INAUGURAL-DISSERTATION

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der Ruprecht-Karls-Universität
Heidelberg

vorgelegt von

Xavier Loizeau

aus Nantes, France

Tag der mündlichen Prüfung:

Oracle and minimax optimality of Bayesian and frequentist methods for linear statistical ill-posed inverse problems under L^2-loss

Betreuer: Jan Johannes
Claudia Schilings

Acknowledgments

I am grateful to the administration of the *Combined Faculty of Natural Sciences and Mathematics* of the Ruprecht-Karls-Universität Heidelberg.

Zusammenfassung

Abstract in german

Abstract

Abstract in English

Contents

1 Introduction **11**

1.1 Statistical ill-posed inverse problems . . . 11

1.1.1 Statistical model . . . 11

1.1.2 (Compact operator) . . . 12

1.1.3 Ill-posed inverse problem with known operator . . . 12

1.1.4 Ill-posed inverse problem with partially known operator . . . 12

1.1.5 (Ill-posed inverse problem with unknown operator) . . . 12

1.1.6 Popular regularisation methods . . . 13

1.2 Frequentist approach . . . 13

1.2.1 Estimation . . . 13

1.2.2 Decision theory . . . 13

1.2.3 Adaptivity . . . 13

1.3 Bayesian approach . . . 13

1.3.1 The Bayesian paradigm . . . 13

1.3.2 Typical priors for non-parametric models . . . 14

1.3.3 The pragmatic Bayesian approach . . . 14

1.3.4 Existing central results . . . 14

1.3.5 Iteration procedure, self informative limit and Bayes carrier . . . 14

1.4 Examples of inverse problems . . . 14

1.4.1 Inverse Gaussian sequence space . . . 14

1.4.2 Circular density deconvolution . . . 15

2 Bayesian interpretation of penalised contrast model selection **23**

2.1 Iterated Gaussian sieve prior . . . 23

2.2 Adaptivity using a hierarchical prior . . . 26

2.3 Proof strategies for contraction rates . . . 28

2.3.1 A moment control based method for contraction rate computation . 29

2.3.2 An exponential concentration inequality based proof for contraction rates of hierarchical sieve priors . . . 31

2.4 First application example: the inverse Gaussian sequence space model . . . 31

2.4.1 Contraction rate for sieve priors . . . 31

2.4.2 Contraction rate for the hierarchical prior . . . 34

2.5 Second application example: the circular density deconvolution model . . . 36

2.6 On the shape of the posterior mean . . . 36

3 Minimax and oracle optimal adaptive aggregation **39**

3.1 Independent observations with unknown noise distribution 39

3.2 Independent observations with unknown noise distribution 39

3.3 Independent observations with known noise density 39

3.4 Absolutely regular observation process with known noise density 39

3.5 Independent observations with unknown noise distribution 39

3.6 Absolutely regular observation process with unknown noise distribution . . 39

A Simulation skim **41**

B R Code **43**

List of notations

Spaces

$(\mathbb{Y}, \mathcal{Y})$	:	the measurable space of observations;
$(\Xi, \mathcal{A})$	:	the parameter space;
$\mathcal{M}([0,1[)$	:	space of probability measures on $[0,1[$
$\mathcal{S}^+(\mathbb{Z})$	:	set of all positive definite, complex valued functions $[f]$ on $\mathbb{Z}$ with $[f](0) = 1$;

Measures and densities

$(\mathbb{P}_{Y\|f}$	$:\mathcal{Y} \to [0,1])_{f\in\Xi}$	:	family to which the data distribution belongs
$\mathbb{P}^X$	$:\mathcal{A} \to [0,1]$	:	distribution of interest;
$\mathbb{P}^\epsilon$	$:\mathcal{A} \to [0,1]$	:	distribution of the noise;
μ	$:\mathcal{A} \to \mathbb{R}_+$	:	a sigma-finite measure, dominating both $\mathbb{P}^X$ and $\mathbb{P}^\epsilon$;
f^X	$:[0,1[\to \overline{\mathbb{R}_+}$	:	density of $\mathbb{P}^X$ with respect to μ;
f^ϵ	$:[0,1[\to \overline{\mathbb{R}_+}$	:	density of $\mathbb{P}^\epsilon$ with respect to μ;
$\mathbb{P}^Y(= \mathbb{P}^X * \mathbb{P}^\epsilon)$	$:\mathcal{A} \to [0,1]$	:	distribution of the observations;
$f^Y(= f^X * f^\epsilon)$	$:[0,1[\to \overline{\mathbb{R}_+}$	:	density of $\mathbb{P}^Y$ with respect to μ;

Random variables

X	$:(\Omega, \mathcal{B}) \to ([0,1[, \mathcal{A})$	:	a random variable with distribution $\mathbb{P}^X$;
ϵ	$:(\Omega, \mathcal{B}) \to ([0,1[, \mathcal{A})$	:	a random variable with distribution $\mathbb{P}^\epsilon$;
$Y(= X \square \epsilon)$	$:(\Omega, \mathcal{B}) \to ([0,1[, \mathcal{A})$	:	a random variable with distribution $\mathbb{P}^Y$;
$X^n(= (X_i^n)_{i\in[\![1,n]\!]})$	$:(\Omega, \mathcal{B}) \to ([0,1[^n, \mathcal{A}^{\otimes n})$	:	a n-vector of i.i.d. replications of X;
$\epsilon^n(= (\epsilon_i^n)_{i\in[\![1,n]\!]})$	$:(\Omega, \mathcal{B}) \to ([0,1[^n, \mathcal{A}^{\otimes n})$	:	a n-vector of i.i.d. replications of ϵ;
$Y^n(= (Y_i^n)_{i\in[\![1,n]\!]})$	$:(\Omega, \mathcal{B}) \to ([0,1[^n, \mathcal{A}^{\otimes n})$	:	a n-vector of i.i.d. replications of Y;

Unary operators

$$
\begin{array}{llll}
 & \mathbb{E}[\cdot] & : & \text{the expectation operator when the distribution is obvious;} \\
\forall \mathbb{P} \text{ distribution} & \mathbb{E}_{\mathbb{P}}[\cdot] & : & \text{the expected value under } \mathbb{P};
\end{array}
$$

$$
\begin{aligned}
{}^*\cdot \quad : \mathcal{M}([0,1[) &\to \mathcal{M}([0,1[); \\
\mathbb{P} &\mapsto {}^*\mathbb{P} = \mathbb{P} * \mathbb{P}^{\epsilon} \\
{}^*\cdot \quad : \mathcal{D}_\mu([0,1[) &\to \mathcal{D}_\mu([0,1[); \\
f &\mapsto {}^*f = f * f^{\epsilon} \\
\forall j \in \mathbb{Z} \qquad e_j(\cdot) \quad : [0,1[&\to \mathbb{C}; \\
x &\mapsto \exp[-2i\pi j x] \\
\mathcal{F}_\mu(\cdot) \quad : \mathcal{D}_\mu([0,1[) &\to \mathcal{S}^+(\mathbb{Z}); \\
f &\mapsto [f] = \left(j \mapsto \int_0^1 f(x) e_j(x) d\mu(x) \right) \\
\mathcal{F}(\cdot) \quad : \mathcal{M}([0,1[) &\to \mathcal{S}^+(\mathbb{Z}); \\
\mathbb{P} &\mapsto [\mathbb{P}] = \left(j \mapsto \int_0^1 e_j(x) d\mathbb{P}(x) \right) \\
\mathcal{F}_\mu^{-1}(\cdot) \quad : \mathcal{S}^+(\mathbb{Z}) &\to \mathcal{D}_\mu([0,1[); \\
[f] &\mapsto f = \left(x \mapsto \sum_{j \in \mathbb{Z}} [f]_j e_j(x) \right) \\
\mathcal{F}^{-1}(\cdot) \quad : \mathcal{S}^+(\mathbb{Z}) &\to \mathcal{M}([0,1[); \\
[\mathbb{P}] &\mapsto \mathbb{P} = \left(A \mapsto \int_A \sum_{j \in \mathbb{Z}} [\mathbb{P}]_j e_j(x) dx \right)
\end{aligned}
$$

Binary operators

$$
\begin{aligned}
\cdot \square \cdot : [0,1[^2 &\to [0,1[\qquad : \quad \text{the modular addition binary operator on the unit segment;} \\
(x,y) &\mapsto x + y - \lfloor x + y \rfloor
\end{aligned}
$$

Miscellaneous

For any set S, subset $s \subseteq S$ we note $\mathbb{1}_s$ the indicatrix function

$$
\begin{aligned}
\mathbb{1}_s : S &\to \{0,1\}; \\
x &\mapsto \begin{cases} 0 \text{ if } x \notin s \\ 1 \text{ if } x \in s \end{cases}
\end{aligned}
$$

Chapter 1

Introduction

1.1 Statistical ill-posed inverse problems

1.1.1 Statistical model

Consider a measurable observation space $(\mathbb{Y}, \mathcal{Y})$, a parameter space Ξ and a family of probability distribution on $(\mathbb{Y}, \mathcal{Y})$ indexed by Ξ, which we denote $(\mathbb{P}_{Y|f})_{f\in\Xi}$. That is to say, for any f in Ξ, $\mathbb{P}_{Y|f}$ is a probability measure.

Throughout this thesis, we are interested in the estimation of the parameter f, under the two paradigmata of frequentist and Bayesian statistics, specified hereafter, and the quantification of the quality of such estimations.
From a frequentist point of view, one specifies a measurable application from $\mathbb{Y}$ to Ξ, called estimator. Ideally, an estimator solely depends of the observations and needs no knowledge whatsoever about the parameter of interest in order to be implemented properly. Once this application specified the duty of the statistician is to study properties such as consistency, rate of convergence, and asymptotic distribution of the estimator. These properties are presented in more details in SECTION 1.2.

In the Bayesian paradigm, one defines a σ-algebra $\mathcal{B}$ on Ξ and a probability distribution $\mathbb{P}_{\boldsymbol{f}}$ on $(\Xi, \mathcal{B})$ called prior distribution which represents the prior knowledge about the parameter, for example gathered by experts or prior experiments. One is then interested in the posterior distribution, that is, the distribution of the parameter of interest given the observations. From a purely Bayesian point of view, being able to define a prior distribution on Ξ and compute (or approach) the posterior distribution is all that is needed as one does not assume existence of a true underlying parameter. However, from a frequentist Bayesian (also called pragmatic Bayesian) approach, described in details in SECTION 1.3, one assumes existence of a true parameter and wonders if the posterior distribution contracts around this true parameter.

We are particularly interested in a specific class of models where Ξ is a subset of a function (or sequence) space. More specifically, let $\mathbb{T}$ and $\mathbb{D}$ be a subsets of $\mathbb{R}$. Denote $\|\cdot\|$ a norm on the space of functions $\mathbb{T} \to \mathbb{D}$. Then Ξ is the space of functions $\{f : \mathbb{T} \to \mathbb{D}, \|f\| < \infty\}$.

We will assume moreover that there exist a measure $\mathbb{P}^\circ$ on $(\mathbb{Y}, \mathcal{Y})$ dominating the family $(\mathbb{P}_{Y|f})_{f\in\Xi}$ and we denote $L : (\Xi \times \mathbb{Y}, \mathcal{B} \otimes \mathcal{Y}) \to \big(\overline{\mathbb{R}}_+, \mathcal{B}(\mathbb{R})\big)$ the likelihood with respect to

$\mathbb{P}^\circ$:

$$\frac{\mathbb{P}_{Y|f}}{\mathbb{P}^\circ}(f,y) = L(f,y).$$

1.1.2 (Compact operator)

We will give particular interest to inverse problems, a family of models where one wants to infer on a parameter f but the data we observe is generated through the distribution with parameter $T(f)$ where T is an operator from Ξ to itself.
Hence we have:

$$\begin{aligned} T : \Xi &\to \Xi \quad ; \\ f &\to T(f) \\ Y &\sim \mathbb{P}_{Y|T(f)}. \end{aligned}$$

These models gathered interest for a long time because many of them have the particularity to be ill-posed in the sense of Hadamard (1902). That is to say, if we build an estimator $\widehat{T(f)}$ of $T(f)$ from the data Y and try to apply T^{-1} to this estimator in order to estimate f, one of the following problems might arise:

- non existence (the equation $T(x) = \widehat{T(f)}$ does not have a solution);
- non unicity (the equation $T(x) = \widehat{T(f)}$ has multiple solutions);
- non stability (the solutions to the equations $T(x) = \widehat{T(f)}$ and $T(x) = \widehat{T(f)} + \epsilon$ are arbitrary far for ϵ arbitrary small with respect to $\|\cdot\|$).

More-Penrose inverse.
Self adjoint.
Orthogonal basis of eigen functions -> notation $(e_j)_{j\in\mathbb{F}}, (\lambda_j)_{j\in\mathbb{F}}$.
$\mathcal{F}$ the application $\Xi \to \Theta, f \mapsto \theta = (\langle f|e_j\rangle)_{j\in\mathbb{F}}$. $\mathcal{F}^{-1}$ the application $\Theta \to \Xi, \theta \mapsto f = (\int_{j\in\mathbb{F}} \theta_j e_j(x))_{x\in\mathbb{T}}$. Plancherel theorem. $\mathbb{P}_{Y|\theta}$, $\mathbb{P}_{\boldsymbol{\theta}}$, $L(\theta, y)$
Compactness, $\lambda_j \to 0$ -> inverse unbounded (condition 3 not check). $\mathcal{F}(T(f)) = \lambda\theta$

1.1.3 Ill-posed inverse problem with known operator

e_j and λ_j known

1.1.4 Ill-posed inverse problem with partially known operator

e_j known but only an estimator $\widehat{\lambda_j}$ of λ_j is available.

1.1.5 (Ill-posed inverse problem with unknown operator)

e_j and λ_j estimated.

1.1.6 Popular regularisation methods

- spectral cut-off
- Tikhonov
- entropy minimisation
- ...?

1.2 Frequentist approach

1.2.1 Estimation

- (M/Z-estimation);
- projection;
- kernel smoother...

1.2.2 Decision theory

- Loss function;
- risk;
- oracle optimality;
- minimax optimality;

1.2.3 Adaptivity

- penalised contrast
- Lepski
- ...

1.3 Bayesian approach

1.3.1 The Bayesian paradigm

- Bayes' theorem;
- prior distribution;
- posterior distribution (include conditions of existence);

1.3.2 Typical priors for non-parametric models

- Gaussian process prior
- Sieve priors (specific case)

$$\mathbb{P}^n_{\boldsymbol{\theta}}(\theta) = \exp\left[-\frac{1}{2}\sum_{|j|\leq m}|\theta_j|^2\right]\cdot\prod_{|j|>m}\delta_0(\theta_j)$$

- Chinese restaurant process
- Dirichlet process

1.3.3 The pragmatic Bayesian approach

- Consistence
- contraction rate
- exact contraction rate
- uniform contraction rate
- oracle optimality
- minimax optimality

1.3.4 Existing central results

- Goshal Van der Vaart
- Nickl

1.3.5 Iteration procedure, self informative limit and Bayes carrier

1.4 Examples of inverse problems

1.4.1 Inverse Gaussian sequence space

Consider the Gaussian process $Y(x)$, defined on $[0,1[$ with constant volatility $\frac{1}{n}$ with n in $\mathbb{N}^\star$ and mean process $f \star g$ where f and g are functions from $[0,1[$ to $\mathbb{R}$. In short, we have $dY(x) = (f \star g)(x)dx + \frac{1}{n}dW(x)$ where W is the Brownian motion. We want to estimate f while observing a realisation of Y. We assume that g is known.
We denote θ and λ respectively the Fourier transforms of f and g respectively.
The likelihood with respect to the standard Brownian motion, noted $\mathbb{P}^\circ$, for this model can be written as follows (see Liptser and Shiryaev (2013))

$$\frac{d\,\mathbb{P}^n_{Y^n|f,g}}{d\,\mathbb{P}^\circ} \propto \exp\left[\int_{[0,1[}\frac{1}{\sqrt{n}}(f\star g)(x)dW(x) - \frac{1}{2}\left\|\frac{f\star g}{\sqrt{n}}\right\|^2\right].$$

We use the fact that the volatility of the process is constant and the properties of the Fourier transform to show that there exist a sequence of independent random variables with standard normal distribution such that the likelihood of the Fourier transform of the process is given by:

$$\frac{d\,\mathbb{P}^n_{Y^n|(\theta,\lambda)}}{d\,\mathbb{P}^\circ} \propto \exp\left[-\frac{1}{2}\sum_{j\in\mathbb{Z}} \frac{(\theta_j\lambda_j - \xi_j)^2}{\sqrt{n}}\right].$$

Therefore, the Fourier transform of the observed process follows a Gaussian process indexed by $\mathbb{Z}$, with mean $\theta \cdot \lambda$ and variance $\frac{1}{n}$.

Note that is the volatility was not constant, we would obtain

$$\frac{d\,\mathbb{P}^n_{Y^n|(\theta,\lambda)}}{d\,\mathbb{P}^\circ} \propto \exp\left[-\frac{1}{2}\sum_{j\in\mathbb{Z}} ((\sigma \star (\theta\lambda))_j - \xi_j)^2\right].$$

The mean process would hence be $\sigma \star (\theta \cdot \lambda)$, which can be rewritten as an inverse problem with a non diagonal operator, more precisely a Toeplitz operator. We do not consider this case in this thesis.

Another motivation for this model is the heat equation. Heat equation + oracle rate for projection estimate + minimax rate (so all notations are introduced before moving on)

1.4.2 Circular density deconvolution

The circular deconvolution model is defined as follows: let X and ϵ be circular random variables (that is to say, taking values in the unit circle, identified to the interval $[0, 1[$) with respective distributions $\mathbb{P}^X$ and $\mathbb{P}^\epsilon$ and densities f^X and f^ϵ with respect to some common and known dominating measure μ on $([0, 1[, \mathcal{A})$. We would hence write for any x in $[0, 1[$, $f^X(x) = \frac{d\mathbb{P}^X}{d\mu}(x)$ for instance.

Definition 1.4.1 Modular addition
From now on we denote by $\Box$ the modular addition on $[0, 1[$. That is to say

$$\forall (x, y) \in [0, 1[^2, \quad x\Box y = x + y[1] = x + y - \lfloor x + y \rfloor.$$

The object of interest is f^X while we only observe identically distributed replications $Y^n = (Y_k)_{k\in[\![1,n]\!]}$ of the random variable Y, defined by $Y := X\Box\epsilon$. We note $\mathbb{P}^Y$ the distribution of the random variable Y and f^Y its density with respect to μ. One would notice that $\mathbb{P}^Y$ and f^Y are respectively given, for any A in $\mathcal{A}$ and y in $[0, 1[$, by $\mathbb{P}^Y(A) = (\mathbb{P}^X * \mathbb{P}^\epsilon)(A) = \int_{[0,1[}\int_{[0,1[} \mathbb{1}_A(x\Box s)d\mathbb{P}^X(x)d\mathbb{P}^\epsilon(s)$ and $f^Y(y) = (f^X * f^\epsilon)(y) = \int_0^1 f^X(y\Box(-s))f^\epsilon(s)d\mu(s)$. Indeed, for any μ-measurable and μ-almost surely bounded function g, we have

$$\begin{aligned}\mathbb{E}\left[g(Y)\right] =& \mathbb{E}\left[g(X\Box\epsilon)\right]\\ =& \int_0^1\int_0^1 g(x\Box s)d\mathbb{P}^X(x)d\mathbb{P}^\epsilon(s)\\ =& \int_0^1\int_0^1 g(y)d\mathbb{P}^X(y\Box(-s))d\mathbb{P}^\epsilon(s)\\ =& \int_0^1 g(y)\int_0^1 d\mathbb{P}^\epsilon(s)d\mathbb{P}^X(y\Box(-s))\\ =& \int_0^1 g(y)\int_0^1 f^X(y\Box(-s))f^\epsilon(s)d\mu(s)d\mu(y);\end{aligned}$$

one should note that the integrals above converge, according to the dominated convergence theorem.
We will thus note $\mathcal{D}_\mu([0,1[)$ the space of densities on $[0,1[$ with respect to μ. Moreover we write indifferently $^*\cdot$ the unary operator which associates to a distribution itself convoluted with $\mathbb{P}^\epsilon$ and the unary operator which associates to a density itself convoluted with f^ϵ. That is to say, given a probability measure $\mathbb{P}$ on $([0,1[,\mathcal{A})$, $^*\mathbb{P}$ is such that, for any A in $\mathcal{A}$, $^*\mathbb{P}_f(A) = (\mathbb{P}^\epsilon * \mathbb{P}_f)(A)$. And for any element f of $\mathcal{D}_\mu([0,1[)$, $^*f$ is such that, for any x in $[0,1[$, $^*f(x) = (f * f^\epsilon)(x)$. The model can therefore at first be written $([0,1[^n,^*\mathbb{P}_f, f\in\mathcal{D}_\mu([0,1[))$, where $\mathbb{P}_f$ is the probability distribution with density f with respect to μ.
I think it should be possible to show that $\mathbb{P}^\epsilon$ does not have to be continuous w.r.t μ and that $\mathbb{P}^Y$ would be anyway. Hence we do not need a density for $\mathbb{P}^\epsilon$ and we can compute the Fourier transform of the distribution anyway.

Reminder 1.4.1 Positive (semi-)definitiveness
A sequence/function $[f]$ from $\mathbb{Z}$ to $\mathbb{C}$ is positive (semi-)definite iff, for any finite subset $\{x_1,\dots,x_n\}$, the Toeplitz matrix $A = (a_{i,j})_{(i,j)\in[\![1,n]\!]^2}$ with $a_{i,j}$ defined by $[f](x_i - x_j)$ is positive (semi-)definite.
In particular, this requires that $[f](x) = \overline{[f](-x)}$, $[f](0) > 0$ and for all x, $[f](x) \le [f](0)$.

Then, by denoting $\mathcal{M}([0,1[)$ the set of all probability measures on $[0,1[$ and $\mathcal{S}^+(\mathbb{Z})$ the set of all positive definite, complex valued, functions $[f]$ on $\mathbb{Z}$ with $[f](0) = 1$, we define the Fourier transform.

Definition 1.4.2 Fourier transform of measures
We denote by $\mathcal{F}$ the Fourier transform operator on measures :

$$\begin{aligned}\mathcal{F}:\quad \mathcal{M}([0,1[) &\to \mathcal{S}^+(\mathbb{Z})\\ \nu\quad &\mapsto \left(j\mapsto \int_0^1 \exp\left[-2i\pi jx\right]d\nu(x)\right).\end{aligned}$$

Notations 1.4.1 Fourier basis functions

As we will operate in the frequency domain for most of the remaining note, it is convenient to use the following notation for the orthonormal basis used in Fourier transform :

$$\forall j \in \mathbb{Z}, \forall x \in [0,1[, \quad e_j(x) := \exp[-2i\pi jx].$$

Remark 1.4.1 It is convenient to note that for any x and s in $[0,1[$ and j in $\mathbb{Z}$, we have $e_j[x\Box s] = e_j[x]e_j[s]$, due to the periodicity of the complex exponential function.

As we are interested in densities of probability distributions dominated by a common measure μ we define the Fourier transform with respect to μ.

Definition 1.4.3 Fourier transform of densities

We denote by $\mathcal{F}_\mu$ the Fourier transform operator of densities with respect to the measure μ :

$$\begin{aligned} \mathcal{F}_\mu : \quad \mathcal{D}_\mu([0,1[) &\to \mathcal{S}^+(\mathbb{Z}) \\ f \qquad\quad &\mapsto \left(j \mapsto \int_0^1 e_j(x) f(x) d\mu(x) \right). \end{aligned}$$

Notations 1.4.2 Fourier transform of useful functions

From now on we adopt the following notations for the functions which will appear regularly :

$$\begin{aligned} \forall j \in \mathbb{Z}, \qquad & \theta_j^\circ := \mathcal{F}_\mu(f^X)(j); \\ & \lambda_j := \mathcal{F}_\mu(f^\epsilon)(j); \\ \forall f \in \mathcal{D}_\mu([0,1[), \forall j \in \mathbb{Z}, \quad & [f](j) := \mathcal{F}_\mu(f)(j). \end{aligned}$$

Obviously, we have

$$
\begin{aligned}
\forall j \in \mathbb{Z}, \mathcal{F}(f^Y)(j) =& \int_0^1 e_j(y)\mathbb{P}^Y(dy) \\
=& \int_0^1 \int_0^1 e_j(x \square s)\mathbb{P}^X(dx)\mathbb{P}^\epsilon(ds) \\
=& \int_0^1 e_j(s) \int_0^1 e_j(x)\mathbb{P}^X(dx)\mathbb{P}^\epsilon(ds) \\
=& \int_0^1 e_j(s)\mathbb{P}^\epsilon(ds) \int_0^1 e_j(x)\mathbb{P}^X(dx) \\
=& \mathcal{F}(\mathbb{P}^\epsilon)(j)\mathcal{F}(\mathbb{P}^X)(j) \\
=& \int_0^1 f^\epsilon(s)e_j(s)d\mu(s) \int_0^1 e_j(x)f^X(x)\mu(dx) \\
=& \mathcal{F}_\mu(f^\epsilon)(j)\mathcal{F}_\mu(f^X)(j) \\
=& \theta_j^\circ \lambda_j
\end{aligned}
$$

so the Fourier transform, exchanges convolution with point-wise product.

The following theorem, which is a special case of Bochner's theorem, allows us to formulate an inverse for the Fourier transform.

Theorem 1.4.1 Herglotz's representation theorem
A function $[f]$ from $\mathbb{Z}$ to $\mathbb{C}$ with $[f](0) = 1$ is semi-definite positive iff there exist μ in $\mathcal{M}([0,1[)$ such that for all j in $\mathbb{Z}$, we have

$$[f](j) = \int_{[0,1[} \exp[-2i\pi jx]d\mu(x).$$

The properties of the set $\mathcal{S}^+(\mathbb{Z})$ can be interpreted as follow :

$$
\begin{aligned}
\mathcal{F}(f)(j) =& \overline{\mathcal{F}(f^Y)(-j)} \quad && f \text{ is real valued;} \\
\mathcal{F}(f)(0) =& 1 && f \text{ integrates at 1;}
\end{aligned}
$$

and $\mathcal{F}(f)$ positive semi-definitive implies the positivity of f.

The Fourier transform being bijective, one can safely write its inversion and we have, for any function $[f]$ in $\mathcal{S}^+$:

$$\begin{aligned}
\forall A \in \mathcal{A}, \qquad & \mathcal{F}^{-1}[f](A) = \int\limits_A \sum_{j\in\mathbb{Z}} [f](j) e_j(x) dx; \\
\forall x \in [0,1[, \quad & \mathcal{F}_\mu^{-1}[f](x) \ = \sum_{j\in\mathbb{Z}} [f](j) e_j(x).
\end{aligned}$$

However, in the most general case, the above mentioned series do not necessarily converge and one would need to consider the densities on our model as Schwartz distributions (see Billingsley (1986)). We avoid this difficulty by assuming the considered distributions dominated by the Lebesgue measure. We hence drop the μ index from now on (and, for example note $\mathcal{D}([0,1[)$ instead of $\mathcal{D}_\mu([0,1))$.
We will hence consider the model written in these terms : $\left([0,1[^n, \mathbb{P}_{[f]}, f \in \mathcal{S}^+(\mathbb{Z})\right)$; where $\mathbb{P}_{[f]}$ is the distribution which admits the density with respect to μ which Fourier transform is $[f]$.

As a concluding note for this section, let us mention the risk we will use to formulate optimality of the different inference methods described there after. For a given, strictly positive real number, we define the usual scalar product on $\mathcal{D}([0,1[)$:

Definition 1.4.4 Scalar product $\langle\cdot|\cdot\rangle_{L^2}$ on $\mathcal{D}([0,1[)$
We define the scalar product

$$\begin{aligned}
\langle\cdot|\cdot\rangle_{L^2} : \quad & \mathcal{D}([0,1[) \times \mathcal{D}([0,1[) \to \overline{\mathbb{R}}. \\
& (f,g) \qquad\qquad \mapsto \int\limits_{[0,1[} f(x)\overline{g(x)} dx
\end{aligned}$$

We obtain with this scalar product the natural L^2 norm :

Definition 1.4.5 L^2-norm $\|\cdot\|_{L^2}$ on $\mathcal{D}([0,1[)$
We define the norm

$$\begin{aligned}
\|\cdot\|_{L^2} : & \mathcal{D}([0,1[) \to \overline{\mathbb{R}_+}. \\
& f \qquad \mapsto \langle f|f\rangle_{L^2}^{1/2} = \left(\int\limits_{[0,1[} |f(x)|^2 dx \right)^{1/2}
\end{aligned}$$

For statistical inference it is generally necessary to assume that the objects of interest have finite norm. We hence define the space $\mathbb{L}^2$:

Definition 1.4.6 Space $\mathbb{L}^2$ of functions
We define the set

$$\mathbb{L}^2 := \{f \in \mathcal{D}([0,1[) : \|f\|_{L^2} < \infty\}.$$

It is common to consider the larger family of norms $\|\cdot\|_{L^p}$ for any number p in $[1,\infty]$ which however do not define an inner product space :

Definition 1.4.7 L^p-norm $\|\cdot\|_{L^p}$ on $\mathcal{D}([0,1[)$
We define the norm

$$\begin{aligned} \|\cdot\|_{L^p} :& \mathcal{D}([0,1[) \to \overline{\mathbb{R}_+}. \\ f \quad & \mapsto \left(\int_{[0,1[} |f(x)|^p dx \right)^{1/p} \end{aligned}$$

Obviously one can define the associated spaces:

Definition 1.4.8 Space $\mathbb{L}^p$ of functions
We define the set

$$\mathbb{L}^p := \{ f \in \mathcal{D}([0,1[) : \|f\|_{L^p} < \infty \}.$$

A last kind of norm which is of interest are the weighted norms. Using a weighted norm as loss function allows to give more interest to some specific features of the functions (high or low frequencies for example).

Definition 1.4.9 $L^p_{\mathfrak{u}}$-norm $\|\cdot\|_{L^p_{\mathfrak{u}}}$ on $\mathcal{D}([0,1[)$
Consider a distribution $\mathfrak{u}$ from $[0,1[$ to $\mathbb{R}$. We define the norm

$$\begin{aligned} \|\cdot\|_{L^r_{\mathfrak{u}}} :& \mathcal{D}([0,1[) \to \overline{\mathbb{R}_+}. \\ f \quad & \mapsto \left(\int_{[0,1[} |(f * \mathfrak{u})(x)|^p dx \right)^{1/p} \end{aligned}$$

In particular, if $\mathfrak{u}$ is the Dirac distribution in 0, we find the definition of $\|\cdot\|_{L^p}$.

We finally define the associated spaces:

Definition 1.4.10 Space $\mathbb{L}^p_{\mathfrak{u}}$ of functions
We define the set

$$\mathbb{L}^p_{\mathfrak{u}} := \left\{ f \in \mathcal{D}([0,1[) : \|f\|_{L^p_{\mathfrak{u}}} < \infty \right\}.$$

Given that we use $\mathcal{S}^+(\mathbb{Z})$ as a parameter space, it is interesting to compare the norms defined above to norms on this space.
For this purpose, we introduce the dot product for sequences.

Definition 1.4.11 Product $\cdot$ on $\mathcal{S}^+(\mathbb{Z})$
We define the bi-linear operator

$$\begin{aligned} \cdot : \quad \mathcal{S}^+(\mathbb{Z})^2 & \to \mathcal{S}^+(\mathbb{Z}) \\ ([f],[g]) \quad & \mapsto [f]\cdot[g] := (j \mapsto [f](j)[g](j)). \end{aligned}$$

We will also use the inner product of $\mathcal{S}^+(\mathbb{Z})$

Definition 1.4.12 Inner product $\langle\cdot|\cdot\rangle_{l^2}$ on $\mathcal{S}^+(\mathbb{Z})$

We define the operator

$$\begin{aligned}\langle\cdot|\cdot\rangle_{l^2}: \quad \mathcal{S}^+(\mathbb{Z})^2 &\to \overline{\mathbb{C}}\\ ([f],[g]) &\mapsto \sum_{j\in\mathbb{Z}}([f]\cdot\overline{[g]})(j).\end{aligned}$$

This leads to the natural l^2-norm

DEFINITION 1.4.13 l^2-NORM $\|\cdot\|_{l^2}$ ON $\mathcal{S}^+(\mathbb{Z})$
We define the norm

$$\begin{aligned}\|\cdot\|_{l^2}: \quad \mathcal{S}^+(\mathbb{Z}) &\to \overline{\mathbb{R}_+}\\ [f] &\mapsto \left(\sum_{j\in\mathbb{Z}}|[f](j)|^2\right)^{1/2}.\end{aligned}$$

It is common to consider the larger family of norms $\|\cdot\|_{l^p}$ for any number p in $[1,\infty]$ which however do not define an inner product space :

DEFINITION 1.4.14 l^p-NORM $\|\cdot\|_{l^p}$ ON $\mathcal{S}^+(\mathbb{Z})$
We define the norm

$$\begin{aligned}\|\cdot\|_{l^p}: \mathcal{S}^+(\mathbb{Z}) &\to \overline{\mathbb{R}_+}.\\ f &\mapsto \left(\sum_{j\in\mathbb{Z}}|[f](j)|^p\right)^{1/p}\end{aligned}$$

A last kind of norm which is of interest are the weighted norms. Using a weighted norm as loss function allows to give more interest to some specific features of the functions (high or low frequencies for example).

DEFINITION 1.4.15 $l^p_{\mathfrak{u}}$-NORM $\|\cdot\|_{l^p_{\mathfrak{u}}}$ ON $\mathcal{S}^+(\mathbb{Z})$
Consider an element $[\mathfrak{u}]$ of $\mathcal{S}^+(\mathbb{Z})$. We define the norm

$$\begin{aligned}\|\cdot\|_{l^p_{[\mathfrak{u}]}}: \mathcal{S}^+(\mathbb{Z}) &\to \mathbb{R}_+.\\ [f] &\mapsto \left(\sum_{j\in\mathbb{Z}}|([f]\cdot[\mathfrak{u}])(j)|^p dx\right)^{1/p}\end{aligned}$$

In particular, if $[\mathfrak{u}]$ is the sequence constantly equal to 1, we find the definition of $\|\cdot\|_{l^p}$.

As previously, we define the spaces associated with these norms:

DEFINITION 1.4.16 SPACES $\mathcal{L}^2, \mathcal{L}^p, \mathcal{L}^p_{[\mathfrak{u}]}$ OF FUNCTIONS
We define the sets

$$\begin{aligned}\mathcal{L}^2 &:= \left\{[f]\in\mathcal{S}^+(\mathbb{Z}) : \|[f]\|_{l^2}<\infty\right\};\\ \mathcal{L}^p_{\mathfrak{u}} &:= \left\{[f]\in\mathcal{S}^+(\mathbb{Z}) : \|[f]\|_{l^p}<\infty\right\};\\ \mathcal{L}^p_{[\mathfrak{u}]} &:= \left\{[f]\in\mathcal{S}^+(\mathbb{Z}) : \|[f]\|_{l^p_{[\mathfrak{u}]}}<\infty\right\}.\end{aligned}$$

We have, for any p in $[1,\infty]$ and f in $\mathcal{D}([0,1[)$.

$$
\begin{aligned}
\|f\|_{\mathfrak{u}}^{r} &= \left(\int_{[0,1[} |(f * \mathfrak{u})(x)|^{p} dx \right)^{1/p} \\
&= \left(\int_{[0,1[} \left| \sum_{j\in\mathbb{Z}} ([f]\cdot[\mathfrak{u}])(j) \cdot e_j(x) \right|^{p} dx \right)^{1/p} \\
&\leq \left(\int_{[0,1[} \sum_{j\in\mathbb{Z}} ((|[f]\cdot[\mathfrak{u}])(j)| \cdot |e_j(x)|)^{p}\, dx \right)^{1/p} \\
&\leq \left(\sum_{j\in\mathbb{Z}} |([f]\cdot[\mathfrak{u}])(j)|^{p} \int_{[0,1[} |e_j(x)|^{r} dx \right)^{1/p} \\
&\leq \left(\sum_{j\in\mathbb{Z}} |([f]\cdot[\mathfrak{u}])(j)|^{p} \cdot 1 \right)^{1/p} \\
&\leq \|[f]\cdot[\mathfrak{u}]\|_{l^{p}} \\
&\leq \|[f]\|_{l^{p}_{[\mathfrak{u}]}}.
\end{aligned}
$$

For the specific case of $p = 2$, the theorem of Plancherel holds and we have

$$
\begin{aligned}
\|f\|_{\mathfrak{u}}^{2} &= \|f * \mathfrak{u}\|^{2} \\
&= \|[f]\cdot[\mathfrak{u}]\|_{l^{2}} \\
&= \|[f]\|_{l^{2}_{[\mathfrak{u}]}}.
\end{aligned}
$$

We hence assume from now on that the parameter of interest has finite norm.

Assumption 1.4.1 The parameter of interest θ° is in $\mathcal{L}^{p}_{[\mathfrak{u}]}$.

Chapter 2

Bayesian interpretation of penalised contrast model selection

In this chapter, we consider the family of Bayesian methods described as "Gaussian sieve priors" in SECTION 1.3.2 as well as an adaptive variant of these priors, the hierarchical sieve priors where the threshold parameter is a random variable with a specified prior distribution. We study their behaviour under two asymptotic, respectively described in SECTION 1.3.3 and SECTION 1.3.5. In SECTION 2.1 we consider the self informative Bayes carrier of Gaussian sieve priors under continuity assumptions for the likelihood and show that its support is contained in the maximum likelihood set. Then, in SECTION 2.2 we show that the distribution of the hyper-parameter in the hierarchical prior contracts around the set of maximisers of a penalised contrast criterion. This section highlights a new link between Bayesian adaptive estimation and the frequentist penalised contrast model selection. In SECTION 2.3, while considering the noise asymptotic, we line out two strategies of proof which allow to obtain contraction rates. The first relies on posterior moment bounding and which, up to our knowledge, is new; the second is specific to the hierarchical sieve prior and is similar to the one used in JOHANNES ET AL. (2016). In SECTION 2.4 we apply this strategies to the specific inverse Gaussian sequence space model. Doing so, we obtain exact contraction rate for the (iterated) Gaussian sieve prior using the first scheme of proof; and the iterated hierarchical prior using the second. This yields optimality for sieve priors with properly chosen threshold parameter; as well as for penalised contrast model selection; and for any iterated version of the hierarchical prior we consider. The most interesting point of this subsection is the novel way to show optimality of the penalised contrast model selection. In SECTION 2.5 we inquire the use of the discussed method to the circular deconvolution model and show that a direct use of those methods is not possible in this context. We give nonetheless some tracks for a fix. Finally, we conclude this chapter with a note about the shape of the posterior mean of the hierarchical prior, motivating the shape of the frequentist estimators we use in CHAPTER 3.

2.1 Iterated Gaussian sieve prior

We consider in this part a statistical model with a functional parameter space as described in SECTION 1.1.1. We adopt a sieve prior as described in SECTION 1.3.2 and first give interest to the asymptotic presented in SECTION 1.3.5.

We first remind the following notations. The parameter space Θ is a function space $\Theta = \{\theta : \mathbb{F} \to \mathbb{I}\}$; with $\mathbb{F}$ a subset of $\mathbb{R}$ and $\mathbb{I}$ a subset of $\mathbb{C}$.
To derive the self informative Bayes carrier we formulate the following hypothesis.

Assumption 2.1.1 Countability assumption
We assume that the set $\mathbb{F}$ is countable.

We equip Θ with the usual $\mathbb{L}^2$ norm that is, $\|\theta\|^2 = \sum_{j\in\mathbb{J}} |\theta_j|^2$ and consider the Borel sigma algebra $\mathcal{B}$ of the topology generated by this $\mathbb{L}^2$ norm.
On the other hand our observation Y take values in the space $(\mathbb{Y}, \mathcal{Y})$ with distribution in the family $(\mathbb{P}_{Y|\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$.
We assume the existence of a function $l : (\Theta, \mathcal{B}) \times (\mathbb{Y}, \mathcal{Y}) \to \mathbb{R}$ such that the likelihood with respect to some reference measures $\mathbb{P}^\circ$ is given by:

$$L(\boldsymbol{\theta}, y) \propto \exp\left[-l(\boldsymbol{\theta}, y)\right].$$

Then, the family of Gaussian sieve priors is indexed by a threshold parameter m in the set of subsets of $\mathbb{J}$, denoted $\mathcal{P}(\mathbb{J})$, and we denote by $\mathbb{P}_{\boldsymbol{\theta}^m}$ the element of this family with index m; moreover, we denote $\boldsymbol{\theta}^m$ a random variable following this distribution. There exists a reference measure $\mathbb{Q}^\circ$ such that the sieve prior with threshold parameter m admits a density of the shape

$$\frac{d\,\mathbb{P}_{\boldsymbol{\theta}^m}}{d\,\mathbb{Q}^\circ}(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2}\sum_{j\in m} |\boldsymbol{\theta}_j|^2\right] \cdot \prod_{j\notin m} \delta_0(\boldsymbol{\theta}_j).$$

If we denote by Θ_m the set $\{\theta \in \Theta : \forall j \notin m, \theta_j = 0\}$, Bayes' theorem gives the following shape for the iterated posterior distribution:

$$\begin{aligned}\frac{d\,\mathbb{P}^\eta_{\boldsymbol{\theta}^m|Y}}{d\,\mathbb{Q}^\circ}(\boldsymbol{\theta}, y) &= \frac{\exp\left[-\left(\frac{1}{2}\sum_{j\in m} |\boldsymbol{\theta}_j|^2 + \eta l(\boldsymbol{\theta}, y)\right)\right] \cdot \prod_{j\notin m} \delta_0(\boldsymbol{\theta}_j)}{\int_{\Theta_m} \exp\left[-\left(\frac{1}{2}\sum_{j\in m} |\mu_j|^2 + \eta l(\mu, y)\right)\right] d\mu} \\ &= \frac{\prod_{j\notin m} \delta_0(\boldsymbol{\theta}_j)}{\int_{\Theta_m} \exp\left[-\frac{1}{2}\sum_{j\in m} \left(|\mu_j|^2 - |\boldsymbol{\theta}_j|^2\right)\right] \exp\left[-\eta\left(l(\mu, y) - l(\boldsymbol{\theta}, y)\right)\right] d\mu}.\end{aligned}$$

The following assumption is also needed to obtain the self informative Bayes carrier.

Assumption 2.1.2 Continuous likelihood asumption
Assume that for any m in $\mathcal{P}(\mathbb{J})$ and y, $\Theta_m \to \mathbb{R}_+, \theta \mapsto l(\theta, y)$ is continuous.

The use of a threshold parameter brings us back to the study of a parametric model and the results from ref Bunke can be used to derive the self informative Bayes carrier.

Theorem 2.1.1 Self informative Bayes carrier for a sieve prior
Under assumption 2.1.1 and assumption 2.1.2 the support of the Bayesian carrier is contained in the set of minimisers of $\theta \mapsto l(\theta, y)$.

PROOF 2.1.1 PROOF OF THEOREM 2.1.1

Let's remind that the definition of continuity gives us:

$$\forall \theta \in \Theta_m, \forall \epsilon \in \mathbb{R}_+^\star, \exists \delta \in \mathbb{R}_+^\star : \forall \mu \in \Theta_m, \|\mu - \theta\| < \delta \Rightarrow |l(\mu, y) - l(\theta, y)| < \epsilon.$$

Then, for any B in $\mathcal{B}$ such that $\inf_{\theta \in B} l(\theta, y) > \inf_{\mu \in \Theta_m} l(\mu, y)$, there exist δ in $\mathbb{R}_+^\star$ and a ball $\mathcal{E}$ of Θ_m of radius δ such that, $\sup_{\mu \in \mathcal{E}} l(\mu, y) < \inf_{\theta \in B} l(\theta, y)$ and hence $\sup_{\mu \in \mathcal{E}} l(\mu, y) - \inf_{\theta \in B} l(\theta, y) < 0$.

Hence we can write

$$\begin{aligned}
\mathbb{P}^\eta_{\boldsymbol{\theta}^m|Y}(B) = & \int_B \frac{\prod_{|j|>m} \delta_0(\boldsymbol{\theta}_j)}{\int_{\Theta_m} \exp\left[-\frac{1}{2} \sum_{|j|\leq m} (|\mu_j|^2 - |\boldsymbol{\theta}_j|^2)\right] \exp\left[-\eta\left(l(\mu, y) - l(\boldsymbol{\theta}, y)\right)\right] d\mu} d\theta \\
\leq & \int_B \frac{\prod_{|j|>m} \delta_0(\theta_j)}{\exp\left[-\eta\left(\sup_{\mu\in\mathcal{E}} l(\mu, y) - \inf_{\theta\in B} l(\theta, y)\right)\right] \int_{\mathcal{E}} \exp\left[-\frac{1}{2} \sum_{|j|\leq m} (|\mu_j|^2 - |\boldsymbol{\theta}_j|^2)\right] d\mu} d\theta \\
\leq & \frac{1}{\exp\left[-\eta\left(\sup_{\mu\in\mathcal{E}} l(\mu, y) - \inf_{\theta\in B} l(\theta, y)\right)\right]} \int_B \frac{\prod_{|j|>m} \delta_0(\theta_j) \exp\left[-\frac{1}{2} \sum_{|j|\leq m} |\boldsymbol{\theta}_j|^2\right]}{\int_{\mathcal{E}} \exp\left[-\frac{1}{2} \sum_{|j|\leq m} |\mu_j|^2\right] d\mu} d\theta \\
\to & 0.
\end{aligned}$$

□

We have hence showed that under the iteration asymptotic, the posterior distribution contracts itself on maximisers of the likelihood, constrained by $\theta_j = 0$ for any $|j| > m$.

Add remark with several maximisers

There is hence a clear link between this type of prior distribution and projection estimators. We will see that, while considering the noise asymptotic, the choice of the threshold is determinant for the quality of the estimation. The choice of the threshold for the projection estimators and for sieve priors should be led in a similar fashion, that is, balancing the bias (small value of the threshold) and the variance (high value of the threshold). As stated previously, the ideal choice of this parameter is however dependent on the parameter of interest and hence not available. It is hence important to inquire adaptive methods for the selection of this parameter. Some methods for the frequentist estimation were outlined in the introduction such as the penalised contrast model selection. In the next section, we introduce the hierarchical sieve prior which consists in modelling the threshold parameter as a random variable. We will show that by selecting the prior distribution for this hyper-parameter properly, the iteration asymptotic gives a Bayesian interpretation to the penalised contrast model selection.

2.2 Adaptivity using a hierarchical prior

We denote $\mathbb{P}_{\boldsymbol{\theta}^M}$ a so called hierarchical prior distribution, described hereafter, and $\boldsymbol{\theta}^M$ a random variable following this prior. Define G a finite subset of $\mathbb{J}$ and $\mathrm{pen} : \mathcal{P}(G) \to \mathbb{R}_+$ a so-called penalty function. The threshold parameter noted m for the sieve prior described in the previous section is now a $\mathcal{P}(G)$-valued random variable denoted M. We note $\mathbb{P}_M$ the distribution of this parameter.

The density of $\mathbb{P}_M$ with respect to the counting measure has the shape

$$\mathbb{P}_M(m) \propto \exp[-\,\mathrm{pen}(m)]\mathbb{1}_{m \subset G}.$$

The dependance structure between the different quantities of the model is then the following:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}^M|M=m} &= \mathbb{P}_{\boldsymbol{\theta}^m};\\
\mathbb{P}_{Y|\boldsymbol{\theta},M} &= \mathbb{P}_{Y|\boldsymbol{\theta}}\,.
\end{aligned}$$

We can then obtain the following form for the posterior distribution of the hyper parameter:

$$\begin{aligned}
\mathbb{P}_{M|Y}(m,y) &\propto \frac{d\,\mathbb{P}_{M,Y}}{d\,\mathbb{P}^\circ}(m,y)\\
&\propto \int_\Theta \frac{d\,\mathbb{P}_{M,Y,\boldsymbol{\theta}^M}}{d\,\mathbb{P}^\circ\; d\,\mathbb{Q}^\circ}(m,y,\theta)d\mathbb{Q}^\circ(\theta)\\
&\propto \int_\Theta \frac{d\,\mathbb{P}_{Y|M,\boldsymbol{\theta}^M}}{d\,\mathbb{P}^\circ}(m,y,\theta)\,\frac{d\,\mathbb{P}_{M,\boldsymbol{\theta}^M}}{d\mathbb{Q}^\circ}(m,\theta)d\mathbb{Q}^\circ(\theta)\\
&\propto \int_\Theta \frac{d\,\mathbb{P}_{Y|\boldsymbol{\theta}^M}}{d\,\mathbb{P}^\circ}(y,\theta)\,\frac{d\,\mathbb{P}_{\boldsymbol{\theta}^M|M}}{d\mathbb{Q}^\circ}(m,\theta)\,\mathbb{P}_M(m)d\mathbb{Q}^\circ(\theta)\\
&\propto \mathbb{P}_M(m)\int_\Theta \frac{d\,\mathbb{P}_{Y|\boldsymbol{\theta}^M}}{d\,\mathbb{P}^\circ}(y,\theta)\,\frac{d\,\mathbb{P}_{\boldsymbol{\theta}^m}}{d\mathbb{Q}^\circ}(m,\theta)d\mathbb{Q}^\circ(\theta)\\
&= \frac{\mathbb{P}_M(m)\int_\Theta \frac{d\,\mathbb{P}_{Y|\boldsymbol{\theta}^M}}{d\,\mathbb{P}^\circ}(y,\theta)\,\frac{d\,\mathbb{P}_{\boldsymbol{\theta}^m}}{d\mathbb{Q}^\circ}(m,\theta)d\mathbb{Q}^\circ(\theta)}{\sum_{j\subset G}\mathbb{P}_M(j)\int_\Theta \frac{d\,\mathbb{P}_{Y|\boldsymbol{\theta}^M}}{d\,\mathbb{P}^\circ}(y,\theta)\,\frac{d\,\mathbb{P}_{\boldsymbol{\theta}^m}}{d\mathbb{Q}^\circ}(j,\theta)d\mathbb{Q}^\circ(\theta)}\\
&= \frac{\exp[-\,\mathrm{pen}(m)]\int_{\Theta_m}\exp[-\frac{1}{2}(2l(y,\theta)+\sum_{k\in m}|\theta_k|^2)]d\mathbb{Q}^\circ(\theta)}{\sum_{j\subset G}\exp[-\,\mathrm{pen}(j)]\int_{\Theta_j}\exp[-\frac{1}{2}(2l(y,\theta)+\sum_{k\in j}|\theta_k|^2)]d\mathbb{Q}^\circ(\theta)}.
\end{aligned}$$

From this, we can deduce the iterated posterior. Indeed, by defining

$$\exp[\Upsilon(Y,m)] := \int_{\Theta_m}\exp[-\frac{1}{2}(2l(y,\theta)+\sum_{k\in m}|\theta_k|^2)]d\mathbb{Q}^\circ(\theta)$$

we have:

$$
\begin{aligned}
\mathbb{P}^{\eta}_{M|Y}(m,y) =& \frac{\mathbb{P}_M(m)\left(\int_{\Theta_m} \exp[-\frac{1}{2}(2l(y,\theta)+\sum_{k\in m}|\theta_k|^2)]d\mathbb{Q}^\circ(\theta)\right)^{\eta}}{\sum_{j\subset J}\mathbb{P}_M(j)\left(\int_{\Theta_j} \exp[-\frac{1}{2}(2l(y,\theta)+\sum_{k\in j}|\theta_k|^2)]d\mathbb{Q}^\circ(\theta)\right)^{\eta}} \\
=& \frac{\exp[-\operatorname{pen}(m)+\eta\Upsilon(Y,m)]}{\sum_{j\subset G}\exp[-\operatorname{pen}(j)+\eta\Upsilon(Y,j)]}\mathbb{1}_{m\subset G} \\
=& \frac{1}{\sum_{j\subset G}\exp\left[\eta\left(\Upsilon(Y,j)-\Upsilon(Y,m)\right)-(\operatorname{pen}(j)-\operatorname{pen}(m))\right]}\mathbb{1}_{m\subset G}
\end{aligned}
$$

and we can deduce the self informative Bayes carrier.

LEMMA 2.2.1 SELF INFORMATIVE BAYES CARRIER OF THE HYPER-PARAMETER IN A HIERARCHICAL SIEVE PRIOR I

The support of the self informative Bayes carrier for the hyper-parameter M is

$$
\arg\max_{m\subset G}\{\Upsilon(Y,m)\}.
$$

Unfortunately, in many practical cases, the choice led by $\arg\max_{m\subset G}\{\Upsilon(Y,m)\}$ is G itself and leads to inconsistent inference (as we will show later). However, if one allows the prior distribution to depend on η and to take the shape $\exp[-\eta\operatorname{pen}(m)]\mathbb{1}_{m\subset G}$, we obtain the following theorem.

THEOREM 2.2.1 SELF INFORMATIVE BAYES CARRIER OF THE HYPER-PARAMETER IN A HIERARCHICAL SIEVE PRIOR II

Using the modified prior which depends on η, the support of the self informative Bayes carrier for the hyper-parameter M is

$$
\arg\max_{m\subset G}\{\Upsilon(Y,m)-\operatorname{pen}(Y,m)\}.
$$

PROOF 2.2.1 PROOF OF THEOREM 2.2.1

For any finite set P of subsets of G such that $\max_{m\in P}\Upsilon(Y,m)-\operatorname{pen}(Y,m) < \max_{k\subset G}\Upsilon(Y,k)-\operatorname{pen}(Y,k)$, we can write

$$
\begin{aligned}
\mathbb{P}^{\eta}_{M|Y}(P) =& \sum_{m\in P}\frac{1}{\sum_{j\subset G}\exp\left[\eta\left(\Upsilon(Y,j)-\Upsilon(Y,m)-(\operatorname{pen}(j)-\operatorname{pen}(m))\right)\right]}\mathbb{1}_{m\subset G} \\
\leq& \frac{Card(P)}{\exp\left[\eta\left(\max_{j\subset G}(\Upsilon(Y,j)-\operatorname{pen}(j))-\max_{m\in P}(\Upsilon(Y,m)-\operatorname{pen}(m))\right)\right]}\mathbb{1}_{m\subset G} \\
\to& 0.
\end{aligned}
$$

□

The posterior distribution for $\boldsymbol{\theta}^M$ itself follows:

$$
\begin{aligned}
\frac{d\mathbb{Q}_{\boldsymbol{\theta}^M|Y}}{d\,\mathbb{P}^\circ}(\theta,y) &\propto \frac{d\,\mathbb{P}_{\boldsymbol{\theta}^M,Y}}{d\mathbb{Q}^\circ\,d\,\mathbb{P}^\circ}(\theta,y)\\
&\propto \sum_{m\subset J}\frac{d\,\mathbb{P}_{\boldsymbol{\theta}^M,Y,M}}{d\mathbb{Q}^\circ\,d\,\mathbb{P}^\circ\,d\,\mathbb{P}^\circ}(\theta,y,m)\\
&\propto \sum_{m\subset J}\frac{d\,\mathbb{P}_{\boldsymbol{\theta}^M|Y,M}}{d\mathbb{Q}^\circ}(\theta,y,m)\frac{d\,\mathbb{P}_{Y,M}}{d\,\mathbb{P}^\circ\,d\,\mathbb{P}^\circ}\\
&\propto \sum_{m\subset J}\frac{d\,\mathbb{P}_{\boldsymbol{\theta}^m|Y}}{d\mathbb{Q}^\circ}(\theta,y,m)\frac{d\,\mathbb{P}_{M|Y}}{d\,\mathbb{P}^\circ}\frac{d\,\mathbb{P}_{Y}}{d\,\mathbb{P}^\circ}(Y)\\
&= \sum_{m\subset J}\frac{d\,\mathbb{P}_{\boldsymbol{\theta}^m|Y}}{d\mathbb{Q}^\circ}(\theta,y,m)\frac{d\,\mathbb{P}_{M|Y}}{d\,\mathbb{P}^\circ}.
\end{aligned}
$$

From this, we can deduce the iterated posterior distribution for $\boldsymbol{\theta}^M$:

$$
\begin{aligned}
\frac{d\mathbb{Q}^\eta_{\boldsymbol{\theta}^M|Y}}{d\,\mathbb{P}^\circ}(\theta,y) &= \sum_{m\subset G}\frac{d\,\mathbb{P}^\eta_{\boldsymbol{\theta}^m|Y}}{d\mathbb{Q}^\circ}(\theta,y,m)\frac{d\,\mathbb{P}^\eta_{M|Y}}{d\,\mathbb{P}^\circ}(m,y)\\
&= \sum_{m\subset G}\frac{\exp\left[-\left(\frac{1}{2}\sum\limits_{j\in m}|\boldsymbol{\theta}_j|^2+\eta l(\boldsymbol{\theta},y)\right)\right]\cdot\prod\limits_{j\notin m}\delta_0(\boldsymbol{\theta}_j)}{\int_{\Theta_m}\exp\left[-\left(\frac{1}{2}\sum\limits_{j\in m}|\mu_j|^2+\eta l(\mu,y)\right)\right]d\mu}\frac{\exp[-\operatorname{pen}(m)+\eta\Upsilon(Y,m)]}{\sum\limits_{j\subset G}\exp[-\operatorname{pen}(j)+\eta\Upsilon(Y,j)]}\mathbb{1}_{m\subset G}
\end{aligned}
$$

And as a consequence, we can deduce the self informative Bayes carrier.

Theorem 2.2.2 Self informative carrier using a hierarchical sieve prior
Denote $\widehat{m} := \underset{m\subset G}{\arg\max}\{\Upsilon(Y,m)-\operatorname{pen}(m)\}$ then the support of the self informative Bayes carrier is contained in $\underset{\theta\in\Theta_m, m\in\widehat{m}}{\arg\max}\{-l(\theta,Y)\}$.

We have hence seen in these two first sections investigated the behaviour of the sieve prior and its hierarchical version under the iterative asymptotic and shown that under some mild assumptions, their self informative Bayes carriers correspond to some constrained maximum likelihood estimator and penalised contrast model selection version of it respectively. We should now investigate the behaviour of these (iterated) priors under the noise asymptotic and define hypotheses under which they behave properly.

2.3 Proof strategies for contraction rates

In this section, we depict two proof strategies for contraction rates. They will be used in the next sections to compute contraction rates for sieve and hierarchical sieve priors respectively.
The first proof relies on moment bounding of the random variable $\|\boldsymbol{\theta}-\theta^\circ\|$ which is more easily interpretable of the norm used is the L^2 norm. The second proof relies on the use of exponential concentration inequalities.

2.3.1 A moment control based method for contraction rate computation

In this section we outline a method to prove contraction rates which requires to bound properly some moments of the posterior distribution. We later use this method in the case of the inverse Gaussian sequence space with a sieve prior. Provided that bounds are available for the required moments, this method barely needs any other assumption on the model. Moreover, it appears that, in the example we display here, it leads to the same rate as the frequentist optimal convergence rate without a logarithmic loss as it is often the case with popular methods.

A limitation is that moments of posterior distributions are not always explicitly available, in particular for non conjugate prior. A consequence is that we were not able to use this method for the deconvolution model nor for computation of contraction rate of the hierarchical prior.

However, we believe that the method could be generalised to wider cases, for example using convergence of distribution in Wasserstein distance implying convergence of moments.

A similar method to obtain lower bounds is described in annex. Unfortunately, it could not be used in any practical case here.

Lemma 2.3.1 Upper bound for posterior expectation

Assume $\max\left\{\mathbb{E}_{\theta^\circ}\left[\mathbb{E}_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right]\right], \sqrt{\mathbb{V}_{\theta^\circ}\left[\mathbb{E}_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right]\right]}\right\} \in \mathcal{O}(\Phi_n(\theta^\circ))$. Then, for any increasing unbounded sequence c_n, we have:

$$\lim_{n\to\infty} \mathbb{P}^n_{\theta^\circ}\left(\mathbb{E}^n_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right] \geq c_n\Phi_n\right) = 0.$$

Proof 2.3.1 Proof of lemma 2.3.1

Define the sequence of random variables $\mathcal{S}_n := \frac{\mathbb{E}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]-\mathbb{E}_{\theta^\circ}\left[\mathbb{E}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\right]}{\sqrt{\mathbb{V}_{\theta^\circ}\left[\mathbb{E}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\right]}}$. This is a sequence of random variables with common expectation 0 and variance 1 and, as such, their distributions for a sequence of tight measures. Hence, for any increasing unbounded sequence c_n and $K_n := \mathbb{E}_{\theta^\circ}\left[\mathbb{E}^n_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right]\right] + c_n\sqrt{\mathbb{V}_{\theta^\circ}\left[\mathbb{E}^n_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right]\right]}$ we can write

$$\begin{aligned}\mathbb{P}^n_{\theta^\circ}\left(\mathbb{E}^n_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right] \geq K_n\right) =& \mathbb{P}^n_{\theta^\circ}\left(S_n \geq \frac{K_n - \mathbb{E}_{\theta^\circ}\left[\mathbb{E}^n_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right]\right]}{\sqrt{\mathbb{V}_{\theta^\circ}\left[\mathbb{E}^n_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right]\right]}}\right)\\ =& \mathbb{P}^n_{\theta^\circ}\left(S_n \geq c_n\right)\end{aligned}$$

which tends to 0 as S_n is tight.

Lemma 2.3.2 Upper bound for posterior variance

Assume $\max\left\{\mathbb{E}_{\theta^\circ}\left[\mathbb{V}_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right]\right], \mathbb{V}_{\theta^\circ}\left[\mathbb{V}_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right]\right]\right\} \in \mathcal{O}(\Phi_n(\theta^\circ))$. Then, for any increasing unbounded sequence c_n, we have:

$$\lim_{n\to\infty} \mathbb{P}^n_{\theta^\circ}\left(\mathbb{V}^n_{\boldsymbol{\theta}|Y^n}\left[\|\boldsymbol{\theta}-\theta^\circ\|\right] \geq c_n\Phi_n\right) = 0.$$

Proof 2.3.2 Proof of lemma 2.3.2

Define the sequence of random variables $\mathcal{S}_n := \frac{\mathbb{V}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]-\mathbb{E}_{\theta^\circ}\left[\mathbb{V}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\right]}{\sqrt{\mathbb{V}_{\theta^\circ}\left[\mathbb{V}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\right]}}$. This is a sequence of random variables with common expectation 0 and variance 1 and, as such, their distributions for a sequence of tight measures. Hence, for any increasing unbounded sequence c_n and $K_n := \mathbb{E}_{\theta^\circ}\left[\mathbb{V}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\right] + c_n\sqrt{\mathbb{V}_{\theta^\circ}\left[\mathbb{V}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\right]}$ we can write

$$\begin{aligned}\mathbb{P}^n_{\theta^\circ}\left(\mathbb{V}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\geq K_n\right) =& \mathbb{P}^n_{\theta^\circ}\left(S_n \geq \frac{K_n - \mathbb{E}_{\theta^\circ}\left[\mathbb{V}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\right]}{\sqrt{\mathbb{V}_{\theta^\circ}\left[\mathbb{V}^n_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\right]}}\right)\\ =&\mathbb{P}^n_{\theta^\circ}\left(S_n\geq c_n\right)\end{aligned}$$

which tends to 0 as S_n is tight.

Theorem 2.3.1 Upper bound

Under the hypotheses of lemma 2.3.1 and lemma 2.3.2 we have for any increasing unbounded sequence c_n

$$\lim_{n\to\infty}\mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}_{\boldsymbol{\theta}|Y^n}\left(\|\boldsymbol{\theta}-\theta^\circ\|>c_n\Phi_n(\theta^\circ)\right)\right]=0.$$

Proof 2.3.3 Proof of theorem 2.3.1

Define the tight sequence of random variables $S_n := \frac{\|\boldsymbol{\theta}-\theta^\circ\|-\mathbb{E}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]}{\sqrt{\mathbb{V}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]}}$. We consider the sequence of events $\Omega_n := \left\{\left\{\mathbb{E}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\geq c_n\Phi_n\right\}\cap\left\{\mathbb{V}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\geq c_n\Phi_n\right\}\right\}$. We have $\mathbb{P}_{\theta^\circ}(\Omega_n)\leq\max\left(\mathbb{P}_{\theta^\circ}(\left\{\mathbb{E}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\geq c_n\Phi_n\right\}),\mathbb{P}_{\theta^\circ}(\left\{\mathbb{V}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]\geq c_n\Phi_n\right\})\right)$ which hence tends to 0. Hence, for $K_n := c_n\Phi_n(1+c_n)$, we can write

$$\begin{aligned}\mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}_{\boldsymbol{\theta}|Y^n}\left(\|\boldsymbol{\theta}-\theta^\circ\|>K_n\right)\right] &= \mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}_{\boldsymbol{\theta}|Y^n}\left(S_n > \frac{K_n-\mathbb{E}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]}{\sqrt{\mathbb{V}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]}}\right)\right]\\ &= \mathbb{E}^n_{\theta^\circ}\left[\mathbb{1}_{\Omega_n}\mathbb{P}_{\boldsymbol{\theta}|Y^n}\left(S_n > \frac{K_n-\mathbb{E}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]}{\sqrt{\mathbb{V}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]}}\right)\right]\\ &\quad+\mathbb{E}^n_{\theta^\circ}\left[\mathbb{1}_{\Omega_n^c}\mathbb{P}_{\boldsymbol{\theta}|Y^n}\left(S_n > \frac{K_n-\mathbb{E}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]}{\sqrt{\mathbb{V}_{\boldsymbol{\theta}|Y^n}[\|\boldsymbol{\theta}-\theta^\circ\|]}}\right)\right]\\ &\leq\mathbb{P}^n_{\theta^\circ}\left(\Omega_n\right)+\mathbb{P}_{\theta^\circ}\left(\Omega_n^c\right)\cdot\mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}_{\boldsymbol{\theta}|Y^n}\left(S_n>\frac{K_n-c_n\Phi_n}{c_n\Phi_n}\right)\right]\\ &\leq\mathbb{P}^n_{\theta^\circ}\left(\Omega_n\right)+\mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}_{\boldsymbol{\theta}|Y^n}\left(S_n>c_n\right)\right].\end{aligned}$$

We can conclude as S_n is a tight sequence, c_n tends to infinity and $\mathbb{P}^n_{\theta^\circ}\left(\Omega_n\right)$ tends to 0.

2.3.2 An exponential concentration inequality based proof for contraction rates of hierarchical sieve priors

2.4 First application example: the inverse Gaussian sequence space model

In this section, we consider the inverse Gaussian sequence space model and use the methodology described in SECTION 2.3 to compute upper bounds of the Gaussian sieve priors described in SECTION 2.1 when applied to this specific model. Doing so, we will notice that it gives us, for a very general case, the same speed as the convergence rate of projection estimators and that, by choosing properly the threshold parameter, we reach the oracle rate of convergence as well as the minimax optimal rate, **without a log-loss**.
Then, using a methodology similar to JOHANNES ET AL. (2016) we show that under some regularity conditions, the iterated hierarchical prior leads to optimal posterior contraction rate. As a consequence, we can conclude about the oracle and minimax optimality of the penalised contrast model selection estimator with a new strategy of proof.

2.4.1 Contraction rate for sieve priors

Considering this model, we use a Gaussian sieve prior for θ as described in SECTION 2.1 and inquire the behaviour of the posterior distribution under the asymptotic $n \to \infty$. To sum up our setting we have:

$$\begin{array}{llll} \text{noise level:} & n & \in & \mathbb{N}; \\ \text{parameter of interest:} & \theta^\circ = (\theta_j^\circ)_{j\in\mathbb{N}} & \in & \mathbb{L}^2(\mathbb{R}^{\mathbb{N}}); \\ \text{convolution operator:} & \lambda = (\lambda_j)_{j\in\mathbb{N}} & \in & \mathbb{R}^{\mathbb{N}}; \\ \text{noise sequence:} & \xi = (\xi_j)_{j\in\mathbb{N}} & \sim_{i.i.d.} & \mathcal{N}(0,1); \\ \text{observation:} & Y^n = (Y_j^n)_{j\in\mathbb{N}} & = & \theta_j^\circ \lambda_j + \frac{1}{\sqrt{n}}\xi_j; \\ \text{threshold sequence:} & m_n & \in & \mathbb{N}^{\mathbb{N}}; \\ \text{prior guess} & \boldsymbol{\theta}^{m_n} = (\boldsymbol{\theta}_j^{m_n})_{j\in\mathbb{N}} & \sim & \mathcal{N}(0,1)\mathbb{1}_{j\leq m_n} + \delta_0 \mathbb{1}_{j<m_n}. \end{array}$$

We are in a conjugated case and the iterated posterior is easily derived. Define for any j in $\mathbb{N}$ and η in $\mathbb{N}^\star$ the quantities

$$\widehat{\theta}_j^{(\eta)} := \frac{n\eta Y_j^n \lambda_j}{1+n\eta\lambda_j^2}; \quad \sigma_j^{(\eta)} := \frac{1}{1+n\eta\lambda_j^2}.$$

Then, for any j in $\mathbb{N}$, the posterior distribution of $\boldsymbol{\theta}_j$ after η iterations is given by

$$\boldsymbol{\theta}_j | Y^{n,\eta} \sim \mathcal{N}(\widehat{\theta}_j^{(\eta)}, \sigma_j^{(\eta)})\mathbb{1}_{j\leq m_n} + \delta_0(\boldsymbol{\theta}_j)\mathbb{1}_{j>m_n}.$$

According to SECTION 2.1, for n fixed, if η tends to infinity, the posterior distribution contracts around the maximiser of the constrained likelihood. Considering the limits of $\widehat{\theta}_j^{(\eta)}$ and $\sigma_j^{(\eta)}$ as η tells us that this maximiser is the projection estimator $\overline{\theta}^{m_n} = (\overline{\theta}_j^{m_n})_{j\in\mathbb{N}} =$

$(\frac{Y_j}{\lambda_j}\mathbb{1}_{j\le m_n})_{j\in\mathbb{N}}$.

We can then compute the quantities appearing in SECTION 2.3 which gives the following results.

COROLLARY 2.4.1 For any θ° in Θ and increasing, unbounded sequence c_n, we have

$$\lim_{n\to\infty} \mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}^{n,(\eta)}_{\boldsymbol{\theta}^{m_n}|Y^n}\left(d^2\left(\theta^\circ, \boldsymbol{\theta}^{m_n}\right) \le c_n \Phi^{m_n}_n\right)\right] = 1.$$

PROOF 2.4.1 We want to find a sequence $(K_n)_{n\in\mathbb{N}}$ (for short, K_n) converging to 0 such that

$$\lim_{n\to\infty} \mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left(\|\boldsymbol{\theta}^{m_n} - \theta^\circ\|^2 \ge K_n\right)\right] = 0.$$

For any n, we define $S^{m_n} := \sum_{j=1}^{m_n} \left(\boldsymbol{\theta}^{m_n} - \theta^\circ\right)^2$. Therefore we have :

$$\mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left(\|\boldsymbol{\theta}^{m_n} - \theta^\circ\|^2 \ge K_n\right)\right] = \mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left(S^{m_n} \ge K_n - \mathfrak{b}_{m_n}\right)\right].$$

By definition, S^{m_n} has finite expectation and strictly positive, finite variance. We define $\mathcal{S}^{m_n} := \frac{S^{m_n} - \mathbb{E}^n_{\boldsymbol{\theta}^{m_n}|Y^n}[S^{m_n}]}{\sqrt{\mathbb{V}^n_{\boldsymbol{\theta}^{m_n}|Y^n}[S^{m_n}]}}$. The sequence of random variables defined this way is tight as their expectations are all equal to 0 and their variances to 1. We now have to look for a contraction rate for this new family of random variables as we have :

$$\mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left(\|\boldsymbol{\theta}^{m_n} - \theta^\circ\|^2 \ge K_n\right)\right] = \mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left(\mathcal{S}^{m_n} \ge \frac{K_n - \mathfrak{b}^2_{m_n} - \mathbb{E}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left[S^{m_n}\right]}{\sqrt{\mathbb{V}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left[S^{m_n}\right]}}\right)\right].$$

We now control the convergence in probability of $\mathbb{E}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left[\mathcal{S}^{m_n}\right]$ and $\mathbb{V}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left[\mathcal{S}^{m_n}\right]$ which are given by

$$\begin{aligned}
\mathbb{E}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left[S^{m_n}\right] =& \sum_{j=1}^{m_n} \frac{\Lambda_j}{n\eta} \cdot \left(\frac{1}{\frac{\Lambda_j}{n\eta}+1}\right)\left(1 + \frac{\left(-\theta^\circ_j + \eta\sqrt{n}\xi_j\lambda_j\right)^2}{\frac{\eta n}{\Lambda_j}\left(\frac{\Lambda_j}{\eta n}+1\right)}\right)\\
\le& \sum_{j=1}^{m_n} \frac{\Lambda_j}{n\eta} + \sum_{j=1}^{m_n} \frac{\Lambda_j^2}{n^2\eta^2}\left(-\theta^\circ_j + \eta\sqrt{n}\xi_j\lambda_j\right)^2;\\
\mathbb{V}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left[S^{m_n}\right] =& 2\sum_{j=1}^{m_n}\left(\frac{\Lambda_j}{n\eta}\cdot\frac{1}{\frac{\Lambda_j}{n\eta}+1}\right)^2\left(1 + 2\frac{\left(-\theta^\circ_j + \eta\sqrt{n}\xi_j\lambda_j\right)^2}{\frac{\eta n}{\Lambda_j}\left(\frac{\Lambda_j}{\eta n}+1\right)}\right)\\
\le& 2\sum_{j=1}^{m_n} \frac{\Lambda_j^2}{n^2\eta^2} + 4\sum_{j=1}^{m_n} \frac{\Lambda_j^3}{n^3\eta^3}\left(-\theta^\circ_j + \eta\sqrt{n}\xi_j\lambda_j\right)^2.
\end{aligned}$$

We now control the stochastic parts of those moments.
Define for some sequence $(u_n)_{n\in\mathbb{N}}$, tending to 0 and any deterministic sequence (a_j) the event $\Omega_{m_n} := \left\{\sum_{j=1}^{m_n} a_j\left(-\theta^\circ_j + \eta\sqrt{n}\xi_j\lambda_j\right)^2 \le u_n\right\}$.

Obviously, $\Omega_{m_n}^c = \left\{ \sum_{j=1}^{m_n} a_j \left(-\theta_j^\circ + \eta\sqrt{n}\xi_j\lambda_j \right)^2 \geq u_n \right\}$ has probability

$$\mathbb{P}_{\theta^\circ}^n \left(\Omega_{m_n}^c \right) = \mathbb{P}_{\theta^\circ}^n \left(\sum_{j=1}^{m_n} \left(-\theta_j^\circ + \eta\sqrt{n}\xi_j\lambda_j \right)^2 \geq u_n \right).$$

In the same spirit as previously, we define the sequence of random variables $T^{m_n} := \sum_{j=1}^{m_n} a_j \left(-\theta_j^\circ + \eta\sqrt{n}\xi_j\lambda_j \right)^2$. We have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^{m_n}|Y^n}^n \left[T^{m_n} \right] &= \sum_{j=1}^{m_n} a_j \frac{\eta^2 n}{\Lambda_j} \left[1 + \frac{\Lambda_j}{n\eta^2} \left(\theta_j^\circ \right)^2 \right] \\
&\leq \sum_{j=1}^{m_n} a_j \frac{\eta^2 n}{\Lambda_j} \left[1 + \frac{\Lambda_1}{\eta^2} \left(\theta_j^\circ \right)^2 \right]; \\
\mathbb{V}_{\boldsymbol{\theta}^{m_n}|Y^n}^n \left[T^{m_n} \right] &= 2 \sum_{j=1}^{m_n} a_j^2 \left(\frac{\eta^2 n}{\Lambda_j} \right)^2 \left[1 + 4 \frac{\Lambda_j}{\eta^2 n} \left(\theta_j^\circ \right)^2 \right] \\
&\leq 2 \sum_{j=1}^{m_n} a_j^2 \left(\frac{\eta^2 n}{\Lambda_j} \right)^2 \left[1 + 4 \frac{\Lambda_1}{\eta^2} \left(\theta_j^\circ \right)^2 \right];
\end{aligned}$$

and the sequence of random variables $\mathcal{T}^{m_n} := \frac{T^{m_n} - \mathbb{E}_{\boldsymbol{\theta}^{m_n}|Y^n}^n [T^{m_n}]}{\sqrt{\mathbb{V}_{\boldsymbol{\theta}^{m_n}|Y^n}^n [T^{m_n}]}}$ is tight. Therefore, $\mathbb{P}_{\theta^\circ}^n \left(\sum_{j=1}^{m_n} \left(-\theta_j^\circ + \eta\sqrt{n}\xi_j\lambda_j \right)^2 \geq u_n \right) = \mathbb{P}_{\theta^\circ}^n \left(\mathcal{T}^{m_n} \geq \frac{u_n - \mathbb{E}_{\boldsymbol{\theta}^{m_n}|Y^n}^n [T^{m_n}]}{\sqrt{\mathbb{V}_{\boldsymbol{\theta}^{m_n}|Y^n}^n [T^{m_n}]}} \right)$. Consider any sequence (c_n) diverging to infinity. Then if

$$\begin{aligned}
u_n &= \sqrt{\mathbb{V}_{\boldsymbol{\theta}^{m_n}|Y^n}^n \left[T^{m_n} \right]} c_n + \mathbb{E}_{\boldsymbol{\theta}^{m_n}|Y^n}^n \left[T^{m_n} \right] \\
&= c_n \cdot \sqrt{2 \sum_{j=1}^{m_n} a_j^2 \left(\frac{\eta^2 n}{\Lambda_j} \right)^2 \left[1 + 4 \frac{\Lambda_j}{\eta^2 n} \left(\theta_j^\circ \right)^2 \right]} + \sum_{j=1}^{m_n} a_j \frac{\eta^2 n}{\Lambda_j} \left[1 + \frac{\Lambda_j}{n\eta^2} \left(\theta_j^\circ \right)^2 \right].
\end{aligned}$$

Then $\mathbb{P}_{\theta^\circ}^n (\Omega_{m_n}^c) \leq \mathbb{P}_{\theta^\circ}^n (\mathcal{T}^{m_n} \geq c_n) \to 0$ as $\mathcal{T}^{m_n}$ is tight.

We can now conclude about the posterior contraction by defining

$$\begin{aligned}
K_n :=& \mathfrak{b}_{m_n}^2 + \sum_{j=1}^{m_n} \frac{\Lambda_j}{n\eta} \cdot \left(\frac{1}{\frac{\Lambda_j}{n\eta} + 1} \right) \left(1 + \frac{u_n}{\frac{\eta n}{\Lambda_j} \left(\frac{\Lambda_j}{\eta n} + 1 \right)} \right) \\
&+ c_n \cdot \sqrt{2 \sum_{j=1}^{m_n} \left(\frac{\Lambda_j}{n\eta} \cdot \frac{1}{\frac{\Lambda_j}{n\eta} + 1} \right)^2 \left\{ 1 + 2 \frac{u_n}{\frac{\eta n}{\Lambda_j} \left(\frac{\Lambda_j}{\eta n} + 1 \right)} \right\}} \\
=& \mathcal{O} \left(c_n \cdot \frac{m_n \overline{\Lambda}_{m_n}}{n\eta} \vee \mathfrak{b}_{m_n}^2 \right)
\end{aligned}$$

Indeed :

$$\begin{aligned}\mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}^n_{\boldsymbol{\theta}^{m_n}|Y^n}\left(\|\boldsymbol{\theta}^{m_n}-\theta^\circ\|^2 \geq K_n\right)\right] &\leq \mathbb{E}^n_{\theta^\circ}\left[\mathbb{1}_{\Omega n}\,\mathbb{P}^n_{\theta^{m_n}|Y^n}\left(\|\boldsymbol{\theta}^{m_n}-\theta^\circ\|^2 \geq K_n\right)\right] + \mathbb{P}_{\theta^\circ}(\Omega n^c)\\ &\leq \mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}^n_{\theta^{m_n}|Y^n}\left(\mathcal{S}^{m_n} \geq c_n\right)\right] + \mathbb{P}_{\theta^\circ}(\Omega^c_n)\end{aligned}$$

Which tends to 0 as $\mathcal{S}^{m_n}$ is a tight sequence of random variables. One could notice that if η diverges to infinity, the sequence c_n cancels and we recover the frequentist $\mathbb{L}_2$ rate of convergence for projection estimators.

Notice that if one selects $m_n = m_n^\circ$ we obtain the oracle rate of convergence of projection estimators.

Corollary 2.4.2 For any increasing and unbounded sequence, we have

$$\lim_{n\to\infty}\inf_{\theta^\circ\in\Theta_{\mathfrak{a}}(r)}\mathbb{E}^n_{\theta^\circ}\left[\mathbb{P}^{n,(\eta)}_{\boldsymbol{\theta}^{m^\star_n}|Y^n}\left(d^2\left(\theta^\circ,\boldsymbol{\theta}\right) \leq c_n\cdot\Psi^\star_n(\Theta_{\mathfrak{a}}(r))\right)\right] = 1;$$

Moreover, if one lets the number of iterations tend to infinity, we observe that the distribution degenerates around the projection estimator as defined in section 2.1:

$$\lim_{\eta\to\infty}\mathbb{P}^{n,(\eta)}_{\boldsymbol{\theta}^m|Y^n} = \delta_{\overline{\theta}^m}.$$

2.4.2 Contraction rate for the hierarchical prior

Let be $G_n := \max\{m \in [\![1,n]\!] : \Lambda_m/n \leq \Lambda_1\}$. We give the following specific shape to the prior for the threshold parameter $\mathbb{P}^n_M(M=m) = \dfrac{\exp\left(-3\cdot\eta\cdot\frac{m}{2}\right)\cdot\prod_{j=1}^m\left(\frac{1}{\sigma_j^{(\eta)}}\right)^2}{\sum_{k=1}^{G_n}\exp\left(-3\cdot\eta\cdot\frac{k}{2}\right)\cdot\prod_{j=1}^k\left(\frac{1}{\sigma_j^{(\eta)}}\right)^2}$ with $\sigma_j^{(\eta)}$ as defined in section 2.4.1.
Hence, for all m in $[\![1,G_n]\!]$, the posterior distribution are characterised by :

$$\mathbb{P}^{n,(\eta)}_{M|Y^n}(m) = \frac{\exp\left[-\frac{1}{2}\left(3m\eta - \|\widehat{\theta}^{m,(\eta)}\|^2_{\sigma^{m,(\eta)}}\right)\right]}{\sum_{k=1}^{G_n}\exp\left[-\frac{1}{2}\left(3k\eta - \|\widehat{\theta}^{k,(\eta)}\|^2_{\sigma^{k,(\eta)}}\right)\right]},$$

and

$$\mathbb{P}^{n,(\eta)}_{\boldsymbol{\theta}^M|Y^n} = \sum_{m\in\mathbb{N}^\star}\mathbb{P}^{n,(\eta)}_{\boldsymbol{\theta}^m|Y^n}\cdot\mathbb{P}^{n,(\eta)}_{M|Y^n}(m);$$

and the posterior mean is then $\widehat{\theta}^{M,(\eta)} := \sum_{m\in\mathbb{N}^\star}\widehat{\theta}^{m,(\eta)}\,\mathbb{P}^{n,(\eta)}_{M|Y^n}(m) = \left(\widehat{\theta}^{(\eta)}_j\cdot\mathbb{P}^{n,(\eta)}_{M|Y^n}(M\geq j)\right)_{j\in\mathbb{N}^\star}$.
As we have seen previously with the sieve priors, the iteration procedure conserves the contraction rate.

Corollary 2.4.3 Under **??** and **??**, if, in addition $\log(G_n)/m_n^\circ \to 0$ as $n\to\infty$ then with $D^\circ := D^\circ(\theta^\circ,\lambda) = \lceil 5L/\kappa^\circ\rceil$ and $K^\circ := 10(2\vee\|\theta^\circ\|^2)L^2(16\vee D^\circ\Lambda_{D^\circ})$ we have, for any

η $(1 \leq \eta < \infty)$:

$$\lim_{n\to\infty} \mathbb{E}^n_{\theta^\circ} \left[\mathbb{P}^{n,(\eta)}_{\boldsymbol{\theta}^M|Y^n} \left(\left(K^\circ\right)^{-1} \Phi^\circ_n \leq \|\theta^\circ - \boldsymbol{\theta}^M\|^2_{l^2} \leq K^\circ \Phi^\circ_n\right)\right] = 1.$$

Corollary 2.4.4 Under **??** and **??**, if, in addition, $\log(G_n)/m^\star_n \to 0$ as $n \to \infty$ then, for any η $(1 \leq \eta < \infty)$

- for all θ° in $\Theta_{\mathfrak{a}}(r)$, with $D^\star := D^\star(\mathfrak{a}, \lambda) = \lceil 5L/\kappa^\star \rceil$ and $K^\star := 16\,(2 \vee r)\, L^2\, (16 \vee D^\star \Lambda_{D^\star})\,(1 \vee r)$, we have
$$\lim_{n\to\infty} \mathbb{E}^n_{\theta^\circ} \left[\mathbb{P}^{n,(\eta)}_{\boldsymbol{\theta}^M|Y^n} \left(\|\theta^\circ - \boldsymbol{\theta}^M\|^2 \leq K^\star \Phi^\star_n\right)\right] = 1;$$

- for any monotonically increasing and unbounded sequence K_n holds
$$\lim_{n\to\infty} \inf_{\theta^\circ \in \Theta_{\mathfrak{a}}(r)} \mathbb{E}^n_{\theta^\circ} \left[\mathbb{P}^{n,(\eta)}_{\boldsymbol{\theta}^M|Y^n} \left(\|\theta^\circ - \boldsymbol{\theta}^M\|^2 \leq K_n \Phi^\star_n\right)\right] = 1.$$

Now, in this adaptive case, we consider the eventuality of letting η tend to infinity. In the spirit of the frequentist model selection method presented in section 2.3, define $\Upsilon_\eta(m) = -\sum_{j=1}^{m} \frac{1}{1+\frac{\Lambda_j}{\eta n}} Y_j^2$ and $E_\eta(m) = \mathrm{pen}(m) + \Upsilon_\eta(m)$.
We see that for all m in $[\![1, G_n]\!]$,

$$\mathbb{P}^{n,(\eta)}_{M|Y^n}(m) = \frac{1}{\sum_{k=1}^{G_n} \exp[-\frac{\eta n}{2}\left(E_\eta(k) - E_\eta(m)\right)]}.$$

If η tends to $+\infty$, for all m, $\Upsilon_\eta(m)$ tends to $\Upsilon(m) := -\sum_{j=1}^{m} (Y_j)^2$ and we define for all m, $E(m) := \mathrm{pen}(m) + \Upsilon(m)$.

Interestingly, if we define the contrast Γ for any sequence $\theta^\star$ in Θ as

$$\Gamma\left(\theta^\star\right) := \sum_{j=1}^{G_n} \left(\theta^\star_j\right)^2 \lambda_j^2 - 2\sum_{j=1}^{G_n} \theta^\star_j \lambda_j Y_j,$$

we see, by differentiating Γ summand-wise, that $\overline{\theta}^{G_n}$ minimises this contrast and that $\Gamma\left(\overline{\theta}^{G_n}\right) = \Upsilon\left(G_n\right)$.

If for all k different from m, $E(k) - E(m) > 0$, then $\mathbb{P}^{n,(\eta)}_{M|Y^{(n)}}(m)$ trivially tends to 1 as η tends to ∞. On the other hand, if there exists k such that $E(k) - E(m) < 0$, then $\mathbb{P}^{n,(\eta)}_{M|Y^n}(m)$ obviously tends to 0 as η tends to ∞. So we see that, similarly to the model selection, this method only selects threshold parameters that minimise a penalised contrast.

Note that for all distinct k and m in $[\![1, G_n]\!]$, we almost surely have $E(k) - E(m) \neq 0$ since $\Upsilon(k) - \Upsilon(m)$ is a random variable with absolutely continuous distribution with respect to Lebesgue measure and hence, $\mathbb{P}_{\theta^\circ}[\{\Upsilon(k) - \Upsilon(m) = \mathrm{pen}(k) - \mathrm{pen}(m)\}] = 0$.
We hence define $\widehat{m} := \underset{m \in [\![1, G_n]\!]}{\arg\min}\{E(m)\}$ and $\overline{\theta}^{\widehat{m}}$ the associated projection estimator. Hence, the self informative Bayes limit is $\overline{\theta}^{\widehat{m}}$ and the self informative Bayes carrier is degenerated on it: $\mathbb{P}^{n,(\infty)}_{\boldsymbol{\theta}^M|Y^n} = \delta_{\overline{\theta}^{\widehat{m}}}$.

We obtain here optimality results both for the self informative limit and self informative Bayes carrier.

THEOREM 2.4.1 Consider $\overline{\theta}^{\widehat{m}}$ the frequentist estimator given by the self-informative limit. Under ??, ?? and the condition that $\limsup_{n\to\infty} \frac{\log\left(\frac{G_n^2}{\Phi_n^\circ}\right)}{m_n^\circ} \leq \frac{5}{9L}$, we have

$$\exists C^\circ \in \mathbb{R}_+^\star : \forall \theta^\circ \in \Theta, \quad \mathbb{E}_{\theta^\circ}^n \left[\|\overline{\theta}^{\widehat{m}} - \theta^\circ\|^2 \right] \leq C^\circ \Phi_n^\circ.$$

This first theorem states that, under our set of assumptions, the self-informative limit reaches the oracle rate of the projection estimators.

THEOREM 2.4.2 Under ??, ?? and the condition that $\limsup_{\epsilon\to 0} \frac{\log(G_n)}{m_n^\circ}$, define $D^\circ := \left\lceil \frac{3}{\kappa^\circ} + 1 \right\rceil$ and $K^\circ := 16L \cdot [9 \vee D^\circ \Lambda_{D^\circ}]$; then, we have for all θ° in Θ,

$$\lim_{n\to\infty} \mathbb{E}_{\theta^\circ}^n \left[\mathbb{P}_{\boldsymbol{\theta}^M|Y^n}^{n,(\infty)} \left((K^\circ)^{-1} \Phi_n^\circ \leq \|\boldsymbol{\theta}^M - \theta^\circ\|^2 \leq K^\circ \Phi_n^\circ \right) \right] = 1.$$

This result states that the self informative Bayes carrier contracts with oracle optimal rate of the sieve priors under our set of assumptions.

THEOREM 2.4.3 Consider $\overline{\theta}^{\widehat{m}}$ the frequentist estimator given by the self-informative limit. Then, under ??, ?? and the condition that $\limsup_{n\to\infty} \frac{\log\left(\frac{G_n^2}{\Phi_n^\star}\right)}{m_n^\star} < \frac{5}{9L}$, we have

$$\exists C^\star \in \mathbb{R}_+^\star : \quad \sup_{\theta^\circ\in\Theta} \mathbb{E}_{\theta^\circ}^n \left[\|\overline{\theta}^{\widehat{m}} - \theta^\circ\|^2 \right] \leq C^\star \Psi_n^\star.$$

This result shows that the self-informative limit converges with minimax optimal rate over Sobolev's ellipsoids under our set of assumptions.

THEOREM 2.4.4 Under ??, ?? and the condition that $\limsup_{n\to\infty} \frac{\log(G_n)}{m_n^\star}$, define $D^\star := \left\lceil \frac{3(1\vee L^\circ)}{\kappa^\star} + 1 \right\rceil$ and $K^\star := 9L\,(1 \vee L^\circ)\, D^\star \Lambda_{D^\star}$; then, we have for all θ° in $\Theta^{\mathfrak{a}}(r)$,

$$\lim_{n\to\infty} \mathbb{E}_{\theta^\circ}^n \left[\mathbb{P}_{\boldsymbol{\theta}^M|Y^n}^{n,(\infty)} \left(\|\boldsymbol{\theta}^M - \theta^\circ\|^2 \leq K^\star \Psi_n^\star \right) \right] = 1,$$

and, for any increasing function K_n such that $\lim_{n\to\infty} K_n = \infty$,

$$\lim_{n\to\infty} \sup_{\theta^\circ\in\Theta^{\mathfrak{a}}(r)} \mathbb{E}_{\theta^\circ}^n \left[\mathbb{P}_{\boldsymbol{\theta}^M|Y^n}^{n,(\infty)} \left(\|\boldsymbol{\theta}^M - \theta^\circ\|^2 \leq K_n \Psi_n^\star \right) \right] = 1.$$

2.5 Second application example: the circular density deconvolution model

2.6 On the shape of the posterior mean

We have hence seen that in a general case, considering the asymptotic iteration , the posterior distribution using a sieve prior contracts around the projection estimator and

while using a hierarchical prior, the posterior contracts around some penalised contrast model selection estimator.
It is also interesting to note that for any number of iteration η, the posterior mean can be written both as a shrinkage and as an aggregation estimator. Indeed, we have shapes
Using the same methodology as in JOHANNES ET AL. (2016) we can obtain, in the Gaussian case, the following optimality properties for the posterior mean of the iterated hierarchical prior. properties

Aggregation estimates gathered a lot of interest, see for example Tsybakov. While considering such estimators, optimality is formulated in the following way:
We can see that our estimator reaches this optimality criterion. We will hence consider this shape of estimator in another inverse problem, the circular density deconvolution.

Chapter 3

Minimax and oracle optimal adaptive aggregation

We inquire in this chapter the properties of aggregation estimators as introduced in SECTION 2.6. We introduce first a skim of proof for oracle and minimax optimality of this kind of estimator before applying it to the inverse Gaussian sequence space and the circular deconvolution models respectively introduced in SECTION 1.4.1 and SECTION 1.4.2, including in presence of dependance and partially known operator.

3.1 Independent observations with unknown noise distribution

3.2 Independent observations with unknown noise distribution

3.3 Independent observations with known noise density

3.4 Absolutely regular observation process with known noise density

3.5 Independent observations with unknown noise distribution

3.6 Absolutely regular observation process with unknown noise distribution

Appendix A

Simulation skim

Appendix B

R Code

Bibliography

Billingsley, P. (1986). *Probability and Measure*. John Wiley and Sons, second edition.

BUNKE, O. and JOHANNES, J. (2005). Selfinformative limits of bayes estimates and generalized maximum likelihood. *Statistics*, 39(6):483–502.

Ghosal, S., Ghosh, J. K., Van Der Vaart, A. W., et al. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531.

Hadamard, J. (1902). Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52.

JOHANNES, J. and SCHWARZ, M. (2012). Adaptive gaussian inverse regression with partially unknown operator.

JOHANNES, J., SIMONI, A., and SCHENK, R. (2016). Adaptive bayesian estimation in indirect gaussian sequence space model.

Liptser, R. S. and Shiryaev, A. N. (2013). *Statistics of random processes: I. General theory*, volume 5. Springer Science & Business Media.

MASSART, P. (2003). Concentration inequalities and model selection. In Springer, editor, *Ecole d'Été de Probabilités de Saint-Flour XXXIII*, volume 1869.