Inverse problems in statistics

Heidelberg, April 2007

Laurent Cavalier

1) Inverse problems

- Introduction
- Linear inverse problems with random noise
- Spectral Theorem
- SVD and sequence space model
- Examples

2) Nonparametric estimation

- Minimax rates
- Classes of functions
- Regularization methods
- Adaptation and oracle inequalities

3) Model selection

- Unbiased risk estimation
- Risk hull method

4) Conclusion

1 Inverse problems

1.1 Introduction

There exist many fields where inverse problems appear, from geology to financial mathematics and biology.

These are problems where we have indirect observations of an object (a function) that we want to reconstruct. From a mathematical point of view, this usually corresponds to the inversion of some operator.

The classical problem is the following : let A be an operator from the Hilbert space H in G :

$$\text{Given } g \in G \text{ find } f \in H \text{ such that } Af = g.$$

This is really an inverse problem in the sense that one has to invert the operator A. A case of major interest is the case of ill-posed problems where the operator is not invertible. The problem is then to handle this inversion in order to obtain a precise reconstruction.

A classical definition is the following (see [12]).

Definition 1 *(Hadamard) A problem is called* **well-posed** *if*

1. *there exists a solution to the problem (existence)*
2. *there is at most one solution to the problem (uniqueness)*
3. *the solution depends continuously on the data (stability)*

A problem which is not well-posed is called **ill-posed**.

If the data space is defined as the set of solutions, existence is clear. However, this could be modified if the data are perturbed by noise. Uniqueness of the solution is not easy to show. In case where it is not garanted by the data, then the set of a priori solutions can be restricted, and the problem is then reformulated.

Nevertheless, the main issue is usually stability. Indeed, suppose A^{-1} exists but is not bounded. Given a noisy version of g called g_ε, the reconstruction $f_\varepsilon = A^{-1}g_\varepsilon$ may be far from the true f.

Until the beginning of the 20th century it was generally believed that for natural problems the solution will always depend continuously on the data. If this was not the case, the mathematical model was believed to be inadequate. Therefore these problems were called ill-posed. Only in the second half of the last century, it was realized that a huge number of problems arising in sciences and technology were ill-posed in any reasonnable framework. This initiated a large amount of research in stable and accurate methods (see for example [16]).

1.2 Linear inverse problems with random noise

The classical framework for inverse problem is the linear inverse problems between two Hilbert spaces.

Let H and G two separable Hilbert spaces. Let A be a known linear bounded operator from the space H to G. The null space, the range and the definition domain of A are denoted by $\mathrm{Ker}(A)$, $R(A)$ and $D(A)$.

Suppose that we have the following observation model

$$Y = Af + \varepsilon\xi, \tag{1}$$

where Y is the observation, f is an unknown element in H, ξ is a stochastic error, ε corresponds to the noise level. Our aim here is to estimate (or reconstruct) the unknown f by use of the observation Y. The idea is that when ε will be small we hope to obtain rather precise results.

The standard framework (due to [28] and [29]), and corresponds to the case of inverse problems with deterministic noise. In this case, the noise ξ is considered as some element in G, with $\|\xi\| \leq 1$. Since the noise is some unknown element of a ball in G, then the results have to be obtained for any possible noise, i.e. for the worst noise.

Our framework is a statistical inverse problem (due to [27]). Indeed we observe a noisy version (with random error) of Af and we want to reconstruct f. Thus, two main difficulties appear :

- Deal with the noise in the observation (statistics).
- Invert the operator A (inverse problems theory)

The stochastic error is a Hilbert-space process, i.e. a bounded linear operator $\xi : G \to L^2(\Omega, \mathcal{A}, P)$ where $(\Omega, \mathcal{A}, P)$ is the underlying probability space.

Thus, for any functions $g_1, g_2 \in G$, define the Hilbert-space random variables $\langle \xi, g_j \rangle$ $j = 1, 2$. By definition $E\langle \xi, g_j \rangle = 0$ and define its covariance Cov_ξ as the bounded linear operator from H in G such that $\langle \mathrm{Cov}_\xi g_1, g_2 \rangle = \mathrm{Cov}(\langle \xi, g_1 \rangle, \langle \xi, g_2 \rangle)$.

The standard hypothesis corresponds to the following assumption.

Definition 2 *We say that ξ is a* **white noise process** *if* $\mathrm{Cov}_\xi = I$ *and the random variables are Gaussian :*

for any functions $g_1, g_2 \in G$, the random variables $\langle \xi, g_j \rangle$ are $\mathcal{N}(0, \|g_j\|^2)$ and $\mathrm{Cov}(\langle \xi, g_1 \rangle, \langle \xi, g_2 \rangle) = \langle g_1, g_2 \rangle$.

See for example [14]

Remark 1 *An important remark is that a white noise, as a Hilbert-space process, is not in general a Hilbert-space random variable, and note that $\|\xi\| = \infty$. One main difference between the determinic and the stochastic approaches of inverse problems is that the random noise is large compared to the deterministic one.*

1.3 Spectral theorem

The Halmos version of the spectral theorem (see [13]) turns out to be very convenient for the construction and the analysis of inverse problems.

Theorem 1 *For any self-adjoint linear operator $A : D(A) \to G$ defined on a dense subset of a separable Hilbert space H there exists, a $\sigma-$ compact space S, a Borel measure Σ on S, a unitary operator $U : H \to L^2(\Sigma)$, and a positive measurable function $\rho : S \to I\!R_+$ such that for all $f \in D(A)$*

$$UAf = \rho \cdot Uf, \ \Sigma \text{ a.e.} \tag{2}$$

Define the multiplication operator $M_\rho : L^2(\Sigma) \to L^2(\Sigma)$, where $M_\rho \varphi = \rho \cdot \varphi$. We can then rewrite (2) as

$$A = U^{-1} M_\rho U. \tag{3}$$

Remark 2 *This very important result means that when working in the good space, any linear operator is equivalent to a multiplication.*

Using the spectral theorem one obtains the equivalent model to (1)

$$UY = U(Af + \varepsilon\xi) = UAf + \varepsilon U\xi = UU^{-1}M_\rho Uf + \varepsilon U\xi,$$

which gives in the spectral domain

$$Z = \rho.\theta + \varepsilon\eta, \tag{4}$$

where $Z = UY$, $\theta = Uf$ and $\eta = U\xi$ which is a white noise in $L^2(\Sigma)$ since U is a unitary operator.

Remark 3 *If A is not self-adjoint then we use A^*A, where A^* is the adjoint of A.*

1.4 SVD and sequence space model

In the special case where A is a compact operator, a well-known version of the spectral theorem states that A has a complete orthogononal system of eigenvectors $\{\varphi_k\}$ with corresponding eigenvalues ρ_k. This is a special case of (2) where $S = N$, Σ is the counting measure, $L^2(\Sigma) = \ell^2(N)$, and $\rho(k) = \rho_k$.

Let A be a injective compact operator then we have

$$A^*Af = \sum_{k=1}^{\infty} \rho_k \langle f, \varphi_k \rangle \varphi_k,$$

where $\rho_k > 0$. Define the normalized image $\{\psi_k\}$ of $\{\varphi_k\}$ by $A\varphi_k = b_k\psi_k$, where $b_k = \sqrt{\rho_k}$. Remark 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

$$A^*\psi_k = b_k^{-1}A^*A\varphi_k = b_k^{-1}b_k^2\varphi_k = b_k\varphi_k.$$

Thus, we have

$$A\varphi_k = b_k\psi_k, \quad A^*\psi_k = b_k\varphi_k.$$

Definition 3 *We say that A admits a* **singular value decomposition (SVD)** *if*

$$A^*Af = \sum_{k=1}^{\infty} b_k^2\theta_k \ \varphi_k,$$

where θ_k are the coefficients of f in $\{\varphi_k\}$, $\{b_k\}$ are the **singular values**.

The SVD is the natural basis for A since it diagonalizes A^*A.

Now make the projection of Y on $\{\psi_k\}$

$$\langle Y, \psi_k\rangle = \langle Af, \psi_k\rangle + \varepsilon\langle \xi, \psi_k\rangle = \langle Af, b_k^{-1}A\varphi_k\rangle + \varepsilon\xi_k = b_k^{-1}\langle A^*Af, \varphi_k\rangle + \varepsilon\xi_k = b_k\theta_k + \xi_k,$$

where $\xi_k = \langle \xi, \psi_k\rangle$.

Since ξ is a white noise $\{\xi_k\}$ is a sequence of i.i.d. standard Gaussian random variables $\mathcal{N}(0,1)$.

Thus, under these assumptions, one has the equivalent discrete sequence observation model derived from (1):

$$y_k = b_k\theta_k + \varepsilon\xi_k, \quad k = 1, 2, \ldots, \tag{5}$$

where y_k stands for $\langle Y, \psi_k\rangle$. This model is called the **sequence space model**. The aim here is to estimate the sequence $\theta = \{\theta_k\}$ by use of the observations $y = \{y_k\}$.

Remark 4 *An important special case is the case where $A = I$. This corresponds to the* **direct model** *where f is directly observed without inverse problems. In this case $b_k = 1$ and the model in (5) corresponds to the classical sequence space model in statistics. The model is then related to the Gaussian white noise model and is very close to nonparametric regression with $\varepsilon = n^{-1/2}$*

One can see the influence of the ill-posedness of the inverse problem when A is compact. Indeed, since b_k are the singular values of a compact operator, then $b_k \to 0$ as $k \to \infty$. Thus, when k increases the "signal" $b_k\theta_k$ is weaker and it is clearly more difficult to estimate θ_k.

Another comment concerns the fact that one wants to estimate $\{\theta_k\}$ and not $\{b_k\theta_k\}$. Thus, one really has to invert $\{b_k\}$, i.e. to invert the operator A.

For this reason, the following equivalent model to (5) is more natural

$$X_k = \theta_k + \varepsilon\sigma_k\xi_k, \quad k = 1, 2, \dots, \tag{6}$$

where $X_k = y_k/b_k$, and $\sigma_k = b_k^{-1} > 0$. Note that $\sigma_k \to \infty$. In this model the aim is to estimate $\{\theta_k\}$ by use of $\{X_k\}$. When k is large the noise in X_k may then be very large, making the estimation difficult.

The sequence space formulation (5) or (6) for statistical inverse problems has been studied in a number of papers (see [9, 20, 18, 8] among others).

Remark 5 *For ill-posed inverse problems we have $b_k \to 0$ and $\sigma_k \to \infty$, as $k \to \infty$. We can see that ill-posed problems are more difficult than the direct model. Indeed, when k is large the estimation is more difficult.*

One can characterize linear inverse problems by the difficulty of the operator, i.e. with our notations, by the behavior of the σ_k. If $\sigma_k \to \infty$, as $k \to \infty$, the problem is ill-posed.

Definition 4 *An inverse problem will be called* **mildly ill-posed** *if the sequence σ_k has a polynomial behaviour when k is large*

$$\sigma_k \approx k^\beta, \ k \to \infty,$$

and **severely ill-posed** *if σ_k tends to ∞ at an exponential rate*

$$\sigma_k \approx \exp(\beta k), \ k \to \infty,$$

where $\beta > 0$ is called the **degree of ill-posedness** *of the inverse problem.*

A special of inverse problems is the **direct model** *where*

$$\sigma_k \approx 1, \ k \to \infty,$$

corresponding to $\beta = 0$.

1.5 Examples

Here are some examples of ill-posed problems where the spectral theorem may be applied. In each case, the SVD or the spectrum can be explicitely computed. However, in several inverse problems the spectral equivalence may not be used.

Moreover, from a practical point of view, methods based on SVD are usually very expensive in term of computations. For these reasons, many populars methods nowadays do not use explicitely the SVD.

On the other hand, even for these methods, the spectral domain is often used in order to deal with the theoretical accuracy of the methods.

1.5.1 Derivation

An example, which does not exactly correspond to our framework, but is very important, is the estimation of a derivative. Suppose that we observe

$$Y = f + \varepsilon\xi, \tag{7}$$

where $H = L^2[0,1]$, f is a periodic C^β function, $\beta \in \mathbf{N}$, in $L^2[0,1]$ and ξ is a white noise. A standard problem in statistics is the estimation of the derivative $D^\beta f = f^{(\beta)}$ of f, or the function f itself when $\beta = 0$.

One may use here the Fourier basis $\varphi_k(x) = e^{2\pi ikx}$, $k \in Z$. Denote by θ_k the Fourier coefficients of f, $\theta_k = \int_0^1 f(x)e^{2\pi ikx}dx$. It is well-known that we then have

$$f^{(\beta)} = \sum_{k=-\infty}^{\infty} (2\pi ik)^\beta \theta_k \varphi_k.$$

We have the following equivalent model in the Fourier domain

$$y_k = \theta_k + \varepsilon\xi_k, \ k \in Z^*,$$

and we want to estimate $\nu_k = \theta_k(2\pi ik)^\beta$. This is equivalent to

$$y_k = (2\pi ik)^{-\beta}\nu_k + \varepsilon\xi_k, \ k \in Z^*.$$

Thus, derivation is a mildly ill-posed inverse problem of degree β.

1.5.2 Heat equation

Consider the following heat equation :

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

where $u(x,t)$ is defined for $x \in [0,1], t \in [0,T]$, and the initial condition f is a 1-periodic function. The problem is then given the temperature $g(x) = u(x,T)$ at time T find the initial temperature $f \in L^2([0,1])$.

The SVD is here again the Fourier basis.

In this case u may be written as

$$g(x) = u(x,T) = \sqrt{2}\sum_{k=1}^{\infty} \theta_k e^{-\pi^2 k^2 T} \sin(k\pi x).$$

The singular values b_k are equal to $e^{-\pi^2 k^2 T/2}$ and the problem is severely ill-posed.

1.5.3 Deconvolution

In this section, we are going to present an example of application (given in [24]) for which the operator is non-compact. The operator considered here is the following

$$Af(t) = r * f(t) = \int_{-\infty}^{\infty} r(t-u)f(u)du,$$

where $r * f$ denotes the convolution through a known filter $r \in L^1(I\!\!R)$, The aim is to reconstruct the unknown function f.

The problem of convolution is one of the most standard inverse problems. The problem of circular convolution, i.e. periodic on $[a,b]$, appears for example in [8] and in Section 2.2. The main difference is that for periodic convolution the operator is compact and the basis of eigenfunctions is the Fourier basis. It seems clear, from a heuristic point of view, that the results could be extended to the case of convolution on $I\!\!R$ by using the Fourier transform on $L^2(I\!\!R)$ instead of the Fourier series.

Suppose that r is symmetric about 0, then

$$\tilde{r}(x) = \int_{-\infty}^{\infty} e^{itx} r(t)dt = \int_{-\infty}^{\infty} \cos(tx)r(t)dt > 0, \ \forall x \in I\!\!R.$$

The operator A is self-adjoint.

Define the Fourier transform as a unitary operator from $L^2(I\!\!R)$ into $L^2(I\!\!R)$ by

$$(Ff)(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} f(t)dt, \ x \in I\!\!R, \ f \in L^1(I\!\!R) \cap L^2(I\!\!R), \tag{8}$$

and its continuous extension on $L^2(I\!\!R)$.

We have that

$$F(r * f)(x) = \tilde{r}(x).(Ff)(x),$$

and then $A = F^{-1}M_{\tilde{r}}F$.

This framework may be extended in a two-dimensional problem. In this case, it may modelize the problem of blurred images. One observes an image which is blurred and one wants to get a good reconstruction. This problem corresponds to a standard applied inverse problem linked to the blurred images of the Hubble satellite (see [10]).

1.5.4 Computerized tomography

Computerized tomography is used in medical image processing and has been studied for a long time ([22]). In medical X-ray tomography one tries to have an image of the internal structure of an object. This image is characterized by a function f. However, there is no direct observations of f. By measuring attenuation of X-rays, one observes cross section of the body.

From the mathematical point of view this problem corresponds to the reconstruction of an unknown function f in $I\!R^d$ based on observations of its Radon transform Rf, i.e. integrals over hyperplanes

$$Rf(s,u) = \int_{v:\langle v,s\rangle = u} f(v)dv,,$$

where $u \in [-1,+1]$, $s \in S^{d-1}, S^{d-1} = \{v \in I\!R^d, \|v\| = 1\}$ is the unit sphere in $I\!R^d$, $\|v\|$ is the Euclidean norm of v, and $\langle \cdot,\cdot\rangle$ is the scalar product in $I\!R^d$. The function $Rf(s,u)$ is defined on the cylinder $\mathcal{Z} = S^{d-1} \times [-1,+1]$. The set of points $v \in I\!R^d$ such that $\langle v,s\rangle = u$ is a hyperplane in $I\!R^d$, characterized by (s,u). The Radon transform is a suitable tool for the problem of tomography, because $Rf(s,u)$ represents the integral of f over the hyperplane $\{v \in I\!R^d, \langle v,s\rangle = u\}$.

In this case, the operator R is compact, and the SVD basis is known for the Radon transform. However, this basis is not easy to compute.

2 Nonparametric estimation

The aim of nonparametric estimation is to estimate (reconstruct) a function f (density or regression funtion) by use of observations. The main difference with parametric statistics is that the function f is not in some parametric family of functions, for example if f is a Gaussian probability density $\{\mathcal{N}(\mu,1),\ \mu \in I\!R\}$.

Instead of a general framework, the problem of nonparametric estimation will be described here in the setting of the sequence space model (5) which is equivalent to the inverse problem with random noise (1).

2.1 Minimax approach

Let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots)$ be an estimator of $\theta = (\theta_1, \theta_2, \ldots)$ based on the data $X = \{X_k\}$. An estimator of θ may be any measurable function of the observation $X = \{X_k\}$.

Then f is estimated by $\hat{f} = \sum_k \hat{\theta}_k \varphi_k$.

The first point is to define the accuracy of some given estimator $\hat{\theta}$. Since an estimator is by definition random, we will measure the squared difference between $\hat{\theta}$ and the true θ, and then take the mathematical expectation.

Define the **mean integrated squared risk (MISE)** of $\hat{f}$ is

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

where the notation $\| \cdot \|$ means the ℓ_2-norm when applied to θ-vectors in the sequence space. Here and later $\mathbf{E}_f$ and $\mathbf{E}_\theta$ denote the expectations w.r.t. Y or $X = (X_1, X_2, \ldots)$ for models (1) and (5) respectively. Analyzing the risk $\mathcal{R}(\hat{f}, f)$ of the estimator $\hat{f}$ is equivalent to analyze the corresponding sequence space risk $\mathbf{E}_\theta \|\hat{\theta} - \theta\|^2$.

The aim would be to find the estimator with the minimum risk. However, the risk of an estimator depends, by definition, on the unknown f or θ.

Suppose that f belongs to some class of function $\mathcal{F}$.

Definition 5 *Define the* **maximal risk** *of the estimator $\hat{f}$ on $\mathcal{F}$ as*

$$\sup_{f \in \mathcal{F}} \mathcal{R}(\hat{f}, f),$$

and the **minimax risk** *as*

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

where the $\inf_{\hat{f}}$ is taken for any estimator of f.

It is usually not possible in nonparametric statistics to find an estimator which attains the minimax risk. A more natural approach is to consider the asymptotic properties, i.e. when the noise level tends to 0, $\varepsilon \to 0$.

Definition 6 *Suppose that some estimator $f^\star$ is such that there exist constants $0 < C_2 \le C_1 < \infty$ with, as $\varepsilon \to 0$*

$$\sup_{f \in \mathcal{F}} \mathcal{R}(f^\star, f) \le C_1 v_\varepsilon^2,$$

where the positive sequence v_ε, $v_\varepsilon \to 0$, is such that

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}} \mathcal{R}(\hat{f}, f) \geq C_2 v_\varepsilon^2,$$

In this case the estimator $f^\star$ is said to be **optimal** *or to* **attain the optimal rate of convergence** *v_ε.*

In the special case where $C_1 = C_2$ the estimator $f^\star$ is said to be **minimax** *or to* **attain the exact constant***.*

An optimal estimator is then an estimator whose risk is of the order of the best possible estimator.

2.2 Classes of functions

An important problem now is to define "natural" classes of functions on $\mathcal{F}$

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

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

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 then will be small.

Remark 6 *Assumptions on the coefficients θ_k will be usually related to some properties (smoothness) on f. One difficulty in using SVD in inverse problems is that the basis $\{\varphi_k\}$ is defined by the operator A, and one has to hope good properties for the coefficients θ_k of f in this specific basis.*

In the special cases where the SVD basis is the Fourier basis, hypothesis on $\{\theta_k\}$ may be precisely written in terms of smoothness for f.

Such classes arise naturally in various inverse problems, they include as special cases the Sobolev classes and classes of analytic functions.

For example in a circular deconvolution problem with the following operator

$$Af(x) = r * f(x), \quad x \in [0, 1],$$

where r is a known 1-periodic convolution kernel in $L_2([0, 1])$. The SVD basis is clearly here the Fourier basis.

Let $\{\varphi_k(t)\}$ be the real trigonometric basis on $[0,1]$:

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

Introduce the **Sobolev classes of functions**

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

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

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

where $\alpha > 0$, $L > 0$.

If α is an integer, this corresponds to the equivalent definition

$$\mathcal{W}(\alpha, L) = \left\{ f : \int_0^1 (f^{(\alpha)}(t))^2 dt \le \pi^{2\alpha} L \right\}$$

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

Consider also the classes of functions

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

where $a_k = \exp(\alpha k)$, $\alpha > 0$, and $L > 0$. This corresponds to the usual classes of analytical functions (functions which admit an analytical continuation into a band of the complex plane).

In this setting of ill-posed inverse problems with compact operator and functions with coefficients in some ellipsoid, several results have been obtained.

There exists a famous result by [23] which exhibits an estimator which is even minimax, i.e. which attains not only the optimal rate, but also the exact constant.

This result has been also generalized to the very specific case of severy ill-posed problems with analytic functions in [11].

The optimal rates of convergence may also be found, for example in [2], [8], and appear in the following table :

Inverse Problem/Functions	Sobolev (poly)	Analytic (expo)
Direct	$\varepsilon^{\frac{4\alpha}{2\alpha+1}}$	$\varepsilon^2(\log\frac{1}{\varepsilon})$
Mildly (poly)	$\varepsilon^{\frac{4\alpha}{2\alpha+2\beta+1}}$	$\varepsilon^2(\log\frac{1}{\varepsilon})^{2\beta+1}$
Severely (expo)	$(\log\frac{1}{\varepsilon})^{-2\alpha}$	$\varepsilon^{\frac{4\alpha}{2\alpha+2\beta}}$

Remark 7 *We may remark that the rates usually depend strongly on the smoothness α of the function f and on the degree of ill-posedness β. When β increases the rates are slower. In the polynomial case with $\beta = 0$ we get the rates of the direct model, i.e. the standard rates for nonparametric estimation. In the standard cases (polynomial/polynomial or exponential/ exponential) the rates are polynomial in ε and slower than ε^2 as usual in nonparametric statistics. The two other cases are very specific problems. In the first case, a very difficult inverse problem with not smooth enough functions. The rate is then logarithmic, and then very slow. In the second case, a mildly ill-posed problem with very smooth functions, the rate is almost the parametric rate ε^2.*

2.3 Regularization methods

We are going to give some examples of regularization method or estimators which are commonly used. All these methods are defined in the spectral domain even if some of them may be computed without using all the spectrum.

Let $\lambda = (\lambda_1, \lambda_2, \ldots)$ be a sequence of nonrandom weights. Every sequence λ defines a **linear estimator** $\hat{\theta}(\lambda) = (\hat{\theta}_1, \hat{\theta}_2, \ldots)$ where

$$\hat{\theta}_k = \lambda_k X_k \text{ and } \hat{f}(\lambda) = \sum_{k=1}^{\infty} \hat{\theta}_k \ \varphi_k.$$

Examples of commonly used weights λ_k are the projection weights $\lambda_k = I(k \leq N)$ where $I(\cdot)$ denotes the indicator function. These weights correspond to the **projection estimator** (also called **truncated SVD** or **spectral cut-off**)

$$\hat{\theta}(N) = \begin{cases} X_k, & k \leq N, \\ 0, & k > N. \end{cases}$$

The value N is called the **bandwidth**.

The truncated SVD is a very simple estimator. It is usually used as a benchmark since it attains the optimal rate of convergence. However, it is not a very precise estimator. From a numerical point of view, it is usually time consuming since, one has to compute all the coefficients X_k.

Define now, the well-known **Tikhonov regularization method** ([29]). In this method one wants to minimize the following functional :

$$\min_g \left\{ \|Ag - Y\|^2 + \gamma \|g\|^2 \right\},$$

where $\gamma > 0$ is some tuning parameter.

The Tikhonov method is very natural. Indeed, the idea is to choose an estimator which, due to the first term will fit the data, and which will be "regular", due to the second term. As we will see in Section 3 the choice of γ is very sensible since it characterizes the balance between the fitting and the smoothness.

This minimum is attained by

$$\hat{f}_\alpha = (A^*A + \gamma I)^{-1} A^* Y,$$

under rather mild assumptions. These estimators may then be expressed in the spectral domain associated with the weights

$$\lambda_k = \frac{1}{1 + \gamma \sigma_k^2}, \ \gamma > 0.$$

Remark 8 *There exist several modified versions of the Tikhonov estimator, with a starting point g_0, the iterated version, or the Hilbert scales approach.*

Define now the L^2-risk of any linear estimators :

$$\mathcal{R}(\hat{f}(\lambda), f) = R(\theta, \lambda) = \mathbf{E}_\theta \sum_k (\hat{\theta}_k(\lambda) - \theta_k)^2 = \sum_{k=1}^{\infty} (1 - \lambda_k)^2 \theta_k^2 + \varepsilon^2 \sum_{k=1}^{\infty} \sigma_k^2 \lambda_k^2.$$

The first term in the RHS is called **bias term** and the second term is called the **stochastic term** or **variance term**. The bias term is linked to the approximation error and measure if the chosen regularization method is a good approximation of the unknown f. On the other hand, the stochastic measure the influence of the random noise and of the inverse problem in the accuracy of the method.

In these lectures, we are going to study mainly the projection estimators. This method is the most simple and can be studied in a very easy way. The risk of a projection estimator with 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 decomposition is very simple. Indeed, we estimate the first N coefficients by their empirical version X_k and the other coefficients by 0. Thus, the bias term measure the influence of the remainder coefficients θ_k $k > N$, and the stochastic term is due to the random noise in the N first coefficients. We can see now that one simple question is how to choose the bandwidth N ?

Remark 9 *Thus, we get to the key-point in nonparametric statistics. We have to choose N in order to balance the bias term and the variance term. As we will see this choice will be difficult since the bias term depends on the unknown f.*

In a standard framework we obtain the following theorem.

Theorem 2 *Consider now 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 $\varepsilon \to 0$*

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

This rate may be shown to be optimal.

We have,

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

Deal with term

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

The variance term is controlled by

$$\varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 = \varepsilon^2 \sum_{k=1}^{N} k^{2\beta} \approx \frac{\varepsilon^2 N^{2\beta+1}}{2\beta+1},$$

when N is large. Thus,

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

If we want to attain the optimal rate of convergence we have to choose N of order $\varepsilon^{-2/(2\alpha+2\beta+1)}$ as $\varepsilon \to 0$. This choice corresponds to the trade-off between the bias term and the variance term.

Remark 10 *Considering the minimax point of view, we may remark that there exists an optimal choice for N which corresponds to the balance between the bias and the variance. However, this choice depends very precisely on the smoothness α and on the degree of ill-posedness of the inverse problem β.*

Even in the case where the operator A (and then its degree β) is known, it has no real meaning to consider that we know the smoothness of the unknown function f.

These remarks lead to the notion of adaptation and also oracle inequalities, i.e. how to choose the bandwidth N without strong a priori assumptions on f.

2.4 Adaptation and oracle inequalities

One of the most important point in nonparametric statistics is then typically linked to the problem of calibrating by the data the tuning parameter (N or γ) in any class of estimators. For example, we have seen that this choice is very sensible if we want to attain the optimal rate of convergence.

Optimization for a given class of functions.

The starting point of the approach 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 he does not know exactly which one. If Θ_α is a smoothness class, this assumption can be interpreted as follows : the statistician knows that the underlying function has some smoothness, but he does not know the degree of smoothness.

Definition 7 *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.*

This notion has been developped, for example in [19].

Remark 11 *Minimax adaptive estimators are really important in statistics from a theoretical and from a practical point of view. Indeed, it implies that these estimators are optimal for any possible parameter in the collection $\mathcal{A}$. From a more practical point of view it garantees a good accuracy of the estimator for a very large choice of functions.*

Thus, we have an estimator which automatically adapt to the unknown smoothness of the underlying function.

Optimization within a given class of estimators (or model selection).

Consider now a linked, but different point of view. Assume that a class of estimators is fixed, i.e. that the class of possible weights Λ is given. Define the **oracle** λ^0 as

$$R(\theta, \lambda^0) = \inf_{\lambda \in \Lambda} R(\theta, \lambda). \tag{9}$$

The oracle corresponds to the best possible choice in Λ, i.e. the one which minimizes the risk. However, this is not an estimator since the risk depends on the unknown θ, the oracle will depend also. For this reason, it is called oracle since it is the best one in the family, but it knows 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 $\theta^\star = \hat{\theta}(\lambda^\star)$ satisfies an **oracle inequality**, for any $\theta \in \ell^2$,

$$\mathbf{E}_\theta \|\theta^\star - \theta\|^2 \leq C \inf_{\lambda \in \Lambda} R(\theta, \lambda) + \ \Delta_\varepsilon, \tag{10}$$

where Δ_ε is some positive remainder term and $C > 0$ (close to 1 if possible). If the remainder term is small, i.e. smaller than the main term $R(\theta, \lambda^0)$ then an oracle inequality proves that the estimator has a risk of the order of the oracle.

We are interested in data-driven methods, and then automatic, which more or less mimic the oracle.

Remark 12 *The oracle approach is in some sense the opposite of the minimax approach. Here, we fix a family of estimators and choose the best one among them. In the minimax approach, on the other hand, one try to get the best accuracy for functions which belong to some function class. Moreover, the oracle inequalities, are true for any θ, and are non asymptotic.*

The oracle approach is often used as a tool in order to obtain adaptive estimators. Indeed, the best estimator in a given class often attains the optimal rate of convergence.

On the other hand, the minimax theory may be viewed as a justification for oracle inequality. Indeed, on may ask if the given family of estimators is satisfying. One possible answer comes from minimax results, which proves that a given family gives optimal estimators.

One may obtain some asymptotic results when $\varepsilon \to 0$. We call an **exact oracle inequality** on the class Λ, as $\varepsilon \to 0$,

$$\mathbf{E}_\theta \|\theta^\star - \theta\|^2 \leq (1 + o(1)) R(\theta, \lambda^0), \tag{11}$$

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 $\theta \in \Theta_0$.

The key assumption in this approach is that $\lambda^\star$ is restricted to take its values in same class Λ that appears in the RHS of (10). A **model selection** interpretation of (10) is the following : in a given class of models Λ we pick the model $\lambda^\star$ that is asymptotically the closest to the true parameter θ in terms of the risk $R(\theta, \lambda)$.

3 Model selection

The framework of model selection is very popular in statistics, and may have several meaning depending on the topics. We consider the model selection approach as the problem of choosing among a given family of models Λ (estimators) the best possible one. This choice should be made based on the data and not due to some a priori information on the unknown function f.

3.1 Unbiased risk estimation

The definition of the oracle in (9) is that the oracle minimizes the risk. Since θ is unknown, the risk is also.

A very natural idea in statistics is to estimate this unknown risk by a function of the observations, and then to minimize this estimator of the risk. A classical approach to this minimization problem is based on the principle of **unbiased risk estimation** (URE) (see [25]). The idea to use this method for data-driven bandwidth choice goes back to [1, 21]. Originally, the URE was proposed in the context of regression estimation. Nowadays, it is used as a basic adaptation tool for many statistical models.

This idea appears also in all the cross-validation techniques.

For inverse problems, this method was studied in [6], where exact oracle inequalities for the mean square risk were obtained.

In this setting, the functional

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

is an unbiased estimator of $R(\theta,\lambda)$:

$$R(\theta,\lambda) = \mathbf{E}_\theta U(X,\lambda), \ \forall\lambda. \tag{12}$$

The principle of unbiased risk estimation suggests to minimize over $\lambda \in \Lambda$ the functional $U(X,\lambda)$ in place of $R(\theta,\lambda)$. This leads to the following data-driven choice of λ:

$$\lambda^\star = \arg\min_{\lambda\in\Lambda} U(X,\lambda). \tag{13}$$

Denote

$$S = \left(\frac{\max_{\lambda\in\Lambda}\sum_{k=1}^{\infty}\sigma_k^4\lambda_k^2}{\min_{\lambda\in\Lambda}\sum_{k=1}^{\infty}\sigma_k^4\lambda_k^2}\right)^{1/2}.$$

We obtain the following oracle inequality.

Theorem 3 *Under Assumptions 1 and 2 in [6]. Assume that Λ is finite with cardinality D. There exit constants $\gamma_1,\gamma_2 > 0$ such that for every $\theta \in \ell_2$ and for the estimator $\theta^\star = (\theta_1^\star,\theta_2^\star,\ldots)$ with $\theta_k^\star = \lambda_k^\star X_k$ we have for B large enough,*

$$\mathbf{E}_\theta\|\theta^\star - \theta\|^2 \le (1+\gamma_1 B^{-1})\min_{\lambda\in\Lambda} R_\varepsilon[\lambda,\theta] + \gamma_2 B\varepsilon^2(\log(DS))^{2\beta+1}. \tag{14}$$

By assuming hypothesis on the behaviour of D and S when ε is large, one may obtain an exact oracle inequality.

Proof. The proof of this theorem may be found in [6].

Remark 13 *One of main difficulties in adaptation or oracle results is that we deal with data-driven choices of N. Thus, the risk of the estimator is very difficult to control since it depends on the observations by X_k but also by the use of a random N. We will see this influence in the proof of Theorem 4.*

3.2 Risk hull method

In order to present the risk hull minimization, which is an improvment of the URE method, we restrict ourselves to the class of projection estimators. In this case, the URE criterion may be written

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

This corresponds in fact to the minimization in N of

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

and then

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

Remark 14 *One may note that even if we have obtained a very precise oracle inequality in Theorem 3, the URE method is in fact not satisfying in simulations. This leads to the idea of choosing the bandwidth N by a more stable approach.*

There exists a more general approach which is very close to the URE. This method is called **method of penalized empirical risk**, and in the context of our problem it provides us with the following bandwidth choice

$$N(X) = \arg\min_{N\geq 1} \bar{R}_{pen}(X,N), \quad \bar{R}_{pen}(X,N) = \Big\{ -\sum_{k=1}^{N} X_k^2 + \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 + \mathrm{pen}(N) \Big\}, \tag{15}$$

where pen(N) is a penalty function. The modern literature on this method is very vast and we refer interested reader to [4]. The main idea at the heart of this approach is that severe penalties permit to improve substantially the performance of URE. However, it should be mentioned that the principal difficulty of this method is related to the choice of the penalty function pen(N). In this context, the URE criterion corresponds to a specific penalty called the Akaike penalty

$$\mathrm{pen}_{ure}(N) = \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2.$$

The idea is usually to choose a heavier penalty, but the choice of such a penalty is a very sensible problem, and as we will see later, specially in the inverse problems context.

In [5], propose a more general approach, called **risk hull minimization** (RHM) which gives a relatively good strategy for the penalty choice. The goal is to present heuristic and mathematical justifications of this method.

The heuristic motivation of the RHM approach is based on the oracle ideology. Suppose there is an oracle which provides us with $\theta_k,\ k = 1, \ldots$, but we are allowed to use only the projection method. In this case the optimal bandwidth is evidently given by

$$N_{or} = \arg\min_N r(X, N), \text{ where } r(X, N) = \|\hat{\theta}(N) - \theta\|^2.$$

This oracle mimimizes the loss and is even better than the oracle of the risk. Let us try to mimic this bandwidth choice. At the first glance this problem seems hopeless since in the decomposition

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

neither θ_k^2 nor ξ_k^2 are really 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. It means that we minimize the following non-random functional

$$l(\theta, N) = \sum_{k=N+1}^{\infty} \theta_k^2 + V(N), \tag{16}$$

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 this function such that

$$\mathbf{E} \sup_N \Big[\varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 \xi_k^2 - V(N) \Big] \le 0, \tag{17}$$

since then we can easily control the risk of any projection estimator with any data-driven bandwidth $\tilde{N}$

$$\mathbf{E}_\theta \|\hat{\theta}(\tilde{N}) - \theta\|^2 \le \mathbf{E}_\theta l(\theta, \tilde{N}). \tag{18}$$

This motivation leads to the following definition :

Definition 8 *A non random function* $\ell(\theta, N)$ *is called* **risk hull** *if*

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

Thus, we can say that $l(\theta, N)$ defined by (16) and (17) is a risk hull. Evidently, we want to have the upper bound (18) as small as possible. So, we are looking for the minimal hull. Note that this hull strongly depends on σ_k^2.

Once $V(N)$ satisfying (17) has been chosen, the minimization of $l(\theta, N)$ can be completed by the standard way using the unbiased estimation. Note that our problem is reduced to minimization of $-\sum_{k=1}^N \theta_k^2 + V(N)$. Replacing the unknown θ_k^2 by their unbiased estimates $X_k^2 - \varepsilon^2\sigma_k^2$, we arrive at the following method of adaptive bandwidth choice

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

In the framework of the empirical risk minimization the RHM can be defined as follows. Let the penalty in (15) be for any $\alpha > 0$

$$\mathrm{pen}(N) = \mathrm{pen}_{rhm}(N) = \varepsilon^2 \sum_{k=1}^N \sigma_k^2 + (1+\alpha) U_0(N), \tag{19}$$

where

$$U_0(N) = \inf \Big\{ t > 0 : \ \mathbf{E}\left(\eta_N I(\eta_N \geq t) \right) \leq \varepsilon^2 \sigma_1^2 \Big\}, \ \text{with } \eta_N = \varepsilon^2 \sum_{k=1}^N \sigma_k^2 (\xi_k^2 - 1). \tag{20}$$

This RHM penalty corresponds in fact to the URE penalty plus some term $(1 + \alpha)U_0(N)$. We have as $N \to \infty$

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

The RHM chooses the bandwidth N_{rhm} according to (15) with the penalty function defined by (19) and (20). The following oracle inequality provides an upper bound for the mean square risk of this approach. Recall that it is assumed that σ_k has a polynomial growth ($\sigma_k = k^\beta$).

Theorem 4 *Let RHM bandwidth choice N_{rhm} according to (15) with the penalty function defined by (19, 20) and $\theta^\star_{rhm}$ the associated projection estimator.*

There exist constants $C_* > 0$ and $\delta_0 > 0$ such that for all $\delta \in (0, \delta_0]$ and $\alpha > 1$

$$\mathbf{E}\,\|\tilde{\theta}^\star_{rhm} - \theta\|^2 \le (1+\delta)\inf_N R_{rhm}(\theta, N) + C_* \varepsilon^2 \left(\frac{1}{\delta^{4\beta+1}} + \frac{1}{\alpha - 1}\right), \qquad (21)$$

where

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

Proof. Many of the details are deleted, in order to keep only the idea behind the risk hull. The correct proof may be found in [5].

As usual in nonparametric statistics we need to prove that $U_0(N) \approx \mathcal{N}(0, 2\varepsilon^4 \sum_{k=1}^N \sigma_k^4)$ and in particular the following exponential inequality

$$\mathbf{P}\{\eta_N > x\} \le \exp\left(-\frac{Cx^2}{\varepsilon^4 \sum_{k=1}^N \sigma_k^4}\right) \quad 0 < x < \frac{\Sigma_N}{\varepsilon^2 \max_{k=1,\dots,N} \sigma_i^2}, \qquad (22)$$

where $\Sigma_N = \varepsilon^4 \sum_{k=1}^N \sigma_k^4$. The proof is now in two parts :

the first part is to prove that

$$l_\alpha(\theta, N) = \sum_{k=N+1}^{\infty} \theta_k^2 + \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 + (1+\alpha) U_0(N) + \frac{C\varepsilon^2}{\alpha}.$$

is a risk hull. Remark that

$$\mathbf{E}_\theta \sup_N \left(r(X, N) - l_\alpha(\theta, N)\right)_+ \le 0,$$

is equivalent to

$$\mathbf{E} \sup_N \left(\eta_N - (1+\alpha)U_0(N)\right)_+ \le \frac{C\varepsilon^2}{\alpha}.$$

We have

$$\mathbf{E} \sup_N \left(\eta_N - (1+\alpha)U_0(N)\right)_+ \le \sum_{N=1}^{\infty} \mathbf{E}\left(\eta_N - (1+\alpha)U_0(N)\right)_+ .$$

The definition of $U_0(N)$ in (20) implies

$$\mathbf{E}\left(\eta_N - U_0(N)\right)_+ \le \varepsilon^2 \sigma_1^2.$$

More precisely, by integrating by parts we obtain

$$\mathbf{E}\left(\eta_h - (1+\alpha)U_0(N)\right)_+ = \int_{(1+\alpha)U_0(N)}^{\infty} \mathbf{P}(\eta_h > x)dx.$$

Due the exponential behaviour of η_N in (22) we want to control

$$\sum_{N=1}^{\infty} \int_{(1+\alpha)U_0(N)}^{\infty} \exp\left(-\frac{Cx^2}{\varepsilon^4 \sum_{k=1}^{N} \sigma_k^4}\right) dx \leq C \sum_{N=1}^{\infty} \sqrt{\Sigma_N} \exp\left(-C(1+\alpha)^2 \log(\Sigma_N)\right).$$

When $x > \Sigma_N/\varepsilon^2 \max \sigma_i^2$, it can be proved rather easily that the integral term is smaller.

Thus for α large enough, since Σ_N is polynomial in N, the term is then

$$\sum_{N=1}^{\infty} \mathbf{E}\left(\eta_N - (1+\alpha)U_0(N)\right)_+ \leq \frac{C\varepsilon^2}{\alpha}.$$

The proof for α small is much more technical.

In the second part we need to prove that we are able to minimize this risk hull based on the data. Since $l_\mu(\theta, N)$ is a risk hull for any $\mu > 0$ we have

$$l_\mu(\theta, N) = \sum_{k=N+1}^{\infty} \theta_k^2 + \varepsilon^2 \sum_{k=1}^{N} \sigma_k^2 + (1+\mu)U_0(N) + \frac{C\varepsilon^2}{\mu}, \tag{23}$$

and therefore

$$\mathbf{E}_\theta \|\hat{\theta}(N_{rhm}) - \theta\|^2 \leq \mathbf{E}_\theta l_\mu(\theta, N_{rhm}). \tag{24}$$

On the other hand, since N_{rhm} minimizes $\bar{R}_{pen}(X, N)$, we have for any integer N

$$\mathbf{E}_\theta \bar{R}_{pen}(X, N_{rhm}) \leq \mathbf{E}_\theta \bar{R}_{pen}(X, N) = R_{rhm}(\theta, N) + \|\theta\|^2. \tag{25}$$

In order to combine the inequalities (24) and (25), we rewrite $l_\mu(\theta, N_{rhm})$ in terms of $\bar{R}_{pen}(X, N_{rhm})$

$$\bar{R}_{pen}(X, N_{rhm}) + \|\theta\|^2 + \frac{C\varepsilon^2\sigma_1^2}{\mu} =$$

$$= l_\mu(\theta, N_{rhm}) - 2\sum_{i=1}^{N_{rhm}} \sigma_k \theta_k \xi_k - \varepsilon^2 \sum_{k=1}^{N_{rhm}} \sigma_k^2(\xi_k^2 - 1) + (\alpha - \mu)U_0(N_{rhm}).$$

Therefore, using this equation, (24) and (25), we obtain that for any integer N

$$\mathbf{E}_\theta \|\hat{\theta}(N_{rhm}) - \theta\|^2 \leq R_{rhm}(\theta, N) + \frac{C\varepsilon^2}{\mu} + 2\mathbf{E}_\theta \varepsilon \sum_{i=1}^{N_{rhm}} \sigma_k \theta_k \xi_k$$

$$+\mathbf{E}_\theta\left[\varepsilon^2\sum_{k=1}^{N_{rhm}}\sigma_k^2(\xi_k^2-1)-(\alpha-\mu)U_0(N_{rhm})\right].$$

The next step is to control the last two terms in the above equation. The first term may be included in the left term and in the remainder term. In the second term, we use again the risk hull.

Remark 15

- *We have an oracle inequality but with a penalty term on the RHS.*
- *We have $U_0(N) = o(\varepsilon^2 \sum_{k=1}^N \sigma_k^2)$ as $N \to \infty$. From an asymptotic point of view, there is no real difference between the URE and the RHM.*
- *However there exist main differences specially in the case of inverse problems. RHM is much more stable than URE.*
- *One of the reason for this unstability is that URE is based on some asymptotic ideas. In inverse problems, usually N is not very large, due to the increasing noise. One has to be very careful with asymptotics.*
- *The penalty $U_0(N)$ may be computed by Monte Carlo simulations.*
- *By use of RHM we obtain, an explicit penalty which comes from the proof of Theorem 4 and may be used directly in the simulations.*

4 Conclusion

4.1 Conclusion

A very promising approach to inverse problems is the statistical framework. It is based on a model where observations contain a random noise. This does not correspond to the standard framework of the work of [28] where the error is deterministic.

The optimal rates of convergence are different in the statistical and deterministic frameworks. Moreover, in ill-posed inverse problems the rates are slower than in the direct model corresponding to the standard nonparametric statistics framework.

One of the major advantages of this statistical approach is that it allows to obtain oracle inequalities and to construct adaptive estimators. The oracle approach is very interesting in inverse problems. Indeed , one can construct procedures in order to choose the best estimator among a given family of regularization methods. From a practical point of view, this choice is usually done by simulations in a very empirical way.

An important remark is that inverse problems are a rather difficult framework, since we have to invert an operator in order to get the reconstruction. A main issue is then to get very precise oracle inequalities. Indeed the ill-posedness of the problem appears in the results, which are then very sensible. Thus, in statistical inverse problems one has to define very precise model selection methods, otherwise, due to the difficulty of the problem, the estimator will not be accurate.

4.2 Open problems

In this lecture, the results have been obtained in a very specific and restrictive model, there exist many different approaches in order to extend the results or to deal with other kind of problems.

Noisy operators. One very restrictive assumption is that the operator A is perfectly known. In case, where we also have noise in the operator one may extend some of the results (see [7]).

RHM for other methods. There exist many other regularization methods (Iterative methods, Landweber, $\nu-$methods...). These methods usually attain the optimal rates of convergence (see [3]). A main point would be to extend the RHM approach to these families of estimators.

Wavelets and sparsity. One of the main drawback is that all these methods are linked to the spectral approach. In many problems, this leads to the Fourier domain. Another very popular approach is based on wavelets (see [9] or [15]). By using wavelets, one may usually deal with functions which are not very smooth, by replacing Sobolev classes by Besov classes.

Nonlinear operators. All the results given here are valid for linear inverse problems. In case of non-linear operators the problem is much more difficult. This framework have been intensively studied in the deterministic context but not that much in the statistical one (see [10]).

References

[1] Akaike H. (1973). *Information theory and an extension of the maximum likelihood principle.* Proc. 2nd Intern. Symp. Inf. Theory, Petrov P.N. and Csaki F. eds. Budapest, 267-281.

[2] Belitser E.N. and Levit B. Ya.(1995) On minimax filtering on ellipsoids. *Math. Meth. Statist.* **4**, 259-273.

[3] Bissantz N., Hohage T., Munk A. and Ruymgaart F. (2003) Convergence rates of general regularizations methods for statistical inverse problems and applications. Preprint University of Dortmund.

[4] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3** 203-268.

[5] Cavalier L. and Golubev G.K. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Annals of Statist.* **34**, 1653-1677.

[6] Cavalier L., Golubev G.K., Picard D. and Tsybakov A.B. (2002). Oracle inequalities in inverse problems. *Annals of Statist.* **30**, 843-874.

[7] Cavalier L. and Hengartner N. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems* **21**, 1345-1361.

[8] Cavalier L. and Tsybakov A.B. (2002). Sharp adaptation for inverse problems with random noise. *Proba. Theory and Rela. Fields* **123**, 323-354.

[9] Donoho D.L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational Harmonic Analysis* **2**, 101-126.

[10] Engl H.W., Hanke M. and Neubauer A. (1996). *Regularization of Inverse Problems*. Kluwer Academic Publishers.

[11] Golubev G.K. and Khasminskii R.Z. (2001) Statistical approach to Cauchy problem for Laplace equation. In: *State of the Art in Probability and Statistics, Festschrift for W.R. van Zwet* (M. de Gunst, C. Klaassen, A. van der Vaart, eds), IMS Lecture Notes Monograph Series **36**, 419-433 .

[12] Hadamard, J. (1932), Le problème de Cauchy et les équations aux dérivées partielles hyperboliques (Herman, Paris).

[13] Halmos P.R. (1963). What does the spectral theorem say? *Amer. Math. Monthly* **70**, 241-247.

[14] Hida T. (1980). *Brownian Motion*. Springer-Verlag, New York-Berlin.

[15] Hoffmann M. and Reiss M. (2005). Nonlinear estimation for linear inverse problems with error in the operator. Manuscript.

[16] Hohage T. (2002). Lecture notes on inverse problems. Lecture in University of Gottingen.

[17] Ibragimov I.A. and Hasminskii R.Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, N.Y. e.a.

[18] Johnstone I.M. (1999). Wavelet shrinkage for correlated data and inverse problems: adaptivity results. Statistica Sinica 9, 51-83.

[19] Lepskii O.V. (1990). One problem of adaptive estimation in Gaussian white noise. *Theory Prob. Appl.* **35**, 459-470.

[20] Mair B. and Ruymgaart F.H. (1996). Statistical estimation in Hilbert scale. *SIAM J. Appl. Math.* **56**, 1424-1444.

[21] Mallows C. L. (1973). Some comments on C_p. *Technometrics* **15** 661-675.

[22] Natterer F. (1986). *The Mathematics of Computerized Tomography*. Teubner, Stuttgart, J. Wiley, Chichester.

[23] Pinsker M.S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Info. Trans.* **16**, 120-133.

[24] Ruymgaart F.H. (2001). A short introduction to inverse statistical inference. Lecture in IHP, Paris.

[25] Stein, C.M. (1981) Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1135-1151.

[26] Stone C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348-1360.

[27] Sudakov V.N., and Khalfin L.A. (1964). Statistical approach to ill-posed problems in mathematical physics, *Soviet Math. Doklady* 157, 1094-1096.

[28] Tikhonov A.V. (1963), Regularization of incorrectly posed problems, *Soviet Math. Doklady* 4, 1624-1627.

[29] Tikhonov A. V. and Arsenin, V. Y. (1977). *Solution of Ill-posed Problems.* Winston & sons.