BAYESIAN MODEL SELECTION AND ESTIMATION WITHOUT MCMC

Sameer K. Deshpande

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

Edward I. George, Universal Furniture Professor, Professor of Statistics

Graduate Group Chairperson

Catherine M. Schrand, Celia Z. Moh Professor of Accounting

Dissertation Committee

Dylan S. Small, Class of 1965 Wharton Professor, Professor of Statistics,

Abraham J. Wyner, Professor of Statistics

Veronika Ročková, Assistant Professor of Statistics and Econometrics

BAYESIAN MODEL SELECTION AND ESTIMATION WITHOUT MCMC

Dedicated to my grandparents:

Eknath & Srila Deshpande and Vasant & Sulabha Joshi

ACKNOWLEDGEMENT

First and foremost, I would like to thank my advisor Ed, for his constant support and encouragement. No matter how busy you may have been, you have always taken time to meet with me to discuss new ideas and problems. I have learned so much from our discussions and always come away inspired to think about new directions. Most importantly, however, through your example, you've shown me how to be a good colleague, mentor, friend, and overall "good guy." For that, I will be forever in your debt.

I would like to thank my committee members, Dylan, Adi, and Veronika. It has been such a pleasure and honor working with each of you over the last several years. Thank you so much for your constant inspiration and encouragement. I look forward to continued collaboration, discussion, and friendship.

Thanks too to the entire Wharton Statistics Department for creating such a warm, familiar, and welcoming environment. To the faculty with whom I have been lucky to interact – thank you for your time, dedication, and so many stimulating conversations. To our incredible staff – thank you for all that you do to keep our department running smoothly and for making the department such a friendly place to work.

I have been blessed to have made so many incredible friends during my time at Wharton.

Matt and Colman – what an incredible journey it has been! Thank you for your constant companionship these last five years. Raiden – it has been a real pleasure working with you these last two years. I always look forward to our near-daily discussions, whether it is about basketball or statistics. Gemma and Cecilia – I've loved getting to work with the two of your over the last year and a half. While I am sad that our weekly "reading group" must come to an end, I am so excited for many years of friendship and collaboration. I cannot imagine what my time at Wharton would be like without each of you.

Thanks are of course due to my parents, with whom none of this would be possible. Words

are simply inadequate to express fully my love, admiration, and gratitude to them.

ABSTRACT

BAYESIAN MODEL SELECTION AND ESTIMATION WITHOUT MCMC

Sameer K. Deshpande

Edward I. George

This dissertation explores Bayesian model selection and estimation in settings where the model space is too vast to rely on Markov Chain Monte Carlo for posterior calculation. First, we consider the problem of sparse multivariate linear regression, in which several correlated outcomes are simultaneously regressed onto a large set of covariates, where the goal is to estimate a sparse matrix of covariate effects and the sparse inverse covariance matrix of the residuals. We propose an Expectation-Conditional Maximization algorithm to target a single posterior mode. In simulation studies, we find that our algorithm outperforms other regularization competitors thanks to its adaptive Bayesian penalty mixing. In order to better quantify the posterior model uncertainty, we then describe a particle optimization procedure that targets several high-posterior probability models simultaneously. This procedure can be thought of as running several "mutually aware" mode-hunting trajectories that repel one another whenever they approach the same model. We demonstrate the utility of this method for fitting Gaussian mixture models and for identifying several promising partitions of spatially-referenced data. Using these identified partitions, we construct an approximation for posterior functionals that average out the uncertainty about the underlying partition. We find that our approximation has favorable estimation risk properties, which we study in greater detail in the context of partially exchangeable normal means. We conclude with several proposed refinements of our particle optimization strategy that encourage a wider exploration of the model space while still targeting high-posterior probability models.

TABLE OF CONTENTS

ACKNOWLEDGEMENT iv

ABSTRACT vi

LIST OF TABLES ix

LIST OF ILLUSTRATIONS x

CHAPTER 1 : Introduction 1

CHAPTER 2 : The Multivariate Spike-and-Slab LASSO 6

2.1 Introduction 6

2.2 Model and Algorithm 11

2.3 Dynamic Posterior Exploration 18

2.4 Full Multivariate Analysis of the Football Safety Data 33

2.5 Discussion 39

CHAPTER 3 : A Particle Optimization Framework for Posterior Exploration . . . 41

3.1 Motivation 41

3.2 A Variational Approximation 42

3.3 Implementation 44

3.4 Mixture Modeling with an Unknown Number of Mixture Components . . . 45

CHAPTER 4 : Identifying Spatial Clusters 58

4.1 Model and Particle Search Strategy 63

4.2 Simulated Example 66

4.3 Discussion 73

CHAPTER 5 : Estimating (Partially?) Exchangeable Normal Means 74

5.1 Whence Partial Exchangeability? 74
5.2 A Multiple Shrinkage Estimator 78
5.3 Approximate Multiple Shrinkage 80
5.4 Towards a Better Understanding of Risk 88

CHAPTER 6 : Conclusion and Future Directions 96
6.1 Next Steps 96

BIBLIOGRAPHY 103

LIST OF TABLES

TABLE 1 : Low-dimensional variable selection and estimation performance . . 26

TABLE 2 : Low-dimensional covariance selection and estimation performance . 27

TABLE 3 : High-dimensional variable selection and estimation performance . . 28

TABLE 4 : High-dimensional covariance selection and estimation performance 29

TABLE 5 : Balance of covariates between football players and controls 36

TABLE 6 : Risk comparison of smoothing within spatial clusters 62

TABLE 7 : Batting Averages from Efron and Morris (1975) 85

LIST OF ILLUSTRATIONS

FIGURE 1 : Examples of spike-and-slab densities 9

FIGURE 2 : Reproduction of Figure 2c in Ročková and George (2016) 19

FIGURE 3 : Stability of the mSSL trajectories 22

FIGURE 4 : Distributions of signal recovered and unrecovered by mSSL-DPE . 31

FIGURE 5 : Estimated residual graphical model in football safety study 38

FIGURE 6 : Gaussian mixture data 54

FIGURE 7 : Partitions of the Gaussian mixture data 55

FIGURE 8 : Log-counts of violent crime in Philadelphia 60

FIGURE 9 : Three spatial partitions of the grid 61

FIGURE 10 : Three specifications of $\boldsymbol{\beta}$ 61

FIGURE 11 : Top nine spatial partitions when cluster means are well-separated 67

FIGURE 12 : Number of unique particles discovered 68

FIGURE 13 : Risk of approximate estimator when clusters well-separated 69

FIGURE 14 : Top nine spatial partitions when the cluster means are not well-separated 71

FIGURE 15 : Risk of approximate estimator when cluster means are not well-separated 72

FIGURE 16 : Comparison of the standard and clustered Lindley estimators . . . 77

FIGURE 17 : Top four partitions 82

FIGURE 18 : Risk of approximate multiple shrinkage estimator 83

FIGURE 19 : Re-analysis of Efron and Morris (1975)'s batting averages 87

FIGURE 20 : Examples of perturbations $\mathbf{Y}^{+}$ 94

FIGURE 21 : Two approximations of a discrete distribution 97

CHAPTER 1 : Introduction

Given a realization of data, the Bayesian paradigm provides a coherent and tremendously flexible framework within which to reason about our uncertainty about the data generating process. A Bayesian analysis starts with a *likelihood* that is based on the distribution of the data conditional on some unknown parameters, which are treated as random variables drawn from a *prior* distribution. Together, the likelihood and prior are combined via Bayes' rule to form the *posterior* distribution, which encapsulates all of our uncertainty about the unknown parameters in light of the observed data. As Gelman et al. (2008) notes, this specification of the likelihood is a "major stumbling block" for Bayesian analyses.

In this thesis, we consider situations where we have a combinatorially large number of potential likelihoods (i.e. generative models) but are uncertain about which might be most appropriate. We consider two general problems: model selection, in which we wish to identify the models that best describe the data, and estimation of model-specific parameters in the presence of this uncertainty. Conceptually, the Bayesian framework provides an easy answer: simply place a prior over the collection of all models and turn the proverbial Bayesian crank to compute the posterior. Moving from the joint distribution of the data and generative model to the posterior requires computing the marginal likelihood for the data. This in turn requires us to sum over the entire model space, which is often not possible.

We focus on two problems where this is the case: multivariate linear regression and mixture modeling. In multivariate linear regression, we aim to use p covariates to predict the values of q correlated outcomes simultaneously. When p and q are large and even greater than the number of observations n, there is great interest in fitting *sparse* models, where only a small number of covariates are used to predict each outcome and only a small number of residuals are conditionally dependent on each other. Formally, this problem reduces to estimating a sparse $p \times q$ matrix of covariate effects and a sparse $q \times q$ matrix of partial

covariances between the residuals. In all, there are $2^{pq+q(q-1)/2}$ different combinations of the supports of these two matrices, leading to a rather high-dimensional model space for even moderately-sized p and q. For instance, in Section 2.4, we examine data from Deshpande et al. (2017), an observational study on the long-term effects of playing high school football. In that study, there were $p = 204$ covariates and $q = 29$ outcomes, yielding a model space of dimension $2^{6332} > 10^{1903}$. Clearly, we cannot expect to explore even a small fraction of this space within a reasonable number of MCMC iterations. Next, we consider clustering and mixture modeling, in which we assume each data point arose from one of several different distributions. *A priori*, the number of mixture components and the allocation of each observation to a component is unknown. We may encode this structure with a partition of the integers $[n] = \{1, 2, \ldots, n\}$. In this problem, the dimension of our model space grows exponentially in the number of observations: for $n = 10$, there are $115,975$ partitions and for $n = 20$, there are 51,724,158,235,372 partitions! In both of these problems, we are unable to evaluate the posterior distribution exactly.

Of course, intractable posteriors are not a particularly new thorn in the side of Bayesian analysis. Following the seminal paper of Gelfand and Smith (1990), Markov Chain Monte Carlo (MCMC) methods have emerged as the "gold standard" approach for summarizing the posterior distribution by simulating random draws from it. Unfortunately, despite its prominence, the viability of MCMC for performing model selection is limited when the model space is combinatorially massive. To paraphrase Jones et al. (2005), for problems of even moderate size, the model space to be explored is so large that a model's frequency in the sample of models visited by the stochastic search cannot be viewed as reflecting its posterior probability. Even worse, many models are not revisited by the Markov chain, which itself may miss large pockets of posterior probability (Scott and Carvalho, 2008). Scott and Carvalho (2008) go even further, noting that they find "little comfort in [the] infinite-runtime guarantees" when using MCMC in large and complex model spaces because "assessing whether a Markov chain over a multimodal space has converged to the stationary distribution is devilishly tough ... [and] ... apparent finite-time convergence can prove to

be a mirage."

In this thesis, we elaborate and extend a line of work initiated by Ročková and George (2016) and Ročková (2017), who use optimization rather than MCMC to rapidly identify promising models in the high-dimensional univariate linear regression setting. In Chapter 2, we build on Ročková and George (2014)'s and Ročková and George (2016)'s deterministic spike-and-slab formulation of Bayesian variable selection to develop a full joint procedure for simultaneous variable and covariance selection problem in multivariate linear regression models. We propose and deploy an Expectation-Conditional Maximization algorithm within a path-following scheme to identify the modes of several posterior distributions. This *dynamic posterior exploration* of several posteriors is in marked contrast to MCMC, which attempts to characterize a single posterior. In simulation studies, we find that our method outperforms regularization competitors and we also demonstrate our method using data from Deshpande et al. (2017), an observational study on the long-term health effects of playing high school football.

While our results are certainly encouraging, the procedure introduced in Chapter 2 only targets a single posterior mode. To begin to explore the posterior uncertainty about the underlying model, in Chapter 3 we revisit Ročková (2017)'s Particle EM for variable selection, which targets several promising models in the univariate linear regression setting simultaneously. We re-derive this procedure in a more general model selection setting and demonstrate the utility of this *particle optimization* procedure for clustering and mixture modeling. At a high-level, this procedure works by running several mode-hunting trajectories through the model space that repel one another whenever they appear headed to the same point. In this way, the procedure aims to identify several high-posterior probability models once rather than a single promising model multiple times, as can happen if we ran independent instantiations of a mode-hunting algorithm from several random starting points.

Chapter 4 describes work that is motivated by a study of the time trend in crime rate

in the city of Philadelphia. In that application, we have data from every census block group in the city and wish to fit a regression model within each in a spatially smooth manner. Intuitively, we might expect that the regression slopes in neighboring block groups are similar. As a result, we wish to borrow strength between adjacent spatial units in a principled manner. A common way of doing this is to use a conditionally auto-regressive prior (see, e.g. Besag, 1974) on the region-specific regression slopes. Doing so, however, introduces a certain global smoothness and may in fact over-smooth across sharp spatial boundaries. Such boundaries exist in complex urban environments as a product of physical barriers (e.g. highways and rivers) or human barriers (e.g. differences in demographics) between adjacent spatial regions. To deal with this possibility, we aim to first partition the spatial regions into clusters with similar trends and then estimate the slopes within each cluster separately. In the context of studying trends in the crime rate, this can result in the under- or over-estimation of crime in individual block groups. We introduce a version of the Bayesian Partition Model of Holmes et al. (1999) in which we induce spatial smoothness within clusters of regions but model each cluster independently. Rather than directly sample from the space of all possible spatial partitions, we deploy the particle optimization method developed in Chapter 3 to identify several promising partitions. We then approximate the marginal posterior mean of the regression slopes using an adaptive combination of conditional posterior means corresponding to the identified partitions. In simulation studies, we see that this adaptive estimator can realize substantial improvements in estimation risk over estimators based on pre-specified partitions of the data. This is on-going work with with Cecilia Balocchi, Ed George, and Shane Jensen.

In Chapter 5, we continue with the theme of approximating posterior expectations in the presence of model uncertainty, focusing on minimax shrinkage estimation of normal means. Unlike in most treatments of this problem, we no longer assume that the means are exchangeable, instead assuming only that there may be groups of means which are similar in value. We encode this *partial exchangeability* structure with a partition on the set $[n]$ and express our initial uncertainty about it with a prior over the space of partitions. We

propose a prior hierarchy conditional on the underlying partition so that the corresponding conditional posterior expectations of the vector of unknown means are minimax shrinkage estimators. We rely on results in George (1986b,a,c) to combine these conditional estimators to form a minimax *multiple shrinkage estimator*. Unfortunately, computing this estimator requires enumeration of all partitions of $[n]$. To approximate the estimator, we use a similar strategy as in Chapter 4: we first identify several promising partitions and then adaptively combine the corresponding estimators. We find in simulation settings that this procedure can sometimes yield substantial improvements in risk relative to possibly mis-specified estimators which assume a particular partial exchangeability structure. We then begin studying the risk of this estimator, in an attempt to determine whether the approximate estimator is still minimax and if not, how far from minimax it is. A central challenge to this study is the fact that the selection of these partitions and the ultimate estimation of the vector of means are not independent.

Despite the promise shown by the particle optimization procedure, we have found that it has a tendency to remain stuck in the vicinity of a dominant posterior mode. While this is not totally unreasonable, it does limit our ability to summarize the posterior over the model space. We propose a set of relaxations of the original optimization objective designed to encourage a wider exploration of the model space in Chapter 6.

CHAPTER 2 : The Multivariate Spike-and-Slab LASSO

2.1. Introduction

We consider the multivariate Gaussian linear regression model, in which one simultaneously regresses $q > 1$ possibly correlated responses onto a common set of p covariates. In this setting, one observes n independent pairs of data $(\mathbf{x}_i, \mathbf{y}_i)$ where $\mathbf{y}_i \in \mathbb{R}^q$ contains the q outcomes and $\mathbf{x}_i \in \mathbb{R}^p$ contains measurements of the covariates. One then models $\mathbf{y}_i = \mathbf{x}_i' B + \varepsilon_i$, with $\varepsilon_1, \ldots, \varepsilon_n \sim \mathrm{N}\left(\mathbf{0}_q, \Omega^{-1}\right)$, independently, where $B = (\beta_{j,k})_{j,k}$ and $\Omega = \left(\omega_{k,k'}\right)_{k,k'}$ are unknown $p \times q$ and $q \times q$ matrices, respectively. The main thrust of this chapter is to propose a new methodology for the simultaneous identification of the regression coefficient matrix B and the residual precision matrix Ω. Our framework additionally includes estimation of B when Ω is known and estimation of Ω when B is known as important special cases.

The identification and estimation of a sparse set of regression coefficients has been extensively explored in the univariate linear regression model, often through a penalized likelihood framework. Perhaps the most prominent method is Tibshirani (1996)'s LASSO, which adds an ℓ_1 penalty to the negative log-likelihood. The last two decades have seen a proliferation of alternative penalties, including the adaptive lasso (Zou, 2006), smoothly clipped absolute deviation (SCAD), (Fan and Li, 2001), and minimum concave penalty (Zhang, 2010). Given the abundance of penalized likelihood procedures for univariate regression, when moving to the multivariate setting, it is very tempting to deploy one's favorite univariate procedure to each of the q responses separately, thereby assembling an estimate of B column-by-column. Such an approach fails to account for the correlations between responses and may lead to poor predictive performance (see, e.g., Breiman and Friedman (1997)). Perhaps more perniciously, in many applied settings one may reasonably believe that some groups of covariates are simultaneously "relevant" to many responses. A response-by-response approach to variable selection fails to investigate or leverage such structural assumptions. This has led to the the block-structured regularization approaches of Turlach et al. (2005), Obozinski et al.

(2011) and Peng et al. (2010), among many others. While these proposals frequently yield highly interpretable and useful models, they do not explicitly model the residual correlation structure, essentially assuming that $\Omega = I$.

Estimation of a sparse precision matrix from multivariate Gaussian data has a similarly rich history, dating back to Dempster (1972), who coined the phrase *covariance selection* to describe this problem. While Dempster (1972) was primarily concerned with estimating the covariance matrix $\Sigma = \Omega^{-1}$ by first sparsely estimating the precision matrix Ω, recent attention has focused on estimating the underlying Gaussian graphical model, G. The vertices of the graph G correspond to the coordinates of the multivariate Gaussian vector and an edge between vertices k and k' signifies that the corresponding coordinates are conditionally dependent. These conditional dependency relations are encoded in the support of Ω. A particularly popular approach to estimating Ω is the graphical lasso (GLASSO), which adds an ℓ_1 penalty to the negative log-likelihood of Ω (see, e.g., Yuan and Lin (2007), Banerjee et al. (2008), and Friedman et al. (2008)).

While variable selection and covariance selection each have long, rich histories, joint variable and covariance selection has only recently attracted attention. To the best of our knowledge, Rothman et al. (2010) was among the first to consider the simultaneous sparse estimation of B and Ω, solving the penalized likelihood problem:

$$\arg\min_{B,\Omega} \left\{ -\frac{n}{2} \log |\Omega| + \frac{1}{2} \mathrm{tr}\left((\mathbf{Y} - \mathbf{X}B)\, \Omega\, (\mathbf{Y} - \mathbf{X}B)' \right) + \lambda \sum_{j,k} |\beta_{j,k}| + \xi \sum_{k \neq k'} \left| \omega_{k,k'} \right| \right\} \quad (2.1)$$

Their procedure, called MRCE for "Multivariate Regression with Covariance Estimation", induces sparsity in B and Ω with separate ℓ_1 penalties and can be viewed as an elaboration of both the LASSO and GLASSO. Following Rothman et al. (2010), several authors have proposed solving problems similar to that in Equation (2.1): Yin and Li (2011) considered nearly the same objective but with adaptive LASSO penalties, Lee and Liu (2012) proposed weighting each $|\beta_{j,k}|$ and $\left|\omega_{k,k'}\right|$ individually, and Abegaz and Wit (2013) replaced the

ℓ_1 penalties with SCAD penalties. Though the ensuing joint optimization problem can be numerically unstable in high-dimensions, all of these authors report relatively good performance in estimating B and Ω. Cai et al. (2013) takes a somewhat different approach, first estimating B in a column-by-column fashion with a separate Dantzig selector for each response and then estimating Ω by solving a constrained ℓ_1 optimization problem. Under mild conditions, they established the asymptotic consistency of their two-step procedure, called CAPME for "Covariate-Adjusted Precision Matrix Estimation."

Bayesians too have considered variable and covariance selection. A workhorse of sparse Bayesian modeling is the spike-and-slab prior, in which one models parameters as being drawn *a priori* from either a point-mass at zero (the "spike") or a much more diffuse continuous distribution (the "slab") (Mitchell and Beauchamp, 1988). To deploy such a prior, one introduces a latent binary variable for each regression coefficient indicating whether it was drawn from the spike or slab distribution and uses the posterior distribution of these latent parameters to perform variable selection. George and McCulloch (1993) relaxed this formulation slightly by taking the spike and slab distributions to be zero-mean Gaussians, with the spike distribution very tightly concentrated around zero. Their relaxation facilitated a straight-forward Gibbs sampler that forms the backbone of their Stochastic Search Variable Selection (SSVS) procedure for univariate linear regression. While continuous spike and slab densities generally preclude exactly sparse estimates, the intersection point of the two densities can be viewed as an *a priori* "threshold of practical relevance." More recently, Ročková and George (2016) took both the spike and slab distributions to be Laplacian, which led to posterior distributions with exactly sparse modes. Under mild conditions, their "spike-and-slab lasso" priors produce posterior distributions that concentrate asymptotically around the true regression coefficients at nearly the minimax rate. Figure 1 illustrates these three different spike-and-slab proposals.

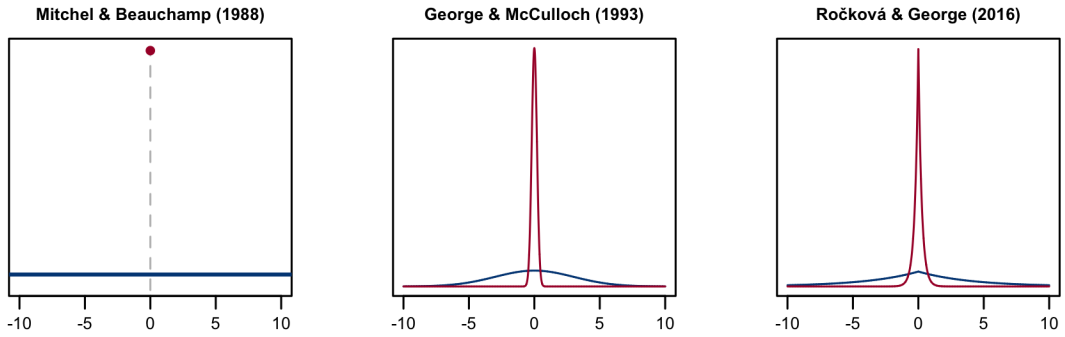


Figure 1: Three choices of spike and slab densities. Slab densities are colored red and spike densities are colored blue. The heavier Laplacian tails of Ročková and George (2016)'s slab distribution help stabilize non-zero parameter more so than George and McCulloch (1993)'s Gaussian slabs.

An important Bayesian approach to covariance selection begins by specifying a prior over the underlying graph G and a hyper-inverse Wishart prior (Dawid and Lauritzen, 1993) on $\Sigma|G$. This prior is constrained to the set of symmetric positive-definite matrices such that off-diagonal entry $\omega_{k,k'}$ of $\Sigma^{-1} = \Omega$ is non-zero if and only if there is an edge between vertices k and k' in G. See Giudici and Green (1999), Roverato (2002), and Carvalho and Scott (2009) for additional methodological and theoretical details on these priors and see Jones et al. (2005) and Carvalho et al. (2007) for computational considerations. Recently, Wang (2015) and Banerjee and Ghosal (2015) placed spike-and-slab priors on the off-diagonal elements of Ω, using a Laplacian slab and a point-mass spike at zero. Banerjee and Ghosal (2015) established the posterior consistency in the asymptotic regime where $(q+s)\log q = o(n)$ where s is the total number of edges in G.

Despite their conceptual elegance, spike-and-slab priors result in highly multimodal posteriors that can slow the mixing of MCMC simulations. This is exacerbated in the multivariate regression setting, especially when p and q are moderate-to-large relative to n. To overcome this slow mixing when extending SSVS to the multivariate linear regression model, Brown et al. (1998) restricted attention to models in which a variable was selected as "relevant" to either all or none of the responses. This enabled them to marginalize out the parameter B

and directly Gibbs sample the latent spike-and-slab indicators. Despite the computational tractability, the focus to models in which a covariate affects all or none of the responses may be unrealistic and overly restrictive. More recently, Richardson et al. (2010) overcame this by using an evolutionary MCMC simulation, but made the equally restrictive and unrealistic assumption that Ω was diagonal. Bhadra and Mallick (2013) placed spike-and-slab priors on the elements of B and a hyper inverse Wishart prior on $\Sigma|G$. To ensure quick mixing of their MCMC, they made the same restriction as Brown et al. (1998): a variable was selected as relevant to all of the q responses or to none of them. It would seem, then, that a Bayesian who desires a computationally efficient procedure must choose between having a very general sparsity structure in B at the expense of a diagonal Ω (à la Richardson et al. (2010)), or a general sparsity structure in Ω with a peculiar sparsity pattern in B (à la Brown et al. (1998) and Bhadra and Mallick (2013)). Although their non-Bayesian counter-parts are not nearly as encumbered, the problem of picking appropriate penalty weights via cross-validation can be computationally burdensome.

In this paper, we attempt to close this gap, by extending the EMVS framework of Ročková and George (2014) and spike-and-slab lasso framework of Ročková and George (2016) to the multivariate linear regression setting. EMVS is a deterministic alternative to the SSVS procedure that avoids posterior sampling by targeting local modes of the posterior distribution with an EM algorithm that treats the latent spike-and-slab indicator variables as "missing data." Through its use of Gaussian spike and slab distributions, the EMVS algorithm reduces to solving a sequence of ridge regression problems whose penalties *adapt* to the evolving estimates of the regression parameter. Subsequent development in Ročková and George (2016) led to the spike-and-slab lasso procedure, in which both the spike and slab distributions were taken to be Laplacian. This framework allows us to "cross-fertilize" the best of the Bayesian and non-Bayesian approaches: by targeting posterior modes instead of sampling, we may lean on existing highly efficient algorithms for solving penalized likelihood problems while the Bayesian machinery facilities adaptive penalty mixing, essentially for free.

Much like Ročková and George (2014)'s EMVS, our proposed procedure reduces to solving a series of penalized likelihood problems. Our prior model of the uncertainty about which covariate effects and partial residual covariances are large and which are essentially negligible allows us to perform selective shrinkage, leading to vastly superior support recovery and estimation performance compared to non-Bayesian procedures like MRCE and CAPME. Moreover, we have found our joint treatment of B and Ω, which embraces the residual correlation structure from the outset, is capable of identifying weaker covariate effects than two-step procedures that first estimate B either column-wise or by assuming $\Omega = I$ and then estimate Ω.

The rest of this paper is organized as follows. We formally introduce our model and algorithm in Section 2.2. In Sections 2.3, we embed this algorithm within a path-following scheme that facilitates *dynamic posterior exploration*, identifying putative modes of B and Ω over a range of different posterior distributions indexed by the "tightness" of the prior spike distributions. We present the results of several simulation studies in Section 2.3.2. In Section 2.4, we re-analyze the data of Deshpande et al. (2017), a recent observational study on the effects of playing high school football on a range of cognitive, behavioral, psychological, and socio-economic outcomes later in life. We conclude with a discussion in Section 2.5.

2.2. Model and Algorithm

We begin with some notation. We let $\|B\|_0$ be the number of non-zero entries in the matrix B and, abusing the notation somewhat, we let $\|\Omega\|_0$ be the number of non-zero, off-diagonal entries in the upper triangle of the precision matrix Ω. For any matrix of covariates effects B, we let $\mathbf{R}(B) = \mathbf{Y} - \mathbf{X}B$ denote the residual matrix whose k^{th} column is denoted $\mathbf{r}_k(B)$. Finally, let $S(B) = n^{-1}\mathbf{R}(B)'\mathbf{R}(B)$ be the residual covariance matrix. In what follows, we will usually suppress the dependence of $\mathbf{R}(B)$ and $S(B)$ on B, writing only $\mathbf{R}$ and S. Additionally, we assume that the columns of $\mathbf{X}$ have been centered and scaled to have mean 0 and Euclidean norm $\sqrt{n}$ and that the columns of $\mathbf{Y}$ have been centered and are on

approximately similar scales.

Recall that our data likelihood is given by

$$p(\mathbf{Y}|B,\Omega) \propto |\Omega|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left((\mathbf{Y}-\mathbf{X}B)\,\Omega\,(\mathbf{Y}-\mathbf{X}B)'\right)\right\}$$

We introduce latent 0–1 indicators, $\boldsymbol{\gamma} = (\gamma_{j,k} : 1 \leq j \leq p, 1 \leq k \leq q)$ so that, independently for $1 \leq j \leq p, 1 \leq k \leq q$, we have

$$\pi(\beta_{j,k}|\gamma_{j,k}) \propto \left(\lambda_1 \mathrm{e}^{-\lambda_1|\beta_{j,k}|}\right)^{\gamma_{j,k}} \left(\lambda_0 \mathrm{e}^{-\lambda_0|\beta_{j,k}|}\right)^{1-\gamma_{j,k}}.$$

Similarly, we introduce latent 0–1 indicators, $\boldsymbol{\delta} = \left(\delta_{k,k'} : 1 \leq k < k' \leq q\right)$ so that, independently for $1 \leq k < k' \leq q$, we have

$$\pi(\omega_{k,k'}|\delta_{k,k'}) \propto \left(\xi_1 \mathrm{e}^{-\xi_1|\omega_{k,k'}|}\right)^{\delta_{k,k'}} \left(\xi_0 \mathrm{e}^{-\xi_0|\omega_{k,k'}|}\right)^{1-\delta_{k,k'}}$$

Recall that in the spike-and-slab framework, the spike distribution is viewed as having *a priori* generated all of the negligible parameter values, permitting us to interpret $\gamma_{j,k} = 0$ as an indication that variable j has an essentially null effect on outcome k. Similarly, we may interpret $\delta_{k,k'} = 0$ to mean that the partial covariance between $\mathbf{r}_k$ and $\mathbf{r}_{k'}$ is small enough to ignore. To model our uncertainty about $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, we use the familiar beta-binomial prior (Scott and Berger, 2010) :

$$\gamma_{j,k}|\theta \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\theta) \qquad \theta \sim \text{Beta}(a_\theta, b_\theta)$$
$$\delta_{k,k'}|\eta \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\eta) \qquad \eta \sim \text{Beta}(a_\eta, b_\eta)$$

where $a_\theta, b_\theta, a_\eta$, and b_η are fixed positive constants, and $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are *a priori* independent. We may view θ and η as measuring the proportion of non-zero entries in B and non-zero off-diagonal elements of Ω, respectively. To complete our prior specification, we place independent exponential $\text{Exp}(\xi_1)$ priors on the diagonal elements of Ω and additionally

restrict the prior on Ω to the cone of symmetric positive definite matrices.

Before proceeding, we take a moment to introduce two functions that will play a critical role in our optimization strategy. Given $\lambda_1, \lambda_0, \xi_1$ and ξ_0, define the functions $p^\star, q^\star : \mathbb{R} \times [0,1] \to [0,1]$ by

$$p^\star(x,\theta) = \frac{\theta\lambda_1 \mathrm{e}^{-\lambda_1|x|}}{\theta\lambda_1 \mathrm{e}^{-\lambda_1|x|} + (1-\theta)\lambda_0 \mathrm{e}^{-\lambda_0|x|}}$$
$$q^\star(x,\eta) = \frac{\eta\xi_1 \mathrm{e}^{-\xi_1|x|}}{\eta\xi_1 \mathrm{e}^{-\xi_1|x|} + (1-\eta)\xi_0 \mathrm{e}^{-\xi_0|x|}}.$$

Letting Ξ denote the collection $\{B, \theta, \Omega, \eta\}$, it is straightforward to verify that $p^\star(\beta_{j,k}, \theta) = \mathbb{E}\left[\gamma_{j,k} | \mathbf{Y}, \Xi\right]$ and $q^\star(\omega_{k,k'}, \eta) = \mathbb{E}\left[\delta_{k,k'} | \mathbf{Y}, \Xi\right]$, the conditional posterior probabilities that $\beta_{j,k}$ and $\omega_{k,k'}$ were drawn from their respective slab distributions.

Integrating out the latent indicators, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, the log-posterior density of Ξ is, up to an additive constant, given by

$$\begin{aligned}
\log \pi(\Xi|\mathbf{Y}) = \frac{n}{2}\log|\Omega| - \frac{1}{2}\mathrm{tr}\left((\mathbf{Y}-\mathbf{X}B)'(\mathbf{Y}-\mathbf{X}B)\,\Omega\right) \\
+ \sum_{j,k} \log\left(\theta\lambda_1 \mathrm{e}^{-\lambda_1|\beta_{j,k}|} + (1-\theta)\,\lambda_0 \mathrm{e}^{-\lambda_0|\beta_{j,k}|}\right) \\
+ \sum_{k,k'} \log\left(\eta\xi_1 \mathrm{e}^{-\xi_1|\omega_{k,k'}|} + (1-\eta)\,\xi_0 \mathrm{e}^{-\xi_0|\omega_{k,k'}|}\right) - \xi_1 \sum_{k=1}^{q} \omega_{k,k} \\
+ (a_\theta - 1)\log\theta + (b_\theta - 1)\log(1-\theta) + (a_\eta - 1)\log\eta + (b_\eta - 1)\log(1-\eta).
\end{aligned} \tag{2.2}$$

Rather than directly sample from this intractable posterior distribution with MCMC, we maximize the posterior density, seeking $\Xi^* = \arg\max\{\log\pi(\Xi|\mathbf{Y})\}$. Performing this joint optimization is quite challenging, especially in light of the non-convexity of the log-posterior density. To overcome this, we use an Expectation/Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993) that treats the only the partial covariance indicators $\boldsymbol{\delta}$ as "missing data." For the E step of this algorithm, we first compute $q^\star_{k,k'} := q^\star(\omega^{(t)}_{k,k'}, \eta^{(t)}) =$

$\mathbb{E}\left[\delta_{k,k'}|\mathbf{Y}, \Xi^{(t)}\right]$ given a current estimate $\Xi^{(t)}$ and then consider maximizing the surrogate objective function

$$\begin{aligned}\mathbb{E}\left[\log \pi(\Xi, \boldsymbol{\delta}|\mathbf{Y})|\Xi^{(t)}\right] = \frac{n}{2}\log|\Omega| - \frac{1}{2}\mathrm{tr}\left((\mathbf{Y}-\mathbf{X}B)'(\mathbf{Y}-\mathbf{X}B)\,\Omega\right) \\ + \sum_{j,k}\log\left(\theta\lambda_1 \mathrm{e}^{-\lambda_1|\beta_{j,k}|} + (1-\theta)\,\lambda_0 \mathrm{e}^{-\lambda_0|\beta_{j,k}|}\right) \\ - \sum_{k,k'}\xi^{\star}_{k,k'}\left|\omega_{k,k'}\right| - \xi_1\sum_{k=1}^{q}\omega_{k,k} \\ + (a_\theta - 1)\log\theta + (b_\theta - 1)\log(1-\theta) \\ + (a_\eta - 1)\log\eta + (b_\eta - 1)\log(1-\eta)\end{aligned}$$

where $\xi^{\star}_{k,k'} = \xi_1 q^{\star}_{k,k'} + \xi_0(1 - q^{\star}_{k,k'})$.

We then perform two CM steps, first updating the pair (B, θ) while holding $(\Omega, \eta) = (\Omega^{(t)}, \eta^{(t)})$ fixed at its previous value and then updating (Ω, η) while fixing (B, θ) at its new value $\left(B^{(t+1)}, \Omega^{(t+1)}\right)$. As we will see shortly, augmenting our log-posterior with the indicators $\boldsymbol{\delta}$ facilitates simple updates of Ω by solving a GLASSO problem. It is worth noting that we do not also augment our log-posterior with the indicators $\boldsymbol{\gamma}$ as the update of B can be carried out with a coordinate ascent strategy despite the non-convex penalty seen in the second line of Equation (2.2).

We are now ready to describe the two CM steps. Holding $(\Omega, \eta) = (\Omega^{(t)}, \eta^{(t)})$ fixed, we update (B, θ) by solving

$$(B^{(t+1)}, \theta^{(t+1)}) = \underset{B,\theta}{\arg\max}\left\{-\frac{1}{2}\mathrm{tr}\left((\mathbf{Y}-\mathbf{X}B)\,\Omega\,(\mathbf{Y}-\mathbf{X}B)'\right) + \log\pi(B|\theta) + \log\pi(\theta)\right\} \tag{2.3}$$

where

$$\pi(B|\theta) = \prod_{j,k}\left(\theta\lambda_1 \mathrm{e}^{-\lambda_1|\beta_{j,k}|} + (1-\theta)\lambda_0 \mathrm{e}^{-\lambda_0|\beta_{j,k}|}\right).$$

and $\pi(\theta) \propto \theta^{a_\theta - 1}(1-\theta)^{b_\theta - 1}$. We do this in a coordinate-wise fashion, sequentially updating

θ with a simple Newton algorithm and updating B by solving the following problem

$$\tilde{B} = \arg\max_{B} \left\{ -\frac{1}{2}\text{tr}\left((\mathbf{Y} - \mathbf{X}B)\,\Omega\,(\mathbf{Y} - \mathbf{X}B)'\right) + \sum_{j,k} \text{pen}(\beta_{j,k}|\theta) \right\} \tag{2.4}$$

where

$$pen(\beta_{j,k}|\theta) = \log\left(\frac{\pi\left(\beta_{j,k}|\theta\right)}{\pi(0|\theta)}\right) = -\lambda_1\left|\beta_{j,k}\right| + \log\left(\frac{p^\star(\beta_{j,k},\theta)}{p^\star(0,\theta)}\right).$$

Using the fact that the columns of $\mathbf{X}$ have norm $\sqrt{n}$ and Lemma 2.1 of Ročková and George (2016), the Karush-Kuhn-Tucker condition tells us that

$$\tilde{\beta}_{j,k} = n^{-1}\left[|z_{j,k}| - \lambda^\star(\tilde{\beta}_{j,k},\theta)\right]_+ \text{sign}(z_{j,k}),$$

where

$$z_{j,k} = n\tilde{\beta}_{j,k} + \sum_{k'} \frac{\omega_{k,k'}}{\omega_{k,k}} \mathbf{x}_j' \mathbf{r}_{k'}(\tilde{B})$$

$$\lambda^\star_{j,k} := \lambda^\star(\tilde{\beta}_{j,k},\theta) = \lambda_1 p^\star(\tilde{\beta}_{j,k},\theta) + \lambda_0(1 - p^\star(\tilde{\beta}_{j,k},\theta)).$$

The form of $\tilde{\beta}_{j,k}$ above immediately suggests a coordinate ascent strategy with soft-thresholding to compute $\tilde{B}$ that is very similar to the one used to compute LASSO solutions (Friedman et al., 2007). As noted by Ročková and George (2016), however, this necessary characterization of $\tilde{B}$ is generally not sufficient. Arguments in Zhang and Zhang (2012) and Ročková and George (2016) lead immediately to the following refined characterization of $\tilde{B}$.

Proposition 1. *The entries in the global mode* $\tilde{B} = \left(\tilde{\beta}_{j,k}\right)$ *in Equation* (2.4) *satisfy*

$$\tilde{\beta}_{j,k} = \begin{cases} n^{-1}\left[|z_{j,k}| - \lambda^\star(\tilde{\beta}_{j,k},\theta)\right]_+ sign\left(z_{j,k}\right) & when\ |z_{j,k}| > \Delta_{j,k} \\ 0 & when\ |z_{j,k}| \leq \Delta_{j,k} \end{cases}$$

where

$$\Delta_{j,k} = \inf_{t>0}\left\{\frac{nt}{2} - \frac{pen(\tilde{\beta}_{j,k},\theta)}{\omega_{k,k}t}\right\}$$

The threshold $\Delta_{j,k}$ is generally quite hard to compute but can be bounded, as seen in the following analog to Theorem 2.1 of Ročková and George (2016).

Proposition 2. *Suppose that* $(\lambda_1 - \lambda_0) > 2\sqrt{n\omega_{k,k}}$ *and* $(\lambda^\star(0,\theta) - \lambda_1)^2 > -2n\omega_{k,k}p^\star(0,\theta)$. *Then* $\Delta^L_{j,k} \leq \Delta_{j,k} \leq \Delta^U_{j,k}$ *where*

$$\Delta^L_{j,k} = \sqrt{-2n\omega^{-1}_{k,k}\log p^\star(0,\theta) - \omega^{-2}_{k,k}d} + \omega^{-1}_{k,k}\lambda_1$$
$$\Delta^U_{j,k} = \sqrt{-2n\omega^{-1}_{k,k}\log p^\star(0,\theta)} + \omega^{-1}_{k,k}\lambda_1$$

where $d = -\left(\lambda^\star(\delta_{c_+},\theta) - \lambda_1\right)^2 - 2n\omega_{k,k}\log p^\star(\delta_{c_+},\theta)$ *and* δ_{c_+} *is the larger root of* $pen''(x|\theta) = \omega_{k,k}$.

Proposition 1 gives us a refined characterization of $\hat{B}$ in terms of element-wise thresholds $\Delta_{j,k}$. Proposition 2 allows us to bound these thresholds and together they suggest a *refined coordinate ascent* strategy for updating our estimate of B. Namely, starting from some initial value B^{old}, we can update $\beta_{j,k}$ with the thresholding rule:

$$\beta^{new}_{j,k} = \frac{1}{n}\left(|z_{j,k}| - \lambda^\star(\beta^{old}_{j,k},\theta)\right)_+ \text{sign}(z_{j,k})\mathbb{I}\left(|z_{j,k}| > \Delta^U_{j,k}\right).$$

Before proceeding, we pause for a moment to reflect on the threshold $\lambda^\star_{j,k}$ appearing in the KKT condition and Proposition 1, which evolves alongside our estimates of B and θ. In particular, when our current estimate of $\beta_{j,k}$ is large in magnitude, the conditional posterior probability that it was drawn from the slab, $p^\star_{j,k}$, tends to be close to one so that $\lambda^\star_{j,k}$ is close to λ_1. On the other hand, if it is small in magnitude, $\lambda^\star_{j,k}$ tends to be close to the much larger λ_0. In this way, as our EM algorithm proceeds, performs *selective shrinkage*, aggressively penalizing small values of $\beta_{j,k}$ without overly penalizing larger values. It is worth pointing out as well that $\lambda^\star_{j,k}$ adapts not only to the current estimate of B but also to the overall level of sparsity in B, as reflected in the current estimate of θ. The adaptation is entirely a product our explicit *a priori* modeling of the latent indicators $\boldsymbol{\gamma}$ and stands in stark contrast to regularization techniques that deploy fixed penalties.

Fixing $(\Omega, \eta) = (\Omega^{(t)}, \eta^{(t)})$, we iterate between the refined coordinate ascent for B and the Newton algorithm for θ until some convergence criterion is reached at some new estimate $(B^{(t+1)}, \theta^{(t+1)})$. Then, holding $(B, \theta) = (B^{(t+1)}, \theta^{(t+1)})$, we turn our attention to (Ω, η) and solving the posterior maximization problem

$$\left(\Omega^{(t+1)}, \eta^{(t+1)}\right) = \arg\max \left\{ \frac{n}{2} \log|\Omega| - \frac{1}{2}\mathrm{tr}\left(S\Omega\right) - \sum_{k<k'} \xi^{\star}_{k,k'} \left|\omega_{k,k'}\right| - \xi_1 \sum_{k=1}^{q} \omega_{k,k} \right.$$
$$\left. + \log\eta \times \left(a_\eta - 1 + \sum_{k<k'} q^{\star}_{k,k'} \right) + \log\left(1-\eta\right) \times \left(b_\eta - 1 + \sum_{k<k} \left(1 - q^{\star}_{k,k'}\right) \right) \right\}.$$

It is immediately clear that there is a closed form update of η:

$$\eta^{(t+1)} = \frac{a_\eta - 1 + \sum_{k<k'} q^{\star}_{k,k'}}{a_\eta + b_\eta - 2 + q(q-1)/2}.$$

For Ω, we recognize the M Step update of Ω as a GLASSO problem.

$$\Omega^{(t+1)} = \arg\max_{\Omega \succ 0} \left\{ \frac{n}{2} \log|\Omega| - \frac{n}{2}\mathrm{tr}\left(S\Omega\right) - \sum_{k<k'} \xi^{\star}_{k,k'} \left|\omega_{k,k'}\right| - \xi_1 \sum_{k=1}^{q} \omega_{k,k} \right\} \tag{2.5}$$

To find $\Omega^{(t+1)}$, rather than using the block-coordinate ascent algorithms of Friedman et al. (2008) and Witten et al. (2011), we use the state-of-art QUIC algorithm of Hsieh et al. (2014), which is based on a quadratic approximation of the objective function and achieves a super-linear convergence rate. Each of these algorithms returns a positive semi-definite $\Omega^{(t+1)}$. Just like with the $\lambda^{\star}_{j,k}$'s, the penalties $\xi^{\star}_{k,k'}$ in Equation (2.5) adapt to the values of the current estimates of $\omega_{k,k'}$ and the overall level of sparsity in Ω, captured by η.

Finally, we note that this proposed framework for simultaneous variable and covariance selection can easily be modified to estimate B when Ω is known and to estimate Ω when B is known.

2.3. Dynamic Posterior Exploration

Given any specification of hyper-parameters $(a_\theta, b_\theta, a_\eta, b_\eta)$ and $(\lambda_1, \lambda_0, \xi_1, \xi_0)$, it is straightforward to deploy the ECM algorithm described in the previous section to identify a putative posterior mode. We may moreover run our algorithm over a range of hyper-parameter settings to estimate the mode of a range of different posteriors. Unlike MCMC, which expends considerable computational effort sampling from a single posterior, this *dynamic posterior exploration* provides a snapshot of several different posteriors. In the univariate regression setting, Ročková and George (2016) proposed a path-following scheme in which they fixed λ_1 and identified modes of a range of posteriors indexed by a ladder of increasing λ_0 values, $\mathcal{I}_\lambda = \left\{\lambda_0^{(1)} < \cdots < \lambda_0^{(L)}\right\}$ with sequential re-initialization to produce a sequence of posterior modes. To find the mode corresponding to $\lambda_0 = \lambda_0^{(s)}$, they "warm started" from the previously discovered mode corresponding to $\lambda_0 = \lambda_0^{(s-1)}$. Early in this path-following scheme, when λ_0 is close to λ_1, distinguishing relevant parameters from negligible is difficult as the spike and slab distributions are so similar. As λ_0 increases, however, the spike distribution increasingly absorbs the negligible values and results in sparser posterior modes. Remarkably, Ročková and George (2016) found that the trajectories of individual parameter estimates tended to stabilize relatively early in the path, indicating that the parameters had cleanly segregated into groups of zero and non-zero values. This is quite evident in Figure 2 (a reproduction of Figure 2c of Ročková and George (2016)), which shows the trajectories of several parameter estimates as a function of λ_0.

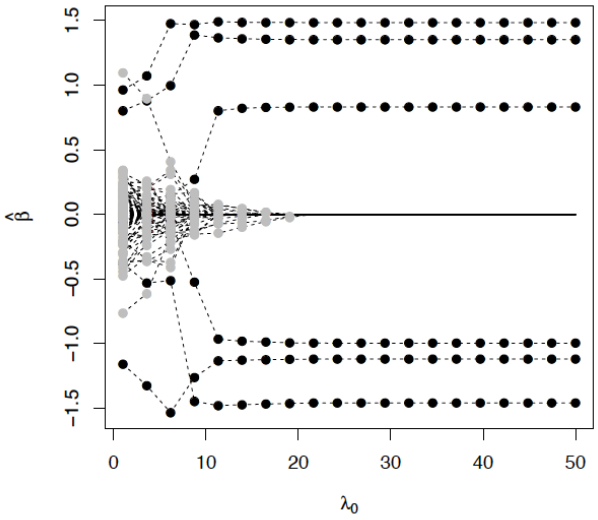


Figure 2: Trajectory of parameter estimates in Ročková and George (2016)'s dynamic posterior exploration.

The stabilization evident in Figure 2 allowed them to focus on and report a single model out of the L that they computed without the need for cross-validation. From a practitioner's point of view, the stabilization of the path-following scheme sidesteps the issue of picking just the right λ_0: one may specify a ladder spanning a wide range of λ_0 values and observe whether or not the trajectories stabilize after a certain point. If so, one may then report any stable estimate and if not, one can expand the ladder to include even larger values of λ_0. It may be helpful to compare dynamic posterior exploration pre-stabilization to focusing a camera lens: starting from a blurry image, turning the focus ring slowly brings an image into relief, with the salient features becoming increasingly prominent. In this way, the priors serve more as filters for the data likelihood than as encapsulations of any real subjective beliefs.

Building on this dynamic posterior exploration strategy for our multivariate setting, we begin by specifying ladders $\mathcal{I}_\lambda = \left\{\lambda_0^{(1)} < \cdots < \lambda_0^{(L)}\right\}$ and $\mathcal{I}_\xi = \left\{\xi_0^{(1)} < \cdots < \xi_0^{(L)}\right\}$ of increasing λ_0 and ξ_0 values. We then identify a sequence $\left\{\hat{\Xi}^{s,t} : 1 \leq s, t \leq L\right\}$, where $\hat{\Xi}^{s,t}$ is an estimate of the mode of the posterior corresponding to the choice $(\lambda_0, \xi_0) = (\lambda_0^{(s)}, \xi_0^{(t)})$,

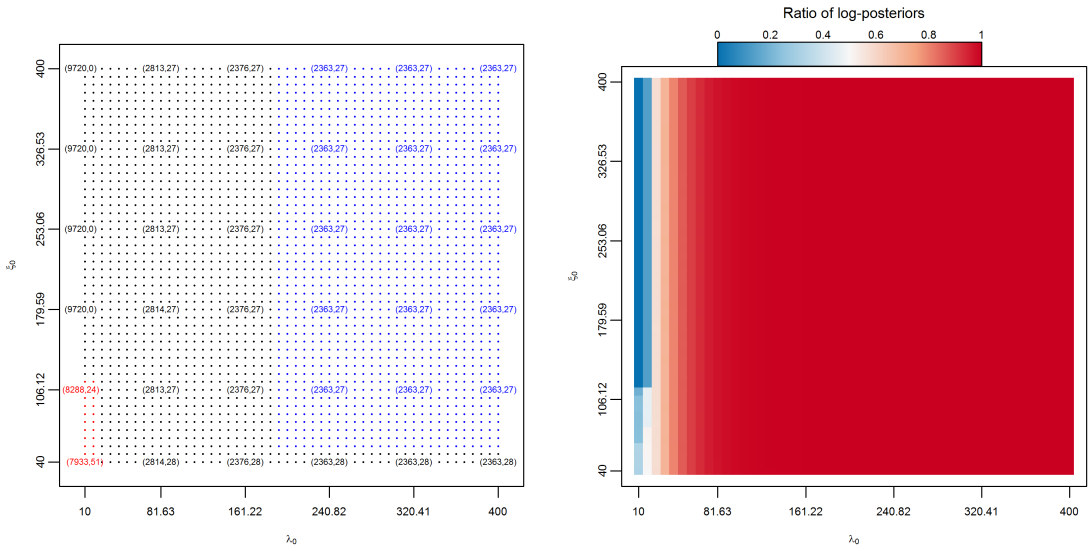
which we denote $\Xi^{s,t*}$. When it comes time to estimate $\Xi^{s,t*}$, we launch our ECM algorithm from whichever of $\hat{\Xi}^{s-1,t}, \hat{\Xi}^{s,t-1}$ and $\hat{\Xi}^{s-1,t-1}$ has the largest log-posterior density, computed according to Equation (2.2) with $\lambda_0 = \lambda_0^{(s)}$ and $\xi_0 = \xi_0^{(t)}$. We implement this dynamic posterior exploration by starting with $B = \mathbf{0}, \Omega = I$ and looping over the λ_0^s values and ξ_0^t values. Proceeding in this way, we propagate a single estimate of Ξ through a series of prior filters indexed by the pair $\left(\lambda_0^{(s)}, \xi_0^{(t)}\right)$.

When λ_0 is close to λ_1, our refined coordinate ascent can sometimes promote the inclusion of many negligible but non-null $\beta_{j,k}$'s. Such a specification combined with a ξ_0 that is much larger than ξ_1, could over-explain the variation in $\mathbf{Y}$ using several covariates, leaving very little to the residual conditional dependency structure and a severely ill-conditioned residual covariance matrix S. In our implementation, we do not propagate any $\hat{\Xi}^{s,t}$ where the accompanying S has condition number exceeding $10n$. While this choice is decidedly arbitrary, we have found it to work rather well in simulation studies. When it comes time to estimate $\Xi^{s,t*}$, if each of $\hat{\Xi}^{s-1,t}, \hat{\Xi}^{s,t-1}$ and $\hat{\Xi}^{s-1,t-1}$ is numerically unstable, we re-launch our EM algorithm from $B = \mathbf{0}$ and $\Omega = I$.

To illustrate this procedure, which we call mSSL-DPE for "Multivariate Spike-and-Slab LASSO with Dynamic Posterior Exploration," we simulate data from the following model with $n = 400, p = 500$, and $q = 25$. We ran mSSL-DPE taking $\mathcal{I}_\lambda$ and $\mathcal{I}_\xi$ to contain 50 evenly spaced points ranging from 1 to n and $0.1n$ and n, respectively. We generate the matrix $\mathbf{X}$ according to a $\mathrm{N}_p(\mathbf{0}_p, \Sigma_X)$ distribution where $\Sigma_X = \left(0.7^{|j-j'|}\right)_{j,j'=1}^{p}$. We construct matrix B_0 with $pq/5$ randomly placed non-zero entires independently drawn uniformly from the interval $[-2, 2]$. This allows us to gauge mSSL-DPE's ability to recover signals of varying strength. We then set $\Omega_0^{-1} = \left(0.9^{|k-k'|}\right)_{k,k'=1}^{q}$ so that Ω_0 is tri-diagonal, with all $\|\Omega_0\|_0 = q - 1$ non-zero entries immediately above the diagonal. Finally, we generate data $\mathbf{Y} = \mathbf{X}B_0 + E$ where the rows of E are independently $\mathrm{N}\left(\mathbf{0}_q, \Omega_0^{-1}\right)$. For this simulation, we set $\lambda_0 = 1, \xi_0 = 0.01n$ and set I_λ and I_ξ to contain $L = 50$ equally spaced values ranging from 1 to n and from $0.1n$ to n, respectively.

In order to establish posterior consistency in the univariate linear regression, Ročková and George (2016) required the prior on θ to place most of its probability in a small interval near zero and recommended taking $a_\theta = 1$ and $b_\theta = p$. This concentrates their prior on models that are relatively sparse. With pq coefficients in B, we take $a_\theta = 1$ and $b_\theta = pq$ for this demonstration. We further take $a_\eta = 1$ and $b_\eta = q$, so that the prior on the underlying residual Gaussian graph G concentrates on very sparse graphs with average degree just less than one. We will consider the sensitivity of our results to these choices briefly in the next subsection.

Figure 3a shows the trajectory of the number of non-zero $\beta_{j,k}$'s and $\omega_{k,k'}$'s identified at a subset of putative modes $\hat{\Xi}^{s,t}$. Points corresponding to numerically unstable modes were colored red and points corresponding to those $\hat{\Xi}^{s,t}$ for which the estimated supports of B and Ω were identical to the estimated supports at $\hat{\Xi}^{L,L}$, were colored blue. Figure 3a immediately suggests a certain stabilization of our multivariate dynamic posterior exploration. In addition to looking at $\left\|\hat{B}\right\|_0$ and $\left\|\hat{\Omega}\right\|_0$, we can look at the log-posterior density of each $\hat{\Xi}^{s,t}$ computed with $\lambda_0 = \lambda_0^{(L)}, \xi_0 = \xi_0^{(L)}$. Figure 3b plots a heat map of the ratio $\frac{\log \pi(\hat{\Xi}^{s,t}|\mathbf{Y})/\pi(\hat{\Xi}^{0,0}|\mathbf{Y})}{\log \pi(\hat{\Xi}^{L,L}|\mathbf{Y})/\pi(\hat{\Xi}^{0,0}|\mathbf{Y})}$. It is interesting to note that this ratio appears to stabilize before the supports did.



(a) (b)

Figure 3: (a)Trajectory of $(\|B\|_0, \|\Omega\|_0)$, (b) Trajectory of $\frac{\log\left(\pi(\hat{\Xi}^{s,t}|\mathbf{Y})/\pi(\hat{\Xi}^{0,0}|\mathbf{Y})\right)}{\log\left(\pi(\hat{\Xi}^{L,L}|\mathbf{Y})/\pi(\hat{\Xi}^{0,0}|\mathbf{Y})\right)}$

The apparent stabilization in Figure 3 allows us to focus on and report a single estimate $\hat{\Xi}^{L,L}$, corresponding to the top-right point in Figure 3a, avoiding costly cross-validation. Of course, this estimate is nearly indistinguishable from the estimates corresponding to the other blue points in Figure 3a and we could just as easily report any one of them. On this dataset, mSSL-DPE correctly identified 2360 out of the 2500 non-zero $\beta_{j,k}$'s with only 3 false positives and correctly identified all 24 non-zero $\omega_{k,k'}$'s in the upper triangle of Ω, again with only 3 false positive identifications. We should point out that there is no general guarantee of stabilization for arbitrary ladders $\mathcal{I}_\lambda$ and $\mathcal{I}_\xi$. However, in all of the examples we have tried, we found that stabilization occurred long before $\lambda_0^{(s)}$ and $\xi_0^{(t)}$ reached $\lambda_0^{(L)} = \xi_0^{(L)} = n$. We should add that once the solutions stabilize, the algorithm runs quite quickly so the excess computations are not at all burdensome.

2.3.1. Faster Dynamic Conditional Posterior Mode Exploration

mSSL-DPE can expend considerable computational effort identifying modal estimates $\hat{\Xi}^{s,t}$ corresponding to smaller values of λ_0 and ξ_0. Although the support recovery performance of

$\hat{\Xi}^{L,L}$ from mSSL-DPE is very promising, one might also consider streamlining the procedure using the following procedure we term mSSL-DCPE for "Dynamic *Conditional* Posterior Exploration." First, we fix $\Omega = I$ and sequentially solve Equation (2.3) for each $\lambda_0 \in I_\lambda$, with warm-starts. This produces a sequence $\left\{\left(\hat{B}^s, \hat{\theta}^s\right)\right\}$ of *conditional* posterior modes of $(B, \theta)\,|\mathbf{Y}, \Omega = I$. Then, holding $(B, \theta) = (\hat{B}_0^L, \hat{\theta}_0^L)$ fixed, we run a modified version of our dynamic posterior exploration to produce a sequence $\left\{\left(\hat{\Omega}^t, \hat{\eta}^t\right)\right\}$ of conditional modes of $(\Omega, \eta)\,|\mathbf{Y}, B = \hat{B}^L$. We finally run our ECM algorithm from $\left(\hat{B}^L, \hat{\theta}^L, \hat{\Omega}^L, \hat{\eta}^L\right)$ with $\lambda_0 = \lambda_0^L$ and $\xi_0 = \xi_0^L$ to arrive at an estimate of $\Xi^{L,L*}$, which we denote $\tilde{\Xi}^{L,L}$. We note that the estimate returned by mSSL-DCPE, $\hat{\Xi}^{L,L}$ usually does not coincide with $\tilde{\Xi}^{L,L}$. This is because, generally speaking, when it comes time to estimate $\Xi^{L,L*}$, mSSL-DPE and mSSL-DCPE launch the ECM algorithm from different starting points.

In sharp contrast mSSL-DPE, which visits several joint posterior modes before reaching an estimate of posterior mode $\Xi^{L,L*}$, mSSL-DCPE visits several conditional posterior modes to reach another estimate of the same mode. On the same dataset from the previous subsection, mSSL-DCPE correctly identified 2,169 of the 2,500 non-zero $\beta_{j,k}$ with 8 false positives and all 24 non-zero $\omega_{k,k'}$'s but with 28 false positives. This was all accomplished in just under 30 seconds, a considerable improvement over the two hour runtime of mSSL-DPE on the same dataset. Despite the obvious improvement in runtime, mSSL-DCPE terminated at a sub-optimal point whose log-posterior density was much smaller than the solution found by mSSL-DPE. All of the false negative identifications in the support of B made by both procedures corresponded to $\beta_{j,k}$ values which were relatively small in magnitude. Interestingly, mSSL-DPE was better able to detect smaller signals than mSSL-DCPE. We will return to this point later in Section 2.3.2.

2.3.2. Simulations

We now assess the performance of mSSL-DPE and mSSL-DCPE on data simulated from two models, one low-dimensional with $n = 100, p = 50$ and $q = 25$ and the other somewhat high-dimensional with $n = 400, p = 500, q = 25$. Just as above, we generate the matrix $\mathbf{X}$

according to a $\mathrm{N}_p(\mathbf{0}_p, \Sigma_X)$ distribution where $\Sigma_X = \left(0.7^{|j-j'|}\right)_{j,j'=1}^{p}$. We construct matrix B_0 with $pq/5$ randomly placed non-zero entires independently drawn uniformly from the interval $[-2, 2]$. We then set $\Omega_0^{-1} = \left(\rho^{|k-k'|}\right)_{k,k'=1}^{q}$ for $\rho \in \{0, 0.5, 0.7, 0.9\}$. When $\rho \neq 0$, the resulting Ω_0 is tri-diagonal. Finally, we generate data $\mathbf{Y} = \mathbf{X}B_0 + E$ where the rows of E are independently $\mathrm{N}\left(\mathbf{0}_q, \Omega_0^{-1}\right)$. For this simulation, we set $\lambda_1 = 1, \xi_1 = 0.01n$ and set $\mathcal{I}_\lambda$ and $\mathcal{I}_\xi$ to contain $L = 10$ equally spaced values ranging from 1 to n and from $0.1n$ to n, respectively.

We simulated 50 datasets according to each model, each time keeping B_0 and Ω_0 fixed but drawing a new matrix of errors E. To assess the support recovery and estimation performance, we tracked the following quantities: SEN (sensitivity), SPE (specificity), PREC (precision), ACC (accuracy), MCC (Matthew's Correlation Coefficient), MSE (mean square error in estimating B_0), FROB (squared Frobenius error in estimating Ω_0), and TIME (execution time in seconds). If we let TP, TN, FP, and FN denote the total number of true positive, true negative, false positive, and false negative identifications made in the support recovery, these quantities are defined as:

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$
$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \qquad \mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

and

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}.$$

Table 1 – 4 reports the average performance, in both low- and high-dimesnional settings, of mSSL-DPE, mSSL-DCPE, Rothman et al. (2010)'s MRCE procedure, Cai et al. (2013)'s CAPME procedure, each with 5-fold cross validation, and the following two competitors:

Sep.L+G: We first estimate B by solving separate LASSO problems with 10-fold cross-validation for each outcome. We then estimate Ω from the resulting residual matrix

using the GLASSO procedure of Friedman et al. (2008), also run with 10-fold cross-validation

Sep.SSL + SSG: We first estimate B column-by-column, deploying Ročková and George (2016)'s path-following SSL along the ladder $\mathcal{I}_\lambda$ separately for each outcome. We then run a modified version of our dynamic posterior exploration that holds B fixed and only updates Ω and η with the ECM algorithm along the ladder $\mathcal{I}_\xi$. This is similar to **Sep.L+G** but with adaptive spike-and-slab lasso penalties rather than fixed ℓ_1 penalties.

In the previous subsection, we took $a_\theta = 1, b_\theta = pq, a_\eta = 1$ and $b_\eta = q$. These hyper-parameters placed quite a lot of prior probability on rather sparse B's and Ω's. Earlier we observed that with such specification we achieved reasonably good support recovery of the true sparse B and Ω. The extent to which our prior specification drove this recovered sparsity is not immediately clear. Put another way, were our sparse estimates of B and Ω truly "discovered" or were they "manufactured" by the prior concentrating on sparse matrices? To investigate this possibility, we ran mSSL-DPE and mSSL-DCPE for the two choices of $(b_\theta, b_\eta) = (1, 1)$ and $(b_\theta, b_\eta) = (pq, q)$, keeping $a_\theta = a_\eta = 1$. In Tables 1 – 4 mSSL-DPE(pq,q) and mSSL-DPE(1,1) correspond to the different settings of the hyper-parameters (b_θ, b_η), with $a_\theta = a_\eta = 1$.

Table 1: Variable selection and estimation performance of several methods in low-dimensional settings. NaN indicates that the specified quantity was undefined, either because no non-zero estimates were returned or because there were truly no non-zero parameters (Simulation 4). MSE has been re-scaled by a factor of 1000 and TIME is measured in seconds

Method	SEN/SPE	PRE/ACC	MCC	MSE	TIME
Simulation 1: $n = 100, p = 50, q = 25, \rho = 0.9$					
mSSL-DPE(pq, q)	0.87 / **1.00**	**1.00** / **0.97**	**0.92**	**1.04**	8.61
mSSL-DPE(1,1)	0.88 / **1.00**	0.99 / **0.97**	**0.92**	1.30	9.17
mSSL-DCPE (pq,q)	0.74 / **1.00**	0.99 / 0.95	0.83	6.59	0.41
mSSL-DCPE(1,1)	0.75 / **1.00**	0.98 / 0.95	0.83	5.61	0.30
MRCE	0.86 / 0.71	0.43 / 0.74	0.47	33.91	1406.71
CAPME	**0.96** / 0.23	0.24 / 0.38	0.20	26.92	139.02
SEP.L+G	0.84 / 0.84	0.57 / 0.84	0.60	17.71	2.58
SEP.SSL+SSG	0.73 / **1.00**	0.98 / 0.94	0.82	8.98	**0.07**
Simulation 2: $n = 100, p = 50, q = 25, \rho = 0.7$					
mSSL-DPE(pq,q)	0.81 / **1.00**	**0.99** / **0.96**	0.87	3.47	2.04
mSSL-DPE(1,1)	0.82 / **1.00**	0.98 / **0.96**	**0.88**	**3.30**	1.72
mSSL-DCPE(pq,q)	0.73 / **1.00**	**0.99** / 0.94	0.82	7.41	0.25
mSSL-DCPE(1,1)	0.74 / **1.00**	**0.99** / 0.95	0.83	6.47	0.18
MRCE	**0.89** / 0.66	0.41 / 0.71	0.45	18.30	1532.75
CAPME	0.86 / 0.75	0.47 / 0.77	0.51	23.94	140.14
SEP.L+G	0.85 / 0.84	0.57 / 0.84	0.60	17.59	2.63
SEP.SSL+SSG	0.73 / **1.00**	**0.99** / 0.94	0.82	8.61	**0.06**
Simulation 3: $n = 100, p = 50, q = 25, \rho = 0.5$					
mSSL-DPE(pq,q)	0.76 / **1.00**	**0.99** / **0.95**	0.84	5.98	1.32
mSSL-DPE(1,1)	0.78 / 0.99	0.97 / **0.95**	**0.85**	**5.53**	1.11
mSSL-DCPE(pq,q)	0.73 / **1.00**	**0.99** / 0.94	0.82	8.33	0.22
mSSL-DCPE(1,1)	0.74 / **1.00**	0.98 / **0.95**	0.83	7.54	0.16
MRCE	**0.91** / 0.66	0.40 / 0.71	0.46	9.86	664.97
CAPME	0.86 / 0.77	0.48 / 0.78	0.52	23.41	144.80
SEP.L+G	0.85 / 0.84	0.57 / 0.84	0.60	171.3	2.92
SEP.SSL+SSG	0.73 / **1.00**	**0.99** / 0.94	0.82	8.36	**0.05**
Simulation 4: $n = 100, p = 50, q = 25, \rho = 0$					
mSSL-DPE(pq,q)	0.73 / **1.00**	**0.99** / 0.94	0.82	8.90	0.54
mSSL-DPE(1,1)	0.75 / **1.00**	0.98 / **0.95**	**0.83**	**8.06**	0.57
mSSL-DCPE(pq,q)	0.72 / **1.00**	**0.99** / 0.94	0.82	9.07	0.13
mSSL-DCPE(1,1)	0.75 / **1.00**	0.98 / 0.95	**0.83**	8.07	0.13
MRCE	**0.90** / 0.66	0.40 / 0.70	0.45	13.11	537.90
CAPME	0.86 / 0.75	0.47 / 0.78	0.51	22.96	144.31
SEP.L+G	0.84 / 0.84	0.57 / 0.84	0.60	17.39	2.38
SEP.SSL+SSG	0.73 / **1.00**	**0.99** / 0.94	0.82	8.54	**0.05**

Table 2: Covariance selection and estimation performance of several methods in the low-dimensional setting. In these settings, the `R` implementation of MRCE returned errors indicating over-fitting. NaN indicates that the specified quantity was undefined, either because no non-zero estimates were returned or because there were truly no non-zero parameters (Simulations 4). TIME is reported in seconds.

Method	SEN/SPE	PREC/ACC	MCC	FROB	TIME
Simulation 1: $n = 100, p = 50, q = 25, \rho = 0.9$					
mSSL-DPE(pq, q)	**1.00** / **1.00**	**0.95** / **1.00**	**0.97**	**116.27**	8.61
mSSL-DPE(1,1)	0.98 / 0.99	0.92 / 0.99	0.94	140.51	9.17
mSSL-DCPE (pq,q)	0.79 / 0.96	0.62 / 0.94	0.67	1151.81	0.41
mSSL-DCPE(1,1)	0.83 / 0.96	0.63 / 0.95	0.69	988.30	0.30
MRCE	0.96 / 0.73	0.24 / 0.75	0.40	669.81	1406.71
CAPME	**1.00** / 0.00	0.08 / 0.08	NaN	2323.08	139.02
SEP.L+G	0.99 / 0.67	0.21 / 0.70	0.37	2521.15	2.58
SEP.SSL+SSG	0.79 / 0.96	0.63 / 0.95	0.68	1468.96	**0.07**
Simulation 2: $n = 100, p = 50, q = 25, \rho = 0.7$					
mSSL-DPE(pq,q)	**1.00** / **1.00**	**1.00** / **1.00**	**1.00**	**8.66**	2.04
mSSL-DPE(1,1)	**1.00** / **1.00**	**1.00** / **1.00**	**1.00**	9.46	1.72
mSSL-DCPE(pq,q)	0.95 / **1.00**	0.94 / 0.99	0.94	28.53	0.25
mSSL-DCPE(1,1)	0.96 / **1.00**	0.95 / 0.99	0.95	21.43	0.18
MRCE	**1.00** / 0.78	0.33 / 0.80	0.50	26.41	1532.75
CAPME	0.98 / 0.42	0.13 / 0.47	0.23	89.81	140.14
SEP.L+G	**1.00** / 0.78	0.29 / 0.80	0.47	141.14	2.63
SEP.SSL+SSG	0.94 / **1.00**	0.95 / 0.99	0.94	40.60	**0.06**
Simulation 3: $n = 100, p = 50, q = 25, \rho = 0.5$					
mSSL-DPE(pq,q)	0.90 / **1.00**	0.98 / **0.99**	**0.94**	**5.62**	1.32
mSSL-DPE(1,1)	0.92 / **1.00**	0.98 / **0.99**	**0.94**	6.13	1.11
mSSL-DCPE(pq,q)	0.28 / **1.00**	**1.00** / 0.94	0.72	23.03	0.22
mSSL-DCPE(1,1)	0.45 / **1.00**	0.99 / 0.96	0.79	17.25	0.16
MRCE	**1.00** / 0.82	0.33 / 0.83	0.51	6.63	664.97
CAPME	0.99 / 0.36	0.12 / 0.41	0.20	15.29	144.80
SEP.L+G	0.98 / 0.83	0.34 / 0.84	0.52	25.38	2.92
SEP.SSL+SSG	0.57 / **1.00**	0.99 / 0.97	0.74	13.78	**0.05**
Simulation 4: $n = 100, p = 50, q = 25, \rho = 0$					
mSSL-DPE(pq,q)	0.73 / **1.00**	**0.99** / 0.94	0.82	8.90	0.54
mSSL-DPE(1,1)	0.75 / **1.00**	0.98 / **0.95**	**0.83**	**8.06**	0.57
mSSL-DCPE(pq,q)	0.72 / **1.00**	**0.99** / 0.94	0.82	9.07	0.13
mSSL-DCPE(1,1)	0.75 / **1.00**	0.98 / 0.95	**0.83**	8.07	0.13
MRCE	**0.90** / 0.66	0.40 / 0.70	0.45	13.11	537.90
CAPME	0.86 / 0.75	0.47 / 0.78	0.51	22.96	144.31
SEP.L+G	0.84 / 0.84	0.57 / 0.84	0.60	17.39	2.38
SEP.SSL+SSG	0.73 / **1.00**	**0.99** / 0.94	0.82	8.54	**0.05**

Table 3: Variable selection and estimation performance of several methods in the high-dimensional setting. In these settings, the `R` implementation of MRCE returned errors indicating over-fitting. TIME is reported in seconds.MSE has been re-scaled by a factor of 1000 and TIME is reported in seconds.

Method	SEN/SPE	PRE/ACC	MCC	MSE	TIME
Simulation 5: $n = 400, p = 500, q = 25, \rho = 0.9$					
mSSL-DPE(pq,q)	**0.95** / **1.00**	**1.00** / **0.99**	**0.97**	**0.24**	841.33
mSSL-DPE(1,1)	**0.95** / **1.00**	0.99 / **0.99**	0.96	0.58	2510.22
mSSL-DCPE(pq,q)	0.88 / **1.00**	0.99 / 0.97	0.92	1.40	27.19
mSSL-DCPE(1,1)	0.89 / **1.00**	0.99 / 0.98	0.92	1.25	23.65
CAPME	**0.95** / 0.54	0.34 / 0.62	0.40	8.56	6991.55
SEP.L+G	0.92 / 0.76	0.48 / 0.79	0.56	10.32	20.48
SEP.SSL+SSG	0.88 / **1.00**	0.98 / 0.97	0.91	2.28	**3.16**
Simulation 6: $n = 400, p = 500, q = 25, \rho = 0.7$					
mSSL-DPE(pq,q)	**0.92** / **1.00**	0.98 / **0.98**	0.94	**0.86**	2082.44
mSSL-DPE(1,1)	**0.92** / **1.00**	0.98 / 0.94	0.94	1.22	2680.39
mSSL-DCPE(pq,q)	0.88 / 1.00	**0.99** / 0.97	**0.97**	1.63	27.89
mSSL-DCPE(1,1)	0.89 / 1.00	0.98 / 0.97	0.92	1.54	23.09
CAPME	0.67 / 0.84	0.53 / 0.81	0.48	110.13	7601.53
SEP.L+G	**0.92** / 0.76	0.48 / 0.79	0.56	10.25	20.88
SEP.SSL+SSG	0.88 / **1.00**	0.98 / 0.97	0.91	2.23	**3.06**
Simulation 7: $n = 400, p = 500, q = 25, \rho = 0.5$					
mSSL-DPE(pq,q)	0.91 / 0.60	0.38 / 0.66	0.42	34.23	3803.50
mSSL-DPE(1,1)	**0.92** / 0.55	0.35 / 0.62	0.38	35.59	3888.59
mSSL-DCPE(pq,q)	0.88 / **1.00**	**0.99** / **0.97**	**0.92**	1.91	23.87
mSSL-DCPE(1,1)	0.89 / 0.99	0.98 / **0.97**	0.91	**1.82**	23.29
CAPME	0.65 / 0.86	0.54 / 0.82	0.48	116.42	7200.09
SEP.L+G	**0.92** / 0.76	0.49 / 0.79	0.56	10.21	20.23
SEP.SSL+SSG	0.88 / **1.00**	0.98 / **0.97**	0.91	2.21	**3.13**
Simulation 8: $n = 400, p = 500, q = 25, \rho = 0$					
mSSL-DPE(pq,q)	0.91 / 0.58	0.35 / 0.64	0.39	36.26	2759.05
mSSL-DPE(1,1)	**0.92** / 0.54	0.33 / 0.62	0.37	36.61	2766.72
mSSL-DCPE(pq,q)	0.88 / **1.00**	**0.98** / **0.97**	**0.91**	2.22	23.11
mSSL-DCPE(1,1)	0.89 / 0.99	0.97 / **0.97**	**0.91**	**2.20**	22.78
CAPME	0.66 / 0.86	0.54 / 0.82	0.48	116.46	7435.93
SEP.L+G	**0.92** / 0.76	0.49 / 0.79	0.56	10.28	19.13
SEP.SSL+SSG	0.88 / **1.00**	**0.98** / **0.97**	**0.91**	2.21	**3.13**

Table 4: Covariance selection and estimation performance of several methods in the high-dimensional setting. In these settings, the `R` implementation of MRCE returned errors indicating over-fitting. NaN indicates that the specified quantity was undefined, either because no non-zero estimates were returned or because there were truly no non-zero parameters (Simulation 8). TIME is reported in seconds.

Method	SEN/SPE	PREC/ACC	MCC	FROB	TIME
Simulation 5: $n = 400, p = 500, q = 25, \rho = 0.9$					
mSSL-DPE(pq,q)	**1.00** / 0.98	**0.85** / **0.99**	**0.91**	**25.88**	841.33
mSSL-DPE(1,1)	0.96 / 0.98	0.83 / 0.98	0.88	126.77	2510.22
mSSL-DCPE(pq,q)	**1.00** / 0.89	0.44 / 0.90	0.63	1228.97	27.19
mSSL-DCPE(1,1)	1.00 / 0.89	0.45 / 0.90	0.63	1066.24	23.65
CAPME	0.00 / **1.00**	NaN / 0.92	NaN	2989.08	6991.55
SEP.L+G	**1.00** / 0.60	0.18 / 0.63	0.32	2684.93	20.48
SEP.SSL+SSG	0.99 / 0.87	0.40 / 0.88	0.59	1959.21	**3.16**
Simulation 6: $n = 400, p = 500, q = 25, \rho = 0.7$					
mSSL-DPE(pq,q)	**1.00** / **1.00**	**0.97** / **1.00**	**0.98**	24.89	2082.44
mSSL-DPE(1,1)	0.99 / 0.99	0.94 / 0.99	0.96	31.13	2680.39
mSSL-DCPE(pq,q)	**1.00** / 0.96	0.71 / 0.97	0.83	14.29	27.89
mSSL-DCPE(1,1)	**1.00** / 0.96	0.72 / 0.97	0.83	**9.97**	23.09
CAPME	0.00 / **1.00**	NaN / 0.92	NaN	286.65	7601.53
SEP.L+G	0.99 / 0.87	0.40 / 0.88	0.58	161.92	20.88
SEP.SSL+SSG	**1.00** / 0.96	0.70 / 0.96	0.82	57.88	**3.06**
Simulation 7: $n = 400, p = 500, q = 25, \rho = 0.5$					
mSSL-DPE(pq,q)	0.05 / **1.00**	NaN / 0.92	NaN	36484.42	3803.50
mSSL-DPE(1,1)	0.02 / **1.00**	0.96 / 0.92	0.98	75456.74	3888.59
mSSL-DCPE(pq,q)	**1.00** / **1.00**	0.97 / **1.00**	0.98	**2.15**	23.87
mSSL-DCPE(1,1)	**1.00** / **1.00**	0.97 / **1.00**	0.98	3.08	23.29
CAPME	0.00 / **1.00**	NaN / 0.92	NaN	87.13	7200.09
SEP.L+G	0.86 / 0.96	0.65 / 0.95	0.72	29.30	20.23
SEP.SSL+SSG	**1.00** / **1.00**	**0.98** / **1.00**	**0.99**	4.34	**3.13**
Simulation 8: $n = 400, p = 500, q = 25, \rho = 0$					
mSSL-DPE(pq,q)	NaN / **1.00**	NaN / **1.00**	NaN	40646.23	2759.05
mSSL-DPE(1,1)	NaN / **1.00**	NaN / **1.00**	NaN	787543.36	2766.72
mSSL-DCPE(pq,q)	NaN / **1.00**	NaN / **1.00**	NaN	1.17	23.11
mSSL-DCPE(1,1)	NaN / **1.00**	NaN / **1.00**	NaN	1.81	22.78
CAPME	NaN / **1.00**	NaN / **1.00**	NaN	24.00	7435.93
SEP.L+G	NaN / 0.99	0.00 / 0.99	NaN	10.28	19.13
SEP.SSL+SSG	NaN / **1.00**	0.00 / **1.00**	NaN	**1.14**	**3.13**

In both the high- and low-dimensional settings, we see immediately that the regularization methods utilizing cross-validation (MRCE, CAPME, and SEP.L+G) are characterized by high sensitivity, moderate specificity, and low precision in recovering the support of both B and Ω. The fact that the precisions of these three methods are less than 0.5 highlights the fact that the majority of the non-zero estimates returned are in fact false positives, a

rather unattractive feature from a practitioner's standpoint! This is not entirely surprising, as cross-validation has a well-known tendency to over-select. In stark contrast are mSSL-DPE, mSSL-DCPE, and SEP.SSL+SSG, which all utilized adaptive spike-and-slab penalties. These methods are all characterized by somewhat lower sensitivity than their cross-validated counterparts but with vastly improved specificity and precision, performing exactly as anticipated by Ročková and George (2016)'s simulations from the univariate setting. In a certain sense, the regularization competitors cast a very wide net in order to capture most of the non-zero parameters, while our methods are much more discerning. So while the latter methods may not capture as much of the true signal as the former, they do not admit nearly as many false positives.

CAPME, SEP.L+G, and SEP.SSL+SSG all estimate B in a column-wise fashion and are incapable of "borrowing strength" across outcomes. MRCE and mSSL-DPE are the only two methods considered that explicitly leverage the residual correlation between outcomes from the outset. As noted above, in the low-dimensional settings, MRCE tended to over-select in B and Ω, leading to rather poor estimates of both matrices. Moreover, in Simulations 5 – 8, the standard `R` implementation of MRCE returned errors indicating over-fitting during the cross-validation. In all but Simulations 7 and 8, mSSL-DPE displayed far superior estimation and support recovery performance than MRCE.

Recall that mSSL-DCPE proceeds by finding a conditional mode $\left(\hat{B}^L, \hat{\theta}^L\right)$ fixing $\Omega = I$, finding a conditional mode $\left(\hat{\Omega}^L, \hat{\eta}^L\right)$ fixing $B = \hat{B}^L$, and then refining these two conditional modes to a single joint mode. It is only in this last refining step that mSSL-DCPE introduces the correlation between residuals to its estimation of B. As it turns out, this final refinement did little to change the estimated support of B, so the nearly identical performance of SEP.SSL+SSG and mSSL-DCPE is not that surprising. Further, the only practical difference between the two procedures is the adaptivity of the penalties on $\beta_{j,k}$: in SEP.SSL+SSG, the penalties separately adapt to the sparsity within each column of B while in mSSL-DCPE, they adapt to the overall sparsity of B.

By simulating the non-zero $\beta_{j,k}$'s uniformly from $[-2, 2]$, we were able to compare our methods' abilities to detect signals of varying strength. Figure 4 super-imposes the distribution non-zero $\beta_{j,k}$'s correctly identified as non-zero with the distribution of non-zero $\beta_{j,k}$'s incorrectly estimated as zero by each of mSSL-DPE, mSSL-DCPE, and SEP.SSL+SSG from a single replication of Simulation 5.

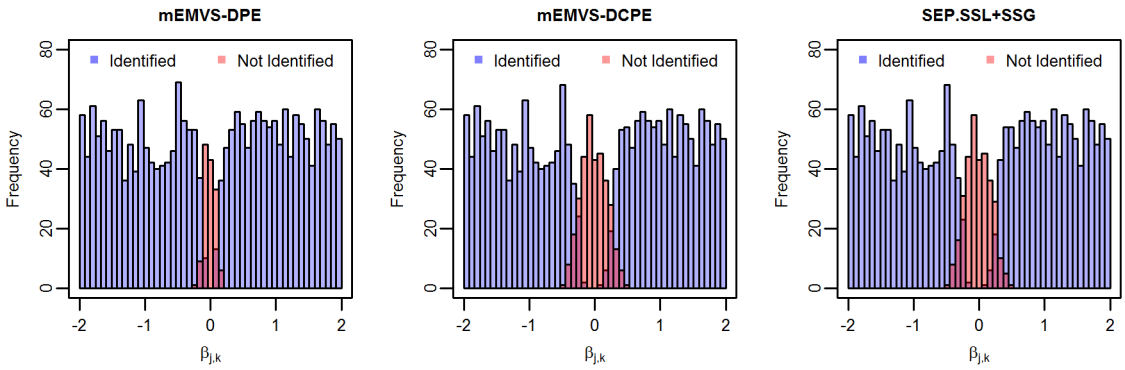


Figure 4: Histograms of non-zero $\beta_{j,k}$ values that are correctly identified as non-zero (blue) and non-zero $\beta_{j,k}$ values incorrectly identified as zero (red). mSSL-DPE demonstrates the greatest acuity in recovering small $\beta_{j,k}$ values.

In this situation, mSSL-DPE displays greater acuity for detecting smaller $\beta_{j,k}$'s than mSSL-DCPE or SEP.SSL+SSG, which are virtually ignorant of the covariance structure of the outcomes. This is very reminiscent of Zellner (1962)'s observation that multivariate estimation of B in seemingly unrelated regressions is asymptotically more efficient than proceeding response-by-response and ignoring the correlation between responses. To get a better sense as to why this may the case, recall the refined thresholding used to update our estimates of $\beta_{j,k}$ in our ECM algorithm:

$$\beta_{j,k}^{new} = \frac{1}{n}\left(|z_{j,k}| - \lambda^{\star}(\beta_{j,k}^{old}, \theta)\right)_{+} \text{sign}(z_{j,k})\mathbb{I}\left(|z_{j,k}| > \Delta_{j,k}\right).$$

The quantity $z_{j,k}$ can be decomposed as

$$z_{j,k} = n\beta_{j,k}^{old} + \mathbf{x}_j'\mathbf{r}_k(B^{old}) + \sum_{k' \neq k} \frac{\omega_{k,k'}}{\omega_{k,k}} \mathbf{x}_j'\mathbf{r}_{k'}(B^{old}).$$

Writing $z_{j,k}$ in this way, we can readily see how $\omega_{k,k'}$ regulates the degree to which our estimate of $\beta_{j,k}$ depends on the outcome $\mathbf{y}_{k'}$: if $\omega_{k,k'}$ is close in value to $\omega_{k,k}$, our estimate of variable j's impact on outcome k will depend almost much as on the residuals $\mathbf{r}_{k'}$ as they do on the residuals $\mathbf{r}_k$. On the other hand, if $\omega_{k,k'} = 0$, then we are unable to "borrow strength" and use information contained in $\mathbf{y}_{k'}$ to help estimate $\beta_{j,k}$. Non-zero values of $\omega_{k,k'}$ in the sum in the above expression may make it easier for some $z_{j,k}$'s corresponding to small $\beta_{j,k}$ values to overcome the thresholds $\Delta^U_{j,k}$ and $\lambda^\star_{j,k}$ in mSSL-DPE, resulting in far fewer false negative identifications in the support of B than mSSL-DCPE.

Recalling that we took $a_\theta = a_\eta = 1$, we do not observe terribly different results from mSSL-DPE and mSSL-DCPE with the different settings of hyper-parameters b_θ and b_η. It is reassuring to see that we still recover rather good sparse estimates when we took $b_\theta = b_\eta = 1$, though we observe more false positive identifications with these settings.

Finally we must address Simulations 7 and 8, in which mSSL-DPE appears to perform exceptionally poorly. On closer inspection, in all of the replications, mSSL-DPE stabilized immediately at a rather dense estimate of B that left very little residual variance and produced a diagonal estimate of Ω with massive entries on the diagonal. As it turns out, the log-posterior evaluated at this estimate with $(\lambda_0, \xi_0) = \left(\lambda_0^{(L)}, \xi_0^{(L)}\right)$ was considerably smaller than the log-posterior evaluated at mSSL-DCPE's estimate. In other words, mSSL-DCPE was able to escape the "dense B – unstable, diagonal Ω" region of the parameter space and navigate to regions of higher posterior density. In Simulation 7, the truly non-zero $\omega_{k,k'}$'s were rather small and in Simulation 8, Ω was the identity. Taken together, these two simulations suggests that when $p > n$, estimating B and Ω jointly with small values of λ_0 can lead to sub-optimal estimates. In practice, we recommend running both mSSL-DPE and mSSL-DCPE and reporting results of whichever estimate has higher log-posterior.

2.4. Full Multivariate Analysis of the Football Safety Data

More than 1 million high school students played American-style tackle football in 2014, but many medical professionals have recently begun questioning the safety of the sport (Bachynski, 2016; Pfister et al., 2016) or called for its outright ban (Miles and Prasad, 2016). Concern over the long-term safety of the sport have been driven partially by studies like Lehman et al. (2012), which found an increased risk of neurodegenerative disease and Guskiewicz et al. (2005, 2007) and Hart Jr et al. (2013), which highlighted associations between concussion history and later-life cognitive impairment and depression.

In a recent observational study, Deshpande et al. (2017) studied the effect of playing high school football on later-life cognitive and mental health using data from the Wisconsin Longitudinal Study (WLS), which has followed 10,317 people since they graduated from a Wisconsin high school in 1957. In addition to an indicator of participation in high school football, the WLS dataset contains a rich set of baseline variables that may be associated with later-life health, including adolescent IQ, percentile rank in high school, and anticipated years of education. Further, the WLS dataset contains many socio-economic outcomes measured in the mid-1970's, when the participants were in their mid-to-late 30's, as well as results from a battery of cognitive, psychological, and behavioral tests conduced in 1993, 2003-05, and 2011, when the subjects were approximately 54, 65, and 72 years of age. Deshpande et al. (2017) took a univariate approach, analyzing each outcome separately, and found no evidence of a harmful effect of playing high school football on any outcome considered, after carefully adjusting for several important confounders.

We now re-visit the dataset of Deshpande et al. (2017) from a full multivariate perspective with mSSL-DPE and mSSL-DCPE. Our more powerful multivariate methodology not only confirms the main findings of their analysis but also provides new insight into the residual inter-dependence of the cognitive, psychological, and socio-economic outcomes that was otherwise unavailable in their univariate analysis.

In order to isolate the effect of playing football, Deshpande et al. (2017) began by creating matched sets containing one football player and one or more control subjects, or one control subject and one or more football players, using full matching with a propensity score caliper. These matched sets optimally balance the distribution of each baseline variable between football players and controls, and were constructed in such a way that the standardized difference in means between the two groups was less than 0.2 standard deviations. They then regressed several standardized cognitive, psychological, behavioral, and socio-economic outcomes onto the indicator of football participation, the baseline covariates, and indicator variables for matched set inclusions. This allowed them to estimate the effect of playing football with the associated partial slope. This combination of full matching and model-based covariate adjustment has been shown to remove biases due to residual covariate imbalance (Cochran and Rubin, 1973; Silber et al., 2001) in an efficient and robust fashion (see, e.g., Rosenbaum, 2002; Hansen, 2004; Rubin, 1973, 1979).

The cognitive outcomes considered included scores on Letter Fluency (LF), Immediate Word Recall (IWR), Delayed Word Recall (DWR) , Digit Ordering (DO), WAIS Similarity (SIM), and Number Series (NS) tests. All of these tests were administered in both 2003 and 2011, except for SIM which was also administered in 1993 and NS which was only administered in 2011. The psychological and behavioral outcomes included scores on the Center for Epidemiological Studies-Depression scale (CES-D), Anger Index (ANG), Hostility Index (HOS), and Anxiety Index (ANX). CES-D and HOS scores were available from 1993, 2003, and 2011, while ANG and ANX scores were available only in 2003 and 2011. The socio-economic and education outcomes included occupational prestige scores (SEI) for jobs held in 1964, 1970, 1974, and 1975, number of weeks worked (WW) in 1974, earnings (EARN) in 1974, and number of years of education completed by 1974.

We now focus on the $n = 448$ subjects with all available outcomes. Of these 448 subjects, 157 played high school football. Following the broad outline of Deshpande et al. (2017), we first matched football players to controls along several baseline covariates using full

matching and a propensity caliper. Table 5 lists these covariates, along with their pre- and post-matching means and standardized differences for the football players and controls. In all we had 157 matched sets, each comprised of a single football player and up to 6 controls, that adequately balanced the distribution of each baseline covariate. We then standardized each of the $q = 29$ outcomes and regressed them onto the $p = 204$ predictors, which included all of the covariates listed in Table 5 as well as indicators of matched set inclusion. Like the simulation study in Section 2.3.2, we ran mSSL-DPE and mSSL-DCPE with $\mathcal{I}_\lambda$ and $\mathcal{I}_\xi$ containing 10 evenly spaced points ranging from 1 to n and $0.1n$ to n, respectively, and set $a_\theta = a_\eta = 1, b_\theta = pq = 5,916$ and $b_\eta = q = 29$.

Table 5: Baseline covariates, along with pre- and post-matching means and standardized differences

		Control Mean		Standardized Differences	
Covariate	FB Mean	Pre-Match	Post-Match	Pre-Match	Post-Match
Occupational Prestige of Job Aspired To	581.97	523.52	555.55	0.25	0.11
High School Size	138.08	179.92	146.24	-0.33	-0.06
High School Rank (quantile)	55.81	44.56	51.94	0.43	0.15
Considered outstanding by teacher	13%	9%	12%	0.13	0.04
Parental Income ($100)	73.19	59.63	59.49	0.19	0.19
Participated in band or orchestra	32%	37%	35%	-0.09	-0.05
Participated in speech or debate	32%	22%	28%	0.25	0.10
Participated in school publications	25%	15%	22%	0.26	0.08
Father was a farmer	26%	22%	23%	0.10	0.07
Planned to serve in military	25%	30%	27%	-0.12	-0.06
Attended Catholic high school	4%	8%	5%	-0.19	-0.03
IQ	105.11	100.40	103.03	0.34	0.15
Father's Education (years)	9.73	9.40	9.50	0.10	0.07
Mother's Education (years)	10.80	10.20	10.66	0.22	0.05
Lived with both parents	89%	91%	91%	-0.07	-0.05
Mother Working in 1957	42%	33%	38%	0.19	0.09
Teachers Encouraged College	63%	45%	57%	0.37	0.12
Parents Encouraged College	66%	59%	62%	0.15	0.09
Had Friend Planning on College	39%	34%	35%	0.11	0.08
Never discussed future plans with parents	3%	2%	2%	0.01	0.04
Sometimes discussed future plans with parents	42%	46%	43%	-0.08	-0.03
Often discussed future plans with parents	56%	52%	55%	0.08	0.02
Family wealth considerably below community average	1%	0%	0%	0.16	0.16
Family wealth somewhat below community average	9%	7%	7%	0.08	0.14
Family wealth considerably around community average	66%	73%	75%	-0.16	-0.20
Family wealth somewhat above community average	22%	19%	16%	0.08	0.14
Family wealth considerably above community average	2%	1%	1%	0.07	0.04
Parents cannot financially support college education	30%	31%	29%	-0.03	0.02
Parents can financially support college education with sacrifice	53%	55%	60%	-0.03	-0.13
Parents can easily financially support college education	17%	14%	12%	0.08	0.15

mSSL-DCPE recovered 9 non-zero $\beta_{j,k}$'s and 41 non-zero $\omega_{k,k'}$'s. mSSL-DPE recovered 14 non-zero $\beta_{j,k}$'s, eight of which were identified by mSSL-DCPE. Additionally, mSSL-DPE

identified 37 of the 41 non-zero entries in $\omega_{k,k'}$'s found by mSSL-DCPE along with several more. On closer inspection, we found that mSSL-DPE's estimated mode had a slightly larger log-posterior value than mSSL-DCPE's. In terms of estimating the effect of playing football on these outcomes, our results comport with Deshpande et al. (2017)'s findings from separate univariate analyses: neither mSSL-DPE nor mSSL-DCPE identified a non-zero $\beta_{j,k}$ corresponding to football participation. Much of the signal uncovered by mSSL-DPE is quite intuitive: adolescent IQ was a relevant predictor of scores on the digits ordering task in 2003 and the WAIS similarity task in 1993, 2003, and 2011, anticipated years of post-secondary education was a strong predictor of actual years of education completed by 1974 and the occupational prestige of subjects' job in 1964, and the occupational prestige of the jobs to which subjects aspired in high school was a relevant predictor of the occupational prestige of the jobs they actually held in 1964, 1970, 1974, and 1975. In addition, mMEVS-DPE also selected several of the indicator variables of matched set membership. These corresponded to matched sets containing subjects with similar covariates and propensity scores who had higher than average CES-D scores in 1993 (i.e. they displayed more depressive symptoms), higher than average earnings in 1974, or higher than average scores on the Anger Index in 2004.

Not only does our multivariate approach confirm the main findings of Deshpande et al. (2017)'s univariate analysis, it also provides an estimate of the residual residual Gaussian graphical model G of the 29 outcomes considered, shown in Figure 5. The edges in G encode conditional dependency between the cognitive, psychological/behavioral, and socio-economic outcomes that remain after we adjust for the measured confounders. G exhibits a very strong community structure, with many more edges between outcomes of the same type (colored in red) than of different type (colored in gray). This is rather interesting, in light of the fact that the implicit prior on G, which made each edge equally likely to appear, did not tend to favor any such structure.

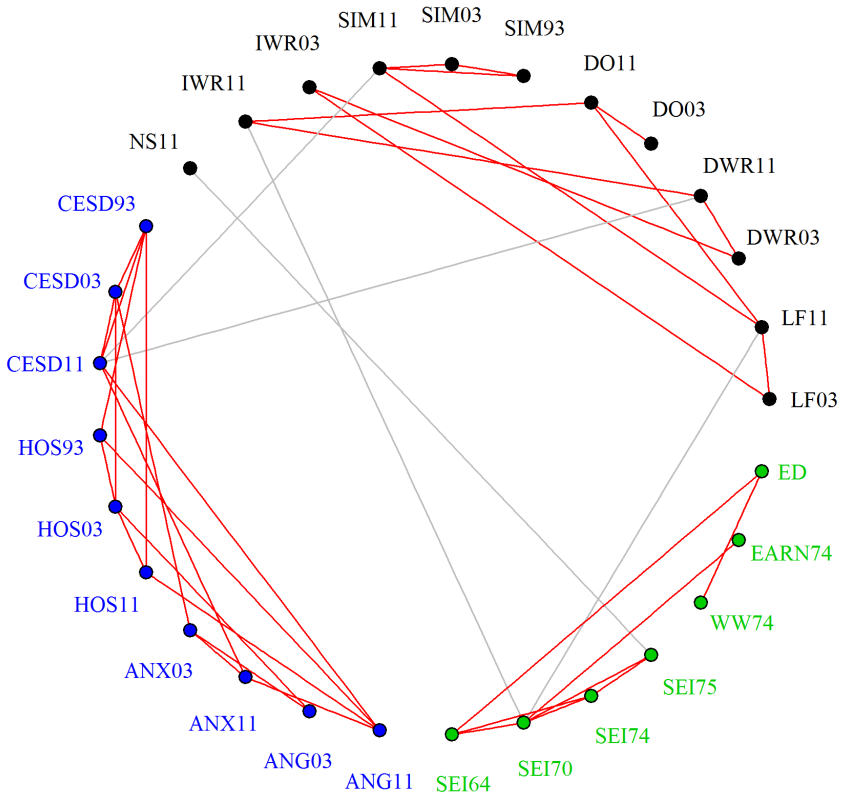


Figure 5: The number following outcome abbreviation indicates the year in which it was measured. Outcomes are colored according to type: cognitive (black), psychological/behavioral (blue), socio-economic / educational (green). Observe that there are many more "within community" edges, colored red, than "between community" edges, colored gray.

Many of the conditional dependence relations represented in G seem intuitive: after adjusting for the covariates listed in Table 5, we see that results from the same cognitive test administered in multiple years tended to be conditionally dependent on each other (see, e.g., the triangle formed by SIM93, SIM03, and SIM11). Additionally, we see that the CES-D scale depression scores and anger, hostility, and anxiety scores from the same year tended to be conditionally dependent as well. Perhaps more interesting are the "between community" links between outcomes of different types, colored in red. After adjusting for covariates, oc-

cupational prestige of the job held in 1975 (SEI75) appears conditionally dependent on the score on the number series task in 2011 (NS11), while the scores on both the CES-D scale and letter fluency test (CESD11 and LF11) are conditionally dependent on the similarity test result in 2011 (SIM11).

2.5. Discussion

In this chapter, we have built on Ročková and George (2014)'s and Ročková and George (2016)'s deterministic spike-and-slab formulation of Bayesian variable selection for univariate linear regression to develop a full joint procedure for simultaneous variable and covariance selection problem in multivariate linear regression models. We proposed and deployed an ECM algorithm within a path-following scheme to identify the modes of several posterior distributions, corresponding to different choices of spike distributions. This dynamic exploration of several posteriors is in marked contrast to MCMC, which attempts to characterize a single posterior. In our simulation experiments and analysis of the football safety data, the modal estimates identified by our dynamic posterior exploration stabilized, allowing us to report a single estimate out of the many we computed without the need for cross-validation. Though there is no general guarantee that these trajectories will stabilize, a figure like Figure 3 provide a useful self-check: if one observes stabilization in the supports of B and Ω and in the log-posterior, one can safely report the final mode identified. On the other hand, if the modal estimates have not stabilized, one can simply add larger values of λ_0 and ξ_0 to the ladders and continue exploring.

To negotiate the dynamically changing multimodal environment, we have focused on modal estimation, at the cost of temporarily sacrificing full uncertainty quantification and posterior inference. Assessing the variability in the estimates of mSSL-DPE remains an important problem. One could run a general MCMC simulation starting from the final mSSL-DPE estimate. Alternatively, the relative speed of our ECM algorithm allows it to be used within Taddy et al. (2016)'s recently proposed bootstrap independent Metropolis-Hasting algorithm.

As anticipated by results in Ročková and George (2014) and Ročková and George (2016), our procedure tends to out-perform procedures that use cross-validation to select regularization penalties. A key driver of the improvement is the hierarchical modeling of the uncertainty of the indicators $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, which allows the penalties $\lambda^{\star}_{j,k}$ and $\xi^{\star}_{k,k'}$ in our ECM algorithm selectively shrink each $\beta_{j,k}$ and $\omega_{k,k'}$. This is in marked contrast to regularization methods that apply the same amount of shrinkage to each $\beta_{j,k}$ and the same amount of shrinkage to each $\omega_{k,k'}$. While we have focused on the simplest setting where the $\boldsymbol{\gamma}$'s and $\boldsymbol{\delta}$'s are treated as exchangeable, it is straightforward to incorporate more thoughtful structured sparsity within our framework. For instance, if the covariates displayed a known grouping structure, we could introduce several θ parameters, one for each group, with little additional computational overhead.

CHAPTER 3 : A Particle Optimization Framework for Posterior Exploration

3.1. Motivation

In Chapter 2, we focused on identifying a single point estimate of (B, Σ) . By focusing solely on identifying the posterior mode, though, we are unable to study the posterior distribution of individual covariates effects $\beta_{j,k}$ or partial covariance $\omega_{k,k'}$. At first glance, this might seem at odds to the Bayesian paradigm, a key strength of which is rigorous quantification of posterior uncertainty. Nevertheless, as Engelhardt and Adams (2014) and Petretto et al. (2010) suggest there is often greater initial interest in identifying promising models and assessing uncertainty about them rather than performing inference on specific parameters within the model. Specifically, in the context of sparse high-dimensional linear regression, Engelhardt and Adams (2014) argue that estimating whether a feature contributes or not is more important than estimating its relative contribution.

Adopting this view, we now seek to quantify the uncertainty about the selected model. A natural first step towards this goal would be to identify several promising models rather than a single mode. To do this in the sparse multivariate regression example, we could simply launch several instances of our ECM algorithm from randomly selected starting points. Unfortunately, there are two immediate limitations of such a strategy. First, our ECM algorithm operates over the continuous space of (B, Ω) rather than in the discrete model space. So while it targets modal values of B and Ω, there is no guarantee that the corresponding supports have the largest marginal posterior probability. Secondly, and perhaps more importantly, if the optimization trajectories are run independently, they run the risk of terminating at the same mode, resulting in potentially massive redundancy.

In the univariate regression setting, Ročková (2017) resolves both of these issues with the Particle EM for variable selection. First, she introduced a "reversed EM", which treated the continuous covariate effects as "missing data," and operated directly in the discrete model space. She then proposed an ensemble optimization framework which amounted to

running L "mutually aware" instantiations of the reversed EM whose trajectories repelled each other if they appeared headed towards the same model. This procedure was able to identify several promising sparse regression models simultaneously and efficiently.

In this chapter, we extend Ročková (2017)'s Particle EM for variable selection to the more general model selection setting and describe a *particle optimization* framework that targets the L models with largest marginal posterior probability. Letting $\boldsymbol{\gamma}$ denote a generic model, we will assume throughout that we are able to compute the marginal likelihood $p(\mathbf{Y}|\boldsymbol{\gamma})$ and the unnormalized prior $\pi(\boldsymbol{\gamma})$. We let $\Gamma_L = \left\{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(L)}\right\}$ be the collection of L models with largest posterior probabilities. For any collection of L not-necessarily distinct models $\Gamma = \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L\}$ and vector $\mathbf{w} = (w_1, \ldots, w_L)$ of non-negative weights summing to one, we will let $q(\cdot|\Gamma, \mathbf{w})$ be the discrete distribution that places probability w_ℓ on the partition $\boldsymbol{\gamma}_\ell$. We will denote the set of all such distributions $q(\cdot|\Gamma, \mathbf{w})$ by $\mathcal{Q}_L$. It should be noted that $\mathcal{Q}_L$ contains all distributions on the model space with at most L atoms, so that $\mathcal{Q}_L \subset \mathcal{Q}_{L+1}$ for all $L \in \mathbb{N}$. Following Ročková (2017), we refer to the $\boldsymbol{\gamma}_\ell$'s as *particles*, Γ as a *particle set*, the w_ℓ's as *importance weights*, and the pair $(\Gamma, \mathbf{w})$ as the *particle system*.

The rest of this chapter is organized as follows. In Section 3.2, we describe a variational approximation of the posterior $\pi(\boldsymbol{\gamma}|\mathbf{Y})$. We then outline a general computational strategy in Section 3.3 and demonstrate our method's utility for Gaussian mixture modeling in Section 3.4.

3.2. A Variational Approximation

The goal of variational inference (see Blei et al., 2017, and references therein) is to approximate an intractable probability density $p(\theta)$ with the density $q^*(\theta)$ within a class $\mathcal{Q}$ of simpler, tractable densities which is closest to $p(\theta)$ in a Kullback-Leibler sense:

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathbb{E}_q \left[\log \frac{p(\theta)}{q(\theta)}\right].$$

Because it relies on optimization rather than sampling, variational inference is widely used as a scalable alternative to MCMC in many Bayesian problems. In our model selection context, the sheer number of models $\boldsymbol{\gamma}$ makes exact evaluation of $\pi(\boldsymbol{\gamma}|\mathbf{Y})$ impossible. Nevertheless, as the following claim demonstrates, we may still identify Γ_L without an exhaustive search of the model space by considering the variational approximation within the class $\mathcal{Q}_L$.

Claim 3. *Suppose* Γ^* *and* $\mathbf{w}^*$ *are the particle set and importance weights corresponding to the variational approximation of* $\pi(\boldsymbol{\gamma}|\mathbf{Y})$ *within the class* $\mathcal{Q}_L$. *Then* $\Gamma^* = \Gamma_L$ *and* $w_\ell^* \propto \pi(\boldsymbol{\gamma}^{(\ell)}|\mathbf{Y})$ *for each* $\ell = 1, \ldots, L$.

Proof. By definition, we know

$$q^* = \arg\min_{q \in \mathcal{Q}} \left\{ \sum_{\boldsymbol{\gamma}} q(\boldsymbol{\gamma}) \log \frac{q(\boldsymbol{\gamma})}{\pi(\boldsymbol{\gamma}|\mathbf{Y})} \right\}, \tag{3.1}$$

where by convention we take $0 \log 0 = 0$. Let q^* denote the optimal approximation $q(\cdot|\Gamma^*, \mathbf{w}^*)$ and suppose, for the sake of contradiction, that the optimal particle set $\Gamma^* \neq \Gamma_L$. Then there is some partition $\boldsymbol{\gamma}^{(\ell)}$ in Γ_L that is not itself contained in Γ^* and some other partition $\boldsymbol{\gamma}_i \in \Gamma^*$ but not in Γ_L with $\pi(\boldsymbol{\gamma}^{(\ell)}|\mathbf{Y}) > \pi(\boldsymbol{\gamma}_i|\mathbf{Y})$. Let $\tilde{\Gamma}$ be the particle set formed from Γ^* by replacing $\boldsymbol{\gamma}_i$ with $\boldsymbol{\gamma}^{(\ell)}$ and let $\tilde{q}$ be the distribution in $\mathcal{Q}_L$ indexed by the particle set $\tilde{\Gamma}$ and importance weights $\mathbf{w}^*$. Then

$$KL(q^*|\pi(\boldsymbol{\gamma}|\mathbf{Y})) - KL(\tilde{q}|\pi(\boldsymbol{\gamma}|\mathbf{Y})) = w_i \log \frac{\pi(\boldsymbol{\gamma}^{\ell}|\mathbf{Y})}{\pi(\boldsymbol{\gamma}_i|\mathbf{Y})} \geq 0,$$

contradicting the optimality of q^*. Having established that $\Gamma^* = \Gamma_L$, simple calculus verifies the claim $w_\ell^* \propto \pi(\boldsymbol{\gamma}^{(\ell)}|\mathbf{Y})$. □

It is not difficult to verify that the problem in Equation (3.1) is equivalent to following penalized optimization problem

$$(\Gamma^*, \mathbf{w}^*) = \arg\max_{(\Gamma, \mathbf{w})} \left\{ \sum_{\boldsymbol{\gamma}_\ell} w_\ell \log \pi(\boldsymbol{\gamma}_\ell, \mathbf{Y}) + H(\Gamma, \mathbf{w}) \right\} \tag{3.2}$$

where $H(\Gamma, \mathbf{w}) = -\mathbb{E}_{q(\cdot|\Gamma,\mathbf{w})}[\log q(\cdot|\Gamma, \mathbf{w})]$ is the entropy of the distribution $q(\cdot|\Gamma, \mathbf{w})$. We pause briefly to reflect on the two terms in Equation (3.2). The first term is, up to an additive constant depending only on $\mathbf{Y}$, an importance-weighted average of the height of the log-posterior at each particle. This term is clearly maximized by setting all particles equal to the MAP model, $\boldsymbol{\gamma}^{(1)}$. The predilection of the first term towards collapsing the particle set to the MAP is tempered by the entropy of the particle system, $H(\Gamma, \mathbf{w})$, which is maximized when all of the particles are distinct and the importance weights are all equal to $\frac{1}{L}$. Returning to the analogy of running several instantiations of the same mode hunting algorithm, $H(\Gamma, \mathbf{w})$ penalizes trajectories from visiting the same model at the same time. It is important to stress at this point that the entropy term H induces only a weak form of repulsion between the trajectories in that it only discourages trajectories from visiting the same model as opposed to models that are close according to some metric. We will return to this point in a bit more detail in Chapter 6.

3.3. Implementation

To solve the problem in Equation (3.2), we proceed in an coordinate-wise fashion, iteratively updating one of the particle set Γ or importance weights $\mathbf{w}$ while holding the other constant until we reach a stationary point. Before proceeding we require additional notation. Given a collection of L particles, $\Gamma = \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L\}$, we let $\Gamma^\star = \{\boldsymbol{\gamma}_1^\star, \ldots, \boldsymbol{\gamma}_{L^\star}\}$ be the collection of the unique $L^\star$ particles contained in the particle set Γ. Moreover, define $\mathbf{p}^\star = (p_1^\star, \ldots, p_{L^\star}^\star)$ to be the cumulative importance weights associated with each of the unique particles with $p_{\ell^\star}^\star = \sum_{\ell=1}^{L} w_\ell \mathbb{I}(\boldsymbol{\gamma}_\ell = \boldsymbol{\gamma}_{\ell^\star}^\star)$. From here, it is not difficult to see that

$$H(\Gamma, \mathbf{w}) = -\sum_{\ell^\star=1}^{L^\star} p_{\ell^\star}^\star \log p_{\ell^\star}^\star$$

Abusing our notation slightly, we will write $H(\boldsymbol{\gamma}, \Gamma_{-\ell}, \mathbf{w})$ be the entropy of the particle system with the particle $\boldsymbol{\gamma}$ replacing $\boldsymbol{\gamma}_\ell$ and keeping the importance weights fixed at $\mathbf{w}$.

We are now ready to describe the iterative updates of Γ and $\mathbf{w}$. Suppose after t iterations,

our current estimates are $\Gamma^{(t)}$ and $\mathbf{w}^{(t)}$. We initialize $\Gamma^{(t+1)} = \Gamma^{(t)}$ and update each particle sequentially, holding the remaining particles fixed at their present values and holding the importance weights fixed at $\mathbf{w}^{(t)}$. For the ℓ^{th} particle, we aim to solve

$$\boldsymbol{\gamma}_\ell^{(t+1)} = \arg\max_{\boldsymbol{\gamma}} \left\{ w_\ell^{(t)} \log \pi(\boldsymbol{\gamma}, \mathbf{Y}) + H\left(\boldsymbol{\gamma}, \Gamma_{-\ell}^{(t)}, \mathbf{w}^{(t)}\right) \right\}$$

where the maximum is taken over all possible models $\boldsymbol{\gamma}$. An exhaustive search over the entire model space would allow us to identify $\boldsymbol{\gamma}_\ell^{(t+1)}$ exactly but such a strategy is obviously infeasible. Instead, we take a simple, local, greedy approach and restrict our search space to a set of candidate models that are close, in some sense, to $\boldsymbol{\gamma}_\ell^{(t)}$. Denoting this set of candidates $\text{Cand}\left(\boldsymbol{\gamma}_\ell^{(t)}\right)$, we set

$$\boldsymbol{\gamma}_\ell^{(t+1)} = \arg\max_{\boldsymbol{\gamma} \in \text{Cand}(\boldsymbol{\gamma}_\ell^{(t)})} \left\{ w_\ell^{(t)} \log \pi(\boldsymbol{\gamma}, \mathbf{Y}) + H(\boldsymbol{\gamma}, \Gamma_{-\ell}^{(t)}, \mathbf{w}) \right\} \tag{3.3}$$

Having swept over the particle set once, we can turn our attention to updating the importance weights. To this end, let $\Gamma^{\star(t+1)}$ be the collection of $L^{\star(t+1)}$ unique particles contained in $\Gamma^{(t+1)}$. Straightforward calculations show that now $p_\ell^{\star(t+1)} \propto \pi(\boldsymbol{\gamma}_{\ell\star}^{(t+1)\star} | \mathbf{Y})$. Finally, we can divide the updated cumulative importance weight $p_{\ell\star}^{\star(t+1)}$ equally between the $w_\ell^{(t+1)}$'s corresponding to the particles $\boldsymbol{\gamma}_\ell^{(t+1)} = \boldsymbol{\gamma}_{\ell\star}^{\star(t+1)}$. In this way, the importance weight w_ℓ reflects the relative importance of the particle $\boldsymbol{\gamma}_\ell$: the larger it is, the more posterior mass there is at $\boldsymbol{\gamma}_\ell$. We continue these iterative updates until we reach a stationary point.

3.4. Mixture Modeling with an Unknown Number of Mixture Components

In Section 2.4, we attempted to estimate the effect of playing high school football on several later-life cognitive, psychological, and socio-economic outcomes and to estimate the residual Gaussian graphical model between these outcomes simultaneously. In doing so, we relied on a single parameter to capture the effect of playing football on any given outcome. This may be insufficient, as certain groups of players may be exposed to higher levels of head trauma and may be more likely to experience later-life dysfunction (Broglio et al., 2011).

A potentially much more reasonable modeling approach would be to introduce separate parameters for each group (i.e. a separate B and Ω for each group of subjects). Unfortunately, our dataset did not contain positional information and we were unable to pre-specify specific groups. Lacking this information, we must rely on the data to not only estimate subgroup-specific parameters but also to identify specific subgroups.

The main thrust of cluster analysis is to divide the sample into smaller subgroups which may plausibly represent homogeneous sub-populations of interest. Clustering often proceeds with a hierarchical mixture model in which we assume that the data $Y_1, \dots Y_n \in \mathbb{R}^q$ are generated according to the following process. First, we draw a partition $\boldsymbol{\gamma} = \{C_1, \cdots C_K\}$ of the integers $[n] = \{1, 2, \dots, n\}$ according to some probability law $\mathcal{P}_{\boldsymbol{\gamma}}$ over the space of all partitions of $[n]$. Then, to each partition element C_k (hereafter referred to as a "cluster"), we associate a parameter θ_k which is itself drawn from some law $\mathcal{P}_\theta(\theta|\boldsymbol{\gamma})$ that depends on $\boldsymbol{\gamma}$. Finally, for each k, the data $\mathbf{Y}_k = \{Y_i : i \in C_k\}$ arise independently from some distribution $\mathcal{P}_y(y|\theta_k, \boldsymbol{\gamma})$ that depends on the unknown cluster-specific parameter θ_k and the partition $\boldsymbol{\gamma}$. In other words, we assume that each datum Y_i has been drawn independently from one of K distinct distributions, each of which represents a homogeneous sub-population of interest. The main goal of clustering is to recover the partition $\boldsymbol{\gamma}$ encoding the original cluster allocations.

If the number of clusters K is known *a priori*, perhaps the most popular approach to clustering is to use the k-means algorithm to identify the partition $\boldsymbol{\gamma} = \{C_1, \dots, C_K\}$ of the integers $[n] = \{1, 2, \dots, n\}$ minimizing the objective

$$\sum_{k=1}^{K} \sum_{i \in C_k} \left\| Y_i - \overline{Y}_k \right\|^2$$

where $\overline{Y}_k = |C_k|^{-1} \sum_{i \in C_k} Y_i$ are the cluster means. The algorithm begins with an initial allocation of the data into K clusters and then alternates between reassigning individual data points to the cluster with closest mean and recomputing the cluster means. The

straightforward implementation, speed, and scalability of the k-means algorithm has made it especially popular in statistics and machine learning. This is in spite of the fact that the algorithm relies on an *a priori* knowledge of the number of clusters. Since one generally does not know the number of sub-populations K *a priori*, it is common to run the k-means algorithm over a range of different K values and to select the one that best fits the data according to some heuristic.

Absent real prior information about the underlying clustering structure, a fully Bayesian approach to clustering begins by assigning positive prior probability to *every* possible partition $\boldsymbol{\gamma}$ and then updating this prior with the data to derive the posterior $\pi(\boldsymbol{\gamma}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})$. Perhaps the most popular choice of prior $\pi(\boldsymbol{\gamma})$ is based on the Dirichlet process of Ferguson (1973), which we review briefly in Section 3.4.2. While conceptually straightforward, the incredible number of partitions renders the posterior analytically intractable and one typically relies on MCMC to approximate the distribution. Summarizing the output of the MCMC for clustering has long been recognized as a challenging problem (see, e.g. Medvedovic and Sivaganesan, 2002; Dahl, 2006; Wade and Ghahramani, 2017). This partly due the fact that set of visited partitions is quite irregular (Lau and Green, 2007) and many authors have instead taken decision-theoretic approaches to summarize the posterior $\pi(\boldsymbol{\gamma}|\mathbf{Y})$. The most intuitive way to identify the MAP partition is by simply computing the posterior probability (up to a normalization constant) of each visited partition and reporting the one with the largest unnormalized probability. This approach is not particularly viable as one cannot hope to explore even a small fraction of the total number of partitions in a practical number of MCMC iterations. Somewhat more recently, several authors have by-passed stochastic search methods entirely, developing optimization algorithms for targeting the MAP partition which are considerably faster than MCMC (see, e.g. Heller and Ghahramani, 2005; Heard et al., 2006; Dahl, 2009). As Lau and Green (2007) argue however, as the dimension of the parameter space increases, posterior modes are increasingly unrepresentative characterizations of the "center" of the posterior distribution. Worse still, according to Dahl (2006), the MAP partition may only be slightly more probable *a posteriori* than the

next best alternative. Though we may regard the MAP as the optimal point estimate with respect to a 0–1 loss, this loss function is far from appealing (Lau and Green, 2007) as it ascribes the same loss to partitions that differ in the allocation of a single index as it does to partitions that differ in the allocation of several indices. Lau and Green (2007) propose partition estimates that minimize a particular loss function due to Binder (1978), which measures the disagreement in cluster assignment between all pairs of indices between two partitions. These approaches, while certainly more principled than MAP estimation, make no attempt to quantify the posterior uncertainty about $\boldsymbol{\gamma}$. A notable exception is the recent paper by Wade and Ghahramani (2017), which constructs point estimates and posterior credible balls of $\boldsymbol{\gamma}$ without resorting to MCMC.

Rather than focusing on a single point estimate to summarize the posterior distribution $\pi(\boldsymbol{\gamma}|\mathbf{Y})$, we may use our particle optimization framework to identify the top partitions *a posteriori* simultaneously. In this setting, we must navigate the space of all partitions of $[n]$, whose dimension scales exponentially in the number of observations n. To demonstrate, we focus on the canonical Gaussian mixture model, which we introduce in the next subsection. The remainder of this chapter is organized as follows. We begin with a brief (and non-exhaustive) review of the celebrated Dirichlet Process and its application to clustering in Section 3.4.2 and describe our search strategy in Section 3.4.3. We conclude with a demonstration of our optimization scheme on simulated data.

3.4.1. The Gaussian Mixture Model

Given a partition $\boldsymbol{\gamma} = \{C_1, \ldots, C_K\}$, we introduce parameters $\theta_k = (\mu_k, \Sigma_k)$ and for each $i \in C_k$, we model $Y_i|\theta_k \sim \mathrm{N}_q(\mu_k, \Sigma_k)$, independently. For each k, let $\mathbf{Y}_{C_k} = \{Y_i : i \in C_k\}$. Conditional on $\boldsymbol{\gamma}$, our likelihood factorizes over the clusters:

$$p(\mathbf{Y}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{k=1}^{K} \prod_{i \in C_k} p(\mathbf{Y}_i|\mu_k, \Sigma_k),$$

where

$$p(Y_i|\mu_k, \Sigma_k) = (2\pi)^{-\frac{q}{2}} |\Sigma_k|^{-1} \exp\left\{-\frac{1}{2}(Y_i - \mu_k)^\top \Sigma_k^{-1}(Y_i - \mu_k)\right\}$$

Models like this, in which observations in different clusters are independent, are known as product partition models and have been explored quite extensively in the literature (see, e.g., Hartigan, 1990; Crowley, 1997).

We now specify a prior on the cluster-specific parameters $\theta_1, \ldots, \theta_k$, so that, given the partition $\boldsymbol{\gamma}$, they are independent. If cluster C_k contains $n_k = |C_k| > 1$ elements, we place an improper prior distribution of the parameter (μ_k, Σ_k) with density

$$\pi(\mu_k, \Sigma_k) = 2^{-\frac{q^2}{2}} \Gamma_q\left(\frac{q}{2}\right)^{-1} |\Sigma|^{-\frac{q+q+1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}\left(\Sigma_k^{-1}\right)}.$$

Note that this corresponds to placing an improper, flat prior on μ_k and an Inverse-Wishart $\mathrm{IW}(I_q, q)$ prior on Σ_k independently. If, on the other hand, cluster C_k contains a single element, we restrict $\mu_k = \mathbf{0}_q$, $\Sigma_k = \tau_k^2 I_q$ and place an improper prior on τ_k^2 with density equal to τ_k^{-2}. Under this model and prior specification, the marginal density of $\mathbf{Y}|\boldsymbol{\gamma}$ will factorize over the clusters as well and so to derive a closed-form expression for $p(\mathbf{Y}|\boldsymbol{\gamma})$, it is enough to do so over each cluster. Straightforward calculations yields

$$p(\mathbf{Y}_{C_k}|\boldsymbol{\gamma}) = \begin{cases} \pi^{-\frac{q(n_k-1)}{2}} \times \frac{\Gamma_q\left(\frac{q+n_k-1}{2}\right)}{\Gamma_q\left(\frac{q}{2}\right)} \times |I_q + S_k|^{-\frac{q+n_k-1}{2}} & \text{if } n_k \geq 2 \\ (2\pi)^{-\frac{q}{2}} \times \left(\frac{S_k}{2}\right)^{-\frac{q}{2}} \times \Gamma\left(\frac{q}{2}\right) & \text{if } n_k = 1 \end{cases}$$

3.4.2. The Dirichlet Process and Bayesian Clustering

Having specified the likelihood $p(\mathbf{Y}|\boldsymbol{\gamma})$ we are ready to turn our attention to the prior $\pi(\boldsymbol{\gamma})$ on the space of all partitions and describe a search strategy. Perhaps the most common prior can be derived from the Dirichlet process of Ferguson (1973), which we now review briefly. The following is based primarily on Chapter 22 of Gelman et al. (2008).

Formally, given some parameter space Θ, base probability measure $\mathcal{P}_0$, and scalar $\eta > 0$, a realization $\mathcal{P}$ of the Dirichlet process $DP(\eta, \mathcal{P}_0)$ satisfies the following property: for any finite partition of $\Theta = A_1 \cup \cdots \cup A_M$,

$$(\mathcal{P}(A_1), \ldots, \mathcal{P}(A_M)) \sim \text{Dirichlet}(\eta\mathcal{P}_0(A_1), \ldots, \eta\mathcal{P}_0(A_M)).$$

One can show that the distribution $\mathcal{P}$ is an almost surely discrete distribution

$$\sum_{h=1}^{\infty} w_h \delta_{\vartheta_h^*}$$

whose atoms ϑ_j^* are drawn independently from the base measure $\mathcal{P}_0$ and whose weights are generated according to the following *stick-breaking* construction:

$$w_h = V_h \prod_{h'<h} (1 - V_{h'}),$$

where the V_h's are i.i.d. Beta$(1, \eta)$ random variables. To use the Dirichlet process in clustering, one begins by associating each datum Y_i with its own parameter ϑ_i and models $\vartheta_i \sim \mathcal{P}$ and $\mathcal{P} \sim \text{DP}(\alpha, \mathcal{P}_0)$. The fact that $\mathcal{P}$ is indexed by a countably infinite number of parameters appears, at least at first, quite problematic from the standpoint of performing posterior calculations. However, we may marginalize out $\mathcal{P}$ to obtain the induced prior on the parameters $(\vartheta_1, \ldots, \vartheta_n)$. We can describe this distribution through a sequence of conditional distributions:

$$\vartheta_i | \vartheta_1, \ldots, \vartheta_{i-1} \sim \left(\frac{\eta}{\eta + i - 1}\right) \mathcal{P}_0(\vartheta_i) + \sum_{j=1}^{i-1} \left(\frac{1}{\eta + i - 1}\right) \delta_{\vartheta_j}. \tag{3.4}$$

In words, ϑ_i is either drawn afresh from the base measure $\mathcal{P}_0$ with probability $\frac{\eta}{\eta+i-1}$ or it is drawn uniformly from the collection $\{\vartheta_1, \ldots, \vartheta_{i-1}\}$ with probability $\frac{i-1}{\eta+i-1}$. The fact that some of the ϑ_i's will coincide with positive prior probability induces a partition of the observations in which Y_i and Y_j belong to the same cluster if and only if $\vartheta_i = \vartheta_j$.

A very useful metaphor for understanding Equation (3.4) is the Chinese Restaurant process, which imagines n customers arriving at a restaurant with infinitely many tables and deciding where to sit. In this metaphor, the first custom sits at the first table, which has dish ϑ_1^*. The next customer then sits at this first table with probability $\frac{\eta}{1+\eta}$ or at a new table with residual probability. In the latter case, this new table is assigned dish ϑ_2^*. The process continues, with the i^{th} customer sitting at the table with dish ϑ_k^* with probability proportional to the number of customers already seated there or at a new table with probability proportional to η. In this metaphor, we may view each customer as observations, tables as clusters, and dishes as cluster-specified parameters.

Using Equation (3.4) and the Chinese Restaurant process metaphor, it is possible to derive the implied prior on the partition $\boldsymbol{\gamma} = \{C_1, \ldots, C_K\}$:

$$\pi(\boldsymbol{\gamma}|\eta) = \frac{\Gamma(\eta)}{\Gamma(\eta+n)}\eta^K \prod_{j=1}^{K} \Gamma(n_k) \tag{3.5}$$

where $n_k = |C_k|$ is the number of indices in cluster k. We follow Casella et al. (2014) and refer to this as the Ewens-Pitman(η) prior. For the remainder of this section and throughout Chaptes 4 and 5, we will use this prior with $\eta = 1$ in our particle optimization method. Before closing this digression, we note in passing that there are many other choices of prior distributions on $\boldsymbol{\gamma}$; see, e.g., Casella et al. (2014) for a more in-depth discussion and comparison of these alternative priors.

3.4.3. Exploring the Space of Partitions

The *transfer distance* between two partitions is defined as the smallest number of transfer of elements from their own cluster to another, possibly empty, cluster to turn one partition into another. For example, the transfer distance between $\boldsymbol{\gamma} = \{\{1,2\},\{3,4\},\{5,6\}\}$ and $\boldsymbol{\gamma}' = \{\{1,6\},\{2,3\},\{4,5\}\}$ is 3. The Chinese Restaurant metaphor of the Dirichlet Process naturally inspires a Gibbs sampler, in which observations are sequentially re-allocated to a new or existing cluster at random. In essence, the Gibbs sampler for Dirichlet pro-

cess mixtures traverses this space by taking random steps of transfer distance one. It is straightforward to adapt this stochastic search strategy to our optimization setting by cycling over the observations and re-allocating each to the new or existing cluster which yields the largest increase in the objective. While simple in principle, such a strategy precludes moving groups of observations simultaneously between clusters. That is, in order to pass between two partitions, say $\{\{1,2,3\},\{4,5,6\}\}$ and $\{\{1,6\},\{2,3\},\{4,5\}\}$, we must pass through an intermediate state like $\{\{1\},\{2,3\},\{4,5,6\}\}$. Unfortunately, such intermediate states typically have low posterior probability and because of its the incremental nature, the Gibbs sampler will generally mix very slowly (Jain and Neal, 2004; Celeux et al., 2000). In our optimization context, similarly restricting our proposed moves to the set of partitions at transfer distance one could lead to entrapment at sub-optimal local modes of the posterior.

An obvious workaround is to simply set $\text{Cand}\left(\boldsymbol{\gamma}^{(t)}\right)$ to be the transfer distance ball around $\boldsymbol{\gamma}^{(t)}$ of radius r. This becomes quickly unwieldy as the number of partitions at a fixed transfer distance can be quite massive (possibly of order $O(n^4)$ for $r = 2$). Instead, we take inspiration from the Split-Merge MCMC of Jain and Neal (2004), which proposed re-allocating multiple observations simultaneously as follows:

1. Randomly select two indices $i, j \in [n]$.
2. If i and j belong to different clusters, propose merging these two clusters.
3. If i and j belong to the same cluster, propose splitting this cluster into two parts, one containing i and the other containing j. A Gibbs sampler restricted to this cluster is used to determine how remaining indices are distributed to the two new clusters

We propose to construct our candidate set as follows. For each cluster C_k, let $i_k^* = \arg\max_{i \in C_k} \left\{\left\|Y_i - \overline{Y}_K\right\|_2\right\}$ be the index in C_k corresponding to the observation Y_i which is furthest from the cluster mean. Additionally, for a partition $\boldsymbol{\gamma}$, for each $i \in [n]$, let $k(i) = \arg\min_{k: i \notin C_k} \left\{\left|Y_i - \overline{Y}_k\right|\right\}$ be the index of the cluster whose overall mean $\overline{Y}_k$ is closest to Y_i among those that do not contain i. We consider the following proposals:

- For each cluster C_k, propose re-allocating i_k^* to an entirely new, singleton cluster
- For each cluster C_k, propose re-allocating i_k^* to the existing cluster $C_{k(i_k^*)}$
- Split C_k into two pieces anchored by two randomly selected indices $i, j \in C_k$. Then propose leaving the two newly created clusters as is and also propose merging each to another existing cluster with closest mean.
- Propose merging each cluster C_k to the existing cluster whose mean is closest.

The first two types of candidates are useful for locally refining our particle system while the latter two types of candidates attempt to take larger jumps across the space of all partitions.

To illustrate the proposed methodology, we generated a dataset consisting of $n = 400$ points in $\mathbb{R}^2$ divided into 5 clusters. Figure 6 shows the data, colored according to the true clustering. Also shown are the mean and covariance matrices of each mixture component as well as 95% ellipses for the each of mixture component. It should be noted from the outset that recovering the true partition structure from this realization of the data is, in general, quite difficult in light of the considerable overlap between the mixture distributions. For instance, we cannot reasonably expect any clustering algorithm to correctly allocate the green point located near (0,-2) in the bulk of the black point cloud or the two red points in the bulk of the blue point cloud.

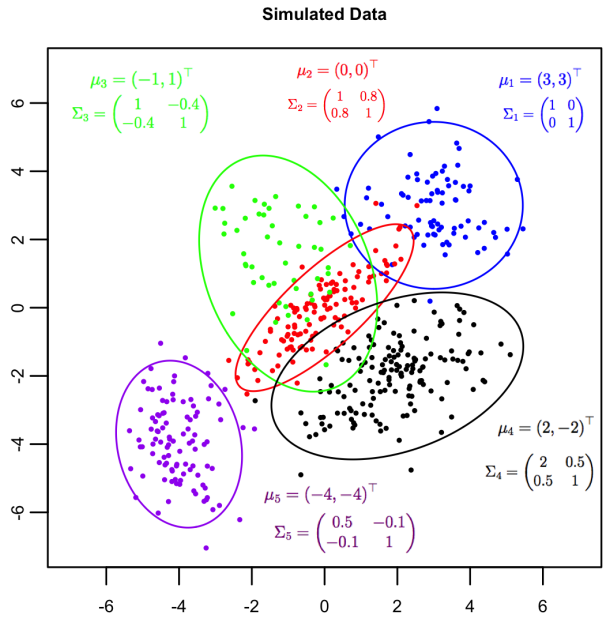


Figure 6: Single realization from a Gaussian mixture model with $K = 5$. Also displayed are the 95% probability ellipses for each mixture component.

We ran our particle optimization procedure with $L = 10$ particles with a Ewens-Pitman(1) prior on $\boldsymbol{\gamma}$. At convergence, we found that only two of partitions identified had non-negligible importance weights. Figure 7 shows these top two partitions, denoted $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, along with the true partition and the one recovered by running k-means with $K = 5$ known.

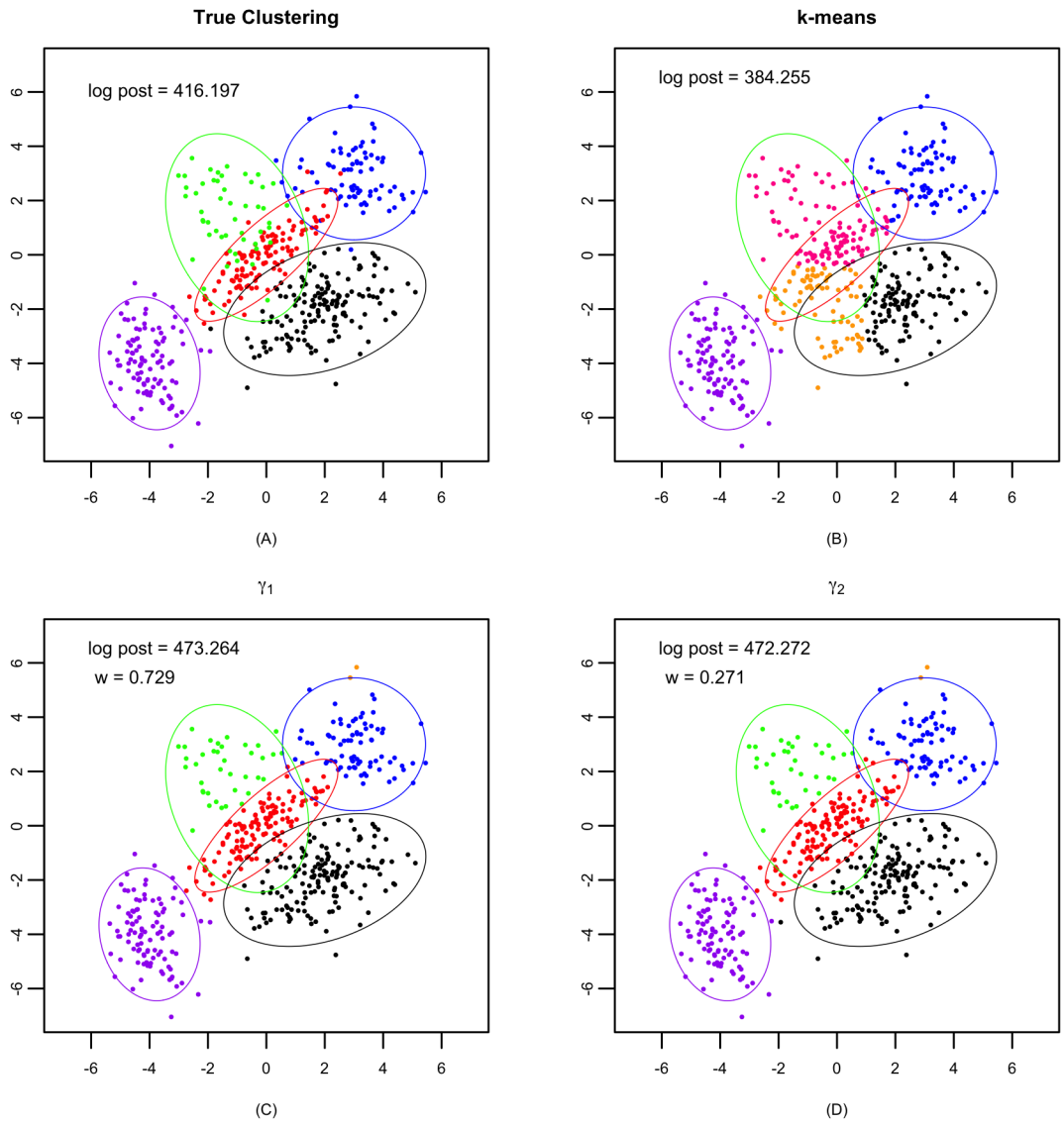


Figure 7: Original partition, k-means estimate, and top two partitions returned by our method.

The original purple, blue, and black clusters were rather well-separated from the other clusters and it is not therefore not surprising that both k-means and our procedure were able to recover them relatively well. Kulis and Jordan (2012) show how the k-means objective can be derived by considering a finite Gaussian mixture model in which every mixture component has covariance matrix $\sigma^2 I$ and letting $\sigma^2 \to 0$. In light of this, it is perhaps

unsurprising that k-means is unable to distinguish differences in the scale parameters of the underlying mixture components. This is most pronounced with the original green and red clusters, which k-means was unable to recover. k-means split the original red cluster into two parts, with one part clustered with most of the data from the original green cluster into the pink cluster of Figure 7 and the other grouped together with some data from the original black cluster to form the orange cluster. It is interesting to note that the partition recovered by k-means has substantially less posterior mass than the true clustering, as evidenced by the difference of 32 in log-posterior mass.

Reassuringly, the top partitions identified by our method are much closer to the original clustering. In light of the overlap between the original clusters, it is not surprising to see that both $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ have higher log-posterior values than the true partition. Interestingly, both of these separate two points originally near the boundary of the blue point cloud into a new, sixth cluster. The primary differences between $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ can be seen at the interface between the original red and black clusters and the interface between the original red and green clusters. The remaining 8 particles all had comparatively negligible importance weights, ranging from 5.072×10^{-6} to 5.842×10^{-17}.

At first glance, these results suggest that the posterior concentrates at $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, with comparatively negligible probability given to the remaining partitions. It could very well be the case, however, that this is an artifact of our particle system's inability to fully explore the space of all partitions; after all, we only consider a small number of proposals at each iteration. In this example, we initialized the particle system with all particles equal to the partition $\{1, 2, \ldots, n\}$ and all importance weights equal to 0.1 Early on, most of the particle updates involved splitting a large cluster into two smaller parts, producing a clustering that was very close to $\boldsymbol{\gamma}_1$. Towards the end, however, all of the updates were local and involved re-allocating individual observations between the existing clusters. In this way, our particle system traversed the space of all partitions by first taking rather large, coarse steps away from $\{1, 2, \ldots, n\}$ and then taking very small, local refining steps. It is not immediately

clear, whether there are other regions of the space of partitions with intermedia posterior probability to which our particle system was simply unable to navigate. Recall that the term $H(\Gamma, \mathbf{w})$ in Equation (3.2) measures the entropy of the particle system and penalizes redundancy among particles. As mentioned earlier, this induces a very weak repulsion between the particles that may be insufficient to push a particle away from a dominant mode and into other interesting parts of the space. In particular, we would like to encourage our particle system to consider moves that trade-off maximizing the posterior and maximizing the distance between particles. We discuss strategies for doing so in Chapter 6.

CHAPTER 4 : Identifying Spatial Clusters

In the previous chapter, we extended Ročková (2017)'s Particle EM for variable selection to a more generic model selection setting, focusing primarily on mixture modeling. The main focus there was to identify several promising partitions $\boldsymbol{\gamma}$ of the data into clusters and to begin to quantify the posterior uncertainty about the true partition. In this chapter and the next, our main interest will be on parameter estimation in the presence of uncertainty about the underlying cluster structure of the data. The work described in this chapter is joint with Cecilia Balocchi, Shane Jensen, and Ed George.

A central organizing principle in spatial data analysis is Waldo Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Bayesian hierarchical modeling provides a powerful and coherent way to actualize this law by facilitating the principled sharing of information between neighboring and nearby spatial units. To illustrate, suppose we observe pairs of data $(\mathbf{y}_1, \mathbf{x}_1), \ldots (\mathbf{y}_n, \mathbf{x}_n)$ from the n spatial regions and consider a simple univariate linear regression model in each unit $\mathbf{y}_i = \beta_i \mathbf{x}_i + \boldsymbol{\varepsilon}_i$ with Gaussian error $\boldsymbol{\varepsilon}_i$. A particularly popular way to induce spatial smoothness among the β_i's is the conditionally auto-regressive (CAR) framework of Besag (1974). This involves placing a multivariate normal distribution on the collection $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ in such a way that β_i is conditionally independent of β_j if and only if regions i and j are not adjacent to one another. There are many variations of Besag (1974)'s CAR model but the following one proposed by Leroux et al. (2000) has proven best in practice (Lee and Mitchell, 2012). Given a constant $\rho \in (0, 1)$ we introduce hyper-parameters μ and σ^2 and model

$$\beta_i | \boldsymbol{\beta}_{-i}, \mu, \sigma^2 \sim \mathrm{N}\left(\frac{(1-\rho)\mu + \rho \sum_{j=1}^n a_{i,j}\beta_j}{1 - \rho + \rho \sum_{j=1}^n a_{i,j}}, \frac{\sigma^2}{1 - \rho + \rho \sum_{j=1}^n a_{i,j}} \right)$$

where $A = (a_{i,j})$ is the adjacency matrix of the spatial units. In this way, the conditional expectation of β_i is rather conveniently expressed as convex combination of the average

value of the neighboring β_j's and the global mean μ. The parameter ρ controls the degree of spatial autocorrelation and is typically fixed (see, e.g., Lee and Mitchell, 2012). From the conditional specification, we can read off the joint distribution of $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|\mu, \sigma^2 \sim \mathrm{N}_n \left(\mu \mathbf{1}_n, \sigma^2 \left[\rho A^* + (1-\rho) I_n \right]^{-1} \right),$$

where $A^* = A - D$ is the unnormalized Laplacian of the adjacency matrix A. While incorporating an additional prior on the global mean μ introduces marginal dependence between all of the β_i's, the covariance structure above still ensures that there is somewhat greater dependence between parameters corresponding to neighboring spatial regions. In this way, this prior introduces a certain global smoothness among the β_i's.

In complex urban settings, however, there are natural or human barriers that can manifest as sharp spatial discontinuities in the data. In such cases, using a single global smoother can be decidedly unappealing. As an example, Figure 8 shows the logarithm of the count of violent crimes per census block group in the city of Philadelphia averaged between 2006 and 2015.

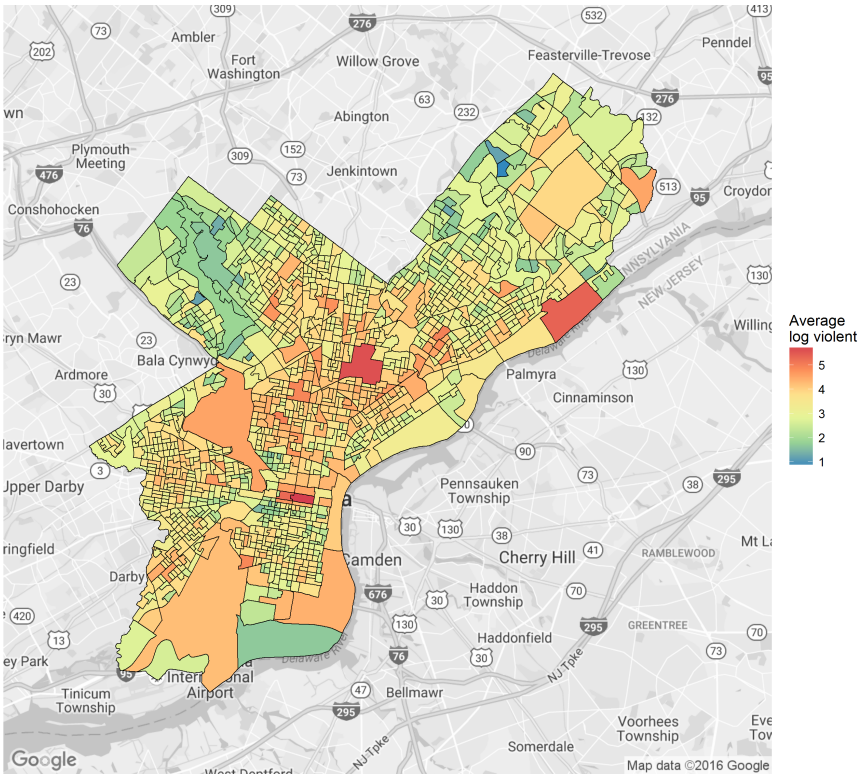


Figure 8: Logarithm of the average number of violent crimes over the ten year period 2006 – 2015 in each census block group in the city of Philadelphia. There are clear spatial discontinuities.

We see that there are some regions with high crime immediately adjacent to many units with much lower crime. Intuitively, over-smoothing these discontinuities with a CAR model could lead to underestimation of the number of crimes in some regions and overestimation in others. A much more sensible approach would be to first partition the regions into several clusters with similar trends and deploy a CAR model within each cluster independently. The following example illustrates how sensibly partitioning the spatial units can substantially improve our estimation of $\boldsymbol{\beta}$ and how inadvertently smoothing over sharp boundaries can substantially bias our estimates.

Consider three spatial partitions $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ of the 10×10 grid in Figure 9, which separate the spatial units into one, two, and three clusters, respectively.

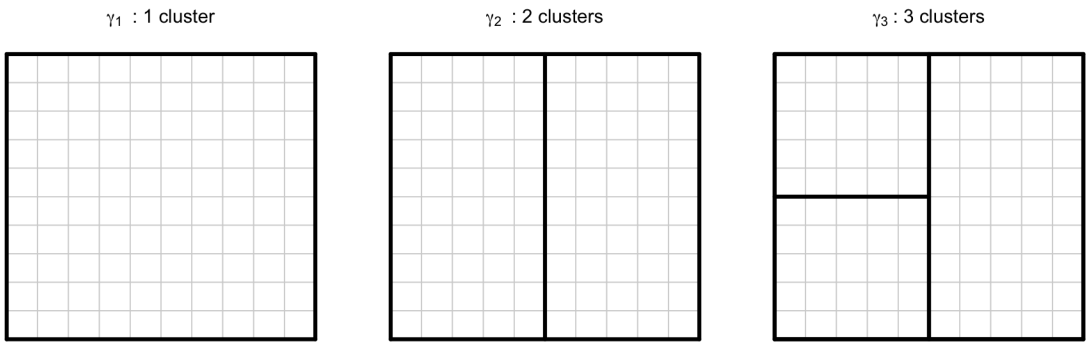


Figure 9: Three spatial partitions of the 10×10 grid with one, two, and three clusters.

We treat $\boldsymbol{\gamma}_3$ as the true spatial partition and generate the β_i's from cluster k according to a CAR model centered at μ_k for $k = 1, 2$, and 3. We consider three different specifications of the cluster means (μ_1, μ_2, μ_3): $(5, 0, -5), (2, 0, -2)$ and $(2, 0, -1)$. Figure 10 shows the values of the β_i's from each setting. For each of these specifications of $\boldsymbol{\beta}$, we generated 100 datasets and evaluated the posterior mean $\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\gamma}_i]$ with respect to the prior on $\boldsymbol{\beta}|\boldsymbol{\gamma}$ specified in Section 4.1 for $i = 1, 2$ and 3.

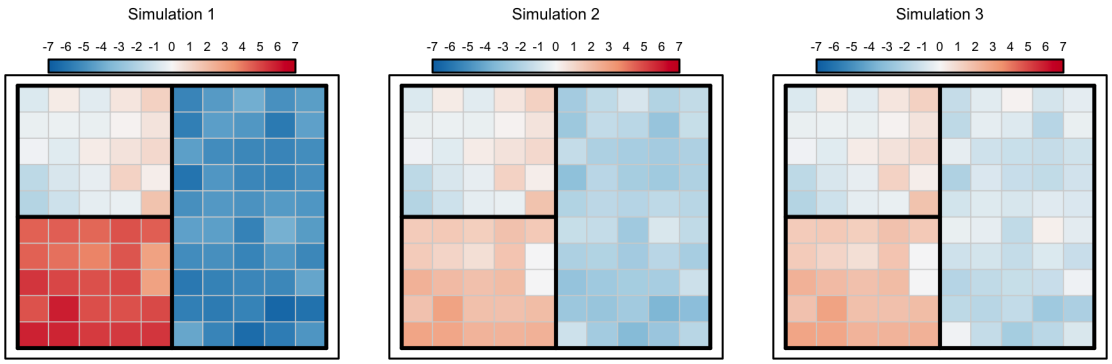


Figure 10: Three specifications of $\boldsymbol{\beta}$ used in our simulation study. Notice that as the cluster means μ_1, μ_2 and μ_3 become less separated, the discontinuities in the β_i's across clusters diminishes

Table 6 shows the approximate risk of each estimator in each of these simulated settings.

Table 6: Approximate risk, averaged over 100 Monte Carlo simulations, of $\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\gamma}_i]$. The correctly specified estimator outperforms the two mis-specified estimators in all three simulations, though the degree of improvement is diminished as the cluster means are made more similar

	Simulation 1	Simulation 2	Simulation 3
$\boldsymbol{\gamma}_1$	1.9100	1.6322	1.6070
$\boldsymbol{\gamma}_2$	1.6368	1.5847	1.5848
$\boldsymbol{\gamma}_3$	**1.5724**	**1.5724**	**1.5724**

In Simulation 1, when the clusters of β_i's are very well-separated, we find that the mis-specified estimators based on the partitions $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ have much larger risk than the correctly specified estimator based on $\boldsymbol{\gamma}_3$. The estimators based on $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ attempt to borrow strength across the cluster borders, shown in bold in Figure 10. This introduces substantial bias in our estimation of the β_i's that sit along these borders. When the cluster means are less well-separated (i.e. Simulations 2 and 3), we see that the risks of misspecified estimators are somewhat closer to the risk of the correctly specified estimator. In these settings, sharing information across the true clusters does not introduce as much bias as it did in Simulation 1.

As seen in the example, knowledge of the underlying spatial partition $\boldsymbol{\gamma}$ can yield improved estimation of $\boldsymbol{\beta}$. In practice, of course, we rarely know $\boldsymbol{\gamma}$ and there may be considerable uncertainty about which of two similar partitions would lead to improved estimation. To model this uncertainty and incorporate it into our estimation of $\boldsymbol{\beta}$, we specify a prior $\pi(\boldsymbol{\gamma})$ over all spatial partitions. Formally, this fully Bayesian approach suggests estimating $\boldsymbol{\beta}$ with the unconditional posterior mean

$$\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}] = \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma}|\mathbf{Y})\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\gamma}].$$

Unfortunately, the number of potential spatial partitions grows rapidly with the number of spatial regions n and evaluating the sum above exactly is impractical. For even moderately large n, the space of possible spatial partitions is so vast that the computational cost of

MCMC is often prohibitive. Instead, we propose a two-step approximation: using our particle optimization framework, we will attempt to identify the L partitions $\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(L)}$ with largest posterior probability $\pi(\boldsymbol{\gamma}|\mathbf{Y})$. We then form the approximate estimator

$$\hat{\boldsymbol{\beta}}_L = \sum_{\ell=1}^{L} w_\ell \mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\gamma}^{(\ell)}]$$

where $w_\ell \propto \pi(\boldsymbol{\gamma}^{(\ell)}|\mathbf{Y})$ and $\sum_\ell w_\ell = 1$. We may view $\hat{\boldsymbol{\beta}}_L$ as a re-normalized truncation of $\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}]$.

4.1. Model and Particle Search Strategy

Before specifying our likelihood model and prior hierarchy of $\boldsymbol{\beta}|\boldsymbol{\gamma}$, we introduce some notation. Let $A \in \mathbb{R}^{n\times n}$ be the adjacency matrix of the spatial units where $a_{ii'} = 1$ whenever regions i and i' are adjacent geographically and 0 otherwise. Let D be the diagonal matrix with $d_{ii} = \sum_{i'} a_{ii'}$ and let $A^* = D - A$ be the unnormalized Laplacian of A. Given a $n \times n$ matrix M and a subset $S \subset [n]$, let M_S be the square sub-matrix of M with rows and columns indexed by the elements of S. In a slight abuse of notation, we will let M_S^* be the unnormalized Laplacian of the square sub-matrix M_S.

4.1.1. Likelihood

Let $\mathbf{y}_i \in \mathbb{R}^T$ be the vector of observed dependent variable and let $\mathbf{x}_i \in \mathbb{R}^T$ be the vector of covariates correspondent to the region i. Assuming that $\mathbf{y}_i$ and $\mathbf{x}_i$ have been re-centered, we consider the simple linear regression model

$$\mathbf{y}_i = \beta_i \mathbf{x}_i + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\varepsilon}_i \sim \mathrm{N}_T\left(\mathbf{0}_T, \sigma_i^2 I_T\right)$, independently in each region. Now suppose that $\boldsymbol{\gamma} = \{C_1, \ldots, C_K\}$ is a partition of the spatial units. We will assume that the residual variances σ_i^2 are constant within clusters but not between clusters and we also allow the β_i's to vary in a spatially smooth fashion between regions in the cluster. We further assume that the K collections

$\boldsymbol{\beta}_{C_k} = \{\beta_i : i \in C_k\}$ are independent, given $\boldsymbol{\gamma}$. With this assumption, it is enough to specify a prior hierarchy for each $\boldsymbol{\beta}_{C_k}$.

To this end, first suppose that C_k consists of $n_k \geq 2$ regions. We introduce parameters μ_{C_k} and $\sigma^2_{C_k}$ and place a CAR prior on $\boldsymbol{\beta}_{C_k}$ so that for each $i \in C_k$we have

$$\beta_i | \boldsymbol{\beta}_{C_k,-i}, \sigma^2_k, \mu_{C_k} \sim \mathrm{N}\left(\frac{(1-\rho)\mu_{C_k} + \rho \sum_{j \in C_k} a_{i,j}\beta_j}{1 - \rho + \rho \sum_{j \in C_k} a_{i,j}}, \frac{a\sigma^2_{C_k}}{1 - \rho + \rho \sum_{j \in C_k} a_{i,j}} \right),$$

where $a > 0$ is some fixed positive constant. The above conditional distribution gives rise to the marginal distribution

$$\boldsymbol{\beta}_{C_k} | \mu_{C_k}, \sigma^2_{C_k}, \boldsymbol{\gamma} \sim \mathrm{N}_{n_k}\left(\mu_{C_k}\mathbf{1}_k, a\sigma^2_{C_k}\left[\rho A^*_{C_k} + (1-\rho)I_{n_k}\right]^{-1} \right).$$

We further model $\mu_{C_k} | \sigma^2_{C_k}, \boldsymbol{\gamma} \sim \mathrm{N}(0, b\sigma^2_{C_k})$, where $b > 0$ is some fixed positive constant. If instead $S_k = \{i_{k,1}\}$ consisted of a single spatial unit, then we simply model $\beta_{i_{k,1}} | \sigma^2 \sim \mathrm{N}\left(0, \sigma^2_k\left(\frac{a}{1-\rho} + b\right)\right)$. Note that this is the marginal distribution under the CAR model described above for a single β_i corresponding to a region that is not connected to any other region within the cluster. To complete our prior specification, we take $\sigma^2_{C_k} \sim$ Inverse Gamma $\left(\alpha, \frac{\nu}{2}\right)$ where $\alpha, \nu > 0$ are fixed hyper parameters. Because our prior for $\boldsymbol{\beta}$ factorizes over the clusters, we know $p(\mathbf{Y}|\boldsymbol{\gamma})$ does as well: $p(\mathbf{Y}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} p(\mathbf{Y}_{C_k}|\boldsymbol{\gamma})$. In this way, our model falls within the class of Bayesian partition models of Holmes et al. (1999) Moreover, by taking advantage of our conditionally conjugate prior specification, we can derive closed-form expressions for $p(\mathbf{Y}_{C_k}|\boldsymbol{\gamma})$ and for the posterior means $\mathbb{E}[\boldsymbol{\beta}_i|\mathbf{Y}, \boldsymbol{\gamma}]$.

4.1.2. Particle Search Strategy

In order to identify the top spatial partitions *a posteriori*, we need to modify the search strategy through the space of partitions from the previous chapter. This is because we wish to restrict our attention only to partitions whose clusters are spatially connected. To facilitate local refinement of the particle set, we consider the following proposals. First, for

each cluster C_k and spatial unit $i \in C_k$, we propose moving i to a new singleton cluster. We term these "island" proposals. Next, for each spatial cluster C_k and spatial unit $i \in C_k$ such that i is adjacent to another cluster $C_{k'}$, we propose moving i to cluster $C_{k'}$. We term these "border" proposals. For both the island and border proposals, if removing spatial unit i from cluster C_k results in a spatially disconnected cluster, we treat the resulting connected components as separate clusters. When the number of spatial clusters n is quite large, instead of exhaustively considering every island or border proposal, we have found it useful to restrict attention to those spatial units i such that the corresponding posterior mean $\mathbb{E}[\beta_i|\mathbf{Y}, \boldsymbol{\gamma}]$ is in the top or bottom 5% within the cluster. In a sense, then, these moves can be viewed as re-allocating outlying parameter estimates.

In addition to local proposals, we consider coarse proposals that allow us to move across the partition space more efficiently. Similar to the previous chapter, we consider "merge" proposals in which we merge each cluster with an adjacent cluster. We once again consider Split-Merge proposals. In the previous chapter, to split cluster C_k, we picked two indices $i, j \in C_k$ at random to anchor the two new clusters and allocated the remaining indices to these two clusters based on how close the data was to Y_i and Y_j. In our spatial clustering setting, we could easily deploy a similar strategy, anchoring the two new clusters at two randomly selected spatial units i and j. Then we can allocate spatial unit $h \in C_k$ to the new cluster anchored by i (resp. j) if the posterior mean $\mathbb{E}[\beta_h|\mathbf{Y}, \boldsymbol{\gamma}]$ is closer to $\mathbb{E}[\beta_i|\mathbf{Y}, \boldsymbol{\gamma}]$ (resp. $\mathbb{E}[\beta_j|\mathbf{Y}, \boldsymbol{\gamma}]$). Unfortunately, the resulting clusters may not be spatially connected.

A popular approach to clustering arbitrary data points $x_1, \ldots, x_n$ is spectral clustering (see von Luxburg, 2007, for a comprehensive review). Spectral clustering starts by forming a similarity matrix, $\mathcal{K} = (k_{ij})$, which measures the pairwise similarity between the points. For instance, with continuous data, it is quite common to use an exponential kernel $k_{ij} = \exp\left\{-\frac{1}{2}\|x_i - x_j\|_2^2\right\}$ to encode similarity. To split the data into K clusters, one first creates an $n \times K$ matrix V using the eigenvectors corresponding to the K smallest non-zero eigenvalues of the Laplacian of $\mathcal{K}$. One then applies a standard clustering method

like k-means to the rows of V. In essence, spectral clustering works by running k-means (or something similar) in a latent space whose dimension is typically substantially smaller than the original dimension of the data.

We propose to use spectral clustering to split C_k into two pieces as follows. First, suppose $C_k = \{i_1, \ldots, i_{n_k}\}$ and let $\hat{\sigma}^2_{\beta,k}$ be the sample variance of the posterior means $\mathbb{E}[\beta_i|\mathbf{Y}, \boldsymbol{\gamma}]$ within cluster C_k. We then construct a *similarity matrix* $\mathcal{K} \in \mathbb{R}^{n_k \times n_k}$ with entries

$$k_{s,t} = \exp\left\{-\frac{(\mathbb{E}[\beta_{i_s}|\mathbf{Y}, \boldsymbol{\gamma}] - \mathbb{E}[\beta_{i_t}|\mathbf{Y}, \boldsymbol{\gamma}])^2}{2\hat{\sigma}^2_{\beta,k}}\right\}.$$

If we were to simply apply spectral clustering using $\mathcal{K}$, there is no guarantee that the resulting clusters will be spatially connected. To enforce connectedness, we modify our similarity matrix by taking the Schur product $\mathcal{K} \circ A_{C_k}$, whose (s,t) entry is equal to zero if spatial units i_s and i_t are not adjacent and equal to $k_{s,t}$ otherwise. We then find the smallest two non-zero eigenvalues of the new similarity matrix $\mathcal{K} \circ A_{C_k}$ and apply k-means to the corresponding matrix V, described above. By zeroing out the similarities between the non-adjacent β_i's, we ensure that the resulting two clusters are spatially connected.

4.2. Simulated Example

To demonstrate our method, we return to Simulation 1 from the example earlier, where the β_i's fell into three well-separated clusters. We run our particle optimization scheme with $L = 100$ particles and a Ewens-Pitman(1) prior on $\boldsymbol{\gamma}$[1]. Figure 11 shows the top 9 unique particles identified, along with the cumulative importance weight and multiplicity of each.

[1]Because our search strategy looks only at partitions whose clusters are spatially connected, our prior is technically the Ewens-Pitman(1) prior restricted to the space of such partitions.

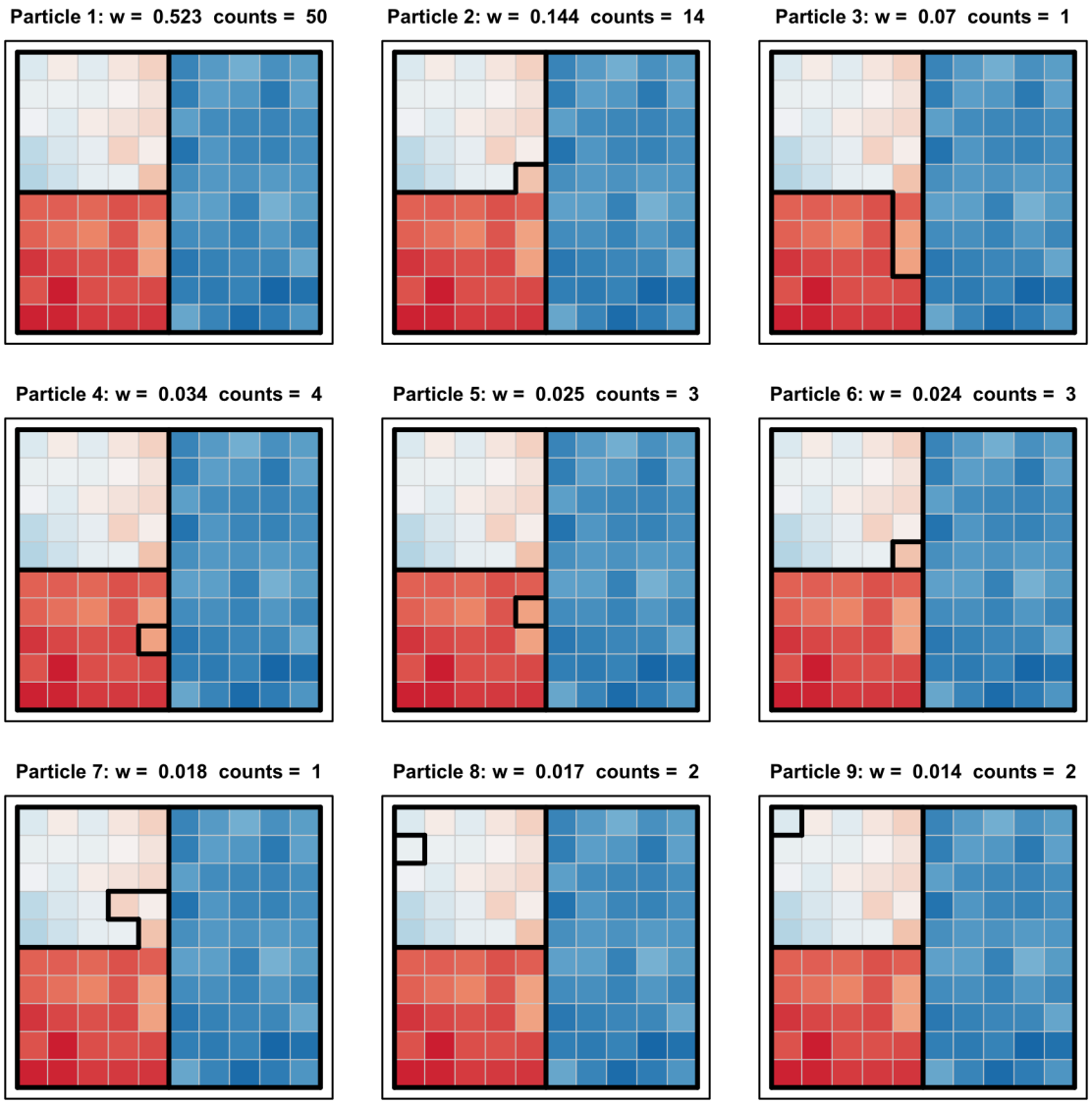


Figure 11: The top nine identified partitions. Observe that 50 of the 100 particles coincided with the true spatial partition

Immediately, we see that the top partition discovered is the true partition. Further, the next eight particles with highest importance weight are all very similar to the true partition. It would appears that our particle system has navigated to a dominant mode and then explored locally. Figure 12 shows a histogram of the number of unique particles discovered when we repeat this simulation 100 times, keeping $\boldsymbol{\beta}$ fixed across simulations.

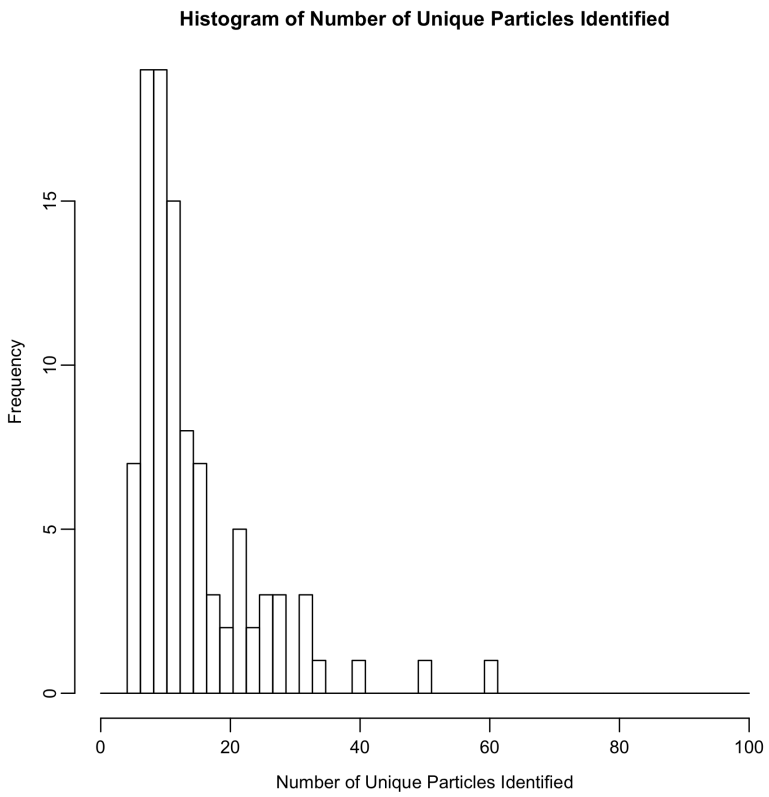


Figure 12: Histogram of the number of unique particles discovered when we initialized with 100 particles.

In the vast majority of these simulations, the top partition discovered was the true partition, with the exceptions differing from the truth in the allocation of only a handful of spatial units. However, we see that in almost all of these replications, there was substantial redundancy in our particle set. This suggests that the entropy penalty did not induce sufficient repulsion to identify 100 unique particles. We will return to this point in Chapter 6.

Turning our attention to $\boldsymbol{\beta}$, Figure 13 plots the risk of our approximate estimator as a function of the level of truncation simulation standard errors. Also shown are the approximate risks of the estimators based on the fixed partitions $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$ used in the earlier example.

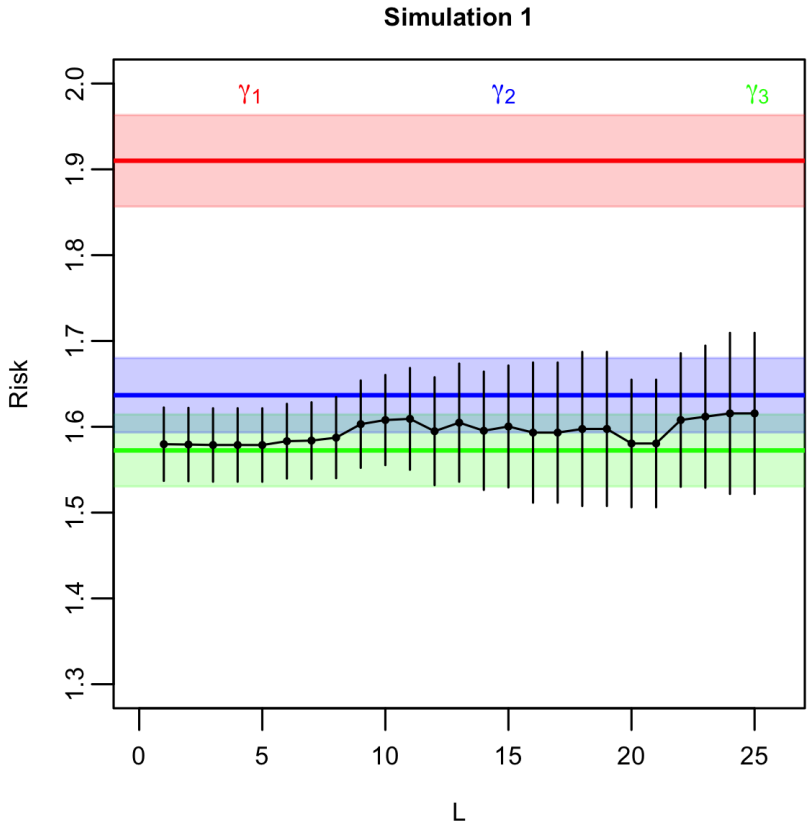


Figure 13: Estimates and 95% confidence intervals for the risk of the approximate estimator (black) as a function of truncation level computed over 100 Monte Carlo simulations when the clusters are well-separated. Also shown are the estimated risk and 95% confidence intervals for estimators based on the mis-specified partitions $\boldsymbol{\gamma}_1$ (red) and $\boldsymbol{\gamma}_2$ (blue) and for the true partition $\boldsymbol{\gamma}_3$ (green).

In this example, when we only use the top partition to estimate $\boldsymbol{\beta}$, our risk appears to nearly coincide with the risk of the oracle estimator. This is because across our simulations, the top partitions discovered was either the true partition used to generate $\boldsymbol{\beta}$ or one that differed only in the allocation of a few spatial units. This variability in the top partition selected explains the small gap between the estimated oracle risk shown in green and the risk of our approximate estimator at truncation level $L = 1$. In a sense, this gap is the price we must pay for using our data twice, once to select the estimator and once again to compute the estimate. As we increase the truncation level and start averaging over many

partition-specific estimators, our risk begins to change. Clearly, if we let the truncation level grow all the way up to the total number of spatial partitions, the risk will converge to the risk of unconditional posterior mean $\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}]$. Interestingly, though, the figure suggests that this convergence may not be monotonic. We also note that as the truncation level increased, the standard error of the estimated risk increased, reflecting the fact that our procedure rarely identified more than 20 unique particles.

Of course, when the means are well separated, this is precisely the behavior we would like to observe: our procedure is able to find the correct (or at least close to correct) partitioning of the spatial regions and our approximate estimator can realize substantial gains in risk over estimators based on mis-specified spatial partitions. This naturally raises the question of how well our procedure performs when the clusters are less well-separated. To probe this, we repeat the simulation study above using the specification of $\boldsymbol{\beta}$ from Simulations 3 in the earlier example. Figures 14 and 15 are the analogs of Figures 11 and 13, respectively.

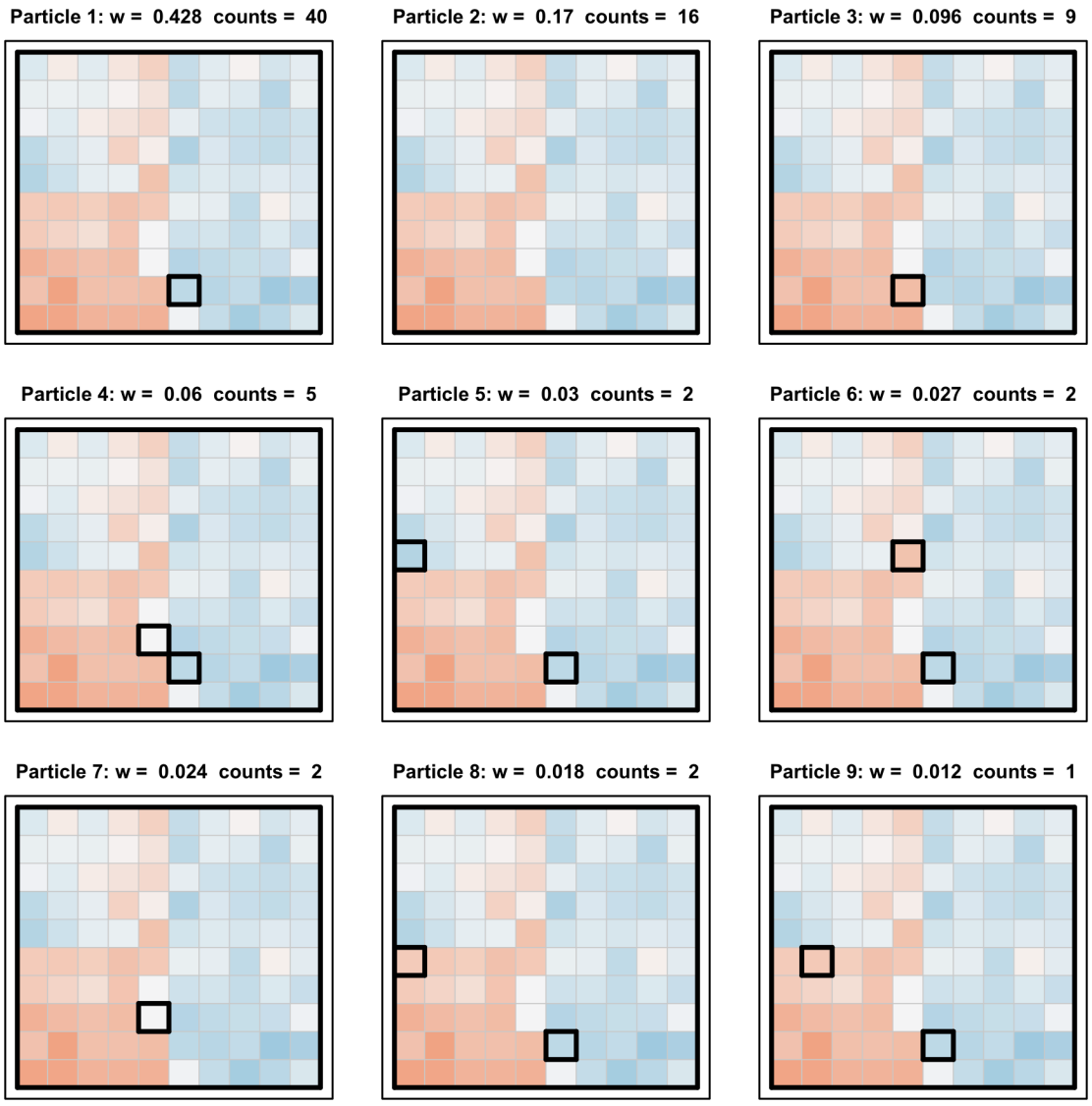


Figure 14: The top nine unique partitions identified in Simulation 3. Moreover the remaining eight partitions shown are all very similar to the true partition, differing only in the allocation of a small number of spatial units.

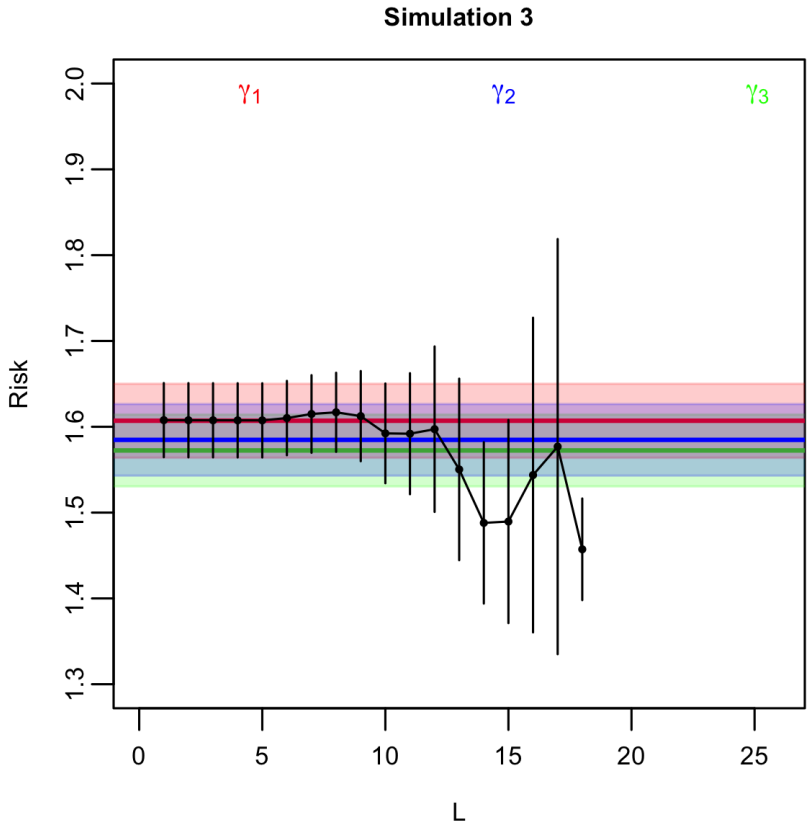


Figure 15: Estimates and 95% confidence intervals for the risk of the approximate estimator (black) as a function of truncation level computed over 100 Monte Carlo simulations when the clusters are not well-separated. Also shown are the estimated risk and 95% confidence intervals for estimators based on the mis-specified partitions $\boldsymbol{\gamma}_1$ (red) and $\boldsymbol{\gamma}_2$ (blue) and for the true partition $\boldsymbol{\gamma}_3$ (green).

In this simulation, since the cluster means are not well-separated, it is hardly surprising to see that the risk of the estimators based on the single partitions $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$ are quite similar. That said, the oracle estimator based on the partition $\boldsymbol{\gamma}_3$ does seem to have a slightly better risk but the apparent difference is within the expected range of Monte Carlo variation. Rather surprisingly, our approximate estimator appears to offer substantial risk reduction when we average over the top $15 - 20$ partitions, though the difference in risk is also within the expected range of simulation-to-simulation variability. Nevertheless, it is suggestive that there might be more interesting partitions of the data that can help estimate

$\boldsymbol{\beta}$. These helpful partitions are generally not *a priori* obvious and it is encouraging that our method seems to be able to identify them.

4.3. Discussion

In this chapter we have proposed a model for spatially smoothing linear regression models fit within individual spatial units. Our model belongs to the class of Bayesian partition models introduced in Holmes et al. (1999). Our prior specification induces spatial smoothness of the regression slopes β_i within clusters through a conditionally auto-regressive prior. Rather than rely on MCMC to estimate the collection of slopes $\boldsymbol{\beta}$ in the presence of the true latent spatial partition $\boldsymbol{\gamma}$, we approximate the marginal posterior mean $\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}]$ by a weighted average of conditional means $\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\gamma}]$ that corresponding to promising spatial partitions. These promising spatial partitions are identified using our particle optimization procedure. In simulation settings, this approximation is seen to have reasonable risk as compared to the conditional posterior mean corresponding to the true underlying partition.

Recall from Figures 11 and 14 that it appeared our particle system navigated to a dominant mode of the posterior and remained in its vicinity. Moreover, we found that many of the particles were redundant; indeed in Simulation 1 with well-separated cluster means, 50 of the 100 particles accumulated at the top partition discovered. While the entropy penalty in our objective function induces a certain degree of repulsion between particles, it was clearly not enough to overcome the gravitational pull of the dominant mode. In Chapter 6, we consider a wider class of optimization problems inspired by Equation (3.2) that will encourage a more complete exploration of the space.

CHAPTER 5 : Estimating (Partially?) Exchangeable Normal Means

In the previous chapter, we proposed approximating the marginal posterior mean $\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}]$ by forming a posterior-weighted average of several conditional posterior means $\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\gamma}]$ corresponding to models $\boldsymbol{\gamma}$ selected using our particle optimization framework. In the context of spatial clustering of linear regression models, we found that these approximate estimators displayed promising risk properties. We observed further that there was a price to pay for using the data twice, once for model selection and once for estimation. In this chapter, we probe this price in more detail using the slightly simpler canonical normal means problem in which we observe $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\beta}, I)$ and are uncertain about which β_i's are similar in value. This uncertainty manifests itself as uncertainty about which β_i's are exchangeable which other β_i's and we review the notion of partial exchangeability in Section 5.1. We then introduce a model for estimating $\boldsymbol{\beta}$ in the presence of uncertainty about the partial exchangeability structure in Section 5.2 and construct our approximate estimator in Section 5.3. We finally study the risk of the approximate estimator in Section 5.4. A central challenge is the fact that the model selection and estimation are not independent.

5.1. Whence Partial Exchangeability?

Consider data hierarchically organized into J groups and suppose that for each group j, we observe n_j independent pairs of data $(y_{1,j}, \mathbf{x}_{1,j}), \ldots, \left(y_{n_j,j}, \mathbf{x}_{n_j,j}\right)$, with outcomes $y_{i,j} \in \mathbb{R}$ and covariates $\mathbf{x}_{i,j} \in \mathbb{R}^p$. Following Gelman et al. (2008), we will sometimes refer to these data as arising from J different "experiments." Consider, for instance, the standard Gaussian linear regression, modeling $y_{i,j} = \mathbf{x}_{i,j}\beta_j + \varepsilon_{i,j}$ where $\varepsilon_{i,j} \sim \mathrm{N}(0, \sigma^2)$ independently for all i, j. We wish to estimate the collection $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_J\}$ under squared error loss in the presence of uncertainty about certain structural assumptions about the individual parameters $\beta_1, \ldots, \beta_J$. We take a Bayesian approach, so that the specific target of our modeling efforts is the posterior mean $\hat{\boldsymbol{\beta}} = \mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}]$ where $\mathbf{Y}$ denotes the collection $\{y_{i,j} : 1 \leq j \leq J, 1 \leq i \leq n_j\}$. Our particular interest in this paper is to estimate $\boldsymbol{\beta}$ in the

presence of uncertain structural assumptions about the group-specific parameters β_j.

To make matters somewhat more precise, consider two rather extreme approaches to estimating $\boldsymbol{\beta}$. We could assume that in fact all of the β_j's are equal to some common value, β_0 say. Estimation would proceed by pooling all of the data across experiments, specifying a single prior on β_0, and computing $\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}]$ in closed form or approximating it via numerical integration or MCMC. While simple, the strong equality restriction may be unpalatable in nearly all applied settings. At the other end of the spectrum, we could of course analyze the data from each experiment separately. Such an approach makes no attempt to exploit any potential homogeneity across experiments. That is, at this extreme, data from experiment j is not permitted to inform estimation of $\beta_{j'}$ for $j \neq j'$.

Bayesian hierarchical modeling seeks a middle ground between these two approaches, enabling us to "borrow strength" across the experiments. Perhaps the simplest hierarchical prior specification begins by introducing hyper-parameters $\beta_0 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}$ and modeling $\beta_1, \ldots, \beta_J | \beta_0 \sim \mathrm{N}_p(\beta_0, \Sigma)$. The prior specification is then completed by fully specifying a prior on (β_0, Σ) . In this specification, we model the β_j's as exchangeable and we may regard each β_j as a noisy measurement of the single hyper-parameter β_0. The hyper-parameter Σ reflects the degree to which the individual β_j's disperse around β_0. The resulting estimates of $\boldsymbol{\beta}$ from this hierarchy can be thought of as a compromise between the two extremes mentioned earlier: we have separate estimates of the β_j's like in the completely separate analyses but these estimates are mutually informative and shrunk towards some common value, β_0. Underlying this simple hierarchical model is the assumption that the β_j's are exchangeable.

As Lindley and Smith wrote in a seminal 1972 paper, the "practical relevance [of the exchangeability assumption] must be assessed before the estimates based on it are used" (Lindley and Smith, 1972). They go on to highlight two situations in which *partial exchangeability* is a more appropriate assumption than exchangeability, writing (emphasis ours)

> if ... our model described the observed yields on n varieties in an agricultural field trial, the exchangeability assumption would be inappropriate if one or more variables were controls and the remainder were experimental. However, the assumption *might* be modified to one of exchangeability within controls and separately within experimental varieties. Similarly with a two-way classification into rows and columns, it might be reasonable to assume separately that the rows and columns were exchangeable. In any application, the particular form of the prior distribution has to be carefully considered.

Hierarchical modeling is now ubiquitous in many fields and, as alluded to in the quote above, careful consideration and exploitation of the exchangeability assumptions is paramount. In the examples put forth in Lindley and Smith (1972), the partial exchangeability structure readily presents itself through auxiliary side information. Increasingly, however, such auxiliary information may be unavailable, introducing considerable uncertainty in the exact partial exchangeability structure. This could have serious ramifications for the eventual estimation. Luckily, the Bayesian framework permits consideration of an exceptionally broad range of partial exchangeability assumptions in a coherent way.

To illustrate this, consider an extreme but important special case of our general problem, the canonical normal means problem. Here we assume that there are no covariates and that $n_j = 1$ so that $y_j \sim \mathrm{N}(\beta_j, \sigma_Y^2)$ independently with σ_Y^2 known. Seminal work in James and Stein (1961) demonstrated that the MLE is inadmissible and is dominated by an estimator that shrinks these estimates uniformly towards zero. Many have subsequently improved on Stein's original shrinkage estimator, with several proposing hierarchical Bayes estimators (see, e.g. Berger and Robert, 1990) and empirical Bayes estimators (see, e.g. Efron and Morris, 1975). One especially important improvement is the Lindley estimator, which is discussed at length in Efron and Morris (1975), that shrinks the MLE to the overall mean $\overline{Y}$ rather than to zero. Common to these estimators is the assumption that the individual means are exchangeable, which permits the "borrowing of strength" across all observations. Stigler (1990) rather evocatively summarized the seemingly paradoxical nature of Stein's result, asking "how can information about the price of apples in Washington and the price of oranges in Florida be used to improve an estimate of the price of French wine?" Of

course, as is well-known, the improvement only manifests itself when the true parameters are very close in value.

In the case of the normal means problem, when partial exchangeability of the β_j's seems more appropriate than exchangeability, we could deploy separate shrinkage estimates to each group of observations. Figure 16 shows two realizations of data (in black), both from a "partially exchangeable normal means" model in which there are 3 groups of means: the first 50 means are equal, the next 25 are equal, and the last 25 are equal. We compare the performance of two estimators on this data. The first, shown in red in Figure 16 is the standard Lindley estimator that shrinks all of the data to the common mean while the second, shown in blue in Figure 16, deploys a Lindely estimator within each group of means separately. Also shown are the estimated risks of both estimators, averaged over 10,000 Monte Carlo simulations.

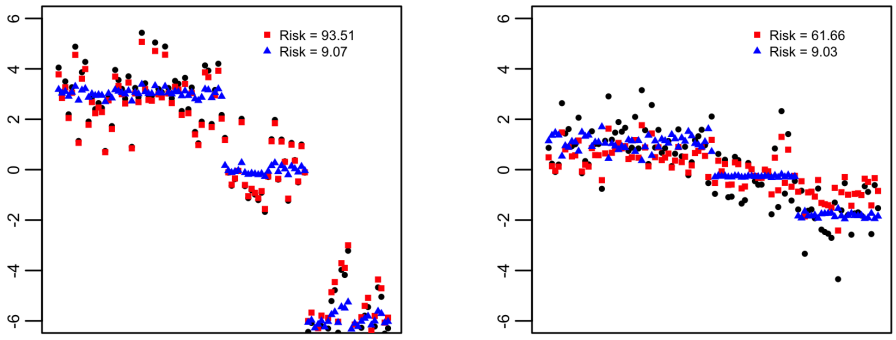


Figure 16: Two realizations of normals means data when the means fall into three groups. The standard Lindley shrinkage estimates of the means are shown in red squares while the estimates obtained by deploying a separate Lindley estimators within each group are shown in blue triangles. When the groups of means are well-separated, the clustered Lindley estimator has substantially better risk than the standard Lindley estimator

5.2. A Multiple Shrinkage Estimator

As is clear from Figure 16, we may realize substantial gains in risk when we correctly exploit the partial exchangeability structure of the underlying normals means. Observe that we may encode this structure with a partition $\boldsymbol{\gamma}$ of the set $[n]$ and place a prior $\pi(\boldsymbol{\gamma})$ over this space to reflect our initial uncertainty about the partial exchangeability of the β_i's. In order to estimate $\boldsymbol{\beta}$ under squared error loss, consider the following prior $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$. For a given partition $\boldsymbol{\gamma} = \{C_1, \ldots, C_K\}$, if the cluster C_k contains $n_k \geq 4$ elements, we introduce hyper-parameters μ_k, σ_k^2 with improper flat hyper-prior $\pi(\mu_k, \sigma_k^2) = 1$ and model $\beta_i \sim \mathrm{N}\left(\mu_k, \sigma_k^2\right)$ independently for $i \in C_k$. If the cluster C_k contains $n_k = 3$ elements, we introduce the single hyper-parameter σ_k^2 with an improper flat hyper-prior $\pi(\sigma_k^2) = 1$ and model $\beta_i \sim \mathrm{N}\left(0, \sigma_k^2\right)$ independently for $i \in P_j$. Finally, if the cluster C_k contains $n_k \leq 2$ elements, we place an improper flat prior on β_i for $i \in P_k$. The prior asserts that β_i and $\beta_{i'}$ are independent given $\boldsymbol{\gamma}$ if the indices i and i' belong to different clusters.

From Equations (3.3) and (3.5) in Stein (1981) and Theorem 1 in George (1986b), we know that the Bayes estimator corresponding to the partition $\boldsymbol{\gamma}$ can be expressed as

$$\delta_{\boldsymbol{\gamma}}(\mathbf{Y}) := \mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\gamma}] = \mathbf{Y} + \nabla \log p(\mathbf{Y}|\boldsymbol{\gamma}).$$

It is not difficult to show that the prior $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ is harmonic, i.e.

$$\Delta\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \sum_{i=1}^{n} \frac{\partial^2}{\partial Y_i^2} \pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) = 0.$$

This in turn implies that the marginal density $m(\mathbf{Y}|\boldsymbol{\gamma})$ is also harmonic and by Corollary 1 in Stein (1981), we can conclude that $\delta_{\boldsymbol{\gamma}}$ is a minimax estimator of $\boldsymbol{\beta}$.

Under this hierarchy, the marginal density $p(\mathbf{Y}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} p(\mathbf{Y}_{C_k}|\boldsymbol{\gamma})$ factorizes over the

clusters and for $i \in C_k$ we compute

$$(\nabla \log p(\mathbf{Y}_{C_k}|\boldsymbol{\gamma}))_i = \begin{cases} -\frac{2\left(Y_i - \overline{Y}_k\right)}{S_k} \times \gamma\left(\frac{n_k-1}{2}, \frac{S_k}{2\sigma_Y^2}\right) / \gamma\left(\frac{n_k-3}{2}, \frac{S_k}{2\sigma_Y^2}\right) & \text{if } n_k \geq 4 \\ -\frac{2Y_i}{S_k} \times \gamma\left(\frac{3}{2}, \frac{S_k}{2\sigma_y^2}\right) / \gamma\left(\frac{1}{2}, \frac{S_k}{2\sigma_y^2}\right) & \text{if } n_k = 3 \\ 0 & \text{if } n_k \leq 2 \end{cases}$$

where $\gamma(s,x) = \int_0^x t^{s-1}\mathrm{e}^{-t}dt$ is the lower incomplete gamma function, $\overline{Y}_k$ is the mean of the observations in cluster C_k, and

$$S_k = \begin{cases} \sum_{i \in C_k} \left(Y_i - \overline{Y}_k\right)^2 & \text{if } n_k \geq 4 \\ \sum_{i \in C_k} Y_i^2 & \text{if } n_k = 3 \end{cases}$$

From these expressions, we see that the estimator $\delta_{\boldsymbol{\gamma}}(\mathbf{Y})$ shrinks the MLE in each cluster separately. Moreover, for clusters with at least four elements, it shrink the MLE to the cluster mean, à la the Lindley estimator, for clusters with three elements, it shrinks to 0, à la the original James-Stein estimator, and for clusters with two or one elements, it performs no shrinkage. As suggested by Figure 16, we can expect $\delta_{\boldsymbol{\gamma}}$ to confer substantial improvements in risk when $\boldsymbol{\gamma}$ is known.

In general, though, $\boldsymbol{\gamma}$ is not known *a priori* and to express our uncertainty about the underlying partial exchangeability structure we may augment our prior specification with a hyper-prior $\pi(\boldsymbol{\gamma})$ over the space of all partitions. The resulting Bayes estimator of $\boldsymbol{\beta}$ under squared error loss is simply a posterior probability-weighted average of all possible $\delta_{\boldsymbol{\gamma}}$'s:

$$\delta_*(\mathbf{Y}) = \mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}] = \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma}|\mathbf{Y})\delta_{\boldsymbol{\gamma}}(\mathbf{Y})$$

As described in George (1986b,a,c), we may write $\delta_*(\mathbf{Y}) = \mathbf{Y} + \nabla \log p^*(\mathbf{Y})$ where

$$p_*(\mathbf{Y}) = \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma}) p(\mathbf{Y}|\boldsymbol{\gamma}).$$

Since p_* is a convex combination of harmonic functions, it too is harmonic, ensuring the minimaxity of the *multiple shrinkage estimator* δ_*. In a sense, δ_* adaptively mixes all of the individual shrinkage estimators $\delta_{\boldsymbol{\gamma}}$ corresponding to individual partitions $\boldsymbol{\gamma}$.

5.3. Approximate Multiple Shrinkage

Since it requires summing over all possible partitions, the multiple shrinkage estimator δ_* cannot be evaluated exactly. Despite the incredible number of terms in this sum, we nevertheless might expect that the vast majority of the $\pi(\boldsymbol{\gamma}|\mathbf{Y})$ are negligible. In particular, if the β_i's fell into well-separated groups, we could expect most of the posterior probability to concentrate at a comparatively small number of partitions. In light of this, letting $\Gamma_L = \left\{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(L)}\right\}$ be the L partitions with largest marginal posterior probability, a natural approximation of δ_* is

$$\delta_{\Gamma_L} = \sum_{\ell=1}^{L} w_\ell \mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\gamma} = \boldsymbol{\gamma}^{(\ell)}]$$

where the w_ℓ's are non-negative weights summing to one with $w_\ell \propto \pi(\boldsymbol{\gamma}^{(\ell)}|\mathbf{Y})$ for $\ell = 1, \ldots, L$. In other words, δ_{Γ_L} approximates δ^* by averaging over the L estimators $\delta_{\boldsymbol{\gamma}} = \mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\gamma}]$'s with the largest weights $\pi(\boldsymbol{\gamma}|\mathbf{Y})$ in the original sum. We may use the particle optimization scheme outlined earlier to estimate Γ_L and form the *approximate multiple shrinkage estimator* $\delta_{\hat{\Gamma}_L}$. We moreover follow the same search strategy as the one outlines in Section 3.4 when considering the Gaussian mixture model.

5.3.1. Simulation Studies

We begin with some simulated examples. Throughout our simulations, we will use $n = 100$ and run our particle optimization framework with $L = 25$ and a standard Ewens-Pitman prior on $\boldsymbol{\gamma}$ with hyper-parameter $\eta = 1$. We consider three different specifications of $\boldsymbol{\beta}$:

Simulation 1 $\beta_1, \ldots, \beta_{50} \sim \text{Uniform}(-5.5, -4.5)$ and $\beta_{51}, \ldots, \beta_{100} \sim \text{Uniform}(4.5, 5.5)$

Simulation 2 $\beta_1, \ldots, \beta_{100} \sim \text{Uniform}(-0.5, 0.5)$

Simulation 3 $\beta_1 = \cdots = \beta_{50} = 0$ and $\beta_{51} = \cdots = \beta_{100} = 2$.

Figure 17 shows the top four partitions discovered when running our particle optimization procedure with these settings in each simulation. In Simulation 1, though we started with 25 particles initially, we find that at convergence, we ultimately find 23 unique particles. It is very reassuring to see that none of the 23 unique partitions identified by our method cluster any of the first 50 observations with any of the second 50 observations. The top two partitions identified both cluster observations 1 – 50 together and splits observations 58 and 84 away from the rest of 51 – 100. The difference between them is slight: the first partition separates observations 58 and 84 into two singleton clusters while the second partition puts them into a single cluster of size two. Despite this difference, we see immediately that they have equal posterior probability under the specified prior on $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. This is entirely due to the choice of $\eta = 1$ in the prior on $\boldsymbol{\gamma}$. We also find that the first and last partitions discovered by our method differ in log-posterior by only 1.17. In other words, the posterior places just over three times the posterior mass on the top partition discovered than on the 23^{rd} partition discovered. Interestingly, the posterior distribution places about 11 times less mass on the true partition we used to generate the data, $\{\{1, \ldots, 50\}, \{51, \ldots, 100\}\}$, than on the top partition identified.

In Simulation 2, when the β_i's are all relatively close to 0, we find that the top partition discovered is exactly the one used to generate the data: $\{1, 2, \ldots, 100\}$. It is interesting to note that the posterior places an almost identical amount of mass on the second best partition, which splits observation 14 (corresponding to the minimum observed Y_i) away from the other 99 observations. Like in Simulation 1, it is reassuring to find that most of the high posterior probability partitions discovered by our method are rather similar to the true underlying partition: the vast majority of observations are grouped together into a single cluster with some of the extreme values split off on their own (see, for instance, panel (F), (G), and (H) in Figure 17).

This is decidedly not the case with Simulation 3, in which the first 50 β_i's were equal to 0 and the next 50 were equal to 2. We find that none of the high posterior probability partitions discovered are at all similar to the partition used to generate the data $\{\{1,\ldots,50\},\{51,\ldots,100\}\}$. In fact, the difference in log-posterior value between the top partition identified and the true partition is 40.5, suggesting that the posterior distribution places about 3.9×10^{17} times more mass on the top partition than on the true partition. This is perhaps not as shocking as it may initially seem. After all, though the underlying means are different, the data does not suggest abandoning the usual exchangeability assumption in favor of the assumption that first 50 β_i's are exchangeable and independent of the second 50. As we will see shortly, however, this can have serious ramifications for estimation.

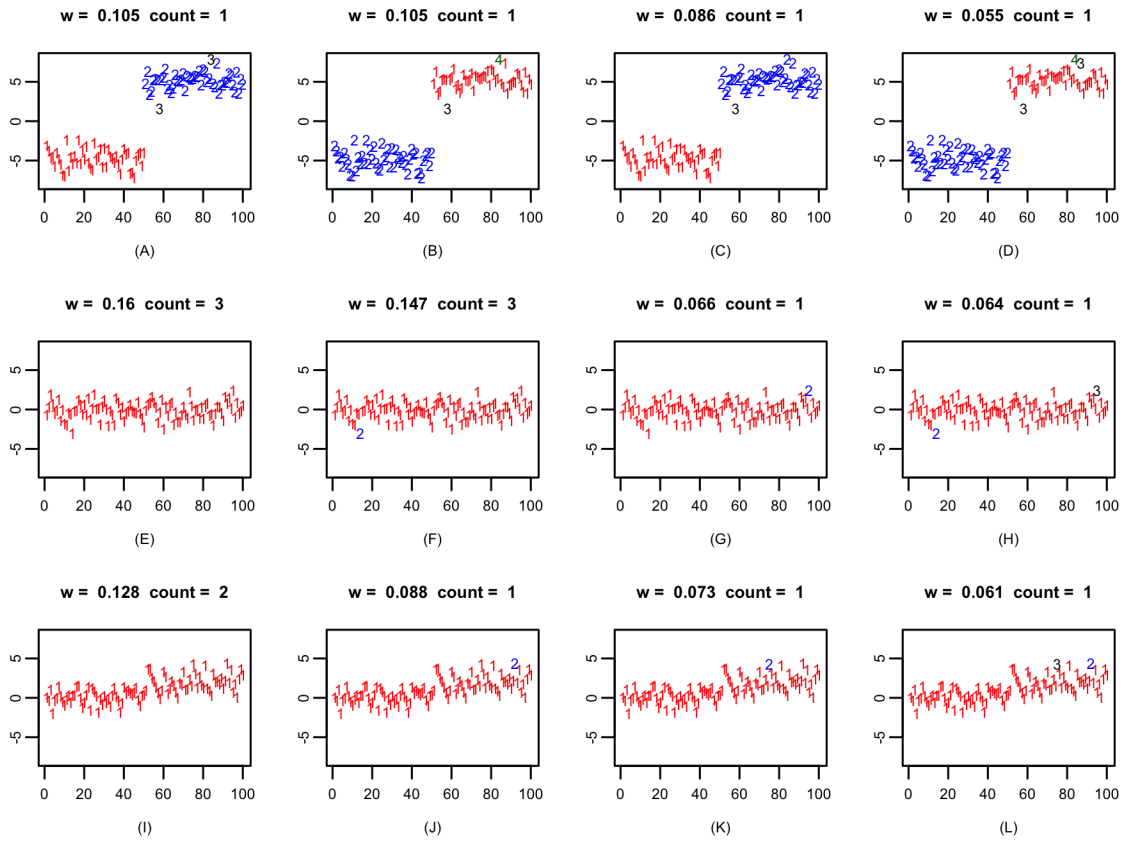


Figure 17: Top partitions discovered in Simulations 1 ((A) – (D)), 2 ((E) – (H)) and 3 ((I) – (J)). Notice that in Simulations 2 and 3, when the means are all quite similar, the top partitions discovered are very close to the partition consisting of a single cluster.

Recall that the approximate multiple shrinkage estimator is a re-weighted truncation of the full multiple shrinkage estimator to the leading L terms of δ_*. Letting L increase, the approximate estimator will converge point-wise to the full estimator and so will its risk. Figure 18 shows the estimated risk (averaged over 100 Monte Carlo simulations) of the approximate multiple shrinkage estimator as a function of the truncation level. We compare this estimated risk to the estimated risk of using the oracle shrinkage estimator, i.e. the one corresponding to the true clustering of the β_i's used to generate the data (green) and to the estimated risk of the usual Lindley estimator (red).

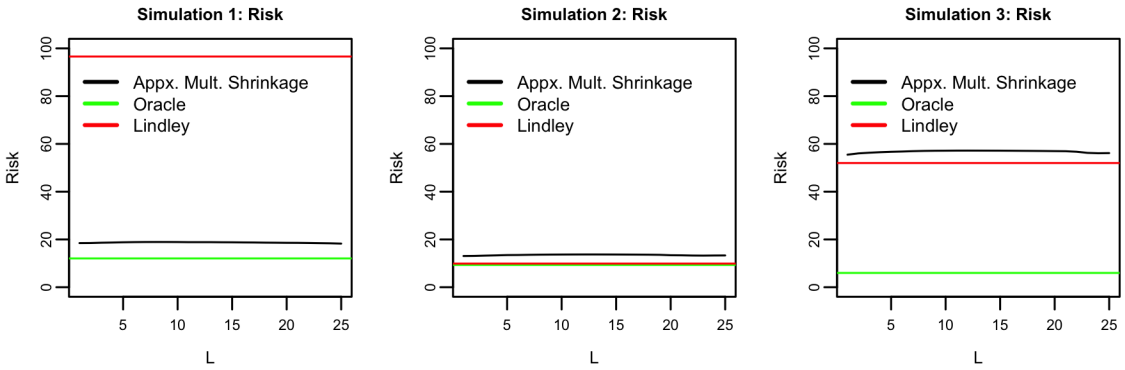


Figure 18: Estimated risk of approximate multiple shrinkage estimator (black) as a function of truncation level. Also shown are the estimated risks of the oracle shrinkage estimator (green) and Lindley estimator. The gap between the green and black curves reflects the price we pay for estimation after selection.

In light of panels (A) – (H) in Figure 17, it is perhaps unsurprising to see that the risk of the approximate multiple shrinkage estimator is quite close to the oracle risk in Simulation 1, when the two groups of means are well-separated, and in Simulation 2, when the means fall into a single homogeneous group. In both of these simulations, the approximate multiple shrinkage estimator appears to provide substantial improvements in risk relative to the MLE precisely because the top partitions discovered are very close to the oracle partition. This ensures that the approximate multiple shrinkage estimator is formed from estimators that can provide substantial risk reduction. As we would expect, the approximate multiple shrinkage estimator has much smaller risk than the Lindley estimator in Simulation 1,

because it is able to adapt to the appropriate partial exchangeability structure underlying the data. In Simulation 2, when the exchangeability assumption is reasonable, the Lindley estimator, of course, has a risk almost exactly equal to the oracle risk.

Simulation 3, however, demonstrates that the approximate multiple shrinkage estimator will not always mimic the oracle estimator. In this setting, it actually appears to have worse risk than the Lindley estimator at all levels of truncation. This is not at all surprising, in light of panels (I) – (L) of Figure 17: the top partitions *a posteriori* kept most of the observations together in a single cluster and the approximate multiple shrinkage estimator was formed with estimators that did not offer nearly the same amount of risk reduction as the oracle. Nevertheless, it is interesting to see that the apparent risk is still less than the minimax risk, though the improvement is not as substantial as it was in Simulations 1 and 2.

It is important to point out that, while the risk of the approximate estimator will converge to the risk of the full estimator as we increase the truncation level L, there are no guarantees about the monotonicity of that convergence. In fact, in the examples in Figure 18, we see that the risk does not appear to monotonic in L. This is a byproduct of the irregularity of the selected partitions. As a somewhat extreme example, consider the realization of data from Simulation 2 in Figure 17. For that dataset, the second best partition *a posteriori* split the 14^{th} observation, corresponding to the smallest observed datapoint, away from the rest. In another realization, when Y_{14} is closer to the center of $\mathbf{Y}$, this partition would surely not be the second best partition *a posteriori*. We also note that in all three of these simulations, regardless of the truncation level, the approximate multiple shrinkage estimator appears to have a risk greater than the oracle risk. We may view the gap between the two risk curves as representing the price we must pay for using our data to select the estimators used to form the approximate multiple shrinkage estimator. We will return to this point in the next section.

5.3.2. Efron & Morris' Batting Averages

To illustrate the performance of the Lindely estimator, Efron and Morris (1975) considered the batting averages of Major League Baseball players in the 1970 baseball season. By April 26, 1970, each player listed in Table 7 had completed exactly 45 at-bats. Efron and Morris (1975) used shrinkage estimation to predict each player's batting average over the remainder of the season from their batting average in the first 45 at-bats. If we let X_i be the observed batting average of player i after $N = 45$ at-bats, we assume $NX_i \sim \text{Bin}(N, p_i)$ where p_i is the true season batting average for player i. If we further define $Y_i = N^{1/2}\text{arcsin}(2X - 1)$, then $Y_i \sim N(\beta_i, 1)$ approximately, where $\beta_i = N^{-1/2}\arcsin(2p_i - 1)$. Table 7 is a reproduction of Table 1 in Efron and Morris (1975) and lists X_i, p_i, Y_i and β_i.

Table 7: Reproduction of Table 1 from Efron and Morris (1975) showing batting performances of 18 players with 45 at-bats through April 26, 1970.

Player	X_i	p_i	At-bats remaining	Y_i	β_i
Clemente	.400	.346	367	-1.35	-2.10
Robinson	378	298	426	-1.66	-2.79
Howard	.356	.276	521	-1.97	-3.11
Johnstone	.333	.222	275	-2.28	-3.96
Berry	.311	.273	418	-2.60	-3.17
Spencer	.311	.270	466	-2.60	-3.20
Kessinger	.289	.263	586	-2.92	-3.32
L. Alvarado	.267	.210	138	-3.26	-4.15
Santo	.244	.269	510	-3.60	-3.23
Swoboda	.244	.23	200	-3.60	-3.83
Unser	.222	.264	277	-3.95	-3.30
Williams	.222	.256	270	-3.95	-3.43
Scott	.222	.303	435	-3.95	-2.71
Petrocelli	.222	.264	538	-3.95	-3.30
E. Rodriguez	.222	.226	186	-3.95	-3.89
Campaneris	.200	.285	558	-4.32	-2.98
Munson	.178	.316	408	-4.70	-2.53
Alvis	.156	.200	70	-5.10	-4.32

Recall that the original James-Stein estimator provides incredible improvements in risk only

when the underlying means are close to zero. In a similar way, the risk of Lindley (or Efron-Morris) estimator is governed by the variance of the β_i's: if they are all rather similar, it realizes substantial risk improvements and if they are highly variable, the estimator performs little shrinkage. Efron and Morris (1975) demonstrated that even with the inclusion of the "unusually good hitter" Roberto Clemente, shrinking towards the overall average $\overline{Y}$ yielded slight improvements over the usual James-Stein estimator. However, an important question remains: can we improve our estimation of the vector $\boldsymbol{\beta}$ by leaving Clemente's extreme observation unshrunk and shrinking the remaining 17 observations towards their overall mean? Moreover, are there other partitions of the data that yield even better shrinkage estimators?

To investigate this possibility, we ran our particle optimization framework with $L = 100$ particles and Ewens-Pitman(1) prior on the underlying partition $\boldsymbol{\gamma}$. In our analysis, we perturbed the data $\mathbf{Y}$ slightly with the addition of $\mathrm{N}(0, 10^{-6})$ noise so that none of the observations were identical. Our method identifies the partition $\{1, 2, \ldots, 18\}$, which groups all of the players together just like Efron and Morris (1975), as the top partition, with an importance weight of 0.43. The partition with the next highest importance weight separated Clemente from the remaining 17 players, who are clustered together. The importance weight of this partition is 0.32, meaning that under our model, it is about 1.3 times less likely *a posteriori* than the top partition discovered. Interestingly, the next partition discovered splits the worst batter, Max Alvis, away from the remaining players, who are similarly clustered together. The importance weight of this partition is 0.25. Figure 19 shows the approximate multiple shrinkage estimates that averages over the top three discovered partitions (green circles) as well as the usual Lindley estimates (blue triangles).

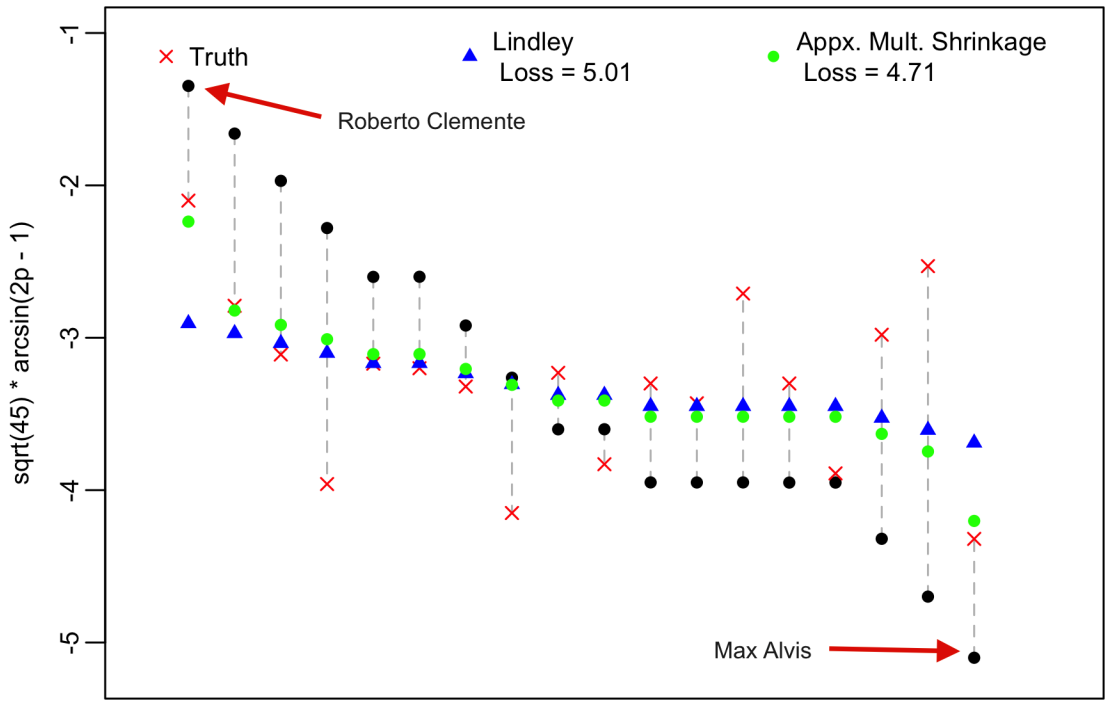


Figure 19: Transformed batting averages over the first 45 at-bats (black) and remainder of season (red). Approximate multiple shrinkage estimates are shown in green while the standard Lindley estimates are shown in blue.

Since each term of our approximate multiple shrinkage estimator clusters the middle 16 players together, it is not surprising to see that the Lindley estimates of their batting averages are very similar to the approximate multiple shrinkage estimates. We find, however, that the Lindley estimator has a squared error loss of 5.01 while our approximate multiple shrinkage estimator has a loss of 4.71. The reduction in loss is entirely driven by the improved estimation of Clemente's and Alvis' transformed batting averages. Indeed, we see that the approximate multiple shrinkage estimates nearly coincide with the true transformed batting averages for both of these players.

5.4. Towards a Better Understanding of Risk

Before investigate the risk of the approximate estimator, we need some additional notation. For a given collection $\Gamma = \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L\}$ of L distinct partitions, let

$$\delta_\Gamma(\mathbf{Y}) = \sum_{\ell=1}^{L} \frac{\pi(\boldsymbol{\gamma}_\ell|\mathbf{Y})}{\sum_{\ell'=1}^{L} \pi(\boldsymbol{\gamma}_{\ell'}|\mathbf{Y})} \delta_{\boldsymbol{\gamma}_\ell}(\mathbf{Y})$$

be the multiple shrinkage estimator that adaptively mixes over the estimators indexed by partitions in Γ. Notice that for each fixed L, the risk of the approximate multiple shrinkage estimator may be decomposed as

$$R(\delta_{\hat{\Gamma}_L}, \boldsymbol{\beta}) = \sum_{\Gamma} \mathbb{E}\left[(\delta_\Gamma(\mathbf{Y}) - \boldsymbol{\beta})^2 \, \mathbb{I}\left(\hat{\Gamma}_L = \Gamma\right)\right],$$

where the sum is taken over all collections Γ of L partitions.

In general, the two terms in the expectation above are not independent. As a result, we cannot easily separate the selection problem (i.e. determining $\hat{\Gamma}_L$) and the estimation problem when studying the risk. To develop further insight into this post-selection inference problem, let us first consider the much simpler but highly idealized setting in which selection can be done independently of estimation. As an example of such a situation, we could imagine that we had two independent realizations of data $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ which are both distributed as $\mathrm{N}(\boldsymbol{\beta}, I)$. We could first use $\mathbf{Y}^{(1)}$ to determine $\hat{\Gamma}_L$ and then evaluate $\delta_{\hat{\Gamma}_L}$ using $\mathbf{Y}^{(2)}$. Denoting this estimator $\delta^{(1,2)}$, it is easy to see that the risk can be written as

$$R(\delta^{(1,2)}, \boldsymbol{\beta}) = \sum_{\Gamma} \mathbb{E}\left[\left(\delta_\Gamma(\mathbf{Y}^{(2)}) - \boldsymbol{\beta}\right)^2 \mathbb{I}\left(\hat{\Gamma}_L(\mathbf{Y}^{(1)}) = \Gamma\right)\right]$$

where the expectation is taken over the joint distribution of $(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. Since $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ are independent, the selection and estimation are independent so the risk can be expressed

as a weighted average of the risk of each multiple shrinkage estimator δ_Γ:

$$R(\delta^{(1,2)}, \boldsymbol{\beta}) = \sum_{\Gamma} R(\delta_\Gamma, \boldsymbol{\beta}) \mathbb{P}(\hat{\Gamma}_L = \Gamma).$$

Under our prior specification for $\boldsymbol{\beta}|\boldsymbol{\gamma}$, we know that δ_Γ is minimax (i.e. $R(\delta_\Gamma, \boldsymbol{\beta}) \leq n$) for each Γ and we can conclude that the idealized estimator $\delta^{(1,2)}$ is minimax.

This argument demonstrates that if we are able to select $\hat{\Gamma}_L$ independently of estimating $\boldsymbol{\beta}$, our approximate estimator could still be minimax. Having established the minimaxity, it is very natural to ask how close the risk of the idealized approximate estimator is to the risk of the full multiple shrinkage estimator. To probe this, let $\hat{r}_{\boldsymbol{\gamma}}$ be the unbiased estimate of risk of the estimator $\delta_{\boldsymbol{\gamma}}$. We will let $w_{\boldsymbol{\gamma}} = \pi(\boldsymbol{\gamma}|\mathbf{Y})$ be the marginal posterior probability of $\boldsymbol{\gamma}$ with respect to the original prior $\pi(\boldsymbol{\gamma})$. Let $\tilde{\pi}(\boldsymbol{\gamma})$ be the restriction of this prior to the pre-specified subset $\hat{\Gamma}_L$ and let $\tilde{w}_{\boldsymbol{\gamma}}$ be the posterior probability of $\boldsymbol{\gamma}$ with respect to $\tilde{\pi}(\boldsymbol{\gamma})$. We may write the full multiple shrinkage estimator $\delta_* = \sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \delta_{\boldsymbol{\gamma}}$ and the approximate estimator $\tilde{\delta}_* = \sum_{\boldsymbol{\gamma}} \tilde{w}_{\boldsymbol{\gamma}} \delta_{\boldsymbol{\gamma}}$. Let $\hat{r}_*$ and $\hat{\tilde{r}}$ be unbiased estimates of the risks of δ_* and $\tilde{\delta}_*$, respectively.

By Theorem 2.9 of Leung and Barron (2006), we have

$$\hat{r}_* = \sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \left[\hat{r}_{\boldsymbol{\gamma}} - \|\delta_* - \delta_{\boldsymbol{\gamma}}\|_2^2 - 2 \left(\nabla \log w_{\boldsymbol{\gamma}}\right)^\top (\delta_* - \delta_{\boldsymbol{\gamma}}) \right],$$

where we take $w_{\boldsymbol{\gamma}} \nabla \log w_{\boldsymbol{\gamma}} = 0$ whenever $w_{\boldsymbol{\gamma}} = 0$.

Since $\delta_{\boldsymbol{\gamma}} = Y + \nabla \log p(\mathbf{Y}|\boldsymbol{\gamma})$, we see

$$\begin{aligned} \sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \left(\nabla \log w_{\boldsymbol{\gamma}}\right)^\top (\delta_* - \delta_{\boldsymbol{\gamma}}) = &\sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \left(\delta_{\boldsymbol{\gamma}} - \mathbf{Y}\right)^\top (\delta_* - \delta_{\boldsymbol{\gamma}}) \\ &+ \sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \left(\nabla \log \pi(\boldsymbol{\gamma})\right)^\top (\delta_* - \delta_{\boldsymbol{\gamma}}) \\ &- \sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \left(\nabla \log p(\mathbf{Y})\right)^\top (\delta_* - \delta_{\boldsymbol{\gamma}}) \end{aligned}$$

Since $\pi(\boldsymbol{\gamma})$ does not depend on $\mathbf{Y}$, the second sum above is equal to zero. Moreover, the third sum above is zero, since $\delta_* = \sum w_{\boldsymbol{\gamma}} \delta_{\boldsymbol{\gamma}}$ and since $\pi(\mathbf{Y})$ does not depend on $\boldsymbol{\gamma}$. It is easy to see that if h is a vector function not depending on $\boldsymbol{\gamma}$ that

$$\sum_{\boldsymbol{\gamma}} w_m h^\top (\delta_* - \delta_{\boldsymbol{\gamma}}) = 0$$

Thus,

$$\sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} (\delta_{\boldsymbol{\gamma}} - \mathbf{Y})^\top (\delta_* - \delta_{\boldsymbol{\gamma}}) = -\sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \|\delta_* - \delta_{\boldsymbol{\gamma}}\|_2^2.$$

Therefore, we can write the unbiased risk estimate of δ_* as

$$\hat{r}_* = \sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \left(\hat{r}_{\boldsymbol{\gamma}} + \|\delta_* - \delta_{\boldsymbol{\gamma}}\|_2^2 \right).$$

This is Corollary 3 of Leung and Barron (2006). We may similarly derive an expression for $\hat{\tilde{r}}$, the unbiased estimate of the risk of the truncated estimator $\tilde{\delta}_*$. An unbiased estimate of the risk difference between $\tilde{\delta}_*$ and δ_* is given by

$$\hat{\tilde{r}}_* - \hat{r}_* = \sum_{\boldsymbol{\gamma}} (w_{\boldsymbol{\gamma}} - \tilde{w}_{\boldsymbol{\gamma}}) \hat{r}_{\boldsymbol{\gamma}} + \sum_{\boldsymbol{\gamma}} \left(w_m \|\delta_* - \delta_{\boldsymbol{\gamma}}\|_2^2 - \tilde{w}_{\boldsymbol{\gamma}} \left\| \tilde{\delta}_* - \delta_{\boldsymbol{\gamma}} \right\|_2^2 \right).$$

Recall that the marginal densities $p(\mathbf{Y}|\boldsymbol{\gamma})$'s are super-harmonic. As a result, we know that $\hat{r}_{\boldsymbol{\gamma}} \leq n$ for each $\boldsymbol{\gamma}$ (Stein, 1981). So we can very trivially bound the first term above by

$$\sum_{\boldsymbol{\gamma}} (w_{\boldsymbol{\gamma}} - \tilde{w}_{\boldsymbol{\gamma}}) \hat{r}_{\boldsymbol{\gamma}} \leq \max_{\boldsymbol{\gamma}} \hat{r}_{\boldsymbol{\gamma}} \times 2\mathrm{TV}(w, \tilde{w}) \leq 2n\mathrm{TV}(w, \tilde{w}),$$

where $\mathrm{TV}(w, \tilde{w})$ is the total variation distance between the two posterior distributions over the model space.

Turning to the second term, it is straightforward to compute

$$\sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \left\| \delta_* - \delta_{\boldsymbol{\gamma}} \right\|^2 = - \left\| \delta_* \right\|_2^2 + \sum_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}} \left\| \delta_{\boldsymbol{\gamma}} \right\|_2^2$$

Thus,

$$\sum_{\boldsymbol{\gamma}} \left(w_{\boldsymbol{\gamma}} \| \delta_* - \delta_{\boldsymbol{\gamma}} \|_2^2 - \tilde{w}_{\boldsymbol{\gamma}} \| \tilde{\delta}_* - \delta_{\boldsymbol{\gamma}} \|_2^2 \right) = \sum_{\boldsymbol{\gamma}} \left(w_{\boldsymbol{\gamma}} - \tilde{w}_{\boldsymbol{\gamma}} \right) \| \delta_{\boldsymbol{\gamma}} \|_2^2 + \| \tilde{\delta}_* \|_2^2 - \| \delta_* \|_2^2$$

We may bound the sum on the right-hand side from above by

$$2\mathrm{TV}(w, \tilde{w}) \times \max_{\boldsymbol{\gamma}} \| \delta_{\boldsymbol{\gamma}} \|_2^2$$

Observe that

$$\| \tilde{\delta}_* \|_2^2 - \| \delta_* \|_2^2 = \sum_{\boldsymbol{\gamma}, \boldsymbol{\gamma}'} \left(\tilde{w}_{\boldsymbol{\gamma}} \tilde{w_{\boldsymbol{\gamma}'}} - w_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}'} \right) \delta_{\boldsymbol{\gamma}}^\top \delta_{\boldsymbol{\gamma}'}$$

which we may bound from above by

$$\max_{\boldsymbol{\gamma}, \boldsymbol{\gamma}'} \left| \delta_{\boldsymbol{\gamma}}^\top \delta_{\boldsymbol{\gamma}'} \right| \sum_{\boldsymbol{\gamma}, \boldsymbol{\gamma}'} \left| \tilde{w}_{\boldsymbol{\gamma}} \tilde{w}_{\boldsymbol{\gamma}'} - w_{\boldsymbol{\gamma}} w_{\boldsymbol{\gamma}'} \right|$$

The sum in this expression is precisely twice the total variation distance between the product measures $w \times w$ and $\tilde{w} \times \tilde{w}$.

Putting everything together, we have

$$\hat{\tilde{r}} - \hat{r} \leq 2\mathrm{TV}(w, \tilde{w}) \times \left(\max_{\boldsymbol{\gamma}} \hat{r}_{\boldsymbol{\gamma}} + \max_{\boldsymbol{\gamma}} \| \delta_{\boldsymbol{\gamma}} \|_2^2 \right) + 2\mathrm{TV}(w \times w, \tilde{w} \times \tilde{w}) \times \max_{\boldsymbol{\gamma}, \boldsymbol{\gamma}'} \left| \delta_{\boldsymbol{\gamma}}^\top \delta_{\boldsymbol{\gamma}'} \right|.$$

We know that $\max_{\boldsymbol{\gamma}} \hat{r}_{\boldsymbol{\gamma}} \leq n$ and

$$\max_{\boldsymbol{\gamma}} \| \delta_{\boldsymbol{\gamma}} \|_2^2 \leq \max_{\boldsymbol{\gamma}, \boldsymbol{\gamma}'} \left| \delta_{\boldsymbol{\gamma}}^\top \delta_{\boldsymbol{\gamma}'} \right|.$$

By Pinsker's Inequality, we know that $TV(w, \tilde{w}) \leq \sqrt{KL(\tilde{w}, w)/2}$ and since $KL(\tilde{w} \times \tilde{w}, w \times w) = 2KL(\tilde{w}, w)$, so we conclude

$$\hat{\tilde{r}}_* - \hat{r}_* \leq (KL(\tilde{w}, w))^{\frac{1}{2}} \times \left(\sqrt{2}n + \left(2 + \sqrt{2}\right) \max_{\gamma, \gamma'} \left| \delta_\gamma^\top \delta_{\gamma'} \right| \right) \tag{5.1}$$

This argument shows that an unbiased estimate of the difference in risk between the approximate estimator and full estimator is governed by the Kullback-Leibler divergence between the distributions $\tilde{w}$ and w. In other words, the more mass posterior that $\hat{\Gamma}_L$ captures, the smaller the upper bound on the risk difference. Importantly, though, it also depends on the alignment of the estimators δ_γ.

5.4.1. Risk via Perturbations

The above derivations are valid so long as the selection of $\hat{\Gamma}_L$ is done independently of the estimation of $\boldsymbol{\beta}$. In practice, we cannot hope to form an estimator like $\delta^{(1,2)}$ from a single realization of the data. However, we may take inspiration from Tian and Taylor (2016) and perturb our data as follows. Given $\mathbf{Y}$ and some fixed constant $\alpha > 0$, we draw $\boldsymbol{\omega} \sim \mathrm{N}(0, I_n)$ and form $\mathbf{Y}^+ = \mathbf{Y} + \alpha\boldsymbol{\omega}$ and $\mathbf{Y}^- = \mathbf{Y} - \alpha^{-1}\boldsymbol{\omega}$. By construction, $\mathbf{Y}^+$ and $\mathbf{Y}^-$ are independent and we can consider using $\mathbf{Y}^+$ to select $\hat{\Gamma}$ and estimate $\boldsymbol{\beta}$ using $\mathbf{Y}^-$. We denote the resulting estimator $\delta_\alpha^{+,-}$. Note that $\mathbf{Y}^+$ and $\mathbf{Y}^-$ do not have unit variance and we must make suitable modification to our selection and estimation procedures to account for this. By an argument virtually identical to the one used to show that $\delta^{(1,2)}$ was minimax, we have

$$R(\delta_\alpha^{+,-}, \boldsymbol{\beta}) = \sum_\Gamma R(\delta_\Gamma(\mathbf{Y}^-), \boldsymbol{\beta}) \mathbb{P}(\hat{\Gamma}(\mathbf{Y}^+) = \Gamma)$$

Since $\mathbf{Y}^- \sim \mathrm{N}\left(\boldsymbol{\beta}, (1 + \alpha^{-2})I\right)$, we know that, by construction $R(\delta_\Gamma(\mathbf{Y}^{(-)}), \boldsymbol{\beta}) \leq n(1+\alpha^{-2})$ for all $\boldsymbol{\beta}$. This means that the estimator $\delta_\alpha^{+,-}$ will never have risk exceeding $n(1 + \alpha^{-2})$.

While this upper bound is somewhat re-assuring, are there $\boldsymbol{\beta}$'s such that $R(\delta_\Gamma(\mathbf{Y}^-), \boldsymbol{\beta}) > n$ for all possible Γ? If so, then this would mean that the risk of $\delta_\alpha^{+,-}$ would exceed n

at such a β, implying that the estimator is not minimax for the original problem. We conjecture that such a $\boldsymbol{\beta}$ can be constructed by taking $\beta_i = c_n \times i$ where c_n is some large constant, possibly depending on n. If c_n is large enough, then all of the β_i's will be well-separated and our particle optimization procedure will identify several partitions that contain only singletons and clusters of size two. According to our prior specification for $\boldsymbol{\beta}|\boldsymbol{\gamma}$, the estimators corresponding to such partitions coincide with the MLE. As a result, the risk will be always be $n\left(1+\alpha^{-2}\right)$.

Though the estimator $\delta_\alpha^{+,-}$ is likely not minimax, its risk will not exceed $n(1+\alpha^{-2})$. Clearly if α is large, then the departure from minimaxity will be small. However, when α is large, it may be harder to detect the clustering structure as the additional noise from $\alpha\boldsymbol{\omega}$ will dominate the signal in $\mathbf{Y}^+$ unless the true clusters of β_i's are very well-separated. When this happens, we cannot reasonably expect to select partitions that are close to the true partition meaning that $\delta_\alpha^{+,-}$ will be formed using estimators that offer little risk reduction relative to the MLE. On the other hand, if we take α to be very small, then we might expect our selection procedure to return reasonable partitions. Unfortunately, when α is small, the signal contained in $\mathbf{Y}^-$ may be dominated by the additional noise $\alpha^-\boldsymbol{\omega}$ in $\mathbf{Y}^-$. As a result, even if we were forming $\delta_\alpha^{+,-}$ from reasonable estimators, the additional noise in $\mathbf{Y}^-$ could potentially inflate our risk well beyond n, the minimax risk of the original problem. Figure 20 illustrates this problem, showing a single realization of data $\mathbf{Y}$ from each of the simulation settings above along with two perturbed data sets.

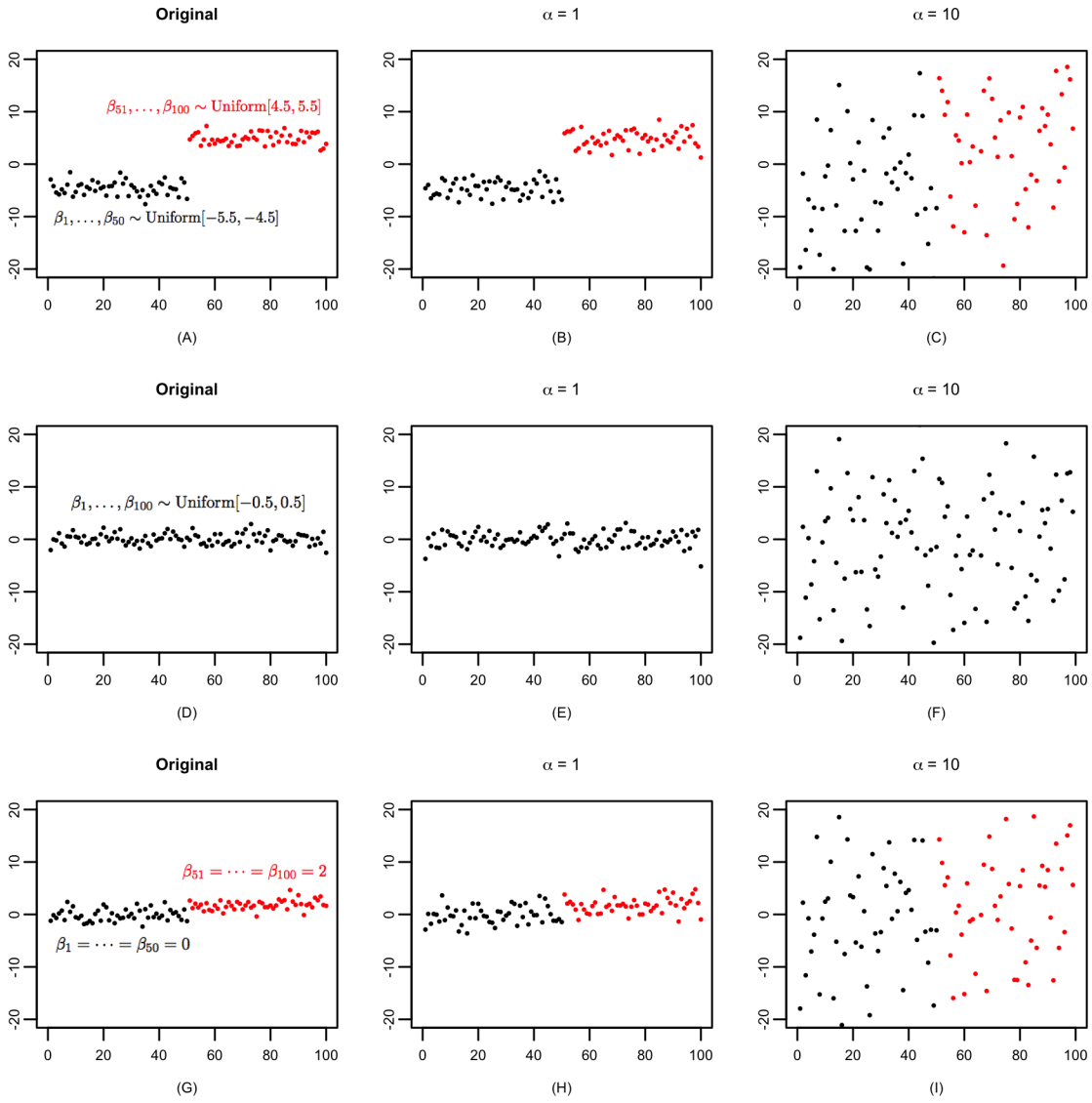


Figure 20: Examples of perturbed datasets

In Simulation 1, when the two groups of β_i's are rather well-separated, the cluster structure is very evident when α is small. In Simulation 3, on the other hand, when the groups of β_i's are not particularly well-separated, even a small amount of additional noise is enough to completely obscure the true clustering structure. Simulation 2 is illuminating in so far as all of the β_i's were initially quite close and the true partition consisted of only one cluster. The additional noise does not suggest any new clustering structure. Unsurprisingly, when

α is large, the perturbed dataset from each of the three simulations are indistinguishable.

CHAPTER 6 : Conclusion and Future Directions

6.1. Next Steps

In the previous three chapters, we have attempted to find several promising partitions of our data by solving the optimization problem in Equation (3.2). In all of our demonstrations, we found that once our particle system was near a dominant posterior mode, nearly all of the split, merge, and Split-Merge proposals had substantially less posterior mass than the local proposals which re-allocated individual observations. This behavior is largely an artifact of the Ewens-Pitman prior, which tends to discourage splitting clusters into two non-singletons sub-clusters and leaving them un-merged with other existing clusters. This is because for any positive integers n_1, n_2, we have $\Gamma(n_1+n_2) > \Gamma(n_1)\Gamma(n_2)$. In other words, proposals which split a cluster into two non-singleton clusters were almost never accepted as any increase in log-likelihood was drowned out by the decrease in log-prior. Across our simulations, we found that split proposals were accepted only when the cluster being split combined sub-clusters whose means were very well-separated. At present, though, we have no general intuition as to the degree of separation required before the gains in log-likelihood realized by a split outpace the influence of the prior. In contrast, local proposals, including those which moved individual observations into their own singleton clusters, were overwhelmingly accepted, since both the log-likelihood and log-prior are relatively insensitive to such single-index updates. As a result, once the particle system navigates near a dominant mode, it tends to remain in the vicinity.

From one perspective, this is perfectly acceptable; after all, we are targeting the collection of partitions which contain the most posterior mass and we should be quite pleased that our particle system does not move to regions of the space that contain substantially less probability. From another perspective, though, the resulting particle system does not seem like a very useful summary of the posterior landscape. It is unclear whether the posterior probability concentrates around a single dominant mode or if there are other pockets of

posterior mass that are "far away" from this mode. If it is the latter, then it would appear that the entropy term in Equation (3.2) provides insufficient repulsion for particles to escape the gravitational pull of a dominant mode and move into the another, potentially slightly sub-optimal region of the space. This points to a fundamental question at the heart of our proposed approach: is the variational approximation of $\pi(\boldsymbol{\gamma}|\mathbf{Y})$ a particularly "good" approximation of the full posterior? Figure 21 illustrates this question, showing a discrete distribution with three very clear pockets of probability.

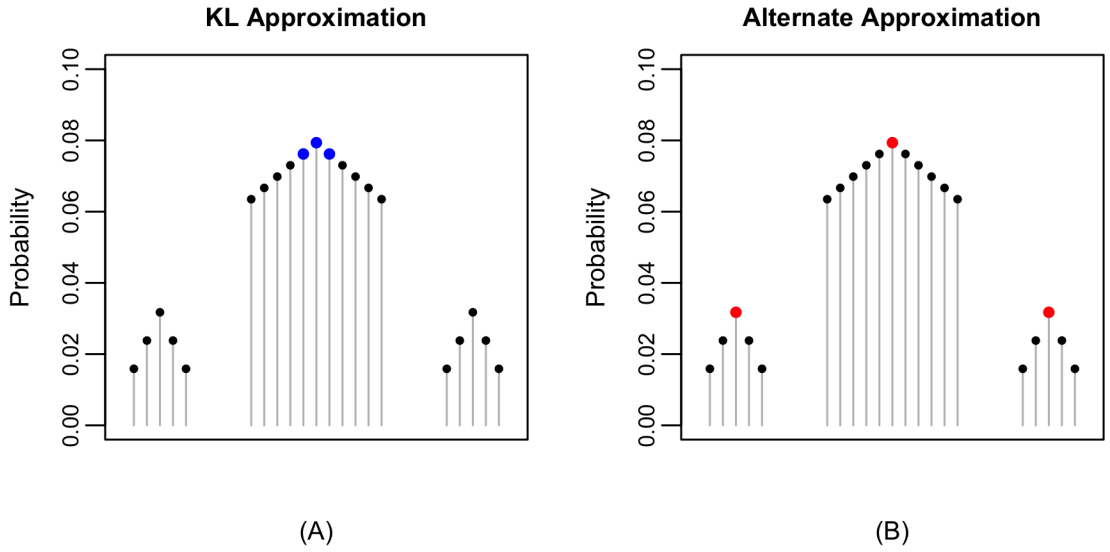


Figure 21: Two approximations of a discrete distribution. (A) shows the best 3-atom approximation in a KL sense, while (B) shows an alternative that better captures the variability

The KL approximation of this distribution within the class $\mathcal{Q}_3$ is shown in blue in Figure 21(A). Though they contain the most posterior mass, by focusing solely on these three points, we gain no information about the pockets of posterior probability located near the other two peaks. One could argue, rather credibly in our opinion, that the approximation shown in Figure 21(B) is much better in this regard. While it captures less of the total posterior probability, it is more representative of the posterior variability. Determining which approximation is better depends, of course, on our ultimate goal: if we simply wish to summarize interesting facets of the posterior, then we would prefer the approximation in

Figure 21(B). On the other hand, if our goal is estimation and model-averaging, as it is in Chapters 4 and 5, the utility of considering estimators corresponding to low posterior probability models is less clear. In light of the idealized upper bound on the unbiased estimate of the difference in risk between the full posterior mean and the truncated estimator in Chapter 5, we would certainly prefer the KL approximation shown in Figure 21(A). Rather than quibble about the superiority of either approximation, we focus on modifying our particle optimization procedure in order to simultaneously summarize the posterior landscape while capturing as much mass as possible.

A simple solution, of course, would be to dramatically increase the number of particles. In reference to Figure 21, if we sought the best approximation within $\mathcal{Q}_{13}$, we would include the two local modes in addition to all of the points in the neighborhood of the global mode. Unfortunately, it is never clear *a priori* how many particles would be enough. A more pernicious problem is our particle system's tendency to remain stuck in the vicinity of a dominant mode. While this behavior is partly due to our particular choice of prior $\pi(\boldsymbol{\gamma})$, it is worth exploring solutions more general than "change the prior." As a first step, we may follow the lead of Ročková (2017) and consider the family of optimization problems indexed by a tuning parameter $\lambda > 0$:

$$(\Gamma^*, \mathbf{w}^*) = \underset{(\Gamma, \mathbf{w})}{\arg\max} \left\{ \sum_{\boldsymbol{\gamma}_\ell} w_\ell \log \pi(\boldsymbol{\gamma}_\ell, \mathbf{Y}) + \lambda H(\Gamma, \mathbf{w}) \right\} \tag{6.1}$$

To solve this problem, we may use exactly the same procedure described in Chapter 3. The only modification comes in updating the importance weights, which are now proportional to $\pi(\boldsymbol{\gamma}_\ell|\mathbf{Y})^{\frac{1}{\lambda}}$. Recall that the first term in the objective is maximized by setting all particles equal to the MAP model $\boldsymbol{\gamma}^{(1)}$ while the entropy term $H(\Gamma, \mathbf{w})$ is maximized by making all of the particles distinct with but with equal importance weights. Now the parameter λ counterbalances these two forces. When λ is large, the price we must pay for redundancy in the particle set is much greater than in our original formulation with $\lambda = 1$. While this

is certainly a step in the right direction, we are still bound by the fact that the entropy $H(\Gamma, \mathbf{w})$ only penalizes **exact** equality in our particle set and does nothing to discourage accepting proposals that are simply *close* to existing particles.

In order to encourage particles to leave the vicinity of a dominant mode and reach other potentially interesting regions of the space, we need to define precisely what it means for two partitions to be "close." Suppose that $\boldsymbol{\gamma} = \{C_1, \ldots, C_K\}$ and $\boldsymbol{\gamma}' = \{C'_1, \ldots, C'_{K'}\}$ are two partitions of $[n]$ into K and K' clusters, respectively. Let N be the $K \times K'$ matrix whose entries $n_{ij} = |C_i \cap C'_j|$ count the number of indices contained in both the i$^{\text{th}}$ cluster C_i of $\boldsymbol{\gamma}$ and the j$^{\text{th}}$ cluster C'_j of $\boldsymbol{\gamma}'$. Binder (1978) introduced a loss function on the space of clusterings that measures the disagreement in all possible pairs of observations between two partitions. A particularly popular version of this loss function, which was studied in both Lau and Green (2007) and and Dahl (2006) is defined as

$$B(\boldsymbol{\gamma}, \boldsymbol{\gamma}') = \frac{1}{2}\left(\sum_{i=1}^{K} |C_i|^2 + \sum_{j=1}^{K'} \left|C'_j\right|^2 - 2\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right).$$

An alternative loss function, the *variation of information*, was introduced by Meilă (2007) and is defined as

$$VI(\boldsymbol{\gamma}, \boldsymbol{\gamma}') = \sum_{i=1}^{K} \frac{|C_i|}{n} \log\left(\frac{|C_i|}{n}\right) + \sum_{j=1}^{K'} \frac{|C'_j|}{n} \log\left(\frac{|C'_j|}{n}\right) - 2\sum_{i=1}^{K}\sum_{j=1}^{K'} \frac{n_{ij}}{n} \log\left(\frac{n_{ij}}{n}\right).$$

It turns out that both $\tilde{B} = \frac{2}{n^2}B$ and VI are bounded metrics on the space of partitions and the partitions at furthest distance under both metrics are the partition consisting of a single cluster and the partition consisting of n singletons clusters (Wade and Ghahramani, 2017). In comparing both metrics, Wade and Ghahramani (2017) point out that if n is an even, square integer, then partitions consisting of two clusters of sizes $\frac{1}{2}(n - \sqrt{n})$ and $\frac{1}{2}(n + \sqrt{n})$ are equally distant under $\tilde{B}$ from these two extremes. They argue that this is rather unappealing, as it implies that the loss of estimating such a partition with only one cluster or with the partition of all singletons is the same, despite the fact that the

former is intuitively much better. For that reason, when constructing posterior credible balls for partitions, Wade and Ghahramani (2017) recommend using VI and we follow their recommendation.

Now that we have a notion of distance between partitions, how can we enable some particles to move away in VI distance from a dominant mode? At a first glance, a natural solution would be to only consider proposals that were far away from the particle being updated in VI distance. This by itself may be sub-optimal as we will be unable to move to high probability partitions that are close to the current particle. That said, this points once again to the tension between optimality and diversity. We instead consider an augmented family of optimization problems:

$$(\Gamma^*, \mathbf{w}^*) = \underset{(\Gamma, \mathbf{w})}{\arg\max} \left\{ \sum_{\boldsymbol{\gamma}_\ell} w_\ell \log \pi(\boldsymbol{\gamma}_\ell, \mathbf{Y}) + \lambda H(\Gamma, \mathbf{w}) + \xi K(\Gamma) \right\} \tag{6.2}$$

where $\xi > 0$ is an additional tuning parameter and $K(\Gamma)$ is a function which encourages distributing the particles throughout the space. We note that we may deploy the same solution strategy as before to solve the problem in Equation (6.2): to update an individual particle, we generate a number of principled proposals and select the one which maximizes the objective, holding all other free parameters fixed. Since the additional penalty $K(\Gamma)$ does not depend on the importance weights $\mathbf{w}$, the update of individual weights w_ℓ's proceeds exactly as before. This is rather attractive: even though for a general (λ, ξ) the solution to Equation (6.2) will no longer be the optimal KL approximation, the importance weights still allow us to assess the relatively posterior probability at each particle.

What should $K(\Gamma)$ look like? Determinantal point processes (DPP) are elegant probabilistic models of negative correlation and are used frequently in machine learning to encourage diversity (see, e.g. Kulesza and Taskar, 2011; Gillenwater et al., 2012; Zou and Adams, 2012; Affandi et al., 2014). Formally, a DPP on a base set Θ is a probability measure on the power set of Θ such that if $\mathcal{T}$ is a random subset of Θ drawn according to $\mathcal{P}$ then for

every $S \subseteq \mathcal{T}$,

$$\mathbb{P}(S \subset \mathcal{T}) = |\mathcal{K}_S|$$

where $\mathcal{K} = (k_{i,j})$ is a symmetric, positive semidefinite similarity matrix whose eigenvalues are between 0 and 1. The matrix $\mathcal{K}$ is known as the marginal kernel. If Θ is discrete and $\vartheta_i, \vartheta_j \in \Theta$ then it can be shown that $\mathbb{P}(\vartheta_i \in \mathcal{T}) = k_{ii}$ and $\mathbb{P}(\vartheta_i, \vartheta_j \in \mathcal{T}) = k_{ii}k_{jj} - k_{ij}^2$. In this way, larger values of k_{ij} imply that ϑ_i and ϑ_j are *less* likely to appear together in the subset $\mathcal{T}$. The class of DPPs is incredibly rich and they have been explored extensively in both the mathematics and machine learning literature. We refer to Kulesza and Taskar (2012) for a much more detailed introduction and review. A particularly useful class of DPPs are *L-ensembles*, first introduced in Borodin and Rains (2005). If L is an arbitrary symmetric matrix with rows and columns are indexed by elements of Θ, an L-ensemble is the measure over the power set of Θ such that the unnormalized probability of sampling a subset $S \subset \Theta$ is proportional to $|L_S|$.

Given an L-ensemble, Kulesza and Taskar (2012) decompose $L_{ij} = q(\vartheta_i)\phi(\vartheta_i)^\top \phi(\vartheta_j)q(\vartheta_j)$ where $\phi : \Theta \to \mathbb{R}^D$ is some "feature mapping" of the elements of Θ in D-dimensional Euclidean space with $D \ll |\Theta|$. The function $q : \Theta \to \mathbb{R}^+$ is viewed as a "quality" function. With this decomposition, we see that the sub-determinant $|L_S|$ is proportional to the volume spanned by the vectors $\{q(\vartheta_i)\phi(\vartheta_i) : i \in S\}$. In this way, the L-ensemble is a probability distribution that places most of its mass on sets S that are diverse and of high quality.

Motivated by L-ensembles, we may take $K(\Gamma)$ to have determinantal form. In particular, we may form a similarity matrix $\mathcal{K}$ whose rows and columns are indexed by partitions of $[n]$ and with entries $k(\boldsymbol{\gamma}, \boldsymbol{\gamma}') = \exp\{-VI(\boldsymbol{\gamma}, \boldsymbol{\gamma}')\}$. Recall from earlier that we denote the set of unique particles in a particle set Γ as $\Gamma^\star$. We may then define

$$K(\Gamma) = \begin{cases} M_n & \text{if } |\Gamma^\star| = 1 \\ \prod_{\gamma^\star_{\ell^\star} \in \Gamma^\star} q(\gamma^\star_{\ell^\star})^2 \times \log|\mathcal{K}_{\Gamma^\star}| & \text{otherwise} \end{cases},$$

where M_n is some large, negative number that depends on n and q is some measure of the "quality" of the particles. By taking M_n large and negative, we strongly discourage letting the particle set collapse onto a single partition. It remains to specify the quality function and an intuitive choice would be to have it depend on the posterior probability of each particle.

While a DPP-inspired $K(\Gamma)$ is very attractive, the choice of quality function is highly non-trivial. In our context, we would ideally like to have it depend on the posterior probability of each particle. As a substantially simpler alternative, we may take $K(\Gamma)$ to be the total pairwise VI distance between the particles. In this way, we explicitly favor particle sets which spread the particles far apart throughout the space.

Note that when $\xi = 0$, the optimization problem reduces to that in Equation (6.1), whose objective function may be suggestively re-written as:

$$\sum_{\boldsymbol{\gamma}_\ell} w_\ell \log \left(\pi(\boldsymbol{\gamma}_\ell, \mathbf{Y})^{\frac{1}{\lambda}} \right) + H(\Gamma, \mathbf{w})$$

Written like this, we immediately recognize the optimal solution as the variational approximation of the *tempered* posterior $\pi(\boldsymbol{\gamma}|\mathbf{Y})^{\frac{1}{\lambda}}$. In Equation (6.2), we have three terms: the importance weighted-averaged height of the log-posterior, the entropy of the particle system, and a measure of the diversity of the particle system with respect to the variation-of-information metric. Each of these terms push the particle set towards different targets and the tuning parameters λ and ξ govern how these forces interact.

Returning to our earlier goal of exploring the space while still capturing the most mass efficiently, we can consider solving Equation (6.2) along a grid of λ and ξ values, much like we did in the Dynamic Posterior Exploration of Chapter 2. In particular, we may initially start by solving the problem with only a small number of particles (say 5 or 10) and with λ and ξ quite large. Then we may gradually reduce the values of λ to 1 and ξ to 0 and solve the new optimization problem. At each step along the way, we may duplicated each

particle so that the size of the particle set expands as our problem becomes closer and closer to the original variational approximation problem in Equation (3.2). Intuitively, the large values of λ and ξ flatten the posterior and promote distribution the particles widely across the space, respectively. Though some of particles may be pushed to regions of substantially less posterior probability, the relative differences in *tempered* posterior probability are much smaller. In this way, the large λ and ξ values work in concert to distribute the particles widely across the space. In a sense, we initially begin by spreading about the space and allow for sojourns into regions of comparatively less posterior probability. Then as we lower the value of λ and ξ, we may to lean on the entropy penalty to stay in the vicinity of potentially strong modes. By doubling the particle set at each stage, we add the ability to explore locally around these emergent modes. The hope is that by the end of the program, we will be solving our original optimization problem (corresponding to $(\lambda, \xi) = (1, 0)$) by initializing our particle set around several potentially promising modes. When it comes time to approximating posterior expectations, we may use the final importance weights to decide over which conditional posterior expectations to average.

BIBLIOGRAPHY

Abegaz, F. and Wit, E. (2013). Sparse time series chain graph models for reconstructing genetic networks. *Biostatistics*, 14:586–599.

Affandi, R. H., Fox, E. B., Adams, R. P., and Taskar, B. (2014). Learning the parameters of determinantal point process kernels. In *Proceedings of the* 31^{st} *International Conference on Machine Learning*.

Bachynski, K. (2016). Tolerable risks? Physicians and youth football. *New England Journal of Medicine*, 374(5):405–407.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Banerjee, S. and Ghosal, S. (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:147–162.

Berger, J. O. and Robert, C. P. (1990). Subjective hierarchical Bayesian estimation of a multivariate normal mean: On the frequentist interface. *Annals of Statistics*, 18(2):617–651.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Bhadra, A. and Mallick, B. K. (2013). Joint high-dimensional Bayesian variabile and covariance selection with an application to eqtl analysis. *Biometrics*, 69:447–457.

Binder, D. (1978). Bayesian cluster analysis. *Biometrika*, 65(1):31–38.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. arXiv:1601.00670.

Borodin, A. and Rains, E. M. (2005). Eynard-Mehta theorem, Schur process, and their Pfaffian analogs. *Journal of Statistical Physics*, 121(3–4):291–317.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society Series B*, 59(1):3 – 54.

Broglio, S. P., Eckner, J. T., Martini, D., Sosnoff, J. J., Kutcher, J. S., and Randolph, C. (2011). Cumulative head impact burden in high school football. *Journal of Neurotrauma*, 28:2069–2078.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society Series B*, 60:627 – 641.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156.

Carvalho, C. M., Massam, H., and West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika*, 94(3):647–659.

Carvalho, C. M. and Scott, J. G. (2009). Objective Bayes model selection in Gaussian graphical models. *Biometrika*, 96(3):497–512.

Casella, G., Moreno, E., and Girón, F. J. (2014). Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis*, 9(3):613–658.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya*, 35(4):417–446.

Crowley, E. M. (1997). Product partition models for normal means. *Journal of the American Statistical Association*, 92(437):192–198.

Dahl, D. B. (2006). Model-based clustering for expression data via a dirichlet process mixture. In Do, K.-A., Müller, P., and Vanucci, M., editors, *Bayesian Inference for Gene Expression and Proteomics*, chapter 10. Cambridge University Press.

Dahl, D. B. (2009). Modal clustering in a class of product partition models. *Bayesian Analysis2009*, 4(2):243–264.

Dawid, A. and Lauritzen, S. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21(3):1272 – 1317.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Deshpande, S. K., Hasegawa, R. B., Rabinowitz, A. R., Whyte, J., Roan, C. L., Tabatabatei, A., Baiocchi, M., Karlawish, J. H., Master, C. L., and Small, D. S. (2017). Association of playing high school football with cognition and mental health later in life. *JAMA Neurology*.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319.

Engelhardt, B. E. and Adams, R. P. (2014). Bayesian structured sparsity from gaussian fields. arXiv:1407.2235.

Fan, J. and Li, R. (2001). Variable selection via noncave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J. H., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2008). *Bayesian Data Analysis*. CRC Press, 3 edition.

George, E. I. (1986a). Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, 81:437–445.

George, E. I. (1986b). A formal Bayes multiple shrinkage estimator. *Communications in Statistics A - Theory and Methods*, 15(7):2099–2114.

George, E. I. (1986c). Minimax multiple shrinkage estimation. *Annals of Statistics*, 14:188–205.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gillenwater, J., Kulesza, A., and Taskar, B. (2012). Discovering diverse and salient threads in document collections. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*.

Giudici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801.

Guskiewicz, K. M., Marshall, S. W., Bailes, J., McCrea, M., Cantu, R. C., Randolph, C., and Jordan, B. D. (2005). Association between recurrent concussion and late-life cognitive impairment in retired professional football players. *Neurosurgery*, 57(4):719–724.

Guskiewicz, K. M., Marshall, S. W., Bailes, J., McCrea, M., Harding Jr, H. P., Matthews, A., Mihalik, J. R., and Cantu, R. C. (2007). Recurrent concussion and risk of depression in retired professional football players. *Medicine and Science in Sports and Exercise*, 39(6):903–910.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618.

Hart Jr, J., Kraut, M. A., Womack, K. B., Strain, J., Didehbani, N., Bartz, E., Conover, H., Mansinghani, S., Lu, H., and Cullum, C. M. (2013). Neuroimaging of cognitive dysfunction and depression in aging retired national football leauge players. *JAMA Neurology*, 70(326-335).

Hartigan, J. (1990). Partition models. *Communications in Statistics A - Theory and Methods*, 19:2745–2756.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes. *Journal of the American Statistical Association*, 101(473):18–29.

Heller, K. A. and Ghahramani, Z. (2005). Bayesian hierarchical clustering. In *Proceedings of the* 22^{nd} *International Conference on Machine Learning Research*.

Holmes, C., Denison, D., and Mallick, B. K. (1999). Bayesian partitioning for classification and regression. Technical report, Imperial College, London.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. (2014). Quic: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15:2911 – 2947.

Jain, S. and Neal, R. M. (2004). A split-merge Markov Chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematics and Statistics*, pages 361–379.

Jones, B., Carvalho, C. M., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400.

Kulesza, A. and Taskar, B. (2011). Structured determinantal point processes. In *Advances in Neural Information Processing Systems 23*.

Kulesza, A. and Taskar, B. (2012). Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3).

Kulis, B. and Jordan, M. I. (2012). Revisiting k-means: New algorithms via Bayesian nonparametrics. In *Proceeings of the* 29^{th} *International Conference on Machine Learning*.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.

Lee, D. and Mitchell, R. (2012). Boundary detection in disease mapping studies. *Biostatistics*, 13(3):415–426.

Lee, W. and Liu, Y. (2012). Simultaneous multiple response regression and inverse covariate matrix estimation via penalized gaussian maximum likelihood. *Journal of Multivariate Analysis*, 111(241-255).

Lehman, E. J., Hein, M. J., Baron, S. L., and Gersic, C. M. (2012). Neurodegenerative causes of death among retired National Football Leauge players. *Neurology*, 79(19):1970 – 1974.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410.

Lindley, D. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society Series B*, 34(1):1–41.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206.

Meilă, M. (2007). Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, 98:873–895.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278.

Miles, S. H. and Prasad, S. (2016). Medical ethics and school football. *The American Journal of Bioethics*, 16(1):6–10.

Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). Support union recovery in high-dimensional multivarate regression. *Annals of Statistics*, 39(1):1–47.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Regularized multivarate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics*, 4(1):53–77.

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenc, M., Hübner, N., Aitman, T. J., Cook, S. A., and Richardson, S. (2010). New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Computational Biology*, 6(4).

Pfister, T., Pfister, K., Hagel, B., Ghali, W. A., and Ronksley, P. E. (2016). The incidence of concussion in youth sports: A systematic review and meta-analysis. *British Journal of Sports Medicine*, 50(5):292–297.

Richardson, S., Bottolo, L., and Rosenthal, J. S. (2010). Bayesian models for sparse regression analysis of high dimensional data. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A. F. M., and WEst, M., editors, *Bayesian Statistics 9*.

Ročková, V. (2017). Particle EM for variable selection. *Journal of the American Statistical Association*.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2016). The spike-and-slab LASSO. *Journal of the American Statistical Association*.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandanavian Journal of Statistics*, 29:391–411.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1):185–203.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366):318–328.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5):2587–2619.

Scott, J. G. and Carvalho, C. M. (2008). Feature-inclusion stochastic search for gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):790–808.

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Evan-Shoshan, O., Chen, W., Zhang, X., and Mosher, R. E. (2001). Multivariate matching and bias reduction in the surgical outcomes study. *Medical Care*, 39(10):1048 – 1064.

Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statistics*, 9(6):1135–1151.

Stigler, S. M. (1990). The 1988 Neyman Memorial Lecture: A Galtonian perspective on shrinkage estimators. *Statistical Science*, 5(1):147–155.

Taddy, M., Lopes, H. F., and Gardner, M. (2016). Scalable semiparametric inference for the means of heavy-tailed distributions. arXiv:1602.08066.

Tian, X. and Taylor, J. (2016). Selective inference with a randomized response. arXiV:1507.06739.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, pages 234–240.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, 47(3):349 – 363.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wade, S. and Ghahramani, Z. (2017). Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*.

Wang, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10:351–377.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Annals of Applied Statistics*, 5(4):2630–2650.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368.

Zhang, C.-H. (2010). Nearly unbaised variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1427.

Zou, J. Y. and Adams, R. P. (2012). Priors for diversity in generative latent variable modeling. In *Advances in Neural Information Processing Systems 25*.