Abstract

Detecting Latent Signal in High Dimensions: Estimation in Gaussian Mixtures and Nonparametric Models

Natalie C. Doss

2020

Detecting hidden signal lying in high-dimensional data is a central objective in statistics. In this thesis, we explore fundamental limits and algorithms for estimating signal in large datasets. This work consists of two main parts.

In the first part (Chapters 2-3), we provide theoretical analyses of Gaussian mixture models in high dimensions. Throughout these sections, we observe signal-plus-noise data where the noise is Gaussian in $\mathbb{R}^d$ and the signal is distributed according to a mixing distribution parametrized by a collection of low-dimensional subspaces of $\mathbb{R}^d$. The mixture signal we wish to learn has been corrupted by high-dimensional noise. We study clustering, parameter estimation, and density estimation under various assumptions on the mixing distribution, and we obtain convergence rates and fast algorithms for estimation.

In the second part, (Chapters 4-5), we study two tasks: individual component analysis when the data are a linearly transformed graphical model, and variational inference when we wish to allow the posterior to have a flexible form. Both tasks require the estimation of a nonparametric density in high dimensions, and for this we turn to an algorithm called score matching that allows for the computationally feasible estimation of a high-dimensional distribution.

Detecting Latent Signal in High Dimensions: Estimation in Gaussian Mixtures and Nonparametric Models

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Natalie C. Doss

Dissertation Directors:
Harrison H. Zhou
Yihong Wu
John Lafferty

December, 2020

Contents

List of Figures

Acknowledgments

I am deeply grateful to my committee, Harrison Zhou, Yihong Wu, and John Lafferty for their incredible support during my studies. I am privileged to have had the opportunity to learn from them during my graduate education, and I truly appreciate the incredible guidance they provided. I thank all the faculty of the Department of Statistics and Data Science at Yale for the outstanding education I received here. Their openness and willingness to discuss problems with students have made this a special experience. I especially thank David Pollard, with whom I had the good fortune to both take classes and pursue independent study early in my graduate career. I am truly grateful for the many discussions with him that deepened my understanding of empirical process theory. I thank Jessi Cisewski for her support on my practical project and helpful conversations beyond that work. I am grateful to the staff of the Yale Center for Research Computing, particularly to IRobert Bjornson, for his support on the computational component of my thesis. I thank Jay Emerson, Susan Wang, Derek Feng, and Cynthia Rush for their unending support and advice during my graduate career. I thank my colleagues Pengkun Yang and Anderson Y. Zhang; it was a privilege to work with them and learn from them. Thank you to Joann DelVecchio, JoAnn Falato, Dawn Hemstock, Karen Kavanaugh, and Elizavette Torres, who were so kind and willing to help me with whatever I needed during my studies, and who made the department feel like home. I thank all my graduate classmates, who commiserated with me when graduate school was tough and made

the good experiences even better. I thank my family. Without their love and support, this would not have been possible.

Chapter 1

Introduction

In this introductory section, we provide an overview of each chapter. We then provide notation that will be used throughout the thesis. We introduce the model that is common to Part 1 of the thesis in Section 1.2 and discuss the algorithm that is common to Part 2 of the thesis in Section 1.3.

2. We explore fundamental limits of estimation in a high-dimensional Gaussian mixture on k means when no lower bound assumptions are placed on the separation or collinearity of the means or weights of the mixing distribution. In this context, clustering is impossible, but mixing distribution and density estimation are possible, and we provide the optimal statistical convergence rates for both tasks, as well as a polynomial time algorithm for mixing distribution estimation. We also provide an extensions in the direction of a density estimation algorithm, maximum likelihood estimation, and estimation in the related problem of estimation in a high-dimensional Gaussian mixture on k subspaces.

3. We study fundamental limits of clustering in high-dimensional Gaussian mixtures on k manifolds, where the manifolds are assumed to be well separated. We explore lower bounds on the amount of separation that is necessary for optimal error rates to be achievable. In particular, we explore the performance of

spectral clustering and its limitations on signal-plus-noise data, and we provide a novel algorithm for the refinement of spectral clustering.

4. We introduce graphical component analysis, a technique for estimating latent signal where linearly transformed signal components are exhibit dependence that can be modeled by a graphical model. We couple score matching with orthogonal matrix optimization to perform a novel type of high-dimensional density estimation.

5. We propose nonparametric variational inference via score matching, a novel type of variational inference that uses a kernel-density based form of score matching to allow for the estimation of a flexible, generative form for the posterior. Our proposed method allows for nonparametric posterior inference while avoiding the optimization of a dual representation of an objective that is so prevalent in the nonparametric density and posterior estimation literature.

1.1 Notation

Let $[n] \triangleq \{1, \ldots, n\}$. Let S^{d-1} and Δ^{d-1} denote the unit sphere and the probability simplex in $\mathbb{R}^d$, respectively. Let $S^{d\times r}$ be the set of matrices in $\mathbb{R}^{d\times r}$ whose columns form an orthonormal basis of some r-dimensional subspace of $\mathbb{R}^d$, i.e., Let $S^{d\times r} = \{V \in \mathbb{R}^{d\times r} : V^\top V = I_r\}$. Because we will work with subspaces of $\mathbb{R}^d$ of dimension less than or equal to r, we will sometimes use $S^{d\times r}$ to indicate matrices V which may have some columns equal to zero, i.e., $V^\top V = diag(I_r, 0, \ldots 0)$.

Let e_j be the vector with a 1 in the jth coordinate and zeros elsewhere. For a matrix A, let $\|A\|_2 = \sup_{x:\|x\|_2=1} \|Ax\|_2$ and $\|A\|_F^2 = \mathrm{tr}(A^\top A)$. For two positive sequences $\{a_n\}, \{b_n\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if there exists a constant C such that $a_n \leq Cb_n$ and we write $a_n \lesssim_k b_n$ and $a_n = O_k(b_n)$ to emphasize that C may depend on a parameter k. For functions $\mathcal{L}(\theta), f(\theta)$, we write $\mathcal{L}(\theta) \overset{c}{=} f(\theta)$ to

indicate that these functions are equal up to terms containing the relevant argument θ.

For $\epsilon > 0$, an ϵ-covering of a set A with respect to a metric ρ is a set $\mathcal{N}$ such that for all $a \in A$, there exists $b \in \mathcal{N}$ such that $\rho(a, b) \leq \epsilon$; denote by $N(\epsilon, A, \rho)$ the minimum cardinality of ϵ-covering sets of A. An ϵ-packing in A with respect to the metric ρ is a set $\mathcal{M} \subset A$ such that $\rho(a, b) > \epsilon$ for any distinct a, b in $\mathcal{M}$; denote by $M(\epsilon, A, \rho)$ the largest cardinality of ϵ-packing sets in A.

For distributions P and Q, let p and q denote their relative densities with respect to some dominating measure μ, respectively. The total variation distance is defined as $\mathsf{TV}(P, Q) \triangleq \frac{1}{2} \int |p(x) - q(x)| \mu(dx)$. If $P \ll Q$, the KL divergence and the χ^2-divergence are defined as $\mathsf{KL}\left(P||Q\right) \triangleq \int p(x) \log \frac{p(x)}{q(x)} \mu(dx)$ and $\chi^2\left(P\|Q\right) \triangleq \int \frac{(p(x)-q(x))^2}{q(x)} \mu(dx)$, respectively. Let $\operatorname{supp}(P)$ denote the support set of a distribution P. Let $\mathcal{L}(U)$ denote the distribution of a random variable U. For a one-dimensional distribution ν, denote the rth moment of ν by $m_r(\nu) \triangleq \mathbb{E}_{U \sim \nu}[U^r]$. Given a d-dimensional distribution Γ, for each $\theta \in \mathbb{R}^d$, we denote

$$\Gamma_\theta \triangleq \mathcal{L}(\theta^\top U), \quad U \sim \Gamma; \tag{1.1.1}$$

in other words, Γ_θ is the pushforward of Γ by the projection $u \mapsto \theta^\top u$; in particular, the ith marginal of Γ is denoted by $\Gamma_i \triangleq \Gamma_{e_i}$, with e_i being the ith coordinate vector. Similarly, for $V \in \mathbb{R}^{d \times k}$, denote

$$\Gamma_V \triangleq \mathcal{L}(V^\top U), \quad U \sim \Gamma. \tag{1.1.2}$$

For a function $f : \mathbb{R} \to \mathbb{R}$, we write $f'(x)$ to mean the derivative of the function f with respect to the argument x. For a function $f : \mathbb{R}^d \to \mathbb{R}$, let $\nabla f \in \mathbb{R}^d = (\partial f/\partial x_1, \dots \partial f/\partial x_d)^\top$ be the gradient of f, $\nabla^2 f$ be the Hessian, and $\triangle f = \sum_{i \leq d} \partial^2 f / \partial x_i^2 = \operatorname{tr}(\nabla^2 f)$.

1.2 Gaussian mixtures

The signal in a high-dimensional dataset is often generated from different sources. For instance, a genetic dataset may contain information from people with different health conditions, or a medical video dataset may contain videos of people performing different exercises. To model such data, we turn to *mixture models*, which posit that the data are generated from a collection of sub-populations, each governed by a different distribution. In chapters 2-3 of this thesis, we will observe data of the form:

$$\begin{gathered} X_i \overset{\text{i.i.d.}}{\sim} P_\Gamma, \text{ for } i = 1, \ldots, n, \text{ where} \\ P_\Gamma = \Gamma * N(0, \sigma^2 I_d), \text{ and } \Gamma = \sum_{j=1}^{k} w_j \Gamma_j. \end{gathered} \tag{1.2.1}$$

In (1.2.1), $*$ denotes the convolution. The latent distribution Γ is the *mixing distribution*. The weights $w_1, \ldots, w_k$ satisfy $w_j \geq 0$ and $\sum_{j=1}^{k} w_j = 1$. For simplicity, we assume $\sigma^2 = 1$. In this work, Γ will be a distribution that is parametrized by k subspaces or manifolds. The distribution Γ may be parametrized by zero-dimensional affine spaces (as in k-means, or the classical Gaussian location mixture model), linear subspaces (as in subspace clustering), or manifolds. We study these scenarios in Chapters 2 and 3.

In this work, crucially, the dimension d will be potentially as large as the sample size n. Data generated according to the model (1.2.1) have the following signal-plus-noise structure:

$$\begin{gathered} X_i = U_i + Z_i, \text{ where} \\ Z_i \sim_{i.i.d.} N(0, \sigma^2 I_d) \text{ and } U_i \sim_{i.i.d.} \Gamma. \end{gathered} \tag{1.2.2}$$

Let ϕ_d denote the standard normal density in d dimensions. Then the density of P_Γ is given by

$$p_\Gamma(x) = \int \sigma^{-d} \phi_d \left((x - u) / \sigma \right) \, d\Gamma(u). \tag{1.2.3}$$

Any signal-plus-noise model of the form (1.2.1) where Γ lies on some low-dimensional subspace of $\mathbb{R}^d$ really has two components: a Gaussian distribution on the irrelevant noise variables, and a Gaussian mixture distribution on the relevant variables. Let Γ be a distribution on a subspace of $\mathbb{R}^d$ of rank at r, and let $V = [v_1, \ldots, v_r]$ be the matrix whose columns for an orthonormal basis for this subspace. Let $V_C = [v_{r+1}, \ldots, v_d]$ be the matrix whose columns are an orthonormal basis of the complement, so $[V, V_C]$ is an othornormal matrix. We have

$$p_\Gamma(x) = \phi_{d-r}(V_C^\top x) p_{\Gamma_V}(V^\top x). \tag{1.2.4}$$

This simple fact will be crucial to the results in Chapters 2-3.

Learning in finite mixture models takes three forms: clustering, parameter estimation, and density estimation. All three problems fall under the guise of unsupervised learning. In well-separated, standard mixture models such as the Gaussian location mixture model, all three aspects of the problem are well understood. What is less understood are optimal rates of convergence and algorithms when the model may be less ideal but more realistic. Chapters 2-3 provide important advances in this direction.

We will see in Chapter 2 that the structure of Gaussian mixtures as in (1.2.4) leads to a parameter estimation rate that contains two components: a subspace estimation component, and a low-dimensional mixing distribution estimation component. In Chapter 3, we will see that dimension reduction techniques aimed at removing the irrelevant noise in (1.2.4) lead to a novel testing algorithm that improves on standard clustering methods to obtain an exponential error rate.

1.3 Score matching

In this thesis, we also explore latent signal detection in contexts that require the estimation of a high-dimensional density. The study of high-dimensional density estimation has a long history in statistics, and the problem is, in general, intractable since it usually requires the computation of a normalizing constant in high dimensions. To circumvent this issue, we will rely on a powerful tool known as score matching, which provides a way to estimate a high-dimensional density without computing the normalizing constant. We now introduce the score matching technique and state a few crucial lemmas about its utility; proofs of these lemmas are in the appendix.

Suppose $x \in \mathbb{R}^d$ is a random variable distributed according to a distribution P with density p. Let q be some estimate of p. Score matching seeks a q that minimizes the Fisher divergence between p and q:

$$\mathcal{L}(q) = \mathbb{E}_p \left\| \nabla \log p - \nabla \log q \right\|_2^2 . \tag{1.3.1}$$

In (5.1.1), the derivatives are taken with respect to the argument x. Clearly, the normalizing constant in q disappears in the objective (1.3.1), but this objective does not appear to be something we can optimize in q since it explicitly involves the true density p. However, it turns out that (1.3.1) simplifies to an objective of the form $\mathbb{E}_p O_q$ where O_q is an operator depending only on the density q. The simplification is due simply to integration by parts and was first noted by [Hyvarinen, 2005]; we state this here for clarity.

Lemma 1.3.1 (The score matching objective can be optimized in q)**.** *Let* $\hat{q} = \operatorname{argmin}_q \mathcal{L}(q)$ *with* $\mathcal{L}(q)$ *as in (5.1.1). Then*

$$\hat{q} = \operatorname*{argmin}_q \int p(x) \left(\frac{1}{2} \left\| \nabla \log q \right\|_2^2 + \triangle \log q \right) \, dx.$$

The objective (1.3.1) simplifies further in the case where q is assumed to be an exponential family. In fact, in this case, the solution for q has a closed form; it requires only the computation of a simple quadratic [Hyvarinen, 2005], as we now state.

Lemma 1.3.2 (Score matching for exponential families.)**.** *Let* $q(x) = \frac{e^{g(x)}}{\int e^{g(x)} dx}$*, and suppose* $g : \mathbb{R}^d \to \mathbb{R}$*, with* $g(x) = \gamma^\top \phi(x)$ *where* $\phi : \mathbb{R}^d \to \mathbb{R}^K$ *and* $\gamma \in \mathbb{R}^K$*. Let* $\mathcal{L}(q)$ *be as in (5.1.1). Let* $\hat{\gamma} = \left(\mathbb{E}_{p(x)} A(x)\right)^{-1} \mathbb{E}_{p(x)} k(x)$*, where*

$$k(x) = \sum_{i \leq d} \frac{\partial^2 \phi(x)}{\partial x_i^2} \in \mathbb{R}^K.$$
$$A(x) = \sum_{i \leq d} \frac{\partial \phi(x)}{\partial x_i} \frac{\partial \phi(x)}{\partial x_i}^\top \in \mathbb{R}^{K \times K}.$$

Then

$$\operatorname*{argmin}_q \mathcal{L}(q) \propto \exp\left(\hat{\gamma}^\top \phi(x)\right). \tag{1.3.2}$$

Lemma 1.3.2 provides the optimal estimator $\operatorname{argmin}_q \mathcal{L}(q)$ for the population score matching loss $\mathcal{L}(q)$; in practice, an empirical version of this loss is used, with the expectations in $\hat{\gamma}$ taken over a sample from P. Because of the simplicity of the solution and utility for estimating densities while avoiding estimation of the normalizing constant, score matching in the form Lemma 1.3.2 has been used in the estimation of an exponential family density when data come from a graphical model [Lin et al., 2016], [Janofsky, 2015], or when the sufficient statistic function lies in a Reproducing Kernel Hilbert Space [Sriperumbudur et al., 2017]. A rate of convergence depending on the smoothness of the density class was obtained in [Sriperumbudur et al., 2017]. This work in fact considers a penalized version of (1.3.1), imposing an ℓ_2 penalty that also leads to a quadratic solution when q belongs to an exponential family.

Notice that the solution in Lemma 1.3.2 requires the computation of $A(x)$, a

Hessian matrix. This computation can be expensive, as noted in [Dai et al., 2018]. Score matching in the form of (1.3.1) is not the only possibility for optimizing a nonparametric density while avoiding computation of the normalizing constant. The works [Vincent, 2011] and [Saremi et al., 2018] propose a type of score matching that avoids the Hessian calculation required in classical score matching. We now discuss this method.

Suppose we observe data x distributed according to a distribution P with density p. Suppose we also observe a noisy form of x, ξ, satisfying $\xi = x + z$, where $z \sim N(0, \sigma^2 I_d)$. Recall that in score matching, we optimize in θ the score matching divergence $\int p(x) \left\| \nabla_x \log p(x) - \nabla \log_x q_\theta(x) \right\|_2^2 dx$. [Vincent, 2011] propose that instead of using the integration-by-parts trick to obtain an optimizeable empirical version of this loss, we instead use an explicity representation of $p(x)$; in fact, a kernel density estimate. Specifically, let us observe data points $x_1, \ldots, x_n \in \mathbb{R}^d$. Let

$$p(\xi) = \frac{1}{n} \sum_{i=1}^{n} K(\xi | x_i),$$

where K is a kernel function. [Vincent, 2011] then uses the kernel-density score matching loss:

$$\mathcal{L}_{kde}(\theta) = \mathbb{E}_{p(\xi)} \left\| \nabla_\xi \log p(\xi) - \nabla_\xi \log q_\theta(\xi) \right\|_2^2 . \tag{1.3.3}$$

It turns out that this can be represented in a simple way that makes it easy to optimize in practice, as we now show.

Lemma 1.3.3. *Let $\mathcal{L}_{kde}(\theta)$ be as defined in (1.3.3). Then optimizing $\mathcal{L}_{kde}(\theta)$ in θ is equivalent to optimizing the loss:*

$$\mathcal{L}_{kde}(\theta) \stackrel{c}{=} \mathbb{E}_{(\xi,x)\sim J} \left\| \nabla_\xi \log K(\xi | x_i) - \nabla_\xi \log q_\theta(\xi) \right\|_2^2 ,$$

where J is the joint distribution with density $\frac{1}{n} \sum_{i=1}^{n} K(\xi | x_i)$, i.e., it is the empirical

expectation (over the empirical density of x) of the conditional density $K(\xi|x)$.

[Saremi et al., 2018] use Lemma 1.3.3 to obtain a simple score matching objective that avoids Hessian computations and is moreover amenable to optimization when the desired form for q_θ is a neural network. Suppose that $K(\xi|x) = \phi_\sigma(\xi - x)$, where ϕ_σ is the Gaussian kernel with bandwidth σ^2, i.e., $K(\xi|x) = \exp\left(-\|\xi - x\|_2^2/2\sigma^2\right)$. Then $\nabla_\xi \log K(\xi|x) = \frac{x-\xi}{\sigma^2}$, and using Lemma 1.3.3, the loss (1.3.3) becomes

$$\mathcal{L}_{kde}(\theta) = \mathbb{E}_{(\xi,x)\sim J} \left\|x - \xi - \sigma^2 \nabla_\xi \log q_\theta(\xi)\right\|_2^2.$$

To approximate this loss, for each observed data point x_i, we generate independent Gaussian samples $\xi_{i1}, \ldots, \xi_{im} \sim_{i.i.d.} N(0, I_d)$. We then use the empirical loss:

$$\mathcal{L}(\theta) = \sum_{i\in[n], j\in[m]} \left\|x_i - \xi_{i,j} - \sigma^2 \nabla_\xi \log q_\theta(\xi_{ij})\right\|_2^2.$$

This loss is optimizeable in θ as long as $\log q_\theta$ is differentiable. [Saremi et al., 2018] let $q_\theta(x) = e^{f_\theta(x)}$ where f_θ is a multilayer perceptron. They optimize for the weights and biases via stochastic gradient descent.

In Chapters 4-5, we will expand on the above-mentioned works to develop two novel algorithms that use score matching in previously unconsidered contexts: graphical component analysis and variational inference. We will show how score matching, both in its classical form as in Lemma 1.3.2, as well as in the kernel-density form Lemma 1.3.3, can be used to estimate high-dimensional densities in these settings.

Chapter 2

Mixtures of means: parameter and density estimation

2.1 Introduction

We now consider the problems of parameter and density estimation in (1.2.1) when Γ is a distribution on k atoms in $\mathbb{R}^d$. This is known as the Gaussian location mixture model; we shall also refer to it as the Gaussian mixture on k means. This is one of the most widely studied mixture models because of its simplicity and wide applicability, but rates of convergence for both parameter and density estimation in this model are not well understood when the mixing distribution centers are both high dimensional and allowed to be arbitrarily close to each other. Consider the k-component Gaussian location mixture model (which we shall refer to as k-GM) in d dimensions:

$$\begin{gathered} X_i \overset{\text{i.i.d.}}{\sim} P_\Gamma, \text{ for } i = 1, \ldots, n, \text{ where} \\ P_\Gamma = \Gamma * N(0, \sigma^2 I_d), \text{ and } \Gamma = \sum_{j=1}^{k} w_j \delta_{\mu_j}. \end{gathered} \tag{2.1.1}$$

In (2.1.1), δ_x denotes the Dirac delta measure at $x \in \mathbb{R}^d$. The unknown parameter is Γ, a discrete distribution on k atoms $\mu_1, \ldots, \mu_k$, each in $\mathbb{R}^d$.

In (2.1.1), a $n^{-1/2}$ convergence rate for parameter estimation is obtained when $d = 1$, $\mu_1, \ldots, \mu_k$ are fixed and well separated, and k is known. A more realistic setting is one in which the mixing components may not be well separated. For instance, some of them may be equal or very close together, so that the model is effectively mis-identified. In a pioneering work in this area, [Chen, 1995] considered k-GM under such weak assumptions. In particular, he showed that when $\Gamma = \frac{1}{3}\delta_{2h} + \frac{2}{3}\delta_{-h}$, the rate for estimating h is $n^{-1/4}$ when the truth is allowed to be in a neighborhood of radius $n^{-1/4}$ around zero.

For general k-GM in one dimension, where k is upper bounded but the mixing distribution atoms may be arbitrarily close, the minimax convergence rate is $n^{-1/(4k-2)}$ when the error is measured in the 1-Wasserstein distance (defined later in this section) on the mixing distribution ([Heinrich and Kahn, 2018], [Wu and Yang, 2019]). When the true number of mixing components k_0 is unknown but upper bounded by some k, a convergence rate of $n^{-1/(4(k-k_0)+2)}$ is obtained for general one dimensional mixtures, including Gaussian mixtures; see [Heinrich and Kahn, 2018] and [Wu and Yang, 2019].

We consider rates of convergence for Gaussian location mixtures when the atoms of Γ may arbitrarily overlap, the weights may be arbitrarily small, and crucially, the dimension d may be as large as the sample size n. This extends [Wu and Yang, 2019] to a high-dimensional setting. As discussed above, it is known that in such a scenario, when d is fixed, the minimax rate for parameter estimation is much slower than $n^{-1/2}$, and the rate for density estimation is parametric. We seek to understand how this extends for arbitrary d.

Several recent works have explored multivariate Gaussian mixture estimation, both in high and low dimensions. For statistical rates, [Ho and Nguyen, 2016, Theorem 1.1] obtained convergence rates for mixing distribution estimation in Wasser-

stein distances for low-dimensional multivariate location-scale Gaussian mixtures, both over- and exact-fitted. Their rates for over-fitted mixtures are determined by algebraic dependencies among a set of polynomial equations whose order depends on the level of overfitting; the rates are potentially much slower than $n^{-1/2}$. The estimator analyzed in [Ho and Nguyen, 2016] is the maximum-likelihood estimator (MLE), which involves non-convex optimization and is typically approximated by the Expectation-Maximization (EM) algorithm.

In the computer science literature, a long line of research starting with [Dasgupta, 1999] has developed fast algorithms for individual parameter estimation in multivariate Gaussian mixtures under fairly weak separation conditions, see, e.g., [Arora and Kannan, 2005, Belkin and Sinha, 2009, Kalai et al., 2010, Moitra and Valiant, 2010, Hsu and Kakade, 2013, Hardt and Price, 2015, Hopkins and Li, 2018]. Since these works focus on individual parameter estimation, some separation assumption on the mixing distribution is necessary.

We now give more background on the problem of estimation in (2.1.1). Denote the probability simplex on k elements by $\Delta_k = \{w = (w_1, \ldots, w_k) : \sum_{j=1}^k w_j = 1, w_j \geq 0\}$. We will let $\mathcal{G}_{k,d}$ be the class of k-atomic distributions supported on a ball of radius R in d dimensions, where, throughout the paper, R is assumed to be an absolute constant. Let $\|\cdot\|_2$ denote the standard Euclidean norm in $\mathbb{R}^d$. Formally,

$$\mathcal{G}_{k,d} \triangleq \left\{ \Gamma = \sum_{j=1}^k w_j \delta_{\mu_j} : \mu_j \in \mathbb{R}^d, \|\mu_j\|_2 \leq R, w = (w_1, \ldots, w_k) \in \Delta_k \right\}.$$

Let

$$P_\Gamma = \Gamma * N(0, I_d) = \sum_{j=1}^k w_j N(\mu_j, I_d)$$

and $\mathcal{P}_{k,d} = \{P_\Gamma : \Gamma \in \mathcal{G}_{k,d}\}$ denotes the collection of k-Gaussian mixtures (k-GMs) whose centers lie in a ball of radius R. Let ϕ_d denote the standard normal density in

d dimensions. Then the density of P_Γ is given by

$$p_\Gamma(x) = \sum_{j=1}^{k} w_j \sigma^{-d} \phi_d \left((x - \mu_j) / \sigma \right),$$

which is a parametric distribution with the $2k-1$ parameters $\mu_1, \ldots, \mu_k$ and $w_1, \ldots, w_k$. In many scenarios, it is possible to estimate the individual weights and location parameters with small error. Under our assumptions, we may be able to estimate some of the components or weights with large error but still have small error in the estimation of the entire mixing distribution. Therefore, it is natural for us to frame the model as in (2.1.1).

We first discuss the problem of parameter estimation in (2.1.1). We use a Wasserstein distance to measure the error on the mixing distribution. For $q \geq 1$, the Wasserstein-q distance (with respect to the Euclidean distance) is defined as

$$W_q(\Gamma, \Gamma') \triangleq \left(\inf_{\nu \in \mathcal{P}(\Gamma, \Gamma')} \mathbb{E}_\nu \left\| U - U' \right\|_2^q \right)^{\frac{1}{q}}, \tag{2.1.2}$$

where the infimum is taken over all couplings of Γ and Γ', i.e., joint distributions of U and U' with marginals Γ and Γ' respectively. The W_1 distance is a natural and commonly used loss function for mixture models and deconvolution problems. It is invariant under permutation of the distributions it compares, so it sidesteps the unidentifiability of mixing distributions. It allows for the development of a meaningful statistical theory in an assumption-free framework. Note that if Γ and Γ' have the same number of atoms and those atoms are well separated and the weights are bounded away from zero, estimation in W_1 distance translates to recovery of the individual parameters up to permutation; see Lemma 1 of [Wu and Yang, 2019]. In the widely studied case of the symmetric 2-GM in which the samples are drawn from $P_\Gamma = \frac{1}{2}N(\mu, I_d) + \frac{1}{2}N(-\mu, I_d)$ with mixing distrubution $\Gamma_\mu = \frac{1}{2}(\delta_{-\mu} + \delta_\mu)$, $W_1(\Gamma_\mu, \Gamma_{\mu'})$

coincides with the commonly used loss function $\min\{\|\mu - \mu'\|_2, \|\mu + \mu'\|_2\}$. [1]

Assume $k \geq 2$ in (2.1.1). A natural conjecture is that the d-dimensional analogue of the $n^{-1/(4k-2)}$-rate of [Wu and Yang, 2019] is $(d/n)^{1/(4k-2)}$. This conjecture is incorrect. The main result of this paper on the minimax rate for estimating Γ in setting (2.1.1) is the following theorem.

Theorem 2.1.1 (Estimating the mixing distribution)**.** *Let P_Γ be the k-GM defined in (2.1.1). Given n i.i.d. samples from P_Γ, the minimax risk of estimating Γ over the class $\mathcal{G}_{k,d}$ satisfies*

$$\inf_{\hat{\Gamma}} \sup_{\Gamma \in \mathcal{G}_{k,d}} \mathbb{E}_\Gamma W_1(\hat{\Gamma}, \Gamma) \asymp_k \left(\frac{d}{n}\right)^{1/4} \wedge 1 + \left(\frac{1}{n}\right)^{1/(4k-2)}, \tag{2.1.3}$$

where the notation $\asymp_k$ means that both sides agree up to constant factors depending only on k. Furthermore, if $n \geq d$, there exists an estimator $\hat{\Gamma}$, computable in $O(nd^2) + O_k(n^{5/4})$ time, and a positive constant C_k, such that for any $\Gamma \in \mathcal{G}_{k,d}$ and any $0 < \delta < \frac{1}{2}$, with probability at least $1 - \delta$,

$$W_1(\hat{\Gamma}, \Gamma) \leq C_k \left(\left(\frac{d}{n}\right)^{1/4} + \left(\frac{1}{n}\right)^{1/(4k-2)} \left(\log \frac{1}{\delta}\right)^{1/2} \right). \tag{2.1.4}$$

We now explain the intuition behind the rate. The location parameters $\mu_1, \ldots, \mu_k$ of Γ span a subspace of $\mathbb{R}^d$ dimension at most k. Let $\{v_1, \ldots, v_k\}$ be an orthonormal basis of this subspace, and let $V = [v_1, \ldots, v_k]$. For each $j \in [k]$, $\mu_j = V\psi_j$, where $\psi_j = V^\top \mu_j \in \mathbb{R}^k$ encodes the coefficients of μ_j in the basis vectors in V. Therefore, we can identify a k-atomic distribution Γ on $\mathbb{R}^d$ with a pair (V, γ), where $\gamma = \sum_{j \in [k]} w_j \delta_{\psi_j}$ is a k-atomic distribution on $\mathbb{R}^k$.

The error in estimating Γ consists of two parts: the high-dimensional component, which results from estimating the subspace on which the centers lie, and the low-

[1] This can be seen directly from the definition of W_1 from (2.1.2); simply take the minimum over all couplings and note that any coupling in this case is a distribution on at most four points. Then $W_1(\Gamma_\mu, \Gamma_{\mu'}) = \min_{0 \leq w \leq 1} w\|\mu - \mu'\|_2 + (1-w)\|\mu + \mu'\|_2$.

dimensional component, which results from estimating the k-atomic distribution in k dimensions. We shall see in Section 2.2 that the error from estimating the subspace is $(d/n)^{1/4}$, while the error from estimating the low-dimensional distribution is $n^{-1/(4k-2)}$. Estimating a high-dimensional Gaussian mixture balances a tradeoff in the error induced by the dimension d and by the number of components k. Indeed, we see that there is a threshold $d^* = n^{(2k-3)/(2k-1)}$ (e.g. $d^* = n^{1/3}$ for $k = 2$), such that for $d > d^*$, the subspace-estimation error, $(d/n)^{1/4}$, dominates the error; otherwise, the low-dimensional error, $(1/n)^{1/(4k-2)}$, does.

As discussed above, the slow rate of $n^{-1/(4k-2)}$ for the low dimensional part of the error can be explained by the weak assumptions placed on the model, making it more difficult to estimate than a standard parametric model. The $(d/n)^{1/4}$ in the subspace estimation component of the error in Theorem 2.1.1 may not be immediately intuitive. If the atoms are bounded away from zero, estimation should be doable at the $(d/n)^{1/2}$ rate, but if not, it is more difficult to estimate the subspace on which the atoms lie. It turns out that this results in the slower rate of $(d/n)^{1/4}$. For a more general statement, see Lemma 2.2.5.

In addition to a minimax rate of convergence, we seek an algorithm providing an estimator that achieves this rate. There are two broad categories of approaches to estimation for mixtures of Gaussians: method-of-moments and maximum likelihood. This paper first provides a polynomial time algorithm based on the former approach. It is derived from the denoised-method-of-moments (DMM) method of [Wu and Yang, 2019], which is an improved variant of the usual method of moments.

The case where d does not grow with n and the case where $d = 1$ are not fundamentally different in terms of the optimal rate of convergence. However, even in the case where d does not grow with n but is greater than one, algorithmic challenges immediately arise. The algorithm presented in this paper solves two distinct

problems presented by the high dimensional mixture: the problem of estimating the subspace (handling a d that may grow with n) and the problem of estimating the low-dimensional mixture which nonetheless is not univariate.

This work also demonstrates that density estimation for high dimensional Gaussian mixtures can be done at the sharp parametric rate. Given distributions P, Q, let p and q denote their relative densities with respect to some dominating measure μ, respectively. The squared Hellinger distance between P and Q is defined as $H^2(P,Q) \triangleq \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mu(dx)$. We now give a brief review of the recent literature on density estimation for Gaussian mixtures, then state our result.

The result in Theorem 2.1.2 below, which follows a long line of research on Gaussian location mixture model density estimation, is the first we know of that avoids logarithmic terms in the numerator. It is known that if $d = 1$ and the mixing distribution support is not finite, i.e., k or R may grow with n, then a $\log n$ term in the numerator of (2.1.5) cannot be avoided, see [Kim, 2014] and [Ibragimov, 2001]. But it is natural to think that in (2.1.1), the sharp parametric rate should be obtainable for density estimation since the mixing distribution support is finite and bounded. Results in this direction when d is fixed are obtained by [Wu and Yang, 2019], who obtain the sharp parametric rate when $d = 1$. In low dimensions, [Ho and Nguyen, 2016, Theorem 2.1] obtained an $O(\sqrt{\log n/n})$-Hellinger guarantee for the MLE Gaussian location-scale mixtures. For high-dimensional location mixtures, the best result till now was the total variation guarantee of $\tilde{O}(\sqrt{kd/n})$ for location mixtures, where $\tilde{O}$ hides polylogarithmic factors, from [Ashtiani et al., 2018].

The algorithm therein runs in time that is exponential in d. To our knowledge, there is no polynomial-time algorithm that achieves the sharp density estimation guarantee in Theorem 2.1.2 (or the slightly suboptimal rate in [Ashtiani et al., 2018]), even for constant k. The works of [Kalai et al., 2010, Moitra and Valiant, 2010] showed that their polynomial-time parameter estimation algorithms also provide density es-

timators without separation conditions, but the resulting rates of convergence are far from optimal. [Feldman et al., 2006, Acharya et al., 2014, Li and Schmidt, 2017] provided polynomial-time algorithms for density estimation with improved statistical performance. In particular, [Acharya et al., 2014] obtained an algorithm that runs in time $\tilde{O}(n^2 d + (d/n)^{k^2})$ and achieves a total variation error of $\tilde{O}((d/n)^{1/4})$. The running time was further improved in [Li and Schmidt, 2017], which achieves the rate $\tilde{O}((d/n)^{1/6})$ for 2-GM.

The above-mentioned works all focus on finite mixtures, which is also the scenario considered in this paper. A related strain of research (e.g., [Genovese and Wasserman, 2000, Ghosal and van der Vaart, 2001, Zhang, 2009, Saha and Guntuboyina, 2017]) studies the so-called *nonparametric mixture model*, in which the mixing distribution Γ may be an arbitrary probability measure.

In this case, the nonparametric maximum likelihood estimator (known as the NPMLE) entails solving a convex (but infinite-dimensional) optimization problem, which, in principle, can be solved by discretization [Koenker and Mizera, 2014]. For statistical rates, it is known that in one dimension, the optimal L_2-rate for density estimation is $\Theta((\log n)^{1/4}/\sqrt{n})$ and the Hellinger rate is at least $\Omega(\sqrt{\log n/n})$ [Ibragimov, 2001, Kim, 2014], which shows that the parametric rate (2.1.5) is only achievable for finite mixture models. For the NPMLE, [Zhang, 2009] proved the Hellinger rate of $O(\log n/\sqrt{n})$ in one dimension; this was extended to the multivariate case by [Saha and Guntuboyina, 2017]. In particular, [Saha and Guntuboyina, 2017, Theorem 2.3] obtained a Hellinger rate of $C_d\sqrt{k(\log n)^{d+1}/n}$ for the NPMLE when the true model is a k-GM. In high dimensions, this is highly suboptimal compared to the parametric rate in (2.1.5), although the dependency on k is optimal.

Theorem 2.1.2 (Density estimation)**.** *Let P_Γ be as in (2.1.1). Then the minimax*

risk of estimating P_Γ *over the class* $\mathcal{P}_{k,d}$ *satisfies:*

$$\inf_{\hat{P}} \sup_{\Gamma \in \mathcal{G}_{k,d}} \mathbb{E}_\Gamma H(\hat{P}, P_\Gamma) \asymp_k \left(\frac{d}{n}\right)^{1/2} \wedge 1. \tag{2.1.5}$$

Furthermore, there exists a proper density estimate $P_{\hat{\Gamma}}$ *and a positive constant* C_k*, such that for any* $\Gamma \in \mathcal{G}_{k,d}$ *and any* $0 < \delta < \frac{1}{2}$*, with probability at least* $1 - \delta$*,*

$$H(P_{\hat{\Gamma}}, P_\Gamma) \leq C_k \left(\frac{d \log(1/\delta)}{n}\right)^{1/2}. \tag{2.1.6}$$

The sharp parametric rate for density estimation follows from local cover number calculations, i.e., to achieve (2.1.5), it is necessary to show that a covering number on a radius-δ ball in Hellinger distance on the densities has size no more than $(\delta/\epsilon)^{dC_k}$. To achieve this for a problem parametrized by some θ in a metric space Θ, it is necessary to show that $H(P_\theta, P_{\theta_*}) \asymp \rho(\theta, \theta_*)$, where ρ is the the relevant metric on the parameter space.

Previous work, e.g., [Ghosal and van der Vaart, 2001] and [Ho and Nguyen, 2016], considered cover numbers in the atoms and weights space, but this results in global, rather than local, cover numbers because closeness in Hellinger distance between two mixture densities of the form (2.1.1) does not imply closeness between the atoms and weights when the atoms and weights may be arbitrarily small; consider, for example, the distribution on $\mathbb{R}$: $\epsilon\delta_1 + (1-\epsilon)\delta_\epsilon$. That is, arguing in the locations and weights space results in upper bounds on the Hellinger distance but not matching lower bounds. For [Ghosal and van der Vaart, 2001], a global entropy could not be avoided since they are in essentially a nonparametric setting, but for our setting, sharper tools are available. In the proof of Theorem 2.1.2, we eschew analyzing cover numbers in the atoms and weights space in favor of covering numbers in the moment space, which turns out to be the most relevant parameter space.

It should be noted that the use of moment tensors in Section 2.3 is theoreti-

cal; we use the moment tensors as a tool to obtain a packing number bound on a small Hellinger ball. Moment tensors have also been used in algorithms for parameter estimation in Gaussian mixtures; in particular, see [Hsu and Kakade, 2013] and [Anandkumar et al., 2014]. Now [Hsu and Kakade, 2013, Theorem 3] translate their tensor-based algorithm into a finite-sample bound for the error in location estimation. Their rates require separation assumptions on the model Γ and are not optimal in general. As shown in [Diakonikolas et al., 2017, Question 1.1], a lower bound on computational time for density estimation is $poly(n, k)$, and this has not been achieved by an proper estimator that also achieves the correct rate of convergence.

Finally, we note that the rate in Theorem 2.1.2 was achieved for $d = 1$ in [Wu and Yang, 2019]. However, the proof therein relates the Hellinger distance to the moments distance, and the concentration of the moments in that work has a non-optimal dependence on k; in fact, it is exponential in k. The result in Theorem 2.1.2, when achieved for $d = 1$, and likewise the one-dimensional MLE result in Lemma 2.7.5, achieve optimal dependence on k.

2.2 Mixing distribution estimation

In this section we present the algorithm that achieves the optimal rate for estimating the mixing distribution in Theorem 2.1.1. The procedure is described in Sections 2.2.1 and 2.2.2. The proof of correctness is given in Sections 2.2.3 and 2.2.4.

2.2.1 Dimension reduction via PCA

Recall that $V \in S^{d\times k}$ is the matrix whose columns form an orthonormal basis for the subspace spanned by the atoms of Γ. An optimal procedure is to first project the data onto a subspace that is "close enough" to the space spanned by the v_j, then to

estimate the mixing distribution γ using the low-dimensional data. We would like to project the data onto the subspace spanned by the columns of V, but since we cannot access it, we use the noisy version. For simplicity, consider a sample of $2n$ observations $X_1, \dots, X_{2n} \overset{\text{i.i.d.}}{\sim} P_\Gamma$. We construct an estimator $\hat{\Gamma}$ of Γ in the following way:

(a) Estimate the subspace V using the first half of the sample. Given $\{X_1, \dots, X_n\}$, let

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top - I_d. \tag{2.2.1}$$

Let $\hat{V} = [\hat{v}_1, \dots, \hat{v}_k] \in \mathbb{R}^{d \times k}$ be the matrix whose columns are the top k orthonormal eigenvectors of $\hat{\Sigma}$.

(b) Project the second half of the sample onto the learned subspace $\hat{V}$:

$$x_i \triangleq \hat{V}^\top X_{i+n}, \quad i = 1, \dots, n. \tag{2.2.2}$$

Thanks to independence, conditioned on $\hat{V}$, $x_1, \dots, x_n$ are iid samples from a k-GM in k dimensions, with mixing distribution

$$\gamma \triangleq \Gamma_{\hat{V}} = \sum_{j=1}^{k} w_j \delta_{\hat{V}^\top \mu_j} \tag{2.2.3}$$

obtained by projecting the original d-dimensional mixing distribution Γ onto $\hat{V}$.

(c) To estimate $\hat{\gamma}$, we apply a multivariate version of the denoised method of moments to $x_1, \dots, x_n$ to obtain a k-atomic distribution on $\mathbb{R}^k$:

$$\hat{\gamma} = \sum_{j=1}^{k} \hat{w}_j \delta_{\hat{\psi}_j}. \tag{2.2.4}$$

This procedure is explained next and detailed in Algorithm 1.

(d) Lastly, we report

$$\hat{\Gamma} = \hat{\gamma}_{\hat{V}^\top} = \sum_{j=1}^{k} \hat{w}_j \delta_{\hat{V}\hat{\psi}_j} \tag{2.2.5}$$

as the final estimate of Γ.

Alternatively, we could achieve a slightly better dimension reduction by first centering the data by subtracting the sample mean, then projecting to a subspace of dimension $k-1$ rather than k. We would then add back the sample mean after obtaining the final estimator. To simplify the presentation, we forgo the centering step.

2.2.2 Estimating a mixing distribution in low dimensions

We now explain how we estimate a k-GM in k dimensions from i.i.d. observations. As mentioned in Section 4.1, the idea is to use many projections to reduce the problem to one dimension. We first present a conceptually simple estimator $\hat{\gamma}^\circ$ with an optimal statistical performance but unfavorable run time $n^{O(k)}$. We then describe an improved estimator $\hat{\gamma}$ that retains the statistical optimality and can be executed in time $n^{O(1)}$.

To make precise the reduction to one dimension, a relevant metric is the *sliced Wasserstein distance* [Rabin et al., 2011], which measures the distance of two d-dimensional distributions by the maximal W_1-distance of their one-dimensional projections:

$$W_1^{\mathsf{sliced}}(\Gamma, \Gamma') \triangleq \sup_{\theta \in S^{d-1}} W_1(\Gamma_\theta, \Gamma'_\theta). \tag{2.2.6}$$

Here we recall that Γ_θ defined in (1.1.1) denotes the projection, or pushforward, of Γ onto the direction θ. Computing the sliced Wasserstein distance can be difficult and in practice is handled by gradient descent heuristics [Rabin et al., 2011]; we will, however, only rely on its theoretical properties. The following result, which is proved in Section 2.2.4, shows that for low-dimensional distributions with few atoms, the full Wasserstein distance and the sliced one are comparable up to constant factors.

Lemma 2.2.1 (Sliced Wasserstein distance)**.** *For any k-atomic distributions Γ, Γ' on $\mathbb{R}^d$,*

$$W_1^{\mathsf{sliced}}(\Gamma, \Gamma') \le W_1(\Gamma, \Gamma') \le k^2\sqrt{d} \cdot W_1^{\mathsf{sliced}}(\Gamma, \Gamma').$$

Having obtained via PCA the reduced samples $x_1, \dots, x_n \sim \gamma * N(0, I_k)$ in (2.2.2), Lemma 2.2.1 suggests the following "meta-procedure": Suppose we have an algorithm (call it a 1-D algorithm) that estimates the mixing distribution based on n i.i.d. observations drawn from a k-GM in one dimension. Then

1. For each $\theta \in S^{k-1}$, since $\langle \theta, x_i \rangle \overset{\text{i.i.d.}}{\sim} \gamma_\theta * N(0,1)$, we can apply the 1-D algorithm to obtain an estimate $\widehat{\gamma_\theta} \in \mathcal{G}_{k,1}$;

2. We obtain an estimate of the multivariate distribution by minimizing a proxy of the sliced Wasserstein distance:

$$\hat{\gamma}^\circ = \operatorname*{argmin}_{\gamma' \in \mathcal{G}_{k,k}} \sup_{\theta \in S^{k-1}} W_1(\gamma'_\theta, \widehat{\gamma_\theta}). \tag{2.2.7}$$

Then by Lemma 2.2.1 (with $d = k$) and the optimality of $\hat{\gamma}^\circ$, we have

$$\begin{aligned} W_1(\hat{\gamma}^\circ, \gamma) &\lesssim_k \sup_{\theta \in S^{k-1}} W_1(\hat{\gamma}^\circ_\theta, \gamma_\theta) \\ &\le \sup_{\theta \in S^{k-1}} W_1(\widehat{\gamma_\theta}, \gamma_\theta) + \sup_{\theta \in S^{k-1}} W_1(\widehat{\gamma_\theta}, \hat{\gamma}^\circ_\theta) \\ &\le 2 \sup_{\theta \in S^{k-1}} W_1(\widehat{\gamma_\theta}, \gamma_\theta). \end{aligned} \tag{2.2.8}$$

Recall that the optimal W_1-rate for k-atomic one-dimensional mixing distribution is $O(n^{-\frac{1}{4k-2}})$. Suppose there is a 1-D algorithm that achieves the optimal rate *simultaneously* for all projections, in the sense that

$$\mathbb{E}\left[\sup_{\theta \in S^{k-1}} W_1(\widehat{\gamma_\theta}, \gamma_\theta)\right] \lesssim_k n^{-\frac{1}{4k-2}}. \tag{2.2.9}$$

This immediately implies the desired

$$\mathbb{E}[W_1(\hat{\gamma}^\circ, \gamma)] \lesssim_k n^{-\frac{1}{4k-2}}. \tag{2.2.10}$$

However, it is unclear how to solve the min-max problem in (2.2.7) where the feasible sets for γ and θ are both non-convex. The remaining tasks are two-fold: (a) provide a 1-D algorithm that achieves (2.2.9); (b) replace $\hat{\gamma}^\circ$ by a computationally feasible version.

Achieving (2.2.9) by denoised method of moments In principle, any estimator for a one-dimensional mixing distribution with exponential concentration can be used as a black box; this achieves (2.2.9) up to logarithmic factors by a standard covering and union bound argument. In order to attain the sharp rate in (2.2.9), we consider the Denoised Method of Moments (DMM) algorithm introduced in [Wu and Yang, 2019], which allows us to use the chaining technique to obtain a tight control of the fluctuation over the sphere.

The DMM method is an optimization-based approach that introduces a denoising step before solving the method of moments equations. For location mixtures, it provides an exact solver to the non-convex optimization problem arising in generalized method of moments [Hansen, 1982]. For Gaussian location mixtures with unit variance, the DMM algorithm proceeds as follows:

(a) Given $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \nu * N(0,1)$ for some k-atomic distribution ν supported on $[-R, R]$, we first estimate the moment vector $\mathbf{m}_{2k-1}(\nu) \triangleq (m_1(\nu), \ldots, m_{2k-1}(\nu))$ by their unique unbiased estimator $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_{2k-1})$, where $\tilde{m}_r = \frac{1}{n}\sum_{i=1}^n H_r(Y_i)$.Then $\mathbb{E}[\tilde{m}_r] = m_r(\nu)$ for all r. This step is common to all approaches based on the method of moments.

(b) In general the unbiased estimate $\tilde{m}$ is not a valid moment vector, in which case

the method-of-moment-equation lacks a meaningful solution. The key idea of the DMM method is to denoise $\tilde{m}$ by its projection onto the space of moments:

$$\hat{m} \triangleq \operatorname{argmin}\{\|\tilde{m} - m\| : m \in \mathcal{M}_r\}, \tag{2.2.11}$$

where the moment space

$$\mathcal{M}_r \triangleq \{\mathbf{m}_r(\pi) : \pi \text{ supported on } [-R, R]\} \tag{2.2.12}$$

consists of the first r moments of all probability measures on $[-R, R]$. The moment space is a convex set and characterized by positive semidefinite constraints (of the associated Hankel matrix); we refer the reader to the monograph [Shohat and Tamarkin, 1943] or [Wu and Yang, 2019, Sec. 2.1] for details. This means that the optimization problem (2.2.11) can be solved efficiently as a semidefinite program (SDP); see [Wu and Yang, 2019, Algorithm 1].

(c) Use Gauss quadrature to find the unique k-atomic distribution $\hat{\nu}$ such that $\mathbf{m}_{2k-1}(\hat{\nu}) = \hat{m}$. We denote the final output $\hat{\nu}$ by $\mathsf{DMM}(Y_1, \ldots, Y_n)$.

The following result shows the DMM estimator achieves the optimal rate in (2.2.9) simultaneously for all one-dimensional projections (for a single θ, this is shown in [Wu and Yang, 2019, Theorem 1]):

Lemma 2.2.2. *For each $\theta \in S^{k-1}$, let $\widehat{\gamma_\theta} = \mathsf{DMM}(\langle\theta, x_1\rangle, \ldots, \langle\theta, x_n\rangle)$ where $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \gamma * N(0, I_k)$ as in (2.2.2). There is a positive constant C_k such that, for any $\delta \in (0, \frac{1}{2})$, with probability at least $1 - \delta$,*

$$\max_{\theta \in S^{k-1}} W_1(\widehat{\gamma_\theta}, \gamma_\theta) \le C_k n^{-1/(4k-2)} \sqrt{\log \frac{1}{\delta}}.$$

Solving (2.2.7) efficiently using marginal estimates We first note that in order to achieve the optimal rate in (2.2.10), it is sufficient to consider any approximate minimizer of (2.2.7) up to an additive error of ϵ, as long as $\epsilon = O(n^{-\frac{1}{4k-2}})$. Therefore, to find an ϵ-optimizer, it suffices to maximize over θ in an ϵ-net (in ℓ_2) of the sphere, which has cardinality $(\frac{1}{\epsilon})^k = n^{O(1)}$, and, likewise, minimize γ over an ϵ-net (in W_1) of $\mathcal{G}_{k,k}$. The W_1-net can be constructed by combining an ϵ-net (in ℓ_2) for each of the k centers and an ϵ-net (in ℓ_1) for the weights, resulting in a set of cardinality $(\frac{1}{\epsilon})^{O(k^2)} = n^{O(k)}$. This naïve discretization scheme leads to an estimator of γ with optimal rate but time complexity $n^{O(k)}$. We next improve it to $n^{O(1)}$.

The key idea is to first estimate the marginals of γ, which narrows down its support set. It is clear that a k-atomic joint distribution is not determined by its marginal distributions, as shown by the example of $\frac{1}{2}\delta_{(-1,-1)} + \frac{1}{2}\delta_{(1,1)}$ and $\frac{1}{2}\delta_{(-1,1)} + \frac{1}{2}\delta_{(1,-1)}$, which have identical marginal distributions. Nevertheless, the support of the joint distribution must be a k-subset of the Cartesian product of the marginal support sets. This suggests that we can select the atoms from this Cartesian product and weights by fitting all one-dimensional projections, as in (2.2.7).

Specifically, for each $j \in [k]$, we estimate the jth marginal distribution of γ by $\widehat{\gamma}_j$, obtained by applying the DMM algorithm on the coordinate projections $\langle e_j, x_1\rangle, \ldots, \langle e_j, x_n\rangle$. Consider the Cartesian product of the support of each estimated marginal as the candidate set of atoms:

$$\mathcal{A} \triangleq \operatorname{supp}(\widehat{\gamma}_1) \times \cdots \times \operatorname{supp}(\widehat{\gamma}_k).$$

Throughout this section, let

$$\epsilon_{n,k} \triangleq n^{-\frac{1}{4k-2}},$$

and fix an $(\epsilon_{n,k}, \|\cdot\|_2)$-covering $\mathcal{N}$ for the unit sphere S^{k-1} and an $(\epsilon_{n,k}, \|\cdot\|_1)$-covering

$\mathcal{W}$ for the probability simplex Δ^{k-1}, such that[2]

$$\max\{|\mathcal{N}|, |\mathcal{W}|\} \lesssim \left(\frac{C}{\epsilon_{n,k}}\right)^{k-1}. \tag{2.2.13}$$

Define the following set of candidate k-atomic distributions on $\mathbb{R}^k$:

$$\mathcal{S} \triangleq \left\{ \sum_{j\in[k]} w_j \delta_{\psi_j} : (w_1, \ldots, w_k) \in \mathcal{W}, \psi_j \in \mathcal{A} \right\}. \tag{2.2.14}$$

Note that $\mathcal{S}$ is a random set which depends on the sample; furthermore, each $\psi_j \in \mathcal{A}$ has coordinates lying in $[-R, R]$ by virtue of the DMM algorithm.

The next lemma shows that with high probability there exists a good approximation of γ in the set $\mathcal{S}$.

Lemma 2.2.3. *Let $\mathcal{S}$ be given in (2.2.14). There is a positive constant C_k such that, for any $\delta \in (0, \frac{1}{2})$, with probability $1 - \delta$,*

$$\min_{\gamma' \in \mathcal{S}} W_1(\gamma', \gamma) \leq C_k n^{-1/(4k-2)} \sqrt{\log \frac{1}{\delta}}. \tag{2.2.15}$$

We conclude this subsection with Algorithm 1, which provides a full description of an estimator for k-atomic mixing distributions in k dimensions. The following result shows its optimality under the W_1 loss:

Lemma 2.2.4. *There is a positive constant C_k such that the following holds. Let $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \gamma * N(0, I_k)$ for some $\gamma \in G_{k,k}$. Then Algorithm 1 produces an estimator $\hat{\gamma} \in \mathcal{G}_{k,k}$ such that, for any $\delta \in (0, \frac{1}{2})$, with probability $1 - \delta$,*

$$W_1(\gamma, \hat{\gamma}) \leq C_k n^{-1/(4k-2)} \sqrt{\log \frac{1}{\delta}}. \tag{2.2.16}$$

[2]This is possible by, e.g., [Rudelson and Vershynin, 2009, Prop. 2.1] and [Ghosal and van der Vaart, 2001, Lemma A.4] for the sphere and probability simplex, respectively.

Algorithm 1: Parameter estimation for k-GM in k dimensions

Input: Dataset $\{x_i\}_{i\in[n]}$ with each point in $\mathbb{R}^k$, order k, radius R.

Output: Estimate $\hat{\gamma}$ of k-atomic distribution in k dimensions.

For $j = 1, \ldots, k$**:**

Compute the marginal estimate $\widehat{\gamma_j} = \mathsf{DMM}(\{e_j^\top x_i\}_{i\in[n]})$;

Form the set $\mathcal{S}$ of k-atomic candidate distributions on $\mathbb{R}^k$ as in (2.2.14) ;

For each $\theta \in \mathcal{N}$**:**

Estimate the projection by $\widehat{\gamma_\theta} = \mathsf{DMM}(\{\theta^\top x_i\}_{i\in[n]})$;

For each candidate distribution $\gamma' \in \mathcal{S}$ **and each direction** $\theta \in \mathcal{N}$**:**

Compute $W_1(\gamma'_\theta, \widehat{\gamma_\theta})$;

Report

$$\hat{\gamma} = \arg\min_{\gamma'\in\mathcal{S}} \max_{\theta\in\mathcal{N}} W_1(\gamma'_\theta, \widehat{\gamma_\theta}). \tag{2.2.17}$$

Remark 1. The total time complexity to compute the estimator (2.2.5) is $O(nd^2) + O_k(n^{5/4})$. Indeed, the time complexity of computing the sample covariance matrix is $O(nd^2)$, and the time complexity of performing the eigendecomposition is $O(d^3)$, which is dominated by $O(nd^2)$ since $d \le n$. By (2.2.13), both $\mathcal{W}$ and $\mathcal{N}$ have cardinality at most $(C/\epsilon_{n,k})^{k-1} = O_k(n^{1/4})$. Each one-dimensional DMM estimate takes $O_k(n)$ time to compute [Wu and Yang, 2019, Theorem 1]. Thus computing the one-dimensional estimator $\widehat{\gamma_\theta}$ for all $\theta = e_i$ and $\theta \in \mathcal{N}$ takes time $O_k(n^{5/4})$. Since both γ'_θ and $\widehat{\gamma_\theta}$ are k-atomic distributions by definition, their W_1 distance can be computed in $O_k(1)$ time. Finally, since $|\mathcal{A}| \le k^k$ and $|\mathcal{S}| \le |\mathcal{W}||\mathcal{A}| = O_k\left(n^{1/4}\right)$, searching over $\mathcal{S} \times \mathcal{N}$ therefore takes time at most $O_k(n^{1/4}) * O_k(n^{1/4}) = O_k\left(n^{1/2}\right)$. Therefore, the overall time complexity of Algorithm 1 is $O_k(n^{5/4})$.

2.2.3 Proof of the main theorem

The proof proceeds as follows. Recall that the estimate $\hat{\Gamma}$ in (2.2.5) is supported on the subspace spanned by the columns of $\hat{V}$, whose projection is $\hat{\gamma}$ in (2.2.4). Similarly, the projection of the ground truth Γ on the space $\hat{V}$ is denoted by $\gamma = \Gamma_{\hat{V}}$ in (2.2.3). Note that both γ and $\hat{\gamma}$ are k-atomic distributions in k dimensions. Let $\hat{H} = \hat{V}\hat{V}^\top$ be the projection matrix onto the space spanned by the columns of $\hat{V}$. By the triangle inequality,

$$\begin{aligned} W_1(\Gamma, \hat{\Gamma}) &\le W_1(\Gamma, \Gamma_{\hat{H}}) + W_1(\Gamma_{\hat{H}}, \hat{\Gamma}) \\ &\le W_1(\Gamma, \Gamma_{\hat{H}}) + W_1(\gamma, \hat{\gamma}). \end{aligned} \tag{2.2.18}$$

We will upper bound the first term by $(d/n)^{1/4}$ (using Lemmas 2.2.5 and 2.2.6 below) and the second term by $n^{-1/(4k-2)}$ (using the previous Lemma 2.2.4).

We first control the difference between Γ and its projection onto the estimated subspace $\hat{V}$. Since we do not impose any lower bound on $\|\mu_j\|_2$, we cannot directly show the accuracy of $\hat{V}$ by means of perturbation bounds such as the Davis-Kahan theorem [Davis and Kahan, 1970]. Instead, the following general lemma bounds the error by the difference of the covariance matrices.

Lemma 2.2.5. *Let $\Gamma = \sum_{j=1}^k w_j \delta_{\mu_j}$ be a k-atomic distribution. Let $\Sigma = \mathbb{E}_{U\sim\Gamma}[UU^\top] = \sum_{j=1}^k w_j \mu_j \mu_j^\top$ with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_d$. Let Σ' be a symmetric matrix and H'_r be the projection matrix onto the subspace spanned by the top r eigenvectors of Σ'. Then*

$$W_1^2(\Gamma, \Gamma_{H'_r}) \le k\left(\lambda_{r+1} + 2\|\Sigma - \Sigma'\|_2\right).$$

And if $r = k$, then $\lambda_{r+1}, \ldots, \lambda_d = 0$ since Σ has rank at most k, so

$$W_1^2(\Gamma, \Gamma_{H'_r}) \le 2k\|\Sigma - \Sigma'\|_2.$$

We will apply Lemma 2.2.5 with Σ' being the sample covariance matrix $\hat{\Sigma}$. The following lemma provides the concentration of $\hat{\Sigma}$ we need to prove the upper bound on the high-dimensional component of the error in Theorem 2.1.1.

Lemma 2.2.6. *Let $\Gamma \in \mathcal{G}_{k,d}$ and $\Sigma = \mathbb{E}_{U\sim\Gamma}[UU^\top]$. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top - I_d$, where $X_1, \dots, X_n \overset{i.i.d.}{\sim} P_\Gamma$. Then there exists a positive constant C such that, with probability at least $1-\delta$,*

$$\|\hat{\Sigma} - \Sigma\|_2 \le C\left(\sqrt{\frac{d}{n}} + k\sqrt{\frac{\log(k/\delta)}{n}} + \frac{\log(1/\delta)}{n}\right).$$

Proof of Theorem 2.1.1. We first show that the estimator (2.2.5) achieves the tail bound stated in (2.1.4), which, after integration, implies the average risk bound in (2.1.3). To bound the first term in (2.2.18), note that the rank of $\Sigma = \mathbb{E}_{U\sim\Gamma}[UU^\top]$ is at most k. Applying Lemmas 2.2.5 and 2.2.6 yields that, with probability $1-\delta$,

$$W_1(\Gamma, \Gamma_{\hat{H}}) \le \sqrt{2Ck}\left(\left(\frac{d}{n}\right)^{1/4} + \left(\frac{k^2\log(k/\delta)}{n}\right)^{1/4} + \sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{2.2.19}$$

where we used the fact that $W_1(\Gamma, \Gamma') \le W_2(\Gamma, \Gamma')$ by the Cauchy-Schwarz inequality. To upper bound the second term in (2.2.18), recall that $\hat{V}$ was obtained from $\{X_1, \dots, X_n\}$ and hence is independent of $\{X_{n+1}, \dots, X_{2n}\}$. Thus conditioned on $\hat{V}$,

$$x_i = \hat{V}^\top X_{i+n} \overset{\text{i.i.d.}}{\sim} \gamma * N(0, I_k), \quad i = 1, \dots, n.$$

Let $\hat{\gamma}$ be obtained from Algorithm 1 with input $x_1, \dots, x_n$. By Lemma 2.2.4, with probability $1-\delta$,

$$W_1(\gamma, \hat{\gamma}) \le C_k n^{-1/(4k-2)}\sqrt{\log\frac{1}{\delta}}. \tag{2.2.20}$$

Note that $(k^2\log(k/\delta)/n)^{1/4} + (\log(1/\delta)/n)^{1/2}$ in (2.2.19) is dominated by $\epsilon = 2C_k n^{-1/(4k-2)}\sqrt{\log\frac{1}{\delta}}$. The desired (2.1.4) follows from combining (2.2.18), (2.2.19), and (2.2.20).

Finally, the lower bound in (2.1.3) is obtained by combining the $\Omega((d/n)^{1/4} \wedge 1)$ lower bound in [Wu and Zhou, 2019, Theorem 10] for the special case of d-dimensional symmetric 2-GM and the $\Omega(n^{-1/(4k-2)})$ lower bound in [Wu and Yang, 2019, Proposition 7] for 1-dimensional k-GM. □

2.2.4 Proofs of supporting lemmas

In this subsection we prove Lemmas 2.2.1–2.2.6.

Proof of Lemma 2.2.1. For the lower bound, simply note that for any $\theta \in S^{d-1}$, $\|U - U'\|_2 \geq |\theta^\top U - \theta^\top U'|$ by the Cauchy-Schwartz inequality. Taking expectations on both sides with respect to the optimal W_1-coupling $\mathcal{L}(U, U')$ of Γ and Γ' yields the lower bound.

For the upper bound, we show that there exists $\theta \in S^{d-1}$ that satisfies the following properties:

1. The projection $y \mapsto \theta^\top y$ is injective on $\mathrm{supp}(\Gamma) \cup \mathrm{supp}(\Gamma')$;
2. For all $y \in \mathrm{supp}(\Gamma)$ and $y' \in \mathrm{supp}(\Gamma')$, we have

$$\|y - y'\|_2 \leq k^2 \sqrt{d} |\theta^\top y - \theta^\top y'|. \tag{2.2.21}$$

This can be done by a simple probabilistic argument. Let θ be drawn from the uniform distribution on S^{d-1}, which fulfills the first property with probability one. Next, for any fixed $x \in \mathbb{R}^d$, we have

$$\mathbb{P}\{|\theta^\top x| < t\|x\|_2\} \leq \frac{2\pi^{d/2}}{\Gamma(d/2)} \frac{\Gamma((d-1)/2)}{2\pi^{((d-1)/2}} \int_{-t}^{t} \left(1 - u^2\right)^{(d-3)/2} \, du < t\sqrt{d}.$$

Let $\mathcal{X} = \{y - y' : y \in \mathrm{supp}(\Gamma), y' \in \mathrm{supp}(\Gamma')\}$, whose cardinality is at most k^2. By a union bound,

$$\mathbb{P}\{\exists\, x \in \mathcal{X} \text{ s.t. } |\theta^\top x| < t \, \|x\|_2\} < k^2 t \sqrt{d},$$

and thus

$$\mathbb{P}\{|\theta^\top x| \geq t\|x\|_2,\ \forall\ x \in \mathcal{X}\} > 1 - k^2 t\sqrt{d}.$$

This probability is strictly positive for $t = 1/(k^2\sqrt{d})$. Thus, there exists $\theta \in S^{d-1}$ such that (2.2.21) holds. Since $\langle\theta, \cdot\rangle$ is injective on the support of Γ and Γ', denote its inverse by $g : \mathbb{R} \to \text{supp}(\Gamma) \cup \text{supp}(\Gamma')$. Then any coupling of the pushforward measures Γ_θ and Γ'_θ gives rise to a coupling of Γ and Γ' in the sense that if $\mathcal{L}(V, V')$ is a coupling of Γ_θ and $\Gamma_{\theta'}$ then $\mathcal{L}(g(V), g(V'))$ is a coupling of Γ and Γ'. By (2.2.21), we have

$$\|g(V) - g(V')\|_2 \leq k^2\sqrt{d}|V - V'|.$$

Taking expectations of both sides with respect to $\mathcal{L}(V, V')$ being the optimal W_1-coupling of Γ_θ and Γ'_θ yields the desired upper bound. □

Next we prove Lemma 2.2.2. Note that a simple union bound here would lead to a rate of $(\log n/n)^{1/(4k-2)}$. To remove the unnecessary logarithmic factors, we use the chaining technique (see the general result in Lemma 2.5.4), which entails proving the concentration of the increments of a certain empirical process.

Note that the sub-Gaussian type concentration in [Wu and Yang, 2019] was proved using a variance bound and then the so-called "median trick" and Chebychev. The median trick works as follows. We split the data into T samples, $\{x_t\}_{t\in[T]}$. We run the algorithm on each sample and take our estimator to be the median of these estimators; we use facts about the concentration of the binomial distribution to show concentration of the resulting median estimator. For the chaining technique, we need to have sub-Gaussian type concentration of the increments $|f_r(\theta_1, x) - f_r(\theta_2, x)|$. We can indeed apply the median trick to the increments $|f_r(\theta_1, x_t) - f_r(\theta_2, x_t)|$ and obtain sub-Gaussian type concentration of the increments. But $med_{t\in[T]}|f_r(\theta_1, x_t) - f_r(\theta_2, x_t)| \neq med(f_r(\theta_1, x_t)) - med(f_r(\theta_2, x_t))$. That is, obtaining sub-Gaussian type concentration of the median of the increments would not result in a statement about concentration

of the actual process $med_{t\in[T]} f_r(\theta, x_t)$ that we use in our estimation. We instead rely on some concentration results for polynomials from the literature.

Proof of Lemma 2.2.2. By the continuity of $\theta \mapsto W_1(\widehat{\gamma}_\theta, \gamma_\theta)$ and the monotone convergence theorem, it suffices to show that there exists a constant C_k such that, for any finite subset $\Theta \subset S^{k-1}$,

$$\mathbb{P}\left[\max_{\theta\in\Theta} W_1(\widehat{\gamma}_\theta, \gamma_\theta) \le C_k n^{-1/(4k-2)} \sqrt{\log \frac{1}{\delta}}\right] \ge 1-\delta. \tag{2.2.22}$$

Throughout the proof, C_k stands for a constant depending only on k whose value may vary from line to line.

Recall that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\gamma$, where $\gamma \in \mathcal{G}_{k,k}$. Define the empirical process

$$\tilde{m}_r(\theta) \triangleq \frac{1}{n}\sum_{i=1}^n H_r(\theta^\top X_i),$$

where H_r is the degree-r Hermite polynomial defined in (2.3.1). Define the centered random process indexed by θ:

$$f_r(\theta) \triangleq \sqrt{n}\left(\tilde{m}_r(\theta) - \mathbb{E}\tilde{m}_r(\theta)\right) = \sqrt{n}\left(\tilde{m}_r(\theta) - m_r(\gamma_\theta)\right). \tag{2.2.23}$$

Let $r \in [2k-1]$. By Lemma 2.5.3, there are positive constants C, c_k such that $\mathbb{P}\{|f_r(\theta_1) - f_r(\theta_2)| \ge \|\theta_1 - \theta_2\|_2 \lambda\} \le C\exp\left(-c_k\lambda^{2/r}\right)$. So we can apply Lemma 2.5.4 (with $\Theta \subseteq S^{k-1}$, $\rho(\theta_1,\theta_2) = \|\theta_1-\theta_2\|_2$, $\epsilon_0 = 2$, and $\alpha = 2/r$). Note that the maximal ϵ-packing of Θ has size $M(\epsilon, S^{k-1}, \|\cdot\|_2) \le (4/\epsilon)^k$. Fix $\theta_0 \in \Theta$. Then, by Lemma 2.5.4, with probability $1 - C\exp(-c_k t^{2/r})$,

$$\max_{\theta\in\Theta} |f_r(\theta) - f_r(\theta_0)| \le C_k\left(t + \int_0^1 (\log(1/u))^{r/2} du\right) = C_k\left(t + \Gamma\left(1+\frac{r}{2}\right)\right).$$

By (2.5.2) in Lemma 2.5.3, $|f_r(\theta_0)| \le C_k t$ with probability $1 - C\exp(-c_k t^{2/r})$. There-

fore, with probability $1 - \frac{\delta}{2k-1}$,

$$\max_{\theta\in\Theta} |\tilde{m}_r(\theta) - m_r(\gamma_\theta)| = \frac{1}{\sqrt{n}} \max_{\theta\in\Theta} |f_r(\theta)| \le C_k \frac{(\log(C_k/\delta))^{\frac{2k-1}{2}}}{\sqrt{n}}.$$

We take a union bound over $r \in [2k-1]$ and obtain that, with probability $1-\delta$,

$$\max_{\theta\in\Theta, r\in[2k-1]} |\tilde{m}_r(\theta) - m_r(\gamma_\theta)| \le C_k \frac{(\log(C_k/\delta))^{\frac{2k-1}{2}}}{\sqrt{n}}. \tag{2.2.24}$$

Recall that for each θ, the DMM estimator results in a k-atomic distribution $\widehat{\gamma}_\theta$, such that $m_r(\widehat{\gamma}_\theta) = \hat{m}_r(\theta)$ for all $r = 1, \dots, 2k-1$, where $(\hat{m}_1(\theta), \dots, \hat{m}_{2k-1}(\theta))$ is the Euclidean projection of $\tilde{m}(\theta) = (\tilde{m}_1(\theta), \dots, \tilde{m}_{2k-1}(\theta))$ onto the moment space $\mathcal{M}_{2k-1}$ (see (2.2.12)). Thus,

$$\max_{r\in[2k-1]} |m_r(\widehat{\gamma}_\theta) - m_r(\gamma_\theta)| \le 2\sqrt{2k-1} \max_{r\in[2k-1]} |\tilde{m}_r(\theta) - m_r(\gamma_\theta)|.$$

By the moment comparison inequality in Lemma 2.5.1, we have

$$W_1(\widehat{\gamma}_\theta, \gamma_\theta) \lesssim_k \max_{r\in[2k-1]} |\tilde{m}_r(\theta) - m_r(\gamma_\theta)|^{1/(2k-1)}.$$

Finally, maximizing both sides over $\theta \in \Theta$ and applying (2.2.24) yields the desired (2.2.22). □

Proof of Lemma 2.2.3. By Lemma 2.2.2, there is a positive constant C_k such that, for any $\delta \in (0, \frac{1}{2})$, with probability $1-\delta$,

$$W_1(\widehat{\gamma}_i, \gamma_i) \le \epsilon \triangleq C_k n^{-1/(4k-2)} \sqrt{\log \frac{1}{\delta}}, \quad \forall\, i \in [k].$$

Let $\gamma = \sum_{j=1}^k w_j \delta_{\mu_j}$. Fix $j \in [k]$. For any $i \in [k]$, by definition of W_1 distance in

(2.1.2),

$$w_j \cdot \min_{x \in \mathrm{supp}(\widehat{\gamma}_i)} |x - e_i^\top \mu_j| \le W_1(\widehat{\gamma}_i, \gamma_i).$$

Thus there exists $\mu_{ji} \in \mathrm{supp}(\widehat{\gamma}_i)$ such that

$$w_j |\mu_{ji} - e_i^\top \mu_j| \le W_1(\widehat{\gamma}_i, \gamma_i) \le \epsilon.$$

Let $\mu'_j = (\mu_{j1}, \ldots, \mu_{jk})^\top \in \mathcal{A}$. Then

$$w_j \left\|\mu'_j - \mu_j\right\|_2 \le \sqrt{k} w_j \left\|\mu'_j - \mu_j\right\|_\infty \le \sqrt{k}\epsilon.$$

Since $\mathcal{W}$ is an $n^{-\frac{1}{4k-2}}$-covering of the probability simplex with respect to $\|\cdot\|_1$, there exists a weights vector $w' = (w'_1, \ldots, w'_k) \in \mathcal{W}$ such that $\|w' - w\|_1 \le \epsilon$.

Consider the distributions $\gamma' \triangleq \sum_{j=1}^k w'_j \delta_{\mu'_j} \in \mathcal{S}$ and $\gamma'' \triangleq \sum_{j=1}^k w_j \delta_{\mu'_j}$. Note that γ and γ'' have the same weights. Using their natural coupling we have $W_1(\gamma, \gamma'') \le \sum_{j=1}^k w_j \|\mu_j - \mu'_j\|_2$. Note that γ'' and γ' have the same support. Using the total variation coupling (see [Gibbs and Su, 2002, Theorem 4]) of their weights w and w' (and the fact that total variation equals half of the ℓ_1-distance), we have $W_1(\gamma'', \gamma') \le R\|w' - w\|_1$. Thus,

$$W_1(\gamma, \gamma') \le \sum_{j=1}^k w_j \|\mu_j - \mu'_j\|_2 + R\|w' - w\|_1 \le C_k \epsilon. \qquad \square$$

Proof of Lemma 2.2.4. Throughout this proof we use the abbreviation $\epsilon \equiv \epsilon_{n,k}$. First,

$$W_1(\hat{\gamma}, \gamma) \precsim_k \sup_{\theta \in S^{k-1}} W_1(\hat{\gamma}_\theta, \gamma_\theta) \precsim_k 2R\epsilon + \max_{\theta \in \mathcal{N}} W_1(\hat{\gamma}_\theta, \gamma_\theta).$$

The first inequality is by the upper bound in Lemma 2.2.1, with $d = k$. The second inequality follows by definition of $\mathcal{N}$, for any $\theta \in S^{k-1}$, there is a $u \in \mathcal{N}$ such that $\|\theta - u\| \le \epsilon$. Now each estimated marginal $\hat{\gamma}_j$ is supported on $[-R, R]$ by nature

of the DMM algorithm, so $\hat{\gamma}$ has atoms on the hypercube $[-R, R]^k$, and similarly for γ. By the Cauchy-Schwarz Inequality and the natural coupling between γ_u, γ_θ, $W_1(\gamma_u, \gamma_\theta) \leq \sqrt{k}R\|\theta - u\|_2$. Now let $\gamma' \in \operatorname{argmin}_{\gamma'' \in \mathcal{S}} W_1(\gamma'', \gamma)$. We have

$$\max_{\theta \in \mathcal{N}} W_1(\hat{\gamma}_\theta, \gamma_\theta) \leq \max_{\theta \in \mathcal{N}} W_1(\hat{\gamma}_\theta, \gamma'_\theta) + \max_{\theta \in \mathcal{N}} W_1(\gamma'_\theta, \gamma_\theta)$$

Now for the first term, we have

$$\begin{aligned}
\max_{\theta \in \mathcal{N}} W_1(\hat{\gamma}_\theta, \gamma'_\theta) &\leq \max_{\theta \in \mathcal{N}} W_1(\hat{\gamma}_\theta, \widehat{\gamma_\theta}) + \max_{\theta \in \mathcal{N}} W_1(\widehat{\gamma_\theta}, \gamma'_\theta) \\
&\leq 2 \max_{\theta \in \mathcal{N}} W_1(\widehat{\gamma_\theta}, \gamma'_\theta) && \text{by definition of } \hat{\gamma} \\
&\leq 2 \max_{\theta \in \mathcal{N}} W_1(\widehat{\gamma_\theta}, \gamma_\theta) + 2 \max_{\theta \in \mathcal{N}} W_1(\gamma_\theta, \gamma'_\theta).
\end{aligned}$$

Putting all this together and applying the lower bound in Lemma 2.2.1, we obtain the deterministic bound

$$W_1(\hat{\gamma}, \gamma) \lesssim_k 2R\epsilon + 2 \max_{\theta \in \mathcal{N}} W_1(\widehat{\gamma_\theta}, \gamma_\theta) + 3W_1(\gamma', \gamma).$$

Applying Lemma 2.2.2 and Lemma 2.2.3 completes the proof. □

It remains to show Lemmas 2.2.5 and 2.2.6.

Proof of Lemma 2.2.5. Let $\mathcal{V}_r'^{\perp}$ be the subspace of $\mathbb{R}^d$ that is orthogonal to the space spanned by the top r eigenvectors of Σ', and let $y_j = \operatorname{argmax}_{x \in \mathcal{V}_r'^{\perp} \cap S^{d-1}} |\mu_j^\top x|$. Then $\|\mu_j - H'_r \mu_j\|_2^2 = (\mu_j^\top y_j)^2$. Furthermore, for each j, $y_j^\top w_j \mu_j \mu_j^\top y_j \leq y_j^\top (\sum_{\ell=1}^k w_\ell \mu_\ell \mu_\ell^\top) y_j = y_j^\top \Sigma y_j$. It remains to bound the latter. Let $\lambda'_1 \geq \ldots \geq \lambda'_d$ be the sorted eigenvalues

of Σ'. Now

$$\begin{aligned}
|y_j^\top \Sigma y_j| &\le |y_j^\top (\Sigma - \Sigma') y_j| + |y_j^\top \Sigma' y_j| \\
&\le \|\Sigma - \Sigma'\|_2 + \lambda'_{r+1} && \text{since } y_j \in \mathcal{V}_r'^{\perp} \\
&\le 2\|\Sigma - \Sigma'\|_2 + \lambda_{r+1},
\end{aligned}$$

where the last step follows from Weyl's inequality [Horn and Johnson, 1991]. By the natural coupling between Γ and $\Gamma_{H'_r}$,

$$W_2^2(\Gamma, \Gamma_{H'_r}) \le \sum_{j=1}^{k} w_j \|\mu_j - H'_r \mu_j\|_2^2 \le k(\lambda_{r+1} + 2\|\Sigma - \Sigma'\|_2). \qquad \square$$

The result follows since $W_1(\Gamma, \Gamma_{H'_r}) \le W_2(\Gamma, \Gamma_{H'_r})$ by the Cauchy-Schwarz Inequality.

Proof of Lemma 2.2.6. Write $X_i = U_i + Z_i$ where $U_i \overset{\text{i.i.d.}}{\sim} \Gamma$ and $Z_i \overset{\text{i.i.d.}}{\sim} N(0, I_d)$ for $i = 1, \ldots, n$. Then

$$\hat{\Sigma} - \Sigma = \left(\frac{1}{n} \sum_{i=1}^{n} U_i U_i^\top - \Sigma \right) + \left(\frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top - I_d \right) + \left(\frac{1}{n} \sum_{i=1}^{n} U_i Z_i^\top + Z_i U_i^\top \right).$$

We upper bound the spectral norms of three terms separately. For the first term, let $\Gamma = \sum_{j=1}^{k} w_j \delta_{\mu_j}$ and $\hat{w}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{U_i = \mu_j\}$. Then $\frac{1}{n} \sum_{i=1}^{n} U_i U_i^\top = \sum_{j=1}^{k} \hat{w}_j \mu_j \mu_j^\top$. Therefore, by Hoeffding's inequality and the union bound, with probability $1 - 2ke^{-2t^2}$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} U_i U_i^\top - \Sigma \right\|_2 \le R^2 \sum_{j=1}^{k} |\hat{w}_j - w_j| \le \frac{R^2 k t}{\sqrt{n}}. \tag{2.2.25}$$

For the second term, by standard results in random matrix theory (see, e.g., [Vershynin, 2012, Corollary 5.35]), and since $d < n$, there exists a positive constant C such that, with

probability at least $1 - e^{-t^2}$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top - I_d \right\|_2 \leq C \left(\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}} + \frac{t^2}{n} \right). \tag{2.2.26}$$

To bound the third term, let $A = \frac{1}{n} \sum_{i=1}^{n} U_i Z_i^\top + Z_i U_i^\top$, and $\mathcal{N}$ be an $\frac{1}{4}$-covering of S^{d-1} of size 2^{cd} for an absolute constant c. Then

$$\|A\|_2 = \max_{\theta \in S^{d-1}} |\theta^\top A \theta| \leq 2 \max_{\theta \in \mathcal{N}} |\theta^\top A \theta| = 4 \max_{\theta \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^{n} (\theta^\top U_i)(\theta^\top Z_i) \right|.$$

For fixed $\theta \in S^{d-1}$, conditioning on U_i, we have $\sum_{i=1}^{n} (\theta^\top U_i)(\theta^\top Z_i) \sim N(0, \sum_{i=1}^{n} (\theta^\top U_i)^2)$. Since $\sum_{i=1}^{n} (\theta^\top U_i)^2 \leq nR^2$, we have

$$\mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} (\theta^\top U_i)(\theta^\top Z_i) \right| > \frac{R\tau}{\sqrt{n}} \right\} \leq \mathbb{P}\left\{ |Z_1| \geq \tau \right\} \leq 2e^{-\frac{\tau^2}{2}}.$$

Therefore, by a union bound, with probability $1 - 2e^{cd - \frac{\tau^2}{2}}$

$$\max_{\theta \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^{n} (\theta^\top U_i)(\theta^\top Z_i) \right| \leq \frac{R\tau}{\sqrt{n}}.$$

By taking $\tau = C(\sqrt{d} + t)$ for some absolute constant C, we obtain that with probability $1 - e^{-t^2}$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} U_i Z_i^\top + Z_i U_i^\top \right\|_2 \leq CR \left(\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}} \right). \tag{2.2.27}$$

□

2.3 Density estimation

In this section we prove the density estimation guarantee of Theorem 2.1.2 for finite Gaussian mixtures. Our strategy is to prove a tight upper bound on the local entropy of Hellinger balls for k-GMs. In Lemma 2.3.6 we show that any ϵ-Hellinger ball in $\mathcal{P}_{k,d}$ can be covered by at most $\left(\frac{C\epsilon}{\delta}\right)^{Cd}$ δ-Hellinger balls, where C only depends on k.

This allows us to invoke the construction of Le Cam and Birgé (see [Le Cam, 1973, Birgé, 1983, Birgé, 1986] to arrive at an estimator that achieves the parametric rate of $O_k(\frac{d}{n})$ in the squared Hellinger loss. We note that the Le Cam-Birgé estimator is a theoretical construction based on (exponentially many) pairwise tests. Finding a computationally efficient proper density estimate that attains the parametric rate in Theorem 2.1.2, or even within logarithmic factors thereof, is open. This problem is much more challenging than estimating the mixing distribution, for which we have constructed a polynomial-time optimal estimator in Section 2.2. In fact, we show in Section 2.3.3, estimation of the mixing distribution can be reduced to proper density estimation both statistically and computationally.

The dimension-free nature of Theorem 2.1.2 depends crucially on the fact that mixture densities in $\mathcal{P}_{k,d}$ can be represented as the product of a mixture density on the relevant variables and a Gaussian density on irrelevant noise variables, as in (1.2.4). This leads to a crucial fact about statistical distances: they can be reduced to statistical distances in a lower dimensional space, as we now state and prove.

Lemma 2.3.1. *Let $\Gamma, \Gamma' \in S$, where S is a subspace of $\mathbb{R}^d$ of dimension at most r. Let $V = [v_1, \ldots, v_r]$ be the matrix whose columns for an orthonormal basis for S. Let $D \in \{\chi^2, KL, H^2\}$. Then*

$$D\left(\Gamma * N(0, I_d), \Gamma' * N(0, I_d)\right) = D(\Gamma_V * N(0, I_r), \Gamma'_V * N(0, I_r).$$

Proof. We prove the fact for the chi-square divergence; proofs for the other divergences are exactly similar. Let $V_C = [v_{r+1}, \ldots, v_d]$ be the matrix whose columns are

an orthonormal basis of the complement, so $[V, V_C]$ is an othornormal matrix.

$$\chi^2\left(\Gamma * N(0, I_d) || \Gamma' * N(0, I_d)\right) = \int_{x \in \mathbb{R}^d} \phi_{d-r}(V_C^\top x) \frac{\left(p_{\Gamma_V}(V^\top x) - p_{\Gamma'_V}(V^\top x)\right)^2}{p_{\Gamma'_V}(V^\top x)} \, dx$$
$$= \int_{x \in \mathbb{R}^r} \frac{\left(p_{\Gamma_V}(x) - p_{\Gamma'_V}(x)\right)^2}{p_{\Gamma'_V}(x)} \, dx.$$

□

2.3.1 Moment characterization of k-atomic Gaussian mixtures

Gaussian mixtures of the form (1.2.1) can be represented in terms of the moments of Γ; it turns out that convergence rates in such mixtures depend crucially on the moment space of Γ. These moments appear because densities like (1.2.3) can be orthogonally decomposed using the *Hermite polynomials*. Let H_j be the degree-j Hermite polynomial, defined via

$$H_j(x) \triangleq j! \sum_{i=1}^{\lfloor j/2 \rfloor} \frac{(-1/2)^i}{i!(j-2i)!} x^{j-2i}, \text{ for } x \in \mathbb{R}. \tag{2.3.1}$$

Its multivariate counterpart is defined as follows. For each multi-index $\mathbf{j} = (j_1, \ldots, j_d) \in \mathbb{Z}_+^d$, define the $\mathbf{j}$th Hermite polynomial as

$$H_{\mathbf{j}}(x) = \prod_{i=1}^{d} H_{j_i}(x_i), \quad x \in \mathbb{R}^d. \tag{2.3.2}$$

which is a degree-$|\mathbf{j}|$ polynomial in x. Recall that for $x, u \in \mathbb{R}$, $\phi(x-u) = \phi(x) \sum_{j \geq 0} H_j(x) \frac{u^j}{j!}$ (see [Abramowitz and Stegun, 1964, 22.9.17]). The multivariate extension of this fact is

$$\phi_d(x-u) = \phi_d(x) \sum_{\mathbf{j} \in \mathbb{Z}_+^d} \frac{H_{\mathbf{j}}(x)}{\mathbf{j}!} \prod_{i=1}^{d} u_i^{j_i}, \quad x, u \in \mathbb{R}^d.$$

And therefore

$$p_\Gamma(x) = \mathbb{E}_{U\sim\Gamma}[\phi_d(x-U)] = \phi_d(x)\sum_{\mathbf{j}\in\mathbb{Z}_+^d}\frac{H_{\mathbf{j}}(x)}{\mathbf{j}!}\underbrace{\mathbb{E}_{U\sim\Gamma}\left[\prod_{i=1}^d U_i^{j_i}\right]}_{m_{\mathbf{j}}(\Gamma)}. \tag{2.3.3}$$

Let $\mathbf{j} = (j_1, \ldots, j_d) \in \mathbb{Z}_+^d$. The $\mathbf{j}$th multivariate moment of Γ is

$$m_{\mathbf{j}}(\Gamma) = \mathbb{E}_{U\sim\Gamma}[U_{j_1}\cdots U_{j_d}]. \tag{2.3.4}$$

Let $|\mathbf{j}| \triangleq j_1 + \ldots j_d$. Then $m_{\mathbf{j}}(\Gamma)$ is the $\mathbf{j}$th entry of the moment tensor $M_{|\mathbf{j}|}(\Gamma)$, which we now define. For any d-dimensional random vector U, its order-ℓ *moment tensor* is

$$M_\ell(\Gamma) \triangleq \mathbb{E}_{U\sim\Gamma}[\underbrace{U\otimes\cdots\otimes U}_{\ell \text{ times}}]. \tag{2.3.5}$$

We will also use the notation $M_\ell(U) = M_\ell(\Gamma)$ where $U \sim \Gamma$. Note that $M_1(\Gamma) = \mathbb{E}[U]$ and $M_2(U - \mathbb{E}[U])$ are the mean and the covariance matrix of U, respectively. For a positive integer j, let $S_{j,d} = \{\mathbf{j} \in \mathbb{Z}_d^+ : |\mathbf{j}| = j\}$. Let $\Delta m_{\mathbf{j}}(\Gamma, \Gamma') = m_{\mathbf{j}}(\Gamma) - m_{\mathbf{j}}(\Gamma')$.

We now recall some basic facts about tensors that will be useful in this work. The *rank* of an order-ℓ tensor $T \in (\mathbb{R}^d)^{\otimes\ell}$ is defined as the minimum r such that T can be written the sum of r rank-one tensors, namely [Kruskal, 1977]:

$$\text{rank}(T) \triangleq \min\left\{r : T = \sum_{i=1}^r \alpha_i u_i^{(1)}\otimes\cdots\otimes u_i^{(\ell)}, \quad u_i^{(j)} \in \mathbb{R}^d, \alpha_i \in \mathbb{R}\right\}, \tag{2.3.6}$$

We will also use the *symmetric rank* [Comon et al., 2008]:

$$\text{rank}_s(T) \triangleq \min\left\{r : T = \sum_{i=1}^r \alpha_i u_i^{\otimes\ell}, \quad u_i \in \mathbb{R}^d, \alpha_i \in \mathbb{R}\right\}. \tag{2.3.7}$$

An order-ℓ tensor T is *symmetric* if $T_{j_1,\ldots,j_\ell} = T_{j_{\pi(1)},\ldots,j_{\pi(\ell)}}$ for all $j_1, \ldots, j_\ell \in [d]$ and all permutations π on $[\ell]$. By definition, moment tensors are symmetric. The Frobenius

norm of a tensor T is defined as $\|T\|_F \triangleq \sqrt{\langle T, T\rangle}$, where the tensor inner product is defined as $\langle S, T\rangle = \sum_{j_1,\ldots,j_\ell \in [d]} S_{j_1,\ldots,j_\ell} T_{j_1,\ldots,j_\ell}$. The spectral norm (operator norm) of a tensor T is defined as

$$\|T\| \triangleq \max\{\langle T, u_1 \otimes u_2 \otimes \cdots \otimes u_\ell\rangle : \|u_i\| = 1, i = 1, \ldots, \ell\}. \tag{2.3.8}$$

Denote the set of d-dimensional order-ℓ symmetric tensors by $\mathbb{S}_\ell(\mathbb{R}^d)$. For a symmetric tensor, the following result attributed to Banach ([Banach, 1938, Friedland and Lim, 2018]) is crucial this work:

$$\|T\| = \max\{|\langle T, u^{\otimes \ell}\rangle| : \|u\| = 1\}. \tag{2.3.9}$$

For $T \in \mathbb{S}_\ell(\mathbb{R}^d)$, if $\mathrm{rank}_s(T) \le r$, then the spectral norm can be bounded by the Frobenius norm as follows [Qi, 2011]:

$$\frac{1}{\sqrt{r^{\ell-1}}} \|T\|_F \le \|T\| \le \|T\|_F . \tag{2.3.10}$$

An important observation is that the moment of the projection of a random vector can be expressed in terms of the moment tensor as follows: for any $u \in \mathbb{R}^d$,

$$m_\ell(\langle X, u\rangle) = \mathbb{E}[\langle X, u\rangle^\ell] = \mathbb{E}[\langle X^{\otimes \ell}, u^{\otimes \ell}\rangle] = \langle M_\ell(X), u^{\otimes \ell}\rangle.$$

Consequently, the difference between two moment tensors measured in the spectral norm is equal to the maximal moment difference of their projections. Indeed, thanks to (2.3.9),

$$\|M_\ell(X) - M_\ell(Y)\| = \sup_{\|u\|=1} |m_\ell(\langle X, u\rangle) - m_\ell(\langle Y, u\rangle)|. \tag{2.3.11}$$

Furthermore, if U is a discrete random variable with a few atoms, then its moment tensor has low rank. Specifically, if U is distributed according to some k-atomic

distribution $\Gamma = \sum_{i=1}^{k} w_i \delta_{\mu_i}$, then

$$M_\ell(\Gamma) = \sum_{i=1}^{k} w_i \mu_i^{\otimes \ell}, \tag{2.3.12}$$

whose symmetric rank is at most k.

We now present a simple result showing that statistical distances between Gaussian mixture distributions of the form (1.2.1) can be upper bounded by distances between the moment tensors of their corresponding mixing distributions. This result also relies on two simple facts that are stated and proved below in Lemma 2.3.10 and Lemma 2.3.11. We include them in this section because they are used elsewhere in Chapters 2-3.

Lemma 2.3.2 (Moment upper bound on statistical distances)**.** *Let Γ, Γ' be distributions supported on $B(0,R) \cap S$ where S is a subspace of $\mathbb{R}^d$ of rank at most r. Then*

$$\chi^2(P_\Gamma || P_{\Gamma'}) \leq C \sum_{j \geq 1} \frac{\|M_j(\Gamma) - M_j(\Gamma')\|_F^2 \, r^j}{j!}. \tag{2.3.13}$$

where the constant C may depend on R and r but does not depend on d.

Proof of Lemma 2.3.2. By Lemma 2.3.11, we may assume without loss of generality that Γ' has mean zero. We also note that the following orthogonality property of the multivariate Hermite polynomials is inherited from that of univariate Hermites: for $Z \sim N(0, I_r)$,

$$\mathbb{E}[H_{\mathbf{j}}(Z) H_{\mathbf{j}'}(Z)] = \mathbf{j}! \mathbf{1}_{\{\mathbf{j}=\mathbf{j}'\}}. \tag{2.3.14}$$

We obtain

$$\begin{aligned}
\chi^2\left(P_\Gamma||P_{\Gamma'}\right) &= \chi^2\left(P_{\Gamma_V}||P_{\Gamma'_V}\right) && \text{by Lemma 2.3.1}\\
&\le e^{R^2/2}\int_{\mathbb{R}^r}\frac{(p_{\Gamma_V}(x)-p_{\Gamma'_V}(x))^2}{\phi_r(x)} && \text{by Lemma 2.3.10}\\
&= e^{R^2/2}\mathbb{E}_{N(0,I_r)}\left(\sum_{\mathbf{j}\in\mathbb{Z}_+^r}\frac{H_{\mathbf{j}}(x)\,\Delta m_{\mathbf{j}}(\Gamma_V,\Gamma'_V)}{\mathbf{j}!}\right)^2 && \text{by the expansion (2.3.3)}\\
&= e^{R^2/2}\sum_{\mathbf{j}\in\mathbb{Z}_+^r}\frac{(\Delta m_{\mathbf{j}}(\Gamma,\Gamma'))^2}{\mathbf{j}!} && \text{by (2.3.14)}\\
&\le e^{R^2/2}\sum_{\ell\ge 1}\frac{\|M_\ell(\Gamma)-M_\ell(\Gamma')\|_{\mathrm{F}}^2}{\ell!}r^\ell,
\end{aligned}$$

where the final step is by the fact that $(|\mathbf{j}|)! \le \mathbf{j}! r^{|\mathbf{j}|}$ for any $\mathbf{j}\in\mathbb{Z}_+^r$. □

The following result gives a characterization of statistical distances (squared Hellinger, KL, or χ^2-divergence) between k-GMs in terms of the moment tensors up to constant factors that do not depend on d.

Theorem 2.3.3 (Moment characterization of statistical distances)**.** *For any pair of k-atomic distributions Γ,Γ' supported on the ball $B(0,R)$ in $\mathbb{R}^d$, for any $D\in\{H^2,\mathsf{KL},\chi^2\}$,*

$$c\max_{\ell\le 2k-1}\|M_\ell(\Gamma)-M_\ell(\Gamma')\|_{\mathrm{F}}^2 \le D(P_\Gamma,P_{\Gamma'}) \le C\max_{\ell\le 2k-1}\|M_\ell(\Gamma)-M_\ell(\Gamma')\|_{\mathrm{F}}^2. \quad (2.3.15)$$

where the constants c,C may depend on k and R but not d.

To prove Theorem 2.3.3 we need a few auxiliary lemmas. The following lemma bounds the difference of higher-order moment tensors of k-atomic distributions using those of the first $2k-1$ moment tensors. The one-dimensional version was shown in [Wu and Yang, 2019, Lemma 10] using polynomial interpolation techniques; however, it is hard to extend this proof to multiple dimensions as multivariate polynomial

interpolation (on arbitrary points) is much less well-understood. Fortunately, this difficulty can be sidestepped by exploiting the relationship between moment tensor norms and projections in (2.3.11).

Lemma 2.3.4. *Let U, U' be k-atomic random variables in $\mathbb{R}^d$. Then for any $j \geq 2k$,*

$$\|M_j(U) - M_j(U')\| \leq 3^j \max_{\ell \in [2k-1]} \|M_\ell(U) - M_\ell(U')\|.$$

Proof.

$$\begin{aligned}
\|M_j(U) - M_j(U')\| &\overset{\text{(a)}}{=} \sup_{\|v\|=1} |m_j(\langle U, v\rangle) - m_j(\langle U', v\rangle)| \\
&\overset{\text{(b)}}{\leq} 3^j \sup_{\|v\|=1} \max_{\ell \in [2k-1]} |m_\ell(\langle U, v\rangle) - m_\ell(\langle U', v\rangle)| \\
&\overset{\text{(c)}}{=} 3^j \max_{\ell \in [2k-1]} \|M_\ell(U) - M_\ell(U')\|,
\end{aligned}$$

where (a) and (c) follow from (2.3.11), and (b) follows from [Wu and Yang, 2019, Lemma 10]. □

The lower bound part of Theorem 2.3.3 can be reduced to the one-dimensional case, which is covered by the following lemma. The proof relies on Newton interpolating polynomials and is deferred till Section 2.3.4.

Lemma 2.3.5. *Let γ, γ' be k-atomic distributions supported on $[-R, R]$. Then for any $(2k-1)$-times differentiable test function h,*

$$H(\gamma * N(0,1), \gamma' * N(0,1)) \geq c \left| \int h d\gamma - \int h d\gamma' \right|, \tag{2.3.16}$$

where c is some constant depending only on k, R, and $\max_{0 \leq i \leq 2k-1} \|h^{(i)}\|_{L_\infty([-R,R])}$.

Proof of Theorem 2.3.3. Since

$$H^2(P,Q) \le \mathsf{KL}(P\|Q) \le \chi^2(P\|Q), \tag{2.3.17}$$

(see, e.g., [Tsybakov, 2009, Section 2.4.1]), it suffices to prove the lower bound for H^2 and the upper bound for χ^2.

Let $U \sim \Gamma$ and $U' \sim \Gamma'$, $X \sim P_\Gamma = \Gamma * N(0, I_d)$ and $X' \sim P_{\Gamma'} = \Gamma' * N(0, I_d)$. Then $\langle \theta, X\rangle \sim P_{\Gamma_\theta}$ and $\langle \theta, X'\rangle \sim P_{\Gamma'_\theta}$. By the data processing inequality,

$$H(P_\Gamma, P_{\Gamma'}) \ge \sup_{\theta \in S^{d-1}} H(P_{\Gamma_\theta}, P_{\Gamma'_\theta}). \tag{2.3.18}$$

Applying Lemma 2.3.5 to all monomials of degree at most $2k-1$, we obtain

$$H(P_\Gamma, P_{\Gamma'}) \ge c \sup_{\theta \in S^{d-1}} \max_{\ell \le 2k-1} |m_\ell(\langle \theta, U\rangle) - m_\ell(\langle \theta, U'\rangle)| = c \max_{\ell \le 2k-1} \|M_\ell(U) - M_\ell(U')\|, \tag{2.3.19}$$

for some constant c depending on k and R, where the last equality is due to (2.3.11). Thus the desired lower bound for Hellinger follows from the tensor norm comparison in (2.3.10).

To show the upper bound, for notational convenience, let $B = V^\top U \sim \nu$ and $B' = V^\top U' \sim \nu'$. For the upper bound, by Lemma 2.3.2,

$$\begin{aligned}
\chi^2(P_\nu \| P_{\nu'}) &\overset{\text{(c)}}{\le} e^{\frac{R^2}{2}} e^{2k} \max_{\ell \in [2k-1]} \|M_\ell(B) - M_\ell(B')\|_{\mathrm{F}}^2 + e^{\frac{R^2}{2}} \sum_{\ell \ge 2k} \frac{(4k^2)^\ell}{\ell!} \|M_\ell(B) - M_\ell(B')\|^2 \\
&\overset{\text{(d)}}{\le} e^{\frac{R^2}{2}} \Bigg(e^{2k} + \underbrace{\sum_{\ell \ge 2k} \frac{(36k^2)^\ell}{\ell!}}_{\le e^{36k^2}} \Bigg) \max_{\ell \in [2k-1]} \|M_\ell(B) - M_\ell(B')\|_{\mathrm{F}}^2,
\end{aligned}$$

where (c) follows from the tensor norm comparison inequality (2.3.10), since the symmetric rank of $M_\ell(B) - M_\ell(B')$ is at most $2k$ for all ℓ; (d) follows from Lemma 2.3.4.

Note that if $\mathbb{E}[B'] \neq 0$, by the shift-invariance of χ^2-divergence, applying Lemma 2.3.11 to $\mu = \mathbb{E}[B']$ (which satisfies $\|\mu\| \leq R$) yields the desired upper bound. □

2.3.2 Local entropy of Hellinger balls

Recall from Section 1.1 that $N(\epsilon, A, \rho)$ the ϵ-covering number of A with respect to ρ, i.e., the minimum size of a ϵ-covering set A_ϵ such that, for any $v \in A$, there exists $\tilde{v} \in A_\epsilon$ with $\rho(v, \tilde{v}) < \epsilon$. The main result of this section is the following lemma, which bounds the covering number of a Hellinger ball in k-GMs. From this the upper bound in Theorem 2.1.2 immediately follows by invoking the Le Cam-Birgé construction ([Birgé, 1986, Theorem 3.1]; for the high-probability bound (2.1.6) see e.g., [Wu, 2017, Theorem 18.3]).

Lemma 2.3.6 (Local entropy of k-GM)**.** *For any $\Gamma_0 \in \mathcal{G}_{k,d}$, let $\mathcal{P}_\epsilon = \{P_\Gamma : \Gamma \in \mathcal{G}_{k,d}, H(P_\Gamma, P_{\Gamma_0}) \leq \epsilon\}$. Then, for any $\delta \leq \epsilon/2$,*

$$N(\delta, \mathcal{P}_\epsilon, H) \leq (\epsilon/\delta)^{c \cdot d}, \tag{2.3.20}$$

where the constant c only depends on k and R.

Proof. Let $\mathcal{M}_\epsilon = \{M(\Gamma) \colon P_\Gamma \in \mathcal{P}_\epsilon\}$, where $M(\Gamma) = (M_1(\Gamma), \ldots, M_{2k-1}(\Gamma))$ consists of the moment tensors of Γ up to degree $2k-1$. Let $c' = \sqrt{c}$ and $C' = \sqrt{C}$ where c and C are from Theorem 2.3.3. To obtain a δ-covering of $\mathcal{P}_\epsilon$, next we show that it suffices to construct a $\frac{\delta}{2C'}$-covering of $\mathcal{M}_\epsilon$ with respect to $\rho(M, M') \triangleq \max_{\ell \leq 2k-1} \|M_\ell - M'_\ell\|_F$ and thus

$$N(\delta, \mathcal{P}_\epsilon, H) \leq N(\delta/(2C'), \mathcal{M}_\epsilon, \rho). \tag{2.3.21}$$

To this end, let $\mathcal{N}$ be the optimal $\frac{\delta}{2C'}$-covering of $\mathcal{M}_\epsilon$ with respect to ρ, and we show that $\mathcal{N}' = \{P_\Gamma : \Gamma = \operatorname{argmin}_{\Gamma' : P_{\Gamma'} \in \mathcal{P}_\epsilon} \rho(M(\Gamma'), M), M \in \mathcal{N}\}$ is a δ-covering of $\mathcal{P}_\epsilon$. For any $P_\Gamma \in \mathcal{P}_\epsilon$, by the covering property of $\mathcal{N}$, there exists moment tensors $M \in \mathcal{N}$

such that $\rho(M, M(\Gamma)) < \frac{\delta}{2C'}$. By the definition of $\mathcal{N}'$, there exists $P_{\tilde{\Gamma}} \in \mathcal{N}'$ such that $\rho(M(\tilde{\Gamma}), M) < \frac{\delta}{2C'}$. Therefore, $\rho(M(\tilde{\Gamma}), M(\Gamma)) < \frac{\delta}{C'}$ and thus $H(P_{\tilde{\Gamma}}, P_\Gamma) < \delta$ by Theorem 2.3.3. It also follows from Theorem 2.3.3 and the fact that Γ_0, Γ are both k-atomic that

$$\mathcal{M}_\epsilon \subseteq M(\Gamma_0) + \{\Delta : \|\Delta_\ell\|_F \leq \epsilon/c', \mathrm{rank}_s(\Delta_\ell) \leq 2k, \forall \ell \leq 2k-1\}, \tag{2.3.22}$$

where $\Delta = (\Delta_1, \ldots, \Delta_{2k-1})$ and $\Delta_\ell \in \mathbb{S}_\ell(\mathbb{R}^d)$. Let $\mathcal{D}_\ell = \{\Delta_\ell \in \mathbb{S}_\ell(\mathbb{R}^d) : \|\Delta_\ell\|_F \leq \epsilon/c', \mathrm{rank}_s(\Delta_\ell) \leq 2k\}$, and $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_{2k-1}$ be the Cartesian product. By monotonicity,

$$N(\delta/(2C'), \mathcal{M}_\epsilon, \rho) \leq N(\delta/(2C'), \mathcal{D}, \rho) \leq \prod_{\ell=1}^{2k-1} N(\delta/(2C'), \mathcal{D}_\ell, \|\cdot\|_F). \tag{2.3.23}$$

The conclusion follows from Lemma 2.3.7. □

Lemma 2.3.7. *Let* $\mathcal{T} = \{T \in \mathbb{S}_\ell(\mathbb{R}^d) : \|T\|_F \leq \epsilon, \mathrm{rank}_s(T) \leq r\}$. *Then, for any* $\delta \leq \epsilon/2$,

$$N(\delta, \mathcal{T}, \|\cdot\|_F) \leq \left(\frac{C\ell\epsilon}{\delta}\right)^{dr} \left(\frac{C\epsilon}{\delta}\right)^{r^\ell}, \tag{2.3.24}$$

for some absolute constant C.

Proof. For any $T \in \mathcal{T}$, $\mathrm{rank}_s(T) \leq r$. Thus $T = \sum_{i=1}^r a_i v_i^{\otimes \ell}$ for some $a_i \in \mathbb{R}$ and $v_i \in S^{d-1}$. Furthermore, $\|T\|_F \leq \epsilon$. Ideally, if the coefficients satisfied $|a_i| \leq \epsilon$ for all i, then we could cover the r-dimensional ϵ-hypercube with an $\frac{\epsilon}{2}$-covering, which, combined with a $\frac{1}{2}$-covering of the unit sphere that covers the unit vectors v_i's, constitutes a desired covering for the tensor. Unfortunately the coefficients a_i's need not be small due to the possible cancellation between the rank-one components (consider the counterexample of $0 = v^{\otimes \ell} - v^{\otimes \ell}$). Next, to construct the desired covering we turn to the Tucker decomposition of the tensor T.

Let $u = (u_1, \ldots, u_r)$ be an orthonormal basis for the subspace spanned by $(v_1, \ldots, v_r)$.

In particular, let $v_i = \sum_{j=1}^r b_{ij} u_j$. Then

$$T = \sum_{\mathbf{j}=(j_1,\ldots,j_\ell)\in[r]^\ell} \alpha_{\mathbf{j}} \underbrace{u_{j_1} \otimes \cdots \otimes u_{j_\ell}}_{\triangleq u_{\mathbf{j}}}, \tag{2.3.25}$$

where $\alpha_{\mathbf{j}} = \sum_{i=1}^r a_i b_{ij_1} \cdots b_{ij_\ell}$. In tensor notation, T admits the following *Tucker decomposition*

$$T = \alpha \times_1 U \cdots \times_\ell U \tag{2.3.26}$$

where the symmetric tensor $\alpha = (\alpha_{\mathbf{j}}) \in \mathbb{S}_\ell(\mathbb{R}^r)$ is called the core tensor and U is a $r \times d$ matrix whose rows are given by $u_1, \ldots, u_r$.

Due to the orthonormality of $(u_1, \ldots, u_r)$, we have for any $\mathbf{j}, \mathbf{j}' \in [r]^\ell$,

$$\langle u_{\mathbf{j}},\, u_{\mathbf{j}'} \rangle = \prod_{i=1}^{\ell} \langle u_{j_i},\, u_{j_i'} \rangle = \mathbf{1}_{\{\mathbf{j}=\mathbf{j}'\}}. \tag{2.3.27}$$

Hence we conclude from (2.3.25) that

$$\|\alpha\|_F = \|T\|_F . \tag{2.3.28}$$

In particular $\|\alpha\|_F \le \epsilon$. Therefore,

$$\mathcal{T} \subseteq \mathcal{T}' \triangleq \left\{ T = \sum_{\mathbf{j}\in[r]^\ell} \alpha_{\mathbf{j}} u_{j_1} \otimes \cdots \otimes u_{j_\ell} : \|\alpha\|_F \le \epsilon, \langle u_i,\, u_j \rangle = \mathbf{1}_{\{i=j\}} \right\}. \tag{2.3.29}$$

Let $\tilde{A}$ be a $\frac{\delta}{2}$-covering of $\{\alpha \in \mathbb{S}_\ell(\mathbb{R}^r) : \|\alpha\|_F \le \epsilon\}$ under $\|\cdot\|_F$ of size $(\frac{C\epsilon}{\delta})^{r^\ell}$ for some absolute constant C; let $\tilde{B}$ be a $\frac{\delta}{2\ell\epsilon}$-covering of $\{(u_1, \ldots, u_r) : \langle u_i,\, u_j \rangle = \mathbf{1}_{\{i=j\}}\}$ under the maximum of column norms of size $(\frac{C\ell\epsilon}{\delta})^{dr}$. Let $\tilde{\mathcal{T}}' = \{\sum_{\mathbf{j}\in[r]^\ell} \tilde{\alpha}_{\mathbf{j}} \tilde{u}_{j_1} \otimes \cdots \otimes \tilde{u}_{j_\ell} : \tilde{\alpha} \in \tilde{A}, \tilde{u} \in \tilde{B}\}$. Next we verify the covering property.

For any $T \in \mathcal{T}'$, there exists $\tilde{T} \in \tilde{\mathcal{T}}'$ such that $\|\alpha - \tilde{\alpha}\|_F \le \frac{\delta}{2}$ and $\max_{i\le r} \|u_i - \tilde{u}_i\| \le$

$\frac{\delta}{2\ell\epsilon}$. Then, by the triangle inequality,

$$\left\|T-\tilde{T}\right\|_F \leq \sum_{\mathbf{j}} |\alpha_{\mathbf{j}}| \left\|u_{j_1}\otimes\cdots\otimes u_{j_\ell} - \tilde{u}_{j_1}\otimes\cdots\otimes\tilde{u}_{j_\ell}\right\|_F + \sum_{\mathbf{j}} |\alpha_{\mathbf{j}}-\tilde{\alpha}_{\mathbf{j}}| \left\|\tilde{u}_{j_1}\otimes\cdots\otimes\tilde{u}_{j_\ell}\right\|_F. \tag{2.3.30}$$

The second term is at most $\|\alpha-\tilde{\alpha}\|_F \leq \delta/2$. For the first term, it follows from the triangle inequality that

$$\|u_{j_1}\otimes\cdots\otimes u_{j_\ell} - \tilde{u}_{j_1}\otimes\cdots\otimes\tilde{u}_{j_\ell}\|_F \leq \sum_{i=1}^{\ell} \|u_{j_1}\otimes\cdots\otimes(u_{j_i}-\tilde{u}_{j_i})\otimes\cdots\otimes\tilde{u}_{j_\ell}\|_F \leq \frac{\delta}{2\epsilon}. \tag{2.3.31}$$

Therefore, the first term is at most $\frac{\delta}{2\epsilon}\|\alpha\|_F \leq \delta/2$. □

2.3.3 Proof of main theorem

Proof of Theorem 2.1.2. The upper bound follows from Lemma 2.3.6 and the LeCam-Birgé estimator [Le Cam, 1973, Birgé, 1983].

For the lower bound, Let $\mathcal{M} = \{M(\Gamma) : \Gamma \in \mathcal{G}_{k,d}\}$, where $M(\Gamma) = (M_1(\Gamma), \ldots, M_{2k-1}(\Gamma))$. By Theorem 2.3.3, we obtain an ϵ_n-cover in KL-divergence on the space of distributions via an cover on the moment tensor space, i.e.,

$$N(\epsilon_n, \mathcal{P}_{k,d}, KL) \leq N(\epsilon_n, \mathcal{M}, \rho),$$

where $\rho(M, M') = \max_{\ell\in[2k-1]} \|M_\ell - M'_\ell\|_F$. By Lemma 2.3.6, $N(\epsilon, \mathcal{M}, \rho) \leq (C/\epsilon)^{c_k d}$, where C, c_k are positive constants. We moreover have from Lemma 2.3.6, and the relationship between packing and cover numbers [3], that $M(\epsilon, \mathcal{P}_{k,d}, H) \leq (C'/\epsilon_{n,H})^{c_k d}$. Thus we have

$$N(\epsilon, \mathcal{P}_{k,d}, KL) \asymp M(\epsilon, \mathcal{P}_{k,d}, H) \asymp \left(\frac{C}{\epsilon}\right)^{c_k d}.$$

We apply Theorem 1 of [Yang and Barron, 1999] as follows. Select $\epsilon_n = \sqrt{d\log(n/d)/n}$,

[3] $N(\epsilon, \mathcal{P}, H) \leq M(\epsilon, \mathcal{P}, H) \leq M(\epsilon/2, \mathcal{P}, H)$; see, e.g., [van der Vaart and Wellner, 1996], page 98, Definition 2.2.3.

so $n\epsilon_n^2 = d\log(n/d)$. And $\log N(\epsilon_n, \mathcal{P}_{k,d}, KL) \asymp c_k d\,(\log(n/d) - \log\log(n/d))$ so we satisfy:

$$n\epsilon_n^2 \geq N(\epsilon_n, \mathcal{P}_{k,d}, KL).$$

Now set $\epsilon_{n,H} = C\sqrt{d/n}$ for an appropriate positive constant C, so that $\log M(\epsilon_{n,H}, \mathcal{P}_{k,d}, H) = \frac{c_k d}{2}\log(C'' n/d)$. Then we have that as needed, $\log M(\epsilon_{n,H}, \mathcal{P}_{k,d}, H) \geq 4n\epsilon_n^2 + 2\log 2$, and we obtain a lower bound on the rate of $\epsilon_{n,H} = C\sqrt{d/n}$. □

The next result shows that optimal estimation of the mixing distribution can be reduced to that of the mixture density, both statistically and computationally, provided that the density estimate is proper, i.e., is a valid k-GM. Note that this does not mean an optimal density estimate $P_{\hat{\Gamma}}$ automatically yields an optimal estimator of the mixing distribution $\hat{\Gamma}$ for Theorem 2.1.1. Instead, we rely on an intermediate step that allows us to estimate the appropriate subspace and then perform density estimation in this low-dimensional space.

Theorem 2.3.8. *Suppose that for every $d \in \mathbb{N}$ and $\Gamma \in \mathcal{G}_{k,d}$, there exists a proper density estimator $P_{\hat{\Gamma}'}$ as a function of $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\Gamma$, such that*

$$\mathbb{E}H(P_{\hat{\Gamma}'}, P_\Gamma) \leq c_k(d/n)^{1/2}, \tag{2.3.32}$$

for some constant c_k. Then there is an estimator $\hat{\Gamma}$ of the mixing distribution Γ and a positive constant C_k such that

$$\mathbb{E}W_1(\tilde{\Gamma}, \Gamma) \leq C_k\left(\left(\frac{d}{n}\right)^{1/4} + \left(\frac{1}{n}\right)^{\frac{1}{4k-2}}\right). \tag{2.3.33}$$

Proof of Theorem 2.3.8. We first construct the estimator $\hat{\Gamma}$ using $X_1, \ldots, X_{2n} \overset{\text{i.i.d.}}{\sim} P_\Gamma$. Let $\hat{\Gamma}' \in \mathcal{G}_{k,d}$ be the estimator from $\{X_i\}_{i\leq n}$ satisfying

$$\mathbb{E}H(P_{\hat{\Gamma}'}, P_\Gamma) \leq c_k\sqrt{d/n}, \tag{2.3.34}$$

for a positive constant c_k, as guaranteed by (2.3.32). Let $\hat{V} \in \mathbb{R}^{d\times k}$ be a matrix whose columns form an orthonormal basis for the space spanned by the atoms of $\hat{\Gamma}'$, $\hat{H} = \hat{V}\hat{V}^\top$, and $\gamma = \Gamma_{\hat{V}}$. Note that conditioned on $\hat{V}$, $\{\hat{V}^\top X_i\}_{i=n+1,\dots,2n}$ is an i.i.d. sample drawn from the k-GM P_γ. Invoking (2.3.32) again, there exists an estimator $\hat{\gamma} = \sum_{j=1}^k \hat{w}_j \delta_{\hat{\psi}_j} \in \mathcal{G}_{k,k}$ such that

$$\mathbb{E}H(P_{\hat{\gamma}}, P_\gamma) \le c_k\sqrt{k/n}. \tag{2.3.35}$$

We will show that $\hat{\Gamma} \triangleq \hat{\gamma}_{\hat{V}^\top} = \sum_{j=1}^k \hat{w}_j \delta_{\hat{V}\hat{\psi}_j}$ achieves the desired rate (2.3.33). Recall from (2.2.18) the risk decomposition:

$$W_1(\Gamma, \hat{\Gamma}) \le W_1(\Gamma, \Gamma_{\hat{H}}) + W_1(\gamma, \hat{\gamma}). \tag{2.3.36}$$

Let $\Sigma = \mathbb{E}_{U\sim\Gamma}[UU^\top]$ and $\hat{\Sigma} = \mathbb{E}_{U\sim\hat{\Gamma}'}[UU^\top]$ whose ranks are at most k. Then $\hat{H}$ is the projection matrix onto the space spanned by the top k eigenvectors of $\hat{\Sigma}$. It follows from Lemma 2.2.5 that $W_1(\Gamma, \Gamma_{\hat{H}}) \le \sqrt{2k\|\Sigma - \hat{\Sigma}\|_2}$. By Lemma 2.3.5 and the data processing inequality of the Hellinger distance,

$$\|\Sigma - \hat{\Sigma}\|_2 = \sup_{\theta\in S^{d-1}} |m_2(\Gamma_\theta) - m_2(\hat{\Gamma}'_\theta)| \le C_k \sup_{\theta\in S^{d-1}} H(P_{\Gamma_\theta}, P_{\hat{\Gamma}'_\theta}) \le C_k H(P_\Gamma, P_{\hat{\Gamma}'}).$$

Therefore, by (2.3.34), we obtain that

$$\mathbb{E}W_1(\Gamma, \Gamma_{\hat{H}}) \le C_k \left(\frac{d}{n}\right)^{1/4}. \tag{2.3.37}$$

We condition on $\hat{V}$ to analyze the second term on the right-hand side of (2.3.36). By Lemmas 2.2.1 and 2.5.1,

$$W_1(\gamma, \hat{\gamma}) \le k^{5/2} \sup_{\theta\in S^{k-1}} W_1(\gamma_\theta, \hat{\gamma}_\theta) \le C_k \sup_{\theta\in S^{k-1}, r\in[2k-1]} |m_r(\gamma_\theta) - m_r(\hat{\gamma}_\theta)|^{\frac{1}{2k-1}}.$$

Again, by Lemma 2.3.5 and the data processing inequality, for any $\theta \in S^{k-1}$ and $r \in [2k-1]$,

$$|m_r(\gamma_\theta) - m_r(\hat{\gamma}_\theta)| \leq C_k H\left(P_{\hat{\gamma}_\theta}, P_{\gamma_\theta}\right) \leq C_k H\left(P_{\hat{\gamma}}, P_{\gamma}\right)$$

Therefore, by (2.3.35), we obtain that

$$\mathbb{E} W_1(\gamma, \hat{\gamma}) \leq C_k \left(\frac{1}{n}\right)^{\frac{1}{4k-2}}. \tag{2.3.38}$$

The conclusion follows by applying (2.3.37) and (2.3.38) in (2.3.36). □

2.3.4 Supporting lemmas and proofs

Lemma 2.3.5 provides a tighter relationship between mixture density estimation error in Hellinger distance and mixing distribution moment estimation than the one provided by Lemma 11 of [Wu and Yang, 2019], allowing for the sharp connection between Hellinger distance and 1-Wasserstein distance provided in Theorem 2.1.2. Note that the relationship between the Hellinger and L_2 distances used in the proof of Lemma 2.3.5 is not necessarily sharp, since it relies on a simple upper bound on the densities in question. For high dimensions, the upper bound on the densities could be large. However, we will deal with k-dimensional densities, so this upper bound is good enough.

We start by recalling the basics of polynomial interpolation in one dimension. For any function h and any set of $m+1$ distinct points $\{x_0, \ldots, x_m\}$, there exists a unique polynomial P of degree at most m, such that $P(x_i) = h(x_i)$ for $i = 0, \ldots, m$. For our purpose, it is convenient to express P in the Newton form (as opposed to the more

common Lagrange form):

$$P(x) = \sum_{j=0}^{m} a_j \prod_{i=0}^{j-1} (x - x_i), \tag{2.3.39}$$

where the coefficients are given by the finite differences of h, namely, $a_j = h[x_0, \ldots, x_j]$, which in turn are defined recursively via:

$$\begin{aligned} h[x_i] &= h(x_i) \\ h[x_i, \ldots, x_{i+r}] &= \frac{h[x_{i+1}, \ldots, x_{i+r}] - h[x_i, \ldots, x_{i+r-1}]}{x_{i+r} - x_i}. \end{aligned}$$

Proof of Lemma 2.3.5. Let $U \sim \gamma$ and $U' \sim \gamma'$. Note that $p_\gamma, p_{\gamma'}$ are bounded above by $\frac{1}{\sqrt{2\pi}}$, so we have

$$H^2(p_\gamma, p_{\gamma'}) = \int \left(\frac{p_\gamma - p_{\gamma'}}{\sqrt{p_\gamma} + \sqrt{p_{\gamma'}}} \right)^2 \geq \frac{\sqrt{2\pi}}{4} \|p_\gamma - p_{\gamma'}\|_2^2.$$

Thus to show (2.3.16), it suffices to show there is a positive constant c such that

$$\|p_\gamma - p_{\gamma'}\|_2 \geq c|\mathbb{E}[h(U)] - \mathbb{E}[h(U')]|. \tag{2.3.40}$$

Next, by suitable orthogonal expansion we can express $\|p_\gamma - p_{\gamma'}\|_2$ in terms of the "moments" of the mixing distributions (see [Wu and Verdú, 2010, Sec. VI]). Let $\alpha_j(y) = \sqrt{\frac{\sqrt{2}\phi(\sqrt{2}y)}{j!}} H_j(\sqrt{2}y)$. Then $\{\alpha_j : j \in \mathbb{Z}_+\}$ form an orthonormal basis on $L^2(\mathbb{R}, \mathrm{d}y)$ in view of (2.3.14). Since p_γ is square integrable, we have the orthogonal expansion $p_\gamma(y) = \sum_{j \geq 0} a_j(\gamma)\alpha_j(y)$, with coefficient

$$a_j(\gamma) = \langle \alpha_j, p_\gamma \rangle = \mathbb{E}[\alpha_j(U + Z)] = \mathbb{E}[(\alpha_j * \phi)(U)] = \frac{1}{2^{\frac{j+1}{2}} \pi^{\frac{1}{4}} \sqrt{j!}} \mathbb{E}[U^j e^{-\frac{U^2}{4}}]$$

where the last equality follows from the fact that [Gradshteyn and Ryzhik, 2007,

7.374.6, p. 803]

$$(\phi * \alpha_j)(y) = \frac{1}{2^{\frac{j+1}{2}}\pi^{\frac{1}{4}}\sqrt{j!}} y^j e^{-\frac{y^2}{4}}.$$

Therefore

$$\|p_\gamma - p_{\gamma'}\|_2^2 = \sum_{j\geq 0} \frac{1}{j!2^{j+1}\sqrt{\pi}} \left(\mathbb{E}[U^j e^{-U^2/4}] - \mathbb{E}[U'^j e^{-U'^2/4}]\right)^2. \tag{2.3.41}$$

In particular, for each $j \geq 0$,

$$|\mathbb{E}[U^j e^{-U^2/4}] - \mathbb{E}[U'^j e^{-U'^2/4}]| \leq \sqrt{j!2^{j+1}\sqrt{\pi}}\|p_\gamma - p_{\gamma'}\|_2. \tag{2.3.42}$$

In view of (2.3.42), to bound the difference $|\mathbb{E}[h(U)]-\mathbb{E}[h(U')]|$ by means of $\|p_\gamma-p_{\gamma'}\|_2$, our strategy is to interpolate $h(y)$ by linear combinations of $\{y^j e^{-y^2/4} : j = 0, \ldots, 2k-1\}$ on all the atoms of U and U', a total of at most $2k$ points. Clearly, this is equivalent to the standard polynomial interpolation of $\tilde{h}(y) \triangleq h(y)e^{y^2/4}$ by a degree-$(2k-1)$ polynomial. Specifically, let $T \triangleq \{t_1, \ldots, t_{2k}\}$ denote the set of atoms of γ and γ'. By assumption, $T \subset [-R, R]$. Denote the interpolating polynomial of $\tilde{h}$ on T by $P(y) = \sum_{j=0}^{2k-1} b_j y^j$. Then

$$\begin{aligned}
|\mathbb{E}[h(U)] - \mathbb{E}[h(U')]| &= |\mathbb{E}[P(U)e^{-U^2/4} - \mathbb{E}[P(U')e^{-U'^2/4}| \\
&\leq \sum_{j=0}^{2k-1} |b_j||\mathbb{E}[U^j e^{-U^2/4}] - \mathbb{E}[U'^j e^{-U'^2/4}]| \\
&\leq \|p_\gamma - p_{\gamma'}\|_2 \sum_{j=0}^{2k-1} |b_j|\sqrt{j!2^{j+1}\sqrt{\pi}}.
\end{aligned}$$

It remains to bound the coefficient b_j independently of the set T. This is given by the next lemma. □

Lemma 2.3.9. *Let h be an m-times differentiable function on the interval $[-R, R]$, whose derivatives are bounded by $|h^{(i)}(x)| \leq M$ for all $0 \leq i \leq m$ and all $x \in [-R, R]$.*

Then for any $m \geq 1$ and $R > 0$, there exists a positive constant $C = C(m, R, M)$, such that the following holds. For any set of distinct nodes $T = \{x_0, \ldots, x_m\} \subset [-R, R]$, denote by $P(x) = \sum_{i=0}^{m} b_j x^j$ the unique interpolating polynomial of degree at most m of h on T. Then $\max_{0 \leq j \leq m} |b_j| \leq C$.

Proof. Express P in the Newton form (2.3.39):

$$P(y) = \sum_{i=0}^{m} h[x_0, \ldots, x_i] \prod_{j=0}^{i-1} (y - x_j)$$

By the intermediate value theorem, finite differences can be bounded by derivatives as follows: (c.f. [Stoer and Bulirsch, 2002, (2.1.4.3)])

$$|h[x_0, \ldots, x_i]| \leq \frac{1}{i!} \sup_{|\xi| \leq R} |h^{(i)}(\xi)|.$$

Let $\prod_{j=0}^{i-1}(y - x_j) = \sum_{j=0}^{i} c_{ij} y^j$. Since $|x_j| \leq R$, $|c_{ij}| \leq C_1 = C_1(R, m)$ all i, j. This completes the proof. □

Lemma 2.3.10. *Let $\Gamma \in \mathbb{R}^d$ be a distribution supported on $B(0, R)$ with mean zero. Let $U \sim \Gamma$. Then*

$$\mathbb{E}_\Gamma \exp\left(x^\top U - \|U\|_2^2/2\right) \geq \exp(-R^2/2) \tag{2.3.43}$$

Proof. By Jensen's Inequality and the fact that $\mathbb{E}_\Gamma Y = 0$, $\mathbb{E}_\Gamma \left(\exp x^\top U - \|U\|_2^2/2\right) \geq \exp\left(-\mathbb{E}_\Gamma(\|U\|_2^2/2\right)$. The result follows because $\|U\|_2 \leq R$. □

Lemma 2.3.11. *For any random vectors X and Y and any deterministic $\mu \in \mathbb{R}^d$,*

$$\|M_\ell(X - \mu) - M_\ell(Y - \mu)\| \leq \sum_{k=0}^{\ell} \binom{\ell}{k} \|M_k(X) - M_k(Y)\| \|\mu\|^{\ell - k}$$

Proof. Using (2.3.11) and binomial expansion, we have:

$$\begin{aligned}\|M_\ell(X-\mu)-M_\ell(Y-\mu)\| &= \sup_{\|u\|=1}|m_\ell(\langle X,u\rangle-\langle\mu,u\rangle)-m_\ell(\langle Y,u\rangle-\langle\mu,u\rangle)|.\\ &\le \sup_{\|u\|=1}\sum_{k=0}^{\ell}\binom{\ell}{k}|m_k(\langle X,u\rangle)-m_k(\langle Y,u\rangle)||\langle\mu,u\rangle|^{\ell-k}\\ &\le \sum_{k=0}^{\ell}\binom{\ell}{k}\|M_k(X)-M_k(Y)\|\|\mu\|^{\ell-k}\end{aligned}$$

where in the step we used the Cauchy-Schwarz inequality. □

2.4 Numerical studies

We now present numerical results. We compare the estimator (2.2.5) to the classical EM algorithm. The algorithm that computes (2.2.5) relies on an exhaustive search and is not meant to be practical, but it turns out that it can be competitive with the EM algorithm both statistically and computationally, as our experiments show.

All simulations are run in Python. The DMM algorithm relies on the CVXPY [Diamond and Boyd, 2016] and CVXOPT ([Andersen et al., 2013]) packages; see Section 6 of [Wu and Yang, 2019] for more details on the implementation of DMM. We also use the Python Optimal Transport package ([Flamary and Courty, 2017]) to compute the d-dimensional 1-Wasserstein distance.

In all experiments, we let $d = 100$ and $\sigma = 1$. We let n range from $10,000$ to $200,000$ in increments of $10,000$. We initialize EM randomly, and our stopping criterion for the EM algorithm is either after 1000 iterations or once the relative change in log likelihood is below 10^{-6}. For the dimension reduction step in the computation of (2.2.5), we first center our data, then do the projection using the top $k-1$ eigenvectors of $\hat{\Sigma}$ in (2.2.1). Thus when $k = 2$, we project onto a one-dimensional subspace and only run DMM once, so the grid search of Algorithm 1 is never invoked.

We note that sample-splitting is used for the estimator (2.2.5) for purposes of analysis only; in the actual experiments, we do not sample split.

When $k = 3$, we project the data to a 2-dimensional subspace after centering. In this case, we need to choose $\mathcal{W}, \mathcal{N}$, the $\epsilon_{n,k}$-nets on the simplex Δ^{k-1} and on the unit sphere S^{k-2}, respectively. Here $\mathcal{W}$ is chosen by discretizing the probabilities and $\mathcal{N}$ is formed by by gridding the angles $\alpha \in [-\pi, \pi]$ and using the points $(\cos\alpha, \sin\alpha)$. Note that here, $|\mathcal{W}| \leq (C_1/\epsilon_{n,k})^{k-1}, |\mathcal{N}| \leq (C_2/\epsilon_{n,k})^{k-2}$. For example, when $n = 10000$, $1/\epsilon_{n,k} \approx 3$. In the experiments, we choose $C_1 = 1, C_2 = 4$. In our experience, an even coarser grid $\mathcal{N}$ can be used and still achieve fairly high accuracy in the well-separated models, while gaining some speed.

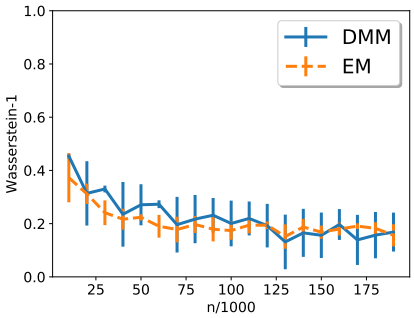
In each row of Fig. 2.1 and Fig. 2.2, we display on the average accuracy, along with the standard deviation over our 10 experiments, in the left-hand plot. We display the average running time in the right-hand plot.

In Fig. 2.1, we compare the performance on the symmetric 2-GM, where the samples are drawn from the distribution $\frac{1}{2}N(\mu, I_d) + \frac{1}{2}N(-\mu, I_d)$. For Fig. 2.1(a), $\mu = 0$, i.e., the components completely overlap. For Fig. 2.1(c) and Fig. 2.1(e), μ is uniformly drawn from the sphere of radius 1 and 2, respectively. In Fig. 2.1(g), the model is $P_\Gamma = \frac{1}{4}N(\mu, I_d) + \frac{3}{4}N(-\mu, I_d)$ where μ is drawn from the sphere of radius 2. That is, the model is the same as the one used in Fig. 2.1(e) except that the weights are uneven. We still run PCA and DMM without a grid search.

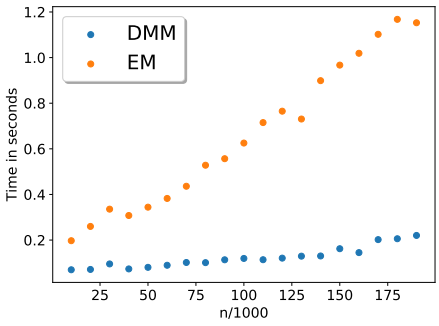
Our algorithm and EM perform similarly for the model with overlapping components; our algorithm is more accurate than EM in the model where $\|\mu\|_2 = 1$, but EM improves as the model components become more separated.

There is little difference in the performance of either algorithm in the uneven weights scenario. In terms of running time, computing (2.2.5) takes about the same time as EM for smaller values of n, but EM slows much more as n increases since it accesses all the samples on each iteration. For the largest sample size in the

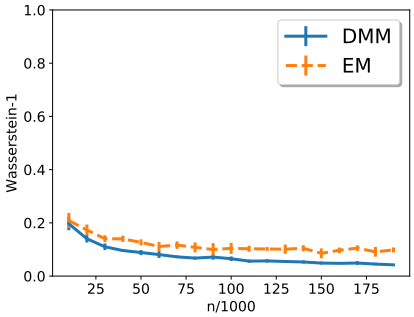
experiments, computing (2.2.5) is about 6 times faster than EM.



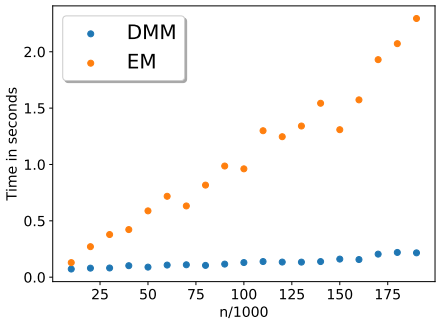
(a) $\mu = 0$



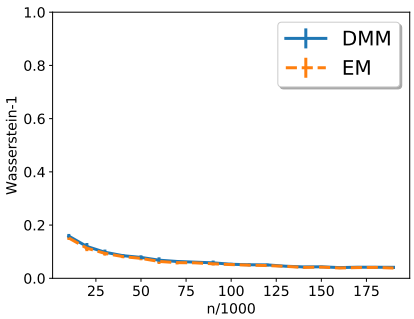
(b) $\mu = 0$



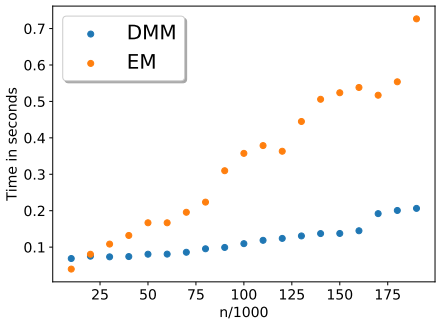
(c) $\|\mu\| = 1$



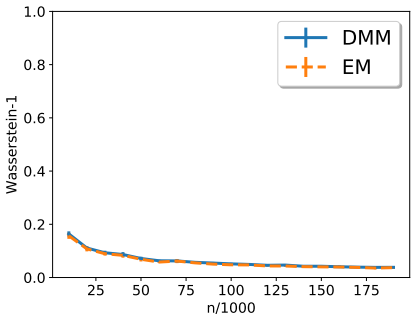
(d) $\|\mu\| = 1$



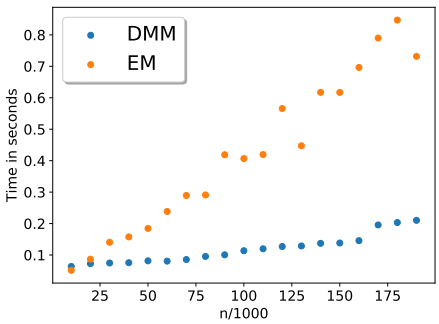
(e) $\|\mu\| = 2$



(f) $\|\mu\| = 2$



(g) $\|\mu\| = 2$

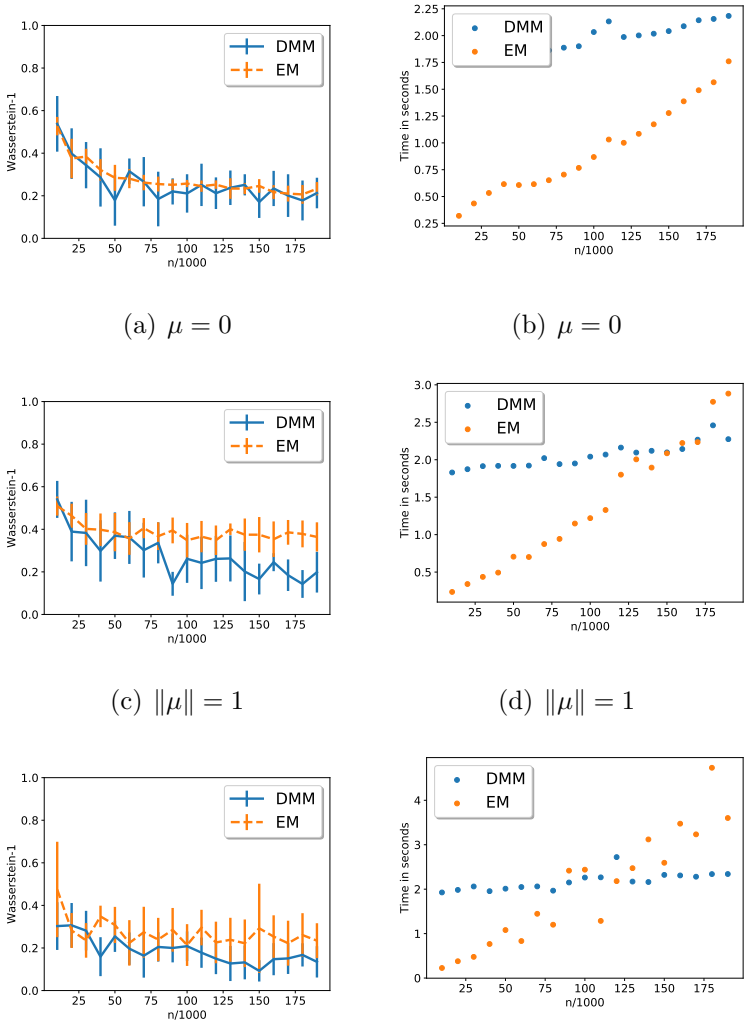


(h) $\|\mu\| = 2$

Figure 2.1: W_1 error and run time of high-dimensional DMM and EM on two-component Gaussian mixtures

In Fig. 2.2, we compare the performance on the 3-GM model $\frac{1}{3}N(\mu, I_d)+\frac{1}{3}N(0, I_d)+\frac{1}{3}N(-\mu, I_d)$. We follow the same pattern of increasing the separation of the components in each experiment. For Fig. 2.2(a), $\mu = 0$, i.e., the components completely overlap. For Fig. 2.2(c) and Fig. 2.2(e), μ is uniformly drawn from the sphere of radius 1 and 2, respectively.

In the experiments where $k = 3$, we see the opposite phenomenon in terms of the relative performance of our algorithm and EM: the former improves more as the centers become more separated. This seems to be because in, for instance, the case where $\mu = 0$, the error in each coordinate for DMM is fairly high, and this is compounded when we select the two-coordinate final distribution. The performance of our algorithm improves rapidly here because as the model becomes more separated, the errors in each coordinate become very small. Note that since we have made the model more difficult to learn by adding a center at 0, the errors are higher than for the $k = 2$ example in every experiment for both algorithms. In terms of running time, for $k = 3$, EM is faster than our algorithm for smaller sample size because of the grid search being invoked for Algorithm 1. But EM slows much more rapidly as the sample size increases and is actually slower than our algorithm for large values of n.



(a) $\mu = 0$ (b) $\mu = 0$

(c) $\|\mu\| = 1$ (d) $\|\mu\| = 1$

(e) $\|\mu\| = 2$ (f) $\|\mu\| = 2$

Figure 2.2: W_1 error and run time of high-dimensional DMM and EM on three-component Gaussian mixtures

2.5 Auxiliary lemmas and proofs

The following moment comparison inequality bound the Wasserstein distance between two univariate k-atomic distributions using their moment differences:

Lemma 2.5.1 ([Wu and Yang, 2019, Proposition 1]). *For any $\gamma, \gamma' \in \mathcal{G}_{k,1}$,*

$$W_1(\gamma, \gamma') \lesssim_k \max_{r \in [2k-1]} |m_r(\gamma) - m_r(\gamma')|^{1/(2k-1)}.$$

Lemma 2.5.2 (Hypercontractivity inequality [Schudy and Sviridenko, , Theorem 1.9]). *Let $Z \sim N(0, I_d)$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a polynomial of degree at most q. Then for any $t > 0$,*

$$\mathbb{P}\{|g(Z) - \mathbb{E}g(Z)| \geq t\} \leq e^2 \exp\left(-\left(\frac{t^2}{C\mathrm{Var}\, g(Z)}\right)^{1/q}\right),$$

where C is a universal constant.

Lemma 2.5.3. *Fix $r \in [2k-1]$. Let $f_r(\theta)$ be the process defined in (2.2.23). Let $\lambda > 0$. There are positive constants C, c_k such that, for any $\theta_1, \theta_2 \in S^{k-1}$,*

$$\mathbb{P}\{|f_r(\theta_1) - f_r(\theta_2)| \geq \|\theta_1 - \theta_2\|_2 \lambda\} \leq C \exp\left(-c_k \lambda^{2/r}\right). \tag{2.5.1}$$

$$\mathbb{P}\{|f_r(\theta_1)| \geq \lambda\} \leq C \exp\left(-c_k \lambda^{2/r}\right). \tag{2.5.2}$$

Proof. Define $\Delta \triangleq \sqrt{n}\left(\tilde{m}_r(\theta_1) - \tilde{m}_r(\theta_2)\right)$. Then, $f_r(\theta_1) - f_r(\theta_2) = \Delta - \mathbb{E}\Delta$. Recall that $\tilde{m}_r(\theta) = \frac{1}{n}\sum_{i=1}^n H_r(\theta^\top X_i)$, where $X_i = U_i + Z_i$ and $U_i \overset{\text{i.i.d.}}{\sim} \gamma$ and $Z_i \overset{\text{i.i.d.}}{\sim} N(0, I_k)$. Conditioning on $U = (U_1, \ldots, U_n)$, we have

$$\mathbb{E}(\Delta|U) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left((\theta_1^\top U_i)^r - (\theta_2^\top U_i)^r\right).$$

Now $|(\theta_1^\top U_i)^r - (\theta_2^\top U_i)^r| \leq rR^r \|\theta_1 - \theta_2\|_2$ since $\|U_i\|_2 \leq R$.

This is by the following argument. Note that $|\theta^\top U_i| \leq R$. Let $f : [-R, R] \to \mathbb{R}$ be defined via $f(x) = x^r$. Then use the univariate mean-value theorem to obtain the upper bound $|f(\theta_1^\top U_i) - f(\theta_2^\top U_i)| = f'(x)(\theta_1^\top U_i) - (\theta_2^\top U_i)$ where $x \in [\theta_1^\top U_i, \theta_2^\top U_i] \in [-R, R]$. Thus $|f(\theta_1^\top U_i) - f(\theta_2^\top U_i)| \leq \sup_{x \in [-R,R]} |f'(x)||\theta_1^\top U_i - \theta_2^\top U_i| \leq rR^{r-1} \|U_i\|_2 \|\theta_1 - \theta_2\|_2 \leq$

$rR^r \|\theta_1 - \theta_2\|_2$. If we wanted to use the multivariate mean-value theorem, we can, but be careful. Let B^d be the unit ball in d dimensions. We have $f : B^d \to \mathbb{R}$ defined via $f(\theta) = (\theta^\top U_i)^r$. We need to define f on the ball, not the sphere, because to apply the mean-value theorem, we must have a convex domain. Then $|f(\theta_1) - f(\theta_2)| \le \left\| \sup_{\|\theta\|_2 \le 1} |\nabla f(\theta)| \right\|_2 \|\theta_1 - \theta_2\|_2 = \sup_{\|\theta\|_2 \le 1} r|(\theta^\top U_i)^{r-1}| \, \|U_i\|_2 \, \|\theta_1 - \theta_2\|_2 \le rR^r \|\theta_1 - \theta_2\|_2$.

By Hoeffding's inequality,

$$\mathbb{P}\{|\mathbb{E}(\Delta|U) - \mathbb{E}\Delta| \ge \|\theta_1 - \theta_2\|_2 \lambda\} \le 2 \exp\left(-\frac{\lambda^2}{2r^2 R^{2r}}\right). \tag{2.5.3}$$

We now condition on U and analyze $|\Delta - \mathbb{E}(\Delta|U)|$. Since Δ is a polynomial of degree r in $Z_1, \ldots, Z_n$, by Lemma 2.5.2,

$$\mathbb{P}\{|\Delta - \mathbb{E}(\Delta|U)| \ge \|\theta_1 - \theta_2\|_2 \lambda\} \le e^2 \exp\left(-\left(\frac{\|\theta_1 - \theta_2\|_2^2}{C\mathrm{Var}(\Delta|U))}\right)^{1/r} \lambda^{2/r}\right). \tag{2.5.4}$$

It remains to upper-bound $\mathrm{Var}\,(\Delta|U)$. We have

$$\mathrm{Var}\,(\Delta|U) = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left(H_r(\theta_1^\top X_i) - H_r(\theta_2^\top X_i)|U_i\right).$$

Since the standard deviation of a sum is no more than the sum of the standard deviations,

$$\sqrt{\mathrm{Var}\left(H_r(\theta_1^\top X_i) - H_r(\theta_2^\top X_i)|U_i\right)} \le \sum_{j=0}^{\lfloor r/2 \rfloor} c_{j,r} \sqrt{\mathbb{E}\left(\left((\theta_1^\top X_i)^{r-2j} - (\theta_2^\top X_i)^{r-2j}\right)^2 |U_i\right)},$$

where $c_{j,r} = (-1/2)^r / j!(r-2j)!$. For any $\ell \le r$, we have $|(\theta_1^\top X)^\ell - (\theta_2^\top X)^\ell| \le \ell \|X\|_2^\ell \|\theta_1 - \theta_2\|_2$ and thus

$$\mathbb{E}\left(\left((\theta_1^\top X_i)^\ell - (\theta_2^\top X_i)^\ell\right)^2 |U_i\right) \le \ell^2 \|\theta_1 - \theta_2\|_2^2 \mathbb{E}(\|X_i\|_2^{2\ell}|U_i).$$

Since $\|X_i\|_2^{2\ell} \le 2^{2\ell-1}\left(\|Z_i\|_2^{2\ell} + \|U_i\|_2^{2\ell}\right)$ and $\mathbb{E}\,\|Z_i\|_2^{2\ell} \le (ck\ell)^\ell$ for a constant c,

$$\mathbb{E}\left(\left((\theta_1^\top X_i)^\ell - (\theta_2^\top X_i)^\ell\right)^2 |U_i\right) \le \|\theta_1 - \theta_2\|_2^2 \cdot \ell^2 2^{2\ell-1}\left(R^{2\ell} + (ck\ell)^\ell\right).$$

We conclude that $\mathrm{Var}\,(\Delta|U) \le C_k\,\|\theta_1 - \theta_2\|_2^2$. Therefore, (2.5.4) holds with probability $1 - e^2\exp(-c_k\lambda^{2/r})$. Then (2.5.1) follows from (2.5.3) and (2.5.4).

The second inequality, (2.5.2), can be proved by a similar application of Hoeffding's Inequality and Lemma 2.5.2. □

The following lemma is adapted from [Pollard, 2016, Section 4.7.1].

Lemma 2.5.4. *Let Θ be a finite subset of a metric space with metric ρ. Let $f(\theta)$ be a random process indexed by $\theta \in \Theta$. Suppose that for $\alpha > 0$, and for $\lambda > 0$, we have*

$$\mathbb{P}\{|f(\theta_1) - f(\theta_2)| \ge \rho(\theta_1,\theta_2)\lambda\} \le C_\alpha \exp\left(-c_\alpha\lambda^\alpha\right), \quad \forall\ \theta_1,\theta_2 \in \Theta. \tag{2.5.5}$$

Let $\theta_0 \in \Theta$ be a fixed point and $\epsilon_0 = \max_{x,y\in\Theta}\rho(x,y)$. Then there is a constant C'_α such that with probability $1 - C_\alpha\exp(-c_\alpha t^\alpha)$,

$$\max_{\theta\in\Theta}|f(\theta) - f(\theta_0)| \le C'_\alpha \int_0^{\epsilon_0/2}\left(t + \log^{\frac{1}{\alpha}}\frac{\epsilon_0|\mathcal{M}(r,\Theta,\rho)|}{r}\right)dr.$$

Proof. We construct an increasing sequence of approximating subsets by maximal packing. Let $\Theta_0 = \{\theta_0\}$. For $i = 0, 1, 2, \ldots,$ let Θ_{i+1} be a maximal subset of Θ containing Θ_i that constitutes an ϵ_{i+1}-packing, where $\epsilon_{i+1} = \epsilon_i/2$. Since Θ is finite, the procedure stops after a finite number of iterations, resulting in $\Theta_0 \subseteq \Theta_1 \subseteq \cdots \subseteq \Theta_m = \Theta$. By definition,

$$N_i \triangleq |\Theta_i| \le M(\epsilon_i,\Theta,\rho).$$

For $i = 0,\ldots,m-1$, define a sequence of mappings $\ell_i : \Theta_{i+1} \to \Theta_i$ by $\ell_i(t) \triangleq$

$\arg\min_{s\in\Theta_i}\rho(t,s)$ (with ties broken arbitrarily). Now

$$\begin{aligned}
\mathbb{P}\left\{\max_{\theta\in\Theta}|f(\theta)-f(L_0(\theta))|\geq\sum_{i=0}^{m-1}\epsilon_i\lambda_i\right\} &\leq \mathbb{P}\left\{\sum_{i=0}^{m-1}\max_{s\in\Theta_{i+1}}|f(s)-f(\ell_i(s))|\geq\sum_{i=0}^{m-1}\epsilon_i\lambda_i\right\}\\
&\leq \mathbb{P}\left\{\sum_{i=0}^{m-1}\max_{s\in\Theta_{i+1}}|f(s)-f(\ell_i(s))|\geq\sum_{i=0}^{m-1}\max_{s\in\Theta_{i+1}}\rho(s,\ell_i(s))\lambda_i\right\}\\
&\leq \sum_{i=0}^{m-1}\mathbb{P}\left\{\max_{s\in\Theta_{i+1}}|f(s)-f(\ell_i(s))|\geq\epsilon_i\lambda_i\right\}\\
&\leq \sum_{i=0}^{m-1}N_{i+1}\exp\left(-c\lambda_i^{\alpha}\right).
\end{aligned}$$

The first inequality is just by the triangle inequality. The second inequality follows from a union bound on the sum. The third inequality is because Θ_i is a maximal ϵ_i-packing, we have $\rho(t,\ell_i(t))\leq\epsilon_i$ for all $t\in\Theta_{i+1}$. The fourth inequality then follows from the assumption (2.5.5). Thus with probability at least $1-\sum_{i=0}^{m-1}N_{i+1}C_\alpha\exp(-c_\alpha\lambda_i^\alpha)$,

$$\max_{\theta\in\Theta}|f(\theta)-f(\theta_0)|\leq\sum_{i=0}^{m-1}\lambda_i\epsilon_i. \tag{2.5.6}$$

Set $\lambda_i=(t^\alpha+\frac{1}{c_\alpha}\log(2^{i+1}N_{i+1}))^{1/\alpha}$. Note that $\epsilon_{i+1}=\epsilon_0 2^{-(i+1)}$. Then $\lambda_i\leq C''_\alpha G(\epsilon_{i+1})$ for a constant C''_α, where $G(r)\triangleq t+(\log(\epsilon_0 M(r,\Theta,\rho)/r))^{1/\alpha}$ is a decreasing function for $r\leq\epsilon_0$. By (2.5.6), there is another constant C'_α such that with probability $1-C_\alpha\exp(-c_\alpha t^\alpha)$,

$$\max_{\theta\in\Theta}|f(\theta)-f(\theta_0)|\leq C'_\alpha\sum_{i=0}^{m-1}G(\epsilon_{i+1})\epsilon_i\leq C'_\alpha\int_{\epsilon_{m+1}}^{\epsilon_1}4G(r)dr\leq 4C'_\alpha\int_0^{\epsilon_1}G(r)dr.\quad\square$$

2.6 Alternative methods

2.6.1 Alternative proofs

In this section, we provide alternative proofs for some of the lemmas used in this work. We use the notation $A \gtrsim B$ to mean $x'Ax \geq x'Bx$ for any $x \in \mathbb{R}^d$.

Alternative proof of Lemma 2.2.5. Let the k subscript denote the k-rank approximation. Let $\hat{H} = \hat{V}_k \hat{V}_k^\top$. Note that $\hat{H}\hat{\Sigma} = \hat{\Sigma}_k$. First,

$$\begin{aligned}\left\|\hat{H}\Sigma - \Sigma\right\|_F &\leq \left\|\hat{H}\Sigma - \hat{\Sigma}_k\right\|_F + \left\|\hat{\Sigma}_k - \Sigma\right\|_F \\ &\precsim_k \left\|\hat{H}\Sigma - \hat{\Sigma}_k\right\|_2 + \left\|\hat{\Sigma} - \Sigma\right\|_2 \quad \text{by Lemma 3.5.8 and since } \hat{\Sigma}, \hat{\Sigma}_k \text{ have rank } k \\ &\precsim_k \left\|\hat{\Sigma} - \Sigma\right\|_2.\end{aligned}$$

(We could also do the above step by going directly to spectral norm in step 1.) So by Lemma 2.2.6, with high probability,

$$\left\|\hat{H}\Sigma - \Sigma\right\|_F \precsim_k (kd/n)^{1/2}.$$

Let $\hat{P} = \hat{H} - I_d$, which is the orthogonal projection operator onto the space orthogonal to that spanned by the vectors in $\hat{V}_k$. We have shown that $\left\|\hat{P}\Sigma\right\|_F \precsim_k (d/n)^{1/2}$. And $\left\|\hat{P}\Sigma\hat{P}\right\|_F \leq \left\|\hat{P}\Sigma\right\|_F \left\|\hat{P}\right\|_F \precsim_k \left\|\hat{P}\Sigma\right\|_F$ since $\left\|\hat{P}\right\|_F = \sqrt{k}$. Now for any $j \in [k]$ and $x \in S^{d-1}$, $x^\top \left(\hat{P}\Sigma\hat{P}\right) x \geq x^\top \left(\hat{P} w_j \mu_j \mu_j^\top \hat{P}\right) x$. So we can apply Lemma 2.6.1 to say

$$w_j \left\|\hat{P}\mu_j\mu_j^\top\hat{P}\right\|_F \leq \left\|\hat{P}\Sigma\hat{P}\right\|_F,$$

which implies the result. □

Lemma 2.6.1. *Let $A, B \in \mathbb{R}^{d\times d}$ be symmetric, positive semi-definite matrices with*

$A \gtrsim B$. *Then*

$$\|A\|_F \geq \|B\|_F .$$

Proof. By definition of the Frobenius norm, we must show that $tr(A^2) \geq tr(B^2)$. Since $A - B \gtrsim 0$, we have $(A-B)^2 \gtrsim 0$, and thus

$$\begin{aligned} tr\left(A^2 + B^2\right) &\geq tr(AB) + tr(BA) \\ tr\left(A^2 + B^2\right) &\geq 2tr(AB), \end{aligned}$$

because $tr(AB) = tr(BA)$ since A, B are symmetric. Subtracting $2tr(B^2)$ from both sides yields:

$$\begin{aligned} &tr\left(A^2 - B^2\right) \geq 2tr(AB) - 2tr(B^2) \Rightarrow \\ &tr\left(A^2 - B^2\right) \geq 2tr\left((A-B)B\right) \Rightarrow \\ &tr\left(A^2 - B^2\right) \geq 2tr\left((A-B)^{1/2}(A-B)^{1/2}B\right) \Rightarrow \\ &tr\left(A^2 - B^2\right) \geq 2tr\left((A-B)^{1/2}B(A-B)^{1/2}\right), \end{aligned}$$

where the last step again follows because $(A-B)^{1/2}$ and B are symmetric. Now $A-B$ is symmetric and positive semi-definite, and B is symmetric and positive semi-definite. Thus the right-hand side is nonnegative, so the desired conclusion follows. □

Lemma 2.6.2. *Let $A \in \mathbb{R}^{n\times p}$ and let A_k be the rank-k approximation of A. Let $P \in \mathbb{R}^{n\times p}$ be rank k. Then*

$$\|A_k - P\|_F^2 \leq 8k \|A - P\|_{OP}^2 .$$

Proof. Since A_k and P are rank k, $A_k - P$ has rank at most $2k$. And for any rank $2k$

matrix B, $\|B\|_F^2 \leq 2k \|B\|_2^2$. So,

$$\begin{aligned}\|A_k - P\|_F^2 &\leq 2k \|A_k - P\|_2^2 \\ &\leq 4k \|A_k - A\|_2^2 + 4k \|A - P\|_2^2 \\ &\leq 8k \|A - P\|_2^2\end{aligned}$$

The third step follows because $\|A - A_k\|_F^2 \leq \|A - P\|_F^2$ for any rank-k matrix P by definition of A_k. □

Another proof of Lemma 3.5.8. The rank-k approximation of A minimizes $\|A_k - A\|_F$ over rank k matrices, so $\|A_k - A\|_F^2 \leq \|P - A\|_F^2$. Adding and subtract P in $\|A_k - A\|_F^2$ yields

$$\|A_k - P\|_F^2 + \|P - A\|_F^2 + 2\langle A_k - P,\, P - A\rangle \leq \|A - P\|_F^2$$

which then yields:

$$\|A_k - P\|_F^2 \leq 2\langle A_k - P,\, A - P\rangle \tag{2.6.1}$$

So

$$\begin{aligned}\|A_k - P\|_F &\leq \sup_{\|U\|_F = 1, rank(U) = k} \langle U,\, Z\rangle \\ &\leq \sqrt{k}\, \|Z\|_2\end{aligned}$$

□

Lemma 2.6.3 is closely related to Lemma 2.3.4.

Lemma 2.6.3. *Let $\mathbf{j} = (j_1, \ldots, j_d) \in \mathbb{Z}_+^d$, and let Γ, Γ' be k-atomic distributions on $\mathbb{R}^d$ with atoms spanning a space of rank r. Then there is a $\theta \in S^{d-1}$ and a constant*

$c_{\mathbf{j}}$ *such that*

$$\Delta m_{\mathbf{j}}(\Gamma, \Gamma') = \Delta m_{|\mathbf{j}|}(\Gamma_{\theta_{\mathbf{j}}}, \Gamma'_{\theta_{\mathbf{j}}}).$$

And $c_{\mathbf{j}} \leq r^{(|\mathbf{j}|-1)/2}$.

Proof. Let $U \in \mathbb{R}^d$. There is a $\theta_{\mathbf{j}} \in S^{d-1}$ and a constant $c_{\mathbf{j}}$ such that $U_1^{j_1} \dots U_d^{j_d} = c_{\mathbf{j}}(\theta_{\mathbf{j}}^\top U)^{\mathbf{j}}$. Thus

$$\Delta m_{\mathbf{j}}(\Gamma, \Gamma') = c_{\mathbf{j}} \Delta m_{|\mathbf{j}|}(\Gamma_{\theta_{\mathbf{j}}}, \Gamma'_{\theta_{\mathbf{j}}}).$$

And the upper bound on $c_{\mathbf{j}}$ holds because of (2.3.10). □

Alternative proof of Theorem 2.3.3. Let $V \in S^{d\times k}$ have columns that form an orthonormal basis of the space spanned by the atoms of Γ, Γ'. Let r be the dimension of this space, and note that $r \leq 2k$. To simplify notation, let $\gamma = \Gamma_V, \gamma' = \Gamma'_V$. Now

$$\begin{aligned}
\chi^2\left(P_\Gamma || P_{\Gamma'}\right) &\leq e^{R^2/2} \sum_{\mathbf{j}\in Z_+^r} \frac{\left(\Delta m_{\mathbf{j}}(\gamma, \gamma')\right)^2}{\mathbf{j}!} && (2.6.2)\\
&= e^{R^2/2} \sum_{\mathbf{j}} \frac{c_{\mathbf{j}}^2 \left(\Delta m_{|\mathbf{j}|}(\gamma_{\theta_{\mathbf{j}}}, \gamma'_{\theta_{\mathbf{j}}})\right)^2}{\mathbf{j}!} && (2.6.3)\\
&\leq e^{R^2/2} \sum_{\mathbf{j}} \frac{c_{\mathbf{j}}^2 \max_{\ell\in[2k-1]} \left(3^{|\mathbf{j}|} \Delta m_\ell(\gamma_{\theta_{\mathbf{j}}}, \gamma'_{\theta_{\mathbf{j}}})\right)^2}{\mathbf{j}!} && (2.6.4)\\
&\leq \max_{\ell\in[2k-1]} \sup_{\theta\in S^{d-1}} \left(\Delta m_\ell(\Gamma_\theta, \Gamma'_\theta)\right)^2 e^{R^2/2} \sum_{\mathbf{j}} \frac{c_{\mathbf{j}}^2 9^{|\mathbf{j}|}}{\mathbf{j}!},
\end{aligned}$$

where (2.6.2) is by Lemma 2.3.2, (2.6.3) is by Lemma 2.6.3, and (2.6.4) is by Lemma 10 of [Wu and Yang, 2019]. And

$$\sum_{\mathbf{j}} \frac{c_{\mathbf{j}}^2 9^{|\mathbf{j}|}}{\mathbf{j}!} \leq \sum_{\mathbf{j}} \frac{r^{|\mathbf{j}|-1} 9^{|\mathbf{j}|}}{\mathbf{j}!} = \left(\sum_{j\geq 1} \frac{r^{j-1} 9^j}{j!}\right)^r,$$

which is a constant depending on r only. $\square$

2.6.2 Failure of the MLE in weights selection

We now provide more discussion on the intuition behind Algorithm 1. Suppose $k = 2$ and the locations do not lie on a single line in $\mathbb{R}^d$. Via the DMM algorithm, we estimate the two one-dimensional distributions $\tilde{\gamma}^{(e_1)}, \tilde{\gamma}^{(e_2)}$ with corresponding atoms $\left(\tilde{\psi}_1^{(e_1)}, \tilde{\psi}_2^{(e_1)}\right)$ and $\left(\tilde{\psi}_1^{(e_2)}, \tilde{\psi}_2^{(e_2)}\right)$. Now how do we combine these atoms into two-dimensional estimates? And what weights should we match with them?

One approach is to simply select one of the k estimated sets of weights as our estimate. But this will not work for two reasons. One is the identifiability issue mentioned above. The second is that we do not have a guarantee on each individual weight. For instance, suppose that on one coordinate, one of the locations has norm smaller than $\epsilon_{n,k}$. Then the DMM algorithm can easily estimate an associated weight that is very far from the truth.

Another naive approach to the above-mentioned problem is to consider every combination of centers into vectors, pair each combination with a set of weights via maximum likelihood, and select the overall best estimator again via maximum likelihood. Alternatively, we might simply pair sets of locations with potential weights from a discretization of the simplex, then select the best of these via maximum likelihood.

To obtain guarantees on a maximum likelihood procedure, we will need obtain an upper bound of $1/n$ on the KL divergence between the mixture densities. But Example 1 shows that there exists a model where one-dimensional estimates cannot be used in any way that would obtain the correct bound on the KL divergence. Note that this does not necessarily mean that this type of MLE procedure for selecting the weights fails, only that we do not have the tools to prove it works.

Example 1. Let $k = 2$ and suppose we have k-atomic distributions in $\mathbb{R}^k$. Let

$\epsilon = n^{-1/6}$. Define

$$\gamma = \frac{2}{3}\delta_{(\epsilon,\epsilon)} + \frac{1}{3}\delta_{(-2\epsilon,-2\epsilon)}.$$

The one-dimensional, coordinate-wise marginal distributions are

$$\gamma_1 = \frac{2}{3}\delta_{\epsilon} + \frac{1}{3}\delta_{-2\epsilon}.$$
$$\gamma_2 = \frac{2}{3}\delta_{\epsilon} + \frac{1}{3}\delta_{-2\epsilon}.$$

Consider the following estimates of each one-dimensional marginal distribution.

$$\hat{\gamma}_1 = \frac{2}{3}\delta_{\epsilon} + \frac{1}{3}\delta_{-2\epsilon}.$$
$$\hat{\gamma}_2 = \frac{2}{3}\delta_{-\epsilon} + \frac{1}{3}\delta_{2\epsilon}.$$

Now $m_r(\hat{\gamma}_1) = m_r(\gamma_1)$ for all r. And $\max_{r\in[2k-1]} |m_r(\hat{\gamma}_2) - m_r(\gamma_2)| = m_3(\hat{\gamma}_2) - m_3(\gamma_2)| \precsim \epsilon^3$. By Lemma 2.5.1, for each one-dimensional coordinate $j \in [k]$, $W_1(\hat{\gamma}_j, \gamma_j) \leq C\epsilon$ for a constant C. So $\hat{\gamma}_1, \hat{\gamma}_2$ are estimates that could be produced by the DMM algorithm.

Let $\mathcal{S}_c$ be as defined as in Algorithm 1. In this example,

$$\mathcal{S}_c = \{(\epsilon, -\epsilon)^\top, (\epsilon, 2\epsilon)^\top, (-2\epsilon, -\epsilon)^\top, (-2\epsilon, 2\epsilon)^\top\}.$$

Define

$$\hat{\gamma} = \sum_{j=1}^{4} w_j \delta_{\mu_j},$$

where $\mu_1, \ldots, \mu_4$ are the vectors in $\mathcal{S}_c$ and $w = (w_1, \ldots, w_4) \in \Delta^3$. We might pick some of the weights to be zero so we have a 2-atomic distribution. But it turns out that no matter which weights we select (even if we use maximum likelihood to select

them), we can show that in this case, $KL(P_\gamma||P_{\gamma'}) \geq \epsilon^4$. For a formal justification, see Lemma 2.6.4.

Lemma 2.6.4. *Let $\hat{\gamma}, \gamma, \epsilon$ be as defined in Example 1. Then there is a positive constant C such that*

$$KL(P_{\hat{\gamma}}||P_\gamma) \geq C\epsilon^4.$$

Proof. By the data processing inequality, $KL(P_{\hat{\gamma}}||P_\gamma) \geq \sup_{\theta \in S^{k-1}} KL(\hat{\gamma}_\theta * N(0,1)||\gamma_\theta * N(0,1))$. Let $\theta = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)^T$. Note that $\gamma_\theta = \delta_0$, so $\gamma_\theta * N(0,1)$ is just the $N(0,1)$ distribution. Define the support set obtained by taking the inner product of θ with each point in $\mathcal{S}_c$ via:

$$\theta^\top \mathcal{S}_c \triangleq \left\{\sqrt{2}\epsilon, -\frac{\epsilon}{\sqrt{2}}, -\frac{4\epsilon}{\sqrt{2}}, \right\}.$$

Then $\hat{\gamma}_\theta = w_1\delta_{\theta^\top \mu_1} + (w_2 + w_3)\delta_{\theta^\top \mu_2} + w_4\delta_{\theta^\top \mu_4}$. Choosing $f(x) = tx^2$ for a constant $t < 1/2$, and using the Donsker-Varadhan representation of the KL-divergence,

$$KL\left(P_{\hat{\gamma}_\theta}||P_{\gamma_\theta}\right) \geq \mathbb{E}_{X \sim P_{\hat{\gamma}_\theta}} tX^2 - \log \mathbb{E}_{X \sim N(0,1)} e^{tX^2} = \mathbb{E}_{X \sim P_{\hat{\gamma}_\theta}} tX^2 + \frac{1}{2}\log(1-2t).$$

Now regardless of which weights w were chosen, $\mathbb{E}_{X \sim P_{\hat{\gamma}_\theta}} tX^2 \propto Ct\epsilon^2 + t$ for a constant C. Let $t = C\epsilon^2/2(C\epsilon^2 + 1)$. This yields a lower bound proportional to ϵ^4. □

2.6.3 Symmetric 2-GM

In this section, we explore a simple, classical example of the model (2.1.1): the symmetric, two-component Gaussian mixture model. Since this model is a subset of the more general model (2.1.1), all results from the previous sections hold. However, in many cases, the estimation procedures, theorems, and proofs can be simplified in the symmetric 2-GM. We give these simplifications here because they illuminate much

of the fundamental ideas from the previous sections, and they also tell us something about a model we will consider in the next chapter, the k-spaces model.

Let

$$P_\mu = \Gamma * N(0, I_d) \text{ where } \Gamma = \frac{1}{2}\delta_\mu + \frac{1}{2}\delta_{-\mu}. \tag{2.6.5}$$

Let $\mathcal{G}_d = \{\Gamma = \frac{1}{2}\delta_\mu + \frac{1}{2}\delta_{-\mu} : \mu \in \mathbb{R}^d, \|\mu\|_2 \leq R\}$. For this model, we will use the loss

$$\ell(\mu, \mu') = \min\left(\|\mu - \mu'\|_2, \|\mu + \mu'\|_2\right), \tag{2.6.6}$$

which is in fact equal to $W_1(\Gamma_\mu, \Gamma_{\mu'})$, as discussed previously. For the model (2.6.5), spectral algorithms achieve the sharp rate for both parameter estimation and density estimation. Let $\hat{\mu}\hat{\mu}^\top$ be the rank-1 approximation of $\hat{\Sigma}$ in (2.2.1). Then it is easy to show that

$$\left\|\hat{\mu}\hat{\mu}^\top - \mu\mu^\top\right\|_2 \leq (d/n)^{1/2}, \tag{2.6.7}$$

with high probability; see for instance, Lemma 3.5.8 and Lemma 2.2.6. We now present a lemma that shows we can directly our error on $\hat{\mu}$ to the covariance gap in (2.6.7).

Lemma 2.6.5. *Let $\mu \in \mathbb{R}^d$ and let $\|\mu\|_2 \leq R$. Let $\hat{\mu} \in \mathbb{R}^d$ satisfy $\left\|\hat{\mu}\hat{\mu}^\top - \mu\mu^\top\right\|_2 \leq \epsilon$. Then*

$$\ell\left(\hat{\mu}, \mu\right) \leq \sqrt{2\epsilon}. \tag{2.6.8}$$

Lemma 2.6.5 shows that the bound in (2.6.7) translates direclty to a bound on $\ell(\hat{\mu}, \mu)$; when μ is arbitrary, a square root factor is lost, but when $\mu \in S^{d-1}$, the rate of estimation is the same for the covariance matrix $\hat{\mu}\hat{\mu}^\top$ and $\hat{\mu}$. This is intuitive, since if we allow $\|\mu\|_2$ to be small, we expect that it is more difficult to estimate its associated direction. (2.6.8) translates to the $(d/n)^{1/4}$ rate on $\ell(\hat{\mu}, \mu)$, i.e., we obtain a parameter

estimator directly from a simple spectral algorithm. Moreover, by Theorem 2.3.3, since both the first and third moment tensors are zero by symmetry, we have

$$H(P_\mu, P_{\mu'})^2 \asymp \left\|\mu\mu^\top - \mu'\mu'^\top\right\|_{\mathrm{F}}^2. \tag{2.6.9}$$

Moreover, in this simple model we are able to obtain a maximum likelihood result for both density estimation and parameter estimation.

Lemma 2.6.6. *Let* $\hat{\mu} = \operatorname{argmax}_{\|\mu\|_2 \le R} \sum_{i=1}^n \log p_\mu(x_i)$. *Then with probability at least* $1-\delta$,

$$\ell(\hat{\mu}, \mu) \le \left(\frac{d\log(1/\delta)}{n}\right)^{1/4}. \tag{2.6.10}$$

Proof. This follows from Lemma 2.6.5 and (2.7.10). □

We now prove Lemma 2.6.5. We moreover provide an alternative proof of (2.6.9). This is a somewhat simpler version of Theorem 2.3.3 that applies to the symmetric 2-GM only.

Proof of Lemma 2.6.5. We first prove (2.6.8). Without loss of generality, let $\|\mu\|_2 \ge \|\hat{\mu}\|_2$. Now $\left\|\hat{\mu}\hat{\mu}^\top - \mu\mu^\top\right\|_2 \le \epsilon$ implies

$$\left|\frac{\langle\hat{\mu},\, \mu\rangle^2}{\|\mu\|_2^2} - \|\mu\|_2^2\right| \le \epsilon. \tag{2.6.11}$$

Now

$$\begin{aligned}
\|\mu\|_2^2 - \epsilon &\le \frac{\langle\hat{\mu},\, \mu\rangle^2}{\|\mu\|_2^2} && \text{by (2.6.11) and the triangle inequality} \\
&\le \frac{|\langle\hat{\mu},\, \mu\rangle|}{\|\mu\|_2^2}\,\|\hat{\mu}\|_2\,\|\mu\|_2 && \text{by Cauchy-Schwarz} \\
&\le |\langle\hat{\mu},\, \mu\rangle| && \text{since } \|\hat{\mu}\|_2 \le \|\mu\|_2.
\end{aligned}$$

And since $\|\hat{\mu}\|_2 \leq \|\mu\|_2$, we also obtain $\|\hat{\mu}\|_2^2 - \epsilon \leq |\langle \hat{\mu}, \mu \rangle|$. Thus

$$\|\hat{\mu}\|_2^2 + \|\mu\|_2^2 - 2\epsilon \leq 2|\langle \hat{\mu}, \mu \rangle|,$$

which implies the result. □

Lemma 2.6.7. *Let $\Gamma, \Gamma_* \in \mathcal{G}_d$. Then for any $D \in \{H^2, \mathsf{KL}, \chi^2\}$,*

$$D(P_\Gamma, P_{\Gamma_*}) \asymp_k \left\| \mu\mu^\top - \mu_*\mu_*^\top \right\|_2.$$

Proof. The lower bound is immediate from Lemma 2.3.5. For the upper bound, let $f(x, \mu) = \frac{1}{2}e^{\langle x, \mu \rangle - \|\mu\|_2^2/2} + \frac{1}{2}e^{\langle x, -\mu \rangle - \|\mu\|_2^2/2}$. By Lemma 2.3.10, it suffices to show that

$$\int_{x \in \mathbb{R}^d} \phi_d(x) \left(f(x, \mu) - f(x, \mu_*) \right)^2 \leq C \left\| \mu\mu^\top - \mu_*\mu_*^\top \right\|_2^2.$$

Let H be the orthogonal projection operator onto the space spanned by μ, μ_*, and let $V \in S_H$. Let $\tilde{\mu} = V^\top \mu, \tilde{\mu}_* = V^\top \mu_*$. By the rotation invariance of the Gaussian distribution, we reduce ourselves to 2-dimensional space, i.e., $\int_{x \in \mathbb{R}^d} \phi_d(x) \left(f(x, \mu) - f(x, \mu_*) \right)^2 = \int_{x \in \mathbb{R}^2} \phi_2(x) \left(f(x, \tilde{\mu}) - f(x, \tilde{\mu}_*) \right)^2 dx$. Now, let $\Gamma = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$. Then we can write $f(x, \tilde{\mu}) = \mathbb{E}_{y \sim \Gamma} e^{\langle x, y\tilde{\mu} \rangle - \|\tilde{\mu}\|_2^2/2}$. Throughout the rest of the proof, $x \in \mathbb{R}^2$. Now

$$f(x, \tilde{\mu}) - f(x, \tilde{\mu}_*) = e^{-\|\tilde{\mu}\|_2^2/2} \mathbb{E}_{y \sim \Gamma} \left(e^{\langle x, y\tilde{\mu} \rangle} - e^{\langle x, y\tilde{\mu}_* \rangle} \right) + \mathbb{E}_{y \sim \Gamma} e^{\langle x, y\tilde{\mu}_* \rangle} \left(e^{-\|\tilde{\mu}\|_2^2/2} - e^{-\|\tilde{\mu}_*\|_2^2/2} \right).$$

For the first term, since Γ is a symmetric distribution, its odd moments are zero. So,

$$\begin{aligned}
\mathbb{E}_{y\sim\Gamma}\left(e^{\langle x,\, y\tilde{\mu}\rangle} - e^{\langle x,\, y\tilde{\mu}_*\rangle}\right) &= \sum_{j\geq 1} \frac{\langle x,\, \tilde{\mu}\rangle^{2j} - \langle x,\, \tilde{\mu}_*\rangle^{2j}}{j!} \\
&\leq \left|\langle x,\, \tilde{\mu}\rangle^2 - \langle x,\, \tilde{\mu}_*\rangle^2\right| \sum_{j\geq 1} \frac{j\left(\max\left(\langle x,\, \tilde{\mu}\rangle^2, \langle x,\, \tilde{\mu}_*\rangle^2\right)\right)^{j-1}}{j!} \\
&\leq \|x\|_2^2 \left\|\mu\mu^\top - \mu_*\mu_*^\top\right\|_2 \sum_{j\geq 0} \frac{\|x\|_2^{2j} R^j}{j!} \qquad (2.6.12) \\
&= \left\|\mu\mu^\top - \mu_*\mu_*^\top\right\|_2 \|x\|_2^2 e^{\|x\|_2^2 R},
\end{aligned}$$

where (2.6.12) follows because $\left\|\tilde{\mu}\tilde{\mu}^\top - \tilde{\mu}_*\tilde{\mu}_*^\top\right\|_2 = \left\|\mu\mu^\top - \mu_*\mu_*^\top\right\|_2$. Thus

$$\begin{aligned}
\int_{x\in\mathbb{R}^2} \phi_2(x)\left(f(x,\mu) - f(x,\mu_*)\right)^2 dx &\leq \left\|\mu\mu^\top - \mu_*\mu_*^\top\right\|_2^2 \int_{x\in\mathbb{R}^2} \|x\|_2^4 e^{2\|x\|_2^2 R} dx \\
&\leq C\left\|\mu\mu^\top - \mu_*\mu_*^\top\right\|_2^2.
\end{aligned}$$

And for the second term, since $e^{\langle x,\, y\tilde{\mu}_*\rangle} > 0$,

$$\begin{aligned}
\mathbb{E}_{y\sim\Gamma} e^{\langle x,\, y\tilde{\mu}_*\rangle}\left(e^{-\|\tilde{\mu}\|_2^2/2} - e^{-\|\tilde{\mu}_*\|_2^2/2}\right) &\leq \mathbb{E}_{y\sim\Gamma} e^{\langle x,\, y\tilde{\mu}_*\rangle} |e^{-\|\tilde{\mu}\|_2^2/2} - e^{-\|\tilde{\mu}_*\|_2^2/2}| \\
&\leq \mathbb{E}_{y\sim\Gamma} e^{\langle x,\, y\tilde{\mu}_*\rangle} |\,\|\tilde{\mu}_*\|_2^2 - \|\tilde{\mu}\|_2^2\,| \\
&\leq \mathbb{E}_{y\sim\Gamma} e^{\langle x,\, y\tilde{\mu}_*\rangle} \left\|\mu\mu^\top - \mu_*\mu_*^\top\right\|_2.
\end{aligned}$$

And $\int \phi_2(x)\left(\mathbb{E}_{y\sim\Gamma} e^{\langle x,\, y\tilde{\mu}_*\rangle}\right)^2 dx \leq C$. □

2.6.4 Discussion

We have focused on the Gaussian location mixture model (2.1.1) in high dimensions, where the variance parameter σ^2 and the number of components k are known, and the centers lie in a ball of bounded radius. Below we discuss weakening these assumptions and other open problems.

Unbounded centers While the assumption of bounded support is necessary for estimating the mixing distribution (otherwise the worst-case W_1-loss is infinity), it is not needed for density estimation [Acharya et al., 2014, Li and Schmidt, 2017, Ashtiani et al., 2018]. In fact, [Acharya et al., 2014] first uses a crude clustering procedure to partition the samples into clusters whose means are close to each other, then zooms into each cluster to perform density estimation. For the lower bound, the worst case occurs when each cluster is equally weighted and highly separated, so that the effective sample size for each component is n/k, leading to the lower bound of $\Omega(\frac{kd}{n})$. Finally, the results of NPMLE in [Ghosal and van der Vaart, 2001, Zhang, 2009, Saha and Guntuboyina, 2017] do not impose bounded assumptions, which is partly responsible for the logarithmic factors in the obtained rates.

Location-scale mixtures We have assumed that the covariance of our mixture is known and common across components. There is a large body of work studying general location-scale Gaussian mixtures, see, e.g., [Moitra and Valiant, 2010, Ho and Nguyen, 2016, Ashtiani et al., 2018]. The introduction of the scale parameters makes the problem significantly more difficult. For parameter estimation, if all clusters share the same unknown scale parameter, the optimal rate is shown to be $n^{-1/(4k)}$ in [Wu and Yang, 2019]; otherwise, the optimal rate remains unknown even in one dimension except for $k = 2$ [Hardt and Price, 2015].

Number of components This work assumes that the parameter k is known and fixed. Since the centers are allowed to overlap arbitrarily, k is effectively an upper bound on the number of components. If k is allowed to depend on n, the optimal W_1-rate is shown in [Wu and Yang, 2019] to be $\Theta(n^{-1/(4k-2)})$ and $\Theta(\frac{\log\log n}{\log n})$ for $k = O(\frac{\log n}{\log\log n})$ and $k = \Omega(\frac{\log n}{\log\log n})$, respectively. Extending this result to the high-dimensional setting of Theorem 2.1.1 is an interesting future direction.

The problem of selecting the mixture order k has been extensively studied. For in-

stance, many authors have considered likelihood-ratio based tests; however, standard asymptotics for such tests may not hold [Hartigan, 1985]. Various workarounds have been considered, including tests inspired by the EM algorithm [Li and Chen, 2010] and quadratic approximation of the log-likelihood ratio [Liu and Shao, 2003], as well as method of moments [Lindsay, 1989], [Dacunha-Castelle and Gassiat, 1997].

Connection to spectral clustering Another major strain of the Gaussian mixture literature is on clustering. As previously noted, under our weak assumptions, clustering is not possible. But we here briefly discuss the connection between the parameter estimation technique used to obtain Theorem 2.1.1 and the classical spectral clustering technique studied in, for instance, [Vempala and Wang, 2004, Kannan et al., 2005, Achlioptas and Mcsherry, 2010, Löffler et al.,]. In spectral clustering, data are first projected to a best subspace, and then a clustering algorithm such as k-means is performed on the projected data. The best subspace in these papers is the same as the one estimated via PCA, and the clustering error analysis depends on the error in the subspace estimation and the error of the clustering method performed on the low-dimensional data, analogously to the analysis of mixing distribution estimation in Theorem 2.1.1. Our specific statement on the rate of estimating the subspace, Lemma 2.2.5, is closely connected to Theorem 3 of [Vempala and Wang, 2004].

Adaptivity The rate in Theorem 2.1.1 is optimal in the worst-case scenario where the centers of the Gaussian mixture can overlap. To go beyond this pessimistic result, in one dimension, [Heinrich and Kahn, 2018] showed that when the atoms of Γ form k_0 well-separated (by a constant) clusters (see [Wu and Yang, 2019, Definition 1] for a precise definition), the optimal rate is $n^{-1/(4(k-k_0)+2)}$, interpolating the rate $n^{-1/(4k-2)}$ in the worst case ($k_0 = 1$) and the parametric rate $n^{-1/2}$ in the best case ($k_0 = k$). Furthermore, this can be achieved adaptively by either the minimum distance estimator [Heinrich and Kahn, 2018, Theorem 3.3] or the DMM algorithm

[Wu and Yang, 2019, Theorem 2].

In the high-dimensional case, by Lemma 2.2.5, the optimal projection $\hat{H}$ preserves separation of the atoms of Γ. Letting $\hat{H}$ be the projection to the optimal subspace of Algorithm 1,

$$\left\|\hat{V}^\top \mu_i - \hat{V}^\top \mu_{i'}\right\|_2 = \left\|\hat{H}\mu_i - \hat{H}\mu_{i'}\right\|_2 \geq \|\mu_i - \mu_{i'}\|_2 - \frac{\left\|\mu_i - \hat{H}\mu_i\right\|_2}{w_i} - \frac{\left\|\mu_{i'} - \hat{H}\mu_{i'}\right\|_2}{w_{i'}}.$$

Let $\lambda = C \max\left((d/n)^{1/4}, n^{-1/(4k-2)}\right)$ for an appropriate positive constant C. And replace weights lower bound assumption of [Wu and Yang, 2019, Definition 1] with

$$w_i = c_k \text{ for all } i \in S_\ell, S_{\ell'}. \tag{2.6.13}$$

Then assuming the location separation condition of [Wu and Yang, 2019, Definition 1] and (2.6.13) hold, we have

$$\left\|\hat{V}^\top \mu_i - \hat{V}^\top \mu_{i'}\right\|_2 \gtrsim \lambda.$$

Therefore, if Γ has k_0 clusters, projecting the atoms of Γ onto the optimal subspace maintains the k_0 clusters. If the dimension of $\hat{V}$ is one, and we simply perform DMM on the projected data $\{V^\top X_i\}_{i\in[n]}$, we will be adaptive to separation on the model locations; rates like that of Theorem 2 of [Wu and Yang, 2019] will hold for the low-dimensional mixing estimation component of the rate in Theorem 2.1.1.

Two complications occur if we try to generalize this to general k-component mixtures. First, Algorithm 1 is not adaptive in general due to the coarseness of the grid search. This can be easily remedied by using an ϵ_n net with $\epsilon_n = n^{-1/2}$ rather than $n^{-1/(4k-2)}$ in Algorithm 1. This results in an algorithm of time $n^{O(k)}$, as discussed in Section 2.2. Alternatively, by the same reasoning as in the proof of Theorem 2.3.8, the density estimator of Theorem 2.1.2 should be able to provide a k-GM that is

adaptive to stronger assumptions.

However, even if we were to use a finer grid search, the analysis of that algorithm depends on the sliced W_1 upper bound of Lemma 2.2.1. In the bound $W_1(\Gamma, \Gamma') \leq C_k \sup_{\theta \in S^{k-1}} W_1(\Gamma_\theta, \Gamma'_\theta)$, it is not clear that the θ used in the upper bound maintains any separation between the locations of Γ, Γ'. Thus it remains an open problem to demonstrate that the estimator in Theorem 2.3.8 is adaptive, and to find a fast algorithm that is adaptive.

2.7 Extensions

In this section, we discuss in detail several directions for future work.

2.7.1 Algorithm for density estimation

In this section, we propose a computationally feasible algorithm for density estimation in (2.1.1). We also conjecture that the algorithm not only runs in optimal time, but produces an estimator that achieves the optimal rate in Theorem 2.1.2. We are currently able to partially prove our conjecture; we explain what we know and what remains to be shown in the conjectures in this section.

What we hope to prove is that statistical distances between Gaussian mixtures on k means can be represented by the covariance distances and low-dimensional moments distances. This would allow us to reprove Theorem 2.1.2 using a different style of counting argument. But more importantly, it would allow us to obtain a fast algorithm for density estimation based on PCA and low-dimensional moment estimation; this algorithm is described in Algorithm 2. Crucially, our proposed algorithm would run in time $O(n^{ck})$. It is known [Diakonikolas et al., 2017] that a density estimation algorithm whose resulting estimator achieves the correct statistical rate must have time at least $n^{O(k)}$; thus, our algorithm would be computationally optimal, and, if

we can show it achieves the correct rate as in Theorem 2.1.2, the estimator resulting from this algorithm would be statistically optimal.

Conjecture 2.7.1. *Let $\Gamma, \Gamma' \in \mathcal{G}_{k,d}$ and let them be symmetric about zero. Let their covariance matrices be Σ, Σ', respectively. Let $V' \in S^{d\times k}$ be a matrix whose columns form an orthonormal basis of the space spanned by the atoms of Γ'. Let $\gamma = \Gamma_{V'}, \gamma' = \Gamma'_{V'}$. For any $D \in \{H^2, \mathsf{KL}, \chi^2\}$,*

$$D\left(P_\Gamma, P_{\Gamma'}\right) \asymp_k \max\left(\underbrace{\|\Sigma - \Sigma'\|_2^2}_{(I)}, \underbrace{\max_{\ell\in[2k-1]} \sup_{\theta\in S^{k-1}} \left(\Delta m_\ell(\gamma_\theta, \gamma'_\theta)\right)^2}_{(II)}\right). \tag{2.7.1}$$

Remark 2. We can already prove that (II) in (2.7.2) is both an upper and lower bound and that (I) is a lower bound. We provide this here. As usual, we upper bound the chi-square divergence and lower bound the Hellinger distance.

Let H' be the orthogonal projection operator onto the space spanned by the k atoms of Γ'. By Theorem 2.3.3 and the triangle inequality,

$$\chi^2\left(\mathbb{P}_\Gamma || \mathbb{P}_{\Gamma'}\right) \precsim_k \underbrace{\max_{\ell\in[2k-1]} \|M_\ell(\Gamma) - M_\ell(\Gamma_{H'})\|_F^2}_{(1)} + \underbrace{\max_{\ell\in[2k-1]} \|M_\ell(\Gamma_{H'}) - M_\ell(\Gamma')\|_F^2}_{(2)}.$$

By (2.3.10), for $U \sim \Gamma, U' \sim \Gamma'$,

$$(2) \precsim_k \max_{\ell\in[2k-1]} \sup_{\theta\in S^{d-1}} \left(\Delta m_\ell(\theta^\top H U, \theta^\top U')\right)^2 = \max_{\ell\in[2k-1]} \sup_{\theta\in S^{k-1}} \left(\Delta m_\ell(\theta^\top V'^\top U, \theta^\top V'^\top U')\right)^2,$$

which gives the second part of the bound.

To show (I) is a lower bound in (2.7.2), we apply the Data Processing Inequality and the second moment in Lemma 2.3.5, as usual:

$$H(P_{\Gamma'}, P_\Gamma) \geq \sup_{\theta\in S^{d-1}} H(P_{\Gamma'_\theta}, P_{\Gamma_\theta}) \geq \sup_{\theta\in S^{d-1}} |m_2(\Gamma'_\theta) - m_2(\Gamma_\theta)| \geq \|\Sigma' - \Sigma\|_2.$$

And showing (II) is a lower bound for $H(P_\Gamma, P_{\Gamma'})$ follows from the Data Processing Inequality and Lemma 2.3.5.

It remains to show that (I) is an upper bound; this is difficult and is the object of current work of the author.

Remark 3. We can already show something slightly weaker than Conjecture 2.7.1; namely, that

$$D\left(P_\Gamma, P_{\Gamma'}\right) \asymp_k \max\left(\|\Sigma - \Sigma'\|_2^4, \max_{r\in[2k-1]} \sup_{\theta\in S^{k-1}} \left(\Delta m_r(\gamma_\theta, \gamma'_\theta)\right)^2\right). \tag{2.7.2}$$

This means that the algorithm in Algorithm 2 yields an estimator that achieves a rate of $(d/n)^{1/4}$ for density estimation. Though this is not the sharp rate, it is sharper than what has been achieved so far by fast algorithms; the best known so far is $(d/n)^{1/4} * \text{polylog}(n)$ from [Acharya et al., 2014].

We now state, as a conjecture, the best size of a local cover on the densities if Conjecture 2.7.1 holds. Conjecture 2.7.2 is a conjecture because it depends on Conjecture 2.7.1; if that conjecture is true, the proof of Conjecture 2.7.2 follows, as we show below.

Conjecture 2.7.2. *Let $\mathbb{P}_\Gamma \in \mathcal{P}_{k,d}$ be the true model. Let $\mathbb{P}_0$ be the true model, and let $\mathcal{P}(\delta) := \{P_{\Gamma'} : H(P_{\Gamma'}, P_\Gamma) \le \delta\}$. Then*

$$\mathcal{N}(\epsilon, \mathcal{P}(\delta), H) \le C\left(\frac{\delta}{\epsilon}\right)^{dk+k^k}.$$

Proof of Conjecture 2.7.2 (assuming Conjecture 2.7.1 holds). Define $\mathcal{G}(\delta) = \{\Gamma' \in \mathcal{G}_{k,d} : H(P_{\Gamma'}, P_\Gamma) \le \delta\}$. Let $V' \in S^{d\times k}$ be a matrix whose columns are an orthonormal basis for the k-dimensional subspace spanned by the atoms of Γ'. By Conjecture 2.7.1

and Lemma 2.6.3,

$$\mathcal{G}(\delta) \subseteq \tilde{\mathcal{G}}(\delta) \triangleq \{\Gamma' : \|\Sigma' - \Sigma\|_2 \leq \delta \text{ and } \max_{r\in[2k-1],\mathbf{r}\in S_{r,k}} |\Delta m_{\mathbf{r}}(\Gamma_{V'}, \Gamma'_{V'})| \leq \delta\}$$

We proceed to form an ϵ-net in the Hellinger metric on $\tilde{\mathcal{G}}(\delta)$. Let $\Gamma' \in \mathcal{G}(\delta)$. Select a covariance matrix, Σ^* satisfying:

$$\|\Sigma' - \Sigma^*\|_2 \leq \epsilon. \tag{2.7.3}$$

This on the covariance space has size no more than $(\delta/\epsilon)^{dk}$. Note that this net defines the kd direction parameters and the second moment parameter. Let V^* be a matrix whose columns form an orthonormal basis for the subspace spanned by the atoms of Γ^*. Let $\gamma' = \Gamma'_{V^*}, \gamma^* = \Gamma^*_{V^*}$. Let $(r_1, r_2) \in S_{2,k}$.

$$\begin{aligned}
|\Delta m_{(r_1,r_2)}(\gamma', \gamma^*)| &\precsim_k \sup_{\theta\in S^{k-1}} |\Delta m_2(\gamma'_\theta, \gamma^*_\theta)| && \text{by Lemma 2.6.3} \\
&= \sup_{\theta\in S^{k-1}} |\theta^\top V^{*\top} \left(\mathbb{E}_{U'\sim\Gamma'} UU'^\top - \mathbb{E}_{U^*\sim\Gamma} U^*U^{*\top}\right) V^*\theta| \\
&\leq \|\Sigma' - \Sigma^*\|_2 .
\end{aligned}$$

So we have guaranteed both a covariance gap and second-order moments gap. It remains to form the rest of the distribution Γ^* by determining the remaining moments. We form a Γ^* in this way to satisfy

$$\max_{r\in[2k-1],\mathbf{r}\in S_{r,k}} |\Delta m_{\mathbf{r}}(\gamma', \gamma^*)| \leq \epsilon. \tag{2.7.4}$$

The resulting net has size no more than $(\delta/\epsilon)^{ck^k}$. The total size of the net is the size of the covariance net times the size of the moments net. □

The result in Conjecture 2.7.1 would provide an alternative way to obtain a cover-

ing number bound on the relevant parameter space (in this case, the covariance and low-dimensional moments space), which provides another way to obtain the rate via the LeCam-Birgé estimator or the MLE.

This result also suggests a fast algorithm for density estimation, which we now detail. The first step of the algorithm proceeds by reducing the dimension of our data from d to k, exactly as described in Section 2.2.1. It then remains to estimate the low-dimensional mixing distribution γ. Let $\epsilon_n = n^{-1/2}$. We form an ϵ_n-net on $\mathcal{G}_{k,k}$ as follows. Let $\mathcal{W}, \mathcal{A}$ be $\|\cdot\|_1, \|\cdot\|_2$ ϵ_n-nets on Δ^{k-1} and $\{(\mu_1, \ldots, \mu_k)^\top : \mu_j \in \mathbb{R}^k, \|\mu_j\|_2 \leq R\}$, respectively. Let $\mathcal{N}$ be an ϵ_n-net on S^{k-1}. We can guarantee that there is a constant C such that

$$\max\{|\mathcal{W}|, |\mathcal{A}|, |\mathcal{N}|\} \leq \left(\frac{C}{\epsilon_n}\right)^k \propto O(n^k). \tag{2.7.5}$$

Define the set of candidate distributions via

$$\mathcal{S} = \mathcal{A} \times \mathcal{W}. \tag{2.7.6}$$

Lemma 2.7.3. *Let $\mathcal{S}$ be as in (2.7.6). With probability at least $1 - \delta$,*

$$\min_{\gamma' \in S} \max_{r \in [2k-1]} \sup_{\theta \in S^{k-1}} |\Delta m_r(\gamma'_\theta, \gamma_\theta)| \leq \left(\frac{\log(1/\delta)}{n}\right)^{1/2}.$$

Proof. Fix $r \in [2k-1], \theta \in S^{k-1}$. Let $\gamma' = \sum_{j=1}^k w'_j \delta_{\mu'_j}, \gamma = \sum_{j=1}^k \delta_{\mu_j}$.

$$\begin{aligned}
|\Delta m_r(\gamma'_\theta, \gamma_\theta)| &\leq \sum_{j=1}^k |w'_j - w_j||(\theta^\top \mu_j)^r| + \sum_{j=1}^k w_j |(\theta^\top \mu'_j)^r - (\theta^\top \mu_j)^r| \\
&\leq R^r \epsilon_n + \sum_{j=1}^k w_j R^{r-1} \left\|\mu'_j - \mu_j\right\|_2 \\
&\precsim_{R,r} \epsilon_n.
\end{aligned}$$

This holds for any r, θ. $\square$

We will rely on vanilla Method-of-Moments (MOM), as described in [Wu and Yang, 2019].

Algorithm 2: Density estimation for k-GM in k dimensions

Input: Dataset $\{x_i\}_{i\in[n]}$ with each point in $\mathbb{R}^k$, order k, radius R.

Output: Estimate $P_{\hat{\gamma}}$ of k-atomic distribution in k dimensions.

Form the set $\mathcal{S}$ of k-atomic candidate distributions on $\mathbb{R}^k$ as in (2.7.6) ;

For each $\theta \in \mathcal{N}$**:**

Estimate the projection by $\widehat{\gamma_\theta} = \text{MOM}\left(\{\theta^\top x_i\}_{i\in[n]}\right)$;

For each candidate distribution $\gamma' \in \mathcal{S}$ **and each direction** $\theta \in \mathcal{N}$**:**

Compute $\max_{r\in[2k-1]} |m_r(\gamma'_\theta) - m_r(\widehat{\gamma_\theta})|$;

Report

$$\hat{\gamma} = \arg\min_{\gamma'\in\mathcal{S}} \max_{\theta\in\mathcal{N}} \max_{r\in[2k-1]} |m_r(\gamma'_\theta) - m_r(\widehat{\gamma_\theta})| \tag{2.7.7}$$

We now state the statistical rate achieved by the estimator computed in Algorithm 2. This is stated as a conjecture (Conjecture 2.7.4) because it depends on Conjecture 2.7.1.

Conjecture 2.7.4. *There exists a proper density estimate $P_{\hat{\Gamma}}$, computable in $O(n^k)$ time, and a positive constant C_k, such that for any $\Gamma \in \mathcal{G}_{k,d}$ and any $\delta \in (0, 1/2)$, with probability $1-\delta$,*

$$H(P_{\hat{\Gamma}}, P_\Gamma) \le C_k \left(\frac{d\log(1/\delta)}{n}\right)^{1/2}.$$

Proof. The bound is achieved by the estimator (2.2.5), this time with $\hat{\gamma}$ computed via Algorithm 2. The covariance matrix gap in the upper bound in Conjecture 2.7.1 is upper bounded by $(d/n)^{1/2}$ by Lemma 2.2.6. The moments gap in Conjecture 2.7.1 is obtained in the following way. Let $\gamma' \in \text{argmin}_{\gamma''\in\mathcal{S}} \sup_{\theta\in S^{k-1}} |\Delta m_r(\gamma'_\theta, \gamma_\theta)|$. Let

$r \in [2k-1]$. Let $\theta^* = \text{argmax}_{\theta \in S^{k-1}} |\Delta m_r(\hat{\gamma}_\theta, \gamma_\theta)|$, which exists since S^{k-1} is compact and the moments gap is continuous. Let $u \in \mathcal{A}$ satisfy $\|\theta^* - u\|_2 \leq \epsilon_n$. Note that since $\hat{\gamma}, \gamma \in \mathcal{G}_{k,d}$ and have atoms bounded by R, $|\Delta m_r(\hat{\gamma}_{\theta^*}, \hat{\gamma}_u)| \leq rR^{r-1} \|\theta^* - u\|_2$ and similarly for γ. By the triangle inequality,

$$\begin{aligned} |\Delta m_r(\hat{\gamma}_{\theta^*}, \gamma_{\theta^*})| &\leq |\Delta m_r(\hat{\gamma}_{\theta^*}, \hat{\gamma}_u)| + |\Delta m_r(\hat{\gamma}_u, \gamma_u)| + |\Delta m_r(\gamma_u, \gamma_{\theta^*})| \\ &\precsim 2rR^{r-1}\epsilon_n \max_{u \in \mathcal{N}} |\Delta m_r(\hat{\gamma}_u, \gamma_u)|. \end{aligned}$$

Now

$$\begin{aligned} \sup_{\theta \in S^{k-1}} |\Delta m_r(\hat{\gamma}, \gamma)| &\precsim_{k,r,R} 2\epsilon_n + \max_{\theta \in \mathcal{N}} |\Delta m_r(\hat{\gamma}_\theta, \gamma_\theta)| \\ &\leq 2\epsilon_n + \max_{\theta \in \mathcal{N}} |\Delta m_r(\hat{\gamma}_\theta, \gamma'_\theta)| + \max_{\theta \in \mathcal{N}} |\Delta m_r(\gamma'_\theta, \gamma_\theta)| \\ &\leq 2\epsilon_n + 2\max_{\theta \in \mathcal{N}} |\Delta m_r(\gamma'_\theta, \widehat{\gamma_\theta})| + \max_{\theta \in \mathcal{N}} |\Delta m_r(\gamma'_\theta, \gamma_\theta)| \text{ by definition of } \hat{\gamma} \\ &\leq 2\epsilon_n + 2\max_{\theta \in \mathcal{N}} |\Delta m_r(\widehat{\gamma_\theta}, \gamma_\theta) + 3\max_{\theta \in \mathcal{N}} |\Delta m_r(\gamma'_\theta, \gamma_\theta)|. \end{aligned}$$

The first term is bounded by ϵ_n with probability at least $1-\delta$ by (2.2.24). The second is bounded ϵ_n with probability at least $1-\delta$ by Lemma 2.7.3. □

2.7.2 Maximum likelihood for the density

A natural approach to any estimation problem is the maximum likelihood estimator (MLE). In mixture models of the form (1.2.1) that we have heretofore considered, the likelihood is non-convex, and typically, the Expectation Maximization (EM) algorithm is used to approximate the MLE. It is nonetheless of interest to understand the performance of the MLE itself, even though it may not be easily computable. For

$X_1, \ldots, X_n \sim_{i.i.d.} P_\Gamma$, define the density MLE:

$$\hat{P}_\Gamma \in \operatorname*{argmax}_{P_\Gamma \in \mathcal{P}_{k,d}} \sum_{i=1}^{n} \log p_\Gamma(X_i). \tag{2.7.8}$$

A rate of convergence for the MLE usually can be found by evaluating the *bracketing entropy integral* [Wong and Shen, 1995, van der Vaart and Wellner, 1996, van de Geer, 2000]. We now introduce the notion of a bracketing number.

Definition 1 (Bracketing Number)**.** Let $\mathcal{F}$ be a class of real-valued functions on a set $\mathcal{X}$. Let $\|f\|_2 = \left(\int f(x)^2\, dx\right)^{1/2}$. A bracket is a pair of functions $[l, u]$ such that $l(x) \leq f(x) \leq u(x)$ for all $x \in \mathcal{X}$. An ϵ-bracket is a bracket $[l, u]$ with $\|u - l\|_2 \leq \epsilon$. The *$\epsilon$-bracketing number* of $\mathcal{F}$ with respect to the $\|\cdot\|_2$-norm, written $N_{[]}\left(\epsilon, \mathcal{F}, \|\cdot\|_2\right)$, is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. The *bracketing entropy* is the log of the bracketing number.

It turns out that the relevant quantity for evaluation of the MLE is the ℓ_2 bracketing on the class of square root densities, or alternatively, the Hellinger bracketing on the class of densities. See Theorem 7.4 of [van de Geer, 2000] for the exact statement. We rely on this theorem to convert the bracketing number of [Maugis and Michel, 2011, Proposition B.4], provided below in (2.7.9), to a rate of convergence for the MLE. Note [Maugis and Michel, 2011] found the global, not local bracketing entropy for Gaussian mixtures of arbitrary dimension. They obtained that there are constants C, c such that

$$\log N_{[]}\left(\epsilon, \mathcal{P}_{k,d}^{1/2}, \|\cdot\|_2\right) \leq Cd \log\left(\frac{cd}{\epsilon}\right). \tag{2.7.9}$$

Applying Theorem 7.4 of [van de Geer, 2000] with the bracketing number in (2.7.9) yields (2.7.10), which states that the MLE in the k-means Gaussian mixture model of (2.1.1) achieves the (nearly) parametric rate. That is, there are constants c, C such

that with probability at least $1 - e^{-cd\log(nd)}$,

$$H(\hat{P}_\Gamma, P_\Gamma) \leq C\left(\frac{d\log(nd)}{n}\right)^{1/2}. \tag{2.7.10}$$

The $\log(nd)$ in (2.7.10) is due to the numerator of d in (2.7.9), as well as to the global nature of the bracketing number. Unfortunately, the methods used in [Maugis and Michel, 2011] are not amenable to an analysis of the local bracketing entropy, which is necessary to achieve the sharp rate of convergence for the MLE for the same reasons as discussed in Section 2.3.2 for the Le Cam-Birgé estimator. An interesting direction for further work is to obtain the sharp rate for the MLE.

Example: MLE for k-GM in one dimension

For the MLE in one dimension, it turns out that we can in fact achieve the sharp rate of convergence, as we state below in Lemma 2.7.5. The MLE for a one-dimensional mixture is evaluated via arguments about the moment space, just as in Theorem 2.1.2. The result Lemma 2.7.5 improves on what was previously known about the MLE for one-dimensional Gaussian mixtures with weak separation conditions and uses the moment technology of this thesis. However, it currently assumes the distributions Γ are centered; this assumption is there for technical purposes, and it will be useful in the future to obtain the result without this condition.

Lemma 2.7.5 (MLE in one dimension)**.** *Let $\mu_1, \ldots, \mu_k \in \mathbb{R}$ satisfy $|\mu_j| \leq R$ for a positive constant R. Let $w_1, \ldots, w_k \geq 0$ and $\sum_{j=1}^k w_j = 1$. Let $\sum_{j=1}^k w_j\mu_j = 0$, and let $\Gamma = \sum_{j=1}^k w_j\delta_{\mu_j}$. Let $\hat{P}_\Gamma$ be the MLE for P_Γ. Then with probability at least $1-\delta$,*

$$H(\hat{P}_\Gamma, P_\Gamma) \leq \left(\frac{k\log(1/\delta)}{n}\right)^{1/2}.$$

This lemma follows directly from Lemma 2.7.7 and Lemma 2.7.6, which are provided below.

Lemma 2.7.6 (Bracketing for k-GM in one dimension)**.** *Let $\Gamma \in \mathcal{G}_{k,1}$. There is a positive constant C such that*

$$N_{[]}(\epsilon, \mathcal{P}_{k,1}^{1/2}(\delta), \|\cdot\|_2) \leq \left(\frac{C\delta}{\epsilon}\right)^{2k-1}.$$

Lemma 2.7.7. *Suppose there is a positive constant C such that*

$$N_{[]}(\epsilon, \mathcal{P}^{1/2}(\delta), \|\cdot\|_2) = \left(\frac{C\delta}{\epsilon}\right)^{d}.$$

Then there is a constant c such that with probability at least $1-\delta$,

$$H(\hat{P}_\Gamma, P_\Gamma) \leq c\left(\frac{d\log(1/\delta)}{n}\right)^{1/2}.$$

The proof of Lemma 2.7.7 is immediate from Theorem 7.4 of [van de Geer, 2000]; we provide it here for clarity.

Proof of Lemma 2.7.7. Note that $\left\|\bar{p}_1^{1/2} - \bar{p}_2^{1/2}\right\|_2 \leq \left\|p_1^{1/2} - p_2^{1/2}\right\|_2 / \sqrt{2}$ by Lemma 4.2 of [van de Geer, 2000]. Therefore we may reduce ourselves to bounding $N_{[]}(\sqrt{2}\epsilon, \mathcal{P}^{1/2}(\delta), \|\cdot\|_2)$. Apply Theorem 7.4 of [van de Geer, 2000] in the following way. Let $\Psi(\delta) := C_1\sqrt{d}\delta$ for an appropriate positive constant C_1. We have

$$\begin{aligned}
\int_{\delta^2/2^{13}}^{\delta} \log\sqrt{N_{[]}(s, \mathcal{P}^{1/2}(\delta), \|\cdot\|_2)}\, ds &\leq \int_{\delta^2/2^{13}}^{\delta} \sqrt{\log\left(\frac{C\delta}{u}\right)^d}\, du \\
&\leq \sqrt{d}\, C\, \delta \int_{\delta/(C2^{13})}^{1/C} \sqrt{\log\frac{1}{t}}\, dt \quad \text{using the substitution } t = u/C\delta. \\
&\leq \sqrt{d}\, C\, \delta \int_{0}^{1/C} \sqrt{\log\frac{1}{t}}\, dt \\
&\leq \Psi(\delta).
\end{aligned}$$

And $\Psi(\delta)/\delta^2$ is nonincreasing in δ. Setting $\delta_{d,n} = (d/n)^{1/2}$, there is a positive constant c such that $c\Psi(\delta_{d,n}) \leq \sqrt{n}\delta_{d,n}^2$. By Theorem 7.4 of [van de Geer, 2000], for any

$\delta \geq \delta_{d,n}$, $\mathbb{P}\{H(\hat{P}_\Gamma, P_\Gamma) \geq \delta\} \leq ce^{-n\delta^2/c^2}$. □

It remains to prove Lemma 2.7.6. We first provide an important supporting lemma for Lemma 2.7.6. A standard approach to computing the bracketing number is to relate it to a cover number in the relevant parameter space via, for instance, Lemma 2.7.11 of [van der Vaart and Wellner, 1996], which is provided here.

Lemma 2.7.8 (Lemma 2.7.11 of [van der Vaart and Wellner, 1996])**.** *Let Θ be a metric space equipped with metric ρ. Let $\theta \in \Theta$. Let $\mathcal{X}$ be a metric space. Let $f_\theta : \mathcal{X} \to \mathbb{R}$ be a function parametrized by θ. Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. Let $N(\epsilon, \Theta, \rho)$ be the ϵ covering number in the ρ metric on Θ. Suppose for every $\theta, \theta_* \in \Theta$, and for all $x \in \mathcal{X}$,*

$$|f_\theta(x) - f_{\theta_*}(x)| \leq \rho(\theta, \theta_*)F(x), \tag{2.7.11}$$

where $F(x)$ is square-integrable under Lebesgue measure. Then

$$N_{[]}\left(\epsilon \left\|F\right\|_2, \mathcal{F}, \left\|\right\|_2\right) \leq N(\epsilon, \Theta, \rho).$$

Proof of Lemma 2.7.6. Let $\mathcal{G}$ be the set of all k-atomic distributions in one dimensions with centers $\mu_1, \ldots, \mu_k$ satisfying $|\mu_j| \leq 1$ for each $j \in [k]$. Let

$$\mathcal{G}(\delta, \Gamma) = \{\Gamma \in \mathcal{G} : H(P_\Gamma, P) \leq \delta\}.$$

By Lemma 2.3.5, for $\Gamma \in \mathcal{G}_\delta(\Gamma)$,

$$\max_{r \in [2k-1]} |m_r(\Gamma) - m_r(\Gamma)| \leq \delta. \tag{2.7.12}$$

A one-dimensional k-atomic distribution is uniquely determined by its first $2k-1$ moments. To create an ϵ-covering on the distributions, we create a net on the moment space. Let $\mathbb{M}(\delta)$ be the set of moment vectors in $\mathbb{R}^{2k-1}$ moments satisfying (2.7.12),

i.e., every vector $m \in \mathbb{M}(\delta)$ satisfies $\|m - m_0\|_\infty \le \delta$. Now $\mathbb{M}(\delta)$ is a subset of the rectangle $[-\delta, \delta]^{2k-1}$, so an ϵ-covering on $\mathbb{M}(\delta)$ has size upper bounded by $(2\delta/\epsilon)^{2k-1}$. Thus $N(\epsilon, \mathcal{G}_\delta(\Gamma), \|\cdot\|_2) \le (2\delta/\epsilon)^{2k-1}$. Let $\Gamma \in \mathbb{M}(\delta)$, and let Γ' be the nearest k-atomic distribution in $\mathcal{N}$. By Lemma 10 of [Wu and Yang, 2019], (2.7.12) implies that we obtain a moments gap for all moments, i.e., for any positive integer r,

$$|m_r(\Gamma) - m_r(\Gamma')| \le 3^r \epsilon. \tag{2.7.13}$$

So

$$|p_\Gamma(x) - p_{\Gamma'}(x)| \le \phi(x) \sum_{r \ge 1} \frac{3^r |H_r(x)||m_r(\Gamma) - m_r(\Gamma')|}{r!} \le \epsilon\phi(x) \sum_{r \ge 1} \frac{3^r |H_r(x)|}{r!}. \tag{2.7.14}$$

Now let m_r, m_r' be the rth moments of Γ, Γ', respectively. Now

$$|\sqrt{p_\Gamma(x)} - \sqrt{p_{\Gamma_*}(x)}| = \frac{|p_\Gamma(x) - p_{\Gamma_*}(x)|}{\sqrt{p_\Gamma(x)} + \sqrt{p_{\Gamma_*}(x)}} \le \phi^{-1/2}(x) \frac{|p_\Gamma(x) - p_{\Gamma_*}(x)|}{e^{xm_1/2 - m_2/4} + e^{xm_1'/2 - m_2'/4}},$$

by Jensen's inequality. We apply Lemma 2.7.8 with

$$F(x) = \phi^{1/2}(x) \frac{\sum_{r=1}^\infty 3^r |H_r(x)|/r!}{e^{xm_1/2 - m_2/4} + e^{xm_1'/2 - m_2'/4}}.$$

Since the distributions Γ are centered, by Lemma 2.3.10, $F(x) \le 2e^{R^2/4} \phi_1^{1/2}(x) \sum_{r \ge 1} 3^r |H_r(x)|/r!$. So $\|F\|_2^2 \le 4e^{R^2/2} \mathbb{E}_{N(0,1)} \left(\sum_{r \ge 1} |H_r(x)|/r! \right)^2$. We have absolute Hermites here so we cannot use their orthogonality; however,

$$\mathbb{E}_{x \sim N(0,1)} \left(\sum_{r \ge 1} \frac{3^r |H_r(x)|}{r!} \right)^2 = \sum_{r \ge 1} \mathbb{E} \frac{9^r H_r^2(x)}{(r!)^2} + \underbrace{\sum_{r \ne l} \mathbb{E} \frac{3^{r+l} |H_r(x) H_l(x)|}{r! l!}}_{(2)}$$

The first term is equal to $\sum_{r \ge 1} \frac{9^r}{r!}$ by evaluating the expected value of the squared

Hermites. For the second term, using Cauchy Schwarz,

$$(2) \leq \sum_{r \neq l} \frac{3^{r+l}\sqrt{\mathbb{E}H_r^2(x)\mathbb{E}H_l^2(x)}}{r!l!} = \sum_{r \neq l} \frac{3^{r+l}}{\sqrt{r!}\sqrt{l!}} \leq \left(\sum_{r \geq 1} \frac{3^r}{\sqrt{r!}}\right)^2 \leq \left(\sum_{r \geq 1} \frac{3^r e^r}{\sqrt{2\pi} r^{r/2+1/4}}\right)^2 < \infty.$$

□

2.7.3 Maximum likelihood for the mixing distribution

In this section, we discuss the problem of obtaining a rate of convergence for the parameter mixing distribution MLE. Let us have data $X_1, \ldots, X_n$ from the model (2.1.1). Define the mixing distribution MLE:

$$\hat{\Gamma} = \underset{\Gamma \in \mathcal{G}_{k,d}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_\Gamma(X_i). \tag{2.7.15}$$

By (2.7.10), the proper (k-GM) density MLE (nearly) achieves the parametric rate $(d\,\mathrm{polylog}(d,n)/n)^{1/2}$. It is interesting to consider whether this means that the derived mixing distribution MLE $\hat{\Gamma}$ achieves the optimal convergence rate stated in Theorem 2.1.1.

Consider first for example the symmetric, 2-GM, when there is a single parameter μ. By (2.7.10) and Lemma 2.3.5, $(d\,\mathrm{polylog}(d,n)/n)^{1/2} \geq H(P_{\hat{\mu}}, P_\mu) \gtrsim \sup_{\theta \in S^{d-1}} |m_2(\theta^\top \hat{\mu}) - m_2(\theta^\top \mu)| = \left\|\hat{\mu}\hat{\mu}^\top - \mu\mu^\top\right\|_2$. By Lemma 2.6.5, this means

$$\ell(\hat{\mu}, \mu) \precsim \left(\frac{d\log(dn)}{n}\right)^{1/4}.$$

Thus in this simple example, it is straightforward to show that the parameter MLE itself achieves the correct rate in Theorem 2.1.1. We now study whether this holds in general for $\hat{\Gamma}$. By (2.7.10) and Theorem 2.3.8, a sample-split version of the MLE does obtain the rate stated in Theorem 2.1.1. It is now of interest to know whether the true MLE, without sample splitting, can achieve the rate. We believe the following

conjecture is true. We state it and provide two methods of proof that seem promising.

Conjecture 2.7.9. *Let $\hat{\Gamma}$ be the MLE as in (2.7.15). Then there is a positive constant C_k such that with probability at least $1-\delta$,*

$$W_1(\hat{\Gamma}, \Gamma) \leq C_k \left(\left(\frac{d \log(1/\delta)}{n} \right)^{1/4} + \left(\frac{\log(1/\delta)}{n} \right)^{1/(4k-2)} \right). \tag{2.7.16}$$

Remark 4. We now provide potential approaches for proving Conjecture 2.7.9.

Let $V = [v_1, \ldots, v_k] \in S^{d \times k}$ be the matrix whose columns form an orthonormal basis for the space spanned by the atoms of the true mixing distribution Γ. Let $H = VV^\top$. Define $\widehat{\Gamma_H} = \operatorname{argmax}_{\Gamma \in \mathcal{G}_{k,d}} \sum_{i=1}^n \log p_\Gamma(HX_i)$, i.e., it is the k-atomic MLE obtained when all the data are projected onto the true direction V. Clearly, this estimator is not something we could compute in practice; we will rely on its theoretical properties. The crucial idea is that since V is the true direction, when data are projected onto this direction and the MLE its computed, its error in Hellinger distance is bounded by the low-dimensional rate, i.e., by (2.7.10),

$$H^2(P_{\widehat{\Gamma_H}}, P_\Gamma) = \tilde{O}\left(\frac{k}{n} \right)^{1/2}. \tag{2.7.17}$$

Now

$$W_1(\hat{\Gamma}, \Gamma) \leq W_1(\hat{\Gamma}, \widehat{\Gamma_H}) + W_1(\widehat{\Gamma_H}, \Gamma). \tag{2.7.18}$$

For the second term,

$$\begin{aligned} W_1(\widehat{\Gamma_H}, \Gamma) &\leq H^{1/(4k-2)}(P_{\widehat{\Gamma_H}}, P_\Gamma) && \text{by Lemma 2.5.1 and Lemma 2.3.5} \\ &\leq \tilde{O}(1/n)^{1/(4k-2)}. && \text{by (2.7.10)} \end{aligned}$$

It remains to tackle the first term in (2.7.18). First,

$$\begin{aligned} H(P_{\hat{\Gamma}}, P_{\widehat{\Gamma_H}}) &\precsim H(P_{\hat{\Gamma}}, P_{\Gamma}) + H(P_{\Gamma}, P_{\widehat{\Gamma_H}}) \\ &= \tilde{O}\left(\frac{d}{n}\right)^{1/2} + \tilde{O}\left(\frac{k}{n}\right)^{1/2} \\ &= \tilde{O}\left(\frac{d}{n}\right)^{1/2}. \end{aligned}$$

It remains to translate this rate from Hellinger to its square root in W_1. This does not seem straightforward, since typically we related W_1 and Hellinger via Lemma 2.3.5 and Lemma 2.5.1, but doing so here would lead to a rate of $(d/n)^{1/(4k-2)}$. The idea is that is as follows. The MLE $\hat{\Gamma}$ is a k-GM with form $\hat{\Gamma} = \sum_{j=1}^{k} \hat{w}_j \delta_{\hat{\mu}_j}$. Let $\hat{\Sigma} = \sum_{j=1}^{k} \hat{w}_j \hat{\mu}_j \hat{\mu}_j^\top$. Let $\hat{V} = [\hat{v}_1, \ldots, \hat{v}_k]$ be the space spanned by the k atoms of $\hat{\Gamma}$. We already know by (2.7.10) and Lemma 2.3.5 that $\left\|\hat{\Sigma} - \Sigma\right\|_2 = \tilde{O}(d/n)^{1/2}$, but this does not necessarily mean that $\hat{V}$ and V are close; for the eigenvalues of $\hat{\Sigma}, \Sigma$ could be small. However, if they are close, it seems that computing the MLE with the same data onto two spaces that are close would yield k-GM's with similar weights, we might be able to employ a natural coupling for the analysis. If the spaces are not close, then $\left\|\hat{\Sigma}\right\|_2, \|\Sigma\|_2$ are small anyway and the rate should hold.

2.7.4 Infinite mixture on a subspace

We now discuss a simple extension of (2.1.1), in which the distribution Γ lies on a bounded span of a vector of arbitrary size, and the parameter of interest is this vector. We will see that the results of this chapter extend in a straightforward way to yield a rate of convergence for estimating the parameter in this problem.

Let $\mathcal{G} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$. Let $\mathbb{Q}$ be a distribution on the interval $[-c, c]$ with mean zero and variance 1. Assume also that $\mathbb{Q}$ is symmetric about zero. We wish to

estimate θ in the model

$$X_1, \ldots, X_n \sim_{i.i.d.} P_\theta, \text{ where} \tag{2.7.19}$$

$$\mathbb{P}_\theta = N(A\theta, \sigma^2 I_d), \text{ and} \tag{2.7.20}$$

$$A \sim \mathbb{Q}. \tag{2.7.21}$$

That is, for each $i \in [n]$,

$$X_i = A_i\theta + Z_i$$

where $A_i \sim \mathbb{Q}$, θ is a fixed vector in $\mathbb{R}^d$, and $Z_i \sim N(0, \sigma^2 I_d)$. Let p_θ be the density of the distribution $\mathbb{P}_\theta$. Let q be the density of the distribution $\mathbb{Q}$. Note that

$$p_\theta(x) = \int_{-c}^{c} \exp\left(-\frac{\|x - a\theta\|_2^2}{2\sigma^2}\right) q(a)\, da.$$

Note that when $\mathbb{Q}$ is the Rademacher(1/2) distribution, this corresponds exactly to the symmetric 2-GM in Section 2.6.3. Our first result is that the convergence rate for estimating θ here is exactly as in the symmetric 2-GM of Section 2.6.3. Since this is in fact an infinite mixture, we cannot directly apply the results from earlier in this chapter. However, similar proofs allow us to obtain the rate here. In Lemma 2.7.10, the estimation of θ is done via a spectral estimator; to compute the estimator, the statistician need not know in advance the exact distribution $\mathbb{Q}$, but the proof does rely on the fact that it is symmetric and bounded on $[-1, 1]$.

Lemma 2.7.10. *Let $X_1, \ldots, X_n \sim_{i.i.d.} P_\theta$ as in (2.7.21). The minimax risk of estimating θ over the class $\mathcal{G}$ satisfies*

$$\inf_{\hat{\theta}} \sup_{\theta} \mathbb{E}_\theta \ell(\hat{\theta}, \theta) \asymp \left(\frac{d}{n}\right)^{1/4}. \tag{2.7.22}$$

And there exists an estimator $\hat{\theta}$, computable in $O(n^3)$ time, such that with probability

at least $1-\delta$,

$$\ell(\hat{\theta},\theta) \asymp \left(\frac{d\log(1/\delta)}{n}\right)^{1/4}. \tag{2.7.23}$$

Proof. We first prove the upper bound in (2.7.23). Let $\hat{\theta} = \sqrt{\hat{\lambda}}\hat{u}$, where $\hat{u}$ is the first unit eigenvector of $X^\top X/n - \sigma^2 I_d$ and $\hat{\lambda}$ is the corresponding eigenvalue. Now $\hat{\theta}\hat{\theta}^\top$ is the rank-1 approximation of $\hat{\Sigma} - \sigma^2 I_d$, so by Lemma 3.5.8, $\left\|\hat{\theta}\hat{\theta}^\top - \theta\theta^\top\right\|_2 \leq \left\|\hat{\Sigma} - \sigma^2 I_d - \theta\theta^\top\right\|_2$. And

$$\hat{\Sigma} - \sigma^2 I_d - \theta\theta^\top = \theta\theta^\top \frac{1}{n}\sum_{i\in[n]} A_i^2 - \theta\theta^\top + + \frac{1}{n}\sum_{i\in[n]} Z_i Z_i^\top - \sigma^2 I_d + \frac{1}{n}\sum_{i\in[n]} A_i(\theta Z_i^\top + Z_i\theta^\top).$$

For the first term, $\left\|\theta\theta^\top \frac{1}{n}\sum_{i\in[n]} A_i^2 - \theta\theta^\top\right\|_2 \leq R^2\left|\frac{1}{n}\sum_{i\in[n]}(A_i^2 - 1)\right|$ since $\|\theta\|_2 \leq R$. Recall that $A_i^2 \leq c^2$ and that $\mathbb{E}A_i^2 = 1$. We apply Hoeffding's Inequality to obtain that with probability at least $1-\delta$,

$$\left\|\theta\theta^\top \frac{1}{n}\sum_{i\in[n]} A_i^2 - \theta\theta^\top\right\|_2 \leq R^2\sqrt{\frac{c^4\log(2/\delta)}{2n}}.$$

For the second term, by the same analysis as in the proof of Lemma 2.2.6, with probability at least $1-e^{-t^2}$,

$$\left\|\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top - \sigma^2 I_d\right\|_2 \leq C\left(\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}} + \frac{t^2}{n}\right). \tag{2.7.24}$$

For the third term, we again imitate the method used in the proof of Lemma 2.2.6. Let $B = \frac{1}{n}\sum_{i=1}^n A_i(\theta Z_i^\top + Z_i\theta^\top)$, and let $\mathcal{N}$ be an $\frac{1}{4}$-covering of S^{d-1} of size 2^{Cd} for an absolute constant C. Then

$$\|B\|_2 = \max_{u\in S^{d-1}} |u^\top B u| \leq 2\max_{u\in\mathcal{N}} |u^\top B u| = 4\max_{u\in\mathcal{N}}\left|\frac{1}{n}\sum_{i=1}^n (A_i u^\top\theta)(u^\top Z_i)\right|.$$

For fixed $u \in S^{d-1}$, conditioning on A_i, we have $\sum_{i=1}^n (A_i u^\top \theta)(u^\top Z_i) \sim N\left(0, \sum_{i=1}^n (A_i u^\top \theta)^2\right)$. Since $\sum_{i=1}^n (A_i u^\top \theta)^2 \le nc^2R^2$,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n (A_i u^\top \theta)(u^\top Z_i) > t\right\} \le \mathbb{P}\left\{Rc\left|\frac{1}{n}\sum_{i=1}^n z_i\right| > t\right\},$$

where $z_i \sim_{i.i.d.} N(0,1)$. By a union bound, with probability at least $1-\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n A_i(\theta Z_i^\top + Z_i\theta^\top)\right\|_2 \le \sqrt{8}Rc\left(\sqrt{\frac{Cd\log 2}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{2.7.25}$$

For the lower bound, we rely on Fano's Lemma. For shorthand, we write $\mathcal{M} = \inf_{\hat\theta}\sup_\theta \mathbb{E}_\theta \left\|\hat\theta - \theta\right\|_2^2$. Let $\mathcal{H} = \{\pm 1\}^d$, the hypercube in d dimensions. By Gilbert-Varshamov, there is a set $\mathcal{V} \subset \mathcal{H}$ such that $|\mathcal{V}| \ge e^{d/8}$ and such that for all $\nu, \nu' \in \mathcal{V}, \|\nu - \nu'\|_2 \ge \sqrt{d}/2$. Now we define our set of parameters $\{\theta(\mathbb{P}_\nu)\}_{\nu\in\mathcal{V}}$. For each $\nu \in \mathcal{V}$, let

$$\theta = \epsilon\nu$$

for ϵ which we will define later. Then for $\theta, \theta' \in \{\theta(\mathbb{P}_\nu)\}_{\nu\in\mathcal{V}}$,

$$\|\theta - \theta'\|_2 = \epsilon\|\nu - \nu'\|_2 \gtrsim \epsilon\sqrt{d}$$

So let

$$\delta = \epsilon\sqrt{d}$$

Now we'll use the global version of Fano's lemma, with $\nu_0 = 0$. By Lemma 2.7.11, $D_{KL}(\mathbb{P}_\theta||\mathbb{P}_0) \precsim \|\theta\|_2^4/\sigma^4$ for any $\theta = \epsilon\nu$. So

$$D_{KL}(P_\nu||P_{\nu_0}) \precsim \frac{\|\theta\|_2^4}{\sigma^4} \le \frac{\epsilon^4\|\nu\|_2^4}{\sigma^4} \le \frac{\epsilon^4 d^2}{\sigma^4}$$

And by our choice of $\mathcal{V}$ and by Gilbert-Varshamov, $\log|\mathcal{V}| \ge d/8$. So by the center

version of Fano's Method,

$$\begin{aligned}\mathcal{M} &\gtrsim \delta^2 \left(1 - \frac{nD_{KL}(P_\nu||P_0)}{\log|\mathcal{V}_0|}\right) \\ &\geq \epsilon^2 d \left(1 - \frac{n\epsilon^4 d^2}{d\sigma^4}\right).\end{aligned}$$

Now choosing ϵ such that $n\epsilon^4 d/\sigma^4 = 1/2$ means choosing

$$\epsilon = \left(\frac{\sigma}{nd}\right)^{1/4}.$$

And then $\delta = \epsilon\sqrt{d} = (\sigma d/n)^{1/4}$. Then the lower bound is

$$\mathcal{M} \gtrsim \delta^2 = \sqrt{\frac{\sigma d}{n}}$$

as needed. □

Lemma 2.7.11. *Let $\mathbb{P}_\theta$ be the $N(A\theta, \sigma^2 I_d)$ model. Then*

$$KL(\mathbb{P}_\theta|\mathbb{P}_0) \precsim \frac{\|\theta\|_2^4}{2\sigma^4}$$

Proof. Let $B \sim \mathbb{Q}$ be independent of A. We will write $\mathbb{E}_A, \mathbb{E}_B$ for shorthand to indicate the expectation over the distribution $\mathbb{Q}$ for random variables A, B. Now

$\frac{p_\theta(x)}{p_0(x)} = \mathbb{E}_A \exp\left(\frac{Ax'\theta - A^2\|\theta\|^2}{2\sigma^2}\right)$.

$$\mathbb{E}_{p_\theta}\left(\frac{p_\theta(x)}{p_0(x)}\right) = \mathbb{E}_B\mathbb{E}_A\mathbb{E}_{X\sim N(B\theta,\sigma^2 I_d)} e^{AX'\theta/\sigma^2 - A^2\|\theta\|^2/2\sigma^2} \text{ by Fubini}$$

$$= \mathbb{E}_B\mathbb{E}_A e^{AB\|\theta\|^2/\sigma^2} \text{using the Gaussian MGF}$$

$$= \sum_{k=0}^{\infty} \mathbb{E}_A\mathbb{E}_B \frac{(AB\,\|\theta\|^2)^k}{\sigma^{2k}k!} \text{ by Taylor-expanding exp and Fubini}$$

$$= \sum_{k=0}^{\infty} \frac{\|\theta\|^{2k}}{\sigma^{2k}k!}\mathbb{E}B^k\mathbb{E}A^k.$$

Since A, B are distributed according to $\mathbb{Q}$, which is symmetric about zero, the odd moments of A, B are zero. Let $m = \max_k \mathbb{E}A^{2k}\mathbb{E}B^{2k}$. Then

$$\begin{aligned}\mathbb{E}_{p_\theta}\left(\frac{p_\theta(x)}{p_0(x)}\right) &= \sum_{k=0}^{\infty} \frac{\|\theta\|^{4k}}{\sigma^{4k}(2k)!}\mathbb{E}B^{2k}\mathbb{E}A^{2k} \\ &\leq \sum_{k=0}^{\infty} \frac{\|\theta\|^{4k}}{\sigma^{4k}2^k k!}\mathbb{E}B^{2k}\mathbb{E}A^{2k} \\ &= \exp\left(\frac{m\,\|\theta\|^4}{2\sigma^4}\right)\end{aligned}$$

where the last step follows because $2^k \leq (2k)*(2k-1)*\ldots,(2k-k+1)$, so $(2k)! \geq 2^k k!$. And $KL(\mathbb{P}_\theta||\mathbb{P}_0) \leq \log \mathbb{E}_{p_\theta}\left(\frac{p_\theta(x)}{p_0(x)}\right)$, which is upper bounded by $C\,\|\theta\|_2^4/\sigma^4$ as we have just shown. □

2.7.5 Mixture of subspaces

The model (2.1.1) is related to a broader model class: a mixture model when the mixing distribution lies on a finite collection of linear subspaces of $\mathbb{R}^d$. For simplicity,

we will allow each subspace $S_1, \ldots, S_k$ to have rank $r \geq 1$. For each $j \in [k]$, let S_j be spanned by the orthonormal basis $\{v_{j1}, \ldots, v_{jr}\}$. We observe data as in (1.2.1), i.e., we observe data from the model

$$X_1, \ldots, X_n \sim_{i.i.d.} P_\Gamma, \text{ where} \tag{2.7.26}$$

$$P_\Gamma = \Gamma * N(0, I_d) \text{ and } \Gamma = \sum_{j=1}^{k} w_j \Gamma_j, \tag{2.7.27}$$

where each Γ_j is a certain uniform distribution on S_j. We have

$$X_i = U_i + Z_i, \text{ where}$$
$$U_i = \sum_{\ell \in [r]} A_{i\ell} V_\ell,$$

where $U_\ell \sim \Psi_\ell = \sum_{j=1}^{k} w_j \delta_{v_{j\ell}}$, and crucially, $U_1, \ldots, U_r$ are coupled so that when $U_1 = v_{j1}, U_2 = v_{j2}$, and so on. For an arbitrary $U \sim \Gamma$, we can also write

$$U = VA \tag{2.7.28}$$

In (2.7.28), $V = [V_1, \ldots, V_r] \in \mathbb{R}^{d \times r}$ and each $V_\ell \sim \Psi_\ell$ and the V_ℓ are coupled as above. We also write $U = VA$ where $V \sim \Psi$, with

$$\Psi = \sum_{j=1}^{k} w_j \delta_{[v_{j1}, \ldots, v_{jr}]},$$

i.e., Ψ is a distribution taking k values, each value being a matrix $W \in S^{d \times r}$, with $V^\top V = I_r$. And $A = (A_1, \ldots, A_r)^\top \in \mathbb{R}^r$, where each $A_\ell \sim_{i.i.d.} Uni(-1, 1)$. We denote the distribution of the vector A via $Uni[-1, 1]^r$. The distribution P_Γ in (2.7.26) has density p_Γ:

$$p_\Gamma(x) = \mathbb{E}_{U \sim \Gamma} \phi(x - U) = \mathbb{E}_{A \sim Uni[-1,1]^r} \mathbb{E}_{V \sim \Psi} \phi(x - VA)$$

Define the parameter and density classes:

$$\mathcal{G}_{k,r,d} = \{\Gamma = \sum_{j=1}^{k} w_j Uni(S_j) : S_j \text{ spanned by } v_{j1}, \ldots, v_{jr}\},$$

$$\mathcal{P}_{k,r,d} = \{p_\Gamma : \Gamma \in \mathcal{G}_{k,r,d}\}.$$

The model (2.7.26) is closely related to the subspace clusters model considered in, e.g., [Vidal, 2009, Elhamifar and Vidal, 2009, Soltanolkotabi et al., 2014]. These works focused on the problem of clustering in such a model, and there has been little work so far on parameter and density estimation in this context. Obtaining rates of convergence for the estimation of Γ and P_Γ in (2.7.26) is an important open problem, and the earlier results of this chapter provide intuition on what these rates might look like. We accordingly state conjectures about density and parameter estimation in the model (2.7.26).

In 2.7.12, we conjecture that density estimation in the model (2.7.26) can be done at the parametric rate of $\sqrt{d/n}$. We anticipate that estimating the density in (2.7.26) will be related to estimating the moment tensors, as in Theorem 2.3.3; this is because densities in the class $\mathcal{P}_{k,r,d}$ have a Hermite expansion similar to those in $\mathcal{P}_{k,d}$.

Conjecture 2.7.12. *Let the model be as in (2.7.26). Then the minimax rate of estimating P_Γ over $\mathcal{P}_{k,r,d}$ satisfies*

$$\inf_{\hat{P}_\Gamma} \sup_{P_\Gamma \in \mathcal{P}_{k,r,d}} \mathbb{E}_\Gamma H(P_\Gamma, P_\Gamma) \asymp \left(\frac{d}{n}\right)^{1/2}.$$

We now turn to our conjecture about parameter estimation. We conjecture that there will be a two-part rate similar to that in Theorem 2.1.1. This is because we could easily perform a similar procedure to Algorithm 1; we can estimate the (kr)-dimensional subspace, project the data to the estimated subspace, then proceed to do mixing estimation on the low-dimensional data. In 2.7.13, we conjecture that the

rate of estimating the subspace should be sharper than in Theorem 2.1.1 because we have full information about the subspaces. This is because (2.7.26) has unit basis vectors, and the coefficients A_ℓ are distributed according to $Uni(-1,1)$; thus we have full information about the subspaces and should be able to estimate them at the parametric rate.

Conjecture 2.7.13. *Let the model be as in (2.7.26). Then the minimax rate of estimating* Γ *over* $\mathcal{G}_{k,r,d}$ *satisfies*

$$\inf_{\hat{\Gamma}} \sup_{\Gamma \in \mathcal{G}_{k,r,d}} E_\Gamma W_1(\hat{\Gamma}, \Gamma) \asymp_{k,r} \left(\frac{d}{n}\right)^{1/2} + \left(\frac{kr}{n}\right)^{1/c_r(4k-2)}.$$

Chapter 3

Mixtures of manifolds: kernel spectral clustering and refinement

3.1 Introduction

Clustering is an important task in statistics, with applications to a wide range of scientific fields. The goal is to separate data into groups such that points within a group are close to each other but points between groups are well-separated. Data analysts often wish to cluster data that lies in a high-dimensional space, and many of the advances in this area over the past decades have relied on the observation that the data, while high-dimensional, truly lies on a low-dimensional subspace of the ambient space.

Consider facial image data consisting of images of people's faces in different positions and lighting. Each data point is a vector in, e.g., $\mathbb{R}^{32\times 32}$, with each vector component representing a pixel's brightness. In a typical example of such a dataset, [Tenenbaum et al., 2000] show that the data can be represented by a three-dimensional manifold, with the dimensions capturing the lighting and the vertical and horizontal inclinations of the face. A good clustering algorithm should be able to sep-

arate a dataset consisting of several faces, each shown in multiple different positions or lighting, into the correct groups, as opposed to naively clustering them according to position or lighting.

A classical technique for handling such data is to find the low-dimensional linear subspace that is nearest to the data, a procedure known as Principal Component Analysis (PCA). This method assumes that the data lie on a single low-dimensional subspace. But in practice, the data may lie on multiple subspaces. If the subspace membership of each point were known, it would be easy to simply do PCA separately on each group of data points. But in general, the clustering may be unknown. Moreover, it is more realistic to assume that the data lie, not exactly on a low-dimensional space, but consist of low-dimensional signal plus some high-dimensional noise.

We consider the problem of clustering data into k clusters, where k is known, when the data are generated from k spaces, $S_1, \ldots, S_k$, embedded in $\mathbb{R}^p$. For instance, each S_j might be a manifold of dimension m_j, where $m_j \ll p$. Let $m = \max_{s\in[k]} m_j$. Suppose we have n data points, $X_1, \ldots, X_n \in \mathbb{R}^p$. Let $\tau \in \{1, \ldots, k\}^n$ be the true assignment vector. The data are truly grouped into k clusters, $C_1, \ldots, C_k$. That is, $C_s = \{i : \tau(i) = s\}$. We assume X_i has the following structure:

$$X_i = M_i + Z_i \tag{3.1.1}$$

in which M_i is distributed according to some distribution on $S_{\tau(i)}$. And $Z_i \sim N(0, I_p)$. All vectors are assumed to be independent. The assumption on the structure of X_i is standard in the literature, for example, see [Karoui, 2010a] and [Karoui and tieng Wu, 2014]. In matrix form, we let $X, M, Z \in \mathbb{R}^{n\times p}$ be the matrices with rows X_i, M_i, Z_i, respectively.

We moreover assume that the spaces on which the signals lie are separated. This is in contrast to, for instance, the subspace clustering problem, in which it is assumed

that the data lie on k linear subspaces which are allowed to overlap. Let

$$d_{st} = \inf_{\tau(i)=s,\tau(j)=t,s\neq t} \|M_i - M_j\|_2$$

And let $d_{\min} = \min_{s\neq t} d_{st}, d_{\max} = \max_{s\neq t} d_{st}$. We assume that M has rank bounded by $r \leq mk$; Lemma 3.5.11 shows why this assumption is reasonable. We also assume that for $i \neq j$ where $\tau(i) = \tau(j), \|M_i - M_j\|_2 \leq b$, where b is some positive constant. And we assume that

$$\max_{\tau(i)=s,\tau(j)=t,s\neq t} \|M_i - M_j\|_2^2 \leq 8b^2 + 4d_{st}^2$$

This is reasonable if we assume that on each manifold there is a point that achieves the minimum distance to each of the other manifolds. For an estimator $\hat{\tau}$ of τ, our loss, or mis-clustering rate, is $\mathcal{L}(\tau, \hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\hat{\tau}(i) \neq \tau(i)\}$.

Spectral methods, which make no assumption about the shapes of the clusters, are widely-used for community detection on such data. These methods depend on a kernel matrix, or on its closely-related Laplacian matrix. The kernel random matrix A, based on $X_1, \ldots, X_n$, is a symmetric matrix with entries $A_{ij} = \mathcal{K}(X_i, X_j)$ where $\mathcal{K}$ is a kernel function.

Let $U_k \in \mathbb{R}^{n\times k}$ be the matrix whose columns are the first k orthonormalized eigenvectors of A. The analyst then applies K-means, with $K = k$, or some other clustering algorithm, to each row of U_k, instead of to the p-dimensional original points. The cluster assigned to row i of U_k is then assigned to X_i. Many variants of this method exist. The K-means algorithm as the clustering step could be replaced by other clustering methods, and indeed, in this paper, we rely on a different algorithm.

For an intuitive understanding of why spectral methods work, it is helpful to turn to graph theory. A graph $G = (V, E)$ is a set of nodes V and an edge set E. The edge set contains pairs of nodes that are connected by a potentially weighted edge.

An adjacency matrix A representing a graph is a matrix with entries A_{ij} equal to the weight of the edge (i, j). Let $L = D - A$, where D is the diagonal matrix with entries D_{ii} equal to the sums of row i of A. L is called the Laplacian matrix, and it can be shown that if G has k connected components, then L has an eigenvalue of 0 with multiplicity k, and that the eigenvectors of the 0 eigenvalue are constant on the indices corresponding to the k connected components of the graph. Thus, performing k-means on the $n \times k$ matrix whose columns are the first k eigenvectors of L - i.e. the eigenvectors corresponding to the 0 eigenvalue - will perfectly separate the connected components of the graph.

We now extend this heuristic to real data. If the data lie, for instance, on a low-dimensional manifold, then the Laplacian of the kernel matrix should capture local information about the manifold. Indeed, [Belkin and Niyogi, 2008] show a connection between the graph Laplacian and the Laplace-Beltrami operator, which is known to provide information about a manifold's connectivity. And [von Luxburg et al., 2008] show consistency of spectral clustering by demonstrating that the eigenvectors of the Laplacian matrix converge uniformly to the eigenfunctions of the Laplacian operator.

A more realistic extension of the low-dimensional manifold setting is one we consider, in which the data are a combination of low-dimensional signal and high-dimensional noise. In this setting, practical algorithms have been developed; see e.g. [Niu et al., 2001] for an algorithm and a performance comparison among several popular methods. The authors of this paper point out that spectral clustering may be sensitive to noisy, irrelevant dimensions in the data, and their algorithm iterates between spectral clustering and searching for the "best" projection of the data onto a low-dimensional space.

There are many ways to formally justify spectral clustering; one method comes from perturbation theory. The graph adjacency matrix A, or the kernel matrix $\mathcal{K}$, is considered to be a noisy, or perturbed, version of a matrix P that is ideal for

clustering. The bound on the mis-clustering rate of spectral clustering then depends crucially on the operator norm of $A - P$.

To our knowledge, [Karoui, 2010a] is the first to develop theory about the kernel random matrix in the signal-plus-noise framework. They compare the kernel matrix formed from the signal-pluse-noise to the pure signal kernel matrix. In this paper and in [Karoui and tieng Wu, 2014], the bounds suggest that the distance between the manifolds, d, must grow with the data dimension, p, in order for spectral clustering to result in few errors on these sorts of data. Yet one of the reasons for the popularity of spectral clustering is that the method is fairly effective in practice even when p is large relative to d.

In this work, we develop a new kernel spectral method that achieves an error rate of $Ck^2 d_{\max}^2/d_{\min}^4$, where $C > 0$ is a constant, thus showing that it does not grow with p. Our bound requires only the weak assumption that $p \leq Cn/(k \log^2(n/k))$, rather than d. Thus we show that spectral clustering can work even if p is large relative to d. To obtain our bound, we propose a new variant of kernel spectral clustering, one that relies on the low-rank approximation of the kernel matrix, along with performing K-lines, instead of K-means, on the resulting low-dimensional matrix. We rely on perturbation theory methods and analyze $\|A - P\|$, but we carefully choose P so that it is both an ideal clustering matrix and is as close as possible to A. This is in contrast to the approach of other authors, who typically compare A to $\mathbb{E}A$ or to the pure signal matrix.

We moreover develop and theoretically justify an algorithm that improves upon spectral clustering by using a dimension-reduction and testing procedure. Our algorithm achieves an error rate that is exponential in d. The algorithm's testing procedure is inspired by the testing procedures of [Zhang and Zhou, 2015]. We provide both theoretical justification for the algorithm, rigorously proving that it achieves the exponential rate. We demonstrate its success on simulated and real data.

This paper is organized as follows. We introduce the model and algorithm in 3.2. We state the major theoretical results in 3.3. Proofs of the main results and key lemmas used are in 3.4.

Notation We now introduce some notation for this section only. We occasionally must refer to points according to the manifold they are generated from. In this case, we write $X_{t,i} = M_{t,i} + Z_{t,i}$ where for $i = 1, \ldots, n/k$, $M_{t,i}$ is uniformly distributed on S_t.

Let $Z_{t,(1)} \leq \ldots \leq Z_{t,(n/k)}$ be the Gaussian noise random variables, sorted according to ℓ_2 norm, from S_t. Since the manifolds are of the same size, note that $\mathbb{E}\left\|Z_{s,(i)}\right\|^2 = \mathbb{E}\left\|Z_{t,(i)}\right\|^2$ for all $t \neq s$. So we often write $\mathbb{E}\left\|Z_{(i)}\right\|^2$, without the other subscript. Let M_{ti} denote the uniform random variables from S_t *associated* with the order statistics $Z_{t,(i)}$. We never sort the uniform random variables themselves.

We write $\|v\|$ for the Euclidean norm of a vector v. We write $\|B\|_{OP}$ and $\|B\|_F$ for the operator and Frobenius norms, respectively, of a matrix B. Let $A_{[i,]}$ denote the ith row of the matrix A.

For a matrix $B \in \mathbb{R}^{n \times p}$, we often make use of the r-rank approximation of B. By this we mean the following. B can be written $B = \sum_{i \leq \min(n,p)} \lambda_i u_i v_i'$. For $r \leq \min(n, p)$, the r-rank approximation of B is defined as

$$\sum_{i=1}^{r} \lambda_i u_i v_i'$$

Many of our matrices will consist of k^2 sub-matrices, each in $\mathbb{R}^{(n/k)\times(n/k)}$. For any such matrix B, we write $(B)_{ij}^{st}$ to indicate the (i,j)th entry of the (s,t) sub-matrix of B. Similarly, for any matrix, we write $(B)_{ij}$ to indicate the (i,j) entry of the matrix B.

Define the Euclidean Distance Kernel matrix:

$$A_{ij}^{(EK)} := \|X_i - X_j\|^2$$

Throughout, we will use the Gaussian Kernel random matrix A with entries:

$$A_{ij}^{(GK)} := \exp\left(-\frac{A_{ij}^{(EK)}}{\rho^2}\right).$$

This kernel is widely used by data analysts and in the literature. However, our result is based almost entirely on work for the Euclidean Distance kernel that is then extended to the Gaussian Kernel result via a Taylor expansion, imitating [Karoui, 2010b].

Now let the matrix $P^{(EK)} \in \mathbb{R}^{n\times n}$ be the matrix consisting of k^2 sub-matrices, each in $\mathbb{R}^{(n/k)\times(n/k)}$, with

$$\begin{aligned}\left(P^{(EK)}\right)_{ij}^{st} &= \mathbb{E}\left\|Z_{s,(i)}\right\|^2 + \mathbb{E}\left\|Z_{t,(j)}\right\|^2 + 4d_{st}^2\mathbf{1}\{s \neq t\} \\ &= \mathbb{E}\left\|Z_{(i)}\right\|^2 + \mathbb{E}\left\|Z_{(j)}\right\|^2 + 4d_{st}^2\mathbf{1}\{s \neq t\}\end{aligned}$$

Since the expectations of the sorted random variables on each manifold are the same. And define,

$$\left(P^{GK}\right)_{ij}^{st} = \exp\left(-\frac{\left(P^{(EK)}\right)_{ij}^{st}}{\rho^2}\right).$$

And let $R^{(GK)} := A^{(GK)} - P^{(GK)}$ and define $R^{(EK)}$ analogously. We will prove our bound on the error rate when we perform spectral clustering on $R^{(GK)}$ using the relation between the Gaussian and Euclidean kernels and a Taylor expansion.

Throughout, denote the χ_p^2 density, distribution, and tail probability functions by f, F, and $\bar{F}$, respectively. Let the hazard rate be denoted by h, where $h = f/\bar{F}$. And let $\{Y_i\}_{i=1}^n \sim_{iid} \chi^2(p)$. Let $\{Y_{(i)}\}_{i=1}^n$ be the order statistics of this sample, with

$Y_{(1)} \leq Y_{(2)} \leq \ldots Y_{(n)}$. Let $S := \sum_{i=1}^{n} \left(Y_{(i)} - \mathbb{E}Y_{(i)}\right)^2$.

In the refinement algorithm, we perform estimation of our manifold points. In this context, though the original points M_i are random, we think of them as fixed, and write μ_i. We let these be the entries of M.

We say that an algorithm achieves a "perfect clustering" on data points if the result of the algorithm is an assignment $\hat{\tau}$ such that there exists a permutation $\pi : [k] \to [k]$ such that $\pi(\hat{\tau}) = \tau$), where τ is the true cluster assignment.

We make use of the following assumptions in our results.

$$d_{\min}^2 \gtrsim \max\left(rk\left(1 + \frac{p}{n}\right), 32r\sigma^2\right) \text{ (distance assumption)}$$

$$n \gtrsim \max\left(\frac{2k\left(d^2 - 16\sigma^2\right)}{128\sigma^2}, 32\log p + 18\right) \text{ (sample size assumption)}$$

$$p \geq 3 \text{ (ambient dimension assumption)}$$

$$\rho \geq \max\left(p + 2\sqrt{3p\log n} + 6\log n, p\log\frac{n}{k}, d_{\min}, \frac{d_{\max}^2}{\sqrt{k}}\right) \text{ (bandwidth assumption)}$$

3.2 Algorithms

The first algorithm is an extension of Lloyd's algorithm for the K-means problem to linear subspaces. We follow [Vidal, 2009] for our exposition. Suppose we have data points $P_1, \ldots, P_n \in \mathbb{R}^n$, and suppose the points truly lie on k different q-dimensional linear spaces. We wish to find τ and the orthogonal projections of the points onto their respective subspaces. The optimization is:

$$\min_{\tau \in [k]^n} \min_{\{H_j\}_{j \leq k}} \sum_{j \leq k} \sum_{\{i : \tau(i) = j\}} \|P_i - H_j P_i\|_2^2$$

For a data matrix P, let $P_{C_j} \in \mathbb{R}^{n_j \times r}$ be the points assigned to cluster j. In our setting, we will apply k-spaces to the rows of k-rank approximation of $A^{(GK)}$. Call this approximation $\hat{P}$. Notice that $\hat{P}$ has n rows, each of dimension n. But as Lemma

3.4.3 shows, each set of rows associated with one cluster is in fact rank 1. So we apply K-spaces with $q = 1$.

Algorithm 3: K-spaces spectral clustering algorithm for the Gaussian Kernel.

Input: Data points $P_1, \ldots, P_n$ arranged in a data matrix $P \in \mathbb{R}^{n \times n}$, number of communities k

Output: clustering result $\hat{\tau}^{init} \in [k]^n$

Initialize $\hat{\tau}^0$;

while *not converged* **do**

$\hat{C}_j^t = \{i : \hat{\tau}^t(i) = j\}$;

$H_j^t = \hat{V}_{\hat{C}_j^t} \hat{V}_{\hat{C}_j^t}^T$ where $P^q_{\hat{C}_j^t} = U_{\hat{C}_j^t} D_{\hat{C}_j^t} V^T_{\hat{C}_j^t}$ is the rank-1 approximation of the nodes in C_j^t ;

$\hat{\tau}^{t+1}(i) = \text{argmin}_j \left\| \hat{P}_i - H_j^t \hat{P}_i \right\|_2^2$;

end

The second algorithm proceeds as follows. We perform an initial clustering, then estimate the manifolds using dimension reduction. Using this manifold estimate, we classify the points of S_1 based on the estimated manifold that they are closest to. Note that we use sample-splitting for the algorithm; the main reason is that it is mathematically tractable; we are able to obtain a sharp bound despite having an initialization that only has $o(1)$ error. In our real data example, we use a single global initializer, then refine on each node. This may work as well or better than the sample-splitting algorithm, but it is harder to analyze. There are several parameters that must be chosen in the algorithm, including ρ, the denominator for the Gaussian kernel, K for the K-nearest neighbors, and more. We propose selecting the parameters

via cross-validation.

Algorithm 4: Refinement Algorithm.

Input: data points $\{X_i\}_{i=1}^n$, number of communities k, effective dimensionality r, and the minimum distance between manifolds $d_{\min}$

Output: clustering result $\hat{\tau} \in [k]^n$

1 **Data splitting** ;

Denote $X^{(1)} = \{X_1, X_2, \ldots, X_{n/2}\}$ and $X^{(2)} = \{X_{n/2+1}, X_{n/2+2}, \ldots, X_n\}$;

2 **Manifold Estimation:** ;

For $j = 1, 2$, let W_j be the matrix containing the first r right-singular vectors of $X^{(j)}$;

For $j = 1, 2$, let $\mathcal{P}_{W_j} = W_j W_j'$;

Let $\hat{M}^{(1)} = X^{(1)}\mathcal{P}_{W_2}$ and $\hat{M}^{(2)} = X^{(2)}\mathcal{P}_{W_1}$;

Let $\hat{M} = [\hat{M}^{(1)}, \hat{M}^{(2)}]^T$. Let $\hat{\mu}_1, \ldots, \hat{\mu}_n$ be the rows of $\hat{M}$;

3 **Initialization** ;

Apply Algorithm Algorithm 3 on $X^{(2)}$ to obtain $\hat{\tau}^{init}(i)$ on $i = n/2 + 1, \ldots, n$;

4 **Refinement:** for $i = 1, 2, \ldots, n/2$, let

$$\hat{\tau}(i) = \underset{s \in [k]}{\operatorname{argmax}} \sum_{\{j : \hat{\tau}^{init}(j) = s\}} 1\{\|\hat{\mu}_i - \hat{\mu}_j\| \le \delta\}, \text{ where } \delta = d_{\min}/4$$

5 **Consensus:** repeat Steps 3-4 but exchange the role of $X^{(1)}$ and $X^{(2)}$; obtain $\hat{\tau}(i)$ for $i = n/2 + 1, \ldots, n$. Combine them together to obtain $\hat{\tau}$.

3.3 Major Results

The error rate of spectral clustering depends on the spectral norm of $A - P$, where typically P is chosen to be a matrix with ideal clustering properties. Typically, $P = EA$ or $P_{ij} = K(\mu_i, \mu_j)$, the signal kernel matrix. If our kernel is the Euclidean distance kernel, then it might be reasonable to imitate [Karoui, 2010a] and use for P

the matrix with entries

$$P_{ij} = \mathbb{E}\left\|Z_i - Z_j\right\|^2 + \left\|M_i - M_j\right\|_2^2 = 2p + \left\|M_i - M_j\right\|_2^2$$

Notice that

$$(A - P)_{ij} \leq \left\|Z_i\right\|^2 - p + \left\|Z_j\right\|^2 - p + Z_i' Z_j$$

Let B be the $n \times n$ matrix with entries $B_{ij} = \left\|Z_i\right\|^2 - p$. Then it is rank 1, so has spectral equal to its Frobenius norm, which is approximately $n^2 p$, since $\left\|Z_i\right\|^2 \backsim \chi_p^2$. This would result in an error rate of about $\frac{k^2}{d^4} p \log n$ for spectral clustering, which requires that p be smaller than d^4. Our method improves upon this by choosing a different P and a different method of clustering.

Theorem 3.3.1. *Suppose the ambient dimension and sample size assumptions hold. Then there is a $c > 0$ such that with probability at least $1 - \frac{1}{n} - \frac{1}{\log^2(n/k)}$,*

$$\left\|R^{(EK)}\right\|_{OP}^2 \leq c(n^2 + p^2 + n^2 b^4 + n^2 d_{\max}^2 + knp \log^2(n/k))$$

Theorem 3.3.2. *Suppose the ambient dimension, sample size, and bandwidth assumptions hold. Let $\hat{\tau}^{init}$ be the clustering obtained fromAlgorithm Algorithm 3 to a dataset of n data points of the form given in (3.1.1). Then there is a constant $c > 0$ such that with probability at least $1 - \frac{2}{n} - \frac{1}{k \log^2(n/k)}$,*

$$\frac{1}{n}\mathcal{L}(\tau, \hat{\tau}^{init}) \leq c\frac{k^2}{d_{\min}^4}\left(1 + \frac{p^2}{n^2} + d_{\max}^2 + b^4 + \frac{kp \log^2(n/k)}{n}\right)$$

Theorem 3.3.3. *Suppose the sample size, ambient dimension, and distance assumptions hold. Let $\hat{\tau}$ be the estimator of τ obtained from Algorithm Algorithm 4. Then*

with probability at least $1 - 1/n - 1/k \log^2(n/k) - \exp\left(-(d^2 - 16\sigma^2)/256\sigma^2\right)$,

$$\frac{1}{n}\mathcal{L}(\tau, \hat{\tau}) \leq 4k \exp\left(-\frac{d^2 - 16\sigma^2 r}{256\sigma^2}\right)$$

3.4 Proofs of the main results

3.4.1 Key Lemmas

Lemma 3.4.1. *Suppose* $P \in \mathbb{R}^{n\times n}$ *has rank* k. *Suppose applying Algorithm Algorithm 3 to* P *yields a perfect cluster assignment. Let the minimum Euclidean distance between two rows of* P *corresponding to different clusters be bounded below by* α^2. *Let* A_k *be the* $k-$*rank approximation of a matrix* $A \in \mathbb{R}^{n\times n}$. *Then Algorithm 3 on the rows of* A_k *results in a* $\hat{\tau}$ *that satisfies*

$$\mathcal{L}(\tau, \tilde{\tau}) \precsim \frac{k \left\|A - P\right\|^2}{\alpha^2}$$

Proof. Define:

$$T := \left\{ i : \left\|(A_k)_{[i,]} - P_{[i,]}\right\|_2^2 < \frac{\alpha^2}{2} \right\}$$

For nodes outside of T, Algorithm Algorithm 3 is not guaranteed to assign the nodes to the correct cluster. Thus

$$\mathcal{L}(\tau, \hat{\tau}) \leq |T^C| = \#\left\{ i : \left\|(A_k)_{[i,]} - P_{[i,]}\right\|_2^2 \geq \frac{\alpha^2}{2} \right\}$$

And $\|A_k - P\|_F^2 = \sum_{i \le n} \left\|(A_k)_{[i,]} - P_{[i,]}\right\|_2^2$, so

$$\begin{aligned} |T^C| &\le \frac{2\|A_k - P\|_F^2}{\alpha^2} \\ &\le \frac{2 * 8k\|A - P\|^2}{\alpha^2} \qquad \text{by Lemma 3.5.8 .} \end{aligned}$$

□

Lemma 3.4.2. *Suppose ρ satisfies the bandwidth assumptions. Let α be the the minimum Euclidean distance between two rows of $P^{(GK)}$. Then there is a constant $c > 0$ such that*

$$\alpha^2 \ge c\frac{n}{k}\frac{d_{\min}^4}{\rho^4}$$

Proof. Suppose without loss of generality that $d_{\min} = d_{12}$. Then

$$\begin{aligned} \alpha^2 &= 2\sum_{j=1}^{n/k}\left(e^{-\frac{\mathbb{E}\left\|Z_{(i)}\right\|^2 + \mathbb{E}\left\|Z_{(j)}\right\|^2}{\rho^2}}\left(e^{-\frac{d^2}{\rho^2}} - 1\right)\right)^2 + \sum_{s,t}\sum_{j=1}^{n/k}\left(e^{-\frac{\mathbb{E}\left\|Z_{(i)}\right\|^2 + \mathbb{E}\left\|Z_{(j)}\right\|^2}{\rho^2}}\left(e^{-\frac{d_{1s}^2}{\rho^2}} - e^{-\frac{d_{2t}^2}{\rho^2}}\right)\right)^2 \\ &\ge \frac{2n}{k}\left(e^{-\frac{2\mathbb{E}\left\|Z_{(n/k)}\right\|^2}{\rho^2}}\right)^2\left(e^{-\frac{d^2}{\rho^2}} - 1\right)^2 \\ &\ge \frac{2n}{k}e^{-\frac{4\mathbb{E}\left\|Z_{(n/k)}\right\|^2}{\rho^2}}\left(\frac{d^2}{\rho^2}\right)^2 \text{ since } e^{-x} \ge 1 - x \end{aligned}$$

By Lemma 3.5.14, $\mathbb{E}\left\|Z_{(n/k)}\right\|^2 \le p\log(n/k)$. Thus $e^{-4\mathbb{E}\left\|Z_{(n/k)}\right\|^2/\rho^2}$ is bounded below by a constant. □

Lemma 3.4.3. *$P^{(GK)}$ is rank k, and Algorithm 3 yields a perfect cluster assignment when applied to $P^{(GK)}$.*

Proof. Recall that $P^{(GK)}$ is composed of k^2 sub-matrices in $\mathbb{R}^{(n/k)\times(n/k)}$. Divide $P^{(GK)}$ into the first $\frac{n}{k}$ rows, the second $\frac{n}{k}$ rows, and so on. Any one of these blocks of rows

is rank 1, since:

$$P^{(GK)}_{[j,]} = \exp\left(\frac{\mathbb{E}\left\|Z_{(i)}\right\|^2 - \mathbb{E}\left\|Z_{(j)}\right\|^2}{\rho^2}\right) * P^{(GK)}_{[i,]}$$

The rows between these blocks are linearly independent, so $P^{(GK)}$ has rank k. And each such cluster of points lies on a 1-dimensional subspace of $\mathbb{R}^n$. Now a sufficient condition for Algorithm 3 to achieve a perfect clustering is that there is a positive distance between the subspaces the points belong to. By Lemma 3.4.2, the distance between the subspaces in $P^{(GK)}$ is at least α^2, a positive constant. □

Lemma 3.4.4. *Consider any $i \in \{1, \ldots, n/2\}$. Let $X_i = \mu_i + Z_i$, and let $\mathcal{P}_{W_2}$ be as in Algorithm 4. Then there is a positive constant C such that with probability at least $1 - 1/n$,*

$$\|\mu_i - \mathcal{P}_{W_2}\mu_i\|_2^2 \leq Ckmr\left(1 + \frac{p}{n} + \beta^2\right)$$

A similar bound holds for $\|\mu_i - \mathcal{P}_{W_1}\mu_i\|_2^2$ when $i \in \{n/2, \ldots, n\}$.

Proof. Assume without loss of generality that $i \in \{1, \ldots, n/2\}$. Let W_2 be as in Algorithm 4, and for simplicity, write $\mathcal{P}_W = W_2 W_2'$, the projection operator from the second sample. Recall that $V \in \mathbb{R}^{n \times p}$ is the matrix of the km unique centers that approximate the manifolds. Now,

$$\begin{aligned}\|V - \mathcal{P}_W V\|_F^2 &\precsim \left\|V - M^{(2)}\right\|_F^2 + \left\|M^{(2)} - \mathcal{P}_W M^{(2)}\right\|_F^2 + \left\|\mathcal{P}_W M^{(2)} - \mathcal{P}_W V\right\|_F^2 \\ &\leq 2\left\|V - M^{(2)}\right\|_F^2 + \left\|M^{(2)} - \mathcal{P}_W M^{(2)}\right\|_F^2\end{aligned}$$

And,

$$\left\|M^{(2)} - \mathcal{P}_W M^{(2)}\right\|_F^2 \precsim \left\|M^{(2)} - \mathcal{P}_W X^{(2)}\right\|_F^2 + \left\|\mathcal{P}_W X^{(2)} - \mathcal{P}_W M^{(2)}\right\|_F^2 \tag{3.4.1}$$

$$\precsim r\left\|M^{(2)} - X^{(2)}\right\|_2^2 + r\left\|\mathcal{P}_W X^{(2)} - \mathcal{P}_W M^{(2)}\right\|_2 \tag{3.4.2}$$

$$\precsim 2r\left\|M^{(2)} - X^{(2)}\right\|_2^2 \tag{3.4.3}$$

$$\precsim 2r(p+n) \tag{3.4.4}$$

Step (3.4.2) follows from Lemma 3.5.8 and the fact that $\mathcal{P}_W X^{(2)} - \mathcal{P}_W M^{(2)}$ is rank r, so its Frobenius norm is bounded by r times its operator norm. And Step (3.4.3) follows since $\left\|\mathcal{P}_W X^{(2)} - \mathcal{P}_W M^{(2)}\right\|_2^2 \leq \|\mathcal{P}_W\|_2^2 \left\|X^{(2)} - M^{(2)}\right\|_2^2$, and the operator norm of an orthogonal projection matrix is 1. The final step is with probability at least $1 - 1/n$, and follows from Lemma 3.5.15. Now we have

$$\begin{aligned}\|V - \mathcal{P}_W V\|_F^2 &\precsim \left\|V - M^{(2)}\right\|_F^2 + r(p+n)\\ &\leq n\beta^2 + r(p+n)\end{aligned}$$

And V has only km unique rows, $v_1, \ldots, v_{km}$, each repeated $n/(km)$ times. Thus $\|V - \mathcal{P}_W V\|_F^2 = \sum_{j=1}^{km} \frac{n}{km} \|v_j - \mathcal{P}_W v_j\|_2^2$, which implies that

$$\sum_{j=1}^{km} \|v_j - \mathcal{P}_W v_j\|_2^2 \precsim \frac{kmn\beta^2}{n} + \frac{kmr(p+n)}{n}$$

And,

$$\begin{aligned}\|\mu_1 - \mathcal{P}_W \mu_1\|_2^2 &\precsim \|\mu_1 - v_1\|_2^2 + \|v_1 - \mathcal{P}_W v_1\|_2^2 + \|\mathcal{P}_W v_1 - \mathcal{P}_W \mu_1\|_2^2\\ &\precsim 2\|\mu_1 - v_1\|_2^2 + \|v_1 - \mathcal{P}_W v_1\|_2^2\\ &\precsim 2\beta^2 + km\beta^2 + (kmr)\left(1 + \frac{p}{n}\right)\end{aligned}$$

And the proof for an $i \in \{n/2+1, \dots n\}$ would be exactly similar, using $W_1 W_1'$ instead of $W_2 W_2'$. □

3.4.2 Proofs of the main theorems

Proof of Theorem 3.3.1. Define the following matrices:

$$\begin{aligned}
R^{(1)} &= \left(\left\| Z_{s,(i)} \right\|^2 - \mathbb{E} \left\| Z_{(i)} \right\|^2 \right)_{ij}^{st} \\
R^{(2)} &= \left(\left\| Z_{t,(j)} \right\|^2 - \mathbb{E} \left\| Z_{(j)} \right\|^2 \right)_{ij}^{st} \\
R^{(3)} &= \left(Z_i' Z_j \right)_{ij} \\
R^{(4)} &= \left(\langle Z_i - Z_j,\, M_i - M_j \rangle \right)_{ij} \\
R^{(5)} &= \left(\|M_{si} - M_{tj}\|_2^2 - d_{st}^2 \mathbf{1}\{s \neq t\} \right)_{ij}^{st}
\end{aligned}$$

We now obtain high-probability bounds on each of the operator norms of these matrices. First,

$$\left\| A^{(EK)} - P^{(EK)} \right\|_2^2 \leq 4 \left(2 \left\| R^{(1)} \right\|_2^2 + \left\| R^{(3)} \right\|_2^2 + \left\| R^{(4)} \right\|_2^2 + \left\| R^{(5)} \right\|_2^2 \right)$$

since $R^{(1)}$ and $R^{(2)}$ have the same norms. Notice that for $R^{(3)}$ and $R^{(4)}$, we have dropped the manifold and order statistic notation since the matrix with entries $\left(Z_i' Z_j \right)_{ij}$ has the same operator norm as the matrix with entries $\left(Z_{s,(i)}' Z_{t,(j)} \right)_{ij}^{st}$, and similarly for the inner product matrix. Now, $R^{(1)}$ has rank 1, so its operator norm is equal to its Frobenius norm. We have:

$$\begin{aligned}
\left\| R^{(1)} \right\|_F^2 &= \sum_{t=1}^{k} \sum_{s=1}^{k} \frac{n}{k} * \left(\sum_{j=1}^{n/k} \left(\left\| Z_{s,(j)} \right\|^2 - \mathbb{E} \left\| Z_{s,(j)} \right\|^2 \right)^2 \right) \\
&= k^2 \frac{n}{k} \left(\sum_{j=1}^{n/k} \left(\left\| Z_{s,(j)} \right\|^2 - \mathbb{E} \left\| Z_{s,(j)} \right\|^2 \right)^2 \right)
\end{aligned}$$

since the noise distribution is the same on all the manifolds. Thus

$$\mathbb{E}\left\|R^{(1)}\right\|_F^2 = kn\mathbb{E}\sum_{j=1}^{n/k}\left(\left\|Z_{s,(j)}\right\|^2 - \mathbb{E}\left\|Z_{s,(j)}\right\|^2\right)^2 \precsim knp\log(n/k). \text{ by (3.5.7)}$$

And,

$$\begin{aligned}
Var\left\|R^{(1)}\right\|_F^2 &= k^2\sum_{s,t=1}^{k} Var\left(\left\|R_1^{st}\right\|_F^2\right) \text{ since the sub-matrices contain independent entries} \\
&= k^4 Var\left\|R_1^{st}\right\|_F^2 \\
&= k^4\frac{n^2}{k^2}Var\left(\sum_{j=1}^{n/k}\left(\left\|Z_{s,(j)}\right\|^2 - \mathbb{E}\left\|Z_{s,(j)}\right\|^2\right)^2\right) \\
&\precsim k^2n^2p^2\log^2(n/k). \text{ by (3.5.8)}
\end{aligned}$$

Using these bounds, along with the Markov Inequality,

$$\begin{aligned}
\mathbb{P}\{\left|\,\|R_1\|_F^2 - \mathbb{E}\,\|R_1\|_F^2\,\right| > knp\log^2(n/k)\} &\leq \frac{Var\,\|R_1\|_F^2}{k^2n^2p^2\log^4(n/k)} \\
&\leq \frac{k^2n^2p^2\log^2(n/k)}{k^2n^2p^2\log^4(n/k)} \\
&= \frac{1}{\log^2(n/k)}
\end{aligned}$$

Now $R^{(3)} = ZZ^\top$. And by a slight adjustment of the analysis in Theorem 5.39 of [Vershynin, 2012],

$$\|Z\| \precsim \sqrt{n} + \sqrt{p} + \log n, \qquad \text{with probability at least } 1 - \frac{1}{n}.$$

Now $\left\|ZZ^\top\right\| \leq \|Z\|^2 \precsim n + p + \log^2 n$. So

$$\left\|ZZ^\top\right\|^2 \precsim n^2 + p^2 + \log^2 n, \qquad \text{with probability at least } 1 - \frac{1}{n}.$$

For $R^{(4)}$, if we condition on $M_i - M_j, \langle Z_i - Z_j,\, M_i - M_j\rangle \sim N\left(0, \|M_i - M_j\|_2^2\right)$. So

$$\begin{aligned}\mathbb{P}\{\left\|R^{(4)}\right\|_2^2 > t\} \le \mathbb{P}\{\left\|R^{(4)}\right\|_F^2 > t\} &\le n^2\mathbb{P}\{N(0,1) > t/2n^2(b^2 + d_{\max}^2)\} \\ &\le n^2 \exp\left(-\frac{t^2}{4n^4(b^4 + d_{\max}^4)}\right)\end{aligned}$$

which leaves us with a bound on the squared operator norm of $n^2(b^2 + d^2)\log(n^2/\delta)$ with probability at least $1 - \delta$. To analyze $R^{(5)}$, we recall that

$$\|M_{si} - M_{tj}\|_2^2 \le 4\left(b_s^2 + d_{st}^2 + b_t^2\right) \le 8b^2 + 4d_{st}^2$$

So

$$\left(\|M_{si} - M_{tj}\|_2^2 - 4d_{st}^2\mathbf{1}\{s \neq t\}\right)^2 \le 64b^4$$

□

Proof of Theorem 3.3.2.

$$\begin{aligned}\frac{1}{n}L(\tau, \tilde{\tau}) &\precsim \frac{k\left\|R^{(GK)}\right\|^2}{nr^2} && (3.4.5)\\ &\le \frac{k^2\rho^4}{n^2 d_{\min}^4}\left\|R^{(GK)}\right\|^2 && (3.4.6)\\ &\precsim \frac{k^2\rho^4}{n^2 d_{\min}^4}\left(\frac{\left\|R^{(EK)}\right\|^2}{\rho^4} + \frac{n^2}{\rho^4} + \frac{n^2 b^8}{\tau^8}\right) && (3.4.7)\\ &\precsim \frac{k^2\rho^4}{n^2 d_{\min}^4}\left(\frac{n^2 + p^2 + n^2(b^4 + d_{\max}^2 + b^2) + nkp\log(n/k)}{\rho^4} + \frac{n^2}{\rho^4} + \frac{n^2(b^8 + d_{\max}^8)}{\rho^8}\right) && (3.4.8)\\ &= \frac{k^2}{d_{\min}^4}\left(c + \frac{p^2}{n^2} + b^4 + d_{\max}^2 + b^2 + \frac{pk\log(n/k)}{n}\right) && (3.4.9)\end{aligned}$$

for a constant $c > 0$, where (3.4.5) follows from Lemma 3.4.1 and Lemma 3.5.8, (3.4.6) follows from Lemma 3.4.2, (3.4.7) follows from Lemma 3.5.9, and (3.4.8) is

by Theorem 3.3.1. □

Proof of Theorem 3.3.3. Let γ^{init} be the error rate from Theorem 3.3.2. Let $\alpha = Ckmr\,(1 + p/n + \beta^2)$. Define the following sets:

$$B = \left\{ \forall i \in \{1, \dots, n/2\}, \|\mu_i - \mathcal{P}_{W_2}\mu_i\|_2^2 \leq \alpha \text{ and } \forall i \in \{n/2+1, \dots, n\}, \|\mu_i - \mathcal{P}_{W_1}\mu_i\|_2^2 \leq \alpha \right\}$$

$$D^{init} = \left\{ i \in \{n/2+1, \dots, n\} : \hat{\tau}^{init}(i) = \tau(i) \right\}$$

$$E^{init} = \left\{ |\left(D^{init}\right)^C| \leq n\gamma^{init} \right\}$$

Let $E := B \cap E^{init}$. By Theorem 3.3.2 and Lemma 3.4.4

$$\mathbb{P}\{E^C\} \leq \mathbb{P}\{B^C\} + \mathbb{P}\{\left(E^{init}\right)^C\} \leq \frac{1}{n} + \frac{1}{k\log^2(n/k)} \tag{3.4.10}$$

By Lemma 3.5.17, we have only to analyze $(k-1)\max_{j\neq\tau(i)} \mathbb{P}\{\hat{\tau}(i) = j \text{ and } E\}$ for a single node i. The error rate will be the same for all $i \in [n]$, so for simplicity, we let $i = 1$. Assume without loss of generality that $\tau(1) = 1$ and that the probability is the same for all j; we will let $j = 2$ below. Let $B_\delta(x)$ be the ball of radius δ around a point x. Define

$$Y_i = \mathbf{1}\{\hat{\mu}_i \in B_\delta(\hat{\mu}_1)\}$$

Now we can bound $\mathbb{P}\{\hat{\tau}(i) = 2 \text{ and } E\}$ above by

$$\mathbb{P}\left\{ \sum_{i\in\hat{C}_2} Y_i - \sum_{i\in\hat{C}_1} Y_i > 0 \text{ and } E \text{ and } \|\mathcal{P}_{W_2}Z_1\|_2 + \|\mathcal{P}_{W_1}Z_i\|_2 \leq d/4 \right\} + 2\mathbb{P}\{\chi_r^2 > d^2/64\sigma^2\}$$

The bound of $2\mathbb{P}\{\chi_r^2 > d/64\sigma^2\}$ follows because both $\|\mathcal{P}_W Z_1\|_2^2$ and $\|\mathcal{P}_W Z_i\|_2^2$ are

distributed as $\sigma^2\chi_r^2$. And by Lemma 3.5.12.

$$\mathbb{P}\{\chi_r^2 \geq d^2/64\sigma^2\} \leq \exp\left(-\frac{d^2 - 64\sigma^2 r}{512\sigma^2}\right)$$

We use the sub-exponential part of the chi-square tail since $d^2/64\sigma^2 \geq 2r$. Let $i \in \hat{C}_2^{init} \cap D^{init}$. To upper bound $\mathbf{1}\{\|\hat{\mu}_i - \hat{\mu}_1\|_2 \leq \delta\}$, we lower bound $\|\hat{\mu}_i - \hat{\mu}_1\|_2$. Now

$$\begin{aligned}
\|\hat{\mu}_i - \hat{\mu}_1\|_2 &= \|\mathcal{P}_{W_1}\mu_i + \mathcal{P}_{W_1}Z_i - \mu_i + \mu_i - \mu_1 + \mu_1 - \mathcal{P}_{W_2}\mu_1 - \mathcal{P}_{W_2}Z_1\|_2 \\
&\geq \|\mu_i - \mu_1\|_2 - \|\mu_i - \mathcal{P}_{W_1}\mu_i\|_2 - \|\mu_1 - \mathcal{P}_{W_2}\mu_1\|_2 - \|\mathcal{P}_{W_1}Z_i\|_2 - \|\mathcal{P}_{W_2}Z_1\|_2 \\
&\geq \|\mu_i - \mu_1\|_2 - 2\sqrt{Ckmr\left(1 + p/n + \beta^2\right)} - d/4,
\end{aligned}$$

where the last step follows by Lemma 3.4.4 and by the fact that $\|\mathcal{P}_{W_2}Z_1\|_2, \|\mathcal{P}_{W_1}Z_i\|_2$ are both bounded above by $d/8$. Now let $i \in \hat{C}_1^{init} \cap D^{init}$. We must lower bound $\mathbf{1}\{\|\hat{\mu}_i - \hat{\mu}_1\| \leq \delta\}$. Similarly to the above, we obtain:

$$\|\hat{\mu}_i - \hat{\mu}_1\| \leq \|\mu_i - \mu_1\| + 2\sqrt{Ckmr\left(1 + p/n + \beta^2\right)} + d/4$$

when $\|\mathcal{P}_{W_2}Z_1\|_2, \|\mathcal{P}_{W_1}Z_i\|_2$ are both bounded above by $d/8$. Now in either case above, if i is not in this set, the above bounds do not hold. But on the set E, there are at most $n\gamma^{init}$ such nodes. Let

$$\begin{aligned}
W_{2i} &= \mathbf{1}\{\|\mu_i - \mu_1\| \leq \delta + 2Ckmr\left(1 + p/n + \beta^2\right) + d/4\} \\
W_{1i} &= \mathbf{1}\{\|\mu_i - \mu_1\| \leq \delta - 2Ckmr\left(1 + p/n + \beta^2\right) - d/4\}
\end{aligned}$$

Now $\mathbb{P}\left\{\sum_{i\in\hat{C}_2} Y_i - \sum_{i\in\hat{C}_1} Y_i > 0 \text{ and } E \text{ and } \|\mathcal{P}_{W_2}Z_1\|_2 + \|\mathcal{P}_{W_1}Z_i\|_2 \leq d/4\right\}$ is upper

bounded by $\mathbb{P}\left\{ \sum_{i\in B_2} W_{2i} - \sum_{i\in B_1} W_{1i} > -n\gamma^{init} \text{ and } E \right\}$. And

$$\mathbb{P}\left\{ \sum_{i\in B_2} W_{2i} - \sum_{i\in B_1} W_{1i} > -n\gamma^{init} \text{ and } E \right\} \leq \mathbb{P}\left\{ \sum_{i\in C_2} W_{2i} - \sum_{i\in C_1} W_{1i} > -2n\gamma^{init} \text{ and } E \right\}$$
$$\leq \mathbb{P}\left\{ \sum_{i\in C_1} W_{1i} < 2n\gamma^{init} \right\}$$

The second step follows because $\sum_{i\in B_2} W_{2i} \leq \sum_{i\in C_2} W_{2i}$ since $B_2 \in C_2$, and $|B_1| \geq |C_1| - n\gamma^{init}$. The final step follows because $\mathbf{1}\{\|\mu_i - \mu_1\| \leq \delta + \sqrt{2Ckmr(1+p/n+\beta^2)} + d/4\} = 0$ for $i \in C_2$ when $\delta \leq d/4$ and $d \geq 2\sqrt{2Ckmr(1+p/n+\beta^2)}$.

Since we have shifted to the non-random sum of the true nodes, $\sum_{i\in C_1} W_{1i} \sim Binomial(n/k, p_1)$, where $p_1 = \max\left(1, \frac{\delta}{b}\right)$. For a random variable Y distributed as $Binomial(n/k, p_1)$, we have the following Chernoff bound for $\lambda \in [0,1]$:

$$\mathbb{P}\{Y < (1-\lambda)\frac{np_1}{k}\} \leq \exp\left(-\frac{\lambda^2 np_1}{k}\right)$$

Recall that we have assumed that $\gamma^{init} = c/k$ for a positive constant c.

$$\mathbb{P}\left\{ Y < n\gamma^{init} \right\} = \mathbb{P}\left\{ Y < \frac{4cn}{k} \right\}$$
$$= \mathbb{P}\left\{ Y \leq \frac{np_1}{k}\frac{4c}{p_1} \right\}$$
$$\leq \exp\left(-\frac{np_1}{k}\left(\frac{p_1 - 4c}{p_1}\right)^2\right)$$

where this Chernoff bound holds because $4c/p_1 \in (0,1)$. Now we can write $p_1 = 1-h$, where h is tiny. Let $c' = (1-h)*\left(\frac{1-h-4c}{1-h}\right)^2$, which is very close to 1. Then our exponent is nc'/k. By our assumption that $n \geq \frac{2k(d^2-16\sigma^2)}{128\sigma^2}$, the desired bound holds. □

3.5 Auxiliary lemmas and proofs

Throughout this section, let $Y_1, \ldots, Y_n \sim_{i.i.d.} \chi^2_p$, and let $Y_{(1)} \leq \ldots \leq Y_{(n)}$ be their order statistics.

3.5.1 Results for order statistics

Lemma 3.5.1. *Assume m is an even integer and $m \leq 6$. Let $n > \max(2m+2, 16m\log p + 4m)$. There is a constant $c > 0$ such that for each i and for any even integer $m \leq 6$,*

$$\mathbb{E}\frac{1}{(h(Y_{(i)}))^m} \leq cp^{m/2}$$

Proof of Lemma 3.5.1. We split the domain of Y into the right tail (III), the middle (II), and the left tail (I). On the right tail, we use Corollary 3.5.6 to bound h. In the middle region, we have a lower bound on h via Lemma 3.5.3. On the left tail, we rely on concentration inequalities for order statistics of the χ^2_p distribution. Let $i \geq n/2$. Let $c_1 > 0$ be as in Lemma 3.5.3 and denote $Y_{(i)}$ by Y to simplify the notation. Now,

$$\mathbb{E}\frac{1}{h(Y)^m} = \underbrace{\mathbb{E}\frac{1}{h(Y)^m}\left\{Y \in [0, p - c_1\sqrt{p})\right\}}_{(I)} + \underbrace{\mathbb{E}\frac{1}{h(Y)^m}\left\{\left|\frac{Y-p}{\sqrt{p}}\right| \leq c_1\right\}}_{(II)}$$
$$+ \underbrace{\mathbb{E}\frac{1}{h(Y)^m}\left\{Y > p + c_1\sqrt{p}]\right\}}_{(III)}.$$

Consider (III). Corollary 3.5.6 says that for some $c_3 > 0$:

$$\begin{aligned}(III) &\leq \mathbb{E}\left(\frac{\bar{F}(Y)}{f(Y)}\right)^m \left\{Y > p + c_1\sqrt{p}\right\} \\ &\leq \mathbb{E}\frac{1}{\left(\frac{1}{2} - \frac{p/2-1}{Y}\right)^m}\left\{Y > p + c_1\sqrt{p}\right\} \\ &\leq \mathbb{E}\frac{1}{\left(\frac{1}{2} - \frac{p/2-1}{p+c_1\sqrt{p}}\right)^m}\left\{Y > p + c_1\sqrt{p}\right\} \\ &\leq c_3 p^{m/2}\end{aligned}$$

For (II), we bound $\bar{F}$ above by 1 and use Lemma 3.5.3 to obtain, for a constant c_2,

$$(II) \leq \mathbb{E}\left(\frac{1}{f(Y)}\right)^m \left\{\left|\frac{Y-p}{\sqrt{p}}\right| \leq c_1\right\} \leq c_2 p^{m/2}. \tag{3.5.1}$$

For (I), we further split this section into three parts:

$$\begin{aligned}(I) &= \mathbb{E}\frac{1}{h(Y)^m}\left\{Y \in (0,1]\right\} + \mathbb{E}\frac{1}{h(Y)^m}\left\{Y \in \left(1, \frac{p}{2}\right)\right\} + \mathbb{E}\frac{1}{h(Y)^m}\left\{Y \in \left(\frac{p}{2}, p - c_1\sqrt{p}\right)\right\} \\ &\triangleq (I_1) + (I_2) + (I_3)\end{aligned}$$

In each, again bound $\bar{F}$ above by 1. For $(I)_1$, we use the density of the ith order statistic and the fact that $F(x) \leq xf(x)$ for $x \in (0,1]$, since the chi-square density is

increasing for $x \leq 1$ when $p \geq 3$.

$$\begin{aligned}
(I)_1 &\leq \int_0^1 \frac{1}{f^m(x)} f(x)(F(x))^{n-i-1}(\bar{F}(x))^i \, dx \\
&\leq \int_0^1 \frac{1}{(f(x))^m} f(x)(xf(x))^{n-i-1} \, dx \\
&\leq \int_0^1 x(f(x))^{n-i-m} \, dx \\
&\leq \int_0^1 x(f(x))^{n/2-m} \\
&\leq 1
\end{aligned}$$

For $(I)_2$, recall that the $f^m(x) = \left(\frac{x^{p-2}e^{-x}}{c_p^2}\right)^{m/2}$ where $c_p = \Gamma(p/2)2^{p/2}$. For $x \in (1, \frac{p}{2}]$, this is bounded below by $\frac{e^{-pm/4}}{c_p^m}$. So:

$$\begin{aligned}
(I)_2 &\leq c_p^m e^{pm/4} \mathbb{P}\left\{ Y \in \left(1, \frac{p}{2}\right) \right\} \\
&\leq c_p^m e^{pm/4} \left(\mathbb{P}\left\{ Y_i \leq \frac{p}{2} \right\} \right)^{i-1} && \text{by Lemma 3.5.13} \\
&\leq c_p^m e^{pm/4} e^{-(i-1)p/16} && \text{since } \mathbb{P}\{Y_i \leq \frac{p}{2}\} \leq e^{-p/16} \text{ by Lemma 3.5.12} \\
&\leq c_p^m e^{pm/4} e^{-p(n/2-1)/16} && \text{since } i \leq n/2 \\
&\leq c_p^m e^{pm/4} e^{-pn/32}
\end{aligned}$$

To bound $c_p^m e^{pm/4}$, recall that $\Gamma(x) \leq x! \leq x^x$ for $x > 2$. So:

$$c_p^m = (\Gamma(p/2))^m \, 2^{pm/2} \leq \left(\frac{p}{2}\right)^{pm/2} 2^{pm/2} = p^{pm/2}$$

Solving for a bound on n:

$$p^{pm/2} e^{pm/4} e^{-pn/32} \leq p^{m/2} \Leftrightarrow n \geq \frac{16m}{p}(p-1)\log p + 4m$$

Thus as long as $n \geq 16m \log p + 4m$, we have the desired conclusion. For $(I)_3$, we rely on Lemma 3.5.3, which tells us that $\frac{1}{(f(x))^m} \precsim p^{m/2} e^{\frac{m(x-p)^2}{2p}}$ if $x \in \left(\frac{p}{2}, p - c_1\sqrt{p}\right)$. So:

$$\begin{aligned}
(I)_3 &\leq p^{m/2} \mathbb{E} e^{\left(\frac{Y-p}{\sqrt{2p}}\right)^2} \left\{ Y \in \left(\frac{p}{2}, p - c_1\sqrt{p}\right) \right\} \\
&\leq p^{m/2} \sum_{j=1}^{\sqrt{p}/c_1} e^{\frac{(-c_1 j)^2}{2}} \mathbb{P}\left\{ Y \in \left(p - (j+1)c_1\sqrt{p}, p - jc_1\sqrt{p}\right] \right\} \\
&\leq p^{m/2} \sum_{j=1}^{\sqrt{p}/c_1} e^{\frac{c_1^2 j^2}{2}} \left(F(p - jc_1\sqrt{p})\right)^{n/2} \\
&\leq p^{m/2} \sum_{j=1} e^{c_1^2 j^2/2} e^{-jn/2} \\
&\leq p^{m/2}
\end{aligned}$$

For $i < n/2$, we divide up the expectation into the same three parts as we did previously. The methods here are essentially the same as those used in the previous part, except in reverse order. From Equation 3.5.1 we have $(II) \leq c_2 p^{m/2}$. Recall from Lemma 3.5.2 that $\tilde{h}$ is increasing and that $\tilde{h}(x) = f(-x)/F(-x)$. This implies that F/f is also increasing. Thus,

$$(I) \leq \mathbb{E} \frac{F^m(p - c_1\sqrt{p})}{f^m(p - c_1\sqrt{p})} 1\left\{ \frac{Y_{(i)} - p}{\sqrt{p}} < -c_1 \right\} \leq f^{-m}(p - c_1\sqrt{p}) \leq c_4 p^{m/2},$$

where the last inequality is by Lemma 3.5.3. And,

$$\begin{aligned}
(III) &\leq \mathbb{E}\left(\frac{1}{f^m(Y_{(i)})}\right) 1\left\{ \frac{Y_{(i)} - p}{\sqrt{p}} > c_1 \right\} && (3.5.2) \\
&\leq p^{m/2} e^{c_1^2/2} e^{-c_1^2(n-i+1)/8} && (3.5.3) \\
&\leq p^{m/2} && (3.5.4)
\end{aligned}$$

where step 3.5.3 follows from Lemma 3.5.3 and Lemma 3.5.13 in Appendix B. □

Lemma 3.5.2. *Let* $S := \sum_{i=1}^{n} \left(Y_{(i)} - \mathbb{E}Y_{(i)}\right)^2$. *Then there are constants* $c_1, c_2 > 0$ *such that, with probability at least* $1 - c_1/\log^2 n$,

$$S \leq c_2 p \log^2 n,$$

Proof of Lemma 3.5.2. For $i \geq \frac{n}{2}$, by Theorem 2.5 of [Boucheron and Thomas, 2012],

$$Var(Y_{(i)}) \leq \frac{2}{n-i+1}\mathbb{E}\frac{1}{(h(Y_{(i-1)}))^2}.$$

Note that [Boucheron and Thomas, 2012] uses the reverse ordering as this, so their result is stated slightly differently. We have reversed their ordering of the order statistics to be more consistent with traditional notation. For $i < \frac{n}{2}$, we use the following transformation. Let $\tilde{Y}_i = -Y_i$ for each $i \in [n]$. Then $\tilde{Y}_i$ has the density $\tilde{f}$ and distribution function $\tilde{F}$ which satisfy $\tilde{f}(x) = f(-x)$ and $\tilde{F}(x) = 1 - F(-x)$. The hazard rate of Y_i have $\tilde{h}(x) = \tilde{f}(x)/(1-\tilde{F}(x)) = f(-x)/F(-x)$, for $x < 0$. $\tilde{h}$ is an increasing function, since:

$$\begin{aligned}
\tilde{h}'(x) &= \frac{f(-x)^2 - f'(-x)F(-x)}{F^2(-x)} \\
&= \frac{f(-x)\Big(f(-x) + (\frac{p/2-1}{x} + \frac{1}{2})F(-x)\Big)}{F^2(-x)} \\
&> 0,
\end{aligned}$$

where we use the fact that $f'(x) = (\frac{p/2-1}{x} - \frac{1}{2})f(x)$. Again by Theorem 2.9 of [Boucheron and Thomas, 2012],

$$Var(Y_{(i)}) = Var(\tilde{Y}_{(n-i+1)}) \leq \frac{2}{i}\mathbb{E}\frac{1}{(\tilde{h}(\tilde{Y}_{(n-i)}))^2}$$

Thus for $i < n/2$,

$$Var(Y_{(i)}) \leq \frac{2}{i}\mathbb{E}\left(\frac{F^2(-\tilde{Y}_{(n-i)})}{f^2(-\tilde{Y}_{(n-i)})}\right) = \frac{2}{i}\mathbb{E}\left(\frac{F^2(Y_{(i+1)})}{f^2(Y_{(i+1)})}\right).$$

In summary, we have:

$$Var(Y_{(i)}) \leq \begin{cases} \frac{2}{n-i+1}\frac{1}{h^2(Y_{(i-1)})}, & i \geq \frac{n}{2}; \\ \frac{2}{i}\frac{1}{h^2(Y_{(i+1)})}, & i < \frac{n}{2}. \end{cases}$$

Combining this with Lemma 3.5.1, we obtain:

$$Var(Y_{(i)}) \leq \begin{cases} \frac{2(c_2+c_3+1)p}{n-i+1}, & i \geq \frac{n}{2}; \\ \frac{2(c_2+c_4+1)p}{i}, & i < \frac{n}{2}. \end{cases}$$

Then for some constant $c_5 > 0$,

$$\mathbb{E}S = \sum_{i=1}^{n} Var(Y_{(i)}) \leq 4(c_2 + c_3 \vee c_4 + 1)p \sum_{i=1}^{n/2} \frac{1}{i} \leq 4(c_2 + c_3 \vee c_4 + 1)p \int_1^{n/2} \frac{1}{x}dx = c_5 p \log(n/2).$$

Note that for any two random variables U and V,

$$\mathbb{E}UV = \mathrm{Cov}(U, V) + \mathbb{E}U\mathbb{E}V \leq \sqrt{VarUVarV} + \mathbb{E}U\mathbb{E}V \tag{3.5.5}$$

So,

$$\begin{aligned} VarS &= \sum_{i=1}^{n} Var(Y_{(i)} - \mathbb{E}Y_{(i)})^2 + \sum_{i\neq j} \mathrm{Cov}((Y_{(i)} - \mathbb{E}Y_{(i)})^2, (Y_{(j)} - \mathbb{E}Y_{(j)})^2) \\ &\leq \sum_{i=1}^{n} Var(Y_{(i)} - \mathbb{E}Y_{(i)})^2 + \sum_{i\neq j} \sqrt{Var(Y_{(i)} - \mathbb{E}Y_{(i)})^2 Var(Y_{(j)} - \mathbb{E}Y_{(j)})^2} \\ &= \left(\sum_{i=1}^{n} \sqrt{Var(Y_{(i)} - \mathbb{E}Y_{(i)})^2}\right)^2 \end{aligned} \tag{3.5.6}$$

It remains to find an upper bound for $Var(Y_{(i)}-\mathbb{E}Y_{(i)})^2$. Let $Z := (Y_{(i)}-\mathbb{E}Y_{(i)})^2, Z_j = (Y^j_{(i)}-\mathbb{E}Y_{(i)})^2$, where $Y^j_{(i)}$ is the ith order statistic of the set $\{Y_1,\ldots,Y_{j-1},Y_{j+1},\ldots,Y_n\}$. Note that $\mathbb{E}Y_{(i)}$ is just a constant that depends on i,n,p.

$$Z_j = \begin{cases} \left(Y_{(i+1)} - \mathbb{E}Y_{(i)}\right)^2, & \text{if } Y_j \le Y_{(i)}; \\ Z, & \text{otherwise} \end{cases}$$

That is, if we remove Y_j from the sample and it is smaller or equal to $Y_{(i)}$, then the ith order statistic of the new sample is now $Y_{(i+1)}$. By Theorem 2.1 of [Boucheron and Thomas, 2012],

$$Var(Y_{(i)} - \mathbb{E}Y_{(i)})^2 \le \sum_{j=1}^{n} \mathbb{E}(Z - Z_j)^2 = i\mathbb{E}\Big((Y_{(i)} - \mathbb{E}Y_{(i)})^2 - (Y_{(i+1)} - \mathbb{E}Y_{(i)})^2\Big)^2.$$

Thus:

$$\begin{aligned} Var\left(Y_{(i)} - \mathbb{E}Y_{(i)}\right)^2 &\le i\mathbb{E}\Big((Y_{(i)} - \mathbb{E}Y_{(i)})^2 - (Y_{(i+1)} - Y_{(i)} + Y_{(i)} - \mathbb{E}Y_{(i)})^2\Big)^2 \\ &\le i\mathbb{E}\left(\left(Y_{(i+1)} - Y_{(i)}\right)^2 + 2(Y_{(i+1)} - Y_{(i)})(Y_{(i)} - \mathbb{E}Y_{(i)})\right)^2 \end{aligned}$$

Let $D_i = Y_{(i+1)} - Y_{(i)}$. Then:

$$\begin{aligned} Var(Y_{(i)} - \mathbb{E}Y_{(i)})^2 &\le i\mathbb{E}\left(D_i^2 + 2D_i(Y_{(i)} - \mathbb{E}Y_{(i)})\right)^2 \\ &= i\left(\mathbb{E}D_i^4 + 4\mathbb{E}D_i^3(Y_{(i)} - \mathbb{E}Y_{(i)}) + 4\mathbb{E}D_i^2(Y_{(i)} - \mathbb{E}Y_{(i)})^2\right) \\ &\triangleq i(A + 4B + 4C). \end{aligned}$$

By (3.5.5),

$$B \le \sqrt{VarD_i^3 Var(Y_{(i)} - \mathbb{E}Y_{(i)})} + \mathbb{E}D_i^3\mathbb{E}(Y_{(i)} - \mathbb{E}Y_{(i)}) \le \sqrt{\mathbb{E}D_i^6 VarY_{(i)}} + 0.$$

and

$$C \leq \sqrt{VarD_i^2 Var(Y_{(i)} - \mathbb{E}Y_{(i)})^2} + \mathbb{E}D_i^2 VarY_{(i)} \leq \sqrt{\mathbb{E}D_i^4 Var(Y_{(i)} - \mathbb{E}Y_{(i)})^2} + \mathbb{E}D_i^2 VarY_{(i)},$$

Combining Lemma 3.5.4 and Lemma 3.5.1, $\mathbb{E}D_i^2 \leq c_1' i^{-2} p, \mathbb{E}D_i^4 \precsim c_1' i^{-4} p^2, \mathbb{E}D_i^6 \leq c_1' i^{-6} p^3$ for some constant $c_1' > 0$. And for $i < n/2, VarY_{(i)} \leq c_1' i^{-1} p$, while for $i \geq n/2, VarY_{(i)} \leq c_1'(n - i + 1)^{-1} p$. So for $i < n/2$,

$$\mathrm{Var}(Y_{(i)} - \mathbb{E}Y_{(i)})^2 \leq i\Big(c_1' i^{-4} p^2 + 4\sqrt{c_1' i^{-4} p^2 Var(Y_{(i)} - \mathbb{E}Y_{(i)})^2} + 4c_1' i^{-3} p^2 + 4\sqrt{c_1' i^{-7} p^4}\Big).$$

By solving this quadratic equation we have

$$\mathrm{Var}(Y_{(i)} - \mathbb{E}Y_{(i)})^2 \leq \frac{c_2' p^2}{i^2},$$

for some constant $c_2' > 0$. Thus continuing from (3.5.6), we have:

$$\begin{aligned} \mathrm{Var}S &\leq 4(c_2' \vee c_3')^2 \left(\sum_{i=1}^{n/2} \frac{p}{i}\right)^2 \\ &\leq c_4' p^2 \log^2 n, \end{aligned}$$

for some constant $c_4' > 0$.

Thus we have shown that

$$\mathbb{E}S \leq cp \log n \tag{3.5.7}$$

and

$$\mathrm{Var}S \leq cp^2 \log^2 n \tag{3.5.8}$$

And by the Markov Inequality:

$$\mathbb{P}\left(|S - \mathbb{E}S| \geq cp\log^2 n\right) \leq \frac{VarS}{c^2p^2\log^4 n} \precsim \frac{1}{\log^2 n}.$$

Thus with probability $1 - 1/\log^2 n, S - \mathbb{E}S \leq |S - \mathbb{E}S| \leq cp\log^2 n$. So $S \leq \mathbb{E}S + cp\log^2 n \leq c_2 p\log^2 n$ by the bound on $\mathbb{E}S$. □

Lemma 3.5.3. *Let $f(x)$ be the density function of $\chi^2(p)$.*

1. *There exists constants $c_1, c > 0$ such that for any $x \in [p - c_1\sqrt{p}, p + c_1\sqrt{p}]$ we have $f(x) \geq c/\sqrt{p}$;*

2. *For $t \in (-\sqrt{p}/2, +\infty), f(p + t\sqrt{p}) \geq c\exp(-t^2/2)/\sqrt{p}$. That is, $f(x) \geq ce^{-\frac{(x-p)^2}{2p}}$.*

3. *For $t \in [\sqrt{p}, +\infty), f(x) \geq \exp(-x/2)$.*

Proof. First, we show that there is a constant $c > 0$ such that $f(p) \geq c/\sqrt{p}$. Recall that

$$f(p) = \frac{p^{p/2-1}e^{-p/2}}{2^{p/2}\Gamma(p/2)} = \frac{(p/2)^{p/2}\, e^{-p/2}}{p\Gamma(p/2)}$$

Now $p \geq 3$ is an integer. First assume that p is even, so $\Gamma(p/2) = \left(\frac{p}{2} - 1\right)!$. By Stirling's bounds,

$$\Gamma(p/2) \leq \left(\frac{p}{2} - 1\right)! \leq e\left(\frac{p}{2} - 1\right)^{p/2-1/2} e^{-p/2-1}$$

Thus, if p is even,

$$f(p) \geq \frac{(p/2)^{p/2}\, e^{-p/2}}{pe(p/2-1)^{p/2-1/2}e^{-p/2-1}} = \frac{(p/2)^{p/2}}{p(p/2-1)^{p/2-1/2}} \geq \frac{c}{\sqrt{p}}$$

since $\frac{p/2}{p/2-1}$ is bounded below by a constant and

$$p\left(\frac{p}{2}-1\right)^{-1/2} = p\left(\frac{2}{p-2}\right)^{1/2} = \sqrt{2}\left(\frac{p^2}{p-2}\right)^{1/2} \leq \sqrt{2}c_2\sqrt{p}$$

where c_2 is some positive constant. Now suppose p is odd. Let f_p denote the density of a χ^2_p random variable. Note that

$$\frac{c}{\sqrt{p}} \leq f_p(p) \leq f_{p-1}(p-1)$$

Therefore, there is a positive constant c_1 such that

$$\frac{c_1}{\sqrt{p-1}} \leq \frac{\sqrt{p-1}}{\sqrt{p}}\frac{c}{\sqrt{p-1}} \leq f_{p-1}(p-1)$$

Now we prove (1).

$$f(p+c_1\sqrt{p})/f(p) \geq \left(\frac{c_1}{\sqrt{p}}+1\right)^{\frac{p}{2}-1}\exp\left(-\frac{c_1\sqrt{p}}{2}\right) \tag{3.5.9}$$

$$= \exp\left((\frac{p}{2}-1)\log\left(\frac{c_1}{\sqrt{p}}+1\right)-\frac{c_1\sqrt{p}}{2}\right) \tag{3.5.10}$$

$$\geq \exp\left((\frac{p}{2}-1)\left(\frac{c_1}{\sqrt{p}}-\frac{c_1^2}{p}\right)-\frac{c_1\sqrt{p}}{2}\right) \quad \text{since } \log(1+x)\geq x-x^2 \tag{3.5.11}$$

$$= \exp\left(-\frac{c_1}{\sqrt{p}}-\frac{c_1^2}{2}+\frac{c_1^2}{p}\right) \tag{3.5.12}$$

$$\geq \exp\left(-\frac{c_1^2}{2}-\frac{1}{4}\right) \tag{3.5.13}$$

$$\triangleq c_3, \tag{3.5.14}$$

In (3.5.13), we simply minimized the function in p and found that the minimum is achieved at $p = 4c_1^2$, with the minimum as given. We also checked, via the second derivative, that this is indeed a minimum. By a similar argument, we have that

$f(p - c_1\sqrt{p})/f(p) \geq c_3$. Thus for any $x \in [p - c_1\sqrt{p}, p + c_1\sqrt{p}]$, $f(x) \geq c_2 c_3/\sqrt{p}$. For 2), we do the same Taylor expansion as above:

$$f(p + t\sqrt{p})/f(p) \geq \exp\Big(- \frac{t}{\sqrt{p}} - \frac{t^2}{2} + \frac{t^2}{p}\Big) \geq \exp(-t^2).$$

For (3), notice that $(2p)^{\frac{p}{2}-1} \geq 2^{p/2}\Gamma(p/2)$ where $\Gamma(p/2) = (p/2)!$ if $p/2$ is an integer and $\Gamma(p/2) \leq \sqrt{2}((p+1)/2)!$ if otherwise. Thus $f(x) \geq \exp(-x/2)$ immediately from its density function. □

Lemma 3.5.4. *Let $Y_{(1)} \leq \ldots \leq Y_{(n)}$ be the order statistics of a sample of n points from a distribution F with a non-decreasing hazard rate h. Let E_i be a standard exponential random variable. Let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the order statistics of a sample of n points from the standard exponential distribution.*

$$\mathbb{E}(Y_{(i+1)} - Y_{(i)})^m \leq \frac{m!}{i^m}\mathbb{E}\frac{1}{h^m(Y_{(i)})}$$

Proof. Let $U(t) = F^-(1 - 1/t)$, as in [Boucheron and Thomas, 2012]. Note that $\frac{d}{dt}U(e^t)) = 1/h(U(e^t))$. So $U \circ \exp$ is concave since h is non-decreasing. Thus:

$$\begin{aligned} U(\exp(X_{(i+1)})) - U(\exp(X_{(i)})) &= U\left(\exp\left(X_{(i)} + X_{(i+1)} - X_{(i)}\right)\right) - U(\exp(X_{(i)})) \\ &\leq \frac{d}{dt}(U \circ \exp)(t)\Big|_{X_{(i)}} \left(X_{(i+1)}) - X_{(i)}\right) \\ &= \frac{1}{h(U(e^{X_{(i)}}))}\left(X_{(i+1)}) - X_{(i)}\right) \end{aligned}$$

Now by Theorem 2.5 of [Boucheron and Thomas, 2012], $Y_{(i)} =_d U(\exp(X_{(i)}))$. Thus:

$$Y_{(i+1)} - Y_{(i)} =_d U(\exp(X_{(i+1)})) - U(\exp(X_{(i)})) \leq \frac{X_{(i+1)} - X_{(i)}}{h\left(U(\exp(X_{(i)}))\right)} =_d \frac{X_{(i+1)} - X_{(i)}}{h\left(Y_{(i)}\right)}$$

And again by Theorem 2.5 of [Boucheron and Thomas, 2012], $X_{(i+1)} - X_{(i)} \sim E_i/i$.

The mth moment of a standard exponential is $m!$. Thus,

$$\mathbb{E}\left(Y_{(i+1)} - Y_{(i)}\right)^m \leq \mathbb{E}\frac{(E_i/i)^m}{h(Y_{(i)})^m} = \frac{m!}{i^m}\mathbb{E}\frac{1}{h(Y_{(i)})^m}$$

□

Lemma 3.5.5. *The hazard function h of a χ^2_p random variable is non-decreasing as long as $p \geq 2$.*

Proof of Lemma 3.5.5. Apply Lemma 3.5.7 to f and $\bar{F}$, which are both defined on $[0, \infty)$ and which satisfy $\lim_{x\to\infty^+} f(x) = \lim_{x\to\infty^+} \bar{F}(x) = 0$. The derivative of $\bar{F}(x)$ is $-f(x)$, which is always negative on $(0, \infty)$. The derivative of f is proportional to $(\frac{p/2-1}{x} - \frac{1}{2})f(x)$. The ratio is thus $\frac{1}{2} - \frac{p/2-1}{x}$, which is increasing in x as long as $p \geq 3$. □

Corollary 3.5.6. *For $\chi^2(p)$ with $p \geq 2$, we have $f \geq (\frac{1}{2} - \frac{p/2-1}{x})\bar{F}$ for $x > p$.*

Proof of Corollary 3.5.6. This is a direct consequence of Lemma 3.5.5 by taking the derivative of h. Note that we have $f' = (\frac{p/2-1}{x} - \frac{1}{2})f$. We prove the lemma by noticing

$$h'(x) = \frac{f(x)(f(x) - (\frac{1}{2} - \frac{p/2-1}{x})\tilde{F}(x))}{(1 - F(x))^2} \geq 0.$$

□

The following lemma is a slight adaptation of Proposition 1.1 in [Pinelis, 2006].

Lemma 3.5.7. *Suppose f and g are two continuously differentiable functions defined on the same interval, (a, b), and that $\lim_{x\to b^+} f(x) = \lim_{x\to b^+} g(x) = 0$. And suppose g is either always positive or always negative on its domain (it cannot change sign and cannot be 0). Then $\frac{f'}{g'}$ increasing implies that $\left(\frac{f}{g}\right)' > 0$.*

Proof. Note that $\left(\frac{f}{g}\right)' = \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}$. The denominator is positive. If remains to show that the numerator is also positive for all $x \in (a, b)$. For $y \in (a, b)$, define

$m_x(y) := g(y)f'(x) - f(y)g'(x)$. Now fix an $x \in (a, b)$ and let $y \in (x, b)$.

$$\frac{\partial}{\partial y} m_x(y) = g'(y)f'(x) - f'(y)g'(x) = g'(y)g'(x)\left(\frac{f'(x)}{g'(x)} - \frac{f'(y)}{g'(y)}\right)$$

Now $g'(y)g'(x) > 0$ since g does not change sign on its domain. And since $y > x$ and $\frac{f'}{g'}$ is increasing, $\left(\frac{f'(x)}{g'(x)} - \frac{f'(y)}{g'(y)}\right) < 0$. Thus $\frac{\partial}{\partial y} m_x(y) < 0$, so $m_x(y)$ is decreasing for $y \in (x, b)$. Moreover, $m_x(y)$ is continuous since g, g', f, f' are. So in fact, $m_x(y)$ is decreasing on $[x, b)$. And:

$$\lim_{y \to b^+} m_x(y) = \lim_{x \to b^+} g(y)f'(x) - \lim_{x \to b^+} f(y)g'(x) = 0$$

That is, $m_x(y)$ is decreasing to 0 on $[x, b)$, which implies $m_x(x) > 0$, as needed. □

3.5.2 Inequalities related to the upper bound

Lemma 3.5.8. *Let $A \in \mathbb{R}^{n \times p}$ and let A_k be the rank k approximation of A. Let P have rank k. Then*

$$\|A_k - P\|_F^2 \leq 8k \|A - P\|^2$$

Proof of Lemma 3.5.8. Since A_k and P are rank k, $A_k - P$ has rank at most $2k$. And for any rank $2k$ matrix A, $\|A\|_F^2 \leq 2k \|A\|_2^2$. So,

$$\begin{aligned}
\|A_k - P\|_F^2 &\leq 2k \|A_k - P\|_2^2 \\
&\leq 4k \|A_k - A\|_2^2 + 4k \|A - P\|_2^2 \qquad \text{by the triangle inequality} \\
&\leq 8k \|A - P\|_2^2
\end{aligned}$$

The third step follows because $\|A - A_k\|_F^2 \leq \|A - P\|_F^2$ for any rank-k matrix P. □

Lemma 3.5.9. *Suppose ρ satisfies the bandwidth assumptions. Then:*

$$\left\|R^{(GK)}\right\|^2 \precsim \left\|\frac{R^{(EK)}}{\rho^2}\right\|^2 + \frac{3n^2}{\rho^4} + \frac{b^8 n^2}{\rho^8}$$

Proof of Lemma 3.5.9. Fix $s, t \in [k]$. Let $i, j \in [n/k]$. By Taylor's Theorem, we know there exists

$$\xi_{ij}^{st} \in \left(0, \frac{\left\|X_{s,(i)} - X_{t,(j)}\right\|^2}{\rho^2}\right)$$

and

$$c_{ij}^{st} \in \left(0, \frac{\mathbb{E}\left\|Z_{s,(i)}\right\|^2 + \mathbb{E}\left\|Z_{t,(j)}\right\|^2 + d_{st}^2\{s \neq t\}}{\rho^2}\right)$$

such that:

$$\left(R^{(GK)}\right)_{ij}^{st} = -\frac{\left(R^{(EK)}\right)_{ij}^{st}}{\rho^2} + \frac{e^{-\xi_{ij}^{st}}}{2}\underbrace{\left(\frac{\left\|X_{s,(i)} - X_{t,(j)}\right\|^2}{\rho^2}\right)^2}_{(I)} - \frac{e^{-c_{ij}^{st}}}{2}\underbrace{\left(\frac{\mathbb{E}\left\|Z_{s,(i)}\right\|^2 + \mathbb{E}\left\|Z_{t,(j)}\right\|^2 + \mathbb{E}\left\|M_{si} - M_{tj}\right\|_2^2}{\rho^2}\right)^2}_{(II)}.$$

Since ξ_{ij}^{st} and c_{ij}^{st} are positive, $e^{-\xi_{ij}^{st}}, e^{-c_{ij}^{st}} \in (0, 1)$. The terms (I) and (II) are real and nonnegative, and the matrices with these entries are symmetric. So Lemma 3.5.10 implies that we can drop the exponential coefficients in analyzing the operator norm. Using the bound $(a+b)^2 \leq a^2 + b^2$ repeatedly,

$$\left\|R^{(GK)}\right\|_2^2 \leq c\left(\frac{\left\|R^{(EK)}\right\|^2}{\rho^2} + \left\|N^{(1)}\right\|^2 + \left\|N^{(2)}\right\|^2 + \left\|N^{(3)}\right\|^2 + \left\|N^{(4)}\right\|^2\right)$$

for some constant $c > 0$, where

$$\begin{aligned}
\left(N^{(1)}\right)_{ij} &= \left(\frac{\|Z_j - Z_j\|^2}{\rho^2}\right)^2 \\
\left(N^{(2)}\right)_{ij} &= \frac{\|M_{s,i} - M_{t,j}\|_2^4}{\rho^4} \\
\left(N^{(3)}\right)_{ij}^{st} &= \left(\frac{\mathbb{E}\left\|Z_{s,(i)}\right\|^2 + \mathbb{E}\left\|Z_{t,(j)}\right\|^2}{\rho^2}\right)^2 \\
\left(N^{(4)}\right)_{ij}^{st} &= \left(\frac{d_{st}^2}{\rho^4}\mathbf{1}\{s \neq t\}\right)
\end{aligned}$$

We proceed to bound each operator norm separately. For $N^{(4)}$, we use the operator norm bound; for the rest, we use the Frobenius norm. First,

$$\begin{aligned}
\mathbb{P}\left\{\left\|N^{(1)}\right\|_F^2 > \frac{n^2}{\rho^4}\right\} &\leq n^2\mathbb{P}\left\{\frac{\|Z_i - Z_j\|^8}{\rho^8} > \frac{1}{\rho^4}\right\} && \text{by a union bound} \\
&= n^2\mathbb{P}\{\|Z_i - Z_j\|^2 > \rho\} \\
&= n^2\mathbb{P}\{\chi_p^2 \geq p + 2\sqrt{3p\log n} + 6\log n\} \\
&\leq n^2\exp\left(-3\log n\right)
\end{aligned}$$

where the third step is since $\|Z_i - Z_j\|^2 \sim 2\chi_p^2$ and since $\rho \geq 2\left(p + 2\sqrt{3p\log n} + 6\log n\right)$. The final step is by Lemma 3.5.12. So with probability $1 - \frac{1}{n}$,

$$\left\|N^{(1)}\right\|^2 \precsim \frac{n^2}{\rho^4}.$$

And for any i, j, $\|M_i - M_j\|\, 2^2 \leq 6b^2 + 3d_{\max}^2$. So there is a constant $c > 0$ such that

$$\left\|N^{(2)}\right\|_F^2 \leq \frac{cn^2(b^8 + d_{\max}^8)}{\rho^8} \leq \frac{cn^2b^8}{\rho^8} + \frac{cn^2}{\rho^4}$$

where the last step is follows since $\rho \geq d_{\max}^2$. And $\mathbb{E}\left\|Z_{(n/k)}\right\|^2 \leq p\log\frac{n}{k}$ by Lemma 3.5.14, so $\rho \geq \mathbb{E}\left\|Z_{s,(i)}\right\|^2 + \mathbb{E}\left\|Z_{t,(j)}\right\|^2$ for all s, t, i, j, by our assumption on ρ. Thus

all squared entries of $N^{(3)}$ are bounded by $\frac{1}{\rho^4}$, so

$$\left\|N^{(3)}\right\|_F^2 \leq \frac{n^2}{\rho^4}.$$

Finally,

$$\left\|N^{(4)}\right\|^2 \leq \frac{n^2 d_{\max}^8}{k^2 \rho^8} \leq \frac{n^2}{\rho^4}.$$

since $\rho \geq d_{\max}^2/\sqrt{k}$. □

Lemma 3.5.10. *If E is a matrix with bounded entries and M is a real symmetric matrix with positive entries, then $\|E \circ M\|_{OP} \leq \max_{i,j} |E_{ij}| \, \|M\|_{OP}$.*

Proof. See Lemma A.5 of [Karoui, 2010b]. □

Lemma 3.5.11. *Let $U' \in \mathbb{R}^{n \times p}$ be a matrix whose rows μ_i' lie on a smooth submanifold M of $\mathbb{R}^p$. Then there exists an m such that U can be well-approximated by an m-rank matrix U, in that:*

$$\|U - U'\|_F^2 \precsim n * o(1/n)$$

Proof. Since M is smooth, we can find m centers such that each μ_i' is within $o(1/n)$ to a μ_i that is a linear combination of these m centers. (E.g., μ_i could just be cv_i for one v_i.) Put these μ_i into the matrix U; then U is rank m. The result holds. □

3.5.3 Standard inequalities in probability

Lemma 3.5.12. *Let $X \sim \chi^2(p)$. Then*

$$\mathbb{P}\Big(|X - p| \geq t\sqrt{p}\Big) \leq \begin{cases} 2\exp(-t^2/8), & \text{if } 0 \leq t \leq \sqrt{p} \\ 2\exp(-\sqrt{p}t/8), & \text{if } t > \sqrt{p}. \end{cases}$$

Proof of Lemma 3.5.12. This is a standard Chernoff bound; see for instance the comments for Lemma 1 in Section 4 of [Laurent and Massart, 2000]. Note that we could also write

$$\mathbb{P}\{X \geq p + 2\sqrt{pt} + 2t\} \leq e^{-t}$$
$$\mathbb{P}\{X \leq p - 2\sqrt{pt}\} \leq e^{-t}$$

□

Lemma 3.5.13. *Let $Y_1, \ldots, Y_n \sim \chi^2_p$, and let $Y_{(1)} \leq \ldots \leq Y_{(n)}$ be the order statistics. Then:*

$$\mathbb{P}\left\{\frac{Y_{(i)} - p}{\sqrt{p}} > t\right\} \leq \exp\left(-(n-i-1)t^2/8\right) \qquad \textit{and}$$
$$\mathbb{P}\left\{\frac{Y_{(i)} - p}{\sqrt{p}} \leq t\right\} \leq \exp\left(-(i-1)t^2/8\right)$$

Proof of Lemma 3.5.13. For the first part, if $Y_{(i)}$ is larger than something, this implies that $n-i$ of the other random variables must be larger than that thing. The argument for the second part is reversed. Then we use Lemma 3.5.12. □

Lemma 3.5.14. *Let $Y_1, \ldots, Y_n \backsim \chi^2_p$ where $p \geq 3$. Then $\mathbb{E}Y_{(n)} \leq p \log n$.*

Proof. Let $M := Y_{(n)}$, the maximum. We use the Chernoff method and let $\theta < 1/2$ so that the MGF of a χ^2_p random variable is defined. Using Jensen's Inequality:

$$e^{\mathbb{E}\theta M} \leq \mathbb{E}e^{\theta M} = \mathbb{E}\max_{i \leq n} e^{\theta Y_i} \leq n\mathbb{E}e^{\theta Y_i} = n(1-2\theta)^{-p/2}$$

where in the final step, we just inserted the χ^2_p moment-generating function. Thus for $\theta < 1/2$:

$$\mathbb{E}M \leq \frac{\log n}{\theta} - \frac{p\log(1-2\theta)}{2\theta}$$

Now, it is well-known that for all $x > 0$, $\frac{-x}{1-x} \leq \log(1-x)$. Thus:

$$\mathbb{E}M \leq \frac{\log n}{\theta} + \frac{p}{1-2\theta}$$

for all $\theta \in (0, 1/2)$. Pick a θ in this range, e.g. $\theta = 1/4$. Then we have a bound of $4 \log n + 2p$, which is bounded above by $8 \max(\log n, p)$. □

Lemma 3.5.15. *Let Z be a $n \times p$ matrix with independent entries, each distributed $N(0, \sigma^2)$. Then with probability at least $1 - 1/n$, there is a $c > 0$ such that*

$$\|Z\|_2^2 \leq c\sigma^2(n+p)$$

Proof. See e.g. the proof of Theorem 5.39 of [Vershynin, 2012]. □

3.5.4 Refinement

Lemma 3.5.17 provides an error rate when sample-splitting was used in the algorithm. Sample splitting is used when the error rate depends on a set E where $\mathbb{P}\{E^C\}$ is large enough so that we don't want it to play a role in the bound. Consider the proof of Lemma 3.5.16. If we instead examined had a separate E_i for each node, and if e.g. $\delta = 1/\log n$, then in the proof, we would have

$$\frac{1}{n}\mathcal{L}(\hat{\tau}, \tau) \precsim \frac{\sum_{i \leq n} \mathbb{P}\{\hat{\tau}(i) \neq \tau(i) \text{ and } E_i\} + \mathbb{P}\{E_i^C\}}{nt}$$

We would then need to choose t to offset the $\delta = 1/\log n$; i.e. the error rate would have an exponential and a logarithmic term, which would be terrible.

Lemma 3.5.16. *[Error rate: ordinary version] Let $\hat{\tau}$ be the result of Algorithm 4.*

Then with probability at least $1 - \sqrt{\sum_{u=1}^{n} \mathbb{P}\{\hat{\tau}(u) \neq \tau(u)\}/n}$,

$$\frac{1}{n}\mathcal{L}(\hat{\tau}, \tau) \leq \left(\frac{\sum_{u=1}^{n} \mathbb{P}\{\hat{\tau}(u) \neq \tau(u)\}}{n} \right)^{1/2}$$

Note that if $\mathbb{P}\{\hat{\tau}(u) \neq \tau(u)\}$ *is the same for all* $u \in [n]$*, then with probability at least* $1 - \sqrt{\mathbb{P}\{\hat{\tau}(u) \neq \tau(u)\}/n}$,

$$\frac{1}{n}\mathcal{L}(\hat{\tau}, \tau) \leq (\mathbb{P}\{\hat{\tau}(u) \neq \tau(u)\})^{1/2}$$

Proof. Let $t > 0$. By the Markov Inequality,

$$\mathbb{P}\{\mathcal{L}(\hat{\tau}, \tau) > nt\} \leq \frac{\sum_{u=1}^{n} \mathbb{P}\{\hat{\tau}(u) \neq \tau(u)\}}{nt}$$

Now choose

$$t = \sqrt{\sum_{u=1}^{n} \mathbb{P}\{\hat{\tau}(u) \neq \tau(u)\}/n}$$

to obtain the result. □

Lemma 3.5.17. *[Error rate: sample-splitting version] Let* $\hat{\tau}$ *be the result of Algorithm Algorithm 4. Let* $s(i)$ *be the sample of node* i*, i.e.,*

$$s(i) = \begin{cases} 1 & \text{if } i \in \{1, \ldots, n/2\} \\ 2 & \text{if } i \in \{n/2 + 1, \ldots, n\} \end{cases}$$

Let $\tau_1 = \{\tau(1), \ldots, \tau(n/2)\}$ *and* $\tau_2 = \{\tau(n/2 + 1), \ldots, \tau(n)\}$*. Define* $\hat{\tau}_1, \hat{\tau}_2$ *similarly. Let* E_1, E_2 *be sets with* $\max_{j \in [2]} \mathbb{P}\{E_j^C\} \leq \delta$.

Then with probability at least $1 - 2\delta - \sqrt{\sum_{u=1}^{n} \mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_{\tau(u)}\}/n}$,

$$\frac{1}{n}\mathcal{L}(\hat{\tau}, \tau) \leq 2 \left(\frac{\sum_{u=1}^{n} \mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_{s(u)}\}}{n} \right)^{1/2}$$

Note that if $\mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_{s(u)}\}$ *is the same for all* $u \in [n]$*, this simplifies to*

$$\frac{1}{n}\mathcal{L}(\hat{\tau}, \tau) \leq 2\left(\mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_{s(u)}\}\right)^{1/2}$$

Proof. Let $t > 0$. Then

$$\begin{aligned}\mathbb{P}\{\mathcal{L}(\hat{\tau}, \tau) > nt\} &\leq \mathbb{P}\{\mathcal{L}(\hat{\tau}_1, \tau_1) > nt/2 \text{ and } E_1\} + \mathbb{P}\{\mathcal{L}(\hat{\tau}_2, \tau_2) > nt/2 \text{ and } E_2\} + \mathbb{P}\{E_1^C\} + \mathbb{P}\{E_2^C\} \\ &\leq \frac{2\sum_{u=1}^n \mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_{s(u)}\}}{nt} + 2\delta\end{aligned}$$

by the Markov Inequality and by the assumption on E_1 and E_2. Choose

$$t = \sqrt{\sum_{u=1}^n \mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_{s(u)}\}/n}$$

to obtain the result. □

Lemma 3.5.18. *[Error rate: leave-one-out version] Let* $\hat{\tau}$ *be the result of Algorithm Algorithm 4. For each* $u \in [n]$*, let* E_u *be a set with* $\mathbb{P}\{E_u^C\} \leq \delta_u \leq \delta$*. Then with probability at least* $1 - \sqrt{\sum_{u=1}^n \mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_u\}/n + \delta}$*,*

$$\frac{1}{n}\mathcal{L}(\hat{\tau}, \tau) \leq \left(\frac{\sum_{u=1}^n \mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_u\}}{n} + \delta\right)^{1/2}$$

If $\mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_u\}$ *is the same for all* $u \in [n]$*, this simplifies to*

$$\frac{1}{n}\mathcal{L}(\hat{\tau}, \tau) \leq \left(\mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_u\} + \delta\right)^{1/2}$$

Proof. Let $t > 0$. Then

$$\begin{aligned}\mathbb{P}\{\mathcal{L}(\hat{\tau}, \tau) > nt\} &\leq \frac{\sum_{u=1}^n \left(\mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_u\} + \delta_u\right)}{nt} \\ &\leq \frac{\sum_{u=1}^n \mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_u\} + n\delta}{nt}\end{aligned}$$

by the Markov Inequality. Choose

$$t = \sqrt{\frac{\sum_{u=1}^{n} \mathbb{P}\{\hat{\tau}(u) \neq \tau(u) \text{ and } E_u\}}{n} + \delta}$$

to obtain the result. □

Lemma 3.5.19. *For any $u \in [n]$,*

$$\mathbb{P}\{\hat{\tau}(u) \neq \tau(u)\} \leq \sum_{j \in ([k] \setminus \tau(u))} \mathbb{P}\{\hat{\tau}(u) = j\} \leq (k-1) \max_{j \in ([k] \setminus \tau(u))} \mathbb{P}\{\hat{\tau}(u) = j\}$$

Lemma 3.5.20. *Let z, v, u be vectors in $\mathbb{R}^p$. Then*

$$\langle z,\, u - v \rangle \geq \|u\|_2^2 - \|v\|_2^2 \;\; \textit{implies that} \; \langle z,\, \frac{u - v)}{\|u - v\|_2} \rangle \geq \|u\|_2 - \|v\|_2$$

Proof. Note that

$$\|u\|_2^2 - \|v\|_2^2 \left(\|u\|_2 - \|v\|_2\right) \left(\|u\|_2 + \|v\|_2\right) \geq \left(\|u\|_2 - \|v\|_2\right) \|u - v\|_2$$

Dividing both sides of the first inequality in the lemma statement by $\|u - v\|_2$ yields the result. □

Chapter 4

Graphical component analysis for latent signal detection

4.1 Introduction

We now study latent signal detection when the observed data are a linear transformation of an underlying signal that we wish to detect. Independent component analysis (ICA) addresses this setting when the signal is generated with independent components, but this independence assumption is often unrealistic. We propose *graphical component analysis* (GCA), which allows for the latent source components to be dependent with the dependence modeled by a sparse graphical model.

Let $A \in \mathbb{R}^{d \times d}$ be invertible, and define $W = A^{-1}$. Let $G = (V, E)$ be a graph with $|V| = d$. Let us observe data according to the model

$$\begin{aligned} &s_i \sim_{i.i.d.} \mathbb{P}_\theta, \text{ for } i = 1, \ldots, n, \text{ and} \\ &x_i = A s_i. \end{aligned} \tag{4.1.1}$$

Assume that $\mathbb{P}_\theta$ has density

$$p_\theta(s) = \exp(\theta^T \phi(s) - \Psi), \tag{4.1.2}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^k, \theta \in \mathbb{R}^k$, and $\Psi = \int_{s \in \mathbb{R}^d} \exp(\theta^T \phi(s))\, ds$. We will also write $g_\theta(s) = \theta^T \phi(s)$. We will place a pairwise graphical model structure on θ, and the aim is to estimate θ, i.e., the joint distribution of the signal; as a byproduct, we will also estimate A.

Estimating a graphical model requires the computation of a normalizing constant, an intractable task if the data dimension is high. To avoid this problem, we rely on the technique of *score matching*, [Hyvarinen, 2005], which allows for the estimation of a graphical model without computation of the normalizing constant. We present an algorithm that jointly estimates the graphical model and mixing transformation, where the graphical model is represented in terms of an orthogonal polynomial basis, and the mixing matrix is estimated using a coordinate descent procedure based on the use of Givens rotations [Shalit and Chechik, 2014]. GCA generalizes both ICA, where the components are independent, and tree component analysis [Bach and Jordan, 2003] (TCA), where they are governed by a graph with no cycles. Our experiments demonstrate how the presence of cycles in the latent graph can result in failures by both ICA and TCA to recover accurate models of the marginal components, while GCA succeeds.

4.1.1 Related literature

In this section we review some related literature. For a tutorial overview of classical ICA and its uses, see [Hyvärinen and Oja, 2000]. Classical ICA poses two main assumptions: that the source components are independent and that the relationship between the source and observed data is linear. An early attempt to relax the indepen-

dence assumption of ICA was TCA of [Bach and Jordan, 2003]; the TCA algorithm imposed a tree graph structure on the latent source, and the algorithm therein relied on the factorization of a tree graph.

Classical ICA algorithms also typically assume a parametric (though usually flexible) form for the source component distributions. [Samworth and Yuan, 2012] propose a nonparametric maximum likelihood method to relax this assumption. They merely assume that the densities on s have log-concave forms; thus, they can use techniques from the literature on log-concave density estimation.

In recent years, several extensions in the direction of non-linear ICA have been proposed [Hyvärinen and Morioka, 2016, Hyvärinen and Morioka, 2017, Hyvärinen et al., 2019]. And recently, [Khemakhem et al., 2020] proposed another method that both allows for some dependence in the source components, as well as a non-linear relationship between the sources and observations.

Our work focuses on a linear relationship between x and s and relaxes ICA to allow for arbitrary graphical model dependence in the components of s; an interesting direction for future work is to relax this further and allow the relationship between x and s to be non-linear while still maintaining the arbitrary dependence structure of s.

The work [Köster et al., 2009] proposes using score matching to estimate an ICA model that allows for Markov Random Field-structured dependence in the data. This model and method appear closely related to ours, and we now dwell briefly on this paper to explain the differences.

Let our observed data be vectors $x \in \mathbb{R}^d$. [Köster et al., 2009] propose to use a Markov random field structure to estimate "feature detectors" in images. They place a graph structure on the observed data points x. This is in contrast to our GCA model, which places the graph structure on the unobserved $s = Wx$.

Let us have a graph $G = (V, E)$ with $|V| = d$. [Köster et al., 2009] treat each

pixel (node) x_j as the center of one clique. Let C be a positive integer and let $2C+1$ be the clique size. Let $x_{j,C}$ be the vector $x_{j,C} = (x_{j-C}, \ldots, x_j, \ldots, x_{j+C})^T$. Let φ be some potential function. The [Köster et al., 2009] model, parametrized by the feature vectors $v_1, \ldots, v_L$, each in $\mathbb{R}^{2C+1}$, is

$$\log p(x) \propto \sum_{l=1}^{L} \sum_{j=C+1}^{p-C} \varphi(v_l' x_{j,C}).$$

Contrast this with our GCA model. Let $W \in \mathbb{R}^{d \times d}$ have rows $w_u \in \mathbb{R}^p$ for $k \in [p]$. Our model is:

$$\log p(x) \propto \sum_{v \in V} \psi_v(w_v^\top x) + \sum_{(u,v) \in E} \psi_{kj}(w_u^\top x, w_v^\top x).$$

Here is a simple example. Suppose that

$$V = \{1, 2, 3, 4\}$$
$$E = \{(1,2), (2,3), (1,3), (2,4), (3,4)\}.$$

Here, $p = |V| = 4$. The cliques are $(1,2,3), (2,3,4)$. The [Köster et al., 2009] model is as follows. They would have $v_1, \ldots, v_l \in \mathbb{R}^3$, since here the clique size is $2C+1 = 3$.

$$\log p_X(x) \propto \sum_{l=1}^{L} \left(\varphi(v_l^\top (x_1, x_2, x_3)) + \varphi(v_l^\top (x_2, x_3, x_4)) \right).$$

If we assume such a graph structure on s, our GCA model would be as follows.

$$\log p_X(x) \propto \varphi(w_1^\top x, w_2^\top x, w_3^\top x) + \varphi(w_2^\top x, w_3^\top x, w_4^\top x),$$

where $\varphi(w_1^\top x, w_2^\top x, w_3^\top x) = \psi_1(w_1^\top x) + \psi_2(w_2^\top x) + \psi_3(w_3^\top x) + \psi_{12}(w_1^\top x, w_2^\top x) + \psi_{23}(w_2^\top x, w_3^\top x) + \psi_{13}(w_1^\top x, w_3^\top x)$, and so on for the other functions.

4.2 Approach

The model (4.1.1) is equivalent to $x = DA\tilde{s}$ where D is a diagonal matrix with non-zero entries and $\tilde{s}$ is the correspondingly-scaled version of s. That is, the model is only identifiable up to scaling. We assume $W^T W = WW^T = I_d$ to handle this problem. Note also for any permutation matrix P, $x = AP^{-1}Ps$, i.e., the model is identifiable up to permutation. We estimate the graphical model up to the equivalence class under permutation of the nodes.

Let the true density of the observed data x be p. Suppose we have a model with density $p_\theta(x)$ and we wish to estimate θ. Recall that the score matching objective is

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbb{E}\left\|\nabla_x \log p(x) - \nabla_x \log p_\theta(x)\right\|_2^2.$$

Recall that due to integration by parts, the score-matching objective simplifies to something we can optimize in θ [Hyvarinen, 2005]:

$$\mathcal{L}(\theta) \stackrel{=}{c} \frac{1}{2}\mathbb{E}\left\|\nabla_x \log p_\theta(x)\right\|_2^2 + \mathbb{E}\,\mathrm{tr}\left(\nabla_x^2 \log p_\theta(x)\right). \tag{4.2.1}$$

In our case, we will use this objective on $p_{W,\theta}(x)$, the observed data density. By the change of variable density formula, the distribution for x has density:

$$p_{W,\theta}(x) = |W| p_\theta(Wx). \tag{4.2.2}$$

In GCA, we will optimize the objective (4.2.1) in W and θ. Using the transformation (4.2.2) and the chain rule,

$$\nabla_x \log p_{\theta,W}(x) = \nabla_x \left(\log p_\theta(Wx) + \log|W|\right) = W^T\left(\nabla \log p_\theta(Wx)\right).$$

And

$$\nabla_x^2 \log p_{\theta,W}(x) = W^T \left(\nabla^2 \log p_\theta(Wx)\right) W.$$

Now $\log p_\theta(x) = g_\theta(x) - \Psi$, and as in the usual score matching, the normalizer disappears when we take the derivative. Plugging into (4.2.1) and using the assumption that $WW^T = I_d$, our objective becomes

$$\mathcal{L}(\theta, W) = \frac{1}{2}\mathbb{E}\left(\nabla g_\theta(Wx)^T \nabla g_\theta(Wx)\right) + \mathbb{E}\,\mathrm{tr}\left(\nabla^2 g_\theta(Wx)\right). \tag{4.2.3}$$

For a function $g : \mathbb{R}^d \to \mathbb{R}$ and for $u, v \in V$, let $\partial_v g = \partial g/\partial s_v$, and $\partial_{uv} g = \partial^2 g/\partial s_u \partial s_v$. And for a vector-valued function $\psi : \mathbb{R}^d \to \mathbb{R}^k$, let $\partial_v \psi = (\partial_v \psi_1, \ldots, \partial_v \psi_m)^T$, and so on. Define $\Delta g(s) = \sum_{v \in V} \partial_{vv}^2 g(s)$. Now $\nabla g = \left(\theta^T \partial_1 \phi, \ldots, \theta^T \partial_p \phi\right)^T$ and $(\nabla^2 g)_{vv} = \langle \theta, \partial_{vv}^2 \phi \rangle$. Define

$$\begin{aligned} C(Wx) &= \sum_{v \in V} \partial_v \phi(Wx) \partial_v \phi(Wx)^T. \\ \xi(Wx) &= \sum_{v \in V} \partial_{vv}^2 \phi(Wx). \end{aligned} \tag{4.2.4}$$

Then we see that our population score matching objective becomes exactly similar to the usual score matching objective, except that now it contains W:

$$\mathcal{L}(\theta, W) = \frac{1}{2}\theta^T \mathbb{E} C(Wx)\theta + \theta^T \mathbb{E}\xi(Wx).$$

To compute this in practice, we use the empirical expectation. Let $\hat{\mathbb{E}} f(x) = \sum_{i \in [n]} f(x_i)/n$. Let $\hat{C}_W = \hat{\mathbb{E}} C(Wx)$ and $\hat{\xi}_W = \hat{\mathbb{E}}\xi(Wx)$. We add in the penalty $\|\theta\|_2^2$ to discourage too many edges. Now define

$$\hat{\mathcal{L}}(\theta, W) = \frac{1}{2}\theta^T \hat{C}_W \theta + \theta^T \hat{\xi}_W + \frac{\lambda}{2}\|\theta\|_2^2$$

Since we have the added constraint that W must be orthogonal, the score-matching

objective becomes:

$$\min_{\theta,W} \hat{\mathcal{L}}(\theta, W) \text{ s.t. } WW^T = I_d$$

Fix W. The objective is quadratic in θ, and there is a closed-form solution:

$$\hat{\theta} = \left(\hat{C}_W + \lambda I_k\right)^{-1} (-\hat{\xi}_W). \tag{4.2.5}$$

Alternatively, as in [Janofsky, 2015], we may impose $\theta_{vu} = \theta_{uv}$, as well as the group sparsity penalty:

$$\hat{\mathcal{L}}(\theta, W) = \frac{1}{2}\theta^T \hat{C}_W \theta + \theta^T \hat{\xi}_W + \sum_{(u,v)\in E} \|\theta_{u,v}\|_2$$

Now this no longer has a closed form solution, but is still a convex optimization. It can be optimized using an alternating direction method of multipliers (ADMM) algorithm; see [Janofsky, 2015] for details. In practice, we in fact use this ADMM algorithm to implement GCA, building on the code of [Janofsky, 2015].

When we have a graphical model on the components of s, g_θ can be expressed as an undirected pairwise graphical model:

$$g_\theta(s) = \sum_{v\in V} f_v(s_v) + \sum_{(u,v)\in E} f_{uv}(s_u, s_v), \tag{4.2.6}$$

where $f_v, f_{u,v}$ are the individual node and pair potential functions. In theory, they could take any nonparametric form. For GCA, we represent the potential functions with a truncated orthogonal basis. Let us have m_1 basis elements for the individual node potential functions and m_2 basis elements for the pairwise potential functions.

Then

$$g_\theta(s) = \sum_{v \in V} \sum_{j \in [m_1]} \theta_v^{(j)} \phi_v^{(j)}(s_v) + \sum_{(u,v) \in E} \sum_{j \in [m_2]} \theta_{u,v}^{(j)} \phi_{uv}^{(j)}(s_u, s_v). \tag{4.2.7}$$

Note that here the number of components of g_θ is $K = m_1 d + m_2 d(d-1)$.

4.3 Algorithm

We iterate over optimizing $\hat{\mathcal{L}}(\theta, W)$ in θ and in W. The optimization in θ has the simple solution given in (4.2.5). To optimize the orthogonal matrix W, we use the Givens rotation method of [Shalit and Chechik, 2014]. This method reduces the problem of optimizing the matrix to $O(d^2)$ sub-problems of optimizing rotations of the matrix. In each rotation, a pair of nodes is selected and the angle of rotation is optimized. This is a one-dimensional optimization which could in theory be done via a grid search. For greater speed, we use a quasi-Newton method. In the implementation of Algorithm 5 in R, we use the optim package with the"L-BFGS-B" algorithm.

For $u, v \in V$, define

$$G_{u,v}(\eta) = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & \cos\eta & \dots & -\sin\eta & \dots & 0 \\ \vdots & & & & & \vdots & \\ 0 & \dots & \sin\eta & \dots & \cos\eta & \dots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix} \tag{4.3.1}$$

That is, it has 1's down the diagonal except for columns and zeros everywhere else except for the (u,u), (v,v), (u,v), and (v,u) entries, where it has $\cos\eta, \cos\eta, -\sin\eta, \sin\eta$,

respectively. Fix a pair (u, v). Write G_η for simplicity.

Algorithm 5: GCA. Define $\hat{C}_W, \hat{\xi}_W, \hat{\mathcal{L}}(\theta, W)$ as in (4.2.4). Let $G_{u,v}(\eta)$ be as defined in (4.6.2).

Input: Dataset $\{x_i\}_{i\in[n]}$, each point in $\mathbb{R}^d$, penalty coefficient $\lambda \in \mathbb{R}$, cutoff $\epsilon \in \mathbb{R}^+$, number of steps T

Output: $\theta \in \mathbb{R}^k$, $W \in \mathbb{R}^{d\times d}$.

Initialize at random a W_0 satisfying $WW^T = I_d$;

While $\hat{\mathcal{L}}(\theta, W) > \epsilon$ **;**

$\theta = \left(\hat{C}_W + \lambda I_k\right)^{-1}\left(-\hat{\xi}_W\right)$ **;**

While $t \in [T]$**, do ;**

Select pair of indices $(u(t), v(t))$ **satisfying** $1 \le u(t) < v(t) \le p$ **;**

$\eta_{t+1} = \text{argmin}_\eta\, \hat{\mathcal{L}}(\theta, W_t G_{u,v}(\eta))$**; do this one-dimensional optimization via quasi-Newton ;**

$W_{t+1} = W_t G_{u,v}(\eta_{t+1})$ **;**

$t = t + 1$ **;**

4.4 Experiments

We now demonstrate empirically the effectiveness of the GCA algorithm. We compare to ICA and TCA on synthetic data generated in the following way. We fix a positive semidefinite matrix $\Omega \in S^+_{d\times d}$ representing the graphical model structure. We generate $Z_i \sim N(0, \Omega^{-1})$ for $i = 1, \ldots, n$. Then for $i = 1, \ldots, n$, we let our source data be

$$S_i = f_\alpha(Z_i), \tag{4.4.1}$$

where $f_\alpha(s) \triangleq sgn(s)|s|^\alpha$, for varying α. Our synthetic source data S_i follow a nonparanormal distribution, as described in [Liu et al., 2009, Liu et al., 2012]. We then transform the S_i using a random orthonormal matrix $A \in \mathbb{R}^{d\times d}$. We generate

soure data with three types of graphical model structures: an independence graph, a tree graph, and a graph containing cycles. We first present, in Fig. 4.1, an example picture showing the performance of ICA, TCA, GCA in a simple case when $d = 3$. Here, we have $\alpha = 1, 2, 3$ in (4.4.1). We compare the recovered sources, i.e., $\hat{W}x$, from each algorithm to the truth in a case where the underlying source graph contains a single cycle. We see that even in this simple case, GCA is able to accurately recover the marginal distributions, while ICA and TCA do not perform as well.

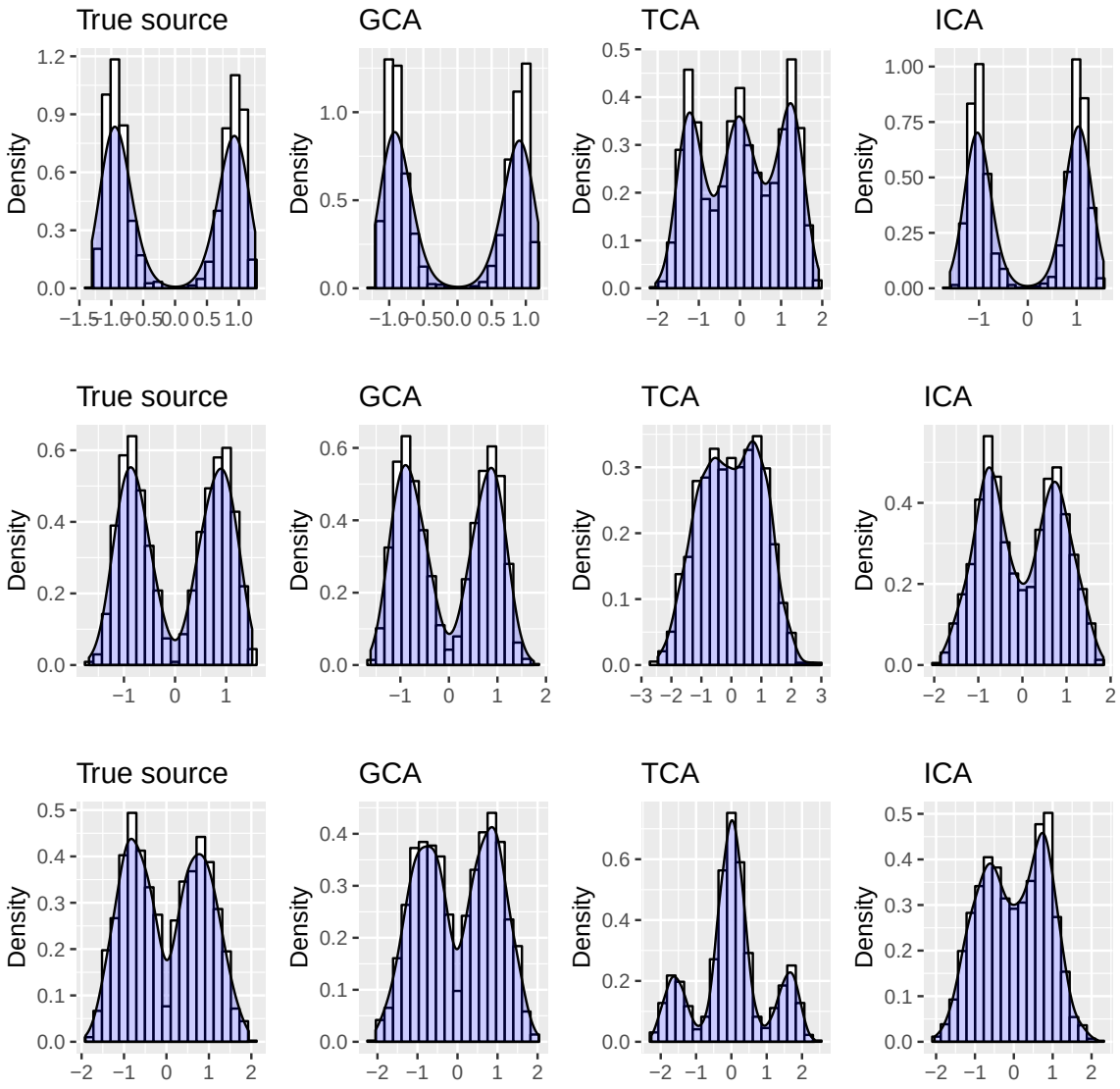


Figure 4.1: Marginal source recovery for GCA, TCA, ICA

We now discuss how to compare the three algorithms in a simulation study. Recall that the three algorithms ICA, TCA, and GCA each produce both an estimate of the de-mixing matrix W and an estimate of the graphical model G on the source components. In the case of ICA, the graphical model is always just the independence model. In the cases of TCA, it is always a tree, and in the case of GCA, any graph structure can be estimated. To compare these algorithms in our preliminary experiments, we use an estimated heldout log likelihood, computed as follows. Given an estimate $\hat{W}$ of W and an estimated graphical model structure $\hat{\theta}$, as well as a set of heldout (test) data x, we compute

$$|\hat{W}|p_{\hat{\theta}}(\hat{W}x_{test}). \tag{4.4.2}$$

If the graphical model governing $\hat{\theta}$ has no connections (independence) or has a tree structure, computing (4.4.2) can be done via kernel density estimation, with a single kernel density estimate for each node in the independence case or pairwise kernel density estimates in the tree case. However, if $\hat{\theta}$ is arbitrary, it is not possible to compute (4.4.2); indeed, this is the motivation for using score matching. Now ICA and TCA *always* estimate independence and tree graphs, respectively, so for these algorithms, we can always estimate (4.4.2). But even if the true θ has a simple structure, GCA can in theory estimate an arbitrary graph, so that (4.4.2) is not computable.

In our preliminary experiments in this section, we circumvent this issue in the following way. For the graph with cycles, we use a graph that contain cycles with no more than three nodes; thus, a three-node kernel density estimate can be used. We found that in general, GCA estimated the graph accurately, so that (4.4.2) could be computed for GCA using the graph it estimated. However, to handle any exceptions, i.e., if GCA estimated a more complex graph, we imposed the following restrictions.

From the $\hat{\theta}$ estimated by GCA, we select the first three closest neighbors with edge weights of at least 1.0 and form a cycle, remove them from the list of potential cycles, then continue in this way till all nodes are accounted for. I in our experiments with cycles of exactly three nodes, we found that GCA almost always selected the correct graph; thus, at least our heldout log likelihood experiments on data with cycles (see Fig. 4.2(c)) demonstrate that GCA can be more effective than ICA or TCA on such data.

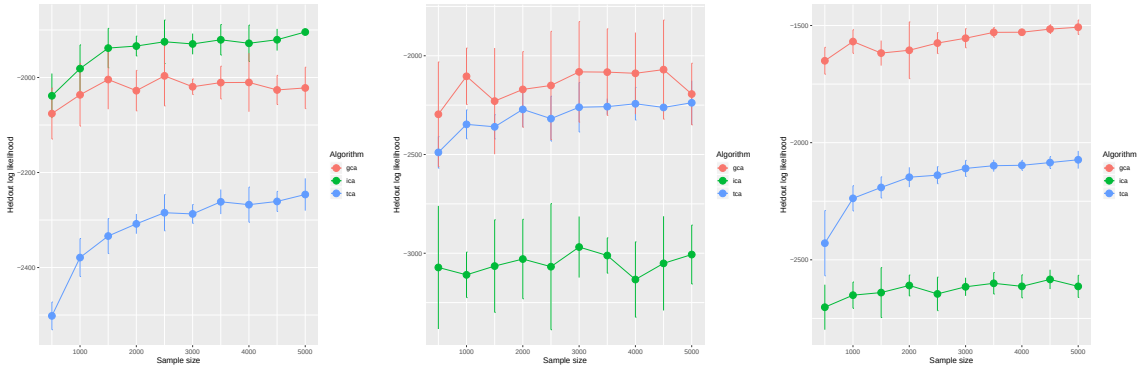
To ensure the above-mentioned method of evaluation is valid, we must see that GCA estimates the graphical model correctly. It is moreover of interest to evaluate the performance of GCA in graphical model estimation. We are able to observe the performance of the graph estimation in the following way.

Recall that W is only identifiable up to permutation; our estimated sources $\hat{W}x$ might correspond to a permutation of the original sources s. This means we wish to estimate the graph up to permutation. Since in our synthetic experiments we can observe the true de-mixing matrix W, we compare $\hat{W}x$ with Wx and select the best node permutation in the following way. We compare the first node of $\hat{W}x$ to each available one in Wx and find the one its distribution is closest to in ℓ_2 norm. We then remove that one from the nodes in Wx and proceed with the next node in $\hat{W}x$. We continue until all nodes are accounted for. This allows us to avoid selecting the best permutation by computing all $d!$ permutations. Note that it relies on our knowledge of the true test source, something we would not have in real-world data.

To compute the best permutation as described above, it is necessary to have unique source marginal distributions. And to make viable the above-mentioned procedure of estimating the graph up to permutation, it is necessary for d to be small. Therefore in our experiments in this section, we let $d = 6$. For larger d, it is difficult to generate data with unique marginal sources that are different enough; thus for larger d, it is difficult to see if the graph estimation is accurate. Therefore the process of

estimating the graph and computing the heldout log likelihood as described above is not necessarily an accurate representation of GCA's performance. We found in our observations that GCA seemed to perform well on higher-dimensional data, but we refrain from showing higher-dimensional experiments till we have found a better way to evaluate the methods. We also note that the current way of approximating the graph estimated by GCA is not ideal for the tree scenario.

In Fig. 4.2, we compare the heldout log likelihoods for synthetic source data s with three types of graphical model structures: independent, tree, and cycles with no more than three nodes, respectively. In these experiments, we let $d = 6$, and the training data sample sizes range from 500 to 5000 in increments of 500; the test set size is always 500. For each sample size value, we run 10 experiments. In all experiments, the matrix W is the same; it was chosen at random in the beginning. We can see that GCA performs slightly worse than ICA on independent data, and surpasses it on tree and cycles data. As mentioned above, for $d = 6$, we have seen that GCA estimates the graph fairly accurately, so the heldout log likelihood calculation in Fig. 4.2(c) is an accurate representation of GCA's performance.

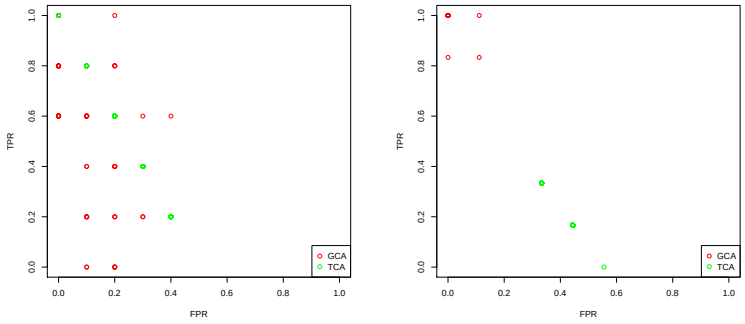


(a) Independence (b) Tree (c) Graph with cycles

Figure 4.2: Heldout log likelihood for GCA, TCA, ICA

In Fig. 4.3, we demonstrate the graph recovery performance of TCA and GCA. There is no need to compare these to ICA, since we know it always estimates an

independence graph. Using our permuted graph, we compare the estimated graph to the true one from Ω. We display the ROC plots in the figures below. In the case of a tree, GCA does not consistently outperform TCA, but in the case of a graph with cycles, it does. Now d is very small here, so GCA in fact performs nearly perfectly in recovering the two cycles in this graph.



(a) Tree (b) Graph with cycles

Figure 4.3: Graph recovery for GCA, TCA

The implementation of GCA used in this section is in R. A major weakness is its slowness; it runs in time $O(d^3)$ due to the Givens rotation. We ran our 300 experiments simultaneously using the Yale High Performance Computing resources; nonetheless, the maximum dimension we can run an experiment on in 24 hours is $d = 15$.

We also note that we experimented on a variety of different data structures (different values of α in (4.4.1)), and found that the performance of TCA varies greatly with different data structures. This is worth further exploration. In the experiments here, we used a different data structure for each node of the signal, with α ranging from 0.5 to 3 in increments of 0.5.

Though the preliminary experiments above are promising, more work is necessary to obtain a better comparison of the performance of these algorithms, especially on higher-dimensional data. In future work, we plan to compare the heldout Fisher

information. The Fisher divergence does not require a normalizing constant. For each algorithm, we will compute the Fisher divergence on a holdout dataset x_{test},

$$\hat{\mathbb{E}}_{x_{test}}\left(\frac{1}{2}\hat{\theta}^\top C(\hat{W}x_{test})\hat{\theta} + \hat{\theta}^\top \xi(\hat{W}x_{test})\right).$$

This method will circumvent the need to compare only on simple graphs, thus allowing us to compare the performance of GCA and the others on more general graphical model structures.

4.5 Kernel density approach

In this section, we propose a computationally more attractive algorithm than the one presented in Section 4.3. As previously noted, the algorithm presented in Section 4.3 runs in time $O(d^3)$. The run time is in fact prohibitively slow in the currently implementation in the R programming language. In this section, we show how the kernel density method of score matching described in Section 1.3 can be used in the context of GCA.

We generate noisy points ξ in the following way. Let $m \in \mathbb{N}$. For each $i \in [n]$, $j \in [m]$,

$$s_i \sim_{i.i.d.} P_S \text{ with density } p_S.$$

$$\xi_{s,i,j} = s_i + z_j, \text{ where } z_j \sim_{i.i.d.} N(0, I_d).$$

Then we have an orthonormal matrix A, and we observe the data:

$$x_i = As_i.$$

$$\xi_{x,i,j} = A\xi_{s,i,j} = As_i + Az_j.$$

Note that since A is orthogonal, $Az_j \sim N(0, I_d)$, so the second term above is simply adding Gaussian noise. We optimize

$$\hat{\mathcal{L}}(\theta, W) = \sum_{i\in[n], j\in[m]} \left\| x_i - \xi_{x,i,j} - \sigma^2 \nabla_\xi \log p_\theta(\xi_{x,i,j}) \right\|_2^2 . \tag{4.5.1}$$

Note that $\nabla_\xi \log p_\theta(\xi_{x,i,j}) = W^\top \nabla_{\xi_s} \log p_\theta(W\xi_{x,i,j})$. We model p_θ as a pairwise graphical model with $\log p_\theta(s) \propto g_\theta(s)$ with g_θ exactly as in (4.2.6). In this case, we let each node and pair potential function $f_v, f_{(u,v)}$ be multilayer perceptrons. For example, let nl be the number of hidden layers and let ℓ be the number of units per hidden layer; we assume we use the same number per layer for simplicity. Let $M_1 \in \mathbb{R}^{\ell\times 2}, M_{nl} \in \mathbb{R}^{1\times\ell}, M_i \in \mathbb{R}^{\ell\times\ell}$ for $i \in \{2, \ldots, nl-1\}$. Let $b_1, \ldots, b_{nl-1} \in \mathbb{R}^\ell, b_{nl} \in \mathbb{R}$. Let $M = (M_1, \ldots, M_{nl})$ and $b = (b_1, \ldots, b_{nl})$, and let $\theta = (M, b)$. Let σ be a nonlinear activation function; in practice we use $\sigma(x) = \tanh(x)$. We have

$$f_{(u,v)}(s_u, s_v) = M_{nl} \ldots \sigma\left(M_2(\sigma\left(M_1(s_u, s_v)^\top + b_1\right) + b_2\right) + \ldots + b_{nl}.$$

Then (4.5.1) is optimized using stochastic gradient descent in TensorFlow. To optimize W, we note that an orthonormal matrix can be represented as a product of orthonormal matrices of the form $G_{u,v}(\eta)$, where η is an angle in $[-\pi, \pi]$ and $(u, v) \in [d]^2$. We adapt the method and code of [Jing et al., 2016] to represent W as a product of orthonormal matrices and to optimize over these angles, as in the previously-described Givens rotation method. It is important to note that [Jing et al., 2016] actually optimize over a dense subset of the space of orthonormal matrices; this subset can be represented as a product of $p \log p$ matrices, whereas in the true space of orthonormal matrices, each element is a product of p^2 matrices. In our preliminary experiments, we use W matrices that actually fall in this dense subset. Our current implementation and preliminary results are available on in the Jupyter notebook in the Github repository Graphical component analysis in TensorFlow.

Algorithm 6: GCA via kernel-density score matching. Define $\hat{\mathcal{L}}(\theta, W)$ as in (4.5.1).

Input: Dataset $\{x_i\}_{i\in[n]}$, each point in $\mathbb{R}^d$, cutoff $\epsilon \in \mathbb{R}^+$, number of steps T

Output: $\theta = (M, b)$, $W \in \mathbb{R}^{d\times d}$.

Initialize at random a W_0 satisfying $WW^T = I_d$ and $\theta \in \Theta$;

While $\hat{\mathcal{L}}(\theta, W) > \epsilon$ **;**

$\theta = \operatorname{argmin} \hat{\mathcal{L}}(W, \theta)$**; Optimize via stochastic gradient descent ;**

While $t \in [T]$**, do ;**

Select pair of indices $(u(t), v(t))$ **satisfying** $1 \le u(t) < v(t) \le p$ **;**

$\eta_{t+1} = \operatorname{argmin}_\eta \hat{\mathcal{L}}(\theta, W_t G_{u,v}(\eta))$ **;**

$W_{t+1} = W_t G_{u,v}(\eta_{t+1})$ **;**

$t = t + 1$ **;**

4.6 Appendix

In this section, we expand on the approach and algorithm used in Section 4.2, providing some details on how the optimization looks in the specific case of a pairwise graphical model.

4.6.1 Pairwise graphical model

Suppose now that we have an undirected pairwise graphical model and that we expand the potential functions with a truncated orthogonal basis. Let us have m_1 basis elements for the individual node potential functions and m_2 basis elements for the

pairwise potential functions. Then

$$\begin{aligned} g(s) &= \sum_{v \in V} \psi_v(s_v) + \sum_{(u,v) \in E} \psi_{u,v}(s_u, s_v) \\ &= \sum_{v \in V} \sum_{k \in [m]} \theta_v^k \phi_v^{(k)}(s_v) + \sum_{(u,v) \in E} \sum_{k \in [m_2]} \theta_{u,v}^k \phi_{uv}^{(k)}(s_u, s_v) \end{aligned}$$

There are a few simple ways to write J, C, ξ, θ, ϕ in such a model. Here is my preferred way. As in [Janofsky, 2015], we define the vectors

$$\begin{aligned} \theta_v &= \left(\theta^{(1)}, \dots, \theta^{(m_1)}\right)^T \in \mathbb{R}^{m_1} \\ \theta_{vu} &= \left(\theta_{vu}^{(1)}, \dots, \theta_{vu}^{(m_2)}\right)^T \in \mathbb{R}^{m_2} \text{ for } u \neq v \\ \theta_{v\cdot} &= \left(-\theta_{v1}^T -, \dots, -\theta_v^T -, \dots, -\theta_{vp}^T -\right)^T \in \mathbb{R}^{m_1 + m_2(p-1)} \\ \theta &= \left(-\theta_{1\cdot} -^T, \dots, -\theta_{p\cdot}^T -\right)^T \in \mathbb{R}^{m_1 p + m_2 p(p-1)} \end{aligned} \tag{4.6.1}$$

Here, we let $\theta_{uv} = \theta_{vu}$. We define the vectors $\phi_v, \phi_{vu}, \phi_{v\cdot}$, and ϕ analogously. Now $\partial_v \phi = (0, \dots, 0, \partial_v \phi_{\cdot v}, 0, \dots, 0)$. Given these definitions, we can write the terms from (4.2.4) via the following. Everything here is a function of Wx, which we drop

for simplicity. Now,

$$J(Wx) = \begin{pmatrix} | & & & \\ \partial_1\phi_{1\cdot} & 0 & & 0 \\ | & & & \\ & | & & \\ 0 & \partial_2\phi_{2\cdot} & 0 & 0 \\ & | & & \\ & & \ddots & \\ & & & | \\ 0 & 0 & 0 & \partial_p\phi_{p\cdot} \\ & & & | \end{pmatrix}$$

and

$$C(Wx) = \begin{pmatrix} \partial_1\phi_{1\cdot}\partial_1\phi_{1\cdot}^T & \dots & \dots \\ \dots & \partial_2\phi_{2\cdot}\partial_2\phi_{2\cdot}^T & \dots \\ \ddots & & \\ \dots & \dots & \partial_p\phi_{p\cdot}\partial_p\phi_{p\cdot}^T \end{pmatrix}$$

and

$$\xi(Wx) = \left(-\partial_{11}\phi_{1\cdot}^T -, \dots - \partial_{pp}\phi_{p\cdot}^T -\right)^T$$

Note that $\theta_{vu} = \theta_{uv}$, and $\theta \in \mathbb{R}^{m_1 p + m_2 p(p-1)}$. Note that we can write the objective as

$$\mathcal{L}(\theta, W, x) = \sum_{v \in V} \theta_{v\cdot}^T \partial_v \phi_{v\cdot}(Wx) \partial_v \phi_{v\cdot}(Wx)^T \theta_{v\cdot} + \theta_{v\cdot}^T \partial_{vv}^2 \phi_{v\cdot}(Wx)$$

4.6.2 W optimization: Givens rotation

The Givens orthogonal rotation algorithm is in Algorithm Algorithm 5. The main thing to figure out is how to find $\text{argmin}_\eta \hat{\mathcal{L}}(WG_{u,v}(\eta))$; for this, we compute the first derivative of the function with respect to η. For $u, v \in V$, define

$$G_{u,v}(\eta) = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & \cos\eta & \dots & -\sin\eta & \dots & 0 \\ \vdots & & & & & \vdots & \\ 0 & \dots & \sin\eta & \dots & \cos\eta & \dots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix} \tag{4.6.2}$$

That is, it has 1's down the diagonal except for columns and zeros everywhere else except for the $(u,u), (v,v), (u,v)$, and (v,u) entries, where it has $\cos\eta, \cos\eta, -\sin\eta, \sin\eta$, respectively. Fix a pair (u,v). Write G_η for simplicity. To simplify notation, let $f_v = \partial_v \phi$ and $g_v = \partial^2_{vv}\phi$. And let $y_\eta := WG_\eta x$. Our objective in η is $\mathbb{E}\mathcal{L}(\theta, WG_\eta, x)$, where

$$\mathcal{L}(\eta, x) = \mathcal{L}(\theta, WG_\eta, x) = \frac{1}{2}\theta^T \left(\sum_{v \in V} f_v(y_\eta) f_v(y_\eta)^T \right) \theta + \theta^T \left(\sum_{v \in V} g_v(y_\eta) \right)$$

Now let $h : \mathbb{R}^p \to \mathbb{R}$, let η be a scalar, and let $y_\eta := y(\eta) = (y_1(\eta), \dots, y_p(\eta))^T \in \mathbb{R}^p$ be a vector depending on η. Then $\partial_\eta h(y_\eta) = \langle \nabla h(y_\eta), \partial_\eta y_\eta \rangle$. For us, $\partial_\eta y_\eta = W \partial_\eta G_\eta x$,

where

$$\partial_\eta G_{u,v}(\eta) = \begin{pmatrix} 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & -\sin\eta & \dots & -\cos\eta & \dots & 0 \\ \vdots & & & & & \vdots & \\ 0 & \dots & \cos\eta & \dots & -\sin\eta & \dots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 0 \end{pmatrix}.$$

This is the matrix whose entries are the derivatives of the entries of G_η with respect to η. Now f_v, g_v are functions from $\mathbb{R}^p \to \mathbb{R}^K$ (so they're vectors in $\mathbb{R}^K$). And

$$\begin{aligned} \partial_\eta f_v(y_\eta) &\triangleq (\partial_\eta f_{v1}(y_\eta), \dots, \partial_\eta f_{vK}(y_\eta))^T \\ &= (\langle \nabla f_{v1}(y_\eta),\, \partial_\eta y_\eta \rangle, \dots, \langle \nabla f_{vK}(y_\eta),\, \partial_\eta y_\eta \rangle)^T \\ &= J_{f_v}(y_\eta) \partial_\eta y_\eta \end{aligned}$$

Note that by the product rule applied to each element of the matrix $f_v(y_\eta) f_v(y_\eta)^T$, $\partial_\eta \left(\theta^T \left(f_v f_v\right)^T\right) \theta\big) = \theta^T \left(\left(\partial_\eta f_v\right) f_v^T + f_v \left(\partial_\eta f_v\right)^T \right) \theta$. Thus

$$\partial_\eta \mathcal{L}(\theta, WG_\eta, x) = \frac{1}{2}\theta^T \left(\sum_{v \in V} J_{f_v}(y_\eta) \partial_\eta y_\eta f_v(y_\eta)^T + f_v(y_\eta) \partial_\eta y_\eta^T J_{f_v}(y_\eta)^T \right) \theta + \theta^T \left(\sum_{v \in V} J_{g_v}(y_\eta) \partial_\eta y_\eta \right)$$

So now assuming we can bring the derivative inside the expectation, $\partial_\eta \mathcal{L}(\theta, WG_\eta) = \mathbb{E} \partial_\eta \mathcal{L}(\theta, WG_\eta, x)$.

Graphical model form for Givens rotation

Let $f_v = \partial_v \phi_{v\cdot}$ and $g_v = \partial^2_{vv} \phi_{v\cdot}$. We can now write our objective in η in the form

$$\mathcal{L}(\theta, WG_\eta, x) = \frac{1}{2} \sum_{v \in V} \theta^T_{v\cdot} f_v(y_\eta) f_v(y_\eta)^T \theta_{v\cdot} + \sum_{v \in V} \theta^T_{v\cdot} g_v(y_\eta)$$

And

$$\partial_\eta \mathcal{L}(\theta, WG_\eta, x) = \frac{1}{2} \sum_{v \in V} \theta^T_{v\cdot} \left(J_{f_v}(y_\eta) \partial_\eta y_\eta f_v(y_\eta)^T + f_v(y_\eta) \partial_\eta y^T_\eta J_{f_v}(y_\eta)^T \right) \theta_{v\cdot} + \sum_{v \in V} \theta^T_{v\cdot} J_{g_v}(y_\eta) \partial_\eta y_\eta$$

Chapter 5

Nonparametric variational inference via score matching

5.1 Introduction

Consider the posterior inference problem: we observe data $x_1, \ldots, x_n \sim_{i.i.d.} \mathbb{P}_z$, a distribution with density $p(x|z)$. We place a prior $p(z)$ on $z \in \mathbb{R}^d$, and we wish to estimate the posterior density $p(z|x)$:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}.$$

Typically $p(x)$ is intractable, so the posterior cannot be directly computed. Markov Chain Monte Carlo (MCMC) methods are a classical tool for this problem, but these methods can be infeasible for large datasets. *Variational inference* has emerged in recent decades as a powerful way to estimate a posterior in high-dimensional settings.

Variational inference posits a family $\mathcal{Q}$ for the posterior density and minimizes a measure of divergence between $q \in \mathcal{Q}$ and the true $p(z|x)$. It is common to choose q to maximize the evidence lower bound, $\mathcal{L}(q) := \mathbb{E}_q \log p(x, z) + H(q)$, which is equivalent to minimizing the KL divergence between the true posterior and q. A

mean-field family, in which $q(z) = \prod_{i \le d} q_i(z_i)$, is often posited for $\mathcal{Q}$. When the likelihood and prior are conjugate, this results in a simple coordinate ascent algorithm in which the updates are computable in closed form. In non-conjugate models, the Laplace method, which assumes a Gaussian family for $\mathcal{Q}$ and expands $\log p(x, z)$ to a quadratic term, is a popular approach; see [Wang and Blei, 2013] for more details.

In these and other standard methods, restrictive families are chosen for $\mathcal{Q}$ because of their computational convenience. But such methods have serious weaknesses. For instance, restricting $\mathcal{Q}$ to be a mean-field family means the posterior estimate will fail to capture dependence among the posterior coordinates and will underestimate the posterior variance; see [Blei et al., 2017] for more discussion. The Gaussian form imposed on q in the Laplace method means the estimate will fail to capture multimodality in the posterior.

We propose *nonparametric variational inference*, in which we assume a nonparametric form for q and optimize it using the score matching objective. Classical score matching seeks a density q to minimize a Fisher divergence, defined via

$$\mathcal{L}(q) = \mathbb{E}_p \left\| \nabla_z \log p(z) - \nabla_z \log q(z) \right\|_2^2 . \tag{5.1.1}$$

This method is appealing for nonparametric density estimation because it does not require the calculation of a normalizing constant; this computation can be prohibitive for nonparametric densities in high dimensions. And it turns out, see Lemma 1.3.1, that (5.1.1) can be expressed as

$$\mathcal{L}(q) = \mathbb{E}_p O(q), \text{ where } O(q) = \left\| \nabla \log q \right\|_2^2 + \Delta q. \tag{5.1.2}$$

When we have samples from the true density p, we can optimize an empirical approximation of (5.1.2) in q. The problem of an intractable normalizing constant that arises in high-dimensional density estimation also appears in posterior inference, and

thus score matching is a promising technique to apply in this context,. However, the crucial difference between density estimation and posterior inference is that in the former, we have access to samples from the truth p but not to the form of p, while in posterior inference we have access to the nonnormalized form of the posterior, i.e., to $p(x, z)$, but not to samples from the posterior. In an exponential family form for q, the cancellation of the normalizing constant $p(x)$ means the objective is still optimizeable in q, but it nonetheless requires estimation of an integral over $p(x, z)$; see Section 5.2 for more discussion. Moreover, we often desire in posterior inference, as in density estimation, to obtain samples from the posterior, but it is not possible to directly sample from a nonparametric exponential family.

To circumvent the issues classical score matching poses in the posterior inference context, we turn to a related objective:

$$\mathcal{L}(q) = \mathbb{E}_q \left\| \nabla \log p - \nabla \log q \right\|_2^2 . \tag{5.1.3}$$

We show in Lemma 5.3.4 that (5.1.3) has a close relationship to the objective (5.1.1). Instead of positing an exponential family form for q, we use a novel score matching variant of [Saremi et al., 2018] to optimize a generative form for q that allows us to sample from the posterior and perform a variational EM algorithm to estimate a model.

Review of recent literature Recent literature has developed methods to allow $\mathcal{Q}$ to be a richer family. For instance, the authors of [Ranganath et al., 2016] use a dual representation of the divergence (5.1.3) and a generative form of q to avoid the use of the density itself in the variational optimization. Their algorithm requires optimization over both the parameters of q and the function in the dual representation. [Li and Turner, 2016] propose a similar method using a generalized dual repre-

sentation of divergences. These methods perform an optimization of the form:

$$\min_{q} \max_{f} \mathbb{E}_q O(f), \tag{5.1.4}$$

where $O(f)$ is an operator on the function f.

The technique of representing a divergence in its dual form has been widely used in the statistics literature for nonparametric density estimation. The classical generative adversarial network (GAN) of [Goodfellow et al., 2014], along with developments such as the Wasserstein-GAN [Arjovsky et al., 2017] and others, have relied on dual representations of the symmetric KL divergence and Wasserstein distance, respectively. Indeed, any f-divergence can be represented using the convex conjugate, or Fenchel dual form (e.g., the Donsker-Varadhan lower bound on the KL divergence). For exponential families, optimizing a representation of the density using a variational lower bound on the normalizing constant was explored in [Dai et al., 2018].

The disadvantage of relying on a dual representation for a divergence is that it requires a procedure that iterates over optimizing a divergence in a function f and minimizing over the density q. This non-convex procedure does not have algorithmic or statistical convergence guarantees. In this work, we propose to optimize the objective (5.1.3) directly, without resorting to a dual representation.

The authors of [Gershman et al., 2012] propose "nonparametric variational inference," which uses a Gaussian mixture form of q and the Delta method to approximate the model. The mixture of Gaussians for q resembles a kernel density estimator from nonparametric estimation. Moreover, this family for q, along with the Delta method, makes the optimization process fairly simple. The Gaussian-mixture family for q is also related to the "mixture of mean-field" family proposed by [Bishop et al., 1998].

5.2 Classical score matching for posterior inference

Suppose we posit an exponential family form form the posterior, i.e., $p(z|x) \propto e^{f(z|x)}$. Suppose further that $f(z|x) = f_x(z) = \gamma^\top \phi(z)$, where $\gamma \in \mathbb{R}^K$ and $\phi : \mathbb{R}^d \to \mathbb{R}^K$. We follow the notation of Section 1.3 except that we replace the argument x therein with z since we are estimating the posterior, a function of z. Letting $A(z), k(z)$ be as defined in Lemma 1.3.2, the objective (5.1.1) simplifies to:

$$\int_z p(z|x) \left(\frac{1}{2} \gamma^\top A(z) \gamma + \gamma^\top k(z) \right) \, dz,$$

which has optimal solution

$$\begin{aligned} \hat{\gamma} &= \left(\int p(z|x) A(z) \, dz \right)^{-1} \int p(z|x) k(z) \, dz \\ &= \left(\int \frac{p(x,z)}{p(x)} A(z) \, dz \right)^{-1} \int \frac{p(z,x)}{p(x)} k(z) \, dz \\ &= \bar{A}^{-1} \bar{k}, \end{aligned} \tag{5.2.1}$$

where

$$\begin{aligned} \bar{A} &= \int p(x,z) A(z) \, dz \\ \bar{k} &= \int p(x,z) k(z) \, dz. \end{aligned}$$

Crucially, the solution (5.2.1) requires only integrals over $p(z,x)$, not over the uncomputable $p(z|x)$, since in the quadratic form, the $p(x)$ cancels out. Now (5.2.1) is something we can calculate. There is still an important difficulty: computing the integrals $\bar{A}, \bar{k}$ over the joint density. We propose importance sampling for this, and our proposed algorithm is described in Algorithm 7.

Algorithm 7: Classical score matching for nonparametric variational inference with an exponential family

Input: Dataset $\{x_i\}_{i\in[n]}$ with each point in $\mathbb{R}^d$, likelihood model $p(x|z)$, prior $p(z)$, orthogonal basis model $\phi(z)$ for posterior, functions $A(z), k(z)$ as defined in Lemma 1.3.2, importance-sampling density $h(z)$

Output: Estimate $p(z|x)$.

Sample $z_1, \ldots, z_m \sim_{i.i.d.} P$ where P is a distribution on $\mathbb{R}^d$ with density h ;

Compute $\bar{A} = \sum_{j=1}^m \frac{p(x,z_j)}{h(z_j)} A(z_j)$;

Compute $\bar{k} = \sum_{j=1}^m \frac{p(x,z_j)}{h(z_j)} k(z_j)$;

Report

$$\hat{\gamma} = \bar{A}^{-1}\bar{k} \tag{5.2.2}$$

$$\log p(z|x) \propto \hat{\gamma}^\top \phi(z). \tag{5.2.3}$$

Several major problems arise in Algorithm 7. First, often $p(x|z)$ is extremely small if n is large, making importance sampling difficult, especially in high dimensional examples.

Second, and more importantly, suppose that $\phi(z)$ is a polynomial; such an assumption is common in nonparametric density estimation. For instance, suppose $\phi(z) = 1 + z + z^2 + \ldots$. Then the first and second derivative functions, $A(z), k(z)$, will contain constant terms. This means that for example, $\bar{k}$ contains terms like $\int p(z, x)1\, dz$, which is precisely $p(x)$. In such an example, we would simply be estimating $p(x)$ via importance sampling. If we are going to do this, we might as well compute $p(z|x)$ directly since approximating $p(x)$ was the major obstacle to directly obtaining the posterior. Note that a polynomial form for ϕ is common, e.g., if the

family is Gaussian; see Example 2.

Finally, in the discussion of Algorithm 7, we implicitly assumed that there is one posterior variable $z \in \mathbb{R}^d$. Suppose we have a slightly more complex model, with data $x_1, \ldots, x_n$ and prior variables $z_1, \ldots, z_n$. Suppose the model is $p(x, z) = \prod_{i=1}^n p(x_i, z_i)$, and suppose we assume a mean-field family for the posterior estimate q, i.e., $q(z) = \prod_{i=1}^n q_i(z_i)$. Suppose further we assume q belongs to an exponential family, so $q(z) = \exp\left(\gamma_1^\top \phi_1(z_1) + \ldots + \gamma_n^\top \phi_n(z_n)\right)$. Then the solution in Algorithm 7 will involve the term $\bar{k}$ satisfying

$$\begin{aligned}
\bar{k} &= \int_z p(x, z) \left(\phi_1''(z_1), \ldots, \phi_n''(z_n)\right)^\top \, dz \\
&= \int_z \prod_{i=1}^n p(x_i, z_i) \left(\phi_1''(z_1), \ldots, \phi_n''(z_n)\right)^\top \, dz_1, \ldots, dz_n \\
&= \left(p(x_2, \ldots, x_n) \int p(x_1, z_1) \phi_1''(z_1) \, dz_1, \ldots, p(x_1, \ldots, x_{n-1}) \int p(x_n, z_n) \phi_1''(z_n) \, dz_n\right)^\top.
\end{aligned}$$

That is, in the mean-field case, we would again have to compute terms of the form $p(x_2, \ldots, x_n)$, which is exactly what we are trying to avoid.

We have implemented Algorithm 7 in a simple, low-dimensional case; see Nonparametric variational inference via score matching. In these experiments, we assume that q is in fact a bounded density. Score matching for a bounded density takes a slightly different form than score matching for an unbounded density; we provide the form and details in Chapter 6.

5.3 Score matching for generative nonparametric variational inference

In this section, we propose an optimization that uses score matching but circumvents some of the issues noted in Section 5.2. We propose to directly optimize (5.1.3)

via kernel density estimation and the reparametrization trick, relying on the kernel-density form of score matching discussed in Section 1.3.

First, we posit a generative form for the posterior q. We generate Gaussian noise, then feed it through a multilayer perceptron f_ϕ to generate samples from q. That is, we draw

$$\begin{aligned} z_{0i} &\sim N(0, I_{d_0}) \\ z_i &= f_\phi(z_{0i}). \end{aligned} \tag{5.3.1}$$

We wish to estimate the parameters ϕ. Here $z_i \in \mathbb{R}^d$, and we will let $d_0 < d$. We will let f_ϕ be a multilayer perceptron with some nonlinear activation function. Since the transformation f_ϕ is not invertible, we cannot use the change of variable density formula to obtain an exact formula for $q(z)$. Thus, it seems impossible to optimize (5.1.3) directly since it requires the form of q in order to compute $\nabla_z \log q(z)$. We will circumvent this issue by representing the posterior q_ϕ via a kernel density representation, as described in Section 1.3.

To run initial tests of our proposal, we generate synthetic data using a variational-EM algorithm to estimate the generative model parameters. We imitate the variational autoencoder of [Kingma and Welling, 2014], but in our case, the variational inference step is done as described above rather than using traditional variational inference. We posit a model

$$p_\theta(x, z) = \prod_{i=1}^n p_\theta(x_i, z_i) = \prod_{i=1}^n p_\theta(x_i|z_i)p(z_i),$$

and we posit an *amortized mean-field* form for the posterior:

$$q_\phi(z|x) = \prod_{i=1}^n q_\phi(z_i|x_i),$$

where the parameter ϕ is common to all data points. We estimate ϕ and θ as described below. To reproduce data samples, we generate from $q_\phi(z_i|x_i)$, then generate a new

sample x according to $p_\theta(x_i|z_i)$. The complete algorithm is provided in Algorithm 8.

Algorithm 8: Nonparametric generative variational inference via score matching

---

Input: Dataset $\{x_i\}_{i\in[n]}$ with each point in $\mathbb{R}^d$, likelihood model $p(x|z)$, prior $p(z)$, stopping criteria T

Output: Estimate $q_\phi(z|x)$ for the posterior.

Initialize ϕ, θ ;

Set $t = 0$;

While $t < T$

Generate $z_{0i} \sim N(0, I_d)$ for $i \in [n]$ and $\xi_{ij} \sim N(0, I_d)$ for $j \in [m]$;

Optimize the posterior via (stochastic) gradient descent:

$\min_\phi \sum_{i\in[n],j\in[m]}^n \|f_\phi(z_{0i}) - \xi_{ij} - \nabla_\xi \log p_\theta(x_i, f_\phi(\xi_{ij}))\|_2^2$;

Optimize the the likelihood via (stochastic) gradient descent:

$\max_\theta \sum_{i=1}^n \log p_\theta(x_i, f(z_{0i}))$;

Report estimates

$$\hat{\phi}, \hat{\theta}. \tag{5.3.2}$$

To reproduce samples:

Select x from dataset ;

Draw $z_0 \sim N(0, I_{d_0})$;

Draw $z(x) = f_{\hat{\phi}(x)}(z_0)$;

Draw $\hat{x} = p_{\hat{\theta}}(x|z(x))$.

To produce new samples:

Draw $x = p_{\hat{\theta}}(x|z)$ where $z \sim N(0, I_d)$.

We run preliminary tests of Algorithm 8 on the MNIST handwritten digit dataset using TensorFlow to perform the optimizations. Each data point x_i in the MNIST dataset lies in $d = 784$ dimensions, with each coordinate representing a pixel, which takes a value in $[0, 1]$. We posit the model, for $i \in [n]$,

$$z_i \sim_{i.i.d.} N(0, I_d)$$
$$x_i = \frac{\exp(\eta^\top \beta(z_i))}{1 + \exp(\eta^\top \beta(z_i))}.$$

Here, the model parameter is $\theta = (\eta, \beta)$, where $\eta \in \mathbb{R}^K$ and $\beta : \mathbb{R}^d \to \mathbb{R}^K$ is a multilayer perceptron. These experiments are available in the Jupyter notebooks in the Github repository Variational autoencoder via score matching and ratio matching.

Ratio matching We propose one other alternative. First, notice the following. For small h,

$$\nabla_z \log q(z) \approx \frac{\log q(z+h) - \log q(z)}{h} = \frac{1}{h} \log \frac{q(z+h)}{q(z)}$$

This is the basis of ratio-matching, which approximates score matching using the above representation of the derivative. Draw samples z_1, z_2 from a sampling form q, as in (5.3.1), and optimize

$$\mathbb{E}_q \left\| \log \frac{q(z_1)}{q(z_2)} - \log \frac{p(x, z_1)}{p(x, z_2)} \right\|_2^2. \tag{5.3.3}$$

This imitates the method of [Hyvarinen, 2007]. To begin, we use a Gaussian form for q. Since we can sample from this, we use the reparametrization trick to estimate the integral in (5.3.3).

5.3.1 Decomposition of the objective

In this section, we show that the score-matching objective (5.1.3) decomposes into two components: a component matching q to p, and a component restricting q to be smooth. This resembles the evidence lower bound of $\mathbb{E}_q \log p(x,z) + H(q)$ used in classical variational inference; the first term matches the posterior estimate to the log likelihood, the second constrains the variance of q.

Lemma 5.3.1. *Suppose* $q(z) = \exp(f(z) - \Psi)$ *where* $\Psi = \int_z \exp(f(z))\,dx$. *Let* $\mathcal{L}(q) = \mathbb{E}_{q(z|x)} \|\nabla_z \log p(z|x) - \nabla_z \log q(z|x)\|_2^2$. *Then*

$$\mathcal{L}(q) =^c \frac{1}{2}\mathbb{E}_q \|\nabla_z \log p(x,z)\|_2^2 + \mathbb{E}_q \Delta \log p(x,z) + \mathbb{E}_q \|\nabla_z f(z)\|_2^2 \tag{5.3.4}$$

In theory, we could directly optimize (5.3.4), since in posterior inference we have the form of $\log p(x,z)$ and therefore we have its gradients. Moreover, $\nabla_z \log q(z)$ does not require the normalizing constant of q, so we could optimize this in q if we had, e.g., an exponential family form for q. The difficulty would be estimating the integral over q when q is from a nonparametric family; there is no general way to do this. In general, to have a viable sampling form for q and an explicit form of $\nabla \log q$, we would need to assume a Gaussian form for q, which we of course wish to avoid. Therefore, we might simply assume a generative form for q and approximate $\nabla \log q$ via a kernel density estimate as previously; optimizing (5.3.4) would thus be an alternative to Algorithm 8.

5.3.2 Connections to the dual representation

The objective (5.1.3) is related to a form of score matching used in [Ranganath et al., 2016] and we now discuss connections between our proposal and that paper and other recent literature. Let $\psi_p(z) = \frac{\partial \log p(z)}{\partial z}$. The authors of [Ranganath et al., 2016] propose

to optimimze following variational objective in q:

$$\sup_{f \in \mathcal{F}} |\mathbb{E}_{q(z)}(O^p_{LS} f)(z)| \tag{5.3.5}$$

where

$$O^p_{LS} f = \psi_p f + \nabla f. \tag{5.3.6}$$

As noted in [Ranganath et al., 2016], the objective (5.3.5) is reasonable since it is zero when $q = p$, as we show in Lemma 5.3.2. [Ranganath et al., 2016] optimize (5.3.5) by iterating between maximizing over f and minimizing over q; a generative multilayer perceptron form similar to ours is used for q, while f is another neural network. This optimization is nonconvex and is not guaranteed to reach some global optimum.

We can relate the objective (5.3.5) to a Stein discrepancy and then to a score matching objective (5.1.3). We state this in Lemma 5.3.3. It turns out that this objective has a close relationship to that in (5.1.1) and (5.3.5), as we state in Lemma 5.3.4. Proofs for these lemmas are provided in Section 5.4.1.

Lemma 5.3.2. *Suppose* $\lim_{z \to \infty} q(z) = \lim_{z \to -\infty} q(z) = 0$. *Then*

$$\mathbb{E}_q \left(\psi_q f + \nabla f \right) = 0$$

Lemma 5.3.3 (Equivalence of score-matching, Stein discrepancy, and OVI)**.** *Let* $\mathcal{F}$ *be a function class, and let* $\mathcal{G}$ *be a class of functions of the form* $g(z) = \psi_p(z) f(z) + f'(z)$, *for* $f \in \mathcal{F}$. *Then*

$$\sup_{g \in \mathcal{G}} |\mathbb{E}_q g| = \sup_{g \in \mathcal{G}} |\mathbb{E}_q g - \mathbb{E}_p g| \qquad \leftarrow \textit{Stein discrepancy} \tag{5.3.7}$$

$$= \mathbb{E}_q (\psi_q - \psi_p)^2. \qquad \leftarrow \textit{Score matching loss} \tag{5.3.8}$$

Lemma 5.3.3 motivates the use of the alternative score matching objective (5.1.3). We now show how it relates to the objective (5.1.1).

Lemma 5.3.4. *Let p, q be densities that are positive everywhere. Let $\mathcal{H}$ be a class of functions, and let $\mathcal{F}$ be a class of functions of the form $f(z) = \frac{p(z)}{q(z)}h(z)$ for $h \in \mathcal{H}$. Then*

$$\sup_{f \in \mathcal{F}} |\mathbb{E}_q O^p_{LS} f| = \sup_{h \in \mathcal{H}} |\mathbb{E}_p O^q_{LS} h|$$

5.4 Appendix

5.4.1 Proofs

Proof of Lemma 5.3.1. We provide the proof in one dimension only.

$$\mathcal{L}(q) = \int q \left(\nabla \log p\right)^2 - 2 \int q (\nabla \log q)(\nabla \log p) + \int q \left(\nabla \log q\right)^2$$

For the middle term, using integration by parts,

$$\int q \frac{q'}{q} (\nabla \log p) = \int q' (\nabla \log p) \stackrel{c}{=} - \int q \nabla^2 \log p.$$

□

Proof of Lemma 5.3.2. First, by the product rule,

$$\mathbb{E}_p \left(\psi_q f + f'\right) = \int \frac{p}{q} (qf)'$$

So $\mathbb{E}_q \left(\psi_q f + f'\right) = \int (qf)' = q(\infty) f(\infty) - q(-\infty) f(-\infty) = 0.$ □

Proof of Lemma 5.3.3. By Lemma 5.3.2, $\mathbb{E}_p \psi_p f + f' = 0$. So

$$\begin{aligned}\sup_{g \in \mathcal{G}} |\mathbb{E}_p g| &= \sup_{f \in \mathcal{F}} |\mathbb{E}_p \left(\psi_q f + f' - \psi_p f - f'\right)| \\ &= \sup_{f \in \mathcal{F}} |\mathbb{E}_p (\psi_q - \psi_p) f| \end{aligned} \tag{5.4.1}$$

And this is maximized when $f(z) = \psi_q(z) - \psi_p(z)$. □

Proof of Lemma 5.3.4. This is clear if we use the representation in (5.4.1). That is,

$$\sup_{f \in \mathcal{F}} |\mathbb{E}_q O^p_{LS} f| = \sup_{f \in \mathcal{F}} \mathbb{E}_q (\psi_q - \psi_p) f = \sup_{h \in \mathcal{H}} \mathbb{E}_q \frac{p}{q} (\psi_q - \psi_p) h = \sup_{h \in \mathcal{H}} \mathbb{E}_p (\psi_q - \psi_p) h = \sup_{h \in \mathcal{H}} |\mathbb{E}_p O^q_{LS} h|$$

Alternatively, we can show this in the following way. We drop the argument z for clarity. The notation p' means the derivative of p with respect to z, and the notation $(ph)'$ means the derivative of the product with respect to z, which is, by the product rule, $p'h + h'p$.

$$\begin{aligned}\sup_{f \in \mathcal{F}} |\mathbb{E}_q O^p_{LS} f| &= \sup_{f \in \mathcal{F}} |\mathbb{E}_q \left(\psi_p f + f'\right)| \\ &= \sup_{h \in \mathcal{H}} \left| \int q \frac{p'}{p} \frac{p}{q} h + \int q \frac{q(ph)' - q'ph}{q^2} \right| \\ &= \sup_{h \in \mathcal{H}} \left| \int p'h + \int (ph)' - \int p \frac{q'}{q} h \right|.\end{aligned}$$

The middle term is zero by Lemma 5.3.2. Adding and subtracting $\int ph'$ yields:

$$\begin{aligned}&= \sup_{h \in \mathcal{H}} \left| \int p'h + \int ph' - \int ph' - \int p \frac{q'}{q} h \right| \\ &= \sup_{h \in \mathcal{H}} \Bigg| \underbrace{\int (ph)'}_{=0 \text{ by Lemma 5.3.2}} - \int p \left(h' + \frac{q'}{q} h \right) \Bigg| \\ &= \sup_{h \in \mathcal{H}} \left| \mathbb{E}_p \psi_q h + \nabla h \right|.\end{aligned}$$

□

5.4.2 Examples

We now provide further discussion of the potential problems with using score matching as in (5.1.1) for posterior inference. We first show how the optimization would look if q is restricted to be Gaussian.

Example 2 (Score-matching with $\mathbb{E}_p$ for posterior inference when q has Gaussian form)**.** Suppose q has form $N(\mu_x, \Sigma_x)$ where $\mu_x \in \mathbb{R}^d, \Sigma_x \in \mathbb{R}^{d\times d}$. Let our prior $p(z)$ be the $N(0, I_d)$ density. Note that we could have $d \gg n$. We have $q(z) \propto e^{g(z)}$ where

$$g(z) = \gamma^\top \phi(z)$$

where $K = 2d$ and

$$\gamma = \left(\Sigma^{-1}, \Sigma^{-1}\mu\right)^\top$$
$$\phi(z) = \left(-\frac{1}{2}zz^\top, z\right)^\top$$

So (for $d = 1$, to keep it simple):

$$\frac{\partial\phi(z)}{\partial z} = (-z, 1)^\top$$
$$\frac{\partial^2\phi(z)}{\partial z^2} = (1, 0)^\top$$

We have $A(z) \in \mathbb{R}^{2d\times 2d}$. Let e.g. z^2 indicate $(z_1^2, \ldots, z_d^2)$. We have

$$A(z) = \begin{pmatrix} diag(z^2) & diag(-z) \\ diag(-z) & diag(1) \end{pmatrix} \text{ and}$$
$$k(z) = (rep(-1, d), rep(0, d))^\top$$

Each submatrix is in $\mathbb{R}^{d\times d}$. We parameterize γ via B; it is some multi-layer non-linear

function with many parameters B. Our objective, written as a sum, is

$$\mathbb{E}_{p(z|x)}\left(\sum_{j\leq d}\gamma_{1j}^2 z_j^2 - \sum_{j\leq d}\gamma_{2j}\gamma_{1j}z_j + \sum_{j\leq d}\gamma_{2j}^2 - \sum_{j\leq d}\gamma_{1j}\right)$$

Note that as noted in [Janofsky, 2015] and [Lin et al., 2016], there are closed-form solutions of μ, Σ. To see it, note that if $q(z)$ is $N(\mu, \Sigma)$,

$$\log q(z) \propto \frac{-(z-\mu)'\Sigma^{-1}(z-\mu)}{2}$$

So the score-matching objective is

$$\frac{1}{2}\left\|\nabla_z \log q(z)\right\|_2^2 + \Delta_z \log q(z) = \frac{1}{2}\left\|\Sigma^{-1}(z-\mu)\right\|_2^2 + tr(\Sigma^{-1})$$

We can directly obtain:

$$\hat{\mu} = \mathbb{E}_{p(z|x)} z$$
$$\hat{\Sigma}_i = \mathbb{E}_{p(z|x)}(z-\hat{\mu})(z-\hat{\mu})^t op$$

While this is estimable for density estimation, it requires the posterior for us. Why does this happen when we can still not require $p(x)$ if we compute the exponential family parameter? Notice that e.g. in the simple $d = 1$ case, the quadratic we'd obtain for score-matching is:

$$\gamma'\mathbb{E}_{p(z|x)}\begin{pmatrix} z^2 & z \\ z & 1 \end{pmatrix}\gamma - \gamma'\mathbb{E}_{p(z|x)}\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Optimizing this allows us to remove the $p(x)$, but notice that then, we optimize in γ

the quadratic $\gamma^\top \bar{A}\gamma^\top - \gamma^\top \bar{k}$, where

$$\bar{k} = \begin{pmatrix} p(x) \\ 0 \end{pmatrix}$$

Now again we can approximate $p(x)$ via importance sampling, but we do have to approximate it either way.

Chapter 6

Appendix

We provide proofs of some auxiliary lemmas that were used in the thesis.

Proof of Lemma 1.3.1.

$$\mathcal{L}(q) \stackrel{c}{=} \int p \left(\nabla \log q\right)^2 - 2 \int p (\nabla \log p)(\nabla \log q).$$

Since $\nabla \log p = p'/p$, the second term is $-2 \int p' \nabla \log q$. Using integration by parts on the second term,

$$\int p' \nabla \log q \stackrel{c}{=} - \int p \nabla^2 q.$$

□

Proof of Lemma 1.3.2. This follows directly from Lemma 1.3.1 and plugging in the form of the exponential family. We have

$$\frac{\partial \phi(x)}{\partial x_i} = \left(\frac{\partial \phi_1(x)}{\partial x_i}, \ldots, \frac{\partial \phi_K(x)}{\partial x_i} \right)^T$$
$$\frac{\partial^2 \phi(x)}{\partial x_i^2} = \left(\frac{\partial^2 \phi_1(x)}{\partial x_i^2}, \ldots, \frac{\partial^2 \phi_K(x)}{\partial x_i^2} \right)^T$$

Since $\log q(x) \propto g(x)$,

$$\begin{aligned}
\frac{1}{2}\left\|\nabla \log q\right\|_2^2 + \triangle \log q &= \sum_{i\leq d}\left(\frac{1}{2}\left(\frac{\partial g(x)}{\partial x_i}\right)^2 + \frac{\partial^2 g(x)}{\partial x_i^2}\right) \\
&= \sum_{i\leq d}\left(\frac{1}{2}\left(\gamma^\top \frac{\partial \phi(x)}{\partial x_i}\right)^2 + \gamma^\top \frac{\partial^2 \phi(x)}{\partial x_i^2}\right) \\
&= \frac{1}{2}\gamma^\top\left(\sum_{i\leq d}\frac{\partial \phi(x)}{\partial x_i}\frac{\partial \phi(x)}{\partial x_i}^\top\right)\gamma + \gamma^\top \sum_{i\leq d}\frac{\partial^2 g(x)}{\partial x_i^2} \\
&= \frac{1}{2}\gamma^\top A(x)\gamma + \gamma^\top k(x).
\end{aligned}$$

So $\mathcal{L}(q) \stackrel{c}{=} \frac{1}{2}\gamma^\top\left(\mathbb{E}_{p(x)}A(x)\right)\gamma + \gamma^\top\left(\mathbb{E}_{p(x)}k(x)\right)$, and the conclusion follows from optimizing this quadratic in γ. □

Proof of Lemma 1.3.3.

$$\mathcal{L}_{kde}(\theta) \stackrel{c}{=} \mathbb{E}_{p(\xi)}\left\|\nabla_\xi \log q_\theta(\xi)\right\|_2^2 - 2\mathbb{E}_{p(\xi)}\langle \nabla_\xi \frac{1}{n}\sum_{i=1}^n K(\xi|x_i),\, \nabla_\xi \log q_\theta(\xi)\rangle. \tag{6.0.1}$$

Writing out the second term:

$$\begin{aligned}
\mathbb{E}_{p(\xi)}\langle \nabla_\xi \frac{1}{n}\sum_{i=1}^n K(\xi|x_i),\, \nabla_\xi \log q_\theta(\xi)\rangle &= \int_\xi \frac{1}{n}\sum_{i=1}^n K(\xi|x_i)\langle \nabla_\xi \frac{1}{n}\sum_{i=1}^n K(\xi|x_i),\, \nabla_\xi \log q_\theta(\xi)\rangle\, d\xi \\
&= \int_\xi \langle \frac{1}{n}\sum_{i=1}^n \nabla_\xi K(\xi|x_i),\, \nabla_\xi \log q_\theta(\xi)\rangle\, d\xi \qquad (6.0.2) \\
&= \frac{1}{n}\sum_{i=1}^n \int_\xi \langle \nabla_\xi K(\xi|x_i),\, \nabla_\xi \log q_\theta(\xi)\rangle\, d\xi.
\end{aligned}$$

where (6.0.2) follows because $\nabla_\xi \sum_{i=1}^n K(\xi|x_i) = \frac{\sum_{i=1}^n \nabla_\xi K(\xi|x_i)}{\sum_{j=1}^n K(\xi|x_j)}$. And so plugging in to (6.0.1),

$$\begin{aligned}
\mathcal{L}_{kde}(\theta) &\stackrel{c}{=} \frac{1}{n}\sum_{i=1}^n \int_\xi \|\nabla_\xi \log q_\theta(\xi)\|_2^2 - 2\frac{1}{n}\sum_{i=1}^n \int_\xi \langle \nabla_\xi K(\xi|x_i),\ \nabla_\xi \log q_\theta(\xi)\rangle\, d\xi \\
&= \frac{1}{n}\sum_{i=1}^n \int_\xi \left(\|\nabla_\xi \log q_\theta(\xi)\|_2^2 - 2\langle \nabla_\xi K(\xi|x_i),\ \nabla_\xi \log q_\theta(\xi)\rangle\right)\, d\xi \\
&\stackrel{c}{=} \mathbb{E}_{(\xi,x)\sim J} \|\nabla_\xi \log K(\xi|x_i) - \nabla_\xi \log q_\theta(\xi)\|_2^2 .
\end{aligned}$$

□

For a bounded density, score matching has a slightly different objective. See [Janofsky, 2015] for more details. We report the basis of bounded-density score matching here because it is in fact what we used for the classical score matching for variational inference implementation in Nonparametric variational inference via score matching.

$$\begin{aligned}
k_1(z) &= \sum_{i\le d} 2(2z_i - 1)z_i(1-z_i)\frac{\partial \phi(z)}{\partial z_i} \\
k_2(z) &= \sum_{i\le d} z_i^2(1-z_i)^2\frac{\partial^2 \phi(z)}{\partial z_i^2} \\
k(z) &= k_1(z) - k_2(z) \\
A(z) &= \sum_{i\le d} \frac{\partial \phi(z)}{\partial z_i}\frac{\partial \phi(z)}{\partial z_i}' z_i^2(1-z_i)^2
\end{aligned}$$

For a bounded density (the second line is for an exponential family), the objective is:

$$\begin{aligned}
h_\gamma(z) &= \sum_{i\le d} \frac{1}{2}\left(\frac{\partial g(z)}{\partial z_i} z_i(1-z_i)\right)^2 - 2(2z_i-1)z_i(1-z_i)\frac{\partial g(z)}{\partial z_i} + z_i^2(1-z_i)^2\frac{\partial^2 g(z)}{\partial z_i^2} \\
&= \sum_{i\le d} \left(\gamma'\frac{\partial \phi(z)}{\partial z_i} z_i(1-z_i)\right)^2 - 2(2z_i-1)z_i(1-z_i)\gamma'\frac{\partial \phi(z)}{\partial z_i} + z_i^2(1-z_i)^2\gamma'\frac{\partial^2 \phi(z)}{\partial z_i^2}
\end{aligned}$$

In one dimension, and for $z \in [0,1]$, this simplifies to the following. Let $\phi : [0,1] \to$

$\mathbb{R}^K$.

$$
\begin{aligned}
h_\gamma(z) &= \frac{1}{2}\left(\frac{\partial g(z)}{\partial z} z(1-z)\right)^2 - 2(2z-1)z(1-z)\frac{\partial g(z)}{\partial z} + z(1-z)\frac{\partial^2 g(z)}{\partial z^2} \\
&= \frac{1}{2} z^2(1-z)^2 \gamma' A(z)\gamma - 2(2z-1)z(1-z)\gamma' k_1(z) + z^2(1-z)^2\gamma' k_2(z)
\end{aligned}
$$

where

$$
\begin{aligned}
k_1(z) &= \left(\frac{\partial \phi_1(z)}{\partial z}, \ldots, \frac{\partial \phi_K(z)}{\partial z}\right)' \\
k_2(z) &= \left(\frac{\partial^2 \phi_1(z)}{\partial z^2}, \ldots, \frac{\partial^2 \phi_K(z)}{\partial z^2}\right)' \\
k(z) &= k_1(z) - k_2(z) \\
A(z) &= k_1(z) k_1(z)'
\end{aligned}
$$

Bibliography

[Abramowitz and Stegun, 1964] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation.

[Acharya et al., 2014] Acharya, J., Jafarpour, A., Orlitksy, A., and Suresh, A. T. (2014). Near-optimal sample estimators for spherical gaussian mixtures. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Neural Information Processing Systems*, pages 1395–1403. Curran Associates, Inc.

[Achlioptas and Mcsherry, 2010] Achlioptas, D. and Mcsherry, F. (2010). On spectral learning of mixtures of distributions.

[Anandkumar et al., 2014] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832.

[Andersen et al., 2013] Andersen, M. S., Dahl, J., and Vandenberghe, L. (2013). Cvxopt: A python package for convex optimization.

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 214–223.

[Arora and Kannan, 2005] Arora, S. and Kannan, R. (2005). Learning mixtures of separated nonspherical gaussians. *Annals of Applied Probability*, 15(1A):69–92.

[Ashtiani et al., 2018] Ashtiani, H., Ben-David, S., Harvey, N. J. A., Liaw, C., Mehrabian, A., and Plan, Y. (2018). Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Neural Information Processing Systems*.

[Bach and Jordan, 2003] Bach, F. R. and Jordan, M. I. (2003). Beyond independent components: trees and clusters. *Journal of Machine Learning Research*, 4:1205–1233.

[Banach, 1938] Banach, S. (1938). Über homogene polynome in ($l^{\{}2\}$). *Studia Mathematica*, 7(1):36–44.

[Belkin and Niyogi, 2008] Belkin, M. and Niyogi, P. (2008). Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Computer Systems Science*, 74(8):1289–1308.

[Belkin and Sinha, 2009] Belkin, M. and Sinha, K. (2009). Learning gaussian mixtures with arbitrary separation.

[Birgé, 1983] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. für Wahrscheinlichkeitstheorie und Verw. Geb.*, 65(2):181–237.

[Birgé, 1986] Birgé, L. (1986). On estimating a density using hellinger distance and some other strange facts. *Probability theory and related fields*, 71(2):271–291.

[Bishop et al., 1998] Bishop, C., Lawrence, N., Jaakkola, T., and Jordan, M. I. (1998). Approximating posterior distributions in belief networks using mixtures. In *Advances in Neural Information Processing Systems*, volume 10, pages 416–422.

[Blei et al., 2017] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

[Boucheron and Thomas, 2012] Boucheron, S. and Thomas, M. (2012). Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17(51):1–12.

[Chen, 1995] Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1):221–233.

[Comon et al., 2008] Comon, P., Golub, G., Lim, L.-H., and Mourrain, B. (2008). Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279.

[Dacunha-Castelle and Gassiat, 1997] Dacunha-Castelle, D. and Gassiat, E. (1997). The estimation of the order of a mixture model. *Bernoulli*, 3(3):279–299.

[Dai et al., 2018] Dai, B., Dai, H., Gretton, A., Song, L., Schuurmans, D., and He, N. (2018). Kernel exponential family estimation via doubly dual embedding. arXiv:1811.02228v1.

[Dasgupta, 1999] Dasgupta, S. (1999). Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, page 634, USA. IEEE Computer Society.

[Davis and Kahan, 1970] Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a peturbation. *SIAM Journal of Numerical Analysis*, 7.

[Diakonikolas et al., 2017] Diakonikolas, I., Kane, D., and Stewart, A. (2017). Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *IEEE 58th Annual Symposium on Foundations of Computer Science*, pages 73–84.

[Diamond and Boyd, 2016] Diamond, S. and Boyd, S. (2016). Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

[Elhamifar and Vidal, 2009] Elhamifar, E. and Vidal, R. (2009). Sparse subspace clustering. In *In CVPR*.

[Feldman et al., 2006] Feldman, J., Servedio, R. A., and O'Donnell, R. (2006). Pac learning axis-aligned mixtures of gaussians with no separation assumption. In Lugosi, G. and Simon, H. U., editors, *Learning Theory*, pages 20–34, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Flamary and Courty, 2017] Flamary, R. and Courty, N. (2017). Pot: Python optimal transport library.

[Friedland and Lim, 2018] Friedland, S. and Lim, L.-H. (2018). Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281.

[Genovese and Wasserman, 2000] Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the gaussian mixture sieve. *Annals of Statistics*, 28(4):1105–1127.

[Gershman et al., 2012] Gershman, S. J., Hoffman, M. D., and Blei, D. M. (2012). Nonparametric variational inference. In *Proceedings of the 29th International Conference on Machine Learning*.

[Ghosal and van der Vaart, 2001] 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. *Annals of Statistics*, pages 1233–1263.

[Gibbs and Su, 2002] Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435.

[Goodfellow et al., 2014] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

[Gradshteyn and Ryzhik, 2007] Gradshteyn, I. and Ryzhik, I. (2007). *Table of Integrals, Series, and Products*. Elsevier, 7 edition.

[Hansen, 1982] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

[Hardt and Price, 2015] Hardt, M. and Price, E. (2015). Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on theory of computing*, pages 753–760.

[Hartigan, 1985] Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 807–810.

[Heinrich and Kahn, 2018] Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, 46(6A):2844–2870.

[Ho and Nguyen, 2016] Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44(6):2726–2755.

[Hopkins and Li, 2018] Hopkins, S. B. and Li, J. (2018). Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pages 1021–1034, New York, NY, USA. Association for Computing Machinery.

[Horn and Johnson, 1991] Horn, T. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. Cambridge, 1 edition.

[Hsu and Kakade, 2013] Hsu, D. and Kakade, S. M. (2013). Learning mixtures of spherical gaussians: moment methods and spectral decompositions. *Fourth Innovations in Theoretical Computer Science*.

[Hyvarinen, 2005] Hyvarinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709.

[Hyvarinen, 2007] Hyvarinen, A. (2007). Some extensions of score matching. *Computational Statistics and Data Analysis*, (51):2499–2512.

[Hyvärinen and Morioka, 2016] Hyvärinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, pages 3765–3773.

[Hyvärinen and Morioka, 2017] Hyvärinen, A. and Morioka, H. (2017). Nonlinear ica of temporally dependent stationary sources. In *The 20th International Conference on Artificial Intelligence and Statistics*.

[Hyvärinen and Oja, 2000] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430.

[Hyvärinen et al., 2019] Hyvärinen, A., Sasaki, H., and Turner, R. E. (2019). Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Con- ference on Artificial Intelligence and Statistics*, pages 859–868.

[Ibragimov, 2001] Ibragimov, I. (2001). *Estimation of analytic functions*, volume Volume 36 of *Lecture Notes–Monograph Series*, pages 359–383. Institute of Mathematical Statistics, Beachwood, OH.

[Janofsky, 2015] Janofsky, E. (2015). *Exponential series approaches for nonparametric graphical models*. PhD thesis, University of Chicago.

[Jing et al., 2016] Jing, L., Shen, Y., Dubcek, T., Peurifoy, J., Skirlo, S. A., Tegmark, M., and Soljacic, M. (2016). Tunable efficient unitary neural networks (EUNN) and their application to RNN. *CoRR*, abs/1612.05231.

[Kalai et al., 2010] Kalai, A. T., Moitra, A., and Valiant, G. (2010). Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on theory of computing*, pages 553–562.

[Kannan et al., 2005] Kannan, R., Salmasian, H., and Vempala, S. (2005). The spectral method for general mixture models. In *18th Annual Conference on Learning Theory (COLT*, pages 444–457.

[Karoui, 2010a] Karoui, N. E. (2010a). On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5):3191–3216.

[Karoui, 2010b] Karoui, N. E. (2010b). The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50.

[Karoui and tieng Wu, 2014] Karoui, N. E. and tieng Wu, H. (2014). Connection graph laplacian methods can be made robust to noise.

[Khemakhem et al., 2020] Khemakhem, I., Monti, R. P., Kingma, D. P., and Hyvärinen, A. (2020). Ice-beem: Identifiable conditional energy-based deep models.

[Kim, 2014] Kim, A. K. H. (2014). Minimax bounds for estimation of normal mixtures. *Bernoulli*, 20(4):1802–1818.

[Kingma and Welling, 2014] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *ICLR*.

[Koenker and Mizera, 2014] Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685.

[Köster et al., 2009] Köster, U., Lindgren, J., and Hyvärinen, A. (2009). Estimating markov random field potentials for natural images. In T. Adali, C. Jutten, J. R. A. B., editor, *Independent Component Analysis and Signal Separation*. Springer, Berlin, Heidelberg.

[Kruskal, 1977] Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics.

[Laurent and Massart, 2000] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338.

[Le Cam, 1973] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. 1(1):38 – 53.

[Li and Schmidt, 2017] Li, J. and Schmidt, L. (2017). Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1302–1382, Amsterdam, Netherlands. PMLR.

[Li and Chen, 2010] Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105(491).

[Li and Turner, 2016] Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Neural Information Processing Systems*.

[Lin et al., 2016] Lin, L., Drton, M., and Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10(1):806–854.

[Lindsay, 1989] Lindsay, B. G. (1989). Moment matrices: applications in mixtures. *Annals of Statistics*, 17(2):722–740.

[Liu et al., 2012] Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

[Liu et al., 2009] Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328.

[Liu and Shao, 2003] Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, 31(3):807–832.

[Löffler et al.,] Löffler, M., Zhang, A. Y., and Zhou, H. H. Optimality of spectral clustering for gaussian mixture model. arXiv:1911.00538.

[Maugis and Michel, 2011] Maugis, C. and Michel, B. (2011). A non-asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:41–68.

[Moitra and Valiant, 2010] Moitra, A. and Valiant, G. (2010). Settling the polynomial learnability of mixtures of gaussians. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102.

[Niu et al., 2001] Niu, D., Dy, J. G., and Jordan, M. I. (2001). Dimensionality reduction for spectral clustering.

[Pinelis, 2006] Pinelis, I. (2006). On l'hopital-type rules for monotonicity. *Journal of Inequalities of Pure and Applied Mathematics*, 7(2).

[Pollard, 2016] Pollard, D. (2016). Lecture notes on empirical processes. http://www.stat.yale.edu/~pollard/Books/Mini/Chaining.pdf.

[Qi, 2011] Qi, L. (2011). The best rank-one approximation ratio of a tensor space. *SIAM journal on matrix analysis and applications*, 32(2):430–442.

[Rabin et al., 2011] Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer.

[Ranganath et al., 2016] Ranganath, R., Tran, D., Altosaar, J., and Blei, D. (2016). Operator variational inference. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 496–504. Curran Associates, Inc.

[Rudelson and Vershynin, 2009] Rudelson, M. and Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62:1707–1739.

[Saha and Guntuboyina, 2017] Saha, S. and Guntuboyina, A. (2017). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. arXiv:1712.02009.

[Samworth and Yuan, 2012] Samworth, R. J. and Yuan, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *Annals of Statistics*, 40(6):2973–3002.

[Saremi et al., 2018] Saremi, S., Mehrjou, A., Schölkopf, B., and Hyvärinen, A. (2018). Deep energy estimator networks. *ArXiv*, abs/1805.08306.

[Schudy and Sviridenko,] Schudy, W. and Sviridenko, M. Concentration and moment inequalities for polynomials of independent random variables.

[Shalit and Chechik, 2014] Shalit, U. and Chechik, G. (2014). Coordinate-descent for learning orthogonal matrices through givens rotations. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–548–I–556. JMLR.org.

[Shohat and Tamarkin, 1943] Shohat, J. A. and Tamarkin, J. D. (1943). *The problem of moments*. Number 1. American Mathematical Soc.

[Soltanolkotabi et al., 2014] Soltanolkotabi, M., Elhamifar, E., and Candes, E. J. (2014). Robust subspace clustering. *The Annals of Statistics*.

[Sriperumbudur et al., 2017] Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017). Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59.

[Stoer and Bulirsch, 2002] Stoer, J. and Bulirsch, R. (2002). *Introduction to Numerical Analysis*. Springer-Verlag, New York, NY, 3rd edition.

[Tenenbaum et al., 2000] Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319.

[Tsybakov, 2009] Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY.

[van de Geer, 2000] van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

[van der Vaart and Wellner, 1996] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Verlag New York, Inc.

[Vempala and Wang, 2004] Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci*, page 2004.

[Vershynin, 2012] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. and Kutyniok, G., editors, *Compressed Sensing, Theory and Applications*, pages 210–268. Cambridge University Press.

[Vidal, 2009] Vidal, R. (2009). A tutorial on subspace clustering.

[Vincent, 2011] Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674.

[von Luxburg et al., 2008] von Luxburg, U., Belkin, M., and Bousquet, O. (2008). Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586.

[Wang and Blei, 2013] Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14:1005–1031.

[Wong and Shen, 1995] Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve mles. *Annals of Statistics*, 23(2):339–362.

[Wu, 2017] Wu, Y. (2017). Lecture notes on information-theoretic methods for high-dimensional statistics. http://www.stat.yale.edu/~yw562/teaching/598/it-stats.pdf.

[Wu and Verdú, 2010] Wu, Y. and Verdú, S. (2010). The impact of constellation cardinality on gaussian channel capacity. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 620–628.

[Wu and Yang, 2019] Wu, Y. and Yang, P. (2019). Optimal estimation of gaussian mixtures via denoised method of moments. *to appear in* The Annals of Statistics.

[Wu and Zhou, 2019] Wu, Y. and Zhou, H. H. (2019). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *arXiv:1908.10935*.

[Yang and Barron, 1999] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27:1564–1599.

[Zhang and Zhou, 2015] Zhang, A. Y. and Zhou, H. H. (2015). Minimax rates of community detection in stochastic block models. *The Annals of Statistics*.

[Zhang, 2009] Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, 19:1297–1318.