SPADES AND MIXTURE MODELS

FLORENTINA BUNEA

ALEXANDRE B. TSYBAKOV

MARTEN H. WEGKAMP

ABSTRACT. This paper studies sparse density estimation via ℓ_1 penalization (SPADES). We focus on estimation in high-dimensional mixture models and nonparametric adaptive density estimation. We show, respectively, that SPADES can recover, with high probability, the unknown components of a mixture of probability densities and that it yields minimax adaptive density estimates. These results are based on a general sparsity oracle inequality that the SPADES estimates satisfy.

MSC2000 Subject classification: Primary 62G08, Secondary 62C20, 62G05, 62G20

Keywords and phrases: Adaptive estimation, aggregation, lasso, minimax risk, mixture models, consistent model selection, nonparametric density estimation, oracle inequalities, penalized least squares, sparsity, statistical learning.

Acknowledgements: The research of Bunea and Wegkamp is partially supported by the National Science Foundation grant DMS 0706829. The research of Tsybakov is partially supported by the grant ANR-06-BLAN-0194 and by the PASCAL Network of Excellence. Part of the research was done while the authors were visiting the Isaac Newton Institute for Mathematical Sciences (Statistical Theory and Methods for Complex, High-Dimensional Data Programme) at Cambridge University during Spring 2008.

1. Introduction

Let $X_1, \ldots, X_n$ be independent random variables with common unknown density f in $\mathbb{R}^d$. Let $\{f_1, \ldots, f_M\}$ be a finite set of functions with $f_j \in L_2(\mathbb{R}^d), j = 1, \ldots, M$, called a dictionary. We consider estimators of f that belong to the linear span of $\{f_1, \ldots, f_M\}$. We will be particularly interested in the case where $M \gg n$. Denote by f_λ the linear combinations

$$\mathsf{f}_\lambda(x) = \sum_{j=1}^{M} \lambda_j f_j(x), \quad \lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M.$$

Let us mention some examples where such estimates are of importance.

Date: January 13, 2009.

- *Estimation in sparse mixture models.* Assume that the density f can be represented as a finite mixture $f = \mathsf{f}_{\lambda^*}$ where f_j are known probability densities and λ^* is a vector of mixture probabilities. The number M can be very large, much larger than the sample size n, but we believe that the representation is sparse, i.e., that very few coordinates of λ^* are non-zero, with indices corresponding to a set $I^* \subseteq \{1, \ldots, M\}$. Our goal is to estimate the weight vector λ^* by a vector $\widehat{\lambda}$ that adapts to this unknown sparsity and to identify I^*, with high probability.
- *Adaptive nonparametric density estimation.* Assume that the density f is a smooth function, and $\{f_1, \ldots, f_M\}$ are the first M functions from a basis in $L_2(\mathbb{R}^d)$. If the basis is orthonormal, a natural idea is to estimate f by an orthogonal series estimator which has the form $\mathsf{f}_{\tilde{\lambda}}$ with $\tilde{\lambda}$ having the coordinates $\tilde{\lambda}_j = n^{-1} \sum_{i=1}^n f_j(X_i)$. However, it is well known that such estimators are very sensitive to the choice of M, and a data-driven selection of M or thresholding is needed to achieve adaptivity (cf., e.g., [30, 21, 6]); moreover these methods have been applied with $M \le n$. We would like to cover more general problems where the system $\{f_j\}$ is not necessarily orthonormal, even not necessarily a basis, M is not necessarily smaller than n, but an estimate of the form $\mathsf{f}_{\widehat{\lambda}}$ still achieves, adaptively, the optimal rates of convergence.
- *Aggregation of density estimators.* Assume now that $f_1, \ldots, f_M$ are some preliminary estimators of f constructed from a training sample independent of $(X_1, \ldots, X_n)$, and we would like to aggregate $f_1, \ldots, f_M$. This means that we would like to construct a new estimator, the aggregate, which is approximately as good as the best among $f_1, \ldots, f_M$ or approximately as good as the best linear or convex combination of $f_1, \ldots, f_M$. General notions of aggregation and optimal rates are introduced in [27, 33]. Aggregation of density estimators is discussed in [31, 29, 28] and more recently in [5] where one can find further references. The aggregates that we have in mind here are of the form $\mathsf{f}_{\widehat{\lambda}}$ with suitably chosen weights $\widehat{\lambda} = \widehat{\lambda}(X_1, \ldots, X_n) \in \mathbb{R}^M$.

In this paper, we suggest a data-driven choice of $\widehat{\lambda}$ that can be used in all the examples mentioned above and also more generally. We define $\widehat{\lambda}$ as a minimizer of an ℓ_1-penalized criterion, that we call SPADES (SPArse Density EStimation). This method was introduced in [11]. The idea of ℓ_1 penalized estimation is widely used in the statistical literature, mainly in linear regression where it is usually referred to as the Lasso criterion [32, 12, 15, 18, 26]. For Gaussian sequence models or for regression with orthogonal design matrix the Lasso is equivalent to soft thresholding [14, 24]. Model selection consistency of the Lasso type linear regression estimators is treated in many papers including [26, 40, 39, 41, 25]. Recently, ℓ_1 penalized methods have been extended to nonparametric regression with general fixed or random design [8, 9, 10, 4], as well as to some classification and other more general prediction type models [22, 23, 35, 7].

In this paper we show that ℓ_1 penalized techniques can also be successfully used in density estimation. In Section 2 we give the construction of the SPADES estimates and we show that they satisfy general oracle inequalities in Section 3. In the remainder of the paper we discuss the implications of these results for two particular problems, identification of mixture components and adaptive nonparametric density estimation. For the application of SPADES in aggregation problems we refer to [11].

Section 4 is devoted to mixture models. A vast amount of literature exists on estimation in mixture models, especially when the number of components is known; see e.g. [36] for examples involving the EM algorithm. The literature on determining the number of mixture components is still developing, and we will focus on this aspect here. Recent works on the selection of the number of components (mixture complexity) are [20, 2]. A consistent selection procedure specialized to Gaussian mixtures is suggested in [20]. The method of [20] relies on comparing a nonparametric kernel density estimator with the best parametric fit of various given mixture complexities. Nonparametric estimators based on the combinatorial density method (see [13]) are studied in [2, 3]. These can be applied to estimating consistently the number of mixture components, when the components have known functional form. Both [20, 2] can become computationally infeasible when M, the number of candidate components, is large. The method proposed here bridges this gap and guarantees correct identification of the mixture components with probability close to 1.

In Section 4 we begin by giving conditions under which the mixture weights can be estimated accurately, with probability close to 1. This is an intermediate result that allows us to obtain the main result of Section 4, correct identification of the mixture components. We show that in identifiable mixture models, if the mixture weights are above the noise level, then the components of the mixture can be recovered with probability larger than $1-\varepsilon$, for any given small ε. Our results are non-asymptotic, they hold for any M and n. Since the emphasis here is on correct component selection, rather than optimal density estimation, the tuning sequence that accompanies the ℓ_1 penalty needs to be slightly larger than the one used for good prediction. The same phenomenon has been noted for ℓ_1 penalized estimation in regression and generalized regression model, see, e.g., [7].

Section 5 uses the oracle inequalities of Section 3 to show that SPADES estimates adaptively achieve optimal rates of convergence (up to a logarithmic factor) simultaneously on a large scale of functional classes, such as Hölder, Sobolev or Besov classes, as well as on the classes of sparse densities, i.e., densities having only a finite, but unknown, number of non-zero wavelet coefficients.

2. Definition of SPADES

Consider the $L_2(\mathbb{R}^d)$ norm

$$\|g\| = \left(\int_{\mathbb{R}^d} g^2(x)\,\mathrm{d}x \right)^{1/2}$$

associated with the inner product

$$< g, h >= \int_{\mathbb{R}^d} g(x)h(x)\,\mathrm{d}x$$

for $g, h \in L_2(\mathbb{R}^d)$. Note that if the density f belongs to $L_2(\mathbb{R}^d)$ and X has the same distribution as X_i, we have, for any $g \in L_2$,

$$< g, f >= \mathbb{E}g(X),$$

where the expectation is taken under f. Moreover

$$\|f - g\|^2 = \|f\|^2 + \|g\|^2 - 2 < g, f >= \|f\|^2 + \|g\|^2 - 2\mathbb{E}g(X). \tag{2.1}$$

In view of identity (2.1), minimizing $\|\mathsf{f}_\lambda - f\|^2$ in λ is the same as minimizing

$$\gamma(\lambda) = -2\mathbb{E}\mathsf{f}_\lambda(X) + \|\mathsf{f}_\lambda\|^2.$$

The function $\gamma(\lambda)$ depends on f but can be approximated by its empirical counterpart

$$\widehat{\gamma}(\lambda) = -\frac{2}{n}\sum_{i=1}^{n} \mathsf{f}_\lambda(X_i) + \|\mathsf{f}_\lambda\|^2.$$

This motivates the use of $\widehat{\gamma} = \widehat{\gamma}(\lambda)$ as the empirical criterion, see, for instance, [6, 30, 37].

We define the penalty

$$\mathrm{pen}(\lambda) = 2\sum_{j=1}^{M} \omega_j |\lambda_j| \tag{2.2}$$

with weights ω_j to be specified later, and we propose the following data-driven choice of λ:

$$
\begin{aligned}
\widehat{\lambda} &= \arg\min_{\lambda\in\mathbb{R}^M} \{\widehat{\gamma}(\lambda) + \text{pen}(\lambda)\} \\
&= \arg\min_{\lambda\in\mathbb{R}^M} \left\{ -\frac{2}{n}\sum_{i=1}^{n} \mathsf{f}_\lambda(X_i) + \|\mathsf{f}_\lambda\|^2 + 2\sum_{j=1}^{M}\omega_j|\lambda_j| \right\}.
\end{aligned}
\tag{2.3}
$$

Our estimator of density f that we will further call the *SPADES estimator* is defined by

$$f^{\spadesuit}(x) = \mathsf{f}_{\widehat{\lambda}}(x),\ \forall x \in \mathbb{R}^d.$$

It is easy to see that, for an orthonormal system $\{f_j\}$, the SPADES estimator coincides with the soft thresholding estimator whose components are of the form $\widehat{\lambda}_j = (1-\omega_j/|\tilde{\lambda}_j|)_+\tilde{\lambda}_j$ where $\tilde{\lambda}_j = n^{-1}\sum_{i=1}^n f_j(X_i)$ and $x_+ = \max(0,x)$. We see that in this case ω_j is the threshold for the jth component of a preliminary estimator $\tilde{\lambda} = (\tilde{\lambda}_1,\dots,\tilde{\lambda}_M)$.

The SPADES estimate can be easily computed by convex programming even if $M \gg n$. It retains the desirable theoretical properties of other density estimators, the computation of which may become problematic for $M \gg n$. We refer to [13] for a thorough overview on combinatorial methods in density estimation, to [34] for density estimation using support vector machines and to [6] for density estimates using penalties proportional to the dimension.

3. Oracle inequalities for SPADES

3.1. **Preliminaries.**

For any $\lambda \in \mathbb{R}^M$, let

$$J(\lambda) = \{j \in \{1,\dots,M\} : \lambda_j \neq 0\}$$

be the set of indices corresponding to non-zero components of λ and

$$M(\lambda) = |J(\lambda)| = \sum_{j=1}^{M} I\{\lambda_j \neq 0\}$$

its cardinality. Here $I\{\cdot\}$ denotes the indicator function. Furthermore, set

$$\sigma_j^2 = \text{Var}(f_j(X_1)), \quad L_j = \|f_j\|_\infty$$

for $1 \le j \le M$, where $\text{Var}(\zeta)$ denotes the variance of random variable ζ and $\|\cdot\|_\infty$ is the $L_\infty(\mathbb{R}^d)$ norm.

We will prove sparsity oracle inequalities for the estimator $\widehat{\lambda} = \widehat{\lambda}(\omega_1,\dots,\omega_M)$, provided the weights ω_j are chosen large enough. We first consider a simple choice:

$$\omega_j = 4L_j r(\delta/2) \tag{3.1}$$

where $0 < \delta < 1$ is a user-specified parameter and

$$r(\delta) = r(M, n, \delta) \quad = \quad \sqrt{\frac{\log(M/\delta)}{n}}. \tag{3.2}$$

The oracle inequalities that we prove below hold with a probability of at least $1-\delta$ and are non-asymptotic: they are valid for all integers M and n. The first of these inequalities is established under a coherence condition on the "correlations"

$$\rho_M(i,j) = \frac{< f_i, f_j >}{\|f_i\|\|f_j\|}, \quad i, j = 1, \ldots, M.$$

For $\lambda \in \mathbb{R}^M$, we define a local coherence number (called *maximal local coherence*) by

$$\rho(\lambda) = \max_{i \in J(\lambda)} \max_{j \neq i} |\rho_M(i,j)|,$$

and we also define

$$F(\lambda) = \max_{j \in J(\lambda)} \frac{\omega_j}{r(\delta/2)\|f_j\|}$$

and

$$G = \max_{1 \le j \le M} \frac{r(\delta/2)\|f_j\|}{\omega_j}.$$

3.2. **Main results.**

Theorem 1. *Assume that $L_j < \infty$ for $1 \le j \le M$. Then with probability at least $1-\delta$ for all $\lambda \in \mathbb{R}^M$ that satisfy*

$$16 G F(\lambda) \rho(\lambda) M(\lambda) \le 1 \tag{3.3}$$

and all $\alpha > 1$ and we have the following oracle inequality:

$$\|f^{\spadesuit} - f\|^2 + \frac{\alpha}{2(\alpha-1)} \sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j - \lambda_j| \le \frac{\alpha+1}{\alpha-1} \|\mathsf{f}_\lambda - f\|^2 + \frac{8\alpha^2}{\alpha-1} \{F(\lambda)G\}^2 r^2(\delta/2) M(\lambda).$$

Note that only a condition on the local coherence (3.3) is required to obtain the result of Theorem 1. However, even this condition can be too strong, because the bound on "correlations" should be *uniform* over $j \in J(\lambda), i \neq j$, cf. the definition of $\rho(\lambda)$. For example, this excludes the cases where the "correlations" can be relatively large for a small number of pairs (i,j) and almost zero for otherwise. To account for this situation, we suggest below another version of Theorem 1. Instead of maximal local coherence, we introduce *cumulative local coherence* defined by

$$\rho_*(\lambda) = \sum_{i \in J(\lambda)} \sum_{j > i} |\rho_M(i,j)|.$$

Theorem 2. *Assume that $L_j < \infty$ for $1 \le j \le M$. Then with probability at least $1-\delta$ for all $\lambda \in \mathbb{R}^M$ that satisfy*

$$16F(\lambda)G\rho_*(\lambda)\sqrt{M(\lambda)} \le 1 \tag{3.4}$$

and all $\alpha > 1$ we have the following oracle inequality:

$$\|f^{\spadesuit} - f\|^2 + \frac{\alpha}{2(\alpha-1)}\sum_{j=1}^{M}\omega_j|\widehat{\lambda}_j - \lambda_j| \le \frac{\alpha+1}{\alpha-1}\|\mathsf{f}_\lambda - f\|^2 + \frac{8\alpha^2}{\alpha-1}\{F(\lambda)G\}^2 r^2(\delta/2)M(\lambda).$$

Theorem 2 is useful when we deal with sparse Gram matrices $\Psi_M = (< f_i, f_j >)_{1\le i,j\le M}$ that have only a small number N of non-zero off-diagonal entries. This number will be called a *sparsity index* of matrix Ψ_M, and is defined as

$$N = |\{(i,j) : i,j \in \{1,\dots,M\},\ i > j \text{ and } \psi_M(i,j) \neq 0\}|,$$

where $\psi_M(i,j)$ is the (i,j)th entry of Ψ_M and $|A|$ denotes the cardinality of a set A. Clearly, $N < M(M+1)/2$. We therefore obtain the following immediate corollary of Theorem 2.

Corollary 1. *Let Ψ_M be a Gram matrix with sparsity index N. Then the assertion of Theorem 2 holds if we replace there (3.4) by the condition*

$$16F(\lambda)N\sqrt{M(\lambda)} \le 1. \tag{3.5}$$

We finally give an oracle inequality, which is valid under the assumption that the Gram matrix Ψ_M is positive definite. It is simpler to use than the above results when the dictionary is orthonormal or forms a frame. Note that the coherence assumptions considered above do not necessarily imply the positive definiteness of Ψ_M. Vice versa, the positive definiteness of Ψ_M does not imply these assumptions.

Theorem 3. *Assume that $L_j < \infty$ for $1 \le j \le M$ and that the Gram matrix Ψ_M is positive definite with minimal eigenvalue larger than or equal to $\kappa_M > 0$. Then, with probability at least $1-\delta$, for all $\alpha > 1$ and all $\lambda \in \mathbb{R}^M$ we have*

$$\|f^{\spadesuit} - f\|^2 + \frac{\alpha}{\alpha-1}\sum_{j=1}^{M}\omega_j|\widehat{\lambda}_j - \lambda_j| \ \le\ \frac{\alpha+1}{\alpha-1}\|\mathsf{f}_\lambda - f\|^2 + \left(\frac{8\alpha^2}{\alpha-1}\right)\frac{G(\lambda)}{n\,\kappa_M}, \tag{3.6}$$

where

$$G(\lambda) \triangleq \sum_{j\in J(\lambda)}\omega_j^2 = \frac{16\log(2M/\delta)}{n}\sum_{j\in J(\lambda)}L_j^2\,.$$

We can consider some other choices for ω_j without affecting the previous results. For instance,

$$\omega_j = 2\sqrt{2}\sigma_j r(\delta/2) + \frac{8}{3} L_j r^2(\delta/2) \tag{3.7}$$

or

$$\omega_j = 2\sqrt{2}T_j r(\delta/2) + \frac{8}{3} L_j r^2(\delta/2) \tag{3.8}$$

with

$$T_j^2 = \frac{2}{n}\sum_{i=1}^n f_j^2(X_i) + 2L_j^2 r^2(\delta/2).$$

yield the same conclusions. These modifications of (3.1) prove useful, for example, for situations where f_j are wavelet basis functions, cf. Section 5. The choice (3.8) of ω_j has an advantage of being completely data-driven.

Theorem 4. *Theorems 1–3 and Corollary 1 hold with the choices (3.7) or (3.8) for the weights ω_j without changing the assertions. They also remain valid if we replace these ω_j by any ω_j' such that $\omega_j' > \omega_j$.*

If ω_j is chosen as in (3.8), our bounds on the risk of SPADES estimator involve the random variables $(1/n)\sum_{i=1}^n f_j^2(X_i)$. These can be replaced in the bounds by deterministic values using the following lemma.

Lemma 1. *Assume that $L_j < \infty$ for $j = 1, \dots, M$. Then*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f_j^2(X_i) \le 2\mathbb{E}f_j^2(X_1) + \frac{4}{3}L_j^2 r^2(\delta/2),\ \forall j = 1, \dots, M\right) \ge 1 - \delta/2. \tag{3.9}$$

From Theorem 4 and Lemma 1 we find that, for the choice of ω_j as in (3.8), the oracle inequalities of Theorems 1–3 and Corollary 1 remain valid with probability at least $1 - 3\delta/2$ if we replace the ω_j in these inequalities by the expressions $2\sqrt{2}\tilde{T}_j r(\delta/2) + (8/3)L_j r^2(\delta/2)$ where $\tilde{T}_j = \left(2\mathbb{E}f_j^2(X_1) + (4/3)L_j^2 r^2(\delta/2)\right)^{1/2}$.

3.3. **Proofs.** We first prove the following preliminary lemma. Define the random variables

$$V_j = \frac{1}{n}\sum_{i=1}^n \{f_j(X_i) - \mathbb{E}f_j(X_i)\}$$

and the event

$$A = \bigcap_{j=1}^M \{2|V_j| \le \omega_j\}. \tag{3.10}$$

Lemma 2. *Assume that $L_j < \infty$ for $j = 1, \ldots, M$. Then for all $\lambda \in \mathbb{R}^M$ we have that, on the event A,*

$$\|f^{\spadesuit} - f\|^2 + \sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j - \lambda_j| \le \|\mathsf{f}_\lambda - f\|^2 + 4 \sum_{j \in J(\lambda)} \omega_j |\widehat{\lambda}_j - \lambda_j|. \tag{3.11}$$

Proof. By the definition of $\widehat{\lambda}$,

$$-\frac{2}{n}\sum_{i=1}^{n} \mathsf{f}_{\widehat{\lambda}}(X_i) + \|\mathsf{f}_{\widehat{\lambda}}\|^2 + 2\sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j| \le -\frac{2}{n}\sum_{i=1}^{n} \mathsf{f}_\lambda(X_i) + \|\mathsf{f}_\lambda\|^2 + 2\sum_{j=1}^{M} \omega_j |\lambda_j|$$

for all $\lambda \in \mathbb{R}^M$. We rewrite this inequality as

$$\begin{aligned}
\|f^{\spadesuit} - f\|^2 \;\; &\le \;\; \|\mathsf{f}_\lambda - f\|^2 - 2 < f, f^{\spadesuit} - \mathsf{f}_\lambda > + \frac{2}{n}\sum_{i=1}^{n} (f^{\spadesuit} - \mathsf{f}_\lambda)(X_i) + 2\sum_{j=1}^{M} \omega_j |\lambda_j| - 2\sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j| \\
&= \;\; \|\mathsf{f}_\lambda - f\|^2 + 2\sum_{j=1}^{M} \left(\frac{1}{n}\sum_{i=1}^{n} f_j(X_i) - \mathbb{E} f_j(X_i) \right) (\widehat{\lambda}_j - \lambda_j) \\
&\quad\; + 2\sum_{j=1}^{M} \omega_j |\lambda_j| - 2\sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j|.
\end{aligned}$$

Then, on the event A,

$$\|f^{\spadesuit} - f\|^2 \;\; \le \;\; \|\mathsf{f}_\lambda - f\|^2 + \sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j - \lambda_j| + 2\sum_{j=1}^{M} \omega_j |\lambda_j| - 2\sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j|.$$

Add $\sum_j \omega_j |\widehat{\lambda}_j - \lambda_j|$ to both sides of the inequality to obtain

$$\begin{aligned}
&\|f^{\spadesuit} - f\|^2 + \sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j - \lambda_j| \\
&\le \|\mathsf{f}_\lambda - f\|^2 + 2\sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j - \lambda_j| + 2\sum_{j=1}^{M} \omega_j |\lambda_j| - 2\sum_{j=1}^{M} \omega_j |\widehat{\lambda}_j| \\
&\le \|\mathsf{f}_\lambda - f\|^2 + 2\sum_{j \in J(\lambda)} \omega_j |\widehat{\lambda}_j - \lambda_j| + 2\sum_{j=1}^{M} \omega_j |\lambda_j| - 2\sum_{j \in J(\lambda)} \omega_j |\widehat{\lambda}_j| \\
&\le \|\mathsf{f}_\lambda - f\|^2 + 4\sum_{j \in J(\lambda)} \omega_j |\widehat{\lambda}_j - \lambda_j|
\end{aligned}$$

where we used that $\lambda_j = 0$ for $j \notin J(\lambda)$ and the triangle inequality. $\square$

For the choice (3.1) for ω_j, we find by Hoeffding's inequality for sums of independent random variables $\zeta_{ij} = f_j(X_i) - \mathbb{E} f_j(X_i)$ with $|\zeta_{ij}| \le 2L_j$ that

$$\mathbb{P}(A) \quad \le \quad \sum_{j=1}^{M} \mathbb{P}\{2|V_j| > \omega_j\} \le 2\sum_{j=1}^{M} \exp\left(-\frac{2n\omega_j^2/4}{8L_j^2}\right) = \delta.$$

Proof of Theorem 1. In view of Lemma 2, we need to bound $\sum_{j\in J(\lambda)} \omega_j |\widehat{\lambda}_j - \lambda_j|$. Set

$$u_j = \widehat{\lambda}_j - \lambda_j, \quad U(\lambda) \quad = \textstyle\sum_{j\in J(\lambda)} |u_j| \|f_j\|, \quad U = \sum_{j=1}^M |u_j|\|f_j\|.$$

Then, by the definition of $F(\lambda)$ and the Cauchy-Schwarz inequality

$$\sum_{j\in J(\lambda)} \omega_j |\widehat{\lambda}_j - \lambda_j| \quad \le \quad rF(\lambda)U(\lambda).$$

Since

$$\sum\sum_{i,j\notin J(\lambda)} < f_i, f_j > u_i u_j \ge 0,$$

we obtain

$$\begin{aligned}
(3.12) \quad \sum_{j\in J(\lambda)} u_j^2 \|f_j\|^2 \quad &= \quad \|f^\spadesuit - \mathsf{f}_\lambda\|^2 - \sum\sum_{i,j\notin J(\lambda)} u_i u_j < f_i, f_j > \\
&\quad -2\sum_{i\notin J(\lambda)} \sum_{j\in J(\lambda)} u_i u_j < f_i, f_j > - \sum\sum_{i,j\in J(\lambda),\ i\neq j} u_i u_j < f_i, f_j > \\
&\le \quad \|f^\spadesuit - \mathsf{f}_\lambda\|^2 + 2\rho(\lambda) \sum_{i\notin J(\lambda)} |u_i| \|f_i\| \sum_{j\in J(\lambda)} |u_j| \|f_j\| \\
&\quad +\rho(\lambda) \sum\sum_{i,j\in J(\lambda)} |u_i||u_j| \|f_i\| \|f_j\| \\
&= \quad \|f^\spadesuit - \mathsf{f}_\lambda\|^2 + 2\rho(\lambda)U(\lambda)U - \rho(\lambda)U^2(\lambda).
\end{aligned}$$

The left-hand side can be bounded by $\sum_{j\in J(\lambda)} u_j^2\|f_j\|^2 \ge U^2(\lambda)/M(\lambda)$ using the Cauchy-Schwarz inequality, and we obtain that

$$U^2(\lambda) \le \|f^\spadesuit - \mathsf{f}_\lambda\|^2 M(\lambda) + 2\rho(\lambda)M(\lambda)U(\lambda)U,$$

which immediately implies

$$U(\lambda) \quad \le \quad 2\rho(\lambda)M(\lambda)U + \sqrt{M(\lambda)}\|f^\spadesuit - \mathsf{f}_\lambda\|. \tag{3.13}$$

Hence, by Lemma 2, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\|f^{\spadesuit}-f\|^2+\sum_{j=1}^{M}\omega_j|\widehat{\lambda}_j-\lambda_j| \\
\le\ & \|\mathsf{f}_\lambda-f\|^2+4\sum_{j\in J(\lambda)}\omega_j|\widehat{\lambda}_j-\lambda_j| \\
\le\ & \|\mathsf{f}_\lambda-f\|^2+4rF(\lambda)U(\lambda) \\
\le\ & \|\mathsf{f}_\lambda-f\|^2+4rF(\lambda)\left\{2\rho(\lambda)M(\lambda)U+\sqrt{M(\lambda)}\|f^{\spadesuit}-\mathsf{f}_\lambda\|\right\} \\
\le\ & \|\mathsf{f}_\lambda-f\|^2+8F(\lambda)\rho(\lambda)M(\lambda)G\sum_{j=1}^{M}\omega_j|\widehat{\lambda}_j-\lambda_j|+4rF(\lambda)\sqrt{M(\lambda)}\|f^{\spadesuit}-\mathsf{f}_\lambda\|,
\end{aligned}$$

where $r=r(\delta/2)$. For all $\lambda\in\mathbb{R}^M$ that satisfy relation (3.3), we find that with probability exceeding $1-\delta$,

$$\begin{aligned}
\|f^{\spadesuit}-f\|^2+\frac{1}{2}\sum_{j=1}^{M}\omega_j|\widehat{\lambda}_j-\lambda_j| &\le \|\mathsf{f}_\lambda-f\|^2+4rF(\lambda)G\sqrt{M(\lambda)}\|f^{\spadesuit}-\mathsf{f}_\lambda\| \\
&\le \|\mathsf{f}_\lambda-f\|^2+2\left\{2rGF(\lambda)\sqrt{M(\lambda)}\right\}\|f^{\spadesuit}-f\| \\
&\quad +2\left\{2rGF(\lambda)\sqrt{M(\lambda)}\right\}\|\mathsf{f}_\lambda-f\|.
\end{aligned}$$

After applying the inequality $2xy\le x^2/\alpha+\alpha y^2$ $(x,y\in\mathbb{R},\ \alpha>0)$ for each of the last two summands, we easily find the claim. $\square$

Proof of Theorem 2. The proof is similar to that of Theorem 1. With

$$U_*(\lambda)=\sqrt{\sum_{j\in J(\lambda)}u_j^2\|f_j\|^2}$$

we obtain now the following analogue of (3.12):

$$\begin{aligned}
U_*^2(\lambda) &\le \|f^{\spadesuit}-\mathsf{f}_\lambda\|^2+2\rho_*(\lambda)\max_{i\in J(\lambda),j>i}|u_i|\|f_i\||u_j|\|f_j\| \\
&\le \|f^{\spadesuit}-\mathsf{f}_\lambda\|^2+2\rho_*(\lambda)U_*(\lambda)\sum_{j=1}^{M}|u_j|\|f_j\| \\
&= \|f^{\spadesuit}-\mathsf{f}_\lambda\|^2+2\rho_*(\lambda)U_*(\lambda)U.
\end{aligned}$$

Hence, as in the proof of Theorem 1, we have

$$U_*(\lambda)\ \le\ 2\rho_*(\lambda)U+\|f^{\spadesuit}-\mathsf{f}_\lambda\|,$$

and using the inequality $U_*(\lambda) \geq U(\lambda)/\sqrt{M(\lambda)}$ we find

$$U(\lambda) \quad \leq \quad 2\rho_*(\lambda)\sqrt{M(\lambda)}U + \sqrt{M(\lambda)}\|f^{\spadesuit} - \mathsf{f}_\lambda\|. \tag{3.14}$$

Note that (3.14) differs from (3.13) only in the fact that the factor $2\rho(\lambda)M(\lambda)$ on the right hand side is now replaced by $2\rho_*(\lambda)\sqrt{M(\lambda)}$. Up to this modification, the rest of the proof is identical to that of Theorem 1. □

Proof of Theorem 3. By the assumption on Ψ_M we have

$$\|\mathsf{f}_\lambda\|^2 = \sum\sum_{1\leq i,j\leq M} \lambda_i\lambda_j \int_{\mathbb{R}^d} f_i(x)f_j(x)\,\mathrm{d}x \geq \kappa_M \sum_{j\in J(\lambda)} \lambda_j^2.$$

By the Cauchy-Schwarz inequality, we find

$$\begin{aligned} 4\sum_{j\in J(\lambda)} \omega_j|\widehat{\lambda}_j - \lambda_j| \quad &\leq \quad 4\sqrt{\sum_{j\in J(\lambda)} \omega_j^2}\sqrt{\sum_{j\in J(\lambda)} |\widehat{\lambda}_j - \lambda_j|^2} \\ &\leq \quad 4\left(\frac{\sum_{j\in J(\lambda)}\omega_j^2}{n\kappa_M}\right)^{1/2} \|f^{\spadesuit} - \mathsf{f}_\lambda\|. \end{aligned}$$

Combination with Lemma 2 yields that, with probability at least $1-\delta$,

$$\begin{aligned} \|f^{\spadesuit} - f\|^2 + \sum_{j=1}^M \omega_j|\widehat{\lambda}_j - \lambda_j| \quad &\leq \quad \|\mathsf{f}_\lambda - f\|^2 + 4\left(\frac{\sum_{j\in J(\lambda)}\omega_j^2}{n\kappa_M}\right)^{1/2} \|f^{\spadesuit} - \mathsf{f}_\lambda\| \\ &\leq \quad \|\mathsf{f}_\lambda - f\|^2 + b\Big(\|f^{\spadesuit} - f\| + \|\mathsf{f}_\lambda - f\|\Big) \end{aligned} \tag{3.15}$$

where $b = 4\sqrt{\sum_{j\in J(\lambda)}\omega_j^2}/\sqrt{n\kappa_M}$. Applying the inequality $2xy \leq x^2/\alpha + \alpha y^2$ $(x, y \in \mathbb{R},\ \alpha > 0)$ for each of the last two summands in (3.15) we get the result. □

Proof of Theorem 4. Write $\bar{\omega}_j = 2\sqrt{2}\sigma_j r(\delta/2) + (8/3)L_j r^2(\delta/2)$ for the choice of ω_j in (3.7). Using Bernstein's exponential inequality for sums of independent random variables $\zeta_{ij} = f_j(X_i) - \mathbb{E}f_j(X_i)$ with $|\zeta_{ij}| \leq 2L_j$, we obtain that

$$\begin{aligned} \mathbb{P}(A^c) \quad &= \quad \mathbb{P}\left(\bigcup_{j=1}^M \{2|V_j| > \bar{\omega}_j\}\right) \leq \sum_{j=1}^M \mathbb{P}\{2|V_j| > \bar{\omega}_j\} \\ &\leq \quad \sum_{j=1}^M \exp\left(-\frac{n\bar{\omega}_j^2/4}{2\mathrm{Var}(f_j(X_1)) + 2L_j\bar{\omega}_j/3}\right) \\ &\leq \quad M\exp(-nr^2(\delta/2)) = \delta/2. \end{aligned} \tag{3.16}$$

Let now ω_j be defined by (3.8). Then, using (3.16), we can write

$$
\begin{aligned}
\mathbb{P}(A^c) &= \mathbb{P}\left(\bigcup_{j=1}^{M}\{2|V_j| > \omega_j\}\right) \\
&\le \sum_{j=1}^{M}\mathbb{P}\{2|V_j| > \bar{\omega}_j\} + \sum_{j=1}^{M}\mathbb{P}\{\bar{\omega}_j > \omega_j\} \\
&\le \delta/2 + \sum_{j=1}^{M}\mathbb{P}\{\bar{\omega}_j > \omega_j\}.
\end{aligned} \tag{3.17}
$$

Define

$$t_j = 2\frac{\mathbb{E}f_j^4(X_1)}{\mathbb{E}f_j^2(X_1)}\frac{\log(2M/\delta)}{n}$$

and note that

$$\frac{2}{n}\sum_{i=1}^{n} f_j^2(X_i) + t_j \le T_j^2.$$

Then

$$
\begin{aligned}
\mathbb{P}\{\bar{\omega}_j > \omega_j\} &= \mathbb{P}\left\{\mathrm{Var}(f_j(X_1)) > T_j^2\right\} \\
&\le \mathbb{P}\{\mathbb{E}f_j^2(X_1) > \frac{2}{n}\sum_{i=1}^{n} f_j^2(X_i) + t_j\} \\
&\le \exp\left(-\frac{n\{\mathbb{E}f_j^2(X_1) + t_j\}^2}{8\mathbb{E}f_j^4(X_1)}\right) \\
&\qquad \text{using Proposition 2.6 in [38]} \\
&\le \exp\left(-\frac{nt_j\mathbb{E}f_j^2(X_1)}{2\mathbb{E}f_j^4(X_1)}\right) \\
&\qquad \text{since } (x+y)^2 \ge 4xy
\end{aligned}
$$

which is less than $\delta/(2M)$. Plugging this in (3.17) concludes the proof. □

Proof of Lemma 1. Using Bernstein's exponential inequality for sums of independent random variables $f_j^2(X_i) - \mathbb{E}f_j^2(X_i)$ and the fact that $\mathbb{E}f_j^4(X_1) \le L_j^2 \mathbb{E}f_j^2(X_1)$ we find

$$\begin{aligned}
& \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f_j^2(X_i) \ge 2\mathbb{E}f_j^2(X_1) + \frac{4}{3}L_j^2 r^2(\delta/2)\right) \\
= \ & \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f_j^2(X_i) - \mathbb{E}f_j^2(X_1) \ge \mathbb{E}f_j^2(X_1) + \frac{4}{3}L_j^2 r^2(\delta/2)\right) \\
\le \ & \exp\left(-\frac{n(\mathbb{E}f_j^2(X_1) + \frac{4}{3}L_j^2 r^2(\delta/2))^2}{2\mathbb{E}f_j^4(X_1) + \frac{4}{3}L_j^2\{\mathbb{E}f_j^2(X_1) + \frac{4}{3}L_j^2 r^2(\delta/2)\}}\right) \\
\le \ & \exp(-nr^2(\delta/2)) = \frac{\delta}{2M}\,,
\end{aligned}$$

which implies the lemma. □

4. Sparse estimation in mixture models

In this section we assume that the true density f can be represented as a finite mixture

$$f(x) = \sum_{j \in I^*} \lambda_j^* f_j(x),$$

where $I^* \subseteq \{1, \dots, M\}$ is unknown, f_j are known probability densities and $\lambda_j^* \neq 0$ for all $j \in I^*$. The focus of this section is on model selection, i.e., on the correct identification of the set I^*. We set $\lambda^* = (\lambda_1^*, \dots, \lambda_M^*)$ where $\lambda_j^* = 0, j \notin I^*$.

For clarity of exposition we consider a simplified version of the general set-up introduced above. We compute the estimates of λ^* via (2.3), with weights defined by (cf. (3.1)):

$$\omega_j = 4Lr, \text{ for all } j,$$

where $r > 0$ is a constant that we specify below, and for clarity of exposition we replaced all $L_j = \|f_j\|_\infty$ by an upper bound L on $\max_{1\le j\le M} L_j$. We assume that all f_j have been standardized to have $\|f_j\| = 1$. Note that under these assumptions condition (3.3) takes the form

$$\rho(\lambda) \le \frac{1}{16M(\lambda)}\,. \tag{4.1}$$

We state (4.1) for the true vector λ^* in the following form.

Condition (A).

$$\rho^* \le \frac{1}{16k^*}$$

where $k^* = |I^*| = M(\lambda^*)$ and $\rho^* = \rho(\lambda^*)$.

The results of Section 3 are valid for any r larger or equal to $r(\delta/2) = \{\log(2M/\delta)/n\}^{1/2}$. They give bounds on the predictive performance of SPADES. As noted in, e.g., [7], for ℓ_1-penalized model selection in regression, the tuning sequence ω_j required for correct selection is typically larger than the one that yields good prediction. We show below that the same is true for selecting the components of a mixture of densities. Specifically, in this section we will take the value

$$r = r(M, n, \delta/(2M)) = \sqrt{\frac{\log(2M^2/\delta)}{n}}. \tag{4.2}$$

We will use the following corollary of Theorem 1, obtained for $\alpha = \sqrt{2}$.

Corollary 2. *Assume that Condition (A) holds. Then with probability at least $1 - \delta/M$ we have*

$$\sum_{j=1}^{M} |\widehat{\lambda}_j - \lambda_j| \leq \frac{4\sqrt{2}}{L} k^* \sqrt{\frac{\log(2M^2/\delta)}{n}}\,. \tag{4.3}$$

Inequality (4.3) guarantees that the estimate $\widehat{\lambda}$ is close to the true λ^* in ℓ_1 norm, if the number of mixture components k^* is substantially smaller than $\sqrt{n}$. We regard this as an intermediate step for the next result that deals with the identification of I^*.

4.1. **Correct identification of the mixture components.** We now show that I^* can be identified with probability close to 1 by our procedure. Let $\hat{I} = J(\hat{\lambda})$ be the set of indices of the non-zero components of $\hat{\lambda}$ given by (2.3). In what follows we investigate when $P(\hat{I} = I^*) \geq 1 - \varepsilon$ for a given $0 < \varepsilon < 1$. Our results are non-asymptotic, they hold for any fixed M and n.

We need two conditions to ensure that correct recovery of I^* is possible. The first one is the identifiability of the model, as quantified by *Condition (A)* above. The second condition requires that the weights of the mixture are above the noise level, quantified by r. We state it as follows.

Condition (B).

$$\min_{j \in I^*} |\lambda_j^*| > 4(\sqrt{2} + 1) r L$$

where $L = \max\left(1/\sqrt{3}, \max_{1 \leq j \leq M} L_j\right)$ and r is given in (4.2).

Theorem 5. *Let $0 < \delta < 1/2$ be a given number. Assume that Conditions (A) and (B) hold. Then $\mathbb{P}(\hat{I} = I^*) \geq 1 - 2\delta(1 + 1/M)$.*

Proof. We begin by noticing that

$$\mathbb{P}(\hat{I} \neq I^*) \leq \mathbb{P}(I^* \not\subseteq \hat{I}) + \mathbb{P}(\hat{I} \not\subseteq I^*),$$

and we control each of the probabilities on the right hand side separately.

Control of $\mathbb{P}(I^* \not\subseteq \hat{I})$. By the definitions of the sets $\widehat{I}$ and I^* we have

$$\begin{aligned} \mathbb{P}(I^* \not\subseteq \hat{I}) &\leq \mathbb{P}(\widehat{\lambda}_k = 0 \text{ for some } k \in I^*) \\ &\leq k^* \max_{k \in I^*} \mathbb{P}(\widehat{\lambda}_k = 0). \end{aligned}$$

We control the last probability by using the characterization (5.9) of $\widehat{\lambda}$ given in Lemma 3 of the Appendix. We also recall that $\mathbb{E} f_k(X_1) = \sum_{j \in I^*} \lambda_j^* \langle f_k, f_j \rangle = \sum_{j=1}^M \lambda_j^* \langle f_k, f_j \rangle$, since we assumed that the density of X_1 is the mixture $f^* = \sum_{j \in I^*} \lambda_j^* f_j$. We therefore obtain, for $k \in I^*$,

$$\begin{aligned} \mathbb{P}\left(\widehat{\lambda}_k = 0\right) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n f_k(X_i) - \sum_{j=1}^M \widehat{\lambda}_j \langle f_j, f_k \rangle\right| \leq 4rL;\ \hat{\lambda}_k = 0\right) \\ &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n f_k(X_i) - \mathbb{E} f_k(X_1) - \sum_{j=1}^M (\widehat{\lambda}_j - \lambda_j^*) \langle f_j, f_k \rangle\right| \leq 4rL;\ \hat{\lambda}_k = 0\right) \\ &\leq \mathbb{P}\left(\left|\lambda_k^* \|f_k\|^2 + \frac{1}{n}\sum_{i=1}^n f_k(X_i) - \mathbb{E} f_k(X_1) - \sum_{j \neq k} (\widehat{\lambda}_j - \lambda_j^*) \langle f_j, f_k \rangle\right| \leq 4rL\right) \\ &\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n f_k(X_i) - \mathbb{E} f_k(X_1)\right| \geq \frac{|\lambda_k^*| \|f_k\|^2}{2} - 2rL\right) \qquad (4.4) \\ &\quad + \mathbb{P}\left(\left|\sum_{j \neq k} (\widehat{\lambda}_j - \lambda_j^*) \langle f_j, f_k \rangle\right| \geq \frac{|\lambda_k^*| \|f_k\|^2}{2} - 2rL\right). \qquad (4.5) \end{aligned}$$

To bound (4.4) we use Hoeffding's inequality, as in the course of Lemma 2. We first recall that $\|f_k\| = 1$ for all k and that, by *Condition (B)*, $\min_{k \in I^*} |\lambda_k^*| \geq 4(\sqrt{2}+1)Lr$, with $r = r(\delta/(2M)) = \{\log(2M^2/\delta)/n\}^{1/2}$. Therefore

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n f_k(X_i) - \mathbb{E} f_k(X_1)\right| \geq \frac{|\lambda_k^*|}{2} - 2rL\right) \qquad (4.6) \\ &\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n f_k(X_i) - \mathbb{E} f_k(X_1)\right| \geq 2\sqrt{2} rL\right) \leq \frac{\delta}{M^2}. \end{aligned}$$

To bound (4.5) notice that, by Conditions (A) and (B),

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j\neq k}(\widehat{\lambda}_j-\lambda_j^*)\langle f_j,f_k\rangle\right|\geq\frac{|\lambda_k^*|}{2}-2rL\right)\\
&\leq\mathbb{P}\left(\sum_{j=1}^{M}|\widehat{\lambda}_j-\lambda_j^*|\geq 32\sqrt{2}rLk^*\right)\\
&\leq\ \mathbb{P}\left(\sum_{j=1}^{M}|\widehat{\lambda}_j-\lambda_j^*|\geq\frac{4\sqrt{2}rk^*}{L}\right)\leq\ \frac{\delta}{M},
\end{aligned}$$

where the penultimate inequality holds since, by definition, $L^2\geq 1/3$ and the last inequality holds by Corollary 4.

Combining the above results we obtain

$$\mathbb{P}(I^*\not\subseteq\hat{I})\quad\leq\quad k^*\frac{\delta}{M^2}+k^*\frac{\delta}{M}\leq\frac{\delta}{M}+\delta.$$

Control of $\mathbb{P}(\hat{I}\not\subseteq I^*)$. Let

$$h(\mu)=-\frac{2}{n}\sum_{i=1}^{n}\sum_{j\in I^*}\mu_j f_j(X_i)+\|\sum_{j\in I^*}\mu_j f_j\|^2+8rL\sum_{j\in I^*}|\mu_j|. \tag{4.7}$$

Let

$$\tilde{\mu}=\underset{\mu\in\mathbb{R}^{k^*}}{\arg\min}\,h(\mu). \tag{4.8}$$

Consider the random event

$$\mathcal{B}=\bigcap_{k\notin I^*}\left\{\left|-\frac{1}{n}\sum_{i=1}^{n}f_k(X_i)+\sum_{j\in I^*}\tilde{\mu}_j\langle f_j,f_k\rangle\right|\leq 4Lr\right\}. \tag{4.9}$$

Let $\bar{\mu}\in\mathbb{R}^M$ be the vector that has the components of $\tilde{\mu}$ given by (4.8) in positions corresponding to the index set I^* and zero components elsewhere. By the first part of Lemma 3 in the Appendix we have that $\bar{\mu}\in\mathbb{R}^M$ is a solution of (2.3) on the event $\mathcal{B}$. Recall that $\widehat{\lambda}$ is a also solution of (2.3). By the definition of the set $\widehat{I}$ we have that $\widehat{\lambda}_k\neq 0$ for $k\in\widehat{I}$. By construction, $\tilde{\mu}_k\neq 0$ for some subset $S\subseteq I^*$. By the second part of Lemma 3 in the Appendix, any two solutions have non-zero elements in the same positions. Therefore $\widehat{I}=S\subseteq I^*$ on $\mathcal{B}$.

Thus,

$$\begin{aligned}
\mathbb{P}(\hat{I} \not\subseteq I^*) &\le \mathbb{P}(\mathcal{B}^c) \\
&\le \sum_{k \notin I^*} \mathbb{P}\left\{ \left| -\frac{1}{n}\sum_{i=1}^n f_k(X_i) + \sum_{j\in I^*} \tilde{\mu}_j \langle f_j, f_k\rangle \right| \ge 4rL \right\} \\
&\le \sum_{k \notin I^*} \mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^n f_k(X_i) - Ef_k(X_1) \right| \ge 2\sqrt{2}rL \right) \\
&\quad + \sum_{k \notin I^*} \mathbb{P}\left(\sum_{j\in I^*} |\tilde{\mu}_j - \lambda_j^*| \left|\langle f_j, f_k\rangle\right| \ge (4-2\sqrt{2})rL \right).
\end{aligned} \tag{4.10}$$

Reasoning as in (4.6) above we find

$$\sum_{k \notin I^*} \mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^n f_k(X_i) - Ef_k(X_1) \right| \ge 2\sqrt{2}rL \right) \le \frac{\delta}{M}.$$

To bound the last sum in (4.10) we first notice that Theorem 1 (if we replace there $r(\delta/2)$ by the larger value $r(\delta/(2M))$, cf. Theorem 4) applies to $\tilde{\mu}$ given by (4.8). In particular

$$\mathbb{P}\left(\sum_{j\in I^*} |\tilde{\mu}_j - \lambda_j^*| \ge \frac{4\sqrt{2}}{L} k^* r \right) \le \frac{\delta}{M}.$$

Therefore, by *Condition (A)*, we have

$$\begin{aligned}
&\sum_{k \notin I^*} \mathbb{P}\left(\sum_{j\in I^*} |\tilde{\mu}_j - \lambda_j^*| \left|\langle f_j, f_k\rangle\right| \ge (4-2\sqrt{2})rL \right) \\
&\quad \le \sum_{k \notin I^*} \mathbb{P}\left(\sum_{j\in I^*} |\tilde{\mu}_j - \lambda_j^*| \ge 32(4-2\sqrt{2})k^* rL \right) \\
&\quad \le \sum_{k \notin I^*} \mathbb{P}\left(\sum_{j\in I^*} |\tilde{\mu}_j - \lambda_j^*| \ge \frac{4\sqrt{2}}{L} k^* r \right) \le \delta,
\end{aligned}$$

which holds since $L^2 \ge 1/3$. Collecting all the bounds above we obtain

$$P(\hat{I} \ne I^*) \le 2\delta + \frac{2\delta}{M},$$

which concludes the proof. ■

4.2. **Example: Identifying true components in mixtures of Gaussian densities.** Consider an ensemble of M Gaussian densities f_j's in $\mathbb{R}^d$ with means μ_j and covariance matrices $\tau_j \mathbb{I}_d$, where $\mathbb{I}_d$ is the unit $d \times d$ matrix. In what follows we show that *Condition*

(A) holds if the means of the Gaussian densities are well separated and we make this precise below. Therefore, in this case, Theorem 5 guarantees that if the weights of the mixture are above the threshold given in Condition B, we can recover the true mixture components with high probability via our procedure.

Recall that *Condition (A)* requires

$$\rho^* = \max_{i\in I^*, j\neq i} \frac{< f_i, f_j >}{\|f_1\|\|f_2\|} \leq \frac{1}{16k^*}.$$

The densities are

$$f_j(x) = \frac{1}{\sqrt{2\pi}\tau_j} \exp\left(-\frac{\|x-\mu_j\|_2}{2\tau_j^2}\right),$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Let $\tau_{\max} = \max_{1\leq j\leq M}\tau_j$ and $D_{\min}^2 = \min_{k\neq j}\|\mu_k - \mu_j\|_2^2$. Via simple algebra we obtain

$$\rho^* \leq \exp\left(-\frac{D_{\min}^2}{4\tau_{\max}^2}\right).$$

Therefore, *Condition (A)* holds if

$$D_{\min}^2 \geq 4\tau_{\max}^2 \log(16k^*).$$

Using this and Theorem 5 we see that SPADES can identifies the true components in a mixture of Gaussian densities if the square Euclidean distance between any two means is large enough as compared to the largest variance of the components in the mixture.

Note that *Condition (B)* on the size of the mixture weights involves the constant L, which in this example can be taken as

$$L = \max\left(\frac{1}{\sqrt{3}},\ \max_{1\leq j\leq M}\|f_j\|_\infty\right) = \max\left(\frac{1}{\sqrt{3}},\ \frac{1}{(\sqrt{2\pi}\tau_{\min})^d}\right),$$

where $\tau_{\min} = \min_{1\leq j\leq M}\tau_j$.

5. SPADES for adaptive nonparametric density estimation

We assume in this section that the density f is defined on a bounded interval of $\mathbb{R}$ that we take without loss of generality to be the interval $[0,1]$. Consider a countable system of functions $\{\psi_{lk},\ l\geq -1, k\in V(l)\}$ in L_2, where the set of indices $V(l)$ satisfies $|V(-1)|\leq C$, $2^l \leq |V(l)| \leq C2^l$, $l\geq 0$, for some constant C, and where the functions psi_{lk} satisfy

$$\|\psi_{lk}\| \leq C_1, \quad \|\psi_{lk}\|_\infty \leq C_1 2^{l/2}, \quad \Big\|\sum_{k\in V(l)} \psi_{lk}^2\Big\|_\infty \leq C_1 2^l, \tag{5.1}$$

for all $l \geq -1$ and for some $C_1 < \infty$. Examples of such systems $\{\psi_{lk}\}$ are given, for instance, by compactly supported wavelet bases, see, e.g., [19]. In this case $\psi_{lk}(x) = 2^{l/2}\psi(2^l x - k)$ for some compactly supported function ψ. We assume that $\{\psi_{lk}\}$ is a frame, i.e., there exist positive constants c_1 and c_2 depending only on $\{\psi_{lk}\}$ such that, for any two sequences of coefficients β_{lk}, β'_{lk},

$$c_1 \sum_{l=-1}^{\infty} \sum_{k \in V(l)} (\beta_{lk} - \beta'_{lk})^2 \leq \Big\| \sum_{l=-1}^{\infty} \sum_{k \in V(l)} (\beta_{lk} - \beta'_{lk})\psi_{lk} \Big\|^2 \leq c_2 \sum_{l=-1}^{\infty} \sum_{k \in V(l)} (\beta_{lk} - \beta'_{lk})^2. \tag{5.2}$$

If $\{\psi_{lk}\}$ is an orthonormal wavelet basis, this condition is satisfied with $c_1 = c_2 = 1$.

Now, choose $\{f_1, \ldots, f_M\} = \{\psi_{lk}, -1 \leq l \leq l_{\max}, k \in V(l)\}$ where $l_{\max}$ is such that $2^{l_{\max}} \asymp n/(\log n)$. Then also $M \asymp n/(\log n)$. The coefficients λ_j are now indexed by $j = (l, k)$, and we set by definition $\lambda_{(l,k)} = 0$ for $(l, k) \notin \{-1 \leq l \leq l_{\max}, k \in V(l)\}$. Assume that there exist coefficients β^*_{lk} such that

$$f = \sum_{l=-1}^{\infty} \sum_{k \in V(l)} \beta^*_{lk}\psi_{lk}$$

where the series converges in L_2. Then Theorem 3 easily implies the following result.

Theorem 6. *Let $f_1, \ldots, f_M$ be as defined above with $M \asymp n/(\log n)$, and let ω_j be given by (3.8) for $\delta = n^{-2}$. Then for all $n \geq 1$, $\lambda \in \mathbb{R}^M$ we have with probability at least $1 - n^{-2}$,*

$$\begin{aligned} \|f^{\spadesuit} - f\|^2 \quad \leq \quad K \Bigg(& \sum_{l=-1}^{\infty} \sum_{k \in V(l)} (\lambda_{(l,k)} - \beta^*_{lk})^2 \\ & + \sum_{(l,k) \in J(\lambda)} \Big[\frac{1}{n} \sum_{i=1}^{n} \psi_{lk}^2(X_i) \frac{\log n}{n} + 2^l \Big(\frac{\log n}{n}\Big)^2 \Big] \Bigg) \end{aligned} \tag{5.3}$$

where K is a constant independent of f.

This is a general oracle inequality that allows one to show that the estimator $f^{\spadesuit}$ attains minimax rates of convergence, up to a logarithmic factor simultaneously on various functional classes. We will explain this in detail for the case where f belongs to a class of functions $\mathcal{F}$ satisfying the following assumption for some $s > 0$.

Condition (C). For any $f \in \mathcal{F}$ and any $l' \geq 0$ there exists a sequence of coefficients $\lambda = \{\lambda_{(l,k)}, -1 \leq l \leq l', k \in V(l)\}$ such that

$$\sum_{l=-1}^{\infty} \sum_{k \in V(l)} (\lambda_{(l,k)} - \beta^*_{lk})^2 \leq C_2 2^{-2l's} \tag{5.4}$$

for a constant C_2 independent of f.

It is well known that Condition (C) holds for various functional classes $\mathcal{F}$, such as Hölder, Sobolev, Besov classes, if $\{\psi_{lk}\}$ is an appropriately chosen wavelet basis, see, e.g., [19] and the references cited therein. In this case s is the smoothness parameter of the class. Moreover, the basis $\{\psi_{lk}\}$ can be chosen so that Condition (C) is satisfied with C_2 independent of s for all $s \le s_{\max}$, where $s_{\max}$ is a given positive number. This allows for adaptation in s.

Under Condition (C) we obtain from (5.3) that, with probability at least $1 - n^{-2}$,

$$(5.5)\quad \|f^{\spadesuit} - f\|^2 \le \min_{l' \le l_{\max}} K \left(C_2 2^{-2l's} + \sum_{(l,k): l \le l'} \left[\frac{1}{n} \sum_{i=1}^{n} \psi_{lk}^2(X_i) \frac{\log n}{n} + 2^l \Big(\frac{\log n}{n}\Big)^2 \right] \right)$$

From (5.5) and the last inequality in (5.1) we find for some constant K', with probability at least $1 - n^{-2}$,

$$(5.6)\qquad \begin{aligned} \|f^{\spadesuit} - f\|^2 &\le \min_{l' \le l_{\max}} K' \left(2^{-2l's} + 2^{l'} \Big(\frac{\log n}{n}\Big) + 2^{2l'} \Big(\frac{\log n}{n}\Big)^2 \right) \\ &= O\left(\Big(\frac{\log n}{n}\Big)^{-2s/(2s+1)} \right) \end{aligned}$$

where the last expression is obtained by choosing l' such that $2^{l'} \asymp (n/\log n)^{1/(2s+1)}$. It follows from (5.6) that $f^{\spadesuit}$ converges with the optimal rate (up to a logarithmic factor) simultaneously on all the functional classes satisfying Condition (C). Note that the definition of the functional class is not used in the construction of the estimator $f^{\spadesuit}$, so this estimator is optimal adaptive in the rate of convergence (up to a logarithmic factor) on this scale of functional classes for $s \le s_{\max}$. Results of such type, and even more pointed (without extra logarithmic factors in the rate and sometimes with exact asymptotic minimax constants) are known for various other adaptive density estimators, see,for instance, [16, 6, 19, 21, 28, 29] and the references cited therein. These papers consider classes of densities that are uniformly bounded by a fixed constant, see the recent discussion in [5]. This prohibits, for example, free scale transformations of densities within a class. Inequality (5.6) does not have this drawback. It allows to get the rates of convergence for classes of unbounded densities f as well.

Another example is given by the classes of sparse densities defined as follows:

$$\mathcal{L}_0(m) = \Big\{ f : [0,1] \to \mathbb{R} : \ f \text{ is a probability density and } \Big|\{j : < f, f_j > \neq 0\}\Big| \le m \Big\}$$

where $m \le M$ is an unknown integer. If $f_1, \ldots, f_M$ is a wavelet system as defined above and $J^* = \{j = (l,k) : < f, f_j > \neq 0\}$, then under the conditions of Theorem 6 for any $f \in \mathcal{L}_0(m)$

we have, with probability at least $1-n^{-2}$,

$$\|f^{\spadesuit}-f\|^2 \;\le\; K\left(\sum_{(l,k)\in J^*}\Big[\frac{1}{n}\sum_{i=1}^n \psi_{lk}^2(X_i)\frac{\log n}{n}+2^l\Big(\frac{\log n}{n}\Big)^2\Big]\right). \tag{5.7}$$

From (5.7), using Lemma 1 and the first two inequalities in (5.1) we obtain the following result.

Corollary 3. *Let the assumptions of Theorem 6 hold. Then, for every $L<\infty$ and $n\ge 1$,*

$$\sup_{f\in\mathcal{L}_0(m)\cap\{f:\,\|f\|_\infty\le L\}} \mathbb{P}\left\{\|f^{\spadesuit}-f\|^2\ge b\left(\frac{m\log n}{n}\right)\right\}\le (3/2)n^{-2},\quad \forall\ m\le M, \tag{5.8}$$

where $b>0$ is a constant depending only on L.

Corollary 3 can be viewed as an analogue for density estimation of the adaptive minimax results for $\mathcal{L}_0$ classes obtained in the Gaussian sequence model [1, 17] and in the random design regression model [10].

Appendix

Lemma 3. *(I) Let $\tilde\mu$ be given by (4.8). Then $\bar\mu=(\tilde\mu,0)\in\mathbb{R}^M$ is a minimizer in $\lambda\in\mathbb{R}^M$ of*

$$g(\lambda)=-\frac{2}{n}\sum_{i=1}^n \mathsf{f}_\lambda(X_i)+\|\mathsf{f}_\lambda\|^2+8Lr\sum_{k=1}^M|\lambda_k|.$$

on the random event $\mathcal{B}$ defined in (4.9).
(II) Any two minimizers of $g(\lambda)$ have non-zero components in the same positions.

Proof. (I). Since g is convex, by standard results in convex analysis, $\bar\lambda\in\mathbb{R}^M$ is a minimizer of g if and only if $0\in D_{\bar\lambda}$ where D_λ is the subdifferential of $g(\lambda)$:

$$D_\lambda=\{w\in\mathbb{R}^M:\ w_k=-\frac{2}{n}\sum_{i=1}^n f_k(X_i)+2\sum_{j=1}^M\lambda_j\langle f_j,f_k\rangle+8rv_k, v_k\in V_k(\lambda_k), 1\le k\le M\}$$

where

$$V_k(\lambda_k)\;=\;\begin{cases}\{L\} & \text{if } \lambda_k>0,\\ \{-L\} & \text{if } \lambda_k<0,\\ [-L,L] & \text{if } \lambda_k=0.\end{cases}$$

Therefore, $\bar{\lambda}$ minimizes $g(\cdot)$ if and only if, for all $1 \le k \le M$,

$$\frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - \sum_{j=1}^{M} \bar{\lambda}_j \langle f_j, f_k\rangle \;=\; 4Lr\,\mathrm{sign}(\bar{\lambda}_k), \quad \text{if } \bar{\lambda}_k \neq 0, \tag{5.9}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - \sum_{j=1}^{M} \bar{\lambda}_j \langle f_j, f_k\rangle\right| \;\le\; 4Lr, \quad \text{if } \;\bar{\lambda}_k = 0. \tag{5.10}$$

We now show that $\bar{\mu} = (\tilde{\mu}, 0) \in \mathbb{R}^M$ with $\tilde{\mu}$ given in (4.8) satisfies (5.9)–(5.10) on the event $\mathcal{B}$ and therefore is a minimizer of $g(\lambda)$ on this event. Indeed, since $\tilde{\mu}$ is a minimizer of the convex function $h(\mu)$ given in (4.7), the same convex analysis argument as above implies that

$$\frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - \sum_{j\in I^*} \tilde{\mu}_j \langle f_j, f_k\rangle \;=\; 4Lr\,\mathrm{sign}(\tilde{\mu}_k), \text{ if } \tilde{\mu}_k \neq 0, \; k \in I^*,$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - \sum_{j\in I^*} \tilde{\mu}_j \langle f_j, f_k\rangle\right| \;\le\; 4Lr, \text{ if } \tilde{\mu}_k = 0, \; k \in I^*.$$

Note that on the event $\mathcal{B}$ we also have

$$\left|\frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - \sum_{j\in I^*} \tilde{\mu}_j \langle f_j, f_k\rangle\right| \le 4Lr, \text{ if } k \notin I^* \text{ (for which } \bar{\mu}_k = 0, \text{ by construction)}.$$

Here $\bar{\mu}_k$ denotes the kth coordinate of $\bar{\mu}$. The above three displays and the fact that $\bar{\mu}_k = \tilde{\mu}_k, k \in I^*$, show that $\bar{\mu}$ satisfies conditions (5.9)–(5.10) and is therefore a minimizer of $g(\lambda)$ on the event $\mathcal{B}$.

(II). We now prove the second assertion of the lemma. In view of (5.9) the index set S of the non-zero components of any minimizer $\bar{\lambda}$ of $g(\lambda)$ satisfies

$$S = \left\{ k \in \{1,\dots,M\} : \; \left|\frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - \sum_{j=1}^{M} \bar{\lambda}_j \langle f_j, f_k\rangle\right| = 4rL \right\}.$$

Therefore, if for any two minimizers $\bar{\lambda}^{(1)}$ and $\bar{\lambda}^{(2)}$ of $g(\lambda)$ we have

$$\sum_{j=1}^{M} (\bar{\lambda}_j^{(1)} - \bar{\lambda}_j^{(2)}) \langle f_j, f_k\rangle = 0, \quad \text{for all } \; k, \tag{5.11}$$

then S is the same for all minimizers of $g(\lambda)$.

Thus, it remains to show (5.11). We use simple properties of convex functions. First, we recall that the set of minima of a convex function is convex. Then, if $\bar{\lambda}^{(1)}$ and $\bar{\lambda}^{(2)}$ are two distinct points of minima, so is $\rho\bar{\lambda}^{(1)} + (1-\rho)\bar{\lambda}^{(2)}$, for any $0 < \rho < 1$. Re-write this convex combination as $\bar{\lambda}^{(2)} + \rho\eta$, where $\eta = \bar{\lambda}^{(1)} - \bar{\lambda}^{(2)}$. Recall that the minimum value of any convex function is unique. Therefore, for any $0 < \rho < 1$, the value of $g(\lambda)$ at $\lambda = \bar{\lambda}^2 + \rho\eta$ is equal to some constant C:

$$\begin{aligned} F(\rho) \quad &\triangleq \quad -\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{M}\left(\bar{\lambda}_j^{(2)} + \rho\eta_j\right) f_j(X_i) + \int \left(\sum_{j=1}^{M}(\bar{\lambda}_j^{(2)} + \rho\eta_j) f_j(x)\right)^2 dx \\ &\quad + 8rL\sum_{j=1}^{M}|\bar{\lambda}_j^{(2)} + \rho\eta_j| = C. \end{aligned}$$

By taking the derivative with respect to ρ of $F(\rho)$ we obtain that, for all $0 < \rho < 1$,

$$\begin{aligned} F'(\rho) \quad &= \quad -\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{M}\eta_j f_j(X_i) + \ 8rL\sum_{j=1}^{M}\eta_j \mathrm{sign}(\bar{\lambda}_j^{(2)} + \rho\eta_j) \\ &\quad + 2\int\left(\sum_{j=1}^{M}(\bar{\lambda}_j^{(2)} + \rho\eta_j) f_j(x)\right)\left(\sum_{j=1}^{M}\eta_j f_j(x)\right) dx = 0. \end{aligned}$$

By continuity of $\rho \mapsto \bar{\lambda}_j^{(2)} + \rho\eta_j$, there exists an open interval in $(0, 1)$ on which $\rho \mapsto \mathrm{sign}(\bar{\lambda}_j^{(2)} + \rho\eta_j)$ is constant for all j. Therefore, on that interval,

$$F'(\rho) = 2\rho\int\left(\sum_{j=1}^{M}\eta_j f_j(x)\right)^2 dx + C'$$

where C' does not depend on ρ. This is compatible with $F'(\rho) = 0, \ \forall\ 0 < \rho < 1$, (cf. (5.12)) only if

$$\sum_{j=1}^{M}\eta_j f_j(x) = 0, \text{for all} \quad x,$$

and therefore

$$\sum_{j=1}^{M}\eta_j\langle f_j, f_k\rangle = 0, \text{for all} \quad k \in \{1, \dots, M\},$$

which is the desired result. This completes the proof of the lemma. □

References

[1] Abramovich,F, Benjamini,Y., Donoho, D.L. and Johnstone, I.M.. (2006). Adapting to unknown sparsity by controlling the False Discovery Rate. *Annals of Statistics* **34** 584-653.

[2] Biau, G. and Devroye, L. (2005). Density estimation by the penalized combinatorial method. *Journal of Multivariate Analysis*, **94**, 196–208.

[3] Biau, G., Cadre, B., Devroye, L. and Györfi, L. (2008). Strongly consistent model selection for densities. *Test*, in press.

[4] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2007). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statist.*, to appear.

[5] Birgé, L. (2008). Model selection for density estimation with L_2 loss. `arXiv:0808.1416`

[6] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien LeCam: Research Papers in Probability and Statistics*, D. Pollard, E. Torgersen, and G. Yang, Eds., pp. 55-87. Springer, New York.

[7] Bunea, F. (2008). Non-asymptotic estimation of sparse patters via the Lasso and the elastic net. *http://www.stat.fsu.edu/flori, Submitted.*

[8] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Aggregation for Gaussian regression. *The Annals of Statistics*, **35** 1674 - 1697.

[9] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2006a). Aggregation and sparsity via ℓ_1-penalized least squares. *Proceedings of 19th Annual Conference on Learning Theory, COLT 2006. Lecture Notes in Artificial Intelligence* **4005** 379–391. Springer-Verlag, Heidelberg.

[10] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Sparsity oracle inequalities for the Lasso. The Electronic Journal of Statistics **1** 169 - 194.

[11] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Sparse density estimation with l1 penalties. *Proceedings of 20th Annual Conference on Learning Theory, COLT 2007. Lecture Notes in Artificial Intelligence* 530 - 544. Springer-Verlag, Heidelberg.

[12] Chen, S., Donoho, D. and Saunders, M. (2001) Atomic decomposition by basis pursuit. *SIAM Review* **43** 129 - 159.

[13] Devroye, L. and Lugosi, G. (2000) *Combinatorial Methods in density estimation*, Springer.

[14] Donoho, D.L. (1995) Denoising via soft-thresholding. *IEEE Trans. Info. Theory* **41** 613-627.

[15] Donoho, D.L., Elad, M. and Temlyakov, V. (2006). Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. *IEEE Trans. on Information Theory* **52** 6–18.

[16] Golubev, G.K. (1992). Nonparametric estimation of smooth probability densties in L_2. *Problems of Information Transmission* **28** 44-54.

[17] Golubev, G.K. (2002). Reconstruction of sparse vectors in white Gaussian noise. *Problems of Information Transmission* **38** 65–79.

[18] Greenshtein, E. and Ritov, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988.

[19] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, Approximation and Statistical Applications.* Lecture Notes in Statistics, vol. 129. Springer, New York.

[20] James, L., Priebe, C. and Marchette, D. (2001). Consistent estimation of mixture complexity. *Annals of Statistics*, **29**, 1281–1296.

[21] Kerkyacharian, G., Picard, D. and Tribouley, K. (1996). L^p adaptive density estimation. *Bernoulli* **2** 229–247.

[22] Koltchinskii, V. (2005). Model selection and aggregation in sparse classification problems. *Oberwolfach Reports* **2** 2663–2667, Mathematisches Forschungsinstitut Oberwolfach.

[23] Koltchinskii, V. (2006). Sparsity in penalized empirical risk minimization. *Submitted.*

[24] Loubes, J. – M. and van de Geer, S. A. (2002). Adaptive estimation in regression, using soft thresholding type penalties. *Statistica Neerlandica* **56** 453 – 478.

[25] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* **2** 90–102.

[26] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436–1462.

[27] Nemirovski, A. (2000). Topics in non-parametric statistics. In P. Bernard, editor, *Ecole d'Eté de Probabilités de Saint-Flour 1998*, volume XXVIII of *Lecture Notes in Mathematics*. Springer, New York.

[28] Rigollet, Ph. (2006). Inégalités d'oracle, agrégation et adaptation. PhD thesis, University of Paris 6.

[29] Rigollet, Ph., and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics* **16** 260–280.

[30] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9** 65 –78.

[31] Samarov, A. and Tsybakov, A. (2007). Aggregation of density estimators and dimension reduction. *Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A. Doksum* (V.Nair, ed.), World Scientific, Singapore e.a., 233–251.

[32] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B.* **58** 267–288.
[33] Tsybakov, A. B. (2003). Optimal rates of aggregation. In *Lecture Notes in Artificial Intelligence*, volume 2777 of *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines*. Springer-Verlag, Heidelberg.
[34] Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory (Information Science and Statistics)*, Springer.
[35] van de Geer, S.A. (2008). High dimensional generalized linear models and the Lasso. *The Annals of Statistics* **26** 225–287.
[36] Wasserman, L. A. (2004) *All of Statistics*, Springer.
[37] Wegkamp, M. H. (1999). Quasi-Universal Bandwidth Selection for Kernel Density Estimators. *Canadian Journal of Statistics* **27** 409 – 420.
[38] Wegkamp, M.H. (2003). Model selection in nonparametric regression. *Annals of Statistics* **31** 252–273.
[39] Zhang, C.H. and Huang, J. (2008). The sparsity and biais of the Lasso selection in high-dimensional linear regression. Annals of Statistics **36** 1567-1594.
[40] Zhao, P. and Yu, B. (2007). On model selection consistency of Lasso. Journal of Machine Learning Research **7** 2541-2567.
[41] Zou, H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association **101** 1418-1429.

Florentina Bunea

Department of Statistics

Florida State University

P.O. Box 3064330

Tallahassee, FL 32306-4330

United States of America

E-mail: bunea@stat.fsu.edu

Alexandre Tsybakov

Laboratoire de Statistique, CREST

3, Avenue Pierre Larousse, 92240 Malakoff

France

and

LPMA (UMR CNRS 7599), Université Paris 6

4, Place Jussieu, 75252 Paris, Cedex 05

France

E-mail: alexandre.tsybakov@upmc.fr

Marten Wegkamp

Department of Statistics

Florida State University
P.O. Box 3064330
Tallahassee, FL 32306-4330
United States of America
E-mail: wegkamp@stat.fsu.edu