Nonparametric statistical inverse problems

To cite this article: L Cavalier 2008 *Inverse Problems* **24** 034004

IOP PUBLISHING INVERSE PROBLEMS

Inverse Problems **24** (2008) 034004 (19pp) doi:10.1088/0266-5611/24/3/034004

Nonparametric statistical inverse problems

L Cavalier

Université Aix-Marseille 1, CMI, 39 rue Joliot-Curie, 13453 Marseille cedex 13, France

E-mail: cavalier@cmi.univ-mrs.fr

Received 7 September 2007, in final form 26 October 2007
Published 23 May 2008
Online at stacks.iop.org/IP/24/034004

Abstract

We explain some basic theoretical issues regarding nonparametric statistics applied to inverse problems. Simple examples are used to present classical concepts such as the white noise model, risk estimation, minimax risk, model selection and optimal rates of convergence, as well as more recent concepts such as adaptive estimation, oracle inequalities, modern model selection methods, Stein's unbiased risk estimation and the very recent risk hull method.

1. Introduction

Loosely speaking, to solve an inverse problem means to recover an object f from indirect noisy observations Y. The object f is usually modeled as a function (or a vector) that has been modified by an operator A; thus, one observes a noisy version of Af. From a mathematical point of view, to solve the inverse problem is to 'invert' the operator A. The problem is that A may not be invertible or nearly so. This is the case of ill-posed problems and is of great practical interest as it arises naturally in many fields such as geophysics, finance, astronomy, biology, etc.

Ill-posed problems are further compounded by the presence of errors (noise) in the data. Statistics enters inverse problems when at least one of the components of the inverse problem (usually the noise) is modeled as stochastic. The question is then to study statistical 'regularization' methods that lead to a meaningful reconstruction despite the noise and ill-posedness.

In our opinion, the inverse problem framework is better known among statisticians than its statistical approach among the inverse problem community. For instance, the latter is well acquainted with the concepts of mean, variance and bias but is less familiar with classical concepts such as the white noise model, risk estimation, minimax risk, model selection and optimal rates of convergence, which we will discuss. In addition to these classical notions we will present some more recent concepts that have been developed since the 1980s such as adaptive estimation, oracle inequalities, modern model selection methods, Stein's unbiased risk estimation and the very recent risk hull method.

0266-5611/08/034004+19$30.00

All the statistical concepts will be defined and discussed in the framework of inverse problems. Although some of the techniques are specific to this field, some may also be used in more general situations. Other statistical methods not discussed in this paper may also have applications to inverse problems, but one should be careful with their application given the intrinsic difficulty and instability of ill-posed problems.

Our objective in this paper is to explain some basic theoretical issues regarding the statistical framework of inverse problems. The paper provides a glimpse of modern nonparametric statistics in the context of inverse problems. Other topics and reviews may be found in [24, 41, 55, 57, 59, 63]. Needless to say, to make our presentation more direct and succinct we will be forced to make some simplifications.

1.1. Linear inverse problems with random noise

As explained above, the data are noisy observations of Af where the unknown function f typically belongs to a space with a natural inner product and A is a linear operator between the two Hilbert spaces H and G. Furthermore, it is usually assumed that A is a bounded operator and the Hilbert spaces are separable.

The standard framework for inverse problems (first proposed in [64, 65]) corresponds to a model with deterministic (and additive) noise, where ξ is considered an element of G with $\|\xi\| \leqslant 1$. That is, the noise is an unknown element within a ball in G. Since no other information is available, inversion estimates are obtained for any possible noise component, i.e., for the worst noise. The approach we consider is statistical (due to [62]); the noise variable ξ is assumed to be random.

We will make the standard assumption of the Gaussian white noise. In the case $G = \mathbb{R}^n$, this just means that ξ is a vector of i.i.d. Gaussian random variables. If ξ is instead a stochastic process, then white noise is defined as a derivative of the Brownian motion (see, for example [34]). In general, ξ is the Gaussian white noise if for any functions $g_1, g_2 \in G$, the random variables $\langle \xi, g_j \rangle$ are $\mathcal{N}(0, \|g_j\|^2)$ with covariance $\mathrm{Cov}(\langle \xi, g_1 \rangle, \langle \xi, g_2 \rangle) = \langle g_1, g_2 \rangle$ (see [34]), where $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$, respectively, denote the norm and the scalar product of H and G.

An important remark is that while the deterministic noise is such that $\|\xi\| \leqslant 1$, in the statistical framework a white noise process is not an element of H because $\|\xi\| = \infty$. In this sense, one difference between the deterministic and the stochastic approaches to inverse problems is that the random noise is large compared to the deterministic noise. This point has already been made in, for example, [17]. The optimal rates of convergence (see section 2.3 for the stochastic noise and [22] for the deterministic one) are then usually different in the two settings. In some special cases, such that severely ill-posed problems or estimation of a functional, the rates are the same. However, statisticians usually consider problems with a rather strong noise.

The standard discrete sample statistical model for linear inverse problems is

$$Y_i = Af(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are observed (we may assume $X_i \in [0, 1]$), f is an unknown function in $L^2(0, 1)$, A is an operator from $L^2(0, 1)$ into $L^2(0, 1)$ and ξ_i are i.i.d. zero-mean Gaussian random variables of variance σ^2.

In theoretical studies, one will often see (1) written as a Gaussian white noise model (i.e., white noise stochastic process)

$$Y = Af + \varepsilon\xi, \tag{2}$$

where ξ is a stochastic error and $\varepsilon > 0$ is the noise level. For example, more formally, the white noise version of the *direct model* $Y_i = f(X_i) + \xi_i$ is written as $Y(\mathrm{d}x) = f(x)\,\mathrm{d}x + \varepsilon\xi(\mathrm{d}x)$,

which can also be written as $Y(x) = \int_0^x f(t)\,dt + \varepsilon W(t)$, where W is the Brownian motion ([30] provides generalizations of this asymptotic equivalence to the non-Gaussian white noise). In the special case where $A = I$, it is proved in [7] that under proper calibration the asymptotics of model (1) as $n \to \infty$ and (2) as $\varepsilon \to 0$ are equivalent with the asymptotics of the latter being easier to derive. In the inverse problem context, model (2) may be seen as an idealized version of (1).

In some cases, A does not have a trivial nullspace. The component of f that is in the nullspace cannot be reconstructed without more information and therefore the regularization of the problem has to include more information about the function to be recovered. See also the related notion of generalized inverse [26]. For the sake of simplicity, we will consider only the estimation of f when A is an injective (but not necessarily stable) operator.

There are several issues to be addressed in order to obtain an estimate of f given the data Y: (i) to deal with the observational noise to derive inversion estimates and assess their properties (statistics), (ii) to invert the operator A (inverse problems theory) and (iii) to derive numerical implementations for practical applications (computational mathematics).

Henceforth, we will be working in the statistical framework and will focus on questions related to (i) and (ii). However, we are also interested in the methods that are computationally efficient.

1.2. SVD and the sequence space model

We start with a discretization of the white noise model obtained by diagonalizing the forward operator A.

Suppose that A^*A (as usual, A^* stands for the adjoint of A) is a compact operator so that it has a complete orthogonal system of eigenvectors $\{\varphi_k\}$ with corresponding eigenvalues ρ_k. We then have the following representation of A^*A:

$$A^*Af = \sum_{k=1}^{\infty} \rho_k \langle f, \varphi_k\rangle \varphi_k = \sum_{k=1}^{\infty} b_k^2 \theta_k \varphi_k, \tag{3}$$

where $b_k = \sqrt{\rho_k}$ and $\theta_k = \langle f, \varphi_k\rangle$. Here, $\rho_k > 0$ because A is injective. If A^*Af can be written as in (3), then we say that A admits a singular value decomposition (SVD) with singular values $\{b_k\}$ with respect to the basis $\{\varphi_k\}$.

Remark 1. The SVD decomposition of a compact operator A is a special case of the spectral theorem (see [33]). The compactness assumption may be relaxed. Indeed any bounded linear operator is unitarily equivalent to a multiplication in some function space.

Remark 2. $\{\varphi_k\}$ is a natural basis for A because it diagonalizes A^*A. Examples of operators with known SVD are included in section 1.3. But note that the SVD has nothing to do with the function to be recovered. More 'flexible' basis functions such as wavelets are useful because even though they do not diagonalize A^*A, they almost do so for a wide variety of operators A. Such bases can also be chosen to adapt to the structure of f (see section 4 and [19, 20]).

The normalized image $\{\psi_k\}$ of $\{\varphi_k\}$ is defined by $A\varphi_k = b_k\psi_k$. Note that

$$\|\psi_k\|^2 = b_k^{-2}\langle A\varphi_k, A\varphi_k\rangle = b_k^{-2}\langle A^*A\varphi_k, \varphi_k\rangle = b_k^{-2}b_k^2\|\varphi_k\|^2 = 1.$$

Moreover, $A^*\psi_k = b_k^{-1}A^*A\varphi_k = b_k\varphi_k$. Thus, we have $A\varphi_k = b_k\psi_k$ and $A^*\psi_k = b_k\varphi_k$.

The coefficients of Y with respect to $\{\psi_k\}$ are

$$\begin{aligned} y_k &= \langle Y, \psi_k\rangle = \langle Af, \psi_k\rangle + \varepsilon\langle \xi, \psi_k\rangle = \left\langle Af, b_k^{-1}A\varphi_k\right\rangle + \varepsilon\xi_k \\ &= b_k^{-1}\langle A^*Af, \varphi_k\rangle + \varepsilon\xi_k = b_k\theta_k + \xi_k, \end{aligned}$$

where $\xi_k = \langle \xi, \psi_k \rangle$. Since ξ is white noise, $\{\xi_k\}$ is a sequence of i.i.d. standard Gaussian random variables $\mathcal{N}(0, 1)$. We thus have an equivalent discrete sequence observation model derived from (2):

$$y_k = b_k \theta_k + \varepsilon \xi_k, \qquad k = 1, 2, \ldots. \tag{4}$$

This model is called a *sequence space model* corresponding to (2). The problem of estimating f given Y has become that of estimating the sequence $\theta = \{\theta_k\}$ given the discrete sequence $y = \{y_k\}$. The sequence space formulation (4) or (5) for statistical inverse problems has been studied in a number of papers (see, for example [15, 18, 38, 46]).

An important example is the case $A = I$; it corresponds to the direct model where f is directly observed. In this case $b_k = 1$ and (4) is the classical sequence space model in statistics [6]. As noted in section 1.1, this model is related to the classical Gaussian white noise model (see [36]) and has a nonparametric regression interpretation.

Since the goal is to estimate $\{\theta_k\}$ and not $\{b_k \theta_k\}$, one has to remove $\{b_k\}$, which is equivalent to inverting the operator A. The effect of the ill-posedness of the inverse problem is clearly seen in the decay of b_k as $k \to \infty$; as k increases the 'signal' $b_k \theta_k$ usually gets weaker and is then more difficult to recover θ_k.

The following model equivalent to (4) is more natural to estimate θ_k:

$$X_k = \theta_k + \varepsilon \sigma_k \xi_k, \qquad k = 1, 2, \ldots, \tag{5}$$

where the new 'data' are $X_k = y_k / b_k$, and $\sigma_k = b_k^{-1} > 0$. This reduces the inverse problem to estimating a multivariate normal mean θ [60] but since $\sigma_k \to \infty$, the signal-to-noise ratio decreases as $k \to \infty$. This does not happen for the direct model for in this case $b_k = 1$. Hence, as expected, the problem of indirect observations is intrinsically more difficult.

The difficulty of a linear inverse problem can be defined by the behavior of σ_k: since $\sigma_k \to \infty$ as $k \to \infty$, the problem is ill-posed but the type of growth of σ_k can be used to define a measure of the ill-posedness. For example, we have the following definition.

Definition 1. *An inverse problem is called mildly ill-posed if the sequence σ_k has a polynomial growth $\sigma_k \approx k^\beta$ as $k \to \infty$ and severely ill-posed if σ_k increases at an exponential rate $\sigma_k \approx \exp(\beta k)$ for some $\beta > 0$. In either case, β is called the degree of ill-posedness of the inverse problem.*

For example, for the direct model we have $\sigma_k \approx 1$ as $k \to \infty$, which corresponds to a mildly ill-posed problem with $\beta = 0$. There are of course inverse problems whose ill-posedness is even worse than the exponential. Section 1.3.3 provides one example.

Definition 1 is used to understand the mathematical complexity of a given inverse problem. In section 1.3.3, we will see that the rates of convergence depend on the degree of ill-posedness as defined by β. But in practice one does not attempt to estimate β from the given data. This is really an ill-posed problem.

Remark 3. There exist more general definitions of the degree of ill-posedness related to the noise structure, smoothness assumptions on f and smoothing properties of A (see [51, 68]). However, for the sake of simplicity, we prefer to deal with this simple notion.

Remark 4. In model (1), one can obtain an equivalent sequence space model but with a finite sequence where $1 \leqslant k \leqslant n$. In this situation, the definitions are understood when $k \to \infty$ with n.

1.3. Examples

From a practical point of view, the SVD is computationally expensive and thus many popular inversion methods try to avoid its explicit use. But even for such methods the spectral domain is often used to study their theoretical performance. Here we present some examples of ill-posed problems where the SVD can be explicitly computed.

1.3.1. Differentiation. An important example is the estimation of a derivative from direct noisy observations. Suppose that we observe

$$Y = f + \varepsilon\xi, \tag{6}$$

where $H = L^2[0, 1]$, f is a periodic m-times continuously differentiable function on $[0, 1]$ (for some $m \in \mathbb{N}$) and ξ is the Gaussian white noise. A standard statistical inverse problem is the estimation of f or its mth derivative $D^m f = f^{(m)}$.

Here we may use the Fourier basis $\varphi_k(x) = \mathrm{e}^{2\pi \mathrm{i}kx}$, $k \in \mathbb{Z}$. Denote by θ_k the Fourier coefficients of f: $\theta_k = \int_0^1 f(x)\,\mathrm{e}^{2\pi \mathrm{i}kx}\,\mathrm{d}x$. The derivative is then given by

$$f^{(m)}(x) = \sum_{k=-\infty}^{\infty} (2\pi \mathrm{i}k)^m \theta_k \varphi_k(x),$$

whose equivalent model in the Fourier domain is $y_k = \theta_k + \varepsilon\xi_k = (2\pi \mathrm{i}k)^{-m} v_k + \varepsilon\xi_k$ for any $k \in \mathbb{Z}^*$, and the goal is now to estimate $v_k = \theta_k (2\pi \mathrm{i}k)^m$. Thus, the estimation of the mth derivative is a mildly ill-posed inverse problem of degree $\beta = m$.

1.3.2. Deconvolution. Deconvolution of a signal is a standard inverse problem that arises frequently in applications. The problem of circular deconvolution, i.e. periodic on $[a, b]$, is considered in, for example, [14, 15, 21, 39].

As an example, consider a circular deconvolution problem defined by the operator

$$Af(x) = g * f(x) = \int_0^1 g(x - t) f(t)\,\mathrm{d}t, \, x \in [0, 1],$$

where g is a known 1-periodic convolution kernel in $L_2([0, 1])$. Here the SVD basis is clearly the Fourier basis. Let $\{\varphi_k(t)\}$ be the real trigonometric basis on $[0, 1]$:

$$\varphi_1(t) \equiv 1, \qquad \varphi_{2k}(t) = \sqrt{2}\cos(2\pi kt), \qquad \varphi_{2k+1}(t) = \sqrt{2}\sin(2\pi kt), \qquad k = 1, 2, \ldots.$$

The observational model is $Y = g * f + \varepsilon\xi$, and the goal is to recover the unknown function f.

This deconvolution framework is easily extended to two dimensions to model the problem of image deblurring, which has important applications to image processing in general and astronomy in particular (e.g., deblurring of images taken by the Hubble space telescope) [22, 67].

1.3.3. Heat equation. Consider the following boundary-value problem for the heat equation (e.g., [23]):

$$\frac{\partial}{\partial t} u(x, t) = \Delta u(x, t), \qquad u(x, 0) = f(x), \qquad u(0, t) = u(1, t),$$

where $u(x, t)$ is defined on $[0, 1] \times [0, T]$, and the initial condition f is a 1-periodic function in $L^2([0, 1])$. The problem is to determine the initial temperature function f given a noisy version of the temperature distribution $u(x, T)$ at time T: $Y = u(, T) + \varepsilon\xi$, where as usual ξ is a Gaussian white noise process.

The SVD for this problem is given again by the Fourier basis and so $u(x, T)$ may be written as $u(x, T) = \sqrt{2}\sum_{k=1}^{\infty} \theta_k e^{-\pi^2 k^2 T} \sin(k\pi x)$. The singular values are $b_k = e^{-\pi^2 k^2 T/2}$ and thus the problem is severely ill-posed; it is even worse than the severely ill-posed case because $b_k \sim e^{-k^2}$.

1.3.4. Computerized tomography. Computerized tomography is used to image the internal structure of an object using indirect measurements. For example, by measuring x-ray attenuation one obtains cross-sectional images of the object. Computerized tomography has many important applications to medical imaging and remote sensing. A general reference is [53].

The object to be imaged is usually modeled as a function f that describes spatial characteristics of the object (e.g., mass). From the mathematical point of view, the inverse problem corresponds to the recovery of the unknown function f in $\mathbb{R}^d$ based on the observations of its Radon transform Rf. This transform is defined by integrals over hyperplanes [53]: let $U = \{x \in \mathbb{R}^2 : \|x\| \leqslant 1\}$ be the unit disc in $\mathbb{R}^2$. Consider the integrals of a function $f : U \to \mathbb{R}$ over all the lines that intersect U. The lines are parametrized by the length $u \in [0, 1]$ of the perpendicular from the origin to the line and by the angle $\varphi \in [0, 2\pi)$ of this perpendicular from the x-axis. Suppose that the function $f(x_1, x_2)$ belongs to $L^2(U)$. The Radon transform Rf of f is defined as

$$Rf(u, \varphi) = \frac{\pi}{2(1-u^2)^{\frac{1}{2}}} \int_{-\sqrt{1-u^2}}^{\sqrt{1-u^2}} f(u\cos\varphi - t\sin\varphi, u\sin\varphi + t\cos\varphi)\,\mathrm{d}t, \tag{7}$$

where $(u, \varphi) \in S = \{(u, \varphi) : 0 \leqslant u \leqslant 1, 0 \leqslant \varphi < 2\pi\}$. Thus, the Radon transform $Rf(u, \varphi)$ is π times the average of f over the line segment (parametrized by (u, φ)) that intersects U. It is natural to consider Rf as an element of $L^2(S, \mu)$, where μ is the measure $\mathrm{d}\mu(u, \varphi) = 2\pi^{-1}(1-u^2)^{\frac{1}{2}}\,\mathrm{d}u\,\mathrm{d}\varphi$.

The Radon transform is a compact operator. Its SVD basis is known but usually not so easy to compute [40]. Its associated observational model is $Y = Rf + \varepsilon\xi$. The recovery of f given Y is an ill-posed problem of degree $\beta = 1/2$ [9, 15, 53].

2. Nonparametric estimation

The problem of finding inversion estimates of a function f from indirect noisy observations depends on the type of information one has about f. We call the problem nonparametric if f is only known to belong to a nonparametric class of functions. For example, the space of twice continuously differentiable functions. An example of a parametric family is the set of polynomials of degree less than d. The coefficients of the polynomials parametrize the family.

We will present some methods to obtain and assess nonparametric inversion estimates in the framework of the sequence space model (4), which is equivalent to the original model (2). Since the operator is diagonalized, it is usually easier to understand the inverse problem and its estimators in the spectral domain.

2.1. Minimax approach

The minimax approach is well known by statisticians, especially those in the nonparametric community. It is a useful tool to define notions of optimality [36, 42, 57, 61].

Let us return to model (5) and let $\widehat{\theta} = \widehat{\theta}(X) = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots)$ be any estimator of $\theta = (\theta_1, \theta_2, \ldots)$ based on the data $X = \{X_k\}$. Since $f = \sum_k \theta_k \varphi_k$, it is natural to estimate f with $\widehat{f} = \sum_k \widehat{\theta}_k \varphi_k$. The obvious question is to determine how good this estimator is.

Since an estimator is by definition random, one way to assess its quality is by computing the expected value of the squared difference between $\widehat{f}$ and the true f. Define the risk function as the *mean integrated squared error (MISE)* of $\widehat{f}$:

$$\mathcal{R}(\widehat{f}, f) = \mathbf{E}_f \|\widehat{f} - f\|^2 = \mathbf{E}_\theta \sum_{k=1}^{\infty} (\widehat{\theta}_k - \theta_k)^2 = \mathbf{E}_\theta \|\widehat{\theta} - \theta\|^2,$$

where the notation $\|\cdot\|$ stands for the ℓ^2-norm when applied to θ-vectors in the sequence space. Henceforth, $\mathbf{E}_f$ and $\mathbf{E}_\theta$ will denote the expectations w.r.t. Y or $X = (X_1, X_2, \ldots)$ for models (2) and (5), respectively. Analyzing the risk $\mathcal{R}(\widehat{f}, f)$ of the estimator $\widehat{f}$ is equivalent to analyzing the corresponding sequence space risk $\mathbf{E}_\theta \|\widehat{\theta} - \theta\|^2$.

Ideally, one would like to find an estimator that minimizes the MISE. However, the risk of an estimator depends, by definition, on the unknown f or θ. So we try instead to minimize an estimate of MISE or to determine an overall measure of the risk such as the minimax risk.

Definition 2. *Suppose we know a priori that f belongs to some class of functions $\mathcal{F}$. The minimax risk on $\mathcal{F}$ is defined as*

$$r_\varepsilon(\mathcal{F}) = \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathcal{R}(\widehat{f}, f),$$

where the infimum is determined over the set of all estimators of f (measurable functions of Y).

In contrast to the parametric case, in nonparametric estimation it is usually not possible to find estimators that attain the minimax risk. One reason for this is the paucity of requirements made on the function f (e.g., f is only assumed to be in a smooth class of functions). A more natural approach is to consider asymptotic properties as the noise level decreases to zero ($\varepsilon \to 0$). As mentioned in the introduction, in the Gaussian white noise model equivalence this is the same as the $n \to \infty$ in the discrete sample model.

It is clear that for any estimator $\widehat{f}_\star$

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathcal{R}(\widehat{f}, f) \leqslant \sup_{f \in \mathcal{F}} \mathcal{R}(\widehat{f}_\star, f).$$

We will call the estimator optimal if we can bracket the above inequality with bounds that decrease to zero with the noise level. More precisely, the definition follows.

Definition 3. *An estimator $\widehat{f}_\star$ is said to be optimal or to attain the optimal rate of convergence v_ε if the positive sequence v_ε converges to zero as $\varepsilon \to 0$ and there are constants $0 < C_2 \leqslant C_1 < \infty$ such that*

$$C_2 v_\varepsilon \leqslant \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathcal{R}(\widehat{f}, f) \leqslant \sup_{f \in \mathcal{F}} \mathcal{R}(\widehat{f}_\star, f) \leqslant C_1 v_\varepsilon$$

as $\varepsilon \to 0$. The estimator $\widehat{f}_\star$ is said to be minimax or to attain the exact constant if $C_1 = C_2$.

An optimal estimator is thus an estimator whose risk is comparable to that of the best possible estimator. Theorem 1 provides examples of optimal estimators.

2.2. Classes of functions

An important problem now is to define 'natural' classes of functions $\mathcal{F}$ that capture the prior information known about the function f (e.g., smoothness). Under regularity conditions, the characteristics of the functions can be translated into properties of the coefficients $\{\theta_k\}$.

Assume that f belongs to the functional class corresponding to ellipsoids Θ in the space of coefficients $\{\theta_k\}$:

$$\Theta = \Theta(a, L) = \left\{ \theta : \sum_{k=1}^{\infty} a_k^2 \theta_k^2 \leqslant L \right\},$$

where $a = \{a_k\}$ is a non-negative sequence that tends to infinity with k, and $L > 0$. This means that for large values of k the coefficients θ_k will be decreasing with k and thus will be small for large k.

Remark 5. The ellipsoid is one of the most natural assumptions as it defines an ℓ^2-behavior of the coefficients. However, more general classes such as ℓ^p-bodies or hyperrectangles may also be used [20].

In the special case where the SVD basis is the Fourier basis, hypotheses on $\{\theta_k\}$ can be precisely written in terms of the smoothness of f. Such classes arise naturally in various inverse problems; they include as special cases the Sobolev classes [68] and classes of analytic functions [37].

For example, consider the circular deconvolution problem of section 1.3.2. Define the *Sobolev* class

$$\mathcal{W}(\alpha, L) = \left\{ f = \sum_{k=1}^{\infty} \theta_k \varphi_k : \theta \in \Theta(\alpha, L) \right\},$$

where $\Theta(\alpha, L) = \Theta(a, L)$ with the sequence $a = \{a_k\}$ such that $a_1 = 0$ and

$$a_k = \begin{cases} (k-1)^\alpha & \text{for } k \text{ odd}, \\ k^\alpha & \text{for } k \text{ even}, \end{cases} \qquad k = 2, 3, \ldots,$$

for $\alpha > 0$, $L > 0$. When α is an integer, this has an equivalent definition in terms of f:

$$\mathcal{W}(\alpha, L) = \left\{ f : \int_0^1 [f^{(\alpha)}(t)]^2 \mathrm{d}t \leqslant \pi^{2\alpha} L \right\},$$

where $f^{(\alpha)}$ denotes the weak derivative of f of order α.

A second choice is the class of *analytic functions* (functions that admit an analytic continuation into a band of the complex plane)

$$\mathcal{A}(\alpha, L) = \left\{ f = \sum_{k=1}^{\infty} \theta_k \varphi_k : \theta \in \Theta(\alpha, L) \right\},$$

where $a_k = \exp(\alpha k)$, $\alpha > 0$ and $L > 0$. Functions in this class are very smooth; they are often used in the context of inverse problems for partial differential equations (as in section 1.3.3). Indeed, in this kind of severely ill-posed problems the functions are often required to be very smooth in order to get reasonable reconstructions (see below).

2.3. Rates of convergence

Some important results have been obtained in the framework of ill-posed inverse problems with compact operators and functions with coefficients in some ellipsoid. For example, optimal rates of convergence have been proved [3, 15]. There is also a famous result by [56] which proves the existence of a minimax estimator (see also [54]). That is, an estimator that attains not only the optimal rate but also the exact constant. This result has been generalized to the case of severely ill-posed problems with analytic functions [11, 28]. In [11, 27], the authors study the problem of the heat equation (see section 1.3.3) with analytic functions.

Table 1. Rates of convergence for three types of inverse problems: direct, mildly ill-posed and severely ill-posed. The two columns refer to the class of functions. The parameter β is defined by the problem while α is defined by the class.

	Sobolev	Analytic
Direct problem	$\varepsilon^{\frac{4\alpha}{2\alpha+1}}$	$\varepsilon^2(\log\frac{1}{\varepsilon})$
Mildly ill-posed	$\varepsilon^{\frac{4\alpha}{2\alpha+2\beta+1}}$	$\varepsilon^2(\log\frac{1}{\varepsilon})^{2\beta+1}$
Severely ill-posed	$(\log\frac{1}{\varepsilon})^{-2\alpha}$	$\varepsilon^{\frac{4\alpha}{2\alpha+2\beta}}$

Table 1 shows some optimal rates of convergence for the sequence space model (5) when the SVD is the Fourier basis and the class of functions is either Sobolev or analytic. We see that the rates of convergence usually depend on the smoothness α of the function f and on the degree of ill-posedness β. When β increases the rates decrease. In the direct model we get the standard rates for nonparametric estimation, slower than ε^2, which is the parametric rate. In the standard cases of mildly ill-posed problems with smooth functions or severely ill-posed problems with very smooth functions, the rates are polynomial in ε. For the severely ill-posed case with functions that are only Sobolev smooth the rate is logarithmic and very slow (the rate depends on β only through the constant). For mildly ill-posed problems with very smooth functions the rate is almost the parametric rate ε^2.

2.4. Regularization methods

The term regularization in inverse problems usually refers to the ways for obtaining meaningful solutions from ill-posed inverse problems. For statisticians regularization is just a family of methods to define reasonable estimators that resolve unidentifiability and instability problems. In image processing, the same tools are just called image reconstruction methods (see, e.g. [22, 24, 41, 55, 63, 67]).

We provide some examples of regularization methods or estimators that are commonly used. These methods will be defined in the spectral domain even if some may be computed without the spectrum. The reason is that the theoretical properties of many regularization methods can be more easily studied in the spectral domain. However, some methods (e.g., Tikhonov regularization and Landweber iteration) are not computed using the SVD but via more computationally efficient techniques [22, 67].

Let $\lambda = (\lambda_1, \lambda_2, \ldots)$ be a sequence of non-random weights. Every sequence λ defines a *linear estimator* $\widehat{\theta}(\lambda) = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots)$, where X_k is as in (5), $\widehat{\theta}_k = \lambda_k X_k$ and

$$\widehat{f}(\lambda) = \sum_{k=1}^{\infty} \widehat{\theta}_k \varphi_k. \tag{8}$$

Examples of commonly used weights λ_k are the projection weights $\lambda_k = I(k \leqslant N)$, where $I(A)$ denotes the usual indicator function of the set A. These weights define the *projection estimator* (also called *truncated SVD* or *spectral cut-off*)

$$\widehat{\theta}(N) = \begin{cases} X_k, & k \leqslant N, \\ 0, & k > N \end{cases}$$

(see, for example [66, 67]). The value N is usually called the *bandwidth*.

The truncated SVD is a natural estimator and is often used as a benchmark as it attains optimal rates of convergence [5]. However, it is a rather rough estimator. Indeed, the weights

can only take the values 0 and 1, i.e., one estimates θ_k by X_k or 0. Also, the estimator is not computationally efficient as it requires computation of the SVD and all the coefficients X_k.

Another popular choice is the well-known *Tikhonov regularization* method [65]. The basic idea is to minimize a data misfit while controlling the regularity of the function. This is achieved by finding an estimator that minimizes the functional

$$\min_g \{\|Ag - Y\|^2 + \gamma \|g\|^2\}, \tag{9}$$

where $\gamma > 0$ is a crucial tuning parameter that controls the balance. We will consider the selection of γ in section 3.3.

Under mild regularity conditions, the minimizer of (9) is $\widehat{f}_\gamma = (A^*A + \gamma I)^{-1}A^*Y$. These estimators may also be defined in the spectral domain using the weights $\lambda_k = 1/\left(1 + \gamma\sigma_k^2\right)$ with $\gamma > 0$. The estimator is then defined as in (8).

Remark 6. There have been many generalizations of Tikhonov regularization: for example to have a starting point g_0 [65], to use an iterated method [22] or to use a Hilbert scales approach, i.e., with a term different from $\|g\|^2$ in (9), for example with $\|Dg\|^2$ instead [51]. See also the general references for regularization given above.

The L^2-risk of any linear estimator (8) with weights λ is defined as

$$\begin{aligned}\mathcal{R}(\widehat{f}(\lambda), f) = R(\theta, \lambda) &= \mathbf{E}_\theta \sum_k (\widehat{\theta}_k(\lambda) - \theta_k)^2 \\ &= \sum_{k=1}^{\infty} (1 - \lambda_k)^2 \theta_k^2 + \varepsilon^2 \sum_{k=1}^{\infty} \sigma_k^2 \lambda_k^2. \end{aligned} \tag{10}$$

As usual, the first sum on the right-hand side is the squared *bias* while the second one is the *variance*. The bias term is related to the approximation error and measures if the estimator provides a good approximation of the unknown f even in the absence of noise. The variance, on the other hand, measures the variability of the inversion estimate introduced by the random noise. Ideally, bias and variance should be small.

Because of their simplicity, we will mainly consider projection estimators to explain the main ideas behind oracle inequalities, minimax risk and the risk hull method. We start with their risk: the risk of a projection estimator of bandwidth N is

$$R(\theta, N) = \sum_{k=N+1}^{\infty} \theta_k^2 + \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2.$$

In this case, the bias–variance decomposition is very simple. Indeed, we estimate the first N coefficients with their empirical versions X_k while the estimates of the rest of the coefficients are set to zero. Thus, the bias measures the influence of the coefficients θ_k with $k > N$, and the stochastic term is due to the random noise in the first N coefficients.

It is evident that the selection of the bandwidth N is crucial. This is a common but important problem in nonparametric statistics: the estimator depends on tuning parameters that have to be chosen to balance bias and variance but we have the problem that the bias depends on the unknown f. Thus, the bias cannot be given exactly but one can often find bounds for the risk, as in the following theorem stated in the framework of model (5).

Theorem 1. *Consider the case where $\sigma_k = k^\beta$ and θ belongs to the ellipsoid $\Theta(\alpha, L)$, where $a_k = k^\alpha$. Then the projection estimator with $N^\star \sim \varepsilon^{-2/(2\alpha+2\beta+1)}$ verifies as $\varepsilon \to 0$*

$$\sup_{\theta \in \Theta(\alpha, L)} R(\theta, N^\star) \leqslant C\varepsilon^{4\alpha/(2\alpha+2\beta+1)}.$$

This rate may be shown to be optimal (see, for example [56]).

Proof. By definition

$$\sup_{\theta\in\Theta(\alpha,L)} R(\theta, N) = \sup_{\theta\in\Theta(\alpha,L)} \sum_{k=N+1}^{\infty} \theta_k^2 + \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2.$$

Note that

$$\sup_{\theta\in\Theta(\alpha,L)} \sum_{k=N+1}^{\infty} \theta_k^2 \leqslant \sup_{\theta\in\Theta(\alpha,L)} \sum_{k=N+1}^{\infty} k^{2\alpha}\theta_k^2 k^{-2\alpha} \leqslant N^{-2\alpha} \sup_{\theta\in\Theta(\alpha,L)} \sum_{k=1}^{\infty} k^{2\alpha}\theta_k^2 \leqslant LN^{-2\alpha}.$$

The variance is controlled by $\varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 = \varepsilon^2 \sum_{k=1}^{N} k^{2\beta} \approx \frac{\varepsilon^2 N^{2\beta+1}}{2\beta+1}$ for large N. Thus,

$$\sup_{\theta\in\Theta(\alpha,L)} R(\theta, N) \leqslant LN^{-2\alpha} + \frac{\varepsilon^2 N^{2\beta+1}}{2\beta+1}.$$

To attain the optimal rate of convergence, N has to be of order $\varepsilon^{-2/(2\alpha+2\beta+1)}$ as $\varepsilon \to 0$. This choice reflects the trade-off between bias and variance of the estimator. One can see this effect in simulation studies: too large an N will overfit the data while too small an N will underfit. □

There is an optimal choice of N that balances the bias and variance in a minimax sense but it depends very precisely on the smoothness α and the degree of ill-posedness β of the operator. And even in the case where the operator A (and thus its degree β) is precisely known, it is unreasonable to assume that we know precisely the smoothness of the unknown function f. We are thus led to the notions of adaptation and oracle inequalities to answer the question of how to choose the bandwidth without making strong assumptions on f.

3. Adaptation and oracle inequalities

One of the most important problems in nonparametric statistics is the data calibration of the tuning parameters (e.g., N or γ) for a chosen class of estimators. For example, we have seen in theorem 1 that this choice is crucial to attain optimal rates of convergence.

3.1. Minimax adaptive estimation

The starting point of *minimax adaptation* is a collection $\mathcal{A} = \{\Theta_\alpha\}$ of classes $\Theta_\alpha \subset \ell_2$. The statistician knows that θ belongs to some member Θ_α of the collection $\mathcal{A}$ but does not know exactly which. When Θ_α is a smoothness class this assumption can be interpreted as follows: the underlying function is known to be smooth but its degree of smoothness is unknown. The notion of minimax adaptivity has been developed to define estimators that 'adapt' to the unknown smoothness of the function [43].

Definition 4. *An estimator $\theta^\star$ is called minimax adaptive on the scale of classes $\mathcal{A}$ if for every $\Theta_\alpha \in \mathcal{A}$ the estimator $\theta^\star$ attains the optimal rate of convergence.*

Minimax adaptive estimators play an important role in nonparametric statistics from both theoretical and practical points of view. Indeed, these estimators are optimal for any parameter α in the collection $\mathcal{A}$. From a more practical point of view, minimax adaptivity guarantees good accuracy of the estimator for a wide choice of functions. Thus, the estimator automatically 'adapts' to the unknown smoothness of the underlying function (see, for example [44]).

Remark 7. Adaptive estimation is one of the main tools from statistics which is used even by the deterministic inverse problems community. There also exist some deterministic techniques of adaptation but the statistical ones are definitely rather popular now. For example, quite recently, several papers used the Lepski procedure of adaptation (see [2, 25, 50, 58]).

3.2. Oracle inequalities

The use of oracle inequalities leads to another approach related to minimax adaptive estimation but not equivalent. The starting point is different as this time the goal is to choose the best possible estimator from a given family of estimators. In the minimax approach, on the other hand, one tries to obtain the best accuracy over all functions in some fixed class. Moreover, as we will see, oracle inequalities are not asymptotic.

Assume a fixed class of estimators. For example, the class of possible weights Λ for the estimators defined by (8). Define the *oracle* λ^0 as

$$R(\theta, \lambda^0) = \inf_{\lambda \in \Lambda} R(\theta, \lambda). \tag{11}$$

The oracle corresponds to the best possible choice in Λ. That is, the one that minimizes the risk. However, this is not an estimator because the risk and the oracle depend on the unknown θ. This is the reason it is called an 'oracle'; it is the best one in the family but it requires knowing the true θ.

The goal is then to find a data-driven sequence of weights $\lambda^\star$ with values in Λ such that the estimator $\theta^\star = \widehat{\theta}(\lambda^\star)$ satisfies the following *oracle inequality* for any $\theta \in \ell^2$:

$$\mathbf{E}_\theta \|\theta^\star - \theta\|^2 \leqslant C_\varepsilon \inf_{\lambda \in \Lambda} R(\theta, \lambda) + \Delta_\varepsilon, \tag{12}$$

where Δ_ε is some positive tolerance term and $C_\varepsilon \geqslant 1$ (the closer to 1 the better it is). If Δ_ε is small (i.e., smaller than $R(\theta, \lambda^0)$), then an oracle inequality guarantees that the estimator has a risk of the same order as that of the oracle. The goal is thus to find data-driven methods that provide an automatic choice of λ and more or less mimic the oracle.

The oracle approach is often used to obtain adaptive estimators. For example, the best estimator in a given class often attains the optimal rate of convergence. On the other hand, minimax theory provides in some sense a justification for the use of oracle inequalities. Indeed, the best possible choice in a given family of estimators is not always satisfactory. Minimax results show that, with a good choice of the tuning parameter, a given family provides optimal estimators.

One may obtain asymptotic results as $\varepsilon \to 0$: an estimator $\theta^\star$ is said to satisfy an *exact oracle inequality* on the class Λ as $\varepsilon \to 0$ if

$$\mathbf{E}_\theta \|\theta^\star - \theta\|^2 \leqslant (1 + o(1)) R(\theta, \lambda^0), \tag{13}$$

for every θ within some large subset $\Theta_0 \subseteq \ell_2$. In other words, the estimator $\theta^\star$ precisely mimics the oracle on Λ for any sequence $\theta \in \Theta_0$.

In the oracle approach, $\lambda^\star$ is usually restricted to take its values in same class Λ that appears in the rhs of (12). A *model selection* interpretation of (12) is the following: in a given class of models Λ, we pick the model $\lambda^\star$ that is the closest to the true parameter θ in terms of the risk $R(\theta, \lambda)$.

3.3. Model selection and risk estimation

The problem of model selection is widespread in statistics and may have different meanings depending on the particular application. We consider model selection as the problem of choosing the 'best' estimator from a family Λ. This selection should be completely data-driven.

3.3.1. Unbiased risk estimation. Since the true parameter θ is unknown, so is the risk. Hence it is natural to look for an estimator that minimizes an estimate of the risk based on the available data. A classical approach to this minimization problem is based on the principle of *unbiased risk estimation* (URE) [60]. The idea of using URE to define a data-driven bandwidth choice goes back to [1, 47]. Originally, the URE was proposed in the context of the regression and cross-validation techniques but nowadays it is used as a basic adaptation tool for many statistical models.

For inverse problems this method was studied in [12] (see also [55] for a discrete sample version), where exact oracle inequalities were obtained. In the framework of linear estimators $\widehat{\theta}(\lambda)$, Stein's risk estimator is

$$U(X,\lambda) = \sum_{k=1}^{\infty}(1-\lambda_k)^2\left(X_k^2 - \varepsilon^2\sigma_k^2\right) + \varepsilon^2\sum_{k=1}^{\infty}\sigma_k^2\lambda_k^2. \tag{14}$$

It is an unbiased estimator of $R(\theta,\lambda)$. That is, $R(\theta,\lambda) = \mathbf{E}_\theta U(X,\lambda)$ for all λ. Note that (14) is obtained by simply using $X_k^2 - \varepsilon^2\sigma_k^2$ as an unbiased estimator of θ_k^2 in (10).

The principle of unbiased risk estimation is then to find the λ that minimizes the risk estimate $U(X,\lambda)$:

$$\lambda^\star = \arg\min_{\lambda\in\Lambda} U(X,\lambda). \tag{15}$$

We now provide an example of oracle inequality obtained when URE is used to estimate N for projection estimators or γ for Tikhonov regularization. Define $S = \left(\max_{\lambda\in\Lambda}\sum_{k=1}^{\infty}\sigma_k^4\lambda_k^2 \big/ \min_{\lambda\in\Lambda}\sum_{k=1}^{\infty}\sigma_k^4\lambda_k^2\right)^{1/2}$. We have the following theorem whose proof can be found in [12].

Theorem 2. *Assume that $\sigma_k = k^\beta$, Λ is finite with cardinality D and the estimator is $\theta^\star = (\theta_1^\star, \theta_2^\star, \ldots)$ with $\theta_k^\star = \lambda_k^\star X_k$. Then under weak conditions (see [12]) there are constants $\gamma_1, \gamma_2 > 0$ such that*

$$\mathbf{E}_\theta\|\theta^\star - \theta\|^2 \leqslant (1+\gamma_1 B^{-1})\min_{\lambda\in\Lambda} R(\theta,\lambda) + \gamma_2 B\varepsilon^2[\log(DS)]^{2\beta+1} \tag{16}$$

for every $\theta \in \ell_2$ and for any $B > 0$ large enough.

In fact by imposing mild restrictions on the behavior of D and S for large ε, one may obtain an exact oracle inequality. This result means, for example, that we can precisely mimic the Tikhonov method with the best possible choice of tuning parameter γ.

Remark 8. One of the problems with adaptation or oracle results is that the data-driven choices of the parameters make the risk of the estimator very difficult to control because the tuning parameter also depends on the same data.

3.4. Risk hull method

Even if one can find precise oracle inequalities as in theorem 2, the URE method is in fact not very satisfying in simulation studies [10] because for some samples the estimator is not precise at all. Thus, the mean risk over all the samples is usually too large. This leads us to the search of more stable data-driven choices of the parameters. To present the main ideas we consider only the class of projection estimators.

The unbiased risk estimator for projection estimators is

$$U(X,N) = \sum_{k=N+1}^{\infty}\left(X_k^2 - \varepsilon^2\sigma_k^2\right) + \varepsilon^2\sum_{k=1}^{N}\sigma_k^2.$$

Minimizing $U(X, N)$ over N is equivalent to minimizing

$$\bar{R}(X, N) = -\sum_{k=N+1}^{\infty} \left(X_k^2 - \varepsilon^2\sigma_k^2\right) + \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 - \|X - \varepsilon\sigma\|^2$$

over N, which is in turn equivalent to minimizing

$$\bar{R}(X, N) = -\sum_{k=1}^{N} X_k^2 + 2\varepsilon^2 \sum_{k=1}^{N} \sigma_k^2. \tag{17}$$

The method of *penalized empirical risk* is a general approach that is very similar to URE. We first define a 'penalized' version of (17):

$$\bar{R}_{\text{pen}}(X, N) = -\sum_{k=1}^{N} X_k^2 + \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 + \text{pen}(N),$$

where $\text{pen}(N)$ is a penalty function. The bandwidth choice is then defined as

$$N(X) = \arg\min_{N \geqslant 1} \bar{R}_{\text{pen}}(X, N). \tag{18}$$

For example, the URE criterion corresponds to a penalized empirical risk with a specific penalty called the URE penalty $\text{pen}_{\text{ure}}(N) = \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2$.

The modern literature on penalized empirical risk is vast; we refer interested reader to [6] or [45] in the inverse problems framework. The main idea at the heart of this approach is that severe penalties may lead to substantial improvements over URE. The problem is that the selection of the penalty function is crucial, especially for inverse problems. The *risk hull minimization* (RHM) method proposed in [10] improves over URE by providing a relatively good strategy for selecting the penalty function (see also [26, 29]).

The heuristic motivation of the RHM approach is based on the oracle ideology. Suppose there is an oracle that gives us $\theta_k, k = 1, \ldots$, but we are only allowed to use the projection method. In this case, the optimal bandwidth is evidently given by $N_{or} = \arg\min_N r(X, N)$, where

$$r(X, N) = \|\widehat{\theta}(N) - \theta\|^2 = \sum_{k=N+1}^{\infty} \theta_k^2 + \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 \xi_k^2.$$

Let us try to mimic this bandwidth choice with a data-driven procedure. At first glance this problem seems hopeless because neither θ_k^2 nor ξ_k^2 are known. However, suppose for a moment that we know all θ_k^2 and try to minimize $r(X, N)$. Since ξ_k^2 are assumed to be unknown, we can use a conservative minimization: we minimize the following non-random functional

$$l(\theta, N) = \sum_{k=N+1}^{\infty} \theta_k^2 + V(N), \tag{19}$$

where $V(N)$ bounds from above the stochastic term $\varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 \xi_k^2$. It seems natural to choose $V(N)$ such that

$$\mathbf{E} \sup_N \left[\varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 \xi_k^2 - V(N)\right] \leqslant 0, \tag{20}$$

because then the risk of any projection estimator with any data-driven bandwidth $\tilde{N}$ can be easily controlled:

$$\mathbf{E}_\theta \|\widehat{\theta}(\tilde{N}) - \theta\|^2 \leqslant \mathbf{E}_\theta l(\theta, \tilde{N}). \tag{21}$$

This argument leads to the following definition: a non-random function $\ell(\theta, N)$ is called a *risk hull* if

$$\mathbf{E}_\theta \sup_N [r(X, N) - \ell(\theta, N)] \leqslant 0.$$

For example, $l(\theta, N)$ as defined by (19) and (20) is a risk hull. Evidently, the upper bound (21) should be as small as possible. So we look for the minimal hull (note that this hull strongly depends on σ_k^2).

Once a function $V(N)$ satisfying (20) has been chosen, the minimization of $l(\theta, N)$ can be done using standard unbiased estimation. The problem reduces to the minimization of $-\sum_{k=1}^N \theta_k^2 + V(N)$. As before, by replacing the unknown θ_k^2 with their unbiased estimates $X_k^2 - \varepsilon^2\sigma_k^2$, we arrive at the following adaptive bandwidth choice:

$$\bar{N} = \arg\min_N \left[-\sum_{k=1}^N X_k^2 + \varepsilon^2 \sum_{k=1}^N \sigma_k^2 + V(N) \right].$$

In the framework of the empirical risk minimization, the RHM can be defined as follows. Let the penalty in (18) be defined as

$$\mathrm{pen}(N) = \mathrm{pen}_{\mathrm{rhm}}(N) = \varepsilon^2 \sum_{k=1}^N \sigma_k^2 + (1+\alpha)U_0(N), \tag{22}$$

where $\alpha > 0$ is a fixed constant (not crucial) and

$$U_0(N) = \inf\left\{ t > 0 : \mathbf{E}\left(\eta_N I(\eta_N \geqslant t)\right) \leqslant \varepsilon^2\sigma_1^2 \right\}, \tag{23}$$

with $\eta_N = \varepsilon^2 \sum_{k=1}^N \sigma_k^2 \left(\xi_k^2 - 1\right)$. The function $U_0(N)$ may be computed by Monte Carlo simulations. The RHM chooses the bandwidth N_{rhm} according to (18) with the penalty function defined by (22) and (23).

The RHM penalty corresponds to the URE penalty plus the term $(1+\alpha)U_0(N)$. As $N \to \infty$, we have

$$U_0(N) \approx \left(2\varepsilon^4 \sum_{k=1}^N \sigma_k^4 \log\left(\frac{\sum_{k=1}^N \sigma_k^4}{2\pi\sigma_1^4} \right) \right)^{1/2}.$$

In particular, $U_0(N) = o\left(\varepsilon^2 \sum_{k=1}^N \sigma_k^2\right)$ as $N \to \infty$. Therefore, asymptotically there is no real difference between the URE and the RHM. But non-asymptotically there are important differences. For example, in applications to ill-posed inverse problems the RHM is much more stable than the URE. Indeed, simulation studies reveal that the RHM is accurate for all the samples, resulting in smaller risks than those of the URE. One reason for this instability is that the URE is based on asymptotic ideas but in many inverse problems N may not be very large. One has to be careful with the asymptotics because of the ill-posedness. Indeed, if a data-driven choice $\tilde{N}$ is larger than the choice of the oracle, then the risk of the estimator may really become large.

The following oracle inequality (whose proof can be found in [10]) provides an upper bound for the mean square risk of the RHM approach. It is assumed that σ_k has a polynomial growth ($\sigma_k = k^\beta$).

Theorem 3. *[10] Let the RHM bandwidth choice N_{rhm} be defined by (18) with the penalty function defined by (22) and (23) and $\theta^\star_{\mathrm{rhm}}$ be the associated projection estimator. Then there are constants $C_* > 0$ and $\delta_0 > 0$ such that for all $\delta \in (0, \delta_0]$ and $\alpha > 1$*

$$\mathbf{E}\|\tilde{\theta}^\star_{\mathrm{rhm}} - \theta\|^2 \leqslant (1+\delta) \inf_N R_{\mathrm{rhm}}(\theta, N) + C_*\varepsilon^2 \left(\frac{1}{\delta^{4\beta+1}} + \frac{1}{\alpha - 1} \right), \tag{24}$$

where

$$R_{\mathrm{rhm}}(\theta, N) = \sum_{k=N+1}^{\infty} \theta_k^2 + \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 + (1+\alpha)U_0(N).$$

Hence we have an oracle inequality but with a penalty term on the risk in the rhs.

The RHM leads to an explicit penalty defined in the proof of theorem 3. Furthermore, this penalty may be used directly; there is no need for empirical calibration through simulations. Further discussion on the risk hull method and numerical results can be found in [10].

4. Summary and discussion

We have considered the statistical approach to ill-posed inverse problems (which is not the standard framework of [64] where the error is modeled as deterministic) and presented some basic theoretical ideas in the context of the white noise model. This model is discretized in the spectral domain using the SVD. The SVD also allows us to define a measure of the ill-posedness of an inverse problem. This measure plays a role in the computation of rates of convergence of the inversion estimates.

To define estimators we used the nonparametric approach so as to make few assumptions on the unknown function. The performance of the estimators is measured by a risk function defined as the mean integrated square error. This risk is made up of bias and variance components. Assessment of the bias is particularly difficult because it depends on the function we want to estimate. One way to proceed then is to consider the worst risk when the function is known to be in a particular class and to look for estimators that minimize such worst risk (minimax estimators). But in nonparametric inverse problems minimax estimators may not exist. Hence we study instead an asymptotic minimax optimality inequality as the noise level goes to zero. These results depend of course on the particular chosen class of functions. We have provided the examples of ellipsoid classes of smooth functions, Sobolev and the class of analytic functions that consist of 'very smooth' functions.

We also considered the question of defining good estimators for ill-posed inverse problems (regularization). The common problem is the dependence of the estimators on tuning parameters such as truncation level of the SVD or the regularization parameter for Tikhonov. Minimax optimality criteria lead to the values with optimal rates of convergence of the estimators. But these optimal parameters are unachievable as they depend on exact knowledge of function smoothness and ill-posedness of the operator. Since the goal is to find data-driven choices of the parameters (adaptive methods), we discussed oracle inequalities and risk estimation.

The oracle gives us the best choice provided we knew the unknown function; it is therefore used as a benchmark. Oracle inequalities are used to make sure the estimators are close to the choice of the oracle. We provided examples of parameter choices based on minimizing risk estimates that lead to (penalized) oracle inequalities.

The idea of minimizing an estimate of the risk is natural. As an example we discussed Stein's unbiased risk estimator (URE). Another well-known application of this estimator is for thresholding of wavelet coefficients [19]. For projection estimators, the idea of estimating the truncation level by minimizing URE is a particular case of the method of penalized empirical risk. This approach performs better than URE provided the penalty function is chosen appropriately. The risk hull method provides one way to find a good penalty function that is guaranteed to improve on URE.

Our discussion was restricted to models based on (2), which is restrictive and strongly linked to the spectral approach based on the SVD. There are many different generalizations of the methods we have discussed.

Noisy operators. In some applications, the forward operator A is only partially known. For example, in astronomical observations the point spread function may be changing due to unknown physical conditions. This very important topic has been the subject of some recent works [13, 14, 35, 48].

RHM for other methods. There are other regularization methods (e.g., the iterative methods, Landweber and ν-methods to name a few) that can attain optimal rates of convergence [5] and have very good properties from a numerical point of view. For example, in a very recent paper, the RHM approach has been extended to these families of estimators [49].

Nondiagonal case. We have intensively used the SVD to diagonalize the operators. Thus, all the methods have been presented for the spectral approach. However, there is a more general situation where the operator cannot be represented by a diagonal matrix. In this case, one may obtain some close results such as optimal rates of convergence and adaptive estimation (see, for example [48, 52]).

Wavelets and sparsity. This framework is strongly related to the previous nondiagonal case. Indeed, wavelet bases have been shown to almost diagonalize many operators and have good adaptability properties. Wavelet thresholding has been used to construct nonlinear adaptive estimators [14, 16, 18, 19, 20, 35, 39]. Wavelets can be used to model functions that may not be very smooth and thus allows the use of the Besov instead of the Sobolev classes. Higher-dimensional wavelet-like bases such as curvelets have also been developed [8].

Nonlinear operators. Inverse problems with nonlinear operators are much more difficult. This framework has been intensively studied in the deterministic context but is not yet well understood in statistics [22]. Some recent papers concerning nonlinear inverse problems are [4, 45].

Acknowledgments

I would like to thank Luis Tenorio for his help in the writing of this manuscript. I would also like to thank the two referees for their interesting comments.

References

[1] Akaike H 1973 Information theory and an extension of the maximum likelihood principle *Proc. 2nd Int. Symp. Inf. Theory* ed P N Petrov and F Csaki (Budapest) p 267
[2] Bauer F and Hohage T 2005 A Lepskij-type stopping rule for regularized Newton methods *Inverse Problems* **21** 1975
[3] Belitser E N and Levit B Y 1995 On minimax filtering on ellipsoids *Math. Methods Stat.* **4** 259
[4] Bissantz N, Hohage T and Munk A 2004 Consistency and rates of convergence for nonlinear Tikhonov regularization with random noise *Inverse Problems* **20** 1773
[5] Bissantz N, Hohage T, Munk A and Ruymgaart F 2007 Convergence rates of general regularization methods for statistical inverse problems and applications *SIAM J. Numer. Anal.* at press
[6] Birgé L and Massart P 2001 Gaussian model selection *J. Eur. Math. Soc.* **3** 203
[7] Brown L D and Low M G 1996 Asymptotic equivalence of nonparametric regression and white noise *Ann. Stat.* **24** 2384

[8] Candès E J and Donoho D L 2002 Recovering edges in ill-posed inverse problems. Optimality of curvelet frames *Ann. Stat.* **30** 784
[9] Cavalier L 1998 Asymptotically efficient estimation in a problem related to tomography *Math. Methods Stat.* **7** 445
[10] Cavalier L and Golubev G K 2006 Risk hull method and regularization by projections of ill-posed inverse problems *Ann. Stat.* **34** 1653
[11] Cavalier L, Golubev G K, Lepskii O and Tsybakov A B 2004 Block thresholding and sharp adaptive estimation in severely ill-posed inverse problems *Theory Prob. Appl.* **48** 426
[12] Cavalier L, Golubev G K, Picard D and Tsybakov A B 2002 Oracle inequalities in inverse problems *Ann. Stat.* **30** 843
[13] Cavalier L and Hengartner N 2005 Adaptive estimation for inverse problems with noisy operators *Inverse Problems* **21** 1345
[14] Cavalier L and Raimondo M 2007 Wavelet deconvolution with noisy eigenvalues *IEEE Trans. Signal Proc.* **55** 2414
[15] Cavalier L and Tsybakov A B 2002 Sharp adaptation for inverse problems with random noise *Probab. Theory Relat. Fields* **123** 323
[16] Cohen A, Hoffmann M and Reiss M 2004 Adaptive wavelet Galerkin method for linear inverse problems *SIAM J. Numer. Anal.* **42** 1479
[17] Donoho D L 1994 Statistical estimation and optimal recovery *Ann. Stat.* **22** 238
[18] Donoho D L 1995 Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition *Appl. Comput. Harmon. Anal.* **2** 101
[19] Donoho D L and Johnstone I M 1994 Ideal spatial adaptation via wavelet shrinkage *Biometrika* **81** 425
[20] Donoho D L and Johnstone I M 1998 Minimax estimation via wavelet shrinkage *Ann. Stat.* **26** 879
[21] Efromovich S 1997 Robust and efficient recovery of a signal passed through a filter and then contaminated by non-Gaussian noise *IEEE Trans. Inform. Theory* **43** 1184
[22] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer)
[23] Engl H W and Rundell W 1995 *Inverse problems in diffusion processes* (Philadelphia: SIAM)
[24] Evans S N and Stark P B 2002 Inverse problems as statistics *Inverse Problems* **18** R55
[25] Goldenshluger A and Pereverzev S V 2000 Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations *Probab. Theory Relat. Fields* **118** 169
[26] Golubev Y 2004 The principle of penalized empirical risk in severely ill-posed problems *Theory Probab. Appl.* **130** 18
[27] Golubev G K and Khasminskii R Z 1999 Statistical approach to some inverse boundary problems for partial differential equations *Problems Inform. Transm.* **35** 51
[28] Golubev G K and Khasminskii R Z 2001 Statistical approach to Cauchy problem for Laplace equation *State of the Art in Probability and Statistics, Festschrift for W.R. van Zwet (IMS Lecture Notes Monograph Series* vol 36*)* ed M de Gunst, C Klaassen and A van der Vaart p 419
[29] Golubev Y and Levit B 2004 An oracle approach to adaptive estimation of linear functionals in a Gaussian model *Math. Methods Stat.* **13** 392
[30] Grama I and Nussbaum M 2002 Asymptotic equivalence for nonparametric regression *Math. Methods Stat.* **11** 1
[31] Groetsch C W 1993 *Inverse Problems in the Mathematical Sciences* (Braunschweig: Vieweg)
[32] Hall P and Horowitz J L 2005 Nonparametric methods for inference in the presence of instrumental variables *Ann. Stat.* **33** 2904
[33] Halmos P R 1963 What does the spectral theorem say? *Am. Math. Monthly* **70** 241
[34] Hida T 1980 *Brownian Motion* (New York: Springer)
[35] Hoffmann M and Reiss M 2007 Nonlinear estimation for linear inverse problems with error in the operator *Ann. Stat.* at press
[36] Ibragimov I A and Hasminskii R Z 1981 *Statistical Estimation: Asymptotic Theory* (New York: Springer)
[37] Ibragimov I A and Hasminskii R Z 1984 On nonparametric estimation of the value of a linear functional in Gaussian white noise *Theory Prob. Appl.* **29** 18
[38] Johnstone I M 1999 Wavelet shrinkage for correlated data and inverse problems: adaptivity results *Stat. Sin.* **9** 51
[39] Johnstone I M, Kerkyacharian G, Picard D and Raimondo M 2004 Wavelet deconvolution in a periodic setting *J. R. Stat. Soc.* B **66** 547 (with discussion 627)
[40] Johnstone I M and Silverman B 1990 Speed of estimation in positron emission tomography and related inverse problems *Ann. Stat.* **18** 251
[41] Kaipio J and Somersalo E 2004 *Statistical and Computational Inverse Problems* (New York: Springer)

[42] Lehmann E L and Casella G 1998 *Theory of Point Estimation* (New York: Springer)
[43] Lepskii O V 1990 One problem of adaptive estimation in Gaussian white noise *Theory Prob. Appl.* **35** 459
[44] Lepskii O V, Mammen E and Spokoiny V G 1997 Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors *Ann. Stat.* **25** 929
[45] Loubes J M and Ludeña C 2007 Model selection for nonlinear inverse problems (in preparation)
[46] Mair B and Ruymgaart F H 1996 Statistical estimation in Hilbert scale *SIAM J. Appl. Math.* **56** 1424
[47] Mallows C L 1973 Some comments on C_p *Technometrics* **15** 661
[48] Marteau C 2006 Regularization of inverse problems with unknown operator *Math. Methods Stat.* **15** 415
[49] Marteau C 2007 Risk hull method for general families of estimators (in preparation)
[50] Mathé P 2006 The Lepskii principle revisited *Inverse Problems* **22** L11
[51] Mathé P and Pereverzev S V 2001 Optimal discretization of inverse problems in Hilbert scales. Regularization of some linear ill-posed problems with discretized random noisy data *Math. Comput.* **75** 1913
[52] Mathé P and Pereverzev S V 2001 Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods *SIAM J. Numer. Anal.* **38** 1999
[53] Natterer F 1986 *The Mathematics of Computerized Tomography* (New York: Wiley)
[54] Nussbaum M 1999 Minimax risk: Pinsker bound *Encyclopedia of Statistical Sciences* (New York: Wiley) p 451
[55] O'Sullivan F 1986 A statistical perspective on ill-posed problems *Stat. Sci.* **1** 502
[56] Pinsker M S 1980 Optimal filtering of square integrable signals in Gaussian white noise *Problems Info. Trans.* **16** 120
[57] Plaskota L 1996 *Noisy Information and Computational Complexity* (Cambridge: Cambridge University Press)
[58] Hamarik U, Palm R and Raus T 2007 Use of extrapolation in regularization methods *J. Inverse Ill-posed Probl.* **15** 277
[59] Ruymgaart F H 2001 A short introduction to inverse statistical inference Lecture in IHP (Paris)
[60] Stein C M 1981 Estimation of the mean of a multivariate normal distribution *Ann. Stat.* **9** 1135
[61] Stone C J 1980 Optimal rates of convergence for nonparametric estimators *Ann. Stat.* **8** 1348
[62] Sudakov V N and Khalfin L A 1964 Statistical approach to ill-posed problems in mathematical physics *Sov. Math.—Dokl.* **157** 1094
[63] Tenorio L 2001 Statistical regularization of inverse problems *SIAM Rev.* **43** 347
[64] Tikhonov A V 1963 Regularization of incorrectly posed problems *Sov. Math.—Dokl.* **4** 1624
[65] Tikhonov A V and Arsenin V Y 1977 *Solution of Ill-posed Problems* (New York: Wiley)
[66] Vogel C R 1986 Optimal choice of a truncation level for the truncated SVD solution of a linear first kind integral equations when data are noisy *SIAM J. Numer. Anal.* **23** 109
[67] Vogel C R 2002 *Computational Methods for Inverse Problems* (Philadelphia: SIAM)
[68] Wahba G 1990 *Spline Models for Observational Data* (Philadelphia: SIAM)