ESAIM: PS 17 (2013) 485–499
DOI: 10.1051/ps/2012003

ESAIM: Probability and Statistics
www.esaim-ps.org

ADAPTIVE HARD-THRESHOLDING FOR LINEAR INVERSE PROBLEMS

Paul Rochet[1]

Abstract. A number of regularization methods for discrete inverse problems consist in considering weighted versions of the usual least square solution. These filter methods are generally restricted to monotonic transformations, *e.g.* the Tikhonov regularization or the spectral cut-off. However, in several cases, non-monotonic sequences of filters may appear more appropriate. In this paper, we study a hard-thresholding regularization method that extends the spectral cut-off procedure to non-monotonic sequences. We provide several oracle inequalities, showing the method to be nearly optimal under mild assumptions. Contrary to similar methods discussed in the literature, we use here a non-linear threshold that appears to be adaptive to all degrees of irregularity, whether the problem is mildly or severely ill-posed. Finally, we extend the method to inverse problems with noisy operator and provide efficiency results in a conditional framework.

Mathematics Subject Classification. 62G08, 62J02.

Received May 2, 2011. Revised November 9, 2011.

1. Introduction

We are interested in recovering an unobservable signal x_0, based on noisy observations of the image of x_0 through a linear operator A. The observation y satisfies the following relation

$$y(t) = Ax_0(t) + \varepsilon(t),$$

where ε is a Gaussian random process representing the noise. This problem is studied in [6, 10, 12] and in many applied fields such as medical imaging in [15] or seismography in [16] for instance. When the measured signal is only available at a finite number of points $t_1, \dots, t_n$, the operator A must be replaced by a discrete version $A_n : x \mapsto (Ax(t_1), \dots, Ax(t_n))^t$, leading to a discrete linear model

$$y = A_n x_0 + \varepsilon,$$

with $y \in \mathbb{R}^n$ (v^t denotes the transpose of v). Difficulties in estimating x_0 occur when the problem is *ill-posed*, in the sense that small perturbations in the observations induce large changes in the solution. This is caused by an ill-conditioning of the operator A_n, reflected by a fast decay of its singular values b_i^2. In such problems, the least square solution, although having a small bias, is generally inefficient due to a too large variance. Hence, *regularization* of the problem is required to improve the estimation. A large number of regularization methods

Keywords and phrases. Inverse problems, singular value decomposition, hard-thresholding.

[1] Institut de Mathématiques de Toulouse, Université Paul Sabatier Toulouse III, 118 route de Narbonne, 31062 Toulouse, France. rochet@math.univ-toulouse.fr

Article published by EDP Sciences

are based on considering weighted versions of the least square estimator. The idea is to allocate low weights λ_i, or *filters*, to the least square coefficients that are highly contaminated with noise, thus reducing the variance, at the cost of increasing the bias at the same time. The most famous filter-based method is arguably the one due to Tikhonov (see [18]), where a collection of filters is indirectly obtained *via* a minimization procedure with ℓ^2 penalization. Tikhonov filters are entirely determined by a parameter τ that controls the balance between the minimization of the ℓ^2 norm of the estimator and the residual. Another well spread filter method that will be given a particular attention, is the *spectral cut-off* discussed in [2,8,9]. One simply considers a truncated version of the least square solution, where all coefficients corresponding to arbitrarily small eigenvalues are removed. Thus, spectral cut-off is associated to binary filters λ_i, equal to 1 if the corresponding eigenvalue b_i exceeds in absolute value a certain threshold τ, and 0 otherwise.

A common feature of spectral cut-off and Tikhonov regularization is the predetermined nature of the filters λ_i, defined in each case as a fixed non-decreasing function $f(\tau, .)$ of the eigenvalues b_i^2, and where only the parameter τ is allowed to depend on the observations. However, in many situations, non-monotonic sequences of filters may seem to be more appropriate. Actually, optimal values for λ_i generally depend on both the noise level, which is determined by the eigenvalue b_i, and the component, say x_i, of x_0 in the direction associated to b_i. A restriction to monotonic collections of filters may turn out to be inefficient in situations where the coefficients x_i are uncorrelated to the singular values of the operator A_n.

Regularization methods involving more general classes of filters have also been treated in the literature. For example, the *unbiased risk estimation* (URE) introduced by Stein in [17] and studied in this context in [6], applies to arbitrary classes of filters, dealing in particular with non-monotonic collections. However, this approach has proven inefficient in cases where the set of possible values for the filters grows exponentially with the number of observations. A suitable alternative is given by the more recent *risk hull method* discussed in [4], although the computation of the estimator may require computationally expensive Monte-Carlo procedures.

Here, we focus on the class of unrestricted binary filters $\lambda_i \in \{0, 1\}$, known as *projection filters*. The computation of the estimator relies on the choice of a proper set of coefficients, $m \subseteq \{1, \dots, n\}$, which increases the number of possibilities compared to the spectral cut-off. A suitable set m is chosen by a hard-thresholding procedure on the observations, which results in selecting only the indices i for which the corresponding observation is greater (in absolute value) than a threshold c_i. While most thresholding methods studied in the literature consider a threshold proportional to the variance (see for instance [1] or [14]), we propose a non-linear threshold that adapts to all degree of irregularity. We show this method to satisfy a non-asymptotic oracle inequality, when the oracle is computed in the class of projection filters. Moreover, we show our estimator to nearly achieve the rate of convergence of the best linear estimator in the maximal class of filters, *i.e.* when no restriction is made on λ_i.

It many actual situations, one may consider that the operator A_n is not known precisely and only an approximation of it is available. Regularization of inverse problems with approximate operator is studied in [5,7,11]. We tackle the problem of estimating x_0 in the situation where we observe independently a noisy version $\hat{b}_i$ of each eigenvalue b_i. We consider a framework where the observations $\hat{b}_i$ are made once and for all, and are thus seen as non-random. We provide a bound on the conditional risk of the estimator, given the values of $\hat{b}_i$, in the form of a conditional oracle inequality.

The paper is organized as follows. We introduce the problem in Section 2. We define our estimator in Section 3, and provide two kinds of oracle inequalities and numerical applications. Section 4 is devoted to an application of the method to inverse problems with noisy operators. The proofs of our results are postponed to Appendix A.

2. Problem setting

Let $(\mathcal{X}, \|.\|)$ be a Hilbert space and $A_n : \mathcal{X} \to \mathbb{R}^n$ $(n > 2)$ a linear operator. We tackle the problem of recovering an unknown signal $x_0 \in \mathcal{X}$ based on the indirect observations

$$y = A_n x_0 + \varepsilon, \tag{2.1}$$

where ε is a Gaussian vector representing the noise. We assume that ε is centered with known covariance matrix $\sigma^2 I$, where I denotes the identity matrix. We endow $\mathbb{R}^n$ with the inner product $\langle u, v\rangle_n = \frac{1}{n} u^t v$ and the associated norm $\|.\|_n$ and we note $A_n^* : \mathbb{R}^n \to \mathcal{X}$ the adjoint of A_n. Let $\mathcal{K}_n$ be the kernel of A_n and $\mathcal{K}_n^\perp$ its orthogonal in $\mathcal{X}$ which we assume to be of dimension n. The fact that A_n is surjective ensures that the observation y provides information in all directions. If this condition is not met, one may simply reduce the dimension of the image in order to make A_n surjective.

Let $\{b_i; \phi_i, \psi_i\}_{i=1,\dots,n}$ be a singular system for the linear operator A_n, that is, $A_n\phi_i = b_i\psi_i$, $A_n^*\psi_i = b_i\phi_i$ and $b_1^2 \geq \ldots \geq b_n^2 > 0$ are the ordered non-zero eigenvalues of the self-adjoint operator $A_n^* A_n$. The ϕ_i's (resp. ψ_i's) form an orthonormal system of $\mathcal{K}_n^\perp$ (resp. $\mathbb{R}^n$). Note that due to the discretized nature of the problem, b_i, ϕ_i, ψ_i all depend on n although the dependency is dropped to ease notations.

The efficiency of the estimator relies first of all on the accuracy of the discrete operator A_n and how "close" it is to the true value A. The convergence of the estimator toward x_0 is subject to the condition that the distance of x_0 to the set $\mathcal{K}_n^\perp$ tends to 0, which is reflected by a proper asymptotic behavior of the design $t_1, \dots, t_n$. This aspect is not discussed here, we consider a framework where we have no control over the design $t_1, \dots, t_n$ and we focus on the convergence of the estimator toward the *best approximate solution* $x^\dagger$, that is, the orthogonal projection of x_0 onto $\mathcal{K}_n^\perp$. Remark that the best approximate solution can also be expressed as the image of $A_n x_0$ through the generalized Moore–Penrose inverse operator $A_n^\dagger = (A_n^* A_n)^\dagger A_n^*$, where $(A_n^* A_n)^\dagger$ denotes the inverse of $A_n^* A_n$, restricted to $\mathcal{K}_n^\perp$. We refer to [8] for more details.

Searching for a solution in the subspace $\mathcal{K}_n^\perp$ allows to reduce the number of regressors to n. Then, estimating $x^\dagger$ can be made using a classical linear regression framework where the number of regressors is equal to the dimension of the observation. Decomposing the observation in the singular basis $\{\psi_i\}_{i=1,\dots,n}$ leads to the following model

$$y_i = b_i x_i + \varepsilon_i, i = 1, \dots, n,$$

where we set $y_i = \langle y, \psi_i\rangle_n$, $\varepsilon_i = \langle \varepsilon, \psi_i\rangle_n$ and $x_i = \langle x_0, \phi_i\rangle$. It now suffices to divide each term by the known singular value b_i to observe the coefficient x_i, up to a noise term $\eta_i := b_i^{-1}\varepsilon_i$. Equivalently, this is obtained by applying the Moore–Penrose inverse $A_n^\dagger$ in the model (2.1). We thus consider the function $z = A_n^\dagger y \in \mathcal{K}_n^\perp$, defined as the inverse image of y through A_n with minimal norm. Identifying z with the vector of its coefficients $z_i = b_i^{-1} y_i$ in the basis $\{\phi_i\}_{i=1,\dots,n}$, we obtain

$$z_i = x_i + \eta_i, \ i = 1, \dots, n. \tag{2.2}$$

In this model, the noise $\eta = (\eta_1, \dots, \eta_n)^t$ is Gaussian with diagonal covariance matrix, as we have $\mathbb{E}(\eta_i\eta_j) = \frac{\sigma^2}{n} b_i^{-1} b_j^{-1} \langle \psi_i, \psi_j\rangle_n$ which is null for all $i \neq j$ and equal to $\sigma_i^2 := \frac{\sigma^2}{n}\ b_i^{-2}$ if $i = j$. This is an heteroscedastic sequential model, with the variances σ_i^2 inversely proportional to the eigenvalues b_i^2. This representation points out the effect of the decay of the singular values b_i on the noise level, making the problem ill-posed. To control the noise with a too large variance σ_i^2, a solution is to consider weighted versions of z. For some filter $\lambda = (\lambda_1, \dots, \lambda_n)^t$, note $\hat{x}(\lambda) \in \mathcal{K}_n^\perp$ the function defined by $\langle \hat{x}(\lambda), \phi_i\rangle = \lambda_i z_i$ for $i = 1, \dots, n$. Filter-based methods aim to cancel out the high frequency noises by allocating low weights to the components z_i corresponding to small singular values. A widely used example is the Tikhonov regularization, with weights of the form $\lambda_i = (1 + \tau\sigma_i^2)^{-1}$ for some $\tau > 0$. The Tikhonov solution can be expressed as the minimizer of the functional

$$\|y - A_n x\|^2 + \frac{\tau\sigma^2}{n}\ \|x\|^2, \ x \in \mathcal{X},$$

which makes the method particularly convenient in cases where the SVD of $A_n^* A_n$ or the coefficients z_i are not easily computable. We refer to [3, 18] for further details.

Another common filter-based method is the *truncated singular value decomposition* or *spectral cut-off* studied in [2, 8, 9]. An estimator of x_0 is obtained as a truncated version of z, where all coefficients z_i corresponding to arbitrarily small singular values are replaced by 0. This approach can be viewed as a principal component analysis, for which only the highly explanatory directions are selected. The spectral cut-off estimator is associated to filters of the form $\lambda_i = \mathbb{1}\{i \leq k\}$, where $\mathbb{1}\{.\}$ denotes the indicator function and k is a bandwidth to be determined. Data-driven methods for selecting suitable values of k are discussed in [3,4,9,19,20].

A natural way to generalize the spectral cut-off procedure is to enlarge the class of estimators by considering non-ordered truncated versions of z, as made in [12,13] or [14] (see also examples 1 and 2 in [6]). This approach reduces to a model selection issue where each model is identified with a set of indices $m \subseteq \{1,\dots,n\}$. Precisely, for m a given model, define $\hat{x}_m \in \mathcal{K}_n^\perp$ and $x_m \in \mathcal{K}_n^\perp$ as the orthogonal projections of z and x_0 respectively onto $\mathcal{X}_m := \operatorname{span}\{\phi_i, i \in m\}$, that is

$$\langle \hat{x}_m, \phi_i \rangle = \begin{cases} z_i & \text{if } \ i \in m, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \langle x_m, \phi_i \rangle = \begin{cases} x_i & \text{if } \ i \in m, \\ 0 & \text{otherwise.} \end{cases}$$

The objective is to find a model m that makes the expected risk $\mathbb{E}\|\hat{x}_m - x_0\|^2$ small. The computation of the estimator no longer relies on the choice of one parameter $k \in \{1,\dots,n\}$ as for spectral cut-off, but on the choice of a set of indices $m \subseteq \{1,\dots,n\}$, which increases the number of possibilities. In particular, this approach allows non-monotonic collections of filters that may perform better than decreasing sequences obtained by spectral cut-off. To see this, write the bias-variance decomposition of the estimator $\hat{x}_m$ for a deterministic model m, $\mathbb{E}\|x_0 - \hat{x}_m\|^2 = \|x_0 - x_m\|^2 + \mathbb{E}\|x_m - \hat{x}_m\|^2$, which follows by

$$\mathbb{E}\|\hat{x}_m - x_0\|^2 = \mathbb{E}\|x_0 - x^\dagger\|^2 + \sum_{i \notin m} x_i^2 + \sum_{i \in m} \sigma_i^2. \tag{2.3}$$

In these settings, it appears that in order to minimize the risk, best is to select indices i for which the component x_i^2 is larger than the noise level σ_i^2. Thus, a proper choice of filter should reasonably depend on both the variance σ_i^2 and the coefficient x_i^2. Consequently, the resulting sequence $\{\lambda_i\}_{i=1,\dots,n}$ has no reason of being a decreasing function of σ_i^2 if some coefficients x_i^2 are large enough to compensate for a large variance.

3. Hard-thresholding regularization

The construction of the projection estimator reduces to finding a proper set m. An optimal value for m (minimizing the risk) is obtained by keeping small simultaneously the bias term $\sum_{i \notin m} x_i^2$ and the variance term $\sum_{i \in m} \sigma_i^2$ in the expression of the risk $\mathbb{E}\|\hat{x}_m - x_0\|^2$. Following this argument, a minimizer of the risk $\mathbb{E}\|\hat{x}_m - x_0\|^2$ is obtained by selecting only the indices i for which the coefficient x_i^2 is larger than the noise level σ_i^2. An optimal model is thus given by $m^* := \{i : \ x_i^2 \geq \sigma_i^2\}$. The coefficients x_i being unknown to the practitioner, the optimal set m^* can not be computed in practical cases. For this reason it is referred to as an *oracle*.

We shall now provide a model $\widehat{m}$ constructed from the available information, that mimics the oracle m^*. Fixing a threshold on the coefficients x_i being impossible, we propose to use a threshold on the coefficients z_i. Precisely, consider the set

$$\widehat{m} = \{i : z_i^2 \geq 4\sigma_i^2 \mu_i\} = \left\{ i : y_i^2 \geq \frac{4\sigma^2 \mu_i}{n} \right\},$$

for $\{\mu_i\}_{i=1,\dots,n}$ a sequence of positive parameters to be chosen and where we recall that $y_i = b_i z_i$. Obviously, the behavior of the resulting estimator $\hat{x}_{\widehat{m}}$ relies on the choice of the sequence $\{\mu_i\}_{i=1,\dots,n}$: the larger the μ_i's, the more sparse is $\hat{x}_{\widehat{m}}$. It must be chosen so that the resulting set $\widehat{m}$ contains only the indices i for which the noise level is small compared to the actual value of x_i, but the only knowledge of the observations z_i and the variances σ_i^2 makes it a difficult task.

A number of thresholding procedures have been studied in the inverse problem literature. In [13], Loubes proposes a ℓ^1-penalization procedure to the inverse problem, corresponding to a soft-thresholding approach with a threshold on y_i^2 of the order $c \, \frac{\log n}{n} \, \sigma^2$, for some $c > 0$. In [1], Abramovich and Silverman discuss an approach based on the decomposition of the observation in a wavelet basis, for which the coefficients can be selected *via* a thresholding criterion. Here again, a threshold of the order $c \, \frac{\log n}{n} \, \sigma^2$ is suggested. For these two approaches, the threshold is a linear function of the variance, which with our notations, corresponds to taking a parameter

$\mu_i = c \log n$ that does not depend on the index i. In Theorem 3.1, we discuss the accuracy of using a non-linear threshold.

3.1. Oracle inequalities

In the definition of $\widehat{m}$, the choice of the parameters μ_i is crucial. Too large values of μ_i will result in an under-adjustment, keeping too few relevant components z_i to estimate x_0. On the contrary, a small value of μ_i increases the probability of selecting a component z_i that is highly affected with noise. Thus, it is essential to find a good balance between these two types of errors.

We introduce the notation $\kappa_n = \sup_{i\in m^*} b_i^{-2}$. Remark that the sequence $\{n^{-1}\kappa_n\}_{n\in\mathbb{N}}$ is bounded by $\|x_0\|^2$. Besides, the condition $\kappa_n = o(n)$ is actually quite mild, as it occurs for instance in the ideal case where the SVD of $A_n^* A_n$ does not depend on n, or if it converges in a weak sense toward the SVD of A^*A (typically, if the sequence $\{b_i, \phi_i\}_{n\in\mathbb{N}}$ converges in $\mathbb{R}\times\mathcal{X}$ as $n\to\infty$, for all i).

For $i = 1, \dots, n$, we note $\gamma_i := \eta_i^2/\sigma_i^2 = n\varepsilon_i^2/\sigma^2$, which have χ^2 distribution with one degree of freedom. Moreover, we use the notation $a \vee b = \max\{a, b\}$.

Theorem 3.1. *For $\theta, \beta > 0$, let $\mu_i = 1 \vee 2\beta \log(\theta^{-1} b_i^{-2})$. The estimator $\hat{x}_{\widehat{m}}$ satisfies*

$$\mathbb{E}\|\hat{x}_{\widehat{m}} - x_0\|^2 \le \|x_{m^*} - x_0\|^2 + \left(1 + 2\sqrt{1 \vee 2\beta\log(\theta^{-1}\kappa_n)}\right)^2 \sum_{i\in m^*} \sigma_i^2 + \frac{2\sigma^2\theta^\beta}{n} \sum_{i\notin m^*} b_i^{2(\beta-1)}.$$

The advantage of the method is that suitable values of θ and β can be chosen prior to the observations, based only on the degree of ill-posedness, *i.e.* on the behavior of the singular values b_i^{-2}. In order to control the residual term, we propose to take θ and β such that

$$\theta^\beta \sum_{i=1}^n b_i^{2(\beta-1)} = o\left(n^\delta\right), \ \forall \delta > 0, \tag{3.1}$$

while θ is chosen sufficiently large to keep the term $\log(\theta^{-1}\kappa_n)$ small. Let us discuss some examples. In the literature, we usually distinguish three main kinds of inverse problems. The problem can be *well-posed*, meaning that the eigenvalues of $(A_n^* A_n)^\dagger$ are bounded, $b_i^{-2} = O(1)$. In this case, one may take $\beta = 1$ and $\theta \sim 1/n$ (or $\theta \sim \log n/n$), recovering usual thresholds of order $\log n$ used in direct problems. If the problem is *mildly ill-posed*, *i.e.* if the singular values grow polynomially, $b_i^{-2} = O(i^{2t})$ for some $t > 0$, we propose to take $\beta = 1 + \frac{1}{t}$ and $\theta \sim \log n$ yielding

$$\theta^\beta \sum_{i=1}^n b_i^{2(\beta-1)} = O\left((\log n)^{\beta+1}\right),$$

thus satisfying (3.1). Finally, if the inverse problem is *severely ill-posed*, *i.e.* if the singular values grow exponentially, $b_i^{-2} = O(\mathrm{e}^{2it})$ for some $t > 0$, the condition (3.1) is fulfilled for $\beta > 1$ and $\theta \sim \log n$. We will show in Section 3.3 that these parameter values provide adaptive estimators.

The estimator $\hat{x}_{\widehat{m}}$ being built using binary filters $\lambda_i \in \{0, 1\}$, it is natural to measure its efficiency by comparing its risk to that of the best linear estimator in this class. Nevertheless, we see in the next corollary that a similar oracle inequality holds if we consider the oracle in the maximal class of filters, that is, allowing the λ_i's to take any real value.

Corollary 3.2. *If θ and β are chosen such that (3.1) holds, the estimator $\hat{x}_{\widehat{m}}$ of Theorem 3.1 satisfies for all $\delta > 0$,*

$$\mathbb{E}\|\hat{x}_{\widehat{m}} - x^\dagger\|^2 \le 2\left(1 + 2\sqrt{1 \vee 2\beta\log(\theta^{-1}\kappa_n)}\right)^2 \inf_{\lambda\in\mathbb{R}^n} \mathbb{E}\|\hat{x}(\lambda) - x^\dagger\|^2 + o\left(n^{-1+\delta}\right).$$

This result is a straightforward consequence of Lemma A.2 in the Appendix, where it is shown that the oracle in the class of binary filters $\lambda_i \in \{0, 1\}$ achieves the same rate of convergence up to a factor 2, as the best filter estimator obtained with non-random values of λ. This results points out that the class of unrestricted binary filters only induces a slight loss of efficiency compared to the maximal class.

3.2. Comparison with unbiased risk estimation and risk hull method

In a general point of view, the estimator $\hat{x}_{\widehat{m}}$ can be obtained *via* a minimization procedure, using a BIC-type criterion for heteroscedastic models,

$$\hat{x}_{\widehat{m}} = \arg\min_{x \in \mathcal{X}} \left\{ \|z - x\|^2 + 4 \sum_{i=1}^{n} \sigma_i^2 \mu_i \mathbb{1}\{\langle x, \phi_i \rangle \neq 0\} \right\}.$$

However, expressing the estimator as the solution to a minimization problem does not ease the computation. The method requires in any case calculation of the SVD of $A_n^* A_n$ and the coefficients z_i, which may be computationally expensive. On the other hand, the computation of the estimator is simple once the decomposition of z in the SVD of $A_n^* A_n$ is known, as it suffices to compare each coefficient z_i^2 to the threshold $4\sigma_i^2 \mu_i$.

Let us compare our approach to general methods dealing with arbitrary classes of filters. First, we discuss the *unbiased risk estimation* (URE) introduced in [17] and studied in [6] in the inverse problem framework. The method constructs an estimator of x_0 *via* the minimization of an unbiased estimate of the risk, over an arbitrary set Λ of filters. When restricted to the class of projection filters $\lambda_i \in \{0, 1\}$, unbiased risk estimation reduces to minimizing over the collection $\mathcal{M}$ of all subsets of $\{1, \ldots, n\}$, the criterion

$$m \mapsto \|z - \hat{x}_m\|^2 + 2 \sum_{i \in m} \sigma_i^2.$$

The minimum is achieved for the set $m = \{i : z_i^2 \geq 2\sigma_i^2\}$, which corresponds to taking $\mu_i = 1/2$. This choice is shown to be unadapted to the class of projection filters $\lambda_i \in \{0, 1\}$ in Proposition 2 in [6], yielding a residual term of the order of the constant.

A good alternative to URE is the *risk hull method* (RHM) discussed in [4]. Rather than considering an unbiased estimate of the risk, the idea of RHM is to find a function $\ell(\lambda)$ that bounds the risk from above, uniformly over the class Λ of filters. So, let $\ell(.)$ be such that

$$\mathbb{E} \sup_{\lambda \in \Lambda} \left\{ \|\hat{x}(\lambda) - x^\dagger\|^2 - \ell(\lambda) \right\} \leq 0. \tag{3.2}$$

The estimator is then defined *via* the minimizer $\tilde{\lambda}$ of $\lambda \mapsto \ell(\lambda)$ over Λ. By the previous inequality, we obtain an upper bound of the risk by

$$\mathbb{E}\|\hat{x}(\tilde{\lambda}) - x^\dagger\|^2 \leq \mathbb{E}\ \ell(\tilde{\lambda}) \leq \min_{\lambda \in \Lambda}\ \mathbb{E}\ \ell(\lambda).$$

The *risk hull* ℓ has to be chosen as small as possible, while still satisfying (3.2), in order to obtain a sharp bound on the risk of the estimator $\hat{x}(\tilde{\lambda})$. An analytic form of the minimal risk hull may be difficult to obtain but it can be computed by Monte-Carlo (see [4]). In the class of projection filters where all filter λ can be canonically identified with a model $m \subseteq \mathcal{M}$, the objective is to find $\ell : \mathcal{M} \to \mathbb{R}$ such that

$$\mathbb{E} \sup_{m \in \mathcal{M}} \left\{ \sum_{i \notin m} x_i^2 + \sum_{i \in m} \eta_i^2 - \ell(m) \right\} \leq 0.$$

Although it is not necessarily minimal, convenient is to consider a risk hull of the form $\ell(m) = \delta + \sum_{i \notin m} x_i^2 + \sum_{i \in m} c_i$, where $\delta \geq 0$ is a tolerance term and the c_i's are such that

$$\mathbb{E} \sup_{m \in \mathcal{M}} \left\{ \sum_{i \in m} (\eta_i^2 - c_i) \right\} = \sum_{i=1}^{n} \mathbb{E}\left[(\eta_i^2 - c_i) \mathbb{1}\{\eta_i^2 \geq c_i\} \right] \leq \delta,$$

in order to recover (3.2). Of course, the true coefficients x_i^2 are unknown, but they can be replaced by their unbiased estimates $z_i^2 - \sigma_i^2$, as suggested in [4]. It appears that taking $c_i \sim c \log n\ \sigma_i^2$ yields a δ of the order $n^{-\alpha} \sum_{i=1}^{n} \sigma_i^2$ for some $\alpha \geq 0$. On the other hand, adding a term $\log \sigma_i^2$ in the expression of c_i enables to obtain a tolerance term δ that does not involve the variances σ_i^2 (see for instance the proof of Lemma A.1), which somehow justifies the choice of the threshold used in Theorem 3.1.

3.3. Simulations

We shall now see numerical applications. We consider an heteroscedastic sequential model,

$$y_i = x_i + \sigma_i \varepsilon_i, \ \ i = 1, \dots, n,$$

with $\varepsilon_i \sim \mathcal{N}(0,1)$. This model illustrates the inverse problem, where the observation is expressed *via* the singular value decomposition of the operator A_n. So, the x_i's stand for the coefficients of x_0 in the singular basis $\{\phi_i\}$, *i.e.* $x_i = \langle x_0, \phi_i \rangle$. The noises ε_i are independently drawn from a standard Gaussian distribution. The variance of the model is determined by the non-decreasing sequence $\{\sigma_i^2\}_{i=1,\dots,n}$ which reflects the decay of the spectrum of $A_n A_n^*$. For now, we do not need to specify the value of basis $\{\phi_i\}$, as it is not directly involved in the model. Consequently, the function of interest x_0 is not fully determined. Nevertheless, this framework covers several possible values for x_0, depending on the underlying value of the operator A_n.

We calculate the risks of the following statistics.

- $\hat{x}_{m^*}$ is the optimal linear projection estimator defined in Section 3.2, obtained with the filters $\lambda_i = \mathbb{1}\{x_i^2 \geq \sigma_i^2\}$;
- $\hat{x}^*_{\text{sco}}$ is the best spectral cut-off estimator, obtained with the filters $\lambda_i = \mathbb{1}\{i \leq k^*\}$, with optimal bandwidth $k^* \in \{0, \dots, n\}$;
- $\hat{x}^*_{\text{lin}}$ is the best estimator computed with linear thresholds, obtained with the filters $\lambda_i = \mathbb{1}\{z_i^2 \geq \tau^* \sigma_i^2\}$, with optimal tuning parameter $\tau^* \geq 0$;
- $\hat{x}^*_{\text{th}}$ is the estimator of Theorem 3.1 obtained with optimal tuning parameter θ^*;
- $\hat{x}_{\widehat{m}}$ is the estimator of Theorem 3.1 obtained with tuning parameter $\theta = \log n$.

For the last two estimators, the value of the parameter β is taken such that $\sum_{i=1}^n b_i^{2(\beta-1)} = O(\log n)$ as suggested in Section 3.1. Precisely, we take $\beta = 1 + t^{-1}$ if the problem is mildly ill-posed with degree of ill-posedness t and we take $\beta = 1.1$ if the problem is severely ill-posed. The risks of the estimators are calculated by Monte Carlo with 10 000 replications of the procedure. We consider two mildly ill-posed situations with a polynomial growth of the variances, $b_i^{-2} \asymp i^2$ and $b_i^{-2} \asymp i^3$, and a severely ill-posed problem with $b_i^{-2} \asymp 2^i$ (for the examples treated here, the notation $b_i^{-2} \asymp u_i$ simply means that $b_i^{-2} = cu_i$ for some positive constant c). For sake of objectivity, the coefficients x_i are randomly drawn from independent centered Gaussian variables $x_i \sim \mathcal{N}(0, v_i^2)$, with variances v_i^2 to be made precise. The coefficients x_i are drawn once and for all and are treated as non-random. This means that the risk of an estimator $\mathbb{E}\|\hat{x} - x_0\|^2$ is to be understood as an expectation conditionally to the values of the x_i's.

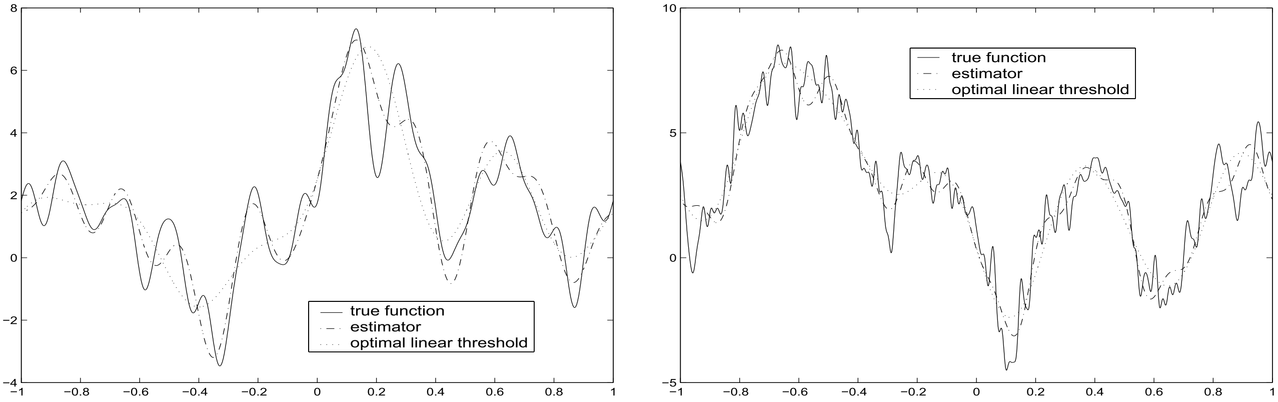
Case 1. The x_i's are drawn beforehand from Gaussian distributions $x_i \sim \mathcal{N}(0, v_i^2)$ with $v_i = \frac{10}{i}$. We consider two sample sizes $n = 50$ and $n = 200$. The risks of the estimator and the oracles are given in the following tables.

$n = 50$	$b_i^{-2} \asymp i^2$	$b_i^{-2} \asymp i^3$	$b_i^{-2} \asymp 2^i$
$\hat{x}_{m^*}$	13.63	15.20	18.30
$\hat{x}^*_{\text{sco}}$	14.08	15.20	18.30
$\hat{x}^*_{\text{th}}$	15.42	15.28	18.60
$\hat{x}^*_{\text{lin}}$	22.53	16.96	22.54
$\hat{x}_{\widehat{m}}$	21.17	16.96	21.92

$n = 200$	$b_i^{-2} \asymp i^2$	$b_i^{-2} \asymp i^3$	$b_i^{-2} \asymp 2^i$
$\hat{x}_{m^*}$	9.84	15.80	14.29
$\hat{x}^*_{\text{sco}}$	13.18	15.90	14.29
$\hat{x}^*_{\text{th}}$	12.53	16.60	14.45
$\hat{x}^*_{\text{lin}}$	17.61	18.21	17.94
$\hat{x}_{\widehat{m}}$	17.79	18.20	17.79

Here, taking $v_i = \frac{10}{i}$ causes an attenuation in the signal corresponding to a decreasing trend of order i^{-2} in the coefficients x_i^2. This results in the function of interest being somehow correlated with the SVD of A_n, which makes the spectral cut-off method particularly efficient. We see that the estimator of Theorem 3.1 with optimal tuning parameter θ^* nearly achieves the same efficiency as the optimal spectral cut-off estimate. We remark moreover a gap between the risk of $\hat{x}^*_{\text{th}}$ and that of the best linear threshold estimate $\hat{x}^*_{\text{lin}}$. Finally, the estimator computed with the arbitrary value $\theta = \log n$ is rather satisfactory, as it performs as well as the best linear threshold. These remarks hold for all degrees of ill-posedness and for the two sample sizes.

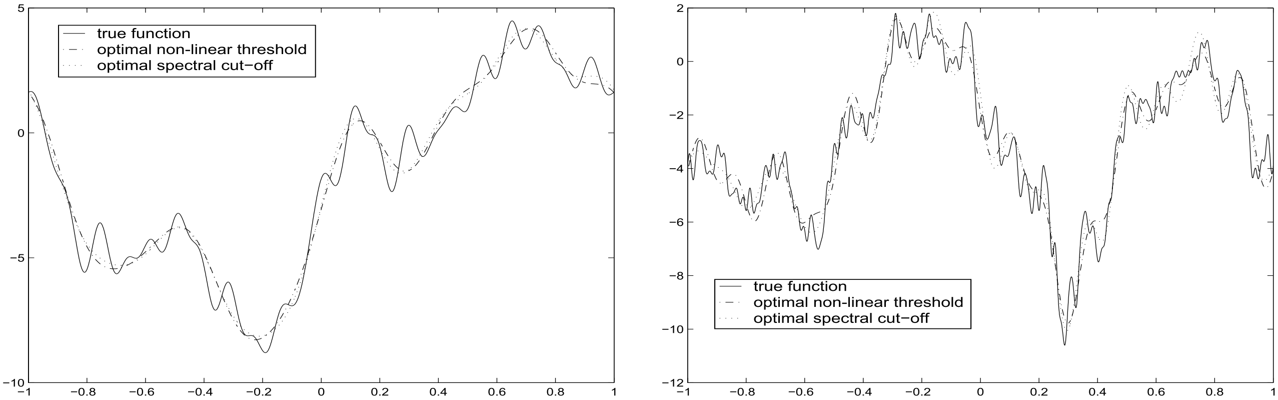
To illustrate these results, we consider an operator $A_n^* A_n$ with eigenfunctions $\{\phi_{2k} = \cos(k\pi.), \phi_{2k+1} = \sin(k\pi.), k \in \mathbb{N}\}$, forming an orthogonal system on $\mathbb{L}^2([-1;1])$. We assume that the coefficients x_i are the decomposition of the signal in this basis. We see in the graphics below a realization of the estimator and the linear threshold oracle when the problem is mildly ill-posed ($b_i^{-2} \asymp i^2$). The simulations are made both for $n = 50$ and $n = 200$.



$n = 50$ $n = 200$

Mildly ill-posed problem: $b_i^{-2} \asymp i^2, x_i \sim i^{-1}$.

On both graphics, we observe that the optimal linear threshold tends to be overly smooth while the non-linear threshold estimator (here obtained with the arbitrary value $\theta = \log n$) matches more the true function. In the following graphics, we compare realizations of the optimal spectral cut-off and the optimal non-linear threshold estimator obtained with optimal tuning parameter θ^*.



$n = 50$ $n = 200$

Mildly ill-posed problem: $b_i^{-2} \asymp i^2, x_i \sim i^{-1}$.

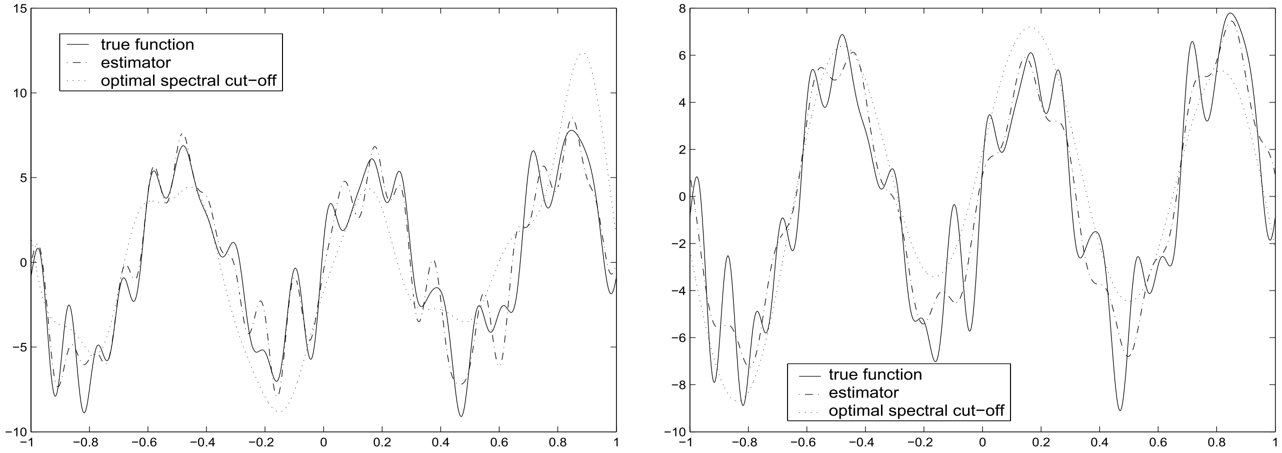
In this example, the coefficients x_i are computed in a way that the sequence $\{x_i\}$ tends to decrease as the variance σ_i^2 grows. As a result, the spectral cut-off procedure appears to be very efficient (with optimal bandwidth) because only the first coefficients are likely to belong in the optimal model m^*. Nevertheless, the above graphics show that the non-linear threshold estimator performs roughly as well with optimal tuning parameter θ^*.

We now consider an example where no condition is imposed on the trend of the sequence $\{x_i\}$. To do so, we draw a uniform random permutation $\rho(.)$ on the indices $i = 1, \ldots, n$ and we compute the coefficients $x_i \sim \mathcal{N}(0, v^2_{\rho(i)})$.

Case 2. The x_i's are drawn beforehand from Gaussian distributions $x_i \sim \mathcal{N}(0, v^2_{\rho(i)})$ with $v_i = \frac{10}{i}$ and $\rho(.)$ is a uniformly drawn permutation on $\{1, \ldots, n\}$. The risks of the estimator and the oracles are given in the following table for $n = 50$.

$n = 50$	$b_i^{-2} \asymp i^2$	$b_i^{-2} \asymp i^3$	$b_i^{-2} \asymp 2^i$
$\hat{x}_{m^*}$	3.59	9.81	16.63
$\hat{x}^*_{\mathrm{sco}}$	8.94	11.21	17.37
$\hat{x}^*_{\mathrm{th}}$	5.41	11.36	18.58
$\hat{x}^*_{\mathrm{lin}}$	5.94	12.89	19.25
$\hat{x}_{\widehat{m}}$	6.55	11.70	19.48

In this table, we observe a significant difference of efficiency between the spectral cut-off and the threshold procedure for a low degree of ill-posedness. This difference vanishes as the regularity decreases, as we see that the risk of the spectral cut-off estimator with optimal bandwidth is roughly equal to that of the non-linear threshold estimator obtained with optimal value of θ in the cases $b_i^{-2} \asymp i^3$ and $b_i^{-2} \asymp 2^i$. Here again, we observe that the efficiency of the estimator obtained for $\theta = \log n$ performs overall as well as the best linear threshold.



$b_i^{-2} \asymp i^2$ $b_i^{-2} \asymp i^3$

Mildly ill-posed problem: $x_i \sim \rho(i)^{-1}, n = 50$.

In this particular case, the spectral cut-off estimator remains overly smooth with optimal bandwidth, especially with a low degree of ill-posedness, as we see in the graphics above. The efficiency is increased using a threshold procedure, even with non-optimal values of the tuning parameter (here again, the estimator is computed taking $\theta = \log n$). We observe that the thresholding procedure becomes roughly as efficient as the spectral cut-off (with optimal tuning parameters) when the degree of ill-posedness increases, making the last coefficients x_i harder to estimate.

4. Regularization with unknown operator

We shall now discuss a situation where the operator A_n is not precisely known and is observed with a noise, independently from y. This situation is studied in [5, 7] or [11]. As in [5], we assume that the eigenvectors ϕ_i and ψ_i are known. This seemingly strong assumption is actually met in many situations, for instance if the problem involves convolution or differential operators which can be decomposed in Fourier basis (see also the examples in [3]). Thus, only the eigenvalues b_i are unknown and we assume they are observed independently of y, with a centered noise ξ_i with known variance $s^2 > 0$:

$$\hat{b}_i = b_i + \xi_i, \ i = 1, \dots, n.$$

The method discussed in this paper is different according to whether the eigenvalues are known exactly or observed with a noise. Thus, we need to assume here that s is positive and the known operator framework can not be seen as a particular case. Moreover, we assume that the ξ_i's are independent and satisfy the two following conditions.

A1. There exist $K, \alpha > 0$ such that $\forall t > 0, \forall i = 1, \dots, n, \ \mathbb{P}(\xi_i^2/s^2 > t) \leq K\mathrm{e}^{-t/\alpha}$.

A2. There exist $C, \delta > 0$ such that $\forall i = 1, \dots, n, \ \min\{\mathbb{P}(\xi_i < -\delta s), \mathbb{P}(\xi_i > \delta s)\} \geq C$.

The condition A1 simply states that the ξ_i's have finite exponential moments. The condition A2 is hardly restrictive, and is fulfilled for instance as soon as the ξ_i's are identically distributed. As we shall see in the sequel, the method requires knowledge of the constant δ (or at least an upper bound for it), but no information on the constants α, K or C is needed to build the estimator.

Knowing the eigenvectors of $A_n^* A_n$ allows us to write the model in the form

$$y_i = b_i x_i + \varepsilon_i, \ i = 1, \dots, n.$$

In our framework where the actual eigenvalues b_i are unknown, a natural estimator of each component x_i is obtained by $\tilde{z}_i = \hat{b}_i^{-1} y_i$, provided that $\hat{b}_i \neq 0$. However, it is clear that this estimate is not satisfactory if $\hat{b}_i$ is far from the true value (consider for instance the extreme case where $\hat{b}_i = 0$ or if $\hat{b}_i$ and b_i are of opposite signs). Actually, the naive estimator $\hat{b}_i^{-1}$ can not be used efficiently to estimate b_i^{-1} because it may have an infinite variance. In [5], the authors fix a threshold w the estimate can not exceed and consider an estimator of b_i^{-1} equal to $\hat{b}_i^{-1}$ if $|\hat{b}_i| > 1/w$ and null otherwise. As we shall see below, we use the same idea here, where the threshold fixed on the $\hat{b}_i$'s is implicitly part of the variable selection process.

We can reasonably assume that null values of $\hat{b}_i$ do not provide any relevant information and can not be used to estimate x_0. Thus, to avoid considering trivial situations, we assume that all $\hat{b}_i$ are non-zero. In all generality, the $\tilde{z}_i$'s can be viewed as noisy observations of x_i by writing

$$\tilde{z}_i = x_i + \tilde{\eta}_i, \ i = 1, \dots, n,$$

with $\tilde{z}_i = \hat{b}_i^{-1} \langle y, \psi_i \rangle_n$ and $\tilde{\eta}_i = \hat{b}_i^{-1}(\varepsilon_i - \xi_i x_i)$, where we recall $\varepsilon_i = \langle \varepsilon, \psi_i \rangle_n$. As in the previous section, we propose a threshold procedure to filter out the observations $\tilde{z}_i$ that are potentially highly contaminated with noise. Here, the noise $\tilde{\eta}_i$ is more difficult to deal with because it depends on the unknown coefficient x_i.

Our objective is to find an optimal variable selection criterion conditionally to the $\hat{b}_i$'s. In order to do so, we consider a framework where the $\hat{b}_i$'s are observed once and for all, and are treated as non-random. Thus, we define as an oracle, a model m_ξ^* minimizing the conditional risk $\mathbb{E}_\xi \|\hat{x}_m - x_0\|^2$, where $\mathbb{E}_\xi(.)$ denotes the expectation knowing $\xi = (\xi_1, \dots, \xi_n)^t$. Following a similar argument as in the previous section, a model minimizing the conditional risk contains only the indices i for which the coefficient x_i^2 is larger than the noise level. Hence, we may define $m_\xi^* = \{i : x_i^2 > \mathbb{E}_\xi(\tilde{\eta}_i^2)\}$. A notable difference here is that the noise $\tilde{\eta}_i$ actually depends on the value x_i. We can calculate the conditional expectation of $\tilde{\eta}_i^2$, given by

$$\mathbb{E}_\xi(\tilde{\eta}_i^2) = \hat{\sigma}_i^2 + \hat{b}_i^{-2} \xi_i^2 x_i^2,$$

where we set $\hat{\sigma}_i^2 = \hat{b}_i^{-2}\sigma^2/n$. After simplifications, it appears that the optimal model conditionally to the ξ_i's can be expressed in the two following equivalent forms

$$m_\xi^* = \left\{ i : 2|\hat{b}_i| > \frac{\sigma^2}{n|b_i|x_i^2} + |b_i| \right\} = \left\{ i : x_i^2 > \frac{\sigma^2}{n(\hat{b}_i^2 - \xi_i^2)},\ |\hat{b}_i| > \frac{|b_i|}{2} \right\}.$$

In the first expression, we see that the oracle selects indices i for which the observation $\hat{b}_i$ exceeds a certain value depending on both x_i and b_i. Interestingly, components $\tilde{z}_i$ corresponding to observations $|\hat{b}_i|$ smaller than half the true eigenvalue $|b_i|$ are not selected in the oracle, regardless of the coefficient x_i. Here again, the optimal model m_ξ^* can not be used in practical cases since it involves the unknown values x_i and ξ_i. We can only try to mimic the optimal threshold, based on the observations $\tilde{z}_i$ and $\hat{b}_i$. Consider the set

$$\widehat{m}_\xi = \left\{ i : \tilde{z}_i^2 > 8\hat{\sigma}_i^2 \nu_i,\ |\hat{b}_i| > \delta s \right\},$$

where $\{\nu_i\}_{i=1,\dots,n}$ are parameters to be chosen and δ is the constant defined in A2. With this definition, only the indices for which the observation $\hat{b}_i$ is larger than a certain value, namely δs, are selected. This conveys the idea discussed in [5], that when b_i is small compared to the noise level, the observation $\hat{b}_i$ is potentially mainly noise. Remark however that in [5], the lower limit for the observed eigenvalues is $s\log^2(1/s)$, while in our method, it is chosen of the same order as the standard deviation s.

Define the set $M = \{i : |b_i| < 2\delta s\}$.

Theorem 4.1. *The threshold estimator obtained with* $\nu_i = 2\log(i\ \hat{b}_i^{-2})$ *satisfies,*

$$\mathbb{E}_\xi \|\hat{x}_{\widehat{m}_\xi} - x^\dagger\|^2 \le \left\{ 9 + 36\log\left(\frac{n\|x^\dagger\|}{\sigma}\right) \vee \frac{4\alpha}{\delta^2}\log n \right\} \mathbb{E}_\xi \|\hat{x}_{m_\xi^*} - x^\dagger\|^2 + \sum_{i\in M} x_i^2 + \Delta(\xi),$$

with

$$\Delta(\xi) = \frac{4\sigma^2(1+\log n)}{n} + 4\sum_{i\notin m_\xi^*} \frac{\xi_i^2 x_i^2}{\delta^2 s^2} \mathbb{1}\{\xi_i^2 > s^2\alpha\log n\}.$$

Moreover, if A1 *holds,* $\mathbb{E}(\Delta(\xi)) = O\left(\frac{\log n}{n}\right)$.

The threshold is chosen in order to control the conditional risk. Inspection of the proof shows that choosing a term $\nu_i = 2\log(i\ \hat{b}_i^{-2})$ involving the index i in the logarithm enables to control the variance regardless of the degree of ill-posedness and the nature of the inverse problem. The main interest of this result lies in the fact that it provides an oracle inequality, conditionally to the $\hat{b}_i$'s. In particular, the conditional oracle $\hat{x}_{m_\xi^*}$ performs better than the minimizer of the expected risk $m \mapsto \mathbb{E}\|\hat{x}_m - x^\dagger\|^2$, since the optimal set m_ξ^* is allowed to depend on the ξ_i's. We see that the estimator $\hat{x}_{\widehat{m}_\xi}$ performs almost as well as the conditional oracle. Indeed, the residual term $\Delta(\xi)$ is independent from ξ with high probability, and its expectation is negligible under A1 as pointed out in the theorem. The non-random term $\sum_{i\in M} x_i^2$ is small if the eigenvalues b_i are observed with a good precision, *i.e.* if the variance s^2 is small. Moreover, this term can be shown to be of the same order as the risk under the condition A2.

Corollary 4.2. *If the conditions* A1 *and* A2 *hold, the threshold estimator defined in Theorem 4.1 satisfies*

$$\mathbb{E}\|\hat{x}_{\widehat{m}_\xi} - x^\dagger\|^2 \le K_1 \log n\ \mathbb{E}\|\hat{x}_{m_\xi^*} - x^\dagger\|^2 + O\left(\frac{\log n}{n}\right),$$

for a constant K_1 *independent from* n *and* s.

With a noisy operator, we manage to provide an estimator that achieves the rate of convergence of the conditional oracle, regardless of the precision of the approximation of the spectrum of A_n. Indeed, the constant K_1 in Corollary 4.2 does not involve the variance s^2 of ξ. Actually, the variance only plays a role in the accuracy of the oracle. The result is non-asymptotic and requires no assumption on s^2.

Appendix A.

A.1. Technical lemmas

Lemma A.1. *For all $\mu_i \geq 0$, we have*

- $\mathbb{E}\left[(\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\}\right] \leq 2\sigma_i^2 \mathrm{e}^{-\mu_i/2}$;
- $\mathbb{E}\left[(x_i^2 - \eta_i^2)\mathbb{1}\{i \notin \widehat{m}\}\right] \leq \left[\left(2\sqrt{\mu_i}+1\right)^2 - 1\right]\sigma_i^2$.

Proof. Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we find that $\eta_i^2 - x_i^2 \leq 2\eta_i^2 - z_i^2/2$. By definition of $\widehat{m}$, we get

$$(\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\} \leq 2\sigma_i^2(\gamma_i - \mu_i)\mathbb{1}\{i \in \widehat{m}\} \leq 2\sigma_i^2(\gamma_i - \mu_i)\mathbb{1}\{\gamma_i \geq \mu_i\},$$

where we used that $X \leq X\mathbb{1}\{X \geq 0\}$. Since γ_i has $\chi^2(1)$ distribution, we get

$$\mathbb{E}\left[(\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\}\right] \leq \frac{2\sigma_i^2}{\sqrt{2\pi}}\int_{\sqrt{\mu_i}}^{\infty}(t^2 - \mu_i)\ \mathrm{e}^{-t/2}\ \mathrm{d}t \leq 2\sigma_i^2\mathrm{e}^{-\mu_i/2}.$$

For the second part write $x_i^2 - \eta_i^2 = z_i^2 - 2\eta_i z_i \leq z_i^2 + 2|\eta_i||z_i|$. Using that $\mathbb{E}|\eta_i| \leq \sigma_i$, we get

$$\mathbb{E}\left[(x_i^2 - \eta_i^2)\mathbb{1}\{i \notin \widehat{m}\}\right] \leq \left[(2\sqrt{\mu_i}+1)^2 - 1\right]\sigma_i^2. \qquad \square$$

Lemma A.2.

$$\inf_{m \in \mathcal{M}} \mathbb{E}\|\hat{x}_m - x^\dagger\|^2 \leq 2 \inf_{\lambda \in \mathbb{R}^n} \mathbb{E}\|\hat{x}(\lambda) - x^\dagger\|^2.$$

Proof. The minimal values of the expected risks can be calculated explicitly in the two classes considered here. Minimizing over $\mathbb{R}^n$ the function $\lambda \mapsto \mathbb{E}\|\hat{x}(\lambda) - x^\dagger\|^2$, we find that the optimal value of λ_i is reached for $\lambda_i^* = x_i^2/(x_i^2 + \sigma_i^2)$. On the other hand, we know that $m \mapsto \mathbb{E}\|\hat{x}_m - x^\dagger\|^2$ reaches its minimum at $m^* = \{i : x_i^2 \geq \sigma_i^2\}$, yielding

$$\inf_{\lambda \in \mathbb{R}^n} \mathbb{E}\|\hat{x}(\lambda) - x^\dagger\|^2 = \sum_{i=1}^n \frac{x_i^2\sigma_i^2}{x_i^2 + \sigma_i^2} \quad \text{and} \quad \inf_{m \in \mathcal{M}} \mathbb{E}\|\hat{x}_m - x^\dagger\|^2 = \sum_{i \in m^*}\sigma_i^2 + \sum_{i \notin m^*} x_i^2.$$

By definition, if $i \in m^*$, $x_i^2/(x_i^2 + \sigma_i^2) \geq \frac{1}{2}$. In the same way, $\sigma_i^2/(x_i^2 + \sigma_i^2) \geq \frac{1}{2}$, for all $i \notin m^*$. We conclude by summing all the terms. $\square$

Lemma A.3. *For all $\nu_i \geq 0$, we have,*

- $\mathbb{E}_\xi\left[(\tilde{\eta}_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}_\xi\}\right] \leq 4\hat{\sigma}_i^2\mathrm{e}^{-\nu_i/2} + \frac{4\xi_i^2 x_i^2}{\delta^2 s^2}$;
- $\mathbb{E}_\xi\left[(x_i^2 - \tilde{\eta}_i^2)\mathbb{1}\{i \notin \widehat{m}_\xi\}\right] \leq 9\hat{\sigma}_i^2\nu_i + 8\mathbb{E}_\xi(\tilde{\eta}_i^2) + x_i^2\mathbb{1}\{|\hat{b}_i| \leq \delta s\}$.

Proof. Remark that $\tilde{\eta}_i^2 = \hat{b}_i^{-2}(\varepsilon_i - \xi_i x_i)^2 \leq 2\hat{b}_i^{-2}\varepsilon_i^2 + 2\hat{b}_i^{-2}\xi_i^2 x_i^2$. Using that $x_i^2 \geq \tilde{z}_i^2/2 - \tilde{\eta}_i^2$, we deduce

$$\tilde{\eta}_i^2 - x_i^2 \leq 4\hat{b}_i^{-2}\varepsilon_i^2 + 4\hat{b}_i^{-2}\xi_i^2 x_i^2 - \frac{\tilde{z}_i^2}{2}.$$

Writing $\widehat{m}_\xi = \{i : \tilde{z}_i^2 > 8\hat{\sigma}_i^2\nu_i,\ |\hat{b}_i| > \delta s\}$, we find

$$(\tilde{\eta}_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}_\xi\} \leq 4\hat{\sigma}_i^2(\gamma_i - \nu_i)\mathbb{1}\{\gamma_i \geq \nu_i\} + 4\hat{b}_i^{-2}\xi_i^2 x_i^2\mathbb{1}\{|\hat{b}_i| > \delta s\},$$

where we recall that $\gamma_i = n\varepsilon_i^2/\sigma^2$. Clearly, $\hat{b}_i^{-2}\mathbb{1}\{|\hat{b}_i| > \delta s\} < \delta^{-2}s^{-2}$ and the result follows using that $\gamma_i \sim \chi^2(1)$. For the second part of the lemma, remark that the complement of $\widehat{m}_\xi$ is $\{i : \tilde{z}_i^2 \leq 8\hat{\sigma}_i^2\nu_i,\ |\hat{b}_i| > \delta s\} \cup \{i : |\hat{b}_i| \leq \delta s\}$. Using that $x_i^2 - \tilde{\eta}_i^2 \leq (1 + a^{-1})\tilde{z}_i^2 + a\tilde{\eta}_i^2$, we get for $a = 8$,

$$(x_i^2 - \tilde{\eta}_i^2)\mathbb{1}\{i \notin \widehat{m}_\xi\} \leq 9\hat{\sigma}_i^2\nu_i + 8\tilde{\eta}_i^2 + x_i^2\mathbb{1}\{|\hat{b}_i| \leq \delta s\}. \qquad \square$$

Lemma A.4. *If* A1 *holds, we have*

$$\xi_i^2 \le s^2\alpha\log n + \xi_i^2 \mathbb{1}\{\xi_i^2 > s^2\alpha\log n\},$$

with $\mathbb{E}\left(\xi_i^2\mathbb{1}\{\xi_i^2 > s^2\alpha\log n\}\right) = O\left(\frac{\log n}{n}\right)$.

Proof. Write $\xi_i^2 \le s^2\alpha\log n\ \mathbb{1}\{\xi_i^2 \le s^2\alpha\log n\} + \xi_i^2\mathbb{1}\{\xi_i^2 > s^2\alpha\log n\}$. To bound the first term, we use the crude inequality $\mathbb{1}\{\xi_i^2 \le s^2\alpha\log n\} \le 1$. For the second term, we have as a consequence of A1,

$$\begin{aligned}\mathbb{E}\left[\xi_i^2\mathbb{1}\{\xi_i^2 > s^2\alpha\log n\}\right] &= \int_0^\infty \mathbb{P}\left(\xi_i^2\mathbb{1}\{\xi_i^2/s^2 > \alpha\log n\} > t\right)\ \mathrm{d}t\\ &= s^2\alpha\log n\ \mathbb{P}(\xi_i^2/s^2 > \alpha\log n) + s^2\int_{\alpha\log n}^\infty \mathbb{P}(\xi_i^2/s^2 > t)\ \mathrm{d}t\\ &\le \frac{K\alpha s^2(1+\log n)}{n}.\end{aligned}$$

□

A.2. Proofs

Proof of Theorem 3.1. Write

$$\|\hat{x}_{\widehat{m}} - x_0\|^2 = \|\hat{x}_{m^*} - x_0\|^2 + \sum_{i\notin m^*}(\eta_i^2 - x_i^2)\mathbb{1}\{i\in\widehat{m}\} + \sum_{i\in m^*}(x_i^2 - \eta_i^2)\mathbb{1}\{i\notin\widehat{m}\}.$$

The objective is to bound the terms $\mathbb{E}[(\eta_i^2 - x_i^2)\mathbb{1}\{i\in\widehat{m}\}]$ and $\mathbb{E}[(x_i^2-\eta_i^2)\mathbb{1}\{i\notin\widehat{m}\}]$ separately. By Lemma A.1, we know that $\mathbb{E}\left[(\eta_i^2 - x_i^2)\mathbb{1}\{i\in\widehat{m}\}\right] \le 2\sigma_i^2\mathrm{e}^{-\mu_i/2}$, which gives if $\mu_i = 2\beta\log\left(\theta^{-1}b_i^{-2}\right)$,

$$\mathbb{E}\left[(\eta_i^2 - x_i^2)\mathbb{1}\{i\in\widehat{m}\}\right] \le \frac{2\sigma^2\theta^\beta}{n}\ b_i^{2(\beta-1)}.$$

By Lemma A.1, the same result holds if $1 \ge 2\beta\log\left(\theta^{-1}b_i^{-2}\right)$ yielding $\mu_i = 1$. On the other hand, if $i\notin\widehat{m}$, then Lemma A.1 warrants

$$\mathbb{E}\left[(x_i^2-\eta_i^2)\mathbb{1}\{i\notin\widehat{m}\}\right] \le \left[(1+2\sqrt{\mu_i})^2 - 1\right]\sigma_i^2.$$

Using that $\mathbb{E}\|\hat{x}_{m^*} - x_0\|^2 = \|x_{m^*} - x_0\|^2 + \sum_{i\in m^*}\sigma_i^2$, we get by summing all the terms

$$\mathbb{E}\|\hat{x}_{\widehat{m}} - x_0\|^2 \le \|x_{m^*} - x_0\|^2 + \left(1 + 2\sqrt{1\vee 2\beta\log(\theta^{-1}\kappa_n)}\right)^2\sum_{i\in m^*}\sigma_i^2 + \frac{2\sigma^2\theta^\beta}{n}\sum_{i\notin m^*}b_i^{2(\beta-1)},$$

where we recall that $\kappa_n = \sup_{i\in m^*}b_i^{-2}$, yielding $\sup_{i\in m^*}\mu_i = 2\beta\log(\theta^{-1}\kappa_n\vee 1)$. □

Proof of Theorem 4.1. The proof starts as in Theorem 3.1. Write

$$\|\hat{x}_{\widehat{m}_\xi} - x^\dagger\|^2 = \|\hat{x}_{m_\xi^*} - x^\dagger\|^2 + \sum_{i\notin m_\xi^*}(\tilde{\eta}_i^2 - x_i^2)\mathbb{1}\{i\in\widehat{m}_\xi\} + \sum_{i\in m_\xi^*}(x_i^2 - \tilde{\eta}_i^2)\mathbb{1}\{i\notin\widehat{m}_\xi\},$$

and the objective is to bound the conditional expectation of each term separately. Using successively Lemmas A.3 and A.4, we get

$$\mathbb{E}_\xi\left[(\tilde{\eta}_i^2 - x_i^2)\mathbb{1}\{i\in\widehat{m}_\xi\}\right] \le \frac{4\sigma^2}{i.n} + \frac{4\xi_i^2x_i^2}{\delta^2s^2} \le \frac{4\alpha\log n}{\delta^2}\ x_i^2 + \Delta_i(\xi),$$

with

$$\Delta_i(\xi) = \frac{4\sigma^2}{i.n} + \frac{4\xi_i^2x_i^2}{\delta^2s^2}\mathbb{1}\{\xi_i^2 > s^2\alpha\log n\}.$$

Using the inequality $\sum_{i=1}^{n}\frac{1}{i}\leq 1+\log n$, we find that $\varDelta(\xi)=\sum_{i\notin m_\xi^*}\varDelta_i(\xi)$ satisfies

$$\mathbb{E}(\varDelta(\xi))\leq\frac{4(1+\log n)(\sigma^2+\delta^{-2}K\alpha\|x^\dagger\|^2)}{n}=O\left(\frac{\log n}{n}\right),$$

by Lemma A.4. On the other hand, Lemma A.3 gives

$$\mathbb{E}_\xi\left[(x_i^2-\tilde{\eta}_i^2)\mathbb{1}\{i\notin\widehat{m}_\xi\}\right]\leq 9\hat{\sigma}_i^2\nu_i+8\mathbb{E}_\xi(\tilde{\eta}_i^2)+x_i^2\mathbb{1}\{|\hat{b}_i|\leq\delta s\}.$$

For all $i\in m_\xi^*$, we know that $|\hat{b}_i|\geq|b_i|/2$. Thus, if $i\in m_\xi^*$, $\mathbb{1}\{|\hat{b}_i|\leq\delta s\}\leq\mathbb{1}\{i\in M\}$, where we recall $M=\{i:|b_i|<2\delta s\}$. We know also that, if $i\in m_\xi^*$, then $\nu_i>0$ and $\hat{b}_i^{-2}\leq nx_i^2/\sigma^2$, yielding $\nu_i=2\log(i\ b_i^{-2})\leq 4\log n+2\log(\|x^\dagger\|^2/\sigma^2)$. Since $\hat{\sigma}_i^2\leq\mathbb{E}_\xi(\tilde{\eta}_i^2)$, we have

$$\mathbb{E}_\xi\left[(x_i^2-\tilde{\eta}_i^2)\mathbb{1}\{i\notin\widehat{m}_\xi\}\right]\leq\left\{36\log\left(\frac{n\|x^\dagger\|}{\sigma}\right)+8\right\}\mathbb{E}_\xi(\tilde{\eta}_i^2)+x_i^2\mathbb{1}\{i\in M\}.$$

The result follows by summing all the terms, in view of the expression of the risk of the oracle

$$\mathbb{E}_\xi\|\hat{x}_{m_\xi^*}-x^\dagger\|^2=\sum_{i\notin m_\xi^*}x_i^2+\sum_{i\in m_\xi^*}\mathbb{E}_\xi(\tilde{\eta}_i^2).\qquad\square$$

Proof of Corollary 4.2. It suffices to show that the term $\sum_{i\in M}x_i^2$ is of the same order as the risk of the oracle. Write

$$\mathbb{E}\|\hat{x}_{m_\xi^*}-x^\dagger\|^2\geq\sum_{i=1}^{n}x_i^2\mathbb{P}(i\notin m_\xi^*)\geq\sum_{i=1}^{n}x_i^2\mathbb{P}(|\hat{b}_i|\leq|b_i|/2).$$

For all $i\in M$, the probability $\mathbb{P}(|\hat{b}_i|\leq|b_i|/2)$ is greater than C as a consequence of A2. We deduce $\sum_{i\in M}x_i^2\leq C^{-1}\mathbb{E}\|\hat{x}_{m_\xi^*}-x^\dagger\|^2$. □

References

[1] F. Abramovich and B.W. Silverman, Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85** (1998) 115–129.

[2] N. Bissantz, T. Hohage, A. Munk and F. Ruymgaart, Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.* **45** (2007) 2610–2636 (electronic).

[3] L. Cavalier, Nonparametric statistical inverse problems. *Inverse Problems* **24** (2008) 034004.

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

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

[6] L. Cavalier, G.K. Golubev, D. Picard and A.B. Tsybakov, Oracle inequalities for inverse problems. *Ann. Statist.* **30** (2000) 843–874.

[7] S. Efromovich and V. Koltchinskii, On inverse problems with unknown operators. *IEEE Trans. Inform. Theory* **47** (2001) 2876–2894.

[8] H.W. Engl, M. Hanke and A. Neubauer, *Regularization of inverse problems*, *Math. Appl.*, vol. 375. Kluwer Academic Publishers Group, Dordrecht (1996).

[9] P.C. Hansen, The truncated SVD as a method for regularization. *BIT* **27** (1987) 534–553.

[10] P.C. Hansen and D.P. O'Leary. The use of the *L*-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.* **14** (1993) 1487–1503.

[11] M. Hoffmann and M. Reiss, Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Statist.* **36** (2008) 310–336.

[12] J.M. Loubes, l1 penalty for ill-posed inverse problems. *Comm. Statist. Theory Methods* **37** (2008) 1399–1411.

[13] J.M. Loubes and C. Ludeña, Adaptive complexity regularization for linear inverse problems. *Electron. J. Stat.* **2** (2008) 661–677.

[14] J.M. Loubes and C. Ludeña, Penalized estimators for non linear inverse problems. *ESAIM: PS* **14** (2010) 173–191.

[15] F. Natterer, *The mathematics of computerized tomography*, *Class. Appl. Math.,* vol. 32. SIAM, Philadelphia, PA (2001). Reprint of the 1986 original.

[16] J.A. Scales and A. Gersztenkorn, Robust methods in inverse theory. *Inverse Problems* **4** (1988) 1071–1091.

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

[18] A.N. Tikhonov and V.Y. Arsenin, *Solutions of ill-posed problems.* V.H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York (1977). Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.

[19] J.M. Varah, On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems. *SIAM J. Numer. Anal.* **10** (1973) 257–267. Collection of articles dedicated to the memory of George E. Forsythe.

[20] J.M. Varah, A practical examination of some numerical methods for linear discrete ill-posed problems. *SIAM Rev.* **21** (1979) 100–111.