Bayesian Linear Modeling in High Dimensions: Advances in Hierarchical Modeling, Inference, and Evaluation

by

Brian L. Trippe

B.A., Columbia University (2016)
MPhil., University of Cambridge (2017)

Submitted to the Computational and Systems Biology Program
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Computational and Systems Biology Program
May 18, 2022

Certified by .....
Tamara Broderick
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Christopher B. Burge
Director, Computational and Systems Biology program

Bayesian Linear Modeling in High Dimensions: Advances in Hierarchical Modeling, Inference, and Evaluation

by

Brian L. Trippe

Submitted to the Computational and Systems Biology Program
on May 18, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Across the sciences, social sciences and engineering, applied statisticians seek to build understandings of complex relationships from increasingly large datasets. In statistical genetics, for example, we observe up to millions of genetic variations in each of thousands of individuals, and wish to associate these variations with the development of disease. For 'high dimensional' problems like this, the languages of linear modeling and Bayesian statistics appeal because they provide interpretability, coherent uncertainty, and the capacity for information sharing across related datasets. But at the same time, high dimensionality introduces several challenges not solved by existing methodology.

This thesis addresses three challenges that arise when applying the Bayesian methodology in high dimensions. A first challenge is how to apply hierarchical modeling, a mainstay of Bayesian inference, to share information between multiple linear models with many covariates (for example, genetic studies of multiple related diseases). The first part of the thesis demonstrates that the default approach to hierarchical linear modeling fails in high dimensions, and presents a new, effective model for this regime. The second part of the thesis addresses the computational challenge presented by Bayesian inference in high dimensions — existing methods demand time that scales super-linearly with the number of covariates. We present two algorithms that permit fast, accurate inferences by leveraging (i) low rank approximations of data or (ii) parallelism across a certain class of Markov chain Monte Carlo algorithms. The final part of the thesis addresses the challenge of evaluation. Modern statistics provides an expansive toolkit for estimating unknown parameters, and a typical Bayesian analysis justifies its estimates through belief in subjective *a priori* assumptions. We address this by introducing a measure of confidence in the new estimate (the 'c-value'), that can diagnose the accuracy of a Bayesian estimate without requiring this subjectivism.

Thesis Supervisor: Tamara Broderick
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

I would first like to thank my advisor, Tamara, for her guidance, support and feedback over the course of my PhD. Tamara encouraged me to pursue the questions that interested me and gave me the freedom to read and learn broadly, but was always there when I needed her to help me narrow in on the more important and practical research directions. As a communicator she has been a source of inspiration and will always be someone to whom I will look up when it comes to crafting technical talks and papers.

Next I would like to thank my committee members, Hilary Finucane, Youssef Marzouk, and Jeff Miller. Hilary has been an amazing mentor and collaborator. I am grateful that she welcomed me into her group and into the world of statistical genetics in the summer of 2019.

My coauthors Jonathan Huggins, Raj Agrawal, Sameer Deshpande, Tin (Stan) Nguyen have all been a pleasure to work with. I learned so much from them and this thesis would not exist without their contributions. Jonathan, especially, was an invaluable mentor in my first year, as was Sameer in 2020 – without him I certainly would have gone mad in that first pandemic year. I am thankful as well to my other collaborators with whom I have worked on projects not in this thesis, Buwei Huang, Erika DeBenedictis, Brian Coventry, Nicholas Bhattacharya, Kevin Yang, David Baker, and Lorin Crawford at Microsoft Research and the University Washington, as well as collaborators Nathan Cheng, Elle Weeks and Jacob Ulirsch at the Broad Institute. And in my last semester it has been exciting to engage with new collaborators Renato Berlinghieri, Jason Yim and Doug Tischer.

I thank my other friends and collaborators in Tamara's group: Trevor, Hannah, Will, Miriam, Ryan, Diana, Yaroslav, Mikołaj, and particularly Lorenzo with whom I found so many shared interests beyond statistics. I am grateful as well to those who mentored me before graduate school, and nurtured my curiosity and interest in science: Naomi Pierce, Marty Chalfie, Chaogu Zheng, James Leighton, Harmen Bussemaker, Mate Lengyel and Richard Turner.

I have been fortunate to have the encouragement of my family, who have supported me and helped me remember that a world exists beyond the university. Lastly, I would like to thank my partner and best friend, Sarah, whose companionship has tempered the lows and accentuated the highs of these years at MIT.

Funding information: The work in this thesis was funded in part by NSF GRFP, an NSF CAREER Award, an ARO YIP Award, a Google Faculty Research Award, a Sloan Research Fellowship, the ONR, and an ARPA-E project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Scientists, social scientists and engineers often seek to understand how an outcome of interest relates to a set of covariates. For example, a geneticist may wish to understand the effects of natural genetic variation on the presence of disease, or a medical practitioner may wish to understand the effect of a patient's history on their future health. In countless settings like this, the ease of modern data collection often yields large sets of covariates for data analysts to parse. While these data should ultimately aid understanding, this "high-dimensionality" adds complication. Because they offer simplicity and interpretability, linear models are extremely widely used across the sciences and social sciences. Unfortunately, when (as in genetics) the number of data points is not substantially larger than the number of covariates, non-trivial inferential uncertainty persists.

Bayesian statistical approaches naturally confront the challenge of inferential uncertainty in high-dimensional linear models in principle, by offering coherent uncertainty and the ability to incorporate expert information and share power across datasets. But realizing these advantages in practice requires methodological choices at three stages an analysis: (1) modeling, (2) inference and (3) evaluation. While effective statistical methods developed for low dimensional problems abound, high-dimensionality introduces challenges across each of these stages not addressed by the established toolkit.

In this thesis, we characterize methodological gaps in the Bayesian toolkit across

modeling, inference and evaluation, and develop new approaches to address them. For modeling, in Chapter 2 we describe how the dominant modeling paradigm for sharing information across related datasets fails for high-dimensional linear models, and introduce a complementary approach suited to high-dimensions. For inference, we argue that existing algorithms are either computationally expensive or inaccurate, and in Chapters 3 and 4 we introduce new ones that are provably fast and accurate. For evaluation, because many complex models do not provide more accurate inferences than simpler baseline approaches, in Chapter 5 we introduce computable criteria for validating the relative accuracy of the outcomes of sophisticated analyses. To ensure that these methods are accurate and reliable, we develop theoretical guarantees to guide their usage.

The contribution of Chapter 2 was completed in collaboration with Hilary Finucane and Tamara Broderick [Trippe et al., 2021b]. The contribution of Chapter 3 was completed in collaboration with Jonathan Huggins, Raj Agrawal, and Tamara Broderick [Trippe et al., 2019]. The contribution of Chapter 4 was completed in equal collaboration with Tin (Stan) Nguyen and is as much his as is mine, as well as with Tamara Broderick [Trippe et al., 2021c]. The contribution of Chapter 5 was completed in collaboration with Sameer Deshpande and Tamara Broderick [Trippe et al., 2021a]. The remainder of this chapter summarizes these contributions.

1.1 New Bayesian models for high-dimensional hierarchical regression

Hierarchical modeling is a mainstay of Bayesian inference that enables information sharing across regression problems on multiple groups of data. Often covariates are shared across multiple related groups, but the effects are typically allowed to vary both by group and by covariate. While standard practice is to model regression parameters (effects) as (1) exchangeable across the groups and (2) correlated to differing degrees across covariates, in Chapter 2, we show that this approach exhibits

poor statistical performance when the number of covariates exceeds the number of groups. For instance, in statistical genetics, one might regress dozens of traits (defining groups) for thousands of individuals (responses) on up to millions of genetic variants (covariates). We argue that when an analyst has more covariates than groups it preferable to instead model effects as (1) exchangeable across *covariates* and (2) correlated to differing degrees across *groups*. To this end, we propose a hierarchical model expressing this alternative perspective. We develop theory that demonstrates that this model produces estimates that are more accurate than the classic approach when the number of covariates dominates the number of groups, and corroborate this result empirically on several high-dimensional multiple regression and classification problems.

1.2 Fast inference with theoretical guarantees

In Bayesian analyses, inference is the computational and algorithmic step of combining the chosen model with observed data to obtain conclusions about unobserved parameters. The computational challenge of inference remains a barrier to wider adoption of many Bayesian methods, especially in high dimensions. A variety of inference methods, such as variational Bayes, can provide fast approximations. However, these approaches can have pathological behaviour that leads to poor accuracy [MacKay, 2003, Turner and Sahani, 2011, Trippe and Turner, 2018, Huggins et al., 2020]. In this work, we develop practical approximate inference methods with provable guarantees on runtime and approximation error.

Low rank approximation for fast inference in generalized linear models: Generalized linear models (GLMs) are widely used across the sciences and social sciences to relate covariates of interest to not only real-valued but also binary and count-valued responses. Unfortunately, existing methods for Bayesian inference in GLMs require running times roughly cubic in parameter dimension, and so are limited to settings with at most tens of thousand parameters. In Chapter 3, we propose

to reduce time and memory costs with a low-rank approximation of the data in an approach we call LR-GLM. When used with the Laplace approximation or Markov chain Monte Carlo, LR-GLM provides a full Bayesian posterior approximation and admits running times reduced by a full factor of the parameter dimension. We rigorously establish the quality of our approximation and show how the choice of rank allows a tunable computational-statistical trade-off. Experiments support our theory and demonstrate the efficacy of LR-GLM on real large-scale datasets.

Recently, [Hauzenberger et al., 2021] extended our approach in an econometrics application to perform efficient inference in time-varying parameter regressions and stochastic volatility models. By leveraging our technical advances, these authors were able to fit models of price inflation in the United States with thousands of parameters for which traditional methods are impractically slow.

Improving accuracy with embarrassingly parallel computation: Computational couplings of Markov chains provide a practical route to unbiased Monte Carlo estimation that can utilize parallel computation [Jacob et al., 2020]. However, these approaches depend crucially on chains meeting after a small number of transitions. For models that assign data into groups, e.g. mixture models, the obvious approaches to couple Gibbs samplers fail to meet quickly. In Chapter 4, we trace this failure to the so-called "label-switching" problem [Jasra et al., 2005]; semantically equivalent relabelings of the groups contribute well-separated posterior modes that impede fast mixing and cause large meeting times. We then demonstrate how to avoid label switching by considering chains as exploring the space of partitions rather than labelings. Using a metric on this space, we employ an optimal transport coupling of the Gibbs conditionals. This coupling outperforms alternative couplings that rely on labelings and, on a real dataset, provides estimates more precise than usual ergodic averages in the limited time regime.

Fast and accurate inference in hierarchical modeling: Scalable inference algorithms are also central to the contributions of Chapter 2. In particular, for our

approach to inference in the hierarchical models we consider relies on an expectation maximization algorithm, we de, which enables efficiently solving an empirical Bayes problem. In that work, we additionally document how to use the conjugate gradient method for solving up to million dimensional linear systems by taking advantage of sparsity, Kronecker structure and good initializations for provably fast convergence.

1.3 Evaluation and model selection

A third challenge in high-dimensional Bayesian analyses is model evaluation. Modern statistics provides an expansive toolkit of methods applicable to high-dimensional problems. However, this abundance often presents practitioners already familiar with one mode of analysis with a difficult challenge: choosing between the output of a familiar method and that of a more complicated method, for example, one that shares information across related datasets. In Chapter 5 we introduce a statistical tool, the "c-value", for computing a measure of confidence that a new estimate is more accurate than a baseline approach. In analogy to how a small p-value provides evidence to reject a null hypothesis, a large c-value provides evidence to replace an old estimate with a new one. For a wide class of problems and estimators, we show how to compute a c-value by first constructing a data-dependent high-probability lower bound on the difference in loss. The c-value is frequentist in nature, but we show that it can provide a validation of Bayesian estimates in real data applications involving hierarchical models and Gaussian processes.

Chapter 2

For high-dimensional hierarchical models, consider exchangeability of effects across covariates instead of across datasets

Abstract

Hierarchical Bayesian methods enable information sharing across regression problems on multiple groups of data. While standard practice is to model regression parameters (effects) as (1) exchangeable across the groups and (2) correlated to differing degrees across covariates, we show that this approach exhibits poor statistical performance when the number of covariates exceeds the number of groups. For instance, in statistical genetics, we might regress dozens of traits (defining groups) for thousands of individuals (responses) on up to millions of genetic variants (covariates). When an analyst has more covariates than groups, we argue that it is often preferable to instead model effects as (1) exchangeable across *covariates* and (2) correlated to differing degrees across *groups*. To this end, we propose a hierarchical model expressing our alternative perspective. We devise an empirical Bayes estimator for learning the degree

of correlation between groups. We develop theory that demonstrates that our method outperforms the classic approach when the number of covariates dominates the number of groups, and corroborate this result empirically on several high-dimensional multiple regression and classification problems.

2.1 Introduction

Hierarchical modeling is a mainstay of Bayesian inference. For instance, in (generalized) linear models, the unknown parameters are *effects*, each of which describes the association of a particular covariate with a response of interest. Often covariates are shared across multiple related groups, but the effects are typically allowed to vary both by group and by covariate. A classic methodology, dating back to Lindley and Smith (1972) [Lindley and Smith, 1972], models the effects as conditionally independent across groups, with a latent (and learnable) degree of relatedness across covariates. From a practical standpoint, the model is motivated by the understanding that it "borrows strength" across different groups [Gelman et al., 2013, Chapter 5.6]. Mathematically, the model is motivated by assuming effects are exchangeable across groups and applying a de Finetti theorem [Lindley and Smith, 1972, Jordan, 2010]. The methodology of Lindley and Smith is ubiquitous when the number of groups is larger than the number of covariates. It is a standard component of Bayesian pedagogy [[Gelman and Hill, 2006, Chapter 13.3]; [Gelman et al., 2013, Chapter 15.4]] and software; e.g. it is used in the mixed modeling package `lme4` [Bates et al., 2015b], which has over 16 million downloads at the time of writing.

Despite its resounding success when there are more groups than covariates, we show in the present work that this standard methodology performs poorly when there are more covariates than groups. To address the many-covariates case, we turn for inspiration to statistical genetics, where scientists commonly learn linear models relating genetic variants (covariates) to traits (corresponding to different groups) across individuals (which each exhibit a response). These applications may exhibit millions of covariates, thousands of responses, and just a handful of groups. In these cases,

[Lee et al., 2012, Bulik-Sullivan et al., 2015, Stephens, 2013, Zhou and Stephens, 2014, Maier et al., 2015, Runcie et al., 2020] use a multivariate Gaussian prior akin to that of Lindley and Smith, but assume conditional independence across *covariates* and prior parameters that encode correlations across *groups*, rather than than the other way around.

As we will see, this alternative modeling approach may be motivated from a Bayesian perspective when one begins from an assumption of a priori exchangeability of the effects across covariates (rather than across groups). This exchangeability assumption is reasonable in statistical genetics, where we have little knowledge to distinguish our expectations about the effects of different genetic variants; we argue this modeling approach can be effective other modern high-dimensional analyses of multiple groups of data (beyond statistical genetics) in which large collections of covariates are frequently treated monolithically, e.g. by applying ridge regression. Namely, when there are more covariates than groups, we propose to model the effects as exchangeable across *covariates* (rather than groups) and learn the degree of relatedness of effects across *groups* (rather than covariates). In what follows, we refer to this framework as *ECov*, for *e*xchangeable effects across *cov*ariates, and distinguish it from *e*xchangeable effects across *groups* or *EGroup*.

While the existing methods in statistical genetics for modeling multiple traits obtain as a special case of ECov, to the best of our knowledge this approach is absent from existing literature on hierarchical Bayesian regression. Brown and Zidek (1980) [Brown and Zidek, 1980] and Haitovsky (1987) [Haitovsky, 1987] form two exceptions, but these two papers (1) consider only the situation in which a single covariate matrix is shared across all groups (or equivalently, for each data point all responses are observed) and (2) include only theory and no empirics. While Lindley and Smith (and others) discuss a priori exchangeability across covariates in the context of analysis of a single group, to our knowledge no other work has pushed this idea forward to share strength across multiple groups.

We suspect that the historical origins of the methodology in statistical genetics may have hindered earlier expansion of this class of models to a wider audience. In

particular, this literature traces back to mixed effects modeling for cattle breeding [Thompson, 1973]; here, an even-earlier notion of the genetic contribution of trait correlation (i.e. "genetic correlation;" see Hazel (1943) [Hazel, 1943]) informs the covariance structure of random effects. Although genetic correlation is now commonly understood to describe the correlation of effects of DNA sequence changes on different traits [Bulik-Sullivan et al., 2015], its provenance predates even the first identification of DNA as the genetic material in 1944 [Avery et al., 1944]. As such, this older motivation obviated the need for a more general justification grounded in exchangeability. See Appendix A.1 for further discussion of related work, including more recent works from within the machine learning community on sharing strength across multiple groups of data.

In the present work, we propose ECov as a general framework for hierarchical regression when the number of covariates exceeds the number of groups. We show that the classic model structure from statistical genetics can be seen as an instance of this framework, much as Lindley and Smith give a (complementary) instance of an EGroup framework. To make the ECov approach generally practical, we devise an accurate and efficient algorithm for learning the matrix of correlations between groups. We demonstrate with theory and empirics that ECov is preferred when the number of covariates exceeds the number of groups, while EGroup is preferred when the number of groups exceeds the number of covariates. Our experiments analyze three real, non-genetic groups in regression and classification, including an application to transfer learning with pre-trained neural network embeddings. We provide proofs of theoretical results in the appendix.

2.2 Exchangeability and its applications to hierarchical linear modeling

We start by establishing the data and model, motivating exchangeability among covariate effects (ECov), and motivating our Bayesian generative model.

Setup and notation. Consider Q groups with D covariates. Let N^q be the number of data points in group q. For the qth group, the $N^q \times D$ real design matrix X^q collects the covariates, and Y^q is the N^q-vector of responses. The nth datapoint in group q consists of covariate D-vector X_n^q and scalar response Y_n^q. We let $\mathcal{D} := \{(X^q, Y^q)\}_{q=1}^Q$ denote the collection of data from all Q groups. We consider the generalized linear model $Y_n^q | X_n^q, \beta^q \overset{indep}{\sim} p(\cdot | X_n^{q\top} \beta^q)$ with unknown D-vector of real effects β^q. We collect all effects in a $D \times Q$ matrix β with (d, q) entry β_d^q. The linear form of the likelihood allows interpretation of β_d^q as the association between the dth covariate and the response in group q. In linear regression, the responses are real-valued and the conditional distribution is Gaussian. In logistic regression, the responses are binary, and we use the logit link. The independence assumption conflicts with some models that one might use, for example in some cases when the different groups partially overlap.

Example. As a motivating non-genetics example, consider a study of the efficacy of microcredit. There are seven famous randomized controlled trials of microcredit, each in a different country [Meager, 2019]. We might be interested in the association between various aspects of small businesses (covariates), including whether or not they received microcredit, and their business profit (response). In this case, the dth element of X_n^q would be the dth characteristic of the nth small business in the qth country, and Y_n^q is the profit of this business. See the experiments for additional examples in rates of policing, web analytics, and transfer learning.

Exchangeable effects across groups (EGroup). To fully specify a Bayesian model, we need to choose a prior over the parameters β. Lindley and Smith assume the effects are exchangeable across groups. Namely, for every Q-permutation σ, $p(\beta^1, \beta^2, \ldots, \beta^Q) = p(\beta^{\sigma(1)}, \beta^{\sigma(2)}, \ldots, \beta^{\sigma(Q)})$. Assuming exchangeability holds for an imagined growing Q and applying de Finetti's theorem motivates a conditionally independent prior. Concretely, Lindley and Smith take $\beta^q \overset{i.i.d.}{\sim} \mathcal{N}(\xi, \Gamma)$, for D-vector ξ and $D \times D$ covariance matrix Γ. The (d, d') entry of Γ captures the degree of relatedness between the effects for covariates d and d'. Both ξ and Γ may be learned in an empirical Bayes procedure. However, when D is large relative to Q, learning

these parameters can present both computational and inferential challenges, as the $O(D^2)$ degrees of freedom in Γ outnumber the $O(DQ)$ effects.

Exchangeable effects across covariates (ECov). We here argue for a complementary approach in settings where $D > Q$. In the microcredit example, notice that $D > Q$ will arise whenever the experimenter records more characteristics of a small business than there are locations with microcredit experiments; that is, $D > 7$ in this particular case. Concretely, let β_d be the Q-vector of effects for covariate d across groups. Then, in the ECov approach, we will assume that effects are exchangeable across covariates instead of across groups. Namely, for every D-permutation σ, $p(\beta_1, \beta_2, \ldots, \beta_D) = p(\beta_{\sigma(1)}, \beta_{\sigma(2)}, \ldots, \beta_{\sigma(D)})$. We will see theoretical and empirical benefits to ECov in later sections, but note that the ECov assumption is often consistent with prior beliefs in high dimensional settings. For instance, regarding microcredit, we may have no prior knowledge about how effects differ for distinct small-business characteristics. And we may a priori believe that different countries could exhibit more similar effects – and wish to learn the degree of relatedness across those countries.

We may apply a similar rationale as Lindley and Smith to motivate a conditionally independent model. Analogous to Lindley and Smith, we propose a Gaussian prior: $\beta_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$. Σ is now a $Q \times Q$ covariance matrix whose (q, q') entry captures the similarity between the effects in the q and q' groups. For simplicity, we restrict to $\mathbb{E}[\beta_d] = 0$; see Appendix A.5.3 for discussion. Another potential benefit to ECov relative to EGroup is that we might expect a statistically easier problem, with $O(Q^2)$ rather than $O(D^2)$ values to learn in the relatedness matrix. We provide a rigorous theoretical analysis in Sections 2.4 and 2.5.

2.3 Our method

We next describe our inference method for specific instances of the exchangeable covariate effects model of Section 2.2. We compute the β posterior and take an empirical Bayes approach to estimate Σ. We find that an expectation maximization (EM) algorithm estimates Σ effectively; Appendix A.1.2 compares our approach to

existing methods for the related problem of estimating Γ for EGroup.

Notation. We identify estimates of β and Σ with hats. For instance, $\hat{\beta}_{\text{LS}}$ is the least squares estimate, with $\hat{\beta}^q_{\text{LS}} := (X^{q\top}X^q)^{-1}X^{q\top}Y^q$. We will sometimes find it useful to stack the columns of β or its estimates into a length DQ vector; we denote such vectors with an arrow; for example, $\vec{\beta} := [\beta^{1\top}, \beta^{2\top}, \ldots, \beta^{Q\top}]^\top$. For a natural number N, we use $I_N, \mathbf{1}_N$, and e_N to denote the $N \times N$ identity matrix, N-vector of ones, and Nth basis vector, respectively. We use $\otimes$ to denote the Kronecker product.

2.3.1 Posterior inference with a Gaussian likelihood

We first consider a Gaussian likelihood: for each group q and observation n, we take $Y^q_n | X^q_n, \beta^q \overset{indep}{\sim} \mathcal{N}(X^{q\top}_n \beta^q, \sigma^2_q)$ where σ^2_q is a group-specific variance. When the relatedness matrix Σ is known, a natural estimate of β is its posterior mean. We obtain the full posterior, including its mean, via a standard conjugacy argument; see Appendix A.2.1:

Proposition 2.3.1. *For each covariate d, let $\beta_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ a priori. For each group q and data point n, let $Y^q_n | X^q_n, \beta^q \overset{indep}{\sim} \mathcal{N}(X^{q\top}_n \beta^q, \sigma^2_q)$. Then $\vec{\beta} | \mathcal{D}, \Sigma \sim \mathcal{N}(\vec{\mu}, V)$ for $\vec{\mu} = V[\sigma_1^{-2} Y^{1\top} X^1, \ldots, \sigma_Q^{-2} Y^{Q\top} X^Q]^\top$ and $V^{-1} = \Sigma^{-1} \otimes I_D + \text{diag}(\sigma_1^{-2} X^{1\top} X^1, \ldots, \sigma_Q^{-2} X^{Q\top} X^Q)$, where $\text{diag}(\sigma_1^{-2} X^{1\top} X^1, \ldots, \sigma_Q^{-2} X^{Q\top} X^Q)$ denotes a $DQ \times DQ$ block-diagonal matrix.*

At first glance, the posterior mean $\vec{\mu}$ for this model might seem to introduce a computational challenge because exact computation of V involves an $O(D^3Q^3)$-time matrix inversion. Our experiments (Section 2.6), however, involve on the order of $DQ \approx 1{,}000$ parameters, so direct inversion of V demands less than a single second. Moreover, in much larger problems $\vec{\mu}$ may still be computed very efficiently using the conjugate gradient algorithm [Nocedal and Wright, 2006, Chapter 5], with convergence in a small number of $O(D^2Q)$ time iterations; see Appendix A.2.2.

Algorithm 1 Expectation Maximization for Exchangeability Among Covariate Effects

// Initialize covariance
$\Sigma^{(0)} \leftarrow I_Q$
// Run EM algorithm
for $t = 0, 1, \ldots$ **do**
    // Expectation step
    $\mu_1, \ldots, \mu_D, V_1, \ldots, V_D \leftarrow$ `E_Step`$(\Sigma^{(t)})$

    // Maximization step
    $\Sigma^{(t+1)} \leftarrow D^{-1} \sum_{d=1}^{D} (\mu_d \mu_d^\top + V_d)$

Return $\Sigma^{(t+1)}$

2.3.2 Empirical Bayes estimation of Σ by expectation maximization

The posterior mean of β in Proposition 2.3.1 requires Σ, which is typically unknown. Accordingly, we propose an empirical Bayes approach of estimating Σ by maximum marginal likelihood:

$$\hat{\beta}_{\text{ECov}} := \mathbb{E}[\beta \mid \mathcal{D}, \hat{\Sigma}] \text{ where } \hat{\Sigma} := \arg\max_{\Sigma \succeq 0} p(\mathcal{D} \mid \Sigma). \tag{2.1}$$

Eq. (2.1) defines a two step procedure. In the first step, we learn the similarity between groups via estimation of Σ. In the second step, we use this similarity to compute an estimate, $\hat{\beta}_{\text{ECov}}$, that correspondingly shares strength. Though we have been unable to identify a general analytic form for $\hat{\Sigma}$, we can compute it with an expectation maximization (EM) algorithm [McLachlan and Krishnan, 2007, Chapter 1.5]. Algorithm 1 summarizes this procedure; see Appendix A.2.3 for details.

2.3.3 Classification with logistic regression

We can extend the approach above to inference for multiple related classification problems. We assume a logistic likelihood; for each q and n, $Y_n^q | X_n^q, \beta^q \overset{indep}{\sim} \text{Bern}[(1 + \exp\{-X_n^{q\top} \beta^q\})^{-1}]$. In the classification case, we cannot use Gaussian conjugacy directly,

Algorithm 2 E-Step: Linear Regression

$\vec{\mu}, V \leftarrow \mathbb{E}[\vec{\beta}|\mathcal{D}, \Sigma], \mathrm{Var}[\vec{\beta}|\mathcal{D}, \Sigma]$
for $d = 1, \ldots, D$ **do**
    $\mu_d \leftarrow (e_d \otimes I_Q)^\top \vec{\mu}$
    $V_d \leftarrow (e_d \otimes I_Q)^\top V (e_d \otimes I_Q)$
Return $\mu_1, \ldots, \mu_D, V_1, \ldots, V_D$

Algorithm 3 E-Step: Logistic Regression

$\vec{\mu}^* \leftarrow \arg\max_{\vec{\beta}} \log p(\vec{\beta}|\mathcal{D}, \Sigma)$
$V \leftarrow -[\nabla^2_{\vec{\beta}} \log p(\vec{\beta}|\mathcal{D}, \Sigma)\big|_{\vec{\beta}=\vec{\mu}^*}]^{-1}$
for $d = 1, \ldots, D$ **do**
    $\mu_d \leftarrow (e_d \otimes I_Q)^\top \vec{\mu}^*$
    $V_d \leftarrow (e_d \otimes I_Q)^\top V (e_d \otimes I_Q)$
Return $\mu_1, \ldots, \mu_D, V_1, \ldots, V_D$

so we apply an approximation. Specifically, we adapt the original E-step in Algorithm 3 by using a Laplace approximation to the posterior [Bishop, 2006, Chapter 4.4]. We approximate the posterior mean of β by the maximum a posteriori value. We leave extensions to other generalized linear models to future work.

2.4 Theoretical comparison of frequentist risk

In this section, we prove theory that suggests ECov has better frequentist risk than EGroup when D is large relative to Q. Analyzing $\hat{\beta}_{\mathrm{ECov}}$ directly is challenging due to its non-differentiability as a function of the data, so we take a multipart approach. First, in Theorem 2.4.2, we show that an ECov estimate based on moment-matching (MM), $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$, dominates least squares, $\hat{\beta}_{\mathrm{LS}}$, when D is large relative to Q; $\hat{\beta}_{\mathrm{LS}}$ in turn dominates $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ (a similar estimator for EGroup). Second, in Theorem 2.4.3, we show that $\hat{\beta}_{\mathrm{ECov}}$ uniformly improves on $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$.

Setup. Take a fixed value of β and an estimator $\hat{\beta}$. We use squared error risk, $\mathrm{R}(\beta, \hat{\beta}) := \mathbb{E}\left[\|\hat{\beta} - \beta\|_F^2 \mid \beta\right]$, as our measure of performance. $\|\cdot\|_F$ is the Frobenius norm of a matrix, and the expectation is over all observations $Y^1, \ldots, Y^Q$ jointly. We

require the following orthogonal design condition.

Condition 2.4.1. For each group q, $\sigma_q^{-2} X^{q\top} X^q = \sigma^{-2} I_D$ for some shared variance σ^2.

Though restrictive, this condition is useful for theory, as other authors have found; see Appendix A.3.1. We empirically demonstrate that our theoretical conclusions apply more broadly in Section 2.6.

ECov vs. EGroup when using moment matching in high dimensions. For ECov, the following estimate for Σ is unbiased under correct prior specification: $\hat{\Sigma}^{\mathrm{MM}} := D^{-1}\hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} - D^{-1}\mathrm{diag}(\sigma_1^2 \|X^{1\dagger}\|_F^2, \ldots, \sigma_Q^2 \|X^{Q\dagger}\|_F^2)$, where $\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix and $\hat{\beta}_{\mathrm{LS}}$ is the least squares estimate. We define $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}} := \mathbb{E}[\beta | \mathcal{D}, \hat{\Sigma}^{\mathrm{MM}}]$ to be the resulting parameter estimate, and define $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ analogously for EGroup; see Appendix A.3.2 for details. While $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ and $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ are naturally defined only when $D \geq Q$ and $D \leq Q$, respectively, we find it informative to compare how their performances depend on D and Q nonetheless.

Before our theorem, a lemma provides concise expressions for the risks of $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ and $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$.

Lemma 2.4.1. *Under Condition 2.4.1 and when* $D \geq Q$, $\mathrm{R}(\beta, \hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}) = \sigma^2 DQ - \sigma^4 D(D - 2 - 2Q)\mathbb{E}[\|\hat{\beta}^\dagger_{\mathrm{LS}}\|_F^2 \mid \beta]$. *Additionally, when* $D \leq Q$, $\mathrm{R}(\beta, \hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}) = \sigma^2 DQ - \sigma^4 Q(Q - 2 - 2D)\mathbb{E}[\|\hat{\beta}^\dagger_{\mathrm{LS}}\|_F^2 \mid \beta]$.

Lemma 2.4.1 reveals forms for the risks of $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ and $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ that are surprisingly simple. The symmetry between the forms and risks of these estimators, however, is intuitive; under Condition 2.4.1, $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ and $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ can be seen as respectively arising from the same procedure applied to $\hat{\beta}_{\mathrm{LS}}$ and its transpose.

With Lemma 2.4.1 in hand, we can now compare the risk of $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$, $\hat{\beta}_{\mathrm{LS}}$, and $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$.

Theorem 2.4.2. *Let Condition 2.4.1 hold. Then (1) if* $D > 2Q + 2$, $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ *dominates* $\hat{\beta}_{\mathrm{LS}}$ *with respect to squared error risk. In particular, for any* β, $\mathrm{R}(\beta, \hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}) < \mathrm{R}(\beta, \hat{\beta}_{\mathrm{LS}})$. *Additionally, (2) if* $D > Q/2 - 1$, $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ *is dominated by* $\hat{\beta}_{\mathrm{LS}}$.

Since $\hat{\beta}_{\mathrm{LS}}$ is minimax [Lehmann and Casella, 2006, Chapter 5], Theorem 2.4.2 implies that $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ has minimax risk in the high-dimensional setting. It follows that,

regardless of how well the ECov prior assumptions hold, $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ will not perform very poorly.

Further improvement with maximum marginal likelihood. The moment based approach analyzed above has a limitation: with positive probability, $\hat{\Sigma}^{\mathrm{MM}}$ is not positive semi-definite (PSD). Though our expression for $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ remains well-defined in this case, this non-positive definiteness obscures the interpretation of $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ as a Bayes estimate. We next show that performance further improves if Σ is instead estimated by maximum marginal likelihood (Eq. (2.1)) and is thereby constrained to be PSD.

Our next lemma characterizes the form of the resulting estimator, $\hat{\beta}_{\mathrm{ECov}}$, and establishes a connection to the positive part James-Stein estimator [Baranchik, 1964].

Lemma 2.4.2. *Assume $D > Q$ and consider the singular value decomposition $\hat{\beta}_{\mathrm{LS}} = V\mathrm{diag}(\lambda^{\frac{1}{2}})U^\top$ where V and U satisfy $V^\top V = U^\top U = I_Q$, and λ is a Q-vector of non-negative reals. Under Condition 2.4.1, Eq.* (2.1) *reduces to $\hat{\Sigma} = U\mathrm{diag}\left[(D^{-1}\lambda - \sigma^2\mathbf{1}_Q)_+\right]U^\top$ and $\hat{\beta}_{\mathrm{ECov}} = V\mathrm{diag}\left[\lambda^{\frac{1}{2}} \odot (\mathbf{1}_Q - \sigma^2 D\lambda^{-1})_+\right]U^\top$, where $(\cdot)_+$ is shorthand for $\max(\cdot, 0)$ element-wise, $\odot$ is the Hadamard (i.e. element-wise) product, and the powers in $\lambda^{\frac{1}{2}}$ and λ^{-1} are applied element-wise.*

Lemma 2.4.2 allows us to see $\hat{\beta}_{\mathrm{ECov}}$ as shrinking $\hat{\beta}_{\mathrm{LS}}$ toward 0 in the direction of each singular vector to an extent proportional to the inverse of the associated singular value. The transition from $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ to $\hat{\beta}_{\mathrm{ECov}}$ is then analogous to the taking the "positive part" of the James-Stein estimator in vector estimation, which provides a uniform improvement in risk [Baranchik, 1964]. Though $\mathrm{R}(\beta, \hat{\beta}_{\mathrm{ECov}})$ is not easily available analytically, we nevertheless find that it dominates its moment-based counterpart.

Theorem 2.4.3. *Assume $D > Q + 1$. Under Condition 2.4.1 $\hat{\beta}_{\mathrm{ECov}}$ dominates $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ with respect to squared error loss, achieving strictly lower risk for every value of β.*

We establish Theorem 2.4.3 using a proof technique adapted from Baranchik [1964]; see also Lehmann and Casella [2006][Thm. 5.5.4]. The standard approach we build upon is complicated by the fact that the directions in which we apply shrinkage are themselves random.

Theorem 2.4.3 provides a strong line of support for using $\hat{\beta}_{\text{ECov}}$ over $\hat{\beta}^{\text{MM}}_{\text{ECov}}$ that does not rely on any assumption of "correct" prior specification; in particular the risk improves without any subjective assumptions on β. We discuss related earlier work in Appendix A.1.4.

2.5 Gains from ECov in the high-dimensional limit

The results of Section 2.4 give a promising endorsement of ECov but face two important limitations. First, the domination results relative to least squares do not directly demonstrate that $\hat{\beta}_{\text{ECov}}$ attains improvements by leveraging similarities across groups in a meaningful way; indeed for a single group (i.e. $Q = 1$) $\hat{\beta}_{\text{ECov}}$ can be understood as a ridge regression estimate [Hoerl and Kennard, 1970], and Theorems 2.4.2 and 2.4.3 provide that $\hat{\beta}_{\text{ECov}}$ dominates $\hat{\beta}_{\text{LS}}$ for $D > 3$. Second, domination results reveal nothing about the size of the improvement or how it depends on any structure of β; intuitively, we should expect better performance when β is in some way representative of the assumed prior. To address these limitations, we analyze the size of the gap between the risk of (1) $\hat{\beta}_{\text{ECov}}$ and (2) our method applied to each group independently (ID), which we denote by $\hat{\beta}_{\text{ID}}$.[1] We will characterize the dependence of this gap on β.

Reasoning quantitatively about the dependence of the risk on the unknown parameter poses significant analytical challenges. In particular, Lemma 2.4.1 shows that $\text{R}(\beta, \hat{\beta}^{\text{MM}}_{\text{ECov}})$ depends on β through $\mathbb{E}[\|\hat{\beta}^{\dagger}_{\text{LS}}\|_F^2|\beta]$; however, $\|\hat{\beta}^{\dagger}_{\text{LS}}\|_F^2$ is the sum of the eigenvalues of a non-central inverse Wishart matrix, a notoriously challenging quantity to work with; see e.g. [Letac and Massam, 2004, Hillier and Kan, 2019]. To regain tractability, we (1) develop an analysis asymptotic in the number of covariates D and (2) shift to a Bayesian analysis in order to sensibly consider a growing collection of covariate effects. In particular, we consider a sequence of regression problems, with parameters $\{\beta_d\}_{d=1}^{\infty}$ distributed as $\beta_d \overset{i.i.d.}{\sim} \pi$ for some distribution π. Accordingly, instead of using the frequentist risk as in Section 2.4, we now use the Bayes risk to

[1] Our approach $\hat{\beta}_{\text{ECov}}$ is well defined in the $Q = 1$ single group case; for each group q, we obtain $\hat{\beta}^q_{\text{ID}}$ by computing $\hat{\beta}_{\text{ECov}}$ on the group $\mathcal{D} = \{(X^q, Y^q)\}$.

measure performance. Specifically, for a group with D covariates and an estimator $\hat{\beta}$, the Bayes risk is $\mathrm{R}_\pi^D(\hat{\beta}) := \mathbb{E}_\pi[\mathrm{R}(\beta, \hat{\beta})]$ where $\mathrm{R}(\beta, \hat{\beta})$ is the usual frequentist risk. In the following, we describe the results of this analysis with proofs and additional details left to Appendix A.4.

For a single metric characterizing the benefits of joint modeling, we will define the *asymptotic gain* as the relative performance between our two estimators of interest here, $\hat{\beta}_{\mathrm{ECov}}$ and $\hat{\beta}_{\mathrm{ID}}$.

Definition 2.5.1. Consider a sequence of datasets of Q regression problems with an increasing number of covariates $D, \{\mathcal{D}_D\}_{D=1}^{\infty}$. Assume that for each group Condition 2.4.1 is satisfied with variance σ^2 and that each $\beta_d \overset{i.i.d.}{\sim} \pi$. The asymptotic gain of joint modeling is $\mathrm{Gain}(\pi, \sigma^2) := \lim_{D\to\infty}(\sigma^2 DQ)^{-1}[\mathrm{R}_\pi^D(\hat{\beta}_{\mathrm{ID}}) - \mathrm{R}_\pi^D(\hat{\beta}_{\mathrm{ECov}})]$.

The factor of $\sigma^2 DQ$ in Definition 2.5.1 puts $\mathrm{Gain}(\pi, \sigma^2)$ on a scale that is roughly invariant to the size and noise level of the problem; for example, $(\sigma^2 DQ)^{-1}\mathrm{R}_\pi^D(\hat{\beta}_{\mathrm{LS}}) = 1$ for any π, D, and Q. In Appendix A.4.5 we discuss how this asymptotic formulation may allow relaxation of Condition 2.4.1 if one considers certain random design matrices; for simplicity, the present analysis considers only fixed designs.

Our next lemma gives an analytic expression for $\mathrm{Gain}(\pi, \sigma^2)$ that provides a starting point for understanding its problem dependence.

Lemma 2.5.1. *Assume* $\tilde{\Sigma} := \mathrm{Var}_\pi[\beta_1]$ *is finite and has eigenvalues* $\lambda_1, \ldots, \lambda_Q$. *If Condition 2.4.1 satisfied asymptotically,* $\mathrm{Gain}(\pi, \sigma^2) = \sigma^2 Q^{-1}[\sum_{q=1}^{Q}(\lambda_q + \sigma^2)^{-1} - \sum_{q=1}^{Q}(\tilde{\Sigma}_{q,q} + \sigma^2)^{-1}]$.

Lemma 2.5.1 reveals that the diagonals and eigenvalues and $\tilde{\Sigma}$ are key determinants of $\mathrm{Gain}(\pi, \sigma^2)$, but does not directly provide an interpretation of when $\hat{\beta}_{\mathrm{ECov}}$ offers benefits over $\hat{\beta}_{\mathrm{ID}}$. Our next theorem demonstrates when an improvement can be achieved from joint modeling.

Theorem 2.5.2. $\mathrm{Gain}(\pi, \sigma^2) \geq 0$, *with equality only when* $\tilde{\Sigma} = \mathrm{Var}_\pi[\beta_1]$ *is diagonal.*

Proof. From Lemma 2.5.1 we see $\mathrm{Gain}(\pi, \sigma^2)$ is the difference between a strictly Schur-convex function applied to the eigenvalues of $\tilde{\Sigma}$ and to its diagonals (since $(x + \sigma^2)^{-1}$

is convex on $\mathbb{R}_+$). By the Schur-Horn theorem, the eigenvalues of $\tilde{\Sigma}$ majorize its diagonals, providing the result. □

Theorem 2.5.2 tells us that $\hat{\beta}_{\text{ECov}}$ succeeds at adaptively learning and leveraging similarities among groups in the high-dimensional limit. In particular, $\text{Gain}(\pi, \sigma^2)$ reduces to zero only when the eigenvalues of $\tilde{\Sigma}$ are arbitrarily close to the entries of its diagonal, which occurs only when the covariate effects are uncorrelated across groups. However, when covariate effects are correlated, we obtain an improvement.

Our next theorem quantifies this relationship through upper and lower bounds.

Theorem 2.5.3. *Let $\lambda^{\downarrow}$ and $\ell^{\downarrow}$ denote the eigenvalues and diagonals of $\tilde{\Sigma}$, respectively, sorted in descending order. Then* $\text{Gain}(\pi, \sigma^2) \leq 2\sigma^2 Q^{-1} \|\lambda\|_2 \|\ell^{\downarrow} - \lambda^{\downarrow}\|_2 / (\lambda_{\min} + \sigma^2)^3$ *and* $\text{Gain}(\pi, \sigma^2) \geq \sigma^2 Q^{-1} \|\ell^{\downarrow} - \lambda^{\downarrow}\|_2^2 / (\lambda_{\max} + \sigma^2)^3$, *where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest, respectively, eigenvalues of $\tilde{\Sigma}$.*

Theorem 2.5.3 allows us to see several aspects of when our method will and will not perform well. First, the presence of $\|\ell^{\downarrow} - \lambda^{\downarrow}\|_2^2$ in both the upper and lower bounds demonstrates that $\text{Gain}(\pi, \sigma^2)$ will be small when the eigenvalues are close to the diagonal entries, with Euclidean distance as an informative metric.

As we find in our next corollary, Theorem 2.5.3 additionally allows us to see that nontrivial gains may be obtained only in an intermediate signal-to-noise regime, where signal is given by the size of the covariate effects and noise is the variance level σ^2. Notably, under Condition 2.4.1, σ^2 relates directly to the variance of $\hat{\beta}_{\text{LS}}$, and is influenced by both the residual variances and the group sizes; see Appendix A.3.1. In particular we interpret $\lambda_{\min}$ as a proxy for signal strength since it captures the magnitude of typical β_d's along their direction of least variation.

Corollary 2.5.4. $\text{Gain}(\pi, \sigma^2) \leq 4\kappa^2 \lambda_{\min} / \sigma^2$ *and* $\text{Gain}(\pi, \sigma^2) \leq 4\kappa^2 (\lambda_{\min} / \sigma^2)^{-1}$, *where $\kappa := \lambda_{\max} / \lambda_{\min}$ is the condition number of $\tilde{\Sigma}$.*

Corollary 2.5.4 formalizes the intuitive result that with enough noise, the little recoverable signal is insufficient to effectively share strength. And furthermore, in the low-noise and high-signal regime $\hat{\beta}_{\text{ID}}$ is very accurate on its own and there is little

need for joint modeling. However, when there is a large gap between the largest and smallest eigenvalues of $\tilde{\Sigma}$, leading κ to be large, the gain could be larger. κ will be large, for example, when the covariate effects are very correlated across groups.

2.6 Experiments

2.6.1 Simulated data

We first conduct simulations, where we can directly control the relatedness among groups and where we know the ground truth values of the parameters. We show that ECov is more accurate than EGroup when covariates outnumber groups, whether effects are correlated across groups or not.

In particular, we simulated covariates, parameters, and responses for $Q = 10$ groups across a range of covariate dimensions. We generated covariate effects as $\beta_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$. We chose Σ so that effects were either correlated (Figure 2-1 Left) or independent (Figure 2-1 Right) across groups; see Appendix A.5 for details. We compare performance of six estimates on these groups. These are estimates assuming EGroup/ECov using moment matching and maximum marginal likelihood to choose Σ/Γ ($\hat{\beta}^{\text{MM}}_{\text{EGroup}}/\hat{\beta}^{\text{MM}}_{\text{ECov}}$ and $\hat{\beta}_{\text{EGroup}}/\hat{\beta}_{\text{ECov}}$, respectively), as well as least squares ($\hat{\beta}_{\text{LS}}$), and ECov applied to each group independently ($\hat{\beta}_{\text{ID}}$).

Figure 2-1 reinforces our theoretical conclusions that (1) $\hat{\beta}_{\text{ECov}}$ is more accurate when covariates outnumber groups and (2) $\hat{\beta}_{\text{EGroup}}$ is more accurate when groups outnumber covariates. Our simulated X matrices are somewhat relaxed from a strict orthogonal design (Appendix A.5), so these experiments suggest that our conclusions hold beyond Condition 2.4.1. Additionally, $\hat{\beta}_{\text{ECov}}$ and $\hat{\beta}_{\text{EGroup}}$ both outperform their moment based counterparts, $\hat{\beta}^{\text{MM}}_{\text{ECov}}$ and $\hat{\beta}^{\text{MM}}_{\text{EGroup}}$.

Even for the simulations with independent effects, Theorem 2.4.2 suggests $\hat{\beta}_{\text{ECov}}$ should still outperform $\hat{\beta}_{\text{LS}}$ and $\hat{\beta}_{\text{EGroup}}$ in the higher dimensional regime, and we see this behavior in the right panel of Figure 2-1. Additionally, in agreement with Theorem 2.5.2, $\hat{\beta}_{\text{ECov}}$ does not improve over $\hat{\beta}_{\text{ID}}$ in the presence of independent effects,

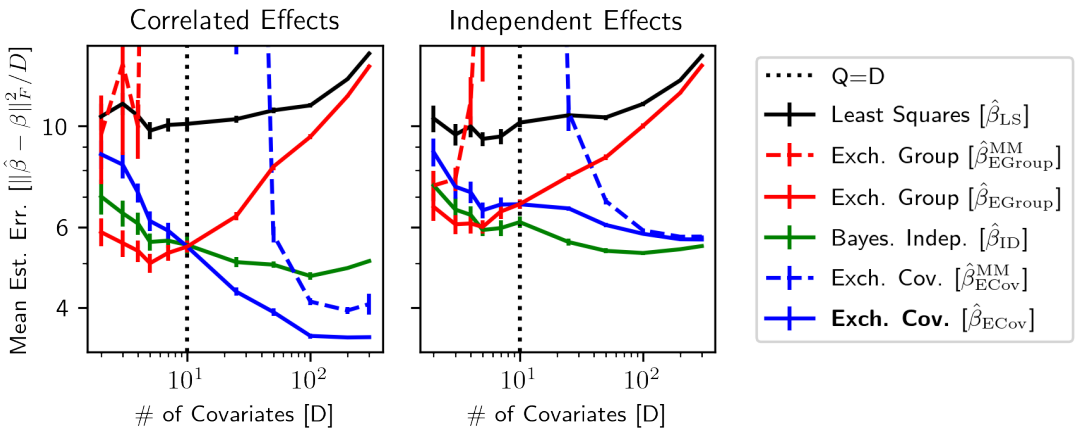


Figure 2-1: Dimension dependence of parameter estimation error in simulation. Covariate effects are either [Left] correlated or [Right] independent across the $Q = 10$ groups. Each point is the mean ± 1SEM across 20 replicates.

and the performances of these two estimators converge as D grows.

2.6.2 Real data

We find that ECov beats EGroup, as well as least squares and independent estimation, across three real groups. We describe the datasets (with additional details in Appendix A.5.4) and then our results.

Community level law enforcement in the United States. Policing rates vary dramatically across different communities, mediating disparate impacts of criminal law enforcement across racial and socioeconomic groups [Weisburd et al., 2019, Slocum et al., 2020]. Understanding how demographic and socioeconomic attributes of communities relate to variation in rates of law enforcement is crucial to understanding these impacts. Linear models provide the desired interpretability. We use a dataset [Redmond and Baveja, 2002] consisting of $D = 117$ community characteristics and their rates of law enforcement (per capita) for different crimes. We consider $Q = 4$ group subsets corresponding to distinct (region, crime) pairs: (Midwest, Robbery), (South, Assault), (Northeast, Larceny), and (West, Auto-theft). This data setup illustrates a small Q and accords with the independent residuals assumption in the likelihood shared by ECov and EGroup (Section 2.2). Across q, N^q represents between

400 and 600 communities.

Blog post popularity. We regress reader engagement (responses) on $D = 279$ characteristics of blog posts (covariates) [Buza, 2014]. We divided the corpus based on an included length attribute into $Q = 3$ groups, corresponding to (1) long posts, (2) short posts, and (3) posts from an earlier corpus with missing length attribute. We hypothesized that the relationships between the characteristics of posts and engagement would differ across these three groups. We randomly downsampled to $N^q = 500$ posts in each group to mimic a low sample-size regime, in which sharing strength is crucial.

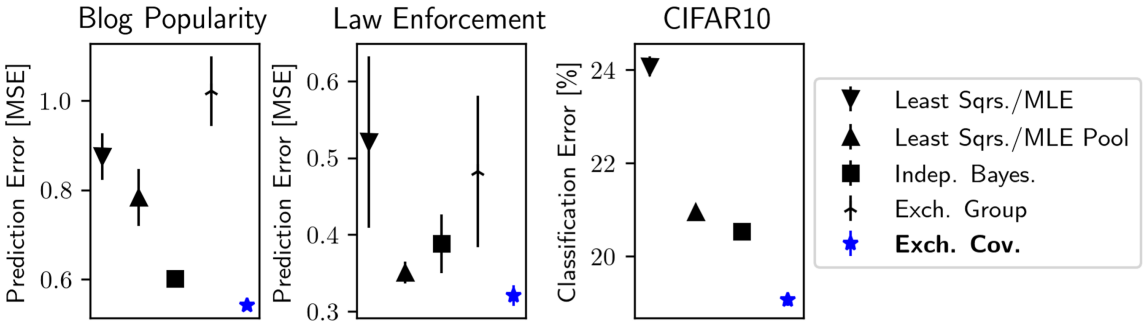


Figure 2-2: Prediction performance on held out data in three applications (mean ± 1SEM across 5-fold cross-validation splits).

Multiple binary classifications using pre-trained neural network embeddings on CIFAR10. Modern machine learning methods have proved very successful on large datasets. Translating this success to smaller datasets is one of the most actively pursued algorithmic challenges in machine learning. It has spurred the development of frameworks from transfer learning [Weiss et al., 2016] to one-shot learning [Vinyals et al., 2016] to meta-learning [Finn et al., 2017]. One common and simple strategy starts with a learned representation (or "embedding") from an expressive neural network fit to a large group. Then one can use this embedding as a covariate vector for classification tasks with few labeled data points.

We take a $D = 128$ dimensional embedding of the CIFAR10 image group [Krizhevsky, 2009, Van Looveren et al., 2019]. We create $Q = 8$ different binary classification tasks using the classes in CIFAR10 (Appendix A.5.4). We downsampled to N^q varying from 100 to 1000 to mimic a setting in which we hope to share strength from large groups

to improve performance on smaller datasets.

Discussion of evaluation and results. In previous sections we have focused on parameter estimation. Here we instead evaluate with prediction error on held-out data since the true parameters are not observed. Specifically we perform 5-fold cross-validation and report the mean squared errors and classification errors on test splits. To reduce variance of out-of-sample error estimates on the applications in which we downsampled, we also evaluate on the additional held-out data. Because the residual variances were unknown, we estimated these for each application and group as $\hat{\sigma}_q^2 := \|P_{X^q}^{\perp} Y^q\|^2/(N_q - D)$, where $P_{X^q}^{\perp} := I_{N_q} - X^q(X^{q\top}X^q)^{-1}X^{q\top}$ (see e.g. [Gelman and Hill, 2006, Chapter 18.1]). All methods ran quickly on a 36 CPU machine; computation of $\hat{\beta}_{\text{ECov}}$, including the EM algorithm, required 2.04 ± 0.64, 6.89 ± 3.19 and 37.14 ± 3.39 seconds (`mean` $\pm$ `st-dev` across splits) on the law enforcement, blog, and CIFAR10 tasks, respectively.

Our results further reinforce the main aspects of our theory. $\hat{\beta}_{\text{ECov}}$ outperformed $\hat{\beta}_{\text{EGroup}}$, independent Bayes estimates ($\hat{\beta}_{\text{ID}}$), and least squares ($\hat{\beta}_{\text{LS}}$) in all applications (at $> 95\%$ nominal confidence with a paired t-test).[2] Additionally, $\hat{\beta}_{\text{ECov}}$ outperformed the baseline of ignoring heterogeneity, pooling groups together, and using the same effect estimates for every group ("Least Sqrs./MLE Pool").

Appendix A.5 includes additional results and comparisons. In particular, we provide the performance of the estimators on each component group for each application. Additionally, we report the performances of (1) stable and computationally efficient moment based alternatives to $\hat{\beta}_{\text{ECov}}$ and $\hat{\beta}_{\text{EGroup}}$ and (2) variants of $\hat{\beta}_{\text{ECov}}$ and $\hat{\beta}_{\text{EGroup}}$ that include a learned (rather than zero) prior mean. Appendix A.5.5 reports the licenses of software we used.

[2] We did not develop an extension akin to Algorithm 3 for EGroup, and so do not report $\hat{\beta}_{\text{EGroup}}$ for CIFAR10. Additionally, we report a maximum likelihood estimate (MLE) instead of $\hat{\beta}_{\text{LS}}$ for CIFAR10.

2.7 Discussion

The Bayesian community has long used hierarchical modeling with priors encoding exchangeability of effects across groups of data (EGroup). In the present work, we have made a case for instead using priors that encode exchangeability across *covariates* (ECov) – in particular, when the number of covariates exceeds the number of groups. We have presented a corresponding concrete model and inference method. We have shown that ECov outperforms EGroup in theory and practice when the number of covariates exceeds the number of groups.

Our approach is, of course, not a panacea. In some settings, a priori exchangeability among covariate effects will be inconsistent with prior beliefs. For example, imagine in the CIFAR10 application if meta-data covariates (such as geo-location and date) were available, in addition to embeddings. Then we might achieve better performance by treating meta-data covariates as distinct from embedding covariates. Additionally, we focused on a Gaussian prior for convenience. In cases where practitioners have more specific prior beliefs about effects, alternative priors and likelihoods may be warranted, though they may be more computationally challenging. Moreover, while relatively interpretable, linear models have their downsides. The linear assumption can be overly simplistic in many applications. It is common to misinterpret effects as causal rather than associative. Both the linear model and squared error loss lend themselves naturally to reporting means, but in many applications a median or other summary is more appropriate; so using a mean for convenience can be misleading.

Many exciting directions for further investigation remain. For example, the covariance Σ may provide an informative measure of task similarity; this similarity measure can be useful in, e.g., meta learning [Jerfel et al., 2019] and statistical genetics [Bulik-Sullivan et al., 2015]. Additionally, we here explored two approaches to choosing the covariance matrices in the empirical Bayes step; more sophisticated approaches to covariance estimation may provided improved performance. It also remains to extend our methodology to other generalized linear models.

Chapter 3

LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

Abstract

Due to the ease of modern data collection, applied statisticians often have access to a large set of covariates that they wish to relate to some observed outcome. Generalized linear models (GLMs) offer a particularly interpretable framework for such an analysis. In these high-dimensional problems, the number of covariates is often large relative to the number of observations, so we face non-trivial inferential uncertainty; a Bayesian approach allows coherent quantification of this uncertainty. Unfortunately, existing methods for Bayesian inference in GLMs require running times roughly cubic in parameter dimension, and so are limited to settings with at most tens of thousand parameters. We propose to reduce time and memory costs with a low-rank approximation of the data in an approach we call LR-GLM. When used with the Laplace approximation or Markov chain Monte Carlo, LR-GLM provides a full Bayesian posterior approximation and admits running times reduced by a full factor of the parameter dimension. We rigorously establish the quality of our approximation

and show how the choice of rank allows a tunable computational–statistical trade-off. Experiments support our theory and demonstrate the efficacy of LR-GLM on real large-scale datasets.

3.1 Introduction

Scientists, engineers, and social scientists are often interested in characterizing the relationship between an outcome and a set of covariates, rather than purely optimizing predictive accuracy. For example, a biologist may wish to understand the effect of natural genetic variation on the presence of a disease or a medical practitioner may wish to understand the effect of a patient's history on their future health. In these applications and countless others, the relative ease of modern data collection methods often yields particularly large sets of covariates for data analysts to study. While these rich data should ultimately aid understanding, they pose a number of practical challenges for data analysis. One challenge is how to discover interpretable relationships between the covariates and the outcome. Generalized linear models (GLMs) are widely used in part because they provide such interpretability – as well as the flexibility to accommodate a variety of different outcome types (including binary, count, and heavy-tailed responses). A second challenge is that, unless the number of data points is substantially larger than the number of covariates, there is likely to be non-trivial uncertainty about these relationships.

A Bayesian approach to GLM inference provides the desired coherent uncertainty quantification as well as favorable calibration properties [Dawid, 1982, Theorem 1]. Bayesian methods additionally provide the ability to improve inference by incorporating expert information and sharing power across experiments. Using Bayesian GLMs leads to computational challenges, however. Even when the Bayesian posterior can be computed exactly, conjugate inference costs $O(N^2D)$ in the case of $D \gg N$. And most models are sufficiently complex as to require expensive approximations.

In this work, we propose to reduce the effective dimensionality of the feature set as a pre-processing step to speed up Bayesian inference, while still performing inference

in the original parameter space; in particular, we show that low-rank descriptions of the data permit fast Markov chain Monte Carlo (MCMC) samplers and Laplace approximations of the Bayesian posterior for the full feature set. We motivate our proposal with a conjugate linear regression analysis in the case where the data are exactly low-rank. When the data are merely approximately low-rank, our proposal is an approximation. Through both theory and experiments, we demonstrate that low-rank data approximations provide a number of properties that are desirable in an efficient posterior approximation method: (1) *soundness:* our approximations admit error bounds directly on the quantities that practitioners report as well as practical interpretations of those bounds; (2) *tunability:* the choice of the rank of the approximation defines a tunable trade-off between the computational demands of inference and statistical precision; and (3) *conservativeness:* our approximation never reports less uncertainty than the exact posterior, where uncertainty is quantified via either posterior variance or information entropy. Together, these properties allow a practitioner to choose how much information to extract from the data on the basis of computational resources while being able to confidently trust the conclusions of their analysis.

3.2 Bayesian inference in GLMs

Suppose we have N data points $\{(x_n, y_n)\}_{n=1}^N$. We collect our covariates, where x_n has dimension D, in the design matrix $X \in \mathbb{R}^{N \times D}$ and our responses in the column vector $Y \in \mathbb{R}^N$. Let $\beta \in \mathbb{R}^D$ be an unknown parameter characterizing the relationship between the covariates and the response for each data point. In particular, we take β to parameterize a GLM likelihood $p(Y \mid X, \beta) = p(Y \mid X\beta)$. That is, β_d describes the effect size of the dth covariate (e.g., the influence of a non-reference allele on an individual's height in a genomic association study). Completing our Bayesian model specification, we assume a prior $p(\beta)$, which describes our knowledge of β before seeing data. Bayes' theorem gives the Bayesian posterior $p(\beta \mid Y, X) = p(\beta)p(Y \mid X\beta)/\int p(\beta')p(Y \mid X\beta')d\beta'$, which captures the updated state of our knowledge after

Table 3.1: Time complexities of naive inference and LR-GLM with a rank M approximation when $D \geq N$.

Method	Naive	LR-GLM	speedup
Laplace	$O(N^2D)$	$O(NDM)$	N/M
MCMC (iter.)	$O(ND)$	$O(NM + DM)$	N/M

observing data. We often summarize the β posterior via its mean and covariance. In all but the simplest settings, though, computing these posterior summaries is analytically intractable, and these quantities must be approximated.

Related work. In the setting of large D and large N, existing Bayesian inference methods for GLMs may lead to unfavorable trade-offs between accuracy and computation; see Appendix B.2 for further discussion. While Markov chain Monte Carlo (MCMC) can approximate Bayesian GLM posteriors arbitrarily well given enough time, standard methods can be slow, with $O(DN)$ time per likelihood evaluation. Moreover, in practice, mixing time may scale poorly with dimension and sample size; algorithms thus require many iterations and hence many likelihood evaluations. Subsampling MCMC methods can speed up inference, but they are effective only with tall data [$D \ll N$; Bardenet et al., 2017].

An alternative to MCMC is to use a deterministic approximation such as the Laplace approximation [Bishop, 2006, Chap. 4.4], integrated nested Laplace approximation [Rue et al., 2009], variational Bayes [VB; Blei et al., 2017], or an alternative likelihood approximation [Huggins et al., 2017, Campbell and Broderick, 2019, Huggins et al., 2016]. However these methods are computationally efficient only when $D \ll N$ (and in some cases also when $N \ll D$). For example, the Laplace approximation requires inverting the Hessian, which uses $O(\min(N, D)ND)$ time (Appendix B.3). Improving computational tractability by, for example, using a mean field approximation with VB or a factorized Laplace approximation can produce substantial bias and uncertainty underestimation [MacKay, 2003, Turner and Sahani, 2011].

A number of papers have explored using random projections and low-rank approximations in both Bayesian [Lee and Oh, 2013, Spantini et al., 2015, Guhaniyogi and Dunson, 2015, Geppert et al., 2017] and non-Bayesian [Zhang et al., 2014, Wang et al.,

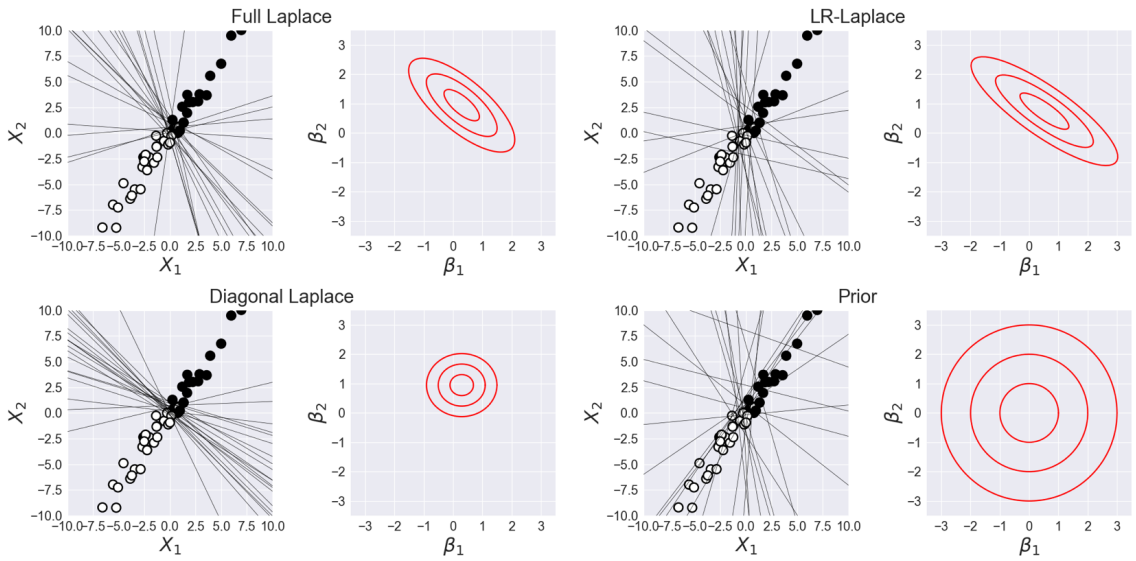


Figure 3-1: LR-Laplace with a rank-1 data approximation closely matches the Bayesian posterior of a toy logistic regression model. In each pair of plots, the left panel depicts the same 2-dimensional dataset with points in two classes (black and white dots) and decision boundaries (black lines) separating the two classes, which are sampled from the given posterior approximation (see title for each pair). In the right panel, the red contours represent the marginal posterior approximation of the parameter β (a bias parameter is integrated out).

2017] settings. The Bayesian approaches have a variety of limitations. E.g., Lee and Oh [2013], Geppert et al. [2017], Spantini et al. [2015] give results only for certain conjugate Gaussian models. And Guhaniyogi and Dunson [2015] provide asymptotic guarantees for prediction but do not address parameter estimation.

See Section 3.6 for a demonstration of the empirical disadvantages of mean field VB, factored Laplace, and random projections in posterior inference.

3.3 LR-GLM

The intuition for our low-rank GLM (LR-GLM) approach is as follows. Supervised learning problems in high-dimensional settings often exhibit strongly correlated covariates [Udell and Townsend, 2019]. In these cases, the data may provide little information about the parameter along certain directions of parameter space. This observation suggests the following procedure: first identify a relatively lower-dimensional

subspace within which the data most directly inform the posterior, and then perform the data-dependent computations of posterior inference (only) within this subspace, at lower computational expense. In the context of GLMs with Gaussian priors, the singular value decomposition (SVD) of the design matrix X provides a natural and effective mechanism for identifying a subspace. We will see that this perspective gives rise to simple, efficient, and accurate approximate inference procedures. In models with non-Gaussian priors the approximation enables more efficient inference by facilitating faster likelihood evaluations.

Formally, the first step of LR-GLM is to choose an integer M such that $0 < M < D$. For any real design matrix X, its SVD exists and may be written as

$$X^\top = U\text{diag}(\lambda)V^\top + \bar{U}\text{diag}(\bar{\lambda})\bar{V}^\top,$$

where $U \in \mathbb{R}^{D\times M}, \bar{U} \in \mathbb{R}^{D\times(D-M)}, V \in \mathbb{R}^{N\times M}$, and $\bar{V} \in \mathbb{R}^{N\times(D-M)}$ are matrices of orthonormal rows, and $\lambda \in \mathbb{R}^M$ and $\bar{\lambda} \in \mathbb{R}^{D-M}$ are vectors of non-increasing singular values $\lambda_1 \geq \cdots \geq \lambda_M \geq \bar{\lambda}_1 \geq \cdots \geq \bar{\lambda}_{D-M} \geq 0$. We replace X with the low-rank approximation $XUU^\top$. Note that the resulting posterior approximation $\tilde{p}(\beta \mid X, Y)$ is still a distribution over the full D-dimensional β vector:

$$\tilde{p}(\beta \mid X, Y) := \frac{p(\beta)p(Y \mid XUU^\top\beta)}{\int p(\beta')p(Y \mid XUU^\top\beta')d\beta'} \tag{3.1}$$

In this way, we cast low-rank data approximations for approximate Bayesian inference as a likelihood approximation. This perspective facilitates our analysis of posterior approximation quality and provides the flexibility either to use the likelihood approximation in an otherwise exact MCMC algorithm or to make additional fast approximations such as the Laplace approximation.

We let *LR-Laplace* denote the combination of LR-GLM and the Laplace approximation. Figure 3-1 illustrates LR-Laplace on a toy problem and compares it to full Laplace, the prior, and diagonal Laplace. Diagonal Laplace refers to a factorized Laplace approximation in which the Hessian of the log posterior is approximated with only its diagonal. While this example captures some of the essence of our pro-

posed approach, we emphasize that our focus in this paper is on problems that are high-dimensional.

3.4 Low-rank data approximations for conjugate Gaussian regression

We now consider the quality of approximate Bayesian inference using our LR-GLM approach in the case of conjugate Gaussian regression. We start by assuming that the data is exactly low rank since it most cleanly illustrates the computational gains from LR-GLM. We then move on to the case of conjugate regression with approximately low-rank data and rigorously characterize the quality of our approximation via interpretable error bounds. We consider non-conjugate GLMs in Section 3.5. We defer all proofs to the Appendix.

3.4.1 Conjugate regression with exactly low-rank data

Classic linear regression fits into our GLM likelihood framework with $p(Y|X,\beta) = \mathcal{N}(Y|X\beta,(\tau I_N)^{-1})$, where $\tau > 0$ is the precision and I_N is the identity matrix of size N. For the conjugate prior $p(\beta) = \mathcal{N}(\beta|0,\Sigma_\beta)$, we can write the posterior in closed form: $p(\beta|Y,X) = \mathcal{N}(\beta|\mu_N,\Sigma_N)$, where $\Sigma_N := (\Sigma_\beta^{-1} + \tau X^T X)^{-1}$ and $\mu_N := \tau\Sigma_N X^\top Y$.

While conjugacy avoids the cost of approximating Bayesian inference, it does not avoid the often prohibitive $O(ND^2 + D^3)$ cost of calculating Σ_N (which requires computing and then inverting Σ_N^{-1}) and the $O(D^2)$ memory demands of storing it. In the $N \ll D$ setting, these costs can be mitigated by using the Woodbury formula to obtain μ_N and Σ_N in $O(N^2D)$ time with $O(ND)$ memory (Appendix B.3). But this alternative becomes computationally prohibitive as well when both N and D are large (e.g., $D \approx N > 20,000$).

Now suppose that X is rank $M \ll \min(D,N)$ and can therefore be written as $X = XUU^T$ exactly, where $U \in \mathbb{R}^{D\times M}$ denotes the top M right singular vectors of X. Then, if $\Sigma_\beta = \sigma_\beta^2 I_D$ and 1_M is the ones vector of length M, we can write (see

Appendix B.4.1 for details)

$$\begin{aligned}\Sigma_N &= \sigma_\beta^2 \left\{ I - U \mathrm{diag}\left(\frac{\tau\lambda \odot \lambda}{\sigma_\beta^{-2} 1_M + \tau\lambda \odot \lambda} \right) U^\top \right\} \\ \text{and} \quad \mu_N &= U \mathrm{diag}\left(\frac{\tau\lambda}{\sigma_\beta^{-2} 1_M + \tau\lambda \odot \lambda} \right) V^\top Y,\end{aligned} \tag{3.2}$$

where multiplication ($\odot$) and division in the diag input are component-wise across the vector λ. Eq. (3.2) provides a more computationally efficient route to inference. The singular vectors in U may be obtained in $O(ND \log M)$ time via a randomized SVD [Halko et al., 2011] or in $O(NDM)$ time using more standard deterministic methods Press et al. [2007]. The bottleneck step is finding λ via $\mathrm{diag}(\lambda \odot \lambda) = U^\top X^\top X U$, which can be computed in $O(NDM)$ time. As for storage, this approach requires keeping only U, λ, and $V^\top Y$, which takes just $O(MD)$ space. In sum, utilizing low-rank structure via Eq. (3.2) provides an order $\min(N, D)/M$-fold improvement in both time and memory over more naive inference.

3.4.2 Conjugate regression with low-rank approximations

While the case with exactly low-rank data is illustrative, real data are rarely exactly low rank. So, more generally, LR-GLM will yield an approximation $\mathcal{N}(\beta|\tilde{\mu}_N, \tilde{\Sigma}_N)$ to the posterior $\mathcal{N}(\beta|\mu_N, \Sigma_N)$, rather than the exact posterior as in Section 3.4.1. We next provide upper bounds on the error from our approximation. Since practitioners typically report posterior means and covariances, we focus on how well LR-GLM approximates these functionals.

Theorem 3.4.1. *For conjugate Bayesian linear regression, the LR-GLM approximation Eq.* (3.1) *satisfies*

$$\|\tilde{\mu}_N - \mu_N\|_2 \leq \frac{\bar{\lambda}_1 \big(\bar{\lambda}_1 \|\bar{U}^\top \tilde{\mu}_N\|_2 + \|\bar{V}^\top Y\|_2 \big)}{\|\tau\Sigma_\beta\|_2^{-1} + \bar{\lambda}_{D-M}^2} \tag{3.3}$$

$$\text{and} \quad \Sigma_N^{-1} - \tilde{\Sigma}_N^{-1} = \tau(X^\top X - UU^\top X^\top X UU^\top). \tag{3.4}$$

In particular, $\|\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1}\|_2 = \tau\bar{\lambda}_1^2$.

The major driver of the approximation error of the posterior mean and covariance is $\bar{\lambda}_1 = \|X - XUU^\top\|_2$, the largest truncated singular value of X. This result accords with the intuition that if the data are "approximately low-rank" then LR-GLM should perform well.

The following corollary shows that the posterior mean estimate is not, in general, consistent for the true parameter. But it does exhibit reasonable asymptotic behavior. In particular, $\tilde{\mu}_N$ is consistent within the span of U and converges to the *a priori* most probable vector with this characteristic (see the toy example in Figure B.5.1).

Corollary 3.4.2. *Suppose* $x_n \overset{i.i.d.}{\sim} p_*$*, for some distribution* p^**, and* $y_n \mid x_n \overset{indep}{\sim} \mathcal{N}(x_n^\top\mu_*, \tau^{-1})$*, for some* $\mu_* \in \mathbb{R}^D$*. Assume* $\mathbb{E}_{p_*}[x_n x_n^\top]$ *is nonsingular. Let the columns of* $U_* \in \mathbb{R}^{D\times M}$ *be the top eigenvectors of* $\mathbb{E}_{p_*}[x_n x_n^T]$*. Then* $\tilde{\mu}_N$ *converges weakly to the maximum a priori vector* $\tilde{\mu}$ *satisfying* $U_*^\top\tilde{\mu} = U_*^\top\mu_*$*.*

In the special case that Σ_β is diagonal this result implies that $\tilde{\mu}_N \xrightarrow{p} U_*U_*^\top\mu_*$ (Appendices B.5.3 and B.6.2). Thus Corollary 3.4.2 reflects the intuition that we are not learning anything about the relation between response and covariates in the data directions that we truncate away with our approach. If the response has little dependence on these directions, $\bar{U}_*\bar{U}_*^\top\mu_* = \lim_{N\to\infty}\tilde{\mu}_N - \mu_*$ will be small and the error in our approximation will be low (Appendix B.5.3). If the response depends heavily on these directions, our error will be higher. This challenge is ubiquitous in dealing with projections of high-dimensional data. Indeed, we often see explicit assumptions encoding the notion that high-variance directions in X are also highly predictive of the response [see, e.g., Zhang et al., 2014, Theorem 2].

Our next corollary captures that LR-GLM never underestimates posterior uncertainty (the *conservativeness* property).

Corollary 3.4.3. *LR-GLM approximate posterior uncertainty in any linear combination of parameters is no less than the exact posterior uncertainty. Equivalently,* $\tilde{\Sigma}_N - \Sigma_N$ *is positive semi-definite.*

[1]This manipulation is purely symbolic. See Appendix B.6.1 for details.

Algorithm 4 LR-Laplace for Bayesian inference in GLMs with low-rank data approximations and zero-mean prior – with computation costs. See Appendix B.8 for the general algorithm.

1:	***Input:*** *prior* $p(\beta) = \mathcal{N}(\mathbf{0}, \Sigma_\beta)$*, data* $X \in \mathbb{R}^{N,D}$*, rank* $M \ll D$*, GLM mapping* ϕ *with* $\vec{\phi}''$ *(see Eq.* (3.5) *and Section 3.5.1)*		
2:	**Pseudo-Code**	3: **Time Complexity**	4: **Mem. Complexity**
5:	*Data preprocessing — M-Truncated SVD*		
6:	$U, \text{diag}(\lambda), V := \text{truncated-SVD}(X^T, M)$	$O(NDM)$	$O(NM + DM)$
7:	*Optimize in projected space, find approximate MAP*		
8:	$\gamma_* := \arg\max_{\gamma \in \mathbb{R}^M} \sum_{i=1}^N \phi(y_i, x_i U\gamma) - \frac{1}{2}\gamma^\top U^\top \Sigma_\beta U\gamma$	$O(NM + DM^2)$	$O(N + M^2)$
9:	$\hat{\mu} = U\gamma_* + \bar{U}\bar{U}^\top \Sigma_\beta U(U^\top \Sigma_\beta U)^{-1}\gamma_*$	$O(DM)$	$O(DM)$
10:	*Compute approximate posterior covariance*		
11:	$W^{-1} := U^\top \Sigma_\beta U - (U^\top X^\top \text{diag}(\vec{\phi}''(Y, XUU^\top \hat{\mu}))XU)^{-1}$	$O(NM^2 + DM)$	$O(NM)$
12:	$\hat{\Sigma} := \Sigma_\beta - \Sigma_\beta UWU^\top \Sigma_\beta$	0 (see note[1])	$O(DM)$
13:	*Compute variances and covariances*		
14:	$\text{Var}_{\hat{p}}(\beta_i) = e_i^\top \hat{\Sigma} e_i$	$O(M^2)$	$O(DM)$
15:	$\text{Cov}_{\hat{p}}(\beta_i, \beta_j) = e_i^\top \hat{\Sigma} e_j$	$O(M^2)$	$O(DM)$

See Figure 3-1 for an illustration of this result. From an approximation perspective, overestimating uncertainty can be seen as preferable to underestimation as it leads to more conservative decision-making. An alternative perspective is that we actually engender additional uncertainty simply by making an approximation, with more uncertainty for coarser approximations, and we should express that in reporting our inferences. This behavior stands in sharp contrast to alternative fast approximate inference methods, such as diagonal Laplace approximations (Appendix B.6.8) and variational Bayes [MacKay, 2003], which can dramatically underestimate uncertainty. We further characterize the conservativeness of LR-GLM in Corollary B.5.1, which shows that the LR-GLM posterior never has lower entropy than the exact posterior and quantifies the bits of information lost due to approximation.

3.5 Non-conjugate GLMs with approximately low-rank data

While the conjugate linear setting facilitates intuition and theory, GLMs are a larger and more broadly useful class of models for which efficient and reliable Bayesian inference is of significant practical concern. Assuming conditional independence of the

observations given the covariates and parameter, the posterior for a GLM likelihood can be written

$$\log p(\beta \mid X, Y) = \log p(\beta) + \sum_{n=1}^{N} \phi(y_n, x_n^\top \beta) + Z \tag{3.5}$$

for some real-valued mapping function ϕ and log normalizing constant Z. For priors and mapping functions that do not form a conjugate pair, accessing posterior functionals of interest is analytically intractable and requires posterior approximation. One possibility is to use a Monte Carlo method such as MCMC, which has theoretical guarantees asymptotic in running time but is relatively slow in practice. The usual alternative is a deterministic approximation such as VB or Laplace. These approximations are typically faster but do not become arbitrarily accurate in the limit of infinite computation. We next show how LR-GLM can be applied to facilitate faster MCMC samplers and Laplace approximations for Bayesian GLMs. We also characterize the additional error introduced to Laplace approximations by low-rank data approximations.

3.5.1 LR-GLM for fast Laplace approximations

The Laplace approximation refers to a Gaussian approximation obtained via a second-order Taylor approximation of the log density. In the Bayesian setting, the Laplace approximation $\bar{p}(\beta \mid X, Y)$ is typically formed at the maximum a posteriori (MAP) parameter: $\bar{p}(\beta \mid X, Y) := \mathcal{N}(\beta \mid \bar{\mu}, \bar{\Sigma})$, where $\bar{\mu} := \arg\max_\beta \log p(\beta \mid X, Y)$ and $\bar{\Sigma}^{-1} := -\nabla^2_\beta \log p(\beta \mid X, Y)|_{\beta=\bar{\mu}}$. When computing and analyzing Laplace approximations for GLMs, we will often refer to vectorized first, second, and third derivatives $\vec{\phi}', \vec{\phi}'', \vec{\phi}''' \in \mathbb{R}^N$ of the mapping function ϕ. For $Y, A \in \mathbb{R}^N$, we define $\vec{\phi}'(Y, A)_n := \frac{\partial}{\partial a}\phi(Y_n, a)|_{a=A_n}$. The higher-order derivative definitions are analogous, with the derivative order of $\frac{\partial}{\partial a}$ increased commensurately.

Laplace approximations are typically much faster than MCMC for moderate/large N and small D, but they become expensive or intractable for large D. In particular, they require inverting a $D \times D$ Hessian matrix, which is in general an $O(D^3)$ time

operation, and storing the resulting covariance matrix, which requires $O(D^2)$ memory.[2]

As in the conjugate case, LR-GLM permits a faster and more memory-efficient route to inference. Here, we say that the *LR-Laplace approximation*, $\hat{p}(\beta \mid X, Y) = \mathcal{N}(\beta \mid \hat{\mu}, \hat{\Sigma})$, denotes the Laplace approximation to the LR-GLM approximate posterior. The special case of LR-Laplace with zero-mean prior is given in Algorithm 4 as it allows us to easily analyze time and memory complexity. For the more general LR-Laplace algorithm, see Appendix B.8.

Theorem 3.5.1. *In a GLM with a zero-mean, structured-Gaussian prior*[3] *and a log-concave likelihood,*[4] *the rank-M LR-Laplace approximation may be computed via Algorithm 4 in $O(NDM)$ time with $O(DM + NM)$ memory. Furthermore, any posterior covariance entry can be computed in $O(M^2)$ time.*

Algorithm 4 consists of three phases: (1) computation of the M-truncated SVD of $X^\top$; (2) MAP optimization to find $\hat{\mu}$; and (3) estimation of $\hat{\Sigma}$. In the second phase we are able to efficiently compute $\hat{\mu}$ by first solving a lower-dimensional optimization for the quantity $\gamma_* \in \mathbb{R}^M$ (Line 8), from which $\hat{\mu}$ is available analytically. Notably, in the common case that $p(\beta)$ is isotropic Gaussian, the expression for $\hat{\mu}$ reduces to $U\gamma_*$ and the full time complexity of MAP estimation is $O(NM + DM)$. Though computing the covariance for each pair of parameters and storing $\hat{\Sigma}$ explicitly would of course require a potentially unacceptable $O(D^2)$ storage, the output of Algorithm 4 is smaller and enables arbitrary parameter variances and covariances to be computed in $O(M^2)$ time. See Appendix B.6.1 for additional details.

3.5.2 Accuracy of the LR-Laplace approximation

We now consider the quality of the LR-Laplace approximate posterior relative to the usual Laplace approximation. Our first result concerns the difference of the posterior means

[2]Notably, as in the conjugate setting, an alternative matrix inversion using the Woodbury identity reduces this cost when $N < D$ to $O(N^2D)$ time and $O(ND)$ memory (Appendix B.3).

[3]For example (banded) diagonal or diagonal plus low-rank, such that matrix vector multiplies may be computed in $O(D)$ time.

[4]This property is standard for common GLMs such as logistic and Poisson regression.

Theorem 3.5.2 (Non-asymptotic)**.** *In a generalized linear model with an α–strongly log concave posterior, the exact and approximate MAP values, $\hat{\mu} = \arg\max_\beta \tilde{p}(\beta \mid X, Y)$ and $\bar{\mu} = \arg\max_\beta p(\beta \mid X, Y)$, satisfy*

$$\|\hat{\mu} - \bar{\mu}\|_2 \leq \frac{\bar{\lambda}_1 \big(\|\vec{\phi}'(Y, X\hat{\mu})\|_2 + \lambda_1 \|\bar{U}^\top \hat{\mu}\|_2 \|\vec{\phi}''(Y, A)\|_\infty \big)}{\alpha}$$

for some vector $A \in \mathbb{R}^N$ such that $A_n \in [x_n^\top U U^\top \hat{\mu}, x_n^\top \hat{\mu}]$.

This bound reveals several characteristics of the regimes in which LR-Laplace performs well. As in conjugate regression, we see that the bound tightens to 0 as the rank of the approximation increases to capture all of the variance in the covariates and $\bar{\lambda}_1 \to 0$.

Remark 3.5.3. For many common GLMs, $\|\vec{\phi}'\|_2$, $\|\vec{\phi}''\|_\infty$, and $\|\vec{\phi}'''\|_\infty$ are well controlled; see Appendix B.6.4. $\|\vec{\phi}'''\|_\infty$ appears in an upcoming corollary.

Remark 3.5.4. The α–strong log concavity of the posterior is satisfied for any strongly log concave prior (e.g., a Gaussian, in which case we have $\alpha \geq \|\Sigma_\beta\|_2^{-1}$) and $\phi(y, \cdot)$ is concave for all y. In this common case, Theorem 3.5.2 provides a computable upper bound on the posterior mean error.

Remark 3.5.5. In contrast to the conjugate case (Corollary 3.4.2), general LR-GLM parameter estimates are not necessarily consistent within the span of the projection. That is, $U^\top \hat{\mu}_N$ may not converge to $U^\top \beta$ (see Appendix B.6.5).

We next consider the distance between our approximation and target posterior under a Wasserstein metric [Villani, 2008]. Let $\Gamma(\hat{p}, \bar{p})$ be the set of all couplings of distributions $\hat{p}$ and $\bar{p}$, i.e. joint distributions $\gamma(\cdot, \cdot)$ satisfying $\hat{p}(\beta) = \int \gamma(\beta, \beta')d\beta'$ and $\bar{p}(\beta) = \int \gamma(\beta', \beta)d\beta'$ for all β. Then the 2-Wasserstein distance between $\hat{p}$ and $\bar{p}$ is defined

$$W_2(\hat{p}, \bar{p}) = \inf_{\gamma \in \Gamma(\hat{p}, \bar{p})} \mathbb{E}_\gamma[\|\hat{\beta} - \bar{\beta}\|_2^2]^{\frac{1}{2}}. \tag{3.6}$$

Wasserstein bounds provide tight control of many functionals of interest, such as means, variances, and standard deviations Huggins et al. [2018]. For example, if

$\xi_i \sim q_i$ for any distribution q_i $(i = 1, 2)$, then $|\mathbb{E}[\xi_1] - \mathbb{E}[\xi_2]| \leq W_2(q_1, q_2)$ and $|\mathrm{Var}[\xi_1]^{\frac{1}{2}} - \mathrm{Var}[\xi_2]^{\frac{1}{2}}| \leq 2W_2(q_1, q_2)$.

We provide a finite-sample upper bound on the 2-Wasserstein distance between the Laplace and LR-Laplace approximations. In particular, the 2-Wasserstein will decrease to 0 as the rank of the LR-Laplace approximation increases since the largest truncated singular value $\bar{\lambda}_1$ will approach zero.

Corollary 3.5.6. *Assume the prior $p(\beta)$ is Gaussian with covariance Σ_β and the mapping function $\phi(y, a)$ has bounded 2nd and 3rd derivatives with respect to a. Take A and α as in Theorem 3.5.2. Then $\bar{p}(\beta)$ and $\hat{p}(\beta)$ satisfy*

$$\begin{aligned} W_2(\hat{p}, \bar{p}) \leq & \sqrt{2}\bar{\lambda}_1 \|\bar{\Sigma}\|_2 \Big\{ c\big[\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2 \|\vec{\phi}''\|_\infty\big] \\ & + (\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty \sqrt{\mathrm{tr}(\hat{\Sigma})} \Big\}, \end{aligned} \tag{3.7}$$

where $c := \big(\|\vec{\phi}'(Y, X\hat{\mu})\|_2 + \lambda_1 \|\bar{U}^\top \hat{\mu}\|_2 \|\vec{\phi}''(Y, A)\|_\infty\big)/\alpha$ and $r := \|U^\top \hat{\mu}\|_\infty \|\vec{\phi}'''\|_\infty + \lambda_1 c \|\vec{\phi}'''\|_\infty$.

When combined with Huggins et al. [2018, Prop. 6.1], this result guarantees closeness in 2-Wasserstein of LR-Laplace to the exact posterior.

We conclude with a result showing that the error due to the LR-GLM approximation cannot grow without bound as the sample size increases.

Theorem 3.5.7 (Asymptotic)**.** *Under mild regularity conditions, the error in the posterior means, $\|\hat{\mu}_n - \bar{\mu}_n\|_2$, converges as $n \to \infty$, and the limit is finite almost surely.*

For the formal statement see Theorem B.6.2 in Appendix B.6.7.

3.5.3 LR-MCMC for faster MCMC in GLMs

LR-Laplace is inappropriate when the posterior is poorly approximated by a Gaussian. This may be the case, for example, when the posterior is multi-modal, a common characteristic of GLMs with sparse priors. To remedy this limitation of LR-Laplace, we

introduce *LR-MCMC*, a wrapper around the Metropolis–Hastings algorithm using the LR-GLM approximation. For a GLM, each full likelihood and gradient computation takes $O(ND)$ time but only $O(NM + DM)$ time with the LR-GLM approximation, resulting in the same $\min(N, D)/M$-fold speedup obtained by LR-Laplace. See Appendix B.7 for further details on LR-MCMC.

3.6 Experiments

We empirically evaluated LR-GLM on real and synthetic datasets. For synthetic data experiments, we considered logistic regression with covariates of dimension $D = 250$ and $D = 500$. In each replicate, we generated the latent parameter from an isotropic Gaussian prior, $\beta \sim \mathcal{N}(0, I_D)$, correlated covariates from a multivariate Gaussian, and responses from the logistic regression likelihood (see Appendix B.1.1 for details). We compared to the standard Laplace approximation, the diagonal Laplace approximation, the Laplace approximation with a low-rank data approximation obtained via random projections rather than the SVD ("Random-Laplace"), and mean-field automatic differentiation variational inference in `Stan` (ADVI-MF).[5]

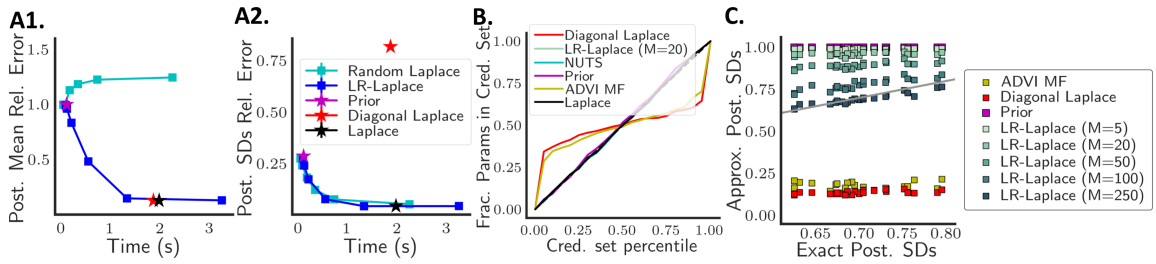


Figure 3-2: *Left*: Error of the approximate posterior (A1.) mean and (A2.) variances relative to ground truth (running NUTS with `Stan`). Lower and further left is better. *Right* (B.): Credible set calibration across all parameters and repeated experiments. (C.): Approximate posterior standard deviations for a subset of parameters. The grey line reflects zero error.

Computational–statistical trade-offs. Figure 3-2A shows empirically the tunable computational–statistical trade-off offered by varying M in our low-rank data

[5]We also tested ADVI using a full rank Gaussian approximation but found it to provide near uniformly worse performance compared to ADVI-MF. So we exclude full-rank ADVI from the presented results.

approximation. This plot depicts the error in posterior mean and variance estimates relative to results from the No-U-Turn Sampler (NUTS) in `Stan` [Hoffman and Gelman, 2014, Carpenter et al., 2017], which we treat as ground truth. As expected, LR-Laplace with larger M takes longer to run but yields lower errors. Random-Laplace was usually faster but provided a poor posterior approximation. Interestingly, the error of the Random-Laplace approximate posterior mean actually increased with the dimension of the projection. We conjecture this behavior may be due to Random-Laplace prioritizing covariate directions that are correlated with directions where the parameter, β, is large.

We also consider predictive performance via the classification error rate and the average negative log likelihood. In particular, we generated a *test* dataset with covariates drawn from the same distribution as the observed dataset and an *out-of-sample* dataset with covariates drawn from a different distribution (see Appendix B.1.1). The computation time vs. performance trade-offs, presented in Figure B.1.1 on the test and out-of-sample datasets, mirror the results for approximating the posterior mean and variances. In this evaluation, correctly accounting for posterior uncertainty appears less important for in-sample prediction. But in the out-of-sample case, we see a dramatic difference in negative log likelihood. Notably, ADVI-MF and diagonal Laplace exhibit much worse performance. These results support the utility of correctly estimating Bayesian uncertainty when making out-of-sample predictions.

Conservativeness. A benefit of LR-GLM is that the posterior approximation never underestimates the posterior uncertainty (see Corollary 3.4.3). Figure 3-2C illustrates this property for LR-Laplace applied to logistic regression. When LR-Laplace misestimates posterior variances, it always overestimates. Also, when LR-Laplace misestimates means (Figure B.1.2), the estimates shrink closer to the prior mean, zero in this case. These results suggest that LR-GLM interpolates between the exact posterior and the prior. Notably, this property is not true of all methods. The diagonal Laplace approximation, by contrast, dramatically underestimates posterior marginal variances (see Appendix B.6.8).

Reliability and calibration. Bayesian methods enjoy desirable calibration

properties under correct model specification. But since LR-Laplace serves as a likelihood approximation, it does not retain this theoretical guarantee. Therefore, we assessed its calibration properties empirically by examining the credible sets of both parameters and predictions. We found that the parameter credible sets of LR-Laplace are extremely well calibrated for all values of M between 20 and 400 (Figure 3-2B and Figure B.1.4). The prediction credible sets were well calibrated for all but the smallest value of M tested ($M = 20$); in the $M = 20$ case, LR-Laplace yielded under-confident predictions (Figure B.1.5). The good calibration of LR-Laplace stood in sharp contrast to the diagonal Laplace approximation and ADVI-MF. Random-Laplace also provided inferior calibration (Figures B.1.4 and B.1.5).

LR-GLM with MCMC and non-Gaussian priors. In Section 3.5.3 we argued that LR-GLM speeds up MCMC for GLMs by decreasing the cost of likelihood and gradient evaluations in black-box MCMC routines. We first examined LR-MCMC with NUTS using `Stan` on the same synthetic datasets as we did for LR-Laplace. In Figures B.1.3 and B.1.6, we see a similar conservativeness and computational–statistical trade-off as for LR-Laplace, and superior performance relative to alternative methods.

We expect MCMC to yield high-quality posterior approximations across a wider range of models than Laplace approximations. For example, for multimodal posteriors and other posteriors that deviate substantially from Gaussianity. We next demonstrate that LR-MCMC is useful in these more general cases. In high-dimensional settings, practitioners are often interested in identifying a sparse subset of parameters that significantly influence responses. This belief may be incorporated in a Bayesian setting through a sparsity-inducing prior such as the spike and slab prior or the horseshoe George and McCulloch [1993], Carvalho et al. [2009]. However, posteriors in these cases may be multimodal, and scalable Bayesian inference with such priors is a challenging, active area of research Guan and Stephens [2011], Yang et al. [2016], Johndrow et al. [2017]. To demonstrate the applicability of low-rank data approximations to this setting, we ran NUTS using `Stan` on a logistic regression model with a regularized horseshoe prior [Carvalho et al., 2009, Piironen and Vehtari, 2017]. In Figure B.1.7,

we see an attractive trade-off between computational investment and approximation error. For example, we obtained relative mean and standard deviation errors of only about 10^{-2} while reducing computation time by a factor of three.

We also applied LR-MCMC to linear regression with the regularized horseshoe prior on a dataset with very correlated covariates and $D = 6{,}238$. However, this sampler exhibited severe mixing problems, both with and without the approximation, as diagnosed by large $\hat{R}$ values in `pyStan`. These issues reflect the innate challenges of high-dimensional Bayesian inference with the horseshoe prior and correlated covariates.

Scalability to large-scale real datasets. Finally, we explored the applicability of LR-Laplace to two real, large-scale logistic regression tasks (Figure 3-3). The first is the UCI Farm-Ads dataset, which consists of $N = 4{,}143$ online advertisements for animal-related topics together with binary labels indicating whether the content provider approved of the ad; there are $D = 54{,}877$ bag-of-words features per ad [Dheeru and Karra Taniskidou, 2017]. As with the synthetic datasets, we evaluated the error in the approximations of posterior means and variances. As a baseline to evaluate this error, we use the usual Laplace approximation because the computational demands of MCMC preclude the possibility of using it as a baseline.

As a second real dataset we evaluated our approach on the Reuters RCV1 text categorization test collection Amini et al. [2009], Chang and Lin [2011]. RCV1 consists of $D = 47{,}236$ bag-of-words features for $N = 20{,}241$ English documents grouped into two different categories. We were unable to compare to the full Laplace approximation due to the high-dimensionality, so we used LR-Laplace with $M = 20{,}000$ as a baseline. For both datasets, we find that as we increase the rank of the data approximation, we incur longer running times but reduced errors in posterior means and variances. Laplace and Diagonal Laplace do not provide the same computation–accuracy trade-off.

Choosing M. Applying LR-GLM requires choosing the rank M of the low rank approximation. As we have shown, this choice characterizes a computational–statistical trade-off whereby larger M leads to linearly larger computational demands, but increases the precision of the approximation. As a practical rule of thumb, we recommend setting M to be as large as is allowable for the given application without

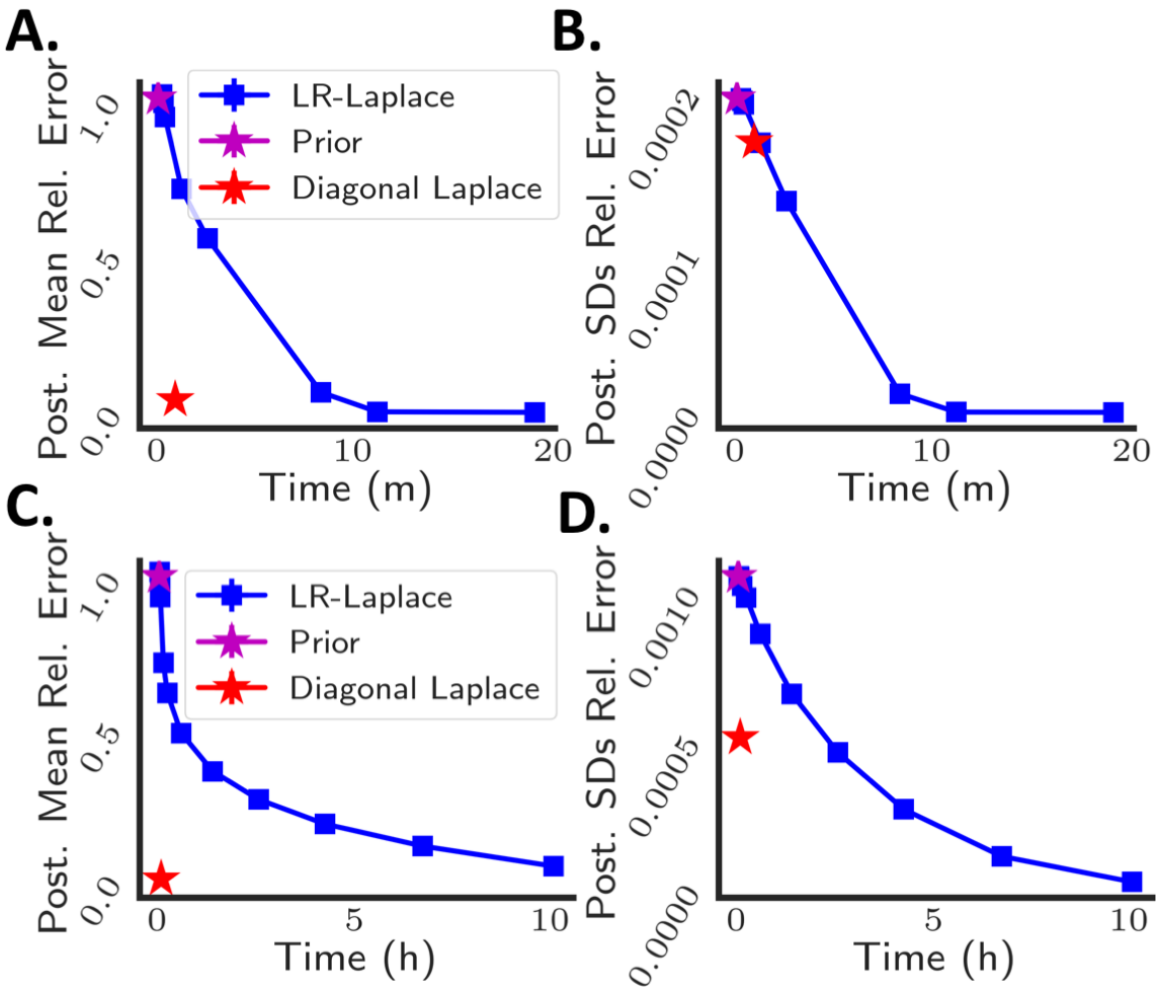


Figure 3-3: LR-Laplace approximation quality on Farm-Ads (top) and RCV-1 (bottom) datasets with varying M. (A.) Farm-Ads error in the posterior mean and (B.) Farm-Ads error in posterior variances (C.) RCV-1 error in posterior mean and (D.) RCV-1 error in posterior variances.

the resulting inference becoming too slow. For our experiments with LR-Laplace, this limit was $M \approx 20{,}000$. For LR-MCMC, the largest manageable choice of M will be problem dependent but will typically be much smaller than 20,000.

3.7 Conclusion

We have shown through theory and experiments that low-rank data approximations can enable efficient, high-quality approximate posterior inference in large scale generalized linear models. Our approximation is transparent; we provide interpretable error bounds

for conjugate Gaussian regression as well as for Bayesian GLMs. We demonstrate an attractive computational–statistical trade-off: increasing the rank of our approximation allows us to achieve higher quality approximations by investing more running time; moreover, we recover the exact likelihood in the limit of a full rank approximation. Lastly, we demonstrated that the error introduced by our approximation errs on the side of conservativeness; that is, we provide approximations that are never less uncertain than the exact posterior. This conservativeness applies to both parameters and predictions. We believe these properties of our low-rank data approximations make them a valuable and practical tool for approximate inference in large-scale Bayesian GLMs.

Chapter 4

Optimal Transport Couplings of Gibbs Samplers on Partitions for Unbiased Estimation

Abstract

Computational couplings of Markov chains provide a practical route to unbiased Monte Carlo estimation that can utilize parallel computation. However, these approaches depend crucially on chains meeting after a small number of transitions. For models that assign data into groups, e.g. mixture models, the obvious approaches to couple Gibbs samplers fail to meet quickly. This failure owes to the so-called 'label-switching' problem; semantically equivalent relabelings of the groups contribute well-separated posterior modes that impede fast mixing and cause large meeting times. We here demonstrate how to avoid label switching by considering chains as exploring the space of partitions rather than labelings. Using a metric on this space, we employ an optimal transport coupling of the Gibbs conditionals. This coupling outperforms alternative couplings that rely on labelings and, on a real dataset, provides estimates more precise than usual ergodic averages in the limited time regime. Code is available at github.com/tinnguyen96/coupling-Gibbs-partition.

4.1 Introduction

Couplings for unbiased Markov chain Monte Carlo. Consider estimating an analytically intractable expectation of a function h of a random variable X distributed according to p, $H^* \coloneqq \int h(X)p(X)dX$. Given a Markov chain $\{X_t\}_{t=0}^{\infty}$ with initial distribution $X_0 \sim p_0$ and evolving according to a transition kernel $X_t \sim T(X_{t-1}, \cdot)$ stationary with respect to p, one option is to approximate H^* with the empirical average of samples $\{h(X_t)\}$. However, while ergodic averages are asymptotically consistent, they are in general biased when computed from finite simulations. As such, one cannot effectively utilize parallelism to reduce error to any desired tolerance.

Computational couplings provide a route to unbiased estimation in finite simulation [Glynn and Rhee, 2014]; in this work we build on the framework of Jacob et al. [2020]. One designs an additional Markov chain $\{Y_t\}$ with two properties. First, $Y_t \,|\, Y_{t-1}$ also evolves using the transition $T(\cdot, \cdot)$, so that $\{Y_t\}$ is equal in distribution to $\{X_t\}$. Secondly, there exists a random *meeting time* $\tau < \infty$ such that the two chains meet exactly at some time τ, $X_\tau = Y_{\tau-1}$, and remain faithful afterwards: for all $t \geq \tau$, $X_t = Y_{t-1}$. Then, one can compute an unbiased estimate of H^* as

$$H_{\ell:m}(X, Y) \coloneqq \underbrace{\frac{1}{m-\ell+1}\sum_{t=\ell}^{m} h(X_t)}_{\text{Usual MCMC average}} + \underbrace{\sum_{t=\ell+1}^{\tau-1} \min\left(1, \frac{t-\ell}{m-\ell+1}\right)\left\{h(X_t) - h(Y_{t-1})\right\}}_{\text{Bias correction}} \tag{4.1}$$

where ℓ is the burn-in length, and m sets a minimum number of iterations [Jacob et al., 2020, Equation 2]. One interpretation of this estimator is as the usual MCMC estimate plus a bias correction. Since $H_{\ell:m}$ is unbiased, we can make the squared error (for estimating H^*) arbitrarily small by simply averaging many estimates computed in parallel. However, the practicality of Eq. (4.1) relies on a coupling that provides sufficiently small meeting times. Large meeting times are doubly problematic: they lead to greater computational cost and higher variance due to the additional terms.

Gibbs samplers over discrete structures and their couplings. Gibbs sampling is a standard inference method for models with discrete structures and tractable conditional distributions. Numerous applications include Bayesian nonparametric clustering using Dirichlet process mixture models [Antoniak, 1974, Neal, 2000], graph coloring for randomized approximation algorithms [Jerrum, 1998], community detection using stochastic block models [Holland et al., 1983, Geng et al., 2019] and computational redistricting [DeFord et al., 2021]. In these cases, the discrete structure is the partition of data into components.

While some earlier works have described couplings of Gibbs samplers, they have not sought to address computational approaches applicable in these settings. For example, Jerrum [1998] uses maximal couplings on labelings to prove convergence rates for graph coloring, and Gibbs [2004] uses a common random number coupling for two-state Ising models. Notably, these approaches rely on explicit labelings and, in our experiments, suffer from large meeting times. We attribute this issue to the label-switching problem [Jasra et al., 2005]; heuristically, many different labelings imply the same partition, and two chains may nearly agree on the partition but require many iterations to change label assignments.

Our contribution. We view the Gibbs sampler as exploring a state-space of partitions rather than labelings (as, for example, in Tosh and Dasgupta [2014]), and define an optimal transport (OT) coupling in this space. We show that our algorithm has a fast run time and empirically validate it in the context of Dirichlet process mixture models [Antoniak, 1974, Prabhakaran et al., 2016] and graph coloring [Jerrum, 1998], where it provides smaller meeting times than the label-based couplings of Jerrum [1998], Gibbs [2004]. We demonstrate the benefits of unbiasedness by reporting estimates of the posterior predictive density and cluster proportions. Our implementation is publicly available at github.com/tinnguyen96/coupling-Gibbs-partition.

4.2 Our Method

4.2.1 Gibbs samplers over partitions

For a natural number N, a *partition* of $[N] := \{1, 2, \ldots, N\}$ is a collection of non-empty disjoint sets $\{A_1, A_2, \ldots, A_k\}$, whose union is $[N]$ [Pitman, 2006, Section 1.2]. We use $\mathcal{P}_N$ to denote the set of all partitions of $[N]$. Throughout, we use π to denote elements of $\mathcal{P}_N$ and Π for a random partition (i.e. a $\mathcal{P}_N$-valued random variable) with probability mass function (PMF) p_Π. Finally π_{-n} and Π_{-n} denote these partitions with data-point n removed. For example, if $\pi = \big\{\{1, 3\}, \{2\}\big\}$, then $\pi_{-1} = \big\{\{3\}, \{2\}\big\}$.

Drawing direct Monte Carlo samples $\Pi \sim p_\Pi$ is often impossible. However, the conditional distributions $p_{\Pi|\Pi_{-n}}$ are supported on at most N partitions. Hence, when p_Π is available up to a proportionality constant, computing and sampling from $p_{\Pi|\Pi_{-n}}$ are tractable operations. A Gibbs sampler exploiting this tractability proceeds as follows. First, a partition π is drawn from an initial distribution p_0 on $\mathcal{P}_N$. For each iteration, we *sweep* through each data-point $n \in [N]$, temporarily remove it from π, and then randomly reassign it to one of the sets within π_{-n} or add it as singleton (that is, as a new group) according the conditional PMF $p_{\Pi|\Pi_{-n}}(\cdot|\pi_{-n})$.

4.2.2 Our approach: optimal coupling of Gibbs conditionals

Algorithm 5 Gibbs Sweep with Optimal Transport Coupling

Input: Target probability mass function (PMF) p_Π. Current partitions π and ν.
for $n = 1, 2, \ldots, N$ **do**
    // Compute Gibbs marginals (PMFs over partitions)
    $q, r \leftarrow p_{\Pi|\Pi_{-n}}(\cdot|\pi_{-n}), p_{\Pi|\Pi_{-n}}(\cdot|\nu_{-n})$

    // Compute and sample from optimal transport coupling
    $[\pi^1, \pi^2, \ldots, \pi^K], [\nu^1, \nu^2, \ldots, \nu^{K'}] \leftarrow \text{support}(q), \text{support}(r)$
    $\gamma^* = \arg\min_{\gamma \in \Gamma(q,r)} \sum_{k=1}^{K} \sum_{k'=1}^{K'} \gamma(\pi^k, \nu^{k'}) \mathrm{d}(\pi^k, \nu^{k'})$
    $\pi, \nu \sim \gamma^*$
Return π, ν

Our coupling encourages the chains to become 'closer' while maintaining the correct marginal evolution. To quantify closeness we use a metric on $\mathcal{P}_N$. While a number of metrics exist [Meilă, 2007, Section 2], for simplicity we chose a classical metric introduced by Mirkin and Chernyi [1970], Rand [1971],

$$\mathrm{d}(\pi, \nu) = \sum_{A \in \pi} |A|^2 + \sum_{B \in \nu} |B|^2 - 2 \sum_{A \in \pi, B \in \nu} |A \cap B|^2, \tag{4.2}$$

which is equivalent to Hamming distance on the adjacency matrices implied by partitions [Mirkin and Chernyi, 1970, Theorems 2-3]. We leave investigation of the impact of metric choice on meeting time distribution to future work.

With the metric in Eq. (4.2), we can formalize an *optimal transport coupling* of two Gibbs conditionals, i.e. the coupling that minimizes the expected distances between the updates. In particular, we let $q := p_{\Pi|\Pi_{-n}}(\cdot|\pi_{-n})$ and $r := p_{\Pi|\Pi_{-n}}(\cdot|\nu_{-n})$ with supports $[\pi^1, \pi^2, \ldots, \pi^K] := \mathrm{support}(q)$ and $[\nu^1, \nu^2, \ldots, \nu^{K'}] := \mathrm{support}(r)$ and define the OT coupling as

$$\gamma^* := \underset{\gamma \in \Gamma(q,r)}{\arg\min} \sum_{k=1}^{K} \sum_{k'=1}^{K'} \gamma(\pi^k, \nu^{k'}) \mathrm{d}(\pi^k, \nu^{k'}), \tag{4.3}$$

where $\Gamma(q, r)$ is the set of all couplings of q and r. Algorithm 5 summarizes this approach.

4.2.3 Efficient computation of optimal couplings

The practicality of our OT coupling depends both on successfully encouraging chains to meet in a small number of steps and on an implementation with computational cost comparable to running single chains. If Algorithm 5 required orders of magnitude more time than the Gibbs sweep of single chains, the extent of parallelism required to place the unbiased estimates from coupled chains on an even footing with standard MCMC could be prohibitive.

In many applications, including those in our experiments, for partitions of size K, the Gibbs conditionals may be computed in $\Theta(K)$ time, and a full sweep through

the N data-points takes $\Theta(NK)$ time for a single chain. At first consideration, an implementation of Algorithm 5 with comparable efficiency might seem infeasible. In particular, when π and ν are of size $O(K)$, Eq. (4.3) requires computing $O(K^2)$ pairwise distances, each of which naively might seem to require at least $O(KN)$ operations — let alone the OT problem.

The following result shows that we can in fact compute this coupling efficiently.

Theorem 4.2.1 (Gibbs Sweep Time Complexity)**.** *Let p_Π be the law of a random N-partition. If for any $\pi \in \mathcal{P}_N, p_{\Pi|\Pi_{-n}}(\cdot|\pi_{-n})$ is computed in constant time, the Gibbs sweep in Algorithm 5 has $O(N\widetilde{K}^3 \log \widetilde{K})$ run time, where $\widetilde{K}$ is the max partition size encountered.*

As a proof of Theorem 4.2.1, we detail an $O(N\widetilde{K}^3 \log \widetilde{K})$ implementation in Appendix C.1.

Theorem 4.2.1 guarantees that the run time of a coupled-sweep is no more than a $O(\widetilde{K}^2 \log \widetilde{K})$ factor slower than a single-sweep. The relative magnitude of $\widetilde{K}$ versus N depends on the target distribution. For the graph coloring distribution, $\widetilde{K}$ is upper bounded by the numbers of available colors. Under the Dirichlet process mixture model (DPMM) prior, with high probability, the size of partition of N data points is within multiplicative factors of $\ln N$ [Arratia et al., 2003, Section 5.2]. We *conjecture* that under most initializations of the Gibbs sampler (such as from the DPMM prior), $\widetilde{K} = O(\ln N)$ with high probability.

Remark 4.2.2. The worst-case run time of Theorem 4.2.1 is attained with Orlin's algorithm [Orlin, 1993] to solve Eq. (4.3) in $O(\widetilde{K}^3 \log \widetilde{K})$ time. However, our implementation uses the simpler network simplex algorithm [Kelly and O'Neill, 1991] as implemented by Flamary et al. [2021]. Although Kelly and O'Neill [1991, Section 3.6] upper bound the worst-case complexity of the network simplex as $O(\widetilde{K}^5)$, the algorithm's average-case performance may be as good as $O(\widetilde{K}^2)$ [Bonneel et al., 2011, Figure 6].

Although Orlin's algorithm [Orlin, 1993] has a better worst-case runtime, convenient public implementations are not available. In addition, our main contribution is the

formulation of the coupling as an OT problem — in principle, the dependence on $\widetilde{K}$ of the runtime in Theorem 4.2.1 inherits from the best OT solver used.

4.3 Empirical Results

In Section 4.3.2, we compare the distribution of meeting times between our partition-based coupling and two label-based couplings: under our coupling, chains meet earlier. In Section 4.3.3, we report unbiased estimates of two estimands of common interest: posterior predictive densities and the posterior mean proportion of data assigned to the largest clusters. But first, we describe the applications and the target distributions under consideration in Section 4.3.1.

4.3.1 Applications

Dirichlet process mixture models. Clustering is a core task for understanding structure in data and density estimation. When the number of latent clusters is a priori unknown, DPMMs [Antoniak, 1974] are a useful tool. The cluster assignments of data points in a DPMM can be described with a Chinese restaurant process, or $\mathrm{CRP}(\alpha, N)$, which is a probability distribution over $\mathcal{P}_N$ with mass $\Pr(\Pi = \pi) = \frac{\alpha^K \prod_{A\in\pi}(|A|-1)!}{\alpha(\alpha+1)\ldots(\alpha+N-1)}$ where K is the number of clusters in π, and $\prod_{A\in\pi}$ iterates through the clusters. We consider a fully conjugate DPMM [MacEachern, 1994],

$$\Pi \sim \mathrm{CRP}(\alpha, N), \quad \mu_A \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \text{ for } A \in \Pi, \quad \mathcal{D}_j \,|\, \mu_A \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_A, \Sigma_1) \text{ for } j \in A. \tag{4.4}$$

The hyper-parameters of Eq. (4.4) are concentration α, cluster prior mean μ_0, observational covariance Σ_1 and cluster covariance Σ_0. For this application, the distribution is the Bayesian posterior, $p_\Pi(\pi) := \Pr(\Pi = \pi \,|\, \mathcal{D})$. The Gibbs conditionals of the posterior $p_{\Pi|\Pi_{-n}}$ can be computed in closed form, using simple formulas for conditioning of jointly Gaussian random variables and the well-known Polya urn scheme [Neal, 2000, Equation 3.7].

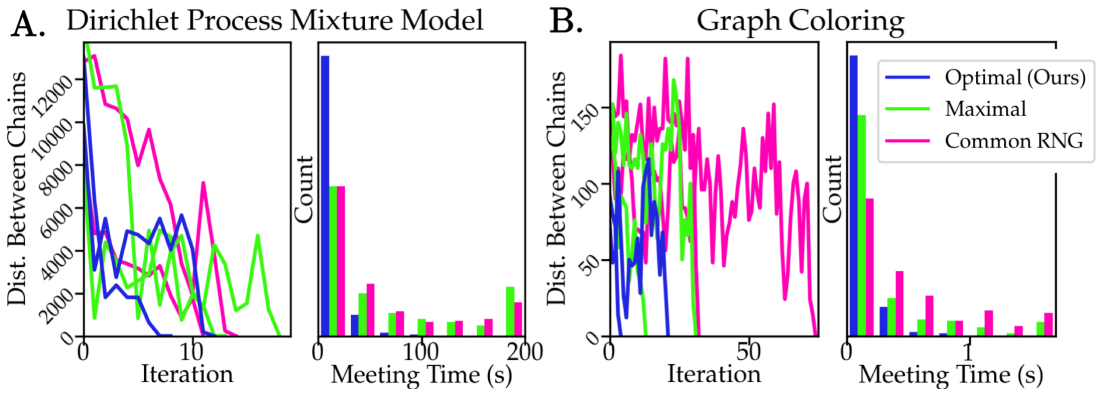


Figure 4-1: Reduced meeting times are achieved by OT couplings of Gibbs conditionals relative to maximal and common random number couplings in applications to (A) DPMM and (B) graph coloring. (A) Left and (B) left show two representative traces of the distance between coupled chains by iteration. (A) Right and (B) right show histograms of meeting times 250 replicate coupled chains.

Graph coloring. Uniform sampling of graph colorings is a problem of fundamental interest in theoretical computer science for its role as a subroutine within fully polynomial randomized approximation algorithms, where samples from the uniform distribution on graph colorings are used to estimate the number of unique colorings [Jerrum, 1998].

Notably, this sampling problem reduces to sampling from the induced distribution on partitions, by choosing an ordering of the sets in the partition and associating it with a random permutation of the set of colors. Accordingly, estimates are just as easily constructed for a Markov chain defined on partitions. See Appendix C.2 for additional details.

4.3.2 Reduced meeting times with OT couplings

Figure 4-1 demonstrates that our approach yields faster couplings than the classical maximal coupling approach [Jerrum, 1998, Section 5], or an analogous coupling using shared common random numbers (see e.g. Gibbs [2004]). In applications to both Bayesian clustering and graph coloring, the distance between coupled chains stochastically decreases to 0 (Figure 4-1 left panels), with our approach leading to meetings after fewer sweeps. Despite the larger per-sweep computational cost, our

OT coupled chains typically meet after a shorter wall-clock time as well. We suspect this improvement comes from avoiding label-switching, which hinders mixing of the maximal and common-RNG coupled chains.

The tightest bounds for mixing time for Gibbs samplers on graph colorings to date [Chen et al., 2019] rely on couplings on labeled representations. Our results suggest better bounds may be attainable by considering convergence of partitions rather than labelings. Reducing the mixing time for Gibbs samplers of DPMM has been a motivation behind collapsed samplers [MacEachern, 1994], but the literature lacks upper bounds on the mixing time.

4.3.3 Unbiased estimation with parallel computation

We adapt the setup from Jacob et al. [2020, Section 3.3]. Fixing a time budget, we run a single chain until time runs out and report the ergodic average. For coupled chains, we attempt as many meetings as possible in this time, and report the average across attempts.

Posterior mean predictive density. The posterior predictive is a key quantity used in model selection [Görür and Rasmussen, 2010], and is of particular interest for DPMMs as it is known to be consistent for the underlying data distribution in total variation distance [Ghosal et al., 1999]. As a proof of concept, we computed unbiased estimates of the posterior predictive distribution of a DPMM (Figure 4-2 A).

We generated $N = 100$ data points from a 10-component Gaussian mixture model in one dimension, with the variance around cluster means equal to 4. We used a DPMM with $\alpha = 1$, $\mu_0 = 0$, $\Sigma_1 = 4.0$, $\Sigma_0 = 9.0$ to analyze the N observations. The solid blue curve is an unbiased estimate of the posterior predictive density. The black dashed curve is the true density of the population. The grey histogram bins the observed data. Because of the finite sample size, the predictive density is not equal to the true density. In Appendix C.3, the difference between the model's predictive density and the true density decreases as sample size N increases.

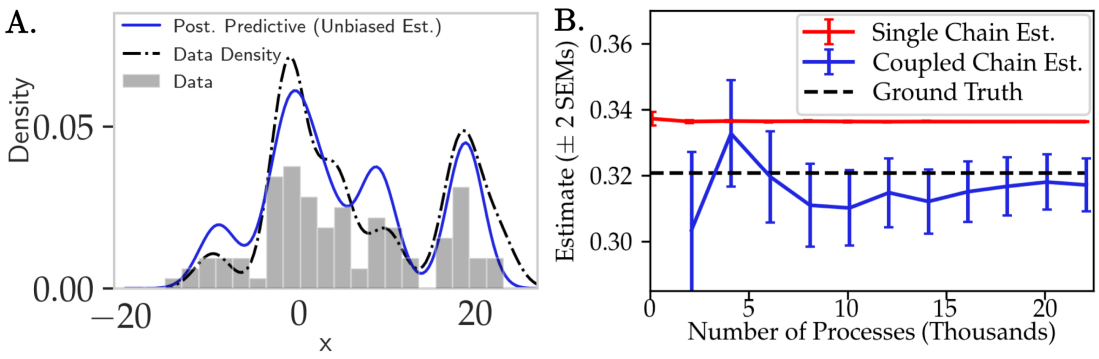


Figure 4-2: Unbiased estimates for Dirichlet process mixture model are obtained using OT coupled chains. (A) Unbiased estimate of the posterior predictive density for a toy problem. (B) Parallelism/accuracy trade-off for single and coupled chain estimators of the posterior mean portion of cells in the largest cluster. Each process is allocated 250 seconds, error bars indicate ±2SEM. Ground truth denotes estimates from very long MCMC chains.

Posterior mean component proportions. A second key quantity of interest in DPMMs is the posterior mean of the proportion of data-points in the largest cluster(s) (e.g. as reported by Liverani et al. [2015]). We lastly explored parallel computation for unbiased estimation of this quantity on a real dataset (Figure 4-2 B). Specifically, we use a subset of the data used by Prabhakaran et al. [2016], who used a DPMM to analyse single-cell RNA-sequencing data obtained from Zeisel et al. [2015] (see Appendix C.2 for details).

Figure 4-2 B presents a series of estimates of the proportion of cells in the largest component, and approximate frequentist confidence intervals. For each number of processes M, we aggregated M independent single and coupled chain estimates, each from a single processor with a 250 second limit. We compare to the 'ground-truth' proportion obtained by MCMC run for 10,000 sweeps. Our results demonstrate the advantage of unbiased estimates in the high-parallelism, time-limited regime; while single-chain estimates have lower variance, coupled chains yield smaller error when aggregated across many processes. In addition, as result of unbiasedness, standard frequentist intervals may be expected to have good coverage. By contrast, we cannot expect such intervals from single chains to be calibrated; indeed, the true value is many standard errors from the single chain estimates (Figure 4-2 B).

However, due to the variance of the unbiased estimates we require a degree of parallelism that may be impractical for most practitioners ($\approx$ 5,000 pairs of chains to attain error comparable to that of as many single chains). Indeed, in our experiments, we simulated this high parallelism by sequentially running batches of 100 processes in parallel. Additionally, the estimation strategy can be finicky: unbiasedness requires coupled chains to meet exactly, and for some models and experiments not shown, we found that some pairs of coupled chains failed to meet quickly. This difficulty is expected for problems where single chains mix slowly, as slow mixing precludes the existence of fast couplings [Jacob, 2020, Chapter 3]. Looking forward, we expect that our work will naturally benefit from advances in parallel-computation software and hardware, such as GPU implementations. Reducing the variance of the unbiased estimates is an open question, and is the target of ongoing work.

Chapter 5

Confidently Comparing Estimators with the c-value

Abstract

Modern statistics provides an ever-expanding toolkit for estimating unknown parameters. Consequently, applied statisticians frequently face a difficult decision: retain a parameter estimate from a familiar method or replace it with an estimate from a newer or more complex one. While it is traditional to compare estimators using risk, such comparisons are rarely conclusive in realistic settings.

In response, we propose the "c-value" as a measure of confidence that a new estimate achieves smaller loss than an old estimate on a given dataset. We show that it is unlikely that a large c-value coincides with a larger loss for the new estimate. Therefore, just as a small p-value provides evidence to reject a null hypothesis, a large c-value provides evidence to use a new estimate in place of the old. For a wide class of problems and estimators, we show how to compute a c-value by first constructing a data-dependent high-probability lower bound on the difference in loss. The c-value is frequentist in nature, but we show that it can provide a validation of shrinkage estimates derived from Bayesian models in real data applications involving hierarchical models and Gaussian processes.

5.1 Introduction

Modern statistics provides an expansive toolkit of sophisticated methodology for estimating unknown parameters. However, the abundance of different estimators often presents practitioners with a difficult challenge: choosing between the output of a familiar method (e.g. a maximum likelihood estimate (MLE)) and that of a more complicated method (e.g. the posterior mean of a hierarchical Bayesian model). From a practical perspective, abandoning a familiar approach in favor of a newer alternative is unreasonable without some assurance that the latter provides a more accurate estimate. Our goal is to determine whether it is safe to abandon a default estimate in favor of an alternative, and to provide an assessment of the degree of confidence we should have in this decision.

Traditionally decisions between estimators are based on risk, the loss averaged over all possible realizations of the data with respect to a likelihood model [Lehmann and Casella, 2006, Chapters 4-5]. We note two limitations of using risk. First, it is rare that one estimator within a given pair will have smaller risk across all possible parameter values. Instead, it is more often the case that one estimator will have smaller risk for some unknown parameter values but larger risk for other parameter values. Second, one estimator may have lower risk than another but incur higher loss on a majority of datasets; see Appendix D.2 for an example in which an estimator with smaller risk has larger loss on nearly 70% of simulated datasets.

In this work we propose a framework for choosing between estimators based on their performance *on the observed dataset* rather than their average performance. Specifically, we introduce the "c-value" ("c" for confidence in the new estimate), which we construct using a data-dependent high-probability lower bound on the difference in loss. We show that it is unlikely that simultaneously the c-value is large and the alternative estimate has larger loss than the default. We then demonstrate how to use the c-value to select between two estimates in a principled, data-driven way. Critically, the c-value requires no assumptions on the unknown parameter; our guarantees hold uniformly across the parameter space. Before presenting our general methodology, we

discuss three motivating examples.

5.1.1 Shrinkage estimates on educational testing data

We revisit Hoff [2021]'s estimates of average student reading ability at several schools in the 2002 Educational Longitudinal Study. These estimates are obtained from a hierarchical Bayesian model intended to "share strength" by partially pooling data across related schools. However, the analysis relied on a simplifying and subjectively chosen prior, and so it was unclear whether the resulting estimates of school-specific parameters were more accurate simpler MLE obtained by considering each school separately. Our analysis is quite confident that Hoff [2021]'s estimates yield smaller square error than the MLE. We additionally consider a more clearly misspecified prior and verify that our methodology does not always favor more complex alternate estimators. Although these estimates have a Bayesian provenance, the use of the c-value to validate them requires neither subjective belief in the prior nor the assumption that it is correctly specified.

5.1.2 Estimating violent crime density at the neighborhood level

Considerable empirical evidence links a community's exposure to violent crime and adverse behavioral, mental, and physical health outcomes among its residents [Buka et al., 2001, Kondo et al., 2018]. Although overall violent crimes rates in the U.S. have decreased over the last two decades, there is considerable variation in time trends at the neighborhood level [Balocchi and Jensen, 2019, Balocchi et al., 2019]. A critical first step in understanding what drives neighborhood-level variation is accurate estimation of the actual amount of violent crime that occurs in each neighborhood.

Typically, researchers rely on the reported counts of violent crime aggregated at small spatial resolutions (e.g. at the census tract level). However, in light of sampling variability due to the relative infrequency of certain crime types in small areas, it is natural to wonder if auxiliary data can be used to improve estimates of violent crime

incidence.

As a second application of our framework, we analyze the number of violent crimes reported per square mile in several neighborhoods in the city of Philadelphia. Our analysis suggests that one can obtain improved estimates of this violent crime density by using a shrinkage estimate that incorporates information about non-violent crime incidence. Further c-value analysis reveals that leveraging spatial information on top of non-violent incidence does not provide additional improvement.

5.1.3 Gaussian process kernel choice: modeling ocean currents

Accurate estimation of ocean current dynamics is critical for forecasting the dispersion of oceanic contaminations [Poje et al., 2014]. While it is commonplace to model ocean flow dynamics at or above the *mesoscale* (roughly 10 km), Lodise et al. [2020] have recently advocated modeling dynamics at both the mesoscale and the *submesoscale* (roughly 0.1–10 km). They specifically proposed a Gaussian process model that accounts for variation across multiple resolutions to estimate ocean currents from positional data taken from hundreds of free-floating buoys.

In a third application of our framework, we find that the multi-resolution procedure produces a large c-value, indicating that accounting for variation across multiple scales enables more accurate estimates than are obtained when accounting only for mesoscale variation.

5.1.4 Organization of the article & contributions

We formally present our general framework and define the c-value in Section 5.2. In Section 5.2.1 we highlight similarities and differences between our framework and existing work on preliminary testing and post-selection inference. Our approach to computing c-values depends on the availability of high-confidence lower bounds on the difference in the losses of the two estimates that holds uniformly across the parameter space. Sections 5.3 to 5.5 provide these bounds for several models and classes of estimators for squared error loss. In Section 5.3, we illustrate our general strategy in

the canonical normal means problem. Then, in Section 5.4, we generalize this strategy to compare affine estimates of normal means with correlated observations. Section 5.5 shows how to extend the framework to cover two nonlinear cases: a nonlinear shrinkage estimator and regularized logistic regression. We provide simulations validating our approach in these settings. We apply our framework to the aforementioned motivating examples in Section 5.6.

5.2 Introducing the c-value

We now describe our approach for quantifying confidence in the statement that one estimate of an unknown parameter is superior to another. We begin by introducing some notation and building up to a definition of the c-value, before stating our main results. This development is very general, and we defer practical considerations to the subsequent sections. We include proofs of the results of this section in Appendix D.1.

Suppose that we observe data y drawn from some distribution that depends on an unknown parameter θ. We consider deciding between two estimates, $\hat{\theta}(y)$ and $\theta^*(y)$, of θ on the basis of a loss function $L(\theta, \cdot)$. Our focus is on asymmetric situations in which $\hat{\theta}(\cdot)$ is a standard or more familiar estimator while $\theta^*(\cdot)$ is a less familiar estimator. For simplicity, we will refer to $\hat{\theta}(\cdot)$ as the default estimator and $\theta^*(\cdot)$ as the alternative estimator.

We next define the "win" obtained by using $\theta^*(y)$ rather than $\hat{\theta}(y)$ as the difference in loss, $W(\theta, y) := L(\theta, \hat{\theta}(y)) - L(\theta, \theta^*(y))$. While a typical comparison based on risk would proceed by taking the expectation of $W(\theta, y)$ over all possible datasets drawn for fixed θ, we maintain focus on the single observed dataset. Notably, the win is positive whenever the alternative estimate achieves a smaller loss than the default estimate. As such, if we knew that $W(\theta, y) > 0$ for the given dataset y and unknown parameter θ, then we would prefer to use the alternative $\theta^*(y)$ instead of the default $\hat{\theta}(y)$.

Since θ is unknown, determining whether $W(\theta, y) > 0$ is impossible. Nevertheless, for a broad class of estimators, we can determine whether the win is positive with

high probability. To start, we construct a lower bound, $b(y, \alpha)$, depending only on the data and a pre-specified level $\alpha \in [0, 1]$, that satisfies for all θ

$$\mathbb{P}_\theta \left[W(\theta, y) \geq b(y, \alpha) \right] \geq \alpha. \tag{5.1}$$

For values of α close to 1, $b(y, \alpha)$ is a high-probability lower bound on the win that holds uniformly across all possible values of the unknown parameter θ. Loosely speaking, if $b(y, \alpha) > 0$ for some α close to 1, then we can be confident that the alternative estimate has smaller loss than the default estimate.

The lower bound $b(y, \alpha)$ allows us to define a precise measure of confidence that $\theta^*(y)$ is superior to $\hat{\theta}(y)$, that we call the c-value,

$$c(y) := \inf_{\alpha \in [0,1]} \left\{ \alpha \mid b(y, \alpha) \leq 0 \right\}. \tag{5.2}$$

The c-value marks a meaningful boundary in the space of confidence levels; it is the largest value such that for every $\alpha < c(y)$, we have confidence α that the win is positive.

Remark 5.2.1. An alternative definition for the c-value is $c^+(y) = \sup_{\alpha \in [0,1]} \{\alpha | b(y, \alpha) \geq 0\}$. Although $c^+(y) = c(y)$ when $b(y, \cdot)$ is continuous and strictly decreasing in α, $c^+(\cdot)$ may be overconfident otherwise. We detail a particularly pathological example in Appendix D.3.

Our first main result formalizes the interpretation of $c(y)$ as a measure of confidence.

Theorem 5.2.2. *Let $b(\cdot, \cdot)$ be any function satisfying the condition in Eq.* (5.1)*. Then for any θ and $\alpha \in [0, 1]$ and $c(y)$ as defined in Eq.* (5.2)*,*

$$\mathbb{P}_\theta \left[W(\theta, y) \leq 0 \text{ and } c(y) > \alpha \right] \leq 1 - \alpha. \tag{5.3}$$

The result follows directly from the definition of $c(\cdot)$ and the condition on $b(\cdot, \cdot)$. Informally, Theorem 5.2.2 assures us that it is unlikely that simultaneously (A) the

c-value is large and (B) $\theta^*(y)$ does not provide smaller loss than $\hat{\theta}(y)$. Just as a small p-value provides evidence to reject a null hypothesis, a large c-value provides evidence to abandon the default estimate in favor of the alternative.

The strategy described above necessarily uses the data twice, once to compute the two estimates and once more to compute the c-value to choose between them. Accordingly, one might justly ask if this "double-dipping" into the dataset is likely to damage the quality of the resulting estimate. To address this question, we formalize this two-step procedure with a single estimator

$$\theta^\dagger(y, \alpha) := \mathbb{1}[c(y) \leq \alpha]\hat{\theta}(y) + \mathbb{1}[c(y) > \alpha]\theta^*(y), \tag{5.4}$$

which picks between the two estimates $\hat{\theta}(y)$ and $\theta^*(y)$ based on the value $c(y)$ and a pre-specified level $\alpha \in [0, 1]$. We can characterize the possible outcomes when using $\theta^\dagger(\cdot, \alpha)$ with a contingency table (Table 5.1), where rows correspond to the estimate with smaller loss, and the columns correspond to the reported estimate.

Table 5.1: Contingency tables with possible outcomes when using the two-staged estimator $\theta^\dagger(\cdot, \alpha)$. By construction, $\theta^\dagger(\cdot, \alpha)$ controls the probability of the shaded event.

	Default reported	Alternative reported
Default has lower loss	Correctly	Incorrect
Alternative has lower loss	Incorrect	Correct

Recalling again that we are interested in an asymmetric situation, we focus on the upper right entry. This entry corresponds to the event that $\theta^\dagger(\cdot, \alpha)$ incurs greater loss than $\hat{\theta}(\cdot)$. Our second main result formalizes that when we use $\theta^\dagger(\cdot, \alpha)$ with α close to 1, the probability of this event is small.

Theorem 5.2.3. *Let $b(\cdot, \cdot)$ be any function that satisfies the condition in Eq. (5.1). Then for any θ and $\alpha \in [0, 1]$,*

$$\mathbb{P}_\theta \left[L\left(\theta, \theta^\dagger(y, \alpha)\right) > L\left(\theta, \hat{\theta}(y)\right) \right] \leq 1 - \alpha. \tag{5.5}$$

When the alternative estimator is less familiar than the default estimator, such reassurance is highly desirable.

Overview of the remainder of the paper. The c-value is useful insofar as the lower bound $b(y, \alpha)$ is sufficiently tight and readily computable. It remains to show that such practical bounds exist. A primary contribution of this work is the explicit construction of these bounds in settings of practical interest. In what follows, we (A) illustrate one approach for constructing and computing $b(y, \alpha)$, (B) explore our proposed bounds' empirical properties on simulated data, and (C) demonstrate their practical utility on real-world data.

5.2.1 Related work

Hypothesis testing, *p*-values, and pre-test estimation. Our proposed c-value bears a resemblance to the p-value in hypothesis testing, but with a few key differences. Indeed, just as a small p-value can provide support to reject a simple null hypothesis in favor of a more complex alternative, a large c-value can provide support for a rejecting a familiar default estimate in favor of a more unfamiliar alternative. Furthermore both tools provide a frequentist notion of confidence based on the idea of repeated sampling. From this perspective, the two-step estimator $\theta^\dagger(\cdot, \alpha)$ resembles a preliminary testing estimator. Preliminary testing links the choice between estimators to the outcome of a hypothesis test for the null hypothesis that θ lies in some pre-specified subspace [Wallace, 1977].

The similarities to hypothesis testing go only so far. Notably, we consider decisions made about a *random* quantity, $W(\theta, y)$. Hypothesis tests, in contrast, concern only fixed statements about parameters, with nulls and alternatives corresponding to disjoint subsets of an underlying parameter space [Casella and Berger, 2002, Definition 8.1.3]. Our approach does not admit an interpretation as testing a fixed hypothesis.

Nevertheless, the connection to p-values can help us understand some limitations of the c-value. First, just as hypothesis tests may incur Type II errors (i.e. failures to reject a false null), for certain models and estimators there may be no $b(\cdot, \cdot)$ that

consistently detects improvements by the alternative estimate. Accordingly, the two stage estimator $\theta^\dagger(\cdot, \alpha)$ does not control the probability that we report the default estimate when the alternative in fact has smaller loss. In such situations, our approach may consistently report the default estimate even though it has larger loss. Second, even if good choices of $b(\cdot, \cdot)$ exist, it could be challenging to derive them analytically. This analytical challenge is reminiscent of difficulties for hypothesis testing in many models, wherein conservative p-values that are stochastically larger than uniform under the null are used when analytic quantile functions are unavailable. Third, we note that it may be tempting to interpret c-values as the conditional probability that the alternative estimate is superior to the default; however, just as it is incorrect to interpret a p-value as a probability that the null hypothesis is true, such an interpretation for a c-value is also incorrect.

Post-selection inference. In recent years, there has been considerable progress on understanding the behavior of inferential procedures that, like $\theta^\dagger(\cdot, \alpha)$, use the data twice, first to select amongst different models and then again to fit the selected model. Important recent work has focused on computing p-values and confidence intervals for linear regression parameters that are valid after selection with the lasso [Lockhart et al., 2014, Lee et al., 2016, Taylor and Tibshirani, 2018] and arbitrary selection procedures [Berk et al., 2013]. Somewhat more closely related to our focus on estimation are Tibshirani and Rosset [2019] and Tian [2020], which both bound prediction error after model selection. Unlike these papers, which study the effects of selection on downstream inference, we effectively perform inference on the selection itself.

5.3 Special case: c-values for estimating normal means

In this section, we derive a bound $b(y, \alpha)$ and compute the c-value a certain class of shrinkage estimators to maximum likelihood estimates (MLE) of the mean of a multivariate normal from a single vector observation (i.e. the normal means problem).

Our goal is to illustrate a simple instance of a general strategy for lower bounding the win that we will later generalize to more complex estimators and models. In Section 5.3.1, we define the model and the estimators that we consider. In Section 5.3.2, we introduce our lower bound $b(\cdot, \cdot)$ and present a theorem that guarantees this bound satisfies Eq. (5.1). Then, in Section 5.3.3, we examine the resulting c-value empirically and study the performance of the estimator $\theta^\dagger(\cdot, \alpha)$ that chooses between the default and alternative estimators based on the c-value (Eq. (5.4)). Several details, including the proof of Theorem 5.3.1, are left to Appendix D.5.

5.3.1 Normal means: notation and estimates

Let $\theta \in \mathbb{R}^N$ be an unknown vector and consider estimating θ from a noisy vector observation $y = \theta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, I_N)$ under squared error loss $L(\theta, \hat{\theta}) := \|\hat{\theta} - \theta\|^2$. For simplicity, we focus on the case of isotropic noise with variance one; we remove this restriction in Section 5.4. For our demonstration, we take the MLE $\hat{\theta}(y) = y$ to be the default estimate. As the alternative estimator, we consider a shrinkage estimator that was first studied extensively by Lindley and Smith [1972],

$$\theta^*(y) = \frac{y + \tau^{-2}\bar{y}\mathbf{1}_N}{1 + \tau^{-2}}$$

where $\mathbf{1}_N$ is the vector of all ones, $\tau > 0$ is a fixed positive constant, and $\bar{y} := N^{-1}\mathbf{1}_N^\top y$ is the mean of the observed y_n's. Operationally, $\theta^*(y)$ shrinks each coordinate of the MLE towards the grand mean $\bar{y}$.

5.3.2 Construction of the lower bound

To lower bound the win, we first rewrite $\theta^*(y) = \hat{\theta}(y) - Gy$ where $G := (1 + \tau^2)^{-1}P_1^\perp$ and $P_1^\perp = I_N - N^{-1}\mathbf{1}_N\mathbf{1}_N^\top$ is the projection onto the subspace orthogonal to $\mathbf{1}_N$. The win in squared error loss may then be written as

$$W(\theta, y) := \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2 = 2\epsilon^\top Gy - \|Gy\|^2. \tag{5.6}$$

Observe that we can compute $\|Gy\|$ directly from our data. As a result, in order to lower bound the win $W(\theta, y)$, it suffices to lower bound $2\epsilon^\top Gy$. As we detail in Appendix D.5.1, $2\epsilon^\top Gy$ follows a scaled and shifted non-central chi-squared distribution,

$$2\epsilon^\top Gy \sim \frac{2}{1+\tau^2}\left[\chi^2_{N-1}(\frac{1}{4}\|P_1^\perp\theta\|^2) - \frac{1}{4}\|P_1^\perp\theta\|^2\right],$$

where $\chi^2_{N-1}(\lambda)$ denotes the non-central chi-squared distribution with $N-1$ degrees of freedom and non-centrality parameter λ. Thus for any $\alpha \in (0,1)$ and any fixed value of $\|P_1^\perp\theta\|^2$,

$$W(\theta, y) \geq \frac{2}{1+\tau^2}F^{-1}_{N-1}(1-\alpha; \frac{1}{4}\|P_1^\perp\theta\|^2) - \frac{\|P_1^\perp\theta\|^2}{2(1+\tau^2)} - \|Gy\|^2 \tag{5.7}$$

with probability α, where $F^{-1}_{N-1}(1-\alpha;\lambda)$ denotes the inverse cumulative distribution function of $\chi^2_{N-1}(\lambda)$ evaluated at $1-\alpha$. Were $\|P_1^\perp\theta\|^2$ known, the right hand side of Eq. (5.7) would immediately provide a valid bound. However since $\|P_1^\perp\theta\|^2$ is not typically known, we use the data to address our uncertainty in this quantity. We obtain our bound by forming a one-sided confidence interval for $\|P_1^\perp\theta\|^2$ that holds simultaneously with Eq. (5.7).

Bound 5.3.1 (Normal means: Lindley and Smith estimate v.s. MLE). Observe $y = \theta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, I_N)$ and consider $\hat{\theta}(y) = y$ vs. $\theta^*(y) = (y + \tau^{-2}\bar{y}\mathbf{1}_N)/(1+\tau^{-2})$. We propose

$$b(y,\alpha) := \inf_{\lambda \in [0, U(y, \frac{1-\alpha}{2})]} \left\{ \frac{2}{1+\tau^2}F^{-1}_{N-1}\left(\frac{1-\alpha}{2}; \frac{\lambda}{4}\right) - \frac{\lambda}{2(1+\tau^2)} - \frac{\|P_1^\perp y\|^2}{(1+\tau^2)^2} \right\} \tag{5.8}$$

as an α-confidence lower bound on the win, where

$$U\left(y, \frac{1-\alpha}{2}\right) := \inf_{\delta > 0}\left\{ \delta \middle| \|P_1^\perp y\|^2 \leq F^{-1}_{N-1}\left(\frac{1-\alpha}{2}; \delta\right)\right\} \tag{5.9}$$

is a high-confidence upper bound on $\|P_1^\perp\theta\|^2$.

Bound 5.3.1 relies on a high-confidence upper bound on $\|P_1^\perp\theta\|^2$, but a two-sided interval could in principle provide a valid bound as well. In Appendix D.5.3 we provide

an intuitive justification for the choice of an upper bound. Theorem 5.3.1 justifies the use of Bound 5.3.1 for computing c-values.

Theorem 5.3.1. *Define* $c(y) := \inf_{\alpha\in[0,1]}\{\alpha | b(y,\alpha) \leq 0\}$ *for* $b(\cdot,\cdot)$ *in Bound 5.3.1. Then* $c(y)$ *is a valid c-value, satisfying the guarantees of Theorems 5.2.2 and 5.2.3.*

Remark 5.3.2 (Computability of the bound). Eq. (5.8) in Bound 5.3.1 can be readily computed. Notably, many standard statistical software packages provide numerical approximation to non-central χ^2 quantiles. Further, the one-dimensional optimization problems in Eqs. (5.8) and (5.9) can be solved numerically.

Remark 5.3.3 (When the variance is unknown). For cases when the noise variance σ^2 is unknown but a confidence interval is available, one can adapt the procedure above by replacing $b(y,\alpha)$ with its infimum with respect to σ^2 over the confidence interval and reducing the confidence level α accordingly.

Remark 5.3.4. The alternative estimator $\theta^*(y)$ considered in this section is the posterior mean of θ corresponding to the hierarchical prior $\theta|\mu \sim \mathcal{N}(\mu\mathbf{1}_N, \tau^2 I_N)$ with further improper hyper-prior on μ. This prior encodes a belief that θ lies close to the one-dimensional subspace spanned by $\mathbf{1}_N$. Using a similar approach to the one above, we can derive lower bounds on the win for a more general class of estimators that shrink the MLE towards a pre-specified D-dimensional subspace. See Appendix D.5.4 for details and an application to a real dataset on which a large computed c-value indicates an improved estimate.

5.3.3 Empirical verification

To explore the empirical properties of Bound 5.3.1, we simulated 500 datasets with $N = 50$ as $y \sim \mathcal{N}(\theta, I_N)$ for each of several values of θ. For each simulated dataset y, we computed the win $W(\theta, y)$, the proposed lower bound $b(y,\alpha)$, and the c-value $c(y)$. Conveniently, for this likelihood, the distributions of $W(\theta, y)$ and $b(y,\alpha)$ depend on θ only through $N^{-\frac{1}{2}}\|P_1^{\perp}\theta\|$. Consequently, we can exhaustively assess how our procedure behaves for different θ by varying this norm. Throughout our simulation study, we

fixed $\tau = 1$. With larger τ, the alternative θ^* behaves more similarly to the default $\hat{\theta}$, but the qualitative properties of the c-value and estimators remain similar.

We first checked that the empirical probability that the win $W(y, \alpha)$ exceeded the bound $b(y, \alpha)$ in Bound 5.3.1 was at least as large as the nominal probability α (Figure 5-1a). Across various choices of $N^{-\frac{1}{2}}\|P_1^{\perp}\theta\|$, we see that $b(\cdot, \alpha)$ is conservative, typically providing higher than nominal coverage. Surprisingly, the gap between the actual and nominal coverages does not seem to depend heavily on θ, suggesting we could potentially obtain a tighter bound by calibrating $b(y, \alpha)$ to its actual coverage.

We next examined the probability that that alternative estimate is selected on the basis of a large c-value but obtains higher loss than the default estimate, Theorem 5.2.3 upper bounds this probability and in Figure 5-1b we confirm it holds in practice across different thresholds α. Figure 5-1b additionally compares our proposed approach to using Stein's unbiased estimate of the risk [Stein, 1981]) of $\theta^*(\cdot)$ to select between the estimates. This approach, which we label "SURE", returns $\theta^*(\cdot)$ if the risk estimate exceeds N and returns $\hat{\theta}(\cdot)$ otherwise, and is akin to the focused information criterion [Claeskens and Hjort, 2003]. However, in contrast to the two stage estimate $\theta^{\dagger}(\cdot, \alpha)$, SURE does not provide tunable control over the probability that the alternative estimate $\theta^*(\cdot)$ is mistakenly returned.

Table 5.2: Contingency tables of simulation outcomes with $\|P_1^{\perp}\theta\|/\sqrt{N} = 1.7$ when using Stein's unbiased risk estimate (SURE), $\theta^{\dagger}(\cdot, \alpha = 0.95)$, or $\theta^{\dagger}(\cdot, \alpha = 0.5)$ to choose between the default and alternative estimates. DLL: **d**efault has **l**ower **l**oss, ALL: **a**lternative has **l**ower **l**oss, DR: **d**efault **r**eported, AR: **a**lternative **r**eported.

	SURE		c-values w/ $\alpha = 0.95$		c-values w/ $\alpha = 0.5$	
	DR	AR	DR	AR	DR	AR
DLL	2%	44%	46%	0%	37%	9%
ALL	36%	18%	54%	0%	54%	0.1%

In the case that $\|P_1^{\perp}\theta\|/\sqrt{N} = 1.7$, choosing based on SURE gives the wrong estimate 80% of the time. Moreover, in the majority of these cases it is the alternative that is incorrectly returned (Table 5.2a, Figure 5-1b). By contrast, the estimator that chooses based on the c-value (with a threshold $\alpha = 0.95$) conservatively returns

the default estimate in every replicate for this $\|P_1^{\perp}\theta\|/\sqrt{N}$ (Figure 5-1c). While this approach provides the estimate with greater loss in 54% of cases, it incorrectly reports the alternative in 0% of cases (Table 5.2 b). This behavior is expected as Theorem 5.2.3 provides an upper bound of $100 * (1 - \alpha)\% = 5\%$. An estimator using the unbiased risk estimate satisfies no such guarantee.

We next checked that our computed c-values successfully detected improvements by the alternative estimate. Recall that the alternative estimate $\theta^*(y)$ shrinks all components of y towards the global mean $\overline{y}$. Further, recall that by construction $\theta^{\dagger}(y, \alpha) = \theta^*(y)$ if and only if $c(y) > \alpha$. Intuitively, then, we would expect the alternative estimator to improve over the MLE and for the two-staged $\theta^{\dagger}(\cdot, \alpha)$ to select $\theta^*(\cdot)$ when θ is close to the subspace spanned by $\mathbf{1}_N$ and $N^{-\frac{1}{2}}\|P_1^{\perp}\theta\|$ is small. Figure 5-1c, which plots the probability that $\theta^{\dagger}(\cdot, \alpha)$ selects $\theta^*(\cdot)$ across different values of θ and α, confirms this intuition; when $N^{-\frac{1}{2}}\|P_1^{\perp}\theta\|$ is small, we very often obtain large c-values and select the alternative estimator.

For completeness, we also considered the risk profile of the two-stage estimator $\theta^{\dagger}(\cdot, \alpha)$ (Figure 5-1d). Specifically, for different choices of θ we computed a Monte Carlo estimate of the expected squared error loss. For the most part, the risk of $\theta^{\dagger}(\cdot, \alpha)$ lies between the risks of $\hat{\theta}(\cdot)$ and $\theta^*(\cdot)$. However, the risk of the two-stage estimator appears to exceed the risks of the default and alternative estimators for a narrow range of values of $\|P_1^{\perp}\theta\|$. While it is tempting to characterize this excess risk as the price we must pay for "double-dipping" into our data, we note that the bump in risk appears to be non-trivial only for very small values of α. Recall again that we recommend choosing $\theta^*(y)$ in place of $\hat{\theta}(y)$ only when $c(y)$ is close to 1. As such, we do not expect this type of risk increase to be much of a concern in practice.

Interpreted together, Figures 5-1c and 5-1d illustrate the conservatism of the two stage approach with $\alpha = 0.95$. For $\|P_1^{\perp}\theta\|$ between 1 and 1.5, $\theta^{\dagger}(\cdot, \alpha)$ only rarely evaluates to $\theta^*(\cdot)$ even though this estimator has lower risk and typically has smaller loss.

Unlike conventional p-values under a null hypothesis, we should not expect the distribution of informative c-values to be uniform; indeed for parameters such that

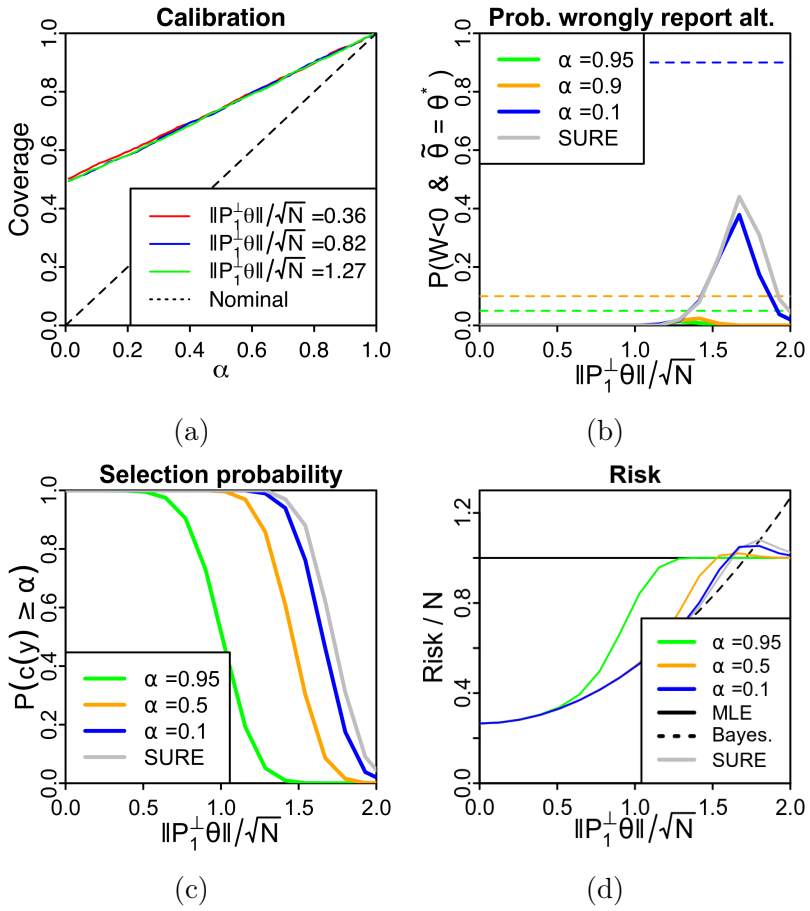


Figure 5-1: Bound calibration and the two-stage estimator for a hierarchical normal model in simulation. (a) Empirical coverage of the lower bound $b(\cdot, \alpha)$ across different levels α. Coverage is nearly identical across the parameter space. (b) Probability that the default has smaller loss but the alternative estimate is selected across the parameter space. (c) Probability of selecting the alternative estimate. Selection probability is higher for lower thresholds α. (d) Risk profiles of the two-stage estimators for different choices of α, as well as the MLE $\hat{\theta}(\cdot)$ and the shrinkage estimator $\theta^*(\cdot)$. Each data point is computed from 500 replicates with $N = 50$.

the win is consistently positive or negative, c-values can concentrate near 1 or 0, respectively.

5.4 Comparing affine estimates with correlated noise

We now generalize the situation described in the previous section in two ways. First, we consider correlated Gaussian noise with covariance Σ, where Σ is any $N \times N$ positive

definite covariance matrix rather than restricting to $\Sigma = I_N$. Second, we let our default and alternative estimates, $\hat{\theta}(y)$ and $\theta^*(y)$, be arbitrary affine transformations of the data y. Though these two estimates take similar functional forms in this section, we remain concerned with asymmetric comparisons wherein $\theta^*(y)$ is less familiar than $\hat{\theta}(y)$.

This situation introduces analytical challenges beyond those encountered in Section 5.3, but we nevertheless obtain an approximate bound that works well in practice. Specifically, for Bound 5.3.1, we used the tractable quantile function of the non-central χ^2 to guarantee exact coverage in Theorem 5.3.1. In the present case, we encounter sums of differently scaled non-central χ^2 random variables, which do not admit analytically tractable quantiles. However, by approximating these sums with Gaussians with matched means and variances, we can proceed in essentially the same manner as in Section 5.3 to derive an approximate lower bound on the win. After introducing the bound, we comment on the key steps in its derivation to highlight the approximations involved, but leave details of intermediate steps to Appendix D.6. We conclude with a non-asymptotic bound on the error introduced by these approximations on the coverage of the proposed bound on the win.

Approximate Bound 5.4.1 (Correlated Gaussian likelihood: arbitrary affine estimates). Observe $y = \theta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \Sigma)$ and consider $\hat{\theta}(y) = Ay + k$ vs. $\theta^*(y) = Cy + \ell$, where $A, C \in \mathbb{R}^{N \times N}$ are matrices and $k, \ell \in \mathbb{R}^N$ are N-vectors. We propose

$$\begin{aligned} b(y, \alpha) = \|\hat{\theta} - y\|^2 - \|\theta^* - y\|^2 + 2\mathrm{tr}[(A - C)\Sigma] + \\ 2z_{\frac{1-\alpha}{2}} \sqrt{U(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2}) + \frac{1}{2}\|\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}\|_F^2} \end{aligned} \tag{5.10}$$

as an approximate high-probability lower bound for the win. In this expression, $\mathrm{tr}[\cdot]$ denotes the trace of a matrix, $G(y) := (A - C)y + (k - \ell)$, $\|\cdot\|_\Sigma$ denotes the Σ quadratic norm of a vector ($\|v\|_\Sigma := \sqrt{v^\top \Sigma v}$), $\|\cdot\|_F$ denotes the Frobenius norm of a

matrix, and z_α denotes the α-quantile of the standard normal.

$$U(\|G(y)\|_\Sigma^2,1-\alpha) := \inf_{\delta>0}\left\{\delta \;\middle|\; \|G(y)\|_\Sigma^2 \le (\delta + \|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_F^2) + z_{1-\alpha}\sqrt{2\|\Sigma^{\frac{1}{2}}(A-C)\Sigma(A-C)^\top\Sigma^{\frac{1}{2}}\|_F^2 + 4\|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2\delta}\right\} \tag{5.11}$$

is an approximate high-confidence upper bound on $\|G(\theta)\|_\Sigma^2$ where $\|\cdot\|_{\mathrm{OP}}$ denotes the L2 operator norm of a matrix.

To derive Approximate Bound 5.4.1 we again start by rewriting the alternative estimate as $\theta^*(y) = \hat{\theta}(y) - G(y)$, where now $G(\cdot)$ is an affine transformation of y, $G(y) := (A-C)y + (k-\ell)$. We next write the squared error win of using $\theta^*(y)$ in place of $\hat{\theta}(y)$ as

$$W(\theta, y) = 2\epsilon^\top G(y) + \left(\|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2\right) \tag{5.12}$$

and observe that it suffices to obtain a high-probability lower bound for this first term. For tractability, we approximate the distribution of $\epsilon^\top G(y)$ by a normal with matched mean and variance. As we will soon see, this approximation is accurate when N is large and $A-C$ is well conditioned; in this case $\epsilon^\top G(y)$ may be written as the sum of many of uncorrelated terms of similar size. The mean and variance may be expressed as

$$\mathbb{E}[\epsilon^\top G(y)] = \mathrm{tr}[(A-C)\Sigma]\ ,\ \mathrm{Var}[\epsilon^\top G(y)] = \|G(\theta)\|_\Sigma^2 + \frac{\|\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}\|_F^2}{2}. \tag{5.13}$$

With these moments in hand, we form a probability α lower bound approximately as

$$W(\theta, y) \ge \|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2 + 2\mathrm{tr}[(A-C)\Sigma] + 2z_{1-\alpha}\sqrt{\|G(\theta)\|_\Sigma^2 + \frac{1}{2}\|\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}\|_F^2}. \tag{5.14}$$

However, as before, in order to use this approximate bound we require a simultaneous upper bound on a norm of a transformation of the unknown parameter, in this

case $\|G(\theta)\|_\Sigma^2$. We compute one by considering the test statistic $\|G(y)\|_\Sigma^2$ and again appealing to approximate normality. In particular we characterize the dependence of the distribution of this statistic on $\|G(\theta)\|_\Sigma^2$ through its mean and variance. We find its mean as

$$\mathbb{E}[\|G(y)\|_\Sigma^2] = \|G(\theta)\|_\Sigma^2 + \|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_F^2 \tag{5.15}$$

and upper bound its variance by

$$\mathrm{Var}[\|G(y)\|_\Sigma^2] \leq 2\|\Sigma^{\frac{1}{2}}(A-C)\Sigma(A-C)^\top\Sigma^{\frac{1}{2}}\|_F^2 + 4\|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2\|G(\theta)\|_\Sigma^2. \tag{5.16}$$

Using the two quantities above and an appeal to approximate normality, we propose the approximate high-confidence upper bound, $U(\|G(y)\|_\Sigma^2, 1-\alpha)$, in Eq. (5.11). As before, by splitting our α across these two bounds we obtain the desired expression, Eq. (5.10) in Approximate Bound 5.4.1.

Approximation Quality. Due to the two Gaussian approximations, Approximate Bound 5.4.1 does not provide nominal coverage by construction. Our next result shows that little error is introduced when N is large enough and the problem is well conditioned.

Theorem 5.4.1 (Berry–Esseen bound)**.** *Let $\alpha \in (0,1)$ and consider $b(\cdot,\alpha)$ in Approximate Bound 5.4.1. If both A and C are symmetric, then*

$$\mathbb{P}_\theta\left[W(\theta,y) \geq b(y,\alpha)\right] \geq \alpha - \frac{10\sqrt{2}}{\sqrt{N}} C_1 \cdot \kappa(\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}})^2 \tag{5.17}$$

where $\kappa(\cdot)$ denotes the condition number of its matrix argument (i.e. the ratio of its largest to smallest singular values) and $C_1 \leq 1.88$ is a universal constant [Berry, 1941, Theorem 1].

Remark 5.4.2. Theorem 5.4.1 is a special case of a more general result that we provide in Appendix D.6.4, which does not require A and C to be symmetric. We highlight this

special case here because the bound takes a simpler form from which the dependence on the conditioning of $A - C$ is clearer, and because this condition is satisfied for many important estimates. Notably A and C are symmetric in all applications discussed in this paper.

Though Theorem 5.4.1 provides an expected $O(N^{-\frac{1}{2}})$ drop in approximation error, the bound itself may be too loose to be useful in practice. In Section 5.6.1 we show in simulation that Approximate Bound 5.4.1 provides sufficient coverage even without this correction. This conservatism likely owes to slack from (A) the operator norm bound in Eq. (5.16) and (B) the union bound ensuring that the confidence interval for $\|G(\theta)\|_\Sigma^2$ and the quantile in Eq. (5.14) hold simultaneously.

We conclude this section with a remark about computation of Approximate Bound 5.4.1.

Remark 5.4.3 (Fast computation of $b(y, \alpha)$). A naive approach to computing $b(y, \alpha)$ in Eq. (5.10) involves finding $U(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2})$ with a binary search. For more rapid computation, we can recognize $U(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2})$ as the root of a quadratic. Specifically, define $\gamma := \|G(y)\|_\Sigma^2 - \|\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}}\|_F^2$, $\eta := z_{\frac{\alpha}{2}}$, $\rho := 2\|\Sigma^{\frac{1}{2}}(A - C)\Sigma(A - C)^\top\Sigma^{\frac{1}{2}}\|_F^2$, and $\nu := 4\|\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2$; then from Eq. (5.11) we have that the δ that achieves the supremum satisfies $\gamma = \delta + \eta\sqrt{\rho + \nu\delta}$. Rearranging, we find that $U(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2})$ is the larger root of $x^2 - (2\gamma + \eta^2\nu)x + (\gamma^2 - \eta^2\rho) = 0$.

5.5 Extending the reach of the c-value

Up to this point, we focused on estimating normal means with fixed affine estimators. Now we extend our c-value framework in two important directions, which we support with both theoretical and empirical results. In Section 5.5.1, we derive c-values for a nonlinear shrinkage estimator of normal means. We then move beyond Gaussian likelihoods in Section 5.5.2 and derive c-values for regularized logistic regression. In contrast to the earlier cases, these introduce nonlinear estimates and non-Gaussian models. To gain analytical tractability, we approximate the estimates by linear transformations of a statistic that is asymptotically Gaussian. This allows us to derive

bounds $b(y, \alpha)$ that we show have the correct coverage in an asymptotic regime. Our approach provides a template that can be followed for other nonlinear estimates and models for which the MLE is asymptotically Gaussian. We defer all proofs and details of synthetic data experiments to Appendices D.7 and D.8.

5.5.1 Empirical Bayes shrinkage estimates

While many Bayesian estimates are affine in the data for fixed settings of prior parameters, when prior parameters are chosen using the data, the resulting *empirical Bayesian* estimates are not affine in general. In this subsection we explore computation of approximate high-confidence lower bounds on the win of empirical Bayesian estimators. In particular, we consider an approach that essentially amounts to ignoring the randomness in estimated prior parameters and computing the bound as if the prior were fixed. For simplicity, we focus on a particularly simple empirical Bayesian estimator for the normal means problem that coincides with the James–Stein estimator [Efron and Morris, 1973]. We find that, in the high-dimensional limit, bounds obtained with this naive approach achieve at least the desired nominal coverage. Finally, we show in simulation that the approximate bound has favorable finite sample coverage properties.

Empirical Bayes for estimation of normal means. Consider a sequence of real-valued parameters $\theta_1, \theta_2, \ldots$, and corresponding observations $y_n \overset{indep}{\sim} \mathcal{N}(\theta_n, 1)$. For each $N \in \mathbb{N}$, let $\Theta_N := [\theta_1, \theta_2, \ldots, \theta_N]^\top$ and $Y_N := [y_1, y_2, \ldots, y_N]^\top$ denote the first N parameters and observations, respectively.

We consider the MLE for Θ_N (i.e. Y_N) as our default, which we denote by $\hat{\Theta}_N(Y_N) = Y_N$, and we take the James–Stein estimate as our alternative; we compare on the basis of squared error loss. We write the James–Stein estimate on the first N data points as $\Theta^*_N(Y_N) := \left(1 - (1 + \hat{\tau}^2_N(Y_N))^{-1}\right) Y_N$, where $\hat{\tau}^2_N(Y_N) := \|Y_N\|^2/(N-2) - 1$. $\Theta^*_N(Y_N)$ corresponds to the Bayes estimate under the prior $\theta_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \hat{\tau}^2_N)$ [Efron and Morris, 1973]. For this comparison, the win is $W_N(Y_N, \Theta_N) := \|\hat{\Theta}_N(Y_N) - \Theta_N\|^2 - \|\Theta^*_N(Y_N) - \Theta_N\|^2$, and Appendix D.7 details the associated bound $b_N(Y_N, \alpha)$ obtained

with Bound D.5.1. In the following theorem, we lower bound the win by applying our earlier machinery for Bayes rules with fixed priors. We find that the desired coverage is obtained in the high-dimensional limit.

Theorem 5.5.1. *For each $N \in \mathbb{N}$, let $\tau_N^2 := N^{-1} \sum_{n=1}^N \theta_n^2$. If the sequence $\tau_1, \tau_2, \ldots$ is bounded, then for any $\alpha \in [0,1]$, $\lim_{N\to\infty} \mathbb{P}\left[W_N(Y_N, \Theta_N) \geq b_N(Y_N, \alpha)\right] \geq \alpha$.*

The key step in the proof of Theorem 5.5.1 is establishing an $O_p(N^{-\frac{1}{2}})$ rate of convergence of $\hat{\tau}_N^2 - \tau_N^2$ to zero; under this condition the empirical Bayes estimate and bound converge to the analogous estimates and bounds computed with the prior variance fixed to τ_N^2. Accordingly, we expect similar results to hold for other models and empirical Bayes estimates when the standard deviations of the empirical Bayes estimates of the prior parameters drop as $O_p(N^{-\frac{1}{2}})$.

Remark 5.5.2. Theorem 5.5.1 easily extends to cover the case in which we consider a sequence of random (rather than fixed) parameters drawn i.i.d. from a Bayesian prior, which is a more classical setup for guarantees of empirical Bayesian methods; see e.g. Robbins [1964]. Specifically, our proof goes through in this Bayesian setting so long as the sequence $\tau_1^2, \tau_2^2, \ldots$ is bounded in probability. This condition is satisfied, for example, when the θ_n are i.i.d. from any prior with a finite second moment.

To check finite sample coverage, we performed a simulation and evaluated calibration of the associated c-values (Figure D.7.1). Despite the empirical Bayes step, the c-values appear to be similarly conservative to those computed with the exact bound in Figure 5-1a. Furthermore, this calibration profile does not appear to be sensitive to the magnitude of the unknown parameter.

5.5.2 Logistic regression

In this subsection we illustrate how to compute an approximate high-confidence lower bound on the win in squared error loss with a logistic regression likelihood. Our key insight is that by appealing to limiting behavior, this non-Gaussian problem may be approached with the same machinery developed in Section 5.4.

Notation and estimates. Consider a collection of M data points with random covariates $X_M := [x_1, x_2, \ldots, x_M]^\top \in \mathbb{R}^{M \times N}$ and responses $Y_M := [y_1, y_2, \ldots, y_M]^\top \in \{1, -1\}^M$. For the mth data point, assume

$$y_m \overset{indep}{\sim} p(\cdot \mid x_m; \theta) := (1 + \exp\{-x_m^\top \theta\})^{-1} \delta_1 + (1 + \exp\{x_m^\top \theta\})^{-1} \delta_{-1}, \tag{5.18}$$

where $\theta \in \mathbb{R}^N$ is an unknown parameter of covariate effects and δ_1 and δ_{-1} denote Dirac masses on 1 and -1, respectively.

In this section, we choose the MLE as our default, $\hat{\theta}(X_M, Y_M) := \arg\max_\theta \log p(Y_M \mid X_M; \theta)$. And we choose our alternative to be a Bayesian maximum a posteriori (MAP) estimate under a standard normal prior ($\theta \sim \mathcal{N}(0, I_N)$):

$$\theta^*(X_M, Y_M) := \arg\max_\theta \left\{ \log p(Y_M \mid X_M; \theta) - \frac{1}{2}\|\theta\|^2 \right\}.$$

While a first choice for a Bayesian estimate might be the posterior mean, the MAP is an effective and widely used alternative to the MLE in practice. Notably, the MAP estimate is easier to compute and is often close to the posterior mean; see Huggins et al. [2018, Proposition 6.2] and Schervish [1995, Theorem 7.116]. In particular, the distance between the posterior mean and the MAP estimate decays at an $O(M^{-1})$ rate with the number of observations M. Furthermore, $\theta^*(X_M, Y_M)$ is also of interest as an L2 regularized logistic regression estimate.

Approximating θ^* by an affine transformation. In moving away from a Gaussian likelihood we forfeit prior-to-likelihood conjugacy. In previous sections, conjugacy provided analytically convenient expressions for Bayes estimates. In order to regain analytical tractability, we appeal to a Gaussian approximation of the likelihood, defined with a second order Taylor approximation of the log likelihood about the MLE. This approximation is equivalent to approximating the distribution of the MLE as $\hat{\theta}(X_M, Y_M) \sim \mathcal{N}(\theta, \tilde{\Sigma}_M)$, where $\tilde{\Sigma}_M := -\nabla_\theta^2 \log p(Y_M \mid X_M; \theta)\big|_{\theta = \hat{\theta}(X_M, Y_M)}$. As such, we regain conjugacy, and we obtain an approximate Bayes estimate as an affine

transformation of the MLE,

$$\tilde{\theta}^*(X_M, Y_M) = \left[I_N + \tilde{\Sigma}_M\right]^{-1} \hat{\theta}(X_M, Y_M). \tag{5.19}$$

As we show in Appendix D.8, $\tilde{\theta}^*(X_M, Y_M)$ is a very close approximation of $\theta^*(X_M, Y_M)$, with distance decreasing at an $O_p(M^{-2})$ rate.

An approximate bound and an asymptotic guarantee. We leverage the form in Eq. (5.19) to compute Approximate Bound 5.4.1 as a lower bound on the win in squared error of using the MAP estimate in place of the MLE. In particular, we take $y := \hat{\theta}(X_M, Y_M)$ as the data in Approximate Bound 5.4.1 (this corresponds to $A = I_N$ and $k = 0$) and approximate the distribution of $\epsilon := \hat{\theta}(X_M, Y_M) - \theta$ as $\mathcal{N}(0, \tilde{\Sigma}_M)$. Further, to compute the bound, we approximate $\theta^*(X_M, Y_M)$ by $\tilde{\theta}^*(X_M, Y_M)$ as in Eq. (5.19), corresponding to $C = \left[I_N + \tilde{\Sigma}_M\right]^{-1}$ and $\ell = 0$.

While the precise coverage of this bound is difficult to analyze, our next result reveals favorable properties in the large sample limit.

Theorem 5.5.3. *Consider a sequence of random covariates $x_1, x_2, \ldots$ and responses $y_1, y_2, \ldots$ distributed as in Eq. (5.18). For each $M \in \mathbb{N}$, let $W_M := \|\hat{\theta}(X_M, Y_M) - \theta\|^2 - \|\theta^*(X_M, Y_M) - \theta\|^2$ be the win of using the MAP estimate in place of the MLE. Finally, let $b_M(\alpha)$ be the level-α approximate bound on W_M described above. If $x_1, x_2, \ldots$ are i.i.d. with finite third moment and with positive definite covariance, then for any $\alpha \in (0, 1)$, $\lim_{M\to\infty} \mathbb{P}_\theta \left[W_M \geq b_M(\alpha)\right] \geq \alpha$.*

Theorem 5.5.3 guarantees that in the large sample limit, $b_M(\cdot)$ has at least nominal coverage. We provide a proof of the theorem and demonstrate its favorable empirical properties in simulation in Appendix D.8.

5.6 Applications

We now demonstrate our approach on the three applications introduced in Section 5.1. Our goal in this section is to demonstrate how one can compute and interpret c-values

in realistic workflows. In analogy to hypothesis testing, where a p-value cutoff of 0.05 is standard for rejecting a null, we require a c-value of at least 0.95 to accept the alternative estimate; with this threshold, we expect to incorrectly reject the default estimate in at most 5% of our decisions. This choice, instead of 0.5 for example, reflects the presumed asymmetry of the comparisons; we require strong evidence to adopt the alternative over the default. For all applications, we provide substantial additional details in Appendix D.9.

5.6.1 Estimation from educational testing data and empirical Bayes

In this section we apply our methodology to a model and dataset considered by Hoff [2021, Section 3.2], in which the goal is to estimate the average student reading ability at different schools in the 2002 Educational Longitudinal Study. At each of $N = 676$ schools, between 5 and 50 tenth grade students were given a standardized test of reading ability. We let $y = [y_1, y_2, \ldots, y_N]^\top$ denote the average scores, and for each school, indexed by n, model $y_n \overset{indep}{\sim} \mathcal{N}(\theta_n, \sigma_n^2)$, where $\theta = [\theta_1, \theta_2, \ldots, \theta_N]^\top$ denotes the school-level means and each σ_n is the school-level standard error; specifically $\sigma_n := \sigma/\sqrt{N_n}$ where σ denotes a student-level standard deviation and N_n is the number of students tested at school N_n. For convenience, we let $\Sigma := \text{diag}([\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2])$ so that we may write $y \sim \mathcal{N}(\theta, \Sigma)$. The goal is to estimate the school-level performances θ.

Following Hoff [2021], we perform small area inference with the Fay-Herriot model [Fay and Herriot, 1979] to estimate θ under the assumption that similar schools may have similar student performances. Specifically, we consider a vector of $D = 8$ attributes of each school $X = [x_1, x_2, \ldots, x_N]^\top$; these include participation levels in a free lunch program, enrollment, and other characteristics such as region and school type. We model the school-level mean as *a priori* distributed as $\theta \sim \mathcal{N}(X\beta, \tau^2 I_N)$ where β is an unknown D-vector of fixed effects and τ^2 is an unknown scalar that describes variation in θ not captured by the covariates. Following Hoff [2021], we take an empirical Bayesian approach and estimate β, τ, and σ with `lme4` [Bates et al.,

2015a]. We then compare the posterior mean — which is affine in y for fixed β, τ, and σ — as an alternative to the MLE as a default; we use Approximate Bound 5.4.1. Specifically, we take $\theta^*(y) := \mathbb{E}[\theta \mid y; \beta, \tau, \sigma] = [I_N + \tau^{-2}\Sigma]^{-1}y + [I_N + \tau^2\Sigma^{-1}]^{-1}X\beta$ and $\hat{\theta}(y) = y$. We compute a large c-value ($c = 0.9926$); its closeness to one strongly suggests that $\theta^*(y)$ is more accurate than $\hat{\theta}(y)$.

We should not always expect to obtain a large c-value for any alternative estimate, however. We next describe a case where we expect the alternative estimate to be less accurate than the default, and we check that we obtain a small c-value. In particular, we now let our alternative estimate be the posterior mean under the same model as above but with the covariates, X, randomly permuted across schools. In this situation, the responses y have no relation to the covariates, and we should not expect an improvement. Indeed, on this dataset we compute a c-value of exactly zero. However, we recall that just as a large p-value in hypothesis testing does not provide evidence that a null hypothesis is true, a small c-value does not provide direct evidence that the alternative estimate is less accurate than the default.

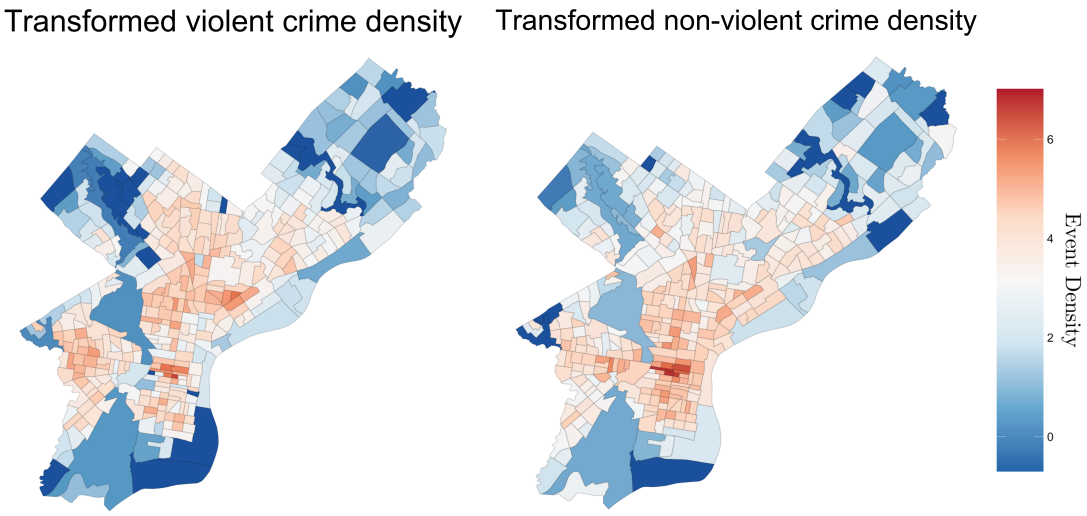
We provide additional details for all parts of this application in Appendix D.9.1. There, we demonstrate in a simulation study that our bounds remain substantially conservative for these estimators and model even with an empirical Bayes step.

5.6.2 Estimating violent crime density in Philadelphia

As a second application, we consider estimating the areal density of violent crimes (i.e. counts per square mile) reported in each of Philadelphia's $N = 384$ census tracts. Following Balocchi et al. [2019], we work with the inverse hyperbolic sine transformed density. Letting y_n be the observed transformed density of reported violent crimes in census tract n, we model $y_n \overset{indep}{\sim} \mathcal{N}(\theta_n, \sigma_y^2)$ where θ_n represents the underlying transformed density and σ_y^2 is the noise variance. While one might interpret θ_n as the true density of violent crime in census tract n, we note that the implicit assumption of zero-mean error in each tract may not be realistic. Namely, systematic biases may impact the rates at which police receive and respond to calls and file incident reports in different parts of the city. Unfortunately, we are unable to probe this possibility

with the available data. Nevertheless, our goal is to estimate the vector of unknown rates, $\theta = [\theta_1, \theta_2, \ldots, \theta_N]^\top$ from $y = [y_1, y_2, \ldots, y_N]^\top$. The observations y are a simple proxy of transformed violent crime density, but they are noisy. So it is natural to wonder if we might obtain a more accurate estimate of θ.

Figure 5-2 plots the transformed densities of both violent and non-violent crimes reported in October 2018 in each census tract. Immediately, we see that, for any particular census tract, the observed densities of the two types of crime are similar. Further, we observe considerable spatial correlation in each plot. It is tempting to use a Bayesian hierarchical model that exploits this structure in order to produce more accurate estimates of θ. In this application, we consider iteratively refining an estimate of θ by (A) incorporating the observed non-violent crime data and then by (B) carefully accounting for the observed spatial correlation. At each step of our refinement, we use a c-value to decide whether to continue. Before proceeding, we make a remark about our sequential approach.



(a) (b)

Figure 5-2: Transformed densities of reported (a) violent and (b) non-violent crimes in each census tract in Philadelphia in October 2018.

Remark 5.6.1. Consider using c-values and a chosen level α to choose one of three estimates (say $\hat{\theta}(y), \theta^*(y)$, and $\theta^\circ(y)$) in two stages. Suppose we first choose $\theta^*(y)$ over $\hat{\theta}(y)$ only if the associated c-value is greater than α. Second, only if we chose $\theta^*(y)$, we

next choose $\theta^\circ(y)$ over $\theta^*(y)$ only if the new c-value associated with those estimates exceeds α. Then a union bound guarantees that $\theta^\circ(y)$ will be incorrectly chosen with probability at most $2(1-\alpha)$.

We begin by seeing if we can improve upon the MLE, $\hat{\theta}(y) = y$, by leveraging the auxiliary dataset of transformed non-violent crimes in each tract, $z_1, z_2, \ldots, z_N$. To this end, we model these auxiliary data analogously to y; in each tract n, we let η_n be the unknown transformed density and independently model $z_n \overset{indep}{\sim} \mathcal{N}(\eta_n, \sigma_z^2)$. We next introduce a hierarchical prior that captures the apparent similarity between θ and η within each tract. Specifically, for each tract n we decompose $\theta_n = \mu_n + \delta_n^y$ and $\eta_n = \mu_n + \delta_n^z$, where μ_n is a shared mean for the transformed densities of violent and non-violent reports and δ_n^y and δ_n^z represent deviations from the shared mean specific to each crime type. Rather than encode explicit prior beliefs about μ_n, we express ignorance in these quantities with an improper uniform prior. Additionally, we model $\delta_n^y, \delta_n^z \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_\delta^2)$. We fix the values of σ_y, σ_z, and σ_δ using historical data. We then compute the posterior mean of θ as an alternative estimate, $\theta^*(y)$. Thanks to the Gaussian conjugacy of this model, $\theta^*(y)$ is affine in the data y, and a closed form expression is available. See Appendix D.9.2 for additional details. The resulting c-value exceeded 0.999, suggesting that we should be highly confident that $\theta^*(y)$ is a more accurate estimate of θ than $\hat{\theta}(y)$.

We next consider additionally sharing strength amongst spatially adjacent census tracts. To this end, consider a second model with spatially correlated variance components: $\theta_n = \mu_n + \delta_n^y + \kappa_n^y$ and $\eta_n = \mu_n + \delta_n^z + \kappa_n^z$. The additional terms $\kappa^y = [\kappa_1^y, \kappa_2^y, \ldots, \kappa_N^y]^\top$ and $\kappa^z = [\kappa_1^z, \kappa_2^z, \ldots, \kappa_N^z]^\top$ capture a priori spatial correlations; we model $\kappa^y, \kappa^z \overset{i.i.d.}{\sim} \mathcal{N}(0, K)$, where K is an $N \times N$ covariance matrix determined by a squared exponential covariance function [Rasmussen and Williams, 2006, Chapter 4] that depends on the distance between the centroids of the census tracts. Once again, we exploit conjugacy in this second hierarchical model to derive the posterior mean $\theta^\circ(y)$ in closed form. As $\theta^\circ(y)$ is also an affine transformation of y, we can use Approximate Bound 5.4.1 to compute the c-value for comparing $\theta^\circ(y)$ to $\theta^*(y)$. The c-value for this comparison is only 0.843, providing much weaker support for

using $\theta^\circ(y)$ over $\theta^*(y)$. Because this c-value is less than 0.95, we conclude our analysis content with $\theta^*(y)$ as our final estimate.

5.6.3 Gaussian process kernel choice: modeling ocean currents

Accurate understanding of ocean current dynamics is important for forecasting the dispersion of oceanic contaminations, such as after the Deepwater Horizon oil spill [Poje et al., 2014]. Lodise et al. [2020] have recently advocated for a statistical approach to inferring ocean currents from observations of free-floating, GPS-trackable buoys. Their approach seeks to provide improved estimates by incorporating variation at the *submesoscale* (roughly 0.1–10 km) in addition to more commonly considered *mesoscale* variation (roughly 10 km and above). In this section we apply our methodology to assess if this approach provides improved estimates relative to a baseline including only mesoscale variation.

In our analysis, we consider a segment of the Carthe Grand Lagrangian Drifter (GLAD) deployment dataset [Özgökmen, 2013]. Specifically, we model a set of 50 buoys with velocities estimated at 3 hour intervals over one day ($N = 400$ observations total). Each observation n consists of latitudinal and longitudinal ocean current velocity measurements $y_n = [y_n^{(1)}, y_n^{(2)}]^\top \in \mathbb{R}^2$ and associated spatio-temporal coordinates $[\text{lat}_n, \text{lon}_n, t_n]$. Following Lodise et al. [2020], we model each measurement as a noisy observation of an underlying time varying vector-field distributed independently as $y_n \overset{indep}{\sim} \mathcal{N}\left(F(\text{lat}_n, \text{lon}_n, t_n), \sigma_\epsilon^2 I_2\right)$, where $F : \mathbb{R}^3 \to \mathbb{R}^2$ denotes the time evolving vector-field of ocean currents and σ_ϵ^2 is the error variance. Our goal is to estimate F at the observation points $\theta := [\theta_1, \theta_2, \ldots, \theta_N]^\top$, where for each $n, \theta_n = [\theta_n^{(1)}, \theta_n^{(2)}]^\top = F(\text{lat}_n, \text{lon}_n, t_n)$.

Following Lodise et al. [2020], we place a Gaussian process prior on F to encode expected spatio-temporal structure while allowing for variation at multiple scales. Specifically, we model $F \sim \mathcal{GP}\left(0, k(\cdot, \cdot)\right)$, where

$$k(\theta_n^{(i)}, \theta_{n'}^{(i)}) = k_1(\theta_n^{(i)}, \theta_{n'}^{(i)}) + k_2(\theta_n^{(i)}, \theta_{n'}^{(i)}), \quad i \in \{1, 2\}. \tag{5.20}$$

Here k_1 and k_2 are squared exponential kernels with spatial and temporal length-scales that reflect mesoscale and submesoscale variations, respectively; see Appendix D.9.3 for details. For simplicity, we model the latitudinal and longitudinal components of F independently. We take the posterior mean of θ under this model as the alternative estimate, $\theta^*(y)$.

As a baseline, we consider an analogous estimate with covariance function $k(\theta_n^{(i)}, \theta_{n'}^{(i)}) = k_1(\theta_n^{(i)}, \theta_{n'}^{(i)}) + k_2(\theta_n^{(i)}, \theta_{n'}^{(i)})\mathbb{1}[n = n']$, which maintains the same marginal variance but excludes submesoscale covariances. We take the posterior mean under this model as the default estimate $\hat{\theta}(y)$. Both $\theta^*(y)$ and $\hat{\theta}(y)$ may be written as affine transformations of y.

Using Approximate Bound 5.4.1, we compute a c-value of 0.99981. This large c-value allows us to confidently conclude that modeling both mesoscale and submesocale variation can yield more accurate estimates of ocean currents than mesocale modeling alone.

5.7 Discussion

We have provided a simple method for quantifying confidence in improvements provided by a wide class shrinkage estimates without relying on subjective assumptions about the parameter of interest. Our approach has compelling theoretical properties, and we have demonstrated its utility on several data analyses of recent interest. However, the scope of the current work has several limitations. The present paper has explored the use of the c-value only for problems of moderate dimensionality (N between 20 and 700). Loosely speaking, we suspect c-values may be underpowered to robustly identify substantial improvements provided by estimates in lower dimensional problems. Further investigation into such dimension dependence is an important direction for future work. In addition, our approach depends crucially on a high-probability lower bound that is inherently specific to the underlying model of the data, a loss function, and the pair of estimators. In the present work, we have shown how to derive and compute this bound for models with general Gaussian likelihoods, when accuracy

may be measured in terms of squared error loss, and when both estimates are affine transformations of the data. We have provided a first step to extending beyond simple Gaussian models with the application to logistic regression; while we have not yet explored the efficacy of this extension on real data, we view our work as an important starting point for generalizing to broader model classes and estimation problems. We believe that further extensions to the classes of models, estimates, and losses for which c-values can be computed provide fertile ground for future work.

It may be possible to construct the bound $b(y, \alpha)$ in a model and loss agnostic approach, using, for example, the parametric bootstrap. Constructing an informative c-value is possible only because in some cases the distribution of the win depends on the unknown parameter only through some low dimensional projection (or at least approximately so). We suspect that this may be the case for some more complex models and estimates too. When this is the case and if these low dimension characteristics are estimated well enough, a parametric bootstrap may present a powerful solution. In particular, one would begin by forming an initial estimate of the parameter, and simulate a collection of bootstrap datasets by sampling data from the likelihood parameterized by the initial estimate, compute the win for each simulated dataset, and return for each $b(y, \alpha)$ the $1 - \alpha$ quantile of this distribution. We expect that this method may work in many important settings – indeed, much of modern statistics and nonlinear methods are predicated on the assumption that low dimensional structure (e.g. sparsity) exists and may be inferred. We leave further development of this more flexible approach, including an investigation of the theoretical properties, to follow-up work.

Appendix A

Exchangeability Supplementary Material

A.1 Additional Related Work

A.1.1 Brown and Zidek details

As discussed in Section 2.1, the papers of Brown and Zidek [1980] and Haitovsky [1987] carry the only references of which we are aware of the idea of exchangeability of effects across covariates for sharing strength among multiple groups of data. We here provide additional discussion on this related prior work. To aid our comparison, we slightly modify their notation to match ours.

In their paper, "Adaptive Multivariate Ridge Regression", Brown and Zidek [1980] consider multiple related regression regression problems with a shared design (i.e. $X := X^1 = X^2 = \cdots = X^Q$) and seek to extend the univariate ridge regression estimator of Hoerl and Kennard [1970] to the multivariate setting. Specifically, the authors propose a class of estimators of the form

$$\hat{\vec{\beta}} = (I_Q \otimes X^\top X + K \otimes I_D)^{-1}(I_Q \otimes X^\top)\vec{Y},$$

where $\vec{Y} := [Y^{1\top}, Y^{2\top}, \cdots, Y^{Q\top}]^\top$, $\otimes$ denotes the Kronecker product, and K is a

$Q \times Q$ ridge matrix which they suggest be chosen by some "adaptive rule" (i.e. that K be a function of the observed data). Notably, this functional form closely resembles our expression for $\mathbb{E}[\vec{\beta}|\mathcal{D}, \Sigma]$ in Proposition 2.3.1, if we take $K = \Sigma^{-1}$.

The authors do not explicitly discuss the interpretation of K^{-1} as the covariance of a Gaussian prior, nor any interpretation for this quantity as capturing any notion of a priori similarity of the regression problems. However, they do point to Bayesian motivations at the outset of the paper. In particular, Brown and Zidek [1980] narrow their consideration of possible methods for choosing K to those which satisfy two criteria:

1. For any K, $\hat{\vec{\beta}}$ correspond to a Bayes estimate.
2. In the case that $X^\top X = I_D$, $\hat{\vec{\beta}}$ correspond to the Efron and Morris [1972b] extension of the James and Stein [1961] estimator to vector observations.[1]

They present four such estimators (derived from existing estimators of a multivariate normal means that dominate the sample mean) and demonstrate conditions under which each of these estimators dominates the least squares estimator for β.

As a further point of connection, the authors claim in the their abstract that their "result is implicitly in the work of Lindley and Smith [1972] although not actually developed there." However, the authors give little support for, or clarification of this claim. In particular, their analysis is entirely frequentist and they provide no explanation for how their proposed estimators for K might be interpreted as reasonable empirical Bayes estimates.

In their short follow-up paper, Haitovsky [1987] elaborates on this Bayesian motivation. The primary focus of Haitovsky [1987] is a matrix normal prior [Dawid, 1981] that captures structure in effects across both groups and covariates. Though this prior is not exchangeable across covariates in general, they note that the special case of where effects are uncorrelated across different covariates satisfies the notion of exchangeability for which we have advocated in this paper.

[1] See Appendix A.1.4 for further discussion of connections to Efron and Morris [1972b].

A.1.2 Methods of inference for Γ in existing work assuming exchangeability of effects across groups.

We here describe several existing approaches for estimating the covariance matrix Γ in the exchangeability of effects among groups model. These existing methods do not translate directly to the exchangeability of effects among covariates model proposed in this paper. However, in principle, one could likely adapt any of them to our setting. We have chosen to use the EM algorithm described in Section 2.3 for its simplicity, efficiency, and stability. We leave the investigation of alternative estimation approaches to future work.

In their initial paper, Lindley and Smith (1972) [Lindley and Smith, 1972] suggest that a fully Bayesian approach would be ideal. They advocate for placing a subjectively specified, conjugate Wishart prior on Γ, and remark that one should ideally consider the posterior of Γ rather than relying on a point estimate. However, in the face of analytic intractability, they propose returning MAP estimates for Γ and β and provide an iterative optimization scheme that they show is stationary at $\hat{\Gamma}, \hat{\beta} = \arg\max \log p(\Gamma, \beta|\mathcal{D})$.

Advances in computational methods since 1972 have given rise to other ways of estimating Γ in this model. Gelfand et al. [1990] describe a Gibbs sampling algorithm for posterior inference. Gelman et al. [2013, Chapter 15 sections 4-5] describe an EM algorithm which returns a maximum a posteriori estimate marginalizing over β, $\hat{\Gamma} = \arg\max p(\Gamma|\mathcal{D}) = \int p(\Gamma, \beta|\mathcal{D})d\beta$; notably, though the updates in our EM algorithm for the case of exchangeability in effects across covariates differ from those in the case of exchangeability among groups, one can see the two algorithms as closely related through their shared dependence on Gaussian conjugacy. Finally, in the software package `lme4`, Bates et al. [2015b] use the maximum marginal likelihood estimate, $\hat{\Gamma} = \arg\max p(\mathcal{D}|\Gamma)$, which they compute using gradient based optimization.

A.1.3 Details on connections to `lme4`

In the notation of `lme4` [Bates et al., 2015b], our paper considers only random effects and no fixed effects. In that work, each vector of random effects, denoted $\mathcal{B}$, corresponds to a length D (q in their notation) column of β (in our notation). Bates et al. [2015b, Equation 3] states the prior derived from Lindley and Smith [1972] that reflects the assumption of exchangeability across groups and captures correlation structure across covariates. This correlation structure is modeled whenever two or more random effects are specified and allowed to vary across groups. In the high dimensional setting (when $D > Q$), however, `lme4` fails to run because the optimization problem associated with empirical Bayes step is ill-conditioned.

A.1.4 Related work on estimation of normal means

As we discuss in Appendix A.3.1, under Condition 2.4.1 and when $\sigma^2 = 1$, we have that

$$\hat{\beta}^q_{\text{LS}} \overset{indep}{\sim} \mathcal{N}(\beta^q, I_D).$$

As such, inference reduces to the "normal means problem", with a matrix valued parameter. Specifically, we can equivalently write

$$\hat{\beta}_{\text{LS}} = \beta + \epsilon,$$

for a random $D \times Q$ matrix ϵ with i.i.d. standard normal entries.

This problem has been studied closely outside of the context of regression. Notably, Efron and Morris [1972a] approach the problem from an empirical Bayesian perspective and recommend an approach analogous to estimating Σ by

$$\hat{\Sigma}^{\text{Ef}} := (D - Q - 1)^{-1} \hat{\beta}^\top_{\text{LS}} \hat{\beta}_{\text{LS}} - I_Q.$$

Efron and Morris [1972a] argue for this estimate because it is unbiased for a transformation of the parameter. In particular, $\hat{\Sigma}^{\text{Ef}}$ satisfies $\mathbb{E}[(I_Q + \hat{\Sigma}^{\text{Ef}})^{-1}] = (I_Q + \Sigma)^{-1}$ when

each $\beta_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$. They show that, among all estimates of the form $\alpha\hat{\beta}_{\mathrm{LS}}^{\top}\hat{\beta}_{\mathrm{LS}} - I_Q$ with real valued α, this factor $\alpha = (D - Q - 1)^{-1}$ is optimal in terms of squared error risk. Notably, this includes the moment estimate $\hat{\Sigma}^{\mathrm{MM}}$ we describe in Section 2.4, which corresponds to $\alpha = D^{-1}$. However, this optimality result does not translate to the associated positive part estimators. In fact, in experiments not shown, we have found that $\hat{\beta}_{\mathrm{ECov}}$ reliably outperforms an analogous positive part variant that estimates Σ by $\hat{\Sigma}^{\mathrm{Ef}}$.

Remark A.1.1. Efron and Morris [1972a, Theorem 5] prove that an analogous positive part estimator is superior to their original estimator in term of "relative savings loss" (RSL). Our domination result in Theorem 2.4.2 is strictly stronger and implies an improvement in RSL as well. Furthermore our proof technique immediately applies to their estimator.

Several other works have noted the dependence of the risk of estimators for the matrix variate normal means problem on the expectations of the eigenvalues of inverse non-central Wishart matrices [Efron and Morris, 1972a, Zidek, 1978, Van Der Merwe and Zidek, 1980]. In all of these cases, the authors did not document attempts to interpret or approximate these difficult expectations.

More recently, Tsukuma [2008] explores a large class of estimators for the matrix variate normal means problems that shrink $\hat{\beta}_{\mathrm{LS}}$ along the directions of its singular vectors in different ways. For subclass of these estimators, Tsukuma [2008][Corollary 3.1] proves a domination result for associated positive part estimators. In the orthogonal design case, $\hat{\beta}_{\mathrm{ECov}}$ can be shown to be a member of this subclass of estimators, providing an alternative route to proving Theorem 2.4.3.

A.1.5 Additional related work on multiple related regressions

Methods for simultaneously estimating the parameters of multiple related regression problems have a long history in statistics and machine learning, with different assumptions and analysis goals leading to a diversity of inferential approaches. Perhaps the most famous is Zellner's landmark paper on seemingly unrelated regressions (SUR)

[Zellner, 1962]. Zellner [1962] addresses the situation where apparent independence of regression problems is confounded by covariance in the errors across Q problems (i.e. 'groups' in our language). In the presence of such correlation in residuals, the parameter may be identified with greater asymptotic statistical efficiency by considering all Q problems together [Zellner, 1962, Zellner and Huang, 1962]. While most work on SUR has taken a purely frequentist perspective in which β is assumed fixed, some more recent works on SUR have considered Bayesian approaches to inference [Blattberg and George, 1991, Chib and Greenberg, 1995, Smith and Kohn, 2000, Griffiths, 2003, Ando and Zellner, 2010]. However these do not address the scenario of interest here, in which we believe *a priori* that there may be some covariance structure in the effects of covariates *across* the regressions, or that some regression problems are more related than others. The setting of the present paper further differs from SUR in that we do not consider correlation in residuals as a possible mechanism for sharing strength between groups, but instead explicitly assume independence in the noise.

Breiman and Friedman [1997] present a distinct, largely heuristic approach to multiple related regression problems where all Q responses are observed for each group, or equivalently each group has the same design. The authors focus entirely on prediction and obviate the need share information across regression problems when forming an initial estimate of β by proposing to predict new responses in each regression with a linear combination of the predictions of linear models defined by the independently computed least squares estimate of each regression problem. However this approach does not consider the problem of estimating parameters, which is a primary concern of the present work.

Reinsel [1985]'s paper, "Mean Squared Error Properties of Empirical Bayes Estimators in a Multivariate Random Effects General Linear Model", considers a mixed effects model in which a linear model for regression coefficients is specified $\beta^q = Ba_q + \lambda_q$ where $a := [a_1, a_2, \ldots, a_Q]$ is a $K \times Q$ known design matrix associated with the regression problems,[2] B is a $D \times K$ matrix of unknown parameters and $[\lambda_1, \lambda_2, \ldots, \lambda_Q]$ is

[2]Notably, though Reinsel [1985] refers to a as a design matrix, it has little relation of the design matrices X^q to which we frequently refer in the present work.

a $D \times Q$ matrix of error terms. These error terms are assumed exchangeable across groups. In contrast to the present work, Reinsel [1985] requires the relatedness between groups to be known a priori through the known design matrix a.

Laird and Ware [1982] consider a random effects model for longitudinal data in which different individuals correspond to different regression problems with distinct parameters. In their construction, covariance structure in the noise is allowed across the observations for each individual, but not across individuals. Additionally, as in Lindley and Smith [1972], the authors model the covariance in effects of different covariates a priori within each regression, but not covariance across regressions.

Brown et al. [1998] propose to use sparse prior for β which encourages a shared sparsity pattern. Conditioned on a binary $D-$vector $\gamma \in \{0,1\}^D$, β is supposed to follow a multivariate normal prior as

$$\vec{\beta} \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma \otimes H_\gamma)$$

where H_γ is a $D \times D$ covariance matrix which expresses that for d such that $\gamma_d = 0$ we expect each $\beta_{d,q}$ to be close to zero. Notably, this is equivalent to the assumption that β follows a matrix-variate multivariate normal distributed as $\beta \sim \mathcal{MN}(0, H_\gamma, \Sigma)$ [Dawid, 1981]. Curiously, and without stated justification, the same Σ is also taken to parameterize the covariance of the residual errors, as well as of an additional bias term. We suspect this restriction is made for the sake of computational tractability. Indeed, [Stephens, 2013] makes similar modeling assumptions for tractability in the context of statistical genetics. In contrast to the present work, the premise of Brown et al. [1998] is sharing strength through similar sparsity patterns and covariance in the residuals, rather than learning and leveraging patterns of similarity in effects of covariates across groups.

Other more recent papers have considered alternative approaches for multiple regression with sparse priors [Bhadra and Mallick, 2013, Lewin et al., 2015, Deshpande et al., 2019]. As one example, Obozinski et al. [2006] estimate parameters across multiple groups with a mixed ℓ_1/ℓ_2 regularized objective that induces sparsity. Yang

et al. [2009], Lee et al. [2010] build on this work by Obozinski et al. [2006] with a focus on applications in genetics. These latter methods may be understood as returning the maximum a posteriori estimate under a Bayesian model. However, in contrast to our approach, the corresponding prior distributions implicit in such perspectives do not capture a priori correlation of effects across groups. Moreover, these methods are of course inappropriate when we do not expect sparsity a priori.

Meta-Learning The popular "Model Agnostic Meta-Learning" (MAML) approach [Finn et al., 2017] can be understood as a hierarchical Bayesian method that treats tasks / groups exchangeably [Grant et al., 2018]. As such, MAML and its variations do not allow tasks to be related to different extents (as our approach does). A few recent works on meta-learning are exceptions; for example, Jerfel et al. [2019] model tasks as grouped into clusters by using a Dirichlet process prior, and Cai et al. [2020] consider a weighted variant of MAML that allows, for a given task of interest, the contribution of data from other tasks to vary. However these works differ from the present paper in their focus on prediction with flexible black-box models, whereas the primary concern of the present is parameter estimation in linear models.

Exchangeability of effects across covariates in the single group context. In the context of regression problems consisting of only a single group (i.e. corresponding to the special case of $Q = 1$) Lindley and Smith [1972] suggest modeling the D scalar covariate effects exchangeable. In particular, they suggest modeling scalar covariate effects as i.i.d. from a univariate Gaussian prior when this exchangeability assumption is appropriate. However, because this development is restricted to analyses of data in a single group, it does not relate to the problem of sharing strength across multiple groups, which is the subject of the present work.

A.2 Section 2.3 supplementary proofs and discussion

A.2.1 Proof of Proposition 2.3.1

Proof. First note that the least squares estimate

$$\hat{\beta}_{\text{LS}} := [(X^{1\top}X^1)^{-1}X^{1\top}Y^1, \ldots, (X^{Q\top}X^Q)^{-1}X^{Q\top}Y^Q]$$

is a sufficient statistic of $\mathcal{D}$ for β, and so $\beta|\mathcal{D}, \Sigma \sim \beta|\hat{\beta}_{\text{LS}}, \Sigma$. As such, it is sufficient to consider the likelihood of $\hat{\beta}_{\text{LS}}$. Let $\hat{\vec{\beta}}_{\text{LS}} := [Y^{1\top}X^1(X^{1\top}X^1)^{-1}, \ldots, Y^{Q\top}X^Q(X^{Q\top}X^Q)^{-1}]$ be the DQ-vector defined by stacking the least squares estimates for each group. Since for each q, we have $\hat{\beta}^q_{\text{LS}}|\beta \overset{indep.}{\sim} \mathcal{N}(\beta^q, \sigma_q^2(X^{q\top}X^q)^{-1})$, we can write $\hat{\vec{\beta}}_{\text{LS}}|\beta \sim \mathcal{N}\left[\vec{\beta}, \text{diag}\left(\sigma_1^2(X^{1\top}X^1)^{-1}, \ldots, \sigma_Q^2(X^{Q\top}X^Q)^{-1}\right)\right]$. Next, that each $\beta_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ a priori implies that we may write $\vec{\beta} \sim \mathcal{N}(0, \Sigma \otimes I_D)$ a priori, where $\otimes$ is the Kronecker product. Then, by Gaussian conjugacy (see e.g. Bishop [2006, Chapter 2.3]), we have that $\vec{\beta}|\mathcal{D} \sim \mathcal{N}(\vec{\mu}, V)$, where

$$\vec{\mu} = V\left[(\Sigma \otimes I_D)^{-1}0 + \text{diag}\left(\sigma_1^2(X^{1\top}X^1)^{-1}, \ldots, \sigma_Q^2(X^{Q\top}X^Q)^{-1}\right)^{-1}\hat{\vec{\beta}}_{\text{LS}}\right]$$

for $V^{-1} = (\Sigma \otimes I_D)^{-1} + \text{diag}\left(\sigma_1^2(X^{1\top}X^1)^{-1}, \ldots, \sigma_Q^2(X^{Q\top}X^Q)^{-1}\right)^{-1}$. Due to the block structure of the matrices above, these simplify to $\vec{\mu} = V\left[\frac{Y^{1\top}X^1}{\sigma_1^2}, \ldots, \frac{Y^{Q\top}X^Q}{\sigma_Q^2}\right]$ and $V^{-1} = \Sigma^{-1} \otimes I_D + \text{diag}(\frac{X^{1\top}X^1}{\sigma_1^2}, \ldots, \frac{X^{Q\top}X^Q}{\sigma_Q^2})$, as desired. □

A.2.2 Efficient computation with the conjugate gradient algorithm

As mentioned in Section 2.3.1, $\vec{\mu} = \mathbb{E}[\vec{\beta}|\mathcal{D}, \Sigma]$ in Proposition 2.3.1 may be computed efficiently using the conjugate gradient algorithm (CG) for solving linear systems. We here describe several properties of CG that make it surprisingly well-suited to this application.

We first note that Proposition 2.3.1 allows us to frame computation of $\vec{\mu}$ as the solution to the linear system

$$A\vec{\mu} = b$$

for $b = \left[Y^{1\top}X^1/\sigma_1^2, \ldots, Y^{Q\top}X^Q/\sigma_Q^2\right]^\top$ and

$$A = \Sigma^{-1} \otimes I_D + \text{diag}\left(\sigma_1^{-2}X^{1\top}X^1, \ldots, \sigma_Q^{-2}X^{Q\top}X^Q\right).$$

A naive approach to computing $\vec{\mu}$ could then be to explicitly compute A^{-1} and report the matrix vector product, $A^{-1}b$. However, as mentioned in Section 2.3.1, since A is a $DQ \times DQ$ matrix, explicitly computing its inverse would require roughly $O(D^3Q^3)$ time. This operation becomes very cumbersome when D and Q are too large; for instance if D and Q are in the hundreds the, DQ is is the tens of thousands.

CG provides an exact solution to linear systems in at most DQ iterations, with each iteration requiring only a small constant number of matrix vector multiplications by A. This characteristic does not provide a complexity improvement for solving general linear systems because for dense, unstructured $DQ \times DQ$ matrices, matrix vector multiplies require $O(D^2Q^2)$ time, and CG still demands $O(D^3Q^3)$ time overall. However this property provides a substantial benefit in our setting. In particular, the special form of A allows computation of matrix vector multiplications in $O(D^2Q)$ rather than $O(D^2Q^2)$ time, and storage of this matrix with $O(D^2Q)$ rather than $O(D^2Q^2)$ memory. Specifically, if $v = [v_1, v_2, \ldots, v_Q]$ is a $D \times Q$ matrix with D-vector columns v_q, for the DQ-vector $\vec{v} = [v_1^\top, v_2^\top, \ldots, v_Q^\top]^\top$ we can compute $A\vec{v}$ as $\texttt{vec}\left(v\Sigma^{-1}\right) + [\sigma_1^{-2}X^{1\top}X^1v_1, \ldots, \sigma_Q^{-2}X^{Q\top}X^Qv_Q]^\top$, where $\texttt{vec}(\cdot)$ represents the operation of reshaping an $D \times Q$ matrix into a DQ-vector by stacking its columns. When $D > Q$, this operation is dominated by the Q $O(D^2)$ matrix-vector multiplications to compute the second term. As such, CG provides an order Q improvement in both time and memory.

Next, CG may be viewed as an iterative optimization method. At each step it provides an iterate which is the closest to the $\vec{\mu}$ on a Krylov subspace of expanding dimension. As such, the algorithm may be terminated after fewer than DQ steps to provide an approximation of the solution. Moreover, the algorithm may be provided with an initial estimate, and improves upon that estimate in each successive iteration. In our case we may readily compute a good initialization. For example, we can

initialize with the posterior mean of the parameter for each group when conditioning on that group alone, i.e. $\vec{\mu}^{(0)} := \left[\mathbb{E}[\beta^1|Y^1]^\top, \ldots, \mathbb{E}[\beta^Q|Y^Q]^\top\right]^\top$.

Finally, the convergence properties of the conjugate gradient algorithm are well understood. Notably the ith iterate of conjugate gradient $\vec{\mu}^{(i)}$ when initialized at $\vec{\mu}^{(0)}$ satisfies

$$\|\vec{\mu}^{(i+1)} - \vec{\mu}\|_A \leq 2\left(\frac{\kappa - 1}{\kappa + 1}\right)^i \|\vec{\mu}^{(0)} - \vec{\mu}\|_A,$$

where $\kappa = \sqrt{\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}}$ is the square root of the condition number of A, and $\|\cdot\|_A$ is the $A-$quadratic norm [Nocedal and Wright, 2006, Chapter 5.1], [Luenberger, 1973]. Since A will often be reasonably well conditioned (note, for example, that $\lambda_{\min}(A) \geq \lambda_{\min}(\Sigma)$), convergence can be rapid. Notably, in an unpublished application the authors encountered (not described in this work) involving $D \approx 20,000$ covariates and $Q \approx 50$ groups, the approximately million dimensional estimate $\vec{\mu}$ was computed in roughly 10 minutes on a 16 core machine.

A.2.3 Expectation maximization algorithm further details

In Sections 2.3.2 and 2.3.3 we introduced EM algorithms for estimating Σ for both linear and logistic regression models. In this subsection we provide a derivation of the updates in Algorithm 1 and discuss computational details of our fast implementation.

Derivations of EM updates for linear regression. Our notation inherits directly from [McLachlan and Krishnan, 2007, Chapter 1.5], to which we refer the reader for context. In our application of the EM algorithm, we take the collection of all covariate

effects β as the 'missing data.' For the expectation (E) step, we therefore require

$$\begin{aligned} Q(\Sigma, \Sigma^{(i)}) :&= \mathbb{E}[\log p(\beta|\Sigma)|\mathcal{D}, \Sigma^{(i)}] \\ &= c + \frac{D}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{d=1}^{D}\mathbb{E}[\beta_d^\top \Sigma^{-1}\beta_d|\mathcal{D}, \Sigma^{(i)}] \\ &= c + \frac{D}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{d=1}^{D}\mathrm{tr}\left(\Sigma^{-1}\mathbb{E}[\beta_d\beta_d^\top|\mathcal{D}, \Sigma^{(i)}]\right) \\ &= c + \frac{D}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{d=1}^{D}\mathrm{tr}\left(\Sigma^{-1}(\mu_d\mu_d^\top + V_d)\right), \end{aligned} \tag{A.1}$$

where c is a constant that does not depend on Σ, $\mu = [\mu_1 \ldots, \mu_D]^\top := \mathbb{E}[\beta|\mathcal{D}, \Sigma^{(i)}]$ and for each d $V_d := (I_Q \otimes e_d)^\top \mathrm{Var}[\vec{\beta}|\mathcal{D}, \Sigma^{(i)}](I_Q \otimes e_d)$. From the last line of Eq. (A.1) we may see that μ and $\{V_d\}_{d=1}^D$, comprise the required posterior expectations.

The solution to the maximization step may then be found by considering a first order condition for maximizing over Σ^{-1} rather than Σ. Observe that $\frac{\partial}{\partial\Sigma^{-1}}Q(\Sigma, \Sigma^{(i)}) = \frac{D}{2}\Sigma - \frac{1}{2}\sum_{d=1}^{D}(\mu_d\mu_d^\top + V_d)$. Setting this to zero we obtain $\Sigma^{(i+1)} = D^{-1}\sum(\mu_d\mu_d^\top + V_d)$. This is the desired update for the M-step provided in Algorithm 2.

Logistic regression EM updates. The updates for the approximate EM algorithm described in Section 2.3 are derived from a Gaussian approximation to the posterior under which the expectation of log prior is taken. In particular we approximate the first line of Eq. (A.1) as

$$\begin{aligned} Q(\Sigma, \Sigma^{(i)}) :&= \mathbb{E}[\log p(\beta|\Sigma)|\mathcal{D}, \Sigma^{(i)}] \\ &= \int p(\beta|\mathcal{D}, \Sigma^{(i)})\log p(\beta|\Sigma)d\beta \\ &\approx \int q^{(i)}(\beta)\log p(\beta|\Sigma)d\beta \end{aligned} \tag{A.2}$$

where $q^{(i)}$ denotes the Laplace approximation to $p(\beta|\mathcal{D}, \Sigma^{(i)})$. Specifically, as we summarized in Algorithm 3, we approximate the posterior mean by the maximum a posteriori estimate, $\vec{\mu}^* := \arg\max_{\vec{\beta}}\log p(\vec{\beta}|\mathcal{D}, \Sigma^{(i)})$, and the posterior variance by $V := -[\nabla^2_{\vec{\beta}}\log p(\vec{\beta}|\mathcal{D}, \Sigma^{(i)})\big|_{\vec{\beta}=\vec{\mu}^*}]^{-1}$. We the let $q^{(i)}$ be the Gaussian density with these

moments. This renders the integral in the last line of Eq. (A.2) tractable, and updates are derived in the same way as in the linear case.

Naively, the approximate EM algorithm for logistic regression could be much more demanding than its counterpart in the linear case. In particular, at each iteration we need to solve a convex optimization problem, rather than linear system. However, in practice the algorithm is only little more demanding because, by using the maximum a posteriori estimate from the previous iteration to initialize the optimization, we can solve the optimization problem very easily. In particular, after the first few EM iterations, only one or two additional Newton steps from this initialization are required.

To simplify our implementation, we used automatic differentiation in `Tensorflow` to compute gradients and Hessians when computing the maximum a posteriori values and Laplace approximations.

Computational efficiency. We have employed several tricks to provide a fast implementation of our EM algorithms. The M-Steps for both linear and logistic regression involve a series of expensive matrix operations. To accelerate this, we used `Tensorflow`[Abadi et al., 2016] to optimize these steps by way of a computational graph representation generated using the `@tf.function` decorator in python. Additionally, we initialize EM with a moment based estimate (see Appendix A.5.2).

A.3 Frequentist properties of exchangeability among covariate effects – supplementary proofs and discussion

A.3.1 Discussion of Condition 2.4.1

The restriction on the design matrices in Condition 2.4.1 places strong limits the immediate scope of our theoretical results. However, as with many statistical assumptions such as Gaussianity of residuals, this condition lends considerable tractability to the problem that enables us to build insights that we can see hold in more relaxed

settings in experiments (see Section 2.6).

Under Condition 2.4.1 estimation of the parameter β may be reduced to a special matrix valued case of the normal means problem with each $\hat{\beta}^q_{\mathrm{LS},d} \sim \mathcal{N}(\beta^q_d, \sigma^2)$. Accordingly, we may recognize σ^2 as a reflection of both the residual variances σ^2_q and sample sizes N_q. In particular, if within each group q the covariates have sample second moment $N_q^{-1}\sum_{n=1}^{N_q} X^q_n X^{q\top}_n = I_D$, and the residual variances and sample sizes are equal (i.e. $\sigma^2_1 = \sigma^2_2 = \cdots = \sigma^2_Q$ and $N^1 = N^2 = \cdots = N^Q$), then $\sigma^2 = \sigma^2_1/N^1$. Additionally, because $\hat{\beta}_{\mathrm{LS}}$ is a sufficient statistic of $\mathcal{D}$ for β, it suffices to consider $\hat{\beta}_{\mathrm{LS}}$ alone, without needing to consider other aspects of $\mathcal{D}$. For these reasons, conditions of this sort are commonly assumed by other authors in related settings (e.g. van Wieringen [2015, Chapters 1.4 and 6.2] and Fan and Li [2001], Golan and Perloff [2002]).

That the trends predicted by our theoretical results persist beyond the limits of Condition 2.4.1 should not be surprising. The likelihood, our estimators and their risks are all continuous in the X^q, and so domination results may be seen to extends via continuity to settings with well-conditioned designs. On the other hand, problems with design matrices that are more poorly conditioned are more challenging for both theory and estimation in practice (see e.g. Brown and Zidek [1980][Example 4.2]).

A.3.2 A proposition on analytic forms of the risks of moment estimators

The following proposition characterizes analytic expressions for the moment based estimators. These expressions provide a starting point for the theory in Section 2.4

Proposition A.3.1. *Assume each* $Y^q_n | X^q_n, \beta^q \sim \mathcal{N}(X^{q\top}_n \beta^q, \sigma^2_q)$ *and define* $\hat{\Sigma}^{\mathrm{MM}} := D^{-1}\hat{\beta}^\top_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}} - D^{-1}\mathrm{diag}(\sigma^2_1\|X^{1\dagger}\|^2_F, \ldots, \sigma^2_Q\|X^{Q\dagger}\|^2_F)$. *Then*

1. *if each* $\beta_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$, $\mathbb{E}[\hat{\Sigma}^{\mathrm{MM}}] = \Sigma$.

Furthermore, under Condition 2.4.1

2. *when* $D \geq Q$, $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}} = \hat{\beta}_{\mathrm{LS}} - \sigma^2 D\hat{\beta}^{\dagger\top}_{\mathrm{LS}}$ *and*

3. when $D \leq Q$, $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}} = \hat{\beta}_{\mathrm{LS}} - \sigma^2 Q \hat{\beta}^{\dagger\top}_{\mathrm{LS}}$,

where $\dagger$ *denotes the Moore-Penrose pseudoinverse of a matrix.*

Proof. We begin with statement (1), that under Condition 2.4.1 and correct prior specification, $\mathbb{E}[\hat{\Sigma}^{\mathrm{MM}}] = \Sigma$.

Recall that $\hat{\Sigma}^{\mathrm{MM}} := D^{-1}\hat{\beta}^{\top}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}} - D^{-1}\mathrm{diag}(\sigma_1^2\|X^{1\dagger}\|_F^2, \ldots, \sigma_Q^2\|X^{Q\dagger}\|_F^2)$. For any fixed β, we have $\mathbb{E}[\hat{\Sigma}^{\mathrm{MM}}|\beta] = D^{-1}\mathbb{E}[\hat{\beta}^{\top}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}}|\beta] - D^{-1}\mathrm{diag}(\sigma_1^2\|X^{1\dagger}\|_F^2, \ldots, \sigma_Q^2\|X^{Q\dagger}\|_F^2)$, and so seek to characterize $\mathbb{E}[\hat{\beta}^{\top}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}}|\beta]$. Note that we may write $\hat{\beta}_{\mathrm{LS}} \stackrel{d}{=} \beta + \epsilon$ for a random $D \times Q$ matrix ϵ with each column q distributed as $\epsilon^q \overset{indep.}{\sim} \mathcal{N}\left[0, \sigma_q^2(X^{q\top}X^q)^{-1}\right]$. As such, for each q we have $\mathbb{E}[\hat{\beta}^{q\top}_{\mathrm{LS}}\hat{\beta}^q_{\mathrm{LS}}|\beta] = \beta^{q\top}\beta^q + \mathbb{E}[\epsilon^{q\top}\epsilon^q]$. Next observe that $\mathbb{E}[\epsilon^{q\top}\epsilon^q] = \mathrm{tr}[\sigma_q^2(X^{q\top}X^q)^{-1}] = \sigma_q^2\|X^{q\dagger}\|_F^2$, where $\dagger$ denotes the pseudo-inverse of a matrix and $\|\cdot\|_F$ is the Frobenius norm. Additionally, for $q \neq q'$, we have $\mathbb{E}[\hat{\beta}^{q\top}_{\mathrm{LS}}\hat{\beta}^{q'}_{\mathrm{LS}}|\beta] = \beta^{q\top}\beta^{q'}$. Putting these together into matrix form, we see $\mathbb{E}[\hat{\beta}^{\top}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}}|\beta] = \beta^{\top}\beta + \mathrm{diag}(\sigma_1^2\|X^{1\dagger}\|_F^2, \ldots, \sigma_Q^2\|X^{Q\dagger}\|_F^2)$, and so $\mathbb{E}[\hat{\Sigma}^{\mathrm{MM}}|\beta] = D^{-1}\beta^{\top}\beta$. Under the additional assumption that for each d, $\beta_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$, we have that $\mathbb{E}[D^{-1}\beta^{\top}\beta] = \Sigma$, and (1) obtains from the law of iterated expectation.

We next prove statement (2), that $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}} := \mathbb{E}[\beta|\mathcal{D}, \hat{\Sigma}^{\mathrm{MM}}] = \hat{\beta}_{\mathrm{LS}} - \sigma^2 D\hat{\beta}^{\dagger\top}_{\mathrm{LS}}$. Consider the singular value decomposition (SVD), $\hat{\beta}_{\mathrm{LS}} = V\mathrm{diag}(\lambda^{\frac{1}{2}})U^{\top}$. Under Condition 2.4.1 substituting this expression into $\hat{\Sigma}^{\mathrm{MM}}$ provides $\hat{\Sigma}^{\mathrm{MM}} = D^{-1}U\mathrm{diag}(\lambda)U^{\top} - \sigma^2 I_Q$. Therefore, Lemma A.3.1 provides that we may write

$$
\begin{aligned}
\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}} &:= \mathbb{E}[\beta|\mathcal{D}, \hat{\Sigma}^{\mathrm{MM}}] \\
&= \hat{\beta}_{\mathrm{LS}} - \hat{\beta}_{\mathrm{LS}}\left[\sigma^{-2}\hat{\Sigma}^{\mathrm{MM}} + I_Q\right]^{-1} \\
&= \hat{\beta}_{\mathrm{LS}} - V\mathrm{diag}(\lambda^{\frac{1}{2}})U^{\top}\left[\sigma^{-2}(D^{-1}U\mathrm{diag}(\lambda)U^{\top} - \sigma^2 I_Q) + I_Q\right]^{-1}U^{\top} \\
&= \hat{\beta}_{\mathrm{LS}} - V\mathrm{diag}\left[\lambda^{\frac{1}{2}} \odot (\sigma^{-2}D^{-1}\lambda)^{-1}\right]U^{\top} \\
&= \hat{\beta}_{\mathrm{LS}} - \sigma^2 DV\mathrm{diag}(\lambda^{-\frac{1}{2}})U^{\top} \\
&= \hat{\beta}_{\mathrm{LS}} - \sigma^2 D\hat{\beta}^{\dagger\top}_{\mathrm{LS}},
\end{aligned}
$$

where $\odot$ is the Hadamard (i.e. elementwise) product, as desired.

We lastly prove (3), that the analogous moment based estimator constructed under

the assumption of a priori exchangeability among groups is $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}} = \hat{\beta}_{\mathrm{LS}} - \sigma^2 Q \hat{\beta}^{\dagger\top}_{\mathrm{LS}}$. We begin by making explicit the assumed model and estimate. Specifically we assume each $\beta^q \overset{i.i.d.}{\sim} \mathcal{N}(0, \Gamma)$ a priori, where Γ is a $D \times D$ covariance matrix.

In this case, we obtain an unbiased moment based estimate of Γ as $\hat{\Gamma}^{\mathrm{MM}} := Q^{-1}\hat{\beta}_{\mathrm{LS}}\hat{\beta}^{\top}_{\mathrm{LS}} - Q^{-1}\sum_{q=1}^{Q}\sigma_q^2 (X^{q\top}X^q)^{-1}$. Following an argument exactly parallel to the one in the proof of (1), we find that under the prior $\beta^q \overset{i.i.d.}{\sim} \mathcal{N}(0, \Gamma)$, we have $\mathbb{E}[\hat{\Gamma}^{\mathrm{MM}}] = \Gamma$. Furthermore, following an argument exactly parallel to the one in the proof of (2), we find that under Condition 2.4.1 the corresponding empirical Bayes estimate $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}} := \mathbb{E}[\beta|\hat{\Gamma}^{\mathrm{MM}}] = \hat{\beta}_{\mathrm{LS}} - \sigma^2 Q \hat{\beta}^{\dagger\top}_{\mathrm{LS}}$. We omit full details to spare repetition. □

Lemma A.3.1. *Under Condition 2.4.1* $\mathbb{E}[\beta|\mathcal{D}, \Sigma] = \hat{\beta}_{\mathrm{LS}} - \hat{\beta}_{\mathrm{LS}}\left[\sigma^{-2}\Sigma + I_Q\right]^{-1}$.

Proof. By Proposition 2.3.1, we have

$$\mathbb{E}[\vec{\beta}|\mathcal{D}, \Sigma] = V\left[\frac{Y^{1\top}X^1}{\sigma_1^2}, \ldots, \frac{Y^{Q\top}X^Q}{\sigma_Q^2}\right] \quad \text{where}$$
$$V^{-1} = \Sigma^{-1} \otimes I_D + \mathrm{diag}(\frac{X^{1\top}X^1}{\sigma_1^2}, \ldots, \frac{X^{Q\top}X^Q}{\sigma_Q^2}).$$

Under Condition 2.4.1, we can simplify this as

$$\begin{aligned}
\mathbb{E}[\vec{\beta}|\mathcal{D}, \Sigma] &= \left[\Sigma^{-1} \otimes I_D + \mathrm{diag}(\frac{X^{1\top}X^1}{\sigma_1^2}, \ldots, \frac{X^{Q\top}X^Q}{\sigma_Q^2})\right]^{-1}\left[\frac{Y^{1\top}X^1}{\sigma_1^2}, \ldots, \frac{Y^{Q\top}X^Q}{\sigma_Q^2}\right] \\
&= \left[\Sigma^{-1} \otimes I_D + \sigma^{-2} I_{DQ}\right]^{-1}\sigma^{-2}\left[\hat{\beta}^1_{\mathrm{LS}}, \ldots, \hat{\beta}^Q_{\mathrm{LS}}\right] \\
&= \left[\sigma^2\Sigma^{-1} \otimes I_D + I_{DQ}\right]^{-1}\left[\hat{\beta}^1_{\mathrm{LS}}, \ldots, \hat{\beta}^Q_{\mathrm{LS}}\right].
\end{aligned}$$

As a result, for each d, $\mathbb{E}[\beta_d|\mathcal{D}, \Sigma] = \left[\sigma^2\Sigma^{-1} + I_Q\right]^{-1}\beta_{\mathrm{LS},d}$ and so, in matrix form, we may write

$$\begin{aligned}
\mathbb{E}[\beta|\mathcal{D}, \Sigma] &= \hat{\beta}_{\mathrm{LS}}\left[\sigma^2\Sigma^{-1} + I_Q\right]^{-1} \\
&= \hat{\beta}_{\mathrm{LS}} - \hat{\beta}_{\mathrm{LS}}\left[I_Q + \sigma^{-2}\Sigma\right]^{-1}.
\end{aligned}$$

□

A.3.3 Proof of Lemma 2.4.1

Proof. We prove the lemma in two parts; first for the case that $D > Q + 1$, and then for the case that $Q \leq D \leq Q + 1$.

Our proof for the case that $D > Q+1$ relies on an expression for the squared error risk for estimators of the form $\hat{\beta} = \hat{\beta}_{\mathrm{LS}} - \sigma^2 c\hat{\beta}_{\mathrm{LS}}^{\dagger\top}$ for real c. In particular, Lemma A.3.2 provides that when $D > Q + 1$ and under Condition 2.4.1,

$$\mathbb{E}[\|\beta - (\hat{\beta}_{\mathrm{LS}} - c\hat{\beta}_{\mathrm{LS}}^{\dagger\top})\|_F^2 \mid \beta] = DQ + \sigma^4 c(c + 2 + 2Q - 2D)\mathbb{E}[\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 \mid \beta].$$

Notably, since under Condition 2.4.1, by Proposition A.3.1 we have that $\hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}} = \hat{\beta}_{\mathrm{LS}} - \sigma^2 D\hat{\beta}_{\mathrm{LS}}^{\dagger\top}$ we obtain $\mathbb{E}[\|\beta - \hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}}\|_F^2 \mid \beta] = \sigma^2 DQ - \sigma^4 D(D - 2Q - 2)\mathbb{E}[\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 \mid \beta]$, as desired.

We next consider $Q \leq D \leq Q + 1$. In this case, both $\mathrm{R}(\beta, \hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}})$ and $\sigma^2 DQ - \sigma^4 D(D - 2Q - 2)\mathbb{E}[\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 \mid \beta]$ are positive infinity. In particular, observe that $\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 = \mathrm{tr}[(\hat{\beta}_{\mathrm{LS}}^{\top}\hat{\beta}_{\mathrm{LS}})^{-1}]$ is the trace of the inverse of a non-central Wishart matrix, which is known to have infinite expectation for $Q \leq D \leq Q + 1$ (see e.g. Hillier and Kan [2019]). Likewise, Lemma A.3.5 reveals that $\mathrm{R}(\beta, \hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}}) = \infty$ as well.

The second assertion of Lemma 2.4.1, that when $D \leq Q$ and under Condition 2.4.1 $\mathbb{E}[\|\beta - \hat{\beta}_{\mathrm{EGroup}}^{\mathrm{MM}}\|_F^2 \mid \beta] = \sigma^2 DQ - \sigma^4 Q(Q - 2D - 2)\mathbb{E}[\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 \mid \beta]$, obtains similarly. Specifically, under these conditions an identical argument to that provided in Lemma A.3.2 provides that

$$\mathbb{E}[\|\beta - (\hat{\beta}_{\mathrm{LS}} - \sigma^2 c\hat{\beta}_{\mathrm{LS}}^{\dagger\top})\|_F^2 \mid \beta] = DQ + \sigma^4 c(c + 2 + 2D - 2Q)\mathbb{E}[\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 \mid \beta]$$

when $D < Q - 1$. The desired expression is then obtained by taking $c = Q$ to reflect $\hat{\beta}_{\mathrm{EGroup}}^{\mathrm{MM}} = \hat{\beta}_{\mathrm{LS}} - \sigma^2 Q\hat{\beta}_{\mathrm{LS}}^{\dagger\top}$, again as specified by Proposition A.3.1. □

Lemma A.3.2. *Let $D > Q+1$ and let $\hat{\beta} = \hat{\beta}_{\mathrm{LS}} - \sigma^2 c\hat{\beta}_{\mathrm{LS}}^{\dagger\top}$. Then under Condition 2.4.1* $\mathbb{E}[\|\beta - \hat{\beta}\|_F^2 \mid \beta] = \sigma^2 DQ + \sigma^4 c(c + 2 + 2Q - 2D)\mathbb{E}[\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 \mid \beta]$.

Proof. The results follows by considering Stein's unbiased risk estimate (SURE) [Lehmann and Casella, 2006, Chapter 4, Corollary 7.2] (restated as Lemma A.3.3)

and making several algebraic simplifications. In order to apply the lemma, we note that under Condition 2.4.1 $\hat{\vec{\beta}}_{\mathrm{LS}} \sim \mathcal{N}(\vec{\beta}, \sigma^2 I_{DQ})$ and $\hat{\vec{\beta}} = \hat{\vec{\beta}}_{\mathrm{LS}} - g(\hat{\vec{\beta}}_{\mathrm{LS}})$ for $g(\hat{\vec{\beta}}_{\mathrm{LS}}) = -\sigma^2 c \cdot \mathtt{vec}(\hat{\beta}_{\mathrm{LS}}^{\dagger\top})$, where $\mathtt{vec}(\cdot)$ represents the operation of reshaping an $D \times Q$ matrix into a DQ-vector by stacking its columns.

We first simplify the sum of partial derivatives in Eq. (A.3) of Lemma A.3.3. Observe that

$$\sum_{n=1}^{DQ} \frac{\partial g_n(\hat{\vec{\beta}}_{\mathrm{LS}})}{\partial \hat{\vec{\beta}}_{\mathrm{LS},n}} = -\sigma^2 c \sum_{d=1}^{D} \sum_{q=1}^{Q} \frac{\partial \hat{\beta}_{\mathrm{LS},d}^{\dagger,q}}{\partial \hat{\beta}_{\mathrm{LS},d}^{q}},$$

where $\hat{\beta}_{\mathrm{LS},d}^{\dagger,q}$ denotes the entry in the qth row and dth column of $\hat{\beta}_{\mathrm{LS}}^{\dagger}$.

Next, letting e_q be the qth basis vector in $\mathbb{R}^Q$, for each q and d we may write

$$\begin{aligned}
\frac{\partial \hat{\beta}_{\mathrm{LS},d}^{\dagger,q}}{\partial \hat{\beta}_{\mathrm{LS},d}^{q}} &= \frac{\partial}{\partial \hat{\beta}_{\mathrm{LS},d}^{q}} \hat{\beta}_{\mathrm{LS},d} (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} e_q \\
&= e_q^{\top} (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} e_q + \hat{\beta}_{\mathrm{LS},d} \frac{\partial}{\partial \hat{\beta}_{\mathrm{LS},d}^{q}} (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} e_q \\
&= e_q^{\top} (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} e_q - \hat{\beta}_{\mathrm{LS},d}^{\top} (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} \left[\frac{\partial}{\partial \hat{\beta}_{\mathrm{LS},d}^{q}} (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}}) \right] (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} e_q \\
&= \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2 - \hat{\beta}_{\mathrm{LS},d}^{\dagger\top} \left[e_q \hat{\beta}_{\mathrm{LS},d}^{\top} + \hat{\beta}_{\mathrm{LS},d} e_q^{\top} \right] (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} e_q \\
&= \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2 - \left[\hat{\beta}_{\mathrm{LS},d}^{\dagger\top} e_q \hat{\beta}_{\mathrm{LS},d}^{\top} (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} e_q + \hat{\beta}_{\mathrm{LS},d}^{\dagger\top} \hat{\beta}_{\mathrm{LS},d} e_q^{\top} (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} e_q \right] \\
&= \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2 - (\hat{\beta}_{\mathrm{LS},d}^{\dagger,q})^2 - \hat{\beta}_{\mathrm{LS},d}^{\dagger\top} \hat{\beta}_{\mathrm{LS},d} \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2,
\end{aligned}$$

where in the fourth and last lines we have used that $e_q^{\top} (\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} e_q = \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2$, as can be seen by observing that $(\hat{\beta}_{\mathrm{LS}}^{\top} \hat{\beta}_{\mathrm{LS}})^{-1} = \hat{\beta}_{\mathrm{LS}}^{\dagger} \hat{\beta}_{\mathrm{LS}}^{\dagger\top}$.

Adding these terms together we find

$$\begin{aligned}
\sum_{d=1}^{D} \sum_{q=1}^{Q} \frac{\partial \hat{\beta}_{\mathrm{LS},d}^{\dagger,q}}{\partial \hat{\beta}_{\mathrm{LS},d}^{q}} &= \sum_{d=1}^{D} \sum_{q=1}^{Q} \left\{ \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2 - (\hat{\beta}_{\mathrm{LS},d}^{\dagger,q})^2 - \hat{\beta}_{\mathrm{LS},d}^{\dagger\top} \hat{\beta}_{\mathrm{LS},d} \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2 \right\} \\
&= D\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 - \|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 - \|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 \sum_{d=1}^{D} \hat{\beta}_{\mathrm{LS},d}^{\dagger\top} \hat{\beta}_{\mathrm{LS},d} \\
&= D\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 - \|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 - \|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 \mathrm{tr}(\hat{\beta}_{\mathrm{LS}}^{\dagger} \hat{\beta}_{\mathrm{LS}}) \\
&= (D - Q - 1)\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2.
\end{aligned}$$

We next note that the regularity condition required by Lemma A.3.3 is satisfied, as demonstrated in Lemma A.3.4, and so we may write

$$\begin{aligned}\mathbb{E}[\|\beta - \hat{\beta}\|_F^2 \mid \beta] &= \sigma^2 DQ + \mathbb{E}[\|g(\hat{\vec{\beta}}_{\mathrm{LS}})\|^2 \mid \beta] - 2\sigma^2 \sum_{d=1}^{D}\sum_{q=1}^{Q} \mathbb{E}[\frac{\partial \hat{\beta}^{\dagger,q}_{\mathrm{LS},d}}{\partial \hat{\beta}^{q}_{\mathrm{LS},d}} \mid \beta] \\ &= \sigma^2 DQ + \sigma^4 c^2 \mathbb{E}[\|\hat{\beta}^{\dagger}_{\mathrm{LS}}\|^2 \mid \beta] - 2\sigma^4 c(D - Q - 1)\mathbb{E}[\|\hat{\beta}^{\dagger}_{\mathrm{LS}}\|_F^2 \mid \beta] \\ &= \sigma^2 DQ + \sigma^4 c(c + 2 + 2Q - 2D)\mathbb{E}[\|\hat{\beta}^{\dagger}_{\mathrm{LS}}\|^2 \mid \beta].\end{aligned}$$

as desired. □

Lemma A.3.3 (Stein's Unbiased Risk Estimate – Lehmann and Casella Corollary 7.2)**.** *Let $X \sim \mathcal{N}(\theta, \sigma^2 I_N)$, and let the estimator $\hat{\theta}$ be of the form $\hat{\theta} = X - g(X)$ where $g(X) = [g_1(X), g_2(X), \ldots, g_N(X)]$ is differentiable. If $\mathbb{E}[|\frac{\partial}{\partial X_n} g_n(X)|] < \infty$ for each $n = 1, \ldots, N$, then*

$$\mathrm{R}(\theta, \hat{\theta}) = \sigma^2 N + \mathbb{E}[\|g(X)\|^2] - 2\sigma^2 \sum_{n=1}^{N} \frac{\partial}{\partial X_n} g_n(X). \tag{A.3}$$

Lemma A.3.4. *Let $D > Q + 1$. Then under Condition 2.4.1 $\mathbb{E}\left[\left|\frac{\partial \hat{\beta}^{\dagger,q}_{\mathrm{LS},d}}{\partial \hat{\beta}^{q}_{\mathrm{LS},d}}\right| \mid \beta\right] \leq \infty$ for each d and q.*

Proof. From our derivation of $\frac{\partial \hat{\beta}^{\dagger,q}_{\mathrm{LS},d}}{\partial \hat{\beta}^{q}_{\mathrm{LS},d}}$ in Lemma A.3.2 we have that

$$\begin{aligned}\frac{\partial \hat{\beta}^{\dagger,q}_{\mathrm{LS},d}}{\partial \hat{\beta}^{q}_{\mathrm{LS},d}} &= \|\hat{\beta}^{\dagger,q}_{\mathrm{LS}}\|^2 - (\hat{\beta}^{\dagger,q}_{\mathrm{LS},d})^2 - \hat{\beta}^{\dagger\top}_{\mathrm{LS},d}\hat{\beta}_{\mathrm{LS},d}\|\hat{\beta}^{\dagger,q}_{\mathrm{LS}}\|^2 \\ &= \|\hat{\beta}^{\dagger,q}_{\mathrm{LS}}\|^2 - (\hat{\beta}^{\dagger,q}_{\mathrm{LS},d})^2 - \|\hat{\beta}^{\dagger,q}_{\mathrm{LS}}\|^2 \mathrm{tr}[(\hat{\beta}^{\top}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}})^{-1}\beta_{\mathrm{LS},d}\beta^{\top}_{\mathrm{LS},d}].\end{aligned}$$

As such we have that

$$\begin{aligned}\left|\frac{\partial \hat{\beta}^{\dagger,q}_{\mathrm{LS},d}}{\partial \hat{\beta}^{q}_{\mathrm{LS},d}}\right| &\leq \|\hat{\beta}^{\dagger,q}_{\mathrm{LS}}\|^2 + |(\beta^{\dagger,q}_{\mathrm{LS},d})^2| + \|\hat{\beta}^{\dagger,q}_{\mathrm{LS}}\|^2 |\mathrm{tr}[(\hat{\beta}^{\top}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}})^{-1}\beta_{\mathrm{LS},d}\beta^{\top}_{\mathrm{LS},d}]| \\ &\leq \|\hat{\beta}^{\dagger,q}_{\mathrm{LS}}\|^2 + \left|\sum_{d'=1}^{D}(\beta^{\dagger,q}_{\mathrm{LS},d'})^2\right| + \|\hat{\beta}^{\dagger,q}_{\mathrm{LS}}\|^2 \left|\mathrm{tr}[(\hat{\beta}^{\top}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}})^{-1}\sum_{d'=1}^{D}\beta_{\mathrm{LS},d'}\beta^{\top}_{\mathrm{LS},d'}]\right|\end{aligned}$$

$$
\begin{aligned}
&= \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2 + \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2 + \|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2 \mathrm{tr}[(\hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}})^{-1} \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}}] \\
&\leq (2+Q)\|\hat{\beta}_{\mathrm{LS}}^{\dagger,q}\|^2 \\
&\leq (2+Q)\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 \\
&= (2+Q)\mathrm{tr}[(\hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}})^{-1}].
\end{aligned}
$$

We next recognize that under Condition 2.4.1, $(\hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}})^{-1}$ is the inverse of a non-central Wishart matrix with non-centrality parameter β. Therefore, from Hillier and Kan [2019, Theorem 1], we have that for $D > Q+1$, $\mathbb{E}\left[\mathrm{tr}\left((\hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}})^{-1}\right) \mid \beta\right] < \infty$. Accordingly, we may conclude that $\mathbb{E}\left[\left|\frac{\partial \hat{\beta}_{\mathrm{LS},d}^{\dagger,q}}{\partial \hat{\beta}_{\mathrm{LS},d}^{q}}\right| \mid \beta\right] \leq \infty$ as desired. □

Lemma A.3.5. *Assume $Q \leq D \leq Q+1$. For any β, $\mathrm{R}(\beta, \hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}}) = \infty$.*

Proof. First observe that we may lower bound $L(\beta, \hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}})$ as

$$
\begin{aligned}
L(\beta, \hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}}) &= \|\hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}} - \beta\|_F^2 \\
&= \|\sigma^2 D \hat{\beta}_{\mathrm{LS}}^{\dagger\top} + \beta - \hat{\beta}_{\mathrm{LS}}\|_F^2 \\
&= \sigma^4 D^2 \|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 + \|\beta - \hat{\beta}_{\mathrm{LS}}\|_F^2 - 2\sigma^2 D \mathrm{tr}\left[-\hat{\beta}_{\mathrm{LS}}^{\dagger}(\beta - \hat{\beta}_{\mathrm{LS}})\right] \\
&\geq \sigma^4 D^2 \|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F^2 + \|\beta - \hat{\beta}_{\mathrm{LS}}\|_F^2 - 2\sigma^2 D \|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F \|\beta - \hat{\beta}_{\mathrm{LS}}\|_F \\
&= (\sigma^2 D \|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F - \|\beta - \hat{\beta}_{\mathrm{LS}}\|_F)^2
\end{aligned}
$$

where the inequality follows from Cauchy-Schwarz. We next consider any constant $c < \sigma^2 D$ and write

$$
\begin{aligned}
\mathrm{R}(\beta, \hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}}) &= \mathbb{E}[L(\beta, \hat{\beta}_{\mathrm{ECov}}^{\mathrm{MM}}) | \beta] \\
&= \mathbb{P}(c\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F \geq \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F) \mathbb{E}[L(\beta, \hat{\beta}_{\mathrm{EGroup}}^{\mathrm{MM}}) \mid \beta, c\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F \geq \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F] \\
&+ \mathbb{P}(c\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F < \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F) \mathbb{E}[L(\beta, \hat{\beta}_{\mathrm{EGroup}}^{\mathrm{MM}}) \mid \beta, c\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F < \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F] \\
&\geq \mathbb{P}(c\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F \geq \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F) \mathbb{E}[L(\beta, \hat{\beta}_{\mathrm{EGroup}}^{\mathrm{MM}}) \mid \beta, c\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F \geq \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F] \\
&\geq \mathbb{P}(c\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F \geq \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F) \\
&\cdot \mathbb{E}[(\sigma^2 D \|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F - \|\beta - \hat{\beta}_{\mathrm{LS}}\|_F)^2 \mid \beta, c\|\hat{\beta}_{\mathrm{LS}}^{\dagger}\|_F \geq \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F]
\end{aligned}
$$

$$\geq \mathbb{P}(c\|\hat{\beta}^{\dagger}_{\mathrm{LS}}\|_F \geq \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F)(\sigma^2 D - c)^2 \mathbb{E}[\|\hat{\beta}^{\dagger}_{\mathrm{LS}}\|_F^2 \mid \beta, c\|\hat{\beta}^{\dagger}_{\mathrm{LS}}\|_F \geq \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F]$$
$$\geq (\sigma^2 D - c)^2 \mathbb{P}(c\|\hat{\beta}^{\dagger}_{\mathrm{LS}}\|_F \geq \|\hat{\beta}_{\mathrm{LS}} - \beta\|_F)\mathbb{E}[\mathrm{tr}[(\hat{\beta}^{\top}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}})^{-1} \mid \beta] = \infty$$

where the last line comes from recognizing $(\hat{\beta}^{\top}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}})^{-1}$ as the inverse of a non-central Wishart matrix, the trace of which has infinite expectation for $Q \leq D \leq Q + 1$. $\square$

A.3.4 Proof of Theorem 2.4.2 and additional details

Proof. The first domination result of Theorem 2.4.2 follows closely from Lemma 2.4.1. Under Condition 2.4.1, $\hat{\beta}_{\mathrm{LS}} \stackrel{d}{=} \beta + \sigma\epsilon$ for a random matrix ϵ with i.i.d. standard normal entries, and so we can see $\mathrm{R}(\beta, \hat{\beta}_{\mathrm{LS}}) = \sum_{d=1}^{D}\sum_{q=1}^{Q}\mathbb{E}[(\sigma\epsilon_d^q)^2] = DQ\sigma^2$. Next, $D > 2Q + 2$ implies that $D - 2 - 2Q > 0$ so that $D(D - 2 - 2Q)\sigma^2\|\hat{\beta}^{\dagger}_{\mathrm{LS}}\|_F^2$ is almost surely positive, and therefore positive in expectation. We therefore obtain the result from Lemma 2.4.1.

We next consider the second domination result. The performance of $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ may be seen to degrade in stages as we transition from a few covariates and many groups regime to a many covariates and few groups regime. When $D < Q/2 - 1$, we can see that $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ has good performance. In fact, by an argument analogous to our proof of the first part of Theorem 2.4.2 above, we can see that $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ dominates $\hat{\beta}_{\mathrm{LS}}$; Specifically, from Lemma 2.4.1 we can recognize $\mathrm{R}(\beta, \hat{\beta}_{\mathrm{LS}}) - \mathrm{R}(\beta, \hat{\beta}_{\mathrm{EGroup}})$ as the expectation of an almost surely positive quantity.

When $D = Q/2 - 1$ we have $Q(Q - 2 - 2D) = 0$, and so regardless of β, the estimators $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ and $\hat{\beta}_{\mathrm{LS}}$ have equal risk, and neither dominates.

Relative performance degrades further in the intermediate regime of $Q/2 - 1 < D < Q - 1$. In this regime, $\mathrm{R}(\beta, \hat{\beta}_{\mathrm{LS}}) - \mathrm{R}(\beta, \hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}) = \sigma^4 Q(Q - 2 - 2D)\mathbb{E}[\|\hat{\beta}^{\dagger}_{\mathrm{LS}}\|_F^2 \mid \beta]$ may be written as the expectation of an almost surely negative quantity, and so $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ is dominated by $\hat{\beta}_{\mathrm{LS}}$.

The situation is even worse when $Q - 1 \leq D \leq Q$; appealing again the they symmetry between $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ and $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$, we can see that by Lemma A.3.5 $\mathrm{R}(\beta, \hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}) = \infty$.

Finally, when $D > Q$ the expression $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}} = \hat{\beta}_{\mathrm{LS}} - \left[\sigma^{-2}\hat{\Gamma}^{\mathrm{MM}} - I_D\right]^{-1} \hat{\beta}_{\mathrm{LS}}$ involves the inverse of a low rank matrix since under Condition 2.4.1, $\hat{\Gamma}^{\mathrm{MM}} = Q^{-1}\hat{\beta}_{\mathrm{LS}}\hat{\beta}_{\mathrm{LS}}^\top - \sigma^2 I_D$. Accordingly we take as our convention $\|\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}\| = \infty$, analogously to defining $\frac{1}{0} = \infty$; as a result $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ has infinite risk in this second regime as well, and we see that this estimator is dominated by $\hat{\beta}_{\mathrm{LS}}$ whenever $D < Q/2 - 1$.

□

With the strong parallels established by Proposition A.3.1 and Lemma 2.4.1 under Condition 2.4.1, we can see that this is not a result of $\hat{\beta}^{\mathrm{MM}}_{\mathrm{EGroup}}$ being singularly bad. Indeed, if we consider the many groups regime with $Q > D$, we can obtain analogous results to demonstrate the superiority of an exchangeability among groups approach.

A.3.5 Proof of Lemma 2.4.2

Proof. We first show that under Condition 2.4.1, $\hat{\Sigma} = U\mathrm{diag}\left[(D^{-1}\lambda - \sigma^2\mathbf{1}_Q)_+\right]U^\top$ is the maximum marginal likelihood estimate of Σ in Eq. (2.1). Our approach is to first derive a lower bound on the negative log likelihood, and then show that this bound is met with equality by the proposed expression.

For convenience, we consider a scaling of the negative log likelihood,

$$-2D^{-1}\ln p(\hat{\beta}_{\mathrm{LS}}|\Sigma) = \ln|\Sigma + \sigma^2 I_Q| + D^{-1}\mathrm{tr}\left[(\Sigma + \sigma^2 I_Q)^{-1}\hat{\beta}_{\mathrm{LS}}^\top\hat{\beta}_{\mathrm{LS}}\right],$$

and are interested in deriving a lower bound on

$$\min_{\Sigma\succeq 0} \ln|\Sigma + \sigma^2 I_Q| + D^{-1}\mathrm{tr}\left[(\Sigma + \sigma^2 I_Q)^{-1}\hat{\beta}_{\mathrm{LS}}^\top\hat{\beta}_{\mathrm{LS}}\right],$$

where the notation $\Sigma \succeq 0$ reflects that the minimum is taken over the space of positive semidefinite matrices.

The problem simplifies if we parameterize the minimization with the eigendecomposition $\Sigma = V^\top\mathrm{diag}(\nu)V$, where V is a $Q \times Q$ matrix satisfying $V^\top V = I_Q$ and ν is a Q-vector of non-negative reals. In particular, if we define $\mathcal{L}(V,\nu) := -2D^{-1}\ln p(\hat{\beta}_{\mathrm{LS}}|\Sigma = V^\top\mathrm{diag}(\nu)V)$ then, leaving the constraints on V and ν implicit,

we have

$$\begin{aligned}
\min_{V,\nu} \mathcal{L}(V,\nu) &= \min_{V,\nu} \ln |V^\top \mathrm{diag}(\nu) V + \sigma^2 I_Q| + D^{-1} \mathrm{tr}\left[(V^\top \mathrm{diag}(\nu) V + \sigma^2 I_Q)^{-1} \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} \right] \\
&= \min_{V,\nu} \ln |V^\top \mathrm{diag}(\nu) V + \sigma^2 I_Q| + D^{-1} \mathrm{tr}\left[(\mathrm{diag}(\nu) + \sigma^2 I_Q)^{-1} V \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V^\top \right] \\
&= \min_{V,\nu} \sum_{q=1}^{Q} \ln(\nu_q + \sigma^2) + D^{-1} \sum_{q=1}^{Q} \frac{1}{\nu_q + \sigma^2} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q \\
&= \min_{V} \sum_{q=1}^{Q} \min_{\nu_q \geq 0} \ln(\nu_q + \sigma^2) + \frac{D^{-1} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q}{\nu_q + \sigma^2}.
\end{aligned}$$

Next, Lemma A.3.6 provides that we may solve the inner optimization problems over ν in the line above analytically to get $\nu^* := \arg\min_\nu \mathcal{L}(V,\nu)$ with entries $\nu_q^* = \max(\sigma^2, D^{-1} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q) - \sigma^2$. Substituting these values in, we obtain

$$\begin{aligned}
\min_{V,\nu} \mathcal{L}(V,\nu) &= \min_{V} \sum_{q=1}^{Q} \ln \left[\max(\sigma^2, D^{-1} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q) \right] + \frac{D^{-1} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q}{\max(\sigma^2, D^{-1} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q)} \\
&= \min_{V} \sum_{q=1}^{Q} \ln \left[\max(\sigma^2, D^{-1} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q) \right] + \sigma^{-2} \min(\sigma^2, D^{-1} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q).
\end{aligned}$$

We can now further simplify the problem by considering the eigendecomposition of $\hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} = U \mathrm{diag}(\lambda) U^\top$, and recognizing that because VU satisfies $(VU)^\top VU = I_Q$ we may write

$$\begin{aligned}
\min_{V,\nu} \mathcal{L}(V,\nu) &= \min_{V} \sum_{q=1}^{Q} \ln \left[\max(\sigma^2, D^{-1} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q) \right] + \sigma^{-2} \min(\sigma^2, D^{-1} V_q^\top \hat{\beta}_{\mathrm{LS}}^\top \hat{\beta}_{\mathrm{LS}} V_q) \\
&= \min_{V} \sum_{q=1}^{Q} \ln \left[\max\left(\sigma^2, V_q^\top \mathrm{diag}(D^{-1}\lambda) V_q\right) \right] + \sigma^{-2} \min\left[\sigma^2, V_q^\top \mathrm{diag}(D^{-1}\lambda) V_q\right].
\end{aligned}$$

Finally, we obtain a lower bound by recognizing $\{V_q^\top \mathrm{diag}(D^{-1}\lambda) V_q\}_{q=1}^Q$ as the diagonals of $D^{-1} V \mathrm{diag}(\lambda) V^\top$ and applying Lemma A.3.7 to obtain that

$$-2D^{-1} \ln p(\hat{\beta}_{\mathrm{LS}} | \Sigma) \geq \sum_{q=1}^{Q} \ln \left[\max(\sigma^2, D^{-1}\lambda_q) \right] + \sigma^{-2} \min(\sigma^2, D^{-1}\lambda_q)$$

for every $\Sigma \succeq 0$.

We next show that this bound is met with equality by $\hat{\Sigma} = U\mathrm{diag}\left[(D^{-1}\lambda - \sigma^2\mathbf{1}_Q)_+\right]U^\top$, the form given in the statement of Lemma 2.4.2. Recognize first that $\hat{\Sigma} + \sigma^2 I_Q = U\mathrm{diag}\left[\max(\sigma^2\mathbf{1}_Q, D^{-1}\lambda)\right]U^\top$. Substituting this expression in, we find

$$\begin{aligned}
-2D^{-1}\ln p(\hat{\beta}_{\mathrm{LS}}|\hat{\Sigma}) &= \ln|\hat{\Sigma} + \sigma^2 I_Q| + D^{-1}\mathrm{tr}\left[(\hat{\Sigma} + \sigma^2 I_Q)^{-1}\hat{\beta}_{\mathrm{LS}}^\top\hat{\beta}_{\mathrm{LS}}\right] \\
&= \ln\left|\mathrm{diag}\left[\max(\sigma^2\mathbf{1}_Q, D^{-1}\lambda)\right]\right| \\
&\quad + D^{-1}\mathrm{tr}\left[\mathrm{diag}\left[\max(\sigma^2\mathbf{1}_Q, D^{-1}\lambda)\right]^{-1}U^\top\hat{\beta}_{\mathrm{LS}}^\top\hat{\beta}_{\mathrm{LS}}U\right] \\
&= \sum_{q=1}^{Q}\ln\left[\max(\sigma^2, D^{-1}\lambda_q)\right] + D^{-1}\lambda_q/\max(\sigma^2, D^{-1}\lambda_q) \\
&= \sum_{q=1}^{Q}\ln\left[\max(\sigma^2, D^{-1}\lambda_q)\right] + \sigma^{-2}\min(\sigma^2, D^{-1}\lambda_q),
\end{aligned}$$

which meets our lower bound. This establishes that the maximum marginal likelihood estimate is $\hat{\Sigma} = U\left[(D^{-1}\lambda - \sigma^2\mathbf{1}_Q)_+\right]U^\top$, as desired.

It now remains to show that, under Condition 2.4.1,

$$\hat{\beta}_{\mathrm{ECov}} = V\mathrm{diag}\left[\lambda^{\frac{1}{2}} \odot (\mathbf{1}_Q - \sigma^2 D\lambda^{-1})_+\right]U^\top.$$

By Lemma A.3.1, we have that $\hat{\beta}_{\mathrm{ECov}} = \hat{\beta}_{\mathrm{LS}} - \hat{\beta}_{\mathrm{LS}}\left[I_Q + \sigma^{-2}\hat{\Sigma}\right]^{-1}$. Substituting in the analytic expression for $\hat{\Sigma}$, recalling the SVD $\hat{\beta}_{\mathrm{LS}} = V\mathrm{diag}(\lambda^{\frac{1}{2}})U^\top$, and rearranging, we obtain

$$\begin{aligned}
\hat{\beta}_{\mathrm{ECov}} &= V\mathrm{diag}(\lambda^{\frac{1}{2}})U^\top - V\mathrm{diag}(\lambda^{\frac{1}{2}})U^\top\left\{I_Q + \sigma^{-2}U\left[(D^{-1}\lambda - \sigma^2\mathbf{1}_Q)_+\right]U^\top\right\}^{-1} \\
&= V\mathrm{diag}\left\{\lambda^{\frac{1}{2}} - \lambda^{\frac{1}{2}}\left[\mathbf{1}_Q + \sigma^{-2}(D^{-1}\lambda - \sigma^2\mathbf{1}_Q)_+\right]^{-1}\right\}U^\top \\
&= V\mathrm{diag}\left\{\lambda^{\frac{1}{2}} \odot \left[\mathbf{1}_Q - \left(\mathbf{1}_Q + (\sigma^{-2}D^{-1}\lambda - \mathbf{1}_Q)_+\right)^{-1}\right]\right\}U^\top \\
&= V\mathrm{diag}\left[\lambda^{\frac{1}{2}} \odot \left(\mathbf{1}_Q - \sigma^2 D\lambda^{-1}\right)_+\right]U^\top,
\end{aligned}$$

as desired. □

Lemma A.3.6. *For any $c > 0$,*

$$\begin{aligned}\nu^* :&= \arg\min_{\nu \geq 0} \ln(\nu + \sigma^2) + \frac{c}{\nu + \sigma^2} \\ &= \max(\sigma^2, c) - \sigma^2\end{aligned}$$

Proof. Define $g(x) := \ln(x+\sigma^2)+c/(x+\sigma^2)$ and $f(x) := g(\sigma^2 x) = \ln(x+1)+\frac{\sigma^{-2}c}{x+1}+\ln\sigma^2$ to lighten notation. Now $\nu^* = \arg\max_{x\geq 0} g(x) = \sigma^2 \arg\max_{x\geq 0} f(x)$. Denote by f' and f'' the first two derivatives of f. Notably, $f'(x) = (x+1)^{-1}\left[1 - \sigma^{-2}c/(x+1)\right]$ and $f''(x) = (x+1)^{-2}\left[2\sigma^{-2}c/(x+1) - 1\right]$. The result may be seen by separately considering the cases of $\sigma^{-2}c < 1$ and $\sigma^{-2}c \geq 1$.

If $\sigma^{-2}c < 1$, then f' is positive on $\mathbb{R}_+$, and so $\arg\min_{x\in\mathbb{R}_+} f(x) = 0$. On the other hand, if $\sigma^{-2}c \geq 1$, then f has a local minimum at $x = \sigma^{-2}c - 1$ (note that $f'(\sigma^{-2}c-1) = 0$, and $f''(\sigma^{-2}c-1) > 0$)). Since this is the only local minimum on $\mathbb{R}_+$, and with the positive second derivative at the this minimum, we can conclude that in this case $\arg\min_{x\in\mathbb{R}_+} f(x) = \sigma^{-2}c - 1$. In either case, we can write $\arg\min_{x\in\mathbb{R}_+} f(x) = \max(1, \sigma^{-2}c) - 1$. Therefore, as desired, we see that $\arg\min_{x\in\mathbb{R}_+} g(x) = \max(\sigma^2, c) - \sigma^2$. □

Lemma A.3.7. *Let A be a $Q \times Q$ Hermitian matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_Q$. Then*

$$\sum_{q=1}^{Q} \ln\left[\max(\sigma^2, A_{q,q})\right] + \sigma^{-2}\min(\sigma^2, A_{q,q}) \geq \sum_{q=1}^{Q} \ln\left[\max(\sigma^2, \lambda_q)\right] + \sigma^{-2}\min(\sigma^2, \lambda_q).$$

Proof. First note that $f(x) = \ln\max(\sigma^2, x) + \min(\sigma^2, x)$ is concave on $\mathbb{R}_+$, and so the vector valued function, $g(x_1, x_2, \ldots, x_N) = \sum_{n=1}^{N} f(x_n)$ is Schur concave. By the Schur-Horn theorem (Theorem A.4.1) the diagonals of A are majorized by its eigenvalues, when each are sorted in descending order. As such $g\left(\mathrm{diag}(A)\right) \geq g\left(\lambda\right)$, as desired. □

A.3.6 Proof of Theorem 2.4.3

Our approach to showing dominance of $\hat{\beta}_{\text{ECov}}$ over $\hat{\beta}^{\text{MM}}_{\text{ECov}}$ parallels the classical approach of Baranchik [1964], to showing that the positive part James-Stein estimator dominates the original James-Stein estimator. In this case, however, our parameter and estimates are matrix-valued, rather than vector-valued. Additionally, we contend with the added complication that the directions along which we apply shrinkage are random.

Proof. To begin, consider again the SVD of the matrix of least squares estimates, $\hat{\beta}_{\text{LS}} = V\text{diag}(\lambda^{\frac{1}{2}})U^\top$. Recall from Proposition A.3.1 that $\hat{\beta}^{\text{MM}}_{\text{ECov}} = \hat{\beta}_{\text{LS}} - \sigma^2 D\hat{\beta}^{\dagger\top}_{\text{LS}}$ under Condition 2.4.1. Because the pseudo-inverse of $\hat{\beta}_{\text{LS}}$ may be written as $\hat{\beta}^{\dagger}_{\text{LS}} = U\text{diag}(\lambda^{-\frac{1}{2}})V^\top$, we rewrite $\hat{\beta}^{\text{MM}}_{\text{ECov}} = V\text{diag}(\lambda^{\frac{1}{2}} - \sigma^2 D\lambda^{-\frac{1}{2}})U^\top$. Comparing this estimate to the expression for $\hat{\beta}_{\text{ECov}}$ in Lemma 2.4.2, $\hat{\beta}_{\text{ECov}} = V\text{diag}\left[\lambda^{\frac{1}{2}} \odot (1 - \sigma^2 D\lambda^{-1})_+\right] U^\top$, we see that the two estimates differ only when $\hat{\beta}^{\text{MM}}_{\text{ECov}}$ "flips the direction" of one or more of the singular values of $\hat{\beta}_{\text{LS}}$. Our strategy to proving the theorem is to show that analogously to the "over-shrinking" of the James-Stein estimator relative to the positive part James-Stein estimator, this "over-shrinking" of singular values increases the loss of $\hat{\beta}^{\text{MM}}_{\text{ECov}}$ in expectation.

For convenience, we define $\rho := \lambda^{\frac{1}{2}} \odot (1 - \sigma^2 D\lambda^{-1})$ and $\rho_+ := \lambda^{\frac{1}{2}} \odot (1 - \sigma^2 D\lambda^{-1})_+$ so that $\hat{\beta}^{\text{MM}}_{\text{ECov}} = V\text{diag}(\rho)U^\top$ and $\hat{\beta}_{\text{ECov}} = V\text{diag}(\rho_+)U^\top$.

To show the desired uniform risk improvement we must show that for any β,

$$\mathbb{E}\left[\text{L}(\beta, \hat{\beta}^{\text{MM}}_{\text{ECov}}) - \text{L}(\beta, \hat{\beta}_{\text{ECov}})\right] > 0, \tag{A.4}$$

where $\text{L}(\beta, \hat{\beta}) = \|\hat{\beta} - \beta\|_F^2$ is squared error loss. We can rewrite this difference in loss as

$$\begin{aligned}
\text{L}(\beta, \hat{\beta}^{\text{MM}}_{\text{ECov}}) - \text{L}(\beta, \hat{\beta}_{\text{ECov}}) &= \|\hat{\beta}^{\text{MM}}_{\text{ECov}} - \beta\|_F^2 - \|\hat{\beta}_{\text{ECov}} - \beta\|_F^2 \\
&= \|\text{diag}(\rho) - V^\top \beta U\|_F^2 - \|\text{diag}(\rho_+) - V^\top \beta U\|_F^2 \\
&= \sum_{q=1}^{Q} (\rho_q - V_q^\top \beta U_q)^2 - (\rho_{+q} - V_q^\top \beta U_q)^2
\end{aligned}$$

$$= \sum_{q=1}^{Q} \rho_q^2 - \rho_{+q}^2 - 2(V_q^\top \beta U_q)(\rho_q - \rho_{+q}),$$

where we here (and in the proof of this theorem only) write V_q and U_q to denotes columns of V and U, rather than rows. Since $\rho_q^2 \overset{a.s.}{\geq} \rho_{+q}^2$, it suffices to show that for any β and each q,

$$\mathbb{E}\left[(V_q^\top \beta U_q)(\rho_q - \rho_{+q})\right] < 0.$$

To show this, we again find an even narrower but easier to prove condition will imply the one above; since ρ_q and ρ_{+q} differ only when $\lambda_q < \sigma^2 D$, it is enough to show that for each $0 < c < \sigma^2 D$

$$\mathbb{E}\left[(V_q^\top \beta U_q)\rho_q | \lambda_q = c\right] < 0. \tag{A.5}$$

If we establish Eq. (A.5), then Eq. (A.4) obtains from the law of iterated expectation. Next, observe that since ρ_q fixed and negative when $\lambda_q = c < \sigma^2 D$, Eq. (A.4) is equivalent to

$$\mathbb{E}\left[V_q^\top \beta U_q | \lambda_q = c\right] > 0.$$

Letting U_{-q} and V_{-q} denote the remaining columns of U and V, respectively, we may write

$$\mathbb{E}\left[V_q^\top \beta U_q | \lambda_q = c\right] = \mathbb{E}\left[\mathbb{E}\left[V_q^\top \beta U_q | \lambda_q = c, U_{-q}, V_{-q}\right]\right]$$

and, again through the law of iterated expectation, see that it will be sufficient to show for every U_{-q} and V_{-q} that $\mathbb{E}\left[V_q^\top \beta U_q | \lambda_q = c, U_{-q}, V_{-q}\right] > 0$.

With all but one column of each of U and V fixed, U_q and V_q are determined up to signs, as unit vectors in the one dimensional subspaces orthogonal to $[\{U^{q'}\}_{q' \neq q}]$ and $[\{V^d\}_{d \neq q}]$. As such, we need only to show

$$\mathbb{P}\left[V_q^\top \beta U_q > 0 | U_{-q}, V_{-q}, \lambda_q = c\right] > \mathbb{P}\left[V_q^\top \beta U_q < 0 | U_{-q}, V_{-q}, \lambda_q = c\right], \tag{A.6}$$

since

$$\mathbb{E}\left[V_q^\top \beta U_q | \lambda_q, U_{-q}, V_{-q}\right] = |V_q^\top \beta U_q| \left\{ \mathbb{P}\left[V_q^\top \beta U_q > 0 | \lambda_q, U_{-q}, V_{-q}\right] - \mathbb{P}\left[V_q^\top \beta U_q < 0 | \lambda_q, U_{-q}, V_{-q}\right] \right\},$$

where, in an abuse of notation, we have moved $|V_q^\top \beta U_q|$ outside the expectation since it is deterministic once we have observed V_{-q} and U_{-q}.

That Eq. (A.6) holds may be seen from considering the conditional probability densities for U_q and V_q, and noting that the density is larger for V_q and U_q such that $V_q^\top \beta U_q$ is positive. In particular, we have that

$$\begin{aligned} \ln p(\hat{\beta}_{\mathrm{LS}} | \beta, U_{-q}, V_{-q}, \lambda) &= -\frac{1}{2}\|\beta - \hat{\beta}\|_F^2 + h \\ &= -\frac{1}{2}\|V^\top \beta U - \mathrm{diag}(\lambda^{\frac{1}{2}})\|_F^2 + h \\ &= -\frac{1}{2}(\lambda_q^{\frac{1}{2}} - V_q^\top \beta U_q)^2 + h' \end{aligned}$$

where h and h' are constants that do not depend on the signs of U_q and V_q. Since $\lambda_q^{\frac{1}{2}}$ is positive with probability one, the conditional probability that $V_q^\top \beta U_q$ is positive is greater than that it is negative. Accordingly, we see that Eq. (A.5) does in fact hold, and the result obtains. □

A.4 Gains from ECov in the high-dimensional limit – supplementary proofs

A.4.1 Proof of Lemma 2.5.1

From the sequence of datasets, $\{\mathcal{D}_D\}_{D=1}^\infty$, we obtain sequences of estimates. To make explicit the dimension dependence, we denote these as explicit functions of the data, e.g. $\{\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D)\}_{D=1}^\infty$ where $\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D)$ denotes $\hat{\beta}_{\mathrm{ECov}}$ in Eq. (2.1) applied to $\mathcal{D}_D$. Furthermore, we consider the entire sequence of datasets and estimates as existing in a single probability space.

We note that Lemma A.4.1 establishes that $\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D)$ and $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}(\mathcal{D}_D)$ coincide almost surely in the high-dimensional limit. As such, the squared error loss of these two estimates coincide almost surely in the limit, and we may write

$$
\begin{aligned}
&\lim_{D\to\infty} D^{-1}\mathrm{R}^D_\pi(\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D)) \\
&\quad= \lim_{D\to\infty} D^{-1}\mathbb{E}\left[\mathbb{E}[\|\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D) - \beta\|_F^2 \mid \beta]\right] \\
&\quad= \lim_{D\to\infty} D^{-1}\mathbb{E}\big[\mathbb{E}[\|\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}(\mathcal{D}_D) - \beta\|_F^2 + \|\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D) - \hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}(\mathcal{D}_D)\|_F^2 + \\
&\quad 2\mathrm{tr}((\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D) - \hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}(\mathcal{D}_D))^\top(\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}(\mathcal{D}_D) - \beta)) \mid \beta]\big] \\
&\quad= \lim_{D\to\infty} \mathbb{E}\left[D^{-1}\mathbb{E}[\|\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}(\mathcal{D}_D) - \beta\|_F^2 \mid \beta]\right] \\
&\quad= \lim_{D\to\infty} \mathbb{E}\left[\sigma^2 Q - \sigma^4(D - 2Q - 2)\mathbb{E}[\|\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\dagger\|_F^2|\beta]\right] \\
&\quad= \sigma^2 Q - \sigma^4 \lim_{D\to\infty} \mathbb{E}[(D - 2Q - 2)\|\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\dagger\|_F^2] \\
&\quad= \sigma^2 Q - \sigma^4 \lim_{D\to\infty} \mathbb{E}[\mathrm{tr}[(\tilde{\Sigma} + \sigma^2 I_Q)^{-1}] + o(1)] \\
&\quad= \sigma^2 Q - \sigma^4 \mathrm{tr}[(\tilde{\Sigma} + \sigma^2 I_Q)^{-1}].
\end{aligned}
$$

The third line comes from linearity of expectation and that $\|\hat{\beta}_{\mathrm{ECov}} - \hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}\| \overset{a.s.}{\to} 0$. The fourth line comes from Lemma 2.4.1. The second to last line comes from Lemma A.4.2.

We next recognize that $\mathrm{tr}[(\tilde{\Sigma} + \sigma^2 I_Q)^{-1}] = \sum_{q=1}^Q (\lambda_q + \sigma^2)^{-1}$, where $\lambda_1, \ldots, \lambda_Q$ are the eigenvalues of $\tilde{\Sigma}$. Accordingly we may write,

$$
\lim_{D\to\infty} D^{-1}\mathrm{R}^D_\pi(\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D)) = \sigma^2 Q - \sigma^4 \sum_{q=1}^Q (\lambda_q + \sigma^2)^{-1}.
$$

Furthermore since we obtain $\hat{\beta}_{\mathrm{ID}}(\mathcal{D}_D)$ by applying $\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D)$ independently to the data in each group, we analogously obtain

$$
\lim_{D\to\infty} D^{-1}\mathrm{R}^D_\pi(\hat{\beta}_{\mathrm{ID}}(\mathcal{D}_D)) = \sigma^2 Q - \sigma^4 \sum_{q=1}^Q (\tilde{\Sigma}_{q,q} + \sigma^2)^{-1}.
$$

Putting these expressions together, we obtain

$$\lim_{D\to\infty} D^{-1}\left[\mathrm{R}_\pi^D(\hat{\beta}_{\mathrm{ID}}(\mathcal{D}_D)) - \mathrm{R}_\pi^D(\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D))\right] = \sigma^4\left[\sum_{q=1}^{Q}(\lambda_q+\sigma^2)^{-1} - \sum_{q=1}^{Q}(\tilde{\Sigma}_{q,q}+\sigma^2)^{-1}\right].$$

Finally, including the additional scaling by $\sigma^{-2}Q^{-1}$ we obtain

$$\mathrm{Gain}(\pi,\sigma^2) = \sigma^2 Q^{-1}\left[\sum_{q=1}^{Q}(\lambda_q+\sigma^2)^{-1} - \sum_{q=1}^{Q}(\tilde{\Sigma}_{q,q}+\sigma^2)^{-1}\right]$$

as desired.

Lemma A.4.1. *Under the conditions of Lemma 2.5.1,*

$$\lim_{D\to\infty} \|\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D) - \hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}(\mathcal{D}_D)\|_F = 0$$

almost surely.

Proof. Note that under the conditions of Lemma 2.5.1, Lemma 2.4.2 provides that $\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D)$ and $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}(\mathcal{D}_D)$ differ only when $\hat{\Sigma}^{\mathrm{MM}}$ is not positive definite; otherwise $\hat{\Sigma}^{\mathrm{MM}} = \hat{\Sigma}$. Since $\hat{\Sigma}^{\mathrm{MM}} = D^{-1}\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\top\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D) - \sigma^2 I_Q$, by Lemma A.4.3 $\hat{\Sigma}^{\mathrm{MM}}$ will be positive definite for all D above some D' almost surely, and so $\hat{\beta}_{\mathrm{ECov}}(\mathcal{D}_D)$ and $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}(\mathcal{D}_D)$ become equal for all D large enough, implying strong convergence. □

Lemma A.4.2. *Under the conditions of Lemma 2.5.1,* $\lim_{D\to\infty} D\|\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\dagger\|_F^2 = \mathrm{tr}[(\tilde{\Sigma}+\sigma^2 I_Q)^{-1}]$ *almost surely.*

Proof. Recall that $\|\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\dagger\|_F^2 = \mathrm{tr}[(\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\top\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D))^{-1}]$. As such, we may write $D\|\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\dagger\|_F^2 = \mathrm{tr}[(D^{-1}\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\top\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D))^{-1}]$. By Lemma A.4.3

$$D^{-1}\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\top\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D) \overset{a.s.}{\to} \tilde{\Sigma} + \sigma^2 I_Q,$$

and so we can see that $D\|\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\dagger\|_F^2 \overset{a.s.}{\to} \mathrm{tr}[(\tilde{\Sigma}+\sigma^2 I_Q)^{-1}]$ as desired. □

Lemma A.4.3. *Under the conditions of Lemma 2.5.1* $\lim_{D\to\infty} D^{-1}\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\top\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D) = \tilde{\Sigma} + \sigma^2 I_Q$ *almost surely.*

Proof. It suffices to show strong convergence element wise, as this implies strong convergence in all other relevant norms. For convenience, let $C^{(D)} := D^{-1}\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\top \hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)$. Note that we may write each entry $C^{(D)}_{q,q'} = \sum_{d=1}^{D} D^{-1}\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^q_d \hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^{q'}_d$ as a sum of D i.i.d. terms. Notably, each term $\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^q_d \cdot \hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^{q'}_d$ is a product of two Gaussian random variables and is therefore sub-exponential with some non-negative parameters (ν, α) (see e.g. Wainwright [2019, Definition 2.7]). As a result, $C^{(D)}$ is then sub-exponential with parameters $(D^{-\frac{1}{2}}\nu, D^{-1}\alpha)$. Therefore, for any constant b satisfying $0 < b < \nu^2/\alpha$, by Wainwright [2019, Proposition 2.9] we have that

$$\mathbb{P}\left[\left|C^{(D)}_{q,q'} - \mathbb{E}[C^{(D)}_{q,q'}]\right| \geq b\right] \leq 2\exp\{-\frac{D}{2}b^2/\nu^2\}.$$

This rapid, exponential decay in tail probability with D implies that for small b,

$$\sum_{D=1}^{\infty} \mathbb{P}\left[\left|C^{(D)}_{q,q'} - \mathbb{E}[C^{(D)}_{q,q'}]\right| \geq b\right] \leq \infty.$$

Therefore, by the Borel-Cantelli lemma we see that $|C^{(D)}_{q,q'} - \mathbb{E}[C^{(D)}_{q,q'}]| \overset{a.s.}{\to} 0$. Since $\mathbb{E}[C^{(D)}] = \tilde{\Sigma} + \sigma^2 I_Q$ for each D, this implies that $\lim_{D\to\infty} D^{-1}\hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D)^\top \hat{\beta}_{\mathrm{LS}}(\mathcal{D}_D) = \tilde{\Sigma} + \sigma^2 I_Q$ almost surely. □

A.4.2 Further discussion of Theorem 2.5.2

We here give further detail related to the proof of Theorem 2.5.2 and introduce additional notation used in the remainder of the section. Recall from Lemma 2.5.1 that $\mathrm{Gain}(\pi, \sigma^2) = \sigma^2 Q^{-1}[\sum_{q=1}^{Q}(\lambda_q + \sigma^2)^{-1} - \sum_{q=1}^{Q}(\tilde{\Sigma}_{q,q} + \sigma^2)^{-1}]$. For convenience, we will use $\ell := \mathrm{diag}(\tilde{\Sigma})^\downarrow$ to denote the Q-vector of diagonal entries of $\tilde{\Sigma}$ sorted in descending order. Similarly, we take λ to be the Q-vector of eigenvalues of $\tilde{\Sigma}$, again sorted in descending order. Next, it is useful to rewrite

$$\mathrm{Gain}(\pi, \sigma^2) = \sigma^2 Q^{-1}\left[\vec{f}(\lambda) - \vec{f}(\ell)\right]$$

where $\vec{f}(x) := \sum_{q=1}^{Q} f(x_q) = \sum_{q=1}^{Q}(\sigma^2 + x_q)^{-1}$ (where $f(x) := (\sigma^2 + x)^{-1}$).

The key theoretical tool used in establishing Theorem 2.5.2 is the Schur-Horn theorem. We state this result below, adapted from Horn [1954, Theorem 5]. The Schur-Horn theorem guarantees that λ majorizes ℓ. In particular, an N-vector a is said to majorize a second N-vector b if $\sum_{n=1}^{N} a_n = \sum_{n=1}^{N} b_n$ and for all $N' \leq N$,

$$\sum_{n=1}^{N'} a_n^{\downarrow} \geq \sum_{n=1}^{N'} b_n^{\downarrow},$$

where for a vector v, we use $v^{\downarrow}$ to denote the vector with the same components as v, sorted in descending order. As captured by Theorem 2.5.2, we can therefore see that $\mathrm{Gain}(\pi, \sigma^2)$ is non-negative for any $\tilde{\Sigma}$ by observing that $\vec{f}$ is Schur-convex (since f is convex).

Theorem A.4.1 (Schur-Horn)**.** *A vector ℓ can be the diagonal of a Hermitian matrix with (repeated) eigenvalues λ if and only if λ majorizes ℓ.*

A.4.3 Proof of Theorem 2.5.3

We here show that $\mathrm{Gain}(\pi, \sigma^2)$ is upper bounded as

$$\begin{aligned}\mathrm{Gain}(\pi, \sigma^2) &\leq \sigma^2 Q^{-1} f''(\lambda_{\min}) \|\lambda\|_2 \|\lambda - \ell\|_2 \\ &= 2\sigma^2 Q^{-1} \|\lambda\|_2 \|\lambda - \ell\|_2 / (\sigma^2 + \lambda_{\min})^3,\end{aligned}$$

and lower bounded as

$$\begin{aligned}\mathrm{Gain}(\pi, \sigma^2) &\geq \frac{1}{2}\sigma^2 Q^{-1} f''(\lambda_{\max}) \|\lambda - \ell\|^2 \\ &= \sigma^2 Q^{-1} \|\lambda - \ell\|^2 / (\sigma^2 + \lambda_{\max})^3,\end{aligned}$$

where $f''(x) := \frac{d^2}{dx^2} f(x)$ where f is as defined in Appendix A.4.2.

We obtain both bounds with quadratic approximations to f. In particular, we define g_α as the 2^{nd} order Taylor approximation of f expanded at α,

$$g_\alpha(x) := f(\alpha) + f'(\alpha)(x - \alpha) + \frac{1}{2} f''(\alpha)(x - \alpha)^2,$$

and note that by Lemma A.4.4

$$\vec{g}_{\lambda_{\max}}(\lambda) - \vec{g}_{\lambda_{\max}}(\ell) \leq \vec{f}(\lambda) - \vec{f}(\ell) \leq \vec{g}_{\lambda_{\min}}(\lambda) - \vec{g}_{\lambda_{\min}}(\ell), \tag{A.7}$$

where $\vec{g}_\alpha(x) := \sum_{q=1}^{Q} g_\alpha(x_q)$.

Proof of upper bound. We obtain the desired upper bound as follows.

Eq. (A.7) and Lemma A.4.5 allow us to see

$$\begin{aligned} \mathrm{Gain}(\pi, \sigma^2) &\leq \sigma^2 Q^{-1} \left[\vec{g}_{\lambda_{\min}}(\lambda) - \vec{g}_{\lambda_{\min}}(\ell)\right] \\ &= \frac{1}{2}\sigma^2 Q^{-1} f''(\lambda_{\min})(\|\lambda\|^2 - \|\ell\|^2). \end{aligned} \tag{A.8}$$

Since f'' is positive on $\mathbb{R}_+$, the problem reduces to upper bounding $\|\lambda\|^2 - \|\ell\|^2$.

In particular, we find

$$\begin{aligned} \|\lambda\|^2 - \|\ell\|^2 &= \langle \lambda + \ell, \lambda - \ell \rangle && (A.9) \\ &\leq \|\lambda + \ell\| \|\lambda - \ell\| && // \text{ by Cauchy-Schwarz} \quad (A.10) \\ &= \sqrt{\|\lambda\|^2 + 2\langle \lambda, \ell \rangle + \|\ell\|^2}\, \|\lambda - \ell\| && (A.11) \\ &\leq \sqrt{\|\lambda\|^2 + 2\|\lambda\|\|\ell\| + \|\ell\|^2}\, \|\lambda - \ell\| && // \text{ by Cauchy-Schwarz} \quad (A.12) \\ &\leq 2\|\lambda\| \|\lambda - \ell\| && // \text{ Since } \|\lambda\| \geq \|\ell\|, \quad (A.13) \end{aligned}$$

where we can see that $\|\lambda\| \geq \|\ell\|$ by noting that $\|\cdot\|^2$ is Schur convex, and again appealing to the Schur-Horn Theorem. The desired upper bound obtains by combining Eqs. (A.8) and (A.9).

Proof of lower bound. We begin as we did for the upper bound. Eq. (A.7) and Lemma A.4.5 allow us to see

$$\begin{aligned} \mathrm{Gain}(\pi, \sigma^2) &\geq \sigma^2 Q^{-1} \left[\vec{g}_{\lambda_{\max}}(\lambda) - \vec{g}_{\lambda_{\max}}(\ell)\right] \\ &= \frac{1}{2}\sigma^2 Q^{-1} f''(\lambda_{\max})(\|\lambda\|^2 - \|\ell\|^2). \end{aligned} \tag{A.14}$$

Since, again, f'' is positive on $\mathbb{R}_+$, the problem reduces to lower bounding $\|\lambda\|^2 - \|\ell\|^2$.

In particular, we would like to show $\|\lambda\|^2 - \|\ell\|^2 \geq \|\lambda - \ell\|^2$. We can arrive at this bound with a particular expansion of $\|\lambda - \ell\|^2$ and using Lemma A.4.6, which again leverages the fact that λ majorizes ℓ. Specifically, we write

$$\begin{aligned}
\|\lambda - \ell\|^2 &= \langle \lambda - \ell, \lambda \rangle - \langle \lambda - \ell, \ell \rangle \\
&= \|\lambda\|^2 - \big[\langle \lambda, \ell \rangle + \langle \lambda - \ell, \ell \rangle\big] \\
&= \|\lambda\|^2 - \|\ell\|^2 - \big[\langle \lambda, \ell \rangle - \langle \ell, \ell \rangle + \langle \lambda - \ell, \ell \rangle\big] \\
&= \|\lambda\|^2 - \|\ell\|^2 - 2\langle \lambda - \ell, \ell \rangle \\
&\leq \|\lambda\|^2 - \|\ell\|^2
\end{aligned} \tag{A.15}$$

where the last line follows from Lemma A.4.6, which provides that $\langle \lambda - \ell, \ell \rangle \geq 0$ since, from the Schur-Horn theorem for any $Q' \leq Q$ $\sum_{q=1}^{Q'} \lambda_q - \ell_q \geq 0$, and ℓ has non-negative, non-increasing entries. We obtain the desired lower bound by combining Eqs. (A.14) and (A.15).

Lemma A.4.4. *Let λ and ℓ be Q-vectors of non-negative reals with non-increasing entries, and let λ majorize ℓ. Consider $\vec{f} : \mathbb{R}^Q \to \mathbb{R}, x \mapsto \sum_{q=1}^{Q} f(x_q) = \sum_{q=1}^{Q} (\sigma^2 + x_q)^{-1}$ (where $f(v) := (\sigma^2 + v)^{-1}$) for any $\sigma^2 > 0$, and define g_α to be the 2^{nd} order Taylor approximation of f expanded at α,*

$$g_\alpha(x) := f(\alpha) + f'(\alpha)(x - \alpha) + \frac{1}{2} f''(\alpha)(x - \alpha)^2.$$

Then

$$\vec{g}_{\lambda_{\max}}(\lambda) - \vec{g}_{\lambda_{\max}}(\ell) \leq \vec{f}(\lambda) - \vec{f}(\ell) \leq \vec{g}_{\lambda_{\min}}(\lambda) - \vec{g}_{\lambda_{\min}}(\ell),$$

where $\vec{g}_\alpha(x) := \sum_{q=1}^{Q} g_\alpha(x_q)$ and $\lambda_{\max} = \lambda_1$ and $\lambda_{\min} = \lambda_Q$ are the largest and smallest entries of λ, respectively.

Proof. If there are indices q for which $\lambda_q = \ell_q$, remove them (they do not affect $\vec{f}(\ell) - \vec{f}(\lambda)$). If all are equal, $\lambda = d$ and so the result is trivial, otherwise we have $Q \geq 2$ entries with $\lambda_q \neq \ell_q$.

We begin with the lower bound; the upper bound follows similarly. For this, it suffices to show $\vec{f}(\lambda) - \vec{f}(\ell) - \big(\vec{g}_{\lambda_{\max}}(\lambda) - \vec{g}_{\lambda_{\max}}(\ell)\big) \geq 0$.

We first express this difference as an inner product

$$\begin{aligned}
&\vec{f}(\lambda) - \vec{f}(\ell) - \big(\vec{g}_{\lambda_{\max}}(\lambda) - \vec{g}_{\lambda_{\max}}(\ell)\big) \\
&\qquad = \sum_{q=1}^{Q} \big[(f - g_{\lambda_{\max}})(\lambda_q) - (f - g_{\lambda_{\max}})(\ell_q)\big] \\
&\qquad = \sum_{q=1}^{Q} (\lambda_q - \ell_q) \left[\frac{(f - g_{\lambda_{\max}})(\lambda_q) - (f - g_{\lambda_{\max}})(\ell_q)}{\lambda_q - \ell_q}\right] \\
&\qquad // \text{ defining each } h_q := \frac{(f - g_{\lambda_{\max}})(\lambda_q) - (f - g_{\lambda_{\max}})(\ell_q)}{\lambda_q - \ell_q} \\
&\qquad = \sum_{q=1}^{Q} (\lambda_q - \ell_q) h_q \\
&\qquad = \langle \lambda - \ell, h \rangle
\end{aligned}$$

where $h = [h_1, h_2, \ldots, h_Q]^\top$.

We will complete our proof by leveraging Lemma A.4.6, which provides that $\langle a, b \rangle \geq 0$ for any Q-vector a satisfying $\sum_{q=1}^{Q} a_q = 0$ and $\sum_{q=1}^{Q'} a_q \geq 0$ for every $Q' \leq Q$, and Q-vector b with non-increasing entries.

It therefore remains only to show that $\lambda - \ell$ and h satisfy the conditions of Lemma A.4.6. Since the entries of λ and ℓ are taken to be in descending order, the condition that $\sum_{q=1}^{Q'} (\lambda - \ell)_q \geq 0$ for any $Q' \leq Q$, follows from the Schur-Horn theorem. Likewise, this theorem provides that $\sum_{q=1}^{Q} \lambda_q = \sum_{q=1}^{Q} \ell_q$, and therefore that $\sum_{q=1}^{Q} (\lambda - \ell)_q = 0$, so that $\lambda - \ell$ meets condition (2) of the lemma.

We next confirm that h has non-increasing entries by considering an expansion of the expressions for each h_q. In particular, observe that

$$\begin{aligned}
h_q &= \frac{(f - g_{\lambda_{\max}})(\lambda_q) - (f - g_{\lambda_{\max}})(\ell_q)}{\lambda_q - \ell_q} \\
&= (\lambda_q - \ell_q)^{-1} \Big\{ f(\lambda_q) - f(\ell_q) - \big[g_{\lambda_{\max}}(\lambda_q) - g_{\lambda_{\max}}(\ell_q)\big] \Big\} \\
&= (\lambda_q - \ell_q)^{-1} \Big\{ \frac{(\sigma^2 + \ell_q) - (\sigma^2 + \lambda_q)}{(\sigma^2 + \ell_q)(\sigma^2 + \lambda_q)}
\end{aligned}$$

$$
\begin{aligned}
&- \left[(\lambda_q - \ell_q) f'(\lambda_{\max}) + \frac{1}{2}((\lambda_q - \lambda_{\max})^2 - (\ell_q - \lambda_{\max})^2) f''(\lambda_{\max})\right] \Big\} \\
&= (\sigma^2 + \lambda_{\max})^{-2} - (\sigma^2 + \ell_q)^{-1}(\sigma^2 + \lambda_q)^{-1} \\
&- \frac{1}{2}(\lambda_q - \ell_q)^{-1}(\sigma^2 + \lambda_{\max})^{-3}\left[\lambda_q^2 - \ell_q^2 - 2\lambda_{\max}(\lambda_q - \ell_q)\right] \\
&= (\sigma^2 + \lambda_{\max})^{-2} - (\sigma^2 + \ell_q)^{-1}(\sigma^2 + \lambda_q)^{-1} - \frac{1}{2}(\sigma^2 + \lambda_{\max})^{-3}\left[\lambda_q + \ell_q - 2\lambda_{\max}\right].
\end{aligned}
$$

Next define $\phi(a,b) = (\sigma^2+\lambda_{\max})^{-2} - (\sigma^2+a)^{-1}(\sigma^2+b)^{-1} - \frac{1}{2}(\sigma^2+\lambda_{\max})^{-3}\left[b + a - 2\lambda_{\max}\right]$, so that for each $q, h_q = \phi(\ell_q, \lambda_q)$. Now, for $q' > q$, we may write

$$
\begin{aligned}
h_{q'} - h_q &= \phi(\ell_{q'}, \lambda_{q'}) - \phi(\ell_q, \lambda_q) \\
&= \int_{\ell_q}^{\ell_{q'}} \frac{\partial}{\partial a}\phi(a, \lambda_q)da + \int_{\lambda_q}^{\lambda_{q'}} \frac{\partial}{\partial b}\phi(\ell_{q'}, b)db.
\end{aligned} \tag{A.16}
$$

Next note that

$$
\frac{\partial}{\partial a}\phi(a,b) = (\sigma^2 + a)^{-2}(\sigma^2 + b)^{-1} - \frac{1}{2}(\sigma^2 + \lambda_{\max})^{-3}
$$

and

$$
\frac{\partial}{\partial b}\phi(a,b) = (\sigma^2 + a)^{-1}(\sigma^2 + b)^{-2} - \frac{1}{2}(\sigma^2 + \lambda_{\max})^{-3}
$$

from which we can see that $\frac{\partial}{\partial a}\phi(a,b)$ and $\frac{\partial}{\partial b}\phi(a,b)$ are positive for $a, b \in [\lambda_{\min}, \lambda_{\max}]$. Accordingly, Eq. (A.16) provides that $h_{q'} - h_q \leq 0$, since $\ell_{q'} \leq \ell_q$ and $\lambda_{q'} \leq \lambda_q$ for $q' > q$, because the entries of ℓ and λ are non-increasing. Therefore $h_{q'} \leq h_q$, completing the proof. □

Lemma A.4.5. *Consider the quadratic function* $\vec{h}(x) = \sum_{q=1}^{Q}(ax_q^2 + bx_q + c)$. *Let* $\lambda, \ell \in \mathbb{R}^Q$ *satisfy* $\sum_{q=1}^{Q}\lambda_q = \sum_{q=1}^{Q}\ell_q$. *Then*

$$
\vec{h}(\ell) - \vec{h}(\lambda) = a(\|\ell\|^2 - \|\lambda\|^2).
$$

Proof. The result follows from the simple algebraic rearrangement below,

$$
\vec{h}(\ell) - \vec{h}(\lambda) = \sum_{q=1}^{Q}(a\ell_q^2 + b\ell_q + c) - (a\lambda_q^2 + b\lambda_q + c)
$$

$$= \sum_{q=1}^{Q} a\ell_q^2 - a\lambda_q^2$$
$$= a(\|\ell\|^2 - \|\lambda\|^2).$$

□

Lemma A.4.6. *Let x be a Q-vector satisfying for each $Q' \leq Q$, $\sum_{q=1}^{Q'} x_q \geq 0$, and let y be a Q-vector with non-increasing entries. If additionally either (1) y has non-negative entries or (2) $\sum_{q=1}^{Q} x_q = 0$ then $\langle x, y\rangle \geq y_Q \sum_{q=1}^{Q} x_q \geq 0$.*

Proof. We first prove the lemma under condition (1) by induction. The base case of $Q = 1$ is trivial; $\langle x, y\rangle = x_1 y_1$ and under (1) x_1 and y_1 are non-negative and under (2) $x_1 = 0$.

Assume the result holds for $Q - 1$. Then

$$\langle x, y\rangle = y_Q x_Q + \langle x_{1:Q-1}, y_{1:Q-1}\rangle \tag{A.17}$$
$$\geq y_Q x_Q + y_{Q-1} \sum_{q=1}^{Q-1} x_q \qquad \text{// by the inductive hypothesis} \tag{A.18}$$
$$\geq y_Q x_Q + y_Q \sum_{q=1}^{Q-1} x_q \qquad \text{// since } y_{Q-1} \geq y_Q \text{ and } \sum_{q=1}^{Q-1} x_q \geq 0 \tag{A.19}$$
$$= y_Q \sum_{q=1}^{Q} x_q \geq 0 \qquad \text{// since } y_Q \text{ and } \sum_{q=1}^{Q} x_q \text{ are non-negative.} \tag{A.20}$$

This provides the desired inductive step, completing the proof under condition (1).

Under condition (2), consider $y' = y - \min_q y_q \mathbf{1}_Q$. Then

$$\langle x, y\rangle = \langle x, y'\rangle + \min_q y_q \langle x, \mathbf{1}_Q\rangle$$
$$= \langle x, y'\rangle.$$

Since y' now has non-negative entries, condition (1) is satisfied and the result follows.

□

A.4.4 Proof of Corollary 2.5.4

We establish the corollary with a brief sequence of upper bounds following from our initial upper bound in Theorem 2.5.2. In particular, the theorem provides

$$\text{Gain}(\pi, \sigma^2) \leq 2\sigma^2 Q^{-1} \|\lambda^{\downarrow}\| \|\ell^{\downarrow} - \lambda^{\downarrow}\| / (\sigma^2 + \lambda_{\min})^3.$$

We begin by simplifying this upper bound. As a first step, note that

$$\begin{aligned} \|\ell^{\downarrow} - \lambda^{\downarrow}\|^2 &= \|\ell\|^2 + \|\lambda\|^2 - 2\langle \ell^{\downarrow}, \lambda^{\downarrow} \rangle \\ &\leq 2\|\lambda\|^2. \end{aligned}$$

As such, we can simplify our upper bound as

$$\begin{aligned} \text{Gain}(\pi, \sigma^2) &\leq 2\sigma^2 Q^{-1} \|\lambda\| \|\ell^{\downarrow} - \lambda^{\downarrow}\| / (\sigma^2 + \lambda_{\min})^3 \\ &\leq 4\sigma^2 Q^{-1} \|\lambda\|^2 / (\sigma^2 + \lambda_{\min})^3 \\ &\leq 4\kappa^2 \lambda_{\min}^2 \sigma^2 / (\sigma^2 + \lambda_{\min})^3 \end{aligned} \tag{A.21}$$

where $\kappa := \lambda_{\max}/\lambda_{\min}$ is the condition number of $\tilde{\Sigma}$.

We then obtain the first bound by noting that

$$\begin{aligned} \lambda_{\min}^2 \sigma^2 / (\sigma^2 + \lambda_{\min})^3 &\leq \lambda_{\min}^2 \sigma^2 / (\sigma^2)^2 / \lambda_{\min} \\ &\leq \lambda_{\min} / \sigma^2 \end{aligned}$$

and the second by noting that

$$\begin{aligned} \lambda_{\min}^2 \sigma^2 / (\sigma^2 + \lambda_{\min})^3 &\leq \lambda_{\min}^2 \sigma^2 / (\lambda_{\min})^3 \\ &\leq \sigma^2 / \lambda_{\min}. \end{aligned}$$

Substituting these expressions into Eq. (A.21) provides the desired expressions in Corollary 2.5.4.

A.4.5 Extensions to random design matrices

The asymptotic formulation in Section 2.5 may allow us to relax Condition 2.4.1. In particular, Theorem 2.5.2 and Theorem 2.5.3 depend on this condition only through Lemma 2.5.1, which provides an analytic expression for the asymptotic gain. We conjecture that this condition may be satisfied for certain sequences of datasets with random design matrices of increasing dimension. For example if for each group q, the number of data points N_D^q grows as $\omega(D^2)$ and if each of the covariates are each distributed as $X_{n,d}^q \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_q^2/(\sigma^2 I_D^2))$, then an asymptotic analogue of Condition 2.4.1 will be satisfied in the sense that $\|\sigma_q^{-2} X^{q\top} X^q - \sigma^2 I_D\|_2$ will be $o(1/\sqrt{D})$ (see e.g. Wainwright [2019, Theorem 6.5]). As a result, we can expect the sequence of estimates $\hat{\beta}_{\text{ECov}}$ to converge to estimates with the simplified form utilized in the proof of Lemma 2.5.1 fast enough that the asymptotic gains are equal in these two cases.

Making this argument rigorous, however, requires contending with convergence of sequences of random variables of changing dimension (recall that we consider $D \to \infty$). This technical aspect complicates the required theoretical analysis because common tools (e.g. continuous mapping theorems) do not apply in this setting. We leave further analysis of $\hat{\beta}_{\text{ECov}}$ with random design matrices to future work.

A.5 Experiments Supplementary Results and Details

A.5.1 Simulations additional details

We here describe the details of the simulated datasets discussed in Section 2.6. For each of the dimensions D and each of the 20 replicates we first generated covariate effects for all $Q = 10$ groups. To do this, we began by setting Σ; for the correlated covariate effects experiments (Figure 2-1 Left) we generating a random $Q \times Q$ matrix of orthonormal vectors U and set $\Sigma = U\text{diag}([2^0, 2^{-1}, \ldots, 2^{Q-1}]^\top)U^\top$, and for independent effects (Figure 2-1 Right) we set $\Sigma = I_Q$. We then simulated covariate effects as $\beta_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$.

We next simulated the design matrices. For each group q, we chose a random number of data points $N^q \sim \text{Pois}(\lambda = 1000)$, and for each data point $n = 1, \ldots, N^q$

sampled $X_n^q \sim \mathcal{N}(0, (1/1000)I_D)$ so that for each group $\mathbb{E}[X^{q\top}X^q] = I_D$. Finally, we generated each response as $Y_n^q \overset{indep}{\sim} \mathcal{N}(X_n^{q\top}\beta^q, 1)$.

For $\hat{\beta}_{\text{EGroup}}$, we estimated the $D \times D$ covariance Γ by maximum marginal likelihood. We did this with an EM algorithm closely related to Algorithm 1. See e.g. Gelman et al. [2013, Chapter 15 sections 4-5] for an explanation of the relevant conjugacy calculations in a more general case that includes a hyper-prior on Γ.

A.5.2 Practical moment estimation for poorly conditioned problems

The moment based estimator (using $\hat{\Sigma}^{\text{MM}}$ in Section 2.4) is unstable in the two real data applications discussed in Section 2.6 due to poor conditioning of the design matrices leading $\hat{\beta}_{\text{LS}}$ to have high variance. To overcome this limitation, we instead used an adapted moment estimation procedure which is less sensitive to this poor conditioning. While, in agreement with Theorem 2.4.3, this approach performs worse than $\hat{\beta}_{\text{ECov}}$ (see Figure A-1) we report it nonetheless because it has lower computational cost and may be appealing for larger scale applications. We describe this approach here. We note however that moment based estimates of the sort we consider here do not naturally extend to logistic regression and so are not reported for our application to CIFAR10.

We first introduce some additional notation. For each group q consider the reduced singular value decomposition $X^q = S^q \text{diag}(\omega^q) R^{q\top}$, where S^q and R^q are $N^q \times D$ and $D \times D$ matrices with orthonormal columns and ω^q is a D-vector of non-negative singular values. Next define for each group $W^q := S^{q\top}X^q$ and $Z^q := S^{q\top}Y^q$, which we may interpret as a $D \times D$ matrix of pseudo-covariates and D-vector of pseudo-responses, respectively. Next define Ω to be the $Q \times Q$ matrix with entries $\Omega_{q,q'} := \text{tr}(W^{q\top}W^{q'})^{-1}$ and $\vec{\sigma}^2 := [\sigma_1^2, \sigma_2^2, \ldots, \sigma_Q^2]^\top$. Lastly, let $Z = [Z^1, Z^2, \ldots, Z^Q]$ be the $D \times Q$ matrix of all pseudo-responses. Our new moment estimator is

$$\hat{\Sigma}^{\text{MM}} := [Z^\top Z - D\text{diag}(\vec{\sigma}^2)] \odot \Omega.$$

We next show hat $\mathbb{E}[\hat{\Sigma}^{\mathrm{MM}}] = \Sigma$ under correct prior and likelihood specification. Note first that if δ is a $D \times Q$ matrix with i.i.d. standard normal entries we may write

$$Z \overset{d}{=} [W^1\beta^1, W^2\beta^2, \ldots, W^Q\beta^Q] + \delta \mathrm{diag}(\vec{\sigma}^2).$$

As such, for each q and q', we have that

$$\begin{aligned}
\mathbb{E}[(Z^\top Z)_{q,q'}] &= \mathbb{E}[Z^{q\top} Z^q] \\
&= \mathbb{E}[\beta^{q\top} W^{q\top} W^{q'} \beta^{q'}] + \mathbb{I}[q = q']\sigma_q^2 D \\
&= \mathrm{tr}(W^{q\top} W^{q'} \mathbb{E}[\beta^{q'} \beta^{q\top}]) + \mathbb{I}[q = q']\sigma_q^2 D \\
&= \Omega^{-1}_{q,q'} \Sigma_{q,q'} + \mathbb{I}[q = q']\sigma_q^2 D.
\end{aligned}$$

Accordingly, we can see that each entry of $\hat{\Sigma}^{\mathrm{MM}}$ has expectation $\mathbb{E}[\hat{\Sigma}^{\mathrm{MM}}_{q,q'}] = \Sigma_{q,q'}$, which establishes unbiasedness.

However, this moment estimate still has the limitation that it evaluates to a non positive semidefinite matrix with positive probability. Under the expectation that, in line with Theorem 2.4.2 the very small and negative eigenvalues of $\hat{\Sigma}^{\mathrm{MM}}$ might lead to over-shrinking, we performed an additional step of clipping these eigenvalues to force the resulting estimate to be reasonably well conditioned. In particular, if our initial estimate had eigendecomposition $\hat{\Sigma}^{\mathrm{MM}} = U\mathrm{diag}(\lambda)U^\top$, we instead used $\hat{\Sigma}^{\mathrm{MM}} = U\mathrm{diag}(\tilde{\lambda})U^\top$, where for each q, we have $\tilde{\lambda}_q = \max(\lambda_q, \lambda_{\max}/100)$ so that the condition number of the modified estimate was at most 100. Though we did not find the performance of the resulting estimates to be very sensitive to this cutoff, we view requirement for these partly subjective implementation choices required to make the $\hat{\beta}^{\mathrm{MM}}_{\mathrm{ECov}}$ effective in practice to be a downside of the approach as compared to $\hat{\beta}_{\mathrm{ECov}}$, which avoids such choices by estimating Σ by maximum marginal likelihood.

Compared to the iterative EM algorithms, which rely on matrix inversions at each iteration, computation of $\hat{\Sigma}^{\mathrm{MM}}$ is much faster. In each of our experiments, computing it requires less than one second.

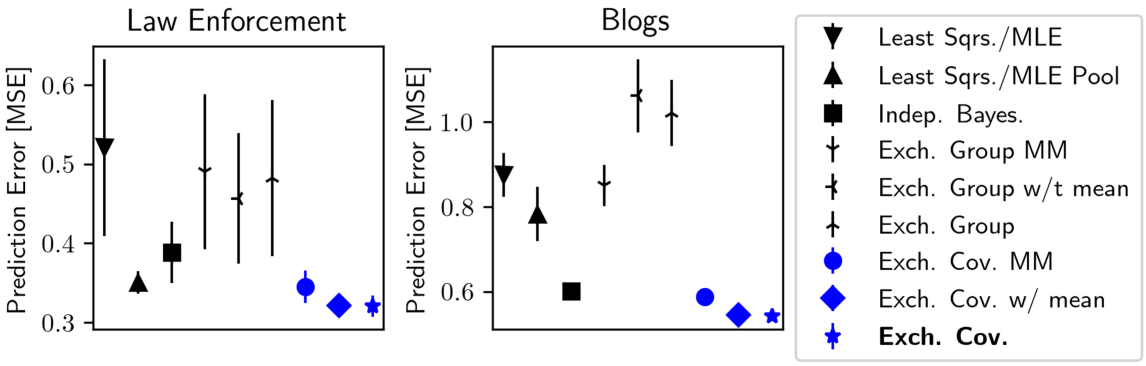


Figure A-1: Performances of additional methods on the law enforcement and blog datasets. Uncertainty intervals are ± 1SEM.

A.5.3 Allowing for non-zero means a priori in hierarchical Bayesian estimates

In the development of our approach in Section 2.2 we imposed the restriction that $\mathbb{E}[\beta_d] = 0$ a priori. Though in general one might prefer to let β have some nontrivial mean (as Lindley and Smith [1972] do in the context of exchangeability of effects across groups) this assumption simplifies the resulting estimators, theory, and notation. When β is permitted to have a non-zero mean, conjugacy maintains and the methodology presented in Section 2.3 may be updated to accommodate the change. While we omit a full explanation of the tedious details of this variation, we include its implementation in our code and the performance of the resulting empirical Bayesian estimators in Figures A-1 to A-3. From these empirical results we see that removing this restriction has little impact on the performance of the resulting estimators. Notably, our results in these figures reveal that the same is true for choosing to include or exclude a prior mean for the exchangeability of effects across groups prior.

A.5.4 Additional details on datasets

In each of the two regression applications, for each component dataset, we mean centered and variance-normalized the responses. Additionally, we Winsorized the responses by group; in particular, we clipped values more than 2 standard deviations from the mean.

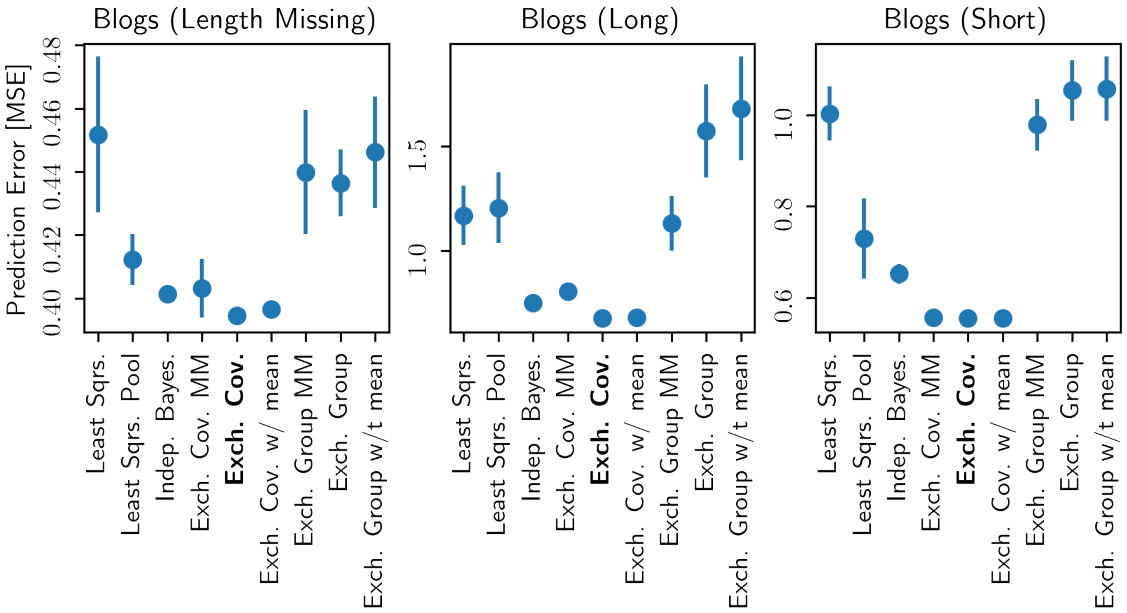


Figure A-2: Performances of methods on the blog dataset, segmented by post type. Uncertainty intervals are ±1SEM.

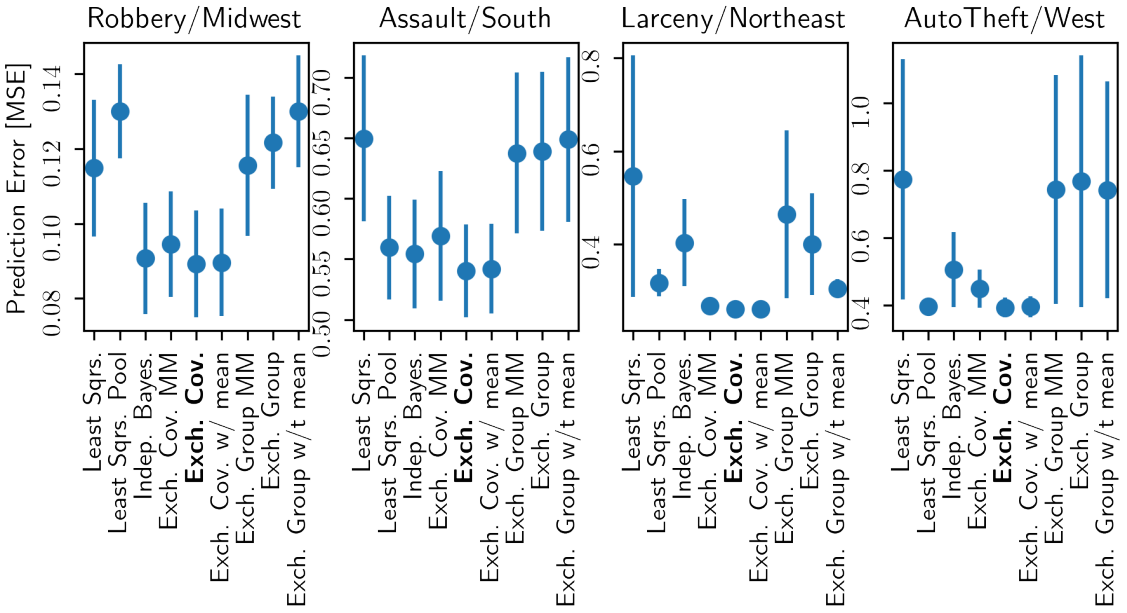


Figure A-3: Performances of methods on the law enforcement dataset, segmented by region and recorded offense categorization. Uncertainty intervals are ±1SEM.

BlogFeedback Data Set details Given the nature of the features included in the blog dataset used in the main text (which are summarizing characteristics rather than readable text), we believe it may be possible to find the blog post that corresponds to a particular data point. But we believe it is unlikely that the dataset directly contains any personally identifiable information. The blog information was obtained by web-crawling on publicly posted pages, so it is unlikely that consent for inclusion of the content into this dataset was obtained.

Communities and Crime Dataset details All data in this dataset was obtained through official channels. This dataset is composed of statistics aggregated at the community level, so it is less likely (though not impossible) to contain personally identifiable information. Since it contains demographic, census, and crime data, it is unlikely to contain offensive content.

CIFAR10 details. For the tasks `car vs. cat`, `car vs. dog`, `truck vs. cat`, and `truck vs. dog` we used $N^q = 100$ data points. For the tasks `car vs. deer`, `car vs. horse`, `truck vs. deer`, and `truck vs. horse` we used $N^q = 1000$ data points.

We generated the pre-trained neural network embeddings using a variational auto-encoder (VAE) [Kingma and Welling, 2013]. We adapted our VAE implementation from `ALIBI DETECT` [Van Looveren et al., 2019], here. See also `notebooks/2021_05_12_CIFAR10_VAE_embeddings.ipynb` for details.

CIFAR10 is composed from a subset of the 80 million tiny images dataset. As is currently acknowledged on the 80 million tiny images website, this larger dataset is known to contain offensive images and images obtained without consent (`https://groups.csail.mit.edu/vision/TinyImages/`). However, given the benign nature of the 10 image classes in CIFAR10, we expect it does not contain offensive or personally identifiable content. These data were also obtained by web-crawling, so it is unlikely that consent for inclusion of the content into this dataset was obtained.

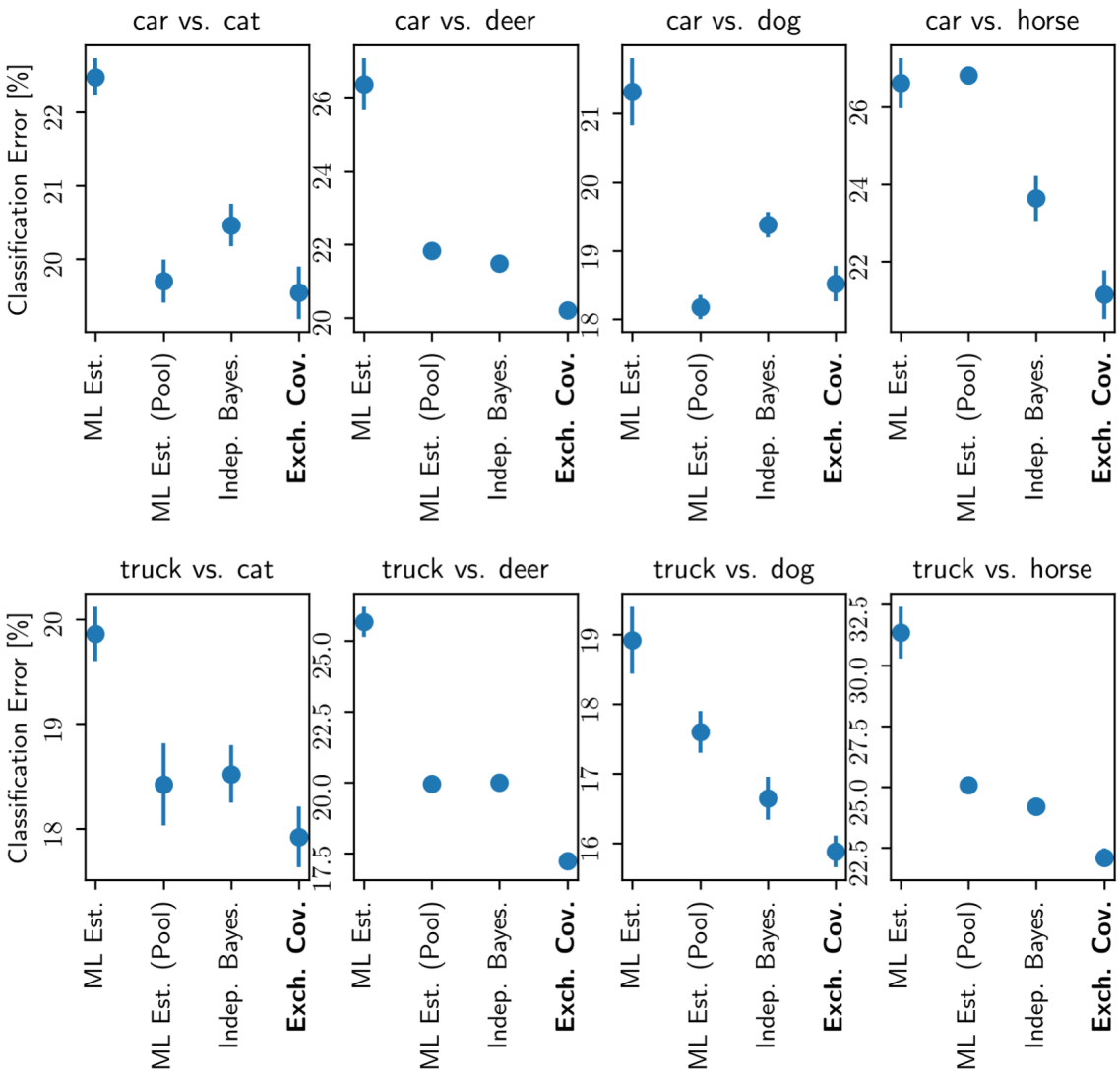


Figure A-4: Performances of methods on CIFAR10 segmented by binary classification task. Uncertainty intervals are ±1SEM.

A.5.5 Software Licenses

We here report the software used to generate our results and their associated licenses.

All of our experiments were implemented in `python`, which is licensed under the PSF license. For ease of reproducibility, ran our experiments and generated our plots IPython in Jupyter notebooks; this software is covered by a modified BSD license.

For our application to transfer learning using CIFAR10, we used a variational auto-encoder implementation adapted from `ALIBI DETECT` [Van Looveren et al., 2019], which uses the Apache licence. Our implementation of our EM algorithm uses `TensorFlow` [Abadi et al., 2016], which is licensed under the MIT license.

We made frequent use of python packages `numpy` and `scipy` and `matplotlib`. These are large libraries with components covered different licenses. See github.com/scipy/scipy/blob/master/LICENSES_bundled.txt for `scipy`, github.com/numpy/numpy/blo for `numpy`, and github.com/matplotlib/matplotlib/tree/master/LICENSE for `matplotlib`.

Appendix B

LR-GLM Supplementary Material

B.1 Additional Experimental Details and Empirical Results

B.1.1 Experimental Details

For all experiments we sampled β from an isotropic Gaussian prior with unit variance. For all synthetic data results we first generated a design matrix by sampling from a zero-mean Gaussian with diagonal covariance Σ with each $\Sigma_{i,i} = 5 * 1.05^{-i}$. We then used a scikit-learn [Pedregosa et al., 2011] implementation of a randomized SVD algorithm due to Halko et al. [2011], computed from two iterations (i.e., passes through X).

To assess robustness, in all experiments we used three or more replicate experiments, defined by independently generated synthetic datasets or train/test splits as well as re-rerunning the randomized truncated SVD.

The performance of the Diagonal Laplace approximation is dependent upon the shape the exact posterior at β^{MAP}. In particular, using a dataset with axis aligned covariance structure gives Diagonal Laplace an unrealistic advantage given that in most real applications we do not believe that low-rank structure will be axis aligned. As such, for all synthetic data experiments presented, we randomly generated a basis

of orthonormal vectors and used this basis to rotate our the design matrix. This rotation preserves the spectral decay of the data but eliminates the axis alignment of the synthetic data.

In all experiments we consider $N = 2{,}500$ training examples. We obtained results on "Out of Sample Data" (in Figures B.1.1 and B.1.5) by sampling X from an alternative distribution over covariates. Specifically, we generated these out-of-sample covariates in the manner described above, but with a different random rotation matrix.

We found MAP estimation using L-BFBS-B to be the most efficient of several available options in the scipy optimize library, and used this method in all MAP estimation and Laplace approximation experiments.

For all Bayesian predictions, we use the probit approximation to the logistic function to enable fast approximation [Bishop, 2006, Chap. 4.5].

B.1.2 Additional Figures

In Figure B.1.1 we present results on prediction performance, in term of classification error, as well as negative log likelihood, reported for "Training", "Test", and "Out of Sample Data". In Figure B.1.2 we report the error of LR-Laplace and Random-Laplace relative to NUTS for estimation of posterior means and variances. We see here that the estimates exhibit behavior increasingly similar to that of the prior as the rank of the approximation, M, decreases. Next, Figure B.1.3 depicts the same error trends for LR-MCMC using NUTS in `Stan`. We report calibration performance of the approximations of interest for credible sets of parameters (Figure B.1.4) as well as for prediction (Figure B.1.5).

We additionally include results analogous to those in the main text for Laplace approximations using low-rank data approximations to perform faster MCMC using NUTS with `Stan` [Carpenter et al., 2017], in Figure B.1.6. Finally, we also here provide the relative error of posterior mean and standard deviation estimation for logistic regression with a regularized horseshoe prior using the LR-MCMC approximation in Figure B.1.7. This experiment uses `Stan` for inference as well.

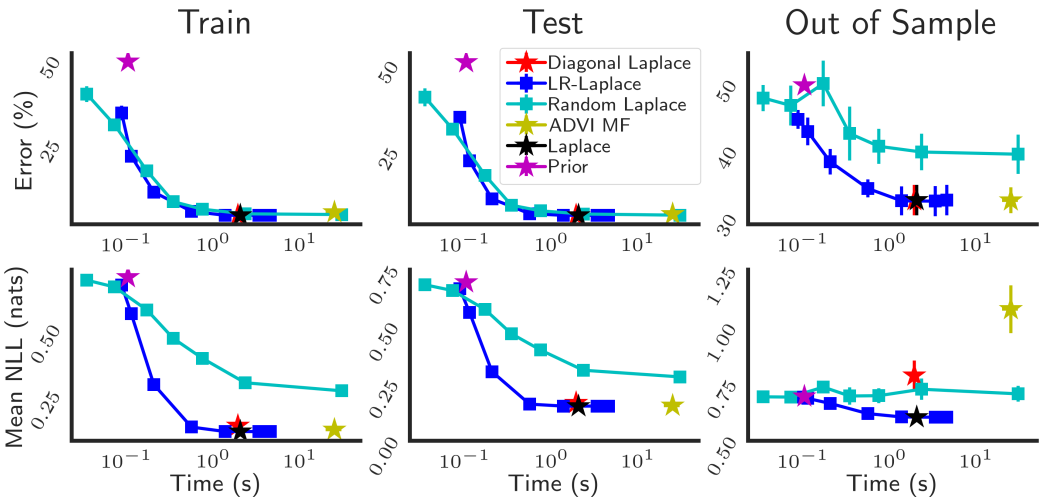


Figure B.1.1: Predictive performance of posterior approximations in Bayesian logistic regression in terms of (Top) classification error and (Bottom) average negative log likelihood (NLL) of responses under approximate posterior predictive distributions on (Left) *train*, (Center) *test* and (Right) *out of sample* datasets. Lower is better.

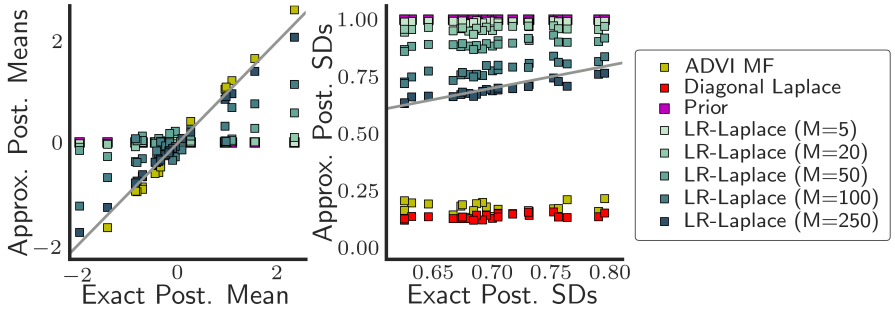


Figure B.1.2: Approximate posterior mean and standard deviation across a parameter subset as M varies. Horizontal axis represents ground truth from running NUTS using `Stan` without the LR-GLM approximation. $D = 250$.

Horseshoe logistic regression experiment

For the logistic regression experiment using a regularized horseshoe prior we used $N = 1{,}000$ data points of dimension $D = 200$. We used ten non-zero effects, each of size 10. Our implementation of the regularized horseshoe and inference in `Stan` closely followed M. Betancourt's "Bayes Sparse Regression" case study.[1] We generated covariates as described in the previous section.

[1] `https://betanalpha.github.io/assets/case_studies/bayes_sparse_regression.html`

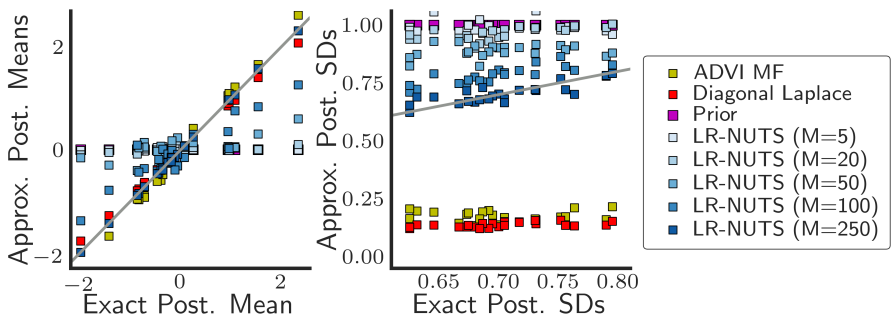


Figure B.1.3: This figure is analogous to Figure B.1.2 but examines the trade-off between computation and accuracy of LR-MCMC using NUTS in `Stan`. $D = 250$.

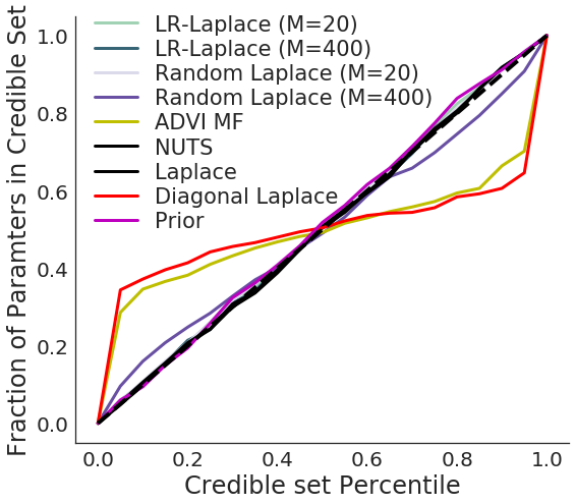


Figure B.1.4: Credible set calibration. The fraction of parameters in the credible sets defined by different lower tail intervals as a function of the approximate posterior probability of parameters taking values in that interval. The black dotted line (on the diagonal) reflects perfect calibration.

B.1.3 Stan Model Code

First we show `Stan` code for Bayesian logistic regression.

```
data {
 int<lower=1> N; // # of data
 int<lower=1> D; // # of covariates
 matrix[N, D] X; // Design matrix
 int<lower=0> y[N]; // labels
 real<lower=0> sigma;
```

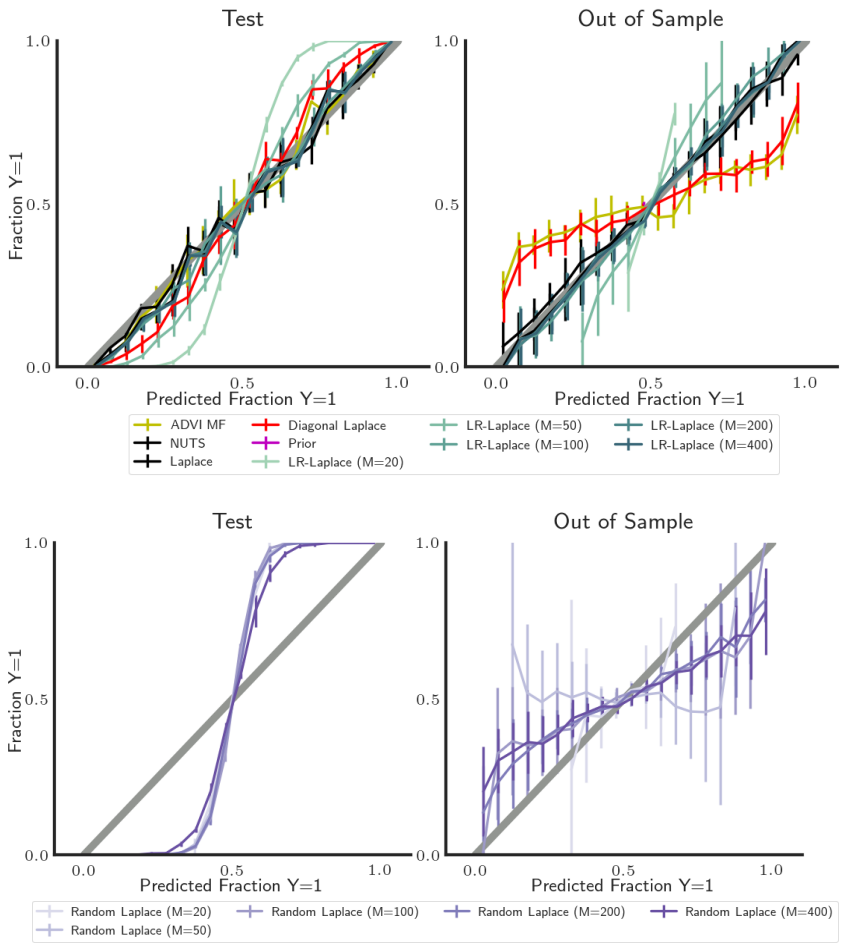


Figure B.1.5: Prediction calibration.

```
}
parameters {
 vector[D] beta;
}
model {
 beta ~ normal(0, sigma);
 y ~ bernoulli_logit(X * beta);
}
```

Second, we show `Stan` code for logistic regression with our low-rank approximation.

```
data {
 int<lower=1> N; // # of data
```

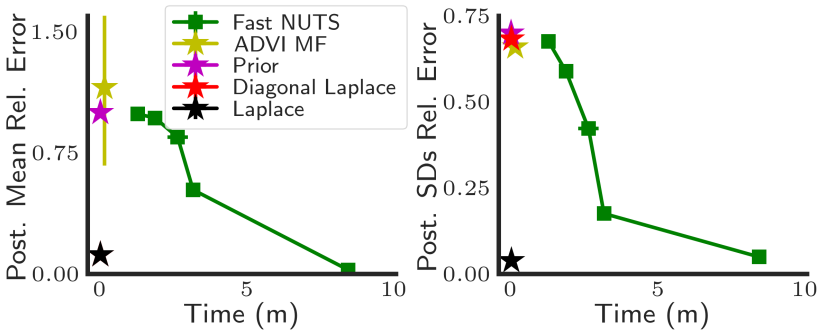


Figure B.1.6: This figure is analogous to Figure 3-2A but assesses LR-MCMC using NUTS in `Stan` rather than LR-Laplace. $D = 250$.

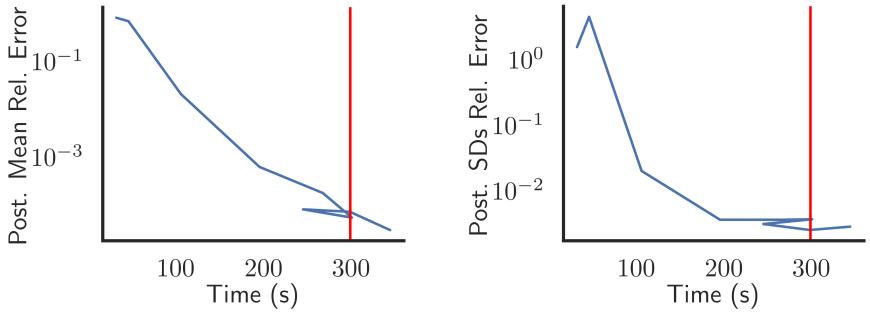


Figure B.1.7: Bayesian logistic regression with a regularized Horseshoe prior using NUTS in `Stan`. The red vertical line indicates the runtime of inference with `Stan` using the exact likelihood.

```
 int<lower=1> D; // # of covariates
 int<lower=1> M; // Projected dimension
 matrix[D, M] U; // Projection matrix
 matrix[N, M] barX; // Projected design matrix
 int<lower=0> y[N]; // labels
 real<lower=0> sigma;
}
parameters {
 vector[D] beta;
}
```

```
transformed parameters {
 vector[M] bar_beta = U' * beta;
}
model {
 beta ~ normal(0, sigma);
 y ~ bernoulli_logit(barX * bar_beta);
}
```

B.2 Related Work on Scalable Bayesian Inference

Developing scalable approximate Bayesian inference for models with many parameters (large D) and many data points (large N) has been active area of research for decades, and researchers have developed a large variety of methods applicable to GLMs. Historically, Markov chain Monte Carlo (MCMC) methods based on the Metropolis-Hastings algorithm Metropolis et al. [1953], Hastings [1970] have been dominant. However MCMC is computationally expensive on large-scale problems in which both D and N are very large. In particular, each likelihood evaluation requires $O(DN)$ time, due to the matrix vector product $X\beta$. Further, estimating posterior covariances uniformly well requires $O(\log D)$ samples [Cai et al., 2010]. Therefore, the total cost of collecting those samples is $O(ND \log D)$ time in the case of perfect, independent Monte Carlo samples. In practice, though, mixing times may also have unfavorable scaling with dimensionality and sample size; these issues can lead to even worse scaling in N and D. Several lines of research have explored the use of subsampling methods to reduce the dependence on N. But these methods either lose the asymptotic guarantees of exact MCMC or fail to provide faster inference in practice due to poor mixing behavior [Bardenet et al., 2017].

Other work has pursued deterministic approximations to the Bayesian posterior. Some of the most widely used of these approximations include (1) the Laplace approximation, which is a Gaussian approximation of the posterior defined locally at the posterior mode, (2) extensions of the Laplace approximation such as the integrated

nested Laplace approximation (INLA) [Rue et al., 2009], and (3) variational Bayes; see, e.g., [Bishop, 2006, Chap. 10] and [Blei et al., 2017]. However, these approaches also scale poorly with dimension in general. The Laplace approximation requires computing and inverting the Hessian of the log posterior which demand $O(ND^2)$ and $O(D^3)$ time respectively, in order to compute approximate posterior means and variances. In the $N \ll D$ setting, this cost can be reduced to $O(N^2D)$ time (Appendix B.3). However, in large-N settings of interest, the $O(N^2D)$ cost can be prohibitive as well. The cost of inference is further compounded when we give a fully Bayesian treatment to model hyperparameters as well as parameters; e.g., INLA requires this heavy computation for each nested approximation. In the face of difficulties posed by high dimensionality, practitioners frequently turn to factorized (or "mean-field") approximations. In the case of VB, the mean-field approach can yield biased approximations that underestimate uncertainty MacKay [2003], Turner and Sahani [2011]. Likewise, factorized Laplace approximations, which approximate the Hessian with only its diagonal elements, similarly underestimate uncertainty (Appendix B.6.8).

Some more recent work has approached scalable approximate inference in generalized linear models with theoretical guarantees on quality in the large-N regime by using likelihood approximations that are cheap to evaluate Huggins et al. [2017], Campbell and Broderick [2019, 2018], Huggins et al. [2016]. But these methods fail to scale well to the large-D case.

More closely related to the present work, Geppert et al. [2017] and Lee and Oh [2013] focus on conjugate Bayesian regression, respectively using random projections and principle component analysis to define low-rank descriptions of the design. Lee and Oh [2013] restrict their consideration to the exactly low-rank case and primarily discuss the asymptotic consistency of the resulting posterior mean without discussing computational considerations. Spantini et al. [2015] use conjugate Bayesian regression as stepping-off point to derive a point estimator for Bayesian inverse problems. Guhaniyogi and Dunson [2015] use random projections for Bayesian GLMs but focus on predictive performance rather than parameter estimation. Outside the Bayesian context, Zhang et al. [2014], Wang et al. [2017], and many others have analyzed ran-

dom projections for regression and classification using, for example, an M-estimation framework.

B.3 Fast matrix inversions in the $N \ll D$ setting

In this section we focus on Gaussian conjugate linear regression with $N \ll D$. In this case, we can detail formulas for more efficient computation of the posterior mean and covariance. We start from the standard expressions for the posterior mean μ_N and covariance Σ_N when the prior is mean zero with covariance Σ_β; see Section 3.3 and Section 3.4.1 for further notation and setup of the model. These expressions are:

$$\Sigma_N^{-1} = \Sigma_\beta^{-1} + \tau X^\top X \tag{B.3.1}$$

$$\mu_N = \tau \Sigma_N X^\top Y. \tag{B.3.2}$$

Using these formulas naively in the $D \gg N$ setting is computationally expensive due to the $O(D^3)$ time cost of matrix inversion and $O(D^2)$ storage cost.

Using the Woodbury matrix identity, $(A^{-1}+UCV)^{-1} = A - AU(C^{-1}+VAU)^{-1}VA$, allows us to write $\Sigma_N = (\Sigma_\beta^{-1} + X^\top(\tau I_N)X)^{-1}$ as

$$\Sigma_N = \Sigma_\beta - \Sigma_\beta X^\top(\tau^{-1} I_N + X\Sigma_\beta X^\top)^{-1} X\Sigma_\beta. \tag{B.3.3}$$

Computing Σ_N via Eq. (B.3.3) requires only $O(DN^2)$ cost for the matrix multiplications and an $O(N^3)$ cost for the matrix inversion. The posterior mean μ_N may then be computed in $O(ND)$ time by multiplying through by $X^\top Y$. These time costs can be significant reductions over the naive $O(D^3)$ cost when $N \ll D$.

Fast inversions for the Laplace approximation to the GLM posterior

We here show that the same approach described above may be used for the Laplace approximation in the context of Bayesian GLMs. We say that we have a GLM

likelihood if we can write

$$p(Y \mid \beta, X) = \sum_{n=1}^{N} \phi(y_n, x_n^\top \beta)$$

for some *mapping function* $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. The Bayesian posterior then becomes

$$\log p(\beta \mid X, Y) = \log p(\beta) + \sum_{n=1}^{N} \phi(y_n, x_n^\top \beta) + Z, \tag{B.3.4}$$

where Z is a typically-intractable log normalizing constant.

Due to the analytic intractability of posterior inference in many common GLMs, approximations are necessary; the Laplace approximation is a particularly widely used approximation and takes the form

$$\bar{p}(\beta) = \mathcal{N}(\beta \mid \bar{\mu}, \bar{\Sigma}), \tag{B.3.5}$$

where $\bar{\mu} := \arg\max_\beta \log p(\beta \mid X, Y)$ and $\bar{\Sigma} := \left(-\nabla^2_\beta \log p(\beta | X, Y)|_{\beta=\bar{\mu}}\right)^{-1}$. However, as in the conjugate case, computing this matrix inverse naively can be expensive in the high-dimensional setting, and we are motivated to consider more computationally efficient routes to evaluate it. In settings when $N \ll D$ and when we have a Gaussian prior $p(\beta) = \mathcal{N}(\beta \mid \mu_\beta, \Sigma_\beta)$, we may take an approach similar to our approach in the conjugate case. We first note

$$\nabla^2_\beta \log p(\beta \mid X, Y)|_{\beta=\bar{\mu}} = -\Sigma_\beta^{-1} + X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))X, \tag{B.3.6}$$

where $\vec{\phi}''(Y, A)$ is a vector in $\mathbb{R}^N$ defined such that for any n in $1, 2, \ldots, N$, $\vec{\phi}''(Y, A)_n := \frac{d^2}{da^2}\phi(y_i, a)|_{a=A_n}$. Applying the same trick to this expression as before, we obtain

$$\bar{\Sigma}_N = \left(-\nabla^2_\beta \log p(\beta \mid X, Y)|_{\beta=\bar{\mu}}\right)^{-1} \tag{B.3.7}$$

$$= \Sigma_\beta - \Sigma_\beta X^\top \left(\mathrm{diag}[-\vec{\phi}''(Y, X\bar{\mu})]^{-1} + X\Sigma_\beta X^\top\right)^{-1} X\Sigma_\beta, \tag{B.3.8}$$

which again can yield computational gains.

It is worth noting however that this route is more computationally efficient only when the prior covariance matrix is structured in some way that allows for fast matrix-vector and matrix-matrix multiplications. This will be the case, for example, if Σ_β is diagonal, block-diagonal, banded diagonal, or diagonal plus a low-rank matrix.

B.4 Conjugate Gaussian regression with exactly low rank design

B.4.1 Derivation of Eq. (3.2)

Here we consider the setting of conjugate Bayesian linear regression, with X exactly low rank and $\Sigma_\beta = \sigma_\beta^2 I_D$, as detailed in Section 3.4.1. We now derive the expressions (Eq. (3.2)) for the mean and covariance of the Gaussian posterior for β in this case. We suppose $X = V\text{diag}(\lambda)U^\top$ for U, V matrices of orthonormal rows and λ a vector. The preceding equation for X will capture low rank structure when $U \in \mathbb{R}^{D\times M}$ for some M with $M \ll \min(D, N)$.

For the covariance, we start from Eq. (B.3.1). Then we can rewrite Σ_N as follows.

$$
\begin{aligned}
\Sigma_N &= \left(\sigma_\beta^{-2} I_D + \tau X^\top X\right)^{-1} \\
&= \left(\sigma_\beta^{-2} I_D + \tau U\text{diag}(\lambda)V^\top V\text{diag}(\lambda)U^\top\right)^{-1} \\
&= \left(\sigma_\beta^{-2} I_D + U\text{diag}(\tau\lambda \odot \lambda)U^\top\right)^{-1} \\
&\quad \text{where } \odot \text{ denotes component-wise multiplication across a vector} \\
&= \sigma_\beta^2 I - \sigma_\beta^2 U(\text{diag}(\tau\lambda \odot \lambda)^{-1} + \sigma_\beta^2 I_M)^{-1} U^\top \sigma_\beta^2 \\
&\quad \text{by the Woodbury matrix identity and } U^\top U = I_M \\
&= \sigma_\beta^2 I - \sigma_\beta^2 U\text{diag}\left\{\left(\frac{1}{\tau\lambda \odot \lambda} + \sigma_\beta^2 1_M\right)^{-1} \sigma_\beta^2\right\} U^\top \\
&\quad \text{where division within 'diag' is component-wise and } 1_M \text{ is the } M-\text{vector of ones} \\
&= \sigma_\beta^2 \left(I_D - U\text{diag}\left\{\frac{\tau\lambda \odot \lambda}{\sigma_\beta^{-2} 1_M + \tau\lambda \odot \lambda}\right\} U^\top\right).
\end{aligned}
$$

Starting from Eq. (B.3.2), we can rewrite the posterior mean as follows.

$$\begin{aligned}
\mu_N &= \tau\Sigma_N X^\top Y \\
&= \tau\sigma_\beta^2 \left(I_D - U\mathrm{diag}\left\{ \frac{\tau\lambda \odot \lambda}{\sigma_\beta^{-2} 1_M + \tau\lambda \odot \lambda} \right\} U^\top \right) U\mathrm{diag}(\lambda) V^\top Y \\
&\quad \text{from the derivation above and substituting for } X \\
&= \tau\sigma_\beta^2 \left(U - U\mathrm{diag}\left\{ \frac{\tau\lambda \odot \lambda}{\sigma_\beta^{-2} 1_M + \tau\lambda \odot \lambda} \right\} \right) \mathrm{diag}(\lambda) V^\top Y \\
&\quad \text{since } U^\top U = I_M \\
&= \tau\sigma_\beta^2 U \left(I_M - \mathrm{diag}\left\{ \frac{\tau\lambda \odot \lambda}{\sigma_\beta^{-2} 1_M + \tau\lambda \odot \lambda} \right\} \right) \mathrm{diag}(\lambda) V^\top Y \\
&= U\mathrm{diag}\left\{ \frac{\tau\lambda}{\sigma_\beta^{-2} 1_M + \tau\lambda \odot \lambda} \right\} V^\top Y.
\end{aligned}$$

B.5 Proofs and further results for conjugate Bayesian linear regression with low-rank data approximations

B.5.1 Proof of Theorem 3.4.1

Recall that for conjugate Gaussian Bayesian linear regression, the exact posterior is $p(\beta \mid X, Y) = \mathcal{N}(\beta \mid \mu_N, \Sigma_N)$, where μ_N and Σ_N are given in Eqs. (B.3.1) and (B.3.2).

Using an orthonormal projection U yields a Gaussian approximate posterior $\tilde{p}(\beta \mid X, Y) = \mathcal{N}(\beta \mid \tilde{\mu}_N, \tilde{\Sigma}_N)$. Recall from Section 3.3 that we obtain this approximate posterior by replacing X with $XUU^\top$. Thus, we can find $\tilde{\mu}_N$ and $\tilde{\Sigma}_N$ by consulting Eqs. (B.3.1) and (B.3.2):

$$\tilde{\Sigma}_N^{-1} = \Sigma_\beta^{-1} + \tau UU^\top X^\top XUU^\top \tag{B.5.1}$$

$$\tilde{\mu}_N = \tilde{\tau}\Sigma_N UU^\top X^\top Y. \tag{B.5.2}$$

Upper bound on the posterior mean approximation error

We will obtain our upper bound on the error of the approximate posterior mean relative to the exact posterior mean by upper bounding the norm of the difference between the gradient of the log posterior with respect to β at the approximate posterior mean, $\tilde{\mu}_N$, and the exact posterior mean, μ_N. Together with the strong convexity of the negative log posterior, this bound will allow us to arrive at the desired upper bound on $\|\mu_N - \tilde{\mu}_N\|_2$.

First, we bound the norm of the gradient difference. To that end, the gradients of the exact log likelihood and the approximate log likelihood are given by

$$\begin{aligned}\nabla_\beta \log p(Y \mid X, \beta) &= \nabla_\beta \left[-\frac{\tau}{2}(X\beta - Y)^\top (X\beta - Y) \right] \\ &= -\tau(X^\top X\beta - X^\top Y)\end{aligned}$$

and

$$\begin{aligned}\nabla_\beta \log \tilde{p}(Y \mid X, \beta) &= \nabla_\beta \left[-\frac{\tau}{2}(XUU^\top \beta - Y)^\top (XUU^\top \beta - Y) \right] \\ &= -\tau(UU^\top X^\top XUU^\top \beta - UU^\top X^\top Y).\end{aligned}$$

We can thus upper bound the norm of the difference between the two log posteriors as follows.

$$\begin{aligned}&\left\|\nabla_\beta \log \tilde{p}(\beta \mid X, Y) - \nabla_\beta \log p(\beta \mid X, Y)\right\|_2 \\ &= \left\|\nabla_\beta \log \tilde{p}(Y \mid X, \beta) - \nabla_\beta \log p(Y \mid X, \beta)\right\|_2 \\ &\quad \text{since the prior is the same in both the exact and approximate model} \\ &\quad \text{and since the normalizing constant has no } \beta \text{ dependence} \\ &= \left\| -\tau\left(UU^\top X^\top XUU^\top \beta - UU^\top X^\top Y\right) + \tau\left(X^\top X\beta - X^\top Y\right) \right\|_2 \\ &= \tau \left\| \left(X^\top X - UU^\top X^\top XUU^\top\right)\beta + UU^\top X^\top Y - X^\top Y \right\|_2 \\ &= \tau \left\| \bar{U}\bar{U}^\top X^\top X\bar{U}\bar{U}^\top \beta - \bar{U}\bar{U}^\top X^\top Y \right\|_2\end{aligned}$$

where $\bar{U}$ (above) as well as $\bar{\lambda}$ and $\bar{V}$ (below) are defined in Section 3.3

$$= \tau \left\| \bar{U} \mathrm{diag}(\bar{\lambda} \odot \bar{\lambda}) \bar{U}^\top \beta - \bar{U} \mathrm{diag}(\bar{\lambda}) \bar{V}^\top Y \right\|_2$$

$$\leq \tau \left(\left\| \bar{U} \mathrm{diag}(\bar{\lambda} \odot \bar{\lambda}) \bar{U}^\top \beta \|_2 + \| \bar{U} \mathrm{diag}(\bar{\lambda}) \bar{V}^\top Y \right\|_2 \right)$$

by the triangle inequality

$$= \tau \left(\left\| \mathrm{diag}(\bar{\lambda} \odot \bar{\lambda}) \bar{U}^\top \beta \|_2 + \| \mathrm{diag}(\bar{\lambda}) \bar{V}^\top Y \right\|_2 \right)$$

since $\|v\|_2^2 = v^\top v$ for a vector v and $U^\top U = I_M$

$$\leq \tau \left(\left\| \mathrm{diag}(\bar{\lambda} \odot \bar{\lambda}) \right\|_{\mathrm{op}} \left\| \bar{U}^\top \beta \|_2 + \| \mathrm{diag}(\bar{\lambda}) \|_{\mathrm{op}} \| \bar{V}^\top Y \right\|_2 \right)$$

by definition of the operator norm in this space

$$= \tau \left(\bar{\lambda}_1^2 \| \bar{U}^\top \beta \|_2 + \bar{\lambda}_1 \| \bar{V}^\top Y \|_2 \right) \tag{B.5.3}$$

Second, we need a result that will let us use the strong convexity of the negative log posterior. We prove the following result in Appendix B.5.2.

Lemma B.5.1. *Let f, g be twice differentiable functions mapping $\mathbb{R}^D \to \mathbb{R}$ and attaining minima at $\beta_f = \arg\min_\beta f(\beta)$ and $\beta_g = \arg\min_\beta g(\beta)$, respectively. Additionally, assume that f is α–strongly convex for some $\alpha > 0$ on the set $\{t\beta_f + (1-t)\beta_g | t \in [0,1]\}$ and that $\|\nabla_\beta f(\beta_g) - \nabla_\beta g(\beta_g)\|_2 = \|\nabla_\beta f(\beta_g)\|_2 \leq c$. Then*

$$\|\beta_f - \beta_g\|_2 \leq \frac{c}{\alpha}. \tag{B.5.4}$$

To use the preceding result, we need a lower bound on the strong convexity constant of the negative log posterior; we now calculate such a bound. We have that μ_N and $\tilde{\mu}_N$ are the maximum a posteriori values of β under $p(\beta|X, Y, \alpha)$ and $\tilde{p}(\beta|X, Y, \alpha)$, respectively; equivalently they minimize the respective negative log of these distributions. For a matrix A, let $\lambda_{\min}(A)$ denote its minimum eigenvalue. The Hessian of the negative log posterior with respect to β is precisely $\Sigma_\beta^{-1} + \tau X^\top X$

everywhere. So the negative log posterior is α–strongly convex, where

$$\alpha = \lambda_{\min}(\Sigma_\beta^{-1} + \tau X^\top X) \geq \lambda_{\min}(\Sigma_\beta^{-1}) + \tau\lambda_{\min}(X^\top X) = \|\Sigma_\beta\|_2^{-1} + \tau\bar{\lambda}_{D-M}^2. \quad \text{(B.5.5)}$$

In the first part of the final equality above, we use that the spectral norm of a matrix inverse is equal to the reciprocal of the minimum eigenvalue of the matrix.

Now we have an upper bound on the norm of the difference in gradients of the negative log posteriors (the same as for the log posteriors, in Eq. (B.5.3)) and a lower bound on the strong convexity constant from Eq. (B.5.5). So we can apply these together with Lemma B.5.1 to find

$$\begin{aligned}
\|\mu_N - \tilde{\mu}_N\|_2 &\leq \frac{\tau\big(\bar{\lambda}_1^2\|\bar{U}^\top\tilde{\mu}_N\|_2 + \bar{\lambda}_1\|\bar{V}^\top Y\|_2\big)}{\alpha} \\
&\quad \text{by Lemma B.5.1 taking } \log p(\beta|X,Y) \text{ and } \log\tilde{p}(\beta|X,Y) \\
&\quad \text{as } f \text{ and } g \text{ respectively, with } c \text{ given by Eq. (B.5.3)} \\
&\leq \frac{\tau\big(\bar{\lambda}_1^2\|\bar{U}^\top\tilde{\mu}_N\|_2 + \bar{\lambda}_1\|\bar{V}^\top Y\|_2\big)}{\|\Sigma_\beta\|_2^{-1} + \tau\bar{\lambda}_{D-M}^2} \\
&\quad \text{by Eq. (B.5.5)} \\
&= \frac{\bar{\lambda}_1\big(\bar{\lambda}_1\|\bar{U}^\top\tilde{\mu}_N\|_2 + \|\bar{V}^\top Y\|_2\big)}{\|\tau\Sigma_\beta\|_2^{-1} + \bar{\lambda}_{D-M}^2}.
\end{aligned}$$

Notably, in the common special case that Σ_β is diagonal, as we saw in Section 3.4.1, $\tilde{\mu}_N$ will be in the span of U, and we will have that $\|\bar{U}^\top\tilde{\mu}_N\|_2 = 0$.

Error in Posterior Precision

The error in the precision matrices for the approximate and exact posteriors in linear regression are particularly straightforward since they do not depend on the responses, Y. In particular, we have

$$\begin{aligned}
\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1} &= (\Sigma_\beta^{-1} + \tau X^\top X) - (\Sigma_\beta^{-1} + \tau UU^\top X^\top XUU^\top) && \text{(B.5.6)} \\
&= \tau X^\top X - \tau UU^\top X^\top XUU^\top && \text{(B.5.7)} \\
&= \tau\bar{U}\bar{U}^\top X^\top X\bar{U}\bar{U}^\top && \text{(B.5.8)} \\
&= \tau\bar{U}\mathrm{diag}(\bar{\lambda}\odot\bar{\lambda})\bar{U}^\top. && \text{(B.5.9)}
\end{aligned}$$

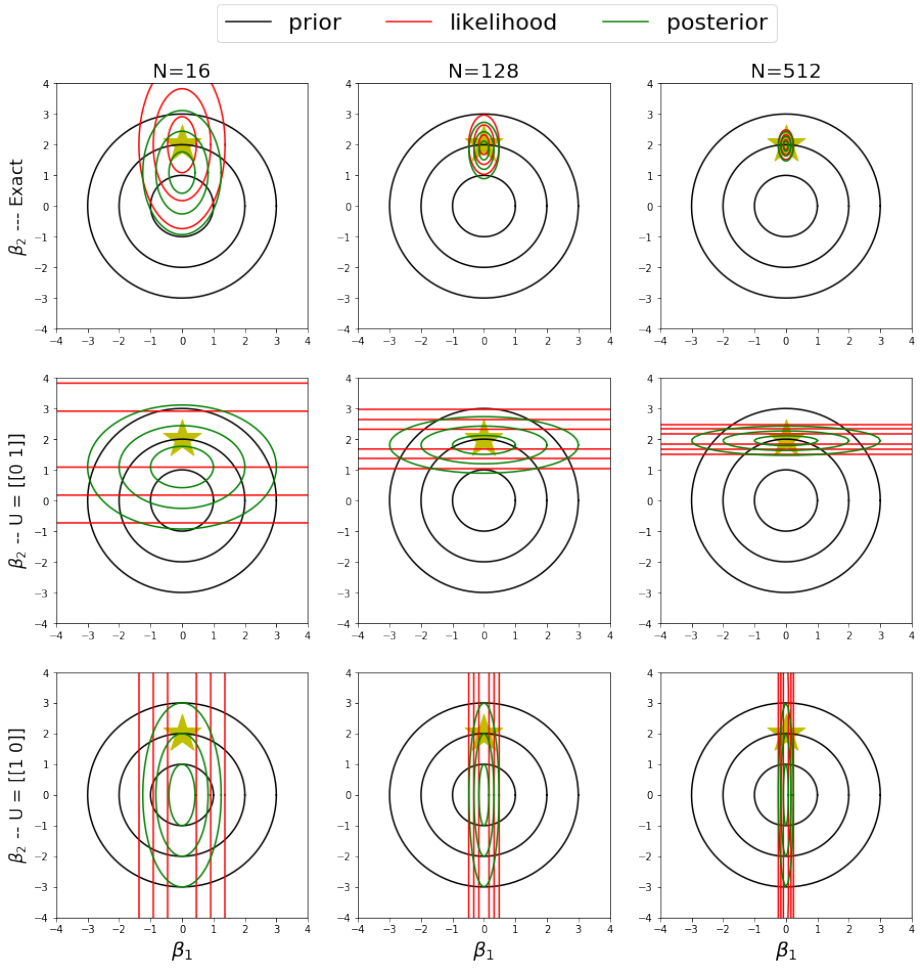


Figure B.5.1: Example of posterior approximations with different projections (characterized by U) for increasing sample sizes. Each plot shows the contours of three densities: the prior, likelihood, and posterior (or approximations thereof). The top row shows the exact posterior. The middle row shows the approximations found by using the best rank-1 approximation to X. The bottom row shows the approximations found using the orthogonal rank-1 approximation. The star is at the parameter value used to generate simulated data for these plots.

Thus, since it is equal to the maximum eigenvalue, the spectral norm of the error in the precisions is precisely $\|\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1}\|_2 = \tau\bar{\lambda}_1^2$.

B.5.2 Proof of Lemma B.5.1

By the fundamental theorem of calculus, we may write

$$\nabla_\beta f(\beta) = \nabla_\beta f(\beta_g) + \int_{t=0}^{1} (\beta - \beta_g)^\top \nabla_\beta^2 f(t\beta + (1-t)\beta_g) dt.$$

Considering the norm of $\nabla_\beta f(\beta)$ and applying the triangle inequality provides that for any β in $\{t\beta_f + (1-t)\beta_g \mid t \in [0,1]\}$,

$$\|\nabla_\beta f(\beta)\|_2 \geq \left\| \int_{t=0}^{1} (\beta - \beta_g)^\top \nabla_\beta^2 f(t\beta + (1-t)\beta_g) dt \right\|_2 - \|\nabla_\beta f(\beta_g)\|_2 \tag{B.5.10}$$

$$\geq \|\beta - \beta_g\|_2 \left\| \int_{t=0}^{1} \nabla_\beta^2 f(t\beta + (1-t)\beta_g) dt \right\|_2 - \|\nabla_\beta f(\beta_g)\|_2 \tag{B.5.11}$$

$$\geq \|\beta - \beta_g\|_2 \alpha - \|\nabla_\beta f(\beta_g)\|_2. \tag{B.5.12}$$

We consider this bound at β_f. Recall we assume that $\|\nabla_\beta f(\beta_g)\|_2 \leq c$. And $\|\nabla_\beta f(\beta_f)\|_2 = 0$ since f is twice differentiable. Therefore, we have that $0 \geq \|\beta_f - \beta_g\|_2 \alpha - c$, and the result follows.

B.5.3 Proof of Corollary 3.4.2

Our approach is to show that

$$\tilde{\mu}_N \xrightarrow{p} \Sigma_\beta U_* (U_*^\top \Sigma_\beta U_*)^{-1} U_*^\top \beta. \tag{B.5.13}$$

We then appeal to the following result, which we prove in Appendix B.5.4:

Lemma B.5.2. *$\tilde{\mu} := \Sigma_\beta U (U^\top \Sigma_\beta U)^{-1} U^\top \beta$ is the vector of minimum Σ_β^{-1}-norm satisfying $U^\top \tilde{\mu} = U^\top \beta$.*

Finally, for any closed $S \subset \mathbb{R}^D$, $\tilde{\mu} = \arg\min_{v \in S} \|v\|_{\Sigma_\beta^{-1}} = \arg\max_{v \in S} -\frac{1}{2} v^\top \Sigma_\beta^{-1} v = \arg\max_{v \in S} \mathcal{N}(0, \Sigma_\beta)$. Therefore, the $\tilde{\mu}$ in Lemma B.5.2 is the maximum a priori vector satisfying the constraint in Lemma B.5.2.

We first turn to proving Eq. (B.5.13). Let $U_N \mathrm{diag}(\lambda^{(N)}) V_N^\top$ denote the M-truncated SVD of the design matrix consisting of N samples $X = (x_1, x_2, \ldots, x_N)$ where $x_i \overset{\text{i.i.d.}}{\sim} p_*$. When the low rank approximation is defined by this SVD, from Eq. (B.5.1) we have that $\tilde{\mu}_N = \tau \tilde{\Sigma}_N U_N U_N^\top X^\top Y$. Noting that $Y = X\beta + \frac{1}{\tau}\epsilon$ for some $\epsilon \in \mathbb{R}^N$ with

$\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, we may expand this out and write:

$$\begin{aligned}
\tilde{\mu}_N &= \tau(\Sigma_\beta^{-1} + U_N U_N^\top X^\top \tau X U_N U_N^\top)^{-1} U_N U_N^\top X^\top (X\beta + \frac{1}{\tau}\epsilon) \\
&= \tau \left\{ \Sigma_\beta^{-1} + U_N \left[\tau \text{diag}(\lambda^{(N)} \odot \lambda^{(N)}) \right] U_N^\top \right\}^{-1} U_N \text{diag}(\lambda^{(N)}) V_N^\top \\
&\quad \left[V_N \text{diag}(\lambda^{(N)}) U_N^\top \beta + \frac{1}{\tau}\epsilon \right] \qquad \text{(B.5.14)} \\
&= \left\{ \Sigma_\beta^{-1} + U_N \left[\tau \text{diag}(\lambda^{(N)} \odot \lambda^{(N)}) \right] U_N^\top \right\}^{-1} U_N \left[\tau \text{diag}(\lambda^{(N)} \odot \lambda^{(N)}) \right] \\
&\quad \left[U_N^\top \beta + \text{diag}(\lambda^{(N)})^{-1} V_N^\top \frac{1}{\tau}\epsilon \right] \qquad \text{(B.5.15)} \\
&= \Sigma_\beta U_N \left[U_N^\top \Sigma_\beta U_N + \tau^{-1} \text{diag}(\lambda^{(N)})^{-2} \right]^{-1} \left[U_N^\top \beta + \text{diag}(\lambda^{(N)})^{-1} V_N^\top \frac{1}{\tau}\epsilon \right] \\
&\xrightarrow{P} \Sigma_\beta U_* (U_*^\top \Sigma_\beta U_*)^{-1} U_*^\top \beta,
\end{aligned}$$

where in the fourth line we use the matrix identity, $(R^{-1} + W^\top Q W)^{-1} W^\top Q = R W^\top (W R W^\top + Q^{-1})^{-1}$ [Petersen and Pedersen, 2008]. Convergence in probability in the last line follows since $\text{diag}(\lambda^{(N)^{-2}}) \xrightarrow{P} 0$ [Vershynin, 2012] and $U_N \xrightarrow{P} U$.

B.5.4 Proof of Lemma B.5.2

We show that $\beta_* = \Sigma_\beta U (U^\top \Sigma_\beta U)^{-1} U^\top \beta$ is the vector of minimum norm satisfying the above constraints in the Hilbert space $\mathbb{R}^D$ with inner product $\langle v_1, v_2 \rangle = v_1^\top \Sigma_\beta^{-1} v_2$ for vectors $v_1, v_2 \in \mathbb{R}^D$.

Define β_* as

$$\beta_* = \underset{v \in \mathbb{R}^D}{\arg\min} \|v\|_{\Sigma_\beta^{-1}} \text{ subject to } U^\top v = U^\top \beta \qquad \text{(B.5.16)}$$

First note that the condition $U^\top \beta_* = U^\top \beta$ may be expressed as a set the M linear constraints

$$\langle \Sigma_\beta U[:,i], \beta_* \rangle = U[:,i]^\top \beta \qquad \text{(B.5.17)}$$

for $i = 1, 2, \ldots, M$. We thereby see that the constraint restricts β_* to the linear variety $\beta + \left[\left\{\Sigma_\beta U[:, i]\right\}_{i=1}^M\right]^\perp$, where $\left[A\right]$ denotes the subspace generated by the vectors of the set A and $\left[A\right]^\perp$ denotes the set of all vectors orthogonal to $\left[A\right]$ (i.e. the orthogonal complement of $\left[A\right]$). By the projection theorem [Luenberger, 1969], β_* is orthogonal to $\left[\left\{\Sigma_\beta U[:, i]\right\}_{i=1}^M\right]^\perp$, or $\beta_* \in \left[\left\{\Sigma_\beta U[:, i]\right\}_{i=1}^M\right]^{\perp\perp} = \left[\left\{\Sigma_\beta U[:, i]\right\}_{i=1}^M\right]$. We can therefore write β_* as a linear combination of the vectors $\left\{\Sigma_\beta U[:, i]\right\}_{i=1}^M$; that is, for some c in $\mathbb{R}^M$

$$\beta_* = \Sigma_\beta U c. \tag{B.5.18}$$

Our constraints in Eq. (B.5.17) then demand that $\langle \Sigma_\beta U[:, i], \Sigma_\beta U c\rangle = U[:, i]^\top \beta$ for each i, or equivalently that $U^\top \Sigma_\beta \Sigma_\beta^{-1} \Sigma_\beta U c = U^\top \beta$. This implies that $c = (U^\top \Sigma_\beta U)^{-1} U^\top \beta$. Plugging this into Eq. (B.5.18) yields $\beta_* = \Sigma_\beta U (U^\top \Sigma_\beta U)^{-1} U^\top \beta$, as desired.

B.5.5 Proof of Corollary 3.4.3

Recall that we wish to show that, for conjugate Bayesian regression, under $\tilde{p}$ the uncertainty (i.e., posterior variance) for any linear combination of parameters, $\mathrm{Var}_{\tilde{p}}[v^\top \beta]$, is no smaller than the exact posterior variance. First, we note that this statement is formally equivalent to stating that $v^\top \tilde{\Sigma}_N v \geq v^\top \Sigma_N v$, or that $E := \tilde{\Sigma}_N - \Sigma_N \succeq 0$ (where $\succeq$ denotes positive definiteness). By Theorem 3.4.1, $\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1} = \bar{U} \mathrm{diag}(\bar{\lambda}^2) \bar{U}^\top \succeq 0$. Since this implies that the inverse of the difference of these matrices is positive definite, we can then see that $(\Sigma_N^{-1} - \tilde{\Sigma}_N^{-1})^{-1} = \tilde{\Sigma}_N (\tilde{\Sigma}_N - \Sigma_N)^{-1} \Sigma_N \succeq 0$. Because, as valid covariance matrices, Σ_N and $\tilde{\Sigma}_N$ are both positive definite, and because inverses and product of positive definite matrices are positive definite, this implies that $\tilde{\Sigma}_N^{-1} \tilde{\Sigma}_N (\tilde{\Sigma}_N - \Sigma_N)^{-1} \Sigma_N \Sigma_N^{-1} = (\tilde{\Sigma}_N - \Sigma_N)^{-1} \succeq 0$. Finally, this implies that $\tilde{\Sigma}_N - \Sigma_N \succeq 0$ as desired.

B.5.6 Information loss due the LR-GLM approximation

We see similar behavior to that demonstrated in Corollary 3.4.3 in the following corollary, which shows that our approximate posterior never has lower entropy than

the exact posterior. Concretely, we look at the reduction of entropy in the approximate posterior relative to the exact posterior [MacKay, 2003], where entropy is defined as:

$$H\big[p(\beta)\big] := \mathbb{E}_p[-\log_2 p(\beta)]$$

Corollary B.5.1. *The entropy $H\big[\tilde{p}(\beta|X,Y)\big]$ is no less than $H\big[p(\beta|X,Y)\big]$. Furthermore, when using an isotropic Gaussian prior $\Sigma_\beta = \sigma_\beta^2 I$, the information loss relative to the exact posterior (in nats) is upper bounded as $H\big[\tilde{p}(\beta|X,Y)\big] - H\big[p(\beta|X,Y)\big] \leq \frac{\tau\sigma_\beta^2}{2}\sum_{i=1}^{D-M}\bar{\lambda}_i^2$.*

This result formalizes the intuition that the LR-GLM approximation reduces the information about the parameter that we are able to extract from the data. Additionally, the upper bound tells us that when U is obtained via an M truncated SVD, at most $\tau\sigma_\beta^2\bar{\lambda}_1^2/2$ additional nats of information would have been provided by using the $M+1$-truncated SVD.

Proof. The entropy of the exact and approximate posteriors are given as:

$$H(p) = -\frac{1}{2}\log|2\pi e\Sigma_N^{-1}| = -\frac{1}{2}\big[D\log 2\pi e + \sum_{i=1}^{D}\log(\sigma_\beta^{-2}+\tau\lambda_i^2)\big]$$

and

$$H(\tilde{p}) = -\frac{1}{2}\log|2\pi e\tilde{\Sigma}_N^{-1}| = -\frac{1}{2}\big[D\log 2\pi e + \sum_{i=1}^{M}\log(\sigma_\beta^{-2}+\tau\lambda_i^2) - \sum_{i=M+1}^{D}\log\sigma_\beta^{-2}\big].$$

Therefore, we conclude that

$$\begin{aligned}
H\big[\tilde{p}(\beta|X)\big] - H\big[p(\beta|X)\big] &= -\frac{1}{2}\sum_{i=1}^{D-M}\log\sigma_\beta^{-2} + \frac{1}{2}\sum_{i=1}^{D-M}\log(\sigma_\beta^{-2}+\tau\bar{\lambda}_i^2) \\
&= \frac{1}{2}\sum_{i=1}^{D-M}\log\frac{\sigma_\beta^{-2}+\tau\bar{\lambda}_i^2}{\sigma_\beta^{-2}} \\
&= \frac{1}{2}\sum_{i=1}^{D-M}\log(1+\frac{\tau}{\sigma_\beta^{-2}}\bar{\lambda}_i^2)
\end{aligned}$$

$$\leq \frac{1}{2} \sum_{i=1}^{D-M} \frac{\tau}{\sigma_\beta^{-2}} \bar{\lambda}_i^2 = \frac{\tau \sigma_\beta^2}{2} \sum_{i=1}^{D-M} \bar{\lambda}_i^2.$$

That $H\big[\tilde{p}(\beta|X)\big] - H\big[p(\beta|X)\big] > 0$ follows from the monotonicity of log, that $\log(1) = 0$, and that $\tau \sigma_\beta^2 \bar{\lambda}_i^2 > 0$ for $i = 1, \ldots, D - M$. □

B.6 Proofs and further results for LR-Laplace in non-conjugate models

In the main text we introduced LR-Laplace as a method which takes advantage of low-rank approximations to provide computational gains when computing a Laplace approximation to the Bayesian posterior. In what follows we verify the theoretical justifications for this approach. Appendix B.6.1 provides a derivation of Algorithm 4 and demonstrates the time complexities of each step, serving as a proof of Theorem 3.5.1. The remainder of the section is devoted to the proofs and discussion of the theoretical properties of LR-Laplace.

B.6.1 Proof of Theorem 3.5.1

Proof of Theorem 3.5.1. The LR-Laplace approximation is defined by mean and covariance parameters, $\hat{\mu}$ and $\hat{\Sigma}$. We prove Theorem 3.5.1 in two parts. First, we show that $\hat{\mu}$ and $\hat{\Sigma}$ do in fact define the Laplace approximation of $\tilde{p}(\beta|X, Y)$, i.e. the construction of $\hat{\mu}$ in Line 9 satisfies $\hat{\mu} = \arg\max_\beta \tilde{p}(\beta|X, Y)$ and that $\hat{\Sigma} = \big(-\nabla_\beta^2 \log \tilde{p}(\beta|X, Y)|_{\beta=\hat{\mu}}\big)^{-1}$. Second, we show that each step of Algorithm 4 may be computed in $O(NDM)$ time with $O(DM + NM)$ storage.

Correctness of $\hat{\mu}$ and $\hat{\Sigma}$

In Line 8, the definition of γ_* implies that $\gamma_* = \arg\max_{\gamma \in \mathbb{R}^M} \tilde{p}_{U^\top \beta|X,Y}(\gamma|X, Y)$ since

$$\begin{aligned}
\log \tilde{p}_{U^\top \beta|X,Y}(\gamma|X, Y) &= \log \tilde{p}_{U^\top \beta}(\gamma) + \log \tilde{p}_{Y|X,U^\top \beta}(Y|X, \gamma) + C \\
&= \log p_{U^\top \beta}(\gamma) + \log p_{Y|X,\beta}(Y|X, U\gamma) + C
\end{aligned}$$

$$
\begin{aligned}
&= \log \mathcal{N}(\gamma | U^\top \mu_\beta, U^\top \Sigma_\beta U) + \sum_{i=1}^N \log p_{y|x,\beta}(y_i | x_i, U\gamma) + C \\
&= -\frac{1}{2}\gamma^\top U^\top \Sigma_\beta U \gamma + \sum_{i=1}^N \phi(y_i, x_i^\top U \gamma) + C',
\end{aligned}
$$

where line 1 uses Bayes' rule, line 2 uses the definition of $\tilde{p}$ in Eq. (3.1), line 3 uses the normality the prior, and the assumed conditional independence of the responses given β, and line 4 follows from the definition of $\phi(\cdot, \cdot)$ and the assumption that $\mu_\beta = 0$. C and C' are constants which do not depend on γ. This together with the following result (proved in Appendix B.6.2) implies that as defined in Line 9 of Algorithm 4, $\hat{\mu} = \arg\max_\beta \tilde{p}(\beta | X, Y)$.

Lemma B.6.1. *Suppose a Gaussian prior $p(\beta) = \mathcal{N}(\mu_\beta, \Sigma_\beta)$, and let*

$$
\gamma_* := \arg\max_{\gamma \in \mathbb{R}^M} \log \tilde{p}_{U^\top \beta | X, Y}(\gamma | X, Y)
$$

. Then $\hat{\mu} := \arg\max_{\beta \in \mathbb{R}^D} \log \tilde{p}(\beta | X, Y)$ may be written as $\hat{\mu} = U\gamma_* + \bar{U}\bar{U}^\top \Sigma_\beta U (U^\top \Sigma_\beta U)^{-1} \gamma_*$.

We now show that as defined in Line 12 of Algorithm 4, $\hat{\Sigma}$ is inverse of the Hessian of the negative log posterior, H. We see this by writing

$$
\begin{aligned}
H :&= \nabla_\beta^2 - \log \tilde{p}(\beta | X, Y)|_{\beta = \hat{\mu}} \\
&= \nabla_\beta^2 - \log \mathcal{N}(\beta | \mu_\beta, \Sigma_\beta)|_{\beta = \hat{\mu}} + \nabla_\beta^2 \sum_{i=1}^N -\phi(y_i, x_i^\top U U^\top \beta)|_{\beta = \hat{\mu}} \\
&= \Sigma_\beta^{-1} + \sum_{i=1}^N -\phi''(y_i, x_i^\top U U^\top \hat{\mu}) x_i U U^\top x_i^\top \\
&= \Sigma_\beta^{-1} + U U^\top X^\top \mathrm{diag}\big(-\vec{\phi}''(Y, XUU^\top \hat{\mu})\big) X U U^\top,
\end{aligned}
$$

where $\vec{\phi}''$ is the second derivative of ϕ. The Woodbury matrix lemma then provides that we may compute $\hat{\Sigma}_N := H^{-1}$ as

$$
\hat{\Sigma}_N = \Sigma_\beta - \Sigma_\beta U \left(U^\top \Sigma_\beta U - \left\{ U^\top X^\top \mathrm{diag}\left[\vec{\phi}''(Y, XUU^\top \hat{\mu}) \right] XU \right\}^{-1} \right)^{-1} U^\top \Sigma_\beta,
$$

which we have written as $\hat{\Sigma} := \Sigma_\beta - \Sigma_\beta U W U^\top \Sigma_\beta$ in Line 12 with $W^{-1} = U^\top \Sigma_\beta U - \left\{ U^\top X^\top \mathrm{diag}\left[\vec{\phi}''(Y, XUU^\top \hat{\mu})\right] XU \right\}^{-1}$.

Time complexity of Algorithm 4

We now prove the asserted time and memory complexities for each line of Algorithm 4.

Algorithm 4 begins with the computation of the M-truncated SVD of $X^\top \approx U \mathrm{diag}(\lambda) V$. As discussed in Section 3.4.1, U may be found in $O(ND \log M)$ time. At the end of this step we must store the projected data $XU \in \mathbb{R}^{N,M}$ and the left singular vectors, $U \in \mathbb{R}^{D,M}$. Which demands $O(NM + DM)$ memory, and the matrix multiply for XU requires $O(NDM)$ time and is the bottleneck step of the algorithm. The matrix V need not be explicitly computed or stored.

The next stage of the algorithm is solving for $\hat{\mu} = \arg\max_\beta \log \tilde{p}(\beta|X, Y)$. This is done in two stages: in Line 8 find $\gamma_* = \arg\max_{\gamma \in \mathbb{R}^M} \log \tilde{p}_{U^\top \beta | X, Y}(\gamma | X, Y)$ as the solution to a convex optimization problem, and in Line 9 find $\hat{\mu}$ as $\hat{\mu} = U\gamma_* + \bar{U}\bar{U}^\top \Sigma_\beta U (U^\top \Sigma_\beta U)^{-1} \gamma_*$. Beginning with Line 8, we note that the function

$$\log \tilde{p}(U^\top \beta | X, Y) = \log p(\beta) + \log \tilde{p}(Y | X, \beta) + c \stackrel{c}{=} \log \mathcal{N}(U^\top \beta | U^\top \mu_\beta, U^\top \Sigma_\beta U) + \sum_{i=1}^{N} \log p(y_i | x_i^\top U U^\top \beta)$$

is a finite sum of functions concave in β and therefore also in $U^\top \beta$. γ_* may therefore be solved to a fixed precision in $O(NM)$ time under the assumptions of our theorem using stochastic optimization algorithms such as stochastic average gradient Schmidt et al. [2017]. In our experiments we use more standard batch convex optimization algorithm (L-BFGS-B Zhu et al. [1997]) which takes at most $O(N^2 M)$ time. This latter upper bound on complexity may be seen from observing each gradient evaluation takes $O(NM)$ time (the cost for the likelihood evaluation, since computing the log prior and its gradient is $O(M^2)$ after computing $U^\top \Sigma_\beta U$ once, which takes $O(DM^2)$ time by assumption) and the number of iterations required can grow up to linearly in the maximum eigenvalue of Hessian, which in turn grows linearly in N Boyd and Vandenberghe [2004].

The second step is computing $\hat{\mu} = U\gamma_* + \bar{U}\bar{U}^\top \Sigma_\beta U (U^\top \Sigma_\beta U)^{-1} \gamma_*$. Given γ_*, this

may be computed in $O(DM)$ time, which one may see by noting that $\bar{U}\bar{U}^\top$ (which we never explicitly compute) may be written as $\bar{U}\bar{U}^\top = (I - UU^\top)$, and finding $\hat{\mu}$ as $\hat{\mu} = U\gamma_* + \Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}\gamma_* - UU^\top\Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}\gamma_*$. By assumption, the structure of Σ_β allows us to compute $U^\top\Sigma_\beta U$ in $O(DM^2)$ time and matrix vector products with Σ_β in $O(D)$ time.

We now turn to the third stage of the algorithm, solving for the posterior covariance $\hat{\Sigma}$, which is represented as an expression of U, Σ_β and W, defined in Line 11. Computing W requires $O(DM)$ and $O(NM^2)$ matrix multiplications (since we have precomputed XU), and two $O(M^3)$ matrix inversions which comes to $O(NM^2 + DM)$ time. The memory complexity of this step is $O(NM)$ since it involves handling XU. Once W has been computed we may use the representation $\hat{\Sigma} = \Sigma_\beta - \Sigma_\beta UWU^\top\Sigma_\beta$ as presented in Line 12. This representation does not entail performing any additional computation (which is why we have written $O(0)$), but as this expression includes U, storing $\hat{\Sigma}$ requires $O(DM)$ memory.

Lastly, we may immediately see that computing posterior variances and covariances takes only $O(M^2)$ time as it involves only indexing into Σ_β and U and $O(M^2)$ matrix-vector multiplies. □

B.6.2 Proof of Lemma B.6.1

We prove the lemma by constructing a rotation of the parameter space by the matrix of singular vectors $[U, \bar{U}]$, in which we have the prior

$$p\Big(\begin{bmatrix} U^\top\beta \\ \bar{U}^\top\beta \end{bmatrix}\Big) = \mathcal{N}\Big(\begin{bmatrix} U^\top\beta \\ \bar{U}^\top\beta \end{bmatrix} \Big| \begin{bmatrix} U^\top\mu_\beta \\ \bar{U}^\top\mu_\beta \end{bmatrix}, \begin{bmatrix} U^\top\Sigma_\beta U, & U^\top\Sigma_\beta\bar{U} \\ \bar{U}^\top\Sigma_\beta U, & \bar{U}^\top\Sigma_\beta\bar{U} \end{bmatrix}\Big).$$

We have that

$$\hat{\mu} := \arg\max_{\beta\in\mathbb{R}^D} \log\tilde{p}(\beta|X, Y)$$

$$
\begin{aligned}
&= [U\,\bar{U}] \underset{U^\top\beta\in\mathbb{R}^M, \bar{U}^\top\beta\in\mathbb{R}^{D-M}}{\arg\max} \log\tilde{p}\Big(\begin{bmatrix} U^\top\beta \\ \bar{U}^\top\beta \end{bmatrix}|X,Y\Big) \\
&= U \underset{U^\top\beta\in\mathbb{R}^M}{\arg\max} \big(\log\tilde{p}(U^\top\beta|X,Y) + \bar{U} \underset{\bar{U}^\top\beta\in\mathbb{R}^{D-M}}{\arg\max} \log\tilde{p}(\bar{U}^\top\beta|U^\top\beta,X,Y)\big) \\
&= U \underset{U^\top\beta\in\mathbb{R}^M}{\arg\max} \log\tilde{p}(U^\top\beta|X,Y) + \\
&\quad \bar{U} \underset{\bar{U}^\top\beta\in\mathbb{R}^{D-M}}{\arg\max} \log\mathcal{N}\big(\bar{U}^\top\beta|\bar{U}^\top\Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}U^\top\beta, \bar{U}\Sigma_\beta\bar{U} - \bar{U}\Sigma_\beta U(U^\top\Sigma_\beta U)U^\top\Sigma_\beta\bar{U}\big) \\
&= U\gamma_* + \bar{U}\bar{U}^\top\Sigma_\beta U(U^\top\Sigma_\beta U)^{-1}\gamma_*.
\end{aligned}
$$

In the second line we simply move to the rotated parameter space. In the third line, we use the chain rule of probability to separate out two terms. To produce the fourth line, we note that since $\tilde{p}(Y|X,\beta) = p(Y|XUU^\top\beta) = \tilde{p}(Y|X,U^\top\beta)$, that Y and $\bar{U}^\top\beta$ are conditionally independent given $U^\top\beta$. We next note that though $\arg\max_{\bar{U}^\top\beta\in\mathbb{R}^{D-M}} \log p(\bar{U}^\top\beta|U^\top\beta)$ depends on $U^\top\beta$, $\max_{\bar{U}^\top\beta\in\mathbb{R}^{D-M}} \log p(\bar{U}^\top\beta|U^\top\beta)$ does not depend $U^\top\beta$. This allows us to use the definition of γ_* to arrive at the fifth line, as desired.

In the special case that Σ_β is diagonal, this expression reduces to $U\gamma_*$. This can be seen by recognizing that $\bar{U}^\top\Sigma_\beta U$ is then $\mathrm{diag}(\mathbf{0})$.

B.6.3 Proof of Theorem 3.5.2

Our approach to proving Theorem 3.5.2 follows a similar approach to that taken to prove Theorem 3.4.1. In particular, we begin by upper bounding the norm of the error of the gradients at the approximate MAP. Noting that the strong log concavity of the exact posterior, which having been assumed to hold globally, must then also hold on $\{t\hat{\mu} + (1-t)\bar{\mu}|t\in[0,1]\}$, we obtain an upper-bound on $\|\hat{\mu}-\bar{\mu}\|_2$ by again applying Lemma B.5.1.

To begin, we first recall that the exact and LR-GLM posteriors may be written as

$$
\log p(\beta|X,Y) = \log p(\beta) + \sum_{n=1}^{N} \phi(y_n|x_n^\top\beta) - \log Z
$$

and

$$\log \tilde{p}(\beta|X,Y) = \log p(\beta) + \sum_{n=1}^{N} \phi(y_n, x_n^\top UU^\top \beta) - \log \tilde{Z}$$

where $\phi(\cdot,\cdot)$ is such that $\phi(y,a) = \log p(y|x^\top\beta = a)$, and Z and $\tilde{Z}$ are the normalizing constants of the exact and approximate posteriors. As a result, the gradients of these log densities are given as

$$\nabla_\beta \log p(\beta|X,Y) = \nabla_\beta \log p(\beta) + X^\top \vec{\phi}'(Y, X\beta)$$

and

$$\nabla_\beta \log \tilde{p}(\beta|X,Y) = \nabla_\beta \log p(\beta) + UU^\top X^\top \vec{\phi}'(Y, XUU^\top \beta),$$

where $\vec{\phi}'(Y, X\beta) \in \mathbb{R}^N$ is such that for each $n \in [N]$, $\vec{\phi}'(Y, X\beta)_n = \frac{d}{da}\phi(y_n, a)|_{a=x_n^\top\beta}$.

And the difference in the gradients is

$$\nabla_\beta \log p(\beta|X,Y) - \nabla_\beta \log \tilde{p}(\beta|X,Y) = X^\top \vec{\phi}'(Y, X\beta) - UU^\top X^\top \vec{\phi}'(Y, XUU^\top\beta). \tag{B.6.1}$$

Appealing to Taylor's theorem, we may write for any β that

$$\phi'(y_n, x_n^\top UU^\top \beta) = \phi'(y_n, x_n^\top \beta) + (x_n^\top UU^\top \beta - x_n^\top \beta)\phi''(y_n, a_n)$$

for some $a_n \in [x_n^\top UU^\top \beta, x_n^\top \beta]$, where $\phi''(y,a) := \frac{d^2}{da^2}\phi(y,a)$.

Using this and introducing vectorized notation for ϕ'' to match that used for $\vec{\phi}'$, we may rewrite the difference in the gradients as

$$\begin{aligned}
&\nabla_\beta \log p(\beta|X,Y) - \nabla_\beta \log \tilde{p}(\beta|X,Y) \\
&= X^\top \vec{\phi}'(Y, X\beta) - UU^\top X^\top \vec{\phi}'(Y, X\beta) - UU^\top X^\top \left[(XUU^\top\beta - X^\top\beta) \circ \vec{\phi}''(Y, A)\right] \\
&= \bar{U}\bar{U}^\top X^\top \vec{\phi}'(Y, X\beta) + UU^\top X^\top \left[(X\bar{U}\bar{U}^\top\beta) \circ \vec{\phi}''(Y, A)\right],
\end{aligned}$$

where $A \in \mathbb{R}^N$ is such that for each $n \in [N]$, $A_n \in [x_n^\top UU^\top \beta, x_n^\top \beta]$, and $\circ$ denotes element-wise scalar multiplication.

We can use this to derive an upper bound on the norm of the difference of the gradients as

$$\begin{aligned}
\|\nabla_\beta \log p(\beta|X,Y) - \nabla_\beta \log \tilde{p}(\beta|X,Y)\|_2 &= \|\bar{U}\bar{U}^\top X^\top \vec{\phi}' + UU^\top X^\top \big[(X\bar{U}\bar{U}^\top \beta) \circ \vec{\phi}''\big]\|_2 \\
&\leq \|\bar{U}\bar{U}^\top X^\top \vec{\phi}'\|_2 + \|UU^\top X^\top \big[(X\bar{U}\bar{U}^\top \beta) \circ \vec{\phi}''\big]\|_2 \\
&\leq \bar{\lambda}_1 \|\vec{\phi}'\|_2 + \lambda_1 \|(X\bar{U}\bar{U}^\top \beta) \circ \vec{\phi}''\|_2 \\
&\leq \bar{\lambda}_1 \|\vec{\phi}'\|_2 + \lambda_1 \bar{\lambda}_1 \|\bar{U}^\top \beta\|_2 \|\vec{\phi}''\|_\infty \\
&= \bar{\lambda}_1 \big(\|\vec{\phi}'\|_2 + \lambda_1 \|\bar{U}^\top \beta\|_2 \|\vec{\phi}''\|_\infty\big),
\end{aligned}$$

where we have written $\vec{\phi}'$ and $\vec{\phi}''$ in place of $\vec{\phi}'(Y, X\beta)$ and $\vec{\phi}''(Y, A)$, respectively, for brevity despite their dependence on β.

Next, let α be the strong log-concavity parameter of $p(\beta|X,Y)$. Lemma B.5.1 then implies that

$$\|\hat{\mu} - \bar{\mu}\|_2 \leq \frac{\bar{\lambda}_1 \big(\|\vec{\phi}'(Y, X\hat{\mu})\|_2 + \lambda_1 \|\bar{U}^\top \hat{\mu}\|_2 \|\vec{\phi}''(Y, A)\|_\infty\big)}{\alpha}$$

as desired, where for each $n \in [N]$, $A_n \in [x_n^\top UU^\top \hat{\mu}, x_n^\top \hat{\mu}]$.

B.6.4 Bounds on derivatives of higher order for the log-likelihood in logistic regression and other GLMs

We here provide some additional support for the claim that in Remark 3.5.3 that the higher order derivatives of the log-likelihood function, ϕ, are well-behaved. For logistic regression (which we explore in detail below), for any y in $\{-1, 1\}$ and a in $\mathbb{R}$, it holds that $|\frac{\partial}{\partial a}\phi(y,a)| \leq 1$ and $|\frac{\partial^2}{\partial^2 a}\phi(y,a)| \leq \frac{1}{4}$. For Poisson regression with $\phi(y,a) = \log \text{Pois}\big(y|\lambda = \log(1+\exp\{a\})\big)$, both $|\frac{\partial}{\partial a}(y,a)|$ and $|\frac{\partial^2}{\partial^2 a}\phi(y,a)|$ are bounded by a small constant factor of y. Additionally, in these cases $|\frac{\partial^3}{\partial a^3}\phi(y,a)|$ is also well behaved, a fact relevant to Corollary 3.5.6. However, for alternative mapping functions for Poisson regression, e.g. defining $\mathbb{E}[y_i|x_i,\beta] = \exp\{x_i^\top \beta\}$, these derivatives will grow exponentially quickly with $x_i^\top \beta$, which illustrates that our provided bounds are

sensitive to the particular form chosen for the GLM likelihood.

We now move to compute explicit upper bounds on the derivatives of the log likelihood in logistic regression. This produces the constants mentioned above, and permits easy computation of upper bounds on the bounds on the approximation error of LR-Laplace provided in Theorem 3.5.2 and Corollary 3.5.6. In particular the logistic regression mapping function [Huggins et al., 2017] is given as

$$\phi(y_n, x_n^\top \beta) = -\log\left(1 + \exp\{-y_n x_n^\top \beta\}\right), \tag{B.6.2}$$

where each $y_n \in \{-1, 1\}$.

The first three derivatives of this mapping function and bounds on their absolute values are as follows:

$$\phi'(y_n, x_n^\top \beta) := \frac{d}{da}\phi(y_n, a)\big|_{a=x_n^\top \beta} = y_n \frac{\exp\{-y x_n^\top \beta\}}{1 + \exp\{-y_n x_n^\top \beta\}} \tag{B.6.3}$$

Notably, $\forall a \in \mathbb{R}, y \in \{-1, 1\}$, $|\phi'(y, a)| < 1$ and

$$\phi''(y_n, x_n^\top \beta) := \frac{d^2}{da^2}\phi(y, a)\big|_{a=x_n^\top \beta} = -(1 + \exp\{x_n^\top \beta\})^{-1}(1 + \exp\{-x_n^\top \beta\})^{-1}. \tag{B.6.4}$$

Furthermore, for any a in $\mathbb{R}$ and y in $\{-1, 1\}$, $-\frac{1}{4} \leq \phi''(y, a) < 0$. This implies that the Hessian of the negative log likelihood will be positive semi-definite everywhere. We additionally have

$$\frac{d^3}{da^3}\phi(y, a) = \phi'''(a) = \frac{\left(\exp\{a\}(\exp(-a) - 1)\right)}{(1 + \exp\{a\})^3} \tag{B.6.5}$$

which for any a in $\mathbb{R}$ satisfies, $-\frac{1}{6\sqrt{3}} \leq \phi'''(a) \leq \frac{1}{6\sqrt{3}}$.

B.6.5 Asymptotic inconsistency of the approximate posterior mean within the span of the projections

Consider a Bayesian logistic regression, in which

$$x_i \sim \mathcal{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0.99 \end{bmatrix} \Big), \quad \beta = \begin{bmatrix} 10 \\ 1000 \end{bmatrix}, \quad y_i \sim \text{Bern}\big((1 + \exp\{x_i^\top \beta\})^{-1}\big).$$

In this setting, a rank 1 approximation of the design will capture only the first dimension of data (i.e. $U_N \to U_* = [1, 0]$). However the second dimension explains almost all of the variance in the responses. As such $y_i | U_*^\top x_i, \beta \overset{d}{\approx} \text{Bern}(1/2)$ and we will get $U_*^\top \beta | X, Y = \beta_1 | X, Y \approx 0.0$ under $\tilde{p}$.

B.6.6 Proof of Corollary 3.5.6

Our proof proceeds via an upper bound on the $(2, \hat{p})$-Fisher distance between $\hat{p}$ and $\bar{p}$ [Huggins et al., 2018]. Specifically, the $(2, \hat{p})$-Fisher distance given by

$$d_{2,\hat{p}}(\hat{p}, \bar{p}) = \left(\int \|\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta)\|_2^2 dp(\beta) \right)^{\frac{1}{2}}. \tag{B.6.6}$$

Given the strong log-concavity of $\bar{p}$, our upper bound on this Fisher distance immediately provides an upper-bound on the 2-Wasserstein distance [Huggins et al., 2018].

We first recall that $\hat{p}$ and $\bar{p}$ are defined by Laplace approximations of $\tilde{p}(\beta | X, Y)$ and $p(\beta | X, Y)$ respectively. As such we have that

$$\log \hat{p}(\beta) \overset{c}{=} -\frac{1}{2}(\beta - \hat{\mu})^\top \big(\Sigma_\beta^{-1} - UU^\top X^\top \text{diag}(\vec{\phi}''(Y, XUU^\top \hat{\mu})) XUU^\top \big)(\beta - \hat{\mu})$$

where $\vec{\phi}''(Y, XUU^\top \hat{\mu})$ is defined as in Algorithm 1 such that

$$\vec{\phi}''(Y, X\beta)_i = \frac{d^2}{da^2} \log p(y_i | x^\top \beta = a)|_{a = x_i^\top \beta}$$

, and

$$\log \bar{p}(\beta) \stackrel{c}{=} -\frac{1}{2}(\beta - \bar{\mu})^\top \big(\Sigma_\beta^{-1} - X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))X\big)(\beta - \bar{\mu}).$$

Accordingly,

$$\nabla_\beta \log \hat{p}(\beta) = -(\beta - \hat{\mu})^\top \big(\Sigma_\beta^{-1} - UU^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, XUU^\top \hat{\mu}))XUU^\top\big)$$

and

$$\nabla_\beta \log \bar{p}(\beta) = -(\beta - \bar{\mu})^\top \big[\Sigma_\beta^{-1} - X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))X\big]$$

To define an upper bound on $d_{2,\hat{p}}(\hat{p}, p)$, we must consider the difference between the gradients,

$$\begin{aligned}
\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta) = &- (\beta - \hat{\mu})^\top \big\{\Sigma_\beta^{-1} - UU^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, XUU^\top \hat{\mu})]XUU^\top\big\} \\
&+ (\beta - \bar{\mu})^\top \big\{\Sigma_\beta^{-1} - X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]X\big\} \\
&= (\hat{\mu} - \bar{\mu})\Sigma_\beta^{-1} + (\beta - \hat{\mu})^\top UU^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, XUU^\top \hat{\mu})]XUU^\top \\
&- (\beta - \bar{\mu})^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]X.
\end{aligned}$$

Appealing to Taylor's theorem, we can rewrite $\vec{\phi}''(Y, XUU^\top \hat{\mu})$ as

$$\begin{aligned}
\vec{\phi}''(Y, XUU^\top \hat{\mu}) &= \vec{\phi}''(Y, X\bar{\mu}) + (XUU^\top \hat{\mu} - X\bar{\mu}) \circ \vec{\phi}'''(Y, A) \\
&= \vec{\phi}''(Y, X\bar{\mu}) + (XUU^\top \hat{\mu} - X\hat{\mu} + X(\hat{\mu} - \bar{\mu})) \circ \vec{\phi}'''(Y, A) \\
&= \vec{\phi}''(Y, X\bar{\mu}) - X\bar{U}\bar{U}^\top \circ \vec{\phi}'''(Y, A) + X(\hat{\mu} - \bar{\mu}) \circ \vec{\phi}'''(Y, A) \\
&= \vec{\phi}''(Y, X\bar{\mu}) + R,
\end{aligned}$$

where the first line follows from Taylor's theorem by appropriately choosing each $A_i \in [x_i^\top UU^\top \hat{\mu}, x_i^\top \bar{\mu}]$, and in the fourth line we substitute in $R := -X\bar{U}\bar{U}^\top \circ \vec{\phi}'''(Y, A) + X(\hat{\mu} - \bar{\mu}) \circ \vec{\phi}'''(Y, A)$.

We now can rewrite the difference in the gradients as

$$\begin{aligned}
\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta) &= (\hat{\mu} - \bar{\mu})\Sigma_\beta^{-1} \\
&\quad + (\beta - \hat{\mu})^\top UU^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]XUU^\top \\
&\quad + (\beta - \hat{\mu})^\top UU^\top X^\top \mathrm{diag}(R)XUU^\top \\
&\quad - (\beta - \bar{\mu})^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))X \\
&= (\hat{\mu} - \bar{\mu})^\top (\Sigma_\beta^{-1} - UU^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]XUU^\top) \\
&\quad + (\beta - \hat{\mu})^\top UU^\top X^\top \mathrm{diag}(R)XUU^\top \\
&\quad - (\beta - \bar{\mu})^\top UU^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))X\bar{U}\bar{U}^\top \\
&\quad - (\beta - \bar{\mu})^\top \bar{U}\bar{U}^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))XUU^\top \\
&\quad - (\beta - \bar{\mu})^\top \bar{U}\bar{U}^\top X^\top \mathrm{diag}(\vec{\phi}''(Y, X\bar{\mu}))X\bar{U}\bar{U}^\top.
\end{aligned}$$

Which is obtained by first writing $X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]X$ in the fourth line as $(UU^\top X^\top + \bar{U}\bar{U}^\top X^\top)\mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})](XUU^\top + X\bar{U}\bar{U}^\top)$, multiplying through and re-arranging the resulting terms.

Given this form of the difference in the gradients, we may upper bound its norm as

$$\begin{aligned}
&\|\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta)\|_2 \\
&\leq \|\hat{\mu} - \bar{\mu}\|_2 \|\Sigma_\beta^{-1} - UU^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]XUU^\top\|_2 \qquad \text{(B.6.7)} \\
&\qquad + \|\beta - \hat{\mu}\|_2 \|UU^\top X^\top \mathrm{diag}(R)XUU^\top\|_2 \\
&\qquad + \|\beta - \bar{\mu}\|_2 \|UU^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]X\bar{U}\bar{U}^\top + \\
&\qquad\quad \bar{U}\bar{U}^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]XUU^\top + \bar{U}\bar{U}^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]X\bar{U}\bar{U}^\top\|_2 \\
&\leq \|\hat{\mu} - \bar{\mu}\|_2 \|\Sigma_\beta^{-1} - UU^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]XUU^\top\|_2 \\
&\qquad + \|\beta - \hat{\mu}\|_2 \|UU^\top X^\top \mathrm{diag}(R)XUU^\top\|_2 \\
&\qquad + \|\beta - \bar{\mu}\|_2 \big\{ \|\bar{U}\bar{U}^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]X\bar{U}\bar{U}^\top\|_2 + \\
&\qquad\quad 2\|\bar{U}\bar{U}^\top X^\top \mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]XUU^\top\|_2 \big\}
\end{aligned}$$

by the triangle inequality.

$$\leq \|\hat{\mu} - \bar{\mu}\|_2 \left\{ \|\Sigma_\beta^{-1}\|_2 + \|U\mathrm{diag}(\lambda)V^\top\|_2 \|\mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]\|_2 \|V\mathrm{diag}(\lambda)U^\top\|_2 \right\}$$

$$+ \|\beta - \hat{\mu}\|_2 \|U\mathrm{diag}(\lambda)V^\top\|_2 \|\mathrm{diag}(R)\|_2 \|V\mathrm{diag}(\lambda)U^\top\|_2$$
$$+ \|\beta - \bar{\mu}\|_2 \big\{ \|\bar{U}\mathrm{diag}(\bar{\lambda})\bar{V}^\top\|_2 \|\mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]\|_2 \|\bar{V}\mathrm{diag}(\bar{\lambda})\bar{U}^\top\|_2 +$$
$$2\|\bar{U}^\top\mathrm{diag}(\bar{\lambda})\bar{V}^\top\|_2 \|\mathrm{diag}[\vec{\phi}''(Y, X\bar{\mu})]\|_2 \|V\mathrm{diag}(\lambda)U^\top\|_2 \big\}$$

by again using the triangle inequality, and decomposing $X^\top$ into $U\mathrm{diag}(\lambda)V^\top + \bar{U}\mathrm{diag}(\bar{\lambda})\bar{V}^\top$.

$$\leq \|\hat{\mu} - \bar{\mu}\|_2 \big(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty\big) + \lambda_1^2 \|\beta - \hat{\mu}\|_2 \|R\|_\infty + (\bar{\lambda}_1^2 + 2\lambda_1\bar{\lambda}_1)\|\beta - \bar{\mu}\|_2 \|\vec{\phi}''\|_2,$$

where in the last line we have shortened $\vec{\phi}''(Y, X\bar{\mu})$ to $\vec{\phi}''$ for convenience.

Next noting that $\|\bar{\mu} - \hat{\mu}\|_2 \leq \bar{\lambda}_1 c$ for $c := \frac{\|\vec{\phi}'(Y, X\hat{\mu})\|_2 + \lambda_1 \|\bar{U}^\top \hat{\mu}\|_2 \|\vec{\phi}''(Y, A)\|_\infty}{\alpha}$, where α is the strong log concavity parameter of $p(\beta|X, Y)$ (which follows from Theorem 3.5.2), we can see that $\|R\|_\infty \leq \bar{\lambda}_1 r$ where $r := (\|U^\top \hat{\mu}\|_\infty \|\vec{\phi}'''(Y, A)\|_\infty + \lambda_1 c \|\vec{\phi}'''(Y, A)\|_\infty)$. That r is bounded follows from the assumption that $\log p(y|x, \beta)$ has bounded third derivatives, an equivalent to a Lipschitz condition on ϕ''. We can next simplify this upper bound to

$$\|\nabla_\beta \log \hat{p}(\beta) - \nabla_\beta \log \bar{p}(\beta)\|_2$$
$$\leq \bar{\lambda}_1 c[\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty] + \lambda_1^2 \bar{\lambda}_1 r \|\beta - \hat{\mu}\|_2 + \bar{\lambda}_1(\bar{\lambda}_1 + 2\lambda_1)\|\beta - \bar{\mu}\|_2 \|\vec{\phi}''\|_\infty \quad \text{(B.6.8)}$$
$$= \bar{\lambda}_1 \big[c(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty) + \lambda_1^2 r \|\beta - \hat{\mu}\|_2 + (\bar{\lambda}_1 + 2\lambda_1)\|\beta - \bar{\mu}\|_2 \|\vec{\phi}''\|_\infty\big]$$
$$\leq \bar{\lambda}_1 \big[c(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty) + \lambda_1^2 r \|\beta - \hat{\mu}\|_2 + (\bar{\lambda}_1 + 2\lambda_1)(\|\hat{\mu} - \bar{\mu}\|_2 + \|\beta - \hat{\mu}\|_2)\|\vec{\phi}''\|_\infty\big]$$

by the triangle inequality.

$$\leq \bar{\lambda}_1 \big[c(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty) + \lambda_1^2 r \|\beta - \hat{\mu}\|_2 + (\bar{\lambda}_1 + 2\lambda_1)(\bar{\lambda}_1 c + \|\beta - \hat{\mu}\|_2)\|\vec{\phi}''\|_\infty\big]$$
$$= \bar{\lambda}_1 \big[c(\|\Sigma_\beta^{-1}\|_2 + \lambda_1^2 \|\vec{\phi}''\|_\infty) + c(\bar{\lambda}_1^2 + 2\lambda_1\bar{\lambda}_1)\|\vec{\phi}''\|_\infty + (\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty)\|\beta - \hat{\mu}\|_2\big]$$
$$= \bar{\lambda}_1 \big[c(\|\Sigma_\beta^{-1}\|_2 + (\lambda_1 + \bar{\lambda}_1)^2 \|\vec{\phi}''\|_\infty) + (\lambda_1^2 r + (\bar{\lambda}_1 + 2\lambda_1)\|\vec{\phi}''\|_\infty)\|\beta - \hat{\mu}\|_2\big].$$

Thus, taking the expectation of this upper bound on the norm squared over β with respect to $\hat{p}$ we get

$$d_{2,\hat{p}}^2(\hat{p}, p)$$

$$\leq \mathbb{E}_{\hat{p}(\beta)}\left(\bar{\lambda}_1^2\left\{c\left[\|\Sigma_\beta^{-1}\|_2+(\lambda_1+\bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]+\left[\lambda_1^2 r+(\bar{\lambda}_1+2\lambda_1)\|\vec{\phi}''\|_\infty\right]\|\beta-\hat{\mu}\|_2\right\}^2\right) \tag{B.6.9}$$

$$\leq 2\bar{\lambda}_1^2\mathbb{E}_{\hat{p}(\beta)}\left\{c^2\left[\|\Sigma_\beta^{-1}\|_2+(\lambda_1+\bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]^2+\left[\lambda_1^2 r+(\bar{\lambda}_1+2\lambda_1)\|\vec{\phi}''\|_\infty\right]^2\|\beta-\hat{\mu}\|_2^2\right\}$$

since $\forall a,b\in\mathbb{R},(a+b)^2\leq 2(a^2+b^2)$

$$=2\bar{\lambda}_1^2\left\{c^2\left[\|\Sigma_\beta^{-1}\|_2+(\lambda_1+\bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]^2+\left[\lambda_1^2 r+(\bar{\lambda}_1+2\lambda_1)\|\vec{\phi}''\|_\infty\right]^2\mathbb{E}_{\hat{p}(\beta)}[\|\beta-\hat{\mu}\|_2^2]\right\}$$

$$=2\bar{\lambda}_1^2\left\{c^2\left[\|\Sigma_\beta^{-1}\|_2+(\lambda_1+\bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]^2+\left[\lambda_1^2 r+(\bar{\lambda}_1+2\lambda_1)\|\vec{\phi}''\|_\infty\right]^2\mathrm{tr}(\hat{\Sigma})\right\}.$$

Next noting that $\bar{p}$ is strongly $\|\bar{\Sigma}\|_2^{-1}$ log-concave, we may apply Theorem B.6.1, stated below, to obtain that

$$W_2(\hat{p},\bar{p})\leq\|\bar{\Sigma}\|_2\sqrt{2\bar{\lambda}_1^2\left\{c^2\left[\|\Sigma_\beta^{-1}\|_2+(\lambda_1+\bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]^2+\left[\lambda_1^2 r+(\bar{\lambda}_1+2\lambda_1)\|\vec{\phi}''\|_\infty\right]^2\mathrm{tr}(\hat{\Sigma})\right\}}$$

$$\leq\sqrt{2}\bar{\lambda}_1\|\bar{\Sigma}\|_2\left\{c\left[\|\Sigma_\beta^{-1}\|_2+(\lambda_1+\bar{\lambda}_1)^2\|\vec{\phi}''\|_\infty\right]+\left[\lambda_1^2 r+(\bar{\lambda}_1+2\lambda_1)\|\vec{\phi}''\|_\infty\right]\sqrt{\mathrm{tr}(\hat{\Sigma})}\right\},$$

which is our desired upper bound.

Theorem B.6.1. *Suppose that $p(\beta)$ and $q(\beta)$ are twice continuously differentiable and that q is α-strongly log concave. Then*

$$W_2(p,q)\leq\alpha^{-1}d_{2,p}(p,q),$$

where W_2 denotes the 2-Wasserstein distance between p and q.

Proof. This follows from Huggins et al. [2018] Theorem 5.2, or similarly from Bolley et al. [2012] Lemma 3.3 and Proposition 3.10. □

B.6.7 Proof of bounded asymptotic error

We here provide a formal statement and proof of Theorem 3.5.7, detailing the required regularity conditions.

Theorem B.6.2 (Asymptotic)**.** *Assume* $x_i \overset{\text{i.i.d.}}{\sim} p_*$ *for some distribution* p_* *such that* $\mathbb{E}_{p_*}[x_i x_i^\top]$ *exists and is non-singular with diagonalization* $\mathbb{E}_{p_*}[x_i x_i^\top] = U_*^\top \operatorname{diag}(\lambda) U_* + \bar{U}_*^\top \operatorname{diag}(\bar{\lambda}) \bar{U}_*$ *such that* $\min(\lambda) > \max(\bar{\lambda})$*. Additionally, for a strictly concave (in its second argument), twice differentiable log-likelihood function* ϕ *with bounded second derivatives (in both arguments) and some* $\beta \in \mathbb{R}^D$*, let* $y_i | x_i \sim \exp\{\phi(y_i, x_i^T \beta)\}$*. Also, suppose that* $\mathbb{E}\|y_i\|_2^2 < \infty$*. Then if* $p(\beta)$ *is log-concave and positive on* $\mathbb{R}^D$*, the asymptotic error (in* N*) of the exact relative to approximate maximum a posteriori parameters,* $\hat{\mu} = \lim_{N\to\infty} \hat{\mu}_N$ *and* $\bar{\mu} = \lim_{N\to\infty} \bar{\mu}_N$ *is finite (where* $\hat{\mu}_N$ *and* $\bar{\mu}_N$ *are the approximate and exact MAP estimates, respectively, after* N *data-points), i.e.,* $\lim_{n\to\infty} \|\hat{\mu}_N - \bar{\mu}_N\|$ *exists and is finite.*

Proof. Before beginning, let $\mathbb{P}$ denote a Borel probability measure on the sample space on which our random variables, $\{x_i\}$ and $\{y_i\}$, are defined such that these random variables are distributed as assumed according to $\mathbb{P}$. In what follows we demonstrate the asymptotic error is finite $\mathbb{P}$-almost surely. To this end, it suffices to show that $\hat{\mu}_N \overset{a.s.}{\to} \hat{\mu}$ and $\bar{\mu}_N \overset{a.s.}{\to} \bar{\mu}$ for some $\hat{\mu}, \bar{\mu}$ in $\mathbb{R}^D$.

Strong convergence of the exact MAP ($\bar{\mu}_N \overset{a.s.}{\to} \bar{\mu}$)

This follows from Doob's consistency theorem [Van der Vaart, 2000, Theorem 10.10]. The only nuance required in the application of this theorem here is that we must accommodate the regression setting. However by constructing a single measure $\mathbb{P}$ governing both the covariates and responses, this simply becomes a special case of the usual theorem for unconditional models.

Strong convergence of the approximate MAP ($\hat{\mu}_N \overset{a.s.}{\to} \hat{\mu}$)

In contrast to the strong consistency of $\bar{\mu}_N$, showing convergence of $\hat{\mu}_N$ requires more work. This is because we cannot rely on standard results such as Bernstein–Von Mises or Doob's consistency theorem, which require correct model specification. Since we have introduced the likelihood approximation $\tilde{p}(y|x,\beta) \neq p(y|x,\beta)$, the vector $\hat{\mu}_N$ is the MAP estimate under a misspecified model.

We demonstrate almost sure convergence in two steps; first we show that $U_*^\top \hat{\mu}_N$ converges almost surely to some $\gamma^* \in \mathbb{R}^M$; then we show that $\hat{\mu}_N = U_* U_*^\top \hat{\mu}_N + \bar{U}_N \bar{U}_N^\top \hat{\mu}_N$ must converge as a result. Since $U_N U_N^\top \overset{a.s.}{\to} U_* U_*^\top$ (as follows from entry-wise almost sure convergence of $\frac{1}{N} X^\top X \to \mathbb{E}_{p_*}[x_i x_i^\top]$ and the Davis–Kahan Theorem [Davis and Kahan, 1970]), this guarantees strong convergence of $\hat{\mu}_N = U_N U_N^\top \hat{\mu}_N + \bar{U}_N \bar{U}_N^\top \hat{\mu}_N$.

Part I: strong convergence of the projected approximate MAP, $U_* \hat{\mu}_N \overset{a.s.}{\to} \gamma^*$

Let $U_* \in \mathbb{R}^{D,M}$ be the top M eigenvectors of $\mathbb{E}_{p_*}[x_i x_i^\top]$, and recall that by assumption for any $y, \phi(y, x_i^T \beta)$ is a strictly concave function of $x_i^\top \beta$, in the sense that for any y and any b, b' in $\mathbb{R}$ and t in $(0,1)$ with $b \neq b'$, $\phi(y, tb + (1-t)b') > t\phi(y,b) + (1-t)\phi(y,b')$. Then by Lemma B.6.2 we have that there is a unique maximizer $\gamma^* = \arg\max_{\gamma \in R^M} \mathbb{E}[\phi(y, x^\top U_* \gamma)]$

We next note that the Hessian of the expected approximate negative log likelihood with respect to γ is positive definite everywhere,

$$\begin{aligned}
&\nabla_\gamma^2 - \mathbb{E}_{y \sim p(y|x,\beta), x \sim p_*}[\phi(y, x^\top U_* \gamma)] \\
&\quad = -\mathbb{E}[\big(\nabla_\gamma \phi'(y, x^\top U_* \gamma)\big) x^\top U_*] = -U_*^\top \mathbb{E}[x \phi''(y, x^\top U_* \gamma) x^\top] U_* \succ 0
\end{aligned}$$

since the strict log concavity and twice differentiability of ϕ ensure that

$$-\mathbb{E}[x \phi''(y, x^\top U_* \gamma) x^\top] \succ 0$$

.

Now consider any compact neighborhood $K \subset \mathbb{R}^M$ containing γ^* as an interior point. Then, by Lemma B.6.3 the set $\mathcal{F} = \{f_\gamma : X \times Y \to \mathbb{R}, (x,y) \mapsto \phi(y, x^\top U_* \gamma) | \gamma \in K\}$ is $\mathbb{P}$-Glivenko–Cantelli. As such $\sup_{f_\gamma \in \mathcal{F}} |\frac{1}{N} \sum_{i=1}^N f_\gamma(x_i, y_i) - \mathbb{E}[f_\gamma(x_i, y_i)]| \overset{a.s.}{\to} 0$, that is to say, the empirical average log-likelihood converges uniformly to its expectation across all $\gamma \in K$. As a result, we have that for $\gamma_N := \arg\max_{\gamma \in K} \log \tilde{p}(U_* \beta = \gamma | X, Y) = \arg\max_{\gamma \in K} \frac{1}{N} \big[\log p(U_*^T \beta = \gamma) + \sum_{i=1}^N \phi(y_i, x_i^\top U_* \gamma) \big]$, $\gamma_N \overset{a.s.}{\to} \gamma^*$.

It remains in this part only to show that convergence of the approximate MAP

parameter within this subset K implies convergence of $U_*^\top \hat{\mu}$, the approximate MAP parameter (across all of $\mathbb{R}^M$). However, this follows immediately from the strict log concavity of the posterior; because $\gamma^* \in K^\circ$, for N large enough each $\gamma_N \in K^\circ$ and we may construct a sub-level set such that $\gamma_N \in C_N \subset K$ such that $\forall \gamma \notin C_N, \log p(U_*^\top \beta = \gamma) + \sum_{i=1}^N \phi(y_i, x_i^\top U_* \gamma) < \log p(U_*^\top \beta = \gamma_N) + \sum_{i=1}^N \phi(y_i, x_i^\top U_* \gamma_N)$.

Part II: convergence of $\bar{U}_* \bar{U}_*^\top \hat{\mu}_N + U_* \gamma$

Using the result of Part I, we can write that $\hat{\mu}_N = U_* U_*^\top \hat{\mu}_N + \bar{U}_* \bar{U}_*^\top \hat{\mu}_N \to U_* \gamma^* + \bar{U}_* \bar{U}_*^\top \hat{\mu}_N$. However, since $\bar{U}_* \bar{U}_*^\top \beta \perp X, Y | U_*^\top \beta$ under $\mathbb{P}$, convergence of $U_*^\top \hat{\mu}_N \to \gamma^*$ implies convergence of $\arg\max_{\bar{U}_*^\top \beta} \tilde{p}(\bar{U}_*^\top \beta | U_*^\top \beta = U_*^\top \hat{\mu}_N, X, Y) = \arg\max_{\bar{U}_*^\top \beta} \tilde{p}(\bar{U}_*^\top \beta | U_*^\top \beta = U_*^\top)$ to some $\bar{U}_*^\top \hat{\mu}_N$ since continuity of $p(\beta)$ and $\tilde{p}(Y|X, \beta)$ imply continuity of the arg-max. Thus both $\hat{\mu}_N$ and $\bar{\mu}_N$ converge, guaranteeing convergence of the asymptotic error. □

Lemma B.6.2. *For any $\phi(\cdot, \cdot)$ which is strictly concave in its second argument, if there is a global maximizer $\beta^* = \arg\max_{\beta \in \mathbb{R}^D} V(\beta) = \mathbb{E}_{x \sim p_*, y \sim p(y|x,\beta)}[\phi(y, x^\top \beta)]$, then there is a unique global maximizer,*

$$\gamma^* = \arg\max_{\gamma \in \mathbb{R}^M} V(U_* \gamma)$$

Proof. We first note that $V(\cdot)$ must have bounded sub-level sets. Thus $W(\cdot) := V(U_* \cdot)$ must also have bounded sub-level sets since $V^{-1}([a, \infty]) = \{\beta | V(\beta) \geq a\} \supset \{\beta | \exists \gamma \in \mathbb{R}^M \text{ s.t. } \beta = U_* \gamma \text{ and } V(U_* \gamma) \geq a\} = U_* W^{-1}([a, \infty])$. Thus, since W is strictly concave and has bounded sub-level sets, it has a unique maximizer. □

Lemma B.6.3. *Let $K \subset \mathbb{R}^M$ be compact and denote by X and Y the domains of the covariates and responses, respectively. Then under the assumptions of Theorem B.6.2, the set $\mathcal{F} = \{f_\gamma : X \times Y \to \mathbb{R}, (x, y) \mapsto \phi(y, x^\top U \gamma) | \gamma \in K\}$ is $\mathbb{P}$-Glivenko–Cantelli.*

Proof. This result follows from Theorem 19.4 in Van der Vaart [2000], and builds from example 19.7 of the same reference; in particular, the condition of bounded second derivatives of ϕ implies that for any $f_\gamma, f_{\gamma'}$ in $\mathcal{F}$ and x in X, y in Y, we have $|f_\gamma(x, y) - f_{\gamma'}(x, y)| \leq C \|x\|_2^2$. The previous condition is sufficient to ensure finite

bracketing numbers, and the result follows. Notably, in keeping with example 19.7 we have that for all x, y and for all γ and γ' in K,

$$\begin{aligned}
\|f_\gamma(x,y) - f_{\gamma'}(x,y)| &= |\int_{x^\top U\gamma'}^{x^\top U\gamma} \phi'(y,a)da\| \\
&= \|\int_{x^\top U\gamma'}^{x^\top U\gamma} \phi'(y, x^\top U\gamma') + \int_{x^\top U\gamma'}^{a} \phi''(y,b)db\ da\| \\
&\leq \|x^\top U(\gamma-\gamma')\phi'(y,x^\top U\gamma')\|_2 + \frac{1}{2}\|x^\top U(\gamma-\gamma')\|_2^2 \sup_{a\in\mathbb{R}} \phi''(y,a) \\
&\leq \|x^\top U\|_2(\|y\|_2 + \|x^\top U\|_2\|\gamma'\|_2)\phi''_{\max}\|\gamma-\gamma'\|_2 \\
&+ \frac{1}{2}\|x^\top U\|_2^2\|\gamma-\gamma'\|_2^2\phi''_{\max} \\
&\leq \big[\frac{3}{2}\|x^\top U\|_2^2 \mathrm{diam}(K)\phi''_{\max} + \|x^\top U\|_2\|y\|_2\phi''_{\max}\big]\|\gamma-\gamma'\|_2 \\
&\leq C(\|x^\top U\|_2^2 + \|x^\top U\|_2\|y\|_2)\|\gamma-\gamma'\|_2,
\end{aligned} \tag{B.6.10}$$

where in the first and second lines we use the fundamental theorem of calculus, and in the fourth and fifth lines we rely on the boundedness of the second derivatives of ϕ and that the compactness subsets of $\mathbb{R}^M$ implies boundedness. In the final line C is an absolute constant.

Finally, we note that $\mathbb{E}_\mathbb{P}\|x^\top U\|_2^2 < \infty$ since $\mathbb{E}_\mathbb{P}\|x^\top U\|_2^2 = \mathbb{E}_\mathbb{P} x^\top UU^\top x < \mathbb{E}_\mathbb{P} x^\top x = \mathrm{Tr}(\mathbb{E}_{p_*} xx^\top) < \infty$, and by Cauchy Schwartz, $\mathbb{E}_\mathbb{P}\|x^\top U\|_2\|y\|_2 \leq \sqrt{\mathbb{E}_\mathbb{P}\|x^\top U\|_2^2 \mathbb{E}_\mathbb{P}\|y\|_2^2} \leq \infty$. This confirms (as in example 19.7 Van der Vaart [2000]) that for all $\epsilon > 0$, the ϵ-bracketing number of $\mathcal{F}$ is finite. By Theorem 19.4 of Van der Vaart [2000], this proves that $\mathcal{F}$ is $\mathbb{P}$-Glivenko-Cantelli. □

B.6.8 Factorized Laplace approximations underestimate marginal variances

We here illustrate that the factorized Laplace approximation underestimates marginal variances. Consider for simplicity the case of a bivariate Gaussian with

$$\Sigma = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

for which the Hessian evaluated anywhere is

$$\Sigma^{-1} = \frac{1}{ac - b^2} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix}.$$

Ignoring off diagonal terms and inverting to approximate Σ_N, as is done by a diagonal Laplace approximation, yields:

$$\tilde{\Sigma} = \begin{bmatrix} a - \frac{b^2}{c} & 0 \\ 0 & c - \frac{b^2}{a} \end{bmatrix}.$$

This approximation reports marginal variances which are lower than the exact marginal variances.

That this approximation underestimates marginal variances in the more general $D > 2$ dimensional case may be easily seen from considering the block matrix inversion of Σ, with blocks of dimension 1×1, $(D-1) \times 1$, $1 \times (D-1)$ and $(D-1) \times (D-1)$, and noting that the Schur complement of a positive definite covariance matrix will always be positive definite.

B.7 LR-MCMC

We provide the LR-MCMC algorithm for performing fast MCMC in generalized linear models with low-rank data approximations.

Algorithm 6 LR-MCMC for Bayesian inference in GLMs with low-rank data approximations.

1: ***Input:*** *prior $p(\beta)$, data $X \in \mathbb{R}^{N,D}$, rank $M \ll D$, GLM mapping ϕ, MCMC transition kernel $q(\cdot,\cdot)$, number of MCMC iterations T. Time and memory complexities that are not included depend on the specific choice of MCMC transition kernel.*

2: **Pseudo-Code**	3: **Time Complexity**	4: **Memory Complexity**
5: *Data preprocessing — M-Truncated SVD*		
6: $U, \text{diag}(\lambda), V := \text{truncated-SVD}(X^\top, M)$	$O(NDM)$	$O(NM + DM)$
7: $X_U = XU$	$O(NM)$	$O(NDM)$
8: *Propose $\beta^{(t)} \in \mathbb{R}^D$, compute likelihood*		
9: $\beta^{(t)} \sim q(\beta^{(t)}, \beta^{(t-1)})$	—	—
10: $\mathcal{L}_t := \sum_{i=1}^N \phi(y_i, x_i^\top U U^\top \beta^{(t)}) + \log p(\beta^{(t)})$	$O(1)$	$O(NM + MD)$
11: *Accept or Reject*		
12: Acceptance probability $p_A := \min\left(1, \frac{\mathcal{L}_t}{\mathcal{L}_{t-1}}\right)$	$O(1)$	$O(1)$
13: Accept $\beta^{(t)}$ with probability p_A	$O(1)$	$O(1)$
14: *Repeat steps 3-6 for T iterations*		

The transition in Line 9 may additionally benefit from the LR-GLM approximation. In particular, widely used algorithms such as Hamiltonian Monte Carlo and the No-U-Turn Sampler rely on many $O(ND)$-time likelihood and gradient evaluations, the cost of which can be reduced to $O(NM + DM)$ with LR-GLM. An implementation of this approximation is given in the `Stan` model in Appendix B.1.3 with performance results in Figures B.1.3 and B.1.6.

B.8 LR-Laplace with non-Gaussian priors

As discussed in the main text, we can maintain computational advantages of LR-GLM even when we have non-Gaussian priors. This admits the procedure provided in Algorithm 7.

In order for this more general LR-Laplace algorithm to be computationally efficient, we still require that the prior have some properties which can accommodate efficiency. In particular Line 11 demands that the Hessian of the prior is computed and inverted,

[2]To keep notation concise we use $\vec{\phi}''_{\hat{\mu}}$ to denote $\vec{\phi}''(Y, XUU^\top\hat{\mu})$

Algorithm 7 LR-Laplace for Bayesian inference in GLMs with low-rank data approximations and twice differentiable prior. Time and memory complexities which are not included depend on the choice of prior and optimisation method, which can be problem specific.

1:	***Input:*** *twice differentiable prior $p(\beta)$, data $X \in \mathbb{R}^{N,D}$, rank $M \ll D$, GLM mapping ϕ with ϕ'' (see Eq.* (3.5) *and Section 3.5.1)*		
2:	**Pseudo-Code**	3: **Time Complexity**	4: **Memory Complexity**
5:	*Data preprocessing — M-Truncated SVD*		
6:	$U, \text{diag}(\lambda), V := \text{truncated-SVD}(X^T, M)$	$O(NDM)$	$O(NM + DM)$
7:	*Optimize to find approximate MAP estimate (in D-dimensional space)*		
8:	$\hat{\mu} := \arg\max_{\mu \in \mathbb{R}^D} \sum_{i=1}^{N} \phi(y_i, x_i U U^\top \mu) + \log p(\beta = \mu)$	—	—
9:	*Compute approximate posterior covariance*[2]		
10:	$\hat{\Sigma}^{-1} := -\nabla^2_\beta \log p_\beta(\hat{\mu}) - UU^\top X^\top \text{diag}(\vec{\phi}''_{\hat{\mu}}) XUU^\top$	—	—
11:	$K := [\nabla^2_\beta \log p(\beta)|_{\beta=\hat{\mu}}]^{-1}$	—	—
12:	$\hat{\Sigma} := -K + KU\big([U^\top X^\top \text{diag}(\vec{\phi}''_{\hat{\mu}}) XU]^{-1} + U^\top KU\big)^{-1} U^\top K$	—	—
13:	*Compute variances and covariances of parameters*		
14:	$\text{Var}_{\hat{p}}(\beta_i) = e_i^\top \hat{\Sigma} e_i$	—	—
15:	$\text{Cov}_{\hat{p}}(\beta_i, \beta_j) = e_i^\top \hat{\Sigma} e_j$	—	—

as will true even in the high-dimensional setting when, for example, the prior factorizes across dimensions. Additionally, properties of the prior such as log concavity will facilitate efficient optimisation in Line 8.

Appendix C

Coupling Supplementary Materials

C.1 Proof of Gibbs Sweep Time Complexity

We here detail our $O(N\widetilde{K}^3 \log \widetilde{K})$ implementation of Algorithm 5. This serves as proof of Theorem 4.2.1.

Note that work in Algorithm 5 may be separated into 2 computationally demanding stages for each of the N data-points, $n \in [N]$; computing the distances between each pair of partitions in the Cartesian product of supports of the Gibbs conditionals $p_{\Pi|\Pi_{-n}}(\cdot|\pi_{-n})$ and $p_{\Pi|\Pi_{-n}}(\cdot|\nu_{-n})$ and solving the optimal transport problem in line 7. As discussed in Remark 4.2.2, the optimal transport problem may be solved in $O(K^3 \log K)$ time, and is the bottleneck step. As such it remains only to show that for each $n \in [N]$, the pairwise distances may also be computed in $O(K^3 \log K)$ time.

Recall that for two partitions $\pi, \nu \in \mathcal{P}_N$ the metric of interest is

$$\mathrm{d}(\pi, \nu) = \sum_{A \in \pi} |A|^2 + \sum_{B \in \nu} |B|^2 - 2 \sum_{(A,B) \in \pi \times \nu} |A \cap B|^2. \tag{C.1.1}$$

However, it is not obvious from this expression alone that fast computation of pairwise distances should be possible. We make this explicit in the following remark.

Remark C.1.1. Given constant $O(1)$ time for querying set membership (e.g. as provided by a standard hash-table set implementation), for $\pi, \nu \in \mathcal{P}_N$, $d(\pi, \nu)$ in Eq. (4.2) may be computed in $O\left(N \min(|\pi|, |\nu|)\right)$ time. If we let $\widetilde{K}$ be the number of groups, so

that $\widetilde{K} \approx \min(|\pi|, |\nu|)$, this gives $O(N\widetilde{K})$ time.

While this is certainly faster than a naive approach relying on the formulation of this metric based on adjacency matrices, it is still not sufficient, as it is a factor of $N\widetilde{K}^2$ slower than the original (recall that we will need to do this for $\widetilde{K}^2$ pairs of clusters assignments).

However we can do better for the Gibbs update by making two observations. First, if we use A_n and B_n to denote the elements of π and ν, respectively, containing data-point n, then for any n we may write

$$d(\pi, \nu) = d(\pi_{-n}, \nu_{-n}) + \left[|A_n|^2 - (|A_n| - n)^2\right] + \left[|B_n|^2 - (|B_n| - n)^2\right] \tag{C.1.2}$$

$$-2\left[|A_n \cap B_n|^2 - (|A_n \cap B_n| - 1)^2\right] \tag{C.1.3}$$

$$= d(\pi_{-n}, \nu_{-n}) + 2\left[|A_n| + |B_n| - 2|A_n \cap B_n|\right]. \tag{C.1.4}$$

Second, the solution to the optimisation problem in Eq. (4.3) is unchanged when we add a constant value to every distance: Using again the notation of Algorithm 5 we let $q := p_{\Pi|\Pi_{-n}}(\cdot|\pi_{-n})$ and $r := p_{\Pi|\Pi_{-n}}(\cdot|\nu_{-n})$ with supports $(\pi_1, \pi_2, \ldots, \pi_K) = \text{support}(q)$ and $(\nu_1, \nu_2, \ldots, \nu_{K'}) = \text{support}(r)$. and rewrite

$$\gamma^* := \arg\min_{\gamma \in \Gamma(q,r)} \sum_{\in \mathcal{P}_N} \sum_{y \in \mathcal{P}_N} d(x, y)\gamma(x, y) \tag{C.1.5}$$

$$= \arg\min_{\gamma \in \Gamma(q,r)} \sum_{x \in \mathcal{P}_N} \sum_{y \in \mathcal{P}_N} (d(x, y) - c)\gamma(x, y) \tag{C.1.6}$$

for any constant c; taking $c = d(\pi_{-n}, \nu_{-n})$ reveals that we need only compute the second term in Eq. (C.1.2).

At first it may seem that this still does not solve the problem, as directly computing the size of the set intersections is $O(N)$ (if cluster sizes scale as $O(N)$). However, Eq. (C.1.5) is just our final stepping stone. If we additionally keep track of sizes of intersections at every step, updating them as we adapt the partitions will take constant time for each update. As such, we are able to form the matrix of pairwise distances in $O(\widetilde{K}^2)$ time. Regardless of N, this moves the bottleneck step to solving the OT

problem which, as discussed in Remark 4.2.2, may be computed in $O(\widetilde{K}^3 \log \widetilde{K})$ time with Orlin's algorithm [Orlin, 1993]. We provide a practical implementation of this approach in our code; see `pairwise_dists()` in `modules/utils.py`.

C.2 Additional Experimental Details

C.2.1 Meeting time distributions

DP mixtures. For each replicate, we simulated $N = 150$ data-points from a $K = 4$ component, 2 dimensional Gaussian mixture model. The target distribution was the posterior of the probabilistic model Eq. (4.4), with $\Sigma_0 = 2.5I_2$, $\Sigma_1 = 2I_2$ and $\alpha = 0.2$. For each replicate true means for the finite mixture were sampled as $\mu_k \sim \mathcal{N}(0, \Sigma_0)$, mixing proportions as $\theta \sim \mathrm{Dir}(\alpha 1_K)$, and each of the $n \in [N]$ observations as $z_n \sim \mathrm{Cat}(\theta)$, $\mathcal{D}_n \sim \mathcal{N}(\mu_{z_n}, \Sigma_1)$. See `notebooks/Coupled_CRP_sampler.ipynb` for complete implementation and details. This code is adapted from github.com/tbroderick/mlss2015_bnp_tutorial/blob/master/ex5_dpmm.R

Graph coloring Let G be an undirected graph with vertices $V = [N]$ and edges $E \subset V \otimes V$, and let $Q = [q]$ be set of q colors. A graph coloring is an assignment of a color in Q to each vertex satisfying that the endpoints of each edge have different colors. We here demonstrate an application of our method to a Gibbs sampler which explores the uniform distribution over valid $q-$colorings of G, i.e. the distribution which places equal mass on ever proper coloring of G.

To employ Algorithm 5, for this problem we need only to characterise the PMF on partitions of the vertices implied by the uniform distribution on its colorings.

A partition corresponds to a proper coloring only if no two adjacent vertices are in the element of the partition. As such, we can write

$$p_{\Pi_N}(\pi) \propto \mathbb{1}\{|\pi| \leq q \text{ and } A(\pi)_{i,j} = 1 \to (i,j) \notin E,\ \forall i \neq j\} \binom{q}{|\pi|} |\pi|!,$$

where the indicator term checks that π can correspond to a proper coloring and the

second term accounts for the number of unique colorings which induce the partition π. In particular it is the product of the number of ways to choose $|\pi|$ unique colors from Q ($\binom{q}{|\pi|} := \frac{q!}{|\pi|!(q-|\pi|)!}$) and the number of ways to assign those colors to the groups of vertices in π.

For the experiments in Figure 4-1, we simulated Erdős-Rényi random graphs with $N = 25$ vertices, and including each possible edge with probability 0.2. We chose a maximum number of colors Q by first initializing a coloring greedily and setting Q as the number of colors used in this initial coloring plus two. See `notebooks/coloring_OT.ipynb` for complete implementation and results. This code is adapted from: github.com/pierrejacob/couplingsmontecarlo/inst/chapter3/3graphcolourings.R

C.2.2 Unbiased estimation

Predictive density in Gaussian mixture data. The true density is a 10-component Gaussian mixture model with known observational noise variance $\sigma = 2.0$. The cluster proportions were generated from a symmetric Dirichlet distribution with mass 1 for all 10-coordinates. The cluster means were randomly generated from $\mathcal{N}(0, 10^2)$. The target DP mixture model had $\alpha = 1$, standard deviation over cluster means 3.0 and standard deviation over observations 2.0. The function of interest is the posterior predictive density

$$\Pr(\mathcal{D}_{N+1} \in dx \,|\, \mathcal{D}_{1:N}) = \sum_{\Pi_{N+1}} \Pr(\mathcal{D}_{N+1} \in dx \,|\, \Pi_{N+1}, \mathcal{D}_{1:N}) \Pr(\Pi_{N+1} \,|\, \mathcal{D}_{1:N}). \quad \text{(C.2.1)}$$

In Eq. (C.2.1), Π_{N+1} denotes the partition of the data $\mathcal{D}_{1:(N+1)}$. To translate Eq. (C.2.1) into an integral over just the posterior over Π_N, the partition of $\mathcal{D}_{1:N}$, we break up Π_{N+1} into (Π_N, Z) where Z is the cluster indicator specifying the cluster of Π_N (or a new cluster) to which $\mathcal{D}_{N+1}$ belongs. Then

$$\Pr(\mathcal{D}_{N+1} \in dx \,|\, \mathcal{D}_{1:N}) = \sum_{\Pi_N} \left[\sum_{Z} \Pr(\mathcal{D}_{N+1} \in dx, Z \,|\, \Pi_N, \mathcal{D}_{1:N}) \right] \Pr(\Pi_N \,|\, \mathcal{D}_{1:N})$$

Each $\Pr(\mathcal{D}_{N+1} \in dx, Z \,|\, \Pi_N, \mathcal{D}_{1:N})$ is computed using the prediction rule for the CRP and Gaussian conditioning. Namely

$$\Pr(\mathcal{D}_{N+1} \in dx, Z \,|\, \Pi_N, \mathcal{D}_{1:N}) = \underbrace{\Pr(\mathcal{D}_{N+1} \in dx \,|\, Z, \Pi_N, \mathcal{D}_{1:N})}_{\text{Posterior predictive of Gaussian}} \times \underbrace{\Pr(Z \,|\, \Pi_N)}_{\text{CRP prediction rule}} .$$

The first term is computed with the function used during Gibbs sampling to reassign data points to clusters. In the second term, we ignore the conditioning on $\mathcal{D}_{1:N}$, since Z and $\mathcal{D}_{1:N}$ are conditionally independent given Π_N.

We ran 10,000 replicates of the time-budgeted estimator using coupled chains, each replicate given a sufficient time budget so that all 10,000 replicates had at least one successful meeting in the allotted time.

Top component proportion in single-cell RNAseq. We extracted $D = 50$ genes with the most variation of $N = 200$ cells. We then take the log of the features, and normalize so that each feature has mean 0 and variance 1. We as our target the posterior of the probabilistic model in Eq. (4.4) with $\alpha = 1.0$, $\mu_0 = 0$, $\Sigma_0 = 0.5$, $\Sigma_1 = 1.3 I_D$. Notably, this is a simplification of the set-up considered by Prabhakaran et al. [2016], who work with a larger dataset and additionally perform fully Bayesian inference over these hyper-parameters. In our experiments, the function of interest is the posterior expected of the proportion of cells in the largest cluster i.e. $\mathbb{E}[\max_{|A| \in \pi} |A|/N \,|\, \mathcal{D}]$.

C.3 More plots of predictive density

C.3.1 Posterior concentration implies convergence in total variation of predictive density

Some references on posterior concentration are Ghosal et al. [1999], Lijoi et al. [2005]. The true data generating process is that there exists some density f_0 w.r.t. Lebesgue measure that generates the data in an iid manner $X_1, X_2, \ldots, X_n$. We use the notation P_{f_0} to denote the probability measure with density f_0. The probabilistic model is that

we have a prior Π over densities f, and observations X_i are conditionally iid given f. Let $\mathcal{F}$ be the set of all densities on $\mathbb{R}$. For any measurable subset A of $\mathcal{F}$, the posterior of A given the observations X_i is denoted $\Pi(A|X_{1:N})$. A strong neighborhood around f_0 is any subset of $\mathcal{F}$ containing a set of the form $V = \{f \in \mathcal{F} : \int |f - f_0| < \epsilon\}$ according to Ghosal et al. [1999]. The prior Π is strongly consistent at f_0 if for any strong neighborhood U,

$$\lim_{n\to\infty} \Pi(U|X_{1:n}) = 1, \tag{C.3.1}$$

holds almost surely for $X_{1:\infty}$ distributed according to $P_{f_0}^{\infty}$.

Theorem C.3.1 (Ghosh and Ramamoorthi [2003, Proposition 4.2.1])**.** *If a prior* Π *is strongly consistent at* f_0 *then the predictive distribution, defined as*

$$\widehat{P}_n(A \mid X_{1:n}) \coloneqq \int_f P_f(A)\Pi(f \mid X_{1:n}) \tag{C.3.2}$$

also converges to f_0 *in total variation in a.s.* $P_{f_0}^{\infty}$

$$d_{TV}\left(\widehat{P}_n, P_{f_0}\right) \to 0.$$

The definition of posterior predictive density in Eq. (C.3.2) can equivalently be rewritten as

$$\widehat{P}_n(A \mid X_{1:n}) = \Pr(X_{n+1} \in A \,|\, X_{1:n}),$$

since $P_f(A) = P_f(X_{n+1} \in A)$ and all the X's are conditionally iid given f.

Theorem C.3.2 (DP mixtures prior is consistent for finite mixture models)**.** *Let the true density be a finite mixture model* $f_0(x) \coloneqq \sum_{i=1}^{m} p_i \mathcal{N}(x|\theta_i, \sigma_1^2)$*. Consider the following probabilistic model*

$$\begin{aligned}
\widehat{P} &\sim \mathrm{DP}(\alpha, \mathcal{N}(0, \sigma_0^2)) \\
\theta_i \mid \widehat{P} &\overset{iid}{\sim} \widehat{P} && i = 1, 2, \ldots, n \\
X_i \mid \theta_i &\overset{indep}{\sim} \mathcal{N}(\theta_i, \sigma_1^2) && i = 1, 2, \ldots, n
\end{aligned}$$

Let $\widehat{P}_n$ be the posterior predictive distribution of this generative process. Then with a.s. P_{f_0}

$$d_{TV}\left(\widehat{P}_n, P_{f_0}\right) \xrightarrow{n\to\infty} 0.$$

Proof of Theorem C.3.2. First, we can rewrite the DP mixture model as a generative model over continuous densities f

$$\begin{aligned} \widehat{P} &\sim \mathrm{DP}(\alpha, \mathcal{N}(0, \sigma_0^2)) \\ f &= \mathcal{N}(0, \sigma_1^2) * \widehat{P} \\ X_i \mid f &\overset{iid}{\sim} f \qquad\qquad i = 1, 2, \ldots, n \end{aligned} \tag{C.3.3}$$

where $\mathcal{N}(0, \sigma_1^2) * \widehat{P}$ is a convolution, with density $f(x) := \int_\theta \mathcal{N}(x - \theta | 0, \sigma_1^2) d\widehat{P}(\theta)$.

The main idea is showing that the posterior $\Pi(f|X_{1:n})$ is strongly consistent and then leveraging Theorem C.3.1. For the former, we verify the conditions of Lijoi et al. [2005, Theorem 1].

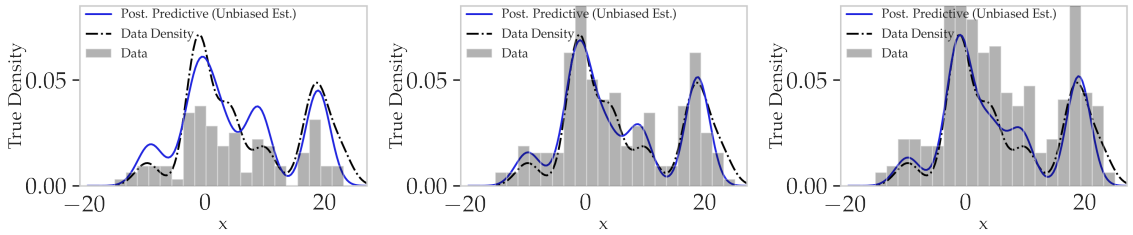
The first condition of Lijoi et al. [2005, Theorem 1] is that f_0 is in the K-L support of the prior over f in Eq. (C.3.3). We use Ghosal et al. [1999, Theorem 3]. Clearly f_0 is the convolution of the normal density $\mathcal{N}(0, \sigma_1^2)$ with the distribution $P(.) = \sum_{i=1}^m p_i \delta_{\theta_i}$. $P(.)$ is compactly supported since m is finite. Since the support of $P(.)$ is the set $\{\theta_i\}_{i=1}^m$ which belongs in $\mathbb{R}$, the support of $\mathcal{N}(0, \sigma_0^2)$, by Ghosh and Ramamoorthi [2003, Theorem 3.2.4], the conditions on P are satisfied. The condition that the prior over bandwidths cover the true bandwidth is trivially satisfied since we perfectly specified σ_1.

The second condition of Lijoi et al. [2005, Theorem 1] is simple: because the prior over $\widehat{P}$ is a DP, it reduces to checking that

$$\int_{\mathbb{R}} |\theta| \mathcal{N}(\theta \mid 0, \sigma_0^2) < \infty$$

which is true.

The final condition trivial holds because we have perfectly specified σ_1: there is actually zero probability that σ_1 becomes too small, and we never need to worry about



(a) $N = 100$. (b) $N = 200$. (c) $N = 300$.

Figure C.3.1: Posterior predictive density for different N. The time budget for each replicate when $N = 100, 200, 300$ is respectively $100, 300, 800$ seconds. We average the results from 400 replicates.

setting γ or the sequence σ_k. □

C.3.2 Predictive density plots for varying N

In Figure C.3.1, the distance between the posterior predictive density and the underlying density decreases as N increases. We sampled a grid $\{u_j\}$ of 150 evenly-spaced points in the domain $[-20, 30]$, and evaluated both the true density and the posterior predictive density on this grid. The distance in question sums over the absolute differences between the evaluations over the grid

$$\text{dist} := \sum_j |f_N(u_j) - f_0(u_j)|.$$

where $f_N(u_j)$ is the posterior predictive density of the N observations under the DPMM at u_j. The distance is meant to illustrate *pointwise* rather than total variation convergence. Although the predictive density converges in total variation to the underlying density, it is only guaranteed that a subsequence of the predictive density converges pointwise to the underlying density.

In Figure C.3.1, each N has a different time budget because for larger N, in general per-sweep time increases and number of sweeps until coupled chains meet also increase.

Appendix D

C-value Supplementary Material

D.1 Appendix

Proof of Theorem 5.2.2

Proof. The result follows directly from the definition of $c(y)$ and the conditions on $b(\cdot, \cdot)$. More explicitly,

$$\begin{aligned}\mathbb{P}_\theta\left[W(\theta, y) \leq 0 \text{ and } c(y) > \alpha\right] &\leq \mathbb{P}_\theta\left[W(\theta, y) \leq 0 \text{ and } b(y, \alpha) > 0\right] \\ &\leq \mathbb{P}_\theta\left[W(\theta, y) < b(y, \alpha)\right] \\ &\leq 1 - \alpha,\end{aligned}$$

where the first line follows from the definition of the c-value and the final line follows from Eq. (5.1). □

Proof of Theorem 5.2.3

Proof. The condition $L(\theta, \theta^\dagger(y, \alpha)) > L(\theta, \hat{\theta}(y))$ can occur only when both (A) $0 > W(\theta, y)$ and (B) $\theta^\dagger(\cdot, \alpha)$ evaluates to $\theta^*(\cdot)$ rather than $\hat{\theta}(\cdot)$. Event (B) implies $c(y) > \alpha$ and therefore $b(y, \alpha) > 0$. By transitivity, $b(y, \alpha) > 0$ and $0 > W(\theta, y) \implies b(y, \alpha) > W(\theta, y)$. By assumption, the event $b(y, \alpha) > W(\theta, y)$ occurs with probability at most $1 - \alpha$. □

D.2 Pitfalls of risk when choosing between estimators

Before proceeding, we require some additional notation and definitions. We denote the risk of an arbitrary estimator $\theta'(\cdot)$ by $R(\theta, \theta') = \mathbb{E}_\theta \left[L\left(\theta, \theta'(y)\right) \right]$. Given two estimators $\theta'(\cdot)$ and $\theta^\dagger(\cdot)$ we say that $\theta'(\cdot)$ *dominates* $\theta^\dagger(\cdot)$ if, for all values of θ, $R(\theta, \theta') \leq R(\theta, \theta^\dagger)$ and $R(\theta, \theta') < R(\theta, \theta^\dagger)$ for at least one value of θ.

If we were able to show that one of $\hat{\theta}(\cdot)$ or $\theta^*(\cdot)$ dominates the other, it would be tempting to always select the dominating estimator. Unfortunately, it is very often the case that neither estimator dominates the other. In other words, it may be the case that $R(\theta, \theta^*) < R(\theta, \hat{\theta})$ for all values of θ in some non-trivial subset of the space Θ_0 but $R(\theta, \theta^*) > R(\theta, \hat{\theta})$ for some $\theta \notin \Theta_0$. Lindley and Smith [1972] provide a simple illustration of this dilemma in the following normal means problem: suppose that we observe an N-vector normally distributed about its mean and with identity covariance, I_N, as $y \sim \mathcal{N}(\theta, I_N)$, and wish to compare the default estimate $\hat{\theta}(y) = y$ of θ and the alternative estimate

$$\theta^*(y) = \frac{y + \overline{y}\mathbf{1}_N/\tau^2}{1 + 1/\tau^2}$$

for a fixed value of $\tau > 0$, where $\overline{y} := N^{-1} \sum_{n=1}^N y_n$ and $\mathbf{1}_N$ is the N-vector of ones. Lindley and Smith [1972] showed that $R(\theta, \theta^*) < R(\theta, \hat{\theta})$ if and only if

$$\|\theta - \overline{\theta}\mathbf{1}_N\|_2 < \sqrt{(N-1)(2+\tau^2)}, \tag{D.2.1}$$

where $\overline{\theta} := N^{-1} \sum_{n=1}^N \theta_n$. Without strong assumptions about the value of θ, which we may be unable or unwilling to make, a simple comparison of risk functions can prove inconclusive. Interestingly, in the setting considered by Lindley and Smith [1972], it is possible to construct θ so that (A) $R(\theta, \theta^*) < R(\theta, \hat{\theta})$ but (B) $\mathbb{P}_\theta[L(\theta, \theta^*(y)) > L(\theta, \hat{\theta}(y))] > 0.5$. In particular, for $N = 2, \tau = 1$, and $\|\theta - \overline{\theta}\mathbf{1}_N\|^2 = 2.999, \theta^*(\cdot)$ has slightly smaller risk than the MLE, but the MLE has smaller loss in 3397 out of 5000 simulated datasets, or about 68% of the time. In other words, even if we were to assume that θ satisfied Eq. (D.2.1), for the majority of datasets y that we might observe, the alternative estimator incurs higher loss than the default. The situation

above highlights an important, but in our mind under-discussed, limitation of risk: the loss averaged over all possible unrealized datasets may not be close to the loss incurred on an observed dataset.

This disagreement between risk and the probability of having smaller loss can be especially pronounced when the distribution of the loss of one of the estimators is heavy-tailed. For example, consider a scalar parameter $\theta = 0$, a deterministic default estimate $\hat{\theta} = 1$, and an alternative estimate distributed as $\theta^* \sim \frac{1}{\alpha}\delta_{\sqrt{\alpha(1+\epsilon)}} + (1 - \frac{1}{\alpha})\delta_0$, where δ_x denotes a Dirac mass on x and $\epsilon > 0$. Then $\theta^*(\cdot)$ has larger risk than $\hat{\theta}(\cdot)$ ($1 + \epsilon$ rather than 1), but has smaller loss with probability $1 - \frac{1}{\alpha}$. By taking $\alpha \to \infty$, we see that $\theta^*(\cdot)$ may have smaller loss than $\hat{\theta}(\cdot)$ with arbitrarily high probability. This example is particularly extreme; our intent is merely to illustrate that large disagreements could, at least in principal, arise in practical settings.

D.3 Defining c-values as a supremum vs. infimum

In this section we describe a pathological model and construction of a lower bound function for which the two possible definitions of the c-value described in Remark 5.2.1 lead to notably different behaviours.

Consider a variant of the normal means problem. Let $\theta \in \mathbb{R}$ be an unknown mean and observe

$$y := \begin{bmatrix} \theta + \epsilon \\ u \end{bmatrix},$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and $u \sim \mathcal{U}([0, 1])$ is a uniform random variable on $[0, 1]$. Note that u is ancillary to θ (i.e. its distribution does not depend on θ). We will construct a pathological $b(y, \alpha)$ that depends on y only through u and will therefore be ancillary to θ as well. We begin by constructing a countably infinite collection of independent uniform random variables from u, indexed by the rationals $\mathbb{Q}$, $S(u) := \{u_r\}_{r\in\mathbb{Q}}$. Such a countably infinite collection may be obtained by segmenting the decimal expansion of u; for example, if we let d_i denote the i^{th} digit of u, we could obtain this sequence

by defining uniform random variables with decimal expansions

$$\begin{aligned} u^1 &:= [d_1, d_2, d_4, d_7, d_{11} \dots], \\ u^2 &:= [d_3, d_5, d_8, d_{12} \dots], \\ u^3 &:= [d_6, d_9, d_{13} \dots], \\ u^4 &:= [d_{10}, d_{14}, \dots], \\ u^5 &:= [d_{15}, \dots], \end{aligned}$$

and so on, and then mapping from $\{u^i\}_{i \in \mathbb{N}}$ to $S(u)$.

Next, define

$$b(y, \alpha) := \begin{cases} (-1)^{\mathbb{1}[u_\alpha < \alpha]} \infty & \text{if } \alpha \in \mathbb{Q} \\ -\infty & \text{otherwise.} \end{cases}$$

For any bounded default and alternative estimators, the win will be finite and the bound $b(y, \alpha)$ holds if an only if it evaluates to $-\infty$. Because $b(y, \alpha) = -\infty$ with probability at least α, even though $b(y, \alpha)$ is ancillary to θ, it still satisfies the condition in Eq. (5.1) for every θ and $\alpha \in [0, 1]$. However, consider two possible definitions of the c-value,

$$c^+(y) := \sup_{\alpha \in [0,1]} \{\alpha | b(y, \alpha) \geq 0\} \text{ vs. } c^-(y) := \inf_{\alpha \in [0,1]} \{\alpha | b(y, \alpha) \leq 0\},$$

where $c^-(y) = c(y)$ is the definition we have chosen in Section 5.2. Note that $c^-(y) \leq c^+(y)$, and that if $b(y, \alpha)$ is continuous and strictly decreasing in α for every y, then $c^-(y) = c^+(y)$. In this almost surely discontinuous case, however, we have that $c^+(y) \overset{a.s.}{=} 1.$ and $c^-(y) \overset{a.s.}{=} 0$. Since estimators exist for which $W(\theta, y) < 0$ with positive probability, the guarantees of Theorems 5.2.2 and 5.2.3 are not met by $c^+(y)$.

In the present paper, $c^-(y) = c^+(y)$ for all bounds considered. Our preference for defining the c-value as $c^-(y)$ derives from simplicity; we may disregard edge cases like the one above, which would complicate our proofs. However for the reason described

in this section, we emphasize that using $c^-(y)$ rather than $c^+(y)$ may have practical implications when these quantities differ.

D.4 Additional related work

Bayesian model checking. Working in a Bayesian context, Box and Hunter [1962] and Box [1980] advocate a dynamic, iterative approach to modeling and inference ("Box's loop" [Blei, 2014]), in which models for an observed dataset are successively proposed, evaluated, and refined. Specifically, one first assesses the suitability of a given Bayesian model with the prior predictive probability of an observed statistic. A sufficiently small *prior predictive p-value* indicates that the observed data is an atypical realization from the proposed model; in this case, one then refines and re-assesses the model. Through this lens, our proposed estimation procedure resembles a single iteration through Box's loop: we discard the default estimator in favor of an alternative if the c-value is sufficiently extreme. However, prior predictive p-values are limited by the requirement for a "good" choice of prior — namely a prior that provides an adequate description of the data. Indeed, one might lack such a prior, but the associated Bayes estimate could still be superior to a given alternative. For example, our approach can choose a Bayes rule with an improper prior over a default estimator, while the prior predictive p-value is undefined. Or our approach can also be used to choose an estimator that is not Bayesian in origin. We emphasize as well that our contribution in the present work is a direct quantification of confidence in the relative performance of estimates that is absent from earlier work.

Non-asymptotic frequentist guarantees for "Bayesian" procedures. As we will demonstrate, our procedure allows a practitioner to benefit from a Bayesian model while still providing frequentist guarantees that do not depend on validity of a Bayesian prior. This flavor is not unique to our proposal.

Most notably, empirical Bayesian methods [Morris, 1983] avoid dependence on certain subjective choices in the specification of the prior by selecting a prior from

the same data used to estimate the parameter. While many frequentist properties of these estimators are typically unavailable due to the difficulty of analysis, in some cases authors have established favorable properties. For example, Efron and Morris [1973] famously show that an empirical Bayesian estimator dominates the maximum likelihood estimate.

Outside of decision theory, Bayesian models have played a role in the construction of smaller confidence intervals with exact frequentist coverage, initially by Pratt [1963] and more recently in the context of empirical Bayes inference by Hoff and Yu [2019] and post-selection inference by Woody et al. [2020]. In the context of hypothesis testing, Hoff [2021] leveraged a similar approach to increase power across multiple hypotheses while maintaining exact coverage. The objectives of these papers are distinct from our own; we do not consider hypothesis testing or forming confidence intervals. But these works are thematically related nonetheless; in particular, their methods can utilize a Bayesian model to provide improved statistical procedures that maintain frequentist guarantees, and they do not rely on subjective assumptions about unknown parameters.

D.5 Additional details related to Section 5.3

D.5.1 Distribution of win term

We here provide a derivation of the distributional form of $2\epsilon^\top Gy$ given in Section 5.3.2. In Section 5.3.2 we found that

$$2\epsilon^\top Gy \sim \frac{2}{1+\tau^2}\left[\chi^2_{N-1}(\frac{1}{4}\|P_1^\perp\theta\|^2) - \frac{1}{4}\|P_1^\perp\theta\|^2\right],$$

where $\chi^2_{N-1}(\lambda)$ denotes the non-central chi-squared distribution with $N-1$ degrees of freedom and non-centrality parameter λ.

Recall that $Gy = (1+\tau^2)^{-1}P_1^\perp(\theta+\epsilon)$. As such we can rewrite

$$2\epsilon^\top Gy = \frac{2}{1+\tau^2}\left[\epsilon^\top P_1^\perp\epsilon + \epsilon^\top P_1^\perp\theta\right]$$

$$
\begin{aligned}
&= \frac{2}{1+\tau^2}\left[(P_1^\perp \epsilon)^\top (P_1^\perp \epsilon) + (P_1^\perp \epsilon)^\top (P_1^\perp \theta)\right] \\
&// \text{ since } P_1^\perp = P_1^\perp P_1^\perp \\
&= \frac{2}{1+\tau^2}\left[\|P_1^\perp \epsilon + \frac{1}{2} P_1^\perp \theta\|^2 - \frac{1}{4}\|P_1^\perp \theta\|^2\right] \\
&// \text{ by completing the square} \\
&= \frac{2}{1+\tau^2}\left[\chi^2_{N-1}(\frac{1}{4}\|P_1^\perp \theta\|^2) - \frac{1}{4}\|P_1^\perp \theta\|^2\right],
\end{aligned}
$$

as desired, where in the last line the degrees of freedom parameter is $N-1$ because $P_1^\perp$ projects into an $N-1$ dimensional subspace of $\mathbb{R}^N$.

D.5.2 Proof of Theorem 5.3.1

We here provide a proof of Theorem 5.3.1.

Proof. The proof amounts to showing that $b(\cdot,\cdot)$ achieves at least nominal coverage, i.e. for any θ and $\alpha \in [0,1]$, $\mathbb{P}\left[W(y,\theta) \geq b(y,\alpha)\right] \geq \alpha$. By construction, $W(\theta,y) \geq b(y,\alpha)$ may be violated only if either (A) $\|P_1^\perp \theta\|^2 \notin [0, U(y, \frac{1-\alpha}{2})]$ or (B) $W(\theta,y) < \frac{2}{1+\tau^2} F^{-1}_{N-1}(\frac{1-\alpha}{2}; \frac{\|P_1^\perp \theta\|^2}{4}) - \frac{\|P_1^\perp \theta\|^2}{2(1+\tau^2)} - \frac{\|P_1^\perp y\|^2}{(1+\tau^2)^2}$. Noticing that $\|P_1^\perp y\|^2 \sim \chi^2_{N-1}(\|P_1^\perp \theta\|^2)$, we can recognize $[0, U(\frac{1-\alpha}{2})]$ as valid confidence interval for $\|P_1^\perp \theta\|^2$ and see that (A) occurs with probability at most $\frac{1-\alpha}{2}$. Next, comparing to Eq. (5.7), we see that (B) represents $2\epsilon^\top G y$ falling below its $\frac{1-\alpha}{2}$ quantile and thus occurs with probability at most $\frac{1-\alpha}{2}$. Therefore the union bound guarantees that $b(y,\alpha)$ obtains at least nominal coverage. □

D.5.3 Why an *upper* bound on $\|P_1^\perp \theta\|^2$?

We here provide justification for the use of a high-confidence upper bound on $\|P_1^\perp \theta\|^2$ in Bound 5.3.1. Recall that Eq. (5.7) provides a lower bound on $W(\theta,y)$ if we can control $\|P_1^\perp \theta\|^2$. However, it is not immediately obvious what sort of control on $\|P_1^\perp \theta\|^2$ will yield the tightest bound; should we have derived a two-sided interval or a lower bound instead of an upper bound? We answer this question by appealing to a normal approximation of the non-central χ^2 for intuition. This approximation will be

close when the degrees of freedom parameter is large. Specifically, by replacing the non-central χ^2 quantile with that of a normal with matched first and second moments we may approximate the lower bound as

$$W(\theta, y) \overset{\sim}{\geq} \frac{2}{1+\tau^2}\left[N - 1 - (\|P_1^\perp \theta\|^2 + 2N - 2)^{\frac{1}{2}} z_\alpha\right] - \frac{\|P_1^\perp y\|^2}{(1+\tau^2)^2}, \tag{D.5.1}$$

where z_α is the α quantile of the standard normal.

Eq. (D.5.1) is monotone decreasing in $\|P_1^\perp \theta\|^2$ for any $\alpha > \frac{1}{2}$. As such, we can expect this quantile to be smallest for large values of $\|P_1^\perp \theta\|^2$, and for this reason seek to find a high-confidence upper bound on $\|P_1^\perp \theta\|^2$. Indeed, in agreement with Eq. (D.5.1) we have found empirically that the infimum in Eq. (5.8) is always achieved at this upper bound, and conjecture that this is true in general.

D.5.4 Shrinking towards an arbitrary subspace

We now show how the approach developed in Section 5.3 immediately extends to a broader class of models in the spirit of those considered by Morris [1983]. In particular, let θ again be an unknown N-vector and $X \in \mathbb{R}^{N \times D}$ be a design matrix where for each n, X_n is a D-vector of covariates associated with θ_n. If we believe that the parameters can be roughly described as scattered around a linear function of these covariates with variance τ^2, we might consider trying to improve our estimates by estimating the linear dependence and interpolating between the sample estimate and the associated linear approximation. Following Morris [1983], we obtain this type of shrinkage with the estimate

$$\theta^*(y) := \frac{y + \tau^{-2} X (X^\top X)^{-1} X^\top y}{1 + \tau^{-2}},$$

which is the posterior mean of the Bayesian model that assumes for each n, $\theta_n \sim \mathcal{N}(X_n^\top \beta, \tau^2)$ a priori. Here β is an unknown D-vector of coefficients that is given an improper uniform prior.

For this setting, we propose the following bound.

Bound D.5.1 (Normal Means: Flexible shrinkage estimate vs. MLE). Observe $y = \theta + \epsilon$

with $\epsilon \sim \mathcal{N}(0, I_N)$ and consider estimates

$$\hat{\theta}(y) = y \quad \text{and} \quad \theta^*(y) := \frac{y + \tau^{-2} X(X^\top X)^{-1} X^\top y}{1 + \tau^{-2}},$$

where τ is a scalar and X is an N by D matrix of covariates. We propose

$$b(y, \alpha) = \inf_{\lambda \in [0, U(y, \frac{1-\alpha}{2})]} \frac{2}{1+\tau^2} F_{N-D}^{-1}\left(\frac{1-\alpha}{2}, \frac{\lambda}{4}\right) - \frac{\lambda}{2(1+\tau^2)} - \frac{\|P_X^\perp y\|^2}{(1+\tau^2)^2} \tag{D.5.2}$$

as a high-probability lower bound on the win. In this expression, $F_{N-D}^{-1}(1-\alpha, \lambda)$ denotes the inverse cumulative distribution function of the non-central χ^2 with $N-D$ degrees of freedom and non-centrality parameter λ evaluated at $1-\alpha$. $P_X^\perp := I_N - X(X^\top X)^{-1} X^\top$ is the projection onto the subspace orthogonal to the column-space of X.

$$U(y, 1-\alpha) := \inf_{\delta > 0} \left\{ \delta \middle| \|P_X^\perp y\|^2 \leq F_{N-D}^{-1}(1-\alpha, \delta) \right\} \tag{D.5.3}$$

is a high-confidence upper bound on $\|P_X^\perp \theta\|^2$.

This bound is identical to Bound 5.3.1 except that it projects to a different subspace, and loses D degrees of freedom in the χ^2 random variables, rather than 1. Indeed, this is a strict generalization, as we obtain our earlier example when we take $X = \mathbf{1}_N$. Bound D.5.1 is also computable (for the same reasons discussed in Remark 5.3.2) and valid, as we see in the next proposition.

Proposition D.5.1. *Eq.* (D.5.2) *in Bound D.5.1 satisfies the conditions of Theorem 5.2.2. In particular, for any* θ *and* $\alpha \in [0, 1]$, $\mathbb{P}_\theta \left[W(y, \theta) \geq b(y, \alpha) \right] \geq \alpha$.

Proof. Proposition D.5.1 follows from an argument very closely analogous to the proof of Theorem 5.3.1. We first rewrite $\theta^*(y)$ as $\theta^*(y) = y - Gy$ for $G := (1+\tau^2)^{-1} P_X^\perp$. Eq. (5.6) then holds exactly as before (i.e. $W(\theta, y) = 2\epsilon^\top Gy - \|Gy\|^2$). The two terms are treated as in Theorem 5.3.1; the only differences are that the norm under consideration is $\|P_X^\perp \theta\|$ rather than $\|P_1^\perp \theta\|$, and the change in degrees of freedom from $N-1$ to $N-D$. □

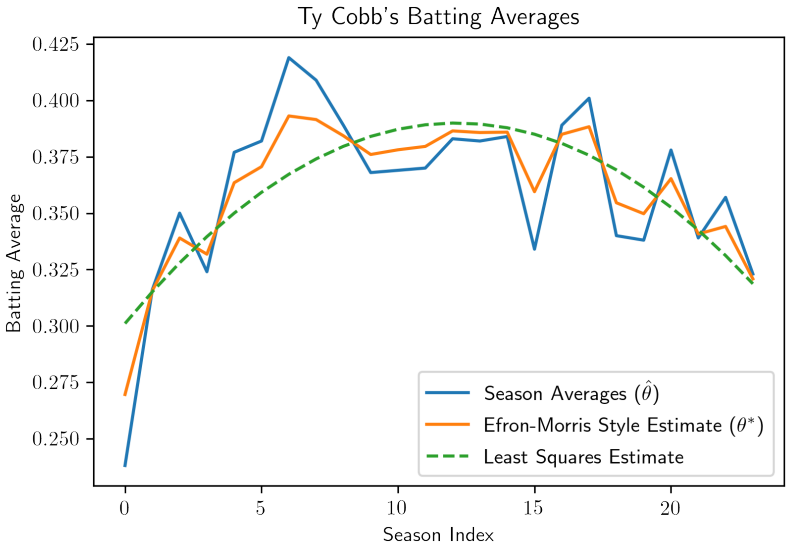


Figure D.5.1: The estimate shrinking towards a quadratic fit provides a significant improvement ($c = 0.953$). The noise and prior standard deviations were set as $\sigma = 0.025$ and $\tau = 0.025$, respectively.

Figure D.5.1 demonstrates an application to Ty Cobb's season batting averages, an example adapted from Morris [1983]. In this analysis, our approach indicates that we should be highly confident ($c = 0.953$) that the alternative estimate, which shrinks the observations towards a quadratic fit of the data, outperforms the MLE . While Morris [1983] provides an argument for estimators of this style based on risk, the present analysis goes a step further by providing a measure of confidence that the estimator improves on this particular dataset. Even though the risk of the estimator $\theta^*(\cdot)$ may be greater than that of $\hat{\theta}(\cdot)$ for many possible θ, this analysis supports the conclusion that for the true unknown θ and observed y, $\theta^*(y)$ is superior.

D.5.5 Distribution of c-values

As mentioned in Section 5.3.3, we do not in general expect to see a uniform distribution of c-values. Figure D.5.2 illustrates the dependence of the distribution of c-values on the parameter, in the same simulations detailed in Figures 5-1a, 5-1c and 5-1d.

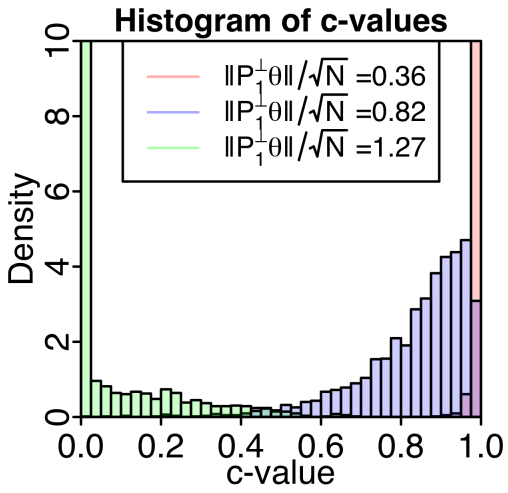


Figure D.5.2: Distribution of c-values across several choices of $N^{-\frac{1}{2}}\|P_1^{\perp}\theta\|$.

D.6 Affine estimators supplementary information

D.6.1 Step by step derivation of Eq. (5.12)

The win of using $\theta^*(y)$ in place of $\hat{\theta}(y)$ may be expressed as

$$\begin{aligned} W(\theta, y) &= \|\hat{\theta}(y) - \theta\|^2 - \|\theta^*(y) - \theta\|^2 \\ &= \left(\|\hat{\theta}(y)\|^2 + \|\theta\|^2 - 2\theta^\top \hat{\theta}(y)\right) - \left(\|\theta^*(y)\|^2 + \|\theta\|^2 - 2\theta^\top \theta^*(y)\right) \\ &= -2\theta^\top G(y) + \left(\|\hat{\theta}(y)\|^2 - \|\theta^*(y)\|^2\right) \\ &\text{// where } G(y) := \hat{\theta}(y) - \theta^*(y) \\ &= 2\epsilon^\top G(y) - 2y^\top G(y) + \left(\|\hat{\theta}(y)\|^2 - \|\theta^*(y)\|^2\right) \\ &= 2\epsilon^\top G(y) + \left(\|\hat{\theta}(y) - y\|^2 - \|\theta^*(y) - y\|^2\right). \end{aligned} \tag{D.6.1}$$

D.6.2 Derivation of Eq. (5.13)

Observe that

$$\begin{aligned} \mathbb{E}[\epsilon^\top G(y)] &= \mathbb{E}[\epsilon^\top G(\theta) + \epsilon^\top (A - C)\epsilon] \\ &= \mathbb{E}[\epsilon]^\top G(\theta) + \mathbb{E}[\mathrm{tr}[(A - C)\epsilon\epsilon^\top]] \\ &= \mathrm{tr}[(A - C)\Sigma] \end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[\epsilon^\top G(y)] &= \mathrm{Var}[\epsilon^\top G(\theta)] + \mathrm{Var}[\epsilon^\top (A-C)\epsilon] \\
&\quad // \text{ since } \epsilon^\top G(\theta) \text{ and } \epsilon^\top (A-C)\epsilon \text{ are uncorrelated} \\
&= (G(\theta))^\top \Sigma (G(\theta)) + 2\mathrm{tr}[\frac{A+A^\top-C-C^\top}{2}\Sigma\frac{A+A^\top-C-C^\top}{2}\Sigma] \\
&= \|G(\theta)\|_\Sigma^2 + \frac{1}{2}\mathrm{tr}[((A+A^\top-C-C^\top)\Sigma)^2] \\
&= \|G(\theta)\|_\Sigma^2 + \frac{1}{2}\|\Sigma^{\frac{1}{2}}(A+A^\top-C-C^\top)\Sigma^{\frac{1}{2}}\|_F^2,
\end{aligned}$$

where $\|\cdot\|_\Sigma$ and $\|\cdot\|_F$ denote the Σ quadratic norm and Frobenius norm, respectively. The third line of the derivation above obtains from recognizing $\mathrm{Var}[\epsilon^\top (A-C)\epsilon]$ as a quadratic form [Mathai and Provost, 1992, Chapter 2].

D.6.3 Derivations of Eqs. (5.15) and (5.16)

Eqs. (5.15) and (5.16) characterize the dependence of the distribution of $\|G(y)\|_\Sigma^2$ on $\|G(\theta)\|_\Sigma^2$ through its mean and variance. Recognizing $\|G(y)\|_\Sigma^2$ as a quadratic form [Mathai and Provost, 1992, Chapter 2], with $G(y) \sim \mathcal{N}\left(G(\theta), (A-C)\Sigma(A-C)^\top\right)$, we find its mean as

$$\begin{aligned}
\mathbb{E}[\|G(y)\|_\Sigma^2] &= G(\theta)^\top \Sigma G(\theta) + \mathrm{tr}[\Sigma((A-C)\Sigma(A-C)^\top)] \\
&= \|G(\theta)\|_\Sigma^2 + \mathrm{tr}[\Sigma^{\frac{1}{2}}(A-C)\Sigma(A-C)^\top\Sigma^{\frac{1}{2}}] \\
&= \|G(\theta)\|_\Sigma^2 + \|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_F^2.
\end{aligned}$$

For the variance, we similarly rely on the known variance of a quadratic form. Starting from that expression, we upper bound the variance as

$$\begin{aligned}\mathrm{Var}[\|G(y)\|_\Sigma^2] &= 2\mathrm{tr}\left[\Sigma\left((A-C)\Sigma(A-C)^\top\right)\Sigma\left((A-C)\Sigma(A-C)^\top\right)\right] + \\ &\quad 4G(\theta)^\top\Sigma\left((A-C)\Sigma(A-C)^\top\right)\Sigma G(\theta) \\ &= 2\|\Sigma^{\frac{1}{2}}(A-C)\Sigma(A-C)^\top\Sigma^{\frac{1}{2}}\|_F^2 + 4\|\left(\Sigma^{\frac{1}{2}}(A-C)^\top\Sigma^{\frac{1}{2}}\right)\Sigma^{\frac{1}{2}}G(\theta)\|_2^2 \\ &\leq 2\|\Sigma^{\frac{1}{2}}(A-C)\Sigma(A-C)^\top\Sigma^{\frac{1}{2}}\|_F^2 + 4\|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2\|G(\theta)\|_\Sigma^2,\end{aligned} \tag{D.6.2}$$

where $\|\cdot\|_{\mathrm{OP}}$ denotes the $L2$ operator norm.

D.6.4 The Berry–Esseen bound: Theorem 5.4.1

We here prove Theorem 5.4.1, a non-asymptotic upper bound on the error introduced by the two Gaussian approximations in Approximate Bound 5.4.1. We begin by restating key notation for convenience. We then state a more general variant of the bound that removes the restriction that the operators A and C be symmetric, and we show how it reduces to the simpler quantity stated in Theorem 5.4.1. Finally, we present a proof of the theorem as well as several supporting lemmas.

Notation and statement of the theorem its more general form. Recall that we are concerned with the coverage of Approximate Bound 5.4.1

$$\begin{aligned}b(y,\alpha) &= \|\hat{\theta}-y\|^2 - \|\theta^*-y\|^2 + 2\mathrm{tr}[(A-C)\Sigma] + \\ &\quad 2z_{\frac{1-\alpha}{2}}\sqrt{U(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2}) + \frac{1}{2}\|\Sigma^{\frac{1}{2}}(A+A^\top-C-C^\top)\Sigma^{\frac{1}{2}}\|_F^2}.\end{aligned}$$

In this equation, $G(y) := (A-C)y + (k-\ell)$, z_α denotes the α-quantile of the standard normal, and

$$U\big(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2}\big) = \inf_{\delta>0}\left\{\delta \;\middle|\; \|G(y)\|_\Sigma^2 \leq (\delta + \|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_F^2) + \right.$$

$$z_{\frac{1-\alpha}{2}}\sqrt{2\|\Sigma^{\frac{1}{2}}(A-C)\Sigma(A-C)^\top\Sigma^{\frac{1}{2}}\|_F^2+4\|\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2\delta}\Bigg\}$$

is a high-confidence upper bound on $\|G(\theta)\|_\Sigma^2$.

For convenience, we introduce

$$\tilde{F}^{-1}(\|G(\theta)\|_\Sigma^2,\alpha) := 2\mathrm{tr}[(A-C)\Sigma]+2z_\alpha\sqrt{\|G(\theta)\|_\Sigma^2+\frac{1}{2}\|\Sigma^{\frac{1}{2}}(A+A^\top-C-C^\top)\Sigma^{\frac{1}{2}}\|_F^2}, \tag{D.6.3}$$

to denote the inverse CDF of our normal approximation to the distribution of $2\epsilon^\top G(y)$ evaluated at α. As such, we may write

$$b(y,\alpha) = \|\hat{\theta}-y\|^2-\|\theta^*-y\|^2+\tilde{F}^{-1}\left(U(\|G(y)\|_\Sigma^2,\frac{1-\alpha}{2}),\frac{1-\alpha}{2}\right).$$

Finally, recall that to prove the theorem we desire to show

$$\mathbb{P}_\theta\left[W(\theta,y)\geq b(y,\alpha)\right]\geq\alpha-\frac{10\sqrt{2}}{\sqrt{N}}C_1\cdot\kappa(\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}})^2$$

for any θ and $\alpha\in[0,1]$, where $C_1<1.88$ is a universal constant, in the case when both A and C are symmetric. We accomplish this by first proving a more general bound holds even in the non-symmetric case,

$$\mathbb{P}_\theta\left[W(\theta,y)\geq b(y,\alpha)\right]\geq\alpha-\frac{5\sqrt{2}}{\sqrt{N}}C_1\left[\kappa(\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}})^2+\kappa(\Sigma^{\frac{1}{2}}(A+A^\top-C-C^\top)\Sigma^{\frac{1}{2}})\right]. \tag{D.6.4}$$

The special case obtains by replacing $A^\top$ and $C^\top$ with A and C, respectively, and noting that $\kappa(M)^2\geq\kappa(M)$ for any matrix, M.

A key tool in this proof is the classic result of Berry [1941], which we restate below.

Theorem D.6.1 (Berry, 1941, Theorem 1)**.** *Let $X_1,X_2,\ldots,X_N$ be random variables. For each $n\in\{1,2,\ldots,N\}$, let σ_n^2 and ρ_n denote the variance and third central moment of X_n, respectively. Define $\lambda_n:=\frac{\rho_n}{\sigma_n^2}$ if $\sigma_n^2>0$ and $\lambda_n=0$ otherwise. Define*

$\sigma^2 := \sum_{n=1}^N \sigma_n^2$ *and* $X := N^{-1}\sum_{n=1}^N X_n$. *Then*

$$\sup_x \left| F_X(x) - \Phi\left(\frac{x - \mathbb{E}[X]}{\sigma}\right)\right| < C_1 \frac{\max_n \lambda_n}{\sigma},$$

where $C_1 \leq 1.88$ *is a universal constant and* $F_X(\cdot)$ *is the cumulative distribution function of* X.

Proof of Theorem 5.4.1 The desired bound may be stated equivalently as, for any $\alpha \in [0,1]$,

$$\mathbb{P}_\theta\left[W(\theta,y) < b(y,\alpha)\right] < (1-\alpha) + \frac{5\sqrt{2}}{\sqrt{N}} C_1 \left[\kappa(\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}})^2 + \kappa(\Sigma^{\frac{1}{2}}(A+A^\top - C - C^\top)\Sigma^{\frac{1}{2}})\right]. \tag{D.6.5}$$

We first rewrite the condition $W(\theta,y) < b(y,\alpha)$ as $2\epsilon^\top G(y) < \tilde{F}^{-1}\left(U(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2}), \frac{1-\alpha}{2}\right)$ (recall Eq. (D.6.1)). Since $\tilde{F}^{-1}$ is monotonically decreasing in its first argument, this condition may occur only if either $2\epsilon^\top G(y) < \tilde{F}^{-1}\left(\|G(\theta)\|_\Sigma^2, \frac{1-\alpha}{2}\right)$ or $\|G(\theta)\|_\Sigma^2 > U(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2})$.

Therefore, by the union bound, we have that

$$\begin{aligned}\mathbb{P}_\theta\left[W(\theta,y) < b(y,\alpha)\right] < \mathbb{P}_\theta&\left[2\epsilon^\top G(y) < \tilde{F}^{-1}\left(\|G(\theta)\|_\Sigma^2, \frac{1-\alpha}{2}\right)\right] \\ &+ \mathbb{P}_\theta\left[\|G(\theta)\|_\Sigma^2 > U(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2})\right].\end{aligned} \tag{D.6.6}$$

Lemmas D.6.1 and D.6.2 provide that $\mathbb{P}_\theta\left[2\epsilon^\top G(y) < \tilde{F}^{-1}\left(\|G(\theta)\|_\Sigma^2, \frac{1-\alpha}{2}\right)\right] < \frac{1-\alpha}{2} + \frac{5\sqrt{2}}{\sqrt{N}} C_1 \kappa(\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}})$ and $\mathbb{P}_\theta\left[\|G(\theta)\|_\Sigma^2 > U(\|G(y)\|_\Sigma^2, \frac{1-\alpha}{2})\right] < \frac{1-\alpha}{2} + \frac{5\sqrt{2}}{\sqrt{N}} C_1 \kappa(\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}})^2$, respectively. Substituting these two bounds into Eq. (D.6.6) we obtain Eq. (D.6.5) as desired.

Lemma D.6.1. *Let* $y = \theta + \epsilon$ *be a random* N*-vector with* $\epsilon \sim \mathcal{N}(0,\Sigma)$. *Let* $\tilde{F}^{-1}$ *be the normal approximation to the inverse CDF of* $2\epsilon^\top G(y)$ *in Eq.* (D.6.3). *Then for*

any $\alpha \in [0,1]$,

$$\mathbb{P}_\theta\left[2\epsilon^\top G(y) < \tilde{F}^{-1}\left(\|G(\theta)\|_\Sigma^2, \alpha\right)\right] < \alpha + \frac{5\sqrt{2}}{\sqrt{N}} C_1 \kappa(\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}).$$

Proof. Note first that for any α we may rewrite

$$\begin{aligned}\mathbb{P}_\theta\left[2\epsilon^\top G(y) < \tilde{F}^{-1}\left(\|G(\theta)\|_\Sigma^2, \alpha\right)\right] &= F\left[\tilde{F}^{-1}\left(\|G(\theta)\|_\Sigma^2, \alpha\right)\right] \\ &= \alpha + \left\{F\left[\tilde{F}^{-1}\left(\|G(\theta)\|_\Sigma^2, \alpha\right)\right] - \tilde{F}\left[\tilde{F}^{-1}\left(\|G(\theta)\|_\Sigma^2, \alpha\right)\right]\right\},\end{aligned}$$

where F and $\tilde{F}$ are the exact and approximate CDFs of $2\epsilon^\top G(y)$, respectively. Recalling that the normal approximation comes from matching moments to $2\epsilon^\top G(y)$, we have that for any v, $\tilde{F}(v) = \Phi(\frac{v - \mathbb{E}[2\epsilon^\top G(y)]}{\sqrt{\mathrm{Var}[2\epsilon^\top G(y)]}})$. Therefore, it will suffice to obtain that for every v,

$$\left|\tilde{F}(v) - F(v)\right| = \left|F(v) - \Phi\left(\frac{v - \mathbb{E}[2\epsilon^\top G(y)]}{\sqrt{\mathrm{Var}[2\epsilon^\top G(y)]}}\right)\right| \leq \frac{5\sqrt{2}}{\sqrt{N}} C_1 \kappa(\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}).$$

We will obtain this result by writing $2\epsilon^\top G(y)$ a sum of independent random variables and using a Berry–Esseen Theorem (Theorem D.6.1) to bound the error of this normal approximation.

Lemma D.6.3 allows us to write $2\epsilon^\top G(y) = 2\epsilon^\top(A-C)\epsilon + 2\left[(A-C)\theta + (k-\ell)\right]^\top \epsilon$ as a shifted sum of N differently-scaled, independent non-central χ^2 random variables. We denote these N random variables by $X_1, X_2, \ldots, X_N$. Lemma D.6.3 additionally tells us that the scaling parameters of these non-central χ^2 random variables will be the eigenvalues of $\Sigma^{\frac{1}{2}}(A + A^\top - C^\top - C)\Sigma^{\frac{1}{2}}$, which we denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$.

To use Theorem D.6.1 we require the ratios of the third to second central moments of these random variables, as well as the variance of the sum. Specifically,

$$\sup_{v \in \mathbb{R}} \left|\Phi(\frac{v - \mathbb{E}[2\epsilon^\top G(y)]}{\sqrt{\mathrm{Var}[2\epsilon^\top G(y)]}}) - F(v))\right| < C_1 \frac{\max_n \frac{\rho(X_n)}{\mathrm{Var}[X_n]}}{\sqrt{\mathrm{Var}[2\epsilon^\top G(y)]}},$$

where for each index n, $\rho(X_n) := \mathbb{E}[(X_n - \mathbb{E}[X_n])^3]$ is the third central moment of X_n,

and $C_1 < 1.88$ is a universal constant.

Conveniently, as we show in Lemma D.6.4, for each n, $\frac{\rho(X_n)}{\mathrm{Var}[X_n]} \leq 10\lambda_n$. Further, since $\sqrt{\mathrm{Var}[2\epsilon^\top G(y)]} > \sqrt{2\sum_{n=1}^N \lambda_n^2} > \sqrt{2N}\lambda_N$ (recall that Eq. (5.13) provides that $\mathrm{Var}[2\epsilon^\top G(y)] = 4\|G(\theta)\|_\Sigma^2 + 2\|\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}\|_F^2$) we may additionally see that

$$\sup_{v\in\mathbb{R}} \left| \Phi\left(\frac{v - \mathbb{E}[2\epsilon^\top G(y)]}{\sqrt{\mathrm{Var}[2\epsilon^\top G(y)]}}\right) - F(v)) \right| < C_1 \frac{10}{\sqrt{2N}} \frac{\max_n \lambda_n}{\min_n \lambda_n}$$
$$= C_1 \frac{5\sqrt{2}}{\sqrt{N}} \kappa\left(\Sigma^{\frac{1}{2}}(A + A^\top - C - C^\top)\Sigma^{\frac{1}{2}}\right)$$

where $\kappa(\cdot)$ denotes the condition number of its matrix argument, as desired. □

Lemma D.6.2. *Let $y = \theta + \epsilon$ be a random N-vector with $\epsilon \sim \mathcal{N}(0, \Sigma)$. Let $U(\|G(y)\|_\Sigma^2, \alpha)$ be the approximate high-confidence upper bound on $\|G(\theta)\|_\Sigma^2$. Then for any $\alpha \in [\frac{1}{2}, 1]$, $\mathbb{P}_\theta\left[\|G(\theta)\|_\Sigma^2 > U(\|G(y)\|_\Sigma^2, 1-\alpha)\right] < 1 - \alpha + \frac{5\sqrt{2}}{\sqrt{N}} C_1 \kappa(\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}})^2$.*

Proof. Our proof of the lemma follows roughly the same approach taken to prove Lemma D.6.1. First note that the condition that $\|G(\theta)\|_\Sigma^2 > U(\|G(y)\|_\Sigma^2, 1-\alpha)$ implies that

$$\|G(y)\|_\Sigma^2 \leq (\|G(\theta)\|_\Sigma^2 + \|\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}}\|_F^2) +$$
$$z_{1-\alpha}\sqrt{2\|\Sigma^{\frac{1}{2}}(A - C)\Sigma(A - C)^\top\Sigma^{\frac{1}{2}}\|_F^2 + 4\|\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}}\|_{\mathrm{OP}}^2\|G(\theta)\|_\Sigma^2}$$
$$\leq \mathbb{E}[G(y)\|_\Sigma^2] + z_{1-\alpha}\sqrt{\mathrm{Var}[G(y)]}$$

for any $\alpha \in [\frac{1}{2}, 1]$, where the first line follows from the definition of $U(\|G(y)\|_\Sigma^2, 1-\alpha)$. The second line follows from the observations that (A) $z_{1-\alpha} < 0$ and (B) the second term in the first line uses an upper bound on the variance of $\|G(y)\|_\Sigma^2$ (Eq. (5.16)).

We now proceed to upper bound the probability of the event in the display equation above. First consider a normal approximation to the distribution of $\|G(y)\|_\Sigma$ with matched moments, and denote its inverse CDF by $F^{\dagger-1}(\theta, \alpha)$. We may then write the

probability of the event above as

$$\mathbb{P}\left[\|G(y)\|_\Sigma^2 \leq \mathbb{E}[G(y)\|_\Sigma^2] + z_\alpha\sqrt{\mathrm{Var}[G(y)]}\right] = F\left[F^{\dagger -1}(\theta,\alpha)\right]$$
$$= \alpha + \left\{F\left[\bar{F}^{-1}(\theta,\alpha)\right] - F^\dagger\left[F^{\dagger -1}(\theta,\alpha)\right]\right\},$$

where $F(\cdot)$ and $F^\dagger(\cdot)$ denote the exact and approximate CDFs of $\|G(y)\|_\Sigma^2$. It will suffice to show that for any v,

$$|F(v) - F^\dagger(v)| \leq \frac{5\sqrt{2}}{\sqrt{N}}\kappa(\Sigma^{\frac{1}{2}}(A-C)\Sigma^{\frac{1}{2}})^2.$$

As in Lemma D.6.1 we obtain this result through the Berry–Esseen theorem. In this case, the variable of interest is $\|G(y)\|_\Sigma^2 = \epsilon^\top(A-C)^\top\Sigma(A-C)\epsilon + 2\epsilon^\top\left[(A-C)\theta + (k-\ell)\right]$. As in this previous lemma, we use Lemma D.6.3 to write this variable as a shifted sum of independent, scaled non-central χ^2 random variables, this time with scaling parameters equal to the eigenvalues $\Sigma^{\frac{1}{2}}(A-C)^\top\Sigma(A-C)\Sigma^{\frac{1}{2}}$. Recognizing that the eigenvalues of the matrix $M^\top M$ are the squares of the singular values of M for any matrix M, we obtain the desired result. □

Lemma D.6.3. *Let X be a random N-vector distributed as $X \sim 2\epsilon^\top A\epsilon + b^\top\epsilon$ where $A \in \mathbb{R}^{N\times N}, b \in \mathbb{R}^N$, and $\epsilon \sim \mathcal{N}(0,\Sigma)$. Then X is distributed as a shifted sum of differently scaled, independent non-central χ^2 random variables. In particular, if we let $U\mathrm{diag}(\lambda)U^\top$ be the eigen-decomposition of $\Sigma^{\frac{1}{2}}(A+A^\top)\Sigma^{\frac{1}{2}}$, then we can write $X \overset{d}{=} \sum_{n=1}^N Y_n - \frac{1}{4}\|\mathrm{diag}(\lambda)^{-1}U^\top\Sigma^{\frac{1}{2}}b\|_2$, where each $Y_n \overset{indep}{\sim} \lambda_n\chi_1^2(\frac{1}{2}\lambda_n^{-1}e_n^\top U^\top\Sigma^{\frac{1}{2}}b)$, where e_n is the n^{th} basis vector.*

Proof. The proof of the lemma proceeds through a long algebraic rearrangement. In particular we rewrite X as

$$\begin{aligned}
X &= 2\epsilon^\top A\epsilon + b^\top\epsilon \\
&= \delta^\top\Sigma^{\frac{1}{2}}(A+A^\top)\Sigma^{\frac{1}{2}}\delta + b^\top\Sigma^{\frac{1}{2}}\delta \\
&// \text{ defining } \delta := \Sigma^{-\frac{1}{2}}\epsilon \text{ so that } \delta \sim \mathcal{N}(0, I_N). \\
&= \delta^\top U\mathrm{diag}(\lambda)U^\top\delta + b^\top\Sigma^{\frac{1}{2}}U\mathrm{diag}(\lambda)^{-\frac{1}{2}}\mathrm{diag}(\lambda)^{\frac{1}{2}}U^\top\delta
\end{aligned}$$

$$
\begin{aligned}
&// \text{ Letting } U\text{diag}(\lambda)U^\top := \Sigma^{\frac{1}{2}}(A + A^\top)\Sigma^{\frac{1}{2}} \text{ be an eigen-decomposition,} \\
&// \text{ with } U^\top U = I_N \text{ and } \lambda \in \mathbb{R}_+^N \\
&\overset{d}{=} \delta^\top \text{diag}(\lambda)\delta + b^\top \Sigma^{\frac{1}{2}} U \text{diag}(\lambda)^{-\frac{1}{2}} \text{diag}(\lambda)^{\frac{1}{2}} \delta \\
&= \sum_{n=1}^{N} (\lambda_n^{\frac{1}{2}} \delta_n + \frac{1}{2}\lambda_n^{-\frac{1}{2}} e_n^\top U^\top \Sigma^{\frac{1}{2}} b)^2 - \frac{1}{4} b^\top \Sigma^{\frac{1}{2}} U \text{diag}(\lambda)^{-1} U^\top \Sigma^{\frac{1}{2}} b \\
&\overset{d}{=} \frac{-b^\top (A + A^\top)^{-1} b}{4} + \sum_{n=1}^{N} \lambda_n \chi_1^2 (\frac{1}{2}\lambda_n^{-1} e_n^\top U^\top \Sigma^{\frac{1}{2}} b),
\end{aligned}
$$

where each e_n denotes the n^{th} basis vector and each of the scaled non-central χ^2 random variables in the last line are independent. □

Lemma D.6.4. *Consider a scaled non-central chi-squared random variable, $X \sim s\chi_1^2(\lambda)$, where s and λ are scaling and non-centrality parameters, respectively. Denote the second and third central moments of X by $\sigma^2 = \text{Var}[X]$ and $\rho = \mathbb{E}\left[(X - \mathbb{E}[X])^3\right]$. Then $\frac{\rho}{\sigma^2} \leq 10s$.*

Proof. Recall that the second and third central moments of the scaled non-central χ^2 have known forms, $\sigma^2 = 2s^2(1 + 2\lambda)$ and $\rho = 8s^3(1 + 3\lambda)$. Therefore we may write

$$
\begin{aligned}
\frac{\rho}{\sigma^2} &= \frac{8s^3(1 + 3\lambda)}{2s^2(1 + 2\lambda)} \\
&\leq 4s\left(\frac{1}{1} + \frac{3\lambda}{2\lambda}\right) \\
&= \frac{4 \cdot 5}{2} s \\
&= 10s,
\end{aligned}
$$

as desired. □

D.7 Empirical Bayes supplementary details

D.7.1 Additional figure

Figure D.7.1 shows the calibration in the simulation experiment described in Section 5.5.1.

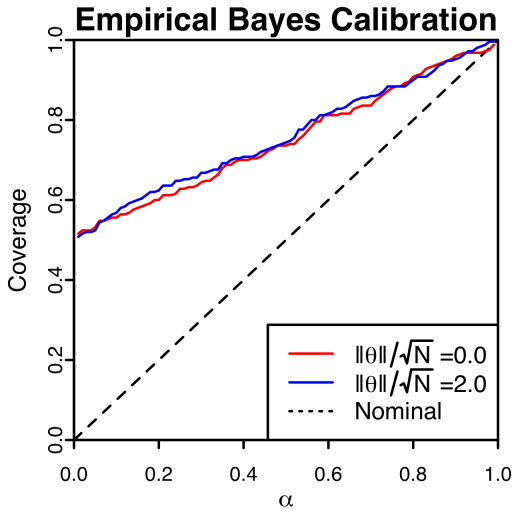


Figure D.7.1: Calibration of approximate high-confidence bounds on the win of an empirical Bayes estimate over the MLE in simulation. Each series depicts calibration for a different choice of the parameter θ ($N = 50$).

D.7.2 Asymptotic coverage of the empirical Bayes estimate

Theorem 5.5.1 shows that we can apply the machinery developed for Bayes rules with fixed priors to lower bound the win with at least the desired coverage asymptotically. We here consider a scaling of win,

$$W_N(\Theta_N, Y_N) := \frac{1}{\sqrt{N}} \left[\|Y_N - \Theta_N\|^2 - \|\Theta_N^*(Y_N) - \Theta_N\|^2 \right].$$

We use a special case of Bound D.5.1 in Appendix D.5.4 with no covariates (i.e. $D = 0$), and we treat the estimate $\hat{\tau}_N^2(Y_N)$ as if it were fixed rather than estimated from the data. For each N, this bound is

$$b_N(Y_N, \alpha) := \frac{1}{\sqrt{N}} \inf_{\lambda \in [0, U(Y_N, \frac{1-\alpha}{2})]} \frac{2}{1+\hat{\tau}_N^2} F^{-1}\left[\chi_N^2(\frac{\lambda}{4}), \frac{1-\alpha}{2} \right] - \frac{\lambda}{2(1+\hat{\tau}_N^2)} - \frac{\|Y_N\|^2}{(1+\hat{\tau}_N^2)^2}$$

where $F^{-1}\left[\chi^2_N(\lambda), 1-\alpha\right]$ denotes the inverse cumulative distribution function of the non-central χ^2 with N degrees of freedom and non-centrality parameter λ, evaluated at $1-\alpha$ and $U(Y_N, 1-\alpha) := \inf_{\delta \geq 0} \left\{ \delta \middle| \|Y_N\|^2 \leq F^{-1}\left[\chi^2_N(\delta), 1-\alpha\right] \right\}$ is a high-confidence upper bound on $\|\theta\|^2$.

For our theorem and its proof, a key quantity is, for each N, the sample second moment for the first N parameters, which we denote by $\tau_N^2 := N^{-1}\sum_{n=1}^N \theta_n^2$. We emphasize, however, that while it may be convenient to describe τ_N^2 as a sample moment, θ is fixed in Theorem 5.5.1 and throughout this analysis.

Proof of Theorem 5.5.1. We prove the theorem by showing that for any α, the gap between the win $W_N(\Theta_N, Y_N)$ and the bound $b_N(Y_N, \alpha)$ computed for the empirical Bayes estimate converges in distribution to the gap between the analogous win and bound computed for the same estimates but with prior variance fixed as $\tau^2 = \tau_N^2$. We denote these latter quantities by $W_N^*(\Theta_N, Y_N)$ and $b_N^*(Y_N, \alpha)$, and note that since τ_N^2 is fixed $\mathbb{P}[W_N^*(\Theta_N, Y_N) \geq b^*(Y_N, \alpha)] \geq \alpha$ by construction (Proposition D.5.1). For convenience, we denote $W_N(\Theta_N, Y_N)$ by W_N, $b_N(Y_N, \alpha)$ by b_N, $W_N^*(\Theta_N, Y_N)$ by W_N^*, and $b_N^*(Y_N, \alpha)$ by b_N^*.

Observe that we can write

$$W_N - b_N = \frac{W_N - b_N}{W_N^* - b_N^*}(W_N^* - b_N^*).$$

By Lemma D.7.4, $W_N^* - b_N^*$ is asymptotically Gaussian, and by Lemma D.7.2 $\frac{W_N - b_N}{W_N^* - b_N^*} \xrightarrow{p} 1$. As a result, the distribution of $W_N - b_N$ approaches the distribution of $W_N^* - b_N^*$ in supremum norm. Since b_N^* obtains the desired coverage by construction, the result follows.

Supporting lemmas.

Lemma D.7.1. *If the sequence τ_N^2 is bounded, then $\tau_N^2 - \hat{\tau}_N^2$ is $O_p(N^{-\frac{1}{2}})$, where $O_p(\cdot)$ denotes stochastic convergence in probability.*

Proof. Note that for each N, $\|Y_N\|^2 \sim \chi^2_N(N\tau_N^2)$. Therefore we have that $\mathbb{E}[\|Y_N\|^2] =$

$N + N\tau_N^2$ and $\mathrm{Var}[\|Y_N\|^2] = 2(N + 2N\tau_N^2)$. So, recalling that $\hat{\tau}_N^2 := \frac{\|Y_N\|^2}{N-2} - 1 = \frac{\|Y_N\|^2-(N-2)}{N-2}$ we may write

$$\begin{aligned}\hat{\tau}_N^2 &= \frac{\|Y_N\|^2 - \mathbb{E}[\|Y_N\|^2]}{N-2} + \frac{(N + N\tau_N^2) - (N-2)}{N-2} \\ &= \frac{\|Y_N\|^2 - \mathbb{E}[\|Y_N\|^2]}{N} + \tau_N^2 + O(\frac{1}{N}).\end{aligned}$$

And so

$$\begin{aligned}|\hat{\tau}_N^2 - \tau_N^2| &\leq \Big|\frac{\|Y_N\|^2 - \mathbb{E}[\|Y_N\|^2]}{N}\Big| + O(\frac{1}{N}) \\ &= \left(\frac{\sqrt{2+4\tau_N^2}}{\sqrt{N}}\right)\Big|\frac{\|Y_N\|^2 - \mathbb{E}[\|Y_N\|^2]}{\sqrt{\mathrm{Var}[\|Y_N\|^2]}}\Big| + O(\frac{1}{N}).\end{aligned}$$

By Chebyshev's inequality, $\frac{\|Y_N\|^2-\mathbb{E}[\|Y_N\|^2]}{\sqrt{\mathrm{Var}[\|Y_N\|^2]}}$ is bounded in probability and we can see that $|\hat{\tau}_N^2 - \tau_N^2|$ is $O_p(N^{-\frac{1}{2}})$. □

Lemma D.7.2. *Let W_N^* and b_N^* denote the win and its bound evaluated for $\tau^2 = \tau_N^2$, rather than the empirical Bayes estimate. Then*

$$\frac{W_N - b_N}{W_N^* - b_N^*} = 1 + \frac{\tau_N^2 - \hat{\tau}_N^2}{1+\hat{\tau}_N^2} = 1 + O_p(\frac{1}{\sqrt{N}}).$$

Proof. Recall that we may decompose W_N as

$$W_N(\Theta_N, Y_N) = \frac{1}{\sqrt{N}}\left[\frac{2}{1+\hat{\tau}_N^2}\epsilon_N^\top Y_N - \frac{1}{(1+\hat{\tau}_N^2)^2}\|Y_N\|^2\right]$$

and that our bound is

$$b_N(Y_N, \alpha) = \frac{1}{\sqrt{N}}\left\{\inf_{\lambda\in[0, U(Y_N, \frac{1-\alpha}{2})]}\frac{2}{1+\hat{\tau}_N^2}F^{-1}\left[\chi_N^2(\frac{\lambda}{4}), \frac{1-\alpha}{2}\right] - \frac{\lambda}{2(1+\hat{\tau}_N^2)} - \frac{\|Y_N\|^2}{(1+\hat{\tau}_N^2)^2}\right\},$$

where $U(Y_N, \alpha)$ does not depend on $\hat{\tau}_N^2$.

As such,

$$W_N - b_N = \frac{2}{\sqrt{N}(1+\hat{\tau}_N^2)} \left\{ \epsilon_N^\top Y_N - \inf_{\lambda \in [0, U(Y_N, \frac{1-\alpha}{2})]} F^{-1}\left[\chi_N^2(\frac{\lambda}{4}), \frac{1-\alpha}{2}\right] + \frac{\lambda}{4} \right\},$$

and we can see that

$$\begin{aligned}\frac{W_N - b_N}{W_N^* - b_N^*} &= \frac{1+\tau_N^2}{1+\hat{\tau}_N^2} \\ &= 1 + \frac{\tau_N^2 - \hat{\tau}_N^2}{1+\hat{\tau}_N^2}.\end{aligned}$$

By Lemma D.7.1 the second term is $O_p(N^{-\frac{1}{2}})$, as desired. □

Lemma D.7.3. *Let $\lambda_1, \lambda_2, \ldots$ be a sequence of reals satisfying, for each N, $N^{-1}\lambda_N < \kappa$ for some constant κ. Let $F^{-1}_{\chi^2_N}(\lambda_N, \alpha)$ denote the inverse CDF of a non-central χ^2 with N degrees of freedom and non-centrality parameter λ_N. Then for any $\alpha \in (0,1)$,*

$$\frac{1}{\sqrt{N}}\left[F^{-1}_{\chi^2_N}(\lambda_N, \alpha) - (N + \lambda_N)\right] = \sqrt{2 + 4\frac{\lambda_N}{N}} z_\alpha + O(\frac{1}{\sqrt{N}}),$$

where z_α is the α-quantile of the standard normal.

Proof. Note that a $\chi^2_N(\lambda_N)$ random variable is equal in distribution to a sum of N i.i.d. $\chi^2_1(N^{-1}\lambda_N)$ random variables. Let $\sigma_N^2 := \mathrm{Var}[\chi^2_1(N^{-1}\lambda_N)] = 2 + 4N^{-1}\lambda_N$ and note that each $\sigma_N^2 \geq 2$. Let $\rho_N := 8 + 24N^{-1}\lambda_N$ be third central moment of these variates and note that each $\rho_N \leq 8 + 24\kappa$.

Let $F_{\chi^2_N(\lambda_N)}(x)$ denote the CDF of a non-central χ^2 random variable with N degrees of freedom and non-centrality parameter λ_N evaluated at x. By the Berry–Esseen theorem [Berry, 1941, Theorem 1], for all x

$$\begin{aligned}\left| F_{\chi^2_N(\lambda_N)}(x) - \Phi\left[\frac{x - (N+\lambda_N)}{\sqrt{2N + 4\lambda_N}}\right] \right| &\leq \frac{C_1 \rho}{\sigma^3 \sqrt{N}} \\ &\leq \frac{C_1(8+24\kappa)}{2^{\frac{3}{2}}\sqrt{N}} \\ &= O(\frac{1}{\sqrt{N}}),\end{aligned}$$

where $C_1 \leq 1.88$ is a universal constant. Since $\Phi(\cdot)$ is continuously differentiable and invertible, we obtain the same convergence rate for the inverse CDFs. That is, for any $\alpha \in (0,1)$,

$$\frac{F^{-1}_{\chi^2_N}(\lambda_N, \alpha) - (N + \lambda_N)}{\sqrt{2N + 4\lambda_N}} - z_\alpha = O(\frac{1}{\sqrt{N}}).$$

Rescaling these terms by $N^{-\frac{1}{2}}\sqrt{2N + 4\lambda_N}$ and rearranging, we find

$$\frac{1}{\sqrt{N}}\left[F^{-1}_{\chi^2_N}(\lambda_N, \alpha) - (N + \lambda_N)\right] = \sqrt{2 + 4\frac{\lambda_N}{N}}z_\alpha + O(\frac{1}{\sqrt{N}})$$

as desired. □

Lemma D.7.4. *Let b^*_N and W^*_N again denote the win and bounds evaluated for the variance $\tau^2 = \tau^2_N$ rather than the empirical Bayes estimate. If the sequence τ^2_N is bounded, then*

$$\frac{(W^*_N - b^*_N) - c_N}{d_N} \to \mathcal{N}(0,1)$$

for some sequences of constants $c_1, c_2, \ldots$ and $d_1, d_2, \ldots$.

Proof. Let κ be such that for all N, $\tau^2_N < \kappa$.

Recall that we may write

$$W^*_N - b^*_N = \frac{2}{\sqrt{N}(1+\tau^2_N)}\left\{\epsilon_N^\top Y_N - \inf_{\lambda \in [0, U(Y_N, \frac{1-\alpha}{2})]} F^{-1}\left[\chi^2_N(\frac{\lambda}{4}), \frac{1-\alpha}{2}\right] + \frac{\lambda}{4}\right\}. \tag{D.7.1}$$

To prove the lemma, we build off of the normal approximation described in Appendix D.5.1. Note first that an application of Chebyshev's inequality provides that $N^{-1}U(Y_N, \frac{1-\alpha}{2}) - \tau^2_N$ is $O_p(N^{-\frac{1}{2}})$, so that $N^{-1}U(Y_N, \frac{1-\alpha}{2}) < \kappa$ with probability approaching 1. Next, by Lemma D.7.3,

$$\frac{1}{\sqrt{N}}\left\{F^{-1}\left[\chi^2_N(\frac{\lambda_N}{4}), \frac{1-\alpha}{2}\right] - \left[\frac{\lambda_N}{4} + N\right]\right\} = \sqrt{2 + \frac{\lambda_N}{N}}z_{\frac{1-\alpha}{2}} + O(\frac{1}{\sqrt{N}}),$$

for any sequence $\lambda_1, \lambda_2, \ldots$ that satisfies, for each N, $N^{-1}\lambda_N < \kappa$.

Notably, since any sequence of λ_N's achieving the infima in Eq. (D.7.1) will satisfy this condition, we may substitute this expression in and rewrite $W_N^* - b_N^*$ as

$$\begin{aligned}
W_N^* - b_N^* &= \frac{2}{1+\tau_N}\left[\frac{\epsilon_N^\top Y_N}{\sqrt{N}} - \sqrt{N}\left\{\inf_{\lambda_N \in [0, U(Y_N, \frac{1-\alpha}{2})]} F^{-1}\left[\chi_N^2(\frac{\lambda_N}{4}), \frac{1-\alpha}{2}\right] - \left[\frac{\lambda_N}{4} + N\right]\right\} - \sqrt{N}\right] \\
&= \frac{2}{1+\tau_N}\left[\frac{\epsilon_N^\top Y_N - N}{\sqrt{N}} - \inf_{\lambda_N \in [0, U(Y_N, \frac{1-\alpha}{2})]} z_{\frac{1-\alpha}{2}}\sqrt{2 + \frac{\lambda_N}{N}} + O_p(\frac{1}{\sqrt{N}})\right] \\
&= \frac{2}{1+\tau_N}\left[\frac{\epsilon_N^\top Y_N - N}{\sqrt{N}} - z_{\frac{1-\alpha}{2}}\sqrt{2 + \frac{U(Y_N, \frac{1-\alpha}{2})}{N}} + O_p(\frac{1}{\sqrt{N}})\right] \\
&= \frac{2}{1+\tau_N}\left[\frac{\epsilon_N^\top Y_N - N}{\sqrt{N}} - z_{\frac{1-\alpha}{2}}\sqrt{2 + \tau_N^2} + O_p(\frac{1}{\sqrt{N}})\right] \\
&\text{// Since } \tau_N^2 - \frac{U(Y_N, \frac{1-\alpha}{2})}{N} \text{ is } O_p(\frac{1}{\sqrt{N}}).
\end{aligned}$$

Finally, note that $\epsilon^\top Y_N$ is approximately normal with mean N and variance $N(2+\tau_N^2)$. Furthermore, the distribution of this quantity approaches that of a normal at the same $O(N^{-\frac{1}{2}})$ rate in the supremum norm (one may make this precise with a Berry–Esseen bound). This allows us to write

$$\begin{aligned}
W_N^* - b_N^* &\sim \frac{2}{1+\tau_N^2}\left[\sqrt{2+\tau_N^2}x - \sqrt{2+\tau_N^2}z_{\frac{1-\alpha}{2}}\right] + O_p(\frac{1}{\sqrt{N}}) \\
&\sim \frac{2\sqrt{2+\tau_N^2}}{1+\tau_N^2}(x - z_{\frac{1-\alpha}{2}}) + O_p(\frac{1}{\sqrt{N}})
\end{aligned}$$

for $x \sim \mathcal{N}(0,1)$. The result obtains by taking $d_N := (2\sqrt{2+\tau_N^2})/(1+\tau_N^2)$ and $c_N := -d_N z_{\frac{1-\alpha}{2}}$, and noting that the lower order term does not influence the limiting distribution of $d_N^{-1}\left[(W_N^* - b_N^*) - c_N\right]$. □

D.8 Logistic regression supplementary material

This section provides supplementary information related to Section 5.5.2. We begin by reviewing notation for convenience in Appendix D.8.1. In Appendix D.8.2 we then provide a proposition demonstrating the asymptotic rate of convergence of the

approximation of the MAP estimate to the exact MAP estimate, as well a proof and supporting lemmas. Appendix D.8.3 then provides a proof of Theorem 5.5.3. Appendix D.8.4 gives additional details on the simulation experiments.

D.8.1 Preliminaries and notation

Consider logistic regression with random N-vector covariates $x_1, x_2, \ldots$ and responses $y_1, y_2, \ldots$, where for each data point m, $y_m \mid x_m, \theta \sim (1 + \exp\{-x_m^\top \theta\})^{-1}\delta_1 + (1 + \exp\{x_m^\top \theta\})^{-1}\delta_{-1}$ for some unknown parameter $\theta \in \mathbb{R}^N$. We use $X_M = [x_1, x_2, \ldots, x_M]^\top$ and $Y_M = [y_1, y_2, \ldots, y_M]^\top$ to denote the first M data points.

One choice of an estimate for θ after observing M observations is the MLE,

$$\hat{\theta}_M := \arg\max_\theta \log p(Y_M \mid X_M, \theta).$$

Another possibility is the MAP estimate under a standard normal prior

$$\theta^*_M := \arg\max_\theta \log p(Y_M \mid X_M, \theta) - \frac{1}{2}\|\theta\|^2.$$

The approach in Section 5.5.2 involves an approximation to this estimate involving a Gaussian approximation to the likelihood, defined by a 2nd order Taylor approximation of the log posterior formed at $\hat{\theta}_M$. In particular, by Bayes' rule, the log posterior is, up to an additive constant,

$$\log p_M(\theta) := \log p(Y_M \mid X_M, \theta) - \frac{1}{2}\|\theta\|^2$$

and we use the approximation

$$\log \tilde{p}_M(\theta) := \log p(Y_M \mid X_M, \hat{\theta}_M) - \frac{1}{2}\|\theta\|^2 - \frac{1}{2}(\theta - \hat{\theta}_M)^\top H_M(\hat{\theta}_M)(\theta - \hat{\theta}_M), \quad \text{(D.8.1)}$$

where $H_M(\hat{\theta}_M) = \nabla^2_\theta - \log p(Y_M \mid X_M, \theta)\big|_{\theta=\hat{\theta}_M}$ is the Hessian of the negative log likelihood, computed at the MLE.

The approximation we use for computing our proposed bound is then the maximizer

of this approximation

$$\tilde{\theta}^*_M := \arg\max_\theta \log \tilde{p}_M(\theta).$$

In Section 5.5.2 we found that we could express $\tilde{\theta}^*_M$ as

$$\tilde{\theta}^*_M = \left[I_N + \tilde{\Sigma}_M\right]^{-1} \hat{\theta}_M,$$

where $\tilde{\Sigma}_M := H_M(\hat{\theta}_M)^{-1}$ is an approximation to the covariance of $\hat{\theta}_M$. This solution may be seen by considering the first order optimality condition (i.e. setting the gradient of $\log \tilde{p}_M(\theta)$ to zero).

D.8.2 Asymptotic approximation quality

We here show that, in the large sample limit, $\tilde{\theta}^*$ provides a very close approximation of the MAP estimate, θ^*.

Proposition D.8.1 (Asymptotic approximation quality)**.** *Consider Bayesian logistic regression with a Gaussian prior $\theta \sim \mathcal{N}(0, I_N)$. Let $x_1, x_2, \ldots$ be a sequence of random i.i.d. covariates satisfying $\mathbb{E}[x_m x_m^\top] \succ 0$ and with bounded third moment, and let $y_1, y_2, \ldots$ be responses distributed as in Eq.* (5.18)*. Denote by $X_M := [x_1, x_2, \ldots, x_M]^\top$ and $Y_M := [y_1, y_2, \ldots, y_M]^\top$ the covariates and labels of the first M data points. Consider the MAP estimate of θ after observing M data points,*

$$\theta^*_M := \arg\max_\theta p(\theta|Y_M, X_M) \text{ and the approximation } \tilde{\theta}^*_M := \left[I_N + \tilde{\Sigma}_M\right]^{-1} \hat{\theta}_M, \tag{D.8.2}$$

where $\hat{\theta}_M := \arg\max_\theta p(Y_M|X_M;\theta)$ and $\tilde{\Sigma}_M := \left[-\nabla^2_\theta \log p(Y_M|X_M;\theta)\big|_{\theta=\hat{\theta}_M}\right]^{-1}$. Then $\|\tilde{\theta}^*_M - \theta^*_M\| \in O_p(M^{-2})$, where O_p denotes stochastic convergence in probability.

The $O_p(M^{-2})$ convergence rate established in Proposition D.8.1 is very fast in comparison to the $O_p(M^{-\frac{1}{2}})$ convergence rate of the MLE, as well as to the $O_p(M^{-1})$ rate of convergence of the MAP to the posterior mean. Notably, this asymptotic rate is consistent with rates observed in simulation (Figure D.8.1a).

Proof. We here show that $\|\theta^*_M - \tilde{\theta}^*_M\|$ is $O_p(M^{-2})$. Our route to proving this relies on Lemma D.8.1 [Trippe et al., 2019, Lemma E.1], which will provide a sequence of bounds on $\|\theta^*_M - \tilde{\theta}^*_M\|$ that depend on the norms of the gradients of $\log p_M(\cdot)$ at $\tilde{\theta}^*_M$, $c_M := \|\nabla_\theta \log p_M(\tilde{\theta}^*_M)\|$, and a sequence of strong log-concavity constants α_M for $\log p_M(\cdot)$ which hold on the interval $\{t\theta^*_M + (1-t)\tilde{\theta}^*_M | t \in [0,1]\}$. In particular, Lemma D.8.1 provides that $\|\theta^*_M - \tilde{\theta}^*_M\| \leq \frac{c_M}{\alpha_M}$ and we obtain the result by showing that α_M grows as $\Omega_p(M)$ and c_M drops as $O_p(M^{-1})$.

We first use Lemma D.8.3 to show that the strong log-concavity constants of $\log p_M$ in a neighborhood of radius ϵ of $\theta, B_\epsilon(\theta)$ grow as $\Omega_p(M)$. This allows us to establish that $\|\tilde{\theta}^*_M - \hat{\theta}_M\|$ is $O_p(M^{-1})$ (Lemma D.8.4). Since both $\hat{\theta}_M$ and θ^*_M converge strongly to θ under these conditions (see e.g. Van der Vaart [2000, Theorem 10.10]), the interval $\{t\theta^*_M + (1-t)\tilde{\theta}^*_M | t \in [0,1]\}$ is then contained within $B_\epsilon(\theta)$ with probability approaching 1. Consequently, the constants of strong log concavity of $\log p_M$ on this interval, which we take as $\alpha_1, \alpha_2, \ldots$, must grow as $\Omega_p(M)$ as well.

Now all that remains is to show that c_M drops as $O_p(M^{-1})$. Recall from above that $\|\tilde{\theta}^*_M - \hat{\theta}_M\|$ is $O(M^{-1})$. This fact and the boundedness of the higher derivatives of $\nabla \log p_M$ will allow us to use Taylor's theorem to obtain the desired rate.

However, before proceeding to a more detailed derivation of this rate, we introduce some additional notation. Let $\phi(y, a)$ denote the GLM mapping function, such that

$$\begin{aligned}\phi(y, a = x^\top \theta) &= \log p(y|x, \theta) \\ &= -\log(1 + \exp\{-yx^\top \theta\})\end{aligned}$$

and note that all higher derivatives with respect to a are bounded. In particular, third derivative satisfies

$$\phi'''(a) := \frac{d^3}{da^3}\phi(y, a) \leq \frac{1}{6\sqrt{3}},$$

where we have dropped y as an argument, because these higher derivatives do not depend on y.

We now proceed to derive a stochastic rate of convergence of $\|\nabla_\theta \log p_M(\tilde{\theta}^*_M)\|$.

We obtain this through a long derivation involving a series of upper bounds.

$$
\begin{aligned}
\|\nabla_\theta \log p_M(\tilde{\theta}^*_M)\| &= \|\nabla_\theta (\log p_M - \log \tilde{p}_M)(\tilde{\theta}^*_M)\| \\
&= \|\nabla_\theta (\log p_M - \log \tilde{p}_M)(\hat{\theta}_M) + (\tilde{\theta}^*_M - \hat{\theta}_M)^\top \nabla^2_\theta (\log p_M - \log \tilde{p}_M)(\theta'_M)\| \\
&\quad // \text{ By Taylor's theorem, for some } \theta'_M \in \{t\hat{\theta}_M + (1-t)\tilde{\theta}^*_M | t \in [0,1]\} \\
&= \|(\tilde{\theta}^*_M - \hat{\theta}_M)^\top \nabla^2_\theta (\log p_M(\theta'_M) - \log \tilde{p}_M(\theta'_M))\| \\
&\quad // \text{ Since } \nabla_\theta \log \tilde{p}_M(\hat{\theta}) = \nabla_\theta \log p_M(\hat{\theta}) \\
&= \|(\tilde{\theta}^*_M - \hat{\theta}_M)^\top \left[\nabla^2_\theta \log p(Y_M|X_M, \theta'_M) - \nabla^2_\theta \log p(Y_M|X_M, \hat{\theta}_M)\right] \| \\
&\quad // \text{ Since } \log \tilde{p}_M \text{ is a second degree approximation defined at } \hat{\theta}_M \\
&\leq \|\tilde{\theta}^*_M - \hat{\theta}_M\| \left[\sum_{m=1}^M \|\nabla^2_\theta \log p(y_m|x_m, \theta'_M) - \nabla^2_\theta \log p(y_m|x_m, \hat{\theta}_M)\|_{\text{OP}}\right] \\
&= \|\tilde{\theta}^*_M - \hat{\theta}_M\| \left[\sum_{m=1}^M \|\theta'_M - \hat{\theta}_M\| \cdot \| \int_{t=0}^1 \frac{\partial}{\partial t} \nabla^2_\theta \log p(y_m|x_m, \theta)\big|_{\theta = t\hat{\theta}_M + (1-t)\theta'_M} \|_{\text{OP}}\right] \\
&\quad // \text{ By the fundamental theorem of calculus} \\
&\leq \|\tilde{\theta}^*_M - \hat{\theta}_M\|^2 \left[\sum_{m=1}^M \| \int_{t=0}^1 \frac{\partial}{\partial t} \nabla^2_\theta \log p(y_m|x_m, \theta)\big|_{\theta = t\hat{\theta}_M + (1-t)\theta'_M} \|_{\text{OP}}\right] \\
&\leq \|\tilde{\theta}^*_M - \hat{\theta}_M\| \left[\sum_{m=1}^M \|x_m\|^3 (\max_a \phi'''(a))\right] \\
&= \frac{1}{6\sqrt{3}} \|\tilde{\theta}^*_M - \hat{\theta}_M\|^2 \left[\sum_{m=1}^M \|x_m\|^3\right] \\
&\leq O_p(\frac{1}{M^2}) O_p(M) = O_p(\frac{1}{M}),
\end{aligned}
$$

where the final line requires that the covariates have bounded third moment. □

Supporting Lemmas

Lemma D.8.1 (Trippe et al., 2019, Lemma E.1)**.** *Let f, g be twice differentiable functions mapping $\mathbb{R}^N \to \mathbb{R}$ and attaining minima at $\theta_f = \arg\min_\theta f(\theta)$ and $\theta_g = \arg\min_\theta g(\theta)$, respectively. Additionally, assume that f is α–strongly convex for*

some $\alpha > 0$ on the set $\{t\theta_f + (1-t)\theta_g | t \in [0,1]\}$ and that $\|\nabla_\theta f(\theta_g) - \nabla_\theta g(\theta_g)\|_2 = \|\nabla_\theta f(\theta_g)\|_2 \leq c$. Then

$$\|\theta_f - \theta_g\|_2 \leq \frac{c}{\alpha}. \tag{D.8.3}$$

Lemma D.8.2 (uniform law of large numbers)**.** *Let $H_M(\theta)$ be as defined in Eq.* (D.8.1) *and define $H(\theta) := \mathbb{E}[\nabla_\theta^2 \log p(y_1|x_1;\theta)]$, where the expectation is taken under the true θ. If $\mathbb{E}[x_1 x_1^\top]$ exists and is positive definite then*

$$\sup_{\theta' \in B_\epsilon(\theta)} \|\frac{1}{M} H_M(\theta') - H(\theta')\|_2 \overset{a.s.}{\to} 0.$$

according to p, where $B_\epsilon(\theta)$ is a closed neighborhood of θ of radius ϵ, for any $\epsilon > 0$.

Proof. Since the each of the M data points $\{(x_m, y_m)\}_{m=1}^\infty$ are i.i.d. by assumption, $M^{-1}H_M$ converges point-wise by the law of large numbers. However, we are additionally interested in uniform convergence; a number of different uniform laws of large numbers suffice for this. Because H is continuously differentiable in θ (recall that for any x_m, $\frac{d^3}{d\theta^3} \log p(y_m|x_m,\theta)$ is bounded) it is therefore Lipschitz continuous on the bounded set $B_\epsilon(\theta)$. As such one can construct a bounded envelope for H on this set, which amounts to a sufficient condition for uniform convergence on B_ϵ, see Van der Vaart [2000, Theorem 19.4 - Glivenko-Cantelli]. We refer the reader to Van der Vaart [2000, Chapter 19] for technical background, and in particular to Van der Vaart [2000, Example 19.8] which walks through an example closely related to the present case. □

Lemma D.8.3. *Consider logistic regression with random covariates, $x_1, x_2, \ldots$. Let $B_\epsilon(\theta)$ be a closed neighborhood of radius $\epsilon > 0$ around θ and for each M define*

$$\alpha_M := \inf_{\theta' \in B_\epsilon(\theta)} \lambda_{min} \left[\nabla_\theta^2 \log p_M(\theta')\right]$$

to be the constant of strong log-concavity constant of $\log p_M(\cdot)$ on $B_\epsilon(\theta)$, where $\lambda_{min}(\cdot)$ denotes the smallest eigenvalue of its matrix argument. If the covariates are i.i.d. and satisfy $\mathbb{E}[x_1 x_1^\top] \succ 0$, then α_M is $\Omega_p(M)$.

Proof. Consider the scaled Hessians of $\log p_M(\cdot)$, $M^{-1}H_M(\cdot)$. By Lemma D.8.2, $M^{-1}H_M(\cdot)$ converges uniformly to its expectation, $H(\theta) := \mathbb{E}[\nabla_\theta^2 \log p(y_1|x_1,\theta)]$ on $B_\epsilon(\theta)$. Since $H(\theta) \succ 0$ on $B_\epsilon(\theta)$, we have that

$$\inf_{\theta' \in B_\epsilon(\theta)} \lambda_{min}(\frac{1}{M} H_M(\theta)) \overset{a.s.}{\to} \inf_{\theta' \in B_\epsilon(\theta)} \lambda_{min}(H_M(\theta)) > 0.$$

Therefore $\alpha_M := \inf_{\theta' \in B_\epsilon(\theta)} \lambda_{min}\left[\nabla_\theta^2 \log p_M(\theta')\right]$ is $\Omega_p(M)$. □

Lemma D.8.4. *Let $\hat{\theta}$ and $\tilde{\theta}^*$ be the MLE and the approximation to the MAP defined in Eq.* (5.19)*, respectively. If the covariates, $x_1, x_2, \ldots$ are i.i.d. and satisfy $\mathbb{E}[x_1 x_1^\top] \succ 0$, then $\|\hat{\theta}_M - \tilde{\theta}^*_M\|$ is $O_p(M^{-1})$.*

Proof. Recall that

$$\tilde{\theta}^*_M = \left[I_N + \tilde{\Sigma}_M\right]^{-1} \hat{\theta}_M,$$

where $\tilde{\Sigma}_M := H_M(\hat{\theta}_M)^{-1}$. Lemma D.8.3 provides that the constants of strong log-concavity for $\log p_M$ grow as $\Omega_p(M)$ in a neighborhood of θ. Therefore, since $\hat{\theta}_M$ converges strongly to θ, we can see that $\lambda_{min}(H_M(\hat{\theta}_M))$ is $\Omega_p(M)$. Next, we rewrite

$$\begin{aligned}
\|\tilde{\theta}^*_M - \hat{\theta}_M\| &= \| \left[I_N + \tilde{\Sigma}_M\right]^{-1} \hat{\theta}_M - \hat{\theta}_M\| \\
&= \| \left[I_N + H_M(\hat{\theta}_M)\right]^{-1} \hat{\theta}_M\| \\
&\leq \| \left[I_N + H_M(\hat{\theta}_M)\right]^{-1} \|_{\text{OP}} \|\hat{\theta}_M\| \\
&\leq \frac{\|\hat{\theta}_M\|}{\lambda_{min}\left(H_M(\hat{\theta}_M)\right)}.
\end{aligned}$$

which one can see is $O_p(M^{-1})$ since $\|\hat{\theta}_M\|$ is bounded in probability. □

D.8.3 Proof of Theorem 5.5.3

Before proving the theorem we begin by explicitly writing out the win and our proposed bound defined in Section 5.5.2. For clarity, we introduce a subscript M to index the size of the dataset on which these quantities are computed. Specifically, recalling that in this case we have $A = I_N$ and $C = (I_N + \tilde{\Sigma}_M)^{-1}$, and noting that therefore

$A - C = (I_N + \tilde{\Sigma}_M^{-1})^{-1}$, we have

$$b_M(\alpha) = 2\mathrm{tr}[(I_N + \tilde{\Sigma}_M^{-1})^{-1}\tilde{\Sigma}_M] + 2z_{\frac{1-\alpha}{2}}\sqrt{U_M(\|G_M(\hat{\theta}_M)\|^2_{\tilde{\Sigma}_M}, \frac{1-\alpha}{2}) + 2\|\tilde{\Sigma}_M^{\frac{1}{2}}(I_N + \tilde{\Sigma}_M^{-1})^{-1}\tilde{\Sigma}_M^{\frac{1}{2}}\|_F^2} - \|\tilde{\theta}_M^* - \hat{\theta}_M\|^2$$

where $G_M(\hat{\theta}_M) := (I_N + \tilde{\Sigma}_M^{-1})^{-1}\hat{\theta}_M$ and

$$U_M(\|G_M(\hat{\theta}_M)\|^2_{\tilde{\Sigma}_M}, 1-\alpha) := \inf_{\delta>0}\Bigg\{\delta \,\Bigg|\, \|G_M(\hat{\theta}_M)\|^2_{\tilde{\Sigma}_M} \leq (\delta + \|\tilde{\Sigma}_M^{\frac{1}{2}}(I_N + \tilde{\Sigma}_M^{-1})^{-1}\tilde{\Sigma}_M^{\frac{1}{2}}\|_F^2) + \tag{D.8.4}$$

$$z_{1-\alpha}\sqrt{2\|\tilde{\Sigma}_M^{\frac{1}{2}}(I_N + \tilde{\Sigma}_M^{-1})^{-1}\tilde{\Sigma}_M(I_N + \tilde{\Sigma}_M^{-1})^{-1}\tilde{\Sigma}_M^{\frac{1}{2}}\|_F^2 + 4\|\tilde{\Sigma}_M^{\frac{1}{2}}(I_N + \tilde{\Sigma}_M^{-1})^{-1}\tilde{\Sigma}_M^{\frac{1}{2}}\|^2_{\mathrm{OP}}\delta}\Bigg\} \tag{D.8.5}$$

is an approximate high-confidence upper bound on $\|G_M(\hat{\theta}_M)\|^2_{\tilde{\Sigma}_M}$. For convenience, we abbreviate $U_M(\|G_M(\hat{\theta}_M)\|^2_{\tilde{\Sigma}_M}, 1-\alpha)$ by U_M.

Next, we recall that we may decompose the win in squared error loss for using θ_M^* in place of $\hat{\theta}_M$ as

$$W_M(\theta) = 2\epsilon_M^\top(I_N + \tilde{\Sigma}_M^{-1})^{-1}\hat{\theta} - \|\theta_M^* - \hat{\theta}_M\|^2,$$

where $\epsilon_M := \hat{\theta}_M - \theta$.

Proof. Proving the theorem amounts to showing that for any θ and $\alpha \in (0,1)$,

$$\lim_{M\to\infty} \mathbb{P}_\theta\left[W_M(\theta) \geq b_M(\alpha)\right] \geq \alpha.$$

Lemma D.8.6 provides that $M^{1.5}(W_M(\theta) - b_M(\alpha))$ converges in distribution to $2\sqrt{\theta^\top H(\theta)^{-3}\theta}(\delta - z_{\frac{1-\alpha}{2}})$, for $\delta \sim \mathcal{N}(0,1)$. Thus for any θ, $\mathbb{P}_\theta\left[W_M(\theta) - b_M(\alpha) > 0\right] \to (1 - \Phi(z_{\frac{1-\alpha}{1}})) = 1 - \frac{1-\alpha}{2} > \alpha$. This establishes that $b_M(\cdot)$ has above nominal coverage asymptotically, as desired. □

Lemma D.8.5. $|U_M - \|\tilde{\Sigma}_M\theta\|^2_{\tilde{\Sigma}_M}|$ *is* $O_p(M^{-3.5})$.

Proof. Recall that we can rearrange Eq. (D.8.4) to see that U_M satisfies

$$\|(I+\tilde{\Sigma}_M^{-1})^{-1}\hat{\theta}_M\|_{\tilde{\Sigma}_M}^2 = U_M + 2\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_F^2 + \sqrt{\|\tilde{\Sigma}_M^4(I_N+\tilde{\Sigma}_M)^2\|_F^2 + 4\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_{\mathrm{OP}}^2 U_M}$$

where we have simplified $\tilde{\Sigma}_M^{\frac{1}{2}}(I_N+\tilde{\Sigma}_M^{-1})^{-1}\tilde{\Sigma}_M^{\frac{1}{2}}$ to $\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}$.

We next further simplify the condition above by replacing two quantities with simplifying approximations plus lower order terms. First note that we may write

$$\begin{aligned}
\|(I+\tilde{\Sigma}_M^{-1})^{-1}\hat{\theta}_M\|_{\tilde{\Sigma}_M}^2 &= \|\tilde{\Sigma}_M\hat{\theta}_M - \tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\hat{\theta}_M\|_{\tilde{\Sigma}_M}^2 \\
&= \|\tilde{\Sigma}_M\hat{\theta}_M\|_{\tilde{\Sigma}_M}^2 + \|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\hat{\theta}_M\|_{\tilde{\Sigma}_M}^2 - 2\hat{\theta}_M^\top\tilde{\Sigma}_M^4(I_N+\tilde{\Sigma}_M)^{-1}\hat{\theta}_M \\
&= \|\tilde{\Sigma}_M(\theta+\epsilon_M)\|_{\tilde{\Sigma}_M}^2 + O_p(M^{-4}) \\
&= \|\tilde{\Sigma}_M\theta\|_{\tilde{\Sigma}_M}^2 + \|\tilde{\Sigma}_M\epsilon_M\|_{\tilde{\Sigma}_M}^2 + 2\epsilon_M^\top\tilde{\Sigma}_M^3\theta + O_p(M^{-4}) \\
&= \|\tilde{\Sigma}_M\theta\|_{\tilde{\Sigma}_M}^2 + O_p(M^{-3.5}).
\end{aligned}$$

Second, we write

$$\begin{aligned}
\sqrt{\|\tilde{\Sigma}_M^4(I_N+\tilde{\Sigma}_M)^2\|_F^2 + 4\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_{\mathrm{OP}}^2 U_M} &= \sqrt{O_p(M^{-8}) + 4\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_{\mathrm{OP}}^2 U_M} \\
&= 2\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_{\mathrm{OP}}\sqrt{U_M} + O_p(M^{-4}).
\end{aligned}$$

As such, we may see that U_M satisfies

$$\begin{aligned}
\|\tilde{\Sigma}_M\theta_M\|_{\tilde{\Sigma}_M}^2 - U_M &= 2\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_F^2 + 2\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_{\mathrm{OP}}\sqrt{U_M} + O_p(M^{-3.5}) \\
&= 2\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_{\mathrm{OP}}\sqrt{U_M} + O_p(M^{-3.5})
\end{aligned} \tag{D.8.6}$$

where we have dropped $2\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_F^2$ since it is $O_p(M^{-4})$.

We next observe that U_M must be $O_p(M^{-3})$. Otherwise, the event that $\|\tilde{\Sigma}_M\theta_M\|_{\tilde{\Sigma}_M}^2 - U_M < 0$ must occur infinitely often (since $\|\tilde{\Sigma}_M\theta\|_{\tilde{\Sigma}_M}^2$ is $O_p(M^{-3})$); in turn, this condition would imply that $\|\tilde{\Sigma}_M^2(I_N+\tilde{\Sigma}_M)^{-1}\|_{\mathrm{OP}}\sqrt{U_M} < 0$ occurs infinitely often, which

provides a contradiction.

Finally, in tangent with Eq. (D.8.6), that U_M is $O_p(M^{-3})$ allows us to see that $\left|U_M - \|\Sigma\theta\|^2_{\tilde{\Sigma}_M}\right|$ is $O_p(M^{-3.5})$, as desired. □

Lemma D.8.6. *Let $\alpha \in (0,1)$ and $\theta \in \mathbb{R}^N$. Consider the sequence of wins, $W_M(\theta)$, and bounds, $b_M(\alpha)$, computed for logistic regression. Then*

$$M^{1.5}(W_M(\theta) - b_M(\alpha)) \xrightarrow{d} 2\sqrt{\theta^\top H(\theta)^{-3}\theta}(\delta - z_{\frac{1-\alpha}{2}}),$$

where $\delta \sim \mathcal{N}(0,1)$.

Proof. We prove the lemma by first writing W_M and b_M using simplifying approximations and lower order terms. The result is obtained by manipulating a scaling of the difference between the two expressions and considering the limit in M.

Note first that we may write

$$\begin{aligned}
W_M(\theta) :&= 2\epsilon^\top(\theta^*_M - \hat{\theta}_M) - \|\theta^*_M - \hat{\theta}_M\|^2 \\
&= 2\epsilon^\top(\tilde{\theta}^*_M - \hat{\theta}_M) - \|\tilde{\theta}^*_M - \hat{\theta}_M\|^2 + O_p(M^{-2}) \\
&= 2\epsilon^\top(I_N + \tilde{\Sigma}_M^{-1})^{-1}\hat{\theta}_M - \|\tilde{\theta}^*_M - \hat{\theta}_M\|^2 + O_p(M^{-2}) \\
&= 2\epsilon^\top\tilde{\Sigma}_M\hat{\theta}_M - \|\tilde{\theta}^*_M - \hat{\theta}_M\|^2 + O_p(M^{-2}) \\
&= 2\epsilon^\top\tilde{\Sigma}_M\theta - \|\tilde{\theta}^*_M - \hat{\theta}_M\|^2 + O_p(M^{-2}).
\end{aligned}$$

Next we write

$$\begin{aligned}
b_M(\alpha) &= 2\mathrm{tr}\left[(I_N + \tilde{\Sigma}_M^{-1})^{-1}\tilde{\Sigma}_M\right] + 2z_{\frac{1-\alpha}{2}}\sqrt{U_M + 2\|\tilde{\Sigma}^2_M(I_N + \tilde{\Sigma}_M)^{-1}\|^2_F} - \|\tilde{\theta}^*_M - \hat{\theta}_M\|^2 \\
&= 2z_{\frac{1-\alpha}{2}}\sqrt{\|\tilde{\Sigma}_M\theta\|^2_{\tilde{\Sigma}_M} + O_p(M^{-3.5})} - \|\tilde{\theta}^*_M - \hat{\theta}_M\|^2 + O_p(M^{-2}) \\
&= 2z_{\frac{1-\alpha}{2}}\|\tilde{\Sigma}_M\theta\|_{\tilde{\Sigma}_M} - \|\tilde{\theta}^*_M - \hat{\theta}_M\|^2 + O_p(M^{-2}).
\end{aligned}$$

where the second line uses Lemma D.8.5.

By considering a scaled difference between these two terms we find,

$$M^{1.5}(W_M(\theta) - b_M(\alpha)) = 2M^{1.5}\epsilon^\top\tilde{\Sigma}_M\theta - 2M^{1.5}z_{\frac{1-\alpha}{2}}\|\tilde{\Sigma}_M\theta\|_{\tilde{\Sigma}_M} + O_p(M^{-\frac{1}{2}})$$

$$\xrightarrow{d} 2M^{1.5}\|\tilde{\Sigma}_M\theta\|_{\tilde{\Sigma}_M}(\delta - z_{\frac{1-\alpha}{2}})$$

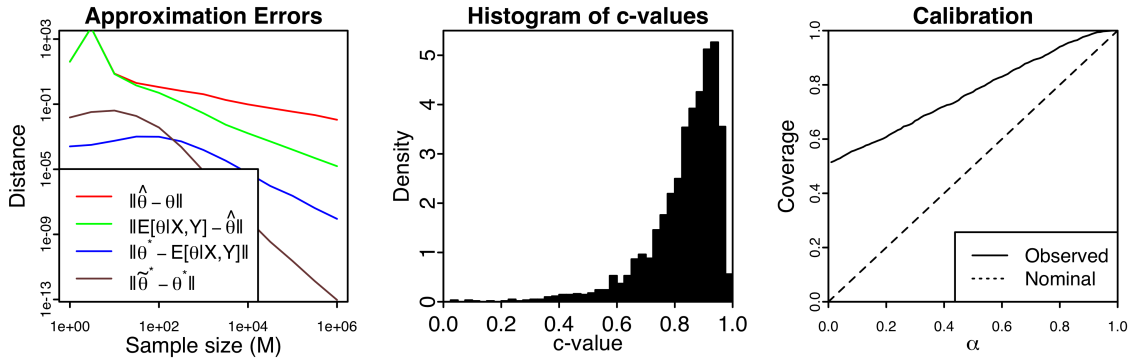
for $\delta \sim \mathcal{N}(0,1)$, by recognizing that ϵ_M is asymptotically normal with mean zero and covariance Σ_M, and therefore that $2\epsilon^\top\tilde{\Sigma}_M\theta$ is asymptotically normal with variance $\|\tilde{\Sigma}_M\theta\|^2_{\tilde{\Sigma}_M}$.

Finally, the result obtains by noting that Lemma D.8.2 implies that

$$M^{1.5}\|\tilde{\Sigma}_M\theta\|_{\tilde{\Sigma}_M} = \sqrt{\theta^\top(H_M(\hat{\theta})/M)^{-3}\theta} \xrightarrow{a.s.} \sqrt{\theta^\top H(\theta)^{-3}\theta}.$$

□

D.8.4 Empirical validation of logistic regression bound in simulation



(a) (b) (c)

Figure D.8.1: c-values for logistic regression in simulation. (a) Empirical rates of convergence of distances amongst various estimates and the true parameter with $N = 2$. In simulation with $N = 25$ and $M = 1000$ (b) c-values are able to detect improvements, sometimes with high confidence (c) the approximate bound has greater than nominal coverage. See Appendix D.8.4 for details.

We here explore the behaviour of our proposed approximation, bound and the associated c-values empirically on simulated data. Figure D.8.1a shows the distance between various estimates and the true parameter for a range of sample sizes in simulation. Due to the log-log scale, the slopes of the series in this plot reflect the polynomial

rates of convergence. Notably we see the fast $O_p(M^{-2})$ rate of convergence of our approximation to the MAP estimate, $\tilde{\theta}^*_M$, to the exact MAP estimate, θ^*_M.

Figure D.8.1b demonstrates that our approach is able to detect improvements (i.e. we can obtain high c-values). Furthermore, our proposed bound has similar coverage properties as in the Gaussian case (Figure D.8.1c). In the experiments for Figures D.8.1b and D.8.1c, we simulated the parameter as $\theta \sim \mathcal{N}(0, \frac{1}{2}I_N)$ and, in each replicate, simulated the covariates for each data point, m, as $x_m \overset{i.i.d.}{\sim} \mathcal{N}(0, N^{-2}I_N)$.

Two of the series in Figure D.8.1a are distances between the posterior mean of θ and other estimates, $\mathbb{E}[\theta|X, Y] = \int p(\theta|X, Y)\theta d\theta$. Because this model is non-conjugate, the estimate does not have an analytic form. As such approximated these quantities with Gauss-Hermite quadrature. For each sample size M, we performed 25 replicate simulations.

In the experiments that went into Figures D.8.1b and D.8.1c, we used $N = 25$ and $M = 1000$. See `logistic_regression_approximations.ipynb` and `logistic_regression_c_values_and_operating_characteristics.ipynb` for details.

D.9 Additional details on applications

In this section, we provide additional details associated with the applications in **??**.

D.9.1 Estimation from educational testing data

Conservatism of c-values with the empirical Bayes step. The application in Section 5.6.1 diverges from the scenarios covered by our theory in Sections 5.3 and 5.4 in its use of the empirical Bayes step to estimate β, τ, and σ. As a result, our theory does provide that $c(y)$ satisfies the guarantee of Theorem 5.2.2. However, given the favorable asymptotic and empirical properties of the empirical Bayes procedure established in Section 5.5.1, we conjectured that the looseness in the lower bound $b(y, \alpha)$ would be sufficiently large to compensate for any error introduced by these departures from the assumptions of our theory. To investigate this, we performed

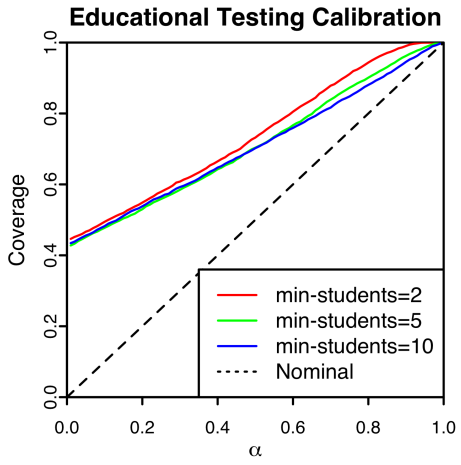


Figure D.9.1: Calibration of the lower bounds $b(y, \alpha)$ in small area inference with an empirical Bayes step (5000 replicates). The coverage on the y-axis is a Monte Carlo estimate of $\mathbb{P}_\theta \left[W(\theta, y) \geq b(y, \alpha) \right]$. Each series corresponds to a set of simulations within which we excluded a different subset of schools based on a minimum number of students tested.

a simulation study in which we used this empirical Bayes step and confirmed that the c-values retained at least nominal coverage (Figure D.9.1). To ensure that the simulated data had similar characteristics to the real data, we simulated 5000 datasets by drawing hypothetical school level means according the assumed generative model with the parameters (β, τ and σ) fit on the real dataset. In each simulation, we re-estimated the fixed effects and variances (again using `lme4`), and computed the associated MLE, Bayes estimates, and bounds across a range of confidence levels. We then computed the empirical coverage of these bounds and found them to be conservative across all tested levels.

Additional preprocessing and calibration details. Hoff [2021] considered only schools at which 2 or more students took the reading test. We excluded an additional 8 schools with fewer than 5 students tested because we expected that the high variance in these observations could introduce too much slack into our bound as result of the poor conditioning of $\Sigma^{\frac{1}{2}}(A - C)\Sigma^{\frac{1}{2}}$ (recall the operator norm bound in Eq. (5.11), derived in Eq. (D.6.2)). Consistent with this hypothesis we computed a c-value of 0.88 when we included these additional schools, and when we further restricted to

the 657 schools with at least 10 students tested we computed a c-value 0.999992. To further validate this hypothesis of increased conservatism we simulated additional datasets with these different thresholds on school size and evaluated the calibration of computed bounds (Figure D.9.1). We observed the coverage for the simulations with smallest threshold was noticeably higher at large α, in agreement with this hypothesis.

D.9.2 Estimation of violent crime rates in Philadelphia

Dependence on the order in which estimates are compared. In Section 5.6.2 we chose to report one among three estimates as described in Remark 5.6.1. We note however that this paradigm is sensitive to the order in which the different estimates are considered. For this set of three models, if we had first compared $\theta^\circ(y)$ as the alternative to $\hat{\theta}(y)$ as the default we would have rejected $\hat{\theta}(y)$ (with $c = 0.99942$), and then again sided against updating our estimate a second time with a low c-value ($c = 0.0$) for comparing $\theta^*(y)$ as the alternative against $\theta^\circ(y)$ as the default. The potential cost of ending up with a worse estimate as a result of considering these estimates in sequence may be understood as a cost of looking at the data an additional time.

Selection of prior parameters from historical data. The parameters $\sigma_\delta^2, \sigma_z^2, \sigma_y^2$ were selected based on historical data. Specifically, we estimated σ_y^2 and σ_z^2 as the averages of the sample variances of the violent and non-violent report rates, respectively, computed within each census block in the preceding years. For the first model described in Section 5.6.2, we then estimated σ_δ^2 using these same historical data to reflect the prior belief that half of the variability across the unknown rates is common across the two response types.

For the second model considered, we selected the signal variance and length scale of this covariance function by drawing hypothetical datasets of crime levels from the prior predictive distributions and selecting those which produced the most reasonable looking patterns. In particular, we chose the length scale to be one sixth of the maximum distance between the centroids of census blocks, and the signal variance

to reflect the prior belief that one third of the variability in the unknown rates was explained by the spatial component. In addition, we choose a smaller value for σ_δ^2 in this second model, so that the total implied variance would be the same. See supplementary code in `Philly_reported_crime_estimation.ipynb` for additional details.

Derivation of θ^* (posterior mean in the first model). As mentioned in the main text, since the prior and likelihoods for this model are independent across each census block we can compute the posterior mean for each block independently.

Let $\pi(\cdot)$ denote the joint density of all variables. Then, since $z_n \perp\!\!\!\perp y_n \big| \theta_n$, we have that

$$\begin{aligned}\pi(\theta_n|y_n, z_n) &\propto \pi(\theta_n|z_n)\pi(y_n|\theta_n, z_n) \\ &= \pi(\theta_n|z_n)\pi(y_n|\theta_n).\end{aligned}$$

Next observe that by construction, $z_n - \theta_n = \epsilon_n^z + \delta_n^z - \delta_n^y \sim \mathcal{N}(0, 2\sigma_\delta^2 + \sigma_z^2)$ and so $\theta_n|z_n \sim \mathcal{N}(z_n, 2\sigma_\delta^2 + \sigma_z^2)$. Since again by construction we have that $y_n|\theta_n \sim \mathcal{N}(\theta_n, \sigma_y^2)$, Gaussian conjugacy provides that

$$\theta_n|y_n, z_n \sim \mathcal{N}(\mathbb{E}[\theta_n|y_n, z_n], \mathrm{Var}[\theta_n|y_n, z_n]),$$

where

$$\begin{aligned}\mathrm{Var}[\theta_n|y_n, z_n] &= \frac{1}{\sigma_y^{-2} + (2\sigma_\delta^2 + \sigma_z^2)^{-1}} \\ &= \frac{\sigma_y^2(2\sigma_\delta^2 + \sigma_z^2)}{\sigma_y^2 + 2\sigma_\delta^2 + \sigma_z^2}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[\theta_n|y_n, z_n] &= \mathrm{Var}[\theta_n|y_n, z_n](\mathrm{Var}[\theta_n|z_n]^{-1}\mathbb{E}[\theta_n|z_n] + \mathrm{Var}[y_n|\theta_n]^{-1}y_n) \\ &= \frac{\sigma_y^2(2\sigma_\delta^2 + \sigma_z^2)}{\sigma_y^2 + 2\sigma_\delta^2 + \sigma_z^2}\left[(2\sigma_\delta^2 + \sigma_z^2)^{-1}z_n + \sigma^{-2}y_n\right]\end{aligned}$$

$$= \frac{2\sigma_\delta^2 + \sigma_z^2}{2\sigma_\delta^2 + \sigma_y^2 + \sigma_z^2} y_n + \frac{\sigma_y^2}{2\sigma_\delta^2 + \sigma_y^2 + \sigma_z^2} z_n$$

as desired.

Analogously, for the second model considered in Section 5.6.2 we find the posterior mean as

$$\theta^\circ(y) = \left[I_N + \sigma_y^2 (2K + 2\sigma_\delta^2 I_N + \sigma_z^2 I_N)^{-1} \right]^{-1} y + \left[I_N + \sigma_y^{-2} (2K + 2\sigma_\delta^2 I_N + \sigma_z^2 I_N) \right]^{-1} z.$$

Additional dataset details. The data considered in this application are counts of police responses categorized as associated with violent crimes and violent crimes in October 2018. These were obtained from opendataphilly.org. The observed data we model are the inverse hyperbolic sine transform of the number of recorded police responses per square mile. For all practical purposes, these values can be interpreted as log densities (see, e.g., Burbidge et al. [1988]).

D.9.3 Gaussian process kernel selection for estimation of ocean currents

We here provide additional details of the Gaussian process covariance functions used in Section 5.6.3. The first covariance function described, which incorporated covariation at two scale is defined, for both the longitudinal and latitudinal components (i in $\{1, 2\}$) and for each pair of buoys n and n', as

$$\begin{aligned} k(\theta_n^{(i)}, \theta_{n'}^{(i)}) =& \sigma_1^2 \exp\left\{ -\frac{1}{2} \left[\frac{(\text{lat}_n - \text{lat}_{n'})^2}{r_{1,\text{lat}}^2} + \frac{(\text{lon}_n - \text{lon}_{n'})^2}{r_{1,\text{lon}}^2} + \frac{(t_n - t_{n'})^2}{r_{1,t}^2} \right] \right\} \\ &+ \sigma_2^2 \exp\left\{ -\frac{1}{2} \left[\frac{(\text{lat}_n - \text{lat}_{n'})^2}{r_{2,\text{lat}}^2} + \frac{(\text{lon}_n - \text{lon}_{n'})^2}{r_{2,\text{lon}}^2} + \frac{(t_n - t_{n'})^2}{r_{2,t}^2} \right] \right\}, \end{aligned}$$

where $\sigma_1^2, r_{1,\text{lat}}, r_{1,\text{lon}}$ and $r_{1,t}$ parameterize the mesoscale variation in currents whereas $\sigma_2^2, r_{2,\text{lat}}, r_{2,\text{lon}}$ and $r_{2,t}$ parameterize the submesoscale variation. As in Lodise et al. [2020], the latitudinal and longitudinal components of F are modeled as a priori

independent. We choose these parameters by maximal marginal likelihood [Rasmussen and Williams, 2006, Chapter 5] on an independent subset of the GLAD dataset. Estimates of the underlying currents are obtained as the posterior mean of F under this model, which we take as the alternative, $\theta^*(y)$.

The second covariance function captures covariation among observations only at the mesoscale. In this case, the Gaussian process prior has covariance function

$$k(\theta_n^{(i)}, \theta_{n'}^{(i)}) = \sigma_1^2 \exp\left\{ -\frac{1}{2}\left[\frac{(\text{lat}_n - \text{lat}_{n'})^2}{r_{1,\text{lat}}^2} + \frac{(\text{lon}_n - \text{lon}_{n'})^2}{r_{1,\text{lon}}^2} + \frac{(t_n - t_{n'})^2}{r_{1,t}^2} \right] \right\} + \sigma_2^2 \mathbb{1}[n = n'],$$

which maintains the same marginal variance but excludes submesoscale covariances. We take the posterior mean under this model as the default estimate $\hat{\theta}(y)$. See `submesoscale_GP_c_value.ipynb` for further implementation details.

Bibliography

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Matthieu Dean, Jeffrey an Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016.

Massih Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views – an application to multilingual text categorization. In *Advances in Neural Information Processing Systems*, 2009.

Tomohiro Ando and Arnold Zellner. Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equations models using a combination of direct Monte Carlo and importance sampling techniques. *Bayesian Analysis*, 5(1), 2010.

Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6), 1974.

Richard Arratia, Andrew D. Barbour, and Simon Tavaré. *Logarithmic Combinatorial Structures: a Probabilistic Approach*, volume 1. European Mathematical Society, 2003.

Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of Experimental Medicine*, 79(2), 1944.

Cecilia Balocchi and Shane T Jensen. Spatial modeling of trends in crime over time in Philadelphia. *The Annals of Applied Statistics*, 13(4), 2019.

Cecilia Balocchi, Sameer K Deshpande, Edward I George, and Shane T Jensen. 'Crime in Philadelphia: Bayesian clustering with particle optimization'. arXiv:1912.00111, 2019.

Alvin J Baranchik. Multiple regression and estimation of the mean of a multivariate normal distribution. Technical report, Stanford University, 1964.

R'emi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1), 2017.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 2015a.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 2015b.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2), 2013.

Andrew C Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1), 1941.

Anindya Bhadra and Bani K Mallick. Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69(2), 2013.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Robert C Blattberg and Edward I George. Shrinkage estimation of price and promotional elasticities: Seemingly unrelated equations. *Journal of the American Statistical Association*, 86(414), 1991.

David M Blei. Build, compute, critique, repeat: data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1(1), 2014.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 2017.

Franccois Bolley, Ivan Gentil, and Arnaud Guillin. Convergence to equilibrium in Wasserstein distance for Fokker–Planck equations. *Journal of Functional Analysis*, 263(8), 2012.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, 2011.

George EP Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A*, 143(4), 1980.

George EP Box and William G Hunter. A useful method for model-building. *Technometrics*, 4(3), 1962.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Leo Breiman and Jerome H Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B*, 59(1), 1997.

Philip J Brown and James V Zidek. Adaptive multivariate ridge regression. *The Annals of Statistics*, 8(1), 1980.

Philip J Brown, Marina Vannucci, and Tom Fearn. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B*, 60(3), 1998.

S L Buka, T L Stichick, I Birdthistle, and FJ Earls. Youth exposure to violence: prevalance, risks, and consequences. *American Journal of Orthopsychiatry*, 71(3), 2001.

Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, Mark J Daly, Alkes L Price, and Benjamin M Neal. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11), 2015.

John B Burbidge, Lonnie Magee, and A Leslie Robb. Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83(401), 1988.

Krisztian Buza. Feedback prediction for blogs. In *Data Analysis, Machine Learning and Knowledge Discovery*. Springer, 2014.

Diana Cai, Rishit Sheth, Lester Mackey, and Nicolo Fusi. Weighted meta-learning. *arXiv preprint arXiv:2003.09465*, 2020.

T Tony Cai, Cun-Hui Zhang, and Harrison H Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4), 2010.

Trevor Campbell and Tamara Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 2018.

Trevor Campbell and Tamara Broderick. Automated scalable Bayesian inference via Hilbert coresets. *Journal of Machine Learning Research*, 20(1), 2019.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming panguage. *Journal of Statistical Software*, 76(1), 2017.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, 2009.

George Casella and Roger L Berger. *Statistical Inference*. Duxbury Pacific Grove, CA, 2002.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 2011.

Sitan Chen, Michelle Delcourt, Ankur Moitra, Guillem Perarnau, and Luke Postle. Improved bounds for randomly sampling colorings via linear programming. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2019.

Siddhartha Chib and Edward Greenberg. Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models. *Journal of Econometrics*, 68(2), 1995.

Gerda Claeskens and Nils Lid Hjort. The focused information criterion. *Journal of the American Statistical Association*, 98(464), 2003.

Chandler Davis and William M Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1), 1970.

A Philip Dawid. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1), 1981.

A Phillip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379), 1982.

Daryl DeFord, Moon Duchin, and Justin Solomon. Recombination: a family of Markov chains for redistricting. *Harvard Data Science Review*, 3 2021. https://hdsr.mitpress.mit.edu/pub/1ds8ptxu.

Sameer K Deshpande, Veronika Ročková, and Edward I George. Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, 2019.

Dua Dheeru and Efi Karra Taniskidou. UCI Machine Learning Repository, 2017.

Bradley Efron and Carl Morris. Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika*, 59(2), 1972a.

Bradley Efron and Carl Morris. Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67(337), 1972b.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors — an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341), 1973.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 2001.

Robert E Fay and Roger A Herriot. Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a), 1979.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*. PMLR, 2017.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie TH Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78), 2021.

Alan E Gelfand, Susan E Hills, Amy Racine-Poon, and Adrian FM Smith. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85(412), 1990.

Andrew Gelman and Jennifer Hill. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

Junxian Geng, Anirban Bhattacharya, and Debdeep Pati. Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526), 2019.

Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423), 1993.

Leo N Geppert, Katja Ickstadt, Alexander Munteanu, Jens Quedenfeld, and Christian Sohler. Random projections for Bayesian regression. *Statistics and Computing*, 27 (1), 2017.

S Ghosal, J K Ghosh, and R V Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27(1), 1999.

J K Ghosh and R V Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics, 2003.

Alison L Gibbs. Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stochastic Models*, 20(4), 2004.

Peter W. Glynn and Chang-han Rhee. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A), 2014.

Amos Golan and Jeffrey M Perloff. Comparison of maximum entropy and higher-order entropy estimators. *Journal of Econometrics*, 107(1-2), 2002.

Dilan Görür and Carl E Rasmussen. Dirichlet process Gaussian mixture models: choice of the base distribution. *Journal of Computer Science and Technology*, 25(4), 2010.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical Bayes. In *International Conference on Learning Representations*, 2018.

William E Griffiths. Bayesian inference in the seemingly unrelated regressions model. In *Computer-Aided Econometrics*. CRC Press, 2003.

Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 2011.

Rajarshi Guhaniyogi and David B Dunson. Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512), 2015.

Yoel Haitovsky. On multivariate ridge regression. *Biometrika*, 74(3), 1987.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2), 2011.

W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 1970.

Niko Hauzenberger, Florian Huber, Gary Koop, and Luca Onorante. Fast and flexible Bayesian inference in time-varying parameter regression models. *Journal of Business & Economic Statistics*, 2021.

Lanoy Nelson Hazel. The genetic basis for constructing selection indexes. *Genetics*, 28(6), 1943.

Grant Hillier and Raymond Kan. Properties of the inverse of a noncentral Wishart matrix. *Available at SSRN 3370864*, 2019.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 1970.

Peter Hoff and Chaoyu Yu. Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, 13(1), 2019.

Peter D Hoff. Smaller p-values via indirect information. *Journal of the American Statistical Association*, 0(0), 2021.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 2014.

Paul W Holland, Kathryn B Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2), 1983.

Alfred Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76(3), 1954.

Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Advances in Neural Information Processing Systems*, 2016.

Jonathan Huggins, Ryan P Adams, and Tamara Broderick. PASS-GLM: polynomial approximate sufficient statistics for scalable Bayesian GLM inference. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Huggins, Mikołaj Kasprzak, Trevor Campbell, and Tamara Broderick. Validated variational inference via practical posterior error bounds. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

Jonathan H Huggins, Trevor Campbell, Mikołaj Kasprzak, and Tamara Broderick. Practical bounds on the error of Bayesian posterior approximations: a nonasymptotic approach. *arXiv preprint arXiv:1809.09505*, 2018.

Pierre E Jacob. Couplings and Monte Carlo. Course Lecture Notes, 2020.

Pierre E Jacob, John O'Leary, and Yves F Atchadé. Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society Series B*, 82 (3), 2020.

W James and Charles Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 1961.

Ajay Jasra, Chris C Holmes, and David A Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 2005.

Ghassen Jerfel, Erin Grant, Thomas L Griffiths, and Katherine Heller. Reconciling meta-learning and continual learning with online mixtures of tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

Mark Jerrum. Mathematical foundations of the Markov chain Monte Carlo method. In *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer, 1998.

James E Johndrow, Paulo Orenstein, and Anirban Bhattacharya. Scalable MCMC for Bayes shrinkage priors. *arXiv preprint arXiv:1705.00841*, 2017.

Michael I Jordan. Bayesian nonparametric learning: Expressive priors for intelligent systems. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 11, 2010.

Damian J Kelly and Garrett M O'Neill. *The minimum cost flow problem and the network simplex solution method*. PhD thesis, Citeseer, 1991.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Michelle C Kondo, Elena Andreyeva, Eugenia C South, John M MacDonal, and Charles C Branas. Neighborhood interventions to reduce violence. *Annual Review of Public Health*, 39, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Paper, University of Toronto*, 2009.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4), 1982.

Jaeyong Lee and Hee-Seok Oh. Bayesian regression based on principal components for high-dimensional data. *Journal of Multivariate Analysis*, 117, 2013.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3), 2016.

Sang Hong Lee, Jian Yang, Michael E Goddard, Peter M Visscher, and Naomi R Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19), 2012.

Seunghak Lee, Jun Zhu, and Eric P Xing. Adaptive multi-task lasso: with application to eQTL detection. *Advances in neural information processing systems*, 2010.

Erich L Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

Guy Letac and Hélene Massam. A tutorial on non central Wishart distributions. *Technical Paper, Toulouse University*, 2004.

Alex Lewin, Habib Saadi, James E Peters, Aida Moreno-Moral, James C Lee, Kenneth GC Smith, Enrico Petretto, Leonardo Bottolo, and Sylvia Richardson. MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues. *Bioinformatics*, 32(4), 2015.

Antonio Lijoi, Igor Prünster, and Stephen G Walker. On consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association*, 100(472), 2005.

Dennis V Lindley and Adrian FM Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B*, 34(1), 1972.

Silvia Liverani, David I Hastie, Lamiae Azizi, Michail Papathomas, and Sylvia Richardson. PReMiuM: An R package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software*, 64(7), 2015.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of Statistics*, 42(2), 2014.

John Lodise, Tamay Özgökmen, Rafael C Gonçalves, Mohamed Iskandarani, Björn Lund, Jochen Horstmann, Pierre-Marie Poulain, Jody Klymak, Edward H Ryan, and Cedric Guigand. Investigating the formation of submesoscale structures along mesoscale fronts and estimating kinematic quantities using Lagrangian drifters. *Fluids*, 5(3), 2020.

David G Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1969.

David G Luenberger. *Introduction to Linear and Nonlinear Programming*. Addison-Wesley Reading, MA, 1973.

Steven N MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation*, 23(3), 1994.

David JC MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Robert Maier, Gerhard Moser, Guo-Bo Chen, Stephan Ripke, Cross-Disorder Working Group of the Psychiatric Genomics Consortium, William Coryell, James B Potash, William A Scheftner, Jianxin Shi, Myrna M Weissman, Christina M Hultman, Mikael Landén, Douglas F Levinson, Kenneth S Kendler, Jordan W Smoller, Naomi R Wray, and S Hong Lee. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *The American Journal of Human Genetics*, 96(2), 2015.

Arakaparampil M Mathai and Serge B Provost. *Quadratic Forms in Random Variables: Theory and Applications*. Dekker, 1992.

Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*, volume 382. John Wiley & Sons, 2007.

Rachael Meager. Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1), 2019.

Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5), 2007.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1953.

B G Mirkin and L B Chernyi. Measurement of the distance between distinct partitions of a finite set of objects. *Automation and Remote Control*, 5, 1970.

Carl N Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381), 1983.

Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2), 2000.

Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2(2.2), 2006.

James B Orlin. A faster strongly polynomial minimum cost flow algorithm. *Operations Research*, 41(2), 1993.

Tamay M Özgökmen. GLAD experiment CODE-style drifter trajectories (low-pass filtered, 15 minute interval records), Northern Gulf of Mexico near DeSoto Canyon, July-October 2012. Harte Research Institute, Texas A&M University-Corpus Christi., 2013. URL `https://data.gulfresearchinitiative.org/data/R1.x134.073:0004`.

Fabian Pedregosa, Ga"el Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and 'Edouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning*, 12, 2011.

Kaare Brandt Petersen and Michael Syskind Pedersen. The Matrix Cookbook. *Technical University of Denmark*, 7(15), 2008.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2), 2017.

Jim Pitman. *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer, 2006.

Andrew C Poje, Tamay M Özgökmen, Bruce L Lipphardt, Brian K Haus, Edward H Ryan, Angelique C Haza, Gregg A Jacobs, AJHM Reniers, Maria Josefina Olascoaga, Guillaume Novelli, Annalisa Griffa, Francisco J Beron-Vera, Shuyi S Chen, Emanuel Coelho, Patrick J Hogan, Albert D Jr Kirwan, Helga S Huntley, and Arthur J Mariano. Submesoscale dispersion in the vicinity of the Deepwater Horizon spill. *Proceedings of the National Academy of Sciences*, 111(35), 2014.

Sandhya Prabhakaran, Elham Azizi, Ambrose Carr, and Dana Pe'er. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, 2016.

John W Pratt. Shorter confidence intervals for the mean of a normal distribution with known variance. *The Annals of Mathematical Statistics*, 1963.

William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 2007.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 1971.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3), 2002.

Gregory C Reinsel. Mean squared error properties of empirical Bayes estimators in a multivariate random effects general linear model. *Journal of the American Statistical Association*, 80(391), 1985.

Herbert Robbins. The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, 35(1), 1964.

Haavard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by Using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2), 2009.

Daniel E Runcie, Jiayi Qu, Hao Cheng, and Lorin Crawford. MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. *BioRxiv*, 2020.

Mark J Schervish. *Theory of Statistics*. Springer Science & Business Media, 1995.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2), 2017.

Lee A Slocum, Beth M Huebner, Claire Greene, and Richard Rosenfeld. Enforcement trends in the city of St. Louis from 2007 to 2017: Exploring variability in arrests and criminal summonses over time and across communities. *Journal of Community Psychology*, 48(1), 2020.

Michael Smith and Robert Kohn. Nonparametric seemingly unrelated regression. *Journal of Econometrics*, 98(2), 2000.

Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6), 2015.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 1981.

Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PloS One*, 8(7), 2013.

Jonathan Taylor and Robert Tibshirani. Post-selection inference for penalized likelihood models. *Canadian Journal of Statistics*, 46(1), 2018.

Robin Thompson. The estimation of variance and covariance components with an application when records are subject to culling. *Biometrics*, 1973.

Xiaoying Tian. Prediction error after model search. *Annals of Statistics*, 48(2), 2020.

Ryan Tibshirani and Saharon Rosset. Excess optimism: how biased is the apparent error of an estimator tuned by SURE? *Journal of the American Statistical Association*, 114(526), 2019.

Christopher Tosh and Sanjoy Dasgupta. Lower bounds for the Gibbs sampler over mixtures of Gaussians. In *International Conference on Machine Learning*, 2014.

Brian L Trippe and Richard Turner. Overpruning in variational Bayesian neural networks. *Advances in Approximate Bayesian Inference*, 2018.

Brian L Trippe, Jonathan H Huggins, Raj Agrawal, and Tamara Broderick. 'LR-GLM: high-dimensional Bayesian inference using low-rank data approximations'. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*. PMLR, 2019.

Brian L Trippe, Sameer K Deshpande, and Tamara Broderick. Confidently comparing estimators with the c-value. *arXiv preprint arXiv:2102.09705 (under review)*, 2021a.

Brian L Trippe, Hilary K Finucane, and Tamara Broderick. For high-dimensional hierarchical models, consider exchangeability of effects across covariates instead of across datasets. *Advances in Neural Information Processing Systems*, 35, 2021b.

Brian L Trippe, Tin D Nguyen, and Tamara Broderick. Optimal transport couplings of Gibbs samplers on partitions for unbiased estimation. *Advances in Approximate Bayesian Inference*, 2021c. Trippe and Nguyen are equal contribution primary authors.

Hisayuki Tsukuma. Admissibility and minimaxity of Bayes estimators for a normal mean matrix. *Journal of Multivariate Analysis*, 99(10), 2008.

Richard E Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. *Bayesian Time Series Models*, 1(3.1), 2011.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal of Mathematical Data Science*, 2019.

A Van Der Merwe and James V Zidek. Multivariate regression analysis and canonical variates. *Canadian Journal of Statistics*, 8(1), 1980.

Aad W Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Arnaud Van Looveren, Giovanni Vacanti, Janis Klaise, and Alexandru Coca. Alibi-Detect: Algorithms for outlier and adversarial instance detection, concept drift and metrics. 2019. URL `https://github.com/SeldonIO/alibi-detect`.

Wessel N van Wieringen. Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*, 2015.

Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3), 2012.

C'edric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

T Dudley Wallace. Pretest estimation in regression: a survey. *American Journal of Agricultural Economics*, 59(3), 1977.

Jialei Wang, Jason D Lee, Mehrdad Mahdavi, Mladen Kolar, and Nathan Srebro. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. *Electronic Journal of Statistics*, 11(2), 2017.

David Weisburd, Malay K Majmundar, Hassan Aden, Anthony Braga, Jim Bueermann, Philip J Cook, Phillip Atiba Goff, Rachel A Harmon, Amelia Haviland, Cynthia Lum, Charles Manski, Stephen Mastrofski, Tracey Meares, Daniel Nagin, Emily Owens, Steven Raphael, Jerry Ratcliffe, and Tom Tyler. Proactive policing: A summary of the report of the national academies of sciences, engineering, and medicine. *Asian Journal of Criminology*, 14(2), 2019.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1), 2016.

Spencer Woody, Oscar Hernan Madrid Padilla, and James G Scott. 'Optimal post-selection inference for sparse signals: a nonparametric empirical-Bayes approach'. arXiv:1810.11042, 2020.

Xiaolin Yang, Seyoung Kim, and Eric Xing. Heterogeneous multitask learning with joint sparsity constraints. *Advances in neural information processing systems*, 22, 2009.

Yun Yang, Martin J Wainwright, and Michael I Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6), 2016.

Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, and Christer Betsholtz. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226), 2015.

Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298), 1962.

Arnold Zellner and David S Huang. Further properties of efficient estimators for seemingly unrelated regression equations. *International Economic Review*, 3(3), 1962.

Lijun Zhang, Mehrdad Mahdavi, Rong Jin, Tianbao Yang, and Shenghuo Zhu. Random projections for classification: A recovery approach. *IEEE Transactions on Information Theory*, 60(11), 2014.

Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4), 2014.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4), 1997.

Jim Zidek. Deriving unbiased risk estimators of multinormal mean and regression coefficient estimators using zonal polynomials. *The Annals of Statistics*, 1978.