchical structure similar to the compound prior considered by [12] and [13] and showing that the corresponding fully data-driven Bayes procedure achieves minimax adaptive inference for the GSSM. The proofs are given in [10].

1.1 Basic model assumptions

Let us consider a Gaussian prior distribution for the parameter $\boldsymbol{\vartheta}$, that is $\{\boldsymbol{\vartheta}_j\}_{j\geqslant 1}$ are independent, normally distributed with prior means $(\theta_j^{\times})_{j\geqslant 1}$ and prior variances $(\varsigma_j)_{j\geqslant 1}$, *i.e.*, $\boldsymbol{\vartheta}_j \sim \mathcal{N}(\theta_j^{\times}, \varsigma_j)$, independent, for $j \in \mathbb{N}$. Standard calculus shows that the posterior distribution of $\boldsymbol{\vartheta}$ given $\boldsymbol{Y} = (\boldsymbol{Y}_j)_{j\geqslant 1}$ is Gaussian, that is, given $\boldsymbol{Y}$, $\{\boldsymbol{\vartheta}_j\}_{j\geqslant 1}$ are conditionally independent, normally distributed random variables with posterior variance $\sigma_j^2 := \mathbb{V}\mathrm{ar}(\boldsymbol{\vartheta}_j \,|\, \boldsymbol{Y}) = (\lambda_j^2\varepsilon^{-1} + \varsigma_j^{-1})^{-1}$ and posterior mean $\theta_j^{\boldsymbol{Y}} := \mathbb{E}[\boldsymbol{\vartheta}_j \,|\, \boldsymbol{Y}] = \sigma_j^2(\varsigma_j^{-1}\theta_j^{\times} + \lambda_j\varepsilon^{-1}\,\boldsymbol{Y}_j)$, for all $j \in \mathbb{N}$. Taking this as a starting point, we construct a sequence of hierarchical prior distributions. To be more precise, let us denote by δ_x the Dirac measure in the point x. Given $m \in \mathbb{N}$, we consider the independent random variables $\{\boldsymbol{\vartheta}_j^m\}_{j\geqslant 1}$ with marginal distributions $\boldsymbol{\vartheta}_j^m \sim \mathcal{N}(\theta_j^{\times}, \varsigma_j)$, $1 \leqslant j \leqslant m$ and $\boldsymbol{\vartheta}_j^m \sim \delta_{\theta_j^{\times}}$, $m < j$, independent, $j \in \mathbb{N}$, resulting in the degenerate prior distribution $P_{\boldsymbol{\vartheta}^m}$. Here, we use the notation $\boldsymbol{\vartheta}^m = (\boldsymbol{\vartheta}_j^m)_{j\geqslant 1}$. Consequently, $\{\boldsymbol{\vartheta}_j^m\}_{j\geqslant 1}$ are conditionally independent given $\boldsymbol{Y}$ and their posterior distribution $P_{\boldsymbol{\vartheta}^m \,|\, \boldsymbol{Y}}$ is Gaussian with mean $\theta_j^{\boldsymbol{Y}}$ and variance σ_j^2 for $1 \leqslant j \leqslant m$ while being degenerate on $\theta_j^{\times}$ for $j > m$. Let $\mathbb{1}_A$ denote the indicator function which takes the value one if the condition A holds true, and the value zero otherwise. Hence, the common Bayes estimate $\widehat{\theta}^m := \mathbb{E}[\boldsymbol{\vartheta}^m \,|\, \boldsymbol{Y}]$ is given for $j \geqslant 1$ by $\widehat{\theta}_j^m := \theta_j^{\boldsymbol{Y}}\,\mathbb{1}_{\{j\,\leqslant\, m\}} + \theta_j^{\times}\,\mathbb{1}_{\{j\,>\,m\}}$.

From a Bayesian point of view, the thresholding parameter m plays the role of a hyperparameter and hence, we may complete the prior specification by introducing a prior distribution on it. Consider a random thresholding parameter $\boldsymbol{M}$ taking its values in $\{1, \ldots, G_\varepsilon\}$ for some $G_\varepsilon \in \mathbb{N}$ with prior distribution $P_{\boldsymbol{M}}$. Both G_ε and $P_{\boldsymbol{M}}$ will be specified in section **1.3**. Conditionally on $\boldsymbol{M}$, the distributions of the random variables $\{\boldsymbol{Y}_j\}_{j\geqslant 1}$ and $\{\boldsymbol{\vartheta}_j^{\boldsymbol{M}}\}_{j\geqslant 1}$ are determined by $\boldsymbol{Y}_j = \lambda_j\,\boldsymbol{\vartheta}^{\boldsymbol{M}} + \sqrt{\varepsilon}\zeta_j$ and $\boldsymbol{\vartheta}_j^{\boldsymbol{M}} = \theta_j^{\times} + \sqrt{\varsigma_j}\eta_j\,\mathbb{1}_{\{1\,\leqslant\, j\,\leqslant\, \boldsymbol{M}\}}$ where $\{\zeta_j, \eta_j\}_{j\geqslant 1}$ are iid. standard normal random variables independent of $\boldsymbol{M}$. Furthermore the posterior mean $\widehat{\theta} := \mathbb{E}[\boldsymbol{\vartheta}^{\boldsymbol{M}} \,|\, \boldsymbol{Y}]$ is the Bayes estimate which satisfies $\widehat{\theta}_j = \theta_j^{\times}$ for $j > G_\varepsilon$ and for all $1 \leqslant j \leqslant G_\varepsilon$ $\widehat{\theta}_j = \theta_j^{\times}\,P(1 \leqslant \boldsymbol{M} \leqslant j-1|\,\boldsymbol{Y}) + \theta_j^{\boldsymbol{Y}}\,P(j \leqslant \boldsymbol{M} \leqslant G_\varepsilon|\,\boldsymbol{Y})$.

1.2 Optimal concentration rate

Conditional on $\boldsymbol{Y}$, the random variables $\{\boldsymbol{\vartheta}_j^m - \theta_{oj}\}_{j=1}^m$ are independent and normally distributed with conditional mean $\theta_j^{\boldsymbol{Y}} - \theta_{oj}$ and conditional variance σ_j^2. The next assertion presents a version of tail bounds for sums of independent squared Gaussian random variables which is due to [3].

Lemma 1.1. *Let $\{X_j\}_{j\geqslant 1}$ be independent and normally distributed r.v. with mean $\alpha_j \in \mathbb{R}$ and standard deviation $\beta_j \geqslant 0$, $j \in \mathbb{N}$. For $m \in \mathbb{N}$ set $S_m := \sum_{j=1}^m X_j^2$ and consider $v_m \geqslant \sum_{j=1}^m \beta_j^2$, $t_m \geqslant \max_{1\leqslant j\leqslant m} \beta_j^2$ and $r_m \geqslant \sum_{j=1}^m \alpha_j^2$. Then for all $c \geqslant 0$ we have*

$$\sup_{m\geqslant 1} \exp\big(\tfrac{1}{4}c(c\wedge 1)(v_m + 2r_m)t_m^{-1}\big) P\big(S_m - \mathbb{E}S_m \leqslant -c(v_m + 2r_m)\big) \leqslant 1;$$
$$\sup_{m\geqslant 1} \exp\big(\tfrac{1}{4}c(c\wedge 1)(v_m + 2r_m)t_m^{-1}\big) P\big(S_m - \mathbb{E}S_m \geqslant \tfrac{3c}{2}(v_m + 2r_m)\big) \leqslant 1.$$

A major step towards establishing a concentration rate of the posterior distribution consists in finding a finite sample bound for a fixed $m \in \mathbb{N}$. We express these bounds in terms of

$$\mathfrak{b}_m := \textstyle\sum_{j>m}(\theta_{oj} - \theta_j^\times)^2, \quad \mathfrak{v}_m := \sum_{j=1}^m \sigma_j^2 \quad \text{with } \sigma_j^2 = (\lambda_j^2\varepsilon^{-1} + \varsigma_j^{-1})^{-1};$$
$$\mathfrak{t}_m := \max_{1\leqslant j\leqslant m} \sigma_j^2 \quad \text{and} \quad \mathfrak{r}_m := \textstyle\sum_{j=1}^m (\mathbb{E}_{\theta_o}[\theta_j^{\boldsymbol{Y}}] - \theta_{oj})^2 = \sum_{j=1}^m \sigma_j^4(\varsigma_j^{-2}(\theta_j^\times - \theta_{oj})^2).$$

The desired convergence to zero of all the aforementioned sequences necessitates to consider appropriate subsequences in dependence of the noise level ε, notably $(\mathfrak{v}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$, $(\mathfrak{t}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$ and $(\mathfrak{r}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$.

Assumption 1.2. *There exist constants $0 < \varepsilon_o := \varepsilon_o(\theta_o, \lambda, \theta^\times, \varsigma) < 1$ and $0 < K := K(\theta_o, \lambda, \theta^\times, \varsigma) < \infty$ such that the prior distribution satisfies the condition $\sup_{0<\varepsilon<\varepsilon_o}(\mathfrak{r}_{m_\varepsilon} \vee m_\varepsilon\mathfrak{t}_{m_\varepsilon})/(\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}) \leqslant K$.*

Proposition 1.3. *Under Assumption 1.2, for all $0 < \varepsilon < \varepsilon_o$ and $0 < c < 1/(8K)$:*

$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 > (4 + (11/2)K)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}]) \leqslant 2\exp(-\tfrac{m_\varepsilon}{36});$$
$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 < (1 - 8\,c\,K)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}]) \leqslant 2\exp(-c^2 m_\varepsilon).$$

Thereby, if we assume in addition that $\mathfrak{v}_{m_\varepsilon} = o(1)$ and $m_\varepsilon \to \infty$ as $\varepsilon \to 0$ then we obtain by the dominated convergence theorem that also $\mathfrak{b}_{m_\varepsilon} = o(1)$. Hence, $(\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$ converges to zero and is indeed a posterior concentration rate.

Theorem 1.4 (Posterior consistency)**.** *Under Assumption 1.2 if $m_\varepsilon \to \infty$ and $\mathfrak{v}_{m_\varepsilon} = o(1)$ as $\varepsilon \to 0$, then*

$$\lim_{\varepsilon\to 0} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}((1 - 8cK)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}] \leqslant \|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 \leqslant (4 + 11K/2)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}]) = 1.$$

The last assertion shows that $(\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon})_{m_\varepsilon\geqslant 1}$ is up to a constant a lower and upper bound of the concentration rate.

Proposition 1.5 (Bayes estimate consistency)**.** *Let the assumptions of Theorem 1.4 be satisfied and $\widehat{\theta}^{m_\varepsilon} := \mathbb{E}[\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}]$ then $\mathbb{E}_{\theta_o}\|\widehat{\theta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 \leqslant (3 + K)[\mathfrak{b}_{m_\varepsilon} \vee \mathfrak{v}_{m_\varepsilon}]$ and consequently $\mathbb{E}_{\theta_o}\|\widehat{\theta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 = o(1)$ as $\varepsilon \to 0$.*

The previous results are obtained under Assumption **1.2**. However, it may be difficult to verify whether a prior specification satisfies such an assumption. Therefore, we now introduce an assumption which states a more precise requirement on the prior variance and that can be more easily verified.

Assumption 1.6. *Define* $\Lambda_j := \lambda_j^{-2}$, $j \geqslant 1$, $\Lambda_{(m)} := \max_{1\leqslant j\leqslant m} \Lambda_j$ *and* $\overline{\Lambda}_m := m^{-1}\sum_{j=1}^m \Lambda_j$, $m \geqslant 1$. *There exist constants* $\varepsilon_o \in (0,1)$ *and* $d > 0$ *such that* $\varsigma_j \geqslant d[\varepsilon^{1/2}\Lambda_j^{1/2} \vee \varepsilon\Lambda_j]$ *for all* $1 \leqslant j \leqslant m_\varepsilon$ *and for all* $0 < \varepsilon < \varepsilon_o$.

If there exists in addition to Assumption **1.6** a strictly positive constant $L := L(\theta_o, \lambda, \theta^\times) < \infty$ such that

$$\sup_{0<\varepsilon<\varepsilon_o} \varepsilon\, m_\varepsilon\, \Lambda_{(m_\varepsilon)}\{\mathfrak{b}_{m_\varepsilon} \vee \varepsilon\, m_\varepsilon\, \overline{\Lambda}_{m_\varepsilon}\}^{-1} \leqslant L \tag{1.1}$$

holds true, then Assumption **1.2** is satisfies with $K := (1 \vee d^{-2}\|\theta_o - \theta^\times\|_{\ell_2}^2)L$.

Corollary 1.7. *Let Assumption 1.6 and* (1.1) *be satisfied, then for all* $0 < \varepsilon < \varepsilon_o$ *and* $0 < c < 1/(8K)$ *we have*

$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 > (4 + (11/2)K)[\mathfrak{b}_{m_\varepsilon} \vee \varepsilon m_\varepsilon \overline{\Lambda}_{m_\varepsilon}]) \leqslant 2\exp(-\tfrac{m_\varepsilon}{36});$$
$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon}\,|\,\boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{m_\varepsilon} - \theta_o\|_{\ell_2}^2 < (1 - 8\,c\,K)(1 + d^{-1})^{-1}[\mathfrak{b}_{m_\varepsilon} \vee \varepsilon m_\varepsilon \overline{\Lambda}_{m_\varepsilon}]) \leqslant 2\exp(-c^2 m_\varepsilon).$$

Under the conditions of the last assertion, the sequence $(\mathfrak{b}_{m_\varepsilon} \vee \varepsilon m_\varepsilon \overline{\Lambda}_{m_\varepsilon})_{m_\varepsilon \geqslant 1}$ provides up to constants a lower and upper bound for the concentration rate. The result implies consistency if $(\mathfrak{b}_{m_\varepsilon} \vee \varepsilon m_\varepsilon \overline{\Lambda}_{m_\varepsilon})_{m_\varepsilon \geqslant 1}$ as $\varepsilon \to 0$ but it does not answer the question of optimality in a satisfactory way. Observe that the rate depends on the parameter of interest θ_o and we could optimize the rate for each θ_o separately, but we are rather interested in a uniform rate over a class of parameters. Given a strictly positive non-decreasing sequence $\mathfrak{a} = \big(\mathfrak{a}_j\big)_{j\geqslant 1}$ with $\mathfrak{a}_1 = 1$ tending to infinity consider for $\theta \in \ell_2$ its weighted norm $\|\theta\|_{\mathfrak{a}}^2 := \sum_{j\geqslant 1} \mathfrak{a}_j \theta_j^2$. We define $\ell_2^{\mathfrak{a}}$ as the completion of ℓ_2 with respect to $\|\cdot\|_{\mathfrak{a}}$. In order to formulate the optimality of the posterior concentration rate let us define

$$m_\varepsilon^\star := m_\varepsilon^\star(\mathfrak{a}, \lambda) := \operatorname*{arg\,min}_{m\geqslant 1}[\mathfrak{a}_m^{-1} \vee \varepsilon\, m\, \overline{\Lambda}_m] \text{ and}$$
$$\mathcal{R}_\varepsilon^\star := \mathcal{R}_\varepsilon^\star\big(\mathfrak{a}, \lambda\big) := [\mathfrak{a}_{m_\varepsilon^\star}^{-1} \vee \varepsilon\, m_\varepsilon^\star\, \overline{\Lambda}_{m_\varepsilon^\star}] \quad \text{for all } \varepsilon > 0.$$

We introduce a further assumption in order to get the next theorem.

Assumption 1.8. *Let* $\mathfrak{a}$ *and* λ *be sequences such that*

$$0 < \kappa := \kappa(\mathfrak{a}, \lambda) := \inf_{0<\varepsilon<\varepsilon_o} \{(\mathcal{R}_\varepsilon^\star)^{-1}[\mathfrak{a}_{m_\varepsilon^\star}^{-1} \wedge \varepsilon\, m_\varepsilon^\star\, \overline{\Lambda}_{m_\varepsilon^\star}]\} \leqslant 1.$$

We illustrate the last assumption for typical choices of the sequences $\mathfrak{a}$ and λ. For two strictly positive sequences $(a_j)_{j\geqslant 1}$ and $(b_j)_{j\geqslant 1}$ we write $a_j \sim b_j$, if $(a_j/b_j)_{j\geqslant 1}$ is bounded away from 0 and infinity.

[P-P] Consider $\mathfrak{a}_j \sim j^{2p}$ and $\lambda_j^2 \sim j^{-2a}$ with $p > 0$ and $a > 0$ then $m_\varepsilon^\star \sim \varepsilon^{-1/(2p+2a+1)}$ and $\mathcal{R}_\varepsilon^\star \sim \varepsilon^{2p/(2a+2p+1)}$.

[E-P] Consider $\mathfrak{a}_j \sim \exp(j^{2p}-1)$ and $\lambda_j^2 \sim j^{-2a}$ with $p > 0$ and $a > 0$ then $m_\varepsilon^\star \sim |\log\varepsilon - \frac{2a+1}{2p}(\log|\log\varepsilon|)|^{1/(2p)}$ and $\mathcal{R}_\varepsilon^\star \sim \varepsilon|\log\varepsilon|^{(2a+1+2s)/(2p)}$.

[P-E] Consider $\mathfrak{a}_j \sim j^{2p}$ and $\lambda_j^2 \sim \exp(-j^{2a}+1)$, with $p > 0$ and $a > 0$ then $m_\varepsilon^\star \sim |\log\varepsilon - \frac{2p+(2a-1)_+}{2a}(\log|\log\varepsilon|)|^{1/(2a)}$ and $\mathcal{R}_\varepsilon^\star \sim |\log\varepsilon|^{-(p-s)/a}$.

In all three cases Assumption **1.8** holds true. We assume in the following that the parameter θ_o belongs to the ellipsoid $\Theta_\mathfrak{a}^r := \{\theta \in \ell_2^\mathfrak{a} : \|\theta - \theta^\times\|_\mathfrak{a}^2 \leqslant r\}$ and therefore, $\mathfrak{b}_m \leqslant \mathfrak{a}_m^{-1} r$. In addition we suppose that

$$\tilde{L} := \tilde{L}(\mathfrak{a}, \lambda) := \sup_{0<\varepsilon<\varepsilon_o} \varepsilon\, m_\varepsilon^\star\, \Lambda_{(m_\varepsilon^\star)} (\mathcal{R}_\varepsilon^\star)^{-1} < \infty. \tag{1.2}$$

We note that under Assumption **1.8** and (1.2) the condition (1.1) is satisfied uniformly for all $\theta_o \in \Theta_\mathfrak{a}^r$ with $L = \tilde{L}/\kappa$.

Theorem 1.9 (Optimal posterior concentration rate)**.** *Suppose that the sequence of prior distributions $(P_{\boldsymbol{\vartheta}^{m_\varepsilon^\star}})_{m_\varepsilon^\star}$ satisfies Assumption **1.6** and let Assumption **1.8** and* (1.2) *be satisfied. Then there exists a constant $K := K(\Theta_\mathfrak{a}^r, \lambda, d, \kappa)$ such that*

$$\lim_{\varepsilon\to 0} \inf_{\theta_o \in \Theta_\mathfrak{a}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon^\star} \mid \boldsymbol{Y}} (\|\boldsymbol{\vartheta}^{m_\varepsilon^\star} - \theta_o\|_{\ell_2}^2 \leqslant K\mathcal{R}_\varepsilon^\star) = 1, \quad \textit{moreover,}$$

if $\Psi_\varepsilon/\mathcal{R}_\varepsilon^\star = o(1)$ as $\varepsilon \to 0$ then $\lim_{\varepsilon\to 0} \sup_{\theta_o \in \Theta_\mathfrak{a}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\varepsilon^\star} \mid \boldsymbol{Y}} (\|\boldsymbol{\vartheta}^{m_\varepsilon^\star} - \theta_o\|_{\ell_2}^2 \leqslant \Psi_\varepsilon) = 0$.

Remark 1.10. *The rate $\mathcal{R}_\varepsilon^\star = \mathcal{R}_\varepsilon^\star(\Theta_\mathfrak{a}^r, \lambda)$ is optimal in a minimax sense. To be more precise, given an estimator $\widehat{\theta}$ of θ let $\sup_{\theta\in\Theta_\mathfrak{a}^r} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|^2$ denote the maximal mean integrated squared error (MISE) over the class $\Theta_\mathfrak{a}^r$. It has been shown in [9] that $\mathcal{R}_\varepsilon^\star$ provides up to a constant a lower bound for the maximal MISE over the class $\Theta_\mathfrak{a}^r$ and that there exists an estimator attaining this rate.*

Proposition 1.11 (Minimax-optimal Bayes estimate)**.** *Let the assumptions of Theorem **1.12** be satisfied and $\widehat{\theta}^{m_\varepsilon^\star} := \mathbb{E}[\boldsymbol{\vartheta}^{m_\varepsilon^\star} \mid \boldsymbol{Y}]$ then there exists a constant $K := K(\Theta_\mathfrak{a}^r, \lambda, d, \kappa)$ such that $\sup_{\theta_o\in\Theta_\mathfrak{a}^r} \mathbb{E}_{\theta_o} \|\widehat{\theta}^{m_\varepsilon^\star} - \theta_o\|_{\ell_2}^2 \leqslant K\mathcal{R}_\varepsilon^\star$.*

1.3 Adaptive Bayesian estimation

We will derive a concentration rate given the aforementioned hierarchical prior distribution. For this purpose set $G_\varepsilon := \max\{m \in \mathbb{N} : \varepsilon\Lambda_{(m)} \leqslant 1\}$ and

$$p_{\boldsymbol{M}}(m) = \frac{\exp(\frac{-3m}{2\varepsilon}) \prod_{j=1}^m (1 + \lambda_j^2 \varsigma_j \varepsilon^{-1})^{1/2}}{\sum_{m'=1}^{G_\varepsilon} \exp(\frac{-3m'}{2\varepsilon}) \prod_{j=1}^{m'} (1 + \lambda_j^2 \varsigma_j \varepsilon^{-1})^{1/2}} \quad \text{for } 1 \leqslant m \leqslant G_\varepsilon\,.$$

Theorem 1.12 (Optimal posterior concentration rate)**.** *Suppose that the sequence of prior distributions* $(P_{\boldsymbol{\vartheta}^{G_\varepsilon}})_{G_\varepsilon}$ *satisfies Assumption 1.6 and in addition that* $m_\varepsilon^\star$ *satisfies Assumption 1.8 and* (1.2)*. Then there exists a constant* $K := K(\Theta_{\mathfrak{a}}^r, \lambda, d, \kappa)$ *such that*

$$\lim_{\varepsilon\to 0} \inf_{\theta_o\in\Theta_{\mathfrak{a}}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{\boldsymbol{M}} \mid \boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{\boldsymbol{M}} - \theta_o\|_{\ell_2}^2 \leqslant K\mathcal{R}_\varepsilon^\star) = 1, \quad \textit{moreover,}$$

if $\Psi_\varepsilon/\mathcal{R}_\varepsilon^\star = o(1)$ *as* $\varepsilon \to 0$ *then* $\lim_{\varepsilon\to 0} \sup_{\theta_o\in\Theta_{\mathfrak{a}}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{\boldsymbol{M}} \mid \boldsymbol{Y}}(\|\boldsymbol{\vartheta}^{\boldsymbol{M}} - \theta_o\|_{\ell_2}^2 \leqslant \Psi_\varepsilon) = 0.$

We shall emphasize that the concentration rate derived from the hierarchical prior coincides with the minimax optimal rate $\mathcal{R}_\varepsilon^\star = \mathcal{R}_\varepsilon^\star\big(\Theta_{\mathfrak{a}}^r, \lambda\big)$ of the maximal MISE over the class $\Theta_{\mathfrak{a}}^r$. In particular this prior does not involve any knowledge of the class $\Theta_{\mathfrak{a}}^r$, therefore, the corresponding Bayes estimate is fully-data driven. The next assertion establishes its minimax-optimality.

Proposition 1.13 (Minimax-optimal Bayes estimate)**.** *Under the assumptions of Theorem 1.12. Consider the Bayes estimate* $\widehat{\theta} := \mathbb{E}[\boldsymbol{\vartheta}^{\boldsymbol{M}} \mid \boldsymbol{Y}]$ *then there exists a constant* $K := K(\Theta_{\mathfrak{a}}^r, \lambda)$ *such that* $\sup_{\theta_o\in\Theta_{\mathfrak{a}}^r} \mathbb{E}_{\theta_o}\|\widehat{\theta} - \theta_o\|_{\ell_2}^2 \leqslant K\mathcal{R}_\varepsilon^\star$ *for all* $\varepsilon > 0$*.*

Our procedure extends and completes the procedure proposed by [13] in two perspectives. First, it allows a prior variance more general than the polynomially decreasing one. Second, in addition to prove minimax-optimality of the Bayes estimator, we prove concentration at the optimal rate of the posterior distribution.

Conclusions and perspectives. We have presented a hierarchical prior leading to a fully data-driven Bayes estimate that is minimax-optimal in an indirect GSSM. Obviously, the concentration rate based on a hierarchical prior in an indirect GSSM possibly with additional noise in the eigenvalues is only one amongst the many interesting questions for further research and we are currently exploring this topic.

Acknowledgements. This work was supported by the IAP research network no. P7/06 of the Belgian Government (Belgian Science Policy), the contract "Projet d'Actions de Recherche Concertées" No 11/16-039 of the "Communauté française de Belgique" and by the "Fonds Spéciaux de Recherche" from the Université catholique de Louvain.

Bibliography

[1] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

[2] E. Belitser and S. Ghosal. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31:536–559, 2003.

[3] L. Birgé. An alternative point of view on Lepski's method. *State of the art in probability and statistics*, *IMS Lecture Notes*, 36:113–133. 2001.

[4] I. Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Statist.*, 2:1281–1299, 2008.

[5] L. Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24:1–19, 2008.

[6] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30:843–874, 2002.

[7] D. Cox. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, 21:903–923, 1993.

[8] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, pages 500–531, 2000.

[9] J. Johannes and M. Schwarz. Adaptive Gaussian inverse regression with partially unknown operator. *Communications in Statistics - Theory and Methods*, Vol. 42, No. 7, pages 1343-1362, 2013.

[10] J. Johannes, R. Schenk, and A. Simoni. Adaptive Bayesian estimation in Gaussian sequence space models. *Discussion paper at Université catholique de Louvain*, 2014.

[11] A. P. Korostelev, and A. B. Tsybakov. Minimax Theory of Image Reconstruction. *Lecture Notes in Statistics. Springer, New York.*, 82, 1993.

[12] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29:687–714, 2001.

[13] L. H. Zhao. Bayesian aspects of some nonparametric problems. *Ann. Statist.*, 28:532–552, 2000.