The Annals of Statistics
2010, Vol. 38, No. 4, 2525–2558
DOI: 10.1214/09-AOS790

SPADES AND MIXTURE MODELS[1]

BY FLORENTINA BUNEA[2], ALEXANDRE B. TSYBAKOV[3], MARTEN H. WEGKAMP[2] AND ADRIAN BARBU

Florida State University, Laboratoire de Statistique, CREST and LPMA, Université Paris 6, Florida State University and Florida State University

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. We offer a data driven method for the choice of the tuning parameter used in the construction of SPADES. The method uses the generalized bisection method first introduced in [10]. The suggested procedure bypasses the need for a grid search and offers substantial computational savings. We complement our theoretical results with a simulation study that employs this method for approximations of one and two-dimensional densities with mixtures. The numerical results strongly support our theoretical findings.

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

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

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

Received January 2009; revised December 2009.

[1]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.

[2]Supported in part by the NSF Grant DMS-07-06829.

[3]Supported in part by the Grant ANR-06-BLAN-0194 and by the PASCAL Network of Excellence.

AMS 2000 subject classifications. Primary 62G08; secondary 62C20, 62G05, 62G20.

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

- *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, that is, that very few coordinates of λ^* are nonzero, with indices corresponding to a set $I^* \subseteq \{1, \dots, M\}$. 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 $\{f_1, \dots, f_M\}$ 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., [6, 27, 37]); 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 $f_1, \dots, f_M$ are some preliminary estimators of f constructed from a training sample independent of $(X_1, \dots, X_n)$, and we would like to aggregate $f_1, \dots, f_M$. This means that we would like to construct a new estimator, the aggregate, which is approximately as good as the best among $f_1, \dots, f_M$ or approximately as good as the best linear or convex combination of $f_1, \dots, f_M$. General notions of aggregation and optimal rates are introduced in [33, 40]. Aggregation of density estimators is discussed in [35, 36, 38] 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 $\widehat{\lambda} = \widehat{\lambda}(X_1, \dots, X_n) \in \mathbb{R}^M$.

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 [14]. 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 [16, 19, 24, 32, 39]. For Gaussian sequence models or for regression with an orthogonal design matrix the Lasso is equivalent to soft thresholding [18, 30]. Model selection consistency of the Lasso type linear regression estimators is treated in many papers including [31, 32, 46–48]. Recently, ℓ_1-penalized methods have been extended to nonparametric regression with general fixed or random design [4, 11–13], as well as to some classification and other more general prediction type models [8, 28, 29, 42].

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 [14].

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, for example, [43] 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 [2, 26]. A consistent selection procedure specialized to Gaussian mixtures is suggested in [26]. The method of [26] 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 [17]) 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 [2, 26] 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 nonasymptotic, 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 linear and generalized regression models; see, for example, [8].

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, that is, densities having only a finite, but unknown, number of nonzero wavelet coefficients.

Section 6.1 offers an algorithm for computing the SPADES. Our procedure is based on coordinate descent optimization, recently suggested by [20]. In Section 6.2 we use this algorithm together with a tuning parameter chosen in a data adaptive manner. This choice employs the generalized bisection method first introduced in [10], a computationally efficient method for constructing candidate tuning parameters without performing a grid search. The final tuning parameter is chosen

from the list of computed candidates by using a 10-fold cross-validated dimension-regularized criterion. The combined procedure works very well in practice, and we present a simulation study in Section 6.3.

2. Definition of SPADES. Consider the $L_2(\mathbb{R}^d)$ norm

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

associated with the inner product

$$\langle g, h \rangle = \int_{\mathbb{R}^d} g(x)h(x)\,dx$$

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 $g \in L_2$,

$$\langle g, f \rangle = \mathbb{E}g(X),$$

where the expectation is taken under f. Moreover,

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

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

$$\widehat{\gamma}(\lambda) = -\frac{2}{n}\sum_{i=1}^{n} \mathsf{f}_\lambda(X_i) + \|\mathsf{f}_\lambda\|^2. \tag{2.2}$$

This motivates the use of $\widehat{\gamma} = \widehat{\gamma}(\lambda)$ as the empirical criterion; see, for instance, [6, 37, 44].

We define the penalty

$$\mathrm{pen}(\lambda) = 2\sum_{j=1}^{M} \omega_j |\lambda_j| \tag{2.3}$$

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

$$\begin{aligned} \widehat{\lambda} &= \arg\min_{\lambda \in \mathbb{R}^M} \{\widehat{\gamma}(\lambda) + \mathrm{pen}(\lambda)\} \\ &= \arg\min_{\lambda \in \mathbb{R}^M} \left\{ -\frac{2}{n}\sum_{i=1}^{n} \mathsf{f}_\lambda(X_i) + \|\mathsf{f}_\lambda\|^2 + 2\sum_{j=1}^{M} \omega_j |\lambda_j| \right\}. \end{aligned} \tag{2.4}$$

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

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

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 $\tilde{\lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_M)$.

The SPADES estimate can be easily computed by convex programming even if $M \gg n$. We present an algorithm in Section 6 below. SPADES retains the desirable theoretical properties of other density estimators, the computation of which may become problematic for $M \gg n$. We refer to [17] for a thorough overview on combinatorial methods in density estimation, to [41] 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

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

be the set of indices corresponding to nonzero 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

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

for $1 \le j \le M$, where $\mathrm{Var}(\zeta)$ 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 $\widehat{\lambda} = \widehat{\lambda}(\omega_1, \ldots, \omega_M)$, 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

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

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

$$\rho_M(i,j)=\frac{\langle f_i, f_j\rangle}{\|f_i\|\|f_j\|}, \qquad i,j=1,\dots,M.$$

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

$$F(\lambda)=\max_{j\in J(\lambda)}\frac{\omega_j}{r(\delta/2)\|f_j\|}=\max_{j\in J(\lambda)}\frac{4L_j}{\|f_j\|}$$

and

$$G=\max_{1\le j\le M}\frac{r(\delta/2)\|f_j\|}{\omega_j}=\max_{1\le j\le M}\frac{\|f_j\|}{4L_j}.$$

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*

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

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

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

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:

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

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

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

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,

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

or

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

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 $\tilde{T}_j = (2\mathbb{E}f_j^2(X_1) + (4/3)L_j^2 r^2(\delta/2))^{1/2}$.

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 $j = 1, \dots, M$. 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

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

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

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

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

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

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

$$u_j = \widehat{\lambda}_j - \lambda_j, \qquad U(\lambda) = \sum_{j\in J(\lambda)} |u_j|\|f_j\|, \qquad U = \sum_{j=1}^{M} |u_j|\|f_j\| r = r(\delta/2).$$

Then, by the definition of $F(\lambda)$,

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

Since

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

we obtain

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

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

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

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

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

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

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

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 $U_*(\lambda) \ge U(\lambda)/\sqrt{M(\lambda)}$, we find

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

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

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

By the Cauchy–Schwarz inequality, we find

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

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

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

where $b = 4\sqrt{\sum_{j\in J(\lambda)}\omega_j^2}/\sqrt{n\kappa_M}$. Applying the inequality $2xy \le x^2/\alpha + \alpha y^2$ $(x, y \in \mathbb{R}, \alpha > 0)$ 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 $|\zeta_{ij}| \le 2L_j$, we obtain

that

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

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

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

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

$$\begin{aligned}\mathbb{P}\{\bar{\omega}_j > \omega_j\} &= \mathbb{P}\{\operatorname{Var}(f_j(X_1)) > T_j^2\}\\ &\le \mathbb{P}\Bigg\{\mathbb{E}f_j^2(X_1) > \frac{2}{n}\sum_{i=1}^{n}f_j^2(X_i) + t_j\Bigg\}\\ &\le \exp\Bigg(-\frac{n\{\mathbb{E}f_j^2(X_1) + t_j\}^2}{8\mathbb{E}f_j^4(X_1)}\Bigg) \qquad \text{using Proposition 2.6 in [45]}\\ &\le \exp\Bigg(-\frac{nt_j\mathbb{E}f_j^2(X_1)}{2\mathbb{E}f_j^4(X_1)}\Bigg) \qquad \text{since } (x+y)^2 \ge 4xy,\end{aligned}$$

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

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

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

which implies the lemma. $\square$

4. Sparse estimation in mixture models. In this section we assume that the true density f can be represented as a finite mixture

$$f(x) = \sum_{j\in I^*} \bar{\lambda}_j p_j(x),$$

where $I^* \subseteq \{1,\dots,M\}$ is unknown, p_j are known probability densities and $\bar{\lambda}_j > 0$ for all $j \in I^*$. We focus in this section on model selection, that is, on the correct identification of the set I^*. It will be convenient for us to normalize the densities p_j by their L_2 norms and to write the model in the form

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

where $I^* \subseteq \{1,\dots,M\}$ is unknown, $f_j = p_j/\|p_j\|$ are known functions and $\lambda_j^* > 0$ for all $j \in I^*$. 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 setup introduced above. We compute the estimates of λ^* via (2.4), with weights defined by [cf. (3.1)]:

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

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 $\max_{1\le j\le M} L_j$. Recall that, by

construction, $\|f_j\| = 1$ for all j. 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 $\rho^* = \rho(\lambda^*)$.

Similar conditions are quite standard in the literature on sparse regression estimation and compressed sensing; cf., for example, [4, 8, 11, 13, 19, 47]. The difference is that those papers use the empirical version of the correlation ρ^* and the numerical constant in the inequality is, in general, different from $1/16$. Note that Condition (A) is quite intuitive. Indeed, the sparsity index k^* can be viewed as the effective dimension of the problem. When k^* increases the problem becomes harder, so that we need stronger conditions (smaller correlations ρ^*) in order to obtain our results. The interesting case that we have in mind is when the effective dimension k^* is small, that is, the model is sparse.

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, for example, [8], 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

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

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*

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

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 $\widehat{I} = J(\widehat{\lambda})$ be the set of indices of the nonzero components of $\widehat{\lambda}$ given by (2.4). In what follows we investigate when $P(\widehat{I} = I^*) \geq 1 - \varepsilon$ for a given $0 < \varepsilon < 1$. Our results are nonasymptotic, 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 $L = \max(1/\sqrt{3}, \max_{1 \leq j \leq M} L_j)$ 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 $\mathbb{P}(\widehat{I} = I^*) \geq 1 - 2\delta(1 + 1/M)$.*

REMARK. Since all λ_j^* are nonnegative, it seems reasonable to restrict the minimization in (2.4) to λ with nonnegative components. Inspection of the proofs shows that all the results of this section remain valid for such a modified estimator. However, in practice, the nonnegativity issue is not so important. Indeed, the estimators of the weights are quite close to the true values and turn out to be positive for positive λ_j^*. For example, this was the case in our simulations discussed in Section 6 below. On the other hand, adding the nonnegativity constraint in (2.4) introduces some extra burden on the numerical algorithm. More generally, it is trivial to note that the results of this and previous sections extend verbatim to the setting where $\lambda \in \Lambda$ with Λ being any subset of $\mathbb{R}^M$. Then the minimization in (2.4) should be performed on Λ, in the theorems of Section 3 we should replace $\lambda \in \mathbb{R}^M$ by $\lambda \in \Lambda$ and in this section λ^* should be supposed to belong to Λ.

PROOF OF THEOREM 5. We begin by noticing that

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

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

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

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

We control the last probability by using the characterization (A.1) 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 $k \in I^*$,

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

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), $\min_{k \in I^*} |\lambda_k^*| \ge 4(\sqrt{2}+1)Lr$, with $r = r(\delta/(2M)) = \{\log(2M^2/\delta)/n\}^{1/2}$. Therefore,

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

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

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

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

Combining the above results, we obtain

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

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

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

Let

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

Consider the random event

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

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 (2.4) on the event $\mathcal{B}$. Recall that $\widehat{\lambda}$ is also a solution of (2.4). By the definition of the set $\widehat{I}$, we have that $\widehat{\lambda}_k \ne 0$ for $k \in \widehat{I}$. By construction, $\tilde{\mu}_k \ne 0$ for some subset $S \subseteq I^*$. By the second part of Lemma 3 in the Appendix, any two solutions have nonzero elements in the same positions. Therefore, $\widehat{I} = S \subseteq I^*$ on $\mathcal{B}$. Thus,

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

Reasoning as in (4.6) above, we find

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

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,

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

Therefore, by Condition (A), we have

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

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

$$P(\widehat{I}\neq I^*)\leq 2\delta+\frac{2\delta}{M},$$

which concludes the proof. □

4.2. *Example*: *Identifying true components in mixtures of Gaussian densities*. Consider an ensemble of M Gaussian densities p_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. The densities are

$$p_j(x)=\frac{1}{(2\pi\tau_j^2)^{d/2}}\exp\Big(-\frac{\|x-\mu_j\|_2^2}{2\tau_j^2}\Big),$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Consequently, $f_j=p_j/\|p_j\|$ with $\|p_j\|=(4\pi\tau_j^2)^{-d/4}$. Recall that Condition (A) requires

$$16\rho^*=16\max_{i\in I^*,j\neq i}|\langle f_i,f_j\rangle|\leq 1/k^*.$$

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

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

Therefore, Condition (A) holds if

$$D_{\min}^2 \geq 4\tau_{\max}^2 \log(16k^*). \tag{4.11}$$

Using this and Theorem 5, we see that SPADES 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

$$L = \max\Bigl(\frac{\sqrt{3}}{3}, \max_{1\leq j\leq M} \|f_j\|_\infty\Bigr) = \max\Bigl(\frac{\sqrt{3}}{3}, (\pi\tau_{\min}^2)^{-d/4}\Bigr),$$

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

REMARK. Often both the location and scale parameters are unknown. In this situation, as suggested by the Associate Editor, the SPADES procedure can be applied to a family of densities with both scale and location parameters chosen from an appropriate grid. By Theorem 1, the resulting estimate will be a good approximation of the unknown target density. An immediate modification of Theorem 5, as in [9], further guarantees that SPADES identifies correctly the important components of this approximation.

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 ψ_{lk} satisfy

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

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, for example, [25]. 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, that is, 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},

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

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

Now, choose $\{f_1, \dots, f_M\} = \{\psi_{lk}, -1 \le l \le l_{\max}, k \in V(l)\}$, 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 $(l,k) \notin \{-1 \le l \le l_{\max}, k \in V(l)\}$. 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 $f_1, \dots, f_M$ 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 $n \ge 1$, $\lambda \in \mathbb{R}^M$ we have, with probability at least $1 - n^{-2}$,*

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

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 $l' \ge 0$ there exists a sequence of coefficients $\lambda = \{\lambda_{(l,k)}, -1 \le l \le l', k \in V(l)\}$ such that

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

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, for example, [25] 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 $1 - n^{-2}$,

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

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

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

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, [6, 22, 25, 27, 35, 36] 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:

$$\mathcal{L}_0(m) = \big\{f : [0,1] \to \mathbb{R} : f \text{ is a probability density and } |\{j : \langle f, f_j\rangle \ne 0\}| \le m\big\},$$

where $m \le M$ is an unknown integer. If $f_1, \dots, f_M$ is a wavelet system as defined above and $J^* = \{j = (l,k) : \langle f, f_j\rangle \ne 0\}$, then under the conditions of Theorem 6 for any $f \in \mathcal{L}_0(m)$ we have, with probability at least $1 - n^{-2}$,

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

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* $n \geq 1$,

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

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, 23] and in the random design regression model [13].

6. Numerical experiments. In this section we describe the algorithm used for the minimization problem (2.4) and we assess the performance of our procedure via a simulation study.

6.1. *A coordinate descent algorithm.* Since the criterion given in (2.4) is convex, but not differentiable, we adopt an optimization by coordinate descent instead of a gradient-based approach (gradient descent, conjugate gradient, etc.) in the spirit of [20, 21]. Coordinate descent is an iterative greedy optimization technique that starts at an initial location $\lambda \in \mathbb{R}^M$ and at each step chooses one coordinate $\lambda_j \in \mathbb{R}$ of λ at random or in order and finds the optimum in that direction, keeping the other variables λ_{-j} fixed at their current values. For convex functions, it usually converges to the global optimum; see [20]. The method is based on the obvious observation that for functions of the type

$$H(\lambda) = g(\lambda) + \omega|\lambda|_1,$$

where g is a generic convex and differentiable function, $\omega > 0$ is a given parameter, and $|\lambda|_1$ denotes the ℓ_1 norm, the optimum in a direction $\lambda_j \in \mathbb{R}$ is to the left, right or at $\lambda_j = 0$, depending on the signs of the left and right partial derivatives of H at zero. Specifically, let g_j denote the partial derivative of g with respect to λ_j, and denote by λ^0_{-j} the vector λ with the jth coordinate set to 0. Then, the minimum in direction j of $H(\lambda)$ is at λ^0_{-j} if and only if $|g_j(\lambda^0_{-j})| < \omega$. This observation makes the coordinate descent become the iterative thresholding algorithm described below.

Coordinate descent.
Given ω, initialize all λ_j, $1 \leq j \leq M$, for example, with $1/M$.

1. Choose a direction $j \in \{1, \ldots, M\}$ and set $\lambda^{\text{old}} = \lambda$.
2. If $|g_j(\lambda^0_{-j})| < \omega$, then set $\lambda = \lambda^0_{-j}$, otherwise obtain λ by line minimization in direction j.

3. If $|\lambda^{\text{old}} - \lambda| > \epsilon$, go to 1, where $\epsilon > 0$ is a given precision level.

For line minimization, we used the procedure `linmin` from Numerical Recipes [34], page 508.

6.2. *Estimation of mixture weights using the generalized bisection method and a penalized cross-validated loss function.* We apply the coordinate descent algorithm described above to optimize the function $H(\lambda)$ given by (2.4), where the tuning parameters ω_j are all set to be equal to the same quantity ω. The theoretical choice of this quantity described in detail in the previous sections may be too conservative in practice. In this section we propose a data driven method for choosing the tuning parameter ω, following the procedure first introduced in [10], which we briefly describe here for completeness.

The procedure chooses adaptively the tuning parameter from a list of candidate values, and it has two distinctive features: the list of candidates is not given by a fine grid of values and the adaptive choice is not given by cross-validation, but by a dimension stabilized cross-validated criterion. We begin by describing the principle underlying our construction of the set of candidate values which, by avoiding a grid search, provides significant computational savings. We use a generalization of the bisection method to find, for each $0 \le k \le M$, a preliminary tuning parameter $\omega = w_k$ that gives a solution $\widehat{\lambda}^k$ with exactly k nonzero elements. Formally, denote by $\widehat{n}(\omega)$ the number of nonzero elements in the λ obtained by minimizing (2.4) with $\omega_j \equiv \omega$ for a given value of the tuning parameter ω. The generalized bisection method will find a sequence of values of the tuning parameter, $w_0, \ldots, w_M$, such that $\widehat{n}(w_k) = k$, for each $0 \le k \le M$. It proceeds as follows, using a queue consisting of pairs (w_i, w_j) such that $\widehat{n}(w_i) < \widehat{n}(w_j) - 1$.

The general bisection method (GBM) for all k.
Initialize all w_i with -1.

1. Choose w_0 very large, such that $\widehat{n}(w_0) = 0$. Choose $w_n = 0$, hence, $\widehat{n}(w_n) = n$.
2. Initialize a queue q with the pair (w_0, w_n).
3. Pop the first pair (a, b) from the queue.
4. Take $w = (a + b)/2$. Compute $k = \widehat{n}(w)$.
5. If $w_k = -1$, make $w_k = w$.
6. If $|\widehat{n}(a) - k| > 1$ and $|a - w| > \alpha$, add (a, w) to the back of the queue.
7. If $|\widehat{n}(b) - k| > 1$ and $|b - w| > \alpha$, add (ω, b) to the back of the queue.
8. If the queue is not empty, go to 3.

This algorithm generalizes the basic bisection method (BBM), which is a well-established computationally efficient method for finding a root $z \in \mathbb{R}$ of a function $h(z)$; see, for example, [15]. We experimentally observed (see also [10] for a detailed discussion) that using the GBM is about 50 times faster than a grid search with the same accuracy.

Our procedure finds the final tuning parameter ω by combining the GBM with the dimension stabilized p-fold cross-validation procedure summarized below. Let D denote the whole data set, and let $D = D_1 \cup \cdots \cup D_p$ be a partition of D in p disjoint subsets. Let $D_{-j} = D \setminus D_j$. We will denote by w_k^j a candidate tuning parameter determined using the GBM on D_{-j}. We denote by I_k^j the set of indices corresponding to the nonzero coefficients of the estimator of λ given by (2.4), for tuning parameter w_k^j on D_{-j}. We denote by $\widehat{\lambda}^{kj}$ the minimizers on D_{-j} of the unpenalized criterion $\widehat{\gamma}(\lambda)$, with respect only to those λ_l with $l \in I_k^j$. Let $L_k^j =: \widehat{\gamma}(\widehat{\lambda}^{kj})$, computed on D_j. With this notation, the procedure becomes the following:

Weight selection procedure.

Given: a data set D partitioned into p disjoint subsets, $D = D_1 \cup \cdots \cup D_p$. Let $D_{-j} = D \setminus D_j$ for all j.

1. For each $1 \le k \le M$ and each fold j of the partition, $1 \le j \le p$:
   Use the GBM to find w_k^j and I_k^j such that $\widehat{n}(w_k^j) = |I_k^j| = k$ on D_{-j}.
   Compute $L_k^j =: \widehat{\gamma}(\widehat{\lambda}^{kj})$, as defined above, on D_j.
2. For each $1 \le k \le M$:
   Compute $L_k =: \frac{1}{p}\sum_{j=1}^{p} L_k^j$.
3. Obtain
$$\widehat{k} = \arg\min_k \left(L_k + 0.5k\frac{\log n}{n} \right).$$
4. With $\widehat{k}$ from Step 3, use the BBM on the whole data set D to find the tuning sequence $w_{\widehat{k}}$ and then compute the final estimators using the coordinate descent algorithm and tuning paramemter $\omega = w_{\widehat{k}}$.

In all the the numerical experiments described below we took the number of splits $p = 10$.

REMARK. We recall that the theoretical results of Section 4.1 show that for correct identification of the mixture components one needs to work with a value of the tuning sequence that is slightly larger than the one needed for good approximations with mixtures of a given density. A good practical approximation of the latter tuning value is routinely obtained by cross-validation; this approximation is, however, not appropriate if the goal is correct selection, when the theoretical results indicate that a different value is needed. Our modification of the cross-validated loss function via a BIC-type penalty is motivated by the known properties of the BIC-type criteria to yield consistent model selection in a large array of models; see, for example, [7] for results on regression models. The numerical experiments presented below show that this is also the case for our criterion in the context of selecting mixture components. The theoretical investigation of this method is beyond the scope of this paper and will be undertaken in future research.

6.3. *Numerical results*. In this subsection we illustrate the performance of our procedure via a simulation study.

6.3.1. *One-dimensional densities*. We begin by investigating the ability of SPADES, with its tuning parameter chosen as above, to (i) approximate well, with respect to the L_2 norm, a true mixture; (ii) to identify the true mixture components. We conducted a simulation study where the true density is a mixture of Gaussian densities with $k^* = 2$ and, respectively, $k^* = 5$ true mixture components. The mixture components are chosen at random from a larger pool of M Gaussians $\mathcal{N}(aj, 1)$, $1 \le j \le M$, where for $k^* = 2$ we take $a = 4$, and for $k^* = 5$ we take $a = 5$. These choices for a ensure that the identifiability condition (4.11) is satisfied. The true components correspond to the first k^* Gaussian densities from our list, and their weights in the true mixture are all equal to $1/k^*$. The maximum size M of the candidate list we considered is $M = 200$, for $k^* = 2$ and $M = 600$, for $k^* = 5$. All the results obtained below are relative to $S = 100$ simulations. Each time, a sample of size n is obtained from the true mixture and is the input of the procedure described in Section 6.2.

We begin by evaluating the accuracy with respect to the L_2 norm of the estimates of f^*. We investigate the sensitivity of our estimates relative to an increase in the dictionary size and k^*. In Figure 1, we plot the median over 100 simulations of $\|f^* - f^{\spadesuit}\|^2$ versus the size M of the dictionary, when the true mixture cardinality is $k^* = 2$ (left panel) and $k^* = 5$ (right panel). For $k^* = 2$ we considered three instances of sample sizes $n = 50, 100, 200$ and we varied M up to 200. For $k^* = 5$ we considered three larger instances of sample sizes $n = 300, 400, 600$ and we varied M up to 600. These experiments provide strong support for our theoretical results: the increase in M does not significantly affect the quality of estimation, and an increase in k^* does. For larger values of k^* we need larger sample sizes to obtain good estimation accuracy.

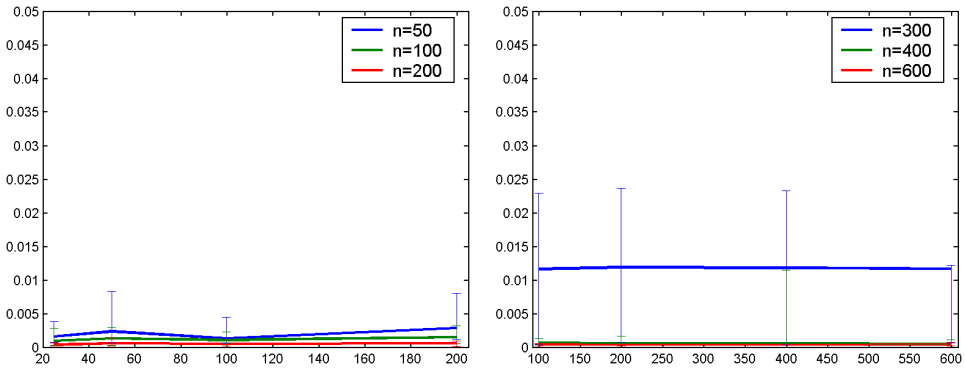

FIG. 1. *Median L_2 error $\|f^* - f^{\spadesuit}\|^2$ for $|I^*| = 2$, respectively, $|I^*| = 5$. The error bars are the 25 and 75 percentiles.*

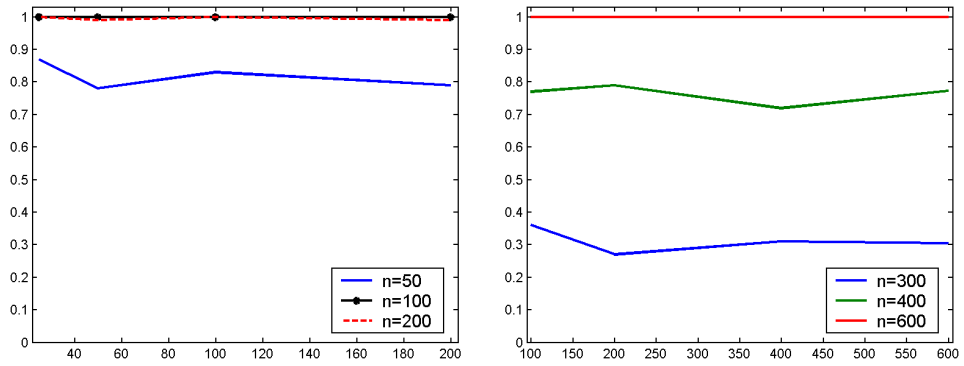

FIG. 2. *Percentage of times* $I^* = \widehat{I}$ *obtained from* 100 *runs, for* $|I^*| = 2$, *respectively,* $|I^*| = 5$.

We next investigated the ability of the SPADES to find the exact mixture components. Figure 2 shows a plot of the percentage of times the exact mixture components were found versus M. We considered the same combinations (n, M) as in Figure 1. Again, observe that the performance does not seriously degrade with the dictionary size M, and is almost unaffected by its increase once a threshold sample size is being used. However, notice that on the difference from the results presented in Figure 1, correct identification is poor below the threshold sample size, which is larger for larger k^*. This is in accordance with our theoretical results: recall Condition (B) of Section 4.1 on the minimum size of the mixture weights. Indeed, we designed our simulations so that the weights are relatively small for $k^* = 5$, they are all equal to $1/k^* = 0.2$, and a larger sample size is needed for their correct identification.

Finally, we evaluated in Figure 3 the dependence of the error and hit rate (i.e., the percentage of times $I^* = \widehat{I}$) on the smallest distance $D_{\min} = \min_{k \neq j} |\mu_k - \mu_j|$ between the means of the densities in the dictionary. The results presented in Fig-

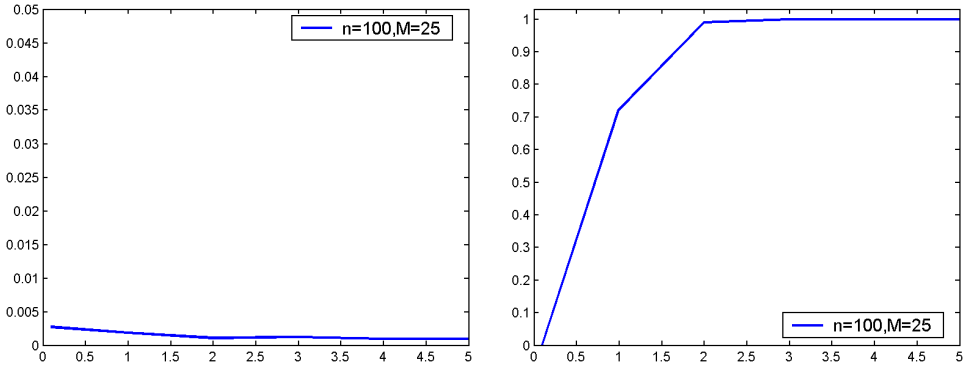

FIG. 3. *Dependence on the distance* $D_{\min} = \min_{k \neq j} |\mu_k - \mu_j|$ *of the* L_2 *error* $\|f^* - f^{\spadesuit}\|^2$ *and the percentage of times* $I^* = \widehat{I}$. *In this example,* $n = 100$, $M = 25$, $|I^*| = 2$.

ures 1 and 2 above were obtained for the value $D_{\min} = 4$, which satisfies the theoretical requirement for correct mixture identification. On the other hand, $D_{\min}$ can be smaller for good L_2 mixture approximation. It is interesting to see what happens when $D_{\min}$ decreases, so that the mixture elements become very close to one another. In Figure 3 we present the simulations for $k^* = 2$, $M = 25$ and $n = 100$, which is sufficient to illustrate this point. We see that, although the L_2 error increases slightly when $D_{\min}$ decreases, the deterioration is not crucial. However, as our theoretical results suggest, the percentage of times we can correctly identify the mixture decreases to zero when the dictionary functions are very close to each other.

6.3.2. *Two-dimensional densities.* In a second set of experiments our aim was to approximate a two-dimensional probability density on a thick circle (cf. the left panel of Figure 5) with a mixture of isotropic Gaussians. A sample of size 2000 from the circle density is shown in the middle panel of Figure 5. We use a set of isotropic Gaussian candidates with covariance $\Sigma = \mathbb{I}_2$ centered at some of the 2000 locations, such that the Euclidean distance between the means of any two such Gaussians is at least 1. We select from these candidate mixture densities in a greedy iterative manner, each time choosing one of the 2000 locations that is at distance at least 1 from each of those already chosen. As a result, we obtain a dictionary of $M = 248$ candidate densities.

The circle density cannot be exactly represented as a finite mixture of Gaussian components. This is a standard instance of many practical applications in Computer Vision, as the statistics of natural images are highly kurtotic and cannot be exactly approximated by isotropic Gaussians. However, in many practical applications a good approximation of an object that reflects its general shape is sufficient and constitutes a first crucial step in any analysis. We show below that SPADES offers such an approximation.

Depending on the application, different trade-offs between the number of mixture components (which relates to the computational demand of the mixture model) and accuracy might be appropriate. For example, in real-time applications a small number of mixture elements would be required to fit into the computational constraints of the system, as long as there is no significant loss in accuracy.

For the example presented below we used the GBM to determine the mixture weights $\widehat{\lambda}^k$, for mixtures with $k = 1, 2, \ldots, 248$ components. Let $\gamma_0 = \min_k \widehat{\gamma}(\widehat{\lambda}^k)$, where we recall that the loss function $\widehat{\gamma}$ is given by (2.2) above. We used the quantity $\widehat{\gamma}(\widehat{\lambda}^k) - \gamma_0$ to measure the accuracy of the mixture approximation. In Figure 4 we plotted $\widehat{\gamma}(\widehat{\lambda}^k) - \gamma_0$ as a function of k and used this plot to determine the desired trade-off between accuracy and mixture complexity. Based on this plot, we selected the number of mixture components to be 80; indeed, including more components does not yield any significant improvement. The obtained mixture is displayed in the right panel of Figure 5. We see that it successfully approximates the circle density with a relatively small number of components.

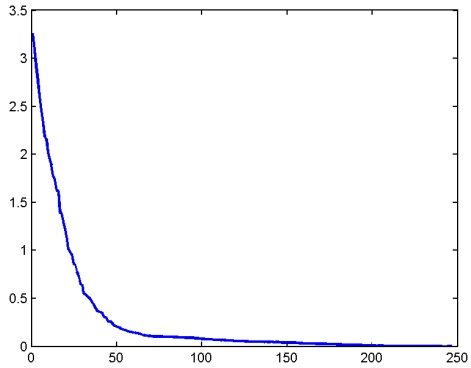

FIG. 4. *Plot of* $\widehat{\gamma}(\widehat{\lambda}^k) - \gamma_0$ *as a function of the mixture components* k.

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 nonzero 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 $g(\lambda)$:

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

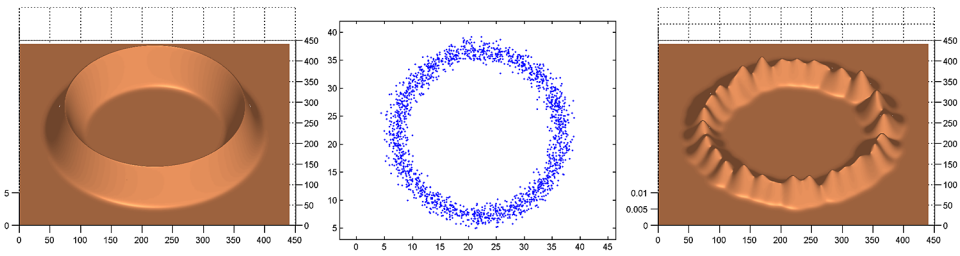

FIG. 5. *A thick circle density, a sample of size* 2000 *from this density and approximations using a mixture of* 80 *isotropic Gaussians.*

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 $1 \le k \le M$,

$$\text{(A.1)} \qquad \frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - \sum_{j=1}^{M} \bar{\lambda}_j \langle f_j, f_k \rangle = 4Lr \operatorname{sign}(\bar{\lambda}_k) \qquad \text{if } \bar{\lambda}_k \neq 0,$$

$$\text{(A.2)} \qquad \left| \frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - \sum_{j=1}^{M} \bar{\lambda}_j \langle f_j, f_k \rangle \right| \le 4Lr \qquad \text{if } \bar{\lambda}_k = 0.$$

We now show that $\bar{\mu} = (\tilde{\mu}, 0) \in \mathbb{R}^M$ with $\tilde{\mu}$ given in (4.8) satisfies (A.1) and (A.2) 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

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

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

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

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

if $k \notin I^*$ (for which $\bar{\mu}_k = 0$, 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 (A.1) and (A.2) 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 (A.1), the index set S of the nonzero 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

$$\text{(A.3)} \qquad \sum_{j=1}^{M} (\bar{\lambda}_j^{(1)} - \bar{\lambda}_j^{(2)}) \langle f_j, f_k \rangle = 0 \qquad \text{for all } k,$$

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

Thus, it remains to show (A.3). 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 $0 < \rho < 1$. Rewrite 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:

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

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

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

By continuity of $\rho \mapsto \bar{\lambda}_j^{(2)} + \rho\eta_j$, there exists an open interval in $(0, 1)$ on which $\rho \mapsto \operatorname{sign}(\bar{\lambda}_j^{(2)} + \rho\eta_j)$ 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 $F'(\rho) = 0, \forall 0 < \rho < 1$, only if

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

and, therefore,

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

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. *Ann. Statist.* **34** 584–653. MR2281879

[2] BIAU, G. and DEVROYE, L. (2005). Density estimation by the penalized combinatorial method. *J. Multivariate Anal.* **94** 196–208. MR2161217

[3] BIAU, G., CADRE, B., DEVROYE, L. and GYÖRFI, L. (2008). Strongly consistent model selection for densities. *TEST* **17** 531–545. MR2470097

[4] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[5] BIRGÉ, L. (2008). Model selection for density estimation with L_2 loss. Available at 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.) 55–87. Springer, New York. MR1462939

[7] BUNEA, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Ann. Statist.* **32** 898–927. MR2065193

[8] BUNEA, F. (2008). Honest variable selection in linear and logistic models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.* **2** 1153–1194. MR2461898

[9] BUNEA, F. (2008). Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* (B. Clarke and S. Ghosal, eds.) **3** 122–137. IMS, Beachwood, OH. MR2459221

[10] BUNEA, F. and BARBU, A. (2009). Dimension reduction and variable selection in case control studies via regularized likelihood optimization. *Electron. J. Stat.* **3** 1257–1287. MR2566187

[11] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[12] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2006). Aggregation and sparsity via ℓ_1-penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory, COLT 2006*. *Lecture Notes in Artificial Intelligence* **4005** 379–391. Springer, Heidelberg. MR2280619

[13] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[14] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Sparse density estimation with ℓ_1 penalties. In *Learning Theory*. *Lecture Notes in Comput. Sci.* **4539** 530–544. Springer, Heidelberg. MR2397610

[15] BURDEN, R. L. and FAIRES, J. D. (2001). *Numerical Analysis*, 7th ed. Brooks/Cole, Pacific Grove, CA.

[16] CHEN, S., DONOHO, D. and SAUNDERS, M. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* **43** 129–159. MR1854649

[17] DEVROYE, L. and LUGOSI, G. (2000). *Combinatorial Methods in Density Estimation*. Springer, New York. MR1843146

[18] DONOHO, D. L. (1995). Denoising via soft-thresholding. *IEEE Trans. Inform. Theory* **41** 613–627. MR1331258

[19] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. MR2237332

[20] FRIEDMAN, J., HASTIE, T., HOFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1** 302–332. MR2415737

[21] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1.
[22] GOLUBEV, G. K. (1992). Nonparametric estimation of smooth probability densties in L_2. *Probl. Inf. Transm.* **28** 44–54. MR1163140
[23] GOLUBEV, G. K. (2002). Reconstruction of sparse vectors in white Gaussian noise. *Probl. Inf. Transm.* **38** 65–79. MR2101314
[24] GREENSHTEIN, E. and RITOV, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988. MR2108039
[25] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, Approximation and Statistical Applications. Lecture Notes in Statistics* **129**. Springer, New York. MR1618204
[26] JAMES, L., PRIEBE, C. and MARCHETTE, D. (2001). Consistent estimation of mixture complexity. *Ann. Statist.* **29** 1281–1296. MR1873331
[27] KERKYACHARIAN, G., PICARD, D. and TRIBOULEY, K. (1996). L^p adaptive density estimation. *Bernoulli* **2** 229–247. MR1416864
[28] KOLTCHINSKII, V. (2005). Model selection and aggregation in sparse classification problems. *Oberwolfach Reports* **2** 2663–2667.
[29] KOLTCHINSKII, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* **45** 7–57.
[30] LOUBES, J.-M. and VAN DE GEER, S. A. (2002). Adaptive estimation in regression, using soft thresholding type penalties. *Statist. Neerlandica* **56** 453–478.
[31] LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087
[32] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363
[33] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics (Saint–Flour, 1998). Lecture Notes in Math.* **1738** 85–277. Springer, Berlin. MR1775640
[34] PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. and FLANNERY, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge Univ. Press, New York. MR2371990
[35] RIGOLLET, P. (2006). Inégalités d'oracle, agrégation et adaptation. Ph.D. thesis, Univ. Paris 6.
[36] RIGOLLET, P. and TSYBAKOV, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280. MR2356821
[37] RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78. MR0668683
[38] SAMAROV, A. and TSYBAKOV, A. (2007). Aggregation of density estimators and dimension reduction. In *Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A. Doksum* (V. Nair, ed.) 233–251. World Scientific, Singapore. MR2416118
[39] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[40] TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines*. *Lecture Notes in Artificial Intelligence* **2777**. Springer, Heidelberg.
[41] VAPNIK, V. N. (1999). *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, New York. MR1367965
[42] VAN DE GEER, S. A. (2008). High dimensional generalized linear models and the Lasso. *Ann. Statist.* **26** 225–287. MR2396809
[43] WASSERMAN, L. A. (2004). *All of Statistics*. Springer, New York. MR2055670
[44] WEGKAMP, M. H. (1999). Quasi-universal bandwidth selection for kernel density estimators. *Canad. J. Statist.* **27** 409–420. MR1704442

[45] WEGKAMP, M. H. (2003). Model selection in nonparametric regression. *Ann. Statist.* **31** 252–273. MR1962506

[46] ZHANG, C. H. and HUANG, J. (2008). The sparsity and biais of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[47] ZHAO, P. and YU, B. (2007). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2567. MR2274449

[48] ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

F. BUNEA
M. WEGKAMP
A. BARBU
DEPARTMENT OF STATISTICS
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FLORIDA 32306-4330
USA
E-MAIL: bunea@stat.fsu.edu
wegkamp@stat.fsu.edu
abarbu@stat.fsu.edu

A. TSYBAKOV
LABORATOIRE DE STATISTIQUE, CREST
92240 MALAKOFF
AND
LPMA (UMR CNRS 7599)
UNIVERSITÉ PARIS 6
75252 PARIS, CEDEX 05
FRANCE
E-MAIL: alexandre.tsybakov@upmc.fr