"The Importance of Being (Asymptotically) Earnest"

Four short adventures in balls and bands for non–parametric models.

[Thesis Proposal]

Pierpaolo Brutti

Introduction.

Over the past decades, adaptive estimation of an unknown regression function has received much attention. What *adaptivity* means in practice is that we are able to find a *single* procedure that performs (essentially) optimally in a minimax sense for a wide range of smoothness assumptions simultaneously, even when the degree of smoothness is *unknown*.

The desire to go beyond estimation and have *honest* confidence statements is what motivates the title. For the sake of consistency with the structure of its noble literary ancestor, I divided thesis work into four Acts, each of them illustrating a particular research topic I am currently exploring, plus a Prologue containing background, and an Epilogue where conclusions are drawn along with the description of possible applications in astrophysics.

The structure of this proposal, instead, is as follows. In Section 1 I provide the basic tools and terminology of modern nonparametric statistics, whereas a review of recently developed techniques to build nonparametric confidence sets is given in Sections 2 and 3 along with some optimality results. In Section 4 I describe my proposed thesis work. Finally the Appendix contains some additional background material.

1 A Primer in Nonparametric Regression.

This section is devoted to the introduction of the basic terminology and review of some recent results concerning the problem of function estimation. I will start by reviewing some concepts in function space, minimax optimality and adaptation. Then I will move quickly to the main theme of this proposal: nonparametric confidence sets (nCS's)

1.1 Function spaces: some notions and notation.

An Hilbert space $\mathscr{H}$ is a complete *normed* space whose norm is indexed by an inner (or scalar) product. Two disjoint subspaces $\mathscr{A}$ and $\mathscr{B}$ of a space $\mathscr{S}$ form a direct sum decomposition of $\mathscr{S}$ if every element of $\mathscr{S}$ can be written *uniquely* as a sum of an element of $\mathscr{A}$ and an element of $\mathscr{B}$. The

notation $\mathscr{S} = \mathscr{A} \oplus \mathscr{B}$ is then used. A (Lebesgue) measurable function $f(\cdot)$ belongs to the Lebesgue space $\mathsf{L}^p(\mathbb{R})$, $p \in [1, +\infty)$ if

$$\left\|f\right\|_{\mathsf{L}^p} \triangleq \left[\int_{\mathbb{R}} \left|f(x)\right|^p \mathrm{d}x\right]^{\frac{1}{p}} < +\infty$$

An example of Hilbert space is the Lebesgue space $\mathsf{L}^2(\mathbb{R})$ of measurable and square integrable functions. A countable subset $\{f_k\}_k$ of functions belonging to the Hilbert space $\mathscr{H} = \mathsf{L}^2(\mathbb{R})$ is a *Riesz basis* if $\forall\ f \in \mathscr{H}$ we have

1. $f(\cdot)$ can be written *uniquely* as
$$f(x) = \sum_k \theta_k f_k(x)$$
2. two positive constants A and B with $A < B$ exist such that
$$A \left\|f\right\|_{\mathsf{L}^2}^2 \leqslant \sum_k |\theta_k|^2 \leqslant B \left\|f\right\|_{\mathsf{L}^2}^2$$

A Riesz basis is an orthogonal basis if the $f_k(\cdot)$ are mutually orthogonal. In this case, $A = B = 1$. So, intuitively, the absolute value of A and B determine the "stability" of the system $\{f_k\}_k$ in reconstructing elements of $\mathscr{H}$ (redundancy, in general, generates unstable reconstructions).

1.2 Pointwise Estimation.

The main goal in this work will be to propose new inferential methods to recover an unknown function from noisy data. For this reason, unless otherwise stated, the two settings I will refer throughout the rest of this section are:

The basic nonparametric regression model:

$$Y_i = f(x_i) + \sigma\, \varepsilon_i, \quad i \in \{1, \dots, n\}, \tag{1.1}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \mathsf{I}_n)$, σ is assumed known, $\{x_i\}_{i\in\{1,\dots,n\}}$ are fixed (equispaced) design points and $f \in \mathsf{L}^2\Big([0,1]\Big)$ is an uknown (smooth) function to be recovered

and, interchangeably, its strict parent:

The *Normal means* model:

$$Z_i = \theta_i + \sigma_n\, \varepsilon_i, \quad i \in \{1, \dots, n\}, \tag{1.2}$$

where, again, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \mathsf{I}_n)$, $\boldsymbol{\theta}_n = [\theta_1, \dots \theta_n]^\intercal \in \mathbb{R}^n$ is a vector of unknown parameters and σ_n, in general equal to $\sigma/\sqrt{n}$, is assumed to be known in advance.

The way to move from one model to the other is via a sequence of basis functions and, heuristically[1], it works as follow:

Step 1: Start from Equation 1.1 and let $\{\psi_j\}_{j\in\mathbb{N}}$ be any orthonormal basis for $\mathsf{L}^2\Big([0,1]\Big)$, the space where $f(\cdot)$ lives. Expand $f(\cdot)$ in this basis and write

$$f(x) = \sum_{j\in\mathbb{N}} \theta_j \psi_j(x), \quad \forall\, x \in [0,1],$$

with $\theta_j = \Big\langle f, \psi_j \Big\rangle_{\mathsf{L}^2} = \int_{[0,1]} f(x)\psi_j(x)\mathrm{d}x$. Of course, truncating the expansion to the n–th term we obtain an approximation to $f(\cdot)$ given by

$$f(x) \approx \sum_{j=1}^{n} \theta_j \psi_j(x), \quad \forall\, x \in [0,1].$$

Step 2: Define the following random variables

$$Z_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \psi_j(x_i), \quad \forall\, j \in \{1,\dots n\},$$

and observe that $\mathbf{Z}_n = [Z_1,\dots,Z_n]^\mathsf{T} \sim \mathcal{N}_n\Big(\boldsymbol{\theta}_n, \sigma_n^2 \mathsf{I}_n\Big)$.

Step 3: To complete the (statistical) "metamorphosis", notice that being L^2 an Hilbert space, squared loss for an estimator $\widehat{f}(x) = \sum_{j\in\mathbb{N}} \widehat{\theta}_j \psi_j(x)$ of $f(x)$ is equivalent, by *Parseval identity*, to squared error loss in sequence–space for $\widehat{\boldsymbol{\theta}}$ estimator of $\boldsymbol{\theta}$

$$\Big\|\widehat{f} - f\Big\|_{\mathsf{L}^2}^2 = \int_{[0,1]} \Big(\widehat{f}(x) - f(x)\Big)^2 \mathrm{d}x = \sum\nolimits_{j\in\mathbb{N}} \Big(\widehat{\theta}_j - \theta\Big)^2 \triangleq \Big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big\|_{\ell^2}^2$$

We can now move on and consider what kind of statistical properties any reasonable pointwise estimator $\widehat{f_n}(\cdot)$ of the regression function $f(\cdot)$ should satisfy to be considered theoretically "good". The following four concepts are those that more frequently appear in contemporary nonparametric statistic:

- **Minimaxity:** Assume that $f(\cdot)$, the function to be estimated, belongs to a given functional space or class of smoothness $\mathcal{F}_{\alpha_0}$ indexed by α_0 known in advance. In the next section we will relax this assumption. Define the L^p*–risk* of an arbitrary estimator T_n based on the sample data as $\mathbb{E}\,\Big\|T_n - f\Big\|_p^p$ with $p \in [1,+\infty]$, and consequently, its L^p*–minimax risk* by

$$\mathsf{R}_n\Big(\mathcal{F}_{\alpha_0}, p\Big) = \inf_{T_n} \sup_{f\in\mathcal{F}_{\alpha_0}} \mathbb{E}\,\Big\|T_n - f\Big\|_p^p,$$

where the infimum is taken over all measurable estimators T_n of $f(\cdot)$.

[1] See [10, 50, 63] for further details and connections with the *white noise model*.

Def. 1.1 *The sequence* $\{a_n\}_n \in \mathbb{N}$ *is called the optimal rate of convergence, (or minimax rate of convergence) on the class* $\mathcal{F}_{\alpha_0}$ *for the* L^p*–risk if* $a_n \sim R_n(\mathcal{F}_{\alpha_0}, p)^{1/p}$*. We say that an estimator* $\widehat{f_n}$ *of* $f(\cdot)$ *attains the optimal rate of convergence if*

$$\sup_{f \in \mathcal{F}_{\alpha_0}} \mathbb{E} \left\| T_n - f \right\|_p^p \sim R_n\Big(\mathcal{F}_{\alpha_0}, p\Big).$$

As an example, in [55, 91], the authors found that, under a squared loss ($p = 2$), the optimal rate of convergence attainable by an estimator when the underlying function belongs to the *Sobolev* class

$$\mathcal{W}^{p,s}(C) = \left\{ f : \left\| f \right\|_p^p + \left\| \tfrac{\mathrm{d}^s}{\mathrm{d}x^s} f(x) \right\|_p^p \leqslant C^2 \right\},$$

is $a_n = -\frac{s}{2s+1}$, hence $\mathsf{R}_n\Big(\mathcal{W}^{p,s}(C), 2\Big) = n^{-\frac{2s}{2s+1}}$, and actual estimators that achieve the minimax bound were also given. More in general, minimax theory under the Normal means model for mean squared error – but also confidence intervals and probabilistic error, see Section 2 – can all be precisely characterized by the so called *modulus of continuity* introduced by Donoho and Liu [37]. Specifically, for any linear functional T and convex parameter space $\mathcal{F}$, the minimax mean squared error is of order $\omega^2\Big(\frac{1}{\sqrt{n}}, \mathcal{F}\Big)$ where the *modulus* $\omega(\epsilon, \mathcal{F})$ is defined by

$$\omega(\epsilon, \mathcal{F}) = \sup_{\{f,g \in \mathcal{F} : \left\| f-g \right\|_{\mathsf{L}^2} \leqslant \epsilon\}} \Big\{ \left| \mathsf{T}f - \mathsf{T}g \right| \Big\},$$

and linear procedure can be constructed which are (*nearly*) minimax in the previous sense. See Ibragimov Hasminskii [55], Donoho and Liu [37] and Donoho [34] for precise results.

- **Maxiset:** The minimax approach has some drawbacks: the choice for the function classes is quite subjective and exhibiting an estimator well adapted to the worst functions of this class seems too pessimistic for practical purposes. Among others, these are the reasons why, DeVore, Kerkyacharian and Picard [26] and Kerkyacharian and Picard [68, 67] focused on an alternative to the minimax setting: the *maxiset* approach which consists in investigating the maximal space – called the *maxiset* – where an estimation procedure achieves a given rate of convergence. For instance, these authors applied the theory to two well known non-parametric procedures: wavelet thresholding and local bandwidth selection. By showing that the maxiset for the first one is included into the maxiset for the second one, they could conclude that local bandwidth selection is at least as good as the thresholding procedures.

  Formally a *maxiset* is defined in great generality as follow

  Def. 1.2 *Consider a sequence of models* $\mathscr{E}_n = \left\{ \mathbb{P}_\theta^n \right\}_{\theta \in \Theta}$*, where* $\mathbb{P}_\theta^n$*'s are probability distribution on the measurable space* Ω_n*, and* Θ *is a generic parameter–set. Consider also a sequence of estimate* $\widehat{g_n}$ *of a function* $g(\theta)$ *of the parameter* θ*, a loss function* $L_n\Big(\widehat{g_n}, g(\theta)\Big)$*, and a rate of convergence* $\{a_n\}_{n \in \mathbb{N}}$ *tending to zero. Then, we define maxiset associated with the sequence* $\{\widehat{g_n}\}_{n \in \mathbb{N}}$*, the loss function* L_n*, the rate* $\{a_n\}_{n \in \mathbb{N}}$ *and the constant* C *as the following set:*

$$MS(\widehat{g_n}, L_n, a_n)[C] = \left\{ \theta \in \Theta : \sup_{n \in \mathbb{N}} \frac{\mathbb{E}_\theta L_n\Big(\widehat{g_n}, g(\theta)\Big)}{a_n} \leqslant C \right\}.$$

Hence, instead of a priori fixing a (functional) set such as a Hölder, Sobolev or Besov ball as it is the case in a minimax framework, in the maxiset framework the parameter set Θ can be as large as the set of bounded, measurable functions.

- **Adaptivity:** It may not be sufficient to know that for $f(\cdot)$ belonging to a given space, the estimator performs well. Indeed, in general we do not know which space the function belongs to. Hence it is of great interest to consider a scale of function classes and to look for an estimator that attains simultaneously the best rates of convergence across the whole scale. For example, the L^q–Sobolev scale is a set of Sobolev function classes indexed by the parameters s and C of a Sobolev class. We now formalize the notion of an adaptive estimator. Let $\mathcal{A}$ be a given set and let $\{\mathcal{F}_\alpha, \alpha \in \mathcal{A}\}$ be the scale of functional classes $\mathcal{F}_\alpha$ indexed by $\alpha \in \mathcal{A}$. Denote $\mathsf{R}_n(\alpha, p)$ the minimax risk over $\mathcal{F}_\alpha$ for the L^p–loss:

$$\mathsf{R}_n(\alpha, p) = \inf_{\widehat{f_n}} \sup_{f \in \mathcal{F}_\alpha} \mathbb{E} \left\| \widehat{f_n} - f \right\|_{\mathsf{L}^p}^p .$$

Def. 1.3 *The estimator $\widehat{f_n}(\cdot)$ is called rate adaptive for the L^p loss and the scale of classes $\mathcal{F}_\alpha$ with $\alpha \in \mathcal{A}$ if for any $\alpha \in \mathcal{A}$ there exists $c_\alpha > 0$ such that*

$$\sup_{f \in \mathcal{F}_\alpha} \mathbb{E} \left\| \widehat{f_n} - f \right\|_{\mathsf{L}^p}^p \leqslant c_\alpha R_n(\alpha, p), \quad \forall\, n \in \mathbb{N}.$$

The estimator $\widehat{f_n}(\cdot)$ is called adaptive up to a logarithmic factor for the L^p loss and the scale of classes $\mathcal{F}_\alpha$ with $\alpha \in \mathcal{A}$ if for any $\alpha \in \mathcal{A}$ there exist $c_\alpha > 0$ and $\gamma = \gamma_\alpha > 0$ such that

$$\sup_{f \in \mathcal{F}_\alpha} \mathbb{E} \left\| \widehat{f_n} - f \right\|_{\mathsf{L}^p}^p \leqslant c_\alpha (\log n)^\gamma R_n(\alpha, p), \quad \forall\, n \in \mathbb{N}.$$

Thus, adaptive estimators have an optimal rate of convergence and behave as if they know in advance in which class the function to be estimated lies.

The theory of adaptive estimation depends strongly on how the risk is measured. When the performance is measured "globally" through, for example, the usual (integrate) L^p loss, then full adaptation can often be achieved [13, 36, 53]. But when the performance is measured at a point, it if often the case that full adaptation is not possible and in general a logarithmic penalty must be paid for not knowing the smoothness (see [11, 14, 39, 74]). The goal is then to construct estimators which, for a range of parameter spaces, are both minimax rate optimal for *integrated* squared error loss and pay only a logarithmic penalty for squared error loss *at each point* (see [13]). To bridge the gap between these two extremes, Cai and Low [18] developed a theory of superefficiency (see below) and adaptation under flexible performance measures, which provides a multiresolution view of risk by cropping a *global* loss outside some neighborhood that shrinks toward a target–point as the sample size increases. Wavelet procedures are also given which adapt rate optimally for given shrinking neighborhoods (see [18]).

- **Superefficiency:** The theory of adaptive estimators is closely connected to that of *superefficient* estimators which in turn depend on how the risk is measured. The concept of *asymptotic efficiency* was invented by Fisher as early as 1922 roughly in the form as we use it for "regular" models today: a sequence of statistics is *efficient* if it tends to a normal distribution with the least possible standard deviation. Successively Cramér, systematizing the topic in the mid '40, introduced the concept of *asymptotic efficiency* of an estimator as the quotient of the inverse Fisher information and the asymptotic variance, and ended up defining an estimator sequence to be *asymptotically efficient* if its asymptotic efficiency equals one: the classical result about the asymptotic efficiency of maximum likelihood estimators (under regularity conditions) saw the light in this same period. But, as defined, the concept of efficiency was too weak. In 1951, in fact, Hodges produced the first example of a *superefficient* estimator sequence: an estimator sequence with efficiency *at least* one for *all* value of the parameter θ, and *more* than one for *some* θ. Although Hodges' example threw doubt on Fisher's result concerning the asymptotic efficiency of maximum likelihood estimators, was Lucien Le Cam, shortly after [69], that settled the controversy showing that, for regular parametric models, a sequence of estimators can be superefficient on at most a Lebesgue null set[2].

  Much of the recent work on nonparametric function estimation can be interpreted as an attempt to construct superefficient estimators with desirable properties such as, adaptivity (see [6, 7]).

  To appreciate the link between superefficiency and adaptation, we need the following definition

  Def. 1.4 *For a parameter space $\mathcal{F}_{\alpha_0}$ we call an estimator $\widehat{f_n}(\cdot)$ superefficient at $f \in \mathcal{F}_{\alpha_0}$ under a loss function $L_n\left(\widehat{f_n}, f\right)$ – for example $L_n\left(\widehat{f_n}, f\right) = \left\|\widehat{f_n} - f\right\|_{L^p}$ – if the risk at f converges faster then the minimax risk, namely*

$$\frac{\mathbb{E}_f L_n\left(\widehat{f_n}, f\right)}{\inf\limits_{\widehat{f_n}} \sup\limits_{f \in \mathcal{F}_{\alpha_0}} \mathbb{E}_f L_n\left(\widehat{f_n}, f\right)} \to 0 \tag{1.3}$$

  Now consider two *nested* function classes $\mathcal{F}_{\alpha_2} \subset \mathcal{F}_{\alpha_1}$. Then, according to Definition 1.3, a fully rate adaptive estimator $\widehat{f_n}(\cdot)$ over these classes w.r.t. the loss function L_n would necessary satisfy

$$\sup_{f \in \mathcal{F}_{\alpha_i}} \mathbb{E}_f L_n\left(\widehat{f_n}, f\right) \asymp \mathsf{R}_n\left(\mathcal{F}_{\alpha_i}, L_n\right) = \inf_{\widehat{f_n}} \sup_{f \in \mathcal{F}_{\alpha i}} \mathbb{E}_f L_n\left(\widehat{f_n}, f\right), \quad \forall\, i \in \{1, 2\}.$$

  The risk of $\widehat{f_n}$ for each $f \in \mathcal{F}_{\alpha_2}$ must then converge faster than the minimax risk over the larger parameter space $\mathcal{F}_{\alpha_1}$. Hence such estimators must be superefficient at each $f \in \mathcal{F}_{\alpha_2}$ with respect to $\mathcal{F}_{\alpha_1}$. The circle is closed.

[2] See [70, 71] for further details.

2 Optimality of Confidence Sets.

- **The Consequences of "Honesty":** A first asymptotic insight might be drawn from the following Theorem due to Li

  Theorem 2.1 ([76], Theorem 2.1). *Consider the Normal means model (1.2) and let*

  $$\mathfrak{B}_n = \left\{ \boldsymbol{\theta}_n \in \mathbb{R}^n : \left\| \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n \right\|_{\mathbb{R}^n}^2 \leqslant d_n \right\},$$

  where $\widehat{\boldsymbol{\theta}}_n$ is an estimator of $\boldsymbol{\theta}_n$ and $d_n(\mathbf{Z}_n)$ is the (random) radius of the ball. Suppose that

  $$\liminf_{n \to +\infty} \inf_{\boldsymbol{\theta}_n \in \mathbb{R}^n} \mathbb{P}_{\boldsymbol{\theta}_n} (\boldsymbol{\theta}_n \in \mathfrak{B}_n) \geqslant 1 - \alpha \tag{2.1}$$

  Then, for any sequence $\{\boldsymbol{\theta}_n\}_{n \in \mathbb{N}}$ and any $C_n \to 0$,

  $$\limsup_{n \to +\infty} \mathbb{P}_{\boldsymbol{\theta}_n} \left(d_n \leqslant C_n \sigma_n n^{\frac{1}{4}} \right) \leqslant \alpha.$$

  For asymptotic confidence procedures, a natural requirement is *uniform coverage*:

  $$\sup_{\boldsymbol{\theta}_n \in \mathbb{R}^n} \left| \mathbb{P}\{\mathfrak{B}_n \ni \boldsymbol{\theta}_n\} - (1 - \alpha) \right| \to 0 \tag{2.2}$$

  so that the coverage error depends only on n and not on $\boldsymbol{\theta}_n$. Theorem 2.1 clearly shows that with no smoothness assumptions on $\boldsymbol{\theta}_n$, any $(1 - \alpha)$ confidence set of the form

  $$\mathfrak{B}_n = \left\{ \boldsymbol{\theta}_n \in \mathbb{R}^n : n^{-1/2} \left\| \boldsymbol{\theta}_n - \widehat{\boldsymbol{\theta}}_n \right\|_2 \leqslant \mathrm{d}_n \right\}$$

  that is asymptotically honest in the sense that

  $$\liminf_{n \to \infty} \inf_{\boldsymbol{\theta}_n \in \mathbb{R}^n} \mathbb{P}_{\boldsymbol{\theta}_n} \left\{ \mathfrak{B}_n \ni \boldsymbol{\theta}_n \right\} \geqslant 1 - \alpha,$$

  must necessarily have $\mathrm{d}_n \geqslant C\, n^{-1/4}$. This implies that, in general, the radius does not *adapt* to the smoothness of the unknown function: the smoothness of $f(\cdot)$ does not affect in any good way the convergence rate of the confidence set. Robins and van der Vaart [86], Juditsky and Lambert–Lacroix [65], and Cai and Low [15] show that some degree of adaptivity is possible but quite restricted.

- **Probabilistic Error and Fixed Radius Sets:** In Section 1.2 we saw how the minimax theory for mean squared error can be characterized by a modulus of continuity. This same quantity plays an important role even when we consider other types of losses strictly related to confidence sets. In fact, moving from mean squared to probabilistic error; that is, to loss functions that measure the probability that the estimator is close to the unknown function, the optimal rates of convergence for estimating a linear functional T over a convex parameter space $\mathcal{F}$ turns out to be $\omega\left(\frac{1}{\sqrt{n}}, \mathcal{F}\right)$. More precisely, results in [34] and Cai and Low [16] show that

- $\forall\ \alpha \in (0,1],\ \exists\ \widehat{\mathsf{T}}$ linear estimator : $\sup_{f\in\mathcal{F}} \mathbb{P}_f\left\{\left|\widehat{\mathsf{T}} - \mathsf{T}f\right| > \frac{3}{2}\omega\left(\frac{z_{\alpha/2}}{\sqrt{n}}, \mathcal{F}\right)\right\} \leqslant \alpha,$
- $\forall\ \widehat{\mathsf{T}}$ linear estimator, $\exists\ \alpha \in (0, \frac{1}{2})$: $\sup_{f\in\mathcal{F}} \mathbb{P}_f\left\{\left|\widehat{\mathsf{T}} - \mathsf{T}f\right| > \frac{1}{2}\omega\left(\frac{z_{\alpha}}{\sqrt{n}}, \mathcal{F}\right)\right\} > \alpha.$

These bounds on probabilistic error have direct consequences for the construction of *fixed* length confidence intervals and in particular show that for *any* given coverage, the shortest fixed length interval has length of order $\omega\left(\frac{1}{\sqrt{n}}, \mathcal{F}\right)$ (see Donoho [34]).

- **Probabilistic Error Adaptivity and Recentered Confidence Sets:** In contrast to mean squared error, *full* adaptation for linear functional under probabilistic error can commonly be achieved. In particular there are many examples where an estimator can be constructed which is simultaneously rate optimal under probabilistic error and for which there do not exist simultaneously minimax rate optimal mean squared error estimators [16]. An interesting question is then whether we can find estimators which have both good mean squared error performance and also optimal probabilistic error. In [17], Cai and Low quantify the penalty that must be pay in probabilistic error when the estimator is optimal in mean squared error, and, viceversa, in mean squared loss when the estimator performs well in probabilistic error. An immediate consequence of these results is that centering confidence intervals on adaptive mean squared error estimators in general yields suboptimal confidence procedures: either the resulting interval has poor coverage probability, or it is unnecessarily long.

3 Available Methods: a Review.

There are several approaches to computing confidence bands and balls. Working in function space, for example, one approach is to use pointwise confidence intervals on a very fine grid of the observation interval. The level of these confidence intervals can be adjusted by the Bonferroni method in order to obtain uniform confidence bands. The gaps between the grid points can be bridged via smoothness conditions on the regression curve. A drawback to the Bonferroni approach is that the resulting intervals will quite often be too long. The reason is that this method does not make use of the substantial positive correlation of the curve estimates at nearby points. Another approach is to consider $\widehat{f_n}(x) - f(x)$ as a stochastic process (in x) and then to derive asymptotic Gaussian approximations to that process. The extreme value theory of Gaussian processes yields the level of these confidence bands. A third approach is based on the *bootstrap*. By resampling one attempts to approximate the distribution of

$$\mathsf{L}_n\left(\widehat{f_n}, f\right) = \left\|\widehat{f_n} - f\right\|_{\mathsf{L}^\infty} = \sup_{x\in[0,1]} \left|\widehat{f_n}(x) - f(x)\right|;$$

which yields lower and upper bands computed as the $(\alpha/2)$ and $(1-\alpha/2)$ quantiles of L_n, respectively. Another bootstrap method is based on approximating the distribution of $\widehat{f_n}(x) - f(x)$ at distinct points x and then to simultaneously correct the pointwise confidence intervals in order to obtain the joint coverage probability $1-\alpha$.

In this section I will briefly review some of the most common methods available in the literature to build reliable confidence bands around a given scatter plot smoother or, alternatively, confidence balls in sequence space for smoothers based on orthogonal expansions. We will see how most of them, in one way or another, are essentially based on some particular estimator of the *loss* associated to the smoother. In the following will be convenient to drop the subscript and write $\boldsymbol{\theta}$ and $\mathbf{Z}$ instead of $\boldsymbol{\theta}_n$ and $\mathbf{Z}_n$.

- **"Classical" methods:** The following are two of the simplest – and probably oldest – methods available to build confidence sets with guaranteed coverage.

  [1]– χ^2 : The classical α–level confidence set for $\boldsymbol{\theta}$ in (1.2) is based on the well–known fact that $\frac{1}{\sigma_n^2}\|\mathbf{Z}-\boldsymbol{\theta}\|^2_{\mathbb{R}^n}$ has a chi–squared distribution with n degrees of freedom, so that an obvious candidate is the following

$$\mathfrak{B}_{\chi^2} = \left\{\boldsymbol{\theta}\in\mathbb{R}^n : \|\mathbf{Z}-\boldsymbol{\theta}\|_{\mathbb{R}^n} \leqslant \sqrt{\sigma_n^2\,\chi^2_{n,\alpha}}\right\} = \left\{\boldsymbol{\theta}\in\mathbb{R}^n : \|\mathbf{Z}-\boldsymbol{\theta}\|^2_{\mathbb{R}^n} \leqslant \sigma_n^2\,\chi^2_{n,\alpha}\right\}, \tag{3.1}$$

  where $\chi^2_{n,\alpha}$ is the upper α–quantile of a χ^2_n. Evidently $\mathbb{P}_{\boldsymbol{\theta}}\Big(\mathfrak{B}_{\chi^2} \ni \boldsymbol{\theta}\Big)$ is exactly $1-\alpha$, for every $\boldsymbol{\theta}\in\mathbb{R}^n$ and, by the triangular array central limit theorem, $\mathfrak{B}_{\chi^2}$ is a *fixed* radius sphere centered at the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} = \mathbf{Z}$ whose radius is approximatively $\sigma_n^2\Big(n + \sqrt{2n}\,\Phi^{-1}(\alpha)\Big)$ for *large* values of n. Thus

$$\lim_{n\to+\infty}\sup_{\|\boldsymbol{\theta}\|^2_{\ell^2}\leqslant Cn}\left|\mathsf{g}_n\left(\mathfrak{B}_{\chi^2},\boldsymbol{\theta}\right) - 2\sigma_n\right| = 0, \quad \forall\ \mathbb{R}_+ \ni C < +\infty,$$

  where, for a given $\boldsymbol{\theta}_0\in\mathbb{R}^n$, $\mathsf{g}_n(\mathfrak{B},\boldsymbol{\theta}_0)$ denotes the *geometrical risk* of the confidence set $\mathfrak{B}$ as a set–valued estimator of $\boldsymbol{\theta}$, and is defined as follow

$$\mathsf{g}_n(\mathfrak{B},\boldsymbol{\theta}_0) \triangleq \tfrac{1}{\sqrt{n}}\mathbb{E}_{\boldsymbol{\theta}_0}\left[\sup_{\boldsymbol{\theta}\in\mathfrak{B}}\|\boldsymbol{\theta}_0-\boldsymbol{\theta}\|_{\mathbb{R}^n}\right]. \tag{3.2}$$

  In other words, the geometrical risk is the expected distance to $\boldsymbol{\theta}_0$ from the most distant point in the confidence set.

  Notice that the approximate critical value for $\mathfrak{B}_{\chi^2}$ can be found directly from the asymptotic normal distribution of the difference

$$\tfrac{1}{\sqrt{n}}\left[\|\mathbf{Z}-\boldsymbol{\theta}\|^2_{\mathbb{R}^n} - n\sigma_n^2\right], \tag{3.3}$$

  which compares the quadratic loss of $\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} = \mathbf{Z}$ with an *unbiased* estimator of its risk. We will see how, moving from fixed to random radius balls, we can improve upon $\mathfrak{B}_{\chi^2}$ within the tight constrained discussed in Section 2.

  [2]– S ´– : Under a standard parametric model

$$Y_i = f(\mathbf{x}_i) + \sigma\,\varepsilon_i = \boldsymbol{\psi}(\mathbf{x}_i)^{\mathsf{T}}\boldsymbol{\theta} + \sigma\,\varepsilon_i, \quad \forall\, i\in\{1,\dots,n\},$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of unknown parameters, $\boldsymbol{\psi}(\mathbf{x}) \in \mathbb{R}^p$ for each $\mathbf{x} \in x \subset \mathbb{R}^d$ is a known vector function of predictors, σ^2 is the unknown variance, and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \mathsf{I}_n)$, Scheffe [88] proposed to use

$$\kappa = \sqrt{p \cdot \mathsf{F}^{\alpha}_{p,n-p}}$$

where $p = \mathsf{rank}(\Psi)$ with $\Psi = \Big[\boldsymbol{\psi}(\mathbf{x}_1), \ldots, \boldsymbol{\psi}(\mathbf{x}_n)\Big]$, and $\mathsf{F}^{\alpha}_{p,n-p}$ is the upper α quantile of an F distribution with $(p, n-p)$ degrees of freedom, to build the following *predictive* confidence bands for $f(\mathbf{x})$:

$$\mathfrak{B}_{\kappa}(\mathbf{x}) = \Big\{ y \in \mathbb{R} : \big|\widehat{f_n}(\mathbf{x}) - y\big| \leqslant \kappa \widehat{\sigma} \|\boldsymbol{\psi}^{\star}(\mathbf{x})\|_{\mathbb{R}^n} \Big\}, \tag{3.4}$$

where $\widehat{\sigma}$ is any consistent estimator of the variance and $\widehat{f_n}(\cdot)$ is the least square *linear* smoother

$$\widehat{f_n}(\mathbf{x}) = \sum_{i=1}^{n} \psi_i^{\star}(\mathbf{x}) Y_i = \langle \boldsymbol{\psi}^{\star}(\mathbf{x}), \mathbf{Y} \rangle_{\mathbb{R}^n},$$

with smoothing vector at $\mathbf{x}$ given by

$$\boldsymbol{\psi}^{\star}(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})^{\mathsf{T}} \Big(\Psi^{\mathsf{T}} \Psi \Big)^{-1} \Psi^{\mathsf{T}}.$$

- **Tube–formula method [78, 93, 94]:** To build *simultaneous* confidence bands around a *linear* smoother

$$\widehat{f_n}(\mathbf{x}) = \sum_{i=1}^{n} \psi_i^{\star}(\mathbf{x}) Y_i = \langle \boldsymbol{\psi}^{\star}(\mathbf{x}), \mathbf{Y} \rangle_{\mathbb{R}^n}, \quad \mathbf{x} \in x \subset \mathbb{R}^d,$$

one solution is the *volume of tubes* formula proposed in [93]. The idea is simply to replace the very *conservative* upper–bound to the non–coverage probability adopted by the Scheffe's method, with a better, more direct one. To this end, we consider the random process inside the non–coverage probability,

$$\mathsf{G}_n(\mathbf{x}) = \frac{\widehat{f_n}(\mathbf{x}) - f(\mathbf{x})}{\|\boldsymbol{\psi}^{\star}(\mathbf{x})\|},$$

then, under the nonparametric regression model (1.1), $\mathsf{G}_n(\mathbf{x})$ converges to a Gaussian process $\mathbb{G}(\mathbf{x})$ on x with mean zero and covariance function $\mathbb{C}ov(\mathbf{x}, \mathbf{x}')$. This leads to a general question about how to find good approximation to

$$\mathbb{P}\left\{ \sup_{\mathbf{x} \in x} \mathbb{G}(\mathbf{x}) > z \right\}, \quad \text{as } z \to \infty. \tag{3.5}$$

Sun [92] has a two-term approximation for a general class of random fields when x does not have boundary:

$$\mathbb{P}\left\{ \sup_{\mathbf{x} \in x} \mathbb{G}(\mathbf{x}) > z \right\} \approx \kappa_0 \Gamma_0(z) + \kappa_2 \Gamma_2(z), \tag{3.6}$$

where $\Gamma_i(\cdot)$ are incomplete Gamma functions depending on the dimension d and z and the κ_i are constants depending on x and $\mathbb{C}ov(\mathbf{x}, \mathbf{x}')$ only.

Now, returning to our original problem, if $d = 1$ and $\widehat{f_n}(\cdot)$ is *unbiased*, then

$$\mathsf{G}_n(\mathbf{x}) = \left\langle \frac{\psi^\star(\mathbf{x})}{\|\psi^\star(\mathbf{x})\|_{\mathbb{R}^n}}, \boldsymbol{\varepsilon} \right\rangle_{\mathbb{R}^n} \triangleq \left\langle \overline{\boldsymbol{\psi}}(\mathbf{x}), \boldsymbol{\varepsilon} \right\rangle_{\mathbb{R}^n},$$

and the critical value κ in the confidence bands (3.4) is such that

$$\alpha \approx \frac{\kappa_0}{\pi}\left(1 + \frac{\kappa^2}{n-p}\right)^{-\frac{n-p}{2}} + \mathcal{E} \cdot \mathbb{P}\left(\left|\mathsf{t}_{n-p}\right| > \kappa\right), \tag{3.7}$$

where the first term is the result of an appropriate integration of Equation 3.5 and an application of the symmetry of $\mathbb{G}(\cdot)$; the second term is from a boundary correction to (3.5) and involves the right tail of a t distribution with $(n-p)$ degrees of freedom, plus the *Euler–Poincaré* characteristic $\mathcal{E}$ of $\{\overline{\boldsymbol{\psi}}(\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$ (see [93] for further details). $\mathcal{E}$ is equal to 1 if $\mathcal{X} = [a, b]$ and $\overline{\boldsymbol{\psi}}(a) \neq \overline{\boldsymbol{\psi}}(b)$, whereas $\mathcal{E} \equiv 0$ if $\overline{\boldsymbol{\psi}}(a) \equiv \overline{\boldsymbol{\psi}}(b)$. In addition, in the case $d = 1$, the following simple computational formula for κ_0 is available:

$$\kappa_0 \approx \int_{\mathcal{X}} \left\| \frac{d}{dx}\overline{\boldsymbol{\psi}}(x) \right\|_{\mathbb{R}^n} dx \approx \sum\nolimits_i \left\| \overline{\boldsymbol{\psi}}(x_i) - \overline{\boldsymbol{\psi}}(x_{i-1}) \right\|_{\mathbb{R}^n},$$

which, geometrically, is the approximate length of the curve $\left\{\overline{\boldsymbol{\psi}}(x)\right\}_{x\in\mathcal{X}}$. Sufficient conditions on the tubular area around $\overline{\boldsymbol{\psi}}(\mathbf{x})$ under which the approximation in (3.7) works well are given in [93]. If these assumptions are violated as, for example, when the domain of interest consists of a few points close to each other, then the approximation symbol "$\approx$" in Equation 3.7 should be replaced by "$\leqslant$".

- **Robins–van der Vaart split–sample method [86]:** This approach makes explicit the basic idea that has driven much of the recent research in nCS's: if I want to build an L^p ball around a "good" estimator of my choice, all I have to do is to find a way to explore the behavior of its L^p loss (not the risk!). More specifically, taking $p = 2$ and identifying "good" with adaptive estimator, we "just" need to control a suitable estimator of a quadratic functional opportunely centered, hoping that, to some extend, the good properties of the pointwise estimator carry over the derived confidence set.

- **Subspace Pre–Testing [3]:** Consider again the nonparametric regression model

$$\mathbf{Y}_n = \mathbf{f}_n + \sigma\,\boldsymbol{\epsilon}_n, \quad \text{where} \quad \mathbf{f}_n = [f(x_1), \ldots, f(x_n)]^{\mathrm{T}}.$$

Baraud procedure constructs L^2–finite–sample confidence balls for $\mathbf{f}_n$ by a sequence of suitable tests. Let $\mathcal{S} = \{\mathsf{S}_k\}_{k\in\mathsf{K}}$ be a collection of subspaces of $\mathbb{R}^n$ and assume $\mathbb{R}^n \in \mathcal{S}$. For each $\mathsf{S}_k \in \mathcal{S}$, let $\widehat{\mathbf{f}}_n = \Pi_{\mathsf{S}_k}\mathbf{Y}_n$ with Π_{S_k} being the orthogonal projection onto S_k, and note that the basic building block of an L^2–ball can be written as

$$\left\|\mathbf{f}_n - \widehat{\mathbf{f}}_n\right\|_2^2 = \left\|(\mathbb{I} - \Pi_{\mathsf{S}_k})\mathbf{f}_n\right\|_2^2 + \sigma^2\left\|\Pi_{\mathsf{S}_k}\boldsymbol{\epsilon}_n\right\|_2^2.$$

Baraud uses $\mathsf{T}_{\mathsf{S}_k} = (\mathbb{I} - \Pi_{\mathsf{S}_k})\mathbf{Y}_n$ to test

$$\mathsf{H}_0 : \mathbf{f}_n \in \mathsf{S}_k \quad \text{vs} \quad \mathsf{H}_1 : \mathbf{f}_n \notin \mathsf{S}_k,$$

and control $\|(\mathbb{I} - \Pi_{\mathsf{S}_k})\mathbf{f}_n\|_2^2$, for each $k \in \mathsf{K}$: it will be small with high probability when we do not reject the null.

- **Wahba's Average Coverage:** This "tangentially" Bayesian technique, was started by Grace Wahba in [97] and further studied in [27, 84, 101, 102]. See [51] and [98] also. Using the same notation adopted for the Sheffe's method, consider a generic linear smoother

$$\widehat{f}_{\lambda_{\mathsf{G}}}(\mathbf{x}_n) = \Psi^{\star}_n(\lambda_{\mathsf{G}})\mathbf{y}_n,$$

where $\mathbf{y}_n = [y_1, \ldots, y_n]^{\mathsf{T}}$, $\mathbf{x}_n = [x_1, \ldots, x_n]^{\mathsf{T}}$, $x_i \in \mathbb{R}$ and $\{\Psi^{\star}_n(\lambda_G)\}_{\lambda_{\mathsf{G}} \in \mathcal{A}}$ with $\mathcal{A} \subset \mathbb{R}_+$, is a family of smoothing matrices indexed by the *global* smoothing parameter λ_{G}. In this context, Wahba first suggested that α–level *pointwise* confidence intervals for the regression curve $f(\cdot)$ at x_i could be constructed using the smoothing spline estimator $\widehat{f}_{\widehat{\lambda}_{\mathsf{G}}}(\cdot)$ in the form

$$\mathfrak{B}_{\widehat{\lambda}_{\mathsf{G}}}(x_i) = \left[\widehat{f}_{\widehat{\lambda}_{\mathsf{G}}}(x_i) - z_{\alpha/2}\sqrt{\widehat{\sigma}^2\left[\Psi^{\star}_n\left(\widehat{\lambda}_{\mathsf{G}}\right)\right]_{(i,i)}}, \widehat{f}_{\widehat{\lambda}_{\mathsf{G}}}(x_i) + z_{\alpha/2}\sqrt{\widehat{\sigma}^2\left[\Psi^{\star}_n\left(\widehat{\lambda}_{\mathsf{G}}\right)\right]_{(i,i)}}\right] \tag{3.8}$$

where $\widehat{\lambda}_{\mathsf{G}}$ is the value of λ_{G} that minimizes (λ_{G}), $\left[\Psi^{\star}_n\left(\widehat{\lambda}_{\mathsf{G}}\right)\right]_{(i,i)}$ is the i–th diagonal element of the smoothing matrix[3], and $\widehat{\sigma}^2$ is the estimator of σ^2 from the smoothing spline estimator given by $\mathtt{RSS}/\left\{n - \mathsf{tr}[\Psi^{\star}_n(\lambda_{\mathsf{G}})]\right\}$ with $\mathtt{RSS}$ equal to the residual sum of squares. Through simulation results, Wahba found that the *average coverage* probability, $\frac{1}{n}\sum_i \mathbb{P}\left[f(x_i) \in \mathfrak{B}_{\widehat{\lambda}_{\mathsf{G}}}(x_i)\right]$, is close to $1 - \alpha$. This result led to the conjecture by Wahba that $\frac{1}{n}\widehat{\sigma}^2\mathsf{tr}\left[\Psi^{\star}_n(\lambda_{\mathsf{G}})\right]$ was related to the *expected average squared error*

$$\mathsf{EASE}(\lambda_{\mathsf{G}}) = \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[f(x_i) - \widehat{f}_{\lambda_{\mathsf{G}}}(x_i)\right]^2\right\}.$$

Nychka [84] proved this conjecture of Wahba's to be true and a more detailed discussion of these pointwise confidence intervals is given therein. Finally in [27] this same strategy for confidence intervals based on a global smoothing parameter λ_{G} is applied to estimates where the smoothing parameter is actually a vector valued function $\boldsymbol{\lambda}_{\mathsf{L}} = [\lambda_1, \ldots, \lambda_n]^{\mathsf{T}}$ that adapts to the local curvature of the function. Accordingly, the pointwise intervals have the same form as in Equation 3.8 except the estimate and standard error are evaluated at the local smoothing parameter $\widehat{\lambda}_i$ rather than $\widehat{\lambda}_{\mathsf{G}}$. If the quantity $\widehat{\sigma}^2\left[\Psi^{\star}_n\left(\widehat{\lambda}_i\right)\right]_{(i,i)}$ is close to the pointwise expected squared error, then one would expect the confidence interval to hold its level uniformly at all points. And in fact the proposed new method called was shown to be superior to the standard use of generalized cross-validation for constructing confidence intervals.

- **Asymptotic and Bootstrap–based Pivot–balls [5, 8]:** Suppose we wish to construct a confidence set for the parametric function $\tau(\theta)$. Classical theory advises to find a *pivot*; that is, a function of the sample $\mathbf{Z}_n$ and of $\tau(\cdot)$ whose distribution under the model does not depend on the unknown parameter θ.

[3] As an aside, the term $\widehat{\sigma}^2\left[\Psi^{\star}_n\left(\widehat{\lambda}_{\mathsf{G}}\right)\right]_{(i,i)}$ results from some nice simplifications for splines in terms of expressing the posterior variance of the estimate as a simple function of the smoothing matrix A. For other smoothers what should be included here is the posterior variance for the estimator

Though important, the exact pivotal technique is rarely available, in particular when dealing with nonparametric model (it already fails to generate confidence intervals for the difference of two normal means in the *Behrens–Fisher* problem). To generalize the pivotal method we can follow two different paths: one rooted in resampling methods, and the other in asymptotic approximations of some kind. In both cases, we start with a function $R_n(\mathbf{Z}_n, \tau)$ with distribution $\mathcal{H}_n(\theta)$. R_n does not need to be a pivot, but it plays an analogous role. Beran calls R_n a *root*. Assume now that by bootstrap or some asymptotic argument, an approximation $\widehat{\mathcal{H}}_n(\theta)$ to $\mathcal{H}_n(\theta)$ is available. At this point we are in position to build confidence sets by analogy with classical pivot–method.

Let $\widehat{\mathcal{H}}_n^{-1}(\alpha)$ denote the α–quantile of our approximation to $\mathcal{H}_n$ and let $\mathcal{T}$ denote the support of $\tau(\cdot)$. Define the approximate confidence set for $\tau(\cdot)$ to be

$$\mathfrak{B}_n = \left\{ t \in \mathcal{T} : R_n(\mathbf{Z}_n, t) \leqslant \widehat{\mathcal{H}}_n^{-1}(\alpha) \right\}.$$

Under suitable conditions on $\widehat{\mathcal{H}}_n$ and on the limit $\mathcal{H}(\cdot)$ of the sequence $\{\mathcal{H}_n\}_n$, the coverage of $\mathfrak{B}_n$ converges to α.

For the basic Normal mean problem the function τ of interest is the signal $\boldsymbol{\theta}_n$ itself. In this setting, Beran (1995), based on Stein (1981), considers L^2–balls centered on the James–Stein estimator $\widehat{\boldsymbol{\theta}}_{\mathsf{JS}}$. This leads *naturally* to a *root* that compares the L^2–loss $\mathsf{L}_n\left(\widehat{\boldsymbol{\theta}}_{\mathsf{JS}}, \boldsymbol{\theta}_n\right)$ of $\widehat{\boldsymbol{\theta}}_{\mathsf{JS}}$ with an unbiased estimator of its *risk* $\widehat{\mathsf{R}}_n\left(\widehat{\boldsymbol{\theta}}_{\mathsf{JS}}\right)$:

$$R_n(\mathbf{Z}_n, \boldsymbol{\theta}_n) = \tfrac{1}{\sqrt{n}}\left[\mathsf{L}_n - \widehat{\mathsf{R}}_n\right],$$

Later, Beran and Dumbgen (1998) extended this approach to confidence sets centered on a larger class of estimators that includes the so called *modulators*:

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = [\lambda_1 \tilde{\theta}_1, \ldots, \lambda_n \tilde{\theta}_n]^{\mathsf{T}},$$

where $1 \geqslant \lambda_1 \geqslant \cdots \geqslant \lambda_n \geqslant 0$, and $\tilde{\boldsymbol{\theta}}_n$ is the maximum likelihood estimator.

4 Proposed Work.

Act I: T W T "F " C B

Wavelet bases are ubiquitous in modern nonparametric statistics starting from the 1994 seminal paper by Donoho and Johnstone [35]. What makes them so appealing to statisticians is their ability to capture the relevant features of smooth signals in a few "big" coefficients at high scales (low frequencies) so that zero thresholding the small ones, results in an effective denoising scheme (see [96]).

In a variety of real–life signals, significant wavelet coefficients often occur in clusters at adjacent scales and locations. Irregularities, like a discontinuity for example, in general tend

to affect the whole block of coefficients corresponding to wavelet functions whose "support" contains them. For this reason it is reasonable to expect that the risk of "blocked" thresholding rules might compare quite favorably with other classical estimators based on level–wise or global thresholds. The literature is filled with successful examples of "horizontally" (within scales) blocked rules derived from both, purely frequentist arguments [13, 19, 52], or Bayesian reasonings of some flavor [1, 21, 100]. Recently, an increasing amount of work has been devoted to study a new class of "vertically" (across scales) blocked or *treed* rules [2, 12, 23, 40, 87], that have proved to be of invaluable help in at least two settings of great importance: the construction of *adaptive* pointwise confidence intervals [85] and the derivation of pointwise estimators of a regression function that adapt rate optimally under what I will call a *focused* performance measure that bridges the gap between pointwise and global risk cropping the latter outside some neighborhood around a target point of interest [18].

The main goal of this section is to explore the possibility of using *treed* or other type of doubly–blocked rules to support the construction of *focused* confidence bands and balls; that is, of possibly adaptive confidence sets with guaranteed coverage only in a reasonably sized neighborhood of a target–point. More specifically, my main interest here will be in obtaining what can be called *maximal* focused bands; that is, adaptive α–level focused bands over the *maximally* allowed neighborhood of a target–point. In fact, although we have a whole catalog of adaptive function estimators considered also practically effective in recovering a function observed in a *low* noise environment, when we come to *global* nonparametric confidence sets, adaptivity is extremely limited (L^p balls with $p < +\infty$, see [15, 76]) or just impossible (uniform bands or L^∞–balls, see [79]), no matter where we center our sets: here adaptive methods do not perform significantly better than others as, for example, fixed bandwidth local polynomial smoothers. But, looking at the way Picard and Tribuley build adaptive *pointwise* confidence interval in their recent paper [85], it seems that some room is left out to rescue the credibility of particular classes of pointwise adaptive estimators in driving the construction of localized (*focused*) balls and bands.

Act II: B T S P –T

In a recent work (see [3] and Section 3 also), Yannick Baraud constructs finite–sample confidence ball for an unknown regression function $f(\cdot)$ observed in Gaussian noise, based on the following sequential testing procedure (pre–testing) inspired by Lepski [75]: given a family of *linear* subspaces $\{\mathcal{S}\}_{\alpha\in\mathcal{A}}$, for each α he tests $f \in \mathcal{S}_\alpha$ and, in case the test does not reject, he builds an α–level L^2–ball with minimal radius centered on the L^2–optimal estimator (under the constraint that $f \in \mathcal{S}_\alpha$). The output of this algorithm is the unique confidence ball having minimal *random* radius. He then goes on comparing his technique with the one based on the James–Stein estimator [76, 90].

Based on this, my goal in this part of the work will be two-fold: on one hand I will show how to employ George's multiple shrinkage setting [43] to inject flexibility into the Li–Stein's procedure so to achieve the performance of Baraud's confidence sets; on the other, I will try to improve upon Baraud's method itself by relaxing the condition on the linearity of the subspaces, via a sequential *treed* testing, reminiscence of the "treed" thresholding rule described in the previous Act. In both cases I will consider two different options to

calibrate the confidence radius: one based on asymptotic arguments and the other on an appropriate bootstrap technique. Possible ways to attack the problem from a finite–sample point of view will also be described.

Act III: A N M P P E

While frequentist methods for nonparametric estimation are flourishing, nonparametric Bayesian estimation methods have a relatively shorter history (see [99] for a nice review). Besides philosophical reasons, there are some practical advantages on using the Bayesian approach: on the one hand it allows to reflect ones prior beliefs into the analysis, on the other, it is, at least in principle, straightforward to apply since inference is based on the posterior distribution only, although computation of this distribution is an important and potentially formidable issue. In fact, due to the lack of useful analytical expressions for the posterior in most curve estimation problems, computation has to be done by some numerical technique, usually by the help of suitable Markov chain Monte-Carlo methods.

From a theoretical point of view, questions relative to consistency and convergence rates of Bayesian procedures have been explored in length [22, 49, 47, 89], whereas the study of adaptivity issue is still in its infancy [4, 46, 54, 103]. Moving to another topic of major concern, the catalog of priors over infinite dimensional spaces available to conduct a non-parametric analysis is growing constantly and a nice review is contained in a recent paper by Choudhuri, Ghosal and Roy [44], but still, it seems that there is a lack in the literature relative to the construction/discovering of what in the "classical" parametric setting has be called *probability matching prior*; that is, of a prior distribution under which specific α–level credible sets are also α–level confidence intervals up to some order of approximation (see [28] for further information). Motivated by such a simple consideration, and by the growing complexity of Bayesian nonparametric models, in this part I will try to (partially) fill in the gap tackling the matching problem from a purely algorithmic point of view: I will consider the "non–informative" prior distributions introduced in [45] and, rewriting the matching conditions as an optimization problem, I will see to what extend the perturbed–ellipsoids methods due to Ghosh and Mukerjee [48] can be adapted to the nonparametric case.

Act IV: REACT –G N

Building upon the already mentioned 1981 Stein's seminal paper [90], in 1998 Rudy Beran and Lutz Dümbgen introduced what later has been called *pivot*–ball in sequence space, to assess the uncertainty around a nonparametric function estimator based on orthogonal expansion. Since then, the basic framework has been extended in several directions [41, 58], but none of them has focused on broadening the noise class this procedure applies to, and this is precisely what I will pursue in this last part of my work. More specifically, I will extend the asymptotic pivoting argument to the case of error distributions belonging to the (natural) exponential family with quadratic variance function (QVEF in the following) introduced by Carl Morris in the early '80 [82, 83] and covering the Gaussian, Poisson, gamma, binomial, negative binomial and generalised hyperbolic secant distributions.

■

Appendix

A Act I: Treed Wavelet Thresholding and "Focused" Bands.

[Background]

A.1 Wavelets, function spaces and smoothing: an overview.

In this section, I give an overview on *multiresolution* analysis, wavelet series and wavelet estimators in the classical setting . By "classical" or "first-generation" wavelets, I mean wavelets whose construction is deeply rooted in Mallat's pyramidal algorithm and were initially designed to analyze signals observed at equispaced design points and with a sample size which is a power of two. This class of wavelets have to be contrasted with the"second-generation" wavelets basis introduced in [95] and recently applied to nonparametric regression with random design [31, 32].

If one wants to analyze a function of time with a series expansion, the first idea that comes probably into one's mind is to use a Fourier series, i.e. decompose the function into sine and cosine at different frequencies. In this process, we hope that only a few coefficients in the series will carry most of the information about the signal. Certain smooth functions admit such an "economical" Fourier expansion. However, for a large range of functions, a good Fourier series approximation requires numerous sine and cosine basis functions. Indeed, the sine functions have a precise frequency but are not localized in time, hence a localized information in the signal like a discontinuity will a.ect all the coe.cients of the series. This drawback lead to look for more efficient bases, that is, bases which are localized both in time and in frequency. We will see here that a wavelet basis offers exactly this property.

A.1.1 Multiresolution analysis.

A natural way to introduce wavelets is through the multiresolution analysis. Given a function $f \in \mathsf{L}^2(\mathbb{R})$, a multiresolution of $\mathsf{L}^2(\mathbb{R})$ will provide a sequence of spaces $\{V_k\}_{j\in\mathbb{Z}}$, such that the projections of $f(\cdot)$ onto these spaces give finer and finer approximations (as $j \to +\infty$) of the function $f(\cdot)$.

Def. A.1 *A multiresolution of $L^2(\mathbb{R})$ is defined as a sequence of closed subspaces $V_j \subset L^2(\mathbb{R})$, with $j \in \mathbb{Z}$, that verifies the following propertis*

1. $\ldots \subset V_{-1} \subset V_0 \subset V_1 \subset \ldots$
2. $\bigcup_{j\in\mathbb{Z}} V_j$ *is dense in* $L^2(\mathbb{R})$ *and* $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$
3. *If $f(x) \in V_0$, then $f(2^j x) \in V_j$, i.e. the spaces V_j are scaled versions of the central space V_0.*
4. *If $f(x) \in V_0$, then $f(x-k) \in V_0$, $\forall\, k \in \mathbb{Z}$, i.e., V_0 (and hence all the V_j) is invariant under translation.*
5. *The exists $\phi(x) \in V_0$ such that $\{\phi(x-k)\}_{k\in\mathbb{Z}}$ is a Riesz basis in V_0.*

I will call "level" of a one of the subspaces V_j. From Definition A.1, it follows that, for fixed j, the set $\{\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)\}_{k\in\mathbb{Z}}$ of scaled and translated version of $\phi(\cdot)$ is a Riesz basis for V_j. Since $\phi \in V_0 \subset V_1$, we can express $\phi(\cdot)$ as a linear combination of $\{\phi_{1,k}\}_{k\in\mathbb{Z}}$:

$$\phi(x) = \sum_{k\in Z} h_k \phi_{1,k}(x) = \sqrt{2}\sum_{k\in Z} h_k \phi(2x-k). \tag{A.1}$$

Equation A.1 is called the *two–scale equation* or *refinement equation*. It is a fundamental equation in since it tells us how to go from a *fine level* V_1 to a *coarser* level V_0. The function $\phi(\cdot)$ is usually called *father wavelet* or *scaling function*.

As said before, the spaces V_j will be used to approximate general functions. This will be done by defining appropriate projections onto these spaces. Since the union of all the V_j is dense in $\mathsf{L}^2(\mathbb{R})$, we are guaranteed that any given function of $\mathsf{L}^2(\mathbb{R})$ can be approximated arbitrary well by such projections. As an example, define the space V_j as

$$V_j = \left\{ f \in \mathsf{L}^2(\mathbb{R}) \,\middle|\, \forall k \in \mathbb{Z},\ f\big|_{[2^{-j}k, 2^{-j}(k+1)]} \text{ is constant} \right\}. \tag{A.2}$$

Then the scaling function $\phi(x) = \mathbb{1}_{[0,1)}(x)$, called *Haar* scaling function, generates by translation and dilations a for the sequence of spaces $\{V_j\}_{j\in\mathbb{Z}}$ defined in Equation A.2, see [30].

A.1.2 The detail space and the wavelet function.

Rather than considering all the nested spaces V_j, it would be more efficient to code only the information needed to go from V_j to V_{j+1}. Hence consider the space W_j which complements V_j in V_{j+1}:

$$V_{j+1} = V_j \oplus W_j, \tag{A.3}$$

The space W_j is not necessarily orthogonal to V_j, but it always contains the *detail* information needed to go from an approximation at resolution j to an approximation at resolution $j+1$. Consequently, by using recursively the Equation A.3, we have for any fixed $j_0 \in \mathbb{Z}$, the decomposition

$$\mathsf{L}^2(\mathbb{R}) = \overline{V_{j_0} \oplus \bigoplus_{j=j_0}^{\infty} W_j}.$$

With the notational convention that $W_{j_0-1} \triangleq V_{j_0}$, the sequence $\{W_j\}_{j\geqslant j_0-1}$ is usually called *multiscale decomposition* ( ), and a function $\psi(\cdot)$ is called *mother wavelet* or simply *wavelet* whenever the set $\{\psi(x-k)\}_{k\in\mathbb{Z}}$ is a Riesz basis of W_0. Since $W_0 \subset V_1$, there also exist a refinement equation for $\psi(\cdot)$ similar to A.1:

$$\psi(x) = \sqrt{2} \sum_{k\in\mathbb{Z}} g_k \phi(2x-k). \tag{A.4}$$

The collection of wavelet functions $\{\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)\}_{(j,k)\in\mathbb{Z}^2}$ is then a Riesz basis for $Lp^2(\mathbb{R})$. One of the main features of the wavelet functions is that they possess a certain number of vanishing moments.

Def. A.2 *A wavelet function $\psi(\cdot)$ has N vanishing moments if*

$$\int x^p \psi(x) \mathrm{d}x = 0, \quad \forall\, p \in \{0, \ldots, N-1\}.$$

The following are three interesting cases of wavelet bases.

- **Orthogonal bases on $\mathbb{R}$:** In an *orthogonal* , the spaces W_j are defined as the orthogonal complement of V_j in V_{j+1}. The following theorem tells us one of the main advantages of such a .

  Theorem A.1 ([30], Theorem 5.1.1). *If a sequence of closed subspaces $\{V_j\}_{j\in\mathbb{Z}}$ in $\mathsf{L}^2(\mathbb{R})$ satisfies Definition A.1, and if, in addition, $\{\phi(x-k)\}_{k\in\mathbb{Z}}$ is an orthogonal basis for V_0, then there exists one function $\psi(\cdot)$ such that $\{\psi(x-k)\}_{k\in\mathbb{Z}}$ forms an orthogonal basis for the orthogonal complement W_0 of V_0 in V_1.*

  An immediate consequence of Theorem A.1 is that $\{\psi_{j,k}\}_{k\in\mathbb{Z}}$ constitutes an orthogonal basis for the orthogonal complement W_j of V_j in V_{j+1}. In this section, let $\mathcal{P}_j$ and $\mathcal{Q}_j$ be respectively the *orthogonal* projection operator onto V_j and onto W_j. The orthogonal expansion

$$f(x) = \mathcal{P}_{j_0}(f)(x) + \sum_{j=j_0}^{\infty} \mathcal{Q}_j(f)(x) = \sum_{k\in\mathbb{Z}} \left\langle f, \phi_{j_0,k} \right\rangle \phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty} \sum_{k\in\mathbb{Z}} \left\langle f, \psi_{j,k} \right\rangle \psi_{j,k}(x),$$

says that a first, coarse approximation of $f(\cdot)$ in V_{j_0} is further refined with the projection of $f(\cdot)$ onto the detail spaces W_j. The simplest orthogonal wavelet system is the *Haar* system associated to the scaling function defined in Section A.1.1. The Haar wavelet given by

$$\psi(x) = 2^{-1/2}\Big[\phi(2x-1) - \phi(2x)\Big] = \mathbb{1}_{\left[\frac{1}{2},1\right)}(x) - \mathbb{1}_{\left[0,\frac{1}{2}\right)}(x),$$

has only one vanishing moment and consequently is optimal only to represent functions having a low degree of regularity, like, for example, β–Hölder function with $\beta \in (0,1)$. Daubechies constructed in [29, 30] compactly supported wavelets which have more than one vanishing moment. Compactly supported wavelets are desirable from a numerical point of view, while having more than one vanishing moment they allow to reconstruct exactly polynomial of higher order. These wavelets cannot, in general, be written in a closed analytic form. However, their graph can be computed with arbitrary high precision using a subdivision scheme algorithm. In addition, compactly supported wavelets cannot, in general, be symmetric. The best we can do is to build what Daubechies called *least asymmetric wavelets* or *Symmlets*. See [30, 53, 96] for other orthogonal systems.

- **Orthogonal bases on [0,1]:** Different procedures are available to build orthogonal wavelets on an interval. Among these, the most popular is probably the one synthesized in [25] by transforming the *boundary* wavelets, i.e. those whose supports overlap $x = 0$ or $x = 1$, into functions having their support stricly contained in $[0,1]$ so to provide the necessary complement to generate a basis for $\mathsf{L}^2([0,1])$. If the mother wavelet has a compact support then there is only a constant number of boundary wavelets at each scale to be modified.

  The main difficulty in implementing these schemes, is to construct boundary–wavelets that keep their vanishing moments. In fact, of the three main procedures available, the one mentioned here is slightly more complicated to construct, but it is also the only capable to produce bases having as many vanishing moments as the original *inside*–wavelets. See [80] for further details and alternative constructions.

  Another construction particularly interesting for its statistical implications, is the one recently introduced by Silverman and Johnstone [63]. More specifically, they use their boundary–modified *coiflets* basis to show that the discrete wavelet transform of finite data from sampled regression models asymptotically provides a close approximation to the wavelet transform of the continuous Gaussian white noise model so that, as a matter of fact, estimating errors in the practical discrete setting need not be larger than those expected in the continuous statistical setting.

- **Biorthogonal bases:** Having an orthogonal puts strong constraints on the construction of a wavelet basis. For example, the Haar wavelet is the only real–valued function which is compactly supported and symmetric. However, if we relax orthogonality for *biorthogonality*, then it becomes possible to have real-valued wavelet bases of fixed but arbitrary *high order* (see Definition A.3) which are symmetric and compactly supported [24].

  In a biorthogonal setting, a *dual* scaling function $\widetilde{\phi}(\cdot)$ and a dual wavelet function $\widetilde{\psi}(\cdot)$ exist. They generate a *dual* with subspaces $\widetilde{V}_j$ and complement spaces $\widetilde{W}_j$ such that

$$\widetilde{V}_j \perp W_j \quad \text{and} \quad V_j \perp \widetilde{W}_j.$$

  In other words,

$$\left\langle \widetilde{\phi}(\cdot), \psi(\cdot - k) \right\rangle = 0 \quad \text{and} \quad \left\langle \phi(\cdot), \widetilde{\psi}(\cdot - k) \right\rangle = 0.$$

  Moreover, the dual functions also have to satisfy

$$\left\langle \widetilde{\phi}(\cdot), \phi(\cdot - k) \right\rangle = \delta_{k,0} \quad \text{and} \quad \left\langle \widetilde{\psi}(\cdot), \psi(\cdot - k) \right\rangle = \delta_{k,0},$$

  where $\delta_{x,y}$ is the *Kronecker simbol*. By construction, the dual scaling and wavelet functions satisfy a refinement equation pretty similar to Equations A.1 and A.4. In this work, we use the following convention: the *dual* will be used to decompose a function (or a signal), while the original, or *primal* reconstructs the function. This yields the following representation of a function $f \in \mathsf{L}^2(\mathbb{R})$

$$f(x) = \sum_{k \in Z} \left\langle f, \widetilde{\phi}_{j_0,k} \right\rangle \phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty} \sum_{k \in Z} \left\langle f, \widetilde{\psi}_{j,k} \right\rangle \psi_{j,k}(x). \tag{A.5}$$

A.1.3 Wavelets as unconditional bases.

The most important basis in analysis has certainly been the trigonometric system. This is so because the resolution of several key problems in physics is particularly simple when formulated in this setting. Unfortunately, the convergence of the corresponding series posed important mathematical problems since Du Bois-Reymond showed in 1873 that the Fourier series of a continuous function may diverge (see [66] where the fine properties of Fourier series and the development of ideas that led to wavelet analysis are described). Is this phenomenon inherent to any orthogonal decomposition? Hilbert posed this problem to his student Alfred Haar, who gave a negative answer in his thesis by constructing in 1909 the wavelet basis named after him (see Section A.1.1). Haar showed that the partial sums of the decomposition of a continuous function in this basis are uniformly convergent. The comparison with the trigonometric system is striking: a basis composed of discontinuous functions is more adapted to the analysis and reconstruction of continuous functions than the trigonometric system, though this system is composed of $\mathscr{C}^\infty$ functions. The Haar basis has another important property which the trigonometric system lacks: Marcinkiewicz showed in 1937 that it is an unconditional basis for the spaces L^p when $1 < p < \infty$; this means that any function of L^p can be written in only one way as $\sum \theta_{j,k}\psi_{j,k}(\cdot)$ and the convergence is unconditional, i.e. does not depend on the order of summation. This result still has important implications in current research.

Of course, since the Haar basis is not composed of continuous functions, it cannot be a basis for spaces of continuous functions. This last remark motivated the search for "smooth" analogs of the Haar basis. The goal was to construct bases of similar algorithmic type, and which would be unconditional for a wide range of function spaces. In 1910, Faber considered on $[0,1]$ the basis composed by 1, x and the primitives of the Haar basis. This *Schauder basis* (so-called because it was rediscovered by Schauder in 1927) has the same algorithmic form as the Haar basis with $\psi(\cdot)$ equal to the primitive of the Haar wavelet, and, as Faber himself proved, it is actually a basis for $\mathscr{C}^0([0,1])$. The price to be paid is that it is no longer a basis for $\mathsf{L}^2([0,1])$. Should one necessarily lose on one hand what has been obtained by the other? In 1928, Franklin showed that this is actually not the case. By applying the Gram-Schmidt orthonormalization procedure to the Schauder basis, he obtained a basis which is simultaneously unconditional for all $\mathsf{L}^p([0,1])$ spaces with $p \in (1,\infty)$, for $\mathscr{C}^0([0,1])$ and for the Sobolev spaces of low regularity. One can go on and iterate one step of integration (which regularizes) and one step of Gram-Schmidt orthonormalization. Doing so, in 1972 Ciesielski constructed bases which are unconditional for a wider and wider range of function spaces on $[0,1]$. Of course, applying the Gram-Schmidt orthonormalization procedure iteratively makes these bases essentially impossible to be computed numerically. Something has been lost along the way. However, algorithmic simplicity and regularity can go together. In 1981, Strömberg had the idea of applying the Gram-Schmidt orthonormalization on the whole line instead of $[0,1]$ only (loosely speaking, one starts the orthonormalization at $-\infty$). Because of the dilation and translation invariance of the real line, this substitute of the Shauder basis now has the exact algorithmic form. In addition, starting the orthonormalization with B-splines of arbitrary high degree, Strömberg was able to construct orthonormal wavelet bases of arbitrary regularity that are actually unconditional for a wide range of Sobolev and Besov spaces. The ultimate perfection was found in 1986 by Yves Meyer and Pierre-Gilles Lemarié, who constructed $\mathscr{C}^\infty$ wavelets $\left\{\psi^{(i)}(\cdot)\right\}_{i\in\{1,\dots,2^d-1\}}$ such that the functions

$$\psi^{(i)}_{j,\mathbf{k}}(\mathbf{x}) \triangleq 2^{\frac{dj}{2}}\psi^{(i)}\left(2^j\mathbf{x}-\mathbf{k}\right), \quad \forall\,(i,j,\mathbf{k}) \in \left\{1,\dots,2^d-1\right\}\times\mathbb{Z}\times\mathbb{Z}^d, \tag{A.6}$$

form an orthonormal basis for $\mathsf{L}^2(\mathbb{R}^d)$, thus

$$f(x) = \sum_{i,j,\mathbf{k}} \theta^{(i)}_{j,\mathbf{k}}\psi^{(i)}_{j,\mathbf{k}}(x), \tag{A.7}$$

where the wavelet coefficients

$$\theta^{(i)}_{j,\mathbf{k}} = \int_{\mathbb{R}^d} 2^{dj}\psi^{(i)}_{j,\mathbf{k}}(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x},$$

are normalized with respect to the L^∞ norm. An easy way to build such multidimensional systems is via the so called *tensor product* method: let $\phi^\circ(x)$ and $\psi^\circ(x)$ be the scale and corresponding wavelet functions in $\mathbb{R}^1$, then define

$$\phi(\mathbf{x}) = \prod_{r=1}^{d}\phi^\circ(x_r).$$

Let S denote the set of 2^d-1 vectors $\mathbf{s}=[s_1,\dots,s_d]^\mathsf{T}$ consisting of zeros and ones excluding the zero vector $[0,\dots,0]^\mathsf{T}$. Define the set of wavelet functions

$$\{\psi^{(i)}(\mathbf{x})\}_{i\in\{1,\dots,2^d-1\}}=\left\{\prod_{r=1}^{d}\psi_{s_r}(x_r)\right\}_{\mathbf{s}\in S}$$

with $\psi^{(0)}(x)=\phi^\circ(x)$ and $\psi^{(1)}(x)=\psi^\circ(x)$. This basis allows one to characterize functions of arbitrary regularity or, by duality, distributions – in the sense of Schwarz – of arbitrary irregularity (see [73]). To be more specific, let $p\in(1,+\infty)$ and $s\geqslant 0$, and define the *Sobolev* space $\mathcal{W}^{p,s}(\mathbb{R}^d)$ as the function space composed by the elements of $\mathsf{L}^p(\mathbb{R}^d)$ whose *fractional derivatives* of order s also belongs to $\mathsf{L}^p(\mathbb{R}^d)$. It can be proved that a function belongs to $\mathcal{W}^{p,s}(\mathbb{R}^d)$ if and only if its wavelet coefficients satisfy the following condition (see [81])

$$f\in\mathcal{W}^{p,s}(\mathbb{R}^d)\Leftrightarrow\left(\sum_{i,j,\mathbf{k}}\left|\theta^{(i)}_{j,\mathbf{k}}\right|^2\left(1+2^{2sj}\right)\mathbb{1}_{\mathcal{k}}(\mathbf{x})\right)^{\frac{1}{2}}\in\mathsf{L}^p. \tag{A.8}$$

where $\mathcal{k}=\mathcal{k}_{j,\mathbf{k}}=\frac{\mathbf{k}}{2^j}+\left[0,\frac{1}{2^j}\right]^d$ denotes the generic dyadic cube of $\mathbb{R}^d$ centered at $\frac{\mathbf{k}}{2^j}$. The contrast with classical Fourier expansions is evident: when $p\neq 2$, there is no characterization of $\mathcal{W}^{p,s}$ by condition on the moduli of the Fourier coefficients.

This characterizations are quite difficult to handle and, in the context of wavelet analysis, Besov spaces are much more commonly in use. The (inhomogeneous) Besov spaces on the unit intervals, $\mathcal{B}^{s,p}_q([0,1])$, consist once again of functions that have a specific degree of smoothness in their derivatives. The parameter r1 can be viewed as a degree of functions inhomogeneity while s is a measure of its smoothness. Roughly speaking, the (not necessarily integer) parameter s indicates the number of functions derivatives, where their existence is required in an Lr1-sense; the

The success of for the two following reasons:

- They are very close to Sobolev spaces, as shown by the following embeddings

$$\forall\,\epsilon>0,\quad\forall\,p\geqslant 1,\quad\forall\,q\in\mathbb{R}\quad\Rightarrow\quad\mathcal{W}^{p,s+\epsilon}\hookrightarrow\mathcal{B}^{s,p}_q\hookrightarrow\mathcal{W}^{p,s-\epsilon}$$

- They have a very simple wavelet characterization (see [81]),

$$f\in\mathcal{B}^{p,s}_q(\mathbb{R}^d)\Leftrightarrow\left(\sum_{i,\mathbf{k}}\left|\theta^{(i)}_{j,\mathbf{k}}2^{\left(s-\frac{d}{p}\right)j}\right|^p\right)^{\frac{1}{p}}\triangleq\eta_j,\quad\text{and}\quad\{\eta_j\}_{j\in\mathbb{Z}}\in\ell^q(\mathbb{R}). \tag{A.9}$$

Note that in all such characterizations, wavelets are assumed to be smooth enough, say, with at least derivatives up to order $[s]+1$ and having fast decay (see [9] for optimal regularity assumptions on the wavelets). In sharp contrast with the Sobolev case, Besov spaces are defined for any $p>0$.

One of the reasons for the success of wavelet decompositions in applications is that they often lead to very sparse representations of signals. This sparsity can be characterized by determining to which Besov spaces $\mathcal{B}^{p,s}_q$ the function considered belongs when p is close to 0. Let us illustrate this assertion by an example. Consider the function $h(x)=\mathbb{1}_{[-1,1]}(x)$ and suppose that the wavelet used is compactly supported on $[-a,a]$. For each j, there are less than $4a$ non–zero wavelet coefficients, so that the wavelet expansion of a function $f(\cdot)$ is extremely sparse. Now, since $h(x)$ is bounded, $\theta_{j,k}\leqslant C$, $\forall\ (j,k)$ thus, using Equation A.9 with $d=1$, we conclude that $h(x)$ actually belongs to $\mathcal{B}^{p,s}_q(\mathbb{R})$ as soon as $\left(s-\frac{1}{p}\right)<0$. Let us check that, conversely, this property is a way to express that the wavelet expansion of a function is sparse. Suppose that a bounded function $f(\cdot)$ satisfies

$$\forall\,p,q>0,\quad\forall\,s<\tfrac{1}{p},\quad f\in\mathcal{B}^{p,s}_q.$$

We claim that $\forall\,a>0$, $\forall\,\epsilon>0$, at each scale j there are less than $C_{\epsilon,a}2^{\epsilon j}$ coefficients of size larger than 2^{-aj}. Indeed, if it was not the case, taking $p=\frac{\epsilon}{2a}$, we get $\sum_k|\theta_{j,k}|^p\to+\infty$ when $j\to+\infty$, hence a contradiction. Here is another illustration of the relationship between sparsity of the wavelet expansion and Besov regularity. Suppose that $f(\cdot)$ belongs to

$$\bigcap_{p>0}\mathcal{B}_p^{d/p,p}(\mathbb{R}^d),$$

Going back to Equation A.9, this condition exactly means that the sequence belongs to ℓ^p for all $p > 0$, which is also equivalent to the fact that the decreasing rearrangement of the sequence has fast decay, which, again, is a way to express sparsity, see [57]. Besov spaces when $p < 1$ are no longer locally convex, which partly explains the difficulties met when using them. Before the introduction of wavelets, these spaces were either characterized by the order of approximation of $f(\cdot)$ by rational functions whose numerator and denominator have a given degree, or equivalently by the order of approximation by splines with *free knots*, see [33] and [57]. However such characterizations were much more difficult to handle, and of difficult use in numerical applications.

A.1.4 Approximation of Functions.

Lets start with a definition of the order of a multiresolution analysis.

Def. A.3 *A multiresolution analysis is said to be of order $N^\star$ if the primal scaling function $\phi(\cdot)$ reproduces polynomials up to degree $N^\star - 1$, i.e.,*

$$\forall p \in 0, \dots, N^\star, \exists C_k \in \mathbb{R} : x^p = \sum_k C_k \phi(x-k).$$

The associated dual wavelet $\widetilde{\psi}(\cdot)$ has then $N^\star$ vanishing moments. It can be proved that the order of a and the regularity of the scaling function are linked: the larger $N^\star$, the higher the regularity of $\phi(\cdot)$. Symmetrically to Definition A.3, the order of the dual is N if $\widetilde{\phi}(\cdot)$ can reproduce polynomials up to degree $N-1$.

The main goal when decomposing a function in a wavelet series is to create sparse representation of the function, that is, to obtain a decomposition where only a few number of detail coefficients are "large" in absolute value, while the majority of the coefficients are close to zero. Near a singularity, large detail coefficients at different levels will be needed to recover the discontinuity. However, between points of singularity, we can hope to have small detail coefficients, in particular if the analyzing wavelet $\widetilde{\psi}_{j,k}$ have a large number of vanishing moments. Indeed, suppose the function $f(\cdot)$ to be decomposed is analytic on the interval $\Im$ without discontinuity. Since $\left\langle x^p, \widetilde{\psi}_{j,k} \right\rangle_{L^2} = 0$ for $p \in \{0, \dots, N^\star - 1\}$, we are sure that the first $N^\star$ terms of a Taylor expansion of $f(\cdot)$ will not give a contribution to the wavelet coefficient $\left\langle f, \widetilde{\psi}_{j,k} \right\rangle_{L^2}$ provided that the support of $\widetilde{\psi}_{j,k}$ does not contains any singularities of the function $f(\cdot)$.

This sparse representation also explains why (first generation) wavelets provide smoothness characterization of function spaces like Sobolev and Besov bodies as suggested in Section A.1.3 (see [38] for further information). Another interesting example of this wavelet–based characterization is the case of β–Hölder functions.

Def. A.4 *The class $\Lambda^\beta(C)$ of Hölder continuous functions is defined as follow:*

1. *if $\beta \leqslant 1$, $\Lambda^\beta(C) = \left\{ f \in L^2 : |f(x) - f(y)| \leqslant C\,|x-y|^\beta \right\}$,*
2. *if $\beta > 1$, $\Lambda^\beta(C) = \left\{ f \in L^2 : \left|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)\right| \leqslant C\,|x-y|^{\beta'} \wedge \left|f^{(\lfloor\beta\rfloor)}(x)\right| < \infty \right\}$, where $\lfloor\beta\rfloor$ is the largest integer less than β and $\beta' = \beta - \lfloor\beta\rfloor$.*

The *global*[4] Hölder regularity of a function can be characterized as follow [20, 30]:

Theorem A.2 *Let $f \in \Lambda^\beta(C)$, and suppose that the (orthogonal) wavelet function $\psi(\cdot)$ has r continuous derivatives and r vanishing moments with $r > \beta$. The*

$$\left|\left\langle f, \psi_{j,k} \right\rangle_{L^2}\right| \leqslant C_1 2^{-j\left(\beta+\frac{1}{2}\right)}.$$

where $C_1 \in \mathbb{R}_+$.

[4] In [56] a *local* Hölder regularity is defined a it is shown how this quantity is related to a brand new family of wavelet–characterized function spaces called *Oscillation* spaces.

B Act II: Bayesian and Treed Subspace Pre–Testing.

[Background]

B.1 Multiple Shrinkage.

The classical James–Stein estimator (JSE) used by Li [76] in constructing his asymptotically optimal confidence set, depends heavily on the specification of a *single* target point toward which the observation vector is shrunk. It is well–known that regardless of the location of the target in relation to the *true* mean, the JSE will dominate the MLE but still meaningful reductions in MSE occur only if this shrinkage target is relatively close to the *true* mean vector. If the shrinkage target is a poor guess, the JSE will provide little or no improvement upon the MLE. There are many scenarios in which an accurate shrinkage target might be available. A reasonable target might be provided by the result of a previous experiment or by some special physical or statistical structure of the problem. In many cases however it is difficult to specify a single accurate point shrinkage target for use in the JSE. In this section I will review an expansion to the basic JSE framework that allows adaptive simultaneous shrinkage toward multiple target points or multiple subspaces that allows for a substantial reductions in risk compared to the MLE not just in the vicinity of a single point. The discussion in this section focuses mainly on the results in [43] and [42].

B.1.1 Shrinkage Toward a Subspace.

In many settings the requirement that the JSE shrink toward a unique target point is way too restrictive: consider, for instance, an observation vector comprising measurements that are believed to be independent and identically distributed. In this case, although there is a strong prior indication of what might be a "reasonable" estimate of the mean $\boldsymbol{\theta} \in \mathbb{R}^p$ – namely, $\widehat{\boldsymbol{\theta}} = \widehat{\theta}\mathbf{1}_p$ – the the standard form of the JSE does not allow incorporation of this knowledge. More generally, if there is reason to believe that the mean belongs to a lower dimensional subspace of $\mathbb{R}^p$, we might want to restrict any estimate to be close to this subspace. Again, however, the JSE in its basic formulation does not has the ability to do this.

Consider again the general problem: we wish to estimate $\boldsymbol{\theta} \in \mathbb{R}^p$ from a vector observation $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\theta}, \sigma^2 \mathsf{I}_p)$. Suppose that instead of specifying a single point as a shrinkage target, we wish to specify an entire subspace $\mathsf{V} \subset \mathbb{R}^p$, that captures some belief about the region of $\mathbb{R}^p$ in which a reasonable estimate might lie. Adapting the JSE, we would like to shrink toward the entire subspace V instead of toward a single point. The modification needed is remarkably simple: let the subspace V have dimension $p - q$ where q is required to be greater than 2. Define a projection operator $\mathcal{P}_\mathsf{V}$ from $\mathbb{R}^p$ into V and a projection operator $\mathcal{P}_\perp = \mathbb{1} - \mathcal{P}_\mathsf{V}$ from $\mathbb{R}^p$ into the orthogonal complement of V in $\mathbb{R}^p$ (which has dimension q), so that

$$\mathbf{x} = \mathcal{P}_\mathsf{V}\mathbf{x} + \mathcal{P}_\perp\mathbf{x} \triangleq \mathbf{x}_\mathsf{V} + \mathbf{x}_\perp.$$

Now, because $\mathbf{x}_\mathsf{V} \in \mathsf{V}$, it corresponds to the component of the observation that conforms to the prior belief indicating V as the subspace in which a reasonable estimate should lie. On the other hand, $\mathbf{x}_\perp$ corresponds to the deviation of the observation from this prior belief. It might then be reasonable to estimate $\boldsymbol{\theta}$ by maintaining the component $\mathbf{x}_\mathsf{V}$, but reducing the magnitude of the orthogonal component $\mathbf{x}_\perp$. In other words, we might want to retain $\mathbf{x}_\mathsf{V}$ and shrink $\mathbf{x}_\perp$ toward $\mathbf{0}_p$. Applying the JSE we get

$$\widehat{\boldsymbol{\theta}}_{\mathrm{JS}[\mathsf{V}]} = \mathbf{x}_\mathsf{V} + \left[1 - \frac{\sigma^2(q-2)}{\|\mathbf{x}_\perp\|_2^2}\right]\mathbf{x}_\perp = \mathcal{P}_\mathsf{V}\mathbf{x} + \left[1 - \frac{\sigma^2(q-2)}{\|\mathbf{x} - \mathcal{P}_\mathsf{V}\mathbf{x}\|_2^2}\right](\mathbf{x} - \mathcal{P}_\mathsf{V}\mathbf{x}). \tag{B.1}$$

Of course, Equation B.1 is an adaptation of the JSE to the case where shrinkage is desired not toward a single point but toward an entire subspace and the standard JSE can be viewed as a special case in which $\mathsf{V} = \{\mathbf{v}\} \in \mathbb{R}^p$, $q \equiv p$ and $\mathcal{P}_\mathsf{V}\mathbf{x} = \mathbf{v}$. The subspace–target JSE just introduced can be trivially modified to yield a dominating positive–part JSE as for the single target point case. The general idea of modifying the JSE to shrink toward a subspace instead of a point appears to have been first proposed by Lindley [77] in 1962.

Shrinkage toward a subspace as prescribed by Equation B.1 adds a great amount of flexibility to the estimation procedure but, as usual, it also comes with a serious drawback: a reduction in the potential improvement over the MLE. It is well–known that the potential savings in MSE increases with the dimensionality of the problem; by shrinking toward a subspace instead of a single point, the dimension of the estimation problem is effectively reduced resulting in a reduction of the potential improvement. Depending on the specific application, this may or may not be an acceptable price to pay.

B.1.2 Shrinkage Toward Multiple Targets.

Suppose that multiple vague and possibly conflicting prior information suggest that any one of the (linear) subspaces in the family $\{\mathsf{V}_k\}_{k\in\{1,\dots,K\}}$ with $\mathsf{V}_k \subset \mathbb{R}^p$ and $\mathsf{dim}(\mathsf{V}_k) = p - q_k$, might be an appropriate shrinkage target for estimation of $\boldsymbol{\theta}$ from $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\theta}, \sigma^2\mathsf{I})$. Denote by $\widehat{\boldsymbol{\theta}}_{\mathsf{JS}[k]}$ the positive–part JSE formed by shrinking toward the $k-$th subspace V_k:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{JS}[k]} = \mathcal{P}_k\mathbf{x} + \left[1 - \frac{\sigma^2(q_k-2)}{\|\mathbf{x}-\mathcal{P}_k\mathbf{x}\|_2^2}\right]_+ (\mathbf{x} - \mathcal{P}_k\mathbf{x}) \tag{B.2}$$

where $\mathcal{P}_k$ is the projection operator from $\mathbb{R}^p$ into V_k and $q_k > 2$. The choice of which of these JSEs to use will have a great impact on the performance of the estimator.

It is natural to investigate whether the K estimators defined by Equation B.2 might be combined to allow for substantial reductions in risk as long as $\boldsymbol{\theta}$ is close to *any* of the V_k. George studies this issue in [43] drawing a parallel to Bayesian estimation to derive a candidate multiple shrinkage estimator. More specifically, consider again the problem of estimating $\boldsymbol{\theta}$ from $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\theta}, \sigma^2\mathsf{I}_p)$ under a prior $\pi_k(\boldsymbol{\theta})$. The Bayesian estimate $\widehat{\boldsymbol{\theta}}_{\mathsf{B}[k]} = \mathbb{E}(\boldsymbol{\theta}|\mathbf{x})$ in this case can be shown[5] to satisfy

$$\widehat{\boldsymbol{\theta}}_{\mathsf{B}[k]} = \mathbf{x} + \sigma^2\nabla\log\left(\mathsf{m}_k(\mathbf{x})\right), \tag{B.3}$$

where $\mathsf{m}_k(\mathbf{x})$ is the marginal density for $\mathbf{x}$, i.e.,

$$\mathsf{m}_k(\mathbf{x}) = \int_{\mathbb{R}^p} \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{x}-\boldsymbol{\theta}\|_2^2\right]\pi_k(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{B.4}$$

One way to interpret an arbitrary estimate $\widehat{\boldsymbol{\theta}}$ that can be put in the form (B.3), then, is as a "pseudo–Bayesian" estimate of $\boldsymbol{\theta}$ generated under a particular marginal density $\mathsf{m}_k(\mathbf{x})$. George shows that each of the JSEs given by (B.2) can be put in the form (B.3) with

$$\mathsf{m}_k(\mathbf{x}) = \begin{cases} \left[\frac{\sigma^2(q_k-2)}{\|\mathbf{x}-\mathcal{P}_k\mathbf{x}\|_2^2}\right]^{\frac{q_k-2}{2}} \exp\left[-\frac{q_k-2}{2}\right], & \text{if } \|\mathbf{x}-\mathcal{P}_k\mathbf{x}\|_2^2 \geqslant \sigma^2(q_k-2); \\ \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathcal{P}_k\mathbf{x}\|_2^2\right], & \text{if } \|\mathbf{x}-\mathcal{P}_k\mathbf{x}\|_2^2 < \sigma^2(q_k-2), \end{cases} \tag{B.5}$$

for $k \in \{1,\dots,K\}$. A few comments regarding Equation B.5 are in order. First of all notice the split based on a "test" very similar to the one used by Baraud to build his confidence regions. As a consequence, each of the marginals naturally encapsulates some soft prior information suggesting or discouraging shrinkage toward a subspace V_k: each $\mathsf{m}_k(\cdot)$ monotonically decreases as $\mathbf{x}$ becomes more distant from the corresponding V_k. This having said, it is worth adding that the $\mathsf{m}_k(\cdot)$'s are not true marginal densities: examination of Equation B.5, in fact, reveals that these "marginals" do not, in general, integrate to 1. Secondly, George does not explicitly assume that these marginals actually arise from appropriately chosen priors by way of Equation B.4. The assumed marginals are essentially convenient constructions used to draw an analogy between James–Stein estimation and Bayesian estimation.

Returning to the problem of trying to combine the shrinkage estimators, suppose that prior information is available regarding how likely each V_k is to be the "correct" target for shrinkage. Specifically, suppose that we were actually able to assign weights $\{w_k\}_{k\in\{1,\dots,K\}}$, satisfying $\sum_{k=1}^K w_k = 1$, to each of the target–space. In the Bayesian analogy, this corresponds to specify a mixture prior $\pi(\boldsymbol{\theta}) = \sum_{k=1}^K w_k\pi_k(\boldsymbol{\theta})$ that produces a marginal density equal to

$$\mathsf{m}(\mathbf{x}) = \sum_{k=1}^K w_k\mathsf{m}_k(\mathbf{x}), \tag{B.6}$$

[5] See [72].

and, consequently, a (pseudo)Bayesian estimate of $\boldsymbol{\theta}$ that represents the most natural combination scheme we may think about is given by

$$\widehat{\boldsymbol{\theta}}_{\mathsf{B}} = \mathbf{x} + \sigma^2 \nabla \log\left(\mathsf{m}(\mathbf{x})\right). \tag{B.7}$$

George shows that $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$ can be expressed in the following enlightening way

$$\widehat{\boldsymbol{\theta}}_{\mathsf{B}} = \sum_{k=1}^{K} \left[\frac{w_k \mathsf{m}_k(\mathbf{x})}{\mathsf{m}(\mathbf{x})}\right] \widehat{\boldsymbol{\theta}}_{\mathsf{B}[k]} \triangleq \sum_{k=1}^{K} \rho_k(\mathbf{x}) \widehat{\boldsymbol{\theta}}_{\mathsf{B}[k]}.$$

Clearly $\sum_{k=1}^{K} \rho_k(\mathbf{x}) = 1$, thus $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$ is just a data dependent[6] convex combination of the individual shrinkage estimators with adaptive weights given by the *relevance* functions $\{\rho_k(\mathbf{x})\}_{k\in\{1,\dots,K\}}$. This is George's multiple shrinkage estimators.

Investigation of the behavior of $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$ reveals many appealing properties. First of all when $\mathbf{x}$ is distant from all of the target–spaces, the shrinkage provided by each of the component estimators is very small and, as a result, the combined estimators is approximately the same as the MLE. Secondly, the *relevance* function $\{\rho_k(\cdot)\}_k$ adapt to $\mathbf{x}$ so to provide the most shrinkage toward the closest V_k and, viceversa, very limited shrinkage toward distant V_k: the *relevance* functions vary proportionally to the corresponding marginal $\mathsf{m}_k(\cdot)$ that, in turn, is largest when $\mathbf{x}$ is close to V_k. Thirdly, the prior weights w_k can be chosen to provide more shrinkage toward lower–dimensional subspaces to reflect the fact that proximity of $\mathbf{x}$ to a low–dimensional subspace V_k is a better validation of prior information than proximity of $\mathbf{x}$ to a high–dimensional subspace[7]. In the absence of other prior information, George proposes to implement this idea using the following prior weights:

$$w_k = C^{\left(\frac{q_k-2}{2}\right)} \mathrm{e}^{\left(\frac{q_k-2}{2}\right)}, \tag{B.8}$$

where $C \geqslant 1$ is an arbitrary constant to be tuned in order to modulate the amount of shrinkage toward smaller subspaces for equal distances. Using this scheme, equal–dimensional subspaces are all weighted similarly and, consequently, the *relevance* function $\rho_k(\cdot)$ represents, in most cases, just the shrinkage coefficient of the $k-th$ component estimator. In addition, regardless of the specific choice of the w_k, the multiple shrinkage estimator behaves approximately like $\widehat{\boldsymbol{\theta}}_{\mathsf{B}[k]}$ in the vicinity of a single V_k. Finally, in view of what is coming up next, it seems interesting to observe that, as a matter of fact, the center of Baraud's is simply a particular case of George's multiple shrinkage estimator; more specifically, it is the one obtained by a winner–take–all quantization of the weight functions. The "winner" in this case, is clearly the estimator that induces the minimal radius and that, in general, coincides with the one associated to the maximum weight (i.e. the one that shrinks toward the nearest subspace).

But, as I said before, there is a price for adaptivity. A naive approach to achieve a good estimate would be to specify a plethora of subspaces V_k as shrinkage targets. This approach is clearly flawed: the more subspace targets we specify, the more unwanted shrinkage toward incorrect subspaces will be performed. More formally, we can examine the reduction in risk from the MLE achieved by $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$ studying its Stein's unbiased risk estimate ( ). George shows that the risk of each component estimator is given by

$$_k(\mathbf{x}) = p\,\sigma^2 - \mathsf{D}_k(\mathbf{x}),$$

where

$$\mathsf{D}_k(\mathbf{x}) = \begin{cases} \frac{\left[\sigma^2(q_k-2)\right]^2}{\|\mathbf{x}-\mathcal{P}_k\mathbf{x}\|_2^2}, & \text{if } \|\mathbf{x}-\mathcal{P}_k\mathbf{x}\|_2^2 \geqslant \sigma^2(q_k-2); \\ 2\sigma^2 q_k - \|\mathbf{x}-\mathcal{P}_k\mathbf{x}\|_2^2, & \text{if } \|\mathbf{x}-\mathcal{P}_k\mathbf{x}\|_2^2 < \sigma^2(q_k-2), \end{cases}$$

whereas, the risk estimate for $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$ turns out to be

$$(\mathbf{x}) = p\,\sigma^2 - D(\mathbf{x}),$$

[6] It is quite striking the similarity of the current estimator to the one obtained within the *hierarchical mixture of experts* framework introduced by Jordan and Jacobs in [64] and further explored, among other papers, in [60, 59, 61, 62].

[7] That is, a point chosen at random is less likely to be near a lower–dimensional subspace than a higher–dimensional

with

$$\mathsf{D}(\mathbf{x}) = \sum_{k=1}^{K} \rho_k(\mathbf{x})\mathsf{D}_k(\mathbf{x}) - \sum_{k=1}^{K}\sum_{\ell=1}^{K} \rho_k(\mathbf{x})\rho_\ell(\mathbf{x}) \left\| \widehat{\boldsymbol{\theta}}_{\mathsf{B}[k]} - \widehat{\boldsymbol{\theta}}_{\mathsf{B}[\ell]} \right\|_2^2. \tag{B.9}$$

This last equation implies that the expected reduction in risk achieved by $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$ is roughly a convex combination of the reductions achieved by each individual $\widehat{\boldsymbol{\theta}}_{\mathsf{B}[k]}$, with the important caveat that the risk is possibly inflated by a factor depending on how far apart the component estimators are. If $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$ is strongly influenced by only one $\widehat{\boldsymbol{\theta}}_{\mathsf{B}[k]}$, then the remaining component, being relatively "next" to each other, will be roughly equal to the MLE, and $\mathsf{D}(\cdot)$ will approximatively equal the corresponding $\mathsf{D}_k(\cdot)$. On the other hand, if $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$ is influenced by two or more components, each of them shrinking toward a different subspace and thus producing widely separated estimates, then Equation B.9 indicates that there will be a price to pay in terms of the expected reduction in risk. An unjustified proliferation of shrinkage targets pooled together into a single estimator like $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$, results in an increased probability that $\mathbf{x}$ will be pulled in many different directions confounding the desirable property characterizing the JSE, namely, a significant reduction in risk near the shrinkage target. Additionally, regardless of the number of shrinkage targets, there is a price to pay for adaptivity even in the ideal case when $\boldsymbol{\theta} \in \mathsf{V}_k$ for some k. The reason behind this is simply that the *relevance* functions $\rho_k(\cdot)$ are, in general, smaller than 1[8] and this clearly implies that there will always be some "pressure" from the other components in pulling the overall estimate out of the "true" V_k, causing $\widehat{\boldsymbol{\theta}}_{\mathsf{B}}$ not to perform quite as well[9] as $\widehat{\boldsymbol{\theta}}_{\mathsf{B}[k]}$.

[8] Except in degenerate case when $w_k = 1$ or $\mathbf{x} \in \mathsf{V}_k$

[9] To fix this, we could design a simple winner–take–all quantizer based on some suitable threshold over the *relevance* functions.

References

[1] F. Abramovich, P. Besbeas, and T. Sapatinas. Empirical Bayes approach to block wavelet function estimation. *Computational Statistics and Data Analysis*, 39:435–451, 2002.

[2] F. Autin, D. Picard, and V. Rivoirard. Maxiset comparisons of procedures, application to choosing priors in a bayesian nonparametric setting. Technical Report PMA-931, Laboratoire de Probabilités et Modéles Aléatoires, Université Paris 7, 2004.

[3] Y. Baraud. Confidence balls in Gaussian regression. *The Annals of Statistics*, 32(2):528–551, 2004.

[4] E. Belitser and S. Ghosal. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *The Annals of Statistics*, 31(2):536–559, 2003.

[5] R. Beran. Stein confidence sets and the bootstrap. *Statistica Sinica*, 5:109–127, 1995.

[6] R. Beran. Superefficient estimation of multivariate trend. *Mathematical Methods of Statistics*, 8:166–180, 1999.

[7] R. Beran. REACT scatterplot smoothers: Superefficiency through basis economy. *Journal of the American Statistical Association*, 95:155–171, 2000.

[8] R. Beran and L. Dümbgen. Modulation of estimators and confidence sets. *The Annals of Statistics*, 26(5):1826–1856, 1998.

[9] G. Bourdaud. Ondelettes et espaces de Sobolev. *Rev. Mat. Iberoam.*, 11(3):477–512, 1995.

[10] L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398, December 1996.

[11] L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, 24(6):2524–2535, 1996.

[12] P. Brutti. Variable bandwidth schemes for local polynomial smoothers via vertical wavelet thresholding. In S. Barber R.G. Aykroyd and K.V. Mardia, editors, *Bioinformatics, Images, and Wavelets*, pages 119–121. Department of Statistics, University of Leeds, 2004.

[13] T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *The Annals of Statistics*, 27:898–924, 1999.

[14] T. Cai. Rates of convergence and adaptation over Besov spaces under pointwise risk. *Statistica Sinica*, 13:881–902, 2003.

[15] T. Cai and M. G. Low. An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840, 2004.

[16] T. Cai and M. G. Low. Adaptation under probabilistic error for estimating linear functionals. *Journal of Multivariate Analysis*, 2005. In Press.

[17] T. Cai and M. G. Low. Adaptive estimation of linear functionals under different performance measures. *Bernoulli*, 11, 2005.

[18] T. Cai and M. G. Low. Nonparametric estimation over shrinking neighborhoods: superefficiency and adaptation. *The Annals of Statistics*, 33(1), 2005.

[19] T. Cai and W. Silverman. Incorporating information on the neighboring coefficients into wavelet estimation. *Sankhyā, Series B*, 63:127–148, 2001. Special issue on wavelets.

[20] T. T. Cai and L. D. Brown. Wavelet shrinkage for nonequispaced samples. *The Annals of Statistics*, 26(5):1783–1799, October 1998.

[21] D. De Canditiis and B. Vidakovic. Wavelet Bayesian block shrinkage via mixtures of normal–inverse gamma priors. Technical Report RT 234/01, Istituto per le Applicazioni del Calcolo , Sezione di Napoli, 2001.

[22] T. Choi and M. J. Schervish. Posterior consistency in nonparametric regression problems under Gaussian process priors. Technical Report 809, Department of Statistics, Carnegie Mellon University, 2004.

[23] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore. Tree approximation and optimal encoding. *Applied Computational and Harmonic Analysis*, 11(2):167–191, 1999.

[24] A. Cohen, I. Daubechies, and J. Feauveau. Bi-orthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45:485–560, 1992.

[25] A. Cohen, I. Daubechies, and P. Vial. Wavelet on the interval and fast wavelets tranforms. *Appl. Comput. Harmon. Anal.*, 1:54–81, 1993.

[26] A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard. Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon*, 11(2):167–191, 2001.

[27] D. J. Cummins, T. G. Filloon, and D. Nychka. Confidence intervals for nonparametric curve estimates: Toward more uniform pointwise coverage. *Journal of the American Statistical Association*, 96:233–246, 2001.

[28] G. S. Datta and R. Mukerjee. *Probability Matching Priors: Higher Order Asymptotics*. Number 178 in Lecture Notes in Statistics. Springer, first edition, 2004.

[29] I. Daubechies. Orthonormales bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 41:909–996, 1988.

[30] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.

[31] V. Delouille, J. Franke, and R. von Sachs. Nonparametric stochastic regression with design-adapted wavelets. *Sankhyā, Series A*, 63(3):328–366, 2001.

[32] V. Delouille, J. Simoens, and R. von Sachs. Smooth design-adapted wavelets for nonparametric stochastic regression. *Journal of the American Statistical Association*, 99(467):643–658, 2004.

[33] R. Devore. Nonlinear approximation. *Acta Numerica*, pages 1–99, 1998.

[34] D. L. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270, 1994.

[35] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(425-455):425–455, 1994.

[36] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinking. *Journal of the American Statistical Association*, 90(1200-1224):1200–1224, 1995.

[37] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence, iii. *The Annals of Statistics*, 19(2):668–701, 1991.

[38] D. E. Edmunds and H. Triebel. *Function Spaces Entropy Numbers and Differential Operators*, volume 120 of *Cambridge Tracts in Mathematics*. Cambridge University Press, first edition, 1996.

[39] S. Efromovich and M. G. Low. Adaptive estimates of linear functionals. *Probability theory and related fields*, 98:261–275, 1994.

[40] P. Fryzlewicz. Bivariate hard thresholding in wavelet function estimation. Technical Report TR-04-03, Department of Mathematics, Imperial College London, UK, 2004.

[41] C. R. Genovese and L. Wasserman. Confidence sets for nonparametric wavelet regression. *The Annals of Statistics*, 33(2), 2005.

[42] E. I. George. Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, 81(394):437–445, 1986.

[43] E. I. George. Minimax multiple shrinkage estimation. *The Annals of Statistics*, 14(1):188–205, 1986.

[44] S. Ghosal, N. Choudhuri, and A. Roy. Bayesian methods for function estimation. Technical report, North Carolina State University Department of Statistics, 2003.

[45] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. *Non-informative priors via sieves and packing numbers*, pages 119–132. Advances in Statistical Decision Theory and Applications. Birkhauser, Boston, 1997.

[46] S. Ghosal, L. Jüri, and A. W. van der Vaart. Bayesian adaptation. In *Acta Appl. Math.*, volume 79 of *Proceedings of the Eighth Vilnius Conference on Probability Theory and Mathematical Statistics. Part II*, pages 165–175, 2003.

[47] S. Ghosal and A. W. van der Vaart. Convergence rates of posterior distributions for non i.i.d. observations. Technical report, North Carolina State University Department of Statistics, 2005.

[48] J. K. Ghosh and R. Mukerjee. On perturbed ellipsoidal and highest posterior density regions with approximate frequentist validity. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(4):761–769, 1995.

[49] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics. Springer, first edition, 2003.

[50] I. Grama and M. Nussbaum. Asymptotic equivalence for nonparametric regression. *Mathematical Methods of Statistics*, 11(1):1–36, 2002.

[51] C. Gu. *Smoothing Spline ANOVA Models*. Springer Series in Statistics. Springer, first edition, 2002.

[52] P. Hall, G. Kerkyacharian, and D. Picard. On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica*, 9:33–50, 1999.

[53] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation and Statistical Applications*. Number 129 in Lecture Notes in Statistics. Springer–Verlag, New York, 1998.

[54] Tzee-Ming Huang. Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics*, 32(4):1556–1593, 2005.

[55] I. A. Ibragimov and R. Z. Hasminskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.

[56] S. Jaffard. Beyond Besov spaces, part 2: Oscillation spaces. *Constructive Approximation*, 2004.

[57] S. Jaffard, Y. Meyer, and R. Ryan. *Wavelet Tools for Science and Technology*. S.I.A.M, 2000.

[58] W. Jang, C. R. Genovese, and L. Wasserman. Nonparametric density estimation and clustering in astronomical sky surveys. Technical Report 797, Carnegie Mellon University, Department of Statistics, 2004.

[59] W. Jiang and M. A. Tanner. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *The Annals of Statistics*, 27:987–1011, 1999.

[60] W. Jiang and M. A. Tanner. On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural Computation*, 11:1183–1198, 1999.

[61] W. Jiang and M. A. Tanner. On the identifiability of mixtures-of-experts. *Neural Networks*, 12:1253–1258, 1999.

[62] W. Jiang and M. A. Tanner. On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models. *IEEE Transactions on Information Theory*, 46(3):1005–1013, May 2000.

[63] I. M. Johnstone and B. W. Silverman. Needles and hay in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32:15941649, 2004.

[64] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[65] A. Juditsky and S. Lambert-Lacroix. Nonparametric confidence set estimation. *Mathematical Methods of Statistics*, 12(4):410–428, 2003.

[66] J. P. Kahane and P. G. Lemarié. *Fourier series and Wavelets*. Gordon Breach, New York, 1996.

[67] G. Kerkyacharian and D. Picard. Minimax or maxisets? *Bernoulli*, 8(2):219–253, 2002.

[68] G. Kerkyacharian and D. Picard. Thresholding algorithms, maxisets and well-concentrated bases. *Test*, 9(2):283–344, 2002.

[69] L. Le Cam. On some asymptotic properties of maximum likelihood estimated and related Bayesian estimate. *University of California, Publications in Statistics*, 1:277–330, 1953.

[70] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, 1986.

[71] L. Le Cam and G. Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer, 2000.

[72] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, New York, second edition, 1998.

[73] P. G. Lemarié and Y. Meyer. Ondelettes et bases hilbertiennes. *Revista Math.Iberoamericana*, 1, 1986.

[74] O. Lepski. Asymptotic minimax estimation with prescribed properties. *Theory of Probability and its Applications*, 34:604–615, 1990.

[75] O. Lepski. How to improve the accuracy of estimation. *Mathematical Methods of Statistics*, 8:441–486, 1999.

[76] K. C. Li. Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008, 1989.

[77] D. V. Lindley. Discussion on professor Stein's paper. *Journal of the Royal Statistical Society, Series B, Methodological*, 24:285–287, 1962.

[78] C. Loader. *Local Regression and Likelihood*. Statistics and Computing. Springer, first edition, 1999.

[79] M. G. Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.

[80] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, second edition, 1998.

[81] Y. Meyer. *Wavelets and Operators*. Cambridge University Press, 1992.

[82] C. N. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65–80, 1982.

[83] C. N. Morris. Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics*, 11(2):515–529, 1983.

[84] D. Nychka. Bayesian "confidence" intervals for smoothing splines. *Journal of the American Statistical Association*, 83:1134–1143, 1988.

[85] D. Picard and K. Tribouley. Adaptive confidence interval for pointwise curve estimation. *The Annals of Statistics*, 28(1):298–335, 2000.

[86] J. Robins and A. van der Vaart. Adaptive nonparametric confidence sets. Technical Report 2004-5, Vrije Universiteit Amsterdam, Stochastics Section, 2004.

[87] J. Romberg, H. Choi, and R. Baraniuk. Bayesian tree-structured image modeling using wavelet domain hidden Markov models. *IEEE Transactions on Image Processing*, 10(7):1056–1068, 2001.

[88] H. Scheffe. *The Analysis of Variance*. John Wiley & Sons, 1959.

[89] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

[90] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

[91] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.

[92] J. Sun. Tail probabilities of the maxima of Gaussian random fields. *The Annals of Probability*, 21:34–71, 1993.

[93] J. Sun and C. Loader. Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics*, 22:1328–1345, 1994.

[94] J. Sun and C. Loader. Robustness of tube formula based confidence bands. *Journal of Computational and Graphical Statistics*, 1997.

[95] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.

[96] B. Vidakovic. *Statistical Modeling by Wavelets*. Wiley-Interscience, New York, 1999.

[97] G. Wahba. Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B, Methodological*, 45:133–150, 1983.

[98] Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS–NSF Regional Conference Series in Applied Mathematics*. SIAM, 1990.

[99] S. G. Walker. Modern Bayesian asymptotics. *Statistical Science*, 19(1):111–117, 2004.

[100] X. W. Wang and A. T. A. Wood. Empirical Bayes block shrinkage of wavelet coefficients via the non–central χ^2 distribution. Technical Report 03–01, University of Nottingham, Division of Statistics, 2003.

[101] Y. Wang and G. Wahba. Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals. *J. Statistical Computation and Simulation*, 51:263–279, 1995.

[102] Y. Wang, G. Wahba, R. Chappell, and C. Gu. Simulation studies of smoothing parameter estimates and Bayesian confidence intervals in Bernoulli SS ANOVA models. *Communications in Statistics, Part B – Simulation and Computation [Split from: @J(CommStat)]*, 24:1037–1059, 1995.

[103] Y. Yang. Minimax rate adaptive estimation over continuous hyper–parameters. *IEEE Transaction on Information Theory*, 47:2081–2085, 2001.