Available online at www.sciencedirect.com

ScienceDirect

C. R. Acad. Sci. Paris, Ser. I 344 (2007) 399–402

COMPTES RENDUS MATHEMATIQUE

http://france.elsevier.com/direct/CRASS1/

Statistics/Probability Theory

ℓ^1 sparsity and applications in estimation

Jean-Michel Loubes

UMR CNRS 5149, Département de mathématiques et de modélisation, université Montpellier 2, 34095 Montpellier, France

Received 22 May 2006; accepted after revision 25 January 2007

Available online 12 March 2007

Presented by Paul Deheuvels

Abstract

In this Note, we study the asymptotic behaviour of a new class of penalized M-estimators, built with an ℓ^1 type penalty. We prove that adding an ℓ^1 constraint enables to construct *adaptive* estimators, in the sense that the estimators converge at the optimal rate of convergence without prior knowledge of the regularity of the function to be reconstructed. Moreover, we show how the usual issues in nonparametric estimation, such as density estimation, estimation of a regression function and inverse problem estimation can be handled with this methodology. ***To cite this article: J.-M. Loubes, C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

Résumé

Contrainte ℓ^1 et applications en estimation non-paramétrique. Nous étudions les propriétés asymptotiques d'une nouvelle classe de M-estimateurs pénalisés par une pénalité de type norme ℓ^1. Nous montrons que nous pouvons ainsi construire des estimateurs adaptatifs, c'est-à-dire convergeant à la vitesse optimale sans connaître la régularité de la fonction à estimer. Nous montrons que ce procédé général s'applique dans le cadre du modèle de régression, des problèmes inverses et pour l'estimation de densités. ***Pour citer cet article : J.-M. Loubes, C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

1. Introduction

Penalized empirical risk minimization procedures have been extensively studied in the nonparametric statistical literature and enable to construct a wide range of estimators, see [6,1] or [5] for a general overview. However, their main drawback is to heavily rely on optimal choices of the trade-off parameters, balancing the two contributions given on the one hand by the loss-function, and on the other hand by the penalty. Indeed, this optimal choice depends on regularity assumptions over the function to be estimated. In this Note, we present a new methodology which consists in minimizing a contrast function together with an ℓ^1 penalty. The sparsity property of the ℓ^1 norm enables to build adaptive estimators, converging at the optimal rate of convergence without any prior regularity assumption.

Sparsity is a familiar notion in statistics and beyond, which expresses the idea that the information of a signal is concentrated in few coefficients. ℓ^p norms track sparsity for $p < 2$, with smaller p giving more stringent measures. So adding an ℓ^1 penalty models the prior constraint that the signal has a sparse representation in the given basis, yet with

E-mail address: jean-michel.loubes@math.univ-montp2.fr.

doi:10.1016/j.crma.2007.01.027

more flexibility than a penalty on the number of nonzero coefficients. Contrary to differentiable penalties ($p \geqslant 2$), for which adaptivity implies selecting the smoothing sequence among a set of possible choices, there is an optimal choice of the trade-off parameter, which does not depend on the unknown regularity of the parameter of interest. Hence ℓ^1 norm penalty is used in estimation in [4], in classification in [3] and in inverse problems in [2].

We first present, in Section 2, the general penalized M-estimation procedure and provide a general inequality which stresses the properties of the ℓ^1 penalty. Then, we apply in Section 3 this methodology in nonparametric estimation.

2. M-estimation procedure with an ℓ^1 penalty

Consider $X_1, \ldots, X_n$ independent random observations with values in a measurable space $\mathcal{X}$. Let P_i be the distribution of X_i, depending on an unknown function f_0, lying in a metric space $\mathcal{F}$ endowed with the metric d. Our aim is to estimate this function f_0. Define $\bar{P} = \frac{1}{n}\sum_{i=1}^n P_i$ and let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ be the empirical distribution. Assume that there exists an orthonormal basis $\varphi_1, \ldots, \varphi_n$, with respect to the empirical measure and let $\|\cdot\|_n$ and $\langle\cdot,\cdot\rangle_n$ be respectively the empirical norm and the empirical scalar product. Then any function $f \in \mathcal{F}$ can be decomposed onto this basis as $f = \sum_{j=1}^n \alpha_j \varphi_j$ with $\alpha_j = \langle f, \varphi_j\rangle_n$. Write also $f_0 = \sum_{j=1}^n \alpha_{j,0}\varphi_j$.

For a loss function $\gamma_f : \mathcal{X} \to \mathbb{R}$ for any $f \in \mathcal{F}$, and a smoothing sequence λ_n^2, define the penalized M-estimator as

$$\hat{f}_n = \arg\min_{f\in\mathcal{F}}\left\{\int \gamma_f \, \mathrm{d}P_n + \lambda_n^2 \|f\|_1\right\}. \tag{1}$$

We want to prove that an ℓ^1 penalty enables us to find an optimal choice of the smoothing sequence without knowing the regularity of the true function f_0.

To find the rate of convergence of the estimator, we need two ingredients. First we need to study the asymptotic behaviour of $\int \gamma_f \, \mathrm{d}\bar{P}$. We assume that there exists a constant c such that

$$\forall f \in \mathcal{F}, \quad \int (\gamma_f - \gamma_{f_0})\, \mathrm{d}\bar{P} \geqslant c d^2(f, f_0). \tag{2}$$

Then, we need to control the behaviour of the empirical process $\sqrt{n}\int \gamma_f \, \mathrm{d}(P_n - \bar{P})$. More precisely we need to prove that there exists a constant C such that

$$\mathbf{P}\left(\left|\int \gamma_f \, \mathrm{d}(P_n - \bar{P})\right| \geqslant C\lambda_n^2 \|f - f_0\|_1\right) \to 0. \tag{3}$$

Let $f_\star \in \mathcal{F}$ be an oracle, i.e. an approximation of the function f_0 whose rate of convergence is known (and depends on the unknown smoothness of f_0). Now, consider the set of indices $\mathcal{J}_n$ of cardinality N_n, and note that the ℓ^1 penalty can be split into two terms $I_N(.)$ and $I_M(.)$ defined by

$$\|f\|_1 = \sum_{j=1}^n |\alpha_j| = \sum_{j\in\mathcal{J}_n} |\alpha_j| + \sum_{j\notin\mathcal{J}_n} |\alpha_j| := I_N(f) + I_M(f).$$

We obtain the following bound with high probability

$$c d^2(\hat{f}_n, f_0) \leqslant d^2(f_\star, f_0) + (1 + C)\lambda_n^2 I_N(f_\star - \hat{f}_n) + 2\lambda_n^2 I_M(f_\star). \tag{4}$$

This decomposition is the key to adaptive estimation. Indeed the distance between the estimator and the true function is bounded by an approximation term $d^2(f_\star, f_0)$, the bias of the oracle and two approximation terms $I_N(f_\star - \hat{f}_n)$ and $I_M(f_\star)$. The first term represents the approximation error for the coefficients in $\mathcal{J}_n$ while the second term stands for the remainder term of the oracle in $\mathcal{J}_n^c$. These two terms are balanced depending on the size of $\mathcal{J}_n$. We point out that this bound provides a control of the estimation error, with respect to the distance given by the observation design. It is not a major drawback since norm equivalence results under entropy conditions can be used to extend this result. We refer to [6] for more references.

3. Application to nonparametric estimation problems

3.1. Nonparametric regression

Consider $y_1, \dots, y_n$ real-valued observations from the standard regression model

$$y_i = \Phi(f_0)(t_i) + \epsilon_i, \quad i = 1, \dots, n, \tag{5}$$

where $f_0 \in \mathcal{F}$ is the function to be recovered, $\Phi : \mathcal{F} \to \mathcal{F}$ is a known selfadjoint operator and ϵ_i are i.i.d. realizations of an observation noise. Note $X_i = (y_i, t_i)$ and take P_n as the distribution of $X_1, \dots, X_n$.

3.2. Direct estimation model

If $\Phi = \mathrm{id}$, (5) is the classical regression model. Take $\gamma_f(y,t) = (y - f(t))^2$. Hence the ℓ^1 penalized estimator can be written as

$$\hat{f}_n = \arg \min_{f = \sum_{j=1}^n \alpha_j \varphi_j} \left\{ \frac{1}{n} \sum_{i=1}^n \left| y_i - f(t_i) \right|^2 + 2\lambda_n^2 \sum_{j=1}^n |\alpha_j| \right\}. \tag{6}$$

This estimator has been studied in [4]. In this case, for $\lambda_n^2 = c\sqrt{\frac{\log n}{n}}$, (4) can be written as

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 \leqslant 4\left(\|\alpha_* - \alpha_0\|_n^2 + 4\lambda_n^4 N_n\right).$$

Theorem 1. *If there exists a smoothness parameter s and a constant M such that $\sum_{j=1}^n \alpha_{j,0}^{2/(2s+1)} \leqslant M$, then we get*

$$\mathbf{P}\left(\|\hat{f}_n - f_0\|_n^2 \geqslant c n^{-2s/(2s+1)}\right) \leqslant c \exp\left[-\frac{\log n}{c^2}\right]. \tag{7}$$

3.3. Inverse model

In the following, suppose that Φ is a nontrivial linear operator. We will denote Φ^* its adjoint. As often Φ is not of full rank, so the singular value decomposition (SVD) is a useful tool. Let $(\lambda_j; \psi_j, \varphi_j)_{j \geqslant 1}$ be a singular system for a linear operator Φ, that is, $\Phi\varphi_j = \lambda_j \psi_j$ and $\Phi^* \psi_j = \lambda_j \varphi_j$; where $\{\lambda_j^2\}_{j \geqslant 1}$ are the nonzero eigenvalues of the selfadjoint operator $\Phi^*\Phi$ (and also of $\Phi\Phi^*$), considered in decreasing order. We can write

$$\Phi f = \sum_{j=1}^n \lambda_j \langle f, \varphi_j \rangle \psi_j, \qquad \Phi^* y = \sum_{j=1}^n \lambda_j \langle y, \psi_j \rangle \varphi_j.$$

Note that for large j, the term $1/\lambda_j$ grows to infinity. Thus, the *high frequency errors* are strongly amplified. This amplification measures the difficulty of the inverse problem, the faster the decay of the eigenvalues, the more difficult is the inverse problem. So we assume that there exists an index t such that $\lambda_j = \mathcal{O}(j^{-t})$ for some t, called the index of ill-posedness of the operator Φ.

Take $\lambda_n = (\mu_j)_{j=1,\dots,n}$ and γ_f such that the corresponding estimator is defined by

$$\hat{f}_n = \arg \min_{f = \sum_{j=1}^n \alpha_j \varphi_j \in \mathcal{F}} \left[\sum_{j=1}^n \left| \left\langle y - \Phi(f), \frac{\psi_j}{\lambda_j} \right\rangle_n \right|^2 + 2 \sum_{j=1}^n \mu_j |\alpha_j| \right] = \sum_{j=1}^n \hat{\alpha}_{j,n} \varphi_j. \tag{8}$$

For a choice $\mu_j = \frac{c}{\lambda_j}\sqrt{\frac{\log n}{n}}$, the bound (4) can be written as

$$\|\hat{f}_n - f_0\|_n^2 \leqslant c\|f_\star - f_0\|_n^2 + 4c\frac{\log n}{n} \sum_{j \in \mathcal{J}_n} \frac{1}{\lambda_j^2}. \tag{9}$$

Theorem 2. *If there are two parameters s and $0 \leqslant p \leqslant 2$ such that*

$$f_0 \in X_{s,p} = \left\{ f = \sum_j \alpha_j \varphi_j,\ \sum_{j=1}^{n} j^{p(s+1/2-1/p)} \alpha_j^p \leqslant 1 \right\},$$

then for ill-posed problems we get

$$\mathbf{P}\left[\|\hat{f}_n - f_0\|_n^2 \geqslant \left(\frac{n}{\log n}\right)^{-4s/(2s+2t+1)} \right] \leqslant c \exp\left[-\frac{\log n}{c^2} \right].$$

Here again, the estimator converges at the minimax rate of convergence for the sets $X_{s,p}$ for ill-posed inverse problems.

3.4. Density estimation

Suppose we observe $X_1, \ldots, X_n$ a random independent sample of X with unknown density $f_0 = \mathrm{d}\mathbf{P}\mathrm{d}\lambda \in \mathcal{F}$, a set of density. To every density $f \in \mathcal{F}$, we associate the variable $\gamma = \log f + b(\gamma)$ lying in the correspondent functional class Γ, with $b(\gamma) = \log \int \mathrm{e}^{\gamma(x)}\, \mathrm{d}\lambda(x)$. So we have $b(\gamma) - b(\gamma_0) = K(f, f_0)$, the Kullback–Leibler information. To apply the previous framework, we first project the model onto a finite approximation space V_{j_1} using a wavelet basis $(\psi_{jk})_{(j,k)}$. The estimator can be written

$$\hat{\gamma}_n = \arg \max_{\gamma = \sum_{j<j_1} \sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}} \left(\frac{1}{n} \sum_{i=1}^{n} \gamma(X_i) - b(\gamma) - \lambda_n^2 \|\gamma\|_1 \right).$$

Set $\gamma_\star$ the projection of γ_0 onto V_{j_1}. For $\lambda_n^2 \geqslant c\sqrt{\frac{\log n}{n}}$ with c a constant, (4) can be written as

$$\frac{\|\hat{\gamma}_n - \gamma_0\|^2}{1 + \mathrm{O}_{\mathbf{P}}(1)} + \lambda_n^2 \|\hat{\gamma}_n\|_1 \leqslant \lambda_n^2 \|\hat{\gamma}_n - \gamma_\star\|_1 + \lambda_n^2 \|\gamma_\star\|_1.$$

The following theorem proves the optimality of the estimation procedure for Besov spaces $B^s_{p\infty}([0,1]), s > 1/p$.

Theorem 3. *Assume that $\exists 0 < C < \infty$, $\sup_{\gamma \in \Gamma} |\gamma| \leqslant C$ and $\gamma_0 \in B^s_{p\infty}([0,1])$, with $s > 1/p$. For $2^{j_1} = \mathrm{O}(\frac{n}{\log n})$*

$$\|\hat{\gamma}_n - \gamma_0\|^2 = \mathrm{O}_{\mathbf{P}}\left(\frac{n}{\log n}\right)^{-2s/(2s+1)}.$$

References

[1] A. Berlinet, G. Biau, L. Rouvière, Optimal L_1 bandwidth selection for variable kernel density estimates, Statist. Probab. Lett. 74 (2005) 116–128.

[2] A. Cohen, M. Hoffmann, M. Reiß, Adaptive wavelet Galerkin methods for linear inverse problems, SIAM J. Numer. Anal. 42 (4) (2004) 1479–1501.

[3] B. Efron, T. Hastie, I. Johnstone, R. Tibschirani, LARS, Ann. Statist. 32 (2) (2004) 407–499.

[4] S. Loubes, J.-M. van de Geer, Statist. Neerlandica 56 (4) (2002) 454–479.

[5] B. Silverman, On the estimation of a probability density function by the maximum penalized likelihood method, Ann. Statist. 10 (1982) 795–810.

[6] S. van de Geer, Applications of Empirical Process Theory, Cambridge Univ. Press, Cambridge, 2000.