Title
Shape-constrained estimation for modern statistical problems

Author
Soloff, Jake

Publication Date
2022

Peer reviewed|Thesis/dissertation

Shape-constrained estimation for modern statistical problems

by

Jake Soloff

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Adityanand Guntuboyina, Co-chair
Professor Michael I. Jordan, Co-chair
Professor Martin J. Wainwright

Spring 2022

Shape-constrained estimation for modern statistical problems

Abstract

Shape-constrained estimation for modern statistical problems

by

Jake Soloff

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Adityanand Guntuboyina, Co-chair

Professor Michael I. Jordan, Co-chair

Shape constraints encode a relatively weak form of prior information specifying the direction of certain relationships in an unknown signal. Classical examples include estimation of a convex function or a monotone density. Shape constraints are often strong enough to dramatically reduce statistical complexity while still yielding flexible, nonparametric estimators. This thesis brings shape constraints to bear on several recent research areas in statistics—distribution-free inference, high-dimensional covariance estimation, empirical Bayes, and multiple hypothesis testing.

Chapter 2 discusses my joint work with Professor Aditya Guntuboyina and Professor Jim Pitman on distribution-free properties of isotonic regression. In this work, we establish a distributional result for the components of the isotonic least squares estimator using its characterization as the derivative of the greatest convex minorant of a random walk. Provided the walk has exchangeable increments, we prove that the slopes of the greatest convex minorant are distributed as order statistics of the running averages. This result implies an exact formula for the squared error risk of least squares in homoscedastic isotonic regression when the true sequence is constant that holds for every exchangeable error distribution.

Chapter 3 discusses my joint work with Professor Aditya Guntuboyina and Professor Michael I. Jordan on sign-constrained precision matrix estimation. We investigate the problem of high-dimensional covariance estimation under the constraint that the partial correlations are nonnegative. The sign constraints dramatically simplify estimation: the Gaussian maximum likelihood estimator is well defined with only two observations regardless of the number of variables. We analyze its performance in the setting where the dimension may be much larger than the sample size. We establish that the estimator is both high-dimensionally consistent and minimax optimal in the symmetrized Stein loss. We also prove a negative

result which shows that the sign-constraints can introduce substantial bias for estimating the top eigenvalue of the covariance matrix.

Chapter 4 discusses my joint work with Professor Aditya Guntuboyina and Professor Bodhisattva Sen on nonparametric empirical Bayes with multivariate, heteroscedastic Gaussian errors. Multivariate, heteroscedastic errors complicate statistical inference in many large-scale denoising problems. Empirical Bayes is attractive in such settings, but standard parametric approaches rest on assumptions about the form of the prior distribution which can be hard to justify and which introduce unnecessary tuning parameters. We extend the nonparametric maximum likelihood estimator (NPMLE) for Gaussian location mixture densities to allow for multivariate, heteroscedastic errors. NPMLEs estimate an arbitrary prior by solving an infinite-dimensional, convex optimization problem; we show that this convex optimization problem can be tractably approximated by a finite-dimensional version. We introduce a dual mixture density whose modes contain the atoms of every NPMLE, and we leverage the dual both to establish non-uniqueness in multivariate settings as well as to construct explicit bounds on the support of the NPMLE.

The empirical Bayes posterior means based on an NPMLE have low regret, meaning they closely target the oracle posterior means one would compute with the true prior in hand. We prove an oracle inequality implying that the empirical Bayes estimator performs at nearly the optimal level (up to logarithmic factors) for denoising without prior knowledge. We provide finite-sample bounds on the average Hellinger accuracy of an NPMLE for estimating the marginal densities of the observations. We also demonstrate the adaptive and nearly-optimal properties of NPMLEs for deconvolution. We apply the method to two astronomy datasets, constructing a fully data-driven color-magnitude diagram of 1.4 million stars in the Milky Way and investigating the distribution of chemical abundance ratios for 27 thousand stars in the red clump.

Chapter 5 discusses my joint work with Daniel Xiang and Professor William Fithian on finite-sample control of the maximum local false discovery rate in multiple hypothesis testing. Despite the popularity of the false discovery rate (FDR) as an error control metric for large-scale multiple testing, its close Bayesian counterpart the local false discovery rate (lfdr), defined as the posterior probability that a particular null hypothesis is false, is a more directly relevant standard for justifying and interpreting individual rejections. However, the lfdr is difficult to work with in small samples, as the prior distribution is typically unknown. We propose a simple multiple testing procedure and prove that it controls the expectation of the maximum lfdr across all rejections; equivalently, it controls the probability that the rejection with the largest p-value is a false discovery. Our method operates without knowledge of the prior, assuming only that the p-value density is uniform under the null and decreasing under the alternative. We also show that our method asymptotically implements the oracle Bayes procedure for a weighted classification risk, optimally trading off between false positives and false negatives. We derive the limiting distribution of the attained maximum lfdr over the rejections, and the limiting empirical Bayes regret relative to the oracle procedure.

To my family.

Contents

List of Figures

Acknowledgments

Reflecting on my time in Berkeley, I am grateful to more people than I could possibly enumerate. First and foremost, I thank my advisors, Professor Adityanand Guntuboyina and Professor Michael I. Jordan, for their encouragement and mentorship. Their influence, guidance, and patience have shaped and broadened my intellectual horizons. In research, Aditya has the remarkable ability to simultaneously see the forest through the trees as well as even the finest details on each leaf and branch. I have learned to appreciate these not as two distinct features of his thinking but rather as a single, 'multi-resolution' process. As a result, he does not shy away from returning to the 'basics' or foundations of statistics with an openness to change his mind when he inevitably uncovers new subtleties. In countless stimulating discussions, he encouraged me to think through these subtleties for myself and find my own path, and yet he consistently makes time to help me work through technical details and provide feedback when I need it. Mike's encouragement and direction has been a vital source underlying of all of my proudest accomplishments in grad school. His faith in students was my original inspiration to become an academic statistician, and I have benefited from that same faith in his classes and group meetings and research collaborations. From the astounding breadth of his expertise, he finds new and surprising connections to point me in new directions, and his incisive and deep thinking leads him to see months ahead of where I am in my work. Thank you both for pushing me to improve and mature my ideas and my writing, and for teaching me to roll up my sleeves and face a problem head-on.

Next, I send my endless thanks to my fantastic collaborators, Professors Jim Pitman, Bodhisattva Sen, William Fithian, Yuting Wei, and Ashwin Pananjady, as well as Bridget L. Ratcliffe, Daniel Xiang, Stephen Bates, and Michael Sklar. Additionally, I am grateful for illuminating conversations with Professors Martin Wainwright, Bin Yu, Peter Bickel, Jon McAuliffe, Chris Paciorek, and Jacob Steinhardt. Thanks also to all of my mentors prior to grad school, including Professors Stefano Bloch, Louis M. Friedler, Stuart Geman, Mike Hughes, Jackson Loper, Richard Evan Schwartz, and Erik Sudderth.

I am grateful to all of the staff and faculty in our department who have cultivated such a warm and welcoming environment. I owe a special debt to La Shana Porlaris, who looked out for me whenever I needed help.

So much of what is learned during grad school is passed down by example or word-of-mouth from senior grad students and postdocs. I found so many role models during my time here, including Rebecca Barter, Zsolt Bartha, Stephen Bates, Alejandra Benitez, Joe Borja, Ahmed El Alaoui, Billy Fang, Ryan Giordano, Nhat Ho, Steve Howard, Kenneth Hung, Sören Künzel, Lihua Lei, Horia Mania, Kellie Ottoboni, Ashwin Pananjady, Yannik Pitcan, Maxim Rabinovich, Aaditya Ramdas, Sujayam Saha, Sara Stoudt, Simon Walter, Yixin Wang, Yuting Wei, Jason Wu, Fanny Yang, Manolis Zampetakis, and Chelsea Zhang. Also, thanks to the many peer graduate students who made me think, including Taejoo Ahn, Olivia Angiuli, Eli Ben-Michael, Akosua Busia, Patrick Chao, Alice Cima, Mihaela Curmei, Tiffany Ding, Yassine El Maazouz, Emily Flanagan, Sara Fridovich-Keil, Avishek Ghosh, Amanda Glazer, Wenshuo Guo, Ella Hiesmayr, Miyabi Ishihara, Adam Jaffe, Hansheng

Jiang, Koulik Khamaru, Hyunsuk Kim, Karl Krauth, Xiao Li, Tianyi Lin, Bryan Liu, Lydia Liu, Romain Lopez, Benji Lu, Yixiang Luo, Tyler Maltba, Eric Mazumdar, Drew Nguyen, Mehdi Ouaki, Reese Pathak, Frank Qiu, Esther Rolf, Facu Sapienza, Dan Soriano, Asher Spector, Jake Spertus, Tiffany Tang, Ryan Theisen, Nilesh Tripuraneni, Alexander Tsigler, Zoe Vernon, Neha Wadia, Serena Wang, Yu Wang, Eric Xia, Chiao-Yu Yang, Michelle Yu, Tijana Zrnic, among many others. Special thanks to my best friend, Sam Kortchmar, and everyone else I've had the privilege of calling my friend over the years.

Finally, no institution has shaped me more than my big, eccentric family. At my best, who I am reflects their generosity, fortitude, joy, and intense commitment to our shared experience. My accomplishments are theirs.

Chapter 1

Introduction

The study of shape-constrained estimation and inference traces back to the mid-twentieth century, when researchers introduced isotonic regression (Ayer et al., 1955; van Eeden, 1956) and monotone density estimation (Grenander, 1956). Both problems constrain an unknown function—either a regression function or a probability density function—to be monotone. Although the target estimand need not be differentiable, it is helpful to view shape restrictions such as monotonicity or convexity as constraining the *sign* of a (first or second) derivative of an unknown function. By contrast, smoothness assumptions in density estimation and regression commonly constrain the *magnitude* of one or more derivatives of a target function. A resulting attraction of shape constraints is that one can often perform constrained maximum likelihood estimation with no explicit regularization to yield tuning-free, nonparametric estimators.

Conversely, quantifying the full benefits of such qualitative restrictions can be a delicate exercise. In recent decades, a resurgence in theoretical research has greatly expanded our mathematical toolkit for analyzing shape constraints. Some important developments include Barber and Samworth (2021), Cai and Low (2015), Cai et al. (2013), Chatterjee (2014), Dümbgen (2003), Dümbgen et al. (2011), Groeneboom et al. (2001), Guntuboyina and Sen (2013), Han et al. (2019), Kim and Samworth (2016), Meyer and Woodroofe (2000), Slawski and Hein (2013), Wei et al. (2019), and Zhang (2002); see also Guntuboyina and Sen (2018) for a recent survey and Groeneboom and Jongbloed (2014) for a general introduction.

As this now rich statistical literature routinely demonstrates, shape-constrained estimation has the potential to alleviate the challenges of nonparametric estimation while preserving the flexibility of light assumptions. Realizing this potential in a wide variety of applications, however, demands innovations on other research fronts. On one important front, designing scalable algorithms for large samples, high-dimensional data, and complex constraints presents many challenges. Koenker and Mizera (2014) approximate nonparametric maximum likelihood estimators with finite-dimensional, convex optimization problems, opening up nonparametric density estimation to a wide range of existing algorithms and off-the-shelf solvers. In multivariate convex regression, Mazumder et al. (2019) illustrate the flexibility of the alternating direction method of multipliers (ADMM), which itself is closely related to

the algorithm of Dykstra (1983) for constrained least squares.

Beyond computation and risk bounds, getting shape constraints to work with other areas of statistics requires a richer appreciation of the varied roles constraints play in framing and simplifying problems. In this dissertation, we investigate the role of shape constraints in four modern research areas, and for the rest of the introduction, we highlight the varied roles of shape constraints in these and other problems.

Problem specification. Likelihood-based estimation often fails in high-dimensional spaces, and shape constraints offer one approach to rescue nonparametric estimation from the limitations of the maximum likelihood criterion over large parameter spaces. A prototypical example comes from density estimation. Let $p_1, \ldots, p_m$ be an iid sample of continuous random variables on $[0, 1]$ with density f^* and corresponding cdf F^*. It is well known that the maximum likelihood estimator (MLE) of F^* over the space of all cdfs is the empirical cdf F_m, defined as $F_m(t) := \frac{1}{m}\sum_{i=1}^{m} 1\{p_i \leq t\}$. The MLE of the pdf f^*, on the other hand, does not exist: essentially, since F_m is discrete, the likelihood of a density f increases without bound as we move outside of the space of all densities on $[0, 1]$ and towards a discrete distribution.

Imposing the additional constraint that f^* is monotone decreasing enables maximum likelihood estimation. To see why, first note that a density f is nonincreasing precisely when the corresponding cdf F is concave. The empirical cdf F_m is not concave, so the constrained MLE $\hat{F}_m$ of the cdf F^* over all *concave cdfs* is a different object than F_m. In fact, the likelihood under the concave MLE $\hat{F}_m$ is lower than the likelihood under the unconstrained MLE F_m, since the latter maximizes the likelihood over all cdfs. Roughly, in order to make the gap due to introducing the concavity constraint as small as possible, $\hat{F}_m$ should be as close to F_m as possible. More precisely, it can be shown (Groeneboom & Jongbloed, 2014, Lemma 2.2) that $\hat{F}_m$ is the least concave majorant of F_m, i.e. the (pointwise) smallest concave function dominating F_m. The least concave majorant $\hat{F}_m$ is a linear spline, so it is 'nearly' differentiable, i.e. differentiable everywhere except at the knots. Since $\hat{F}_m$ is concave the slope of the line to the left of a given knot is larger than the slope of the line to the right, and indeed the Grenander (1956) estimator $\hat{f}_m$ of a monotone density is given by the left derivative of the least concave majorant $\hat{F}_m$. The empirical cdf F_m plays the role of a 'default' estimator of F^* across statistics, e.g. in the nonparametric bootstrap, so its scope and limitations as a plug-in estimator are well understood. The Grenander estimator $\hat{f}_m$ plays a similar role as a 'default' density estimator of a monotone density.

The Grenander estimator $\hat{f}_m$ is somewhat special among shape-constrained density estimators, in that it has this exact geometric characterization, especially amenable to precise asymptotic theory. Moreover, there are many equivalent ways of thinking about the problem of monotone density estimation that connect it to other types of shape constraints. For instance, Khintchine's theorem states that the collection of monotone densities on $[0, 1]$ coincides with the collection of densities of uniform scale mixtures, i.e. distributions of the form $\text{Unif}(0, \Theta)$ where $\Theta \sim G^*$ is an arbitrary probability measure on $[0, 1]$. Hence, the Grenander estimator $\hat{f}_m$ implicitly describes an estimator $\widehat{G}_m$ of the mixing measure G^*,

known as the nonparametric maximum likelihood estimator or NPMLE of G^*. Since $\hat{F}_m$ is a (concave) linear spline, the corresponding density $\hat{f}_m$ is a piecewise constant (decreasing) density with discontinuities at the knots of $\hat{F}_m$; equivalently, the NPMLE $\widehat{G}_m$ is a discrete distribution supported on the knots of $\hat{F}_m$. These statements are equivalent, but the latter is remarkable for its own reason: $\widehat{G}_m$ maximizes the likelihood of the uniform scale mixture over *all* probability measures supported on $[0, 1]$. Hence, whereas the Grenander estimator $\hat{f}_m$ is a constrained MLE over the class of monotone densities, $\widehat{G}_m$ is an *unconstrained* NPMLE where the likelihood is that of a uniform scale mixture. The key takeaway is that the mixture structure can facilitate fully nonparametric modeling of unobserved heterogeneity, a point which surprisingly predates even the Grenander estimator (Robbins, 1950). The idea to use the NPMLE of a mixing distribution to model a prior distribution is foundational to nonparametric empirical Bayes, as we shall see in Chapter 4.

More recent results have begun to uncover multivariate settings with flexible 'default' estimators. Suppose we aim to estimate the density f^* of i.i.d. observations $X_1, \dots, X_n \in \mathbb{R}^p$ under the constraint that $\log f^*$ is concave: Cule et al. (2010) showed that the MLE is well-defined almost surely once $n > p$. Another remarkable result, due to Slawski and Hein (2015), establishes a setting where shape constraints lead to a well-defined constrained MLE even in very high-dimensional settings $p \gg n$. Instead of estimating the density of $X_1, \dots, X_n \in \mathbb{R}^p$, the goal is to estimate the precision matrix $\Theta^* := (\Sigma^*)^{-1}$ under a multivariate Gaussian likelihood $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma^*)$. When $p > n$, the MLE of Θ^* is not defined, since the sample covariance matrix $S = \frac{1}{n}\sum_{i=1}^n X_i X_i^\mathsf{T}$ is rank-deficient. However, under the additional sign-constraints $\Theta^*_{jk} \leq 0$ for $j \neq k$, Slawski and Hein (2015) showed that the constrained MLE $\widehat{\Theta}$ is almost surely well-defined for any p once $n > 1$. The sign-constraints on the off-diagonal entries of Θ^* provide a model of positive dependence that has been widely used in economics and actuarial sciences (Karlin & Rinott, 1983). The fact that the constrained MLE $\widehat{\Theta}$ is well-defined is nontrivial, but Lauritzen et al. (2019) give a beautiful proof, exploiting a connection to single-linkage clustering. We investigate the statistical properties of this estimator $\widehat{\Theta}$ in high-dimensional settings in Chapter 3.

It is worth noting that reasonable shape constraints do not always impose sufficient structure to produce well-defined solutions. A classical example is the estimation of a unimodal density, even with univariate data, where the likelihood is unbounded as in the unconstrained density estimation example. In Chapter 4 we show that certain NPMLEs of multivariate mixing distributions can have infinitely many solutions.

Finally, we note that shape restrictions have been leveraged to specify a problem structure beyond producing well-defined constrained MLEs. In causal inference, for example, Angrist and Imbens (1994) leveraged a monotonicity assumption on compliance to identify the local average treatment effect under two-sided noncompliance.

Structured decisions. Shape-enforcement is often viewed as a post-processing step to improve upon a base estimator (Bonakdarpour et al., 2018). Obozinski et al. (2008) followed this approach to yield logically consistent predictions in protein function annotation. In this domain, various data sources are combined to produce probabilistic predictions of protein

function. The attributes that an individual protein may satisfy fit into the gene ontology (GO), a dictionary of terms together with a topology describing their logical relations. For instance, 'nucleic acid binding' and 'protein binding' are two attributes that have 'binding' as a parent attribute. Probabilistic predictions are made on a term-by-term basis, e.g. the probability that a given protein is nucleic acid binding is estimated independently of its other attributes, so this base set of predictions need not respect the GO structure—e.g. 'binding' could be reported as less likely than one of its child attributes. These probabilistic predictions can be 'reconciled' by projecting them onto the set of probabilistic predictions respecting the GO partial order. When the initial predictions are reconciled using Euclidean projection, this is known as isotonic regression with respect to a partial order. Obozinski et al. (2008) find empirically that the isotonized predictions almost always have higher precision than the base predictions, whereas more heuristic reconciliation methods and Bayesian networks often perform worse than the base predictions.

In empirical Bayes applications, downstream decisions often satisfy desirable properties as a byproduct of the mixture constraint. Efron and Hastie (2016, Chapter 6) describe the example of an insurance company needing to estimate the number of claims a policy holder will make next year based on the number of claims they made this year. The dataset is simply a collection of nonnegative counts $Y_1, \dots, Y_n$, and from the law of rare events it is reasonable to model each policy holder's number of claims Y_i as a Poisson random variable with rate parameter λ_i, independent across i. The rate parameters λ_i likely vary across policy holders $i = 1, \dots, n$, and we can model this unobserved heterogeneity by placing a common prior on the rate parameters

$$\lambda_i \overset{\text{iid}}{\sim} G^*, \text{ for } i = 1, \dots, n,$$

where G^* is a probability measure supported on the nonnegative reals $\mathbb{R}_+$. Marginally, the counts Y_i are iid with Poisson mixture pmf p_{G^*}, where

$$p_{G^*}(y) \coloneqq \int e^{-\lambda} \lambda^y / y! \, \mathrm{d}G^*(\lambda).$$

Robbins (1956) showed that the Bayes estimator $\delta_{G^*}(y) \coloneqq \mathbb{E}[\lambda_i \mid Y_i = y]$ can be written directly in terms of the marginal pmf

$$\delta_{G^*}(y) = \frac{(y+1) p_{G^*}(y+1)}{p_{G^*}(y)}.$$

It is not hard to show that, no matter the choice of prior G^*, the posterior mean δ_{G^*} is non-decreasing in the number of counts. This is a desirable property for the insurance company: any non-monotone decision rule penalizes some policy holders for making fewer claims than others. In practice, however, G^* is unknown so the Bayes estimator is unavailable. Robbins (1956) proposed to plug-in the empirical probability mass function (pmf) $\hat{p}_n(y) = \frac{\#\{i : Y_i = y\}}{n}$ and take

$$\hat{\delta}(y) = \frac{(y+1)\hat{p}_n(y+1)}{\hat{p}_n(y)}$$

for any y such that $\hat{p}_n(y) > 0$, and define $\hat{\delta}(y) = 0$ otherwise. This estimator remains the 'default' choice for empirical Bayes denoising of count data, despite severely violating monotonicity: the largest count $y_{\max} = \max_{i=1:n} Y_i$ is assigned a predicted value of $\hat{\delta}(y_{\max}) = 0$. Brown et al. (2013) patched this deficiency of Robbins' estimator by isotonizing $\hat{\delta}$, among other post-processing proposals. By contrast, if we estimate p_{G^*} by any Poisson mixture $p_{\widehat{G}_n}$, for instance by minimizing $D(\hat{p}_n, p_G)$ over all probability measures G on $\mathbb{R}_+$, for some divergence D, the resulting plug-in procedure is automatically monotone. For instance, if D is the KL divergence, then $\widehat{G}_n$ is the Poisson mixture NPMLE. In this example, utilizing the shape constraint in an end-to-end fashion spares us from post-processing adjustments to achieve the desired structure of a decision rule.

Sharper inferences. In other cases, the desired structure of optimal decision rules delivers a tractable shape constraint. In Chapter 5, we study multiple hypothesis testing within a Bayes two-groups model

$$p_i \mid H_i = h \overset{\text{ind}}{\sim} f_h, \qquad \text{with} \qquad H_i \overset{\text{iid}}{\sim} \text{Bern}(1 - \pi_0), \qquad \text{for } i = 1, \dots, m,$$

where $H_i = 0$ if the ith hypothesis is null and $H_i = 1$ otherwise. The p-values p_i follow a density $f_0 \coloneqq 1_{[0,1]}$ under the null and f_1 under the alternative, and the null proportion is $\pi_0 \in [0, 1]$. Let $f \coloneqq \pi_0 + (1 - \pi_0) f_1$ denote the common mixture density of the p-values, and let $F(t) \coloneqq \int_0^t f(u)\, du$ denote the corresponding cumulative distribution function (cdf). Sun and Cai (2007) showed that optimally trading off false positives and false negatives is achieved by rejecting hypothesis with local false discovery rate (lfdr, Efron et al., 2001)

$$\text{lfdr}(t) \coloneqq \mathbb{P}\left(H_i = 0 \mid p_i = t\right) = \frac{\pi_0}{f(t)}$$

falling below some level $q \in [0, 1]$, depending our relative tolerance for Type I and Type II errors. In other words, the optimal procedures reject the null hypothesis for $p_i \in A_q^*$, where the rejection regions A_q^* have the form

$$A_q^* \coloneqq \{t : \text{lfdr}(t) \leq q\}.$$

In most cases, we seek to reject all the p-values falling below some threshold, meaning the rejection region is an interval of the form $[0, t]$. If we posit that the optimal rejection regions also have this form $A_q^* = [0, \tau_q^*]$, note that τ_q^* is nondecreasing as a function of q, simply because the rejection regions A_q^* are automatically nested as the tolerance q increases. From this, it follows that lfdr is nondecreasing, or equivalently that the mixture density f is nonincreasing. The assertion that it is optimal to reject p-values below some threshold is thus equivalent to the assertion that f is monotone.

In other words, f being monotone nonincreasing means that smaller p-values represent stronger evidence against the null, and if the optimal decision rule rejects small p values then f must be monotone. Fundamental methods in multiple hypothesis testing, such as the Benjamini and Hochberg (1995, BH) procedure, are thresholding procedures, so such

procedures are best suited to settings where f is monotone. The BH procedure builds a rejection region of the form

$$A_q^{\mathrm{BH}} := \left\{ t : \widehat{\mathrm{Fdr}}(t) \leq q \right\} \qquad \text{where} \qquad \widehat{\mathrm{Fdr}}(t) = \frac{1}{F_m(t)}.$$

This procedure is well-motivated from an empirical Bayes perspective (see, e.g., Efron et al., 2001) but it does not directly target the optimal rejection regions A_q^*. Moreover, enforcing the shape restriction that the true cdf F is concave yields little benefit, since the concave MLE $\hat{F}_m$ is extremely close to F_m at least for large m (Kiefer & Wolfowitz, 1976).

The real advantage of the monotonicity assumption is that we may leverage the Grenander estimator $\hat{f}_m$ as our 'default' estimator of a monotone density to directly target the optimal rejection region:

$$A_q := \left\{ t : \widehat{\mathrm{lfdr}}(t) \leq q \right\} \qquad \text{where} \qquad \widehat{\mathrm{lfdr}}(t) = \frac{1}{\hat{f}_m(t)}.$$

The shape constraint thus affords us the opportunity to approach a much more ambitious inferential target in a fully nonparametric manner. We explore the properties of this testing procedure in much greater detail in Chapter 5.

We close this section with a cautionary result showing that, in some cases, shape constraints can have much worse statistical properties than the base estimators they are designed to improve upon. In particular, we show in the high-dimensional covariance estimation problem in Chapter 3 that the constrained MLE of the covariance matrix $\widehat{\Sigma} = \widehat{\Theta}^{-1}$ can be *much worse* than the sample covariance matrix S for estimating the top eigenvalue. Note that the sample covariance S is itself inconsistent in high-dimensional regimes, but the top-eigenvalue of $\widehat{\Sigma}$ diverges at an even faster rate. However, we argue that because $\widehat{\Theta}$ essentially represents a projection of S under a Bregman divergence known as Stein's loss, it is more natural to study the high-dimensional behavior of $\widehat{\Theta}$ under that loss function.

Chapter 2

Distribution-free isotonic regression

2.1 Introduction

Isotonic regression with homoscedastic errors refers to the problem of estimating a monotone sequence $\theta_1^* \leq \cdots \leq \theta_n^*$ based on a noisy observation vector Y, assumed to be an additive perturbation of $\theta^* = (\theta_1^*, \ldots, \theta_n^*)$

$$Y = \theta^* + \sigma Z,$$

where the components $Z_1, \ldots, Z_n$ of Z are assumed to have zero mean and unit variance. It is commonly assumed that $Z_1, \ldots, Z_n$ are independent and identically distributed (i.i.d.) but we work with the more general assumption of exchangeability in this chapter. A natural estimator for θ^* in this setting is the isotonic Least Squares Estimator (LSE), defined as

$$\hat{\theta} := \Pi_{\mathcal{M}^n}(Y) := \operatorname*{argmin}_{\theta \in \mathcal{M}^n} \|Y - \theta\|_2^2,$$

where $\|\cdot\|_2$ denotes the usual Euclidean norm on $\mathbb{R}^n$ and $\mathcal{M}^n := \{\theta \in \mathbb{R}^n : \theta_1 \leq \cdots \leq \theta_n\}$ is the monotone cone of length n nondecreasing sequences. As $\mathcal{M}^n$ is a closed convex cone, $\hat{\theta}$ as defined above exists uniquely; it can also be computed in $O(n)$ time by the pool adjacent violators algorithm (Brunk et al., 1972; Grotzinger & Witzgall, 1984).

One approach to evaluating the statistical properties of $\hat{\theta}$ is to measure the risk, or expected deviation of $\hat{\theta}$ from θ^*. Indeed, the risk provides a convenient summary of the accuracy of $\hat{\theta}$ and many papers on isotonic regression have focused on obtaining bounds for the risk of $\hat{\theta}$ (see e.g., Bellec, 2018; Guntuboyina and Sen, 2018; Zhang, 2002). In this chapter, we primarily consider the normalized mean squared error:

$$R(\hat{\theta}, \theta^*) := \frac{1}{n}\mathbb{E}_{\theta^*}\|\hat{\theta} - \theta^*\|_2^2.$$

A key quantity in understanding $R(\hat{\theta}, \theta^*)$ is

$$\delta_n(\mu) := \mathbb{E}_{Z \sim \mu}\|\Pi_{\mathcal{M}^n}(Z)\|_2^2,$$

where μ denotes the law of the noise vector Z. Indeed, it is clear that

$$\frac{n}{\sigma^2} R(\hat{\theta}, \theta^*) = \delta_n(\mu) \qquad \text{when } \theta_1^* = \cdots = \theta_n^*.$$

When $\theta_1^* \leq \cdots \leq \theta_n^*$ are not all equal, let $(A_1, \ldots, A_k)$ be the coarsest partition of $\{1, \ldots, n\}$ such that θ^* is constant on each A_i. It has been shown (Bellec, 2018; Fang & Guntuboyina, 2017; Oymak & Hassibi, 2016) that

$$\frac{n}{\sigma^2} R(\hat{\theta}, \theta^*) \begin{cases} \leq \delta_{n_1}(\mu_{A_1}) + \cdots + \delta_{n_k}(\mu_{A_k}) & \text{for every } \sigma > 0 \\ \to \delta_{n_1}(\mu_{A_1}) + \cdots + \delta_{n_k}(\mu_{A_k}) & \text{as } \sigma \downarrow 0 \end{cases}, \tag{2.1}$$

where μ_{A_i} denotes the marginal distribution of $(Z_j)_{j \in A_i}$ and $n_i = |A_i|$ is the length of the i^{th} block for all $i = 1, \ldots, k$. We emphasize that (2.1) holds for arbitrarily dependent $Z_1, \ldots, Z_n$ with zero mean and finite variance. It was also shown by Bellec (2018) that $\delta_n(\mu)$ also bounds the risk of the isotonic LSE in misspecified settings where θ^* does not lie in $\mathcal{M}^n$.

The quantity $\delta_n(\mu)$ therefore crucially controls the risk of the isotonic LSE. The goal of this chapter is to explicitly determine $\delta_n(\mu)$ for every $n \geq 1$ under the additional assumption that Z is exchangeable. Specifically, under the assumption of exchangeability, we show in Corollary 2.6 that, for all n,

$$\delta_n(\mu) = \rho n + (1 - \rho) H_n, \tag{2.2}$$

where $H_n := 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ is the n^{th} harmonic number, $\rho = \text{Cor}(Z_1, Z_2)$ is the pairwise correlation, and $\sigma^2 = 1$. Combined with (2.1), our result provides a sharp, non-asymptotic bound on the risk of isotonic regression for *any* exchangeable noise vector. In the special case when $Z_1, \ldots, Z_n$ are i.i.d. with zero mean and unit variance, $\rho = 0$ and thus (2.2) gives:

$$\delta_n(\otimes_{i=1}^n \eta) = H_n \qquad \text{for every probability measure } \eta. \tag{2.3}$$

Here η is the common distribution of the independent variables $Z_1, \ldots, Z_n$.

Previously, the formula (2.3) was known when η is the standard Gaussian probability measure on $\mathbb{R}^n$. This was observed by Amelunxen et al. (2014) who proved it by observing first that when $\mu = \otimes_{i=1}^n \eta$ and η is the standard Gaussian measure, the formula

$$\mathbb{E} \|\Pi_K(Z)\|_2^2 = \sum_{k=0}^n k \, \nu_k(K) \tag{2.4}$$

holds for every closed convex cone $K \subseteq \mathbb{R}^n$ where $\nu_k(K)$ is the k^{th} intrinsic volume of K. When $K = \mathcal{M}^n$ is the monotone cone, the right hand side in equation (2.4) can be shown to be equal to H_n by using the fact that the generating function $s \mapsto \sum_{k=0}^n s^k \nu_k(\mathcal{M}^n)$ can be computed in closed form. Amelunxen et al. (2014) used the theory of finite reflection groups (Coxeter & Moser, 2013) to obtain the exact expression for this generating function. However, the exact expression for $\sum_{k=0}^n s^k \nu_k(\mathcal{M}^n)$ can already be found in the classical

literature on isotonic regression (see Theorem 2.4.2 in Robertson et al. (1988) or Section 8 of Sparre-Andersen (1954)).

The above proof does not work for non-Gaussian η mainly because the expression (2.4) does not hold for general η. In fact, the best available result on $\delta_n(\otimes_{i=1}^n \eta)$ for non-Gaussian η is in equation (2.11) of Zhang (2002), who proved the asymptotic result:

$$\delta_n(\otimes_{i=1}^n \eta) = (1 + o(1))(1 + \log n) \qquad \text{as } n \to \infty.$$

This bound gives the right behavior as the right hand side of equation (2.3) but only as $n \to \infty$. We improve this result by proving for every $n \geq 1$ that $\delta_n(\otimes_{i=1}^n \eta)$ is always equal to the n^{th} harmonic number H_n for every probability measure η having mean 0 and variance 1.

We prove (2.2) by developing a precise characterization of the marginal distribution of each individual component $(\Pi_{\mathcal{M}^n}(Z))_k$ of $\Pi_{\mathcal{M}^n}(Z)$. Specifically, as long as Z is exchangeable, we show in Theorem 2.2 that $(\Pi_{\mathcal{M}^n}(Z))_k$ has the same distribution as $\bar{Z}_{(k)}$, the k^{th} order statistic of the running averages $\bar{Z}_j = \frac{Z_1 + \cdots + Z_j}{j}$. We prove Theorem 2.2 in Section 2.2, using a characterization of the components of the isotonic LSE as the left-hand slopes of the greatest convex minorant of the random walk with increments $Z_1, \ldots, Z_n$. This result and its continuous-time analogue may be of independent interest outside the study of isotonic regression, so in Section 2.2 we also address consequences for the greatest convex minorant of a stochastic process with exchangeable increments. The order statistics of the running averages $\{\bar{Z}_k\}_{k=1}^n$ can be fairly complicated even when Z is Gaussian; however, Theorem 2.2 easily implies results such as (2.2). In Section 2.3, we detail some risk calculations for isotonic regression and its variants which all follow from Theorem 2.2.

2.2 Main result

Let $S_k = \sum_{i=1}^k Z_i$ denote the partial sums for $k = 1, \ldots, n$, started at $S_0 = 0$. Identify the random walk $\{S_k\}_{k=0}^n$ with its *cumulative sum diagram* $S : [0, n] \to \mathbb{R}$, where $S(k) = S_k$ for integers $k = 0, \ldots, n$ and linearly interpolated between integers. Let $C : [0, n] \to \mathbb{R}$ denote the *greatest convex minorant* (GCM) of S, i.e. the greatest convex function that lies below S. See Figure 2.1 for a depiction of the GCM of S. With this notation, we now recall the graphical representation of the isotonic LSE as given in Theorem 1.2.1 of Robertson et al. (1988).

Lemma 2.1. *For any vector Z, the isotonic LSE $\Pi_{\mathcal{M}^n}(Z)$ is given by the left-hand slopes of the greatest convex minorant of the cumulative sum diagram. For all $k = 1, \ldots, n$*

$$(\Pi_{\mathcal{M}^n}(Z))_k = C(k) - C(k-1) = \partial_- C(k).$$

For the remainder of this section let

$$\Delta_k := \partial_- C(k) = \min_{k \leq v \leq n} \max_{0 \leq u < k} \frac{S_v - S_u}{v - u} \tag{2.5}$$

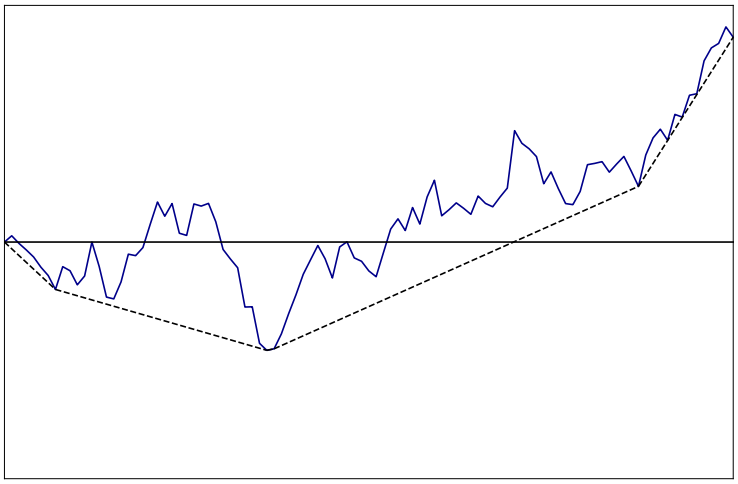


Figure 2.1: Solid blue curve is the cumulative sum diagram S of increments $Z_1, \ldots, Z_n$; dashed black curve is the greatest convex minorant C of S.

denote the left-hand slope of the GCM at k, so $\Delta = (\Delta_1, \ldots, \Delta_n)$ is equal to $\Pi_{\mathcal{M}^n}(Z)$ by the lemma. In particular, when $k = 1$ we have $\Delta_1 = \min_{1 \le v \le n} \frac{S_v}{v}$. When $k = n$, we have $\Delta_n = \max_{0 \le u < n} \frac{S_n - S_u}{n-u}$, and if $(Z_n, \ldots, Z_1) \stackrel{d}{=} (Z_1, \ldots, Z_n)$ then $\Delta_n \stackrel{d}{=} \max_{1 \le u \le n} \frac{S_u}{u}$. Our next result generalizes this observation, showing that the k^{th} slope Δ_k is equal in distribution to the k^{th} smallest running average if Z is exchangeable.

Theorem 2.2. *Suppose $Z = (Z_1, \ldots, Z_n)$ is exchangeable. Let $\bar{Z}_k := \frac{1}{k} \sum_{i=1}^k Z_i$ denote the k^{th} running average for $k = 1, \ldots, n$ and let $\bar{Z}_{(1)} \le \cdots \le \bar{Z}_{(n)}$ denote their order statistics. Then*

$$\Delta_k \stackrel{d}{=} \bar{Z}_{(k)} \tag{2.6}$$

marginally for all $k = 1, \ldots, n$.

Proof. As before, let S_k denote the k^{th} partial sum. Let M be the last argmin of the sequence $\{S_i\}_{i=0}^n$, and let N be the amount of time the walk is nonpositive $N := \sum_{i=1}^n 1(S_i \le 0)$. We will use Corollary 11.14 of Kallenberg (2006), due to Sparre-Andersen, which says $M \stackrel{d}{=} N$ as long as Z is exchangeable.

Note that the slope of the GCM switches from nonpositive to positive at time M, since the horizontal line with intercept S_M minorizes the GCM and touches it at time M. Hence, no matter the sequence of increments Z_i, there is the identity of events

$$(\Delta_k \le 0) = (M \ge k). \tag{2.7}$$

Also, for the time N that the walk is nonpositive, since $S_i \leq 0$ if and only if $\bar{Z}_i \leq 0$, there is the identity of events

$$(\bar{Z}_{(k)} \leq 0) = (N \geq k).$$

The equality in distribution $M \stackrel{d}{=} N$ then implies

$$\mathbb{P}(\Delta_k \leq 0) = \mathbb{P}(\bar{Z}_{(k)} \leq 0).$$

If the sequence $\{Z_i\}$ is modified to $\{Z_i - z\}$ for some fixed z, the modified sequence is exchangeable, and the values of Δ_k and $\bar{Z}_{(k)}$ for the modified sequence are just $\Delta_k - z$ and $\bar{Z}_{(k)} - z$. Applying the above identity to the modified sequence gives

$$\mathbb{P}(\Delta_k \leq z) = \mathbb{P}(\Delta_k - z \leq 0) = \mathbb{P}(\bar{Z}_{(k)} - z \leq 0) = \mathbb{P}(\bar{Z}_{(k)} \leq z).$$

So Δ_k and $\bar{Z}_{(k)}$ have the same cumulative distribution function, hence the same distribution. □

The proof of Theorem 2.2 generalizes to the setting where $S : [0,1] \to \mathbb{R}$ is a continuous-time stochastic process. Knight (1996) showed that the analogous distributional identity $M \stackrel{d}{=} N$ holds when S has exchangeable increments and $S(0) = 0$. Hence, by a similar proof, we find that the slope $\Delta(p)$ of the greatest convex minorant of S at time $p \in [0,1]$ has the same distribution as the p^{th} percentile point of the occupation measure for the process $(\frac{S(t)}{t}, 0 \leq t \leq 1)$. We record this result as the following corollary.

Corollary 2.3. *Let S denote a real-valued càdlàg stochastic process on $[0,1]$ with exchangeable increments, such that $S(0) = 0$. Define $\Delta(t)$ as the slope of the greatest convex minorant of S at t, and let $F : \mathbb{R} \to [0,1]$ denote the (random) cdf associated with the occupation measure of $(\frac{S(t)}{t}, 0 \leq t \leq 1)$,*

$$F(x) = \lambda(\{t \in [0,1] : S(t) \leq tx\}), \tag{2.8}$$

where λ denotes Lebesgue measure. Then

$$\Delta(p) = \inf_{p \leq v \leq 1} \sup_{0 \leq u < p} \frac{S(v) - S(u)}{v - u} \stackrel{d}{=} F^{-1}(p) \tag{2.9}$$

marginally for all $p \in [0,1]$.

See Abramson et al. (2011) for a general study of convex minorants of random walks and processes with exchangeable increments. In the special cases where S is a standard Brownian motion or Brownian bridge on the unit interval, Carolan and Dykstra (2001) derive the distribution of the slope $\Delta(p)$, jointly with the process $S(p)$ and its convex minorant at p, for a fixed value $p \in [0,1]$. Given our corollary, their explicit formula for the slope $\Delta(p)$ provides the distribution of $F^{-1}(p)$, giving new information about the occupation measure of $(\frac{S(t)}{t}, 0 \leq t \leq 1)$ for Brownian motion and Brownian bridge. The distribution of the p^{th} percentile point of the occupation measure for $(S(t), 0 \leq t \leq 1)$ has been obtained under the same generality as Corollary 2.3: see the introduction of Dassios (2005) and references therein.

2.3 Consequences for isotonic regression

Since the identity of Theorem 2.2 holds marginally, it allows us to simplify expectations of functions that are additive in the components of $\Pi_{\mathcal{M}^n}(Z)$. By Lemma 2.1, the k^{th} component $(\Pi_{\mathcal{M}^n}(Z))_k = \Delta_k$, which by Theorem 2.2 is equal in distribution to $\bar{Z}_{(k)}$. Hence, as long as Z is exchangeable,

$$\sum_{k=1}^{n} \mathbb{E}h((\Pi_{\mathcal{M}^n}(Z))_k) = \sum_{k=1}^{n} \mathbb{E}h(\bar{Z}_{(k)}) = \sum_{k=1}^{n} \mathbb{E}h(\bar{Z}_k). \tag{2.10}$$

Taking $h(x) = |x|^p$, we obtain our first corollary.

Corollary 2.4. *Suppose $Z = (Z_1, \ldots, Z_n)$ is exchangeable. For $p > 0$,*

$$\mathbb{E}\|\Pi_{\mathcal{M}^n}(Z)\|_p^p = \sum_{k=1}^{n} \mathbb{E}\left|\frac{1}{k}\sum_{i=1}^{k} Z_i\right|^p, \tag{2.11}$$

provided $\mathbb{E}|Z_1|^p < \infty$.

Remark 2.5. *Viewed through its graphical representation, $\Delta_k = C(k) - C(k-1)$ is the left-derivative of the GCM C at k, so when the power $p = 1$, equation (2.11) yields the discrete arc-length formula*

$$\sum_{k=1}^{n} \mathbb{E}|C(k) - C(k-1)| = \mathbb{E}\|\Pi_{\mathcal{M}^n}(Z)\|_1 = \sum_{k=1}^{n} \frac{1}{k}\mathbb{E}|S_k| \tag{2.12}$$

Closely related to this formula is the identity of Spitzer and Widom (1961), which takes $\tilde{Z}_1, \ldots, \tilde{Z}_n$ to be a sequence of i.i.d. random variables in $\mathbb{R}^2$ (or the complex plane $\mathbb{C}$) with finite variance. If $\tilde{S}_k = \sum_{i=1}^{k} \tilde{Z}_i$ is the partial sum and $\tilde{L}_n$ is the length of the perimeter of the convex hull conv$(0, \tilde{S}_1, \ldots, \tilde{S}_n)$, then

$$\mathbb{E}\tilde{L}_n = 2\sum_{k=1}^{n} \frac{1}{k}\mathbb{E}\|\tilde{S}_k\|. \tag{2.13}$$

These formulas connect the geometry of the convex hull of a random walk to the magnitudes of the running means.

Consider the case when $p = 2$. Since Z is exchangeable, every pair of components has the same correlation ρ. If we further assume Z_1 has zero mean and unit variance, the right hand side of equation (2.11) can be computed explicitly

$$\mathbb{E}\left(\frac{1}{k}\sum_{i=1}^{k} Z_i\right)^2 = \rho + \frac{1-\rho}{k}.$$

Summing over k yields our next result.

Corollary 2.6. *Suppose $Z \sim \mu$ is an exchangeable random vector with zero mean, unit variance, and pairwise correlation ρ. Then*

$$\delta_n(\mu) = \rho n + (1-\rho)H_n.$$

This result should be contrasted with other distribution-free identities, namely

$$\mathbb{E}\|Z\|_2^2 = n \text{ and } \mathbb{E}\|\bar{Z}_n 1_n\|_2^2 = 1,$$

provided Z has i.i.d. components with zero mean and unit variance. In particular, suppose we observe $Y = \theta^* + \sigma Z$ where Z has i.i.d. components with zero mean and unit variance, but it turns out that $\theta^* = c1_n$ is constant. If we know θ^* is constant, we can estimate it by a constant sequence $\bar{Y}1_n$ and pay a price of $\frac{\sigma^2}{n}$ in risk (normalized mean squared error). If we know nothing about the structure of θ^* and use $\hat{\theta} = Y$, the risk σ^2 is quite large by comparison. The monotone sequence estimate resides in the middle, with a much smaller risk of $\frac{H_n\sigma^2}{n}$ and knowledge only about the relative order.

Theorem 2.2 characterizes the distribution of a component of the isotonic LSE $\hat{\theta}$ when the underlying sequence θ^* is constant. When $\theta^* \in \mathcal{M}^n$ is not constant, Theorem 2.2 can be applied to characterize the distribution of a component $\hat{\theta}_i$ in the low noise limit $\sigma \downarrow 0$. In this limit, the distribution depends only on flat regions of θ^*:

Corollary 2.7. *Suppose $Y = \theta^* + \sigma Z$, for some $\theta^* \in \mathcal{M}^n$, and let $\hat{\theta} = \Pi_{\mathcal{M}^n}(Y)$ denote the isotonic LSE. Let $(A_1, \ldots, A_k)$ be the coarsest partition of $\{1, \ldots, n\}$ such that θ^* is constant on each A_j, and suppose Z is exchangeable on each of these blocks. If an index $i \in \{1, \ldots, n\}$ belongs to the j^{th} block, then letting $t_i = i + 1 - \min_{s \in A_j} s$ and $X = (Z_s)_{s \in A_j}$, we have*

$$\frac{\hat{\theta}_i - \theta_i^*}{\sigma} \xrightarrow{d} \bar{X}_{(t_i)} \text{ as } \sigma \downarrow 0. \tag{2.14}$$

Proof. As $\sigma \downarrow 0$, the ratio $\frac{\hat{\theta}-\theta^*}{\sigma}$ tends to the directional derivative $D_Z\Pi_{\mathcal{M}^n}(\theta^*)$. Lemma 4.6 in Zarantonello (1971) shows that this derivative exists and equals the projection of Z onto the tangent cone $T_{\mathcal{M}^n}(\theta^*)$. Hence

$$\frac{\hat{\theta} - \theta^*}{\sigma} \to \Pi_{T_{\mathcal{M}^n}(\theta^*)}(Z) \text{ as } \sigma \downarrow 0. \tag{2.15}$$

From the tangent cone computation in Bellec (2018), we have

$$\left(\Pi_{T_{\mathcal{M}^n}(\theta^*)}(Z)\right)_i = \left(\Pi_{\mathcal{M}^{n_i}}(X)\right)_{t_i},$$

where $n_i = |A_{j_i}|$. Finally, by Theorem 2.2, $\left(\Pi_{\mathcal{M}^{n_i}}(X)\right)_{t_i} \overset{d}{=} \bar{X}_{(t_i)}$. □

We explained in Section 2.1 how risk calculations when $\theta^* = 0$ generalize to MSE bounds that are sharp in the low noise limit for arbitrary θ^*. For example, when $\theta^* \in \mathcal{M}^n$ has k

constant pieces, then (2.1), Corollary 2.6 and the fact that $H_l \leq \log(el)$ for every $l \geq 1$ imply that

$$R(\hat{\theta}, \theta^*) \leq \frac{k\sigma^2}{n} \log\left(\frac{en}{k}\right) \tag{2.16}$$

whenever $Z_1, \ldots, Z_n$ are i.i.d. with mean zero and unit variance. The bound (2.16) should be compared with the risk of the structure-respecting estimator that averages over the constant blocks and achieves a risk of exactly $\frac{k\sigma^2}{n}$ when the blocks are all of size $\frac{n}{k}$. If $\theta^* \in \mathbb{R}^n$ is not necessarily in $\mathcal{M}^n$, then Corollary 2.6, together with the results of Bellec (2018), implies that

$$R(\hat{\theta}, \theta^*) \leq \inf_{\theta \in \mathcal{M}^n} \left(\frac{1}{n}\|\theta - \theta^*\|^2 + \sigma^2 \frac{k(\theta)}{n} \log\left(\frac{en}{k(\theta)}\right)\right),$$

where $k(\theta)$ is the number of constant pieces of the vector θ. These formulae (with the leading constant of 1 in front of the $\frac{k\sigma^2}{n} \log \frac{en}{k}$ term on the right hand side) were previously only known when the distribution of $Z_1, \ldots, Z_n$ was standard Gaussian.

Define the L^p-risk of the isotonic LSE

$$R^{(p)}(\hat{\theta}, \theta^*) = \frac{1}{n}\mathbb{E}\|\hat{\theta} - \theta^*\|_p^p$$

so that $R(\hat{\theta}, \theta^*) = R^{(2)}(\hat{\theta}, \theta^*)$. We can similarly employ Theorem 2.2 to explicitly calculate the L^p-risk of the isotonic LSE $\hat{\theta}$ when θ^* is constant and Z is Gaussian:

Corollary 2.8. *Suppose $Z \sim \mathcal{N}(0, I_n)$. Then for any $p > 0$,*

$$\mathbb{E}\|\Pi_{\mathcal{M}^n}(Z)\|_p^p = H_{n,p/2}\mathbb{E}|Z_1|^p = H_{n,p/2}\sqrt{\frac{2^p}{\pi}}\Gamma\left(\frac{p+1}{2}\right),$$

where $H_{n,m} = \sum_{k=1}^n \frac{1}{k^m}$.

Proof. Note $\mathbb{E}\left|\frac{1}{k}\sum_{i=1}^k Z_i\right|^p = \left(\frac{2}{k}\right)^{p/2} \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$ and apply the theorem. □

Corollary 2.8 should similarly be contrasted with the following identities when $Z \sim \mathcal{N}(0, I_n)$:

$$\mathbb{E}\|Z\|_p^p = n\mathbb{E}|Z_1|^p \text{ and } \mathbb{E}\|\bar{Z}1_n\|_p^p = n^{1-p/2}\mathbb{E}|Z_1|^p$$

respectively. In particular, when $p > 2$, the bound $H_{n,p/2} < \sum_{k=1}^\infty \frac{1}{k^{p/2}} < \infty$ holds for all n, which is to say $\mathbb{E}\|\Pi_{\mathcal{M}^n}(Z)\|_p^p$ is bounded when $p > 2$ whereas $\mathbb{E}\|Z\|_p^p$ grows without bound as n grows.

When θ^* is constant and $Z \sim \mathcal{N}(0, I_n)$, the L^p risk of isotonic regression is

$$R^{(p)}(\hat{\theta}, \theta^*) = \frac{H_{n,p/2}}{n}\sigma^p\mathbb{E}|Z_1|^p. \tag{2.17}$$

When $1 \leq p \leq 2$, Theorem 2.3 of Zhang (2002) shows an asymptotic result for the L^p risk on constant θ^* that agrees with equation (2.17).

The continuous-time distributional identity in Corollary 2.3 applies to the asymptotic distribution of the isotonic least squares estimator. A standard model for studying the asymptotic behavior of isotonic regression is

$$\theta_k^* = f^*\left(\frac{k}{n}\right)$$

where $f^* : [0,1] \to \mathbb{R}$ is nondecreasing. We observe Y, a noisy version of θ^*, and calculate $\hat{\theta}$ by projecting Y onto the monotone cone. The function estimate $\hat{f}$ is defined by $\hat{f}\left(\frac{k}{n}\right) = \hat{\theta}_k$ and linearly interpolated between design points. Here, as before, the dependence on n in $\theta^* \in \mathcal{M}^n$ is suppressed, but now we are interested in the behavior of isotonic least squares $\hat{f}(p)$ at a fixed point $p \in [0,1]$ as $n \to \infty$.

Define the partial sum process $S^{(n)} : [0,1] \to \mathbb{R}$ by $S^{(n)}(k/n) = \frac{Y_1 + \cdots + Y_k}{\sqrt{n}}$, linearly interpolated between design points. When the function $f^* \equiv c$ is constant, the quantity

$$\sqrt{n}(\hat{f}(p) - f^*(p))$$

is given by the left-derivative of the greatest convex minorant of $S^{(n)}$ at p. By the invariance principle, this converges in distribution to the left-derivative of the greatest convex minorant of standard Brownian motion $B = (B(t), 0 \le t \le 1)$ at t_0. This asymptotic result is well known and a similar result was noted for the Grenander estimator by Carolan and Dykstra (1999), where Brownian motion is replaced with a Brownian bridge. Corollary 2.3 relates this asymptotic distribution to the percentile points of the occupation measure for $(\frac{B(t)}{t}, 0 \le t \le 1)$.

Finally, Corollary 2.6 on the projection onto $\mathcal{M}^n$ extends over to that of the set of nonnegative monotone sequences $\mathcal{M}_+^n = \mathcal{M}^n \cap \mathbb{R}_+^n$. Theorem 1 of Németh and Németh (2012) observes that the projection of Z onto $\mathcal{M}_+^n$ is given by $\Pi_{\mathcal{M}_+^n}(Z) = \Pi_{\mathcal{M}^n}(Z)_+$, the element-wise positive part of the projection onto $\mathcal{M}^n$. Hence the distributional identity Theorem 2.2 yields a similar set of identities for nonnegative isotonic regression.

Corollary 2.9. *For any exchangeable noise vector Z,*

$$(\Pi_{\mathcal{M}_+^n}(Z))_k \stackrel{d}{=} (\bar{Z}_{(k)})_+ \tag{2.18}$$

Provided $\mathbb{E}|Z_i|^p < \infty$,

$$\mathbb{E}\|\Pi_{\mathcal{M}_+^n}(Z)\|_p^p = \sum_{k=1}^n \mathbb{E}\left(\frac{1}{k}\sum_{i=1}^k Z_i\right)_+^p, \tag{2.19}$$

Furthermore, if Z is symmetric with unit variance, the generalized statistical dimension of the monotone cone is

$$\mathbb{E}\|\Pi_{\mathcal{M}_+^n}(Z)\|_2^2 = \frac{\rho n + (1-\rho)H_n}{2}, \tag{2.20}$$

where ρ is the pairwise correlation.

Proof. Equation (2.19) follows from equation (2.10) by taking $h(x) = (x)_+^p$. When $Z_i \overset{d}{=} -Z_i$ is symmetric with unit variance,

$$\mathbb{E}\left(\frac{1}{k}\sum_{i=1}^{k} Z_i\right)_+^2 = \frac{1}{2}\mathbb{E}\left(\frac{1}{k}\sum_{i=1}^{k} Z_i\right)^2 = \frac{1}{2}\left(\rho + \frac{1-\rho}{k}\right).$$

Summing over k yields equation (2.20). □

Equation (2.20) is also shown by Amelunxen et al. (2014) in the special case $Z \sim \mathcal{N}(0, I_n)$ using the theory of finite reflection groups. The identity (2.19) allows us to show equation (2.20) for a much wider variety of noise vectors, and as before also allows us to obtain relations for the expected L^p norms of the projection of the noise vector. All of our exact formulae follow from the distributional identity in Theorem 2.2, which exploits the geometric characterization of the isotonic LSE in Lemma 2.1. An interesting open question is whether similar characterizations— such as for convex regression (Groeneboom et al., 2001)—may yield exact non-asymptotic risk calculations in other shape-constrained estimation problems.

Chapter 3

Sign-constrained precision matrix estimation

3.1 Introduction

Consider the problem of estimating a $p \times p$ covariance matrix Σ^* and its inverse $\Theta^* := (\Sigma^*)^{-1}$ from an $n \times p$ data matrix X whose rows are independently distributed according to the multivariate normal distribution $\mathcal{N}(0, \Sigma^*)$ with mean zero and covariance matrix Σ^*. The maximum likelihood estimator (MLE) of Θ^* is given by

$$\tilde{\Theta} := \underset{\Theta \in \mathcal{S}^{p\times p}_{\succeq 0}}{\operatorname{argmin}} \left\{ \langle \Theta, S \rangle - \log\det \Theta \right\}, \tag{3.1}$$

where $\mathcal{S}^{p\times p}_{\succeq 0}$ denotes the set of all $p \times p$ symmetric, positive semi-definite (PSD) matrices, $\langle \Theta, S \rangle := \operatorname{tr}(\Theta^{\mathsf{T}} S)$ denotes the Frobenius inner product, and S is the sample covariance matrix, defined as

$$S := n^{-1} X^{\mathsf{T}} X. \tag{3.2}$$

It is well known that $\tilde{\Theta}$ exists if and only if S is nonsingular, in which case $\tilde{\Theta} = S^{-1}$. In particular, in the high-dimensional setting where $p > n$, the MLE does not exist, since the minimum in (3.1) is not finite. Slawski and Hein (2015) observed, however, that if the optimizer in (3.1) is constrained to lie in the set of $p \times p$ positive semidefinite matrices with *nonpositive off-diagonal entries*, then, with probability one, the optimum is well-defined and attained for all $n \geq 2$ regardless of the value of p. Specifically, let

$$\mathcal{M}^{p\times p} := \left\{ \Theta \in \mathcal{S}^{p\times p}_{\succeq 0} : \Theta_{jk} \leq 0 \text{ for } j \neq k \right\},$$

and observe that it is the convex cone of symmetric *M-matrices*, an important class of matrices appearing in many contexts (see, e.g., Berman & Plemmons, 1994, Chap. 6). Slawski and Hein (2015) proved that the optimizer

$$\widehat{\Theta} := \underset{\Theta \in \mathcal{M}^{p\times p}}{\operatorname{argmin}} \left\{ \langle \Theta, S \rangle - \log\det \Theta \right\}, \tag{3.3}$$

exists uniquely as long as, in the observed sample, no two variables are perfectly positively correlated (i.e., $S_{jk} < \sqrt{S_{jj}S_{kk}}$ for all $j \neq k$) and no variable is constant (i.e., $S_{jj} > 0$ for all j). Both conditions hold with probability one under the assumed Gaussian model for $n \geq 2$, and thus, unlike the unconstrained MLE in (3.1), the estimator (3.3) is well-defined even in the high-dimensional regime.

The constrained MLE $\widehat{\Theta}$ presents an elegant, tuning-free method for estimating precision matrices which works for $n \geq 2$ and all values of p under the assumption $\Theta^* \in \mathcal{M}^{p\times p}$. Efficient algorithms for computing $\widehat{\Theta}$ are given in Slawski and Hein (2015) and Lauritzen et al. (2019). Note that the precision matrix having nonpositive off-diagonal entries Θ^*_{jk} is equivalent to nonnegative partial correlations $-\Theta^*_{jk}/\sqrt{\Theta^*_{jj}\Theta^*_{kk}}$ (Bølviken, 1982). Examples of practical covariance estimation problems with nonnegative partial correlations abound (see, e.g., Agrawal et al., 2019; Lake & Tenenbaum, 2010; Slawski & Hein, 2015). More generally, Karlin and Rinott (1983) showed that for the normal distribution the condition that the precision matrix belongs to $\mathcal{M}^{p\times p}$ is equivalent to multivariate total positivity of order two (MTP_2). MTP_2 is a strong form of positive dependence (Colangelo et al., 2005) that has been widely used in auction theory (Milgrom & Weber, 1982), actuarial sciences (Denuit et al., 2006), and educational evaluation and policy analysis (Chade et al., 2014).

There is growing interest in $\widehat{\Theta}$ in the graph signal processing literature (Egilmez et al., 2017; Pavez et al., 2018; Pavez & Ortega, 2016), where M-matrices are known as *Generalized Graph Laplacians* (GGL). Indeed, every graph Laplacian is a diagonally dominant M-matrix, and conversely every M-matrix $\Theta \in \mathcal{M}^{p\times p}$ can be viewed as a generalized graph Laplacian, in the sense that it has a sparse *edge-incidence factorization* $\Theta = VV^\mathsf{T}$, where $V \in \mathbb{R}^{p\times p(p+1)/2}$ has at most two nonzero entries per column, whereas positive semidefinite matrices that have other sign patterns typically require dense factorizations (Boman et al., 2005). This connection to nonnegative weighted graphs has led to a host of other application areas in image processing and network analysis.

This chapter investigates the statistical properties of $\widehat{\Theta}$ as an estimator of the unknown precision matrix Θ^* in the high-dimensional regime. Even though $\widehat{\Theta}$ exists uniquely for all $n \geq 2$ regardless of the value of p, rigorous results have not yet been proved for the accuracy of $\widehat{\Theta}$ in the high-dimensional regime. In the classical low dimensional asymptotic regime where p is fixed and $n \to \infty$, Slawski and Hein (2015) apply standard results for M-estimators to show consistency of $\widehat{\Theta}$. More recently, Lauritzen et al. (2019) provide an elegant perspective on $\widehat{\Theta}$ and a bound on the support graph $G(\widehat{\Theta}) = \{(j,k) : \widehat{\Theta}_{jk} < 0\}$, and Wang et al. (2019) develop a consistent estimator of $G(\Theta^*)$.

The study of consistency and optimality properties of $\widehat{\Theta}$ requires fixing an appropriate loss function. Because $\widehat{\Theta}$ is defined via maximum likelihood, it is natural to work with the *Stein loss*:

$$L^{\text{s}}(\Theta, \Theta^*) := \frac{1}{p}\langle \Theta, \Sigma^* \rangle - \frac{1}{p}\log\det\Theta\Sigma^* - 1, \tag{3.4}$$

which, up to scaling by p, is the Kullback-Leibler divergence between multivariate mean zero normal distributions with precision matrices Θ and Θ^* respectively. The Stein loss

has a long history of application in covariance matrix estimation (Dey & Srinivasan, 1985; Donoho et al., 2018; James & Stein, 1961; Ledoit & Wolf, 2018; Stein, 1975, 1986). In this chapter, we work with the *symmetrized Stein loss* (alternatively known as the *divergence loss*), defined as

$$L^{\text{ssym}}(\Theta, \Theta^*) := \frac{L^{\text{s}}(\Theta, \Theta^*) + L^{\text{s}}(\Theta^*, \Theta)}{2} = \frac{1}{2p} \langle \Theta - \Theta^*, \Sigma^* - \Sigma \rangle , \tag{3.5}$$

where $\Sigma = \Theta^{-1}$. Note that $L^{\text{ssym}}(\Theta, \Theta^*)$ is symmetric and $2L^{\text{ssym}}(\Theta, \Theta^*)$ clearly dominates both the Stein loss and the reversed Stein loss $L^{\text{s}}(\Theta^*, \Theta)$ (which is also known as the *entropy loss*). Properties of L^{ssym} are further discussed in Section 3.2.

We use the $1/p$ scaling in the loss function (3.5) because, as explained by Ledoit and Wolf (2018), this is necessary for consistency in the high-dimensional regime where the number of variables p may be much larger than the sample size n. Indeed, in the simple case where Θ^* is known to be diagonal, the natural estimator is the diagonal matrix $\hat{\Theta}^{\text{DIAG}}$ with diagonal entries $1/S_{jj}, j = 1, \ldots, p$ (where S is the sample covariance matrix defined in (3.2)). It is easy to see that $\left\langle \hat{\Theta}^{\text{DIAG}} - \Theta^*, \Sigma^* - \hat{\Sigma}^{\text{DIAG}} \right\rangle$ is of the order p/n which will be far from zero in the high-dimensional regime where $p > n$.

We present results on the performance of $\widehat{\Theta}$ in the symmetrized Stein loss in Section 3.2. Our main result in Theorem 3.1 implies that $L^{\text{ssym}}(\widehat{\Theta}, \Theta^*)$ converges to zero as long as $\log p = o(n)$. This implies high-dimensional consistency of $\widehat{\Theta}$. Moreover, the rate of convergence is $\sqrt{\frac{\log p}{n}}$, which we prove in Theorem 3.2 is optimal in the minimax sense. Thus $\widehat{\Theta}$ is minimax optimal in the high-dimensional regime under the symmetrized Stein loss. Our results provide rigorous support for the assertion that the nonpositive off-diagonal constraint provides strong implicit regularization in the high-dimensional regime. In Theorem 3.4, we also lower bound the loss $L^{\text{ssym}}(\widehat{\Theta}, \Theta^*)$ which implies that the $\sqrt{n}$ rate is not an artifact of our analysis even when the true precision matrix Θ^* is diagonal.

High-dimensional consistency with the rate $\sqrt{\frac{\log p}{n}}$ has appeared previously in many papers on covariance and precision matrix estimation—see for instance Cai et al. (2011), Ravikumar et al. (2011), Rothman et al. (2008), Sun and Zhang (2013), and Yuan (2010) and Cai et al. (2016b) for a review of rates in structured covariance estimation. Most of these results are for estimators that use explicit regularizers, such as the ℓ_1 penalty in the Graphical Lasso (Banerjee et al., 2008; Friedman et al., 2008; Mazumder & Hastie, 2012), which is crucially exploited by the proof techniques and assumptions employed in these papers. By contrast, the regularization induced by the assumption $\Theta^* \in \mathcal{M}^{p\times p}$ is implicit and we consequently use different arguments relying on careful use of the KKT conditions underlying the optimization (3.3). Our analysis identifies a bound relating the entries of an M-matrix to its spectrum, providing new insight into the simplifying structure of the convex cone $\mathcal{M}^{p\times p}$.

The symmetrized Stein loss has the additional symmetry property of invariance under inversion: $L^{\text{ssym}}(\widehat{\Sigma}, \Sigma^*) = L^{\text{ssym}}(\widehat{\Theta}, \Theta^*)$ where $\widehat{\Sigma} := \widehat{\Theta}^{-1}$. This means that $\hat{\Sigma}$ is also a high-

dimensionally-consistent estimator of Σ^*. The choice of the loss function is quite crucial here. In Section 3.3, using the Perron-Frobenius theorem and a careful analysis of the entry-wise positive part S_+ of the sample covariance, we prove a negative result which shows that, for the maximum eigenvalue, $\widehat{\Sigma}$ can be much worse as an estimator of Σ^* compared to the sample covariance matrix S. This result indicates that enforcing the sign-constraints can exacerbate bias in the estimation of the top eigenvalue.

The chapter is organized as follows: Section 3.2 contains our main results establishing optimality of $\widehat{\Theta}$, Section 3.3 establishes suboptimality under the spectral norm, and Section 3.4 has a discussion which touches upon some related issues including misspecification (where $\Theta^* \notin \mathcal{M}^{p\times p}$), estimation of correlation matrices and connections to shape-restricted regression. Finally Section 3.5 contains proofs of all the results of the chapter.

3.2 Symmetrized Stein loss: consistency and optimality

This section contains our results on the high-dimensional consistency and optimality of $\widehat{\Theta}$ under the symmetrized Stein loss L^{ssym} defined in (3.5). We start by describing some basic properties of L^{ssym}.

The expected value of the objective in (3.3), $\langle \Theta, \Sigma^* \rangle - \log\det\Theta$, agrees up to factors depending only on Σ^* with the *Stein loss* (3.4), which is also a matrix Bregman divergence (Dhillon & Tropp, 2008), proportional to the Kullback-Leibler (KL) divergence between centered multivariate Gaussian distributions: $\frac{2}{p}D(\mathcal{N}(0,\Sigma)\|\mathcal{N}(0,\Sigma^*))$. It is well known that the KL divergence is not symmetric. When the inputs to the divergence are reversed, the resulting Bregman divergence is also known as the *entropy loss*, $L^{\text{ent}}(\Theta,\Theta^*) := L^{\text{s}}(\Theta^*,\Theta)$. The sum of these loss functions dominates each, and conveniently does not directly involve any determinants. Following Ledoit and Wolf (2018), we define $L^{\text{ssym}} = \frac{L^{\text{s}}+L^{\text{ent}}}{2}$ to be the average of the two loss functions. Commonly known as the *symmetrized Stein loss* or *divergence loss*, L^{ssym} is equal to the Jeffreys (1946) divergence between two centered multivariate Gaussian distributions, divided by p. Definition (3.5) entails a number of useful and important properties for the symmetrized Stein loss:

(i) (Nonnegativity) $L^{\text{ssym}}(\Theta,\Theta^*) \geq 0$, with equality if and only if $\Theta = \Theta^*$.

(ii) (Symmetry) $L^{\text{ssym}}(\Theta,\Theta^*) = L^{\text{ssym}}(\Theta^*,\Theta)$.

(iii) (Invariance under inversion) $L^{\text{ssym}}(\Theta,\Theta^*) = L^{\text{ssym}}(\Sigma,\Sigma^*)$.

(iv) (Invariance under congruent transformations) For all $p\times p$ nonsingular matrices P, we have the scale-invariance property:

$$L^{\text{ssym}}(\Theta,\Theta^*) = L^{\text{ssym}}(P^{\mathsf{T}}\Theta P, P^{\mathsf{T}}\Theta^* P) \tag{3.6}$$

The symmetrized Stein loss thus induces a natural geometry on the space of PSD matrices—see Moakher and Batchelor (2006) for a review and comparison to other geometries. We emphasize that triangle inequality fails to hold for both L^{ssym} and $\sqrt{L^{\text{ssym}}}$. As a loss, L^{ssym} treats the dual problems of estimating the covariance matrix and the precision matrix equally. It can also be shown that the symmetrized Stein loss is equivalent to the squared Frobenius norm when the input matrices Θ and Θ^* have bounded spectra.

In terms of the eigenvalues $(\lambda_j)_{j=1}^p$ of $\Theta\Sigma^*$, the symmetrized Stein loss is simply the goodness-of-fit measure

$$L^{\text{ssym}}(\Theta, \Theta^*) = \frac{1}{p}\sum_{j=1}^{p} \frac{(\lambda_j - 1)^2}{2\lambda_j}. \tag{3.7}$$

This alternative representation provides further insight into the normalization of the loss (3.5) with a factor of p. The symmetrized Stein loss is the expectation of the function $\lambda \mapsto \frac{(\lambda-1)^2}{2\lambda}$ with respect to the empirical spectral distribution of $\Theta\Sigma^*$. This expectation measures how far the spectrum of $\Theta\Sigma^*$ deviates from a point mass at one, which is the spectrum of the identity I_p. In asymptotic settings where $p = p(n) \to \infty$ as $n \to \infty$, a natural consistency criterion checks whether this expectation converges to zero.

Our analysis of the symmetrized Stein loss $L^{\text{ssym}}(\widehat{\Theta}, \Theta^*)$ involves the maximum population correlation between any two variables:

$$\max_{j \neq k} \frac{\Sigma^*_{jk}}{\sqrt{\Sigma^*_{jj}\Sigma^*_{kk}}}.$$

We assume that the above quantity is strictly less than 1 which is clearly necessary for Σ^* to be nonsingular i.e., for Θ^* to exist. Our bound on $L^{\text{ssym}}(\widehat{\Theta}, \Theta^*)$ will involve the quantity:

$$\gamma(\Sigma^*) := \left(1 - \max_{j \neq k} \frac{\Sigma^*_{jk}}{\sqrt{\Sigma^*_{jj}\Sigma^*_{kk}}}\right)^{-1}.$$

It is natural for $\gamma(\Sigma^*)$ to enter the analysis in light of the existence result of Slawski and Hein (2015) which states that the maximum sample correlation must be less than one in order for the estimator $\widehat{\Theta}$ to be well-defined. Note that $\gamma(\Sigma^*)$ is the smallest $\gamma \geq 1$ such that

$$\max_{j \neq k} \frac{\Sigma^*_{jk}}{\sqrt{\Sigma^*_{jj}\Sigma^*_{kk}}} \leq 1 - \gamma^{-1} < 1. \tag{3.8}$$

Because $\gamma(\Sigma^*)$ is defined in terms of population correlations, it is scale-invariant. Note that L^{ssym} also has this scale invariance property (see (3.6)).

Theorem 3.1. *Let $S = n^{-1}X^\mathsf{T}X$ denote the sample covariance matrix based on data matrix $X \in \mathbb{R}^{n\times p}$ with i.i.d. $\mathcal{N}(0, \Sigma^*)$ rows, where $\Theta^* = (\Sigma^*)^{-1} \in \mathcal{M}^{p\times p}$. For all $n \geq c_1\gamma^2(\Sigma^*)\log p$,*

the MLE $\widehat{\Theta}$ *defined in* (3.3) *satisfies*

$$L^{ssym}(\widehat{\Theta}, \Theta^*) \leq c_2 \gamma(\Sigma^*) \sqrt{\frac{\log p}{n}}, \tag{3.9}$$

with probability at least $1 - c_3 p^{-2}$. *Here* c_1, c_2, c_3 *are universal positive constants.*

Theorem 3.1 states that $\widehat{\Theta}$ is high-dimensionally consistent in the symmetrized Stein loss L^{ssym} as long as $\log p = o(n)$. We prove Theorem 3.1 in Section 3.5, deriving a basic inequality from the first order optimality conditions for (3.3) and showing that concentration of the intrinsic noise $\|S - \Sigma^*\|_\infty$ is sufficient to control the basic inequality. Crucially, we use the fact that every M-matrix $\Theta \in \mathcal{M}^{p \times p}$ is up to diagonal scaling equivalent to a diagonally dominant matrix (see Berman & Plemmons, 1994, Chap. 6, Property M_{34}).

We emphasize that the result holds without additional assumptions on the underlying precision matrix such as sparsity. Consistency in the symmetrized Stein loss is a strong guarantee compared to the recent literature on optimal shrinkage of the sample covariance S under high-dimensional asymptotics (Donoho et al., 2018; Ledoit & Wolf, 2018), where the symmetrized Stein loss L^{ssym} converges to a nonzero limit under the asymptotic regime $p/n \to \alpha > 0$ as $n \to \infty$. By contrast, for the constrained MLE the loss $L^{\text{ssym}}(\widehat{\Theta}, \Theta^*)$ converges in probability to zero whenever $\log p = o(n)$.

Since the upper bound (3.9) depends only on the true precision matrix Θ^* through the population quantity $\gamma(\Sigma^*)$, Theorem 3.1 actually bounds the worst case risk obtained from the divergence loss over all M-matrices Θ^* with $\gamma(\Sigma^*)$ bounded. It is natural to question whether the $\sqrt{n}$ rate is improvable. Our next result shows that, in the high-dimensional setting where p grows superlinearly in n, the minimax rate over the class of M-matrices with $\gamma(\Sigma^*) \leq \gamma$ matches the $\sqrt{\frac{\log p}{n}}$ rate from Theorem 3.1.

Theorem 3.2. *Let* $X \in \mathbb{R}^{n \times p}$ *have i.i.d.* $\mathcal{N}(0, \Sigma^*)$ *rows, and suppose the number of variables* p *satisfies* $c_1 n^\beta \leq p \leq \exp(c_2 n)$. *For every* $\gamma > 1$, *we have*

$$\inf_{\breve{\Theta} = \breve{\Theta}(X)} \sup_{\substack{\Theta^* \in \mathcal{M}^{p \times p} \\ \gamma(\Sigma^*) \leq \gamma}} \mathbb{E} L^{ssym}(\breve{\Theta}, \Theta^*) \geq c_\gamma \sqrt{\frac{\log p}{n}}. \tag{3.10}$$

Here $c_1, c_2 > 0$ *and* $\beta > 1$ *are universal constants and* $c_\gamma > 0$ *is a constant depending only on* γ.

Paired with Theorem 3.1, this result implies that $\widehat{\Theta}$ is minimax optimal in the symmetrized Stein loss over M-matrices with correlations bounded away from one. Our proof adapts the construction of Cai et al. (2016a), Theorem 4.1, which lower bounds the minimax risk in the spectral norm over a parameter set of sparse precision matrices of the form $I + \varepsilon A$, where ε depends on problem parameters p and n, and A is an adjacency matrix. A key aspect of this approach is to allow for different perturbations over the rows and columns of A, in order to recover the $\sqrt{n}$ rate (Kim, 2020).

The M-matrix constraint provides implicit regularization and is crucial for achieving the minimax rate $\sqrt{\frac{\log p}{n}}$. If this constraint is dropped, it is impossible for any estimator to achieve a rate better than $\sqrt{\frac{p}{n}}$ when $p > n$. This follows from the next result where we prove a minimax lower bound of $\sqrt{\frac{p}{n}}$ for the L^{ssym} loss function over the entire class $\mathcal{S}_{\succeq 0}^{p\times p}$ of positive semidefinite matrices when $p > n$. On the other hand, $\mathcal{M}^{p\times p}$ is much larger than diagonal matrices because the minimax rate of estimation over the class $\mathcal{D}_{+}^{p\times p}$ of positive diagonal matrices in the L^{ssym} loss function is $1/n$ (this is also proved in the next result). In summary, the class of M-matrices acts as a strong high-dimensional regularizer while being considerably larger than the class of all positive diagonal matrices.

Proposition 3.3. *Fix p and $n > 2$. The minimax risk in the symmetrized Stein loss over diagonal precision matrices satisfies*

$$\inf_{\hat{\Theta}=\hat{\Theta}(X)} \sup_{\Theta^* \in \mathcal{D}_{+}^{p\times p}} \mathbb{E} L^{ssym}(\hat{\Theta}, \Theta^*) \asymp \frac{1}{n}. \tag{3.11}$$

The minimax risk in the symmetrized Stein loss over PSD matrices satisfies

$$\inf_{\hat{\Theta}=\hat{\Theta}(S)} \sup_{\Theta^* \in \mathcal{S}_{\succeq 0}^{p\times p}} \mathbb{E} L^{ssym}(\hat{\Theta}, \Theta^*) \gtrsim \min\left\{\frac{p}{n}, \sqrt{\frac{p}{n}}\right\}. \tag{3.12}$$

Theorem 3.2 implies that the $\sqrt{n}$ rate of Theorem 3.1 cannot be improved in worst case over the entire class $\mathcal{M}^{p\times p}$. In the next result, we prove that the $\sqrt{n}$ rate for $\widehat{\Theta}$ cannot be improved even when the truth Θ^* lies in the class $\mathcal{D}_{+}^{p\times p}$ of positive diagonal matrices. In other words, this shows that $\widehat{\Theta}$ does not adapt to the minimax rate over $\mathcal{D}_{+}^{p\times p}$.

Theorem 3.4. *Suppose $\Theta^* \in \mathcal{D}_{+}^{p\times p}$ is a positive diagonal matrix and $c_1 p \geq \sqrt{n}$. Then*

$$L^{ssym}(\widehat{\Theta}, \Theta^*) \geq \frac{c_1}{2\sqrt{n}}, \tag{3.13}$$

with probability at least $1 - 3p\exp\left(-c_2(n \wedge p)\right)$, where c_1 and c_2 are universal positive constants.

3.3 Spectral norm: suboptimality

In this section, we prove a negative result which implies that $\widehat{\Theta}$ and $\widehat{\Sigma}$ can be suboptimal for estimating spectral quantities of Θ^* and Σ^* respectively. Consider the case when $\Sigma^* = I_p$ and consider estimation of the top eigenvalue $\lambda_{\max}(\Sigma^*) = 1$. The performance of the sample covariance matrix S is well understood. Indeed, in the asymptotic setting $p/n \to \alpha > 0$, Geman (1980) proved that

$$\lambda_{\max}(S) \to (1 + \sqrt{\alpha})^2,$$

in probability as $n \to \infty$. This implies that S is inconsistent for the estimation of $\lambda_{\max}(\Sigma^*)$ when p/n converges to a positive constant. Our next result proves that $\widehat{\Sigma} = \widehat{\Theta}^{-1}$ is also inconsistent for estimating $\lambda_{\max}(\Sigma^*)$ and, more interestingly, its performance is *substantially worse* compared to S. Specifically, in the same asymptotic setting where $p/n \to \alpha > 0$, we have

$$\lambda_{\max}(\widehat{\Sigma}) \to \infty \tag{3.14}$$

in probability as $n \to \infty$. Thus the introduction of the sign constraints make the resulting covariance matrix estimator $\widehat{\Sigma}$ much worse compared to S for estimating the principal eigenvalue. This should be contrasted with the high-dimensional minimax optimality results from the previous section in the symmetrized Stein loss.

Theorem 3.5. *Suppose $\Sigma^* = I_p$ and $p \geq 17$. Then*

$$\lambda_{\max}(\widehat{\Sigma}) \geq 1 + c_1 \frac{p}{\sqrt{n}}, \tag{3.15}$$

with probability at least $1 - 3p \exp\left(-c_2(n \wedge p)\right)$, for some universal positive constants c_1, c_2.

Note that when $p/n \to \alpha > 0$, the right hand side of (3.15) diverges to ∞ which proves (3.14).

The proof of Theorem 3.5 is crucially based on following dual formulation to the constrained MLE (3.3) (see, e.g., Slawski & Hein, 2015):

$$\widehat{\Sigma} = \underset{\substack{\Sigma \in \mathcal{S}^{p\times p}_{\succeq 0} \\ \Sigma \geq S,\, D_\Sigma = D_S}}{\operatorname{argmax}} \det \Sigma, \tag{3.16}$$

where the second constraint $\Sigma \geq S$ is an entry-wise inequality. This fact and the well-known observation that the inverse of an M-matrix is entry-wise nonnegative (see Berman & Plemmons, 1994, Chap. 6, Property N_{38}) together imply that $\widehat{\Sigma}_{jk} \geq S_{jk} \vee 0$ for all j, k. This allows us to prove Theorem 3.5 by a careful analysis of the entry-wise positive part matrix S_+ of S.

Theorem 3.5 implies minimax suboptimality of $\widehat{\Sigma}$ in the spectral norm $|||\cdot|||_2$. To see this, note that, for every $K > 0$, the sample covariance S satisfies the worst case risk bound

$$\sup_{\substack{\Sigma^* \in \mathcal{S}^{p\times p}_{\succeq 0} \\ \lambda_{\max}(\Sigma^*) \leq K}} \mathbb{E}|||S - \Sigma^*|||_2 \leq CK\left(\sqrt{\frac{p}{n}} + \frac{p}{n}\right),$$

where $C > 0$ is a universal constant (see, e.g., Wainwright, 2019, Example 6.3). By contrast, Theorem 3.5 implies

$$\sup_{\substack{\Theta^* \in \mathcal{M}^{p\times p} \\ \lambda_{\max}(\Sigma^*) \leq K}} \mathbb{E}|||\widehat{\Sigma} - \Sigma^*|||_2 \geq \mathbb{E}_{\Sigma^* = KI_p}\left[\lambda_{\max}(\widehat{\Sigma}) - K\right] \geq cK\frac{p}{\sqrt{n}}$$

for $n \gtrsim \log p$. Hence $\widehat{\Sigma}$ is minimax suboptimal in the spectral norm for most choices of p and n.

Theorem 3.5 also implies inconsistency in spectral norm for the precision matrix. Since $\lambda_{\max}(\widehat{\Sigma}) = \frac{1}{\lambda_{\min}(\widehat{\Theta})}$, we have

$$\lambda_{\min}(\widehat{\Theta}) \leq \frac{1}{1 + c_1\alpha\sqrt{n}},$$

with probability at least $1 - 3p\exp(-c_2(\alpha \wedge 1)n)$, where $\alpha = p/n$. As $n \to \infty$, the upper bound approaches zero: the minimum eigenvalue of $\widehat{\Theta}$ poorly estimates that of Θ^*. We record this as a separate corollary.

Corollary 3.6. *Suppose $\Sigma^* = I_p$ and $p = \alpha n \geq 17$. Then*

$$|||\widehat{\Theta} - \Theta^*|||_2 \geq 1 - \lambda_{\min}(\widehat{\Theta}) \geq \frac{1}{1 + 1/(c_1\alpha\sqrt{n})}, \tag{3.17}$$

with probability at least $1 - 3\alpha n e^{-c_2 n(\alpha\wedge 1)}$. Hence, $\widehat{\Theta}$ is inconsistent in the spectral norm as $n \to \infty$ and $p/n \to \alpha$.

3.4 Discussion

In this chapter, we establish the possibility of tuning-free estimation of a large precision matrix Θ^* based only on the knowledge that it is an M-matrix i.e., it has nonpositive off-diagonal entries. Our main contribution is to identify a loss—namely, the symmetrized Stein loss—in which $\widehat{\Theta}$ is both high-dimensionally consistent and minimax optimal. As the form (3.7) for the symmetrized Stein loss suggests, the quantity $L^{\text{ssym}}(\Theta, \Theta^*)$ is an average measure of closeness across all of the eigenvalues. The estimator $\widehat{\Theta}$ is inadequate, however, for estimating the extreme eigenvalues when p is large relative to n, and our other main result establishes that $\widehat{\Sigma}$ is minimax suboptimal in the spectral norm, even relative to the usual sample covariance matrix S. For the remainder of this section, we discuss some aspects that are naturally connected to our main results.

Misspecification. In practice, the assumption that all partial correlations are nonnegative may not hold exactly. Slawski and Hein (2015) empirically evaluate the impact of misspecification on the estimator $\widehat{\Theta}$, defining the *attractive part* $\Theta^\bullet \in \mathcal{M}^{p\times p}$ of the population precision $\Theta^* \notin \mathcal{M}^{p\times p}$ as the population analogue of the Bregman projection (3.3) with S replaced by Σ^*. Under the symmetrized Stein loss, a straightforward extension of Theorem 3.1 shows that $\widehat{\Theta}$ targets the attractive part $\Theta^\bullet$ even under misspecification.

Theorem 3.7. *Let $S = n^{-1}X^\mathsf{T}X$ denote the sample covariance based on $X \in \mathbb{R}^{n\times p}$ with i.i.d. $\mathcal{N}(0, \Sigma^*)$ rows. Define the attractive part $\Theta^\bullet \in \mathcal{M}^{p\times p}$ of the model as*

$$\Theta^\bullet := \operatorname*{argmin}_{\Theta\in\mathcal{M}^{p\times p}} \{\langle\Theta, \Sigma^*\rangle - \log\det\Theta\}.$$

For all $n \geq c_1 \gamma^2(\Sigma^\bullet) \log p$, *the MLE* $\widehat{\Theta}$ *defined in* (3.3) *satisfies*

$$L^{ssym}(\widehat{\Theta}, \Theta^\bullet) \leq c_2 \gamma(\Sigma^\bullet) \sqrt{\frac{\log p}{n}},$$

with probability at least $1 - c_3 p^{-2}$. *Here* c_1, c_2, c_3 *are universal positive constants.*

Estimating the correlation matrix. One may also be interested, under the same nonnegative partial correlations assumption, in estimating the population correlation matrix $\Gamma^* := D_{\Sigma^*}^{-1/2} \Sigma^* D_{\Sigma^*}^{-1/2}$ and its inverse $\Omega^* = (\Gamma^*)^{-1} = D_{\Sigma^*}^{1/2} \Theta^* D_{\Sigma^*}^{1/2}$ (here D_{Σ^*} denotes the diagonal matrix whose diagonal is equal to that of Σ^*). It is natural to use $\widehat{\Omega} := D_S^{1/2} \widehat{\Theta} D_S^{1/2}$ to estimate Ω^*. One can check that $\widehat{\Omega}$ satisfies

$$\widehat{\Omega} = \underset{\Omega \in \mathcal{M}^{p \times p}}{\operatorname{argmin}} \left\{ \langle \Omega, R \rangle - \log \det \Omega \right\}.$$

because the optimization problem is equivariant with respect to diagonal scaling (see Lauritzen et al., 2019, Lemma 2.5). The high-dimensional consistency result of Theorem 3.1 also holds for $\widehat{\Omega}$ as an estimator of the inverse correlation matrix Ω^*. This follows from an argument analogous to the proof of Theorem 3.1, with the tail bound for $\|S - \Sigma^*\|_\infty$ replaced by the corresponding tail bound on $\|R - \Gamma^*\|_\infty$ (see, e.g., Sun & Zhang, 2013, Lemma 19).

Non-Gaussian observations. We state Theorems 3.1 and 3.7 under the Gaussian assumption for simplicity and to remain consistent with other results in this chapter. In general, the upper bound depends on the tail behavior of $\|S - \Sigma^*\|_\infty$—see Lemma 3.8. A similar result holds when the rows of X are i.i.d. with σ-sub-Gaussian components. As Ravikumar et al. (2011) note, estimators of the form (3.3) are motivated via maximum likelihood yet remain sensible for non-Gaussian X. For general X, the estimator $\widehat{\Theta}$ is motivated as a Bregman projection of S with respect to the Stein loss.

Modifying $\widehat{\Theta}$. Although we focus on properties of the tuning-free estimator $\widehat{\Theta}$, additional processing such as thresholding $\widehat{\Theta}$ or pre-processing the sample covariance S may produce an estimator that is high-dimensionally consistent in the spectral norm. The tuning-free covariance estimate $\widehat{\Sigma}$ may also prove more useful for spectral analysis when the true covariance is a dense matrix. For instance, in the equicorrelation model where Σ^* has unit diagonal and every off-diagonal entry equal to $r \in (0, 1)$, the entry-wise inequalities in (3.16) may introduce less bias.

Related problems. Karlin and Rinott (1983), who pioneered the connection between M-matrices and MTP_2, also considered repulsive models where the covariance matrix $\Sigma^* \in \mathcal{M}^{p \times p}$ has nonpositive off-diagonal, in which case all marginal and partial correlations are nonpositive. This also defines an interesting model class which may similarly simplify estimation in high-dimensional problems. Note, however, that the constraint set $\{\Theta : \Theta^{-1} \in \mathcal{M}^{p \times p}\}$ of symmetric inverse-M matrices is nonconvex, presenting potential difficulties for maximum likelihood estimation.

Connection to shape-restricted regression. As a subset of the $p \times p$ symmetric positive-semidefinite matrices, the M-matrices $\mathcal{M}^{p \times p}$ form a closed, convex cone determined

only by sign constraints. The sign constraints on the precision matrix are analogous to a shape constraint in shape-restricted regression, enabling the use of likelihood techniques without explicit regularization. In particular, one can define the Bregman projection $\widehat{\Theta}$ of S onto $\mathcal{M}^{p\times p}$ (Lauritzen et al., 2019; Slawski & Hein, 2015). This work thus represents a first foray into the study of shape constraints for high-dimensional precision matrix estimation, inspired by results on regularization-free prediction in high-dimensional linear models via nonnegative least squares (Slawski & Hein, 2013).

3.5 Proofs

3.5.1 Proofs of Theorems 3.1 and 3.7

We first introduce two lemmas needed in the proof of Theorem 3.1. Following previous results on sparse precision matrix estimation (see, e.g., Cai et al., 2011; Ravikumar et al., 2011; Sun & Zhang, 2013), we rely on concentration of the entry-wise maximum deviation $\|S-\Sigma^*\|_\infty = \max_{j,k}|S_{jk}-\Sigma^*_{jk}|$ in the high-dimensional regime. A key technical tool in our analysis is the following lemma, which follows from an application of Bernstein's inequality.

Lemma 3.8. *Jankova and Van De Geer, 2015, Lemma 6 Suppose $X \in \mathbb{R}^{n\times p}$ has i.i.d. $\mathcal{N}(0,\Sigma^*)$ rows and let $S = n^{-1}X^\mathsf{T}X$. For any $t > 2$,*

$$\mathbb{P}\left(\|S-\Sigma^*\|_\infty \geq 2\|\Sigma^*\|_\infty\left[\sqrt{\frac{2t\log p}{n}}+\frac{t\log p}{n}\right]\right) \leq \frac{2}{p^{t-2}}.$$

Proof. Let $\alpha = e_j$ and $\beta = e_k$ denote the standard basis vectors. Lemma 6 of Jankova and Van De Geer (2015) provides

$$\mathbb{P}\left(\alpha^\mathsf{T}(S-\Sigma^*)\beta \geq 2\|\Sigma^*\|_\infty\left[\sqrt{\frac{2x}{n}}+\frac{x}{n}\right]\right) \leq 2e^{-x}.$$

Taking a union bound over $j \leq k$ and setting $x = \log p^t$ yields the claim. □

The next lemma records a distinctive property of M-matrices, corresponding to the fact that M-matrices are generalized diagonally dominant (Plemmons, 1977).

Lemma 3.9. *Every M-matrix $\Theta \in \mathcal{M}^{p\times p}$ satisfies $\|\Theta\|_1 := \sum_{i,j}|\Theta_{ij}| \leq 2\mathrm{tr}(\Theta)$.*

Proof. Since Θ is symmetric PSD, there are vectors $\theta_1,\ldots,\theta_p$ such that $\Theta_{ij} = \langle\theta_i,\theta_j\rangle$. Moreover, since Θ has nonpositive off-diagonal entries, $\langle\theta_i,\theta_j\rangle \leq 0$ for $i \neq j$. Hence

$$\|\Theta\|_1 = \sum_i\|\theta_i\|_2^2 - \sum_{i\neq j}\langle\theta_i,\theta_j\rangle = 2\sum_i\|\theta_i\|_2^2 - \left\|\sum_i\theta_i\right\|_2^2 \leq 2\sum_i\|\theta_i\|_2^2 = 2\mathrm{tr}(\Theta). \qquad \square$$

An illustrative example is the one-parameter family of $p \times p$ symmetric matrices $A_x = (1-x)I_p + x1_p1_p'$ (where $1_p = \sum_{j=1}^p e_j$ is the all ones vector) with unit diagonal and every off-diagonal equal to x. Its eigenvalues are $1-x$ (with multiplicity $p-1$) and $1+(p-1)x$. Thus A_x is PSD if and only if $x \in \left[-\frac{1}{p-1}, 1\right]$, whereas A_x is an M-matrix if and only if $x \in \left[-\frac{1}{p-1}, 0\right]$. Finally, note $\|A_x\|_1 = p + p(p-1)|x|$ and $\mathrm{tr}(A_x) = p$. This example shows Lemma 3.9 is tight. For general PSD matrices, the element-wise ℓ_1-norm can be as large as p times the trace, but for M-matrices it can be at most twice as large.

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. For any positive diagonal matrix $D \in \mathcal{D}_+^{p\times p}$,

$$\begin{aligned} L^{\mathrm{ssym}}(\widehat{\Theta}(S), \Theta^*) &= L^{\mathrm{ssym}}(D\widehat{\Theta}(S)D, D\Theta^*D) \\ &= L^{\mathrm{ssym}}(\widehat{\Theta}(D^{-1}SD^{-1}), D^{-1}\Sigma^*D^{-1}), \end{aligned}$$

where the first step uses the fact that L^{ssym} is invariant under congruent transformations, and the second step uses the scale-invariance of the program (3.3). With a sample covariance S based on Gaussian observations, the loss $L^{\mathrm{ssym}}(\widehat{\Theta}, \Theta^*)$ has the same distribution for covariance matrices of the form $\{D^{-1}\Sigma^*D^{-1}\}_{D\in\mathcal{D}_+^{p\times p}}$. In particular, taking $D = D_{\Sigma^*}^{1/2}$, we may assume without loss of generality that Σ^* is *normalized*; i.e., Σ^* has unit diagonal or equivalently Σ^* equals the population correlation matrix Γ^*.

Let $f(\Theta) = \langle \Theta, S\rangle - \log|\Theta|$. Since the estimator solves the constrained convex optimization problem $\widehat{\Theta} = \arg\min_{\Theta\in\mathcal{M}^{p\times p}} f(\Theta)$, it is characterized by $\langle \nabla f(\widehat{\Theta}), \Theta - \widehat{\Theta}\rangle \geq 0$, for all $\Theta \in \mathcal{M}^{p\times p}$, where $\nabla f(\Theta) = S - \Theta^{-1}$. Hence

$$\langle S - \widehat{\Sigma}, \Theta^* - \widehat{\Theta}\rangle \geq 0.$$

Rearranging yields the basic inequality

$$L^{\mathrm{ssym}}(\widehat{\Theta}, \Theta^*) \leq \frac{1}{2p}\left\langle S - \Sigma^*, \Theta^* - \widehat{\Theta}\right\rangle.$$

Let $A := \|S - \Sigma^*\|_\infty$. Using Hölder's inequality, we have:

$$L^{\mathrm{ssym}}(\widehat{\Theta}, \Theta^*) \leq \frac{A}{2p}\left\|\Theta^* - \widehat{\Theta}\right\|_1.$$

Now applying the triangle inequality and Lemma 3.9 to the element-wise ℓ_1-norm,

$$L^{\mathrm{ssym}}(\widehat{\Theta}, \Theta^*) \leq \frac{A}{p}\left(\mathrm{tr}(\Theta^*) + \mathrm{tr}(\widehat{\Theta})\right).$$

Since we have assumed without loss of generality that $\Sigma^* = \Gamma^*$,

$$\mathrm{tr}(\widehat{\Theta}\Sigma^*) = \mathrm{tr}(\widehat{\Theta}) + \sum_{j \neq k} \widehat{\Theta}_{jk}\Gamma^*_{jk} \geq \left(1 - \max_{j \neq k} \Gamma^*_{jk}\right) \mathrm{tr}(\widehat{\Theta})$$
$$p = \mathrm{tr}(\Theta^*\Sigma^*) = \mathrm{tr}(\Theta^*) + \sum_{j \neq k} \Theta^*_{jk}\Gamma^*_{jk} \geq \left(1 - \max_{j \neq k} \Gamma^*_{jk}\right) \mathrm{tr}(\Theta^*),$$

where we have again used Lemma 3.9, along with the facts that $\widehat{\Theta}_{jk}$ and Θ^*_{jk} are nonpositive for $j \neq k$ and $\Sigma^* \geq 0$ entry-wise (see Berman & Plemmons, 1994, Chap. 6, Property N_{38}). Combining the last three displays and using the characterization of $\gamma(\Sigma^*)$ in (3.8), we get

$$\begin{aligned} L^{\mathrm{ssym}}(\widehat{\Theta}, \Theta^*) &\leq \frac{\gamma(\Sigma^*)A}{p}\left(p + \mathrm{tr}(\widehat{\Theta}\Sigma^*)\right) \\ &\leq \gamma(\Sigma^*)A\left(3 + 2L^{\mathrm{ssym}}(\widehat{\Theta}, \Theta^*)\right). \end{aligned}$$

On the event $E = \left\{2\gamma(\Sigma^*)A \leq \frac{1}{2}\right\}$, we have $L^{\mathrm{ssym}}(\widehat{\Theta}, \Theta^*) \leq 6\gamma(\Sigma^*)A$. Applying Lemma 3.8 with $t = 4$, the event $E' = \left\{A \leq 2\sqrt{\frac{8\log p}{n}} + \frac{8\log p}{n}\right\}$ occurs with probability at least $1 - 2/p^2$.

To guarantee $E' \subset E$, we require

$$2\sqrt{\frac{8\log p}{n}} + \frac{8\log p}{n} \leq \frac{1}{4\gamma(\Sigma^*)},$$

which is equivalent to

$$\frac{\log p}{n} \leq 2 + \frac{1}{4\gamma(\Sigma^*)} - \sqrt{4 + \gamma^{-1}(\Sigma^*)}.$$

Using $\gamma(\Sigma^*) \geq 1$, it is straightforward to check that the right hand side above is at least $\frac{1}{72\gamma^2(\Sigma^*)}$. Hence, as long as $n \geq 72\gamma^2(\Sigma^*)\log p$,

$$L^{\mathrm{ssym}}(\widehat{\Theta}, \Theta^*) \leq 6\gamma(\Sigma^*)\left(2\sqrt{\frac{8\log p}{n}} + \frac{8\log p}{n}\right)$$

with probability at least $1 - 2/p^2$. Since $\gamma(\Sigma^*) \geq 1$, the $\sqrt{\frac{8\log p}{n}}$ dominates the $\frac{8\log p}{n}$ term. In particular, we have $\sqrt{\frac{8\log p}{n}} \leq \frac{1}{3}$, so $L^{\mathrm{ssym}}(\widehat{\Theta}, \Theta^*) \leq 28\gamma(\Sigma^*)\sqrt{\frac{2\log p}{n}}$ with probability at least $1 - 2/p^2$. □

Proof of Theorem 3.7. Since the attractive part $\Theta^\bullet$ is an M-matrix, from the first order optimality conditions for $\widehat{\Theta}$,

$$\langle S - \widehat{\Sigma}, \Theta^\bullet - \widehat{\Theta}\rangle \geq 0.$$

Using the first order optimality conditions for $\Theta^\bullet$ and the fact that $\widehat{\Theta} \in \mathcal{M}^{p\times p}$,

$$\langle \Sigma^* - \Sigma^\bullet, \widehat{\Theta} - \Theta^\bullet \rangle \geq 0.$$

Adding these and rearranging yields the basic inequality

$$L^{\text{ssym}}(\Theta^\bullet, \widehat{\Theta}) \leq \frac{1}{2p} \langle S - \Sigma^*, \Theta^\bullet - \widehat{\Theta} \rangle.$$

The rest of the proof proceeds as the proof of Theorem 3.1, substituting Θ^* with $\Theta^\bullet$. □

3.5.2 Proof of Theorem 3.2

Proof of Theorem 3.2. As in Cai et al. (2016a, Proof of Theorem 4.1), we consider precision matrices of the form

$$\Theta = \begin{bmatrix} I_{\lceil p/2 \rceil} & \varepsilon A \\ \varepsilon A^\mathsf{T} & I_{\lfloor p/2 \rfloor} \end{bmatrix}, \tag{3.18}$$

where A is a sparse binary matrix with k nonzero entries per row and at most $2k$ nonzero entries per column, for some positive integer k and some ε to be chosen later. As long as $\varepsilon < 0$ and $2k|\varepsilon| < 1$, the matrix Θ is a diagonally dominant M-matrix. Its inverse is given by the Neumann series

$$\begin{aligned} \Sigma = \Theta^{-1} &= \sum_{m=0}^{\infty} (-\varepsilon)^m \begin{bmatrix} 0 & A \\ A^\mathsf{T} & 0 \end{bmatrix}^m \\ &= \sum_{m=0}^{\infty} \varepsilon^{2m} \begin{bmatrix} (AA^\mathsf{T})^m & -\varepsilon A (A^\mathsf{T}A)^m \\ -\varepsilon A^\mathsf{T}(AA^\mathsf{T})^m & (A^\mathsf{T}A)^m \end{bmatrix} \end{aligned}$$

From the last display, it is clear that $D_\Sigma \geq I_p$, so $\max_{j\neq k} \Gamma_{jk} \leq \max_{j\neq k} \Sigma_{jk}$ where Γ is the correlation matrix corresponding to Σ. Furthermore, by triangle-inequality, the largest off-diagonal entry of the top left diagonal block is at most

$$\left\| \sum_{m=0}^{\infty} \varepsilon^{2m} (AA^\mathsf{T})^m \right\|_{\infty,\text{off}} \leq \sum_{m=1}^{\infty} \varepsilon^{2m} \left\| (AA^\mathsf{T})^m \right\|_{\infty,\text{off}},$$

where we use that the first term $m = 0$ has zero off-diagonal. This yields

$$\left\| (AA^\mathsf{T})^m \right\|_{\infty,\text{off}} \leq \left\| (AA^\mathsf{T})^m \right\|_2 \leq (2k)^{2m}.$$

By similar bounds on the other blocks of Σ, it can be shown that

$$\max_{j\neq k} \Gamma_{jk} \leq \max_{j\neq k} \Sigma_{jk} \leq \frac{2k|\varepsilon|}{1 - (2k\varepsilon)^2}.$$

A simple sufficient condition to guarantee $\gamma(\Sigma) \le \gamma$ is thus $4k|\varepsilon| \le (1-\gamma^{-1}) \wedge \frac{1}{2}$.

By the Geršgorin circle theorem, the spectrum of Θ lies in the range $[0, 2]$. Further constraining the supremum in (3.10) to $\lambda_{\max}(\Theta) \le 2$, by Cai and Zhou (2012, Eq. (54)), we have:

$$\inf_{\breve{\Theta}} \sup_{\substack{\Theta \in \mathcal{M}^{p\times p} \\ \gamma(\Sigma)\le\gamma}} \mathbb{E} L^{\text{ssym}}(\breve{\Theta}, \Theta) \ge \frac{1}{4} \inf_{\breve{\Theta}} \sup_{\substack{\Theta \in \mathcal{M}^{p\times p} \\ \gamma(\Sigma)\le\gamma \\ \lambda_{\max}(\Theta)\le 2}} \mathbb{E} \frac{\|\breve{\Theta} - \Theta\|_F^2}{p},$$

so it suffices to lower bound the minimax rate in the Frobenius norm.

Now let $\mathcal{A}$ denote the set of all $\lceil p/2 \rceil \times \lfloor p/2 \rfloor$ binary matrices with k nonzero entries per row and at most $2k$ nonzero entries per column, and $\mathcal{B} = \{0,1\}^{\lceil p/2 \rceil}$. Finally, let e denote a vector of ones of length $\lfloor p/2 \rfloor$. Given $A \in \mathcal{A}$ and $b \in \mathcal{B}$, the matrix $(b \otimes e) \circ A$ has the same shape as A, where the j^{th} row is nonzero if and only if $b_j = 1$. Let

$$\mathcal{F} = \left\{ \Theta_{A,b} = \begin{bmatrix} I_{\lceil p/2 \rceil} & \varepsilon(b \otimes e) \circ A \\ \varepsilon(b^\mathsf{T} \otimes e^\mathsf{T}) \circ A^\mathsf{T} & I_{\lfloor p/2 \rfloor} \end{bmatrix} : A \in \mathcal{A}, b \in \mathcal{B} \right\}.$$

As we have shown, $\mathcal{F} \subset \{\Theta \in \mathcal{M}^{p\times p} : \gamma(\Sigma) \le \gamma, \lambda_{\max}(\Theta) \le 2\}$. By Cai and Zhou, 2012, Lemma 3

$$\inf_{\breve{\Theta}} \max_{\Theta \in \mathcal{F}} \mathbb{E} \frac{\|\breve{\Theta} - \Theta\|_F^2}{p} \ge \frac{1}{32} \left[\min_{\substack{A, A' \in \mathcal{A}, b, b' \in \mathcal{B} \\ b \ne b'}} \frac{\|\Theta_{A,b} - \Theta_{A',b'}\|_F^2}{H(b,b')} \right] \left[\min_{1 \le j \le \lceil p/2 \rceil} \|\bar{P}_{j,0} \wedge \bar{P}_{j,1}\| \right],$$

where H denotes the Hamming distance and $\|\bar{P}_{j,0} \wedge \bar{P}_{j,1}\|$ denotes the total variation affinity between the measures $\bar{P}_{j,0}$ and $\bar{P}_{j,1}$, where $\bar{P}_{j,i}$ is the uniform mixture over $\mathcal{N}(0, \Theta_{A,b}^{-1})$ over all $A \in \mathcal{A}$ and all $b \in \mathcal{B}$ such that $b_j = i$.

For the first term, fix A, A' and $b \ne b'$. For j such that $b_j \ne b'_j$, if say $b_j = 0$, the j^{th} row of $(b \otimes e) \circ A$ is zero and the j^{th} row of $(b' \otimes e) \circ A'$ has k nonzero entries. Hence

$$\min_{\substack{A, A' \in \mathcal{A}, b, b' \in \mathcal{B} \\ b \ne b'}} \frac{\|\Theta_{A,b} - \Theta_{A',b'}\|_F^2}{H(b,b')} \ge \min_{\substack{A, A' \in \mathcal{A}, b, b' \in \mathcal{B} \\ b \ne b'}} \frac{2\sum_{j : b_j \ne b'_j} k\varepsilon^2}{H(b,b')} = 2k\varepsilon^2.$$

In particular, we have shown

$$\inf_{\breve{\Theta}} \sup_{\substack{\Theta \in \mathcal{M}^{p\times p} \\ \gamma(\Sigma)\le\gamma}} \mathbb{E} L^{\text{ssym}}(\breve{\Theta}, \Theta) \ge ck\varepsilon^2 \min_{1 \le j \le \lceil p/2 \rceil} \|\bar{P}_{j,0} \wedge \bar{P}_{j,1}\|.$$

Finally, the same argument of (Cai et al., 2016a, proof of Lemma 4.5) with $\varepsilon = c'\sqrt{\frac{\log p}{n}}$ can be used to show $\min_{1 \le j \le \lceil p/2 \rceil} \|\bar{P}_{j,0} \wedge \bar{P}_{j,1}\| \ge c'' > 0$, yielding

$$\inf_{\breve{\Theta}} \sup_{\substack{\Theta \in \mathcal{M}^{p\times p} \\ \gamma(\Sigma)\le\gamma}} \mathbb{E} L^{\text{ssym}}(\breve{\Theta}, \Theta) \ge cc''k\varepsilon^2 = c_\gamma \varepsilon. \qquad \square$$

3.5.3 Proof of Proposition 3.3

Proof of Proposition 3.3. Let $\tilde{\Sigma}^{\mathrm{DIAG}} = c \cdot D_S$. Since $S_{11} = \frac{1}{n}\sum_{i=1}^n X_{i1}^2$ for $X_{i1} \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma^*_{11})$,

$$\mathbb{E}\left[L^{\mathrm{ssym}}(\tilde{\Theta}^{\mathrm{DIAG}}, \Theta^*)\right] = \frac{1}{2}\mathbb{E}\left[\frac{\Sigma^*_{11}}{cS_{11}} + \frac{cS_{11}}{\Sigma^*_{11}} - 2\right] = \frac{1}{2}\left[\frac{1}{c}\frac{n}{n-2} + c - 2\right].$$

The minimum is achieved at $c = \sqrt{\frac{n}{n-2}}$, but taking $c = 1$ suffices to prove the minimax rate (3.11) is upper bounded by $\frac{C}{n}$.

Now consider a prior G on $\mathcal{D}_+^{p\times p}$ over which the components Θ^*_{jj} are i.i.d. Lower bound the minimax risk by the Bayes risk with respect to G:

$$\begin{aligned}\inf_{\hat{\Sigma}} \sup_{\Sigma^* \in \mathcal{D}_+^{p\times p}} \mathbb{E}L^{\mathrm{ssym}}(\hat{\Sigma}, \Sigma^*) &\geq \inf_{\hat{\Sigma}} \mathbb{E}_G L^{\mathrm{ssym}}(\hat{\Sigma}, \Sigma^*) \\ &= \inf_{\hat{\Sigma}_{11}} \mathbb{E}_G L^{\mathrm{ssym}}(\hat{\Sigma}_{11}, \Sigma^*_{11}).\end{aligned}$$

If $G = [\mathrm{Gamma}(a,b)]^{\otimes n}$, such that $\Theta^*_{jj} \overset{\text{iid}}{\sim} \mathrm{Gamma}(a,b)$ under G, then combining with the likelihood we have:

$$S_{11} \mid \Theta^*_{11} \sim \mathrm{Gamma}(a,b).$$

By conjugacy, the posterior is readily seen to be

$$\Theta^*_{11} \mid S_{11} = s \sim \mathrm{Gamma}\left(a + \frac{n}{2}, b + \frac{ns}{2}\right).$$

Thus, for $n > 2$,

$$\begin{aligned}\mathbb{E}_G\left[L^{\mathrm{ssym}}(\mathfrak{d}, \Sigma^*_{11}) \mid S_{11} = s\right] &= \frac{1}{2}\mathbb{E}_G\left[\mathfrak{d}\Theta^*_{11} + \frac{\Sigma^*_{11}}{\mathfrak{d}} - 2 \mid S_{11} = s\right] \\ &= \frac{\mathfrak{d}}{2}\frac{a + n/2}{b + ns/2} + \frac{1}{2\mathfrak{d}}\frac{b + ns/2}{a + n/2 - 1} - 1.\end{aligned}$$

This is minimized at $\mathfrak{d}^* = \frac{b+ns/2}{\sqrt{(a+n/2)(a+n/2-1)}}$, giving a Bayes risk of

$$\mathbb{E}_G\left[L^{\mathrm{ssym}}(\mathfrak{d}^*, \Sigma^*_{11})\right] = \sqrt{\frac{a + n/2}{a + n/2 - 1}} - 1.$$

Letting $a \downarrow 0$, we find

$$\begin{aligned}\inf_{\hat{\Sigma}} \sup_{\Sigma^* \in \mathcal{D}_+^{p\times p}} \mathbb{E}L^{\mathrm{ssym}}(\hat{\Sigma}, \Sigma^*) &\geq \sqrt{1 + \frac{2}{n-2}} - 1 \\ &= \frac{1}{n-2} + o(n^{-1}),\end{aligned}$$

as $n \to \infty$. This proves the minimax rate (3.11) on $\mathcal{D}_+^{p\times p}$.

To prove the lower bound (3.12) on $\mathcal{S}_{\succeq 0}^{p\times p}$, place an inverse Wishart prior $\Sigma^* \sim \mathcal{W}^{-1}(\Sigma_0, \nu)$ on the covariance matrix. By conjugacy,

$$\Sigma^* \mid S \sim \mathcal{W}^{-1}(\Sigma^*, \Sigma_0 + nS, \nu + n).$$

As long as $\nu + n > p + 1$, the posterior loss can be written in closed form as

$$\mathbb{E}\left[L^{\text{ssym}}(\hat{\Theta}, \Theta^*) \mid S\right] = \frac{1}{2p}\left[(\nu+n)\text{tr}(\hat{\Sigma}(\Sigma_0 + nS)^{-1}) + \frac{\text{tr}(\hat{\Theta}(\Sigma_0 + nS))}{\nu + n - p - 1} - 2p\right],$$

which is minimized at $\hat{\Theta} = \sqrt{(\nu+n)(\nu+n-p-1)}(\Sigma_0 + nS)^{-1}$, yielding a Bayes risk of

$$\mathbb{E}\left[L^{\text{ssym}}(\hat{\Theta}, \Theta^*) \mid S\right] = \sqrt{\frac{\nu+n}{\nu+n-p-1}} - 1,$$

independent of Σ_0. Setting $\nu = p + 1$,

$$\inf_{\hat{\Theta}=\hat{\Theta}(S)} \sup_{\Theta^* \succeq 0} \mathbb{E}L^{\text{ssym}}(\hat{\Theta}, \Theta^*) \geq \sqrt{1 + \frac{p+1}{n}} - 1,$$

Finally, use $\sqrt{1+x} - 1 \geq (\sqrt{2}-1)(x \wedge \sqrt{x})$ for any $x \geq 0$. □

3.5.4 Proof of Theorem 3.4

Proof of Theorem 3.4. This proof uses Theorem 3.5 which is proved in the next subsection. Since $\Sigma^* \in \mathcal{D}_+^{p\times p}$, as in the proof of Theorem 3.1 we have

$$L^{\text{ssym}}(\widehat{\Theta}(S), \Theta^*) = L^{\text{ssym}}\left(\widehat{\Theta}(D_{\Sigma^*}^{-1/2} S D_{\Sigma^*}^{-1/2}), I_p\right).$$

In particular, due to scale invariance of both the estimator and the loss, the symmetrized Stein loss $L^{\text{ssym}}(\widehat{\Theta}, \Theta^*)$ has the same distribution for all diagonal matrices $\Sigma^* \in \mathcal{D}_+^{p\times p}$. We thus assume with no loss of generality that $\Sigma^* = I_p$.

Let $f(t) = t + t^{-1} - 2$ for $t > 0$. By (3.7) and nonnegativity of the function f,

$$L^{\text{ssym}}(\widehat{\Theta}, I_p) = \frac{1}{p}\sum_{j=1}^{p} f(\lambda_j(\widehat{\Theta})) \geq \frac{f(\lambda_{\max}(\widehat{\Theta}))}{p}.$$

For $t > 1$, $f'(t) > 0$, so by Theorem 3.5,

$$L^{\text{ssym}}(\widehat{\Theta}, I_p) \geq \frac{1}{p} f\left(1 + c_1 \frac{p}{\sqrt{n}}\right) = \frac{c_1}{\sqrt{n}}\left[1 - \frac{1}{1 + c_1 \frac{p}{\sqrt{n}}}\right],$$

with probability at least $1 - 3p\exp(-c_2(n \wedge p))$. If $c_1 p \geq \sqrt{n}$, this implies $L^{\text{ssym}}(\widehat{\Theta}, I_p) \geq \frac{c_1}{2\sqrt{n}}$, completing the proof. □

3.5.5 Proof of Theorem 3.5

The most technically involved part of the proof is a lower bound on the row sums of the positive part S_+ of the sample covariance matrix, which we include as a separate lemma.

Lemma 3.10. *Under the conditions of Theorem 3.5,*

$$\sum_{j=1}^{p}(S_+)_{pj} \geq 1 + c_0\frac{p}{\sqrt{n}},$$

with probability at least $1 - 3\exp(-c_1(n \wedge p))$*, for some universal positive constants* c_0, c_1*.*

We give the proof of Theorem 3.5 assuming the above lemma and then prove the lemma subsequently.

Proof of Theorem 3.5. Since $\widehat{\Sigma}$ is an inverse M-matrix, it is entry-wise nonnegative; i.e., $\widehat{\Sigma} \geq 0$. Combining this with the first constraint $\widehat{\Sigma} \geq S$ in the dual formulation (3.16), we have that $\widehat{\Sigma} \geq S_+ \geq 0$, where S_+ is the entry-wise positive part of the sample covariance matrix S. The Perron-Frobenius theorem Berman and Plemmons, 1994, Corollary 1.5 gives

$$\lambda_{\max}(\widehat{\Sigma}) \geq \lambda_{\max}(S_+).$$

Thus, we want to show that $\lambda_{\max}(S_+)$ is more severely biased than $\lambda_{\max}(S)$. To this end, we apply another standard result from the spectral theory of nonnegative matrices Berman and Plemmons, 1994, Theorem 2.35:

$$\lambda_{\max}(S_+) \geq \min_k \sum_j (S_+)_{jk}.$$

By Lemma 3.10, $\sum_j (S_+)_{jk} \geq 1 + c_0\frac{p}{\sqrt{n}}$ with probability at least $1 - 3e^{-c_1(n\wedge p)}$ for each fixed k, so by a union bound,

$$\min_k \sum_j (S_+)_{jk} \geq 1 + c_0\frac{p}{\sqrt{n}}$$

with probability at least $1 - 3pe^{-c_1(n\wedge p)}$. Combining the last three displays gives the desired lower bound on $\lambda_{\max}(\widehat{\Sigma})$. □

We now prove the key lemma on the row sums of S_+.

Proof of Lemma 3.10. For $u > 0$, write

$$\mathbb{P}\left\{\sum_{j=1}^{p}(S_+)_{pj} \leq 1 + u\right\} \leq \mathbb{P}\left\{S_{pp} \leq 1 - u\right\} + \mathbb{P}\left\{\sum_{j<p}(S_{pj})_+ \leq 2u\right\}.$$

To bound the first term, note that $nS_{pp} \sim \chi^2_n$ and the following standard chi-squared lower tail bound (see e.g., Laurent and Massart (2000, inequality (4.4))):

$$\mathbb{P}\left\{\frac{\chi^2_n}{n} \leq 1-u\right\} \leq \exp\left(\frac{-nu^2}{4}\right) \tag{3.19}$$

gives

$$\mathbb{P}\left\{S_{pp} \leq 1-u\right\} \leq \exp\left(-\frac{nu^2}{4}\right). \tag{3.20}$$

To bound the second term, notice that conditionally on $X_{ip}, i = 1, \ldots, n$,

$$S_{pj}, j = 1, \ldots, p-1 \Big| X_{ip}, i = 1, \ldots, n \overset{\text{i.i.d}}{\sim} N\left(0, \frac{1}{n^2}\sum_{i=1}^n X_{ip}^2\right).$$

Thus, conditionally on $X_{ip}, i = 1, \ldots, n$, we can write $S_{pj} = AZ_j$ for $j = 1, \ldots, p-1$ where

$$A^2 := \frac{1}{n^2}\sum_{i=1}^n X_{ip}^2 \quad \text{and} \quad Z_1, \ldots, Z_{p-1} \overset{\text{i.i.d}}{\sim} N(0,1).$$

We can therefore write (using the notation $\mathbb{P}^|$ for probability conditioned on $X_{ip}, i = 1, \ldots, n$)

$$\begin{aligned}\mathbb{P}^|\left\{\sum_{j<p}(S_{pj})_+ \leq 2u\right\} &= \mathbb{P}^|\left\{\sum_{j<p}(Z_j)_+ \leq \frac{2u}{A}\right\} \\ &= \mathbb{P}^|\left\{\frac{1}{p-1}\sum_{j<p}((Z_j)_+ - c) \leq \frac{2u}{(p-1)A} - c\right\},\end{aligned}$$

where $c := \mathbb{E}(Z_1)_+ = (2\pi)^{-1/2}$ is a universal constant. We now note that

$$(z_1, \ldots, z_{p-1}) \mapsto \frac{1}{p-1}\sum_{j<p}(z_j)_+$$

is a Lipschitz function with Lipschitz constant $(p-1)^{-1/2}$. Thus by the usual concentration inequality for Lipschitz functions of Gaussian random vectors (see, e.g., Wainwright, 2019, Theorem 2.26), we obtain

$$\mathbb{P}^|\left\{\frac{1}{p-1}\sum_{j<p}((Z_j)_+ - c) \leq \frac{2u}{(p-1)A} - c\right\} \leq \exp\left(-\frac{(p-1)}{2}\left(c - \frac{2u}{(p-1)A}\right)^2\right),$$

assuming that $c > 2u/(A(p-1))$. In particular, for $c > 4u/(A(p-1))$, we get

$$\mathbb{P}^|\left\{\frac{1}{p-1}\sum_{j<p}((Z_j)_+ - c) \leq \frac{2u}{(p-1)A} - c\right\} \leq \exp\left(-\frac{(p-1)c^2}{8}\right).$$

We have thus proved

$$\mathbb{P}^{|}\left\{\sum_{j<p}(S_{pj})_+ \leq 2u\right\} \leq \exp\left(-\frac{(p-1)c^2}{8}\right) + I\left\{c \leq 4u/(A(p-1))\right\}.$$

Taking an expectations on both sides of this expression, we obtain

$$\mathbb{P}\left\{\sum_{j<p}(S_{pj})_+ \leq 2u\right\} \leq \exp\left(-\frac{(p-1)c^2}{8}\right) + \mathbb{P}\left\{A \leq \frac{4u}{(p-1)c}\right\}.$$

Note now that $n^2A^2 \sim \chi^2_n$ and thus

$$\mathbb{P}\left\{A \leq \frac{4u}{(p-1)c}\right\} = \mathbb{P}\left\{\frac{\chi^2_n}{n} - 1 \leq \frac{16u^2n}{(p-1)^2c^2} - 1\right\}.$$

We now make the choice $u = \frac{(p-1)c}{4\sqrt{2}\sqrt{n}}$, which gives (via (3.19))

$$\mathbb{P}\left\{A \leq \frac{4u}{(p-1)c}\right\} = \mathbb{P}\left\{\frac{\chi^2_n}{n} - 1 \leq \frac{-1}{2}\right\} \leq \exp\left(-\frac{n}{16}\right).$$

We have thus proved

$$\mathbb{P}\left\{\sum_{j<p}(S_{pj})_+ \leq \frac{(p-1)c}{2\sqrt{2}\sqrt{n}}\right\} \leq \exp\left(-\frac{(p-1)c^2}{8}\right) + \exp\left(-\frac{n}{16}\right).$$

Combining this with (3.20) and using $c = (2\pi)^{-1/2}$, we obtain

$$\mathbb{P}\left\{\sum_{j=1}^{p}(S_+)_{pj} \leq 1 + \frac{(p-1)}{8\sqrt{\pi n}}\right\} \leq \exp\left(-\frac{(p-1)^2}{256\pi}\right) + \exp\left(-\frac{p-1}{16\pi}\right) + \exp\left(-\frac{n}{16}\right).$$

For $p \geq 17$ the first term is of lower order; i.e., $\exp\left(-\frac{(p-1)^2}{256\pi}\right) \leq \exp\left(-\frac{p-1}{16\pi}\right)$. □

Chapter 4

Shrinkage for multivariate, heteroscedastic data

4.1 Introduction

Consider a d-dimensional ($d \geq 1$), heteroscedastic normal observation model

$$X_i \mid \theta_i^* \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i^*, \Sigma_i), \qquad \text{with} \qquad \theta_i^* \overset{\text{iid}}{\sim} G^*, \qquad \text{for } i \in \{1, \ldots, n\}, \tag{4.1}$$

where $(\Sigma_i)_{i=1}^n$ is a known sequence of $d \times d$ positive-definite covariance matrices, and the underlying mean vectors $(\theta_i^*)_{i=1}^n$ are additionally assumed to be drawn from a common prior G^*, where G^* belongs to the collection $\mathcal{P}(\mathbb{R}^d)$ of all probability measures on $\mathbb{R}^d$. In settings where G^* is known, model (4.1) fully specifies a Bayesian model; this chapter studies the common empirical Bayes setting where G^* must be estimated. The main goal of the chapter is to nonparametrically estimate G^* and the sequence $(\theta_i^*)_{i=1}^n$ from the observed data $(X_i, \Sigma_i)_{i=1}^n$.

Empirical Bayes methods for the normal sequence model (4.1) have been studied extensively in the univariate, homoscedastic setting where $d = 1$ and $\Sigma_i \equiv \sigma^2$ (see, e.g., Efron (2012, 2014), Efron and Morris (1972a, 1972b, 1973a, 1973b), James and Stein (1961), and Morris (1983) as well as Johnstone (2019) for a manuscript on estimation in Gaussian sequence models). Numerous methods extend empirical Bayes to the univariate, heteroscedastic case (see Banerjee et al., 2021; Jiang et al., 2011; Jiang, 2020; Tan, 2016; Weinstein et al., 2018; Xie et al., 2012, and references therein). Relatively little attention has been given to the general case of the present chapter.

Model (4.1) naturally arises in the analysis of astronomy data, where often a calibrated measurement error distribution comes attached to each observation, and typically these errors are heteroscedastic (Kelly, 2012); also see e.g. Akritas and Bershady (1996), Hogg et al. (2010), Anderson et al. (2018). The first part of model (4.1) indicates that the target sequence $(\theta_i^*)_{i=1}^n$ has, due to measurement error, been corrupted by additive, zero-mean

Gaussian noise, i.e.

$$X_i = \theta_i^* + \epsilon_i, \qquad \text{where} \quad \epsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \Sigma_i), \qquad \text{for } i = 1, \ldots, n.$$

Interestingly, the Σ_i's above, which typically differ across i, are known in many applications where the measurement process is well-characterized. In many situations it is assumed that θ_i^* is itself random and independent of ϵ_i for all i. Although each observation has a different error distribution, the n observations are tied together by the assumption that the θ_i^*'s are i.i.d. from some distribution G^*, yielding model (4.1). By allowing for arbitrary prior distributions $G^* \in \mathcal{P}(\mathbb{R}^d)$, model (4.1) captures a range of important structural assumptions on the underlying sequence $(\theta_i^*)_{i=1}^n$: for instance, the clustering problem (where the terms of $(\theta_i^*)_{i=1}^n$ take on at most k^* distinct values) corresponds to discrete G^*, and sparse modeling (where most of the $(\theta_i^*)_{i=1}^n$ are zero) corresponds to $G^*(\{0\}) \approx 1$. The model also accommodates more complex manifold-like structures (see e.g. Figure 4.1) as well as substantially more heterogeneous sequences (e.g. G^* heavy tailed).

Our motivating example for model (4.1) involves the construction of a precise stellar color-magnitude diagram. A color-magnitude diagram (CMD) is a scatter plot of stars, displaying their absolute magnitude (luminosity) versus color (surface temperature) to provide a cross-sectional view of stellar evolution. The continued expansion of available stellar measurements has made purely statistical models such as model (4.1) increasingly attractive for denoising. One common approach, known as *Extreme Deconvolution* (XD) (Bovy et al., 2011), assumes

$$G^* = \sum_{j=1}^{K} \alpha_j^* \mathcal{N}(\mu_j^*, V_j^*) \tag{4.2}$$

and estimates the parameters $(\alpha_j^*, \mu_j^*, V_j^*)_{j=1}^K$ via the *Expectation-Maximization* (EM) algorithm with split-and-merge operations designed to avoid local optima. For instance, Anderson et al. (2018) applied XD to build a low-noise CMD with $n \approx 1.4$ million de-reddened stars from the Gaia TGAS catalogue. The XD assumption (4.2) that the prior G^* is itself a mixture of K-Gaussians has a number of drawbacks. Although the class of Gaussian location-scale mixtures is flexible for large K, the choice of K requires tuning; violations of assumption (4.2) for fixed K induce bias in the estimation. To our knowledge, no theoretical results for the statistical properties of XD are available, making it difficult to quantify the misspecification error. Moreover, the class of all probability distributions of the form (4.2) is nonconvex for finite K, so even split-and-merge techniques employed within EM do not guarantee convergence to the global maximizer of the likelihood.

To avoid these difficulties, we extend the Kiefer and Wolfowitz (1956) nonparametric maximum likelihood estimator (NPMLE) to incorporate multivariate and heteroscedastic errors. An NPMLE is any $\widehat{G}_n \in \mathcal{P}(\mathbb{R}^d)$ which maximizes the marginal likelihood of the observations $(X_i)_{i=1}^n$. Marginally, the observations are independent, and the i^{th} observation X_i is distributed according to a Gaussian location mixture with density

$$f_{G^*,\Sigma_i}(x) \coloneqq \int \varphi_{\Sigma_i}(x - \theta) \, \mathrm{d}G^*(\theta), \qquad \text{for } x \in \mathbb{R}^d, \tag{4.3}$$

where $\varphi_{\Sigma_i}(x) := \frac{1}{\sqrt{\det(2\pi\Sigma_i)}} \exp\left(-\frac{1}{2}x^\mathsf{T}\Sigma_i^{-1}x\right)$ denotes the density of $\mathcal{N}(0, \Sigma_i)$. Hence an NPMLE is any maximizer

$$\widehat{G}_n \in \operatorname*{argmax}_{G \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{n} \sum_{i=1}^{n} \log f_{G,\Sigma_i}(X_i). \tag{4.4}$$

In contrast to the parametric model used in XD, the nonparametric domain $\mathcal{P}(\mathbb{R}^d)$ is convex, so $\widehat{G}_n$ solves a convex optimization problem, and tools from convex optimization may be leveraged to find principled approximations to $\widehat{G}_n$ (Kim et al., 2020; Koenker & Mizera, 2014).

Given an estimate $\widehat{G}_n$ of the prior G^*, empirical Bayes imitates the optimal Bayes analysis, known as the *oracle* (Efron, 2019). If G^* were known, optimal denoising of θ_i^* would be achieved through the posterior distribution $\theta_i^* \mid X_i$. It is well known, for instance, that the oracle posterior mean

$$\hat{\theta}_i^* \coloneqq \mathbb{E}_{G^*}\left[\theta_i^* \mid X_i\right], \text{ where } \theta_i^* \sim G^* \text{ and } X_i \mid \theta_i^* \sim \mathcal{N}(\theta_i^*, \Sigma_i) \tag{4.5}$$

minimizes the squared error Bayes risk

$$\mathbb{E}_{G^*}\|\mathfrak{d}_i(X_i) - \theta_i^*\|_2^2$$

over *all* measurable functions $\mathfrak{d}_i : \mathbb{R}^d \to \mathbb{R}^d$. The NPMLE (4.4) yields a fully data-driven, empirical Bayes estimate of the oracle posterior mean via

$$\hat{\theta}_i \coloneqq \mathbb{E}_{\widehat{G}_n}\left[\theta_i^* \mid X_i\right], \text{ where } \theta_i^* \sim \widehat{G}_n \text{ and } X_i \mid \theta_i^* \sim \mathcal{N}(\theta_i^*, \Sigma_i). \tag{4.6}$$

Figure 4.1 shows the $d = 2$ dimensional dataset of Anderson et al. (2018), where each observation has a known error distribution and may be modeled as multivariate normal after a suitable transformation. The noise in the raw CMD of Figure 4.1 obscures many known features of stellar evolution, rendering the raw CMD unreliable for downstream parallax inference. The right panel of Figure 4.1 displays the empirical Bayes posterior means $(\hat{\theta}_i)_{i=1}^n$ based on the NPMLE. The substantial shrinkage of our method reveals many recognizable features of the CMD, such as the red clump and a narrow red giant branch in the upper-right region of the plot, as well as the binary sequence tail distinct from the main sequence tail in the bottom-center region. The NPMLE $\widehat{G}_n$ and corresponding posterior means $(\hat{\theta}_i)_{i=1}^n$ offer a powerful approach to shrinkage estimation under minimal assumptions.

The idea of using the NPMLE to estimate a prior distribution, due to Robbins (1950), has seen a resurgence in recent years (Deb et al., 2021; Dicker & Zhao, 2016; Efron, 2019; Feng & Dicker, 2018; Gu & Koenker, 2016; Jiang, 2020; Jiang & Zhang, 2009, 2010; Kim et al., 2020; Koenker & Gu, 2017; Koenker & Mizera, 2014; Polyanskiy & Wu, 2020; Saha & Guntuboyina, 2020a). These advancements, taken together, have begun to establish the NPMLE as a formidable approach to shrinkage estimation both in theory and in practice. All this prior work has focused on either the univariate setting $d = 1$ or the homoscedastic

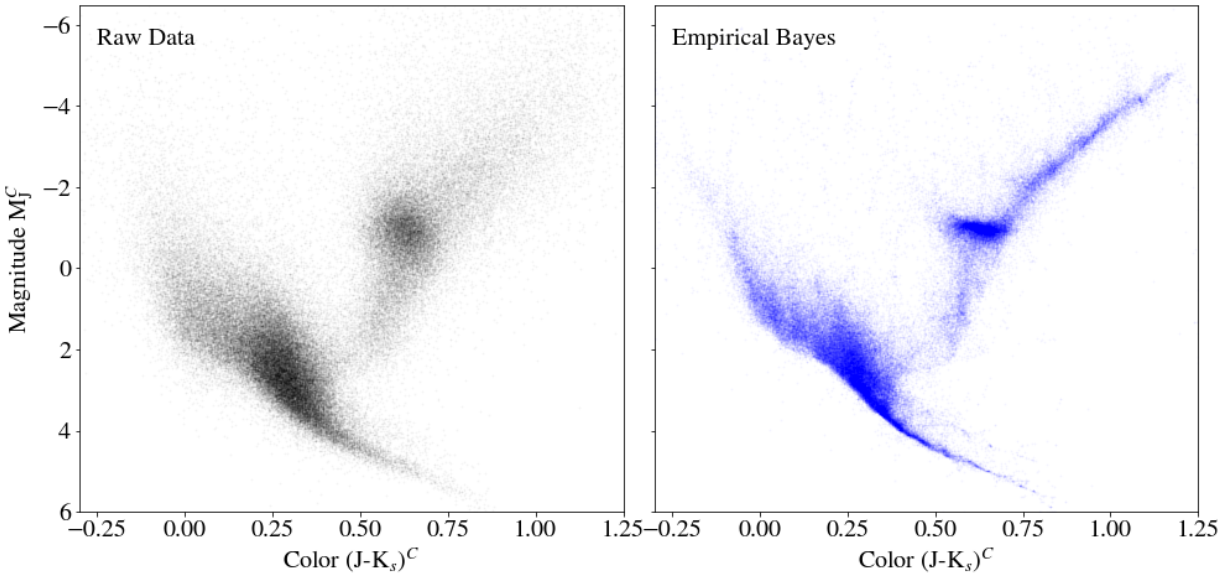


Figure 4.1: A noisy color-magnitude diagram (CMD) corresponding to the observations X_i in model (4.1), with corresponding fully-nonparametric denoised estimates $\hat{\theta}_i$ in the right panel. To avoid overplotting, we display a subsample of $n = 10^5$ stars.

setting $\Sigma_i \equiv \Sigma$, however. Our work extends the NPMLE to the practically important and more general setting of multivariate and heteroscedastic errors, uncovering a number of important differences.

Basic properties of the NPMLE that are well-understood in the univariate, homoscedastic setting (Lindsay, 1995) have not received careful attention in more complex settings. We verify in Lemma 4.1 that a solution $\widehat{G}_n$ exists for every instance of the optimization problem (4.4), and we record the first-order optimality conditions characterizing the solution set. Similar to the univariate, homoscedastic setting, there exists a solution $\widehat{G}_n$ which is discrete with at most n atoms, and the sequence of fitted values $\hat{L} \equiv (\hat{L}_1, \ldots, \hat{L}_n) = (f_{\widehat{G}_n, \Sigma_i}(X_i))_{i=1}^n$ is unique, i.e. every solution $\widehat{G}_n$ has the same sequence of fitted likelihood values $\hat{L}$.

An important contribution of Lemma 4.1 is our reinterpretation of the characterizing system of inequalities in terms of a natural 'dual' mixture density $\widehat{\psi}_n$. Specifically, $\widehat{\psi}_n$ is a heteroscedastic, n-component mixture density—a convex combination of Gaussian bumps centered at the datapoints $\mathcal{N}(X_i, \Sigma_i)$ with weights inversely proportional to $\hat{L}_i$ for $i = 1, \ldots, n$—such that the support of every NPMLE $\widehat{G}_n$ is contained in the set of the global maximizers of $\widehat{\psi}_n$. This observation has a number of important consequences that we explore in detail in Section 4.2; in particular, tools from algebraic statistics for studying the modes of Gaussian mixtures (Améndola et al., 2020; Ray & Lindsay, 2005) translate directly into results on

the support set. We leverage this connection to establish that $\widehat{G}_n$ is not necessarily unique when $d > 1$, even in the homoscedastic case. This finding is distinctive from the univariate, homoscedastic case where it is known that (4.4) has a unique solution for every problem instance (Lindsay & Roeder, 1993). Our counterexample in Lemma 4.2 appears to be new and seems to invalidate prior claims of strict concavity of the log-likelihood (Koenker & Gu, 2017; Marriott, 2002). Whereas the fitted values $\hat{L}$ are always unique, our counterexample also demonstrates that the empirical Bayes posterior means $(\hat{\theta}_i)_{i=1}^n$ are not necessarily unique. In light of the non-uniqueness of $\widehat{G}_n$, a natural question is whether there exist non-discrete solutions: we rule out this possibility in Corollary 4.3, however, showing every solution is indeed discrete with a finite number of atoms.

The problem of computing a solution $\widehat{G}_n$ is complicated by the presence of multivariate, heteroscedastic errors. The main difficulty in general is that the NPMLE solves an infinite-dimensional optimization problem. Since $\widehat{G}_n$ may be taken to be discrete with at most n atoms, a solution can in principle be found with a finite mixture model. In particular, defining the set of discrete distributions with at most $k \geq 1$ atoms,

$$\mathcal{P}_k(\mathbb{R}^d) = \left\{ \sum_{j=1}^k w_j \delta_{a_j} : \sum_j w_j = 1, w \geq 0, a_j \in \mathbb{R}^d, j = 1, \ldots, k \right\},$$

maximum likelihood solutions over $\mathcal{P}_n(\mathbb{R}^d)$ are also NPMLEs. Hence, the EM algorithm can be applied to optimize $(w_j, a_j)_{j=1}^n$, as first observed by Laird (1978), though EM over discrete distributions is prohibitively slow for moderately large n and suffers from the same nonconvexity issue as XD. Many algorithms (Böhning, 1985; Lesperance & Kalbfleisch, 1992; Liu & Zhu, 2007; Wang, 2007) have been proposed for finding approximate solutions to the optimization problem (4.4); Koenker and Mizera (2014) identified a convex, finite-dimensional, highly scalable approximation. Instead of maximizing the log-likelihood of the data $\frac{1}{n}\sum_{i=1}^n \log f_{G,\Sigma_i}(X_i)$ over $G \in \mathcal{P}_n(\mathbb{R}^d)$, the idea is to maximize the log-likelihood over $\mathcal{P}(\mathcal{A})$, the collection of all probability measures supported on a finite set $\mathcal{A} \subset \mathbb{R}^d$. If $\mathcal{A}$ has $m > 0$ elements, then $\mathcal{P}(\mathcal{A})$ is isometric to the $m-1$ dimensional simplex $\Delta_{m-1} := \{w \in \mathbb{R}_+^m : \sum_j w_j = 1\}$, and maximizing the likelihood corresponds to optimizing over the mixing proportions w, which is a convex optimization problem. When $d = 1$, it is straightforward to see that $\widehat{G}_n$ is supported on the range of the data $[X_{(1)}, X_{(n)}]$, so Koenker and Mizera (2014) proposed taking $\mathcal{A}$ to discretize this range. Jiang and Zhang (2009, Proposition 5) bounded the discretization error in $d = 1$ dimension, establishing that optimizing the weights w via EM can lead to a good approximation once $m \asymp (\log n)\sqrt{n}$. Dicker and Zhao (2016) further justified the discretization scheme in $d = 1$ dimension by showing the discretized NPMLE is statistically indistinguishable from $\widehat{G}_n$ once the analyst uses at least $m = \lfloor \sqrt{n} \rfloor$ atoms.

The discretization approach naturally extends to multivariate, heteroscedastic settings, but to our knowledge, no principled recommendations are available for choosing $\mathcal{A} \subset \mathbb{R}^d$ in general. Feng and Dicker (2018) recommended taking $\mathcal{A}$ to be a grid over a compact region containing the data. We address the key questions of how to choose this compact region and

how the discretization error depends on the fineness of the grid. For choosing a compact region to discretize, a natural desideratum is that the region should contain the support of $\widehat{G}_n$. To this end, in Corollary 4.3 we present compact support bounds on the NPMLE in terms of the data $(X_i, \Sigma_i)_{i=1}^n$. When $d = 1$ our support bounds reduce to the range of the data, reaffirming the original suggestion of Koenker and Mizera (2014), and when $d > 1$ but the errors are homoscedastic, it suffices to discretize the convex hull of $(X_i)_{i=1}^n$. Interestingly, with multivariate and heteroscedastic errors, the support of the NPMLE can lie outside the convex hull of $(X_i)_{i=1}^n$, so a different region known as the ridgeline manifold $\mathcal{M}$ of $(X_i, \Sigma_i)_{i=1}^n$ is needed. Fortunately, this region $\mathcal{M} \subset \mathbb{R}^d$ is compact, and the NPMLE over $\mathcal{P}(\mathcal{M})$ agrees with the NPMLE over $\mathcal{P}(\mathbb{R}^d)$. This justifies the choice of $\mathcal{A}$ as a $\delta > 0$ cover of $\mathcal{M}$, and in Proposition 4.5, we verify that as $\delta \downarrow 0$, the log-likelihood of the discretized NPMLE approaches that of the NPMLE. We prove a quantitative bound on the gap for fixed δ, providing some guidance on how the discretization error depends on the fineness of the grid.

Our principled and efficient method of computation facilitates simulation studies assessing the performance of the empirical Bayes estimate $\hat{\theta}_i$ in a setting where we can actually compare to the oracle Bayes estimate $\hat{\theta}_i^*$. Figure 4.2 illustrates the method on simulated data. The means θ_i^* were drawn i.i.d. from a circle of radius two, and the data $X_i \mid \theta_i^*$ were drawn according to (4.1) using a variety of diagonal covariance matrices $\Sigma_i = \begin{bmatrix} \sigma_{1,i}^2 & 0 \\ 0 & \sigma_{2,i}^2 \end{bmatrix}$, taking each $\sigma_{j,i}^2 \in (1/2, 3/4)$. Visually, it is clear that the empirical Bayes estimates improve upon the observations by shrinking towards the underlying circle; the corresponding mean squared errors were $\frac{1}{n}\sum_{i=1}^n \|\hat{\theta}_i - \theta_i^*\|_2^2 = 0.87$ and $\frac{1}{n}\sum_{i=1}^n \|X_i - \theta_i^*\|_2^2 = 1.46$, respectively. The oracle, which minimizes the mean squared error in expectation, attained an error of $\frac{1}{n}\sum_{i=1}^n \|\hat{\theta}_i^* - \theta_i^*\|_2^2 = 0.84$. While the oracle cannot be computed in practice because G^* is unknown, this value sets a benchmark in simulations to which we may compare the performance of bona fide estimators. The empirical Bayes estimates not only track well with this benchmark; the individual estimates also track remarkably well with the oracle. In our simulation, the *regret*—defined as the mean squared error between the estimator $(\hat{\theta}_i)_{i=1}^n$ and the oracle $(\hat{\theta}_i^*)_{i=1}^n$—was $\frac{1}{n}\sum_{i=1}^n \|\hat{\theta}_i - \hat{\theta}_i^*\|_2^2 = 0.03$. Whereas $\hat{\theta}_i$ is a function of the observed data, the oracle $\hat{\theta}_i^*$ makes optimal use of the unknown prior G^*, making the similarity between the two especially striking.

This striking similarity between $\hat{\theta}_i$ and $\hat{\theta}_i^*$ affirms the empirical Bayes adage that "*large data sets of parallel situations carry within them their own Bayesian information*" (Efron & Hastie, 2016). However, the setting of Figure 4.2 is complicated by the fact the situations are not directly parallel, in that each observation X_i has a distinct error distribution. Even in heteroscedastic settings, the extent to which we glean prior information for the purpose of denoising is captured by the empirical Bayes regret $\frac{1}{n}\sum_{i=1}^n \|\hat{\theta}_i - \hat{\theta}_i^*\|_2^2$. Theorem 4.8 develops a detailed profile of the finite-sample regret properties of the NPMLE for denoising. We show that under certain tail conditions on G^* the regret is bounded by a rate that is nearly parametric in n, i.e. $\frac{1}{n}$ up to logarithmic multiplicative factors. The regret still converges at a slower, nonparametric rate under less structured conditions, where G^* may have heavy tails.

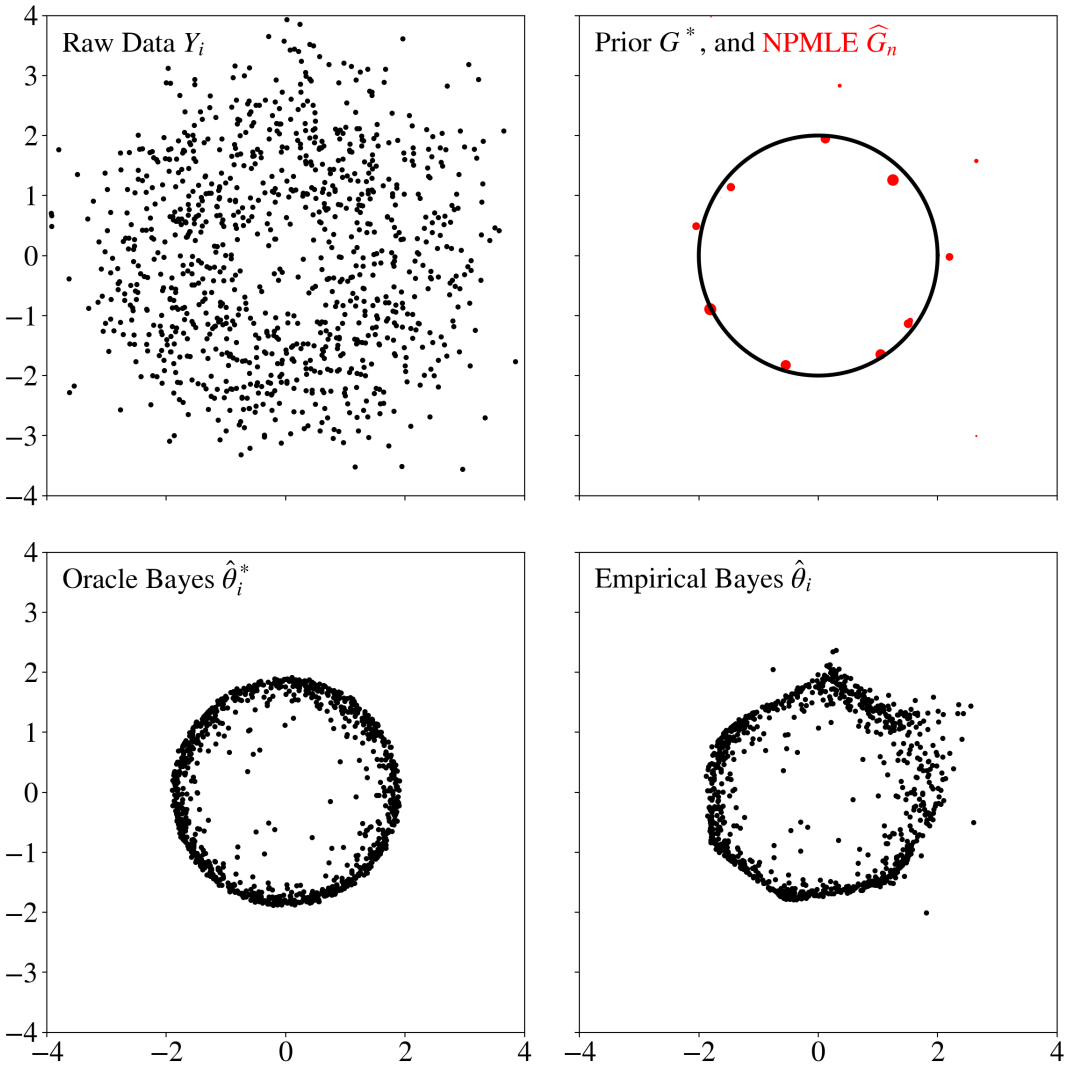


Figure 4.2: Toy data of size $n = 1,000$ and $d = 2$. Top: observations X_i (left) were generated by adding heteroscedastic Gaussian errors to the underlying means $\theta_i^* \overset{\text{iid}}{\sim} G^*$ (right), generated IID uniformly from a circle of radius 2. Our discrete estimate $\widehat{G}_n$ of the prior is shown in red over the prior G^* in black. Bottom: a comparison of oracle Bayes $\hat{\theta}_i^*$ (left) based on knowledge of the prior distribution G^* and empirical Bayes $\hat{\theta}_i$ (right), a function of the observed data.

Furthermore, when G^* possesses finer structure, such as the clustering problem where G^* is a discrete measure with k^* atoms, we prove that the regret is bounded from above by $\frac{k^*}{n}$ up to logarithmic multiplicative factors in n. The clustering case is particularly remarkable, as

the NPMLE is completely tuning-free, with no knowledge of k^*, yet $\widehat{G}_n$ performs essentially as well as any estimator which knows the number of clusters k^*. Thus, Theorem 4.8 demonstrates that the NPMLE effectively discovers structure when available and also effectively learns when structure is unavailable. Theorem 4.8 generalizes the regret bounds of Saha and Guntuboyina (2020a) and Jiang (2020) who analyzed the homoscedastic $\Sigma_i \equiv \Sigma$ setting and the univariate $d = 1$ setting, respectively. These papers in turn built upon Jiang and Zhang (2009) who studied the univariate, homoscedastic setting.

A key ingredient in the analysis of the regret is a more explicit representation of the estimator $(\hat{\theta}_i)_{i=1}^n$ and oracle $(\hat{\theta}_i^*)_{i=1}^n$. The oracle posterior mean (4.5) has the following alternative expression, known as Tweedie's formula (Banerjee et al., 2021; Dyson, 1926; Efron, 2011; Robbins, 1956):

$$\hat{\theta}_i^* = X_i + \Sigma_i \frac{\nabla f_{G^*,\Sigma_i}(X_i)}{f_{G^*,\Sigma_i}(X_i)}. \tag{4.7}$$

Similarly, our plug-in estimate can be written as

$$\hat{\theta}_i = X_i + \Sigma_i \frac{\nabla f_{\widehat{G}_n,\Sigma_i}(X_i)}{f_{\widehat{G}_n,\Sigma_i}(X_i)}. \tag{4.8}$$

Tweedie's formula clarifies that under model (4.1) the posterior means only depend on the prior G^* via the marginal likelihood $f_{G^*,\Sigma_i}(X_i)$ and its gradient. Jiang and Zhang (2009) first leveraged this observation to relate the empirical Bayes regret to the problem of estimating the marginal density. In heteroscedastic problems, there are n different marginal densities, $(f_{G^*,\Sigma_i})_{i=1}^n$, to estimate, and corresponding estimators $(f_{\widehat{G}_n,\Sigma_i})_{i=1}^n$. We show in Theorem 4.6 and Corollary 4.7 that the NPMLE achieves similar adaptive rates in the density estimation problem under an appropriate average Hellinger distance across all $i = 1, \dots, n$ estimands $(f_{G^*,\Sigma_i})_{i=1}^n$.

Whereas most recent work has focused on properties of $\widehat{G}_n$ for density estimation and denoising, the NPMLE is potentially much more generally applicable as a plug-in estimate of the prior. To expand our understanding of its applicability, we present the first analysis of the deconvolution error for the NPMLE. Whereas density estimation captures the problem of describing the observations $(X_i)_{i=1}^n$, deconvolution is the equally natural problem of interpreting the infinite-dimensional parameter G^*. We study the accuracy of the NPMLE under a Wasserstein distance $W_2(\widehat{G}_n, G^*)$. The Wasserstein distance is particularly useful for this problem since $\widehat{G}_n$ and G^* are typically mutually singular; in particular, G^* may be absolutely continuous whereas $\widehat{G}_n$ is always discrete. The Wasserstein distance will be discussed in detail in Section 4.3. We show in Theorem 4.10 that $\widehat{G}_n$ attains the minimax rate of deconvolution, which happens to be a very slow, logarithmic rate $\frac{1}{\log n}$. Inspired by the richness of the density estimation and denoising results, we hint at some of the adaptation properties of the NPMLE under the Wasserstein loss; Theorem 4.12 shows that when $G^* = \delta_\mu$ is a point mass distribution, the Wasserstein rate improves dramatically to $n^{-1/4}$ up to logarithmic factors.

The rest of the chapter is organized as follows: Section 4.2 systematically addresses basic properties of the NPMLE, including existence, discreteness, and non-uniqueness; Section 4.2.2 gives a full account of the approximate computation of NPMLEs. Section 4.3 establishes finite-sample risk bounds on the accuracy of $\widehat{G}_n$ as an estimator of G^* for the purposes of density estimation, denoising and deconvolution. In Section 4.4, we apply the method to astronomy data to construct a fully data driven color-magnitude diagram of 1.4 million stars and compare our method to extreme deconvolution where it has previously been applied (Anderson et al., 2018). We also apply the method to chemical abundance data for a smaller subset of stars that has previously been analyzed by Ratcliffe et al. (2020). Section 4.5 concludes with some discussion of future work. The proofs are in Section 4.6.

4.2 Computational properties

4.2.1 Characterization and basic properties

In this section, we establish some basic properties of solutions to the nonparametric maximum likelihood problem (4.4), including existence, non-uniqueness, discreteness of solutions $\widehat{G}_n$, invariance under certain transformations, and bounds on the support. These results provide a foundation both for computing $\widehat{G}_n$ (Section 4.2.2) and for understanding its statistical properties (Section 4.3). Our first result extends the well-known characterization of $\widehat{G}_n$ for univariate, homoscedastic errors (Lindsay, 1995, Theorems 18-21) to our more general setting.

Lemma 4.1. *Problem* (4.4) *attains its maximum: there exists a discrete solution* $\widehat{G}_n$ *with at most* n *atoms, and the vector* $\hat{L} \equiv (\hat{L}_1, \ldots, \hat{L}_n) = (f_{\widehat{G}_n, \Sigma_i}(X_i))_{i=1}^n$ *of fitted likelihood values is unique. Moreover,* $\widehat{G}_n \in \mathcal{P}(\mathbb{R}^d)$ *solves* (4.4) *if and only if*

$$D(\widehat{G}_n, \vartheta) \leq 0 \text{ for all } \vartheta \in \mathbb{R}^d, \text{ where } D(G, \vartheta) \coloneqq \frac{1}{n} \sum_{i=1}^n \frac{\varphi_{\Sigma_i}(X_i - \vartheta)}{f_{G,\Sigma_i}(X_i)} - 1.$$

The support of any $\widehat{G}_n$ *is contained in the zero set* $\mathcal{Z} \coloneqq \{\vartheta : D(\widehat{G}_n, \vartheta) = 0\}$*; the zero set* $\mathcal{Z}$ *is equal to the set of global maximizers of the* n*-component, heteroscedastic* dual *mixture density*

$$\widehat{\psi}_n(\vartheta) \coloneqq \sum_{i=1}^n \left(\frac{\hat{L}_i^{-1}}{\sum_{\iota=1}^n \hat{L}_\iota^{-1}} \right) \varphi_{\Sigma_i}(X_i - \vartheta).$$

We prove Lemma 4.1, along with all results in this section, in Section 4.6.1. The first statement of the lemma guarantees the existence of a discrete solution, which we typically write as $\widehat{G}_n = \sum_{j=1}^{\hat{k}} \hat{w}_j \delta_{\hat{a}_j}$ (here $\hat{w}_j \geq 0$, $\sum_j \hat{w}_j = 1$ and $\hat{a}_j \in \mathbb{R}^d$), with $\hat{k} \leq n$ providing an upper bound on the complexity of at least one solution. This implies that $\widehat{G}_n$ may be taken to be the maximum likelihood solution to a $\hat{k}$-component, heteroscedastic Gaussian

mixture model where $\hat{k}$ is selected in a data dependent manner. Since finite mixture models are nested by the number of components and $\hat{k} \leq n$, we may also say in general that $\widehat{G}_n$ is the maximum likelihood solution to an n-component, heteroscedastic Gaussian mixture model.

The bound $\hat{k} \leq n$ is tight: for each $n \geq 1$, there are sequences of observations $(X_i)_{i=1}^n$ and covariances $(\Sigma_i)_{i=1}^n$ such that the smallest number of components $\hat{k}$ of any solution $\widehat{G}_n$ to (4.4) is precisely n (see, e.g., Lindsay, 1995, p. 116). However, in practice, the number of components is typically much smaller than n. For instance, in the univariate, homoscedastic case, Polyanskiy and Wu (2020) established a much stronger bound of $\hat{k} = O_P(\log n)$ under certain conditions on the prior distribution G^*.

The last part of Lemma 4.1 states that the atoms of $\widehat{G}_n$ occur at the global maximizers of the n-component Gaussian mixture $\widehat{\psi}_n$, which has component distributions of the form $\mathcal{N}(X_i, \Sigma_i)$ for $i = 1, \ldots, n$ with weights inversely proportional to fitted likelihoods $\hat{L}$. Results on the modes of Gaussian mixtures (e.g. Améndola et al., 2020; Dytso et al., 2019; Ray & Lindsay, 2005) thus provide information about the support of the NPMLE; in particular, our next two results exploit this connection to yield novel results on the NPMLE.

In the univariate $d = 1$ and homoscedastic setting $\Sigma_i \equiv \sigma^2$, it is additionally known that (4.4) has a *unique* solution $\widehat{G}_n$ for all observations $X_1, \ldots, X_n$ (Lindsay & Roeder, 1993). This means that, for every dataset $X_1, \ldots, X_n$ and every variance level $\sigma^2 > 0$, there is a unique probability measure $\widehat{G}_n \in \mathcal{P}(\mathbb{R})$ such that $\hat{L}_i = f_{\widehat{G}_n, \sigma^2}(X_i)$ for all i, where $\hat{L}$ is the unique vector of optimal likelihoods from Lemma 4.1. We observe, however, that uniqueness of the solution $\widehat{G}_n$ may not hold when $d > 1$, even with isotropic covariances $\Sigma_i \equiv \sigma^2 I_d$.

Lemma 4.2. *Let $d = 2$, $n = 3$ and $X_1 = (0, 1)$, $X_2 = (\frac{\sqrt{3}}{2}, -\frac{1}{2})$, $X_3 = (-\frac{\sqrt{3}}{2}, -\frac{1}{2})$. Then (4.4) with data $(X_i)_{i=1}^3$, covariances $\Sigma_i \equiv \sigma^2 I_2$ and $\sigma^2 = 3/(\log 256)$ has infinitely many solutions of the form*

$$\widehat{G}_n = \alpha \delta_0 + (1 - \alpha) \frac{1}{3} \sum_{i=1}^{3} \delta_{X_i/2}$$

where $\alpha \in [0, 1]$.

Figure 4.3 illustrates the counterexample given in Lemma 4.2. A key observation in the proof of Lemma 4.2 is that the dual mixture $\widehat{\psi}_n = f_{H, \sigma^2 I_2}$ can be written explicitly as a homoscedastic mixture with uniform mixing distribution $H = \frac{1}{3} \sum_{i=1}^3 \delta_{X_i}$ over the observations $(X_i)_{i=1}^3$. This set-up closely follows a construction, due to Duistermaat (see Améndola et al., 2020), exhibiting an isotropic, homoscedastic Gaussian mixture with more modes than components. Duistermaat used the same component locations X_i but took $\sigma^2 = 0.53$ to obtain an example of a three-component mixture of isotropic, homoscedastic Gaussians such that the mixture has four modes. By specifically choosing $\sigma^2 = \frac{3}{\log 256} \approx 0.54$, the height of the mixture $\widehat{\psi}_n = f_{H, \sigma^2 I_2}$ is equal at all four modes, i.e. all four modes are global maximizers, and the modes are located at $\{X_1/2, X_2/2, X_3/2, 0\}$. By Lemma 4.1 any NPMLE must be

supported on these modes. Representing the fitted values $\hat{L} = (f_{\widehat{G}_n,\sigma^2 I_2}(X_i))_{i=1}^3$ by a probability measure $\widehat{G}_n = \sum_{j=1}^3 \hat{w}_j \delta_{X_j/2} + \hat{w}_4 \delta_0$ supported on the global modes is equivalent to finding a set of weights $\hat{w} \in \mathbb{R}_+^4$ such that $\sum_{j=1}^4 \hat{w}_j = 1$ and $\hat{w}$ solves the under-determined linear system $\hat{L} = A\hat{w}$, where A is a 3×4 matrix given by

$$A_{ij} = \begin{cases} \varphi_{\sigma^2 I_2}(X_i - X_j/2) & j \leq 3 \\ \varphi_{\sigma^2 I_2}(X_i) & j = 4. \end{cases}$$

Finally, we also note that although the fitted likelihoods $f_{\widehat{G}_n,\sigma^2 I_2}(X_i)$ are unique, the posterior means $\hat{\theta}_i$ in this example differ for the solutions $\widehat{G}_n$ given in Lemma 4.2.

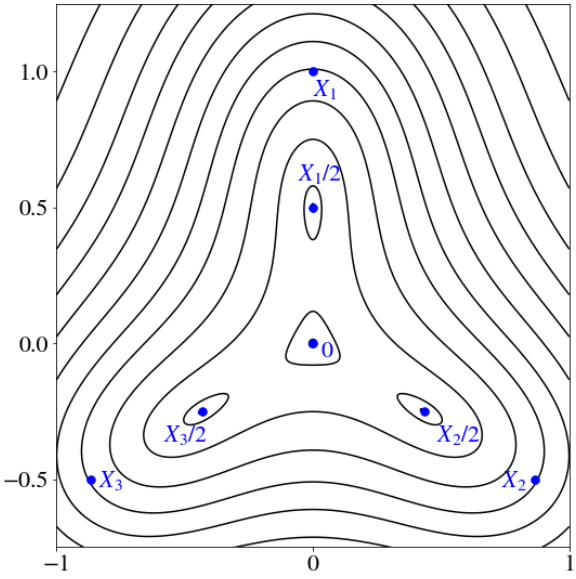


Figure 4.3: Level sets of the dual mixture density $\widehat{\psi}_n = f_{H,\sigma^2 I_2}$ where $n = 3$ and $H = \frac{1}{3}\sum_{i=1}^3 \delta_{X_i}$ is uniform over the vertices of the larger equilateral triangle $\triangle X_1 X_2 X_3$. With $\sigma^2 = \frac{3}{\log 256}$, the dual mixture density $\widehat{\psi}_n$ has four global modes.

Although the NPMLE searches over all probability measures $G \in \mathcal{P}(\mathbb{R}^d)$ supported on $\mathbb{R}^d$, it is useful algorithmically to reduce the search space to probability measures supported on a compact subset of $\mathbb{R}^d$. By Lemma 4.1, to restrict the support of the NPMLE it suffices to bound the maximizers $\mathcal{Z}$ of the n-component Gaussian mixture $\widehat{\psi}_n$. Ray and Lindsay (2005, Theorem 1) showed that all critical points of a Gaussian mixture $\widehat{\psi}_n(\vartheta) =$

$\sum_{i=1}^n \left(\frac{\hat{L}_i^{-1}}{\sum_{\iota=1}^n \hat{L}_\iota^{-1}}\right) \varphi_{\Sigma_i}(X_i - \vartheta)$ belong to the ridgeline manifold

$$\begin{aligned} \mathcal{M} &\coloneqq \left\{ x^*(\alpha) : \alpha \in \mathbb{R}_+^n, \sum_{i=1}^n \alpha_i = 1 \right\}, \text{ where} \\ x^*(\alpha) &\coloneqq \left(\sum_{i=1}^n \alpha_i \Sigma_i^{-1} \right)^{-1} \sum_{i=1}^n \alpha_i \Sigma_i^{-1} X_i. \end{aligned} \tag{4.9}$$

In general, the ridgeline manifold $\mathcal{M}$ is a compact subset of $\mathbb{R}^d$ which does not depend on the weights $\left(\frac{\hat{L}_i^{-1}}{\sum_{\iota=1}^n \hat{L}_\iota^{-1}}\right)_{i=1}^n$. In the univariate case $d = 1$, the rigdeline manifold $\mathcal{M} = [X_{(1)}, X_{(n)}]$ is simply the range of the data, so the univariate NPMLE is constrained to be supported on this range. In the multivariate setting, we may further simplify $\mathcal{M}$ depending on certain shape restrictions on the covariance matrices.

Corollary 4.3. *Every solution to* (4.4) *is discrete with a finite number of atoms, supported on the ridgeline manifold* $\mathcal{M}$ *defined in* (4.9)*. Depending on the values of* (Σ_i) *we further bound the support as follows:*

(i) (Homoscedastic) If $\Sigma_i = \Sigma$ *for all* i*, or if* $\Sigma_i = c_i\Sigma$ *are proportional up to a sequence* (c_i) *of positive scalars, the ridgeline manifold* $\mathcal{M}$ *is the convex hull of the data* $\operatorname{conv}(\{X_1, \ldots, X_n\})$*.*

(ii) (Diagonal Covariances) If Σ_i *is a diagonal matrix for every* i*, the ridgeline manifold* $\mathcal{M}$ *is contained in the axis-aligned minimum bounding box of the data*

$$\prod_{j=1}^d \left[\min_{i \in \{1,\ldots,n\}} X_{ij}, \max_{i \in \{1,\ldots,n\}} X_{ij} \right],$$

where $X_i = (X_{i1}, \ldots, X_{id})$ *for all* i*.*

(iii) (General Covariances) Let $\overline{k} \geq \underline{k} > 0$ *be chosen such that* $\underline{k} I_d \preceq \Sigma_i \preceq \overline{k} I_d$ *for all* i*, where* $A \preceq B$ *means* $B - A$ *is a symmetric positive semidefinite matrix. Choose* $r > 0$ *and* $x_0 \in \mathbb{R}^d$ *such that* $\|X_i - x_0\|_2 \leq r$ *for all* i*. Then the ridgeline manifold* $\mathcal{M}$ *is contained in the ball*

$$\mathbb{B}_{\kappa r}(x_0) \coloneqq \left\{ y \in \mathbb{R}^d : \|y - x_0\|_2 \leq \kappa r \right\}$$

where $\kappa = \overline{k}/\underline{k}$*.*

The first part of Corollary 4.3 in general gives the smallest possible convex body over which the support of $\widehat{G}_n$ can be constrained independently of $\{\Sigma_i\}$. To see that the first part is tight, consider a fixed set of observations $(X_i)_{i=1}^n$ and isotropic covariance matrices

$\Sigma = \sigma^2 I_d$; as σ is made arbitrarily small, the support of $\widehat{G}_n$ approaches the set of observations $(X_i)_{i=1}^n$ (Lindsay, 1995). Therefore, in general the convex hull is the smallest convex body containing the support in the homoscedastic setting and more generally the setting of proportional covariance matrices. By contrast, the convex hull of the data is in general too small to capture the support of $\widehat{G}_n$ in the heteroscedastic setting. Figure 4.4 presents one example with diagonal covariances where the support of $\widehat{G}_n$ is pushed towards the corners of the minimum axis-aligned bounding box of the data. Thus, the above discussion and Figure 4.4 indicate that both parts (i) and (ii) of Corollary 4.3 give the tightest possible convex support bounds in their respective special cases.

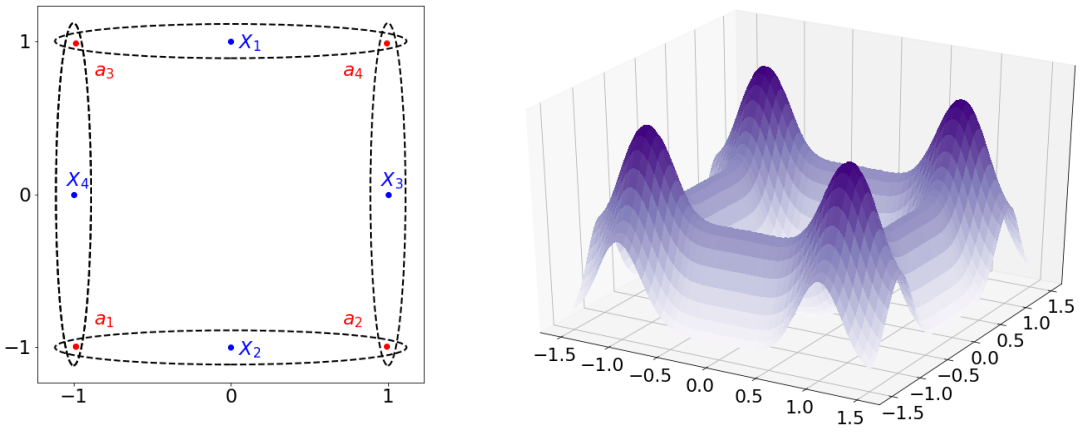


Figure 4.4: Left: An example of observations $X_1 = (0, 1)$, $X_2 = (0, -1)$, $X_3 = (1, 0)$, and $X_4 = (-1, 0)$ (blue points) with diagonal covariances $\Sigma_1 = \Sigma_2 = \begin{bmatrix} 5 & 0 \\ 0 & .05 \end{bmatrix}$ and $\Sigma_3 = \Sigma_4 = \begin{bmatrix} .05 & 0 \\ 0 & 5 \end{bmatrix}$ (dashed ellipses), where the NPMLE is supported on atoms $a_1, \ldots, a_4$ (red points) well outside the convex hull of the data, and near the corners of the minimum axis-aligned bounding box. Right: The mixture $\widehat{\psi}_n(\vartheta) = \frac{1}{4}\sum_{i=1}^4 \varphi_{\Sigma_i}(X_i - \vartheta)$ only has modes at the atoms $a_1, \ldots, a_4$, so no NPMLE is supported within the convex hull of the data.

We close this section with a brief discussion on how the NPMLE behaves under certain simple transformations of the data $(X_i, \Sigma_i)_{i=1}^n$. Given a map $T : \mathbb{R}^d \to \mathbb{R}^d$, let $T_\# G \in \mathcal{P}(\mathbb{R}^d)$ denote the pushforward of $G \in \mathcal{P}(\mathbb{R}^d)$ given by $T_\# G(B) = G(T^{-1}(B))$, for any Borel set $B \subseteq \mathbb{R}^d$. In other words, if $V \sim G$, then $T_\# G$ is the distribution of $T(V)$.

Lemma 4.4. *Fix a dataset $(X_i, \Sigma_i)_{i=1}^n$, a point $x_0 \in \mathbb{R}^d$ and a $d \times d$ orthogonal matrix U_0. Consider the transformed dataset $(X_i', \Sigma_i')_{i=1}^n$ where $\Sigma_i' = U_0 \Sigma_i U_0^\mathsf{T}$ and $X_i' = T(X_i)$ for*

$i = 1, \ldots, n$, with $T(x) = U_0 x + x_0$. Then

$$f_{T_\# G, \Sigma_i'}(X_i') = f_{G,\Sigma_i}(X_i)$$

for all $i = 1, \ldots, n$ and all $G \in \mathcal{P}(\mathbb{R}^d)$.

Lemma 4.4 is a straightforward consequence of the change of variables formula, but it has a number of useful corollaries. In particular, if $\widehat{G}_n \in \mathcal{P}(\mathbb{R}^d)$ is an NPMLE for the dataset $(X_i, \Sigma_i)_{i=1}^n$, then $T_\# \widehat{G}_n$ is an NPMLE for the modified dataset $(X_i', \Sigma_i')_{i=1}^n$, and the fitted likelihood values are the same, i.e.

$$f_{T_\# \widehat{G}_n, \Sigma_i'}(X_i') = f_{\widehat{G}_n, \Sigma_i}(X_i),$$

for all $i = 1, \ldots, n$. Thus, an NPMLE $\widehat{G}_n = \sum_{j=1}^{\hat{k}} \hat{w}_j \delta_{\hat{a}_j}$ is equivariant under translations $T(y) = y + x_0$: if every observation is shifted by some fixed $x_0 \in \mathbb{R}^d$, then the modified NPMLE $T_\# \widehat{G}_n = \sum_{j=1}^{\hat{k}} \hat{w}_j \delta_{\hat{a}_j + x_0}$ simply shifts every atom by x_0. Similarly, the NPMLE is equivariant under orthogonal transformations, which explains why the fitted likelihood values are all equal in the rotationally symmetric toy datasets presented in Figure 4.3 and Figure 4.4.

4.2.2 Grid approximation

The NPMLE solves a convex optimization problem (4.4) that is *infinite-dimensional* in the sense that the decision variable G ranges over all probability measures on $\mathbb{R}^d$. Many numerical methods for approximately computing the NPMLE have been considered—including EM (Laird, 1978), vertex direction and exchange methods (Böhning, 1985), semi-infinite methods (Lesperance & Kalbfleisch, 1992), constrained-Newton methods (Wang, 2007), and hybrid methods (Böhning, 2003; Liu & Zhu, 2007)—typically described for the special case of univariate and homoscedastic errors. In this section, we discuss our strategy for computing the NPMLE as well as the challenges of scaling the computation to large datasets.

We follow the approach of Koenker and Mizera (2014), who approximated the infinite-dimensional problem by constraining the support of G to a large finite set. For a nonempty, closed set $\mathcal{A} \subseteq \mathbb{R}^d$, define a support-constrained NPMLE as any solution

$$\widehat{G}_n^{\mathcal{A}} \in \operatorname*{argmax}_{G \in \mathcal{P}(\mathcal{A})} \frac{1}{n} \sum_{i=1}^n \log f_{G,\Sigma_i}(X_i), \tag{4.10}$$

where $\mathcal{P}(\mathcal{A})$ denotes the set of probability measures supported on $\mathcal{A}$. In particular, $\widehat{G}_n = \widehat{G}_n^{\mathbb{R}^d}$ by definition, and by Corollary 4.3 we may write $\widehat{G}_n = \widehat{G}_n^{\mathcal{M}}$ for a compact subset $\mathcal{M}$ defined explicitly in terms of the data.

We now describe our strategy for choosing the discretization set $\mathcal{A}$. Fix $\delta > 0$. Let $\mathcal{H}$ denote a covering of $\mathcal{M}$ by closed hypercubes of width δ, i.e.

$$\mathcal{H} = \left\{ x_j + [-\delta/2, \delta/2]^d : j \in \{1, \ldots, J\} \right\}$$

for some set of points $x_1, \ldots, x_J \in \mathbb{R}^d$ such that $\mathcal{M} \subseteq \bigcup_{j=1}^J \left(x_j + [-\delta/2, \delta/2]^d\right)$. Now define the discretized support $\mathcal{A}$ to be the set of corners of hypercubes in $\mathcal{H}$; specifically, for each hypercube $x_j + [-\delta/2, \delta/2]^d$ in $\mathcal{H}$, the point $x_j + \frac{\delta}{2}v \in \mathcal{A}$ for every $v \in \{-1, 1\}^d$. Because $\mathcal{M}$ is compact, $\mathcal{A}$ is a finite set which we denote by $\{a_j\}_{j=1}^m$. Constraining the NPMLE to this finite set of atoms $a_1, \ldots, a_m$ yields a finite-dimensional convex optimization problem over the mixing proportions. That is, the solution to (4.10) can be written as $\widehat{G}_n^{\mathcal{A}} = \sum_{j=1}^m \tilde{w}_j \delta_{a_j}$, where

$$\tilde{w} \in \operatorname*{argmax}_{w \in \Delta_{m-1}} \frac{1}{n} \sum_{i=1}^n \log \left(\sum_{j=1}^m L_{ij} w_j \right), \tag{4.11}$$

and $L_{ij} = \varphi_{\Sigma_i}(X_i - a_j)$ encodes an $n \times m$ kernel matrix. The EM algorithm (Dempster et al., 1977) can be used to optimize directly over the mixing proportions $\tilde{w}$. While this approach was advocated by Lashkari and Golland (2008) and Jiang and Zhang (2009), EM can be prohibitively slow (Koenker & Mizera, 2014; Redner & Walker, 1984). A crucial observation made by Koenker and Mizera (2014) is that (4.11) is a (finite-dimensional) convex optimization problem, enabling the use of a wide array of tools from modern convex optimization; they proposed solving the dual to (4.11) using an interior point solver, and Koenker and Gu (2017) provided an R implementation to solve univariate problems. Kim et al. (2020) proposed sequential quadratic programming to solve a variant of the primal problem directly, demonstrating superior scalability with the sample size n. Our implementation uses the **MOSEK** library (MOSEK ApS, 2019) for Python.

To justify the grid approximation, some consideration of the discretization error is warranted. Our next result shows that as $\delta \downarrow 0$, the log-likelihood of the discretized NPMLE approaches that of the (unconstrained) NPMLE; moreover, the bound on the gap depends on known quantities, so it can be used to guide a suitable choice of δ.

Proposition 4.5. *Let $\mathcal{M} \subset \mathbb{R}^d$ denote any compact set such that every solution (4.4) is supported on $\mathcal{M}$. Suppose the diameter of the set $\mathcal{M}$ is at most D, the minimum eigenvalue of each Σ_i is at least $\underline{k}$, and fix $\delta \in \left(0, \sqrt{\frac{3}{4d}}\underline{k}D^{-1}\right)$. Let $\mathcal{H}$ denote a cover of $\mathcal{M}$ by closed hypercubes of width δ, and let $\mathcal{A}$ denote the set of corners of hypercubes in $\mathcal{H}$. Every approximate NPMLE $\widehat{G}_n^{\mathcal{A}}$ satisfies*

$$\sup_{G \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{n} \sum_{i=1}^n \log f_{G,\Sigma_i}(X_i) - \frac{1}{n} \sum_{i=1}^n \log f_{\widehat{G}_n^{\mathcal{A}},\Sigma_i}(X_i) \leq d\underline{k}^{-2} \left(2D^2 + \frac{1}{2}\right) \delta^2. \tag{4.12}$$

We prove Proposition 4.5 in Section 4.6.1. Proposition 4.5 shows that we can tractably approximate the NPMLE via a finite-dimensional, convex optimization problem. As we show in Section 4.3, our theoretical results on the statistical properties of the NPMLE hold for any approximate solution $\widehat{G}_n^{\mathcal{A}}$ which places nearly as high likelihood on the observations as the global optimizer, in the sense of (4.12). Hence for δ sufficiently small we can guarantee that the discretization error is negligible.

Dicker and Zhao (2016) showed in the univariate, homoscedastic case that a finely discretized NPMLE is statistically indistinguishable from the NPMLE for the purpose of density estimation. However, their analysis of the discretization error makes use of the modeling assumptions (4.1) and is statistical in nature, so their theoretical results provide little guidance on how much error is incurred due to discretization for a fixed dataset. Our result aligns more closely with and in fact essentially generalizes Jiang and Zhang (2009, Proposition 5), which bounded the optimality gap for a particular algorithm, discretization scheme and fixed dataset. The main difference between our result and Jiang and Zhang (2009, Proposition 5) is that the latter analyzed the EM algorithm for the mixing proportions (4.11), whereas by using a black-box, second-order optimization method to solve for the mixing proportions $\tilde{w}$, we can solve for the discretized NPMLE $\widehat{G}_n^{\mathcal{A}}$ much more accurately.

4.3 Statistical properties

The NPMLE $\widehat{G}_n$ applies as a plug-in estimator of the prior distribution G^* for many purposes. The traditional statistical setting is density estimation, where working in a Gaussian mixture model greatly simplifies the problem of estimating the marginal density of each observation X_i. In particular, $f_{\widehat{G}_n,\Sigma_i}$ is a natural, tuning-free estimate of the true marginal density f_{G^*,Σ_i}. Another problem setting—at the heart of empirical Bayes methodology—is to imitate the Bayesian inference we would conduct if we knew G^*. Denoising, using $(\hat{\theta}_i)_{i=1}^n$ as plug-in estimators of the true posterior means $(\hat{\theta}_i^*)_{i=1}^n$, represents the most basic instantiation. Finally, often we wish to compare $\widehat{G}_n$ to the prior G^* directly. Since we are estimating the prior given observations from a convolution model $X_i \overset{\text{ind}}{\sim} f_{G^*,\Sigma_i}$, deconvolution refers to the problem of estimating G^*.

In this section, we establish that the NPMLE is well-suited for all three disparate targets of estimation: the marginal densities $(f_{G^*,\Sigma_i})_{i=1}^n$, the oracle posterior means $(\hat{\theta}_i^*)_{i=1}^n$ and the prior G^*. In this section, we allow for the possibility that $\widehat{G}_n$ is an approximate NPMLE, with the exact conditions being given in each theorem. Throughout this section, we use the standard notation $X \lesssim_{p,q} Y$ to mean $X \leq C_{p,q}Y$ for some positive constant $C_{p,q} > 0$ depending only on problem parameters p, q.

4.3.1 Density estimation: average Hellinger accuracy

As the distribution of X_i varies with i, we consider the density estimation quality of the NPMLE (4.4) in terms of the average squared Hellinger distance, i.e. for $G, H \in \mathcal{P}(\mathbb{R}^d)$,

$$\bar{h}^2(f_{G,\bullet}, f_{H,\bullet}) \coloneqq \frac{1}{n}\sum_{i=1}^n h^2(f_{G,\Sigma_i}, f_{H,\Sigma_i}),$$

where $h^2(f,g) = \frac{1}{2}\int\left(\sqrt{f}-\sqrt{g}\right)^2$ denotes the usual squared Hellinger distance between a pair of densities f, g. In the homoscedastic case where $\Sigma_i \equiv \Sigma$, our proposed loss func-

tion $\bar{h}^2(f_{G,\bullet}, f_{H,\bullet}) = h^2(f_{G,\Sigma}, f_{H,\Sigma})$ agrees with the usual squared Hellinger distance. Our first result bounds the average squared Hellinger accuracy $\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet})$ of the NPMLE. In order to accommodate general heteroscedastic Σ_i, we state our results in terms of uniform upper and lower bounds on the spectra of all of the matrices, i.e. $\underline{k} I_d \preceq \Sigma_i \preceq \overline{k} I_d$ for all i. To state the result, some additional notation is needed. We fix a positive scalar $M \geq \sqrt{10\overline{k} \log n}$ and a nonempty compact set $S \subset \mathbb{R}^d$. Define the rate function controlling the squared Hellinger distance

$$\varepsilon_n^2(M, S, G^*) \coloneqq \mathrm{Vol}(S^{\overline{k}^{1/2}}) \frac{M^d}{n} (\log n)^{d/2+1} + \inf_{q \geq (d+1)/(2 \log n)} \left(\frac{2\mu_q}{M} \right)^q \log n, \tag{4.13}$$

where μ_q denotes the q^{th}-moment of $\mathfrak{d}_S(\vartheta) \coloneqq \inf_{s \in S} \|\vartheta - s\|_2$ under $\vartheta \sim G^*$, and $S^a \coloneqq \{y : \mathfrak{d}_S(y) \leq a\}$ denotes the a-enlargement of the set S. Note that we have suppressed the dependence of ε_n^2 on the upper bound $\overline{k}$.

The following result states that $\varepsilon_n^2(M, S, G^*)$ bounds the rate in average Hellinger accuracy both with high probability and in expectation. The scalar $M \geq \sqrt{10\overline{k} \log n}$ and compact set $S \neq \emptyset$ are free parameters. Note that the first term on the right-hand side of (4.13) is increasing in M and S, whereas the second is decreasing in each. In principle, then, we may tune the values of M and S to optimize the rate function $\varepsilon_n^2(M, S, G^*)$. Later in this section, we discuss a number of special cases where a more explicit rate can be obtained.

Theorem 4.6. *Suppose $X_i \overset{\text{ind}}{\sim} f_{G^*,\Sigma_i}$ where $\underline{k} I_d \preceq \Sigma_i \preceq \overline{k} I_d$ for all i. Any (approximate) solution $\widehat{G}_n \in \mathcal{P}(\mathbb{R}^d)$ of (4.4) satisfying*

$$\sup_{G \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{n} \sum_{i=1}^n \log f_{G,\Sigma_i}(X_i) - \frac{1}{n} \sum_{i=1}^n \log f_{\widehat{G}_n,\Sigma_i}(X_i) \lesssim_{d,\overline{k},\underline{k}} \varepsilon_n^2(M, S, G^*) \tag{4.14}$$

satisfies

$$\mathbb{P}\left(\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \gtrsim_{d,\overline{k},\underline{k}} t^2 \varepsilon_n^2(M, S, G^*) \right) \leq 2n^{-t^2}, \tag{4.15}$$

for all $t \geq 1$, provided $n > \max(e\underline{k}^{-d/2}, (2\pi)^{d/2})$. Moreover,

$$\mathbb{E}\left[\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \right] \lesssim_{d,\overline{k},\underline{k}} \varepsilon_n^2(M, S, G^*). \tag{4.16}$$

We prove Theorem 4.6 in Section 4.6.2. Our proof extends Theorem 2.1 of Saha and Guntuboyina (2020a) on the multivariate, homoscedastic case $\Sigma_i \equiv I_d$ and Theorem 4 of Jiang (2020) on the univariate, heteroscedastic case $d = 1$, which in turn build upon Theorem 1 of Zhang (2009) on the univariate, homoscedastic case. The general theory on rates of convergence for maximum likelihood estimators (van de Geer, 2000; Wong & Shen, 1995) can in principle be used to bound $\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet})$. Our proof technique deviates from the general

theory by directly bounding the likelihood $f_{\widehat{G}_n,\Sigma_i}(x)$ for x outside some pre-specified domain (controlled by the choice of set S), and then covering the set of densities $\{f_{G,\bullet} : G \in \mathcal{P}(\mathbb{R}^d)\}$ within the domain in the L_∞ metric.

Theorem 4.6 provides a sharp bound in many special cases of G^*. For a given G^* we need to optimize over the choices of $M \geq \sqrt{10\overline{k}\log n}$ and the nonempty compact set $S \subset \mathbb{R}^d$ to obtain the smallest value of the rate function $\varepsilon_n^2(M, S, G^*)$. Our next result performs this calculation for various assumptions on the prior G^*.

Corollary 4.7. *Suppose $X_i \overset{\text{ind}}{\sim} f_{G^*,\Sigma_i}$ where $\underline{k} I_d \preceq \Sigma_i \preceq \overline{k} I_d$ for all i. Suppose $\widehat{G}_n \in \mathcal{P}(\mathbb{R}^d)$ is any approximate NPMLE such that*

$$\sup_{G\in\mathcal{P}(\mathbb{R}^d)} \frac{1}{n}\sum_{i=1}^n \log f_{G,\Sigma_i}(X_i) - \frac{1}{n}\sum_{i=1}^n \log f_{\widehat{G}_n,\Sigma_i}(X_i) \lesssim_{d,\overline{k},\underline{k}} \frac{(\log n)^{d+1}}{n}. \tag{4.17}$$

(i) (Discrete support) If $G^* = \sum_{j=1}^{k^*} w_j^* \delta_{a_j^*}$, then

$$\mathbb{E}\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \lesssim_{d,\overline{k},\underline{k}} \frac{k^*}{n}(\log n)^{d+1}.$$

(ii) (Compact support) If G^* has compact support S^*, then

$$\mathbb{E}\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \lesssim_{d,\overline{k},\underline{k}} \frac{\mathrm{Vol}\left(S^* + \mathbb{B}_{\overline{k}^{1/2}}(0)\right)}{n}(\log n)^{d+1},$$

where $\mathbb{B}_r(x) \coloneqq \{y : \|x-y\|_2 \leq r\}$ denotes the d-dimensional ball of radius r centered at x.

(iii) (Simultaneous moment control) Suppose that there is a compact $S^* \subset \mathbb{R}^d$ and $\alpha \in (0,2]$, $K \geq 1$ such that $\mu_q \coloneqq \mathbb{E}_{\vartheta\sim G^*}\left[\mathfrak{d}^q(\vartheta, S)\right]^{1/q} \leq K q^{1/\alpha}$ for all $q \geq 1$ (recall $\mathfrak{d}_S(\vartheta) \coloneqq \inf_{s\in S}\|\vartheta - s\|_2$ as above). Then

$$\mathbb{E}\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \lesssim_{\alpha,K,d,\overline{k},\underline{k}} \frac{\mathrm{Vol}\left(S^* + \mathbb{B}_{\overline{k}^{1/2}}(0)\right)}{n}(\log n)^{\frac{2+\alpha}{2\alpha}d+1}.$$

(iv) (Finite q^{th} moment) Suppose that there is a compact $S^* \subset \mathbb{R}^d$ and $\mu, q > 0$ such that $\mu_q \leq \mu$. Then

$$\mathbb{E}\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \lesssim_{\mu,q,d,\overline{k},\underline{k}} \left(\frac{\mathrm{Vol}\left(S^* + \mathbb{B}_{\overline{k}^{1/2}}(0)\right)}{n}\right)^{\frac{q}{q+d}} (\log n)^{\frac{q}{2q+2d}d+1}.$$

Given the general result in Theorem 4.6, Corollary 4.7 follows directly from the calculations of Saha and Guntuboyina (2020a) in Corollary 2.2 and Theorem 2.3. Corollary 4.7 captures an important adaptation property of the NPMLE. The cases $(i) - (iv)$ described in the result are nested in the sense that (i) implies (ii), (ii) implies (iii), and (iii) implies (iv); consequently the rates get progressively worse as our assumptions weaken. This means that the NPMLE, despite searching over all probability measures $\mathcal{P}(\mathbb{R}^d)$, obtains better rates when structure is present in the prior G^*.

Most strikingly, when G^* has discrete support with k^* support points, the rate in (i) is $\frac{k^*}{n}$ up to logarithmic factors *without assuming any knowledge of* k^*. This rate matches the minimax rate over all discrete distributions with at most k^* support points (Saha & Guntuboyina, 2020a), meaning we could not expect to do much better even if k^* were known. In the extreme case where $k^* = 1$, the observations actually come from a simple Gaussian, i.e. $f_{G^*,\Sigma_i}(x) = \varphi_{\Sigma_i}(x - a_1^*)$ with common mean $a_1^* \in \mathbb{R}^d$, so our result says we don't lose much in the rate when we model the density with a mixture even when it turns out to be a simple Gaussian. Similarly, in (ii), the rate adapts to the size of the support S^* without prior knowledge of this support or even a bound on its size. Up through simultaneous moment control (iii), the dimension d only affects the rate as a function of n through the logarithmic factor. Hence, the NPMLE avoids the usual curse of dimensionality to some extent, while still achieving consistency in the heavier tailed setting (iv). The logarithmic factors in our bounds might be reduced slightly but cannot be eliminated as they are present in the minimax lower bounds (Kim & Guntuboyina, 2020).

4.3.1.1 Implications for the Discretization Rate

Theorem 4.6 establishes that up to a multiplicative constant (depending only on the dimension d and bounds $\underline{k}, \overline{k}$ on the eigenvalues of the covariance matrices) the quantity $\varepsilon_n^2(M, S, G^*)$ controls the average Hellinger accuracy $\mathbb{E}[\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet})]$ of the NPMLE. This also holds for approximate solutions to the optimization problem (4.4) that, in accordance with (4.14), place nearly as much likelihood on the data as does a global maximizer. It is natural to compare the requirement (4.14) with our computational guarantee on the discretization error (4.12) from Proposition 4.5. The free parameter which controls the discretization error is the resolution $\delta > 0$, which represents the width of the hypercubes we use to cover the ridgeline manifold $\mathcal{M}$ or any of its outer-approximations from Corollary 4.3. Thus, in order to satisfy the main requirement of Theorem 4.6, we need to take δ such that

$$\varepsilon_n^2(M, S, G^*) \gtrsim_{d,\overline{k},\underline{k}} d\underline{k}^{-2} \left(2D^2 + \frac{1}{2}\right) \delta^2.$$

Observe from the definition of ε_n^2 that $\varepsilon_n^2(M, S, G^*) \gtrsim_{d,\overline{k},\underline{k}} \frac{(\log n)^{d+1}}{n}$ for all $M \geq \sqrt{10\overline{k} \log n}$ and all compact S. Absorbing additional terms depending on d,$\overline{k}$, and $\underline{k}$ and assuming for

simplicity that $D > \frac{1}{2}$, choosing δ such that

$$D^2\delta^2 \lesssim_{d,\overline{k},\underline{k}} \frac{(\log n)^{d+1}}{n} \tag{4.18}$$

suffices for the discretized NPMLE to be statistically indistinguishable from a global maximizer.

The inequality (4.18) gives a preliminary bound on the rate at which the discretization level δ should decrease with n. Still, recall from Proposition 4.5 that D denotes the diameter of the ridgeline manifold $\mathcal{M}$, so D does depend on n. To sketch the dependence, let us consider a representative example where G^* has sub-Gaussian tails and all of the Σ_i's are diagonal. In this case, by Corollary 4.3 part (ii), the ridgeline manifold $\mathcal{M}$ is contained in the axis-aligned minimum bounding box of the data

$$\prod_{j=1}^{d}\left[\min_{i\in\{1,\ldots,n\}} X_{ij}, \max_{i\in\{1,\ldots,n\}} X_{ij}\right].$$

Due to the tail condition, the length of each side of this hyper-rectangle grows like $\sqrt{\log n}$ with high probability up to multiplicative factors depending on $\overline{k}$: hence, the diameter D also scales like $\sqrt{\log n}$ with high probability up to multiplicative factors depending on $\overline{k}$ and d. We have thus shown that it suffices to discretize at a resolution of $\delta \asymp \sqrt{\frac{(\log n)^d}{n}}$. The number of points in our covering $\mathcal{A}$ is of order $m \asymp \left(\frac{n}{(\log n)^d}\right)^{d/2}$. In the univariate case $d = 1$, this slightly improves the finding of Theorem 2 of Dicker and Zhao (2016), who showed that an $m = \sqrt{n}$-discretization of the range of the data $[X_{(1)}, X_{(n)}]$ suffices for the same rate in Hellinger distance. Their bound on the large-deviation probability is also logarithmic, i.e. $O\left(\frac{1}{\log n}\right)$ whereas our equation (4.15) is polynomial in n. Our analysis also clarifies that the sense in which we need approximate NPMLE (4.14) is through the likelihood of the observations, relative to the global optimum, which could be useful for comparing alternative approaches to approximating the NPMLE.

4.3.2 Denoising: an oracle inequality

In this section we turn to the problem of estimating the oracle posterior means $(\hat{\theta}_i^*)_{i=1}^n$; see (4.5). We evaluate the performance of $(\hat{\theta}_i)_{i=1}^n$ (see (4.6)) as an estimator for $(\hat{\theta}_i^*)_{i=1}^n$ using the mean squared error risk measure:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\hat{\theta}_i - \hat{\theta}_i^*\|_2^2.$$

Since $\hat{\theta}_i^*$ is the optimal estimator of θ_i^* given model (4.1), the above mean squared error quantifies the price of misspecifying G^* with the data-driven estimator $\widehat{G}_n$. Hence, this loss is also known as the per-instance *empirical Bayes regret*.

Our next result states that the rate function $\varepsilon_n^2(M, S, G^*)$ governing the Hellinger accuracy (see (4.13)) also upper bounds the regret, up to additional logarithmic factors. We provide the same special cases of the rate as those stated in Corollary 4.7.

Theorem 4.8. *Suppose* $X_i \overset{\text{ind}}{\sim} f_{G^*,\Sigma_i}$ *where* $\underline{k} I_d \preceq \Sigma_i \preceq \overline{k} I_d$ *for all* i. *Let* $\widehat{G}_n$ *denote any approximate NPMLE satisfying* (4.17). *Fix some* $M \geq \sqrt{10\bar{k} \log n}$ *and a nonempty, compact set* $S \subset \mathbb{R}^d$. *Define* $\varepsilon_n^2(M, S, G^*)$ *as in* (4.13). *For all* $n \geq 5\underline{k}^{-d/2} \vee (2\pi)^{d/2}$,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\hat{\theta}_i - \hat{\theta}_i^*\|_2^2 \lesssim_{d,\overline{k},\underline{k}} \varepsilon_n^2(M, S, G^*)(\log n)^{(d/2-1)\vee 3}. \tag{4.19}$$

In particular, consider the following special cases for G^*:

(i) (Discrete support) If $G^* = \sum_{j=1}^{k^*} w_j^* \delta_{a_j^*}$, *then*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\hat{\theta}_i - \hat{\theta}_i^*\|_2^2 \lesssim_{d,\overline{k},\underline{k}} \frac{k^*}{n}(\log n)^{d+((d/2)\vee 4)}.$$

(ii) (Compact support) If G^* *has compact support* S^*, *then*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\hat{\theta}_i - \hat{\theta}_i^*\|_2^2 \lesssim_{d,\overline{k},\underline{k}} \frac{\mathrm{Vol}\left(S^* + \mathbb{B}_{\overline{k}^{1/2}}(0)\right)}{n}(\log n)^{d+((d/2)\vee 4)}.$$

(iii) (Simultaneous moment control) Suppose that there is a compact $S^* \subset \mathbb{R}^d$ *and* $\alpha \in (0, 2]$, $K \geq 1$ *such that* $\mu_q := \mathbb{E}_{\vartheta\sim G^*}\left[\mathfrak{d}^q(\vartheta, S^*)\right]^{1/q} \leq K q^{1/\alpha}$ *for all* $q \geq 1$. *Then*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\hat{\theta}_i - \hat{\theta}_i^*\|_2^2 \lesssim_{\alpha,K,d,\overline{k},\underline{k}} \frac{\mathrm{Vol}\left(S^* + \mathbb{B}_{\overline{k}^{1/2}}(0)\right)}{n}(\log n)^{\frac{2\alpha d}{2+\alpha}+((d/2)\vee 4)}.$$

(iv) (Finite q^{th} *moment) Suppose that there exists a compact* $S^* \subset \mathbb{R}^d$ *and* $\mu, q > 0$ *such that* $\mu_q \leq \mu$. *Then*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\hat{\theta}_i - \hat{\theta}_i^*\|_2^2 \lesssim_{\mu,q,d,\overline{k},\underline{k}} \left(\frac{\mathrm{Vol}\left(S^* + \mathbb{B}_{\overline{k}^{1/2}}(0)\right)}{n}\right)^{\frac{q}{q+d}} (\log n)^{\frac{qd}{2q+2d}+((d/2)\vee 4)}.$$

Theorem 4.8 shows that the denoising problem shares the adaptation features as the density estimation problem. Since we have assumed $\underline{k} I_d \preceq \Sigma_i \preceq \overline{k} I_d$ for all $i = 1, \ldots, n$, the same set of results also hold for the scaled regret $\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\hat{\theta}_i - \hat{\theta}_i^*)^\mathsf{T}\Sigma_i^{-1}(\hat{\theta}_i - \hat{\theta}_i^*)$.

Remark 4.9. *(On the proof of Theorem 4.8 in Section 4.6.3) Our proof extends Theorem 3.1 of Saha and Guntuboyina (2020a) on the multivariate, homoscedastic case $\Sigma_i \equiv I_d$ and Theorem 1 of Jiang (2020) on the univariate, heteroscedastic case $d = 1$, which in turn build upon Theorem 5 of Jiang and Zhang (2009) on the univariate, homoscedastic case. Jiang and Zhang (2009) and Jiang (2020) used a related notion of regret*

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\hat{\theta}_i - \theta_i^*\|_2^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\hat{\theta}_i^* - \theta_i^*\|_2^2}.$$

Tweedie's formula relates the oracle (4.7) *and empirical Bayes* (4.8) *posterior means to the corresponding marginal likelihoods, so the density estimation results of the previous section turn out to be useful for proving Theorem 4.8 as well. In particular, we consider Bayes rules for priors in a covering of the Hellinger ball*

$$\left\{G \in \mathcal{P}(\mathbb{R}^d) : \bar{h}^2(f_{G,\bullet}, f_{G^*,\bullet}) \lesssim_{d,\bar{k},\underline{k}} t^2\varepsilon_n^2(M, S, G^*)\right\},$$

which, by Theorem 4.6, contains $\widehat{G}_n$ with high probability. For a fixed prior G, the denominator in the correction factor of Tweedie's formula

$$X_i + \Sigma_i \frac{\nabla f_{G,\Sigma_i}(X_i)}{f_{G,\Sigma_i}(X_i)},$$

namely $f_{G,\Sigma_i}(X_i)$, can be small. To avoid dividing by near-zero quantities, we regularize the above Bayes rule by replacing the denominator with $\max\{f_{G,\Sigma_i}(X_i), \rho\}$ for a small positive ρ. To handle heteroscedastic errors, we show that Tweedie's formula, even its regularized form, is equivariant under scale transformations.

4.3.3 Deconvolution: estimating the prior

We turn to the fundamental question of how well $\widehat{G}_n$ estimates G^*. This is known as the deconvolution problem and has received much attention in the statistical literature (Meister, 2009). Indeed, the original consistency results (Kiefer & Wolfowitz, 1956; Pfanzagl, 1988) for the NPMLE focused on weak convergence of $\widehat{G}_n$ to G^* as $n \to \infty$. While most prior work on deconvolution has focused on deconvolution with homoscedastic error distributions, Delaigle and Meister (2008) allowed for heteroscedastic errors but relied on kernel estimators which contain additional smoothing parameters. By contrast, the NPMLE provides a tuning-free estimate of the mixing distribution G^*, yet to our knowledge, non-asymptotic bounds on the rate of convergence for $\widehat{G}_n$ in the deconvolution problem are not known.

In practice, the true prior G^* may not be discrete even though $\widehat{G}_n$ always is, and even if both distributions are discrete, their supports will typically differ. Our loss function must allow for comparisons of probability measures with potentially disjoint supports. Nguyen

(2013) established that a natural loss for this problem is the Wasserstein distance from the theory of optimal transport

$$W_2^2(G, H) \coloneqq \min_{(U,V)\in\Pi_{G,H}} \mathbb{E}\|U - V\|_2^2,$$

where $G, H \in \mathcal{P}(\mathbb{R}^d)$ are two probability measures and $\Pi_{G,H}$ denotes the set of couplings of G and H, i.e. joint distributions over $(U, V) \in \mathbb{R}^{2d}$ such that $U \sim G$ and $V \sim H$. Indeed, even the likelihood criterion is intimately related to the Wasserstein distance: in the homoscedastic case $\Sigma_i \equiv \sigma^2 I_d$, it is known that the NPMLE (4.4) equivalently solves an entropic-regularized optimal transport problem (Rigollet & Weed, 2018).

Nguyen (2013) connected the deconvolution error $W_2^2(G, H)$ to the density estimation error between the mixtures, i.e. $h^2(f_{G,I_d}, f_{H,I_d})$ in a homoscedastic Gaussian deconvolution setting. By leveraging similar techniques as well as the support bounds of Corollary 4.3, we arrive at the following upper bound on the deconvolution error.

Theorem 4.10. *Suppose $X_i \overset{\text{ind}}{\sim} f_{G^*,\Sigma_i}$ where $\underline{k} I_d \preceq \Sigma_i \preceq \overline{k} I_d$ and Σ_i is a diagonal matrix for each i. Suppose further that $G^*([-L, L]^d) = 1$ for some $L \geq 0$. Let $\widehat{G}_n$ denote any approximate NPMLE supported on the minimum axis-aligned bounding box of the data satisfying* (4.17)*. Then there is a function $n(d, \overline{k}, \underline{k}, L)$ such that, for all sample sizes n with $n \geq n(d, \overline{k}, \underline{k}, L)$,*

$$W_2^2(G^*, \widehat{G}_n) \lesssim_{d,\overline{k}} \frac{1}{\log n},$$

with probability at least $1 - \frac{4d}{n^8}$.

Theorem 4.10 (proved in Section 4.6.4) upper bounds the rate of convergence under the Wasserstein distance by the extremely slow logarithmic rate $\frac{1}{\log n}$. It is well known that the smoothness of the Gaussian errors makes the deconvolution more difficult; in fact, the logarithmic rate is minimax optimal (Dedecker & Michel, 2013).

Remark 4.11. *(On Theorem 4.10) To our knowledge, Theorem 4.10 is novel, and the rate of convergence for the NPMLE under a Wasserstein distance has not been studied previously. The structure of the proof follows the proof of Theorem 2 of Nguyen (2013). To deal with the fact that $\widehat{G}_n$ and G^* are typically singular, we convolve each with a distribution with full support but low variance. Compared to our results on the density estimation and denoising problems, Theorem 4.10 makes additional assumptions on the problem structure, specifically that the covariance matrices are diagonal and that G^* is compactly supported. Many practical applications satisfy the diagonal covariances restriction, including both of our applications in Section 4.4.*

A common feature to our results on density estimation and denoising have been that the NPMLE adapts to the complexity of G^*. It is reasonable to conjecture, then, that

in the deconvolution problem, $\widehat{G}_n$ will also enjoy some adaptation properties under the Wasserstein distance. We close this section with a sharper result on the Wasserstein rate in the special case where the observations are drawn from Gaussian distributions with common mean $\mu \in \mathbb{R}^d$.

Theorem 4.12. *Suppose* $X_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu, \Sigma_i)$, *i.e.* $X_i \overset{\text{ind}}{\sim} f_{G^*,\Sigma_i}$ *where* $G^* = \delta_\mu$ *and* $\underline{k} I_d \preceq \Sigma_i \preceq \overline{k} I_d$ *for all* $i = 1, \ldots, n$. *Let* $\widehat{G}_n$ *denote any approximate NPMLE satisfying* (4.17) *and supported on* $\mathbb{B}_{\kappa r}(\bar{X})$ *where* $\kappa = \overline{k}/\underline{k}$, $r = \max_i \|X_i - \bar{X}\|_2$, *and* $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. *Then*

$$W_2(\widehat{G}_n, G^*) \lesssim_{d,\overline{k},\underline{k}} t^{3/2} \frac{(\log n)^{(d+3)/4}}{n^{1/4}}$$

with probability at least $1 - 3n^{-t^2}$ *for all* $t \geq 1$.

If the approximate NPMLE $\widehat{G}_n$ of Theorem 4.12 is selected according to the strategy described in Section 4.2.2, then by Corollary 4.3 part (*iii*) its support will be contained within the ball $\mathbb{B}_{\kappa r}(\bar{X})$. This additional assumption on the support of the approximate NPMLE is needed to have some control over the moments of $\widehat{G}_n$.

Up to logarithmic factors, the $n^{1/4}$-rate in Theorem 4.12 agrees with Corollary 4.1 of Ho and Nguyen (2016) for the MLE of an overfitted mixture. Specifically, their result compared the MLE of k-component finite Gaussian mixture to a true mixing distribution G^* with $k^* < k$ components. Wu and Yang (2020) and Doss et al. (2020) also derived the $n^{1/4}$-rate for a different estimator under a different Wasserstein metric. All of these previous results were restricted to the homoscedastic setting. In our setting, $k^* = 1$ and $k = \hat{k}$ is the data-dependent order of the NPMLE. The best known bound on $\hat{k}$ is logarithmic in n (Polyanskiy & Wu, 2020), whereas Ho and Nguyen (2016) required k to be fixed as $n \to \infty$. When k^* is known, a faster $n^{1/2}$-rate is possible (Heinrich & Kahn, 2018) and is achieved by the MLE in a well-specified finite mixture model, i.e. setting $k = k^*$ (Ho & Nguyen, 2016).

While the slower $n^{1/4}$-rate appears to be the price of flexibility of the NPMLE, Theorem 4.12 establishes that the NPMLE indeed adapts to structure in G^*. Our analysis is greatly simplified by the assumption $G^* = \delta_\mu$, since there is only one coupling between $\widehat{G}_n$ and G^*. We leave for future work the important question of the extent to which $\widehat{G}_n$ adapts to more general distributions G^*.

4.4 Applications

4.4.1 Color-magnitude diagram

In this section, we continue our discussion of denoising the color-magnitude diagram (CMD) from Section 4.1. Our modeling strategy is closely related to the work of Anderson et al. (2018). To compare our method to extreme deconvolution (Bovy et al., 2011), we use the same stellar sample, relaxing only their assumption that the prior G^* is a mixture of

Gaussians; by contrast, we allow G^* to be an arbitrary probability measure. Specifically, we assume that after a suitable transformation of the color and magnitude measurements, the pair, denoted $X_i \in \mathbb{R}^2$, come from a two-dimensional Gaussian mixture f_{G^*,Σ_i} with known covariance Σ_i.

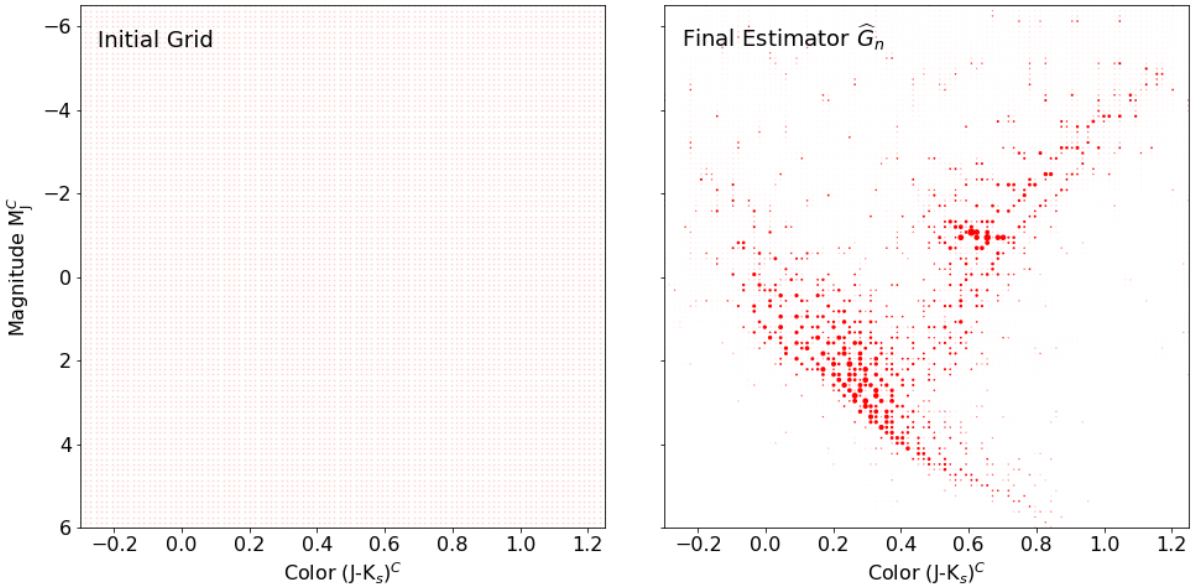


Figure 4.5: Initial grid (left) of $m = 10^4$ support points and estimated prior $\widehat{G}_n$ (right) where the area of each atom is proportional to its weight.

Figure 4.1 in Section 4.1 shows the plot of the observed data X_i (left) and estimated posterior means $\hat{\theta}_i$ (right), the latter constituting the denoised CMD. Contrasting our CMD with theirs (Anderson et al., 2018, Figure 7), which we do not depict here, it appears that ours performs more shrinkage overall. Our CMD has rather sharp tails in the bottom of the plot (i.e. the main sequence) and the top right (i.e. the tip of the red-giant branch) as well as a definitive cluster in the center-right (i.e. the red clump).

There are also important differences between the NPMLE and extreme deconvolution in the estimated prior $\widehat{G}_n$. Figure 4.5 shows the initial and final iterates in the computation of the NPMLE. It is clear that we are using a discrete distribution to model the prior, and since all of the covariance matrices Σ_i are diagonal, by Corollary 4.3 we have restricted the support points to lie in the minimum axis-aligned bounding box of the data. By contrast, extreme deconvolution models the prior as itself a Gaussian mixture, so the estimated prior (Anderson et al., 2018, Figure 4) actually is supported on all of $\mathbb{R}^2$.

4.4.2 Chemical abundance ratios

Our second data set is taken from the Apache Point Observatory Galactic Evolution Experiment survey (APOGEE); see Majewski et al. (2017), Abolfathi et al. (2018). We examine chemical abundance ratios for the red clump (RC) stars given in the DR14 APOGEE red clump catalog; see Ratcliffe et al. (2020) where this data set has been studied. Following the pre-processing in Ratcliffe et al. (2020) to remove the outliers with anomalous abundance measurements, the data set contains $n = 27,238$ observations. We pick $d = 2$ features from the 19 dimensions, namely, [Si/Fe]-[Mg/Fe].

In Figure 4.6 we plot the observed data (top left) and estimated posterior means using Gaussian denoising under the estimated prior $\widehat{G}_n$ (top right). The initial grid (bottom left) of $m = 10^4$ support points and estimated prior $\widehat{G}_n$ (bottom right), where the area of each atom is proportional to its weight, is also provided. The denoised data reveals a very interesting structure — it shows that the variables [Si/Fe] and [Mg/Fe] are strongly correlated, especially, the observations for the upper right cluster of stars could be lying on one dimensional manifold; something that is not at all visible when plotting the original data.

4.5 Concluding remarks

In this chapter we study the NPMLE $\widehat{G}_n$ as an estimator of a prior distribution G^* in the presence of multivariate, heteroscedastic measurement errors. We resolve a number of basic questions on the existence, uniqueness, discreteness, and support of the NPMLE, where in several cases the answers differ significantly from the traditional univariate, homoscedastic setting. Our analysis identifies a dual mixture density $\widehat{\psi}_n$ with Gaussian $\mathcal{N}(X_i, \Sigma_i)$ components at each observation, whose modes contain the atoms of the NPMLE. Our characterization implies that the NPMLE is supported on the ridgeline manifold $\mathcal{M}$, which is a compact subset of $\mathbb{R}^d$ defined in terms of the observations $(X_i)_{i=1}^n$ and corresponding covariance matrices $(\Sigma_i)_{i=1}^n$. This support reduction allows us to approximate the NPMLE by a finite-dimensional convex optimization over the mixing proportions, and we develop a novel approach to bounding the discretization error, justifying the gridding scheme proposed by Koenker and Mizera (2014). Our real data applications show that this approach is viable for practical astronomy problems. Our theoretical results in Section 4.3 provide strong justification for using the NPMLE in a variety of contexts—estimating the prior, marginal densities, and oracle posterior means.

We conclude by outlining some possible future research directions. Computation remains an important barrier for large-scale applications. Specifically, for problems with a large number of samples, e.g. $n \gg 10^6$, some additional forms of approximation are warranted, such as stochastic optimization or binning via coresets (see also Ritchie and Murray (2019) on approaches for scaling Extreme Deconvolution to large datasets). Further, our result on the discretization error suggests that discretization becomes infeasible in moderate-dimensions,

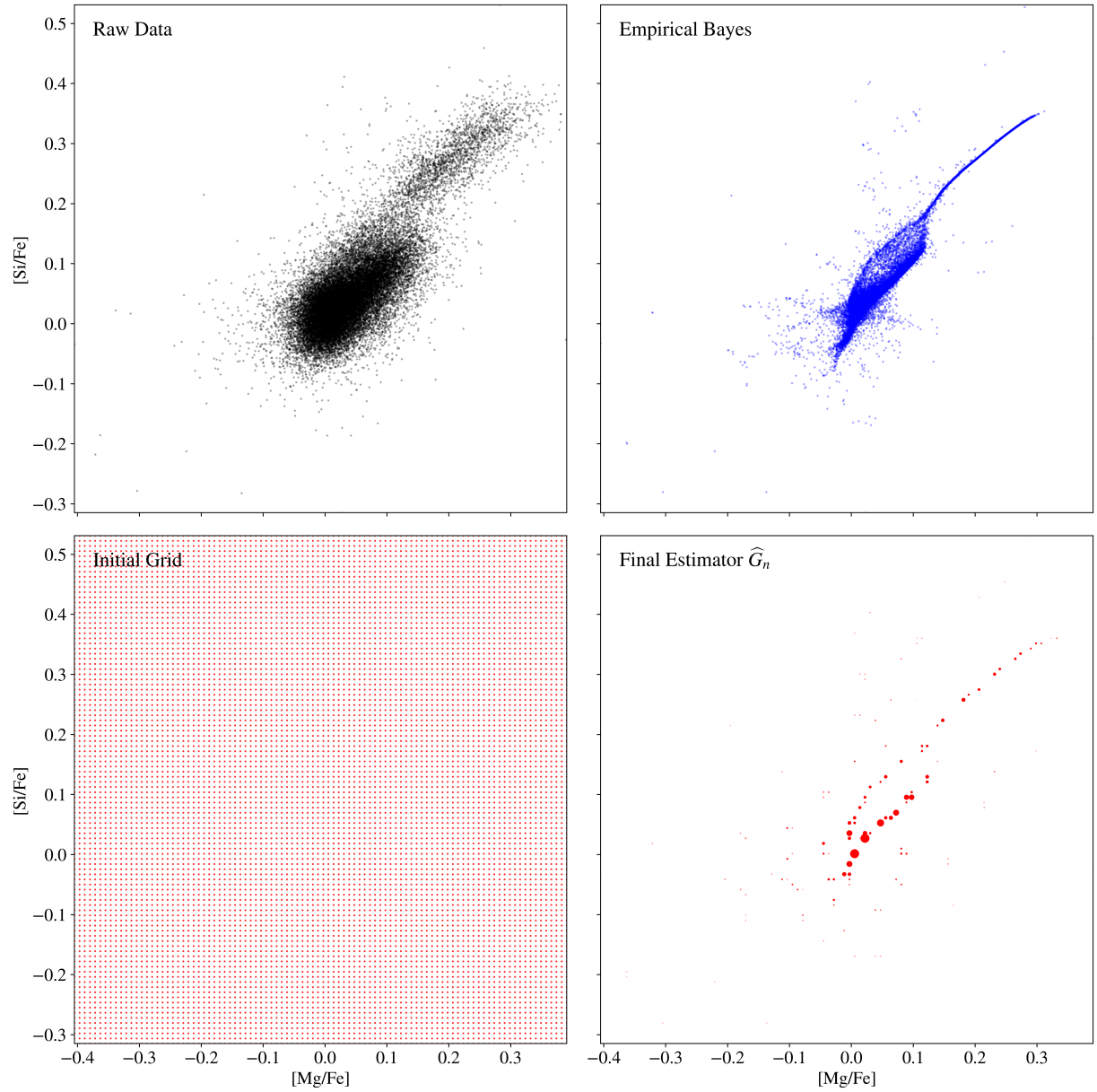


Figure 4.6: Top: Observed data (left) and estimated posterior means (right); Bottom: Initial grid (left) of $m = 10^4$ support points and estimated prior $\widehat{G}_n$ (right) where the area of each atom is proportional to its weight.

where the number of atoms needs to grow roughly like a polynomial in the number of dimensions, e.g. $m = O(\delta^{-d})$. This limitation, which is common to many forms of discretization across applied mathematics, highlights the need for grid-free methods for computation of

the NPMLE in high-dimensions. The connection to entropic-regularized optimal transport established by Rigollet and Weed (2018) represents one possible direction for grid-free methods.

Next, while our framework allows the prior G^* to be arbitrary, the underlying assumption—that the means (θ_i^*) are *identically distributed*—can sometimes be difficult to justify for heteroscedastic observations. The IID assumption reflects the belief that the observation covariance Σ_i is uninformative for the corresponding mean θ_i^*. This assumption led to reasonable results in our applications but may be problematic in other settings. In the univariate, heteroscedastic case, Weinstein et al. (2018) proposed grouping observations with similar variances and applying a spherically symmetric estimator separately within each group. Their approach is capable of capturing dependence between θ_i and σ_i^2, at the expense of not sharing information across groups. Furthermore, to our knowledge, the grouping approach has not been extended to multivariate settings where binning the set of covariance matrices is more difficult. Thus, in multivariate settings there remains the important problem of how to model the relationship between θ_i and Σ_i.

Finally, there remain a number of open statistical questions for future work. Our analysis of the denoising problem focuses on estimating the posterior mean based on the unknown prior G^*, but there are numerous inferential goals one could target with an approximate prior. The analyst might summarize the empirical posteriors using a different functional, such as the posterior median or the posterior mean of some transformed parameter. This question warrants a more general analysis evaluating the quality of the empirical posterior distributions for the true, unknown posteriors.

4.6 Proofs

4.6.1 Proofs of results in sections 4.2 and 4.2.2

4.6.1.1 Proof of Lemma 4.1

The following uses similar techniques as Section 5.2 of Lindsay (1995), which contains a subset of our result in the homoscedastic case.

Proof of Lemma 4.1. By convexity, the first-order optimality condition for $\widehat{G}_n$ is

$$D(\widehat{G}_n, G) \leq 0 \text{ for all } G \in \mathcal{P}(\mathbb{R}^d)$$

where

$$\begin{aligned} D(\widehat{G}_n, G) &:= \lim_{\alpha \downarrow 0} \frac{\frac{1}{n}\sum_{i=1}^n [\log f_{(1-\alpha)\widehat{G}_n + \alpha G, \Sigma_i}(X_i) - \log f_{\widehat{G}_n, \Sigma_i}(X_i)]}{\alpha} \\ &= \frac{1}{n}\sum_{i=1}^n \frac{1}{f_{\widehat{G}_n,\Sigma_i}(X_i)} \left(f_{G,\Sigma_i}(X_i) - f_{\widehat{G}_n,\Sigma_i}(X_i) \right) = \frac{1}{n}\sum_{i=1}^n \frac{f_{G,\Sigma_i}(X_i)}{f_{\widehat{G}_n,\Sigma_i}(X_i)} - 1 \end{aligned}$$

When $G = \delta_\vartheta$ is a point mass we write $D(\widehat{G}_n, \vartheta)$ instead of $D(\widehat{G}_n, G)$. It suffices to check $D(\widehat{G}_n, \vartheta) \le 0$ for all $\vartheta \in \mathbb{R}^d$ because $D(\widehat{G}_n, G) = \int D(\widehat{G}_n, \vartheta)\,\mathrm{d}G[\vartheta]$.

For the first part of the Lemma, define $\mathcal{C} \coloneqq \left\{(f_{G,\Sigma_i}(X_i))_{i=1}^n : G \in \mathcal{P}(\mathbb{R}^d)\right\} \cup \{0\}$. Observe that

$$\mathcal{C} = \operatorname{conv}(\mathcal{L}), \text{ where } \mathcal{L} \coloneqq \left\{(\varphi_{\Sigma_i}(X_i - \vartheta))_{i=1}^n : \vartheta \in \mathbb{R}^d\right\} \cup \{0\}.$$

Since $\vartheta \mapsto (\varphi_{\Sigma_i}(X_i - \vartheta))_{i=1}^n$ is continuous and $\lim_{\|\vartheta\|_2 \to \infty}(\varphi_{\Sigma_i}(X_i - \vartheta))_{i=1}^n = 0$, the set $\mathcal{L}$ is closed, and by boundedness of the Gaussian likelihood, $\mathcal{L}$ is compact. Hence $\mathcal{C} \subset \mathbb{R}^n$ is convex and compact, and $f(L) = \frac{1}{n}\sum_{i=1}^n \log L_i$ is strictly concave over $\mathcal{C}$. Thus, f attains its maximum at a unique (nonzero) boundary point $\hat{L} \in \partial\mathcal{C}$. Observe $\mathcal{C} = \operatorname{conv}\left\{(\varphi_{\Sigma_i}(X_i - \vartheta))_{i=1}^n : \vartheta \in \mathbb{R}^d\right\}$: by Carathéodory's theorem, any boundary point $\hat{L} \in \partial\mathcal{C}$ can be written as $\hat{L}_i = \sum_{j=1}^{\hat{k}} \hat{w}_j \varphi_{\Sigma_i}(X_i - \hat{a}_j)$ for some $\hat{k} \le n$.

Suppose $B \subset \operatorname{supp}(\widehat{G}_n)$ is contained in the support of the NPMLE. Given $\widehat{G}_n(B) > 0$, define a new probability measure $\widehat{G}_n^B$ via $\widehat{G}_n^B(A) \coloneqq \frac{\widehat{G}_n(A \cap B)}{\widehat{G}_n(B)}$. Since $\widehat{G}_n = \alpha_0 \widehat{G}_n^B + (1-\alpha_0)\widehat{G}_n^{B^\mathsf{c}}$ for $\alpha_0 = \widehat{G}_n(B)$, the mixture

$$G_\alpha = (1-\alpha)\widehat{G}_n + \alpha \widehat{G}_n^B$$

remains a valid probability measure for $\alpha \ge -\frac{\alpha_0}{1-\alpha_0}$. Since $\alpha = 0$ maximizes the log-likelihood of G_α over a range $\alpha \in [-\frac{\alpha_0}{1-\alpha_0}, 1]$ including both negative and positive values, the derivative of the log-likelihood is zero at $\alpha = 0$, i.e.

$$0 = D(\widehat{G}_n, \widehat{G}_n^B) = \int D(\widehat{G}_n, \vartheta)\,\mathrm{d}\widehat{G}_n^B[\vartheta],$$

so $\widehat{G}_n^B(\mathcal{Z}) = 1$ for all $B \subset \operatorname{supp}(\widehat{G}_n)$ such that $\widehat{G}_n(B) > 0$. This implies $\widehat{G}_n(B \cap \mathcal{Z}) = \widehat{G}_n(B)$ for all measurable B, from which we may conclude $\mathcal{Z} \supseteq \operatorname{supp}(\widehat{G}_n)$. Finally, observe that

$$D(\widehat{G}_n, \vartheta) = \frac{1}{n}\sum_{i=1}^n \hat{L}_i^{-1}\varphi_{\Sigma_i}(X_i - \vartheta) - 1 = \left(\frac{1}{n}\sum_{i=1}^n \hat{L}_i^{-1}\right)\widehat{\psi}_n(\vartheta) - 1,$$

so $D(\widehat{G}_n, \vartheta) \le 0$ is equivalent to $\widehat{\psi}_n(\vartheta) \le \left(\frac{1}{n}\sum_{i=1}^n \hat{L}_i^{-1}\right)^{-1}$. This proves the last statement of the Lemma, that $\mathcal{Z}$ is equal to the set of global maximizers of $\widehat{\psi}_n$. □

4.6.1.2 Proof of Lemma 4.2

Proof of Lemma 4.2. By Lemma 4.4, the fitted values $\hat{L}_1 = \hat{L}_2 = \hat{L}_3$ are equal. By Lemma 4.1, the atoms of $\widehat{G}_n$ occur at the global modes of $\widehat{\psi}_n = f_{H,\sigma^2 I_2}$, where $H = \frac{1}{3}\sum_{i=1}^3 \delta_{X_i}$. Since $\hat{L}_1 = \hat{L}_2 = \hat{L}_3$, the fitted values are also equal to the global maximum of $\widehat{\psi}_n$, i.e.

$$\hat{L}_i = \max_x f_{H,\sigma^2 I_2}(x) = \frac{2^{2/3}\log 2}{3\pi}$$

for each $i = 1, 2, 3$. Note that $\hat{L}_i = f_{\delta_0, \sigma^2 I_2}(X_i)$ for all X_i, so $\widehat{G}_n = \delta_0$ is an NPMLE. Now let $\widehat{G}'_n = \frac{1}{3}\sum_{i=1}^n \delta_{X_i/2}$. It suffices to check the fitted values of $\widehat{G}'_n$ at the observations. For $i = 1$,

$$\begin{aligned} f_{\widehat{G}'_n, \sigma^2 I_2}(X_1) &= \frac{1}{3}\sum_{i=1}^3 \varphi_{\sigma^2 I_2}(X_1 - X_i/2) \\ &= \frac{4\log 2}{9\pi}(2^{-(4/3)(1/4)} + 2^{-(4/3)(7/4)} + 2^{-(4/3)(7/4)}) = \frac{2^{2/3}\log 2}{3\pi} = \hat{L}_1. \end{aligned}$$

Similarly, for $i = 2$,

$$\begin{aligned} f_{\widehat{G}'_n, \sigma^2 I_2}(X_2) &= \frac{1}{3}\sum_{i=1}^3 \varphi_{\sigma^2 I_2}(X_2 - X_i/2) \\ &= \frac{4\log 2}{9\pi}(2^{-(4/3)(7/4)} + 2^{-(4/3)(1/4)} + 2^{-(4/3)(7/4)}) = \frac{2^{2/3}\log 2}{3\pi} = \hat{L}_2, \end{aligned}$$

and, for $i = 3$,

$$\begin{aligned} f_{\widehat{G}'_n, \sigma^2 I_2}(X_3) &= \frac{1}{3}\sum_{i=1}^3 \varphi_{\sigma^2 I_2}(X_3 - X_i/2) \\ &= \frac{4\log 2}{9\pi}(2^{-(4/3)(7/4)} + 2^{-(4/3)(7/4)} + 2^{-(4/3)(1/4)}) = \frac{2^{2/3}\log 2}{3\pi} = \hat{L}_3. \end{aligned}$$

This verifies that $\widehat{G}'_n = \frac{1}{3}\sum_{i=1}^n \delta_{X_i/2}$ is also an NPMLE, so every convex combination $\alpha\widehat{G}_n + (1-\alpha)\widehat{G}'_n$ is an NPMLE. □

4.6.1.3 Proof of Corollary 4.3

Proof of Corollary 4.3. We have already observed that $\mathcal{Z} \subset \mathcal{M}$ (Ray & Lindsay, 2005). Observe that $\mathcal{M}$ is compact as it is the continuous image of the simplex, a compact set. Since any real-analytic function has a finite number of zeros, $\mathcal{Z}$ is finite. Hence any NPMLE $\widehat{G}_n$ is discrete with a finite number of atoms.

In the proportional covariances case $\Sigma_i = c_i\Sigma$, we have

$$\begin{aligned} x^*(\alpha) &= \left(\sum_{i=1}^n \alpha_i \Sigma_i^{-1}\right)^{-1} \sum_{i=1}^n \alpha_i \Sigma_i^{-1} X_i \\ &= \sum_{i=1}^n \frac{\alpha_i/c_i}{\sum_{\iota=1}^n \alpha_\iota/c_\iota} X_i \end{aligned}$$

As α ranges over the simplex, so does $\left(\frac{\alpha_i/c_i}{\sum_{\iota=1}^n \alpha_\iota/c_\iota}\right)_{i=1}^n$. Thus $\mathcal{M} = \mathrm{conv}(\{X_1, \ldots, X_n\})$, proving (i). If each Σ_i is diagonal, letting $x_j^*(\alpha)$ denote the j^{th} coordinate of $x^*(\alpha) \in \mathbb{R}^d$,

$$x_j^*(\alpha) = \sum_{i=1}^n \frac{\alpha_i(\Sigma_i)_{jj}}{\sum_{i'=1}^n \alpha_{i'}(\Sigma_{i'})_{jj}} X_{ij} \in \left[\min_{i\in\{1,\ldots,n\}} X_{ij}, \max_{i\in\{1,\ldots,n\}} X_{ij}\right],$$

proving (ii). For (iii), using concavity of the minimum eigenvalue,

$$\begin{aligned}\|x^*(\alpha) - x\|_2 &= \left\| \left(\sum_{i=1}^n \alpha_i \Sigma_i^{-1} \right)^{-1} \sum_{i=1}^n \alpha_i \Sigma_i^{-1} (X_i - x) \right\|_2 \\ &\leq \left\| \left(\sum_{i=1}^n \alpha_i \Sigma_i^{-1} \right)^{-1} \right\|_2 \left\| \sum_{i=1}^n \alpha_i \Sigma_i^{-1} (X_i - x) \right\|_2 \\ &\leq \left(\sum_{i=1}^n \alpha_i \overline{k}^{-1} \right)^{-1} \sum_{i=1}^n \alpha_i \underline{k}^{-1} \|X_i - x\|_2 \leq \kappa r\end{aligned}$$

so $\mathcal{M} \subseteq \mathbb{B}_{\kappa r}(x)$. □

4.6.1.4 Proof of Lemma 4.4

Proof of Lemma 4.4. By the change of variables formula,

$$\begin{aligned}f_{T_\# G, \Sigma_i'}(X_i') &= \int \varphi_{U_0 \Sigma_i U_0^\intercal}(U_0 X_i + x_0 - \theta) \, \mathrm{d} T_\# G(\theta) \\ &= \int \varphi_{U_0 \Sigma_i U_0^\intercal}(U_0 X_i + x_0 - T(\theta)) \, \mathrm{d} G(\theta) \\ &= \int \varphi_{\Sigma_i}(X_i - \theta) \, \mathrm{d} G(\theta) = f_{G, \Sigma_i}(X_i),\end{aligned}$$

completing the proof. □

4.6.1.5 Proof of Proposition 4.5

Proof of Proposition 4.5. Write $\widehat{G}_n = \sum_{j=1}^{\hat{k}} \hat{w}_j \delta_{\hat{a}_j}$, and for each $j \in [\hat{k}]$, let $C_j \in \mathcal{H}$ such that $\hat{a}_j \in C_j$. Next, define a positive measure H_j supported on the corners of C_j such that $H_j(C_j) = \hat{w}_j$ and

$$\int_{C_j} u \, \mathrm{d} H_j(u) = \hat{w}_j \hat{a}_j = \int_{C_j} u \, \mathrm{d} \widehat{G}_n^j(u), \tag{4.20}$$

where $\widehat{G}_n^j \coloneqq \hat{w}_j \delta_{\hat{a}_j}$. Now fix $u \in C_j$ and $i \in [n]$, and let $x_j = \Sigma_i^{-1/2}(X_i - \hat{a}_j)$ and $t = \Sigma_i^{-1/2}(u - \hat{a}_j)$. By the moment identity (4.20) and by Jiang and Zhang (2009, A.27),

$$\begin{aligned}&\int_{C_j} \varphi_{\Sigma_i}(X_i - u) \, \mathrm{d} \widehat{G}_n^j(u) - \int_{C_j} \varphi_{\Sigma_i}(X_i - u) \, \mathrm{d} H_j(u) \\ &\qquad \leq \int_{C_j} \langle x_j, t \rangle^2 \varphi_{\Sigma_i}(X_i - u) \, \mathrm{d} \widehat{G}_n^j(u) + \int_{C_j} \left(e^{\|t\|_2^2/2} - 1 \right) \varphi_{\Sigma_i}(X_i - u) \, \mathrm{d} H_j(u) \\ &\qquad \leq \underline{k}^{-2} D^2 d \delta^2 \int_{C_j} \varphi_{\Sigma_i}(X_i - u) \, \mathrm{d} \widehat{G}_n^j(u) + \left(e^{\underline{k} d \delta^2/2} - 1 \right) \int_{C_j} \varphi_{\Sigma_i}(X_i - u) \, \mathrm{d} H_j(u).\end{aligned}$$

Let $H = \sum_{j=1}^{\hat{k}} H_j$. Summing the above inequality over j,

$$f_{\widehat{G}_n,\Sigma_i}(X_i) - f_{H,\Sigma_i}(X_i) \leq \underline{k}^{-2} D^2 d\delta^2 f_{\widehat{G}_n,\Sigma_i}(X_i) + \left(e^{\underline{k} d\delta^2/2} - 1\right) f_{H,\Sigma_i}(X_i).$$

Since H is supported on $\mathcal{A}$, by optimality of $\widehat{G}_n^{\mathcal{A}}$,

$$\prod_{i=1}^{n} f_{\widehat{G}_n^{\mathcal{A}},\Sigma_i}(X_i) \geq \prod_{i=1}^{n} f_{H,\Sigma_i}(X_i).$$

Combining our findings,

$$\prod_{i=1}^{n} f_{\widehat{G}_n^{\mathcal{A}},\Sigma_i}(X_i) \geq e^{-n\underline{k}^{-2} d\delta^2/2} \left(1 - \underline{k}^{-2} D^2 d\delta^2\right)^n \prod_{i=1}^{n} f_{\widehat{G}_n,\Sigma_i}(X_i).$$

Using the elementary inequality $1 - x \geq e^{-2x}$ for $x \leq 3/4$, we obtain

$$\prod_{i=1}^{n} f_{\widehat{G}_n^{\mathcal{A}},\Sigma_i}(X_i) \geq \exp\left(-n\underline{k}^{-2} d\delta^2/2 - 2n\underline{k}^{-2} D^2 d\delta^2\right) \prod_{i=1}^{n} f_{\widehat{G}_n,\Sigma_i}(X_i).$$

for $\delta \leq \sqrt{\frac{3}{4d}} \underline{k} D^{-1}$. □

4.6.2 Proof of Theorem 4.6

The following notation will be used throughout this section:

1. $\mathbb{B}_r(x) = \{y \in \mathbb{R}^d : \|x - y\|_2 \leq r\}$ denotes a closed ball in $\mathbb{R}^d$.
2. For a positive integer m, let $[m] = \{1, \ldots, m\}$.
3. Given a pseudo-metric space (M, ρ) and $\varepsilon > 0$, let $N(\varepsilon, M, \rho)$ denote the ε-covering number, i.e. the smallest positive integer N such that there exist $x_1, \ldots, x_N \in M$ such that
$$M \subset \bigcup_{i=1}^{N} \{y : \rho(y, x_i) \leq \varepsilon\}.$$
Any such a set $\{x_i\}_{i=1}^N$ is known as an ε-net or ε-cover of M under the pseudo-metric ρ. When M is a subset of Euclidean space we write $N(\varepsilon, M)$ instead of $N(\varepsilon, M, \|\cdot\|_2)$.
4. We use the shorthand $f_{G,\bullet} = (f_{G,\Sigma_i})_{i=1}^n$, the matrices $\Sigma_1, \ldots, \Sigma_n$ being viewed as fixed. Let
$$\mathbb{F} = \left\{f_{G,\bullet} : G \in \mathcal{P}(\mathbb{R}^d)\right\}.$$
5. For $S \subset \mathbb{R}^d$ and $M > 0$, S^M denotes the M-enlargement $S^M = \{x \in \mathbb{R}^d : \mathfrak{d}_S(x) \leq M\}$.

6. Define the semi-norm

$$\|f_{G,\bullet} - f_{H,\bullet}\|_{\infty,S^M} \coloneqq \max_{1\le i\le n} \sup_{x\in S^M} |f_{G,\Sigma_i}(x) - f_{H,\Sigma_i}(x)|.$$

Similarly, define

$$\|f_{G,\bullet} - f_{H,\bullet}\|_{\nabla,S^M} \coloneqq \max_{1\le i\le n} \sup_{x\in S^M} |\nabla f_{G,\Sigma_i}(x) - \nabla f_{H,\Sigma_i}(x)|.$$

Our proof generalizes and builds upon prior techniques for analyzing the Hellinger accuracy of the NPMLE (Jiang, 2020; Saha & Guntuboyina, 2020a; Zhang, 2009). The basic structure of our argument is to recognize, given the approximation (4.14) in the likelihood, that we may trivially rewrite the large deviation probability for the NPMLE as a joint probability

$$\mathbb{P}\bigg(\bar{h}(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \gtrsim_{d,\bar{k},\underline{k}} t\varepsilon_n\bigg) = \mathbb{P}\bigg(\bar{h}(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \gtrsim_{d,\bar{k},\underline{k}} t\varepsilon_n, \prod_{i=1}^n \frac{f_{\widehat{G}_n,\Sigma_i}(X_i)}{f_{G^*,\Sigma_i}(X_i)} \ge \exp\left(-c_{d,\bar{k},\underline{k}} n\varepsilon_n^2\right)\bigg).$$

If $\widehat{G}_n$ were a fixed probability measure G_0 such that $\bar{h}(f_{G_0,\bullet}, f_{G^*,\bullet}) \gtrsim_{d,\bar{k},\underline{k}} t\varepsilon_n$, the right-hand side of the last display similarly simplifies as

$$\begin{aligned}
&\mathbb{P}\bigg(\bar{h}(f_{G_0,\bullet}, f_{G^*,\bullet}) \gtrsim_{d,\bar{k},\underline{k}} t\varepsilon_n, \prod_{i=1}^n \frac{f_{G_0,\Sigma_i}(X_i)}{f_{G^*,\Sigma_i}(X_i)} \ge \exp\left(-c_{d,\bar{k},\underline{k}} n\varepsilon_n^2\right)\bigg) \\
&\quad = \mathbb{P}\bigg(\prod_{i=1}^n \frac{f_{G_0,\Sigma_i}(X_i)}{f_{G^*,\Sigma_i}(X_i)} \ge \exp\left(-c_{d,\bar{k},\underline{k}} n\varepsilon_n^2\right)\bigg).
\end{aligned}$$

Since $\widehat{G}_n$ is not fixed, we first approximate it using a covering argument, and then bound the right-hand side of the previous display using Markov's inequality.

Proof of Theorem 4.6. Suppose for some γ_n the NPMLE satisfies

$$\prod_{i=1}^n \frac{f_{\widehat{G}_n,\Sigma_i}(X_i)}{f_{G^*,\Sigma_i}(X_i)} \ge \exp\left((\beta-\alpha)n\gamma_n^2\right) \text{ for some } 0<\beta<\alpha<1.$$

We bound the probability

$$\mathbb{P}\bigg(\bar{h}(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \ge t\gamma_n\bigg)$$

for $t > 1$.

Take $\{f_{H_j,\bullet}\}_{j=1}^N \subset \mathbb{F}$ to be an η-net of $\mathbb{F}$ under $\|\cdot\|_{\infty,S^M}$. For each j, let $H_{0,j}$ be a distribution satisfying

$$\|f_{H_{0,j},\bullet} - f_{H_j,\bullet}\|_{\infty,S^M} \le \eta \text{ and } \bar{h}(f_{H_{0,j},\bullet}, f_{G^*,\bullet}) \ge t\gamma_n$$

and $J = \{j \in [N] : H_{0,j} \text{ exists}\}$. By construction of the η-net, there is $j^* \in [N]$ such that

$$\|f_{H_{j^*},\bullet} - f_{\widehat{G}_n,\bullet}\|_{\infty,S^M} \leq \eta.$$

On the event $\{\bar{h}(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \geq t\gamma_n\}$, the NPMLE $\widehat{G}_n$ acts as a witness that $j^* \in J$, so by the triangle inequality

$$\|f_{H_{0,j^*},\bullet} - f_{\widehat{G}_n,\bullet}\|_{\infty,S^M} \leq 2\eta. \tag{4.21}$$

This gives

$$f_{\widehat{G}_n,\Sigma_i}(x) \leq \begin{cases} f_{H_{0,j^*},\Sigma_i}(x) + 2\eta, & \text{if } x \in S^M \\ \frac{1}{\sqrt{(2\pi)^d|\Sigma_i|}}, & \text{otherwise.} \end{cases}$$

Defining $v(x) = \eta 1_{x\in S^M} + \eta\left(\frac{M}{\mathfrak{d}_S(x)}\right)^{d+1} 1_{x\notin S^M}$, we have

$$\exp((\beta-\alpha)nt^2\gamma_n^2) \leq \max_{j\in J}\left[\prod_{i=1}^n \frac{f_{H_{0,j},\Sigma_i}(X_i) + 2v(X_i)}{f_{G^*,\Sigma_i}(X_i)}\right] \cdot \left[\prod_{i:X_i\notin S^M} \frac{1}{\sqrt{(2\pi)^d|\Sigma_i|}\cdot 2v(X_i)}\right] \tag{4.22}$$

on the event $\{\bar{h}(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \geq t\gamma_n\}$. Hence

$$\mathbb{P}\left(\bar{h}(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \geq t\gamma_n\right) \tag{4.23}$$

$$\leq \mathbb{P}\left(\max_{j\in J}\prod_{i=1}^n \frac{f_{H_{0,j},\Sigma_i}(X_i) + 2v(X_i)}{f_{G^*,\Sigma_i}(X_i)} \geq \exp(-\alpha nt^2\gamma_n^2)\right) \tag{4.24}$$

$$+ \mathbb{P}\left(\prod_{i:X_i\notin S^M} \frac{1}{\sqrt{(2\pi)^d|\Sigma_i|}\cdot 2v(X_i)} \geq \exp(\beta nt^2\gamma_n^2)\right) \tag{4.25}$$

By a union bound and Markov's inequality, the first term (4.24) is bounded by

$$e^{\alpha nt^2\gamma_n^2/2}\sum_{j\in J}\mathbb{E}\prod_{i=1}^n \sqrt{\frac{f_{H_{0,j},\Sigma_i}(X_i) + 2v(X_i)}{f_{G^*,\Sigma_i}(X_i)}} \tag{4.26}$$

Writing out the expectation,

$$\begin{aligned}\prod_{i=1}^n \mathbb{E}\sqrt{\frac{f_{H_{0,j},\Sigma_i}(X_i) + 2v(X_i)}{f_{G^*,\Sigma_i}(X_i)}} &= \exp\left(\sum_{i=1}^n \log\mathbb{E}\sqrt{\frac{f_{H_{0,j},\Sigma_i}(X_i) + 2v(X_i)}{f_{G^*,\Sigma_i}(X_i)}}\right) \\ &\leq \exp\left(\sum_{i=1}^n\left\{\int\sqrt{f_{H_{0,j},\Sigma_i} + 2v}\sqrt{f_{G^*,\Sigma_i}} - 1\right\}\right) \\ &\leq \exp\left(-\frac{nt^2\gamma_n^2}{2} + n\sqrt{2\int v}\right)\end{aligned}$$

Putting together the pieces, the first term (4.24) is bounded by

$$\begin{aligned}\mathbb{P}\bigg(\max_{j\in J}\prod_{i=1}^{n}\frac{f_{H_{0,j},\Sigma_i}(X_i)+2v(X_i)}{f_{G^*,\Sigma_i}(X_i)}\geq e^{-\alpha nt^2\gamma_n^2}\bigg)\\ \leq\exp\left(-(1-\alpha)\,\frac{nt^2\gamma_n^2}{2}+\log N+n\sqrt{2\int v}\right)\end{aligned}\tag{4.27}$$

For the second term (4.25), observe by Markov's inequality

$$\begin{aligned}&\mathbb{P}\bigg(\prod_{i:X_i\notin S^M}\frac{1}{\sqrt{(2\pi)^d|\Sigma_i|}\cdot 2v(X_i)}\geq\exp(\beta nt^2\gamma_n^2)\bigg)\\ &\leq\exp\left(-\frac{\beta nt^2\gamma_n^2}{2\log n}\right)\mathbb{E}\left\{\prod_{i:X_i\notin S^M}\left|\frac{1}{\sqrt{|2\pi\Sigma_i|}\cdot 2v(X_i)}\right|\right\}^{1/2\log n}\\ &=\exp\left(-\frac{\beta nt^2\gamma_n^2}{2\log n}\right)\mathbb{E}\left\{\prod_{i=1}^{n}\left(\frac{\mathfrak{d}_S(X_i)}{|2\pi\Sigma_i|^{1/(2d+2)}\cdot(2\eta)^{1/(d+1)}M}\right)^{1_{\mathfrak{d}_S(X_i)\geq M)}}\right\}^{(d+1)/2\log n}\end{aligned}$$

To reduce clutter write $a=\frac{1}{\underline{k}^{d/(2d+2)}\eta^{1/(d+1)}M}$ and $\lambda=\frac{d+1}{2\log n}$. The above expectation is further upper bounded by

$$\begin{aligned}\mathbb{E}\left\{\prod_{i=1}^{n}(a\mathfrak{d}_S(X_i))^{1_{\mathfrak{d}_S(X_i)\geq M}}\right\}^{\lambda}&=\prod_{i=1}^{n}\mathbb{E}\,(a\mathfrak{d}_S(X_i))^{\lambda 1_{\mathfrak{d}_S(X_i)\geq M}}\\ &\leq\prod_{i=1}^{n}\left(1+a^\lambda\mathbb{E}\left[\mathfrak{d}_S(X_i)^\lambda 1_{\mathfrak{d}_S(X_i)\geq M}\right]\right)\\ &\leq\exp\left(a^\lambda\sum_{i=1}^{n}\mathbb{E}\left[\mathfrak{d}_S(X_i)^\lambda 1_{\mathfrak{d}_S(X_i)\geq M}\right]\right)\\ &\leq\exp\left(na^\lambda\left\{C_dM^{d+\lambda-2}\overline{k}^{1-d/2}e^{-M^2/(8\overline{k})}+M^\lambda\left(\frac{2\mu_q}{M}\right)^q\right\}\right)\end{aligned}$$

The last inequality follows from Lemma 4.13. Note we need

$$\frac{d+1}{2(1\wedge q)}\leq\log n,$$

to ensure $\lambda\leq 1\wedge q$. Taking $M\geq\sqrt{8\overline{k}\log n}$, we have $e^{-M^2/(8\overline{k})}\leq\frac{1}{n}$, so

$$\mathbb{E}\left\{\prod_{i=1}^{n}(a\mathfrak{d}_S(X_i))^{1_{\mathfrak{d}_S(X_i)\geq M}}\right\}^{\lambda}\leq\exp\left((aM)^\lambda\left[C_dM^{d-2}\overline{k}^{1-d/2}+n\left(\frac{2\mu_q}{M}\right)^q\right]\right)$$

Noting $(aM)^\lambda = \left(\underline{k}^{d/2}\eta\right)^{-1/(2\log n)}$, choose $\eta = \frac{n^{-2}}{\underline{k}^{d/2}}$, so $(aM)^\lambda = e$. We directly apply Lemma A.7 of Saha and Guntuboyina (2020b) for the integral

$$\int v \leq C_d \eta \mathrm{Vol}(S^M).$$

To bound the metric entropy, i.e. $\log N$ where N denotes the size of our η-net $\{f_{H_j,\bullet}\}_{j=1}^N \subset \mathbb{F}$, we apply Lemma 4.16

$$\log N = \log N\left(\eta, \mathbb{F}, \|\cdot\|_{\infty, S^M}\right) \leq C_d N\left(u, (S^M)^u\right)\left(\log \frac{c_{d,\overline{k},\underline{k}}}{\eta}\right)^2,$$

where the scalar u in the above display corresponds to a used in the lemma. Assuming $4n \geq (2\pi)^{d/2}$,

$$u = \sqrt{-2\overline{k}\log\left(((2\pi\underline{k})^{d/2}\frac{\eta}{4}\right)} \geq \sqrt{2\overline{k}\log n}$$

Similarly $u \leq \sqrt{6\overline{k}\log n}$, so

$$N\left(u, (S^M)^u\right) \leq N\left(\sqrt{2\overline{k}\log n}, (S^M)^{\sqrt{6\overline{k}\log n}}\right) \leq C_{d,\overline{k}}\mathrm{Vol}(S^{2M})(\log n)^{-d/2}$$

Combining our findings,

$$\begin{aligned}
&\mathbb{P}\left(\bar{h}(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \geq t\gamma_n\right) \\
&\quad\leq \exp\left(-(1-\alpha)\frac{nt^2\gamma_n^2}{2} + C_{d,\overline{k},\underline{k}}(\log n)^{d/2+1}\mathrm{Vol}(S^{2M}) + C_d\sqrt{\underline{k}^{-d/2}\mathrm{Vol}(S^M)}\right) \\
&\qquad+ \exp\left(-\frac{\beta}{\log n}\frac{nt^2\gamma_n^2}{2} + C_d M^{d-2}\overline{k}^{1-d/2} + en \inf_{q \geq (d+1)/(2\log n)}\left(\frac{2\mu_q}{M}\right)^q\right)
\end{aligned}$$

for any $t > 1$. Absorbing the dependence on $d, \underline{k}$ and $\overline{k}$ into constants, take $\varepsilon_n^2 = \varepsilon_n^2(M, S, G^*)$ such that

$$\begin{aligned}
&\max\left\{(\log n)^{d/2+1}\mathrm{Vol}(S^{2M}), \sqrt{\mathrm{Vol}(S^M)}, M^{d-2}, en \inf_{q \geq (d+1)/(2\log n)}\left(\frac{2\mu_q}{M}\right)^q\right\} \\
&\quad\lesssim_{d,\overline{k},\underline{k}} n\varepsilon_n^2(M, S, G^*)
\end{aligned}$$

If we then take $\gamma_n^2 = \frac{C_{d,\overline{k},\underline{k}}\varepsilon_n^2(M,S,G^*)}{4\min(1-\alpha,\beta)}$,

$$\mathbb{P}\left(\bar{h}(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}) \geq t\gamma_n\right) \leq 2\exp\left(-\frac{(1-\alpha)\wedge\beta}{4\log n}nt^2\gamma_n^2\right) \qquad \square$$

This proves (4.15). To prove (4.16), integrate the tail from (4.15),

$$\begin{aligned}\mathbb{E}\frac{\bar{h}^2(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet})}{\gamma_n^2} &\le 1 + \int_1^\infty \mathbb{P}\left(\frac{\bar{h}^2(\widehat{G}_n, G^*)}{\gamma_n^2} \ge s\right) \mathrm{d}s \\ &\le 1 + \int_1^\infty 4tn^{-t^2}\,\mathrm{d}t = 1 + \frac{2}{n\log n} \le 3\end{aligned}$$

for $n > 1$, completing the proof.

We now state and prove the lemmas needed in the proof of Theorem 4.6.

Lemma 4.13. *Let $\theta^* \sim G^*$ and $Z \sim \mathcal{N}(0, I_d)$ independently, and $Y = \theta^* + \Sigma^{1/2}Z$, where $\underline{k} I_d \preceq \Sigma \preceq \overline{k} I_d$. Then*

$$\mathbb{E}\left[\mathfrak{d}_S(Y)^\lambda 1_{\mathfrak{d}_S(Y)\ge M}\right] \le C_d M^{d+\lambda-2}\overline{k}^{1-d/2} e^{-M^2/(8\overline{k})} + M^\lambda \left(\frac{2\mu_q}{M}\right)^q,$$

for any $\lambda \in (0, 1\wedge q]$, where μ_q is the q^{th}-moment of $\mathfrak{d}_S(\theta^*)$ under $\theta^* \sim G^*$.

Proof. Since distance $\mathfrak{d}_S$ is 1-Lipschitz,

$$\mathbb{E}\left[\mathfrak{d}_S(Y)^\lambda 1_{\mathfrak{d}_S(Y)\ge M}\right] \le \mathbb{E}\left[(2\|\Sigma^{1/2}Z\|_2)^\lambda 1_{2\|\Sigma^{1/2}Z\|_2\ge M}\right] + \mathbb{E}\left[(2\mathfrak{d}_S(\theta^*))^\lambda 1_{2\mathfrak{d}_S(\theta^*)\ge M}\right] \tag{4.28}$$

For the first term on the RHS of (4.28),

$$\begin{aligned}\mathbb{E}\left[(2\|\Sigma^{1/2}Z\|_2)^\lambda 1_{2\|\Sigma^{1/2}Z\|_2\ge M}\right] &\le M^\lambda \mathbb{E}\left[\left(\frac{\|\Sigma^{1/2}Z\|_2}{M/2}\right)^\lambda 1_{\|\Sigma^{1/2}Z\|_2\ge M/2}\right] \\ &\le 2M^{\lambda-1}\mathbb{E}\left[\|\Sigma^{1/2}Z\|_2 1_{\|\Sigma^{1/2}Z\|_2\ge M/2}\right] \\ &\le 2M^{\lambda-1}\overline{k}^{1/2}\mathbb{E}\left[\|Z\|_2 1_{\|Z\|_2\ge M/(2\overline{k}^{1/2})}\right] \\ &\le 2C_d M^{\lambda-1}\overline{k}^{1/2}\left(\frac{M}{\overline{k}^{1/2}}\right)^{d-1} e^{-M^2/(8\overline{k})} \\ &= C_d M^{d+\lambda-2}\overline{k}^{1-d/2} e^{-M^2/(8\overline{k})}\end{aligned}$$

The penultimate inequality uses $\|\Sigma^{1/2}Z\|_2 \le \overline{k}^{1/2}\|Z\|_2$, and the last inequality directly uses Lemma A.6 of Saha and Guntuboyina (2020b).

Since $\lambda < q$, applying Hölder to the second term on the RHS of (4.28) yields

$$\mathbb{E}\left[(2\mathfrak{d}_S(\theta^*))^\lambda 1_{2\mathfrak{d}_S(\theta^*)\ge M}\right] \le M^\lambda \left(\frac{2\mu_q}{M}\right)^q \qquad \square$$

Lemma 4.14. *(Moment matching, part i) Let $G, H \in \mathcal{P}(\mathbb{R}^d)$. Suppose $A \subset \mathbb{R}^d$ is such that*

$$\mathbb{B}_a(x) \subseteq A \subseteq \mathbb{B}_{ca}(x)$$

for some $c \geq 1$, and that

$$\int_A \theta_1^{k_1} \cdots \theta_d^{k_d} dG(\theta) = \int_A \theta_1^{k_1} \cdots \theta_d^{k_d} dH(\theta), \text{ for } k_1, \ldots, k_d \in [2m+1],$$

for some $m \geq 1$. Then

$$\max_{1 \leq i \leq n} |f_{G,\Sigma_i}(x) - f_{H,\Sigma_i}(x)| \leq \frac{1}{(2\pi \underline{k})^{d/2}} \left(\frac{ec^2a^2}{2\overline{k}(m+1)} \right)^{m+1} + \frac{e^{-a^2/(2\overline{k})}}{(2\pi \underline{k})^{d/2}}.$$

Proof. For each $i \in [n]$, write

$$f_{G,\Sigma_i}(x) - f_{H,\Sigma_i}(x) = \int_A \varphi_{\Sigma_i}(x-\theta)(dG(\theta) - dH(\theta)) + \int_{A^{\mathsf{c}}} \varphi_{\Sigma_i}(x-\theta)(dG(\theta) - dH(\theta))$$

On A^{c}, $\|x - \theta\|_2 \geq a$, so

$$\varphi_{\Sigma_i}(x-\theta) \leq \frac{e^{-a^2/(2\overline{k})}}{(2\pi \underline{k})^{d/2}}.$$

Write the pdf as $\varphi_{\Sigma_i}(z) = P_i(z) + R_i(z)$ where P_i is a polynomial of degree $2m$ and the remainder R_i satisfies

$$|R_i(z)| \leq (2\pi \underline{k})^{-d/2} \left(\frac{e\|z\|_2^2}{2\overline{k}(m+1)} \right)^{m+1}$$

By hypothesis, $\int_A P_i(x-\theta)(\mathrm{d}G(\theta) - \mathrm{d}H(\theta))$, so

$$\begin{aligned} \left| \int_A \varphi_{\Sigma_i}(x-\theta)(\mathrm{d}G(\theta) - \mathrm{d}H(\theta)) \right| &\leq \left| \int_A R_i(x-\theta)(\mathrm{d}G(\theta) - \mathrm{d}H(\theta)) \right| \\ &\leq \frac{1}{(2\pi \underline{k})^{d/2}} \left(\frac{ec^2a^2}{2\overline{k}(m+1)} \right)^{m+1} \end{aligned}$$

completing the proof. □

Lemma 4.15. *(Moment matching, part ii) For any $G \in \mathcal{P}(\mathbb{R}^d)$, there is a discrete distribution H supported on S^a with at most*

$$l \coloneqq (2\lfloor 13.5a^2/\overline{k} \rfloor + 2)^d N(a, S^a) + 1$$

atoms such that

$$\|f_{G,\bullet} - f_{H,\bullet}\|_{\infty, S^a} \leq \left(1 + \frac{1}{\sqrt{2\pi}} \right) (2\pi \underline{k})^{-d/2} e^{-a^2/(2\overline{k})}.$$

Proof. The idea is to choose H to match moments, and then apply the previous lemma. The proof is identical to Lemma D.3 of Saha and Guntuboyina (2020b), except that we take $m \coloneqq \lfloor \frac{27a^2}{2\overline{k}} \rfloor$. □

Lemma 4.16. *There exists positive constants C_d and $c_{d,\overline{k},\underline{k}}$ depending on $d, \overline{k}, \underline{k}$ alone such that for every compact set $S \subset \mathbb{R}^d$, $M > 0$ and $\eta \in (0, e^{-1} \wedge 4(2\pi\underline{k})^{-d/2})$, we have*

$$\log N(\eta, \mathbb{F}, \|\cdot\|_{\infty,S}) \leq C_d N(a, S^a) \left(\log \frac{c_{d,\overline{k},\underline{k}}}{\eta} \right)^{d+1} \tag{4.29}$$

Proof. The idea here is to take $f_{G,\bullet} \in \mathbb{F}$ (induced by some $G \in \mathcal{P}(\mathbb{R}^d)$), approximate G by a discrete distribution H, and then further approximate that discrete distribution with another discrete distribution over a fixed set of atoms and weights. So let $G \in \mathcal{P}(\mathbb{R}^d)$, and apply the previous Lemma to obtain a discrete distribution H supported on S^a with at most l atoms such that

$$\|f_{G,\bullet} - f_{H,\bullet}\|_{\infty,S^a} \leq \left(1 + \frac{1}{\sqrt{2\pi}}\right) (2\pi\underline{k})^{-d/2} e^{-a^2/(2\overline{k})}.$$

Let $\mathcal{C}$ denote a minimal ζ-net of S^a, and let H' approximate each atom of H with its closest element from $\mathcal{C}$. Writing $H = \sum_j w_j \delta_{a_j}$ and $H' = \sum w_j \delta_{b_j}$, we have

$$\begin{aligned}
\|f_{H,\bullet} - f_{H',\bullet}\|_{\infty,S^a} &= \max_{i\in[n]} \sup_{x\in S^a} |f_{H,\Sigma_i}(x) - f_{H',\Sigma_i}(x)| \\
&\leq \max_{i\in[n]} \sup_{x\in S^a} \sum_j w_j |\varphi_{\Sigma_i}(x - a_j) - \varphi_{\Sigma_i}(x - b_j)| \\
&\leq \zeta \max_{i\in[n]} \sup_z \|\nabla \varphi_{\Sigma_i}(z)\|_2 = \zeta \max_{i\in[n]} \sup_z \varphi_{\Sigma_i}(z) \|\Sigma_i^{-1} z\|_2 \\
&\leq \zeta \underline{k}^{-1} (2\pi\underline{k})^{-d/2} \max_{i\in[n]} \sup_t \exp\left(-t^2/2\overline{k}\right) t \\
&\leq \zeta \underline{k}^{-1} (2\pi\underline{k})^{-d/2} (\overline{k}/e)^{1/2}
\end{aligned}$$

Let $\mathcal{D}$ denote a minimal ξ-net of Δ^{l-1} in the ℓ_1 norm, and approximate the weights w by their closest element $v \in \mathcal{D}$. Writing $H'' = \sum_j v_j \delta_{b_j}$,

$$\begin{aligned}
\|f_{H',\bullet} - f_{H'',\bullet}\|_{\infty,S^a} &= \max_{i\in[n]} \sup_{x\in S^a} |f_{H',\Sigma_i}(x) - f_{H'',\Sigma_i}(x)| \\
&\leq \max_{i\in[n]} \sup_{x\in S^a} \sum_j |w_j - v_j| |\varphi_{\Sigma_i}(x - b_j)| \leq (2\pi\underline{k})^{-d/2} \xi.
\end{aligned}$$

Applying triangle inequality to the past three displays,

$$\|f_{G,\bullet} - f_{H'',\bullet}\|_{\infty,S^a} \leq (2\pi\underline{k})^{-d/2} \left[2e^{-a^2/(2\overline{k})} + \zeta \underline{k}^{-1} (\overline{k}/e)^{1/2} + \xi \right] \tag{4.30}$$

Letting $\xi = (2\pi\underline{k})^{d/2}\frac{\eta}{4}$, $\zeta = \xi\underline{k}(\overline{k}/e)^{-1/2}$, and $a = \sqrt{2\overline{k}\log\xi^{-1}}$ yields $\|f_{G,\bullet} - f_{H'',\bullet}\|_{\infty,S^a} \leq \eta$. In order to take a as such we need $\xi < 1$, or equivalently $\eta < 4(2\pi\underline{k})^{-d/2}$.

The number of possible H'' is

$$|\mathcal{C}| \cdot |\mathcal{D}| = N(\xi, \Delta^{l-1})\binom{N(\zeta, S^a)}{l} \leq \left[\left(1 + \frac{2}{\xi}\right)\frac{eN(\zeta, S^a)}{l}\right]^l$$

From the previous Lemma, $l \geq N(a, S^a)/\overline{k}^d$, so

$$\frac{N(\zeta, S^a)}{l} \leq \overline{k}^d\frac{N(\zeta, S^a)}{N(a, S^a)} \leq \overline{k}^d\left(1 + \frac{a}{\zeta}\right)^d = \overline{k}^d\left(1 + \kappa\frac{2}{\sqrt{e}\xi^{3/2}}\right)^d \leq C_d\frac{\overline{k}^{2d}}{\underline{k}^{d+3d^2/4}}\left(\frac{1}{\eta}\right)^{3p/2}$$

Thus,

$$\log N(\eta, \mathbb{F}, \|\cdot\|_{\infty,S}) \leq C_d N(a, S^a)\log\left(\frac{1}{\underline{k}^{d/2}}\left(\frac{\overline{k}^{2d}}{\underline{k}^{3(d/2+1)}} \vee 1\right)\frac{e}{\eta}\right)^{d+1} \qquad \square$$

4.6.3 Proof of Theorem 4.8

Throughout the proof we will group sequences of the form $\theta_1, \ldots, \theta_n$ into $n \times d$ matrices θ, so that, for instance, the regret $\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\hat{\theta}_i - \hat{\theta}_i^*\|_2^2$ in the statement of the theorem may be rewritten as the expected squared Frobenius norm $\frac{1}{n}\mathbb{E}\|\hat{\theta} - \hat{\theta}^*\|_F^2$, where $\|\theta\|_F^2 = \sum_{i=1}^n\sum_{j=1}^d \theta_{ij}^2$. Additionally, we use the same notation introduced at the start of Section 4.6.2.

4.6.3.1 Regularizing the Bayes rule

In evaluating $\hat{\theta}$, an apparent difficulty is that the denominator in Tweedie's formula can be arbitrarily small. However, since $\widehat{G}_n$ is an approximate NPMLE, we show that the likelihood is lower bounded at each of the observations. In accordance with (4.17), we write that

$$\frac{1}{n}\sum_{i=1}^n \log f_{\widehat{G}_n,\Sigma_i}(X_i) - \sup_{G\in\mathcal{P}(\mathbb{R}^d)}\frac{1}{n}\sum_{i=1}^n \log f_{G,\Sigma_i}(X_i) \geq -q, \tag{4.31}$$

for some $q > 0$. Following Jiang and Zhang, 2009, Proof of Proposition 2, for a fixed $j \in [n]$ choose $G = n^{-1}\delta_{X_j} + (1 - n^{-1})\widehat{G}_n$. Then by the previous display,

$$\begin{aligned}\prod_{i=1}^n f_{\widehat{G}_n,\Sigma_i}(X_i) &\geq e^{-nq}\prod_{i=1}^n f_{G,\Sigma_i}(X_i) \\ &\geq e^{-nq}n^{-1}\varphi_{\Sigma_i}(0)(1 - n^{-1})^{n-1}\prod_{i:i\neq j} f_{\widehat{G}_n,\Sigma_i}(X_i).\end{aligned}$$

Cancelling terms for $i \in [n] \setminus \{j\}$, we conclude

$$f_{\widehat{G}_n,\Sigma_j}(X_j) \geq e^{-nq-\log n}\frac{1}{e\sqrt{|2\pi\Sigma_j|}}. \tag{4.32}$$

Given this, it is natural to define the regularized empirical Bayes and oracle Bayes rules

$$\hat{\theta}_{\rho,i} = X_i + \Sigma_i \frac{\nabla f_{\widehat{G}_n,\Sigma_i}(X_i)}{f_{\widehat{G}_n,\Sigma_i}(X_i) \vee (\rho/\sqrt{|\Sigma_i|})} \tag{4.33}$$

$$\hat{\theta}^*_{\rho,i} = X_i + \Sigma_i \frac{\nabla f_{G^*,\Sigma_i}(X_i)}{f_{G^*,\Sigma_i}(X_i) \vee (\rho/\sqrt{|\Sigma_i|})}. \tag{4.34}$$

By the lower bound (4.32) we know that $\hat{\theta}_\rho = \hat{\theta}$ when $\rho \le \rho_0 \coloneqq e^{-nq-\log n}\frac{1}{e(2\pi)^{d/2}}$. In particular,

$$\|\hat{\theta} - \hat{\theta}^*\|_F = \|\hat{\theta}_\rho - \hat{\theta}^*\|_F \le \|\hat{\theta}_\rho - \hat{\theta}^*_\rho\|_F + \|\hat{\theta}^*_\rho - \hat{\theta}^*\|_F. \tag{4.35}$$

The first term $\|\hat{\theta}_\rho - \hat{\theta}^*_\rho\|_F$ represents the regret between regularized rules, which prevents the denominator in Tweedie's formula from blowing up. The second term represents the cost of introducing a small amount of regularization in the oracle Bayes rule.

4.6.3.2 Regularization error of oracle Bayes

Let us first consider the second term $\|\hat{\theta}^*_\rho - \hat{\theta}^*\|_F$ on the RHS of the bound (4.35). Fixing $i \in [n]$, let G^*_i denote the distribution of $\xi_i = \Sigma_i^{-1/2}\theta^*_i$ where $\theta^*_i \sim G^*$. Then we may write $X_i = \Sigma_i^{1/2}X_i$ where $X_i \sim f_{G^*_i,I_d}$. Note how the scale change affects the terms in Tweedie's formula:

$$\begin{aligned}
f_{G^*,\Sigma_i}(X_i) &= \mathbb{E}_{\vartheta_i\sim G^*}\left[\frac{1}{\sqrt{|2\pi\Sigma_i|}}\exp\left(-\frac{1}{2}(X_i-\vartheta_i)'\Sigma_i^{-1}(X_i-\vartheta_i)\right)\right] \\
&= \frac{1}{\sqrt{|\Sigma_i|}}\mathbb{E}_{\xi_i\sim G^*_i}\left[\varphi_{I_d}(\Sigma_i^{-1/2}X_i - \xi_i)\right] = \frac{1}{\sqrt{|\Sigma_i|}} f_{G^*_i,I_d}(X_i) \\
\nabla f_{G^*,\Sigma_i}(X_i) &= \mathbb{E}_{\vartheta_i\sim G^*}\left[\Sigma_i^{-1}(\vartheta_i - X_i)\frac{1}{\sqrt{|2\pi\Sigma_i|}}\exp\left(-\frac{1}{2}(X_i-\vartheta_i)'\Sigma_i^{-1}(X_i-\vartheta_i)\right)\right] \\
&= \frac{1}{\sqrt{|\Sigma_i|}}\Sigma_i^{-1/2}\nabla f_{G^*_i,I_d}(X_i)
\end{aligned} \tag{4.36}$$

In particular, Tweedie's formula, even in its regularized form, is scale equivariant:

$$\hat{\theta}^*_{\rho,i} = \Sigma_i^{1/2}\left(X_i + \frac{\nabla f_{G^*_i,I_d}(X_i)}{f_{G^*_i,I_d}(X_i)\vee\rho}\right) \tag{4.37}$$

In this form, Saha and Guntuboyina (2020a, Lemma 4.3) directly applies. Specifically, defining

$$\Delta(G,\rho) \coloneqq \int \left(1 - \frac{f_{G,I_d}}{f_{G,I_d}\vee\rho}\right)^2 \frac{\|\nabla f_{G,I_d}\|_2^2}{f_{G,I_d}},$$

for any $\rho \leq \rho_0$ and for all compact sets $S_1, \ldots, S_n \subset \mathbb{R}^d$,

$$\begin{aligned}\mathbb{E}\|\hat{\theta}^*_\rho - \hat{\theta}^*\|_F^2 &= \sum_{i=1}^n \mathbb{E}\|\hat{\theta}^*_{\rho,i} - \hat{\theta}^*_i\|_2^2 \leq \bar{k}\sum_{i=1}^n \Delta(G^*_i, \rho) \\ &\leq \bar{k}\sum_{i=1}^n \left\{ C_d N\left(\frac{4}{L(\rho)}, S_i\right) L^d(\rho)\rho + dG^*_i(S_i^{\mathsf{c}}) \right\},\end{aligned} \tag{4.38}$$

where $L(\rho) \coloneqq \sqrt{-\log((2\pi)^d\rho^2)}$ and N denotes the usual covering number in the Euclidean norm. Choosing $\rho = (2\pi)^{-d/2}/n$ and $S_i = \Sigma_i^{-1/2}S^M$,

$$\mathbb{E}\|\hat{\theta}^*_\rho - \hat{\theta}^*\|_F^2 \leq \bar{k}n\left\{ C_d N\left(\frac{4}{\sqrt{\log n}}, \Sigma_i^{-1/2}S^M\right)\frac{(\log n)^{d/2}}{n} + dG^*((S^M)^{\mathsf{c}}) \right\}.$$

Let $x_1, \ldots, x_m$ denote a t-net of S^M. Let $y \in S_i$ and $x = \Sigma_i^{1/2}y$. There is some j s.t. $\|x_j - x\|_2 \leq t$. Let $y_j = \Sigma_i^{-1/2}x_j$. Then

$$t^2 \geq (x_j - x)'(x_j - x) = (y_j - y)'\Sigma_i(y_j - y) \geq \underline{k}\|y_j - y\|_2^2$$

so $y_1, \ldots, y_m$ is a $t/\underline{k}^{1/2}$-net of S_i. This shows $N(t/\underline{k}^{1/2}, S_i) \leq N(t, S^M)$. By Saha and Guntuboyina (2020b, Lemma F.6) and Markov's inequality,

$$\mathbb{E}\|\hat{\theta}^*_\rho - \hat{\theta}^*\|_F^2 \leq C_d\bar{k}n\left\{ \underline{k}^{-d/2}\mathrm{Vol}\left(S^1\right)M^d\frac{(\log n)^d}{n} + \inf_{q \geq (d+1)/2\log n}\left(\frac{2\mu_q}{M}\right)^q \right\}. \tag{4.39}$$

4.6.3.3 Regret of regularized rules

Now we consider the first term $\|\hat{\theta}_\rho - \hat{\theta}^*_\rho\|_F$ on the RHS of the bound (4.35). First, we will introduce some additional notation. For $\delta > 0$ let $A_\delta = \left\{\bar{h}^2\left(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}\right) \leq \delta\right\}$. Given a compact set $S \subset \mathbb{R}^d$, define another metric

$$m^S(G, G') \coloneqq \max_{i\in[n]} \sup_{x:\mathfrak{d}_S(x)\leq M} \left\| \frac{\Sigma_i\nabla f_{G,\Sigma_i}(x)}{f_{G,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} - \frac{\Sigma_i\nabla f_{G',\Sigma_i}(x)}{f_{G',\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} \right\|_2$$

Let $G^{(1)}, \ldots, G^{(N)}$ denote a minimal η^*-covering of $\left\{G : \bar{h}^2(f_{G,\bullet}, f_{G^*,\bullet}) \leq \delta\right\}$ in the metric m^S. For $j \in [N]$ similarly define an $n \times d$ matrix $\hat{\theta}^{(j)}_\rho$ where the i^{th} row is given by $X_i + \Sigma_i\frac{\nabla f_{G^{(j)},\Sigma_i}(X_i)}{f_{G^{(j)},\Sigma_i}(X_i)\vee(\rho/\sqrt{|\Sigma_i|})}$. We bound the regret as $\|\hat{\theta}_\rho - \hat{\theta}^*_\rho\|_F \leq \sum_{t=1}^4 \zeta_t$, where

$$\begin{aligned}\zeta_1 &\coloneqq \|\hat{\theta}_\rho - \hat{\theta}^*_\rho\|_F 1_{A_\delta^{\mathsf{c}}} \\ \zeta_2 &\coloneqq \left(\|\hat{\theta}_\rho - \hat{\theta}^*_\rho\|_F - \max_{j\in[N]}\|\hat{\theta}^{(j)}_\rho - \hat{\theta}^*_\rho\|_F\right)_+ 1_{A_\delta} \\ \zeta_3 &\coloneqq \max_{j\in[N]}\left(\|\hat{\theta}^{(j)}_\rho - \hat{\theta}^*_\rho\|_F - \mathbb{E}\|\hat{\theta}^{(j)}_\rho - \hat{\theta}^*_\rho\|_F\right)_+ \\ \zeta_4 &\coloneqq \max_{j\in[N]}\mathbb{E}\|\hat{\theta}^{(j)}_\rho - \hat{\theta}^*_\rho\|_F\end{aligned} \tag{4.40}$$

We will control the second moment of each ζ_t. Here's our rough overview. ζ_1 uses Theorem 4.6 to show the NPMLE places small probability on A_δ^c; ζ_2 uses the fact that (on A_δ) the cover $\{G^{(j)}\}$ must have some element that is close to the NPMLE in m^S; ζ_3 follows from Gaussian concentration of measure; and ζ_4 bounds each expectation individually and uses closeness in Hellinger.

Bounding $\mathbb{E}\zeta_1^2$

By the scaled Tweedie's formula (4.37)

$$\|\hat{\theta}_{\rho,i} - \hat{\theta}^*_{\rho,i}\|_2^2 \le \bar{k} \left\| \frac{\nabla f_{\widehat{G}_i, I_d}(X_i)}{f_{\widehat{G}_i, I_d}(X_i) \vee \rho} - \frac{\nabla f_{G_i^*, I_d}(X_i)}{f_{G_i^*, I_d}(X_i) \vee \rho} \right\|_2^2$$

Saha and Guntuboyina (2020b, Lemma F.1) provides

$$\mathbb{E}\zeta_1^2 \le 4\bar{k} n \log\left(\frac{(2\pi)^d}{\rho^2}\right) \mathbb{P}(A_\delta^\mathsf{c}). \tag{4.41}$$

By Theorem 4.6, there is a constant $C_{d,\underline{k},\bar{k}} > 0$ such that $\delta = C_{d,\underline{k},\bar{k}} \varepsilon_n^2(M, S, G^*)$ satisfies $\mathbb{P}(A_\delta^\mathsf{c}) \le 2/n$. Hence

$$\mathbb{E}\zeta_1^2 \le 48\bar{k} \log(n). \tag{4.42}$$

Bounding $\mathbb{E}\zeta_2^2$

Observe

$$\begin{aligned}
\zeta_2^2 &\le 1_{A_\delta} \min_{j \in [N]} \|\hat{\theta}_\rho - \hat{\theta}_\rho^{(j)}\|_F^2 \\
&= 1_{A_\delta} \min_{j \in [N]} \sum_{i=1}^n \left\| \frac{\Sigma_i \nabla f_{\widehat{G}_n, \Sigma_i}(X_i)}{f_{\widehat{G}_n, \Sigma_i}(X_i) \vee (\rho/\sqrt{|\Sigma_i|})} - \frac{\Sigma_i \nabla f_{G^{(j)}, \Sigma_i}(X_i)}{f_{G^{(j)}, \Sigma_i}(X_i) \vee (\rho/\sqrt{|\Sigma_i|})} \right\|_2^2
\end{aligned}$$

On A_δ, we may take j such that $m^S(\widehat{G}_n, G^{(j)}) \le \eta^*$. For each i, consider two cases, where $X_i \in S^M$ and where $X_i \notin S^M$. When $X_i \in S^M$ bound the above $\|\cdot\|_2$ by the supremum over all $x \in S^M$. When $X_i \notin S^M$ bound the regularized rules as before. This yields

$$\zeta_2^2 \le 1_{A_\delta} \left(\#\{i : X_i \in S^M\} (\eta^*)^2 + \#\{i : X_i \notin S^M\} 4\bar{k} \log\left(\frac{(2\pi)^d}{\rho^2}\right) \right) \tag{4.43}$$

so in particular

$$\mathbb{E}\zeta_2^2 \le n(\eta^*)^2 + 4\bar{k} \log\left(\frac{(2\pi)^d}{\rho^2}\right) \sum_{i=1}^n \mathbb{P}\left(\mathfrak{d}_S(X_i) \ge M\right). \tag{4.44}$$

To bound the probabilities on the RHS, write $X_i = \theta_i + \Sigma^{1/2} Z_i$. By Lemma 4.13, taking $\lambda \downarrow 0$,

$$\mathbb{E}\zeta_2^2/n \le (\eta^*)^2 + 4\bar{k} \log\left(\frac{(2\pi)^d}{\rho^2}\right)\left(C_d \frac{M^{d-2}}{n}\bar{k}^{1-d/2} + \inf_{q \ge (d+1)/2 \log n}\left(\frac{2\mu_q}{M}\right)^q\right). \tag{4.45}$$

Bounding $\mathbb{E}\zeta_3^2$

Fix $j \in [N]$. Let $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ and $\xi_i^* \sim G_i^*$ and $\xi_i^{(j)} \sim G_i^{(j)}$, where $G_i^{(j)}$ denotes the scale change of $G^{(j)}$ by $\Sigma_i^{-1/2}$. In accordance with (4.37), we write

$$\begin{aligned}
&\|\hat{\theta}_\rho^{(j)} - \hat{\theta}_\rho^*\|_F \\
&\quad = \left(\sum_{i=1}^n \left\| \Sigma_i^{1/2}\left(\xi_i^{(j)} + Z_i + \frac{\nabla f_{G_i^{(j)}, I_d}(\xi_i^{(j)} + Z_i)}{f_{G_i^{(j)}, I_d}(\xi_i^{(j)} + Z_i) \vee \rho}\right) - \Sigma_i^{1/2}\left(\xi_i^* + Z_i + \frac{\nabla f_{G_i^*, I_d}(\xi_i^* + Z_i)}{f_{G_i^*, I_d}(\xi_i^* + Z_i) \vee \rho}\right)\right\|_2^2\right)^{1/2}
\end{aligned}$$

Call the RHS above $F(Z)$. Then

$$\begin{aligned}
|F(Z) - F(Z')| &= \left| \|\hat{\theta}_\rho^{(j)}(Z) - \hat{\theta}_\rho^*(Z)\|_F - \|\hat{\theta}_\rho^{(j)}(Z') - \hat{\theta}_\rho^*(Z')\|_F \right| \\
&\le \|\hat{\theta}_\rho^{(j)}(Z) - \hat{\theta}_\rho^{(j)}(Z')\|_F + \|\hat{\theta}_\rho^*(Z) - \hat{\theta}_\rho^*(Z')\|_F
\end{aligned}$$

Focusing on the second term on the RHS,

$$\begin{aligned}
&\|\hat{\theta}_\rho^*(Z) - \hat{\theta}_\rho^*(Z')\|_F \\
&\quad \le \bar{k}^{1/2} \sqrt{\sum_{i=1}^n \left\| \left(\xi_i^* + Z_i + \frac{\nabla f_{G_i^*, I_d}(\xi_i^* + Z_i)}{f_{G_i^*, I_d}(\xi_i^* + Z_i) \vee \rho}\right) - \left(\xi_i^* + Z_i' + \frac{\nabla f_{G_i^*, I_d}(\xi_i^* + Z_i')}{f_{G_i^*, I_d}(\xi_i^* + Z_i') \vee \rho}\right)\right\|_2^2} \\
&\quad \le \bar{k}^{1/2} \sqrt{\sum_{i=1}^n \left\| \left(Z_i + \frac{\nabla f_{G_i^* - \xi_i^*, I_d}(Z_i)}{f_{G_i^* - \xi_i^*, I_d}(Z_i) \vee \rho}\right) - \left(Z_i' + \frac{\nabla f_{G_i^* - \xi_i^*, I_d}(Z_i')}{f_{G_i^* - \xi_i^*, I_d}(Z_i') \vee \rho}\right)\right\|_2^2}
\end{aligned}$$

Saha and Guntuboyina (2020b, Proof of Lemma F.3) then gives

$$\|\hat{\theta}_\rho^*(Z) - \hat{\theta}_\rho^*(Z')\|_F \le \bar{k}^{1/2} L^2(\rho) \|Z - Z'\|_F,$$

where as before $L(\rho) \coloneqq \sqrt{-\log((2\pi)^d \rho^2)}$. The same argument applies to $\|\hat{\theta}_\rho^{(j)}(Z) - \hat{\theta}_\rho^{(j)}(Z')\|_F$. Hence F is $2\bar{k}L^2(\rho)$-Lipschitz. By concentration of Lipschitz functions of Gaussians and a union bound

$$\mathbb{P}\left(\zeta_3^2 \ge x\right) \le N \exp\left(-\frac{x^2}{8\bar{k}L^4(\rho)}\right).$$

Integrating the tail gives

$$\mathbb{E}\zeta_3^2 \le 8\bar{k}L^4(\rho) \log(eN). \tag{4.46}$$

Bounding $\mathbb{E}\zeta_4^2$

Again by the scaled Tweedie's formula (4.37),

$$\mathbb{E}\|\hat{\theta}_\rho^{(j)} - \hat{\theta}_\rho^*\|_F \le \sqrt{\mathbb{E}\|\hat{\theta}_\rho^{(j)} - \hat{\theta}_\rho^*\|_F^2} \le \sqrt{\bar{k}\sum_{i=1}^n \mathbb{E}_{X_i\sim f_{G_i^*,I_d}} \left\| \frac{\nabla f_{G_i^{(j)},I_d}(X_i)}{f_{G_i^{(j)},I_d}(X_i)\vee\rho} - \frac{\nabla f_{G_i^*,I_d}(X_i)}{f_{G_i^*,I_d}(X_i)\vee\rho}\right\|_2^2}$$

Saha and Guntuboyina (2020a, Lemma E.1) bounds the above expectation, yielding

$$\left(\mathbb{E}\|\hat{\theta}_\rho^{(j)} - \hat{\theta}_\rho^*\|_F\right)^2 \le C_d\bar{k}\sum_{i=1}^n \max\left\{\left(\frac{L^2(\rho)}{2}\right)^3, \left|\log h\left(f_{G_i^*,I_d}, f_{G_i^{(j)},I_d}\right)\right|\right\} h^2\left(f_{G_i^*,I_d}, f_{G_i^{(j)},I_d}\right). \tag{4.47}$$

By a change of variables,

$$h^2\left(f_{G_i^*,I_d}, f_{G_i^{(j)},I_d}\right) = h^2\left(f_{G^*,\Sigma_i}, f_{G^{(j)},\Sigma_i}\right).$$

Using the shorthand $h_i^2 = h^2\left(f_{G^*,\Sigma_i}, f_{G^{(j)},\Sigma_i}\right)$ and using $\rho = (2\pi)^{-d/2}/n$,

$$\begin{aligned}\left(\mathbb{E}\|\hat{\theta}_\rho^{(j)} - \hat{\theta}_\rho^*\|_F\right)^2 &\le C_d\bar{k}\sum_{i=1}^n \max\left\{(\log n)^3, -\log h_i\right\} h_i^2 \\ &= C_d\bar{k}\left(\sum_{i:(\log n)^3\ge -\log h_i} (\log n)^3 h_i^2 + \sum_{i:(\log n)^3 < -\log h_i} -(\log h_i)h_i^2\right) \\ &\le C_d\bar{k}\left(n(\log n)^3\delta + \sum_{i:(\log n)^3<\log h_i^{-1}} (\log h_i^{-1})h_i^2\right),\end{aligned} \tag{4.48}$$

where in the last step we used $\frac{1}{n}\sum_{i=1}^n h_i^2 = \bar{h}^2(f_{G^*,\bullet}, f_{G^{(j)},\bullet}) \le \delta$. To bound the second term, note for $n \ge 6$, $(\log n)^3 \ge 3\log n$, implying $h_i \le n^{-3}$ for all i such that $(\log n)^3 < \log h_i^{-1}$. Since $h_i \log h_i^{-1} \le e^{-1}$ for all $h_i \in [0,1]$,

$$\sum_{i:(\log n)^3<\log h_i^{-1}} (\log h_i^{-1})h_i^2 \le \sum_{i:(\log n)^3<\log h_i^{-1}} \frac{1}{en^3} \le \frac{1}{en^2}.$$

The first term dominates, so

$$\mathbb{E}\zeta_4^2 \le C_d\bar{k}n(\log n)^3\delta. \tag{4.49}$$

Bounding the metric entropy $\log N$

We will actually bound the larger covering number $\log N(\eta^*, \mathcal{P}(\mathbb{R}^d), m^S)$ of the space of all probability measures $\mathcal{P}(\mathbb{R}^d)$ in the metric m^S. For any measure G we let G_i denote the measure scaled by $\Sigma_i^{-1/2}$ as in the scaled Tweedie formula. For $G, H \in \mathcal{P}(\mathbb{R}^d)$,

$$\begin{aligned}
m^S(G,H) &\coloneqq \max_{i\in[n]} \sup_{x:\mathfrak{d}_S(x)\le M} \left\| \frac{\Sigma_i \nabla f_{G,\Sigma_i}(x)}{f_{G,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} - \frac{\Sigma_i \nabla f_{H,\Sigma_i}(x)}{f_{H,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} \right\|_2 \\
&\le \max_{i\in[n]} \sup_{x:\mathfrak{d}_S(x)\le M} \left\| \frac{\Sigma_i \nabla f_{G,\Sigma_i}(x)}{f_{G,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} - \frac{\Sigma_i \nabla f_{G,\Sigma_i}(x)}{f_{H,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} \right\|_2 \\
&\quad + \max_{i\in[n]} \sup_{x:\mathfrak{d}_S(x)\le M} \left\| \frac{\Sigma_i \nabla f_{G,\Sigma_i}(x)}{f_{H,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} - \frac{\Sigma_i \nabla f_{H,\Sigma_i}(x)}{f_{H,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} \right\|_2 \\
&\le \max_{i\in[n]} \sup_{x:\mathfrak{d}_S(x)\le M} \left\| \frac{\Sigma_i \nabla f_{G,\Sigma_i}(x)}{f_{G,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} \right\|_2 \\
&\quad \times \frac{\left| f_{G,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|}) - f_{H,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|}) \right|}{f_{H,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} \\
&\quad + \max_{i\in[n]} \sup_{x:\mathfrak{d}_S(x)\le M} \left\| \frac{\Sigma_i(\nabla f_{G,\Sigma_i}(x) - \nabla f_{H,\Sigma_i}(x))}{f_{H,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} \right\|_2
\end{aligned}$$

For the first term, by Saha and Guntuboyina (2020b, Lemma F.1),

$$\left\| \frac{\Sigma_i \nabla f_{G,\Sigma_i}(x)}{f_{G,\Sigma_i}(x) \vee (\rho/\sqrt{|\Sigma_i|})} \right\|_2 = \left\| \Sigma_i^{1/2} \frac{\nabla f_{G_i^*,I_d}(\Sigma_i^{-1/2}x)}{f_{G_i^*,I_d}(\Sigma_i^{-1/2}y) \vee \rho} \right\|_2 \le \bar{k}^{1/2} L(\rho).$$

Replacing $f \vee (\rho/\sqrt{|\Sigma_i|})$ with $\rho/\sqrt{|\Sigma_i|}$ in the denominator can only make the denominator smaller, so

$$\begin{aligned}
m^S(G,H) &\le \bar{k}^{1/2}\rho^{-1}L(\rho) \max_{i\in[n]} \sup_{x:\mathfrak{d}_S(x)\le M} \sqrt{|\Sigma_i|}\, |f_{G,\Sigma_i}(x) - f_{H,\Sigma_i}(x)| && (4.50)\\
&\quad + \rho^{-1} \max_{i\in[n]} \sup_{x:\mathfrak{d}_S(x)\le M} \sqrt{|\Sigma_i|}\, \|\Sigma_i(\nabla f_{G,\Sigma_i}(x) - \nabla f_{H,\Sigma_i}(x))\|_2 && (4.51)\\
&\le \bar{k}^{d/2+1/2}\rho^{-1}L(\rho)\|f_{G,\bullet} - f_{H,\bullet}\|_{\infty,S^M} + \bar{k}^{d/2+1}\rho^{-1}\|f_{G,\bullet} - f_{H,\bullet}\|_{\nabla,S^M} && (4.52)
\end{aligned}$$

In particular, letting

$$\eta^* = \left(\bar{k}^{d/2+1/2}L(\rho) + \bar{k}^{d/2+1}\right)\frac{\eta}{\rho}$$

we have

$$\log N(\eta^*, \mathcal{P}(\mathbb{R}^d), m^S) \le \log N(\eta/2, \mathbb{F}, \|\cdot\|_{\infty,S^M}) + \log N(\eta/2, \mathbb{F}, \|\cdot\|_{\nabla,S^M}).$$

We already have a bound on $\log N(\eta/2, \mathbb{F}, \|\cdot\|_{\infty,S^M})$ in Lemma 4.16, and we bound the other term similarly in Lemma 4.17 below. Combining these bounds,

$$\log N(\eta^*, \mathcal{P}(\mathbb{R}^d), m^S) \le C_d N(a, S^{M+a}) \left(\log \frac{c_{d,\underline{k},\bar{k}}}{\eta}\right)^{d+1},$$

where $a = \sqrt{-2\bar{k}\log\left(\sqrt{\underline{k}\wedge 1}\frac{(2\pi\underline{k})^{d/2}}{5}\eta\right)}$. Take $\eta = \rho/n = (2\pi)^{-d/2}/n^2$.

$$a = \sqrt{4\bar{k}\log n + 2\bar{k}\log\left(\frac{5}{\sqrt{\underline{k}\wedge 1}\underline{k}^{d/2}}\right)} \in \left[\sqrt{2\bar{k}\log n}, \sqrt{6\bar{k}\log n}\right],$$

provided $1/n \le \frac{5}{\sqrt{\underline{k}\wedge 1}\underline{k}^{d/2}} \le n$. Hence by Saha and Guntuboyina (2020b, Lemma F.6) (see the argument on page 6) gives

$$\log N \le c_{d,\underline{k},\bar{k}}(\log n)^{1+d/2}\mathrm{Vol}(S^{2M}).$$

Lemma 4.17. *For all compact $S \subset \mathbb{R}^d$, $M > 0$ and $\eta > 0$ sufficiently small,*

$$\log N(\eta, \mathbb{F}, \|\cdot\|_{\nabla,S^M}) \le C_d N(a, S^a)\left(\log\frac{c_{d,\underline{k},\bar{k}}}{\eta}\right)^{d+1}$$

where $a = \sqrt{2\bar{k}\log\frac{c'_{d,\bar{k},\underline{k}}}{\eta}}$.

Proof. Fix $G \in \mathcal{P}(\mathbb{R}^d)$. By Lemma 4.15, there is a discrete measure H supported on S^a with at most

$$l := \left(2\lfloor 13.5a^2/\bar{k}\rfloor + 2\right)^d N(a, S^a) + 1$$

atoms such that

$$\|f_{G,\bullet} - f_{H,\bullet}\|_{\nabla,S^a} \le a\left(1 + \frac{3}{\sqrt{2\pi}}\right)(2\pi\underline{k})^{-d/2}e^{-a^2/(2\bar{k})}$$

Now let $\mathcal{C}$ denote a minimal α-net of S^a. Write $H = \sum_j w_j\delta_{a_j}$, and define $H' = \sum_j w_j\delta_{b_j}$ where $b_j \in \mathcal{C}$ is the closest element to a_j. Then

$$\begin{aligned}
\|\nabla f_{H,\Sigma_i}(x) - \nabla f_{H',\Sigma_i}(x)\|_2 &\le \sum_j w_j \|\nabla\varphi_{\Sigma_i}(x-a_j) - \nabla\varphi_{\Sigma_i}(x-b_j)\|_2 \\
&\le \underline{k}^{-1/2}|\Sigma_i|^{-1/2}\sum_j w_j \left\|\nabla\varphi\left(\Sigma_i^{-1/2}(x-a_j)\right) - \nabla\varphi\left(\Sigma_i^{-1/2}(x-b_j)\right)\right\|_2 \\
&\le \frac{\underline{k}^{-d/2-1/2}\alpha}{(2\pi)^{d/2}}\left[1 + \frac{2}{e} + \frac{\alpha}{\sqrt{\underline{k}e}}\right]
\end{aligned}$$

Now let $\mathcal{D}_+^{p\times p}$ denote a minimal β-net of Δ_{l-1} under $\|\cdot\|_1$. Let $H'' = \sum_j w_j'\delta_{b_j}$ where $\|w' - w\|_1 \le \beta$. Then

$$\|\nabla f_{H',\Sigma_i}(x) - \nabla f_{H'',\Sigma_i}(x)\|_2 \le \beta \sup_u \|\nabla \varphi_{\Sigma_i}(u)\|_2 \le \frac{\underline{k}^{-d/2-1/2}\beta}{(2\pi)^{d/2}\sqrt{e}}.$$

By triangle inequality,

$$\|f_{G,\bullet} - f_{H'',\bullet}\|_{\nabla,S^M} \le (2\pi\underline{k})^{-d/2}\left[\left(1+\frac{3}{\sqrt{2\pi}}\right)ae^{-a^2/(2\bar{k})} + \frac{\alpha}{\sqrt{\underline{k}}}\left[1+\frac{2}{e}+\frac{\alpha}{\sqrt{\underline{k}e}}\right] + \frac{\beta}{\sqrt{\underline{k}e}}\right]$$

Taking $a = \sqrt{2\bar{k}\log\alpha^{-1}} \ge 1$ and $\alpha = \beta = \sqrt{k\wedge 1}\frac{(2\pi\underline{k})^{d/2}}{5}\eta$,

$$\|f_{G,\bullet} - f_{H'',\bullet}\|_{\nabla,S^M} \le \frac{5a\alpha}{\sqrt{k\wedge 1}}(2\pi\underline{k})^{-d/2} = a\eta$$

The proof is completed following same steps as the proof of Lemma 5. □

4.6.3.4 Putting together the pieces

Combining (4.39), (4.42), (4.45), (4.46), and (4.49) and pulling out any constants depending on $d, \underline{k}$, or $\bar{k}$,

$$\begin{aligned}
\mathbb{E}\|\hat{\theta} - \hat{\theta}^*\|_F^2/n &\le (5/n)\left[\mathbb{E}\|\hat{\theta}_\rho^* - \hat{\theta}^*\|_F^2 + \sum_{t=1}^4 \mathbb{E}\zeta_t^2\right] \\
&\le c_{d,\underline{k},\bar{k}}\Bigg(\varepsilon_n^2(M,S,G^*)(\sqrt{\log n})^{d-2} \\
&\qquad + \frac{\log n}{n} \\
&\qquad + (\eta^*)^2 + \varepsilon_n^2(M,S,G^*) \\
&\qquad + \log(eN)\frac{(\log n)^2}{n} \\
&\qquad + \varepsilon_n^2(M,S,G^*)(\sqrt{\log n})^6\Bigg) \\
&\le c_{d,\underline{k},\bar{k}}\varepsilon_n^2(M,S,G^*)(\sqrt{\log n})^{(d-2)\vee 6}
\end{aligned} \tag{4.53}$$

This completes the proof of Theorem 4.8.

4.6.4 Proofs of Theorems 4.10 and 4.12

Proof of Theorem 4.10. We will relate the Wasserstein distance to the average Hellinger distance, so we rely on the tools of Nguyen (2013, proof of theorem 2). Fix a symmetric

density K whose Fourier transform $\widetilde{K}$ is bounded with support on $[-1,1]^d$. For any $\delta > 0$ define the scaled kernel $K_\delta(x) = \frac{1}{\delta^d} K(x/\delta)$. By the triangle inequality,

$$W_2(G^*, \widehat{G}_n) \le W_2(G^*, G^* * K_\delta) + W_2(G^* * K_\delta, \widehat{G}_n * K_\delta) + W_2(\widehat{G}_n * K_\delta, \widehat{G}_n).$$

For the first and third terms, bound the minimum over all couplings by the strong coupling:

$$W_2^2(G, G * K_\delta) = \min_{\theta \sim G, \theta' \sim G, \varepsilon \sim K} \mathbb{E}\|\theta - (\theta' + \delta\varepsilon)\|_2^2 \le \delta^2 \mathbb{E}_{\varepsilon \sim K}\|\varepsilon\|_2^2,$$

where the inequality follows from choosing the coupling where $\theta = \theta'$ almost surely. Letting $m_2(K) = \mathbb{E}_{\varepsilon \sim K}\|\varepsilon\|_2^2$ denote the second moment of the (unscaled) kernel, we have

$$W_2(G^*, \widehat{G}_n) \le 2\sqrt{m_2(K)}\delta + W_2(G^* * K_\delta, \widehat{G}_n * K_\delta). \tag{4.54}$$

For the second term, Villani (2008, Theorem 6.15) yields

$$W_2^2(G^* * K_\delta, \widehat{G}_n * K_\delta) \le 2 \int \|x\|_2^2 \, \mathrm{d}\left|G^* * K_\delta - \widehat{G}_n * K_\delta\right|(x)$$

By Nguyen (2013, Lemma 6), for any $s > 2$ such that $m_s(K) = \mathbb{E}\|\varepsilon\|_2^s < \infty$,

$$\begin{aligned}
W_2^2(G^* * K_\delta, \widehat{G}_n * K_\delta) &\le 4 \left\|G^* * K_\delta - \widehat{G}_n * K_\delta\right\|_{L_1}^{(s-2)/s} R^{2/s} \\
&\le 4 \left[2\mathrm{Vol}(B_1)^{s/(d+2s)} R^{d/(d+2s)} \left\|G^* * K_\delta - \widehat{G}_n * K_\delta\right\|_{L_2}^{2s/(d+2s)}\right]^{(s-2)/s} R^{2/s} \\
&= 4 \cdot 2^{(s-2)/s} \mathrm{Vol}(B_1)^{(s-2)/(2s+d)} \\
&\qquad \times R^{d(s-2)/(s(d+2s))+2/s} \left\|G^* * K_\delta - \widehat{G}_n * K_\delta\right\|_{L_2}^{2(s-2)/(d+2s)} \\
&\le 8\sqrt{\mathrm{Vol}(B_1)} \cdot R^{d(s-2)/(s(d+2s))+2/s} \left\|G^* * K_\delta - \widehat{G}_n * K_\delta\right\|_{L_2}^{2(s-2)/(d+2s)}
\end{aligned}$$

where $R \coloneqq \mathbb{E}_{\theta^* \sim G^*, \varepsilon \sim K}\|\theta^* + \delta\varepsilon\|_2^s + \mathbb{E}_{\theta \sim \widehat{G}_n, \varepsilon \sim K}\|\theta + \delta\varepsilon\|_2^s$.

For moments in the term R, use $\mathbb{E}\|\theta + \delta\varepsilon\|_2^s \le 2^s \left(\mathbb{E}\|\theta\|_2^s + \delta^s m_s(K)\right)$, so

$$R \le 2^s(m_s(G^*) + m_s(\widehat{G}_n) + 2\delta^s m_s(K)).$$

The quantity $m_s(K)$ is regarded as a constant depending only on $s > 2$ and d. By assumption, the support of $\widehat{G}_n$ is contained in the minimum bounding box of the observations, which is further contained in $[-U, U]^d$ where $U = \max_{i,j}|X_{ij}| \le L + \max_{i,j}|X_{ij} - \theta_{ij}^*|$. Since $X_{ij} - \theta_{ij}^* \overset{\text{ind}}{\sim} \mathcal{N}(0, (\Sigma_i)_{jj})$, we have by a standard concentration argument that

$$U \le L + 4\sqrt{k \log n}$$

with probability at least $1 - \frac{2d}{n^8}$. Hence, with the same probability

$$m_s(\widehat{G}_n) = \mathbb{E}_{\widehat{G}_n}\|\theta\|_2^s \leq d^{s/2}\mathbb{E}_{\widehat{G}_n}\|\theta\|_\infty^s \leq d^{s/2}\left(L + 4\sqrt{\overline{k}\log n}\right)^s.$$

This same bound holds for $m_s(G^*)$.

For the $\|\cdot\|_{L_2}$ norm $\left\|G^* * K_\delta - \widehat{G}_n * K_\delta\right\|_{L_2}$, let $g_\delta^{(i)}$ denote the inverse Fourier transform of $\widetilde{K_\delta}/\widetilde{\varphi_{\Sigma_i}}$, so that $G * K_\delta = f_{G,\Sigma_i} * g_\delta^{(i)}$. Hence, by Proposition 8.49 of Folland (1999), we have for each $i = 1, \ldots, n$,

$$\left\|G^* * K_\delta - \widehat{G}_n * K_\delta\right\|_{L_2} \leq 2d_{TV}(f_{G^*,\Sigma_i}, f_{\widehat{G}_n,\Sigma_i})\|g_\delta^{(i)}\|_{L_2}.$$

Using Plancherel's theorem and the fact that $\widetilde{K}$ is bounded on its support of $[-1,1]^d$,

$$\begin{aligned}
\|g_\delta^{(i)}\|_{L_2}^2 &= \int_{\mathbb{R}^d} \frac{\widetilde{K}(\delta\omega)^2}{\widetilde{\varphi_{\Sigma_i}}(\omega)^2}\,\mathrm{d}\omega \leq C_d \int_{[-1/\delta,1/\delta]^d} \widetilde{\varphi_{\Sigma_i}}(\omega)^{-2}\,\mathrm{d}\omega \\
&= C_d \int_{[-1/\delta,1/\delta]^d} \exp(\omega'\Sigma_i\omega)\,\mathrm{d}\omega = C_d \prod_{j=1}^d \int_{-1/\delta}^{1/\delta} \exp((\Sigma_i)_{jj}\omega_j^2)\,\mathrm{d}\omega_j \\
&\leq C_d \prod_{j=1}^d e^{2(\Sigma_i)_{jj}\delta^{-2}} \int_{-1/\delta}^{1/\delta} \exp(-(\Sigma_i)_{jj}\omega_j^2)\,\mathrm{d}\omega_j \leq C_d \left(\frac{\pi}{\underline{k}}\right)^{d/2} e^{2d\overline{k}\delta^{-2}}.
\end{aligned}$$

Averaging over $i = 1, \ldots, n$,

$$\begin{aligned}
\left\|G^* * K_\delta - \widehat{G}_n * K_\delta\right\|_{L_2} &\leq C_d \underline{k}^{-d/2} e^{2d\overline{k}\delta^{-2}} \cdot \frac{1}{n}\sum_{i=1}^n d_{TV}(f_{G^*,\Sigma_i}, f_{\widehat{G}_n,\Sigma_i}) \\
&\leq C_d \underline{k}^{-d/2} e^{2d\overline{k}\delta^{-2}} \overline{h}(G^*, \widehat{G}_n).
\end{aligned}$$

Combining our calculations following (4.54), we have

$$\begin{aligned}
W_2(G^*, \widehat{G}_n) \leq C_{d,s} \inf_{\delta\in(0,1)} \Bigg\{ \delta + \Bigg(2^s \left(d^{s/2}\left(L + 4\sqrt{\overline{k}\log n}\right)^s + 2\delta^s m_s(K)\right)\Bigg)^{3d/(2d+4s)} \\
\times \left(\underline{k}^{-d/2} e^{2d\overline{k}\delta^{-2}} \overline{h}(G^*, \widehat{G}_n)\right)^{(s-2)/(d+2s)} \Bigg\}.
\end{aligned} \tag{4.55}$$

Assume n is large enough that $4\sqrt{\overline{k}\log n} \geq L$ and $2^s d^{s/2}\left(4\sqrt{\overline{k}\log n}\right)^s \geq 2m_s(K)$, so

$$W_2(G^*, \widehat{G}_n) \leq C_{d,s} \inf_{\delta\in(0,1)} \left\{ \delta + \left(\overline{k}\log n\right)^{3sd/(4(d+2s))} \left(\underline{k}^{-d/2} e^{2d\overline{k}\delta^{-2}} \overline{h}(G^*, \widehat{G}_n)\right)^{(s-2)/(d+2s)} \right\}. \tag{4.56}$$

Choosing $\delta^{-2} = -\frac{1}{4\overline{k}d}\log \bar{h}(G^*, \widehat{G}_n)$ (provided $\delta < 1$) and $s = d + 2$ yields

$$W_2^2(G^*, \widehat{G}_n) \le C_d \left\{ \frac{\overline{k}d}{-\log \bar{h}(G^*, \widehat{G}_n)} + \left(\overline{k}\log n\right)^{d/2} \left(\underline{k}^{-d}\bar{h}(G^*, \widehat{G}_n)\right)^{1/12} \right\}. \tag{4.57}$$

$\varepsilon_n^2(M, S, G^*)$ defined in (4.13), with $S = [-L, L]^d$ and $M = \sqrt{10\overline{k}\log n}$ gives

$$\varepsilon_n^2 = \frac{\left(4\sqrt{10}\left(L^2 \vee \overline{k}\right)\right)^d}{n}(\log n)^{d+1}.$$

By Theorem 4.6,

$$W_2^2(G^*, \widehat{G}_n) \le C_d \left\{ \frac{\overline{k}d}{\log n - \log C_{d,\overline{k},\underline{k},L} t^2 (\log n)^{d+1}} + \left(\overline{k}\log n\right)^{d/2} \left(C_{d,\overline{k},\underline{k},L} t^2 \frac{(\log n)^{d+1}}{n} \right)^{1/24} \right\},$$

with probability at least $1 - 2n^{-t^2}$. Take $t^2 = 8$. For n sufficiently large the first term dominates, $\delta < 1$, and $\log n - \log C_{d,\overline{k},\underline{k},L} 8(\log n)^{d+1} \ge (\log n)/2$. □

Proof of Theorem 4.12. Take $\mu = 0$ by location equivariance (see Lemma 4.4). Write $\widehat{G}_n = \sum_{j=1}^{\hat{k}} \hat{w}_j \delta_{\hat{a}_j}$. Since $G^* = \delta_0$ is a point mass,

$$W_2^2(\widehat{G}_n, G^*) = \mathbb{E}_{\vartheta \sim \widehat{G}_n} \|\vartheta\|_2^2 = \sum_{j=1}^{\hat{k}} \hat{w}_j \|\hat{a}_j\|_2^2.$$

We relate this to the marginal density $f_{\widehat{G}_n, \bullet}$ via

$$\begin{aligned}
\int f_{\widehat{G}_n, \Sigma_i}(x)\|x\|_2^2 \,\mathrm{d}x &= \sum_{j=1}^{\hat{k}} \hat{w}_j \int \varphi_{\Sigma_i}(x - \hat{a}_j)\|x\|_2^2 \,\mathrm{d}x \\
&= \sum_{j=1}^{\hat{k}} \hat{w}_j \int \varphi_{\Sigma_i}(x)(\|x\|_2^2 + \|\hat{a}_j\|_2^2 + 2\langle x, \hat{a}_j\rangle) \,\mathrm{d}x \\
&= \int f_{G^*, \Sigma_i}(x)\|x\|_2^2 \,\mathrm{d}x + W_2^2(\widehat{G}_n, G^*).
\end{aligned}$$

Hence for any $i \in \{1, \ldots, n\}$,

$$\begin{aligned}
W_2^2(\widehat{G}_n, G^*) &= \int (f_{\widehat{G}_n, \Sigma_i}(x) - f_{G^*, \Sigma_i}(x))\|x\|_2^2 \,\mathrm{d}x \\
&\le 2h(f_{\widehat{G}_n, \Sigma_i}, f_{G^*, \Sigma_i}) \left(\int \left(f_{\widehat{G}_n, \Sigma_i}(x) + f_{G^*, \Sigma_i}(x) \right) \|x\|_2^4 \,\mathrm{d}x \right)^{1/2}.
\end{aligned}$$

Averaging over $i \in \{1, \dots, n\}$,

$$W_2^2(\widehat{G}_n, G^*) \le 2\bar{h}\left(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}\right) \max_{i=1:n}\left(\int \left(f_{\widehat{G}_n,\Sigma_i}(x) + f_{G^*,\Sigma_i}(x)\right)\|x\|_2^4 \,\mathrm{d}x\right)^{1/2}.$$

Applying Theorem 4.6 with $S = \{0\}$ and $M = \sqrt{10\overline{k}\log n}$,

$$\bar{h}^2\left(f_{\widehat{G}_n,\bullet}, f_{G^*,\bullet}\right) \lesssim_{d,\overline{k},\underline{k}} t^2 \frac{(\log n)^{d+1}}{n}$$

with probability at least $1 - 2n^{-t^2}$ for all $t \ge 1$.

For the remaining terms,

$$\int f_{G^*,\Sigma_i}(x)\|x\|_2^4 \,\mathrm{d}x = \mathbb{E}\|X_i\|_2^4 = \mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\|\Sigma_i^{1/2}Z\|_2^4 \le \overline{k}^2\mathbb{E}_{A\sim\chi^2_d}A^2 = \overline{k}^2 d(d+2)$$

and

$$\begin{aligned}
\int f_{\widehat{G}_n,\Sigma_i}(x)\|x\|_2^4 \,\mathrm{d}x &= \sum_j \hat{w}_j \int \varphi_{\Sigma_i}(x)\|x + \hat{a}_j\|_2^4 \,\mathrm{d}x \\
&\le 8\sum_j \hat{w}_j \int \varphi_{\Sigma_i}(x)\left(\|x\|_2^4 + \|\hat{a}_j\|_2^4\right)\mathrm{d}x \\
&\le 8\overline{k}^2 d(d+2) + 8\sum_j \hat{w}_j\|\hat{a}_j\|_2^4.
\end{aligned}$$

By our assumption on the support that each $\hat{a}_j \in \mathbb{B}_{\kappa r}(\bar{X})$, each $\hat{a}_j$ equivalently satisfies

$$\|\hat{a}_j - \bar{X}\|_2^4 \le (\overline{k}/\underline{k})^4 \max_i \|X_i - \bar{X}\|_2^4.$$

Noting that $X_i - \bar{X} \sim \mathcal{N}\left(0, (1-2n^{-1})\Sigma_i + n^{-1}\bar{\Sigma}\right)$ with $\bar{\Sigma} = n^{-1}\sum_{j=1}^n \Sigma_j$, we bound

$$\begin{aligned}
\sum_j \hat{w}_j\|\hat{a}_j\|_2^4 &\le 8\left(\|\bar{X}\|_2^4 + \max_j \|\hat{a}_j - \bar{X}\|_2^4\right) \\
&\le 8\left(\|\bar{X}\|_2^4 + \kappa^4 \max_{i\in[n]} \|X_i - \bar{X}\|_2^4\right) \\
&\le_{\mathrm{st}} 8\overline{k}^2\left(n^{-2}A_0^2 + \kappa^4 \max_{i\in[n]} A_i^2\right) \\
&\le 16\frac{\overline{k}^6}{\underline{k}^4}\max_{i=0:n} A_i^2,
\end{aligned}$$

where $A_0, A_1, \dots, A_n \sim \chi^2_d$ are possibly dependent, and $\le_{\mathrm{st}}$ denotes stochastic inequality.

For $t \geq 1$, we use the following tail bound (see Laurent & Massart, 2000, Lemma 1)

$$\mathbb{P}\left(\max_{i=0:n} A_i^2 \geq 60t^2(\log n)^2\right) \leq n^{-t^2},$$

where we have used the assumption in Theorem 4.6 that $n \geq (2\pi)^{d/2}$ to eliminate the dependence on d. We have thus shown that

$$\max_{i=1:n}\left(\int \left(f_{\widehat{G}_n,\Sigma_i}(x) + f_{G^*,\Sigma_i}(x)\right)\|x\|_2^4 \,\mathrm{d}x\right)^{1/2}$$
$$\leq \left(9\overline{k}^2 d(d+2) + 400\frac{\overline{k}^6}{\underline{k}^4}t^2(\log n)^2\right)^{1/2} \lesssim \frac{\overline{k}^3}{\underline{k}^2}t(\log n)$$

with probability at least $1 - n^{-t^2}$. Combining with a union bound over our earlier estimate,

$$W_2(\widehat{G}_n, G^*) \lesssim_{d,\overline{k},\underline{k}} t^{3/2}\frac{(\log n)^{(d+3)/4}}{n^{1/4}}$$

with probability at least $1 - 3n^{-t^2}$ for all $t \geq 1$. □

Chapter 5

Local false discovery rate control

5.1 Introduction

A common goal in applications of multiple hypothesis testing is to identify a relatively short list of candidate "discoveries" that are sufficiently promising to undertake some costly further action. In scientific applications, for example, each discovery may be the focus of a follow-up experiment, which wastes resources if the apparent discovery was only a mirage. The *false discovery rate* (FDR, Benjamini & Hochberg, 1995) has become a cornerstone of modern large-scale multiple testing because it directly measures the rate of this wastage:

> [T]he proportion of errors in the pool of candidates is of great economical significance since follow-up studies are costly, and thus avoiding multiplicity control is costly. Indeed, the FDR criterion is economically interpretable; when considering a potential threshold, the adjusted FDR gives the proportion of the investment that is about to be wasted on false leads. (Reiner et al., 2003)

An analyst who controls FDR at level $q = 5\%$, then, is willing to waste resources following up on one false discovery in exchange for every nineteen real discoveries.

Carrying this reasoning further, however, we can apply the same cost-benefit analysis to each individual rejection, not only to the list of rejections taken as a whole. In economic terminology, we should consider not only the *average utility* of our entire rejection set, but also the *marginal utility* of each rejection we make, since we always have the option to exclude any rejection that is not individually promising. For example, in Section 5.4 we reproduce the simulations of Benjamini and Hochberg (1995) and find in some settings that, even while the Benjamini–Hochberg (BH) procedure controls FDR at level $q = 5\%$, the *last discovery* (i.e. the discovery with the largest p-value) is false more than 30% of the time. In such settings, unless we are willing to suffer one false discovery for every two true discoveries, we would be better served by excluding the last rejection from the BH rejection set. More generally, to decide where to set our rejection threshold, we should ask about the proportion of false leads among the incremental rejections that we would add or remove by raising or lowering it.

The likelihood that an individual discovery is a false lead is called its *local false discovery rate* (lfdr, Efron et al., 2001). For $i = 1, \dots, m$, let $H_i = 0$ if the ith hypothesis is null and $H_i = 1$ otherwise, and consider the simple *Bayesian two-groups model*

$$p_i \mid H_i = h \overset{\text{ind}}{\sim} f_h, \qquad \text{with} \qquad H_i \overset{\text{iid}}{\sim} \text{Bern}(1 - \pi_0), \qquad \text{for } i = 1, \dots, m, \tag{5.1}$$

where $f_0 \coloneqq 1_{[0,1]}$ and f_1 are densities (null and alternative, respectively) supported on the unit interval $[0, 1]$, and the null proportion is $\pi_0 \in [0, 1]$. Let $f \coloneqq \pi_0 + (1 - \pi_0) f_1$ denote the common mixture density of the p-values in model (5.1), and let $F(t) \coloneqq \int_0^t f(u)\, du$ denote the corresponding cumulative distribution function (cdf). The lfdr is then defined as the posterior probability that $H_i = 0$, conditional on the observed p-value p_i:

$$\text{lfdr}(t) \coloneqq \mathbb{P}\left(H_i = 0 \mid p_i = t\right) = \frac{\pi_0}{f(t)}. \tag{5.2}$$

If we knew the problem parameters π_0 and f_1, then the definition (5.2) would neatly solve the problem posed above: we should reject only those hypotheses whose lfdr is below the break-even threshold of our cost-benefit tradeoff. Concretely, let $\lambda > 0$ define the ratio between the cost of each false discovery and the benefit of each true discovery. Then the utility of making R rejections, of which V are false discoveries, is proportional to $(R - V) - \lambda V$, and a simple calculation shows that we should reject the ith hypothesis if and only if $\text{lfdr}(p_i) \leq \alpha \coloneqq \frac{1}{1+\lambda}$.

We will usually work under the additional assumption that $f_1(t)$ is nonincreasing in t, or equivalently that $\text{lfdr}(t)$ is nondecreasing, so that smaller p-values represent stronger evidence against the null. This assumption, common in multiple testing (see, e.g., Genovese & Wasserman, 2004; Langaas et al., 2005; Strimmer, 2008), lets us restrict our attention to procedures that reject all p-values below a given threshold: if f_1 is nonincreasing then rejecting when $\text{lfdr}(p_i) \leq \alpha$ is equivalent to rejecting when p_i is sufficiently small.

In practice, π_0 and f_1 are typically unknown and must be estimated from the data, and many estimators have been proposed; see e.g. Aubert et al. (2004), Efron (2004, 2008), Efron et al. (2001), Liao et al. (2004), Muralidharan (2010), Patra and Sen (2016), Pounds and Cheng (2004), Pounds and Morris (2003), Robin et al. (2007), Scheid and Spang (2004), Stephens (2017), and Strimmer (2008). To the best of our knowledge, however, there are no known finite-sample lfdr control guarantees for multiple testing procedures based on these methods. By contrast, simple, robust, and well-known methods like the Benjamini–Hochberg (BH) procedure of Benjamini and Hochberg (1995) enjoy finite-sample FDR control without requiring the analyst to model the p-value distribution.

In this chapter, we introduce a new error control metric that measures the lfdr of a multiple testing procedure's least promising rejection. We represent a generic multiple testing method as a function $\mathcal{R}(p_1, \dots, p_m)$ returning an index set $\mathcal{R} \subseteq \{1, \dots, m\}$, where hypothesis i is rejected if and only if $i \in \mathcal{R}$. We say the procedure's *max-lfdr* is

$$\text{max-lfdr}(\mathcal{R}) \coloneqq \mathbb{E}\left[\max_{i \in \mathcal{R}} \text{lfdr}(p_i)\right], \tag{5.3}$$

defining the maximum as zero if no rejections are made. If f_1 is nonincreasing, then the max-lfdr of $\mathcal{R}$ coincides with the probability that the last rejection is a false discovery.

We also introduce a simple multiple testing procedure, which we call the *support line* (SL) procedure, that provably controls the max-lfdr under mild assumptions. Define the p-value order statistics $p_{(1)} \leq \cdots \leq p_{(m)}$, and let $p_{(0)} = 0$ by convention. Then our procedure rejects p-values up to the last (and a.s. unique) minimizer

$$R_q \coloneqq \operatorname*{argmin}_{k=0,\ldots,m} \; p_{(k)} - \frac{qk}{m}. \tag{5.4}$$

That is, we reject $\mathcal{R}_q \coloneqq \{i : \; p_i \leq \tau_q\}$, for the threshold $\tau_q = p_{(R_q)}$. Under the two-groups model (5.1), with nonincreasing f_1, we show in Theorem 5.1 that

$$\text{max-lfdr}(\mathcal{R}_q) = \pi_0 q.$$

Our method can be implemented without knowing π_0 or f_1, apart from the shape constraint, and bears a close relationship to the BH procedure, which replaces R_q in (5.4) with

$$R_q^{\text{BH}} \coloneqq \max \left\{ k \in \{0, \ldots, m\} : p_{(k)} \leq \frac{qk}{m} \right\},$$

rejecting $\mathcal{R}_q^{\text{BH}} \coloneqq \{i : \; p_i \leq \tau_q^{\text{BH}}\}$, for $\tau_q^{\text{BH}} = qR_q^{\text{BH}}/m \geq p_{(R_q^{\text{BH}})}$. Because $R_q \leq R_q^{\text{BH}}$, the BH method makes at least as many rejections as the SL method, and both methods make at least one rejection if and only if $p_{(k)} \leq \frac{qk}{m}$ for some $k \geq 1$; however, as we will argue, in general, the SL method should be run with a strictly larger q than we would use for BH. The left panel of Figure 5.1 illustrates the relationship between the two methods by reproducing the familiar plot of the BH procedure as an operation on the order statistics $p_{(1)}, \ldots, p_{(m)}$.

5.1.1 Multiple testing and the weighted classification loss

To formalize our analysis above, define the per-instance *weighted classification loss*:

$$L_\lambda(H, \mathcal{R}) \coloneqq \frac{(1+\lambda)V - R}{m}. \tag{5.5}$$

This loss can be derived, up to additive and multiplicative constants, by viewing each of the m hypotheses as a binary classification problem, where we incur a cost c_1 for each type I error or false discovery ($i \in \mathcal{R}$, but $H_i = 0$), and cost c_2 from each type II error or false non-discovery ($i \notin \mathcal{R}$, but $H_i = 1$). If the total number of non-nulls is $m_1 = \sum_i H_i$, then there are $m_1 - (R - V)$ false non-discoveries, so the total loss over all m instances is

$$c_1 V + c_2(m_1 - (R - V)) = c_2 m \cdot L_\lambda(H, \mathcal{R}) + c_2 m_1,$$

where $\lambda = c_1/c_2$ is the ratio between the two misclassification costs. L_λ as defined in (5.5) is normalized so that rejecting nothing incurs zero loss, and each true discovery has value $1/m$.

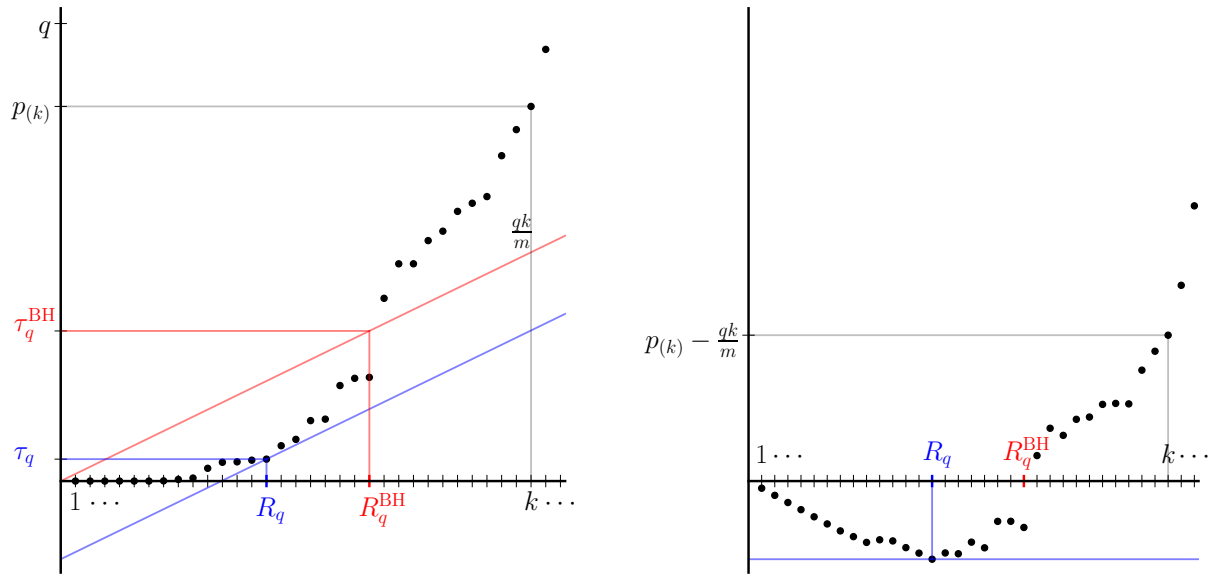


Figure 5.1: Left: The order statistics $p_{(k)}$ of the p-values as a function of the index k, shown in black. The BH procedure, in red, finds the largest index R_q^{BH} such that $p_{(R_q^{\mathrm{BH}})}$ falls below the ray of slope q/m; by contrast, our procedure finds the (last and almost surely unique) boundary point $(R_q, p_{(R_q)})$ of the supporting line of slope q/m. Right: The same plot with the ray through the origin of slope q/m subtracted off. The black dots represent a running estimate (5.7) of the weighted classification loss (5.5), which our procedure minimizes. BH(q) finds the largest threshold where the estimated loss is negative.

Under the two-groups model (5.1), Sun and Cai (2007, Theorem 2) show that the corresponding Bayes risk $\mathbb{E}L_\lambda(H, \mathcal{R})$ is minimized by the oracle procedure

$$\mathcal{R}^* := \{i : \ \mathrm{lfdr}(p_i) \leq \alpha\}, \quad \text{where} \quad \alpha = \frac{1}{1+\lambda}. \tag{5.6}$$

The ratio λ specifies the "break-even exchange rate" at which we are willing to trade true discoveries for false leads; e.g., if $\lambda = 19$ then we are willing to suffer a single false discovery for exactly 19 true discoveries, and we should reject a hypothesis only if its lfdr falls below the break-even tolerance $\alpha = 0.05$. If f_1 is nonincreasing, then the oracle procedure reduces to thresholding p-values at a fixed threshold

$$\mathcal{R}^* = \{i : \ p_i \leq \tau^*\}, \quad \text{for} \quad \tau^* := \max\{t \in [0,1] : \ \mathrm{lfdr}(t) \leq \alpha\},$$

with $\tau^* = 0$ if no such threshold exists.

Our method can be directly interpreted as minimizing an empirical proxy of the weighted classification loss. For a candidate threshold $t \in [0, 1]$, the expected number of null p-values

below the threshold is $m\pi_0 t$. If π_0 is known, we obtain a running estimator of the loss:

$$\hat{L}_\lambda(t;\pi_0) = \frac{(1+\lambda)m\pi_0 t - mF_m(t)}{m} = (1+\lambda)\left(\pi_0 t - \alpha F_m(t)\right), \tag{5.7}$$

where $F_m(t)$ represents the empirical cumulative distribution function (ecdf) of the p-values:

$$F_m(t) = \frac{1}{m}\sum_{i=1}^{m} 1\{p_i \le t\}.$$

Because $\hat{L}_\lambda(t;\pi_0)$ is increasing between successive order statistics, it is minimized at one of the order statistics, or at $p_{(0)} = 0$:

$$\operatorname*{argmin}_{k=0,1,\ldots,m} \hat{L}_\lambda(p_{(k)};\pi_0) \;=\; \operatorname*{argmin}_{k=0,1,\ldots,m} \pi_0 p_{(k)} - \frac{\alpha k}{m}.$$

Comparing the last expression to the definition of our procedure in (5.4), we see that $\hat{L}_\lambda(t;\pi_0)$ is minimized at $t = \tau_q$ for $q = \alpha/\pi_0$. By Theorem 5.1, we then have exactly max-lfdr$(\mathcal{R}_q) = \alpha$.

By contrast, τ_q^{BH} for $q = \alpha/\pi_0$ is the largest value of t that gives $\hat{L}_\lambda(t;\pi_0) = 0$, the same loss we would achieve by rejecting nothing at all. In other words, the BH procedure at level α/π_0 only aims to break even; to do better, we should run BH at a strictly smaller level $q < \alpha/\pi_0$, viewing q as a tuning parameter as in Neuvial and Roquain (2012).

To select q for our SL procedure when π_0 is unknown, we can either conservatively bound $\pi_0 \le 1$ and run the procedure at $q = \alpha$, or estimate π_0 and use $q = \alpha/\hat{\pi}_0$. To avoid confusion, we will always use the notation q to represent our method's tuning parameter, and reserve $\alpha = \frac{1}{1+\lambda}$ to represent the true target lfdr, defined in terms of the cost ratio λ.

Our procedure can alternatively be derived as a plug-in maximum likelihood estimator (MLE) of the oracle procedure $\mathcal{R}^*$, where we estimate $f(t)$ using Grenander's nonparametric MLE for a nonincreasing density (Grenander, 1956):

$$\hat{f}_m \coloneqq \operatorname*{argmax}_{\substack{g:[0,1]\to\mathbb{R}_+ \\ \text{nonincreasing density}}} \frac{1}{m}\sum_{i=1}^{m} \log g(p_i). \tag{5.8}$$

As we will see in Section 5.3.2, τ_q is also the largest value $t \in [0,1]$ for which $\hat{f}_m(t) \ge q^{-1}$. Thus, if we run our procedure at $q = \alpha/\pi_0$, we have

$$\mathcal{R}_{\alpha/\pi_0} = \left\{ i :\; \hat{f}_m(p_i) \ge (\alpha/\pi_0)^{-1} \right\} = \left\{ i :\; \frac{\pi_0}{\hat{f}_m(p_i)} \le \alpha \right\}.$$

As above, if π_0 is unknown, we can either estimate it or conservatively bound $\pi_0 \le 1$.

The relationship between our method and the Grenander estimator is convenient for asymptotic analysis because the latter is very well studied; see the book by Groeneboom

and Jongbloed (2014) for a thorough treatment. The Grenander estimator has previously been considered for estimating the lfdr (Strimmer, 2008) as well as for estimating the null proportion π_0 (Langaas et al., 2005). While $\hat{f}_m$ may be efficiently computed via the pool adjacent violators algorithm (Robertson et al., 1988), the definition in (5.4) is usually preferred for computational purposes.

5.1.2 The max-lfdr and the FDR

The max-lfdr in (5.3) and the FDR are two different error criteria that both appeal to the logic of trading off true and false discoveries. The key difference is that the FDR, defined as

$$\mathrm{FDR}(\mathcal{R}) \coloneqq \mathbb{E}\left[\frac{V}{R} \cdot 1\{R > 0\}\right],$$

measures the likelihood that a *randomly selected* rejection is null, whereas the max-lfdr instead measures the likelihood that the *least promising* rejection is null. In both cases the event in question is deemed not to have occurred if $R = 0$, so that under the global null (all $H_i = 0$, almost surely), both criteria reduce to the probability of making a single rejection.

Throughout this section, we will restrict our attention to procedures that reject the R hypotheses with the smallest p-values. That is, we assume a procedure $\mathcal{R}$ rejects $H_{(1)}, \ldots, H_{(R)}$, where $H_{(k)}$ represents the hypothesis corresponding to $p_{(k)}$. If f_1 is nonincreasing, then the procedure's *last rejection* $H_{(R)}$ is the least promising, and the max-lfdr can be equivalently characterized as the probability that the last rejection is a false discovery:

$$\text{max-lfdr}(\mathcal{R}) = \mathbb{E}\left[\text{lfdr}\left(p_{(R)}\right) \cdot 1\{R > 0\}\right] = \mathbb{P}\left\{H_{(R)} = 0, R > 0\right\}. \tag{5.9}$$

If $\text{max-lfdr}(\mathcal{R}) > \alpha = \frac{1}{1+\lambda}$, then we can improve $\mathcal{R}$ by excluding its last discovery.[1] Let $\mathcal{R}^{-1}$ denote the procedure that makes one fewer rejection than $\mathcal{R}$, meaning it rejects $H_{(1)}, \ldots, H_{(R-1)}$ if $R > 0$, and makes no rejections if $R = 0$. Then we have

$$\begin{aligned}
\mathbb{E}[L_\lambda(H, \mathcal{R}) - L_\lambda(H, \mathcal{R}^{-1})] &= \frac{1}{m}\mathbb{E}\left[(1+\lambda)1\{H_{(R)} = 0, R > 0\} - 1\{R > 0\}\right] \\
&= \frac{1+\lambda}{m}\left(\text{max-lfdr}(\mathcal{R}) - \alpha\mathbb{P}\{R > 0\}\right),
\end{aligned}$$

which is positive if $\text{max-lfdr}(\mathcal{R}) > \alpha$. The converse, that dropping the last rejection does not improve the risk if $\text{max-lfdr}(\mathcal{R}) \leq \alpha$, is almost true if $\mathbb{P}\{R > 0\} \approx 1$. Under the global null, however, any procedure is improved by making fewer rejections.

This thought experiment — what if we dropped the last rejection? — is at the heart of our motivation for proposing the max-lfdr as an error criterion. Even when a rejection set's

[1] Without the shape constraint on f_1, max-lfdr $> \alpha$ still implies that the analyst could improve the procedure by removing the least promising rejection, which may not be the same as the last rejection. However, this improvement is only feasible if the analyst can recognize which rejection is least promising.

average quality is high, the rejections near the threshold may be recognizably bad bets. In that case, we are better off "trimming the fat" from our rejection set until all of the rejections that remain are individually worth following up on.

Because max-lfdr$(\mathcal{R}) \leq$ FDR$(\mathcal{R})$, controlling the max-lfdr is more conservative than controlling FDR at the same level q, in most cases considerably so. From this, it is tempting to conclude that max-lfdr control is an inherently more conservative goal than FDR control, but this conclusion would be mistaken. An analyst whose break-even exchange rate is $\lambda = 9$ and break-even tolerance is $\alpha = 0.1$, for example, would never choose a method with a 10% FDR; the resulting rejection set would be no better on average than rejecting nothing at all, so there would be no point in collecting the data in the first place. Thus, an analyst who is satisfied with a 10% FDR must have a larger break-even tolerance, say $\alpha = 0.2$ or 0.3.

By the same token, it would be unfair to evaluate the risk under L_λ of the BH procedure at level $q = \alpha = \frac{1}{1+\lambda}$, since an analyst whose break-even tolerance is α would want to control FDR at a strictly smaller level q, like $\alpha/2$ or $\alpha/10$. However, as we show in Section 5.3.1, the performance of BH(q) with such *a priori* choices of q can depend sensitively on the unknown alternative density f_1.

5.1.3 Outline and contributions

In Section 5.2, we state and prove our main result, that max-lfdr$(\mathcal{R}_q) = \pi_0 q$ under the Bayesian two-groups model with nonincreasing f_1, applying a result of Takács (1967). Even without monotonicity of f_1, we have $\mathbb{P}\{H_{(R_q)} = 0, R_q > 0\} = \pi_0 q$, but monotonicity ensures that the lfdr is not out of control for rejections in the interior of the rejection region. We also prove max-lfdr control for an adaptive method that estimates π_0 from the data in the same way as the procedure of Storey (2002).

In Section 5.3, we investigate our method's asymptotic performance relative to the oracle procedure $\mathcal{R}^*$. Extending asymptotic results for the Grenander estimator, we show that our method's attained lfdr threshold, lfdr(τ_q), concentrates at a rate $m^{-1/3}$ around its expectation $\pi_0 q$, giving an explicit formula for its asymptotic distribution. We also show that our method's asymptotic regret relative to the oracle shrinks at the rate $m^{-2/3}$. Section 5.4 illustrates our results with selected simulations, and Section 5.5 concludes.

5.2 Finite-sample max-lfdr control

5.2.1 Main result

Our main result is that our procedure $\mathcal{R}_q$ controls the max-lfdr at exactly $\pi_0 q$.

Theorem 5.1. *Suppose $p_1, \ldots, p_m$ follow the Bayesian two-groups model* (5.1)*, with $f_0 = 1_{[0,1]}$. For the procedure defined in* (5.4)*, we have*

$$\mathbb{E}\left[\text{lfdr}\left(p_{(R_q)}\right) \cdot 1\{R_q > 0\}\right] = \mathbb{P}\left\{H_{(R_q)} = 0, R_q > 0\right\} = \pi_0 q. \tag{5.10}$$

If f_1 is nonincreasing, then we have

$$\text{max-lfdr}(\mathcal{R}_q) = \pi_0 q.$$

The familiar optional-stopping arguments from the FDR control literature, introduced by Storey et al. (2004), do not seem to apply to our procedure, since the minimizer R_q of the sequence $p_{(k)} - qk/m$ for $k = 0, \ldots, m$ is not a stopping time. We instead prove Theorem 5.1 via a conditioning argument, which crucially relies on the fact that each null p-value has exactly a q/m chance of being the last rejection $p_{(R_q)}$:

Lemma 5.2. *Fix $p_1, \ldots, p_{m-1} \in [0, 1]$ and let $p_m \sim \text{Unif}(0, 1)$. Then $\mathbb{P}\{p_{(R_q)} = p_m\} = q/m$.*

Given Lemma 5.2, the proof of Theorem 5.1 is straightforward:

Proof of Theorem 5.1. Because the (H_i, p_i) pairs are independent and identically distributed, we can decompose the probability in (5.10) as

$$\begin{aligned}
\mathbb{P}\left\{H_{(R_q)} = 0, R_q > 0\right\} &= \sum_{i=1}^{m} \mathbb{P}\left\{H_i = 0, p_{(R_q)} = p_i\right\} \\
&= m\mathbb{P}\left\{H_m = 0, p_{(R_q)} = p_m\right\} \\
&= \pi_0 m\mathbb{P}\left\{p_{(R_q)} = p_m \mid H_m = 0\right\} \\
&= \pi_0 q,
\end{aligned}$$

where the last step comes from conditioning on $p_1, \ldots, p_{m-1}$ and applying Lemma 5.2. If $f_1(t)$ is nonincreasing, then $\text{lfdr}(t)$ is nondecreasing, so that $\max_{i \in \mathcal{R}_q} \text{lfdr}(p_i) = \text{lfdr}(p_{(R_q)})$ almost surely, completing the argument. □

We now turn to proving Lemma 5.2. Because p_m is uniform, the probability statement is equivalent to a showing that, for any fixed $p_1, \ldots, p_{m-1} \in [0, 1]$, the set of "winning values" $p_m \in [0, 1]$, for which $\tau_q(p_1, \ldots, p_m) = p_m$, has Lebesgue measure q/m. To prove this fact, we rely on a useful result of Takács (1967), which we state next:

Lemma 5.3. *Takács, 1967, Theorem 1 Let $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ denote a nondecreasing step function with $\varphi(0) = 0$. Assume that, for some positive q, we have $\varphi(u + q) = \varphi(u) + \varphi(q)$ for all $u \geq 0$, and define*

$$\delta(u) = 1\{v - \varphi(v) \geq u - \varphi(u) \text{ for all } v \geq u\},$$

Then we have

$$\int_0^q \delta(u)\mathrm{d}u = (q - \varphi(q))_+ .$$

Lemma 5.2 is proved by designing a function φ for which the corresponding indicator $\delta(p_m)$ in Lemma 5.3 checks whether p_m is the last rejection when we run our method on $(p_i)_{i=1}^m$.

Proof of Lemma 5.2. Let $F_{m-1}(t) = \frac{1}{m-1}\sum_{i=1}^{m-1} 1\{p_i \le t\}$ denote the ecdf of $p_1, \dots, p_{m-1}$, and define a new function φ on $[0, q]$

$$\varphi(v) \coloneqq \begin{cases} qF_{m-1}(v)\frac{m-1}{m} & \text{if } v < q \\ q\frac{m-1}{m} & \text{if } v = q. \end{cases}$$

Next, extend φ to a nondecreasing step function on all of $\mathbb{R}_+$ by $\varphi(kq + v) = k\varphi(q) + \varphi(v)$ for all positive integers k and $v \in [0, q]$.

Now let $F_m(t) = \frac{1}{m}\sum_{i=1}^{m} 1\{p_i \le t\}$. If $\tau_q = p_{(R_q)} = p_m$ then we have $p_m - \frac{qR_q}{m} \le p_{(0)} - q\frac{0}{m} = 0$, so we may restrict our attention to $p_m \le q$. On the range $v \in [p_m, q)$ we have

$$mF_m(v) = 1 + (m-1)F_{m-1}(v), \quad \text{so} \quad \varphi(v) = qF_m(v) - \frac{q}{m}.$$

On the range $v \in [q, q + p_m)$, we have

$$mF_m(v-q) = (m-1)F_{m-1}(v-q), \quad \text{so} \quad \varphi(v) = qF_m(v-q) - \frac{q}{m} + q.$$

Letting $\delta(p_m) \coloneqq 1\{p_m = \tau_q(p_1, \dots, p_m)\}$,

$$\begin{aligned} \delta(p_m) &= 1\{v - qF_m(v) \ge p_m - qF_m(p_m) \text{ for all } v \in [0,1]\} \\ &= 1\left\{v - \varphi(v) \ge p_m - \varphi(p_m) \text{ for all } v \in [p_m, q + p_m)\right\} \\ &= 1\left\{v - \varphi(v) \ge p_m - \varphi(p_m) \text{ for all } v \ge p_m\right\}, \end{aligned}$$

where the last step follows from the fact that $\varphi(v) > \varphi(v-q)$ for all $v \ge q + p_m$. We have checked the conditions of Lemma 5.3 Takács, 1967, Theorem 1, from which we conclude

$$\mathbb{P}(\tau_q = p_m) = \int_0^q \delta(p_m)\mathrm{d}p_m = (q - \varphi(q))_+ = \frac{q}{m}. \qquad \square$$

To convey some intuition for our result, Figure 5.2 depicts an illustrative example, highlighting in green the "winning values" of p_m such that $\hat{\tau}_q = p_m$.

Remark 5.4. *Because the set of "winning values" in Lemma 5.2 is a subset of* $[0, q]$ *with Lebesgue measure* q/m*, we can trivially extend the result to conclude* $\mathbb{P}\{p_{(R_q)} = p_m\} \le q/m$*, if* p_m *is drawn from any density* f_0 *with* $f_0(t) \le 1$ *for all* $t \in [0, q]$*. Likewise, we can extend Theorem 5.1 to show that* $\text{max-lfdr}(\mathcal{R}_q) \le \pi_0 q$ *with a more general null density* f_0*, as long as* $\text{lfdr}(t)$ *is nondecreasing and* $f_0(t) \le 1$ *for all* $t \in [0, q]$*.*

5.2.2 Estimating π_0

Theorem 5.1 parallels the exact FDR guarantee $\text{FDR}(\mathcal{R}_q^{\text{BH}}) = \pi_0 q$ for the BH procedure. If we bound $\pi_0 \le 1$, we can run our method at level $q = \alpha$ and ensure that we conservatively

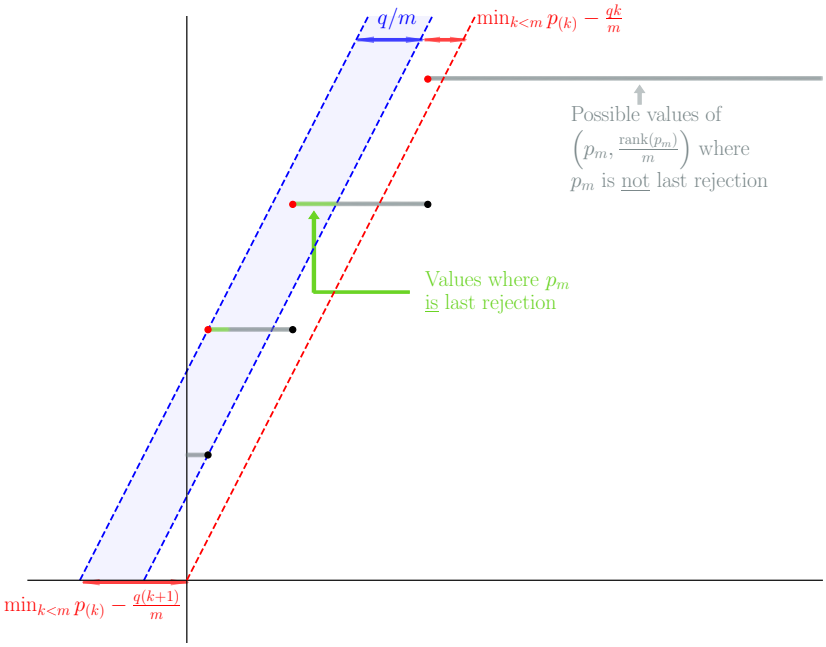


Figure 5.2: Intuition for Lemma 5.2. Black points represent the empirical cdf (scaled by $\frac{m-1}{m}$) of $p_1, \ldots, p_{m-1}$; red points represent how the empirical cdf gets shifted after adding a point p_m to its left. Adding a point p_m can shift the supporting line by at most $\frac{q}{m}$, and each possible shift in $[0, q/m]$ corresponds to precisely one p_m where p_m becomes the new support point.

control max-lfdr at $\pi_0\alpha$, but our method will be overly conservative. In this section, we consider estimating π_0 using the Storey (2002) estimator of the null proportion, defined as

$$\hat{\pi}_0^\zeta := \frac{1 + \#\{i : p_i > \zeta\}}{(1-\zeta)m}, \tag{5.11}$$

modifying an estimator originally proposed by Schweder and Spjøtvoll (1982).

Our next result shows that plugging in $\hat{\pi}_0^\zeta$ and running a modification of our procedure at level $q = \alpha/\hat{\pi}_0^\zeta$ controls max-lfdr at level α in finite samples:

Theorem 5.5. *Suppose $p_1, \ldots, p_m$ follow the Bayesian two-groups model (5.1), with $f_0 = 1_{[0,1]}$ and f_1 nonincreasing. Fix $\zeta \in (0,1)$, and define a modified version of our SL procedure that only examines order statistics below ζ:*

$$R_q^\zeta := \underset{k \geq 0:\, p_{(k)} \leq \zeta}{\text{argmin}} \; \hat{\pi}_0^\zeta p_{(k)} - \frac{qk}{m}, \tag{5.12}$$

and $\mathcal{R}_q^\zeta = \{i : \ p_i \leq p_{(R_q^\zeta)}\}$. *Then we have*

$$\text{max-lfdr}\left(\mathcal{R}_q^\zeta\right) \ = \ q \cdot \frac{(1-\zeta)\pi_0}{1-F(\zeta)} \cdot (1-F(\zeta)^m) \ \leq \ q.$$

The proof of Theorem 5.5 is deferred to the Appendix. The method $\mathcal{R}_\alpha^\zeta$ coincides with $\mathcal{R}_{\alpha/\hat{\pi}_0^\zeta}$, our original procedure applied at the corrected level $\hat{q} = \alpha/\hat{\pi}_0^\zeta$, whenever $\tau_{\hat{q}} \leq \zeta$. Since we usually have $\tau_{\hat{q}} \ll 0.5 \leq \zeta$, the two methods are identical for all practical purposes.

In the next section, we will investigate the asymptotic regret of methods that estimate π_0. In particular, we will show that this estimation error is asymptotically negligible if it shrinks at a faster rate than $m^{-1/3}$. We can indeed achieve this with $\hat{\pi}_0^\zeta$ if f_1 has two continuous derivatives in a neighborhood of 1, with $f_1'(1) = f_1(1) = 0$. By Taylor's theorem, we have

$$1 - F(\zeta) \ = \ (1-\zeta)\pi_0 + \frac{(1-\pi_0)f_1''(\xi)}{6}(1-\zeta)^3,$$

for some $\xi \in [\zeta, 1]$. Assuming $\pi_0 \in (0,1)$ and taking $\zeta = 1 - m^{-1/5}$, we then have

$$m^{2/5}\left(\hat{\pi}_0^\zeta - \pi_0\right) \ \sim \ m^{2/5}\left(\frac{1 + \text{Binom}\left(m, 1-F(\zeta)\right)}{(1-\zeta)m} - \pi_0\right) \ \xrightarrow{d} \ \mathcal{N}\left(\frac{(1-\pi_0)f_1''(1)}{6}, \pi_0\right), \tag{5.13}$$

with subgaussian errors for finite m, so the results in Section 5.3.3 generally apply. See Genovese and Wasserman (2004) and Patra and Sen (2016) for a discussion of estimators for π_0.

5.3 Asymptotic regret analysis

In this section, we study our procedure's empirical Bayes regret under the weighted classification risk $\mathbb{E}\left[L_\lambda(H, \mathcal{R})\right]$, where the expectation is taken over $H_1, \ldots, H_m$ and $p_1, \ldots, p_m$ according to (5.1), and L_λ is defined as in (5.5). Throughout this section we will be considering a sequence of problems with $m \to \infty$.

A fundamental result of Sun and Cai (2007) is that the oracle (5.6) minimizes the weighted classification risk over all procedures, thus representing a benchmark against which we can compare methods that are feasible without *a priori* knowledge the lfdr. In the empirical Bayes literature (see, e.g., Efron, 2019), the price of our ignorance of the model parameters is measured by the *regret*, or average excess risk, given by the optimality gap

$$\text{Regret}_m(\mathcal{R}) \coloneqq \mathbb{E}\left[L_\lambda(H, \mathcal{R}) - L_\lambda\left(H, \mathcal{R}^*\right)\right]. \tag{5.14}$$

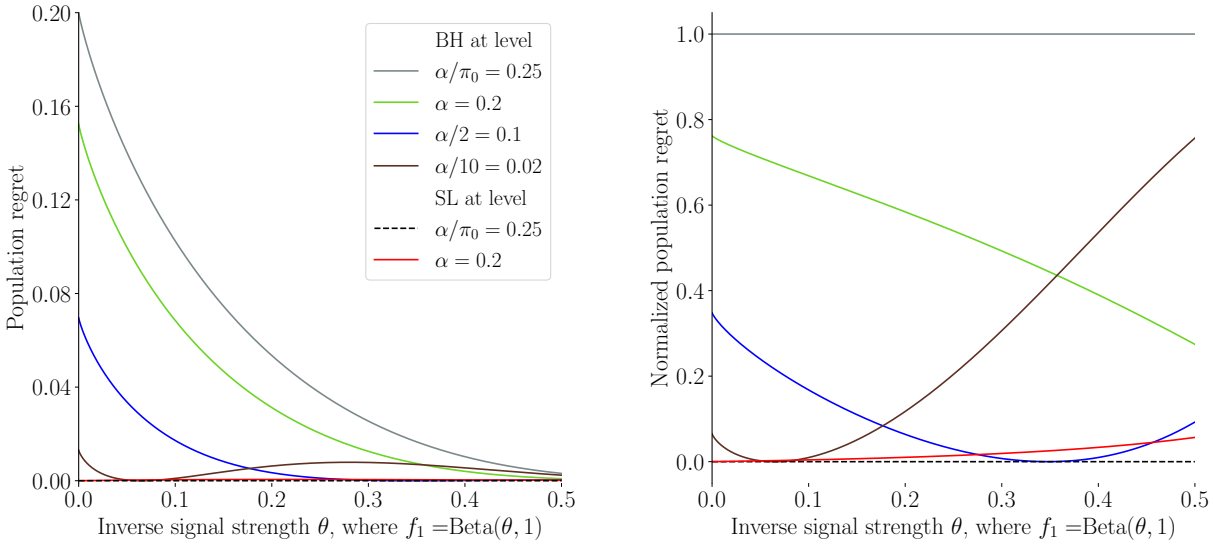


Figure 5.3: Left: The fixed-threshold regret $\rho(t)$ (5.15) with Beta alternatives $f_1(t) = \theta t^\theta$ as a function of $\theta \in [0, .5]$. Right: a normalized version $\rho(t)/\rho(0)$, such that BH at level α/π_0 has unit normalized regret, identical to the regret of the procedure that rejects nothing. The null proportion is $\pi_0 = 0.8$ and the cost-benefit ratio is $\lambda = 4$.

5.3.1 Population regret

Before tackling the more delicate problem of calculating the regret for procedures with data-dependent p-value rejection thresholds, we first investigate the regret of fixed-threshold methods. For $t \in [0, 1]$, let $\mathcal{R}_t^{\text{Fix}} := \{i : \ p_i \le t\}$, and note that the oracle method is $\mathcal{R}^* = \mathcal{R}_{\tau^*}^{\text{Fix}}$. We introduce the function $\rho(t)$ to represent the regret of this method, which is free of m:

$$\rho(t) \ := \ \text{Regret}_m(\mathcal{R}_t^{\text{Fix}}) \ = \ F(\tau^*) - F(t) - \frac{\pi_0}{\alpha}(\tau^* - t). \tag{5.15}$$

If $\text{lfdr}(\tau^*) = \alpha$, then we also have $f(\tau^*) = \pi_0/\alpha$, and $\rho(t)$ is simply the error of the first-order Taylor expansion of F around τ^*, also known as the Bregman divergence associated with $-F$. If f is continuously differentiable between t and τ^*, then

$$\rho(t) \ = \ \frac{-f'(\xi_t)}{2}\,(t - \tau^*)^2\,, \qquad \text{for some } \xi_t \text{ between } t \text{ and } \tau^*. \tag{5.16}$$

Since F is concave, $\rho(t) \ge 0$. Finally, we can also rewrite (5.15) as an integral

$$\rho(t) \ = \ \int_\tau^{\tau^*} \big(1 - \alpha^{-1}\text{lfdr}(t)\big)\, \mathrm{d}F(t). \tag{5.17}$$

This form for the regret underscores the relationship between the lfdr and the regret, and will prove useful for analyzing the regret with data-dependent thresholds.

We can evaluate ρ to investigate the regret of population versions of our procedure and the BH procedure, i.e. versions of the procedures with rejection thresholds chosen using the true cdf F in place of the empirical cdf F_m. The population BH threshold at an arbitrary level $q \in (0,1)$ is found by intersecting F with the ray of slope q^{-1}, i.e.

$$t_q^{\text{BH-POP}} \coloneqq \max\left\{t \in [0,1] : \ F(t) - t/q = 0\right\}.$$

By comparison, the population version of our procedure τ_q is

$$t_q \coloneqq \max\left\{t \in [0,1] : \ f(t) \le q^{-1}\right\},$$

which coincides with the oracle threshold τ^* when $q = \alpha/\pi_0$. Note that t_q is equivalent to the population BH threshold $t_{q'}^{\text{BH-POP}}$ at the lower level

$$q' = \frac{t_q}{F(t_q)}. \tag{5.18}$$

Thus, there is always *some* value q' for which the BH procedure approximately reproduces the oracle, namely $t_{\alpha/\pi_0}/F(t_{\alpha/\pi_0})$, but generally we cannot use it unless we know f_1 and π_0.

To illustrate the population regret in a concrete example, we consider a parametric alternative distribution

$$f_1(t;\theta) \coloneqq \theta t^{\theta-1} \qquad \text{for some } \theta \in (0,1),$$

which is a Beta$(\theta, 1)$ density. This form is called a *Lehmann alternative* in the multiple testing literature (see, e.g., Pounds & Morris, 2003). In this case, the population procedures at level $q \in (0,1)$ use rejection thresholds

$$t_q \ = \ \left(\frac{q^{-1} - \pi_0}{(1-\pi_0)\theta}\right)^{-\frac{1}{1-\theta}}, \quad \text{and} \quad t_q^{\text{BH-POP}} \ = \ \left(\frac{q^{-1} - \pi_0}{1-\pi_0}\right)^{-\frac{1}{1-\theta}}.$$

Furthermore, the threshold equivalence (5.18) gives

$$q' \ = \ \frac{\theta q}{1 - (1-\theta)\pi_0 q} \ \approx \ \theta q,$$

where the approximation holds for small values of q. Thus, the correspondence between q and q' depends on the parameter θ, which controls the signal strength under the alternative. For small values of θ, the signal is very strong, and the "correct" choice of q' is much smaller than the desired max-lfdr level α, but for weaker signals (larger θ), we should choose q' closer to α. Without knowing the signal strength in advance, it is difficult to know at what values of q' the BH method will perform well.

In Figure 5.3 we plot the population regret for various choices of the level of the procedure, $\pi_0 = 0.8$ and $\lambda = 4$ and varying the parameter θ. The population version of our procedure

at level $\frac{\alpha}{\pi_0}$ with $\alpha = \frac{1}{1+\lambda} = 0.2$ is the oracle (5.6), so it achieves zero regret, while the conservative version of our procedure with $q = \alpha$ performs quite well for all values of the alternative parameter θ. In this example, the asymptotic error incurred from conservatively bounding π_0 by one in the procedure is small compared to the error incurred by using BH(q') at an *ad hoc* value. The BH procedure at levels $\frac{\alpha}{\pi_0}$ or α incurs substantial asymptotic regret by comparison. In particular, note that the BH(α/π_0) procedure incurs the same asymptotic regret as the procedure that rejects nothing; i.e. $\rho(t^{\text{BH-POP}}_{\alpha/\pi_0}) = \rho(0)$. If we run BH at a lower level like $\alpha/2$, $\alpha/10$, or $\alpha/100$, we can do well for some range of θ values, but struggle at other parts of the parameter space. No single level for BH dominates in terms of regret, so for the classification risk it is more appropriate to view the BH level as a tuning parameter (Neuvial & Roquain, 2012).

5.3.2 Relationship of our method to the Grenander estimator

Since the marginal density f appears in the denominator of the lfdr, bounding $\pi_0 \leq 1$ and plugging in Grenander's estimator $\hat{f}_m$ (defined in (5.8)) gives the conservative estimate

$$\widehat{\text{lfdr}}(t) \coloneqq \frac{1}{\hat{f}_m(t)}, \quad t \in [0,1].$$

Similar to how the BH procedure chooses an interval $[0,t]$ as large as possible subject to a constraint on an estimate of the FDP, the rejection threshold of the SL procedure can be equivalently expressed as

$$\tau_q = \underset{p_{(0)},\ldots,p_{(m)}}{\text{argmax}} \left\{ \frac{qk}{m} - p_{(k)} \right\} = \sup \left\{ t \in [0,1] : \widehat{\text{lfdr}}(t) \leq q \right\}, \tag{5.19}$$

taking the convention that $\sup \emptyset \equiv 0$. The equivalence in (5.19) is illustrated in Figure 5.4. Let $\hat{F}_m$ denote the least concave majorant of the empirical cdf F_m, plotted as a dotted blue line in the left panel of Figure 5.4. By definition of $\widehat{\text{lfdr}}(t)$, the supremum on the right hand side is equal to the largest t for which $\frac{\mathrm{d}}{\mathrm{d}t}(q\hat{F}_m(t) - t) = q\hat{f}_m(t) - 1 \geq 0$, which corresponds to the maximizer of the function $q\hat{F}_m(t) - t$, illustrated for example in the right panel of Figure 5.4. $\hat{F}_m \geq F_m$ implies

$$q\hat{F}_m(t) - t \geq qF_m(t) - t, \quad t \in [0,1],$$

with equality at the knots of $\hat{F}_m$, and since the maximizer of the left hand side occurs at a knot of $\hat{F}_m$, it is also the maximizer of the right hand side, i.e. the argmax of $\frac{qk}{m} - p_{(k)}$.

We can again compare this result with the BH(q) threshold, given by

$$\tau_q^{\text{BH}} = \max_{k=0,\ldots,m} \left\{ p_{(k)} : \frac{qk}{m} - p_{(k)} \geq 0 \right\} = \sup \left\{ t \in [0,1] : F_m(t) \geq q^{-1}t \right\},$$

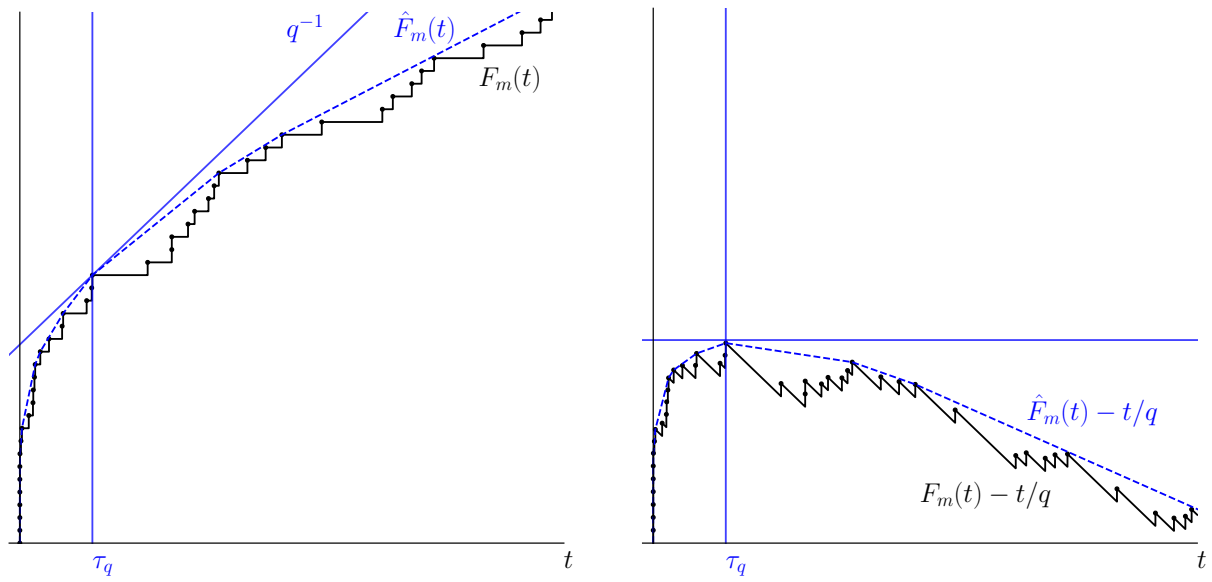


Figure 5.4: Left: empirical cdf F_m and its least concave majorant $\hat{F}_m$. The support line of slope q^{-1} touches both curves at the decision threshold τ_q. Right: the same plot with the line t/q subtracted off.

which is the largest t for which the ray $q^{-1}t$ lies below the ecdf $F_m(t)$. Our procedure instead finds the last intersection of the graph of F_m with a support line of slope q^{-1}, since

$$\widehat{\text{lfdr}}(t) \le q \iff \hat{f}_m(t) \ge q^{-1}.$$

This relationship is illustrated in the left panel of Figure 5.4.

5.3.3 Asymptotic behavior of our procedure

Equation (5.16) suggests that, when f is sufficiently regular near τ^*, the regret is closely related to the squared error of the rejection threshold. Our main result in this section establishes cube-root asymptotics for the behavior of our procedure $\mathcal{R}_q$ with $q = \alpha/\hat{\pi}_0$, where $\hat{\pi}_0$ consistently estimates π_0; if π_0 is known, then the results apply directly with $\hat{\pi}_0 = \pi_0$.

We derive limiting distributions for the threshold τ_q, the lfdr at the threshold, and the regret of $\mathcal{R}_q$. All three are given in terms of Chernoff's distribution (Chernoff, 1964), which is defined as the distribution of the maximizer Z of a standard two-sided Brownian motion $W = (W(t))_{t\in\mathbb{R}}$ with parabolic drift:

$$Z = \operatorname*{argmax}_{t\in\mathbb{R}} W(t) - t^2. \tag{5.20}$$

The random variable Z has a density with respect to the Lebesgue measure on $\mathbb{R}$ that is symmetric about zero. Dykstra and Carolan (1999) suggest approximating the density and cdf of Z by those of $\mathcal{N}(0, (.52)^2)$. This approximation can be somewhat crude but gives a rough sense for the distribution of Z. Groeneboom and Wellner (2001) provide much more accurate numerical methods to compute the density, cdf, quantiles and moments of Z.

Theorem 5.6. *Suppose $p_1, \ldots, p_m$ follow the Bayesian two-groups model (5.1), with $\pi_0 \in (0,1)$, $f_0 = 1_{[0,1]}$, and f_1 nonincreasing. For $q \in (0, \pi_0^{-1})$, assume additionally that*

(i) there is a unique value $t_q \in (0,1)$ for which $f(t_q) = q^{-1}$,

(ii) f is continuously differentiable in a neighborhood of t_q with $f'(t_q) < 0$, and

(iii) $\hat{q}$ is any random variable with $m^{1/3}(\hat{q} - q) \xrightarrow{p} 0$ as $m \to \infty$.

Then we have, as $m \to \infty$,

$$m^{1/3}(\tau_{\hat{q}} - t_q) \xrightarrow{d} \left(\frac{q}{4} \cdot f'(t_q)^2\right)^{-1/3} Z, \qquad \text{and} \tag{5.21}$$

$$m^{1/3} \cdot \frac{\mathrm{lfdr}(\tau_{\hat{q}}) - \pi_0 q}{\pi_0 q} \xrightarrow{d} \left(4q^2 \cdot |f'(t_q)|\right)^{1/3} Z. \tag{5.22}$$

where Z follows Chernoff's distribution defined in (5.20). Further, suppose that

$$\mathbb{P}\{m^{-1/3}(\hat{q} - q) > \varepsilon\} = o\left(m^{-2/3}\right), \quad \text{for all } \varepsilon > 0. \tag{5.23}$$

Then we also have $m^{1/3}\mathbb{E}[\tau_{\hat{q}}] \to t_q$. In addition,

$$m^{2/3}\mathrm{Var}(\tau_{\hat{q}}) \to \left(\frac{q}{4} \cdot f'(t_q)^2\right)^{-2/3} \mathrm{Var}(Z), \qquad \text{and} \tag{5.24}$$

$$m^{2/3}\mathrm{Var}\left(\frac{\mathrm{lfdr}(\tau_{\hat{q}}) - \pi_0 q}{\pi_0 q}\right) \to \left(4q^2 \cdot |f'(t_q)|\right)^{2/3} \mathrm{Var}(Z), \tag{5.25}$$

where $\mathrm{Var}(Z) \approx 0.26$.

The proof of Theorem 5.6 is deferred to Appendix 5.6.2. It is well-known that the Grenander estimator $\hat{f}_m$ estimates f at a cube root rate pointwise, away from zero, but this result, due to (Rao, 1969), is too weak to describe the behavior of our procedure. We rely on a stronger version of this result due to Dümbgen et al. (2016) that approximates the local behavior of the Grenander estimator near t_q.

The distributional result (5.22) complements our result from Theorem 5.1, by showing that $\mathrm{lfdr}(\tau_q) = \max_{i \in \mathcal{R}_q} \mathrm{lfdr}(p_i)$ is not only controlled in expectation, but also concentrates at rate $m^{-1/3}$ around its expectation. In particular, because $\mathbb{P}\{Z \geq 1\} \approx 0.05$, we have

$$\frac{\mathrm{lfdr}(\tau_q) - \pi_0 q}{\pi_0 q} \leq m^{-1/3} \left(4q^2 \cdot |f'(t_q)|\right)^{1/3},$$

with roughly 95% probability in large samples. For example, suppose we use $q = 0.2$, so $f(t_q) = 5$, and suppose that $f'(t_q) = -50$. Then, whereas Theorem 5.1 guarantees $\mathbb{E}[\text{lfdr}(\tau_q)] \leq 0.2$ exactly, the asymptotic estimate from Theorem 5.6 bounds the 95th percentile of $\text{lfdr}(\tau_q)$ at 0.24 if $m = 1000$, or at 0.21 if $m = 64,000$.

To understand why the error is of order $m^{-1/3}$, consider fixed q and recall that the threshold τ_q maximizes the stochastic process

$$U(t) \coloneqq F_m(t) - F_m(t_q) - \frac{t - t_q}{q}.$$

Because $f(t_q) = q^{-1}$, we have for t near t_q,

$$F(t) - F(t_q) \;\approx\; \frac{t - t_q}{q} + \frac{f'(t_q)}{2}(t - t_q)^2.$$

Introducing the local parameterization $t = t_q + m^{-a}h$ for $a > 0$ leads to

$$U(t_q + m^{-a}h) \;\approx\; -\frac{|f'(t_q)|}{2} \cdot \frac{h^2}{m^{2a}} + \mathcal{N}\left(0, \frac{h}{qm^{a+1}}\right).$$

Setting $a = 1/3$ balances the mean and variance, giving

$$m^{2/3}U(t_q + m^{-1/3}h) \;\xrightarrow{d}\; -\frac{|f'(t_q)|}{2}h^2 + \mathcal{N}\left(0, \frac{h}{q}\right).$$

Under this local scaling, $U(t)$ converges to a Brownian motion with parabolic drift, and its maximizer τ_q converges to Chernoff's distribution. Theorem 5.6 applies a more careful version of this argument, replacing $F_m(t)$ with its LCM $\hat{F}_m(t)$ and using a result of Dümbgen et al. (2016) to characterize the process $\hat{f}_m(t)$ under the same local scaling. The corresponding results for $\text{lfdr}(\tau_q)$ follow from first-order Taylor expansion of $\text{lfdr}(t) = \pi_0/f(t)$ around t_q.

By specializing Theorem 5.6 to $q = \alpha/\pi_0$ and $\hat{q} = \alpha/\hat{\pi}_0$, we obtain the limiting regret for our procedure with a known or accurately estimated null proportion.

Theorem 5.7. *Suppose $p_1, \ldots, p_m$ follow the Bayesian two-groups model (5.1), with $\pi_0 \in (0, 1)$, $f_0 = 1_{[0,1]}$, and f_1 nonincreasing. Assume additionally that*

(i) there is a unique value $\tau^* \in (0, 1)$ for which $\text{lfdr}(\tau^*) = \frac{\pi_0}{f(\tau^*)} = \alpha$,

(ii) f is continuously differentiable in a neighborhood of τ^* with $f'(\tau^*) < 0$, and

(iii) $\hat{\pi}_0$ is any estimator of π_0 with $\mathbb{P}\left\{m^{1/3}(\hat{\pi}_0 - \pi_0) > \varepsilon\right\} = o\left(m^{-2/3}\right)$ for all $\varepsilon > 0$.

Then we have, as $m \to \infty$,

$$m^{2/3}\text{Regret}_m(\mathcal{R}_{\alpha/\hat{\pi}_0}) \;\to\; \left(\frac{\alpha^2}{2\pi_0^2} \cdot |f'(\tau^*)|\right)^{-1/3} \text{Var}(Z), \tag{5.26}$$

where Z follows Chernoff's distribution defined in (5.20), and $\text{Var}(Z) \approx 0.26$.

Theorems 5.6–5.7 deal with the regret for $\pi_0 \in (0, 1)$. Under the global null, represented in the Bayesian model by $\pi_0 = 1$, the behavior is different and the regret is simply $\lambda \mathbb{E}V$, which is $O(m^{-1})$, as we see next.

Proposition 5.8. *Suppose* $(p_i)_{i=1}^m$ *follow a two-groups model* (5.1) *with* $f_0 = 1_{[0,1]}$ *and* $\pi_0 = 1$, *i.e.* $H_i = 0$ *for all* i *and* $p_i \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$. *Then as* $m \to \infty$, *we have*

$$m \operatorname{Regret}_m(\mathcal{R}_q) \to \lambda \sum_{k=1}^{\infty} \mathbb{P}\left\{U_k \le q\right\}, \qquad \text{for } U_k \sim \text{Gamma}(k, k),$$

which is finite for every $q \in [0, 1)$.

Proposition 5.8 is closely related to results derived in Finner and Roters (2001).

5.4 Numerical results

This section highlights our main results on simulation experiments. We adapt a simulation setting of Benjamini and Hochberg (1995) to the two-groups model (5.1). Specifically, define the alternative density

$$f_1(t) = \frac{\frac{1}{4}\sum_{i=1}^{4} \phi\left(\bar{\Phi}^{-1}(t) - 5\frac{i}{4}\right)}{\phi\left(\bar{\Phi}^{-1}(t)\right)} \qquad \text{for } 0 \le t \le 1, \tag{5.27}$$

where ϕ and $\bar{\Phi}$ denote the density and survival function of the standard Gaussian distribution. Concretely, a non-null p-value $p_i \sim f_1$ can be constructed by first taking $Y_i \sim \mathcal{N}(\mu_i, 1)$ where μ_i is drawn uniformly at random from the set $\{5\frac{i}{4} : i = 1, 2, 3, 4\}$; then, $p_i = \bar{\Phi}(Y_i)$ is a one sided p-value for the null-hypothesis that $\mu_i = 0$. We use a null proportion of $\pi_0 = 0.75$. Figure 5.5 shows the mixture density and corresponding lfdr.

We repeatedly sampled from the above two-groups model with $m = 64$ hypotheses. Figure 5.5 shows the FDR (left panel) and max-lfdr (right panel) for both our procedure and the BH procedure, at conservative level $q = \alpha$ and estimated level $\hat{q} = \alpha/\hat{\pi}_0^{\zeta}$. The BH procedure, shown in red, achieves FDR exactly $\pi_0 q$, whereas the max-lfdr can be much larger. By contrast, our procedure, shown in blue, conservatively controls FDR substantially below the level $\pi_0 q$ but has max-lfdr equal to $\pi_0 q$.

Figure 5.6 shows a log-log plot of the regret as a function of the sample size m. The red curve shows the regret of our uncorrected procedure $\mathcal{R}_\alpha$ for $\alpha = 0.05$, which asymptotically tends to $\rho(t_\alpha)$ and hence asymptotically incurs some non-vanishing regret described in Section 5.3.1. The blue curve shows the regret of the corrected procedure $\mathcal{R}_{\alpha/\pi_0}$ with known π_0. For larger samples, the simulated regret closely matches the asymptotic prediction from (5.26), shown in black. The green curve (which is nearly indistinguishable from the blue curve) shows the corrected procedure with an estimated null proportion $\hat{\pi}_0^{\zeta}$ based on (5.11) with $\zeta = 1 - m^{-1/5}$.

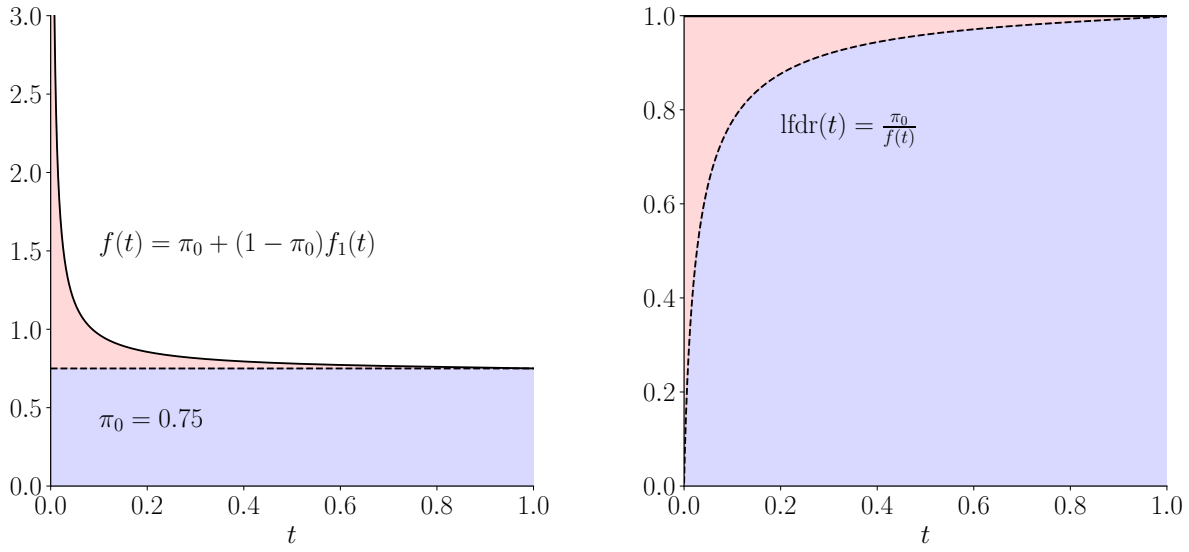
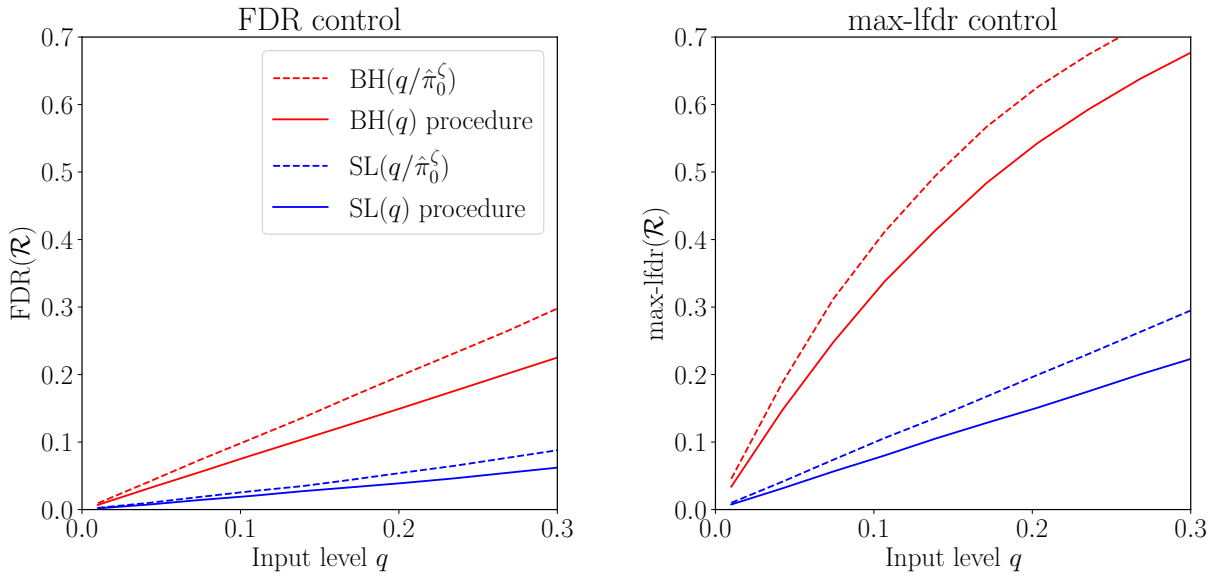


Figure 5.5: Above: Mixture density f (left) and lfdr (right), with alternative density f_1 defined in (5.27) and null proportion $\pi_0 = 0.75$. Note f_1 diverges as $t \downarrow 0$. Below: Comparison of FDR control (left) and max-lfdr control (right) on simulated data. The estimate of the null proportion is (5.11) with $\zeta = 0.5$.



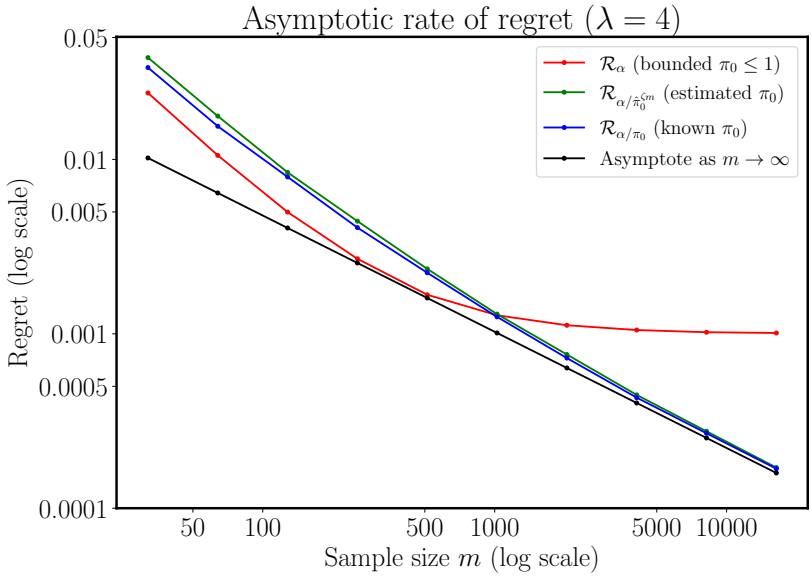


Figure 5.6: A log-log plot of the regret (5.14) as a function of the sample size. The black line shows the asymptotic prediction (5.26) of Theorem 5.7. For this simulation, the alternative density f_1 is defined in (5.27), cost-benefit ratio $\lambda = 19$ and null proportion $\pi_0 = 0.75$.

5.5 Discussion

In this chapter we introduced a new error criterion, the max-lfdr, which modifies the FDR by redirecting attention away from the average quality of the rejection set and toward the rejections that are close to the rejection boundary. Despite the seeming difficulty of measuring the quality of a single rejection, we also introduce a simple new multiple testing procedure that controls the max-lfdr at level $\pi_0 q$ in finite samples, where q is a tuning parameter and π_0 is the null proportion. We assume only that the data follow a Bayesian two-groups model in which smaller p-values reflect stronger evidence against the null. We find that our method is better able than the BH method to adapt to the unknown problem structure, and to perform well without knowledge of the true underlying distribution.

The BH procedure owes its enduring utility for FDR control in part to its versatility beyond this basic setting, however. It is known to still control FDR, for instance, when the null p-values are super-uniform and under certain forms of positive dependence, two of many possible extensions that we leave open for our procedure.

Another seeming advantage of the FDR criterion is that it requires no Bayesian assumptions, whereas the max-lfdr is only defined with reference to a Bayesian model. A possible avenue for generalizing the max-lfdr to frequentist settings is to work with its characterization as the probability that the last rejection is a false discovery. Indeed, our proof of

Theorem 5.1 implies that max-lfdr is controlled even conditional on $H_1, \ldots, H_m$. This is initially puzzling: if each H_i is fixed, then how can we speak of the probability that the last rejection is a false discovery? The answer is that $H_{(R)}$ is random even if $H_1, \ldots, H_m$ are fixed, since its index is random. We leave further development of the frequentist connection to the max-lfdr to future work.

5.6 Proofs

5.6.1 Proofs of results from Section 5.2

Proof of Theorem 5.5. As in the proof of Theorem 5.1, we have

$$\text{max-lfdr}\left(\mathcal{R}_q^\zeta\right) \;=\; \mathbb{P}\left\{H_{(R_q^\zeta)} = 0,\, R_q^\zeta > 0\right\} \;=\; m\,\mathbb{P}\left\{H_m = 0,\, p_{(R_q^\zeta)} = p_m\right\}.$$

Define the σ-field $\mathcal{F} = \sigma\left(p_1, \ldots, p_{m-1}, H_m, 1\{p_m \le \zeta\}\right)$. We restrict our attention to the event $A = \{H_m = 0, p_m \le \zeta\}$, since the event $\{H_m = 0,\, p_{(R_q^\zeta)} = p_m\}$ cannot occur except on A. On A, which is $\mathcal{F}$-measurable, we have $p_m/\zeta \mid \mathcal{F} \sim U[0,1]$.

Let $m^\zeta = \#\{i :\ p_i \le \zeta\}$, which is also $\mathcal{F}$-measurable. If $j_1 \le \cdots \le j_{m^\zeta} = m$ are the indices of the p-values that are below ζ, define the modified p-values $p_i^\zeta = p_{j_i}/\zeta$, for $i = 1, \ldots, m^\zeta$. Because the order statistics of $\zeta p_1^\zeta, \ldots, \zeta p_{m^\zeta}^\zeta$ are also the first m^ζ order statistics of $p_1, \ldots, p_m$, the quantity R_q^ζ defined in (5.12) can be rewritten as

$$\begin{aligned} R_q^\zeta \;&=\; \operatorname*{argmin}_{k=0,\ldots,m^\zeta} \zeta p_{(k)}^\zeta - \frac{q}{\hat\pi_0^\zeta}\cdot\frac{k}{m} \\ &=\; \operatorname*{argmin}_{k=0,\ldots,m^\zeta} p_{(k)}^\zeta - \frac{q^\zeta k}{m^\zeta}, \qquad \text{for } q^\zeta = \frac{q m^\zeta}{\zeta\hat\pi_0^\zeta m}. \end{aligned}$$

Applying Lemma 5.2, we have

$$\mathbb{P}\left\{H_m = 0,\, p_{(R_q^\zeta)} = p_m \mid \mathcal{F}\right\} = \frac{q^\zeta}{m^\zeta}\cdot 1_A = \frac{q}{\zeta\hat\pi_0^\zeta m}\cdot 1_A$$

Marginalizing over $\mathcal{F}$, and noting that $\mathbb{P}(A) = \pi_0\zeta$, we obtain

$$\begin{aligned} \mathbb{P}\left\{H_m = 0,\, p_{(R_q^\zeta)} = p_m\right\} \;&=\; \frac{q}{m}\cdot\mathbb{E}\left[\frac{\pi_0}{\hat\pi_0^\zeta} \mid A\right] \\ &=\; \frac{q}{m}\cdot\frac{(1-\zeta)\pi_0}{1-F(\zeta)}\cdot\mathbb{E}\left[\frac{(1-F(\zeta))m}{1+\#\{i<m:\ p_i>\zeta\}}\right] \\ &=\; \frac{q}{m}\cdot\frac{(1-\zeta)\pi_0}{1-F(\zeta)}\cdot(1-F(\zeta)^m) \\ &\le\; \frac{q}{m}, \end{aligned}$$

completing the proof. The final inequality is a standard binomial identity:

$$\begin{aligned}
\mathbb{E}\left[\frac{\beta m}{1+\text{Binom}(m-1,\beta)}\right] &= \sum_{k=0}^{m-1} \frac{\beta m}{1+k}\binom{m-1}{k}\beta^k(1-\beta)^{m-1-k} \\
&= \sum_{k=0}^{m-1}\binom{m}{k+1}\beta^{k+1}(1-\beta)^{m-(k+1)} \\
&= \sum_{j=1}^{m}\binom{m}{j}\beta^{j}(1-\beta)^{m-j} \\
&= \mathbb{P}\{\text{Binom}(m,\beta)\geq 1\} \\
&= 1-(1-\beta)^m. \qquad \square
\end{aligned}$$

5.6.2 Proofs of results from Section 5.3

Proof of Theorem 5.6. Our proof will use the *switching relation* that states, for any $t \in (0,1)$, we have almost surely

$$\tau_{\hat{q}} \leq t \iff \hat{f}_m(t) \leq \hat{q}^{-1}.$$

We will work with a local expansion of $\hat{f}_m(t)$ around t_q using the local parameterization $t = t_q + m^{-1/3}h$. Using $f(t_q) = q^{-1}$, the switching relation becomes

$$m^{-1/3}(\tau_{\hat{q}} - t_q) \leq h \iff \hat{f}_m\left(t_q + m^{-1/3}h\right) - f(t_q) \leq \hat{q}^{-1} - q^{-1}.$$

Now let W denote a standard two-sided Brownian motion, and let $\mathbb{S}_{a,b}$ denote the process of left derivatives of the least concave majorant of $X_{a,b}(t) = aW(t) - bt^2$, where $a = \sqrt{f(t_q)}$ and $b = |f'(t_q)|/2$. Under our regularity assumptions, Dümbgen et al. (2016) show

$$m^{1/3}\left(\hat{f}_m\left(t_q + m^{-1/3}h\right) - f(t_q)\right) \Rightarrow \mathbb{S}_{a,b}(h)$$

in the Skorokhod topology on $D[-K,K]$ for every finite $K > 0$. Since $m^{1/3}(\hat{q}^{-1} - q^{-1}) \xrightarrow{p} 0$ by assumption, we have

$$\mathbb{P}\left\{m^{1/3}\left(\tau_{\hat{q}} - t_q\right) \leq h\right\} \to \mathbb{P}\left\{\mathbb{S}_{a,b}(h) \leq 0\right\}.$$

Observe that $\mathbb{S}_{a,b}(h) \leq 0$ iff $t^*_{a,b} \leq h$, where $t^*_{a,b}$ is the (a.s. unique) maximizer of $X_{a,b}$ (note the maximizer $t^*_{a,b}$ is always a knot in the concave majorant since the horizontal line with intercept $X_{a,b}(t^*_{a,b})$ is a supporting line intersecting $(t^*_{a,b}, X_{a,b}(t^*_{a,b}))$). Combining this observation with the previous display, we have

$$m^{1/3}\left(\tau_{\hat{q}} - t_q\right) \xrightarrow{d} t^*_{a,b} \overset{d}{=} (b/a)^{-2/3} Z = \left(\frac{q}{4}\cdot f'(t_q)^2\right)^{-1/3} Z,$$

proving (5.21). Next we turn to the lfdr asymptotics. By Taylor's theorem,

$$m^{1/3}\left(\text{lfdr}(\tau_{\hat{q}}) - \pi_0 q\right) = \text{lfdr}'(\omega) \cdot m^{1/3}\left(\tau_{\hat{q}} - t_q\right)$$

for some ω between $\tau_{\hat{q}}$ and t_q. Using

$$\text{lfdr}'(t_q) \;=\; \frac{-\pi_0 f'(t_q)}{f(t_q)^2} \;=\; \pi_0 q^2 \cdot |f'(t_q)|,$$

and applying the continuous mapping theorem and Slutsky's theorem, we obtain

$$\text{lfdr}'(\omega) \cdot m^{1/3}\left(\tau_{\hat{q}} - t_q\right) \;\xrightarrow{d}\; \text{lfdr}'(t_q) \cdot \left(\frac{q}{4} \cdot f'(t_q)^2\right)^{-1/3} Z \;=\; \pi_0 q \cdot \left(4q^2 \cdot |f'(t_q)|\right)^{1/3} Z,$$

proving (5.22). Next, under the strengthened assumption (5.23), fix $\varepsilon > 0$ and define the event

$$A_\varepsilon = \left\{|\hat{q} - q| \le m^{-1/3}\varepsilon,\ |\tau_{\hat{q}} - t_q| \le m^{-2/9}\right\}, \tag{5.28}$$

and the truncated random variable

$$Y_m = m^{1/3}(\tau_{\hat{q}} - t_q) \cdot 1_{A_\varepsilon},$$

We will show that $\mathbb{P}(A_\varepsilon^c) = o\left(m^{-2/3}\right)$. As a result, Y_m has the same limit in distribution as $m^{1/3}(\tau_{\hat{q}} - t_q)$. If we can show that the sequence Y_m^2 is uniformly integrable, we will have convergence of its mean and variance to the mean and variance of its limiting distribution. Then, because

$$\mathbb{E}\left[\left(m^{1/3}(\tau_{\hat{q}} - t_q) - Y_m\right)^2\right] \;\le\; m^{2/3}\mathbb{P}(A_\varepsilon^c) \;\to\; 0,$$

we will have the same limiting mean and variance for $m^{1/3}(\tau_{\hat{q}} - t_q)$.

To show that $\mathbb{P}(A_\varepsilon^c) = o\left(m^{-2/3}\right)$, let $q_1 = q - m^{-1/3}\varepsilon$ and $q_2 = q + m^{-1/3}\varepsilon$ and assume that m is sufficiently large that $m^{-1/3}\varepsilon \le m^{-2/9}/2$, and

$$f'(t) \le f'(t_q)/2, \qquad \text{for all } \; t \in [t_q - m^{-2/9},\ t_q + m^{-2/9}].$$

As a result, for all $t \ge t_{q_2} + m^{-2/9}/2$, we have

$$\begin{aligned} F(t) - F(t_{q_2}) - \frac{t - t_{q_2}}{q_2} \;&\le\; F(t_{q_2} + m^{-2/9}/2) - F(t_{q_2}) - \frac{m^{-2/9}}{2q_2} \\ &\le\; \frac{f'(t_q)}{16} \cdot m^{-4/9} \end{aligned}$$

Then, since $\tau_{\hat{q}} \leq \tau_{q_2}$ a.s. on A_ε, we have

$$\begin{aligned}
\mathbb{P}\left\{\tau_{\hat{q}} > t_q + m^{-2/9},\ A_\varepsilon\right\} &\leq \mathbb{P}\left\{\tau_{q_2} > t_{q_2} + m^{-2/9}/2\right\} \\
&\leq \mathbb{P}\left\{\sup_{t \geq t_{q_2} + m^{-2/9}/2} F_m(t) - F_m(t_{q_2}) - \frac{t - t_{q_2}}{q_2} \geq 0\right\} \\
&\leq \mathbb{P}\left\{\sup_{t \geq t_{q_2} + m^{-2/9}/2} F_m(t) - F(t) - (F_m(t_{q_2}) - F(t_{q_2})) \geq \frac{|f'(t_q)|}{16} \cdot m^{-4/9}\right\} \\
&\leq \mathbb{P}\left\{\sup_{t \in [0,1]} |F_m(t) - F(t)| \geq \frac{|f'(t_q)|}{32} \cdot m^{-4/9}\right\} \\
&\leq C_{\mathrm{DKW}} \exp\left\{-\frac{f'(t_q)^2}{512} \cdot m^{1/9}\right\},
\end{aligned}$$

where C_{DKW} is the constant for the Dvoretzky–Kiefer–Wolfowitz inequality. An analogous argument yields the same bound for $\mathbb{P}\{\tau_{\hat{q}} \leq t_q - m^{-2/9}\}$. □

Proof of Theorem 5.7. Define $q = \alpha/\pi_0$ and $\hat{q} = \alpha/\hat{\pi}_0$, and let $\Delta \subseteq \{1, \ldots, m\}$ denote the symmetric difference between the two rejection sets:

$$\Delta = \begin{cases} \{R_{\hat{q}} + 1, \ldots, R^*\} & \text{if } R_{\hat{q}} < R^* \\ \{R^* + 1, \ldots, R_{\hat{q}}\} & \text{if } R_{\hat{q}} > R^* \\ \emptyset & \text{if } R_{\hat{q}} = R^* \end{cases}.$$

Then we have

$$L_\lambda(H, \mathcal{R}_{\hat{q}}) - L_\lambda(H, \mathcal{R}^*) = \frac{1}{m}\left(R^* - R_{\hat{q}} + \frac{\operatorname{sgn}(R_{\hat{q}} - R^*)}{\alpha} \sum_{i \in \Delta} (1 - H_i)\right).$$

Conditional on F_m, we have $H_i \overset{\text{ind}}{\sim} \mathrm{Bern}(1 - \mathrm{lfdr}(p_{(i)}))$, giving conditional expectation

$$\begin{aligned}
\Gamma_m &\coloneqq \mathbb{E}\left[L_\lambda(H, \mathcal{R}_{\hat{q}}) - L_\lambda(H, \mathcal{R}^*) \mid F_m\right] \\
&= \frac{1}{m}\left(R^* - R_{\hat{q}} + \frac{\operatorname{sgn}(R_{\hat{q}} - R^*)}{\alpha} \sum_{i \in \Delta} \mathrm{lfdr}(p_{(i)})\right) \\
&= \int_{\tau_{\hat{q}}}^{\tau^*} \left(1 - \alpha^{-1}\mathrm{lfdr}(t)\right) dF_m(t) \\
&= \rho(\tau_{\hat{q}}) + \alpha^{-1} \int_{\tau_{\hat{q}}}^{\tau^*} (\alpha - \mathrm{lfdr}(u))\,(dF_m(u) - dF(u))
\end{aligned}$$

Define the same truncation event A_ε as in (5.28).

$$A_\varepsilon = \left\{ |\hat{q} - q| \le m^{-1/3}\varepsilon,\ |\tau_{\hat{q}} - \tau^*| \le m^{-2/9} \right\}.$$

Then, because $|\Gamma_m| \le \alpha^{-1}$ we have

$$\begin{aligned} &\left| \text{Regret}_m(\mathcal{R}_{\hat{q}}) - \mathbb{E}\left[\rho(\tau_{\hat{q}}) 1_{A_\varepsilon}\right] \right| \\ &\qquad \le\ \alpha^{-1}\mathbb{E}\left[\left| \int_{\tau_{\hat{q}}}^{\tau^*} (\alpha - \text{lfdr}(u))\left(dF_m(u) - dF(u)\right) \right| 1_{A_\varepsilon} \right] + \alpha^{-1}\mathbb{P}\left(A_\varepsilon^c\right). \end{aligned} \tag{5.29}$$

We showed in the proof of Theorem 5.6 that $\mathbb{P}\left(A_\varepsilon^c\right) = o\left(m^{-2/3}\right)$. Furthermore,

$$\begin{aligned} m^{2/3}\mathbb{E}\left[\rho(\tau_{\hat{q}}) 1_{A_\varepsilon}\right] \ &=\ \mathbb{E}\left[\frac{f'(\xi_{\tau_{\hat{q}}})}{2} \cdot m^{2/3}(\tau_{\hat{q}} - \tau^*)^2 \cdot 1_{A_\varepsilon} \right] \\ &\to\ \frac{f'(\tau^*)}{2}\left(\frac{\alpha}{4\pi_0} \cdot f'(\tau^*)^2 \right)^{-2/3} \text{Var}(Z) \\ &=\ \left(\frac{\alpha^2}{2\pi_0^2} \cdot |f'(\tau^*)| \right)^{-1/3} \text{Var}(Z), \end{aligned}$$

where we have used the fact that $f'(\xi_{\tau_{\hat{q}}})$ is uniformly close to $f'(\tau^*)$ on A_ε. □

Proof of Proposition 5.8. Since $H_i = 0$ for all i

$$L_\lambda(H, \mathcal{R}_\alpha) - L_\lambda(H, \mathcal{R}_\alpha^{\text{OPT}}) = \frac{\lambda R_\alpha}{m}.$$

Recall R_α is the argmax of the random walk $k \mapsto \alpha\frac{k}{m} - p_{(k)}$, which has exchangeable increments. We will use Corollary 11.14 of Kallenberg (2006), due to Sparre-Andersen, that, by exchangeability, the number of rejections R_α is equal in distribution to the time the walk stays positive:

$$R_\alpha \overset{d}{=} P_\alpha \coloneqq \sum_{k=1}^{m} 1\left\{ p_{(k)} \le \alpha\frac{k}{m} \right\}.$$

Under the global null, the regret thus has mean

$$\begin{aligned} m\mathbb{E}\left[L_\lambda(H, \mathcal{R}_\alpha) - L_\lambda(H, \mathcal{R}_\alpha^{\text{OPT}})\right] = \lambda\mathbb{E}R_\alpha &= \lambda\sum_{k=1}^{m}\mathbb{P}\left\{ p_{(k)} \le \alpha\frac{k}{m} \right\} \\ &\to \lambda\sum_{k=1}^{\infty}\mathbb{P}_{U_k \sim \text{Gamma}(k,k)}\left\{U_k \le \alpha\right\}, \end{aligned}$$

where the last step follows from the law of rare events. □

Bibliography

Abolfathi, B. et al. (2018). The fourteenth data release of the Sloan Digital Sky Survey: First spectroscopic data from the extended Baryon Oscillation Spectroscopic Survey and from the second phase of the Apache Point Observatory Galactic Evolution Experiment. *The Astrophysical Journal Supplement Series*, *235*(2), 42.

Abramson, J., Pitman, J., Ross, N., & Bravo, G. U. (2011). Convex minorants of random walks and Lévy processes. *Electronic Communications in Probability*, *16*, 423–434.

Agrawal, R., Roy, U., & Uhler, C. (2019). Covariance matrix estimation under total positivity for portfolio selection. *arXiv preprint arXiv:1909.04222*.

Akritas, M. G., & Bershady, M. A. (1996). Linear regression for astronomical data with measurement errors and intrinsic scatter. *The Astrophysical Journal*, *470*(2), 706.

Amelunxen, D., Lotz, M., McCoy, M. B., & Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, *3*(3), 224–294.

Améndola, C., Engström, A., & Haase, C. (2020). Maximum number of modes of Gaussian mixtures. *Information and Inference: A Journal of the IMA*, *9*(3), 587–600.

Anderson, L., Hogg, D. W., Leistedt, B., Price-Whelan, A. M., & Bovy, J. (2018). Improving Gaia parallax precision with a data-driven model of stars. *The Astronomical Journal*, *156*(4), 145.

Angrist, J., & Imbens, G. (1994). Identification and estimation of local average treatment effects.

Aubert, J., Bar-Hen, A., Daudin, J.-J., & Robin, S. (2004). Determination of the differentially expressed genes in microarray experiments using local fdr. *BMC bioinformatics*, *5*(1), 1–9.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, 641–647.

Banerjee, O., Ghaoui, L. E., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine learning research*, *9*(Mar), 485–516.

Banerjee, T., Fu, L. J., James, G. M., & Sun, W. (2021). Nonparametric empirical Bayes estimation on heterogeneous data. http://faculty.marshall.usc.edu/gareth-james/Research/Nest%20Biometrika.pdf

Barber, R. F., & Samworth, R. J. (2021). Local continuity of log-concave projection, with applications to estimation under model misspecification. *Bernoulli*, *27*(4), 2437–2472.

Bellec, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, *46*(2), 745–780.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Berman, A., & Plemmons, R. J. (1994). *Nonnegative matrices in the mathematical sciences*. SIAM.

Böhning, D. (1985). Numerical estimation of a probability measure. *Journal of statistical planning and inference*, *11*(1), 57–69.

Böhning, D. (2003). The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. *Statistics and Computing*, *13*(3), 257–265.

Bølviken, E. (1982). Probability inequalities for the multivariate normal with non-negative partial correlations. *Scandinavian Journal of Statistics*, 49–58.

Boman, E. G., Chen, D., Parekh, O., & Toledo, S. (2005). On factor width and symmetric H-matrices. *Linear Algebra and its Applications*, *405*, 239–248.

Bonakdarpour, M., Chatterjee, S., Barber, R. F., & Lafferty, J. Prediction rule reshaping. In: *International conference on machine learning*. PMLR. 2018, 630–638.

Bovy, J., Hogg, D. W., & Roweis, S. T. (2011). Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. *Annals of Applied Statistics*, *5*(2B), 1657–1677.

Brown, L. D., Greenshtein, E., & Ritov, Y. (2013). The Poisson compound decision problem revisited. *Journal of the American Statistical Association*, *108*(502), 741–749.

Brunk, H. D., Barlow, R. E., Bartholomew, D. J., & Bremner, J. M. (1972). *Statistical inference under order restrictions; the theory and application of isotonic regression*. Wiley, New-York.

Cai, T. T., Liu, W., & Zhou, H. H. (2016a). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Annals of Statistics*, *44*(2), 455–488.

Cai, T. T., & Low, M. G. (2015). A framework for estimation of convex functions. *Statistica Sinica*, 423–456.

Cai, T. T., Low, M. G., & Xia, Y. (2013). Adaptive confidence intervals for regression functions under shape constraints. *The Annals of Statistics*, *41*(2), 722–750.

Cai, T. T., Ren, Z., & Zhou, H. H. (2016b). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, *10*(1), 1–59.

Cai, T. T., & Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics*, *40*(5), 2389–2420.

Cai, T., Liu, W., & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, *106*(494), 594–607.

Carolan, C., & Dykstra, R. (1999). Asymptotic behavior of the Grenander estimator at density flat regions. *Canadian Journal of Statistics*, *27*(3), 557–566.

Carolan, C., & Dykstra, R. (2001). Marginal densities of the least concave majorant of Brownian motion. *Ann. Statist.*, *29*(6), 1732–1750. https://doi.org/10.1214/aos/1015345960

Chade, H., Lewis, G., & Smith, L. (2014). Student portfolios and the college admissions problem. *Review of Economic Studies*, *81*(3), 971–1002.

Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *Annals of Statistics*, *42*(6), 2340–2381.

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, *16*(1), 31–41.

Colangelo, A., Scarsini, M., & Shaked, M. (2005). Some notions of multivariate positive dependence. *Insurance: Mathematics and Economics*, *37*(1), 13–26.

Coxeter, H. S. M., & Moser, W. O. J. (2013). *Generators and relations for discrete groups* (Vol. 14). Springer Science & Business Media.

Cule, M., Samworth, R., & Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(5), 545–607.

Dassios, A. (2005). On the quantiles of Brownian motion and their hitting times. *Bernoulli*, *11*(1), 29–36.

Deb, N., Saha, S., Guntuboyina, A., & Sen, B. (2021). Two-component mixture model in the presence of covariates. *Journal of the American Statistical Association*, 1–15.

Dedecker, J., & Michel, B. (2013). Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension. *Journal of Multivariate Analysis*, *122*, 278–291.

Delaigle, A., & Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli*, *14*(2), 562–579.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Denuit, M., Dhaene, J., Goovaerts, M., & Kaas, R. (2006). *Actuarial theory for dependent risks: Measures, orders and models*. John Wiley & Sons.

Dey, D. K., & Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Annals of Statistics*, *13*(4), 1581–1591.

Dhillon, I. S., & Tropp, J. A. (2008). Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, *29*(4), 1120–1146.

Dicker, L. H., & Zhao, S. D. (2016). High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference. *Biometrika*, *103*(1), 21–34.

Donoho, D., Gavish, M., & Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of Statistics*, *46*(4), 1742–1778. https://doi.org/10.1214/17-AOS1601

Doss, N., Wu, Y., Yang, P., & Zhou, H. H. (2020). Optimal estimation of high-dimensional Gaussian mixtures. *arXiv preprint arXiv:2002.05818*.

Dümbgen, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli*, *9*(3), 423–449.

Dümbgen, L., Samworth, R., & Schuhmacher, D. (2011). Approximation by log-concave distributions, with applications to regression. *The Annals of Statistics*, 702–730.

Dümbgen, L., Wellner, J. A., & Wolff, M. (2016). A law of the iterated logarithm for Grenander's estimator. *Stochastic processes and their applications*, *126*(12), 3854–3864.

Dykstra, R., & Carolan, C. (1999). The distribution of the argmax of two-sided Brownian motion with quadratic drift. *Journal of Statistical Computation and Simulation*, *63*(1), 47–58.

Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, *78*(384), 837–842.

Dyson, F. (1926). A method for correcting series of parallax observations. *Monthly Notices of the Royal Astronomical Society*, *86*, 686.

Dytso, A., Yagli, S., Poor, H. V., & Shitz, S. S. (2019). The capacity achieving distribution for the amplitude constrained additive Gaussian channel: An upper bound on the number of mass points. *IEEE Transactions on Information Theory*, *66*(4), 2006–2022.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, *99*(465), 96–104.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical science*, 1–22.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, *106*(496), 1602–1614.

Efron, B. (2012). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction* (Vol. 1). Cambridge University Press.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science*, *29*(2), 285.

Efron, B. (2019). Bayes, oracle Bayes and empirical Bayes. *Statistical Science*, *34*(2), 177–201.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference* (Vol. 5). Cambridge University Press.

Efron, B., & Morris, C. (1972a). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika*, *59*(2), 335–347.

Efron, B., & Morris, C. (1972b). Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. *Journal of the American Statistical Association*, *67*(337), 130–139.

Efron, B., & Morris, C. (1973a). Combining possibly related estimation problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *35*(3), 379–402.

Efron, B., & Morris, C. (1973b). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, *68*(341), 117–130.

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association*, *96*(456), 1151–1160.

Egilmez, H. E., Pavez, E., & Ortega, A. (2017). Graph learning from data under Laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, *11*(6), 825–841.

Fang, B., & Guntuboyina, A. (2017). On the risk of convex-constrained least squares estimators under misspecification. *arXiv preprint arXiv:1706.04276*.

Feng, L., & Dicker, L. H. (2018). Approximate nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions. *Computational Statistics & Data Analysis*, *122*, 80–91.

Finner, H., & Roters, M. (2001). On the false discovery rate and expected type I errors. *Biometrical Journal*, *43*(8), 985–1005.

Folland, G. B. (1999). *Real analysis: Modern techniques and their applications* (Vol. 40). John Wiley & Sons.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Geman, S. (1980). A limit theorem for the norm of random matrices. *Annals of Probability*, *8*, 252–261.

Genovese, C., & Wasserman, L. (2004). A stochastic process approach to false discovery control. *The annals of statistics*, *32*(3), 1035–1061.

Grenander, U. (1956). On the theory of mortality measurement: Part II. *Scandinavian Actuarial Journal*, *1956*(2), 125–153.

Groeneboom, P., Jongbloed, G., & Wellner, J. A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *The Annals of Statistics*, *29*(6), 1653–1698.

Groeneboom, P., & Jongbloed, G. (2014). *Nonparametric estimation under shape constraints* (Vol. 38). Cambridge University Press.

Groeneboom, P., & Wellner, J. A. (2001). Computing Chernoff's distribution. *Journal of Computational and Graphical Statistics*, *10*(2), 388–400.

Grotzinger, S. J., & Witzgall, C. (1984). Projections onto order simplexes. *Applied mathematics and Optimization*, *12*(1), 247–270.

Gu, J., & Koenker, R. (2016). On a problem of Robbins. *International Statistical Review*, *84*(2), 224–244.

Guntuboyina, A., & Sen, B. (2013). Covering numbers for convex functions. *IEEE Transactions on Information Theory*, *59*(4), 1957–1965.

Guntuboyina, A., & Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science*, *33*(4), 568–594.

Han, Q., Wang, T., Chatterjee, S., & Samworth, R. J. (2019). Isotonic regression in general dimensions. *The Annals of Statistics*, *47*(5), 2440–2471.

Heinrich, P., & Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, *46*(6A), 2844–2870.

Ho, N., & Nguyen, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, *10*(1), 271–307.

Hogg, D. W., Myers, A. D., & Bovy, J. (2010). Inferring the eccentricity distribution. *The Astrophysical Journal*, *725*(2), 2166.

James, W., & Stein, C. Estimation with quadratic loss. In: *Proceedings of the fourth berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics*. Berkeley, Calif.: University of California Press, 1961, 361–379. https://projecteuclid.org/euclid.bsmsp/1200512173

Jankova, J., & Van De Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, *9*(1), 1205–1229.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453–461.

Jiang, J., Nguyen, T., & Rao, J. S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, *106*(494), 732–745.

Jiang, W. (2020). On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means. *Electronic Journal of Statistics*, *14*(1), 2272–2297.

Jiang, W., & Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, *37*(4), 1647–1684.

Jiang, W., & Zhang, C.-H. Empirical Bayes in-season prediction of baseball batting averages. In: *Borrowing strength: Theory powering applications–a Festschrift for Lawrence D. Brown*. Institute of Mathematical Statistics, 2010, pp. 263–273.

Johnstone, I. M. (2019). Gaussian estimation: Sequence and wavelet models. http://www-stat.stanford.edu/~imj/.

Kallenberg, O. (2006). *Foundations of modern probability*. Springer Science & Business Media.

Karlin, S., & Rinott, Y. (1983). M-matrices as covariance matrices of multinormal distributions. *Linear Algebra and its Applications*, *52*, 419–438.

Kelly, B. C. Measurement error models in astronomy. In: *Statistical challenges in modern astronomy v*. Springer, 2012, pp. 147–162.

Kiefer, J, & Wolfowitz, J. (1976). Asymptotically minimax estimation of concave and convex distribution functions. *34*, 73–85.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

Kim, A. K. H. (2020). Obtaining minimax lower bounds: A review. *Journal of the Korean Statistical Society*, 1–29.

Kim, A. K., & Guntuboyina, A. (2020). Minimax bounds for estimating multivariate Gaussian location mixtures. *arXiv preprint arXiv:2012.00444*.

Kim, A. K., & Samworth, R. J. (2016). Global rates of convergence in log-concave density estimation. *The Annals of Statistics*, *44*(6), 2756–2779.

Kim, Y., Carbonetto, P., Stephens, M., & Anitescu, M. (2020). A fast algorithm for maximum likelihood estimation of mixture proportions using sequential quadratic programming. *Journal of Computational and Graphical Statistics*, *29*(2), 261–273.

Knight, F. B. (1996). The uniform law for exchangeable and Lévy process bridges. *Astérisque*, (236), 171–188.

Koenker, R., & Gu, J. (2017). REBayes: Empirical Bayes mixture methods in R. *Journal of Statistical Software*, *82*(8), 1–26.

Koenker, R., & Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association*, *109*(506), 674–685.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*(364), 805–811.

Lake, B., & Tenenbaum, J. (2010). Discovering structure by learning sparse graphs. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 778–783.

Langaas, M., Lindqvist, B. H., & Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(4), 555–572.

Lashkari, D., & Golland, P. Convex clustering with exemplar-based models. In: *Advances in neural information processing systems*. 2008, 825–832.

Laurent, B., & Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 1302–1338.

Lauritzen, S., Uhler, C., & Zwiernik, P. (2019). Maximum likelihood estimation in Gaussian models under total positivity. *Annals of Statistics*, *47*(4), 1835–1863.

Ledoit, O., & Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli*, *24*(4B), 3791–3832.

Lesperance, M. L., & Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association*, *87*(417), 120–126.

Liao, J., Lin, Y., Selvanayagam, Z. E., & Shih, W. J. (2004). A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics*, *20*(16), 2694–2701.

Lindsay, B. G. Mixture models: Theory, geometry and applications. In: *Nsf-cbms regional conference series in probability and statistics*. JSTOR. 1995.

Lindsay, B. G., & Roeder, K. (1993). Uniqueness of estimation and identifiability in mixture models. *Canadian Journal of Statistics*, *21*(2), 139–147.

Liu, L., & Zhu, Y. (2007). Partially projected gradient algorithms for computing nonparametric maximum likelihood estimates of mixing distributions. *Journal of Statistical Planning and Inference*, *137*(7), 2509–2522.

Majewski, S. R. et al. (2017). The Apache Point Observatory Galactic Evolution Experiment (APOGEE). *The Astronomical Journal*, *154*(3), 94.

Marriott, P. (2002). On the local geometry of mixture models. *Biometrika*, *89*(1), 77–93.

Mazumder, R., Choudhury, A., Iyengar, G., & Sen, B. (2019). A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, *114*(525), 318–331.

Mazumder, R., & Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, *6*, 2125.

Meister, A. (2009). *Deconvolution problems in nonparametric statistics* (Vol. 193). Springer Science & Business Media.

Meyer, M., & Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *The annals of Statistics*, *28*(4), 1083–1104.

Milgrom, P. R., & Weber, R. J. (1982). A theory of auctions and competitive bidding. *Econometrica*, *50*, 1089–1122.

Moakher, M., & Batchelor, P. G. Symmetric positive-definite matrices: From geometry to applications and visualization. In: *Visualization and processing of tensor fields*. Springer, 2006, pp. 285–298.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American statistical Association*, *78*(381), 47–55.

MOSEK ApS. (2019). Mosek optimization suite. http://docs.mosek.com/9.0/intro.pdf

Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, 422–438.

Németh, A. B., & Németh, S. Z. (2012). How to project onto the monotone nonnegative cone using pool adjacent violators type algorithms. *arXiv preprint arXiv:1201.2343*.

Neuvial, P., & Roquain, E. (2012). On false discovery rate thresholding for classification under sparsity. *The Annals of Statistics*, *40*(5), 2572–2600.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, *41*(1), 370–400.

Obozinski, G., Lanckriet, G., Grant, C., Jordan, M. I., & Noble, W. S. (2008). Consistent probabilistic outputs for protein function prediction. *Genome Biology*, *9*(1), S6.

Oymak, S., & Hassibi, B. (2016). Sharp MSE bounds for proximal denoising. *Foundations of Computational Mathematics*, *16*(4), 965–1029.

Patra, R. K., & Sen, B. (2016). Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(4), 869–893.

Pavez, E., Egilmez, H. E., & Ortega, A. (2018). Learning graphs with monotone topology properties and multiple connected components. *IEEE Transactions on Signal Processing*, *66*(9), 2399–2413.

Pavez, E., & Ortega, A. Generalized Laplacian precision matrix estimation for graph signal processing. In: *2016 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2016, 6350–6354.

Pfanzagl, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: Mixtures. *Journal of Statistical Planning and Inference*, *19*(2), 137–158.

Plemmons, R. J. (1977). M-matrix characterizations. I—Nonsingular M-matrices. *Linear Algebra and its Applications*, *18*(2), 175–188.

Polyanskiy, Y., & Wu, Y. (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*.

Pounds, S., & Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics*, *20*(11), 1737–1745.

Pounds, S., & Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, *19*(10), 1236–1242.

Rao, B. P. (1969). Estimation of a unimodal density. *Sankhyā: The Indian Journal of Statistics, Series A*, 23–36.

Ratcliffe, B. L., Ness, M. K., Johnston, K. V., & Sen, B. (2020). Tracing the assembly of the Milky Way's disk through abundance clustering. *The Astrophysical Journal*, *900*(2), 165.

Ravikumar, P., Wainwright, M. J., Raskutti, G., & Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, *5*, 935–980.

Ray, S., & Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *The Annals of Statistics*, *33*(5), 2042–2065.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, *26*(2), 195–239.

Reiner, A., Yekutieli, D., & Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, *19*(3), 368–375.

Rigollet, P., & Weed, J. (2018). Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, *356*(11-12), 1228–1235.

Ritchie, J. A., & Murray, I. (2019). Scalable extreme deconvolution. *arXiv preprint arXiv:1911.11663*.

Robbins, H. A generalization of the method of maximum likelihood-estimating a mixing distribution. In: *Annals of mathematical statistics. 21.* (2). 1950, 314–315.

Robbins, H. Asymptotically subminimax solutions of compound statistical decision problems. In: *Proceedings of the second berkeley symposium on mathematical statistics and probability*. The Regents of the University of California. 1951.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 157–163.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference*. John Wiley & Sons Ltd.

Robin, S., Bar-Hen, A., Daudin, J.-J., & Pierre, L. (2007). A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational statistics & data analysis*, *51*(12), 5483–5493.

Rothman, A. J., Bickel, P. J., Levina, E., & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, *2*, 494–515.

Saha, S., & Guntuboyina, A. (2020a). On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *Annals of Statistics*, *48*(2), 738–762.

Saha, S., & Guntuboyina, A. (2020b). Supplement to "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising". https://doi.org/10.1214/19-AOS1817SUPP

Scheid, S., & Spang, R. (2004). A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *1*(3), 98–108.

Schweder, T., & Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, *69*(3), 493–502.

Shorack, G. R., & Wellner, J. A. (2009). *Empirical processes with applications to statistics*. SIAM.

Slawski, M., & Hein, M. (2013). Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, *7*, 3004–3056.

Slawski, M., & Hein, M. (2015). Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, *473*, 145–179.

Soloff, J. A., Guntuboyina, A., & Jordan, M. I. (2020). Covariance estimation with nonnegative partial correlations. *arXiv preprint arXiv:2007.15252*.

Soloff, J. A., Guntuboyina, A., & Pitman, J. (2019). Distribution-free properties of isotonic regression. *Electronic Journal of Statistics*, *13*(2), 3243–3253.

Soloff, J. A., Guntuboyina, A., & Sen, B. (2021). Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood. *arXiv preprint arXiv:2109.03466*.

Soloff, J. A., Xiang, D., & Fithian, W. (2022). The edge of discovery: Controlling the local false discovery rate at the margin.

Sparre-Andersen, E. (1954). On the fluctuations of sums of random variables II. *Mathematica Scandinavica*, 195–223.

Spitzer, F., & Widom, H. (1961). The circumference of a convex polygon. *Proceedings of the American Mathematical Society*, *12*(3), 506–509.

Stein, C. Estimation of a covariance matrix. In: *39th annual meeting ims, Atlanta, GA*. 1975.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, *34*(1), 1373–1403.

Stephens, M. (2017). False discovery rates: A new deal. *Biostatistics*, *18*(2), 275–294.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 479–498.

Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(1), 187–205.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC bioinformatics*, *9*(1), 1–14.

Sun, T., & Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, *14*(1), 3385–3418.

Sun, W., & Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, *102*(479), 901–912.

Takács, L. On combinatorial methods in the theory of stochastic processes (L. M. Le Cam & J. Neyman, Eds.). In: In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (L. M. Le Cam & J. Neyman, Eds.). Ed. by Le Cam, L. M., & Neyman, J. *3*. University of California Press, 1967, 431–447.

Tan, Z. (2016). Steinized empirical Bayes estimation for heteroscedastic data. *Statistica Sinica*, 1219–1248.

van de Geer, S. (2000). *Empirical processes in M-estimation* (Vol. 6). Cambridge university press.

van Eeden, C. (1956). Maximum likelihood estimation of ordered probabilities, 2. *Stichting Mathematisch Centrum. Statistische Afdeling*, (S 196/56).

Villani, C. (2008). *Optimal transport: Old and new* (Vol. 338). Springer Science & Business Media.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

Wang, Y. (2007). On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 185–198.

Wang, Y., Roy, U., & Uhler, C. (2019). Learning high-dimensional Gaussian graphical models under total positivity without tuning parameters. *arXiv preprint arXiv:1906.05159*.

Wei, Y., Wainwright, M. J., & Guntuboyina, A. (2019). The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *The Annals of Statistics*, *47*(2), 994–1024.

Weinstein, A., Ma, Z., Brown, L. D., & Zhang, C.-H. (2018). Group-linear empirical Bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association*, *113*(522), 698–710.

Wong, W. H., & Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, *23*(2), 339–362.

Wu, Y., & Yang, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. *The Annals of Statistics*, *48*(4), 1981–2007.

Xie, X., Kou, S., & Brown, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, *107*(500), 1465–1479.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, *11*, 2261–2286.

Zarantonello, E. H. Projections on convex sets in Hilbert space and spectral theory: Part i. projections on convex sets: Part ii. spectral theory. In: *Contributions to nonlinear functional analysis*. 1971, pp. 237–424.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *The Annals of Statistics*, *30*(2), 528–555.

Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, 1297–1318.