The Annals of Statistics
2020, Vol. 48, No. 2, 738–762
https://doi.org/10.1214/19-AOS1817
© Institute of Mathematical Statistics, 2020

ON THE NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATOR FOR GAUSSIAN LOCATION MIXTURE DENSITIES WITH APPLICATION TO GAUSSIAN DENOISING

BY SUJAYAM SAHA* AND ADITYANAND GUNTUBOYINA**

*Department of Statistics, University of California, Berkeley, *sujayam@berkeley.edu; **aditya@stat.berkeley.edu*

We study the nonparametric maximum likelihood estimator (NPMLE) for estimating Gaussian location mixture densities in d-dimensions from independent observations. Unlike usual likelihood-based methods for fitting mixtures, NPMLEs are based on convex optimization. We prove finite sample results on the Hellinger accuracy of every NPMLE. Our results imply, in particular, that every NPMLE achieves near parametric risk (up to logarithmic multiplicative factors) when the true density is a discrete Gaussian mixture without any prior information on the number of mixture components. NPMLEs can naturally be used to yield empirical Bayes estimates of the oracle Bayes estimator in the Gaussian denoising problem. We prove bounds for the accuracy of the empirical Bayes estimate as an approximation to the oracle Bayes estimator. Here our results imply that the empirical Bayes estimator performs at nearly the optimal level (up to logarithmic factors) for denoising in clustering situations without any prior knowledge of the number of clusters.

1. Introduction. In this paper, we study the performance of the nonparametric maximum likelihood estimator (NPMLE) for estimating a Gaussian location mixture density in multiple dimensions. We also study the performance of the empirical Bayes estimator based on the NPMLE for estimating the oracle Bayes estimator in the problem of Gaussian denoising.

By a Gaussian location mixture density in $\mathbb{R}^d$, $d \geq 1$, we refer to a density of the form

$$f_G(x) := \int \phi_d(x-\theta)\, dG(\theta) \tag{1.1}$$

for some probability G on $\mathbb{R}^d$ where $\phi_d(z) := (2\pi)^{-d/2} \exp(-\|z\|^2/2)$ is the standard d-dimensional normal density ($\|z\|$ is the usual Euclidean norm of z). Note that f_G is the density of the random vector $X = \theta + Z$ where θ and Z are independent d-dimensional random vectors with θ having distribution G (i.e., $\theta \sim G$) and Z having the Gaussian distribution with zero mean and identity covariance matrix (i.e., $Z \sim N(0, I_d)$). We let $\mathcal{M}$ to be the class of all Gaussian location mixture densities, that is, densities of the form f_G as G varies over all probability measures on $\mathbb{R}^d$.

Given n independent d-dimensional data vectors $X_1, \ldots, X_n$ (throughout the paper, we assume that $n \geq 2$) generated from an unknown Gaussian location mixture density $f^* \in \mathcal{M}$, we study the problem of estimating f^* from $X_1, \ldots, X_n$. This problem is fundamental to the area of estimation in mixture models to which a number of books (see, e.g., Everitt and Hand [21], Titterington, Smith and Makov [55], Lindsay [34], Böhning [7], McLachlan and Peel [41], Schlattmann [50]) and papers have been devoted. We focus on the situation where d is small or moderate, n is large and where no specific prior information is available

Received December 2017; revised October 2018.
MSC2010 subject classifications. 62G07, 62C12, 62C10.
Key words and phrases. Density estimation, Gaussian mixture model, convex optimization, Hellinger distance, metric entropy, rate of convergence, model selection, adaptive estimation, convex clustering.

about the mixing measure corresponding to f^*. Consistent estimation in the case where d is comparable in size to n needs simplifying assumptions on f^* (such as that the mixing measure is discrete with a small number of atoms and that it is concentrated on a set of sparse vectors in $\mathbb{R}^d$) which we do not make in this paper. Let us also note here that we focus on the problem of estimating f^* and not on estimating the mixing measure corresponding to f^*.

There are two well-known likelihood-based approaches to estimating Gaussian location mixtures: (a) the first approach involves fixing an integer k and performing maximum likelihood estimation over $\mathcal{M}_k$ which is the collection of all densities $f_G \in \mathcal{M}$ where G is discrete and has at most k atoms, and (b) the second approach involves performing maximum likelihood estimation over the entire class $\mathcal{M}$. This results in the nonparametric maximum likelihood estimator (NPMLE) for f^* and is the focus of this paper.

The first approach (maximum likelihood estimation over $\mathcal{M}_k$ for a fixed k) is quite popular. However, it suffers from the two well-known issues: choosing k is nontrivial and, moreover, maximizing likelihood over $\mathcal{M}_k$ results in a nonconvex optimization problem. This nonconvex algorithm is usually approximately solved by the EM algorithm (see, e.g., Dempster, Laird and Rubin [14], McLachlan and Krishnan [42], Watanabe and Yamaguchi [58]). Recent progress on obtaining a theoretical understanding of the behaviour of the nonconvex EM algorithm has been made by Balakrishnan et al. [2]. Analyzing these estimators for data-dependent choices of k is well known to be difficult. Maugis and Michel [39] (see also Maugis-Rabusseau and Michel [40]) proposed a penalization likelihood criterion to choose k by suitably employing the general theory of nonasymptotic model selection via penalization due to Birgé and Massart [5], Barron, Birgé and Massart [3] and Massart [37] and, moreoever, Maugis and Michel [39] established nonasymptotic risk properties of the resulting estimator. The computational aspects of their estimator are quite involved however (see Maugis and Michel [38]) as their estimators are based on solving multiple nonconvex optimization problems.

The present paper studies the second likelihood-based approach involving nonparametric maximum likelihood estimation of f^*. This method is unaffected by nonconvexity and the need for choosing k. Formally, by an NPMLE, we mean any maximizer $\hat{f}_n$ of $\sum_{i=1}^n \log f(X_i)$ as f varies over $\mathcal{M}$:

$$\hat{f}_n \in \operatorname*{argmax}_{f \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \log f(X_i). \tag{1.2}$$

Note that because the maximization is over the entire class $\mathcal{M}$ of all Gaussian location mixtures (and not on any nonconvex subset such as $\mathcal{M}_k$), the optimization in (1.2) is a convex problem. Indeed, the objective function in (1.2) is concave in f and the constraint set $\mathcal{M}$ is a convex class of densities.

The idea of using NPMLEs for estimating mixture densities has a long history (see, e.g., the classical references Kiefer and Wolfowitz [27], Lindsay [32–34], Böhning [7]). The optimization problem (1.2) and its solutions have been studied by many authors. It is known that maximizers of $f \mapsto \sum_{i=1}^n \log f(X_i)$ exist over $\mathcal{M}$ which implies that NPMLEs exist. Maximizers are nonunique, however, so there exist multiple NPMLEs. Nevertheless, for every NPMLE $\hat{f}_n$, the values $\hat{f}(X_i)$ for $i = 1, \ldots, n$ are unique (this is essentially because the objective function in the optimization (1.2) only depends on f through the values $f(X_1), \ldots, f(X_n)$). Proofs of these basic facts can be found, for example, in Böhning [7], Chapter 2.

There exist many algorithms in the literature for approximately solving the optimization (1.2) (note that though (1.2) is a convex optimization problem, it is infinite-dimensional which is probably why exact algorithms seem to be unavailable). These algorithms range from: (a) vertex direction methods and vertex exchange methods (see the review papers Böhning [6],

Lindsay and Lesperance [35] and the references therein), (b) EM algorithms (see Laird [29] and Jiang and Zhang [25]) and (c) modern large-scale interior point methods (see Koenker and Mizera [28] and Feng and Dicker [22]). Most of these methods focus on the case $d = 1$ and involve maximizing the likelihood over mixture densities where the mixing measure is supported on a fixed fine grid in the range of the data. The algorithm of Koenker and Mizera [28] is highly scalable (relying on the commercial convex optimization library Mosek [43]) and can obtain an approximate NPMLE efficiently even for large sample sizes (n of the order 100,000). See Section 5 for more algorithmic and implementation details as well as some simulation results.

Let us now describe the main objectives and contributions of the current paper. Our first goal is to investigate the theoretical properties of NPMLEs. In particular, we study the accuracy of $\hat{f}_n$ as an estimator of the density f^* from which the data $X_1, \ldots, X_n$ are generated. We shall use, as our loss function, the squared Hellinger distance

$$\mathfrak{H}^2(f, g) := \int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx, \tag{1.3}$$

which is one of the most commonly used loss functions for density estimation problems. We present a detailed analysis of the risk, $\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f^*)$, of every NPMLE (the expectation here is taken with respect to $X_1, \ldots, X_n$ distributed independently according to f^*). The other common loss function used in density estimation is the total variation distance. The total variation distance is bounded from above by a constant multiple of $\mathfrak{H}$ so that upper bounds for risk under the squared Hellinger distance automatically imply upper bounds for risk in squared total variation distance.

Our results imply that, for a large class of true densities $f^* \in \mathcal{M}$, the risk of every NPMLE $\hat{f}_n$ is parametric (i.e., n^{-1}) up to multiplicative factors that are logarithmic in n. In particular, our results imply that when $f^* \in \mathcal{M}_k$ for some $1 \le k \le n$, then every NPMLE has risk k/n up to a logarithmic multiplicative factor in n. It is not hard to see that the minimax risk over $\mathcal{M}_k$ is bounded from below by k/n which implies therefore that every NPMLE is nearly minimax over $\mathcal{M}_k$ (ignoring logarithmic factors in n) for every $k \ge 1$. This is interesting because NPMLEs do not use any a priori knowledge of k. The price in squared Hellinger risk that is paid for not knowing k in advance is only logarithmic in n. Our results are nonasymptotic and the bounds for risk over $\mathcal{M}_k$ hold even when $k = k(n)$ grows with n. Our results also imply that NPMLEs have parametric risk (again up to multiplicative logarithmic factors) when the mixing measure of f^* is supported on a fixed compact subset of $\mathbb{R}^d$. Note that we have assumed that the covariance matrix of every Gaussian component of mixture densities in the class $\mathcal{M}$ is the identity matrix. Our results can be extended to the case of arbitrary and unknown covariance matrices provided a lower bound on the eigenvalues is available (see Theorem 2.5) (on the other hand, when no a priori information on the covariance matrices is available, it is well known that likelihood-based approaches are infeasible). These results are described in Section 2.

Previous results on the Hellinger accuracy of NPMLEs were due to Zhang [63] (see also Ghosal and van der Vaart [23] for related results) who dealt with the univariate ($d = 1$) case. Here the Hellinger accuracy was analyzed under conditions on the moments of the mixing measure corresponding to f^*. The accuracy of NPMLEs in the interesting case when $f^* \in \mathcal{M}_k$ does not appear to have been studied previously even in $d = 1$. We study the Hellinger risk of NPMLEs for all $d \ge 1$ and also under a much broader set of assumptions on f^* compared to existing papers.

We would like to mention here that numerous papers have appeared in the theoretical computer science community establishing rigorous theoretical results for estimating densities in $\mathcal{M}_k$. For example, the papers Daskalakis and Kamath [13], Suresh et al. [52], Bhaskara et al.

[4], Chan et al. [10, 11], Acharya et al. [1], Li and Schmidt [31] have results on estimating densities in $\mathcal{M}_k$ with rigorous bounds on the error in estimation. The estimation error is mostly measured in terms of the total variation distance which is smaller (up to constant multiplicative factors) compared to the Hellinger distance used in the present paper. Their sample complexity results imply rates of estimation of k/n up to logarithmic factors in n for densities in $\mathcal{M}_k$ in terms of the squared total variation distance and hence these results are comparable to our results for the NPMLE. The estimation procedures used in these papers range from (a) hypothesis selection over a set of candidate estimators via an improved version of the Scheffé estimate ([13, 52]; see Devroye and Lugosi [15], Chapter 6, for background on the Scheffé estimate), (b) reduction to finding sparse solutions to a nonnegative linear systems [4] and (c) fitting piecewise polynomial densities ([1, 10, 11, 31]; these papers have the sharpest results). These methods are very interesting and, remarkably, come with precise time complexity guarantees. They are not based on likelihood maximization, however, and in our opinion, conceptually more involved compared to the NPMLE. An additional minor difference between our work and this literature is that k is taken to be a constant (and sometimes even known) in these papers while we allow $k = k(n)$ to grow with n and, moreover, the NPMLE does not need any prior knowledge of k.

Let us now describe briefly the proof techniques underlying our risk results for the NPMLEs. Our technical arguments are based on standard ideas from the literature on empirical processes for assessing the performance of maximum likelihood estimators (see van der Vaart and Wellner [56], Wong and Shen [60], Zhang [63]). These techniques involve bounding the covering numbers of the space of Gaussian location mixture densities. For each compact subset $S \subseteq \mathbb{R}^d$, we prove covering number bounds for $\mathcal{M}$ under the supremum distance (L_∞) on S. Our bounds can be seen as extensions of the one-dimensional covering number results of Zhang [63] (which are themselves enhancements of corresponding results in Ghosal and van der Vaart [23]). The covering number results of Zhang [63] can be viewed as special instances of our bounds for the case when $S = [-M, M]$. The extension to arbitrary compact sets S is crucial for dealing with rates for densities in $\mathcal{M}_k$. For proving the final Hellinger risk bounds of $\hat{f}_n$ from these L_∞ covering numbers, we use appropriate modifications of tail arguments from Zhang [63]. A sketch of these ideas is given in Section 4.1.

The second goal of the present paper is to use NPMLEs to yield empirical Bayes estimates in the Gaussian denoising problem. By Gaussian denoising, we refer to the problem of estimating vectors $\theta_1, \ldots, \theta_n \in \mathbb{R}^d$ from independent d-dimensional observations $X_1, \ldots, X_n$ generated as

$$X_i \sim N(\theta_i, I_d) \quad \text{for } i = 1, \ldots, n. \tag{1.4}$$

The naive estimator in this denoising problem simply estimates each θ_i by X_i. It is well known that, depending on the structure of the unknown $\theta_1, \ldots, \theta_n$, it is possible to achieve significant improvement over the naive estimator by using information from X_j, $j \neq i$ in addition to X_i for estimating θ_i. An ideal prototype for such information sharing across observations is given by the *oracle Bayes* estimator which will be denoted by $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ and is defined in the following way:

$$\hat{\theta}_i^* := \mathbb{E}(\theta | X = X_i) \quad \text{where } \theta \sim \bar{G}_n \text{ and } X|\theta \sim N(\theta, I_d)$$

and $\bar{G}_n$ is the empirical measure corresponding to the true set of parameters $\theta_1, \ldots, \theta_n$. In other words, $\hat{\theta}_i^*$ is the posterior mean of θ given $X = X_i$ under the model $X|\theta \sim N(\theta, I_d)$ and the prior $\theta \sim \bar{G}_n$. This is an oracle estimator that is infeasible in practice as it uses information on the unknown parameters $\theta_1, \ldots, \theta_n$ via their empirical measure $\bar{G}_n$. It has the

important well-known property (see, e.g., Robbins [46]) that $\hat{\theta}_i^*$ can be written as $T^*(X_i)$ for each $i = 1, \ldots, n$ where $T^* : \mathbb{R}^d \to \mathbb{R}^d$ minimizes

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|T(X_i) - \theta_i\|^2\right] \tag{1.5}$$

over all possible functions $T : \mathbb{R}^d \to \mathbb{R}^d$. Estimators for $\theta_1, \ldots, \theta_n$ which are of the form $T(X_1), \ldots, T(X_n)$ for a single nonrandom function T are known as separable estimators and the best separable estimator is given by $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$. We shall show that $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ can be estimated accurately by a natural estimator constructed using any NPMLE (1.2) based on $X_1, \ldots, X_n$.

To motivate the estimator, observe first that it is well known (see, e.g., Robbins [47], Brown [8], Stein [51], Efron [18]) that $\hat{\theta}_i^*$ has the following alternative expression as a consequence of Tweedie's formula:

$$\hat{\theta}_i^* = X_i + \frac{\nabla f_{\bar{G}_n}(X_i)}{f_{\bar{G}_n}(X_i)}, \tag{1.6}$$

where $f_{\bar{G}_n}$ is the Gaussian location mixture density with mixing measure $\bar{G}_n$ (defined as in (1.1)). From the above expression, it is clear that the oracle Bayes estimator can be estimated from the data $X_1, \ldots, X_n$ provided one can estimate the Gaussian location mixture density, $f_{\bar{G}_n}$, from the data $X_1, \ldots, X_n$. For this purpose, as insightfully observed in Jiang and Zhang [25], any NPMLE, $\hat{f}_n$, as in (1.2) can be used. Indeed, if $\hat{f}_n$ denotes any NPMLE based on the data $X_1, \ldots, X_n$, then Jiang and Zhang [25] argued that $\hat{f}_n$ is a good estimator for $f_{\bar{G}_n}$ under (1.4) so that $\hat{\theta}_i^*$ is estimable by

$$\hat{\theta}_i := X_i + \frac{\nabla \hat{f}_n(X_i)}{\hat{f}_n(X_i)}. \tag{1.7}$$

This yields a completely tuning-free solution to the Gaussian denoising problem (note, however, that the noise distribution is assumed to be completely known as $N(0, I_d)$). This is the general maximum likelihood empirical Bayes estimator of Jiang and Zhang [25] who proposed it and studied its theoretical properties in detail for estimating sparse univariate normal means. To the best of our knowledge, the properties of the estimator (1.7) for multidimensional denoising problems have not been previously explored. More generally, the empirical Bayes approach to the Gaussian denoising problem goes back to Robbins [45, 46, 48]. The effectiveness of nonparametric empirical Bayes estimators for estimating sparse normal means has been explored by many authors including Johnstone and Silverman [26], Brown and Greenshtein [9], Jiang and Zhang [25], Donoho and Reeves [17], Koenker and Mizera [28] but most work seems restricted to the univariate setting. On the other hand, there exists prior work on parametric empirical Bayes methods in the multivariate Gaussian denoising problem (see, e.g., [19, 20]) but the role of nonparametric empirical Bayes methods in multivariate Gaussian denoising does not seem to have been explored previously.

We perform a detailed study of the accuracy of $\hat{\theta}_i$ in (1.7) as an estimator of the oracle Bayes estimator $\hat{\theta}_i^*$ for $i = 1, \ldots, n$ in terms of the following squared error risk measure:

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|\hat{\theta}_i - \hat{\theta}_i^*\|^2\right], \tag{1.8}$$

where the expectation is taken with respect to $X_1, \ldots, X_n$ generated independently according to (1.4). The risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ depends on the configuration of the unknown parameters $\theta_1, \ldots, \theta_n$ and we perform a detailed study of the risk for natural configurations of the points

$\theta_1, \ldots, \theta_n \in \mathbb{R}^d$. Our results imply that, under natural assumptions on $\theta_1, \ldots, \theta_n$, the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is bounded by the parametric rate $1/n$ up to logarithmic multiplicative factors. For example, when the number of distinct vectors among $\theta_1, \ldots, \theta_n$ equals $k = k(n)$ for some $k \leq n$ (an assumption which makes sense in clustering situations), we prove that the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is bounded from above by the parametric rate k/n up to logarithmic multiplicative factors in n. This result is especially remarkable because the estimator (1.7) is tuning free and does not have knowledge of k. We also prove that the analogous minimax risk over this class is bounded from below by k/n implying that the empirical Bayes estimate is minimax up to logarithmic multiplicative factors. Our result also implies that when $\theta_1, \ldots, \theta_n$ take values in a bounded region on $\mathbb{R}^d$, then also the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is nearly parametric. Summarizing, our results imply that, under a wide range of assumptions on $\theta_1, \ldots, \theta_n$, the empirical Bayes estimator $\hat{\theta}_i$ performs comparably to the oracle Bayes estimator $\hat{\theta}_i^*$ for denoising. We also prove some results about denoising in the heteroscedastic setting where the data $X_1, \ldots, X_n$ are independently generated according to $X_i \sim N(\theta_i, \Sigma_i)$ for more general unknown covariance matrices $\Sigma_1, \ldots, \Sigma_n$. These results are in Section 3. The results and the proof techniques are inspired by the arguments of Jiang and Zhang [25] who studied the univariate denoising problem under sparsity assumptions. We generalize their arguments to multidimensions; a sketch of our proof techniques is provided in Section 4.2.

In addition to theoretical results, we also present simulation evidence for the effectiveness of $\hat{\theta}_i$ in the Gaussian denoising problem in Section 5 (where we also present some implementation and algorithmic details for computing approximate NPMLEs). Here, we illustrate the performance of (1.7) for denoising when the true parameter vectors $\theta_1, \ldots, \theta_n$ take values in certain natural regions in $\mathbb{R}^2$. We also numerically analyze the performance of (1.7) in clustering situations when $\theta_1, \ldots, \theta_n$ take k distinct values for some small k (these results are given in Section G of the Supplementary Material [49]). Here we compare the performance of (1.7) to other natural procedures such as k-means with k selected via the gap statistic (see Tibshirani, Walther and Hastie [54]). We argue that (1.7) performs efficiently in terms of the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$. In terms of a purely clustering based comparison index (such as the Adjusted Rand Index), we argue that the performance of (1.7) is still reasonable.

The rest of the paper is organized in the following manner. In Section 2, we state our results on the Hellinger accuracy of NPMLEs for estimating Gaussian location mixture densities. Section 3 has statements of our results on the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ in the denoising problem. An overview of the key ideas in the proofs of the main results is given in Section 4. Section 5 has algorithmic details and simulation evidence for the effectiveness of (1.7) for denoising. Due to space constraints, complete proofs of all the results in the paper are given in the Supplementary Material [49]. Specifically, proofs for results in Section 2 are given in Section A while proofs for Section 3 are in Section B of [49]. Some additional observations on the heteroscedastic Gaussian denoising problem are also in the Supplementary Material (see Section C of [49]). Metric entropy results for multivariate Gaussian location mixture densities play a crucial rule in the proofs of the main results; these results are proved in [49], Section D. [49], Section E, contains the statement and proof for a crucial ingredient for the proof of the main denoising theorem. Finally, additional technical results needed in the proofs of the main results are collected in [49], Section F, together with their proofs while additional simulation results are in [49], Section G.

2. Hellinger accuracy of NPMLE. For data $X_1, \ldots, X_n$, let $\hat{f}_n$ be any NPMLE defined as in (1.2). In this section, we study the accuracy of $\hat{f}_n$ in terms of the squared Hellinger distance (defined in (1.3)). All the results in this section are fully proved in the Supplementary Material [49], Section A, while Section 4.1 contains a sketch of the key ideas in the proof of Theorem 2.1 (which is the main result of this section).

For investigations into the performance of $\hat{f}_n$, it is most natural to assume that the data $X_1, \ldots, X_n$ are independent observations having common density $f^* \in \mathcal{M}$ in which case we seek bounds on $\mathfrak{H}^2(\hat{f}_n, f^*)$. However, following Zhang [63], we work under the more general assumption that $X_1, \ldots, X_n$ are independent but not identically distributed and that each X_i has a density that belongs to the class $\mathcal{M}$. This additional generality will be used in Section 3 for proving results on the empirical Bayes estimator (1.7) for the Gaussian denoising problem.

Specifically, we assume that $X_1, \ldots, X_n$ are independent and that each X_i has density f_{G_i} for some probability measures $G_1, \ldots, G_n$ on $\mathbb{R}^d$. This distributional assumption on the data $X_1, \ldots, X_n$ includes the following two important special cases: (a) $G_1, \ldots, G_n$ are all identically equal to G (say): in this case, the observations $X_1, \ldots, X_n$ are identically distributed with common density $f^* = f_G \in \mathcal{M}$, and (b) Each G_i is degenerate at some $\theta_i \in \mathbb{R}^d$: here each data point X_i is normal with $X_i \sim N_d(\theta_i, I_d)$ and this has been referred to as the compound decision setting by Robbins.

We let $\bar{G}_n := (G_1 + \cdots + G_n)/n$ to be the average of the probability measures $G_1, \ldots, G_n$. In the case when $G_1, \ldots, G_n$ are all identically equal to G, then clearly $\bar{G}_n = G$. On the other hand, when each G_i is degenerate at some $\theta_i \in \mathbb{R}^d$ (i.e., the compound decision setting), then $\bar{G}_n$ equals the empirical measure corresponding to $\theta_1, \ldots, \theta_n$.

Under the above *independent but not identically distributed* assumption on $X_1, \ldots, X_n$, it has been insightfully pointed out by Zhang [63] that every NPMLE $\hat{f}_n$ based on $X_1, \ldots, X_n$ (defined as in (1.2)) is really estimating $f_{\bar{G}_n}$. Note that $f_{\bar{G}_n}$ denotes the average of the densities of $X_1, \ldots, X_n$.

In this section, we shall prove bounds for the accuracy of any NPMLE $\hat{f}_n$ as an estimator for $f_{\bar{G}_n}$ under the Hellinger distance that is, for $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$. In order to state our main theorem, we need to introduce the following notation. For nonempty sets $S \subseteq \mathbb{R}^d$, we define the function $\mathfrak{d}_S : \mathbb{R}^d \to [0, \infty)$ by

$$\mathfrak{d}_S(x) := \inf_{u \in S} \|x - u\| \quad \text{for } x \in \mathbb{R}^d, \tag{2.1}$$

where $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^d$. Also for $S \subseteq \mathbb{R}^d$, we let

$$S^1 := \{x : \mathfrak{d}_S(x) \le 1\}. \tag{2.2}$$

Our bound on $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$ will be controlled by the following quantity. For every probability measure G on $\mathbb{R}^d$, every nonempty compact set $S \subseteq \mathbb{R}^d$ and every $M \ge \sqrt{10 \log n}$, let $\epsilon_n(M, S, G)$ be defined via

$$\begin{aligned} \epsilon_n^2(M, S, G) := {} & \mathrm{Vol}(S^1) \frac{M^d}{n} (\sqrt{\log n})^{d+2} \\ & + (\log n) \inf_{p \ge \frac{d+1}{2 \log n}} \left(\frac{2 \mu_p(\mathfrak{d}_S, G)}{M} \right)^p, \end{aligned} \tag{2.3}$$

where S^1 is defined in (2.2) and $\mu_p(\mathfrak{d}_S, G)$ is defined as the moment

$$\mu_p(\mathfrak{d}_S, G) := \left(\int_{\mathbb{R}^d} (\mathfrak{d}_S(\theta))^p \, dG(\theta) \right)^{1/p} \quad \text{for } p > 0.$$

Note that the moments $\mu_p(\mathfrak{d}_S, G)$ quantify how the probability (under G) decays as one moves away from the set S.

The next theorem proves that $\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n})$ is bounded (with high probability and in expectation) by a constant (depending on d) multiple of $\epsilon_n^2(M, S, \bar{G}_n)$ for every estimator $\hat{f}_n$ having the property that the likelihood of the data at $\hat{f}_n$ is not too small compared to the likelihood at $f_{\bar{G}_n}$ (made precise in inequality (2.4)). Every NPMLE trivially satisfies this condition

(as it maximizes likelihood) but the theorem also applies to certain approximate likelihood maximizers.

The bound given the following theorem holds for every compact set $S \subseteq \mathbb{R}^d$ and $M \geq \sqrt{10 \log n}$. As will be seen later in this section, under some simplifying assumptions on $\bar{G}_n$, our bound for $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$ can be optimized over S and M to produce an explicit bound.

THEOREM 2.1. *Let $X_1, \ldots, X_n$ be independent random vectors with $X_i \sim f_{G_i}$ and let $\bar{G}_n := (G_1 + \cdots + G_n)/n$. Fix $M \geq \sqrt{10 \log n}$ and a nonempty compact set $S \subseteq \mathbb{R}^d$ and let $\epsilon_n(M, S, \bar{G}_n)$ be defined via (2.3). Then there exists a positive constant C_d (depending only on d) such that for every estimator $\hat{f}_n$ based on the data $X_1, \ldots, X_n$ satisfying*

$$\prod_{i=1}^{n} \frac{\hat{f}_n(X_i)}{f_{\bar{G}_n}(X_i)} \geq \exp\left[\frac{C_d(\beta - \alpha)}{\min(1-\alpha, \beta)} n \epsilon_n^2(M, S, \bar{G}_n)\right] \tag{2.4}$$

for some $0 < \beta \leq \alpha < 1$, we have

$$\mathbb{P}\left\{\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq \frac{t \epsilon_n(M, S, \bar{G}_n)\sqrt{C_d}}{\sqrt{\min(1-\alpha, \beta)}}\right\} \leq 2n^{-t^2} \tag{2.5}$$

for every $t \geq 1$ and, moreover,

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq \frac{4C_d}{\min(1-\alpha, \beta)} \epsilon_n^2(M, S, \bar{G}_n). \tag{2.6}$$

Theorem 2.1 asserts that the risk $\mathbb{E}\mathfrak{H}^2(\hat{f}_n, \bar{f}_{\bar{G}_n})$ is bounded from above by a constant (depending on d, α and β) multiple of $\epsilon_n^2(M, S, \bar{G}_n)$ for every $M \geq \sqrt{10 \log n}$ and compact subset $S \subseteq \mathbb{R}^d$. This is true for every estimator $\hat{f}_n$ satisfying (2.4). Every NPMLE satisfies (2.4) with $\alpha = \beta = 0.5$ (note that the right-hand side of (2.4) is always less than or equal to one because $\beta \leq \alpha$).

Theorem 2.1 is novel to the best of our knowledge. When $d = 1$ and S is taken to be $[-R, R]$ for some $R \geq 0$, then the conclusion given by Theorem 2.1 appears implicitly in Zhang [63], proof of Theorem 1. The presence of an arbitrary compact set S allows the derivation of interesting adaptation results for discrete mixing distributions (as will be clear from the special cases of Theorem 2.1 that are given below). Such results cannot be derived if the arbitrary S is replaced by only a box or a ball such as $[-R, R]$ as in the univariate result of Zhang [63]. Indeed, suppose that $\bar{G}_n$ is a discrete measure gives equal probability to the two points R and $-R$ for a large value of R. Then the bound of Zhang [63] gives a multiplicative factor involving R in the risk bounds which make them quite suboptimal when R is large. On the other hand, Theorem 2.1 applied with $S = \{-R, R\}$ gives a near-parametric risk bound (see Theorem 2.3 below). One can further think of the support of $\bar{G}_n$ being a collection of discrete points, curves and regions (all the while being bounded) for general $d \geq 2$, where a direct extension of Zhang's result would produce an upper bound directly proportional to the volume of the bounding box of the shapes mentioned above; while our result will depend on the total volume of the *fattenings* of each of the shapes described above. In cases where the total fattened volume is a constant while the separation between the different shapes increases as a function n, our result will yield a tighter upper bound (as a negative power of n) than Zhang's result and its naive multidimensional extension.

Our proof of Theorem 2.1 (given in Section A of the Supplementary Material [49]) is greatly inspired by Zhang [63], proof of Theorem 1. An overview of this proof is provided in Section 4.1 where we explain the main ideas as well as points of departure between our proof and the arguments in Zhang [63], proof of Theorem 1.

To get the best rate for $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$ from Theorem 2.1, we need to choose M and S so that $\epsilon_n(M, S, \bar{G}_n)$ is small. These choices obviously depend on $\bar{G}_n$ and in the next result, we describe how to choose M and S based on reasonable assumptions on $\bar{G}_n$. This leads to explicit rates for $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$. Note that, more generally, Theorem 2.1 implies that $\hat{f}_n$ is consistent (in the Hellinger distance) for $f_{\bar{G}_n}$ provided $\bar{G}_n$ is such that

$$\inf_{S \text{ compact}, M \geq \sqrt{10 \log n}} \epsilon_n(M, S, \bar{G}_n) \to 0 \quad \text{as } n \to \infty.$$

For simplicity, we shall assume, for the next result, that $\hat{f}_n$ is an NPMLE so that (2.4) is satisfied with $\alpha = \beta = 0.5$. We shall also only state the results on the risk $\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n})$.

COROLLARY 2.2. *Let $X_1, \ldots, X_n$ be independent random vectors with $X_i \sim f_{G_i}$ and let $\bar{G}_n := (G_1 + \cdots + G_n)/n$. Let $\hat{f}_n$ be an NPMLE based on $X_1, \ldots, X_n$ defined as in (1.2). Below C_d denotes a positive constant depending on d alone.*

1. *Suppose that $\bar{G}_n$ is supported on a compact subset S of $\mathbb{R}^d$. Then*

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq C_d \frac{\mathrm{Vol}(S^1)}{n} (\log n)^{d+1}. \tag{2.7}$$

2. *Suppose there exists a compact subset $S \subseteq \mathbb{R}^d$ and real numbers $0 < \alpha \leq 2$ and $K \geq 1$ such that*

$$\mu_p(\mathfrak{d}_S, \bar{G}_n) \leq K p^{1/\alpha} \quad \text{for all } p \geq 1. \tag{2.8}$$

Then

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq C_d \frac{\mathrm{Vol}(S^1)(K e^{1/\alpha})^d}{n} (\sqrt{\log n})^{(2d/\alpha)+d+2}. \tag{2.9}$$

3. *Suppose there exists a compact set $S \subseteq \mathbb{R}^d$ and real numbers $\mu > 0$ and $p > 0$ such that $\mu_p(\mathfrak{d}_S, \bar{G}_n) \leq \mu$. Then there exists a positive constant $C_{d,\mu,p}$ (depending only on d, μ and p) such that*

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq C_{d,\mu,p} \left(\frac{\mathrm{Vol}(S^1)}{n} \right)^{\frac{p}{p+d}} (\sqrt{\log n})^{\frac{2d+2p+dp}{p+d}}. \tag{2.10}$$

Corollary 2.2 is a generalization of Zhang ([63], Theorem 1), as the latter result can be seen as a special case of Corollary 2.2 for $d = 1$ and $S = [-R, R]$ for some $R \geq 0$. The fact that S can be arbitrary in Corollary 2.2 allows us to deduce the following important adaptation results of NPMLEs for estimating Gaussian mixtures whose mixing measures are discrete. These results are, to the best of our knowledge, novel.

THEOREM 2.3 (Near parametric risk for discrete Gaussian mixtures). *Let $X_1, \ldots, X_n$ be independent random vectors with $X_i \sim f_{G_i}$ and let $\bar{G}_n := (G_1 + \cdots + G_n)/n$. Let $\hat{f}_n$ be an NPMLE based on $X_1, \ldots, X_n$ defined as in (1.2). Then there exists a positive constant C_d depending only on d such that whenever $\bar{G}_n$ is a discrete probability measure that is supported on a set of cardinality k, we have*

$$\mathbb{E}\mathfrak{H}^2(\hat{f}_n, f_{\bar{G}_n}) \leq C_d \left(\frac{k}{n} \right) (\log n)^{d+1}. \tag{2.11}$$

Note that (2.11) directly follows from (2.7). Indeed, when $\bar{G}_n$ is supported on a finite set S of cardinality k, we can apply inequality (2.7) to this S. It is easy to see then that $\mathrm{Vol}(S^1) \le C_d k$ which proves (2.11).

The significance of Theorem 2.3 is the following. The right-hand side of (2.11) is the parametric risk k/n up to an additional multiplicative factor that is logarithmic in n. This inequality shows important adaptation properties of NPMLEs. When the true unknown Gaussian mixture $f_{\bar{G}_n}$ is a discrete mixture having k Gaussian components, then every NPMLE nearly (up to logarithmic factors) achieves the parametric squared Hellinger risk k/n. For a fixed k, it is well known that fitting a k-component Gaussian mixture via maximum likelihood is a nonconvex problem that is usually solved by the EM algorithm. On the other hand, NPMLE is given by a convex optimization algorithm, does not require any prior specification of k and still achieves the k/n rate (up to logarithmic factors) when the truth is a k-component Gaussian mixture. We would also like to stress here that in Theorem 2.3 (and all other results in the paper), k is allowed to grow with n (we can write $k(n)$ instead of k but we are sticking to k for simplicity of notation).

Note that Theorem 2.3 applies to the case of independent but not identically distributed $X_1, \ldots, X_n$ which is more general compared to the i.i.d. assumption. This implies, in particular, that (2.11) also applies to the case when $X_1, \ldots, X_n$ are i.i.d. having density $f^* \in \mathcal{M}$. In this case, we have

$$\sup_{f^* \in \mathcal{M}_k} \mathbb{E}\mathfrak{H}^2(\hat{f}_n, f^*) \le C_d \left(\frac{k}{n}\right)(\log n)^{d+1}. \tag{2.12}$$

The interesting aspect of this inequality is that it holds for every $k \ge 1$ and that the estimator $\hat{f}_n$ does not know or use any information about k.

It is straightforward to prove a minimax lower bound over $\mathcal{M}_k$ that complements Theorem 2.3. The following result proves that the minimax risk over $\mathcal{M}_k$ is bounded from below by a constant multiple of k/n. This implies that the NPMLE is minimax optimal over $\mathcal{M}_k$ ignoring logarithmic factors of n. Moreover, this optimality is adaptive since MLE does not require knowledge of k. This minimax lower bound is stated for the i.i.d. case which implies that it holds for the more general independent but not identically distributed case as well.

LEMMA 2.4. *For $k \ge 1$, let*

$$\mathcal{R}(\mathcal{M}_k) := \inf_{\tilde{f}} \sup_{f \in \mathcal{M}_k} \mathbb{E}_f \mathfrak{H}^2(\tilde{f}, f),$$

where $\mathbb{E}_f$ denotes expectation when the data $X_1, \ldots, X_n$ are independent observations drawn from the density f. Then there exists a universal positive constant C such that

$$\mathcal{R}(\mathcal{M}_k) \ge C\frac{k}{n} \quad \text{for every } 1 \le k \le n. \tag{2.13}$$

Inequality (2.12) and Lemma 2.4 together imply that every NPMLE $\hat{f}_n$ is minimax optimal up to logarithmic factors in n over the class $\mathcal{M}_k$ for every $k \ge 1$. This optimality is adaptive since the NPMLE requires no information on k. The logarithmic term in (2.12) are likely suboptimal but we are unable to determine the exact power of $\log n$ in (2.12).

So far we have studied estimation of Gaussian location mixture densities where the covariance matrix of each Gaussian component is fixed to be the identity matrix. We next show that the same estimator (NPMLE defined as in (1.2)) can be modified to estimate arbitrary Gaussian mixtures (where the covariance matrices can be different from identity) provided

a lower bound on the eigenvalues of the covariance matrices is available. Suppose that h^* is the Gaussian mixture density

$$h^*(x) := \sum_{j=1}^{k} w_j \phi_d(x; \mu_j, \Sigma_j) \quad \text{for } x \in \mathbb{R}^d, \tag{2.14}$$

where $k \geq 1$, $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ and $\Sigma_1, \ldots, \Sigma_k$ are $d \times d$ positive definite matrices. Here $\phi_d(\cdot; \mu, \Sigma)$ denotes the d-variate normal density with mean μ and covariance matrix Σ. Suppose $\sigma^2_{\min}$ and $\sigma^2_{\max}$ are two positive numbers that are, respectively, smaller and larger than all the eigenvalues of $\Sigma_1, \ldots, \Sigma_k$, that is,

$$\sigma^2_{\min} \leq \min_{1 \leq j \leq k} \lambda_{\min}(\Sigma_j) \leq \max_{1 \leq j \leq k} \lambda_{\max}(\Sigma_j) \leq \sigma^2_{\max}. \tag{2.15}$$

Consider the problem estimating h^* from i.i.d. observations $Y_1, \ldots, Y_n$. It turns out that for every NPMLE $\hat{f}_n$ computed as in (1.2) based on the data $X_1 := Y_1/\sigma_{\min}, \ldots, X_n := Y_n/\sigma_{\min}$ can be converted to a very good estimator for h^* via

$$\hat{h}_n(x) := \sigma^{-d}_{\min} \hat{f}_n\big(\sigma^{-1}_{\min} x\big) \quad \text{for } x \in \mathbb{R}^d. \tag{2.16}$$

Our next result shows that the squared Hellinger risk of $\hat{h}_n$ is bounded from above by (k/n) up to a logarithmic factor in n provided that $\sigma_{\max}/\sigma_{\min}$ is bounded by a constant. This result implies that applying the NPMLE to $Y_i/\sigma_{\min}$ leads to a very accurate estimator even for heteroscedastic normal observations.

THEOREM 2.5. *Let $Y_1, \ldots, Y_n$ be independent and identically distributed observations having density h^* defined in (2.14). Consider the estimator $\hat{h}_n$ for h^* defined in (2.16). Then*

$$\mathbb{E}\mathfrak{H}^2\big(\hat{h}_n, h^*\big) \leq C_d \left(\frac{k}{n}\right) \big(\max(1, \tau)\big)^d (\log n)^{d+1}, \tag{2.17}$$

where $\tau := \sqrt{\frac{\sigma^2_{\max}}{\sigma^2_{\min}} - 1}$.

Theorem 2.5 shows that $\hat{h}_n$ achieves near parametric risk k/n (up to logarithmic factors in n) provided τ is bounded from above by a constant. Note that this estimator $\hat{h}_n$ uses knowledge of $\sigma^2_{\min}$ but does not use knowledge of any other feature of h^* including the number of components k. In particular, this is an estimation procedure which (without knowing the value of k) achieves nearly the k/n rate for k-component well-conditioned Gaussian mixtures provided a lower bound $\sigma^2_{\min}$ on eigenvalues is known a priori.

It is natural to compare Theorem 2.5 to the main results in Maugis and Michel [39] where an adaptive procedure is developed for estimating k-component Gaussian mixtures at the rate k/n (up to a logarithmic factor) without prior knowledge of k. The estimator of Maugis and Michel [39] is very different from ours. They first fit m-component Gaussian mixtures for different values of m and then select one of these estimators by optimizing a penalized model-selection criterion. Thus, their procedure is based on solving multiple nonconvex optimization problems. Also, Maugis and Michel [39] impose upper and lower bounds on the means and the eigenvalues of the covariance matrices of the components of the mixture densities. On the contrary, our method is based on convex optimization and we only need a lower bound on the eigenvalues of the covariance matrices (no bounds on the means are necessary). On the flip side, the result of Maugis and Michel [39] has much better logarithmic factors compared to Theorem 2.5 and it is also stated in the form of an oracle inequality.

3. Application to Gaussian denoising. In this section, we explore the role of the NPMLE for estimating the oracle Bayes estimator in the Gaussian denoising problem. All the results in this section are proved in the Supplementary Material [49], Section B.

The goal is to estimate unknown vectors $\theta_1, \ldots, \theta_n \in \mathbb{R}^d$ in the compound decision setting where we observe independent random vectors $X_1, \ldots, X_n$ such that $X_i \sim N(\theta_i, I_d)$ for $i = 1, \ldots, n$. The Oracle estimator is $\hat{\theta}_i^*, i = 1, \ldots, n$ which is given by (1.6) where $\bar{G}_n$ is the empirical measure corresponding to $\theta_1, \ldots, \theta_n$.

It is natural to estimate the oracle Bayes estimator by the Empirical Bayes estimator $\hat{\theta}_i$ which is defined as in (1.7) for $i = 1, \ldots, n$. Here $\hat{f}_n$ is any NPMLE based on $X_1, \ldots, X_n$ (defined as in (1.2)). We will gauge the performance of $\hat{\theta}_i, i = 1, \ldots, n$ as an estimator for $\hat{\theta}_i^*, i = 1, \ldots, n$ in terms of the squared error risk measure $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ defined in (1.8).

The main theorem of this section is given below. This is stated in a form that is similar to the statement of Theorem 2.1. It proves that, for every compact set $S \subseteq \mathbb{R}^d$ and $M \geq \sqrt{10 \log n}$, the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is bounded from above by $\epsilon_n^2(M, S, \bar{G}_n)$ (defined via (2.3)) up to the additional logarithmic multiplicative factor $(\log n)^{\max(d,3)}$. This additional logarithmic factor is a consequence of our proof technique.

THEOREM 3.1. *Let $X_1, \ldots, X_n$ with independent random vectors with $X_i \sim N(\theta_i, I_d)$ for $i = 1, \ldots, n$. Let $\bar{G}_n$ denote the empirical measure corresponding to $\theta_1, \ldots, \theta_n$. Let $\hat{f}_n$ denote an NPMLE based on $X_1, \ldots, X_n$ defined as in (1.2). Let $\hat{\theta}_1, \ldots, \hat{\theta}_n$ be as defined in (1.7) and let $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ be as in (1.6). Also, let $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ be as in (1.8). There exists a positive constant C_d (depending only on d) such that for every nonempty compact set $S \subseteq \mathbb{R}^d$ and $M \geq \sqrt{10 \log n}$, we have*

$$\mathfrak{R}_n\big(\hat{\theta}, \hat{\theta}^*\big) \leq C_d \epsilon_n^2(M, S, \bar{G}_n)(\sqrt{\log n})^{\max(d-2,6)}.$$

REMARK 3.1. For the case of $d = 1$, Jiang and Zhang ([25], Theorem 5), established a related result on the risk of $\hat{\theta}_i$ in comparison to $\hat{\theta}_i^*$. The risk used therein is

$$\left[\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n |\hat{\theta}_i - \theta_i|^2\right)\right]^{1/2} - \left[\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n |\hat{\theta}_i^* - \theta_i|^2\right)\right]^{1/2}. \tag{3.1}$$

Jiang and Zhang [25] investigated the above risk in the case where $d = 1$ and $S = [-R, R]$ for some $R \geq 0$. The statement of Theorem 3.1 and its proof, as well as the following corollary, are inspired by Jiang and Zhang [25], proof of Theorem 5.

Under specific reasonable assumptions on $\bar{G}_n$, it is possible to choose M and S explicitly which leads to the following result that is analogous to Corollary 2.2.

COROLLARY 3.2. *Consider the same setting and notation as in Theorem 3.1. Below C_d denotes a positive constant depending on d alone.*

1. *For every compact set $S \subseteq \mathbb{R}^d$ containing all the points $\theta_1, \ldots, \theta_n$, we have*

$$\mathfrak{R}_n\big(\hat{\theta}, \hat{\theta}^*\big) \leq C_d \frac{\mathrm{Vol}(S^1)}{n}(\sqrt{\log n})^{\max(3d, 2d+8)}. \tag{3.2}$$

2. *For every compact subset $S \subseteq \mathbb{R}^d$ and real numbers $0 < \alpha \leq 2$ and $K \geq 1$ satisfying (2.8), we have*

$$\mathfrak{R}_n\big(\hat{\theta}, \hat{\theta}^*\big) \leq C_d \frac{\mathrm{Vol}(S^1)(K e^{1/\alpha})^d}{n}(\sqrt{\log n})^{\max(\frac{2d}{\alpha}+2d, \frac{2d}{\alpha}+d+8)}. \tag{3.3}$$

3. *Suppose $S \subseteq \mathbb{R}^d$ is compact and real numbers $\mu > 0$ and $p > 0$ are such that $\mu_p(\mathfrak{d}_S, \bar{G}_n) \leq \mu$. Then there exists a positive constant $C_{d,\mu,p}$ (depending only on d, μ and p) such that*

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) \leq C_{d,\mu,p}\left(\frac{\mathrm{Vol}(S^1)}{n}\right)^{\frac{p}{p+d}} (\sqrt{\log n})^{\frac{2d+2p+dp}{p+d}+\max(d-2,6)}. \tag{3.4}$$

Corollary 3.2 has interesting consequences. Inequality (3.2) states that when $\bar{G}_n$ is supported on a fixed compact set S, then the risk $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ is parametric up to logarithmic multiplicative factors in n. This is especially interesting because $\hat{\theta}_1, \ldots, \hat{\theta}_n$ do not use any knowledge of S.

Corollary 3.2 also leads to the following result which gives an upper bound for $\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*)$ when $\theta_1, \ldots, \theta_n$ are clustered into k groups.

PROPOSITION 3.3. *Consider the same setting and notation as in Theorem 3.1. Suppose that $\theta_1, \ldots, \theta_n$ satisfy*

$$\max_{1 \leq i \leq n} \min_{1 \leq j \leq k} \|\theta_i - a_j\| \leq R \tag{3.5}$$

for some $a_1, \ldots, a_k \in \mathbb{R}^d$ and $R \geq 0$. Then

$$\mathfrak{R}_n(\hat{\theta}, \hat{\theta}^*) \leq C_d(1+R)^d \left(\frac{k}{n}\right)(\sqrt{\log n})^{\max(3d,2d+8)}. \tag{3.6}$$

The assumption (3.5) means that $\theta_1, \ldots, \theta_n$ can be grouped into k balls each of radius R centered at the points $a_1, \ldots, a_k$. When R is not large, this implies $\theta_1, \ldots, \theta_n$ can be clustered into k groups. In particular, when $R = 0$, the assumption (3.5) implies that $\theta_1, \ldots, \theta_n$ take only k distinct values. In words, Proposition 3.3 states that when $\theta_1, \ldots, \theta_n$ are clustered into k groups, then $\hat{\theta}_1, \ldots, \hat{\theta}_n$ estimate $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ in squared error loss with accuracy k/n up to logarithmic multiplicative factors in n. The notable aspect about this result is that the estimator does not use any knowledge of k and is tuning-free. It is well known in the clustering literature that choosing the optimal number of clusters is challenging (see, e.g., Tibshirani, Walther and Hastie [54]). It is therefore helpful that $\hat{\theta}_1, \ldots, \hat{\theta}_n$ achieves nearly the k/n rate in (3.5) without explicitly getting into the pesky problem of estimating k. Moreover, $\hat{\theta}_1, \ldots, \hat{\theta}_n$ is given by convex optimization (on the other hand, one usually needs to deal with nonconvex optimization problems for solving clustering-type problems even if the number of clusters k is known).

There exist techniques for estimating the number of clusters and subsequently employing algorithms for minimizing the k-means objective (notably, the "gap statistic" of Tibshirani, Walther and Hastie [54]). However, we are not aware of any result analogous to Proposition 3.3 for such techniques. There also exist other techniques for clustering based on convex optimization such as the method of convex clustering (see, e.g., Lindsten, Ohlsson and Ljung [36], Hocking et al. [24], Chen et al. [12]) which is based on a fused lasso-type penalized optimization. This method requires specification of tuning parameters. While interesting theoretical development exists for convex clustering (see, e.g., Radchenko and Mukherjee [44], Zhu et al. [64], Tan and Witten [53], Wu et al. [61], Wang et al. [57]), to the best of our knowledge, a result similar to Proposition 3.3 is unavailable.

It is straightforward to see that it is impossible to devise estimators that achieve a rate that is faster than k/n for the risk measure $\mathfrak{R}_n$. We provide a proof of this via a minimax lower bound in the following lemma. The logarithmic factors can probably be improved in Proposition 3.3 but we are unable to do so at the present moment. For the lower bound,

let $\Theta_{n,d,k}$ denote the class of all n-tuples $(\theta_1, \ldots, \theta_n)$ with each $\theta_i \in \mathbb{R}^d$ and such that the number of distinct vectors among $\theta_1, \ldots, \theta_n$ is equal to k. Equivalently, $\Theta_{n,d,k}$ consists of all n-tuples $(\theta_1, \ldots, \theta_n)$ whose empirical measure is supported on a set of cardinality k. The minimax risk for estimating $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ with $(\theta_1, \ldots, \theta_n) \in \Theta_{n,d,k}$ in squared error loss from the observations $X_1, \ldots, X_n$ can be defined as

$$\mathcal{R}^*(\Theta_{n,d,k}) := \inf_{\tilde{\theta}_1, \ldots, \tilde{\theta}_n} \sup_{(\theta_1, \ldots, \theta_n) \in \Theta_{n,d,k}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \|\tilde{\theta}_i - \hat{\theta}_i^*\|^2\right].$$

The following result proves that $\mathcal{R}^*(\Theta_{n,d,k})$ is at least Ck/n for a universal positive constant C.

LEMMA 3.4. *Let $\Theta_{n,d,k}$ and $\mathcal{R}^*(\Theta_{n,d,k})$ be defined as above. There exists a universal positive constant C such that*

$$\mathcal{R}^*(\Theta_{n,d,k}) \geq C\frac{k}{n} \quad \text{for every } 1 \leq k \leq n. \tag{3.7}$$

Lemma 3.4, together with Proposition 3.3, implies that $\hat{\theta}_1, \ldots, \hat{\theta}_n$ is nearly minimax optimal (up to logarithmic multiplicative factors) for estimating $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ over the class $\Theta_{n,d,k}$. Moreover, this optimality is adaptive over k because the estimator does not use any knowledge of k.

Before closing this section, let us remark that Theorem 3.1 can be generalized to work with certain kinds of heteroscedasticity in the Gaussian observations. Concretely, consider the problem of heteroscedastic Gaussian denoising where the goal is to estimate $\theta_1, \ldots, \theta_n$ from independent observations $X_1, \ldots, X_n$ generated according to

$$X_i \sim N(\theta_i, \Sigma_i) \tag{3.8}$$

for some *unknown* covariance matrices $\Sigma_1, \ldots, \Sigma_n$. We work with the assumption that $\Sigma_i - I_d$ is positive semidefinite (or equivalently, $\lambda_{\min}(\Sigma_i) \geq 1$) for each $i = 1, \ldots, n$. If $\Sigma_i - \sigma_{\min}^2 I_d$ is positive semidefinite for some other known positive constant $\sigma_{\min}^2$, then one can reduce this to the previous case by simply scaling the observations $X_1, \ldots, X_n$ by $\sigma_{\min}^2$.

Note that we are considering the setting where $\Sigma_1, \ldots, \Sigma_n$ are unknown (satisfying $\Sigma_i - I_d$ is positive semidefinite). This is different from the setting where $\Sigma_1, \ldots, \Sigma_n$ are exactly known and there has been previous work in empirical Bayes estimation under this latter assumption (see, e.g., Xie, Kou and Brown [62] and Weinstein et al. [59]).

Under the assumption that $\Sigma_i - I_d$ is positive semidefinite, it is clear that (3.8) is equivalent to the statement that $X_i \sim f_{G_i^0}$ where G_i^0 is the $N(\theta_i, \Sigma_i - I_d)$ distribution (here we take $N(\theta_i, \Sigma_i - I_d)$ to be the Dirac probability measure centered at θ_i if $\Sigma_i = I_d$). Therefore, as we have seen in Section 2, the estimator $\hat{f}_n$ based on $X_1, \ldots, X_n$ (defined as in (1.2)) will be an accurate estimator of $f_{\bar{G}_n^0}$ where

$$\bar{G}_n^0 := \frac{1}{n}\sum_{i=1}^{n} N(\theta_i, \Sigma_i - I_d) \tag{3.9}$$

under reasonable assumptions on $\theta_1, \ldots, \theta_n$ provided $\sigma_{\max}$ is not too large (here $\sigma_{\max}^2$ is any upper bound on $\max_{1 \leq i \leq n} \lambda_{\max}(\Sigma_i)$). As a result, it is reasonable to believe that $\hat{\theta}_1, \ldots, \hat{\theta}_n$ (defined in (1.7)) will be close to $\breve{\theta}_1^*, \ldots, \breve{\theta}_n^*$ where

$$\breve{\theta}_i^* := X_i + \frac{\nabla f_{\bar{G}_n^0}(X_i)}{f_{\bar{G}_n^0}(X_i)} \quad \text{for } i = 1, \ldots, n. \tag{3.10}$$

The next result rigorizes this intuition. Note that $\breve{\theta}_i^*$ is also given by

$$\breve{\theta}_i^* = \mathbb{E}(\theta | X = X_i) \quad \text{where } \theta \sim \bar{G}_n^0 \text{ and } X|\theta \sim N(\theta, I_d). \tag{3.11}$$

Intuitively, it makes sense that $\hat{\theta}_i$ estimates $\breve{\theta}_i^*$ because an observation $X \sim N(\theta_0, \Sigma)$ (with $\Sigma - I_d$ being positive semidefinite) can also be thought of as being generated from $X|\theta \sim N(\theta, I_d)$ with $\theta \sim N(\theta_0, \Sigma - I_d)$. However, it should be noted that $\breve{\theta}_1^*, \ldots, \breve{\theta}_n^*$ is not the best separable estimator for $\theta_1, \ldots, \theta_n$ in the heteroscedastic setting and this is explained later in this section (after Proposition C.1).

THEOREM 3.5. *Let $X_1, \ldots, X_n$ be independent random vectors with $X_i \sim N(\theta_i, \Sigma_i)$ for some covariance matrices $\Sigma_1, \ldots, \Sigma_n$ with $\Sigma_i - I_d$ being positive semidefinite for every i. Suppose $\sigma^2_{\max}$ is such that $\max_{1 \le j \le k} \lambda_{\max}(\Sigma_j) \le \sigma^2_{\max}$ where $\lambda_{\max}(\Sigma_j)$ denotes the largest eigenvalue of Σ_j. Let $\hat{\theta}_1, \ldots, \hat{\theta}_n$ be as defined in (1.7) and $\breve{\theta}_1^*, \ldots, \breve{\theta}_n^*$ be as defined in (3.10). Then there exists a positive constant C_d (depending only on d) such that for every nonempty compact set $S \subseteq \mathbb{R}^d$ and $M \ge \sqrt{10 \log n}$, we have*

$$\mathfrak{R}_n(\hat{\theta}, \breve{\theta}^*) := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\hat{\theta}_i - \breve{\theta}_i^*\|^2\right] \le C_d \sigma^2_{\max} \epsilon_n^2(M, S, \bar{G}_n^0)(\sqrt{\log n})^{\max(d-2,6)},$$

where $\epsilon_n(M, S, \bar{G}_n^0)$ is as defined in (2.3).

Note that Theorem 3.5 generalizes Theorem 3.1. Indeed, Theorem 3.1 is the special case of Theorem 3.5 when $\Sigma_i = I_d$ for each i because, in this special case, $\sigma^2_{\max} = 1$ and $\bar{G}_n^0$, as defined in (3.9), precisely equals the empirical measure corresponding to $\theta_1, \ldots, \theta_n$. Theorem 3.5 leads to corollaries that are similar to those derived from Theorem 3.1 (see, e.g., Proposition C.1 in the Supplementary Material [49] which is the analogue of Proposition 3.3 for the heteroscedastic setting.

We would like to remark here that Theorem 3.5 is of limited interest unless the heteroscedasticity is mild (by mild, we mean that $\sigma^2_{\max}$ can be chosen to be close to 1). This is because the oracle estimator $\breve{\theta}_i^*$ (defined in (3.10)) is different from the best separable estimator (recall the best separable estimator is given by $T^*(X_i), i = 1, \ldots, n$ where T^* minimizes (1.5) over all functions $T : \mathbb{R}^d \to \mathbb{R}^d$). A description of the best separable estimator along with some results on the discrepancy between the best separable estimator and (3.10) is given in the Supplementary Material [49], Section C.

4. Proof ideas. In this section, we provide a broad overview of the proofs of our main results, Theorem 2.1 and Theorem 3.5. Full proofs of these theorems, of the remaining results in the paper as well as statements and proofs of the supporting results that are used in the proofs are given in the Supplementary Material [49].

4.1. *Proof overview of Theorem 2.1.* Every estimator satisfying (2.4) is an approximate MLE. Therefore the general theory of the rates of convergence of maximum likelihood estimators from, say, van der Vaart and Wellner [56], Wong and Shen [60] can be used to bound $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n})$. This general theory requires bounds on the covering numbers of the underlying class of densities (covering numbers are formally defined at the beginning of Section A in the Supplementary Material [49]). In our particular context, we need to bound covering numbers of the class $\mathcal{M}$ (which consists of all densities of the form f_G as G varies over all probability measures on $\mathbb{R}^d$). Our main covering number result for $\mathcal{M}$ is stated next.

For compact $S \subseteq \mathbb{R}^d$, let $\|\cdot\|_S$ and $\|\cdot\|_{S,\nabla}$ denote pseudonorms given by

$$\|f\|_S := \sup_{x \in S} |f(x)| \quad \text{and} \quad \|f\|_{S,\nabla} := \sup_{x \in S} \|\nabla f(x)\|$$

for densities $f \in \mathcal{M}$. These naturally lead to two pseudometrics on $\mathcal{M}$ and we shall denote the η-covering numbers of $\mathcal{M}$ under these pseudometrics by $N(\eta, \mathcal{M}, \|\cdot\|_S)$ and $N(\eta, \mathcal{M}, \|\cdot\|_{S,\nabla})$, respectively. The following theorem, which could be of independent interest, gives upper bounds for $N(\eta, \mathcal{M}, \|\cdot\|_S)$ and $N(\eta, \mathcal{M}, \|\cdot\|_{S,\nabla})$. We let $S^a := \{x : \mathfrak{d}_S(x) \le a\}$ for $S \subseteq \mathbb{R}^d$ and $a > 0$ and use $N(a, S^a)$ to denote the a-covering number (in the usual Euclidean distance) of the set S^a.

THEOREM 4.1. *There exists a positive constant C_d depending on d alone such that for every compact set $S \subseteq \mathbb{R}^d$ and $0 < \eta \le \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\sqrt{e}}$, we have*

$$\log N(\eta, \mathcal{M}, \|\cdot\|_S) \le C_d N(a, S^a) |\log \eta|^{d+1} \tag{4.1}$$

and

$$\log N(\eta, \mathcal{M}, \|\cdot\|_{S,\nabla}) \le C_d N(a, S^a) |\log \eta|^{d+1}, \tag{4.2}$$

where a is defined as

$$a := \sqrt{2 \log \frac{2\sqrt{2\pi}}{(2\pi)^{d/2}\eta}}. \tag{4.3}$$

To the best of our knowledge, Theorem 4.1 (proved in the Supplementary Material [49], Section D) is novel, although certain special cases (such as when $d = 1$ and S is a closed interval) are known previously (see the Supplementary Material [49], Remark D.1). The generalization for arbitrary compact sets S is crucial for our results. Only the first assertion (inequality (4.1)) is required for the proof of Theorem 2.1; the second assertion involving gradients is needed for the proof of Theorem 3.5.

Let us now sketch the proof of Theorem 2.1 assuming Theorem 4.1. The reader is welcome to read the full proof in the Supplementary Material. As mentioned previously, our proof is inspired from Zhang ([63], proof of Theorem 1) and differences between our proof and the arguments of [63] are pointed out at the end of this subsection.

For simplicity, in this section, let us assume that $\hat{f}_n$ is an NPMLE so that (2.4) holds for $\alpha = \beta = 0.5$. The full proof (in the Supplementary Material [49]) applies to estimators satisfying (2.4) for arbitrary $0 < \beta \le \alpha < 1$. Note first that trivially (for every $t \ge 1$ and $\gamma_n > 0$)

$$\mathbb{P}\{\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \ge t\gamma_n\} = \mathbb{P}\left\{\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \ge t\gamma_n, \prod_{i=1}^n \frac{\hat{f}_n(X_i)}{f_{\bar{G}_n}(X_i)} \ge 1\right\}.$$

The right-hand side above can be easily controlled if $\hat{f}_n$ were nonrandom. To deal with randomness, we cover $\mathcal{M}$ to within some $\eta > 0$ in $L^\infty(S^M)$ (where $S^M := \{x : \mathfrak{d}_S(x) \le M\}$). From this cover, it is possible to deduce the existence of a collection of nonrandom densities h_{0j}, $j \in J$ in $\mathcal{M}$ for some finite set J with cardinality at most the right-hand side of (4.1) such that $\mathfrak{H}(h_{0j}, f_{\bar{G}_n}) \ge t\gamma_n$ and such that the inequality

$$\prod_{i=1}^n \hat{f}_n(X_i) \le \max_{j \in J} \prod_{i: X_i \in S^M} \{h_{0j}(X_i) + 2\eta\} \prod_{i: X_i \notin S^M} (2\pi)^{-d/2}$$

holds whenever $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \ge t\gamma_n$. From here, it can be shown that for every function $v : \mathbb{R}^d \to (0, \infty)$ with $v(x) = \eta$ for $x \in S^M$, we have

$$\prod_{i=1}^n \frac{\hat{f}_n(X_i)}{f_{\bar{G}_n}(X_i)} \le \max_{j \in J} \prod_{i=1}^n \frac{h_{0j}(X_i) + 2v(X_i)}{f_{\bar{G}_n}(X_i)} \prod_{i: X_i \notin S^M} \frac{(2\pi)^{-d/2}}{2v(X_i)}$$

on the event $\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n$. We take

$$v(x) := \begin{cases} \eta & \text{if } x \in S^M, \\ \eta\Big(\dfrac{M}{\mathfrak{d}_S(x)}\Big)^{d+1} & \text{otherwise.} \end{cases} \tag{4.4}$$

The inequality above implies that

$$\mathbb{P}\{\mathfrak{H}(\hat{f}_n, f_{\bar{G}_n}) \geq t\gamma_n\} \leq \sum_{j\in J} \mathbb{P}\left\{\prod_{i=1}^n \frac{h_{0j}(X_i) + 2v(X_i)}{f_{\bar{G}_n}(X_i)} \geq e^{-nt^2\gamma_n^2/2}\right\} + \mathbb{P}\left\{\prod_{i:X_i\notin S^M} \frac{(2\pi)^{-d/2}}{2v(X_i)} \geq e^{nt^2\gamma_n^2/2}\right\}.$$

The first term on the right-hand side above is now controlled by standard arguments for bounding likelihood ratio deviations in terms of Hellinger distances (note that $\mathfrak{H}(h_{0j}, f_{\bar{G}_n}) \geq t\gamma_n$). For the second term, we use Markov's inequality and the following moment inequality (proved in Section F of the Supplementary Material [49]) applied to the Lipschitz function $g(x) := \mathfrak{d}_S(x)$.

LEMMA 4.2. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim f_{G_i}$ and $\bar{G}_n := (G_1 + \cdots + G_n)/n$. Let $g : \mathbb{R}^d \to [0, \infty)$ be a 1-Lipschitz function, that is, $g(x) - g(y) \leq \|x - y\|$ for all $x, y \in \mathbb{R}^d$. Also let $\mu_p(g)$ denote the pth moment of g under the measure $\bar{G}_n$, that is,*

$$\mu_p(g) := \left(\int_{\mathbb{R}^d} g(\theta)^p \, d\bar{G}_n(\theta)\right)^{1/p}.$$

There then exists a positive constant C_d depending only on d such that

$$\mathbb{E}\left\{\prod_{i=1}^n |ag(X_i)|^{I\{g(X_i)\geq M\}}\right\}^\lambda \leq \exp\left\{C_d a^\lambda M^{\lambda+d-2} + (aM)^\lambda n\Big(\frac{2\mu_p(g)}{M}\Big)^p\right\} \tag{4.5}$$

for every $a > 0$, $M \geq \sqrt{8\log n}$ and $0 < \lambda \leq \min(1, p)$.

Further, there exists a positive constant C_d depending only on d such that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{P}\big[g(X_i) \geq M\big] \leq C_d \frac{M^{d-2}}{n} + \inf_{p\geq \frac{d+1}{2\log n}} \Big(\frac{2\mu_p(g)}{M}\Big)^p \tag{4.6}$$

for any $M \geq \sqrt{8\log n}$.

The differences between our proof and that of Zhang ([63], proof of Theorem 1) are the metric entropy result (Theorem 4.1), the breakup of the likelihood ratio into the sets S^M and $(S^M)^c$, the choice of $v(\cdot)$ function in (4.4) and the moment control in Lemma 4.2. Zhang [63] proved special cases of these ingredients for $d = 1$ and $S = [-R, R]$ for some R while our argument applies to every S. As remarked previously, it is crucial to allow S to be arbitrary for obtaining adaptation results to discrete mixtures.

4.2. *Proof overview of Theorem 3.5.* A complete proof of Theorem 3.5 is given in Section B.5 of the Supplementary Material [49]. This subsection gives an overview of the main ideas. Let us now introduce the following notation. Let $\mathbf{X}$ denote the $d \times n$ matrix whose columns are the observed data vectors $X_1, \ldots, X_n$. For a density $f \in \mathcal{M}$, let $T_f(\mathbf{X})$ denote the $d \times n$ matrix whose ith column is given by the $d \times 1$ vector:

$$X_i + \frac{\nabla f(X_i)}{f(X_i)} \quad \text{for } i = 1, \ldots, n.$$

With this notation, we can clearly rewrite $\mathfrak{R}_n(\hat{\theta}, \breve{\theta}^*)$ as

$$\mathfrak{R}_n\big(\hat{\theta}, \breve{\theta}^*\big) = \mathbb{E}\bigg(\frac{1}{n}\big\|T_{\hat{f}_n}(\mathbf{X}) - T_{f_{\bar{G}_n^0}}(\mathbf{X})\big\|_F^2\bigg),$$

where $\|\cdot\|_F$ denotes the usual Frobenius norm for matrices.

Now for $f \in \mathcal{M}$ and $\rho > 0$, let $T_f(\mathbf{X}, \rho)$ be the $d \times n$ matrix whose ith column is given by the $d \times 1$ vector:

$$X_i + \frac{\nabla f(X_i)}{\max(f(X_i), \rho)} \quad \text{for } i = 1, \ldots, n.$$

The first important observation is that for $\rho_n := (2\pi)^{-d/2}/n$, we have $T_{\hat{f}_n}(\mathbf{X}, \rho_n) = T_{\hat{f}_n}(\mathbf{X})$ and this follows from classical results about the NPMLE. This allows us to write

$$\begin{aligned}
\mathfrak{R}_n\big(\hat{\theta}, \breve{\theta}^*\big) &= \mathbb{E}\bigg(\frac{1}{n}\big\|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X})\big\|_F^2\bigg) \\
&\le 2\mathbb{E}\bigg(\frac{1}{n}\big\|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X}, \rho)\big\|_F^2\bigg) \\
&\quad + 2\mathbb{E}\bigg(\frac{1}{n}\big\|T_{f_{\bar{G}_n^0}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X})\big\|_F^2\bigg).
\end{aligned}$$

Using the following lemma (proved in Section F of the Supplementary Material [49]), the second term above is bounded by $(\sqrt{\log n})^{\max(d-2,0)}\epsilon_n^2(M, S, \bar{G}_n^0)$.

LEMMA 4.3. *Fix a probability measure G on $\mathbb{R}^d$ and let $0 < \rho \le (2\pi)^{-d/2}/\sqrt{e}$. Let $L(\rho) := \sqrt{-\log((2\pi)^d\rho^2)}$. Then there exists a positive constant C_d such that for every compact set $S \subseteq \mathbb{R}^d$, we have*

$$\begin{aligned}
\Delta(G, \rho) &:= \int\bigg(1 - \frac{f_G}{\max(f_G, \rho)}\bigg)^2 \frac{\|\nabla f_G\|^2}{f_G} \\
&\le C_d N\bigg(\frac{4}{L(\rho)}, S\bigg)L^d(\rho)\rho + dG\big(S^c\big).
\end{aligned} \tag{4.7}$$

We thus focus attention on the first term in the above bound for $\mathfrak{R}_n(\hat{\theta}, \breve{\theta}^*)$:

$$A(\hat{f}_n) := \mathbb{E}\bigg(\frac{1}{n}\big\|T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X}, \rho)\big\|_F^2\bigg).$$

Now if $\hat{f}_n$ were nonrandom, the above term can be bounded from above by a generalization (to $d \ge 1$) of Jiang and Zhang ([25], Theorem)3, which bounds $A(f)$ in terms of $\mathfrak{H}^2(f, f_{\bar{G}_n^0})$ for nonrandom $f \in \mathcal{M}$. We have stated this general result as Theorem E.1 and proved it in Section E of the Supplementary Material [49]. Of course, this result cannot be directly

used here because $\hat{f}_n$ is random. However, Theorem 2.1 implies that $\hat{f}_n$ belongs with high probability (specifically with probability at least $1-(2/n)$) to the set

$$E_n := \{f \in \mathcal{M} : \mathfrak{H}(f, f_{\bar{G}_n^0}) \le C_d \epsilon_n(M, S, \bar{G}_n^0)\},$$

where C_d is the constant obtained from Theorem 2.1. The idea therefore is to cover the space E_n to within η by deterministic densities $f_{G_1}, \ldots, f_{G_N}$. For this covering, we use the metric

$$\sup_{x:\mathfrak{d}_S(x)\le M} \left\| \frac{\nabla f(x)}{\max(f(x), \rho_n)} - \frac{\nabla g(x)}{\max(g(x), \rho_n)} \right\|.$$

Covering numbers in this metric are given in Corollary D.1 in the Supplementary Material [49] and this result is derived as a corollary of our main covering number result in Theorem 4.1. With these deterministic densities, we bound $A(\hat{f}_n)$ via $A(\hat{f}_n) \le 4\sum_{i=1}^n (\mathbb{E}\zeta_{in}^2/n)$ where

$$\begin{aligned}
\zeta_{1n} &:= \left\| T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X}, \rho_n) \right\|_F I\{\hat{f}_n \notin E_n\},\\
\zeta_{2n} &:= \Big(\left\| T_{\hat{f}_n}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X}, \rho_n) \right\|_F \\
&\qquad - \max_{1\le j\le N} \left\| T_{f_{G_j}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X}, \rho_n) \right\|_F \Big)_+ I\{\hat{f}_n \in E_n\},\\
\zeta_{3n} &:= \max_{1\le j\le N} \big(\left\| T_{f_{G_j}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X}, \rho_n) \right\|_F \\
&\qquad - \mathbb{E}\left\| T_{f_{G_j}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X}, \rho_n) \right\|_F \big)_+,\\
\zeta_{4n} &:= \max_{1\le j\le N} \mathbb{E}\left\| T_{f_{G_j}}(\mathbf{X}, \rho_n) - T_{f_{\bar{G}_n^0}}(\mathbf{X}, \rho_n) \right\|_F.
\end{aligned}$$

Each of these terms is controlled to finish the proof of Theorem 3.5 in the following way:

1. $\mathbb{E}\zeta_{1n}^2/n$ is bounded by $C_d\epsilon_n^2(M, S, \bar{G}_n^0)$ because $\mathbb{P}\{\hat{f}_n \notin E_n\} \le (2/n)$ and the fact that $T_f(\mathbf{X}, \rho_n)$ can be bounded by a term involving ρ_n alone (this result is stated (and proved) as Lemma F.1 in [49]).
2. $\mathbb{E}\zeta_{2n}^2/n$ is bounded by $C_d\epsilon_n^2(M, S, \bar{G}_n^0)$ using the fact that $f_{G_1}, \ldots, f_{G_N}$ form a covering of E_n.
3. $\mathbb{E}\zeta_{3n}^2/n$ is bounded by $C_d\sigma_{\max}^2\epsilon_n^2(M, S, \bar{G}_n^0)(\log n)^2$ using measure concentration properties of Gaussian random variables and an upper bound on N which is given by the covering number result in Theorem D.1.
4. $\mathbb{E}\zeta_{4n}^2/n$ is bounded by $C_d\epsilon_n^2(M, S, \bar{G}_n^0)(\log n)^3$ by Theorem E.1 (in [49]) which bounds $A(f)$ in terms of $\mathfrak{H}(f, f_{\bar{G}_n^0})$ for every nonrandom f.

The structure of the proof and the main ideas are very similar to that of Jiang and Zhang [25], proof of Theorem 5. Other than the fact that our arguments hold for $d \ge 2$ and arbitrary compact sets S, additional differences between our proof and [25], proof of Theorem 5, are as follows. The breakdown of the risk $\mathfrak{R}_n(\hat{\theta}, \breve{\theta}^*)$ into various terms is different as the authors of [25] work with the discrepancy measure (3.1) while we work directly with the discrepancy between $\hat{\theta}$ and $\breve{\theta}^*$. Our argument for $T_{\hat{f}_n}(\mathbf{X}) = T_{\hat{f}_n}(\mathbf{X}, \rho_n)$ (given in inequality (B.9) near the beginning of the proof of Theorem 3.5) is more direct compared to the corresponding argument in [25], Proposition 2. Our measure concentration result (see Lemma F.3) involves $X_i \sim N(\theta_i, \Sigma_i)$ and not Gaussian random vectors with identity covariance as in [25], Proposition 4. Our control of $\mathbb{E}\zeta_{5n}^2/n$ (via Lemma 4.3) is different from and probably more direct compared to the corresponding argument in [25], Theorem 3(ii).

5. Implementation details and some simulation results. In this section, we shall discuss some computational details concerning the NPMLE and also provide numerical evidence for the effectiveness of the estimator (1.7) based on the NPMLE for denoising.

For the optimization problem (1.2), it can be shown that $\hat{f}_n$ exists and is nonunique. However $\hat{f}_n(X_1), \ldots, \hat{f}_n(X_n)$ are unique and they solve the finite dimensional optimization problem

$$\begin{aligned} &\operatorname{argmax} \sum_{i=1}^{n} \log f_i \\ &\text{s.t. } (f_1, \ldots, f_n) \in \operatorname{Conv}\big\{\big(\phi(X_1 - \theta), \ldots, \phi(X_n - \theta)\big) : \theta \in \mathbb{R}^d\big\}, \end{aligned} \tag{5.1}$$

where Conv above stands for convex hull. The constraint set in the above problem, however, involves every $\theta \in \mathbb{R}^d$. A natural way of computing an approximate solution is to fix a finite data-driven set $F := \{a_1, \ldots, a_m\} \subseteq \mathbb{R}^d$ and restrict the infinite convex hull to the convex hull over θ belonging to this set. This leads to the problem

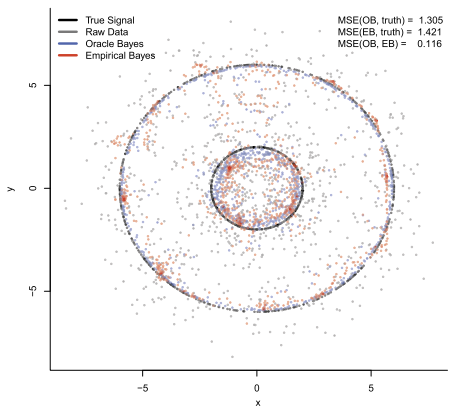
$$\begin{aligned} &\operatorname{argmax} \sum_{i=1}^{n} \log f_i \\ &\text{s.t. } (f_1, \ldots, f_n) \in \operatorname{Conv}\big\{\big(\phi_d(X_1 - \theta), \ldots, \phi_d(X_n - \theta)\big) : \theta \in F\big\}. \end{aligned} \tag{5.2}$$

This can also be seen as an approximation to (1.2) where the densities $f \in \mathcal{M}$ are restricted to have atoms in $\{a_1, \ldots, a_m\} \subseteq \mathbb{R}^d$. (5.2) is a convex optimization problem over the probability simplex in m dimensions and can be solved using many algorithms (e.g., standard interior point methods as implemented in the software, Mosek, can be used here).

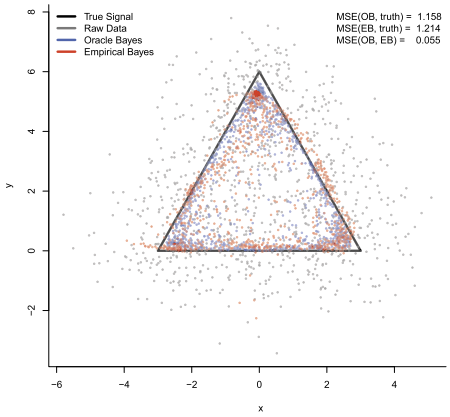
The effectiveness of (5.2) as an approximation to (1.2) depends crucially on the choice of $\{a_1, \ldots, a_m\}$. For $d = 1$, Koenker and Mizera [28] propose the use of a uniform grid within the range $[\min_{1 \le i \le n} X_i, \max_{1 \le i \le n} X_i]$ of the data. Dicker and Zhao [16] discuss this approach in more detail and recommend the choice $m := [\sqrt{n}]$. They also prove (see [16], Theorem 2) that the resulting approximate MLE, $\tilde{f}_n$, has a squared Hellinger accuracy, $\mathfrak{H}^2(\tilde{f}_n, f_0)$, of $O_p((\log n)^2/n)$ when the mixing measure corresponding to f_0 has bounded support. For $d \ge 1$, Feng and Dicker [22] recommend taking a regular grid in a compact region containing the data. They also mention that empirical results seem "fairly insensitive" to the choice of m.

A proposal for selecting $\{a_1, \ldots, a_m\}$ that is different from gridding is the so called "exemplar" choice where one takes $m = n$ and $a_i = X_i$ for $i = 1, \ldots, n$. This choice is proposed in Böhning [7] for $d = 1$ and in Lashkari and Golland [30] for $d \ge 1$. This avoids gridding which can be problematic in multiple dimensions. Also, this method is computationally feasible as long as n is moderate (up to a few thousands) but becomes expensive for larger n. In such instances, a reasonable strategy is to take $a_1, \ldots, a_m$ as a random subsample of the data $X_1, \ldots, X_n$. For fast implementations, one can also extend the idea of Koenker and Mizera [28] by binning the observations and weighting the likelihood terms in (1.2) by relative multinomial bin counts.

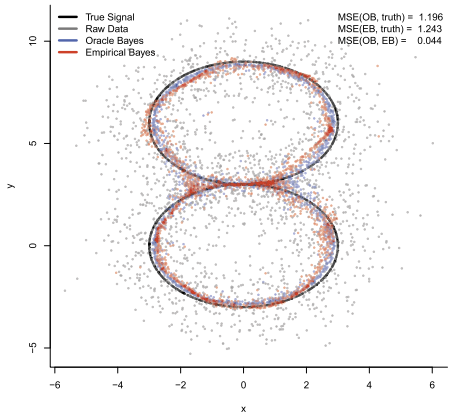
We shall now provide some graphical evidence of the effectiveness of the NPMLE for denoising. For our plots, the NPMLE is approximately computed via the algorithm (5.2) where $a_1, \ldots, a_m$ are chosen to be the data points $X_1, \ldots, X_n$ with $m = n$ (i.e., we follow the exemplar recommendation of [7] and [30]). We use the software, Mosek, to solve (5.2). The results of this paper do not apply directly to these approximate NPMLEs and extending them is the subject of future work. We argue, however, via simulations that these approximate NPMLEs work well for denoising.



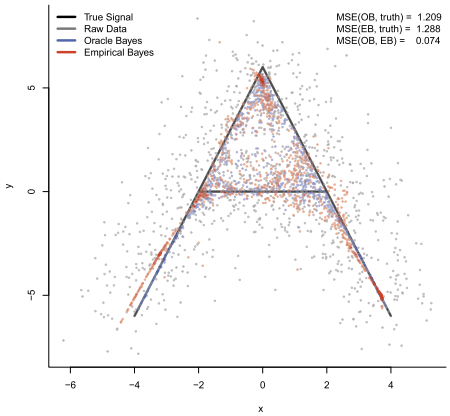
(a) *Two circles*: $n = 1000$. Half of $\{\theta_i\}_{i=1}^n$ are drawn uniformly at random from each of the concentric circles of radii 2 and 6, respectively.



(b) *Triangle*: $n = 999$. A third of $\{\theta_i\}_{i=1}^n$ are drawn uniformly at random from each edge of the triangle with vertices $(-3, 0)$, $(0, 6)$ and $(3, 0)$



(c) *Digit* 8: $n = 1000$. Half of $\{\theta_i\}_{i=1}^n$ are drawn uniformly at random from each of the circles of radii 3 centered at $(0, 0)$ and $(0, 6)$, respectively.



(d) *Letter* A: $n = 1000$. A fifth of $\{\theta_i\}_{i=1}^n$ are drawn uniformly at random from each of the line segments joining the points $(-4, -6)$, $(-2, 0)$, $(0, 6)$, $(2, 0)$ and $(4, 6)$ so as to form the letter A.

FIG. 1. *Illustrations of denoising using the empirical Bayes estimates* (1.7).

In Figure 1, we illustrate the performance of $\hat{\theta}_1, \ldots, \hat{\theta}_n$ (defined as in (1.7)) for denoising when the true vectors $\theta_1, \ldots, \theta_n$ take values in a bounded region of $\mathbb{R}^2$. The plots refer to these estimates as the empirical Bayes estimates and the quantities (1.6) as the oracle Bayes estimates.

In each of the four subfigures in Figure 1, we generate n vectors $\theta_1, \ldots, \theta_n$ from a bounded region in $\mathbb{R}^d$ for $d = 2$: they are generated from two concentric circles in the first subfigure, a triangle in the second subfigure, the digit 8 in the third subfigure and the uppercase letter A in the last subfigure. In each of these cases, the empirical measure $\bar{G}_n$ is supported on a bounded region so that Corollary 3.2 yields the near parametric rate $1/n$ up to logarithmic multiplicative factors in n for every NPMLE. In each of the subfigures in Figure 1, we plot the true parameter values $\theta_1, \ldots, \theta_n$ in black, the data $X_1, \ldots, X_n$ (generated independently according to $X_i \sim N(\theta_i, I_2)$) are plotted in gray, the oracle Bayes estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ are plotted

in blue while the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_n$ are plotted in red. The mean squared discrepancies

$$\frac{1}{n}\sum_{i=1}^{n}\|\hat{\theta}_i^* - \theta_i\|^2, \qquad \frac{1}{n}\sum_{i=1}^{n}\|\hat{\theta}_i - \theta_i\|^2 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\|\hat{\theta}_i^* - \hat{\theta}_i\|^2$$

are given in each figure in the legend at the upper-right corner. Note that the third MSE is much smaller than the other two in each subfigure.

As can be observed from Figure 1, the empirical Bayes estimates (1.7) approximate their targets (1.6) quite well. The most noteworthy fact is that the estimates (1.7) do not require any knowlege of the underlying structure in $\bar{G}_n$, for instance, concentric circles, or triangle or a letter of the alphabet etc. We should also note here that the noise distribution here is completely known to be $N(0, I_d)$ which implies, in particular, that there is no unknown scale parameter representing the noise variance.

We have also done numerical simulations for illustrating the denoising performance of $\hat{\theta}_1, \ldots, \hat{\theta}_n$ in the case when $\theta_1, \ldots, \theta_n$ have a clustering structure. Due to space constraints, these results have been moved to Section G of the Supplementary Material [49].

Acknowledgments. The second author was supported by NSF CAREER Grant DMS-16-54589.

SUPPLEMENTARY MATERIAL

Supplement to "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising" (DOI: 10.1214/19-AOS1817SUPP; .pdf). This supplementary material contains proofs of all results in the main paper as well as some observations on the heteroscedastic Gaussian denoising problem.

REFERENCES

[1] ACHARYA, J., DIAKONIKOLAS, I., LI, J. and SCHMIDT, L. (2017). Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms* 1278–1289. SIAM, Philadelphia, PA. MR3627812 https://doi.org/10.1137/1.9781611974782.83

[2] BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

[3] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028 https://doi.org/10.1007/s004400050210

[4] BHASKARA, A., SURESH, A. and ZADIMOGHADDAM, M. (2015). Sparse solutions to nonnegative linear systems and applications. In *Artificial Intelligence and Statistics* 83–92.

[5] BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam* 55–87. Springer, New York. MR1462939

[6] BÖHNING, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *J. Statist. Plann. Inference* **47** 5–28. MR1360956 https://doi.org/10.1016/0378-3758(94)00119-G

[7] BÖHNING, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications*: *Meta-Analysis*, *Disease Mapping and Others*. *Monographs on Statistics and Applied Probability* **81**. CRC Press/CRC, Boca Raton, FL. MR1684363

[8] BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Stat.* **42** 855–903. MR0286209 https://doi.org/10.1214/aoms/1177693318

[9] BROWN, L. D. and GREENSHTEIN, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1685–1704. MR2533468 https://doi.org/10.1214/08-AOS630

[10] CHAN, S.-O., DIAKONIKOLAS, I., SERVEDIO, R. A. and SUN, X. (2012). Learning mixtures of structured distributions over discrete domains. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms* 1380–1394. SIAM, Philadelphia, PA. MR3202986

[11] CHAN, S.-O., DIAKONIKOLAS, I., SERVEDIO, R. A. and SUN, X. (2014). Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing* 604–613. ACM, New York.

[12] CHEN, G. K., CHI, E. C., RANOLA, J. M. O. and LANGE, K. (2015). Convex clustering: An attractive alternative to hierarchical clustering. *PLoS Comput. Biol.* **11** e1004228.

[13] DASKALAKIS, C. and KAMATH, G. (2014). Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Conference on Learning Theory* 1183–1213.

[14] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

[15] DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation*. *Springer Series in Statistics*. Springer, New York. MR1843146 https://doi.org/10.1007/978-1-4613-0125-7

[16] DICKER, L. H. and ZHAO, S. D. (2016). High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference. *Biometrika* **103** 21–34. MR3465819 https://doi.org/10.1093/biomet/asv067

[17] DONOHO, D. and REEVES, G. (2013). Achieving Bayes mmse performance in the sparse signal+ Gaussian white noise model when the noise level is unknown. In *Information Theory Proceedings* (*ISIT*), 2013 *IEEE International Symposium on* 101–105. IEEE, New York.

[18] EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860 https://doi.org/10.1198/jasa.2011.tm11181

[19] EFRON, B. and MORRIS, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59** 335–347. MR0334386 https://doi.org/10.1093/biomet/59.2.335

[20] EFRON, B. and MORRIS, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4** 22–32. MR0394960

[21] EVERITT, B. S. and HAND, D. J. (1981). *Finite Mixture Distributions*: *Monographs on Applied Probability and Statistics*. CRC Press, London. MR0624267

[22] FENG, L. and DICKER, L. H. (2016). Nonparametric maximum likelihood inference for mixture models via convex optimization. Preprint. Available at arXiv:1606.02011.

[23] GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329 https://doi.org/10.1214/aos/1013203453

[24] HOCKING, T. D., JOULIN, A., BACH, F. and VERT, J.-P. (2013). Clusterpath an algorithm for clustering using convex fusion penalties. In *28th International Conference on Machine Learning*, pp. 1.

[25] JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467 https://doi.org/10.1214/08-AOS638

[26] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135 https://doi.org/10.1214/009053604000000030

[27] KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906. MR0086464 https://doi.org/10.1214/aoms/1177728066

[28] KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742 https://doi.org/10.1080/01621459.2013.869224

[29] LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. MR0521328

[30] LASHKARI, D. and GOLLAND, P. (2008). Convex clustering with exemplar-based models. In *Advances in Neural Information Processing Systems* 825–832.

[31] LI, J. and SCHMIDT, L. (2015). A nearly optimal and agnostic algorithm for properly learning a mixture of k gaussians, for any constant k. Preprint. Available at arXiv:1506.01367.

[32] LINDSAY, B. G. (1983). The geometry of mixture likelihoods: A general theory. *Ann. Statist.* **11** 86–94. MR0684866 https://doi.org/10.1214/aos/1176346059

[33] LINDSAY, B. G. (1983). The geometry of mixture likelihoods. II. The exponential family. *Ann. Statist.* **11** 783–792. MR0707929 https://doi.org/10.1214/aos/1176346245

[34] LINDSAY, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics* i–163. IMS, Hayward, CA.

[35] LINDSAY, B. G. and LESPERANCE, M. L. (1995). A review of semiparametric mixture models. *J. Statist. Plann. Inference* **47** 29–39. MR1360957 https://doi.org/10.1016/0378-3758(94)00120-K

[36] LINDSTEN, F., OHLSSON, H. and LJUNG, L. (2011). *Just Relax and Come Clustering!*: *A Convexification of K-Means Clustering*. Linköping Univ. Electronic Press.

[37] MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math*. **1896**. Springer, Berlin. MR2319879

[38] MAUGIS, C. and MICHEL, B. (2011). Data-driven penalty calibration: A case study for Gaussian mixture model selection. *ESAIM Probab. Stat*. **15** 320–339. MR2870518 https://doi.org/10.1051/ps/2010002

[39] MAUGIS, C. and MICHEL, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat*. **15** 41–68. MR2870505 https://doi.org/10.1051/ps/2009004

[40] MAUGIS-RABUSSEAU, C. and MICHEL, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM Probab. Stat*. **17** 698–724. MR3126158 https://doi.org/10.1051/ps/2012018

[41] MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

[42] MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1417721

[43] MOSEK (2015). The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28), 17.

[44] RADCHENKO, P. and MUKHERJEE, G. (2014). Consistent clustering using an ℓ_1 fusion penalty. Preprint. Available at arXiv:1412.0753.

[45] ROBBINS, H. (1950). A generalization of the method of maximum likelihood-estimating a mixing distribution. *Ann. Math. Stat*. **21** 314–315.

[46] ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 131–148. Univ. California Press, Berkeley and Los Angeles. MR0044803

[47] ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 157–163. Univ. California Press, Berkeley and Los Angeles. MR0084919

[48] ROBBINS, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Stat*. **35** 1–20. MR0163407 https://doi.org/10.1214/aoms/1177703729

[49] SAHA, S. and GUNTUBOYINA, A. (2020). Supplement to "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising." https://doi.org/10.1214/19-AOS1817SUPP.

[50] SCHLATTMANN, P. (2009). *Medical Applications of Finite Mixture Models*. Springer, Berlin.

[51] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist*. **9** 1135–1151. MR0630098

[52] SURESH, A. T., ORLITSKY, A., ACHARYA, J. and JAFARPOUR, A. (2014). Near-optimal-sample estimators for spherical Gaussian mixtures. In *Advances in Neural Information Processing Systems* 1395–1403.

[53] TAN, K. M. and WITTEN, D. (2015). Statistical properties of convex clustering. *Electron. J. Stat*. **9** 2324–2347. MR3411231 https://doi.org/10.1214/15-EJS1074

[54] TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **63** 411–423. MR1841503 https://doi.org/10.1111/1467-9868.00293

[55] TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, Chichester. MR0838090

[56] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[57] WANG, B., ZHANG, Y., SUN, W. W. and FANG, Y. (2018). Sparse convex clustering. *J. Comput. Graph. Statist*. **27** 393–403. MR3816274 https://doi.org/10.1080/10618600.2017.1377081

[58] WATANABE, M. and YAMAGUCHI, K. (2003). *The EM Algorithm and Related Statistical Models*. CRC Press, Boca Raton, FL.

[59] WEINSTEIN, A., MA, Z., BROWN, L. D. and ZHANG, C.-H. (2018). Group-linear empirical Bayes estimates for a heteroscedastic normal mean. *J. Amer. Statist. Assoc*. **113** 698–710. MR3832220 https://doi.org/10.1080/01621459.2017.1280406

[60] WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist*. **23** 339–362. MR1332570 https://doi.org/10.1214/aos/1176324524

[61] WU, C., KWON, S., SHEN, X. and PAN, W. (2016). A new algorithm and theory for penalized regression-based clustering. *J. Mach. Learn. Res*. **17** Paper No. 188, 25. MR3567456

[62] XIE, X., KOU, S. C. and BROWN, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *J. Amer. Statist. Assoc*. **107** 1465–1479. MR3036408 https://doi.org/10.1080/01621459.2012.728154

[63] ZHANG, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statist. Sinica* **19** 1297–1318. MR2536157

[64] ZHU, C., XU, H., LENG, C. and YAN, S. (2014). Convex optimization procedure for clustering: Theoretical revisit. In *Advances in Neural Information Processing Systems* 1619–1627.