BAYESIAN NONPARAMETRIC METHODS FOR HIGH-DIMENSIONAL DATA

David C. Kessler

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2013

Approved by:

Dr. David B. Dunson
Dr. Amy H. Herring
Dr. Stephanie M. Engel
Dr. Hongtu Zhu
Dr. Fei Zou

Abstract

DAVID C. KESSLER: Bayesian Nonparametric Methods for High-Dimensional Data
(Under the direction of Dr. David B. Dunson and Dr. Amy H. Herring)

Bayesian nonparametric (BNP or NP Bayes) methods have enjoyed great strides forward in recent years. BNP methods embody the belief that inference is best driven by the data itself with minimal assumptions about the underlying model; this approach has motivated a wide variety of BNP techniques that have met with with much success.

In the first dissertation paper, we address a long-standing complaint about the nonparametric priors used in BNP analyses, that they do not necessarily reflect the analyst's prior belief or intention, and so are not really Bayesian. In fact, it can be demonstrated that a supposedly uninformative nonparametric prior framework is actually very informative about certain aspects of the distribution it models. We develop a novel method to incorporate prior information about functionals of the unknown distribution, replacing undesirable induced priors on those functionals with prior distributions that reflect real prior belief. We show that the new prior enjoys the support characteristics of the original prior, and we demonstrate with examples the effect of the marginal prior on the quality of inference.

In the second and third dissertation papers, we address challenges in the analysis of high-dimensional data, with a focus on density regression. Many areas of inquiry, particularly in genetics research, are concerned with the modeling of a continuous physical trait as some function of a very large set of predictors. In most cases the number of predictors p is much larger than the number of observations n. In addition, the response to be modeled may have a nontrivial conditional distribution. In the second dissertation paper we develop a solution for this problem in the context of uncorrelated observations, and apply the technique to a problem in molecular epidemiology. In the third dissertation paper we expand the method to address correlated observations. We illustrate the utility of the proposed method in an application to a family-based data from a whole-genome linkage analysis of a neurological condition.

To my mother, Ann Munro Kennedy. I wish you could be here to see this.

To my daughter Ella, who put up with a lot of fatherly absences and now thinks that "data work" is the thing to do. Little Bear, you deserve a lot of credit for helping me learn patience and perseverance.

Most of all, to my lovely wife Kate, without whose support and encouragement I would not have been able to make it through this. Here's to the future, my dear.

Acknowledgments

Many thanks go to my advisor, Dr. David Dunson, whose training as an endurance athlete must have been a valuable asset during these years. He has worked with me through many difficult parts of this odyssey and I have gotten great benefit from his creativity, his continued support, and most of all his faith that I could finish.

Thanks to my committee: Dr. Stephanie Engel, Dr. Amy Herring, Dr. Hongtu Zhu, and Dr. Fei Zou. I have greatly appreciated their participation in this process, their helpful suggestions, and their availability to discuss details of their work.

Hearty thanks to the good friends I made at UNC, particularly my brothers-in-arms, Ryan May and Vonn Walter, and my original academic advisor, Dr. Larry Kupper.

The people in the UNC Department of Biostatistics have always worked hard to make this easier: my gratitude to Melissa Hobgood, Veronica Stallings, Vera Bennett, Evie McKee, Debbie Quach, David Hill, and Scott Zentz.

Much of this work was done with the support of Training Grant T32ES007018 from the National Institute of Environmental Health Sciences, National Institutes of Health.

SAS has given me employment, flexibility, and encouragement. I would particularly like to thank Maura Stokes, Randy Tobias, and Bob Rodriguez.

My collaborators, Dr. Peter Hoff of the University of Washington, Dr. Jack Taylor of NIEHS, and Dr. Allison Ashley-Koch of Duke University were crucial to my progress. Peter Hoff in particular gave me hope that things would work out. Dr. Eric Lock and Heidi Cope at Duke University were instrumental in getting me access to data. I would also like to thank the individuals and families who participated in the studies that produced the data.

My in-laws, Ed and Gale Unterberg, have been unwavering in their support.

My family: John Kessler, Mary Clare Kessler, and a long list of Unterbergs, Curtises, Moburgs, Munros, DeZwartes, Cookes, Larimers, and many others encouraged me to take this leap and learned to stop asking when I would be done.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Literature Review and Motivation

1.1.1 Marginally Specified Priors for Nonparametric Bayes Analysis

In the first paper, we demonstrate a technique for introducing marginal prior information into nonparametric Bayes (NP Bayes) analyses. Many real-world data analysis situations are not well-suited to a description that is governed by a finite-dimensional parameter; this has led to the development of a rich class of NP Bayes methods. These approaches aim to obtain inference under a prior that has support on the entire space of relevant probability distributions (Ferguson 1973). These methods have been applied to a variety of problems, including quantitative trait loci mapping (Zou et al. 2010), density estimation (Muller et al. 1996), regression and classification (Neal 1999), image segmentation (Sudderth and Jordan 2008), speaker diarization (Fox et al. 2011), and functional data analysis (Petrone et al. 2009). This diversity in application reflects the utility of NP Bayes methods in modern statistical practice.

NP Bayes techniques require some introduction, since the “nonparametric” label is somewhat misleading. Much contemporary NP Bayes research traces its origin to Ferguson’s 1973 paper, which presented the idea of the Dirichlet process (DP) prior. This paper was titled "A Bayesian Analysis of Some Nonparametric Problems;" in the NP Bayes setting, the problem itself is nonparametric, not the analysis. Nonparametric analysis addresses problems like density estimation and flexible regression models; in both cases the "parameter" to be estimated has infinite or very high dimension. The attraction of NP Bayes analysis and nonparametric analysis in general is that structural assumptions about the quantity to be modeled or estimated are significantly more flexible.

The careful development of NP Bayes techniques has made them successful in the areas already mentioned. Nevertheless, the complexity of the high- or infinite-dimensional parameter of interest can make the "Bayes" portion of NP Bayes more challenging. When the parameter is a complete distribution, a prior on that parameter that is chosen for large support and convenience in computation can induce priors on functionals of that distribution that are not consistent with actual information.

We are motivated by situations in which there is reliable prior information about marginal aspects of an unknown distribution. For example, consider a demographic survey which collects a small, detailed sample of the population. Assume that an earlier, large-scale census surveyed a much larger sample of the population, but recorded observations for many fewer population characteristics. The two surveys, small and large, are not identical and so are not driven by exactly the same underlying distribution, but they overlap on a few measured quantities. The Bayesian approach is to use the available prior information from the earlier, larger survey on those overlapping quantities to inform inference in the newer survey. If we are using a nonparametric prior for the distribution of interest in the smaller survey, we may not be able to directly manipulate that nonparametric prior to accommodate this prior information without inducing undesirable behavior into other aspects of our nonparametric prior. We may be able to come up with an *ad hoc* prior distribution for our situation, but we cannot guarantee that this bespoke prior will have the same support as generally applied nonparametric priors.

Density estimation provides a good example of the utility of NP Bayes approaches and further illustrates the challenges with the inclusion of prior information. In the nonparametric Bayes treatment of density estimation, the practicioner need not be restricted by a specific parametric form such as a multivariate normal. The early work of Ferguson (1973; 1974), Blackwell and MacQueen (1973) and Antoniak (1974) established theoretical properties for the Dirichlet process (DP) prior, a prior over probability measures. For a given sample space $\mathcal{Y}$, a DP prior over distributions on $\mathcal{Y}$ is parameterized in terms of a "base measure" Q_0 on $\mathcal{Y}$ and a "concentration parameter" α. One limiting aspect of the DP prior is that it produces random measures that are almost surely discrete, making it less suitable for modeling continuous outcomes directly. To address this, Antoniak (1974), Lo (1984), and Ferguson (1983) presented the idea of a Dirichlet process mixture model (DPMM), where the DP prior serves as the prior for a mixing distribution; this provides a more appropriate method for modeling the distribution of continuous quantities. In that setting, the data are assumed to come from a population with density $p(y|Q) = \int p(y|\psi)Q(d\psi)$, where $\{p(y|\psi) : \psi \in \Psi\}$ is a simple parametric family. As Q is discrete with probability 1, the resulting model for the population distribution is a countably infinite mixture model, where the parameters in the component measures are determined by Q_0, and the number of components with non-negligible weights is increasing in α.

Sampling from the posterior distribution under such a prior is problematic due to its complexity; MCMC techniques were developed in Escobar (1994) and Escobar and West (1995). These methods performed well for conjugate DPMs; Kleinman and Ibrahim (1998a;b) demonstrated the use of DPMs to model the distribution of random effects in the generalized linear mixed model.

In cases where the base measure Q_0 is not conjugate to the component likelihood, the integrations required in the method of Escobar and West can significantly expand computation time. Ishwaran and James (2001) developed techniques for an alternative treatment of the DPM, based upon the "stick-breaking" representation for the DP that was developed by Sethuraman (1994). The Ishwaran and James approach avoids integrating over the random mixing measure unlike in the Escobar and West method. This makes non-conjugate base measures in DPM models more feasible than in the Pólya-urn style of sampler developed in the earlier works. Sampling methods for DPM models evolved with the introduction of the retrospective sampler (Papaspiliopoulos and Roberts 2008), the slice sampler (Walker 2007; Kalli et al. 2011) and the exact block Gibbs sampler (Yau et al. 2011). These continuing improvements in sampling techniques have prompted the application of the DPM model to an expanding variety of problems.

In the case of the DPMM, the choice of α and Q_0 will have a significant effect on the prior for the population density, and potentially on posterior inference. Many applications include priors for the base measure (Escobar and West 1995; Muller et al. 1996) and incorporate estimation of Q_0 and α into the posterior inference. Other approaches have addressed the challenge of specifying Q_0 by applying empirical Bayes techniques to develop a point estimate for Q_0 (McAuliffe et al. 2006). Although it is common to give the base measure an over-dispersed form in an attempt to avoid an unduly informative prior, such an approach is actually highly informative in favoring allocation to a single cluster unless α is appropriately adapted (Bush et al. 2010). The particular case of the DP prior illustrates the general challenge of incorporating prior information in a nonparametric setting. The results of Yamato (1984) and Lijoi and Regazzini (2004) can be extended to adjust α and Q_0 in normal DPMMs so that the induced prior expectation and variance of the population mean can be approximately specified although specification beyond the population mean is problematic.

Moala and O'Hagan (2010) proposed a method to update a Gaussian process (GP) prior with expert assessments of the mean and other aspects of an unknown density. As with the Dirichlet process prior, the GP prior requires specification of the mean and covariance functions that characterize the GP. These provide a base for the prior in the same way that the Q_0 base measure does for the Dirichlet process prior. In the Moala and O'Hagan approach, elicitation of these quantities is derived from expert assessments of quantiles of the unknown distributions.

Many NP Bayes methods for finite sample spaces are built on the Dirichlet distribution (DD); in this case the unknown parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{|\mathcal{Y}|-1})$ can be interpreted as the parameter of a multinomial distribution. The DD prior is nonparametric in the sense that it has support on the entire $(|\mathcal{Y}| - 1)$-dimensional simplex. Such a prior on a distribution $\boldsymbol{\pi}$ depends on concentration parameters $\alpha_j, j = 1, \ldots, |\mathcal{Y}|$. Large values for these parameters result in a prior concentrated near the center of the simplex, while small values concentrate the prior at the vertices of the simplex.

This highlights the general challenge with Bayesian methods and the thornier challenge with nonparametric Bayes methods, the elicitation of a prior. In the usual parametric Bayes analysis, we are concerned with specifying priors for parameters that have some readily interpretable effect on the overall model. This can be daunting in complex parametric models; for example, log-linear models for high-dimensional contingency tables have a large number of parameters that require prior specifications. In the nonparametric Bayes case, the parameter of interest has either high or infinite dimension, and elicitation of a meaningful prior is an even more difficult task.

An increase in dimensionality can quickly become a challenge in the analysis of multivariate data, whether those data are discrete or continuous. Simply considering the pairwise interactions for $p-$dimensional continuous data gives some sense of the scale, since the number of such interactions increases as the square of the dimension of the observations. In the case of categorical data the problem can be even more daunting, particularly if we wish to include all possible interactions. For example, genetics data is sometimes presented as multivariate unordered categorical data, with each element in an observation indicating one of four nucleotides. A sequence of 10 nucleotides then needs $4^{10} \approx 10^6$ parameters to describe a saturated model, while a sequence of only twice as many nucleotides needs $4^{20} \approx 10^{12}$ parameters. This is clearly intractable for any parametric treatment of this problem that attempts to address the complete dependency structure.

Frequently, simplifying assumptions are made about the extent of meaningful correlations. For example, the copula Gaussian graphical model presented in Dobra and Lenkoski (2011) eliminates some of this complexity through the use of the copula and the accompanying transformation of categorical variables, but the graphical model introduces additional simplification of the covariance structure in the transformed space so that elements of the Gaussian covariance are constrained to be zero. More familiar approaches for multivariate unordered categorical data, such as the log-linear model in either frequentist or Bayesian analysis, might also discard higher-order interactions pre-emptively. In the case of maximum likelihood analysis, a complex model applied to a moderately-sized data set will certainly result in empty cells in such a high-dimensional multiway contingency table; this will mean that asymptotic assumptions may not hold (Fienberg and Rinaldo 2007). In Bayesian analysis, the sheer number of models possible under this scenario may make it extremely difficult for any one model to appear better than another. Prior specification may also be problematic, though there exist more sophisticated methods for prior choice in this setting (Massam et al. 2009).

An NP Bayes approach to this scenario was proposed by Dunson and Xing (2009). Instead of modeling the complex interactions between all variables directly, they allow this interaction structure to be unknown, and model the joint distribution of all variables as a DP mixture of product multinomials. In effect, this treats the observations as if they come from a mixture of subpopulations. Within each subpopulation the variables are conditionally independent. Marginalizing over the mixing measure for these subpopulations induces dependence between the several variables, without making strong assumptions about the nature of that dependence. This substantially reduces the number of parameters within the model; in the 20 nucleotide example, the number of parameters is linear in the number of mixture components. This approach results in a sparse structure for the data, but Bhattacharya and Dunson (2012) pointed out that increasing the dimension of the problem could result in a proliferation of components, since the entire observation vector is used to weight the cluster allocations. As an alternative, Bhattacharya and Dunson proposed the simplex factor model, a more general solution that potentially avoids this difficulty.

These underline the difficulty with prior elicitation in nonparametric approaches. While it is possible to introduce simplifying assumptions and more nuanced models, the challenge of making an appropriate prior specification remains. Our goal in the first dissertation paper is to retain the large support provided by popular nonparametric priors but to allow the introduction of prior information where such prior information is available, replacing induced priors with the desired prior.

1.1.2 Density Regression With Many Interacting Predictors

In the second and third dissertation papers, we are motivated by efforts to link quantitative traits with genetic and other factors. In many cases, the quantitative traits have nontrivial distributions, even when conditioning on many predictors, and it is unappealing to assume a smooth parametric form for these conditional densities. Furthermore, such data commonly has a large p, or number of observed predictors per observation, relative to n, the total number of observations. Because many phenotypes are associated with more than one locus, the interaction between multiple loci can be just as important as their separate action. In these settings of "large p, small n", the additional complexity of interactions makes standard methods intractable, and we wish to develop methods that can address predictor selection at the same time as we consider conditional densities of nontrivial shape. Finally, we wish to expand these models to accommodate correlated responses, as are found in family-based studies of complex phenotypes.

A common scenario involves measurements of many single nucleotide polymorphisms (SNPs) and a continuous, or quantitative, phenotype. In many treatments of this problem, the strategy is to assess each locus or SNP independently with appropriate controls for overall false discovery rate (FDR). If it could always be believed that traits of interest are governed by single factors, then this would be an acceptable approach. However, it seems unlikely that these traits are so simply explained, and so consideration of multi-factor associations is desirable.

As discussed in Hoggart et al. (2008), simultaneous consideration of multiple SNPs can have many advantages. Due to the colossal number of possible multi-factor interactions, this consideration of many factors simultaneously usually assumes a model where the factors enter additively, each contributing some part to the quantitative trait. One common approach to the problem is the Lasso (Tibshirani 1996), which imposes an L_1 penalty and limits the overall sum of coefficients in the additive model, leading to a sparse set of contributing factors. A corresponding method is ridge regression (Hoerl and Kennard 1970), which uses an L_2 penalty, allowing smaller individual contributions to the response from a larger number of predictors. There are many other methods with the same general approach but using different penalty structures; for example, the elastic net (Zou and Hastie 2005) combines L_1 and L_2 penalties to gain advantages of both the Lasso and ridge regression.

While each of these methods has advantages, they all assume strictly additive behavior for the separate factors, with interactions not considered. Practicioners in genetics are well aware of the possibility for interaction between SNPs, and there are many methods in the literature to address the more biologically plausible scenario of multi-factor interaction. Lou et al. (2007) developed a generalized multi factor dimensionality reduction combinatorial algorithm. Chen et al. (2007) proposed a forest-based approach to identify gene-gene interactions. Zou et al. (2010) used a Gaussian process prior for the regression function that incorporated variable selection, enabling selection of a subset of interacting SNPs impacting the mean of the distribution of a quantitative trait. Yi (2010) surveys statistical approaches for identifying genetic interactions in high-dimensional settings, including genome-wide association studies (GWAS). Cordell (2009) reviews methods for selecting interactions between genetic loci contributing to human disease.

None of these approaches directly addresses density regression; we wish to acknowledge the possibility that the conditional density of quantitative traits given genetic factors may vary in more than just mean location as a function of those factors. This can be of considerable interest if the conditional mean does not change appreciably but the variance or skewness do, so that certain combinations of factors influence the probability of more extreme forms of the trait.

This assessment of behavior outside of mean shifts has prompted the development of quantile regression (Koenker and Bassett 1978) methods. These have been used in diverse areas of genetics research, including the assessment of copy number variation (Eilers and de Menezes 2005) and the analysis of age-dependent gene expression (Ho et al. 2009). Quantile regression works with selected quantiles of the response of interest, but we are more concerned with a characterization of the entire response distribution, and so wish to develop techniques for density regression. One motivating example concerns the distribution of body mass index (BMI), a measure of health that has certain clinically important cut points. In particular, individuals with a BMI in excess of 30 are classified as obese and at higher risk for many debilitating health consequences. A quantile regression approach can only estimate whether a specific quantile is above or below the point of interest. Density regression lets us model the entire conditional distribution so that we can identify the quantile corrresponding to the obesity cutpoint or any other point of interest.

Density regression has enjoyed extensive treatment in the Bayesian literature. The hierarchical mixture of experts (HME) model (Jordan and Jacobs 1994) gives one of the most accessible forms of density regression, representing an arbitrary conditional density as a convex combination of continuous kernels. The expectation maximization (EM) approach (Dempster et al. 1977) provides an attractive computational framework for estimating the parameters in these models via maximum a posteriori estimates. Models like the HME typically assume a specific or maximum number of kernels for the representation. The NP Bayes literature has developed a number of methods based upon DP mixtures, where the number of components is one of the parameters to be estimated. DP mixtures have been adapted in many different ways to the density regression problem, bringing predictors into the mean functions for the components, the mixing weights, or both.

One approach to conditional density estimation is described in Muller et al. (1996). In this approach, one estimates of the joint density of the response and the predictors and then derives a conditional density. This approach has been developed in several other settings (Shahbaba and Neal 2009; Hannah et al. 2011; Dunson and Xing 2009) and works well in those applications. One drawback to these joint estimation methods is the need to estimate the entire joint distribution when we are interested only in a specific conditional distribution. The additional effort to estimate what becomes a high-dimensional nuisance parameter is not appealing when we have a particular question in mind from the beginning.

One challenge to all of these methods is the "curse of dimensionality", or the increasing difficulty in obtaining a parsimonious description of the conditional distribution in the presence of more and more predictors. (Chung and Dunson 2009) proposed methods of predictor selection in these models, and methods for dimensionality reduction have also been proposed (Tokdar et al. 2010; Reich et al. 2011), but these encounter difficulty with larger values of p and do not directly address correlated data.

Random forests (Breiman 2001) provide an approach to density estimation that also provides support for predictor selection. Nevertheless, the form of the random forest makes it difficult to assess the role of specific predictors and their influence on the response. Also, random forests and related ensemble methods do not explicitly address correlated observations.

This question of correlated data in the presence of many potentially informative predictors has motived considerable recent research due to the direct relevance for family-based studies of genetic influences on quantitative traits. Most of this has centered around the adaptation of the standard linear mixed model (LMM) to high-dimensional predictor sets. Listgarten et al. (2012) and the related Lippert et al. (2011) propose an LMM-based approach that uses SNPs to derive a realized relationship matrix between the individuals in the study. They leverage work by Hayes et al. (2009) demonstrating the advantages of this form of relationship matrix over that derived from typical pedigree analysis, and develop algorithms to address the large number of SNPs. In the same direction, Rakitsch et al. (2013) developed the "LMM-Lasso", an adaptation of the L_1 penalty method to situations involving correlated observations. These approaches address the important situation of correlated observations but do not consider the possibility that the conditional density may have a nontrivial form depending on particular combinations of SNPs.

Our goal in the second and third papers is to develop general techniques for conditional density regression in settings with many predictors. In addition to addressing predictor selection, we wish to produce flexible representations for the conditional density that are suitable for the investigation of behavior outside of simple mean shifts. We also derive approaches for correlated data, motivated by studies involving family-based data.

Chapter 2

Marginally Specified Priors for Nonparametric Bayesian Estimation

2.1 Introduction

Many real-world data analysis situations do not lend themselves well to simple statistical models indexed by a finite-dimensional parameter. This has led to the development of a rich class of nonparametric Bayesian (NP Bayes) methods, the general idea of which is to obtain inference under a prior that has support on the entire space of relevant probability distributions (Ferguson 1973). These methods have been applied to a variety of problems, such as density estimation (Muller et al. 1996), image segmentation (Sudderth and Jordan 2008), speaker diarization (Fox et al. 2011), regression and classification (Neal 1999), functional data analysis (Petrone et al. 2009) and quantitative trait loci mapping (Zou et al. 2010) to name only a few. This breadth of applications reflects the utility of NP Bayes methods in modern statistical data analysis.

Many NP Bayes methods are built upon either the Dirichlet distribution (DD) for finite sample spaces or the Dirichlet process (DP) (Ferguson 1973) for infinite sample spaces. For the latter case, the body of work on parameter estimation (Escobar 1994), density estimation and inference (Escobar and West 1995) and the steady improvement in sampling methods (Escobar 1994; Walker 2007; Yau et al. 2011; Kalli et al. 2011) have all made the DP prior an attractive choice for many applications. For a given sample space $\mathcal{Y}$, a DD or DP prior over distributions on $\mathcal{Y}$ is parameterized in terms of a "base measure" Q_0 on $\mathcal{Y}$ and a "concentration parameter" α. Although samples from the DP prior are discrete with probability one, this prior is nonparametric in the sense that it has weak support on the set of all distributions having the same support as Q_0. Analogously, the DD prior is nonparametric in the sense that it has support on the entire $(|\mathcal{Y}| - 1)$-dimensional simplex. For both the DD and DP, a large value of α corresponds to a prior concentrated near Q_0. For the DP, a small α results in distributions with probability mass concentrated on only a few points, drawn independently from Q_0. For the DD, a small α can result in mass being concentrated near the vertices of the simplex.

For many NP Bayes methods, the DP is used as a prior for a mixing distribution in a mixture model: The data are assumed to come from a population with density $p(y|Q) = \int p(y|\psi)Q(d\psi)$, where $\{p(y|\psi) : \psi \in \Psi\}$ is a simple parametric family. A DP prior on Q results in a Dirichlet process mixture model (DPMM) (Lo 1984; Escobar and West 1995; MacEachern and Müller 1998). As Q is discrete with probability 1, the resulting model for the population distribution is a countably infinite mixture model, where the parameters in the component measures are determined by Q_0, and the number of components with non-negligible weights is increasing in α.

Clearly, the choice of α and Q_0 will have a significant effect on the prior for the population density, and potentially on posterior inference. Many applications include priors for the base measure (Escobar and West 1995; Muller et al. 1996) and incorporate estimation of Q_0 and α into the posterior inference. Other approaches have addressed the challenge of specifying Q_0 by applying empirical Bayes techniques to develop a point estimate for Q_0 (McAuliffe et al. 2006). Although it is common to give the base measure an over-dispersed form in an attempt to avoid an unduly informative prior, such an approach is actually highly informative in favoring allocation to a single cluster unless α is appropriately adapted Bush et al. (2010). In many applications, the base measure is given an overdispersed form in an attempt to avoid an unduly informative prior. Of course, doing so precludes the incorporation of prior information into the inference. The particular case of the DP prior illustrates the general challenge of incorporating prior information in a nonparametric setting. The results of Yamato (1984) and Lijoi and Regazzini (2004) can be extended to adjust α and Q_0 in normal DPMMs so that the induced prior expectation and variance of the population mean can be approximately specified (as will be discussed further in Section 3), although specification beyond the population mean is problematic. Moala and O'Hagan (2010) proposed a method to update a Gaussian process (GP) prior with expert assessments of the mean and other aspects of an unknown density. As with the Dirichlet process prior, the GP prior requires specification of the mean and covariance functions that characterize the GP. These provide a base for the prior in the same way that the Q_0 base measure does for the Dirichlet process prior. In the Moala and O'Hagan approach, elicitation of these quantities is derived from expert assessments of quantiles of the unknown distributions.

In this paper, we propose a very general method that allows for the combination of an arbitrary prior on a finite set of functionals with a nonparametric prior on the remaining aspects of the high- or infinite-dimensional unknown parameter. In the next section we show how such a partially informative prior distribution can be constructed from the combination of any prior distribution on the functionals of interest with the conditional distribution of the parameter given the functionals under a canonical nonparametric prior. We show that the resulting marginally specified prior (MSP) inherits desirable features from the canonical prior: The MSP will generally share the support of the canonical prior, and posterior approximation under the MSP can typically be made via small modifications to any Markov chain Monte Carlo algorithm applicable under the canonical prior.

In Section 3 we illustrate the use of the marginally specified prior in the context of multivariate density estimation using normal DPMMs. In an example, we show that existing approaches to incorporate prior information on mean and covariance into DPMMs lead to poor density estimates relative to marginally specified priors unless the parametric base model is an accurate approximation.

In Section 4 we examine the important problem of NP Bayes analysis of large sparse contingency tables in the presence of prior information on the margins. In this context, we develop a marginally specified prior from a canonical NP Bayes approach. In an example, we illustrate how canonical NP Bayes methods designed to be informative on the margins result in poor performance in terms of margin-free functionals (such as dependence functions). In contrast, a marginally specified prior accommodates prior information about the population margins while being minimally informative about other aspects of the population, resulting in strong performance in terms of both marginal and margin-free aspects of the population. A discussion of the results and directions for future research follows in Section 5.

2.2 Marginally specified priors: Construction and computation

We consider the general problem of Bayesian inference for a parameter f belonging to a high- or infinite-dimensional space $\mathcal{F}$. For example, Section 3 considers multivariate density estimation over the space of all densities on $\mathbb{R}^p$ with respect to Lebesgue measure, and Section 4 considers the high-dimensional space of multiway contingency tables. In general, Bayesian inference for f is based on a posterior distribution $\pi(f \in A|y)$ derived from a sampling model $\{p(y|f) : f \in \mathcal{F}\}$ and a prior distribution π defined on a σ-algebra $\mathcal{A}$ of $\mathcal{F}$. In many high-dimensional problems there are only a few classes of priors for which posterior inference is tractable. Typically, practitioners choose a member π_0 of such a class based on support considerations and the feasibility of posterior approximation, rather than how well it accurately represents any information we have about specific features of f. In this section, we show how to construct a nonparametric prior π_1 that is informative about specific features of f, but has the same support as π_0 and is "close" to π_0 in terms of Kullback-Leibler divergence. We also show how MCMC approximation methods for π_0 can be modified to obtain posterior inference under π_1.

2.2.1 Construction of a marginally specified prior

Let $\theta = \theta(f)$ be a function of f, such as a population mean of $p(y|f)$, variance, marginal probability vectors or some finite set of functionals, and let Θ be the range of θ. Any prior distribution π_0 on $(\mathcal{F}, \mathcal{A})$ induces a prior distribution P_0 on $(\Theta, \mathcal{B})$ defined by

$$P_0(B) = \pi_0(\{f : \theta(f) \in B\}), \tag{2.1}$$

for each $B \in \mathcal{B}$. If π_0 is chosen for computational convenience, the induced prior P_0 need not show substantial agreement with available prior information P_1 for the functional $\theta(f)$. In some cases a prior π_0 selected from a computationally feasible class will make the induced prior P_0 similar to P_1: The results of Lijoi and Regazzini (2004) and Yamato (1984) provide some guidance for Dirichlet process priors if the functionals are means, but in general this will be difficult. Furthermore, depending on the structure of the nonparametric class, selecting π_0 in order to match P_0 to P_1 will result in π_0 being inappropriate for other aspects of f. We present an example in Section 2.3 to illustrate a case where making π_0 highly informative about $\theta(f)$ also makes it highly informative about other aspects of f.

Suppose a nonparametric prior π_0 has been identified that is viewed as reasonable in some respects, such as being computationally feasible and having a large support, but does not represent available prior information P_1 about θ. The information in P_1 can be accommodated by replacing P_0, the θ-margin of π_0, with the desired margin P_1. Specifically, a marginally specified prior (MSP) π_1 for f is obtained by combining the conditional distribution of f given θ with our desired marginal distribution P_1 for θ, so that

$$\pi_1(A) = \int \Lambda_0(A|\theta)P_1(d\theta)\ \forall A \in \mathcal{A}, \tag{2.2}$$

where $\Lambda_0(A|\theta)$ is the conditional probability of A given θ under π_0. A prior π_1 constructed this way should have the desired marginal distribution P_1 over$(\Theta, \mathcal{B})$, and if $P_1 \ll P_0$, should also have the same support as π_0, since the conditional probabilities under π_1 should match those under π_0.

Such a construction is straightforward if f is finite dimensional. Accommodation of nonparametric problems where f is potentially infinite dimensional requires some additional mathematical detail. We consider the case where $\mathcal{A}$ are the Borel sets of a Hausdorff space $\mathcal{F}$, and $\theta : \mathcal{F} \to \Theta$ is a measurable map with respect to a σ-algebra $\mathcal{B}$ on Θ. Let the prior π_0 be a regular probability measure on $(\mathcal{F}, \mathcal{A})$, and let P_0 be the induced prior distribution on $(\Theta, \mathcal{B})$, i.e., for all $B \in \mathcal{B}$, $P_0(B) = \pi_0(\{f : \theta(f) \in B\})$.

Example (Dirichlet process mixture model): Recall the Dirichlet process mixture model prior, defined by

$$p(y|Q) = \int p(y|\psi)Q(d\psi)$$
$$Q \sim \mathrm{DP}(\alpha, Q_0),$$

where $\{p(y|\psi), \psi \in \mathbb{R}^p\}$ is a collection of absolutely continuous probability densities over some Euclidean space $\mathcal{Y}$ and Q_0 is an absolutely continuous probability measure over $\mathbb{R}^p$. The random mixing measure Q has a representation as an infinite weighted sum of point mass measures, $Q \overset{d}{=} \sum w_k \delta_{\psi_k}$, where $\psi = \{\psi_1, \psi_2, \ldots\}$ are an infinite i.i.d. sample from Q_0, and $w_k = v_k \prod_{j<k}(1-v_j)$, with $v = \{v_1, v_2, \ldots\}$ are an infinite i.i.d. sample from a beta$(1, \alpha)$ distribution. Therefore the prior over Q can be represented as a prior over $f = (\psi, v) \in \mathbb{R}^\infty$. This space, with the usual product topology, is Hausdorff. Now let θ be a moment of $p(y|Q)$, so that

$$\begin{aligned}\theta(f) &= \int g(y)p(y|Q)dy \\ &= \sum_{k=1}^{\infty}[v_k \prod_{j<k}(1-v_j)] \int g(y)p(y|\psi_k)dy.\end{aligned}$$

The function θ is Borel measurable as long as $p(y|\psi)$ is measurable in ψ for each $y \in \mathcal{Y}$.

Returning to the marginally specified prior given by (2.2), note that $\pi_0(A|\theta)$ is not well defined on null sets of P_0. To make (2.2) meaningful, we restrict attention to informative prior distributions such that P_1 is dominated by P_0. Under this condition and the conditions on $(\mathcal{F}, \mathcal{A})$ and θ given above, the measure π_1 on $\mathcal{A}$ is well defined and the θ-marginal of π_1 is given by P_1.

Theorem 1. *Let $\Lambda_0(\cdot|\cdot) : \mathcal{A} \times \Theta \to [0,1]$ be a conditional probability function for π_0 given θ and let P_1 be a probability measure on $(\Theta, \mathcal{B})$ such that $P_1 \ll P_0$. Then $\pi_1 : \mathcal{A} \to [0,1]$, defined by*

$$\pi_1(A) = \int \Lambda_0(A|\theta)P_1(d\theta),$$

1. *is a probability measure over $\mathcal{A}$;*

2. *satisfies $\pi_1(\{f : \theta \in B\}) = P_1(B)$ for each $B \in \mathcal{B}$;*

3. *is dominated by π_0 with Radon-Nikodym derivative*

$$\frac{d\pi_1}{d\pi_0}(f) = \frac{dP_1}{dP_0}(\theta(f)).$$

For notational economy, we have used θ to represent both an element of Θ and as the function mapping $\mathcal{F}$ to Θ, depending on the context. A proof of the Theorem is provided in the appendix.

The MSP π_1 constructed above is dominated by π_0, but ideally we would like it to have the same support as π_0. Since π_1 and π_0 share conditional distributions, intuitively it seems that π_1 should have reduced support relative to π_0 only if P_1 has reduced support relative to P_0. This result can be shown with the aid of the Radon-Nikodym derivative given above, which implies that $\pi_1(A)$ can be computed as

$$\pi_1(A) = \int_A \frac{p_1(\theta(f))}{p_0(\theta(f))} \pi_0(df) = E_{\pi_0}[1(f \in A)\tfrac{p_1(\theta)}{p_0(\theta)}],$$

where p_1 and p_0 are densities of P_1 and P_0 with respect to some common dominating measure (which could be taken to be P_0, for example). Based on this identity, we have the following result:

Lemma 1. *Suppose $P_1 \ll P_0 \ll P_1$. Then $\pi_1 \ll \pi_0 \ll \pi_1$.*

Proof. It is clear from the definition of π_1 that $\pi_1 \ll \pi_0$. To show $\pi_0 \ll \pi_1$, let $A \in \mathcal{A}$ be a set such that $\pi_1(A) = 0$. We will show that $P_0 \ll P_1$ implies $\pi_0(A) = 0$. Let $B_j = \{\theta : p_j(\theta) > 0\}$ and $A_j = \{f : \theta(f) \in B_j\}$ so that $\pi_j(A_j) = P_j(B_j) = 1$ for $j \in \{0, 1\}$. We have

$$\begin{aligned} 0 = \pi_1(A) &= \pi_1(A \cap A_1) \\ &= E_{\pi_0}[1(A \cap A_1)\tfrac{p_1}{p_0}] \\ &= E_{\pi_0}[1(A \cap A_0 \cap A_1)\tfrac{p_1}{p_0}]. \end{aligned} \tag{2.3}$$

Since $p_1/p_0 > 0$ on $A_0 \cap A_1$, (2.3) implies that $\pi_0(A \cap A_0 \cap A_1) = 0$. Since $\pi_0(A_0) = 1$, we have $\pi_0(A \cap A_1) = \pi_0(A) - \pi_0(A \cap A_1^c) = 0$. Since $0 = \pi_1(A_1^c) = P_1(B_1^c)$ and $P_0 \ll P_1$, we must have $0 = P_0(B_1^c) = \pi_0(A_1^c)$, and so $\pi_0(A) = 0$. $\square$

We also note that π_1 has a characterization as the prior distribution that is closest to π_0 in terms of Kullback-Leibler divergence, among priors with θ-marginal density equal to p_1. This follows from re-expressing the probability measures π_1 and π_0 in terms of densities with respect to a common dominating product measure, so that

$$\pi_k(A \cap \theta^{-1}B) = \int_B \int_A \lambda_k(f|\theta)p_k(\theta)\ \mu(df) \times \nu(d\theta)$$

for $k \in \{0, 1\}$. The Kullback-Leibler divergence is then

$$D(\pi_1||\pi_0) = E_{\pi_1}[\ln \tfrac{\lambda_1(f|\theta)p_1(\theta)}{\lambda_0(f|\theta)p_0(\theta)}] = E_{\pi_1}[\ln \tfrac{\lambda_1(f|\theta)}{\lambda_0(f|\theta)}] + E_{\pi_1}[\ln \tfrac{p_1(\theta)}{p_0(\theta)}].$$

Fixing p_1, the divergence is is minimized by setting $\lambda_1(f|\theta) = \lambda_0(f|\theta)$ for θ a.e. P_1, i.e. matching the conditional distributions, giving $D(\pi_1||\pi_0) = D(P_1||P_0)$.

Lemma 2. *Let $P_1 \ll P_0$. Then among probability measures π_1 on $(\mathcal{F}, \mathcal{A})$ with θ-marginal equal to P_1, the Kullback-Leibler divergence of π_0 from π_1 is minimized when $\pi_1(A) = \int \Lambda_0(A|\theta)P_1(d\theta)$ for all $A \in \mathcal{A}$ and θ a.e. P_1.*

A more detailed derivation of this result is given in the appendix.

2.2.2 Posterior approximation under MSPs

Let $\{p(y|f) : f \in \mathcal{F}\}$ be a dominated statistical model, i.e. a family of probability densities with respect to a common measure. Given a prior distribution π, inference for $f \in \mathcal{F}$ proceeds via the conditional probability distribution $\pi(\cdot|y) : \mathcal{A} \to [0, 1]$, or alternatively the conditional density $\pi(f|y)$, given by

$$\pi(f|y) = \frac{p(y|f)\pi(f)}{p(y)} \equiv \frac{p(y|f)\pi(f)}{\int\ p(y|f')\pi(f')\mu(df')},$$

where $\pi(f)$ denotes the density of π with respect to a dominating measure μ. This represents the conditional measure in that $\int_A \pi(f|y)\mu(df)$ is a version of the conditional probability $\pi(A|y)$ for each $A \in \mathcal{A}$.

For practical reasons the most commonly used priors are those for which there exist straightforward Gibbs samplers or Metropolis-Hastings algorithms for posterior approximation. In many cases, simple modifications to these algorithms will allow for the incorporation of informative priors over functionals of interest. To illustrate, suppose that under prior π_0 we have a Gibbs sampler for a high dimensional parameter f. Recall that the Gibbs sampler can be viewed as a Metropolis-Hastings algorithm for which the proposals are accepted with probability one. From this perspective, a Gibbs sampler for approximating the posterior density $\pi_0(f|y)$ is constructed from proposal distributions with densities $J(f^*|f,y)$ that are proportional to the posterior density, so that

$$\frac{J(f^*|f,y)}{J(f|f^*,y)} = \frac{\pi_0(f^*|y)}{\pi_0(f|y)}. \tag{2.4}$$

For example, decomposing f as $\{f_1,\ldots,f_K\}$, the full conditional distribution $\pi_0(f_k|f_{-k},y)$ is one such proposal distribution.

Posterior approximation of $\pi_1(f|y)$ can proceed by using the proposal distributions of the Gibbs sampler for $\pi_0(f|y)$, but adjusting the acceptance probability. Specifically, the algorithm for approximating $\pi_1(f|y)$ proceeds by iteratively simulating proposals f^* from distributions of the form $J(f^*|f,y)$ which satisfy (2.4), and accepting each proposal f^* with probability $1 \wedge r_{\text{MH}}$, where

$$\begin{aligned}
r_{\text{MH}} &= \frac{\pi_1(f^*|y)}{\pi_1(f|y)} \times \frac{J(f|f^*,y)}{J(f^*|f,y)} \\
&= \frac{\pi_1(f^*|y)}{\pi_1(f|y)} \times \frac{\pi_0(f|y)}{\pi_0(f^*|y)} \\
&= \frac{p(y|f^*)\pi_1(f^*)}{p(y|f)\pi_1(f)} \times \frac{p(y|f)\pi_0(f)}{p(y|f^*)\pi_0(f^*)} = \frac{\pi_1(f^*)/\pi_0(f^*)}{\pi_1(f)/\pi_0(f)}.
\end{aligned}$$

Let the θ-marginal distribution of π_0 be P_0, and let π_1 be a marginally specified prior based on π_0 and a θ-marginal distribution $P_1 \ll P_0$. Let p_0 and p_1 be the densities of P_0 and P_1 with respect to a common dominating measure. By Theorem 1, $\pi_1(f)/\pi_0(f) = p_1(\theta)/p_0(\theta)$ and the acceptance ratio simplifies to

$$\frac{p_1(\theta^*)/p_0(\theta^*)}{p_1(\theta)/p_0(\theta)}.$$

Similarly, an approximation algorithm for $\pi_1(f|y)$ can be constructed from a Metropolis-Hastings algorithm for $\pi_0(f|y)$ via the same adjustment. Suppose we have a proposal distribution $J(f^*|f,y)$ such that the acceptance ratio r^0_{MH} for π_0 is computable:

$$r^0_{\text{MH}} = \frac{\pi_0(f^*|y)}{\pi_0(f|y)} \frac{J(f|f^*,y)}{J(f^*|f,y)}$$

The Metropolis-Hastings algorithm for approximating $\pi_1(f|y)$ using $J(f^*|f,y)$ has acceptance ratio

$$\begin{aligned}
r_{\text{MH}} &= \frac{\pi_1(f^*|y)}{\pi_1(f|y)} \frac{J(f|f^*,y)}{J(f^*|f,y)} \\
&= \frac{\pi_1(f^*|y)}{\pi_1(f|y)} \frac{\pi_0(f|y)}{\pi_0(f^*|y)} r^0_{\text{MH}} \\
&= \frac{p_1(\theta^*)/p_0(\theta^*)}{p_1(\theta)/p_0(\theta)} r^0_{\text{MH}}.
\end{aligned}$$

These results show that an MCMC approximation to $\pi_1(f|y)$ can be constructed from an MCMC algorithm for $\pi_0(f|y)$ as long as the ratio $p_1(\theta)/p_0(\theta)$ can be computed. The value of $p_1(\theta)$ for each $\theta \in \Theta$ is presumably available as p_1 is our desired prior distribution for θ. In contrast, obtaining a formula for $p_0(\theta)$ will be difficult in some settings. In situations where the dimension of θ is moderate, one simple solution is to obtain a Monte Carlo estimate of p_0 based on samples of f from π_0. Specifically, we can obtain an i.i.d. sample $\{\theta_i = \theta(f_i), i = 1, \ldots, S\}$ from $f_1, \ldots, f_S \sim$ i.i.d. π_0, and then approximate p_0 with a kernel density estimate or flexible parametric family. The method of approximation will depend on the nature of θ; the approaches just described are appropriate when $p_0(\theta)$ is absolutely continuous with respect to Lebesgue measure. Note that this can be done before the Markov chain is run, so that the same estimate of p_0 is used for each iteration of the algorithm.

In situations where obtaining a reliable estimate of p_0 is not feasible, it is still possible to induce a prior p_1 that is approximately equal to a target prior $\tilde{p}_1$, as long as p_0 is chosen to be flat compared to $\tilde{p}_1$. This can be done by replacing p_0, the θ-marginal density of π_0, with $p_1(\theta) \propto p_0(\theta)\tilde{p}_1(\theta) = Kp_0(\theta)\tilde{p}_1(\theta)$. This defines a valid probability density as long as $p_0\tilde{p}_1$ is integrable, which is the case, for example, if either density is bounded. In terms of the MCMC approximation to the resulting marginally specified prior π_1, the adjustment to the acceptance ratio is then

$$\frac{p_1(\theta^*)/p_0(\theta^*)}{p_1(\theta)/p_0(\theta)} = \frac{\tilde{p}_1(\theta^*)}{\tilde{p}_1(\theta)},$$

which is presumably computable as $\tilde{p}_1$ is the desired prior density. In this setting, $\tilde{p}_1$ contains the marginal prior information and p_1 takes on a form with computational convenience.

The proposed algorithm is closely related to importance sampling (IS) methods described in the literature. Besag et al. (1995) detail an IS-based approach for assessing prior sensitivity. In this development, an existing MCMC chain $\{\theta^{(t)}\}$ is weighted using the ratios $\tilde{h}(\theta^{(t)})/h(\theta^{(t)})$, where $h(\cdot)$ is the original prior used to produce the sample and $\tilde{h}(\cdot)$ is an alternative prior. The similiarity with our proposed method and its use of ratios of the marginally specified prior p_1 to the induced prior p_0 is clear; one important distinction is that our method replaces an induced prior on functionals with an elicited prior on those functionals, rather than substituting a prior in the main specification.

2.3 Density estimation with marginally adjusted DPMM

Perhaps the most commonly used NP Bayes procedure is the Dirichlet process mixture model, or DPMM (Lo 1984; Escobar and West 1995; MacEachern and Müller 1998). The DPMM consists of a mixture model along with a Dirichlet process prior for the mixing distribution. The population density to be estimated and the prior can be expressed as

$$\begin{aligned} p(y|Q) &= \int p(y|\psi)Q(d\psi) \\ Q &\sim \mathrm{DP}(\alpha Q_0), \end{aligned}$$

where α and Q_0 are hyperparameters of the Dirichlet process prior, with Q_0 typically chosen to be conjugate to the parametric family of mixture component densities, $\{p(y|\psi) : \psi \in \Psi\}$, to facilitate posterior calculations. In this section we show how to obtain posterior approximations under a marginally specified prior π_1 based on a DPMM. The approach is illustrated with the specific case of multivariate density estimation, for which we take the parametric family to be the class of multivariate normal densities. In an example analysis of the well-known bivariate dataset on eruption times of the Old Faithful geyser, we construct a prior distribution π_1 based on the multivariate normal DPMM with a marginally specified informative prior on the marginal means and variances. Here, we use a parametric approximation for the induced joint distribution p_0 of these specific functionals θ. Inference under π_1 is compared to inference under two standard DPMMs, one where the hyperparameters are chosen to be informative about θ and another where the hyperparameters are noninformative.

2.3.1 Posterior approximation

Given a sample $y_1, \ldots, y_n \sim$ i.i.d. $p(y|Q)$, posterior approximation for conjugate DPMMs is often made with a Gibbs sampler that iteratively simulates values of a function that associates data indices to the atoms of Q. In a DPMM, since Q is discrete with probability one, a given mixture component (atom of Q) may be associated with multiple observations. Let $g : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be the unknown mixture component membership function, so that $g_i = g_j$ means that y_i and y_j came from the same mixture component. Note that g can always be expressed as a function that maps $\{1, \ldots, n\}$ onto $\{1, \ldots, K\}$, where $K \leq n$. Inference for conjugate DPMMs often proceeds by iteratively sampling each g_i from its full conditional distribution $p(g_i|y_1, \ldots, y_n, g_{-i})$ (Bush and MacEachern 1996). Additional features of Q and $p(y|Q)$ can be simulated given $g_1, \ldots, g_n$ and the data.

This standard algorithm for DPMMs can be modified to accommodate a marginally specified prior distribution on a parameter $\theta = \theta(Q)$. Let $f = \{g, \theta\}$ and let π_0 be the prior density on f induced by the Dirichlet process on Q. Our marginally specified prior is given by $\pi_1(f) = \pi_0(f)p_1(\theta)/p_0(\theta)$, where p_0 is the density for θ induced by π_0 and p_1 is the informative prior density. An MCMC approximation to $\pi_1(f|y_1, \ldots, y_n)$ can be obtained via the procedure outlined in Section 2.2. Given a current state of the Markov chain $f = \{\theta, g_k, g_1, \ldots, g_{k-1}, g_{k+1}, \ldots, g_n\} = \{\theta, g_k, g_{-k}\}$, the next state is determined as follows:

1. Generate a proposal $f^* = \{\theta^*, g_k^*, g_{-k}\}$ from $\pi_0(\theta, g_k|g_{-k}, y) = \pi_0(g_k|g_{-k}, y)\pi_0(\theta|g, y)$ by

   (a) generating $g_k^* \sim \pi_0(g_k|g_{-k}, y)$;

(b) generating $\theta^* \sim \pi_0(\theta | g_k^*, g_{-k}, y)$.

2. Set the value of the next state of the chain to f^* with probability

   $1 \wedge [p_1(\theta^*)/p_0(\theta^*)]/[p_1(\theta)/p_0(\theta)]$,

   otherwise let the next state equal the current state.

This procedure is iterated over values of $k \in \{1, \ldots, n\}$, possibly in random order, and repeated until the desired number of simulations of f is obtained. Note that steps 1.(a) and 1.(b) compose a standard Gibbs sampler for the DPMM in which posterior inference for θ is provided, although typically we would only simulate θ once per complete update of $g_1, \ldots, g_n$. The algorithm for the marginally specified prior π_1 requires that θ be simulated with each proposed value of g_k so that the acceptance probability in step 2 can be calculated.

Implementing the steps of this MCMC algorithm involves two non-trivial computations: simulation of θ from $\pi_0(\theta|g, y)$, and calculation of $p_0(\theta)$ in order to obtain the acceptance probability. General methods for the latter were discussed in Section 2.2. For the former, we suggest using a Monte Carlo approximation to Q based upon a representation of Dirichlet processes due to Pitman (1996). Let K be the number of unique values of $g_1, \ldots, g_n$ and let n_k be the number of observations i for which $g_i = k$. If Q_0 is conjugate, then the parameter values $\psi_{(1)}, \ldots, \psi_{(K)}$ corresponding to the mixture components can generally be easily simulated. Corollary 20 of Pitman (1996) gives the conditional distribution of Q given $\psi_{(1)}, \ldots, \psi_{(K)}$ and counts $n_1, \ldots, n_K$ as

$$\{Q(H)|\psi_{(1)}, \ldots, \psi_{(K)}, n_1, \ldots, n_K\} \overset{d}{=} \gamma \sum_{k=1}^{K} 1(\psi_{(k)} \in H) w_k + (1-\gamma)\tilde{Q}(H),$$

where $\gamma \sim \text{Beta}(n, \alpha)$, $w \sim \text{Dirichlet}(n_1, \ldots, n_K)$ and $\tilde{Q} \sim \text{DP}(\alpha Q_0)$. A Monte Carlo approximation to Q, and therefore any functional of Q, can be obtained via simulation of a large number S of ψ-values from Q. To do this, we first simulate γ and $w_1, \ldots, w_K$ from their beta and Dirichlet full conditional distributions. From these values we sample cluster memberships for a sample of size S from Q using a multinomial$(S, \{\gamma w_1, \ldots, \gamma w_K, 1 - \gamma\})$ distribution. Note that the count s for the $K+1$st category represents the number of ψ-values that must be simulated from $\tilde{Q}$. To obtain the sample from $\tilde{Q}$ we run a Chinese restaurant process of length s, and then generate the unique ψ-values from Q_0 for each partition. This can generally be done quickly for two reasons: First, the expected number of samples needed from $\tilde{Q}$ is only $S\alpha/(\alpha + n)$. For example, with $S = 1000$, $n = 30$ and $\alpha = 1$, we expect to only need about $s = 32$ simulations from $\tilde{Q}$. Second, the number of unique values in a sample of size s from $\tilde{Q}$ is only of order $\log s$, which will generally be manageably small.

The marginal sampler we describe above has advantages in terms of efficiency and convergence rates (MacEachern 1994). However, because it does marginalize out the random measure Q, we must use the embedded Pitman method to draw samples from $\theta(Q)$ in order to evaluate the Metropolis-Hastings ratio. An alternative approach is to use a stick-breaking representation that does not integrate out the random measure. We can then use a slice sampler (Kalli et al. 2011) or exact block Gibbs sampler (Yau et al. 2011) and compute $\theta(Q)$ without needing an embedded sampling step, but at the possible expense of lower efficiency in the sampler.

2.3.2 Example: Old Faithful eruption times

The Old Faithful dataset consists of 272 bivariate observations of eruption times and waiting times between eruptions, both measured in minutes. To illustrate and evaluate the MSP methodology we construct two subsets of these data: a random sample of size $n_0 = 30$ from which we obtain prior information and a second, non-overlapping random sample of size $n = 30$ representing our observed data. The random samples were obtained by setting the random seed in R (version 2.14.0) to 1, sampling the prior dataset, and then sampling the observed dataset from the remaining observations. The observed sample had marginal means $(2.97, 64.2)$ and marginal variances $(1.29, 206.7)$. The prior sample had marginal means $(3.54, 71.9)$ and marginal variances $(1.24, 134.9)$. For the purpose of this example, we view the full dataset of 272 observations as the true population. A scatterplot of the observed data and marginal density estimates are shown in Figure A.1. The observed dataset consisting of $n = 30$ observations clearly captures the bimodality of the population. However, the marginal plots indicate that the sample has overrepresented one of the modes.

Suppose our knowledge of the prior sample is limited to the bivariate marginal sample means $m_0 \in \mathbb{R}^2$ and sample variances $v_0 \in (R^+)^2$. In such a situation it would be desirable to construct a prior density p_1 over the unknown population marginal means m and variances v based on the values of m_0, v_0 and n_0, and combine this information with the information in our fully observed sample to improve our inference about the population. Incorporating this information with conjugate priors would be straightforward if our sampling model were bivariate normal, but it is difficult in the context of a DPMM. Proposition 5 of Yamato (1984) indicates that if the base measure Q_0 in the Dirichlet process prior is multivariate normal(μ_0, Σ_0), then the induced prior distribution on the mean $\int xQ(dx)$ is approximately multivariate normal$(\mu_0, \Sigma_0/[\alpha+1])$. This result is not directly applicable to the multivariate normal DPMM for two reasons, one being that Q represents the mixing distribution and not the population distribution, and the other being that in the conjugate multivariate normal DPMM the parameter ψ in the mixture component consists not just of a mean μ but also a covariance matrix Σ. Specifically, in the conjugate p-variate normal DPMM, the density q_0 of the base measure Q_0 for $\psi = (\mu, \Sigma)$ is given by

$$q_0(\mu, \Sigma) = \text{normal}_p(\mu : \mu_0, \Sigma/\kappa_0) \times \text{inverse-Wishart}(\Sigma : S_0^{-1}, \nu_0) \tag{2.5}$$

where the functions on the right-hand side are the multivariate normal and inverse-Wishart densities respectively, the latter being parameterized so that $\mathrm{E}[\Sigma] = S_0/(\nu_0 - p - 1)$. Given a choice for α it is possible to obtain values of the hyperparameters $(\mu_0, \kappa_0, S_0, \nu_0)$ so that the induced prior distributions on the population mean $m(Q) = \int\int yp(y|\psi)Q(d\psi)dy$ and variance $V(Q) = \int\int yy^T p(y|\psi)Q(d\psi)dy - m(Q)m(Q)^T$ have the following properties:

$$\mathrm{E}[m(Q)] = m_0 \ , \ \ \mathrm{Var}[m(Q)] = \frac{V_0}{n_0 - p - 1} \approx V_0/n_0 \ , \ \ \mathrm{E}[V(Q)] = \frac{n_0 + \alpha + 1}{n_0}\frac{n_0 V_0}{n_0 - p - 1} \approx V_0 \tag{2.6}$$

Here, m_0 is the desired prior mean and V_0 is the desired prior covariance matrix, derived from the marginal prior information. Within the context of the DPMM, it is difficult to specify the prior on $V(Q)$ separately from that on $m(Q)$. We construct three different nonparametric prior distributions for a comparative analysis of the Old Faithful data:

- Informative DPMM π_0^I: The base measure density q_0 is as in (2.5) with $(\mu_0 = m_0, \kappa_0 = n_0/(\alpha+1), \nu_0 = n_0, S_0 = \nu_0 V_0)$, where the diagonal of V_0 is v_0, the marginal variances from the prior sample, and the correlation is equal to the sample correlation from the observed data. This results in a prior on Q satisfying (2.6), thereby utilizing the prior information.

- Noninformative DPMM π_0^N: The base measure density q_0 is as in (2.5) with $(\mu_0 = \bar{y}, \kappa_0 = 1/10, \nu_0 = p + 2 = 4, S_0 = S_y)$, where $\bar{y}$ is the sample mean from the $n = 30$ values in the observed sample, and S_y is the sample covariance matrix. This prior does not use information from the prior sample, and is designed to promote relative diffuseness of the induced prior on the marginal population means and variances. Note that using sample moments for the hyperparameters weakly centers the prior around the observed data. We can view this as a type of “unit information” prior (Kass and Wasserman 1995).

- Marginally specified prior π_1: Letting $\theta = (m_1, m_2, v_1, v_2)$ be the unknown population means and marginal variances, we construct a marginally specified prior by replacing the θ-margin of π_0^N with $p_1(\theta)$, a product of two univariate normal and two inverse-gamma densities, chosen to match the prior on θ induced by π_0^I as closely as possible. Figure A.2 compares p_1 with kernel density estimates of the marginal priors induced by π_0^I.

Thus π_0^I and π_1 have very similar θ-margins, but otherwise π_1 matches the more diffuse prior π_0^N. We can give π_1 any θ-margin we wish, but matching the margins of π_0^I and π_1 facilitates comparison. The hyperparameter α was set to 1 for all of the above prior distributions. In order to evaluate the Metropolis-Hastings ratios when approximating the posterior distribution under π_1, we found that a skewed multivariate t-distribution provided a very accurate approximation to the joint distribution of the marginal means and log variances induced by π_0^N. Via a change of variables, this provides an accurate approximation to $p_0(\theta)$, with which the acceptance probability is computed for approximation of $\pi_1(f|y)$. Figure A.3 gives an assessment of the adequacy of this approximation, comparing a smoothed density estimate of random draws from the approximated p_0 with a smoothed density estimate of random draws from the true p_0 induced by π_N^0.

We ran Markov chains of length 25,000 under each prior, with parameter values being saved every 10th iteration, resulting in 2500 simulated values of each parameter with which to make posterior approximations. The chains showed no evidence of non-stationarity and mixed well under each prior: Based on the dependent MCMC sequences of length 2500, the equivalent number of independent observations of θ (i.e., the effective sample sizes) were estimated as above 2000 for each element of θ and under each prior. We did sample from the posterior under π_1 using a stick-breaking representation and a slice sampler. The results were not markedly different from those obtained using the marginal sampler. This slice sampling approach required dependent MCMC sequences of length $550,000$ to achieve an effective sample size of 2500; computational time per independent sample was 95% that of the marginal sampler.

Posterior predictive distributions under the three priors are shown in Figure A.4. The informative DPMM provides a poor representation of the population distribution, given in light gray contours. This is primarily a result of having to set the κ_0 hyperparameter to be moderately large ($\kappa_0 = 15$) in order to obtain the desired informative prior variance for the population mean $m = (m_1, m_2)$. Unfortunately, setting this parameter so high means that values of μ in the mixture model are tightly concentrated around m_0, and so the multimodality is not captured. In contrast, the posteriors under the noninformative DPMM π_0^I and the MSP π_1 are able to capture the multimodality of the population.

Figure A.5 gives marginal density estimates under the different priors. The figure suggests that the posterior under π_1 is better at representing the underlying population than the posteriors under the other priors. Recall that the observed sample contains an unrepresentative number of low-valued observations. The posterior under the non-informative prior π_0^N uses only the observed data and thus is equally unrepresentative of the population. In contrast, π_1 is able to use some information from the prior sample, and is therefore more representative of the population.

Finally, the marginal posterior distributions of the marginal parameters m and $\log v$ are given in Figure A.6. The priors are given in gray and the resulting posterior distributions are given in black. The population values based upon the full set of 272 observations are given by gray vertical lines. Across all parameters, π_1 gives posteriors that are most concentrated around the population means. Note that the difference between the priors and the posteriors under π_0^I is not that large. We conjecture that this is primarily a result of the fact that under π_0^I, most observations are estimated as coming from the same mixture component, thereby overestimating the entropy, when in fact the data are bimodal. In contrast, π_1 is able to recognize the bimodality and obtain improved estimates of the marginal densities.

In this example, we have shown that efforts to make the canonical DPMM informative in terms of marginal means and variances leads to poor density estimates, while a noninformative DPMM leads to suboptimal estimates of functionals due to its inability to incorporate prior information. In contrast, a marginally specified prior is able to both incorporate prior information and provide accurate density estimation.

2.4 Marginally specified priors for contingency table data

Even when multivariate categorical data include only moderate numbers of variables and categories, large or full models that allow for complex or arbitrary multivariate dependence can involve a very large number of parameters. For example, a full model for the $2 \times 3 \times 2 \times 8 \times 12$-way contingency table data we consider later in this section requires a 1151-dimensional parameter. One Bayesian approach to the analysis of such data is via model selection among reduced log-linear models (Dawid and Lauritzen 1993; Dobra and Massam 2010). However, model selection can be difficult even for moderate numbers of variables and categories, due to the large number of models with low posterior probability and the resulting difficulty in completely exploring the model space. An alternative NP Bayes approach is provided by Dunson and Xing (2009), who developed a prior based on a Dirichlet process mixture of product multinomial distributions. Such a prior has full support on the parameter space but concentrates prior mass near simple submodels. One drawback to this approach is the lack of a straightforward method for the incorporation of the type of marginal prior information that is frequently available for categorical data.

In this section we consider an alternative NP Bayes approach based on a marginal adjustment to a standard Dirichlet prior distribution. This approach is computationally straightforward and allows for the incorporation of prior information on specific functionals of the unknown population distribution, such as the univariate marginals.

2.4.1 The canonical Dirichlet prior

Multivariate categorical data consist of observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$, for which $y_{ij} \in \{1, 2, \ldots, d_j\}$ for $j = 1, \ldots, p$. A $p-$way contingency table is a common representation for such data, in which each cell of the table indicates the count of observations $\boldsymbol{y}_i$ such that $y_{i1} = c_1, \ldots, y_{ip} = c_p$ for a specific response vector $c = (c_1, \ldots, c_p)$. The sampling model for a contingency table can be expressed as a multinomial distribution, where for each cell $c \in \mathcal{C} = \{c : 1 \leq c_j \leq d_j, j = 1, \ldots, p\}$ we define $f_c \equiv \Pr(y_{i1} = c_1, \ldots, y_{ip} = c_p)$. The full model of all distributions for the data can then be indexed by the parameter $f = \{f_c : c \in \mathcal{C}\}$, which lies in the $(\prod d_j - 1)$-dimensional simplex. Given n i.i.d. observations, the likelihood is $L(f|y_1, \ldots, y_n) = \prod_{c_1=1}^{d_1} \times \cdots \times \prod_{c_p=1}^{d_p} f_c^{\sum_i 1(y_{i1}=c_1, \cdots, y_{ip}=c_p)}$, for which a standard conjugate prior is the Dirichlet distribution with hyperparameter $\alpha \in (\mathbb{R}^+)^{\prod d_j}$. This is a nonparametric prior in the sense that it gives full support on the space of possible values of f.

The Dirichlet prior is an appealing choice computationally because of its conjugacy, but this convenience can have undesirable side effects. In particular, choosing an uninformative Dirichlet prior for f induces substantial informativeness about the marginals $\{\theta_1, \ldots, \theta_p\}$, where $\theta_j = \{\theta_{j1}, \ldots, \theta_{jd_j}\} = \{\Pr(y_{ij} = 1|f), \ldots, \Pr(y_{ij} = d_j|f)\}$. For example, setting $\alpha_c = 1$ for each cell $c \in \mathcal{C}$ results in a uniform prior distribution for f, often used as a default prior distribution in the absence of prior information. However, the induced prior on the marginals $\theta_1, \ldots, \theta_p$ is highly informative: The marginalization properties of the Dirichlet distribution result in $\theta_j \sim \text{Dir}(\prod_{k \neq j} d_k, \ldots, \prod_{k \neq j} d_k)$, which is generally highly concentrated around the uniform distribution on $\{1, \ldots, d_j\}$. On the other hand, it is reasonably straightforward to choose values of α_c to induce particular marginal Dirichlet priors on the θ_j's, although each marginal prior must have the same concentration. However, this approach to constructing an informative prior for the margins necessarily induces a prior over the remaining aspects of f, such as the dependence structure, that could be undesirably informative.

2.4.2 A marginally specified prior

To overcome these undesirable features of the Dirichlet prior, we construct a nonparametric prior on f based upon a Dirichlet distribution with a low total concentration, but with the induced marginal priors for $\theta_1, \ldots, \theta_p$ replaced with informative priors to reflect known information. Specifically, our prior for f takes the form

$$\begin{aligned} \pi_1(f) &= \pi_0(f|\theta) \times p_1(\theta) \\ &= \pi_0(f|\theta) \times \prod_{j=1}^{p} p_{1j}(\theta_j), \end{aligned}$$

where $\pi_0(f)$ is a Dirichlet$(\alpha_0, \ldots, \alpha_0)$ distribution on the $(\prod d_j - 1)$-dimensional simplex and p_{1j} is an informative Dirichlet distribution on $(d_j - 1)$-dimensional simplex. Recall from Section 2 that the marginally specified prior π_1 is the closest distribution in Kullback-Leibler divergence to π_0 that has the desired priors on $\theta_1, \ldots, \theta_p$. Also note that the methodology does not require that these induced priors be Dirichlet, although making them so will facilitate comparison to an informative Dirichlet prior distribution on f in the example data analysis that follows.

Estimation of f via the posterior distribution $\pi_1(f|\boldsymbol{y})$ can proceed via an MCMC algorithm. As in the previous section, we modify an MCMC algorithm for simulating from $\pi_0(f|y)$, the posterior under the canonical nonparametric prior, in order to obtain simulations from $\pi_1(f|y)$, the posterior under the marginally specified prior. Our particular MCMC scheme relies on the representation of a Dirichlet-distributed random variable as a set of independent gamma variables scaled to sum to one. That is, if $Z_c \sim \text{gamma}(\alpha_c, 1)$ and $f_c = Z_c / \sum Z_{c'}$, then $f \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{|\mathcal{C}|})$. We employ an MCMC algorithm that is based upon simulating proposed values of $\{\ln Z_c : c \in \mathcal{C}\}$ from a normal distribution centered at the current values. Because of the high dimension of the parameter f, proposing changes to every element of f simultaneously results in low acceptance rates. To avoid this problem, at each iteration of the algorithm we propose changes to randomly chosen subvectors of f. The steps in a single iteration of the MCMC algorithm are then as follows:

1. Generate a proposal $\{f^*, \theta_1^*, \ldots, \theta_p^*\}$:
   (a) randomly sample a set of cells $\mathcal{C}' \subset \mathcal{C}$;
   (b) simulate proposals $\{\log Z_c^* : c \in \mathcal{C}'\} = \{\log Z_c : c \in \mathcal{C}'\} + \epsilon$, $\epsilon \sim \text{normal}(0, \delta I)$;
   (c) compute the corresponding f^* and marginal probabilities $\theta_1^*, \ldots, \theta_p^*$.
2. Compute the acceptance ratio $r = r_0 r_1$ from r_0, the acceptance ratio for f under π_0, and r_1, the marginal prior ratio:

$$r_0 = \frac{p(y|f^*)\pi_0(Z^*)}{p(y|f)\pi_0(Z)} \prod_c (Z_c^*/Z_c) \quad , \quad r_1 = \frac{p_1(\theta^*)/p_0(\theta^*)}{p_1(\theta)/p_0(\theta)}.$$

3. Accept $f^*, \theta_1^*, \ldots, \theta_p^*$ with probability $1 \wedge r$.

Note that the ratio r_0 includes the Jacobian of the transformation from Z to $\ln Z$, as the proposal distribution is symmetric on the log-scale. The number of cells $|\mathcal{C}'|$ to update at each step and the variance parameter δ in the proposal distribution can be adjusted to achieve target acceptance rates.

As mentioned above, we take p_1 to be a product of Dirichlet densities representing prior information about the margins $\theta_1, \ldots, \theta_p$. To calculate r_1 we must also compute the corresponding joint distribution p_0 of $\theta_1, \ldots, \theta_p$ under the Dirichlet distribution π_0 on f. We approximate p_0 by the product of the prior marginal densities of $\theta_1, \ldots, \theta_p$ under π_0, each of which are Dirichlet. However, we note that the θ_j's are only approximately independent of each other under π_0.

2.4.3 Example: North Carolina PUMS data

We evaluate the performance of the marginally specified prior and several associated priors in terms of their performance under the scenario of a researcher with accurate prior information about the marginal distributions of the p categorical variables. Our scenario is based on data from the Public Use Microdata Sample (PUMS) of the American Community Survey, a yearly demographic and economic survey. We consider data on gender (male, female: $d_1 = 2$), citizenship (native, naturalized, non-citizen: $d_2 = 3$), primary language spoken (English, other: $d_3 = 2$), class of worker ($d_4 = 8$), and mode of transportation to work ($d_5 = 12$) from 40,769 survey participants. The latter two variables are each dominated by a single category, "employee of private company" (63.75%) for worker class and "car, truck or van" (91.97%) for transportation. These classifications yield a five-way contingency table with $|\mathcal{C}| = 1,152$ cells. From these data we constructed a true joint distribution $\tilde{f}$ and marginal frequencies $\tilde{\theta}$ by filling out the multiway contingency table with the PUMS data, replacing zero counts in the contingency table with small fractional counts, and normalizing the resulting counts to produce a probability distribution over $|\mathcal{C}|$. We then simulated smaller datasets of various sample sizes from $\tilde{f}$, and obtained posterior estimates for each under three different prior distributions:

- Informative Dirichlet prior π_0^I: A Dirichlet distribution with parameter $\alpha_I f_0^I$, where $\alpha_I = |\mathcal{C}|$ and f_0^I is in the $(|\mathcal{C}| - 1)$-simplex. Using the method of Csiszar (1975), the prior mean f_0^I of f was chosen to be the frequency vector closest in Kullback-Leibler divergence to the uniform distribution on $|\mathcal{C}|$ among those with margins equal to $\tilde{\theta}$. The induced marginal prior on each θ_j is then $\text{Dir}(|\mathcal{C}|\tilde{\theta}_j)$, which has prior expectation

$\tilde{\theta}_j$ as desired. Note that the concentration hyperparameter α_I is the same as that for a uniform prior on the simplex.

- Noninformative Dirichlet prior π_0^N: A Dirichlet distribution with parameter $\alpha_N f_0^N$, where $\alpha_N = \sqrt{|\mathcal{C}|}$ and $f_0^N = \{1/|\mathcal{C}|, \ldots, 1/|\mathcal{C}|\}$. This prior has the same prior expectation as the uniform prior on the $(|\mathcal{C}|-1)$-simplex, but a smaller prior concentration by a factor of $\sqrt{|\mathcal{C}|}$.

- Marginally specified prior π_1: Constructed by replacing the marginal prior for θ induced by π_0^N with the marginal prior under π_0^I. To compute acceptance ratios, we have used a product of independent Dirichlets corresponding to the marginal distributions induced by π_0^N to approximate p_0. The adequacy of the approximation to p_0 is assessed in Figure A.7. through a comparison of smoothed density estimates of random draws from the approximated p_0 with smoothed density estimates of random draws from the true p_0 induced by π_N^0.

We used the true joint distribution $\tilde{f}$ to generate 200 replicate data sets of sizes $n \in \{$ 100, 1000, 5000, 10000, 20000, 40000 $\}$. The π_0^I and π_0^N priors are conjugate to the multinomial likelihood, and so their posterior distributions are available in closed form. For estimation under π_1, the MCMC algorithm described above was run for 3×10^6 iterations for each simulated dataset. The acceptance rate varied with the sample size n, from 89% at $n = 100$ down to 63% at $n =$10,000. Effective sample sizes corresponding to thinned Markov chains based on every 500th iterate were obtained and were found to be around 1000 (based on thinned chains of length 6000).

For each simulated dataset and prior we obtain posterior mean estimates $(\hat{f}, \hat{\theta})$ which we compare to the true values $(\tilde{f}, \tilde{\theta})$ used to generate the simulated data. To evaluate $\hat{\theta}$, we use an average of the absolute value of the Kullback-Leibler divergence between the true marginal distributions $\{\tilde{\theta}_1, \ldots \tilde{\theta}_p\}$ and the estimated marginal distributions $\{\hat{\theta}_1, \ldots \hat{\theta}_p\}$:

$$M = \frac{1}{p} \sum_{j=1}^{p} \left| \sum_{c=1}^{d_j} \tilde{\theta}_{jc} \; \ln\left(\hat{\theta}_{jc}/\tilde{\theta}_{jc}\right) \right|.$$

Smaller values of M indicate better performance with respect to this marginal metric.

To assess the performance of $\hat{f}$ on aspects of f other than the marginal distributions, we compared the true and estimated values of the local dependence functions (LDFs) of the $\binom{p}{2}$ separate two-way marginal distributions. These LDFs describe the two-way dependencies among the variables, and are invariant to changes in the marginal distributions (Goodman 1969). The LDFs are formed from cross-product ratios of f as follows: Letting $f^{j_1,j_2}_{c_1,c_2} = \Pr(y_{j_1} = c_1, y_{j_2} = c_2 | f)$, we define

$$LDF^{j_1,j_2}_{c_1,c_2}(f) = \ln\left(\frac{f^{j_1,j_2}_{c_1,c_2} \; f^{j_1,j_2}_{c_1+1,c_2+1}}{f^{j_1,j_2}_{c_1,c_2+1} \; f^{j_1,j_2}_{c_1+1,c_2}} \right).$$

For each simulated dataset and prior distribution, we computed the average squared error between $LDF^{j_1,j_2}_{c_1,c_2}(\hat{f})$ and $LDF^{j_1,j_2}_{c_1,c_2}(\tilde{f})$ as

$$L = \binom{p}{2}^{-1} \sum_{j_1<j_2} \frac{1}{(d_{j_1}-1)(d_{j_2}-1)} \sum_{c_1=1}^{d_{j_1}-1} \sum_{c_2=1}^{d_{j_2}-1} (LDF^{j_1,j_2}_{c_1,c_2}(\hat{f}) - LDF^{j_1,j_2}_{c_1,c_2}(\tilde{f}))^2.$$

Smaller values of L indicate better performance in terms of representing the two-way dependence structure of the true distribution $\tilde{f}$.

Figure A.8 shows the M and L performance metrics for each prior and simulated dataset, with the averages over simulations at each sample size joined by lines. The sample sizes are displayed ordinally, with a slight horizontal shift for each prior so that the results under different priors can be distinguished. π_0^I and π_1 outperform those under π_0^N, as these former two priors were designed to have correct prior expectations for θ. The initial non-monotonic trend in the performance of π_0^I with sample size is due to the fact that π_0^I has exactly correct prior expectation. If the sample size were zero, then M would be zero as well. In contrast, the second plot in Figure A.8 indicates that π_0^I provides poor estimates of the dependence functions: At all sample sizes, this prior underperforms compared to the other two, demonstrating the cost of making π_0^I directly informative about the marginals. On the other hand, π_0^N and π_1 have very comparable performance in terms of estimation of the dependence functions. These comparisons, using both the marginal and margin-free performance metrics, highlight the desirable properties of the marginally specified prior formulation: A marginally specified prior π_1 is able to represent prior information about specific functionals θ of the high-dimensional parameter f without being overly informative about other aspects of the parameter.

2.5 Discussion

Nonparametric priors for a high-dimensional parameter f based on Dirichlet processes or Dirichlet distributions do not easily facilitate partial prior information about arbitrary functionals $\theta = \theta(f)$. Attempts to make such priors informative about θ can make the prior undesirably informative about other aspects of f.

In this article, we have presented a simple solution to this problem, via construction of a marginally specified prior (MSP) that can induce a target marginal prior on a functional θ, but is otherwise as close as possible to a given canonical "noninformative" nonparametric prior. We have provided general posterior approximation schemes for such priors, based on simple modifications to standard MCMC routines for canonical nonparametric priors. In two examples we have shown that the MSP behaves as anticipated: Given accurate prior information, the MSP provides improved estimation for θ as compared to "noninformative" priors, while providing similar or better estimation performance for other aspects of the unknown parameter f.

One barrier to the adoption of MSPs is that the posterior approximation schemes we have presented require that the ratio $p_1(\theta)/p_0(\theta)$ be computable, where p_1 is the desired informative prior for θ and p_0 is the prior induced on θ by a canonical prior π_0. Generally, p_0 will not have a closed form, and so must be approximated numerically or otherwise. If the dimension of θ is small, it should generally be feasible to approximate p_0 with a kernel density estimate, or by a simple parametric family. If θ is high-dimensional, then other approximation strategies will be required, such as approximating the joint density of θ as a product density (i.e. assuming independence of subvectors of θ) or perhaps by using mixture models. The latter strategy is more flexible than the former, but it doubles the modeling efforts in any given problem by requiring one to estimate p_0 before estimating f.

Chapter 3

Learning Phenotype Densities Conditional on Many Interacting Predictors

3.1 Introduction

Many areas of research are concerned with learning the distribution of a response conditional on numerous categorical (discrete) predictors. The predictors that have actual importance for the characterization of this distribution are not usually known in advance, and in many cases hundreds or thousands of predictors are associated with each response. In addition, it is frequently the case that the predictors interact in complex ways. Methods that attempt to consider each potential interaction can quickly become mired in the enormous space of models. For example, in a moderate-dimensional case involving $p = 40$ categorical predictors, each with $d_j = 4$ possible realizations, considering all possible levels of interaction leads to a space of $4^{40} \approx 10^{24}$ possible models. Parallelization and technical tricks may work for smaller examples, but data sparsity and the sheer volume of models force us to consider different approaches. In addition, approaches to learning conditional densities that are based on mean regression do not always consider the variation in form of the density. That is, the conditional density may vary in more than just location, as illustrated in Figure B.1. Methods that score well for measures based upon mean square prediction error (MSPE) may fall short on other important questions. Figure B.2 illustrates a synthetic case where distinct combinations $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ of predictors have $E(y|\boldsymbol{x}^{(1)}) = E(y|\boldsymbol{x}^{(2)})$, but $P(y > c|\boldsymbol{x}^{(1)}) \gg P(y > c|\boldsymbol{x}^{(2)})$. Such differences in the predicted probability of extreme observations can be of considerable interest in environmental, financial, and health outcomes settings. In the work that follows, we present a novel nonparametric Bayes (NPB) approach to learning conditional densities that makes use of a conditional tensor factorization to characterize the conditional distribution given the predictor set, allowing for complex interactions between the predictors. The particular form assumed for the conditional density gives rise to an attractive predictor selection procedure, providing support for distinct predictor selection steps. This addresses the challenges of high-dimensional data and produces conditional density estimates that allow assessment of tail risks and other complex quantities.

3.2 Approach

The primary goal for our work is to model the conditional density $f(y|\boldsymbol{x})$, where the form of this density for the response y changes flexibly with the predictor vector $\boldsymbol{x}$. There is a large body of work devoted to this idea of density regression in settings involving $\boldsymbol{x}$ of dimension $p \leq 30$, and such models have provided many options for that situation. We wish to develop techniques for problems involving much larger p, and ideally to scenarios where $p > 1,000$. While several techniques exist for this high-dimensional setting, they can result in black-box models that do not motivate understanding of the effect of a particular predictor on the response. We want to provide a method that performs variable selection, assesses the probability of a predictor's inclusion in the model, and provides easily interpretable estimates of the impact of different predictors.

This classically nonparametric problem has been addressed with variations on the finite mixture model,

$$f(y) = \sum_{h=1}^{K} \pi_h \mathcal{K}(y;\ \theta_h). \tag{3.1}$$

When predictors are incorporated into the model, this becomes the basic hierarchical mixture of experts model (HME, Jordan and Jacobs (1994)). In this representation, K represents the number of contributing parametric kernels $\mathcal{K}(;\theta_h)$ distinguished by parameters θ_h. The π_h provide the weights in this convex combination of kernels, where $\sum_{h=1}^{K} \pi_h = 1$ and $(\pi_1, \ldots, \pi_K) \in \mathcal{S}_{K-1}$, the $K-1$ probability simplex. The most straightforward forms rely on a prespecified K and include the predictors $\boldsymbol{x}$ in a linear model for the mean. HME methods in the frequentist literature have often relied on expectation maximization (EM) (Dempster et al. 1977) techniques, which can suffer from overfitting (Bishop and Svensén 2003). EM approaches in the Bayesian literature seek to avoid this; Waterhouse et al. (1996) employed EM to find maximum a posteriori (MAP) estimates, using the inherent Bayesian penalty against complexity to regulate those estimates. In addition, the Bayesian framework allows the quantification of uncertainty about the parameters in the model.

Nonparametric Bayes (NPB) methods, such as Dirichlet process (DP), prompted techniques like that in Muller et al. (1996), which induced flexible conditional regression through joint modeling of the response and predictors. Subsequent methods included the predictors in π_h and/or θ_h via dependent Dirichlet Process (DDP) mixtures. De Iorio et al. (2004) proposed an ANOVA DDP model with fixed weights $\{\pi_h\}$ that used a small number of categorical predictors to index random distributions for the response. Griffin and Steel (2006) developed an ordered DDP, where the predictor vectors were mapped to specific permutations of the weights $\{\pi_h\}$, yielding different density estimates for different predictor vectors. Reich and Fuentes (2007) and Dunson and Park (2008) employed the kernel stick-breaking process to allow predictors to influence the weights. Chung and Dunson (2009) presented a further alternative in the probit stick-breaking process, which uses a probit transform of a real-valued function of the predictors to incorporate them into the weights. Methods that use joint modeling of response and predictors (Shahbaba and Neal 2009; Hannah et al. 2011; Dunson and Xing 2009) are popular and can work well under many circumstances, but estimation of the marginal distribution of the predictors is a burden.

While the discrete mixture approach (both finite and infinite) has provided the bulk of techniques for Bayesian density regression, there are notable exceptions. For example, Tokdar et al. (2010) developed a technique based upon logistic Gaussian processes. Jara and Hanson (2011) presented an approach using mixtures of transformed Gaussian processes.

These and other methods of Bayesian density regression have proven successful, but as data sets have grown in size and complexity, these approaches encounter difficulties. One particular challenge derives from the so-called "curse of dimensionality" - that is, as we consider problems in higher and higher dimensions, where we consider larger and larger predictor vectors, the complexity of interaction between these explanatory variables grows explosively and data sets may only sparsely fill the associated space. This is even more daunting when we consider discretely valued predictors, since we must consider the factorial combinations of those levels.

The associated challenges of variable selection and dimensionality reduction have been explored in Bayesian density regression. Dimensionality reduction has a goal similar to that of variable selection, that of finding a minimal set of predictors that account for variation in the response. The logistic Gaussian process approach of Tokdar et al. (2010) includes a subspace projection method to reduce the dimension of the predictor space. Reich et al. (2011) developed a technique for Bayesian sufficient dimensionality reduction based upon a prior for a central subspace. While all of these approaches have demonstrated their utility, they do not scale easily beyond $p = 30$ predictors.

There are also techniques like the random forest (Breiman 2001) that aim to find parsimonious models for density estimation involving a large number of predictors. One disadvantage to this type of "black box" method is in interpreting the impact of specific predictors on the response. Bayesian additive regression trees (BART) (Chipman et al. 2006; 2010) focus on modeling the conditional mean and assume a common residual distribution. As previously noted, there are many questions that require learning about more than just the conditional mean of the response. Another flexible approach is the Bayes network (BN), which considers the predictors and the response on equal footing to develop a parsimonious network linking all variables (Pearl 1999; Cowell et al. 1999; Lauritzen 1992). The conditional distribution of the response given the predictors can be derived from such a model, using developed BN techniques for mixed continuous and discrete data (Lauritzen 1992; Moral et al. 2001; Langseth et al. 2012). One aspect of using a BN for conditional density estimation is that the BN estimates a joint density for all of the predictors and the response in order to arrive at a conditional density. If the conditional density is of primary interest, the effort required to estimate what amounts to a huge-dimensional nuisance parameter is unattractive.

To address these disparate challenges, we propose an approach based upon a conditional tensor factorization (CTF) for the mixing weights. As in the DDP and certain of the kernel stick-breaking methods, the predictors influence the mixing weights for this CTF model. The conditional tensor factorization facilitates borrowing of information across different profiles in a flexible representation of the unknown density. We focus our attention on situations involving continuous responses and categorical predictors.

3.3 Methods

We consider a univariate response y and a vector of p categorical predictors $\boldsymbol{x} = (x_1, \ldots, x_p)$, where the j^{th} predictor x_j can take values $1, \ldots, d_j$. We would like a model that can flexibly accommodate conditional densities that change in complex ways with changes in the predictor vector. In addition, we must consider situations where $p \gg n$. In this setting, there may be very few or no exemplars for certain predictor vectors. This sparsity can derail methods that rely on the complete predictor vector $\boldsymbol{x}$ for learning about the conditional distribution of the response. To address this, we propose a Tucker-style factorization with the following general model for the conditional density $f(y|x)$:

$$f(y|\boldsymbol{x}) = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \pi_{h_1,\cdots,h_p}(\boldsymbol{x})\, \lambda(y; \theta_{h_1,\cdots,h_p})$$

where

$$\pi_{h_1,\cdots,h_p}(\boldsymbol{x}) = \prod_{j=1}^{p} \pi_{h_j}^{(j)}(x_j). \tag{3.2}$$

Each $\pi^{(j)}$ can be visualized as a matrix, where the row indexed by x_j contains weights for the combinations of the observed predictor value x_j and the latent predictor values $h_j = 1, \ldots, k_j$. The weights for a particular value of x_j are constrained to be $\in [0, 1]$, and $\sum_{h_j=1}^{k_j} \pi_{h_j}^{(j)}(x_j) = 1$. The number of latent predictors, p, is the same as the number of observed predictors. The actual dimension of the latent space will be determined by the vectors $\pi_{h_j}^{(j)}$. While this does resemble the general HME, an important distinction is in the treatment of the weights $\pi_{h_1,\cdots,h_p}(\boldsymbol{x})$ as a tensor factorization and the assumed form of the kernels $\lambda(y; \theta_{h_1,\cdots,h_p})$, which do not have direct dependency on the predictor vector $\boldsymbol{x}$. This is similar in spirit to the classification approach proposed by Yang and Dunson (2012).

Tucker decompositions (Tucker 1966) and other kinds of decompositions have appeared in the machine learning literature before. Xu et al. (2012) developed an "infinite" Tucker decomposition, making use of latent Gaussian processes rather than explicit treatment of tensors and matrices; in comparison, the proposed method uses the Tucker decomposition to characterize the mapping of predictors into weights. Other factorizations have been used for similar problems; Hoff (2011) presented a reduced-rank approach for table data, but this approach focused on the development of estimates for the mean of a continuous response. Chu and Ghahramani (2009) derive an approach for partially observed multiway data based upon a Tucker decomposition; their objective is to learn about the latent factors driving observations rather than the characterization of the response distribution or variable selection.

The collection across $j = 1, \dots, p$ forms a "soft" clustering from the $d_1 \times \cdots \times d_p$ dimensional space of the observed $\boldsymbol{x}$ to a potentially smaller $k_1 \times \cdots \times k_p$-dimensional space. That is, a predictor vector $\boldsymbol{x}$ is not exclusively associated with a single kernel, but rather with all $k_1 \times \cdots \times k_p$ kernels through the corresponding weights. This form for the mixing weights allows borrowing of information across different combinations of $h_1, \dots, h_p$. Learning about the density conditional on a sparsely observed predictor vector $\boldsymbol{x}^{(*)}$ does not rely exclusively on observations with that predictor vector; instead, each observation contributes some information. The impact of non-matching predictor vectors is governed by the set of maps $\pi^{(j)}$, rather than some hard classification. In settings of extreme sparsity, where most predictor vectors are not represented, this is an attractive property. This uses many fewer parameters than a full factorial representation, and is still flexible enough to represent complex conditional distributions. Finally, we assume normal kernels for the λ, yielding:

$$\begin{aligned} & f(y_i|\boldsymbol{x}_i) \\ & = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \Big\{ N(y_i; \theta_{h_1,\cdots,h_p}, \tau^{-1}_{h_1,\cdots,h_p}) \\ & \qquad \times \prod_{j=1}^{p} \pi^{(j)}_{h_j}(x_{ij}) \Big\} \end{aligned} \tag{3.3}$$

This resembles other mixture-based approaches to density estimation as originally specified in (3.1), but the proposed model for the weights provides the desired support for sparsity and information borrowing previously discussed.

We consider two primary tasks in learning the conditional distribution. The first is to identify those predictors which provide the most information about the response, and the second is to learn the form of the conditional distribution given the set of informative predictors. Both tasks will be influenced by our prior assumptions about uncertainty in the model parameters, quantified as prior distributions. For computational convenience, we employ conjugate priors where possible. The model proposed in (3.3) can be augmented to give a complete-data likelihood assuming a specific classification vector z_i for each observation, kernel mean parameters $\theta_{h_1,\cdots,h_p}$, kernel precision parameters $\tau_{h_1,\cdots,h_p}$ and the soft-clustering parameters $\pi^{(j)}$:

$$\begin{aligned} & \prod_{i=1}^{N} \prod_{h_1=1}^{k_1} \cdots \prod_{h_p=1}^{k_p} \Big\{ N(y_i; \theta_{h_1,\cdots,h_p}, \tau^{-1}_{h_1,\cdots,h_p}) \times \\ & \qquad \prod_{j=1}^{p} \pi^{(j)}_{h_j}(x_{ij}) \Big\}^{1[z_i=(h_1,\cdots,h_p)]} \end{aligned} \tag{3.4}$$

The dimension of the full vectors $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ will be denoted by M, where $M = k_1 \times k_p$.

3.3.1 Prior Structure

1. $\theta_{h_1,\cdots,h_p} \sim N(0, \tau_0^{-1})$.

2. $\tau_{h_1,\cdots,h_p} \sim \text{Gamma}(\delta_t/2, \gamma_t/2)$

3. $\pi^{(j)}(x_j) = (\pi_1^{(j)}(x_j), \ldots, \pi_{k_j}^{(j)}(x_j)) \sim$
   $\text{Dir}(\frac{1}{k_j}, \ldots, \frac{1}{k_j})$
   for $j = 1, \ldots, p$ and $x_j = 1, \ldots, d_j$

4. $\tau_0 \sim \text{Gamma}(\delta_0/2, \gamma_0/2)$

The final set of parameters, the $k_1, \ldots, k_p$, present a particular challenge. Since each k_j can take on the values $1, \ldots, d_j$, the resulting discrete space can be immense, and including these as parameters in the sampler is not an attractive option. Instead, we develop a stochastic search variable selection (SSVS) step that makes use of a "hard" clustering to evaluate different k_j values.

3.3.2 Full Conditionals

Given the augmented likelihood in (3.4), the assumed prior distributions and fixed values $k_1, \ldots, k_p$, the full conditional distributions are:

1. $\theta_{h_1,\cdots,h_p}|\cdots \sim N(\mu^*_{h_1,\cdots,h_p}, (\tau^*_{h_1,\cdots,h_p})^{-1})$, where:
   $\tau^*_{h_1,\cdots,h_p} = \tau_0 + \tau_{h_1,\cdots,h_p} \sum_{i=1}^N 1[z_i = (h_1, \cdots, h_p)]$
   $\mu^*_{h_1,\cdots,h_p} =$
   $\{\tau_{h_1,\cdots,h_p} \sum_{i=1}^N y_i 1[z_i = (h_1, \cdots, h_p)]\}/\tau^*_{h_1,\cdots,h_p}$

2. $\tau_{h_1,\cdots,h_p}|\cdots \sim \text{Gamma}(\delta^*/2, \gamma^*/2)$, where: $\delta^* = \delta_t + \sum_{i=1}^N 1[z_i = (h_1, \cdots, h_p)]$
   $\gamma^* = \gamma_t + \sum_{i=1}^N 1[z_i = (h_1, \cdots, h_p)](y_i - \theta_{h_1,\cdots,h_p})^2$

3. $\tau_0 | \cdots \sim \text{Gamma}([\delta_0 + M]/2, [\gamma_0 + \boldsymbol{\theta}^T \boldsymbol{\theta}]/2)$

4. $(\ \pi_1^{(j)}[x_j], \ldots, \pi_{k_j}^{(j)}[x_j]\ ) | \cdots$

$$\begin{aligned} \sim \text{Diri}(1/k_j + \sum_{i=1}^{N} 1[z_{ij} = 1],\ \ldots,& \\ 1/k_j + \sum_{i=1}^{N} 1[z_{ij} = k_j])& \end{aligned}$$

5. $\Pr[z_i = z_{jm}^* \equiv (h_1, \ldots, h_{j-1}, m, h_{j+1}, \ldots, h_p)] | \propto$

$$\phi\Big[(y_i - \theta_{z_{jm}^*})\sqrt{\tau_{z_{jm}^*}}\Big] \times \pi_m^{(j)}(x_{ij})$$

for $m = 1, \ldots, k_j$ within each $j = 1, \ldots, p$.

The updates for $\boldsymbol{\theta}$, $\boldsymbol{\tau}$ and $\boldsymbol{\pi}^{(j)}$ can be done blockwise. The z_i can updated blockwise at each position j. In updating the z_i or classification vectors, we consider each j separately and draw updates according to the usual finite mixture model approach. The conditional posterior probability of assignment to the different levels $1, \ldots, k_j$ are determined by normalizing the weighted likelihoods at each level. The weights come from the probability tensor $\boldsymbol{\pi}$, indexed by $\boldsymbol{x}_i$, and the likelihoods from the response y_i and the atoms θ and τ.

3.3.3 Predictor Selection and Parameter Estimation

Predictor selection is of paramount importance in settings with a large number of predictors. As discussed above, the sheer number of possible interactions of predictors makes the development of full models infeasible. The form of the weights shown in (3.3) provides an attractive method for predictor selection that we now develop.

To learn appropriate values for $k_1, \ldots, k_p$, we use a predictor selection step based upon a special form of the $\pi^{(j)}$. This special form of the mapping in (3.2) results if exactly one of the elements of $\pi^{(j)}(x_j)$ is equal to 1, with the other $k_j - 1$ elements equal to zero. This gives a "hard" clustering of each predictor vectors $\boldsymbol{x}_i$ to exactly one element of the $M-$dimensional space outlined above. Given a particular clustering and the prior structure outlined above, we can approximate a marginal likelihood for that clustering; these marginal likelihoods provide calibrated measures of different clusterings that drive a stochastic search. We make the simplifying assumption that $\tau_0 = \tau$ and retain the Gamma$(\delta_t/2, \gamma_t/2)$ prior for τ. This gives an exact form for the marginal likelihood of one group within the hard clustering. There will be $M = k_1 \times \cdots \times k_p$ such groups, indexed by m. The log marginal likelihood for the m^{th} group is then:

$$\frac{N_m}{2}\, log(\pi) - \frac{1}{2} log(N_m + 1) + log\, \Gamma\Big(\frac{N_m + \delta_t}{2}\Big) - log\, \Gamma\Big(\frac{\delta_t}{2}\Big)$$
$$+ \frac{\delta_t}{2}\, log(\gamma_t) - \frac{1}{2}(N_m + \delta_t)\, log(Y_m^T Y_m - \frac{(Y_m^T J_{N_m})^2}{N_m + 1} + \gamma_t),$$

where Y_m is the vector of responses and N_m is the number of observations in group m. The product of these M approximated marginal likelihoods drives a stochastic search through the space of clusterings. For each $j = 1, \ldots, p$ we consider "split" moves that result in an increase of k_j by one, and "merge" moves that result in a decrease of k_j by one. For example, if the current mapping for observed predictor j from observed labels to latent labels is $\{1 \to 1, 2 \to 1, 3 \to 2, 4 \to 2\}$, one "split" move would be to change the mapping so that $4 \to 3$, increasing k_j from 2 to 3. The available "merge" move is to change the mapping so that $3 \to 1$ and $4 \to 1$, decreasing k_j from 2 to 1. We select from all available moves with equal probability. We repeat this process for some number of iterations. After discarding burn-in iterations, we use the remaining set to compute inclusion probabilities for each $j = 1, \ldots, p$. These are the proportion of iterations in which the separate k_j are greater than 1, and indicate the importance of the corresponding predictor to the conditional distribution. This stochastic search approach is similar to the presentation in George and McCulloch (1997).

In the first stage, we examine each of the p predictors in isolation. Since it is then feasible (for $d_j \leq 5$) to encapsulate the entire stochastic search of corresponding split and merge moves in a discrete time Markov chain, this step proceeds very quickly. This can be done in an embarrassingly parallel fashion, but experimentation at $p = 5000$ where $d_j = 4$ for all j showed that the computation of each inclusion probability required 0.3s and so serial computation was not overly burdensome. We did investigate a marginal likelihood computation that made fewer simplifying assumptions and relied on numerical approximations. This approach did not produce materially different results and gave a tenfold increase in computational time.

We use the inclusion probabilities from this single-site pass to reorder the predictors in decreasing order of inclusion probability. We also impose a cutoff from the first stage, including only those predictors with inclusion probability greater than some value, typically 0.5. The cutoff may also be determined by a limit on the size of the space we wish to consider or for computational convenience. The re-ordering before the second stage of variable selection combats the tendency of the stochastic search to jump from simple clusterings to complex clusterings with similar or slightly degraded marginal likelihoods. If the best candidates from the first-pass search are considered before weaker candidates, the second-pass search performs better. The second stage of variable selection uses a sequential stochastic search variable selection, proceeding for a moderate number of iterations to produce a second set of inclusion probabilities. This uses the same approximated marginal likelihood approach as in the individual predictor assessment. Predictors with inclusion probabilities exceeding the cutoff value of 0.5 are then used in the Gibbs sampling step.

The Gibbs sampler produces a posterior sample according to the steps detailed in section 3.3.2. Each element from this MCMC sample defines a model that we can use to produce predicted values and intervals around predicted values for a test set.

3.4 Simulation Study

To assess the variable selection and prediction performance of the CTF, we conducted a simulation study, varying the number of training observations $N \in \{300, 500, 1000, 1500\}$ and using a consistent ground truth to produce simulated data sets with total number of predictors $p = 1000$. In each case, the true model was based on three predictors at positions 30, 201 and 801, each with $d_j = 4$ levels and including three-way interactions among these predictors. The combination of predictor values is associated with the mean of an underlying Gaussian, and simulated using a common residual variance τ.

For each of 20 training sets, we produced selected predictor sets and posterior samples based upon the models defined by those sets and the assumed prior structure. We then used the derived models to make predictions for 20 validation sets drawn from the same underlying true distribution. As competitor methods we used random forests (RF) and quantile regression random forests (QRF) (Meinshausen 2006); these are implemented in the `randomForest` and `quantregForest` packages in `R`. BART as implemented in the `BayesTree` package was unable to run to completion on any of the training sets, though we were able to use BART with the real data example in Section 3.5. We chose these methods as competitors for different reasons. RF and QRF include predictor selection directly, and QRF directly addresses the idea of coverage proportion. BART is another MCMC based approach, but it does not directly address variable selection, allowing us to investigate the impact of the large predictor space. We considered the use of Bayes networks, but the implicit cost in estimating the joint distribution of predictors and response made this unattractive.

To compare the methods, we used two metrics. We computed mean square prediction error (MSPE) as the average squared difference between the response value predicted by the model for a predictor vector from the validation set and the actual response value for that observation. We defined coverage proportion (COV) as the proportion of times that the 95% prediction interval for an observation in the validation set included the actual response value, averaged over the intervals for each posterior sample. When comparing performance with that of the competitors, we attempted to give those competitors what advantages we could. In the case of RF, this meant that we did two passes over the training data. The first pass identified important variables using the `importance` method in the `randomForest` package. We used the "mean decrease in accuracy" style of importance; this measurement is derived from the impact of permuting out-of-bag data for each tree in the forest. We then fed those variables identified as important as a preselected set into a second run of RF. This generally improved the MSPE performance of RF. An analogous method was not available for QRF, so we could not treat that method in the same manner. In each of the 20 cases for $p = 1000$ and training $N = 500$, the CTF outperformed RF on mean square prediction error and showed comparable 95% coverage proportions to those derived from QRF; this is summarized in Figure B.4. The CTF and RF showed comparable accuracy in identifying important predictors, but RF tended to include many unimportant predictors. In contrast, the CTF produced no false positive results, identifying the correct subset of predictors in each case. This performance is particularly attractive given the large number of possible interactions in the original predictor set. Both RF and QRF may have suffered due to the strong interactions present in these simulated data.

3.5 Molecular Epidemiology Application

To illustrate the utility of this approach, we apply it to a real-world dataset and compare its performance to that of the same competitor methods (RF, QRF, and BART). The dataset concerns DNA damage to instances of different cell lines when exposed to environmental chemicals. The exposure types are hydrogen peroxide (H2O2) and methyl methane sulfonate (MMS), and the remainder of the predictor set is genotype information on 49,428 single nucleotide polymorphisms (SNPs). Rodriguez et al. (2009) provides extensive details on the original experiments. 100 separate instances of each of 90 cell lines were exposed to each chemical and examined at each of 3 time points (before treatment, immediately after treatment, and a longer time after treatment). The nature of the measurement is destructive; at the desired time interval, comet assay was performed on each cell and the Olive tail moment (Olive et al. 1991) recorded; this assesses the amount of DNA damage in the cell, with higher measurements indicating more damage. The cells from each line are genetically identical, but the resulting distribution of Olive tail moment (OTM) has a different shape for each cell line. In addition, these distributions are different at the separate time points; generally, the Olive tail moments are smallest (least damage) before exposure to the chemical, largest (most damage) immediately after exposure, and somewhere in-between after a longer recovery time.

To develop an appropriate response, we computed empirical quantiles at percentiles $(1/32, 2/32, \ldots, 31/32)$ for each cell line at each of the three time points and then derived a single-number summary w_{ij} to tie these three quantile vectors together for cell line i and exposure j. The summary measure $w_{ij} \in (0,1)$ is the value that minimizes

$$\sum_{h=17}^{31} \left| w_{ij} Q_{ij,N,h} + (1 - w_{ij}) Q_{ij,L,h} - Q_{ij,A,h} \right| \tag{3.5}$$

Here, $Q_{ij,N,h}$ indicates the $h/32^{th}$ quantile for the i^{th} cell line's Olive tail moment distribution at the "**N**o treatment" time, with corresponding quantities for the "**L**ater" time point and the "immediately **A**fter" time point. The use of only the higher quantiles reflects our desire to learn more about the extremes of DNA repair. We used a logit transform to derive our final response $y_{ij} = log(\frac{w_{ij}}{1-w_{ij}})$; this is appropriate for the assumptions of the model. Negative values of the response indicate that the OTM distribution long after treatment is closer to the distribution right after treatment; positive values indicate that the "long after" distribution is closer to the distribution before treatment.

The researchers genotyped the cell lines at 49,428 individual SNPs, each of which had previously been associated with some aspect of DNA repair. Given the small number of cell lines and the fact that many individuals have two copies of the major allele for these SNPs, many of the SNP profiles were identical and many also had no individuals with two copies of the minor allele. We recoded the genotypes so that 1 indicated at most one copy of the major allele and 2 indicated two copies of the major allele. After recoding, we reduced the predictor set to those SNPs with distinct profiles, leaving 23,210 SNPs for analysis.

We used leave-one-out cross-validation to assess the performance of the CTF against that of the three competitors RF, QRF, and BART. Each model from the CTF is represented by an MCMC chain, so for each iterate we developed expected values and 95% prediction intervals for the left-out observation. We ran the variable selection chain for 5,000 burn-in iterations and computed inclusion probabilities from 10,000 samples. We ran the MCMC chain for 40,000 burn-in iterations and retained a sample of 20,000 iterations. Autocorrelation diagnostics indicated an effective sample size of 15,000. We used the same burn-in and posterior sample sizes for BART. As in the simulation study, we used the results from a first run of RF to seed a final run of RF.

The CTF showed consistent selection of the treatment (H2O2 or MMS) as the most important predictor and selected a set of four SNPs (IGFBP5, TGFBR3, CHC1L, XPA) as predictors; information about these SNPS is summarized in Table B.1. In contrast, RF chose the treatment variable in only 56 of the 180 cross-validation scenarios and did not consistently identify any other predictors. The CTF has a higher computational time requirement and took approximately twenty times as long as RF or QRF to estimate a model. Nevertheless, the improved performance is attractive.

When we recoded the SNPs as binary variables, there were many blocks of SNPs with identical profiles. We included only one representative from each block in the analysis. Of the four SNPs included in the model, two (TGFBR3 and CHC1L) were the representatives for multi-SNP blocks. The SNPs that shared profiles with these two representatives are summarized in Tables B.2 through B.5.

Comparison with the competitor methods showed patterns similar to the simulation study; Table B.6 compares the results from each method. The interactions between the treatment and the various SNPs may be weak enough that they do not contribute to the same elevated MSPE that RF demonstrated in the simulation study. Even though the MSPE for RF was close to that for the CTF, the CTF was able to achieve lower MSPE while not sacrificing coverage performance.

Figure B.5 shows estimated conditional densities given varying levels of the treatment and of the IGFBP5 SNP while holding the other three SNPs at the "Zero/One Copy" level, and illustrates how the conditional density changes in more than the conditional mean when the predictor vector changes. In this case, the interaction between MMS treatment and two copies of the major allele for this IGFBP5 SNP tightens the density markedly, while it has a more muted impact on the conditional mean. The change is less dramatic under the exposure to H2O2. In this setting, the shift in the mean response as treatment and genetic profile change is less interesting than the difference in conditional variance; under treatment with H2O2, the mean response is slightly different than under treatment with MMS, but the tail probabilities are noticeably different.

3.6 Conclusion

We have presented a novel method for flexible conditional density regression in the common case of a continuous response and categorical predictors. The simulation study and real data example suggest that this conditional tensor factorization method can have better performance than other modeling tools when there is substantial interaction between the predictors of interest. The CTF does have a higher computational time requirement than the competitor methods, but the improvement in prediction accuracy and coverage still make the CTF an attractive method.

A particularly appealing aspect of the CTF is predictor selection, which finds low-dimensional structure in the high-dimensional predictor set. This reduction to more parsimonious models yields a succinct description of the ways in which the phenotype varies given exposure and SNPs. Finally, a distinct advantage of the CTF is its ability to produce these conditional density estimates. This property of the CTF provides insight beyond a simple conditional expectation and makes it possible to answer more complex questions about the relationship between the response and the predictors.

Chapter 4

Nonparametric Selection of Interacting Polymorphisms Predictive of Quantitative Traits From Family Data

4.1 Introduction

4.1.1 Motivation and Proposed Approach

Large p, small n problems have become commonplace in statistical analyses of genetic data. In settings with very large p, such as studies of single nucleotide polymorphisms (SNPs), there is a multi-layered challenge, as it is necessary to develop scalable methods, which can address the computational and statistical curse of dimensionality, while leading to interpretable results that are not overly subject to false discoveries. There is a rich literature addressing problems of this type. The most common approach relies on independent screening with false discovery rate (FDR) control, but incorporating SNPs simultaneously can have substantial advantages Hoggart et al. (2008). This is typically done within an additive generalized linear model, with a penalty incorporated to allow $p \gg n$. For example, L_1

penalties lead to Lasso and sparse estimation (Tibshirani 1996), while L_2 penalties lead to ridge regression procedures that allow large numbers of SNPs with small coefficients (Hoerl and Kennard 1970), and the elastic net combines L_1 and L_2 penalties (Zou and Hastie 2005).

The assumption in such approaches is that SNPs have an additive relationship with the (transformed) mean of the response phenotype. Clearly, this assumption is very restrictive and it is natural biologically to expect interactions among the SNPs and with environmental exposures, with the density of a quantitative trait not simply shifting in mean as the important factors vary but changing in variance and shape. The focus of this article is on developing a practically implementable statistical method that allows this degree of flexibility, while additionally accommodating variable selection, covariate adjustments and dependence arising within families.

There is a considerable statistical and bioinformatics literature focused on identifying interacting predictors of a quantitative trait from among a very large number of candidates. Lou et al. (2007) developed a generalized multi factor dimensionality reduction combinatorial algorithm. Chen et al. (2007) proposed a forest-based approach to identify gene-gene interactions. Zou et al. (2010) incorporated variable selection within a Gaussian process prior for the regression function, enabling selection of a subset of interacting SNPs impacting the mean of the distribution of a quantitative response. Yi (2010) provide an overview of statistical approaches for identifying genetic interactions in high-dimensional settings, such as genome-wide association studies. Cordell (2009) reviews methods for selecting interactions between genetic loci contributing to human disease.

These methods lack the density regression flexibility, which is a key aspect we are focused on. In particular, our hypothesis is that genetic variants and environmental factors impacting the phenotype density do not simply shift the mean in a simple way, but may impact the variance and shape. Such changes in variance, skewness, and higher moments of the phenotype density seem a natural consequence of genetic heterogeneity. By incorporating this flexibility within our statistical model, we aim to increase power to detect important SNPs and environmental factors. There is a literature on density regression, reviewed below, but current methods fail to produce a variable selection procedure that can adjust for family dependence and scale to very large p.

We propose a model based upon a tensor factorization for the weights on a set of normal kernels. This factorization depends on soft clusterings of observed categorical predictors into a lower-dimensional space. This approach implicitly accounts for interactions and gives a flexible form for the density conditional on predictors and random effects. This conditional tensor factorization for correlated data (CTFC) method accommodates the possibility that the density of the response conditional on the predictors may change in ways that are more complex than a simple shift in conditional mean.

This is an important consideration in contexts where real concern is over behavior in the tails of the conditional distribution. For example, if the quantitative trait is a measure of health like body mass index (BMI), it may be the case that certain combinations of SNP variability have minimal impact on the average BMI, but that the conditional probability of being obese vary widely with these same combinations. Figure C.1 uses data drawn from the 2001-2002 National Health and Nutrition Examination Survey (NHANES) to illustrate distributions for adults of different educational attainments; this is not variation associated with SNPs, but the concept is the same. The two groups have very similar median BMI but noticeably different 90^{th} percentile BMI.

Such considerations have motivated research into quantile regression (Koenker and Bassett 1978). Quantile regression has been used in the modeling of quantitative traits as a function of genetic variation; Ho et al. (2009) used a range of quantile regression techniques to find age-dependent gene expression patterns, and the technique has been used in the analysis of comparative genomic hybridization data (Eilers and de Menezes 2005). In these applications, the practitioner chooses specific quantiles to focus on. In contrast, the density regression approach models all quantiles simultaneously. Our approach addresses predictor selection and density estimation in separate stages, and combines deterministic approximations with Markov Chain Monte Carlo (MCMC) techniques for inference. The proposed method makes no parametric assumptions about the form of this conditional density, instead using an adaptable nonparametric form. An important component is selection of those predictors with the most influence on the response allowing $p \gg n$.

4.1.2 Background on Density Regression

There is a rich literature on estimation of the conditional density of a response variable $y \in \mathcal{Y}$ given predictors $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathcal{X}$. For example, the hierarchical mixtures of experts (HME) model (Jordan and Jacobs 1994) lets

$$f(y|x) = \sum_{h=1}^{K} \pi_h(x)\mathcal{K}(y; x,\ \theta_h). \tag{4.1}$$

Here, K represents the number of contributing parametric kernels. Each kernel $\mathcal{K}(y; x, \theta_h)$ is distinguished by its parameters θ_h and consequently the manner in which the predictors x influence the response. This is a convex combination of kernels, weighted by the $\pi_h(x)$, where $\sum_{h=1}^{K} \pi_h(x) = 1$ and $\{\pi_1(x), \ldots, \pi_K(x)\} \in \mathcal{S}_{K-1}$, the $K-1$ probability simplex. In the simplest form, K is pre-specified and the predictors $\boldsymbol{x}$ enter into a linear model for the mean. HME implementations have often relied on expectation maximization (EM) (Dempster et al. 1977) techniques, which can suffer from overfitting (Bishop and Svensén 2003). Bayesian approaches seek to avoid this; for example, Waterhouse et al. (1996) developed maximum a posteriori (MAP) algorithm, using the inherent Bayesian penalty against complexity to regulate those estimates.

Much of the work in nonparametric Bayes (NPB) methods has centered on the Dirichlet process (DP). Muller et al. (1996) jointly model the response and predictors to induce flexible conditional densities. More recent extensions of this idea have been proposed by Shahbaba and Neal (2009) and Hannah et al. (2011) among others. An unappealing attribute is the need to estimate a potentially high-dimensional nuisance parameter corresponding to the marginal distribution of x. Alternative models have been defined directly for the conditional density using dependent Dirichlet process (DDP) mixtures. De Iorio et al. (2004) proposed an ANOVA DDP model with fixed weights $\{\pi_h\}$ that used a small number of categorical predictors to index random distributions for the response. Griffin and Steel (2006) developed an ordered DDP, where the predictor vectors were mapped to specific permutations of the weights $\{\pi_h\}$, yielding different density estimates for different predictor vectors. Predictor-dependent stick-breaking processes have also been proposed, including kernel stick-breaking Dunson and Park (2008) and probit stick-breaking (Chung and Dunson 2009).

For moderate p, these and other methods of density regression have been successful. As p increases, one encounters the curse of dimensionality. Some methods are available for variable selection (Chung and Dunson 2009) and dimensionality reduction (Tokdar et al. 2010; Reich et al. 2011), but these methods do not scale much beyond $p = 30$ and do not account for correlation in the data, such as occurs in family studies. There is a literature on generalizing penalization methods to adjust for dependence. Rakitsch et al. (2013) propose "LMM-Lasso", extending Lasso to correlated responses. Lippert et al. (2011) develops a realized relationship matrix (RRM) based upon a limited number of SNPs, and then introduces a streamlined optimization method of estimation. This approach and the related approach in Listgarten et al. (2012) are based on the idea that a linear mixed model (LMM) with no fixed effects, where the correlation structure is derived from the SNPs, is equivalent to a regression of those SNPs on the response of interest. Zhang et al. (2010) instead compressed samples into a smaller set of clusters, where clustering is based on kinship of individuals. Our focus is instead of including dependence within a scalable nonparametric Bayes density regression approach.

4.2 Methods

We consider situations where we have observations from N different groups (families). The i^{th} group has n_i members, each with a response observation y_{ij} and a vector of predictors $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$. The s^{th} element, x_{ijs}, is assumed to take on a value in $\{1, \ldots, d_s\}$. We first present the conditional density for the response y_{ij}. Given a family-specific random effect vector $\boldsymbol{b}_i$ and the subject-specific feature vector $\boldsymbol{x}_{ij}$, we assume

$$f(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i) = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \Big\{ N(y_{ij}; \theta_{h_1\cdots h_p} + b_{ij}, \tau_{h_1\cdots h_p}) \Big\} \times \Big\{ \prod_{s=1}^{p} \pi_{h_s}^{(s)}(x_{ijs}) \Big\} \tag{4.2}$$

This defines a mixture of normal kernels, where the kernels are indexed by $h_1, \cdots, h_p$ and the weight given to a particular kernel is driven by the predictor vector $\boldsymbol{x}_{ij}$ through the selections from the $\pi^{(s)}$ for that predictor vector. The representation in (4.2) is notable in that the predictors appear only in the expression for the weight on the kernel indexed by $h_1, \cdots, h_p$. The parameter $\boldsymbol{\pi}^{(s)}$ is a collection of d_s vectors, where the c^{th} such vector is $\{\pi_1^{(s)}(c), \ldots, \pi_{h_s}^{(s)}(c)\}$ and $\sum_{\ell=1}^{h_s} \pi_\ell^{(s)}(c) = 1$. One special case results if each vector $\pi^{(s)}(c)$ has exactly one entry equal to 1. Under this configuration, each predictor vector $\boldsymbol{x}_{ij}$ maps exclusively to a single kernel parameterized by the $(\theta_{h_1\cdots h_p}, \tau_{h_1\cdots h_p})$ pair. This "hard" clustering defines a typical mean-shift scenario through a full-interaction model, and the conditional densities are straightforward Gaussians. In contrast, when the $\pi^{(s)}(c)$ give nontrivial weights to each latent level, this "soft" clustering drives a more flexible form for the conditional density.

The representation in (4.2) assumes knowledge of the $k_1, \ldots, k_p$, whose product $\mathcal{M} = \prod_{s=1}^{p} k_s$ determines the number of kernels used in the conditional mean. Given this assumption, we can introduce prior distributions for the remaining parameters and proceed with the estimation necessary for the density regression. Without this information, we will be required to search over a very large discrete space of potential models defined by the possible combinations of the k_j. In a low-dimensional situation, we might do the predictor selection through the introduction of priors on inclusion for each predictor, but in high-dimensional situations this would require a much larger parameter set and longer computational time.

An important observation is that if $k_s = 1$ for a particular predictor, that predictor does not participate in the model. The resulting vectors $\pi^{(s)}(c)$ for each $c \in \{1, \ldots, d_j\}$ will simply be scalar values of 1, indicating that different values of c produce no change in the weighting term, and hence have no influence on the conditional distribution. This aspect of the representation in (4.2) provides a pathway to predictor selection. Instead of including the predictor selection alongside sampling of the other parameters in the model, we undertake predictor selection as a separate first step.

4.2.1 Predictor Selection

The central approach in our predictor selection method is the identification of those k_s that we set equal to 1 and so remove from the model. As outlined above, an important special case results when exactly one entry in each vector $\boldsymbol{\pi}^{(s)}(c), c \in \{1, \ldots, d_s\}$ is equal to 1, resulting in a "hard" mapping of the observed predictor vector $\boldsymbol{x}_{ij}$ to a selection vector $\boldsymbol{z}_{ij}$. Given different sets of such "hard" mappings, including that for $k_s = 1$, we can evaluate the associated marginal likelihoods for each under a particular prior structure and identify models with the highest marginal likelihoods. Under a particular set of $\{k_1, \ldots, k_p\}$, the dimension of the parameter $\boldsymbol{\theta}$ is $\mathcal{M} = \prod_{s=1}^{p} k_s$.

To evaluate the marginal likelihood under a particular "hard" mapping, we couple the sampling model in (4.2) with the following prior structure:

1. $\tau \sim \text{Gamma}(\delta_t/2, \gamma_t/2)$

2. $\boldsymbol{\theta} \sim N(0, \tau_0^{-1} I_{\mathcal{M}})$

3. $\tau_0 \sim \text{Gamma}(\delta_0/2, \gamma_0/2)$

4. $b_i \sim N(0, \eta^{-1} R_i)$, where R_i is a correlation matrix specific to the i^{th} family.

5. $\eta \sim \text{Gamma}(\delta_e/2, \gamma_e/2)$

We assume that the entries in the correlation matrices R_i are defined by the kinship between pairs of individuals in the family. We assume $R_i(i, j) = 2^{-d(i,j)}$, where the kinship metric $d(i, j)$ is derived from the degree of relationship: $d(i, j) = \infty$ for unrelated individuals, $d(i, j) = 1$ for first-degree relatives (parent and child, full siblings), and $d(i, i) = 0$ for self-kinship.

For the specific deterministic mapping defined by the k_s and the $\boldsymbol{\pi}^{(s)}$, the p-dimensional predictor vector $\boldsymbol{x}_{ij}$ will map to an $\mathcal{M}$-dimensional vector $\boldsymbol{z}_{ij}$, where exactly one element of $\boldsymbol{z}_{ij}$ is equal to 1 and all other elements are zero. The resulting likelihood conditional on the random effects $\boldsymbol{b}_i$ is then

$$\prod_{i=1}^{N} N(\boldsymbol{Z}_i \boldsymbol{\theta} + \boldsymbol{b}_i, \tau^{-1} I_{n_i}) \tag{4.3}$$

Here, $\boldsymbol{Z}_i$ is the stacked set of $\boldsymbol{z}_{ij}$. After multiplying by the specified priors for $\boldsymbol{\theta}$, τ, τ_0, η, and each $\boldsymbol{b}_i$, we can integrate out $\boldsymbol{\theta}$ and the $\boldsymbol{b}_i$, leaving the marginal likelihood expression:

$$\begin{aligned}
\int\!\!\int\!\!\int 2\pi^{\frac{N}{2}} \Big\{ \prod_{i=1}^{N} |\Sigma_i|^{-\frac{1}{2}} \Big\} \Big(\frac{\gamma_t}{2}\Big)^{\frac{\delta_t}{2}} \Big(\frac{\gamma_0}{2}\Big)^{\frac{\delta_0}{2}} \Big(\frac{\gamma_e}{2}\Big)^{\frac{\delta_e}{2}} \Big[\Gamma\Big(\frac{\delta_t}{2}\Big)\Gamma\Big(\frac{\delta_0}{2}\Big)\Gamma\Big(\frac{\delta_e}{2}\Big)\Big]^{-1} \\
\times\, \tau^{\frac{\delta_t}{2}-1} \tau_0^{\frac{\delta_0}{2}-1} \eta^{\frac{\delta_e}{2}-1} e^{-\frac{1}{2}(\tau\gamma_t + \tau_0\gamma_0 + \eta\gamma_e)} \\
\times\, |V|^{-\frac{1}{2}} \quad \times\, e^{\frac{1}{2}(W^T V^{-1} W - \sum_{i=1}^{N} Y_i^T \Sigma_i^{-1} Y_i)}\, d\tau\, d\tau_0\, d\eta \qquad (4.4)
\end{aligned}$$

where

$$W = \sum_{i=1}^{N} Z_i^T \Sigma_i^{-1} Y_i\,; \quad V = \tau_0 I_{\mathcal{M}} + \sum_{i=1}^{N} Z_i^T Z_i\,; \quad \Sigma_i = \tau^{-1} I_{n_i} + \eta^{-1} R_i$$

This final integral is not available in closed form; we use the method of Laplace to approximate the marginal likelihood. We log-transformed the precision parameters τ, τ_0, η so that we could work in $\mathbb{R}^3$ and use unconstrained optimization to find the mode and Hessian for computation of the approximate value. Note that there is one difference between the form assumed here and that shown in (4.2); we use a common τ. Using different $\tau_{h_1 \cdots h_p}$ would substantially complicate the Laplace approximation due to the need to integrate over a higher dimensional space.

It will not be feasible to evaluate marginal likelihoods for each possible mapping, except in very low-dimensional cases where $\mathcal{M}$ is small. Instead, we develop a stochastic search method to traverse the possible model space. Assuming we start with a particular "hard" mapping with an associated marginal likelihood ML_c, we move sequentially through the features from $s = 1, \ldots, p$. For the s^{th} feature, if the current maximum mapped index k_s is less than the maximum possible index d_s, we consider new maps such that $k_s^* = k_s + 1$. Depending on the value of d_s (the maximum value for the observed x_{ijs}), there are several possibilities. For instance, if the current map is $\{1 \to 1, 2 \to 1, 3 \to 2\}$, where $k_s = 2$ and $d_s = 3$, a new map could specify $2 \rightarrow 3$, with $k_s = 3$. If the current $k_s > 1$, we also consider new maps such that $k_s^* = k_s - 1$. Using the same example map, a new map could set $3 \rightarrow 1$, so that $k_s = 1$. After drawing a candidate map from these possibilities, we evaluate the marginal likelihood ML_p for this proposed map and accept the move with probability $1 \wedge (ML_p/ML_c)$. At the conclusion of a pass through the p predictors, we examine the current set of maps and increment a count c_s by one if the current map has $k_s > 1$, indicating that the s^{th} predictor has some influence on the response. After T such tours through the predictor set, we can identify those predictors such that the inclusion proportion $p_s = c_s/T$ is greater than some threshold, usually 0.5. We then proceed with estimation of the parameters associated with the model that includes only those predictors.

The order of evaluation of the predictors can be crucial. In simulation, we noticed that even a predictor designed to be very important in prediction of the response might not be identified if it was at a large offset s and so was evaluated very late in the search. This appeared to be due to rough alignment of other predictors with these intentionally important features, coupled with the random nature of the search. In some cases, a disadvantageous move would be accepted; the search would never escape the resulting state and would never identify the important predictors.

To combat this tendency, we implement a first-stage process to assign each predictor a preliminary score that we can use to preferentially order the predictors for the stochastic search. This first-stage process uses exactly the same core method as the full search, but with a single predictor. If d_s is moderate ($d_s < 6$), the number of possible mappings is small enough that we can feasibly evaluate marginal likelihoods for each mapping and consider them in a compact discrete-time Markov chain. This "collapsed" stochastic search uses the same idea of proposed transitions and acceptance probabilities, but arrives at inclusion proportions p_s without using random acceptance steps. For example, for a dichotomous predictor, there are only two distinct mappings from the two-dimensional space of observed values; map A takes 1 → 1 and 2 → 2, and map B takes both 1 and 2 to 1. Once we have computed marginal likelihoods for each mapping, we can collapse the stochastic search framework into a manageably-sized transition matrix. In this case, we have:

$$P = \begin{bmatrix} p_{AA} & p_{AB} \\ p_{BA} & p_{BB} \end{bmatrix} \tag{4.5}$$

In this notation, p_{ij} is the probability of accepting the transition from map i to map j, where $p_{ij} = 1 \wedge (ML_j/ML_i)$. The form of P^m as $m \to \infty$ gives the long-run proportion of times that a corresponding stochastic search will spend in each state. Since map B is equivalent to a model in which $k_s = 1$ for the corresponding predictor, the complement of the proportion of time that the search spends in this state yields an inclusion probability p_s for the predictor. We do this for each predictor and use the set of p_s to define an order of evaluation in the full stochastic search. This first stage of predictor selection can be completely parallelized, because each predictor is considered in isolation. It is then possible to use emerging techniques for massively parallel computation, such as GPU exploitation and general cloud-based computing, to complete this stage in a greatly reduced time. After this first stage is complete, we sort the predictors by these first-stage inclusion probabilities to guide the second stage of predictor selection. Any predictor with a first-stage inclusion probability p_s below a certain threshold (usually 0.5) is not included in the second stage. In the notation of (4.2), $k_s = 1$ for those predictors not included, indicating that each observed level maps to a single latent level and the predictor does not inform the response. This type of multistage approach to predictor selection has been investigated in the genetics literature previous; Marchini et al. (2005) provides an assessment of this approach in simulation, and Hoh et al. (2000) describes an application to cohort study data. As in much of the literature, these association studies have investigated discrete outcomes like disease status rather than complex quantitative traits. Furthermore, they do not consider interactions more complex than two-way.

4.2.2 Density Estimation

Once we have selected a set of predictors, we implement a Gibbs sampler sized according to the selected predictor set and approximate a draw from the posterior distribution using the full conditional distributions. The resulting draws of the several parameters give us an approximation to the posterior distributions of the various conditional densities. We use the same prior structure as in the marginal likelihood calcuations in the predictor selection step, but allow the residual precision parameters $\tau_{h_1\cdots h_p}$ to be component-specific, with prior distributions $\tau_{h_1\cdots h_p} \sim \text{Gamma}(\delta_t/2, \gamma_t/2)$ such that the prior expectation of $\tau_{h_1\cdots h_p}$ is δ_t/γ_t. In addition, we place Dirichlet priors on each vector $\pi^{(s)}_{h_s}$. Finally, we introduce vectors $\boldsymbol{s}_{ij} = (s_{ij1}, \ldots, s_{ijp})$ to indicate latent assignments for each observation. This augmentation scheme gives a complete-data likelihood that facilitates updating of the other parameters in the model:

$$\prod_{h_1=1}^{k_1}\cdots\prod_{h_p=1}^{k_p}\prod_{i=1}^{N}\prod_{j=1}^{n_i}\left\{N(y_{ij};\theta_{h_1\cdots h_p}+b_{ij},\tau^{-1}_{h_1\cdots h_p})\right\}^{1[\boldsymbol{s}_{ij}=(h_1,\cdots,h_p)]} \tag{4.6}$$

$\boldsymbol{s}_{ij} = (h_1, \ldots, h_p)$ indicates that the j^{th} individual in the i^{th} family is associated with the kernel Under this prior specification, and using augmentation variables s_{ij} to indicate assignment to one of the $\mathcal{M}$ groups, we update parameters according to the following steps. In these full conditional distributions, $\mathcal{S}_{h_1,\cdots,h_p}$ indicates the set of observations such that $\boldsymbol{s}_{ij} = (h_1, \ldots, h_p)$.

1. $\Pr(s_{ijs} = h_s) \propto \pi^{(j)}_{h_s}(x_{ijs}) \times N(y_{ij}; \theta_{s_{ij1},\cdots,s_{ij(s-1)},h_s,s_{ij(s+1)},\cdots,s_{ijp}} + b_{ij}, \tau^{-1}_{s_{ij1},\cdots,s_{ij(s-1)},h_s,s_{ij(s+1)},\cdots,s_{ijp}})$

2. $\theta_{h_1\cdots h_p} \,|\, \cdots \sim N(\mu_*, \tau_*^{-1})$, where

$\tau_* = \tau_0 + \tau_{h_1 \cdots h_p} \cdot \sum_{\mathcal{S}_{h_1, \cdots, h_p}}(1)$

$\mu_* = [\tau_{h_1 \cdots h_p} \cdot \sum_{\mathcal{S}_{h_1, \cdots, h_p}}(y_{ij} - b_{ij})]\tau_*^{-1}$

3. $\tau_{h_1 \cdots h_p} \mid \cdots \sim$ Gamma$(\delta^*/2, \gamma^*/2)$, where

   $\delta^* = \delta_t + \sum_{\mathcal{S}_{h_1, \cdots, h_p}}(1)$

   $\gamma^* = \gamma_t + \sum_{\mathcal{S}_{h_1, \cdots, h_p}}(y_{ij} - \theta_{h_1 \cdots h_p} - b_{ij})^2$

4. $\eta \mid \cdots \sim$ Gamma$([\delta_e + \sum_i n_i]/2, [\gamma_e + \sum_i \boldsymbol{b}_i^T R_i^{-1} \boldsymbol{b}_i]/2)$

5. $\tau_0 \mid \cdots \sim$ Gamma$([\delta_0 + \mathcal{M}]/2, [\gamma_0 + \boldsymbol{\theta}^T \boldsymbol{\theta}]/2)$

6. $\boldsymbol{b}_i \mid \cdots \sim N(m_i, V_i)$, where

   $V_i = [diag(\tau_{s_{i1}}, \ldots, \tau_{s_{i,n_i}}) + \eta * R_i^{-1}]^{-1}$ and

   $m_i = V_i \times (diag(\tau_{s_{i1}}, \ldots, \tau_{s_{i,n_i}})[Z_i(\theta_{s_{i1}}, \ldots, \theta_{s_{i,n_i}})^T - \boldsymbol{y}_i])$

7. $(\pi_1^{(s)}(c), \ldots, \pi_{k_s}^{(s)}(c)) \sim$ Diri$[1/k_s + \sum_{i,j} 1(x_{ijs} = c) \times 1(s_{ijs} = 1) \ldots,$

   $1/k_s + \sum_{i,j} 1(x_{ijs} = 1) \times 1(s_{ijs} = k_s)]$

We can then combine these parameters to get conditional density estimates given different predictor vectors, and produce conditional expectations and prediction intervals.

4.3 Simulation Study

To evaluate the performance of the proposed method, we conducted a simulation study, varying conditions of data generation under the structure outlined in (4.2). In this study, we simulated $p = 500$ predictors, each with $d_j = 4$ levels. Of these simulated predictors, three actually had influence on the response through a full-interaction model. The precision parameters were set as $\tau = 1.0$ and $\eta = 2.0$, and the 64 θ values spaced across $(-10, 10)$. We used a varying number of families, $N \in \{300, 400, 500, 700, 1200\}$, as training sets and used a consistent family size $n_i = 4$, with a correlation structure corresponding to two parents and two children. We did 25 replicate sets at each value of N. For each result, we did out-of-sample comparisons with another set drawn from the same ground truth. We compared results with the Lasso (Tibshirani 1996). The Lasso does not deal explicitly with mixed effects models, but provided a useful comparison with respect to predictor selection and general estimation. In this case the simulated model is close to a linear additive model and the Lasso should not be at an obvious disadvantage. We used the `glmnet` R package to do the Lasso comparisons.

The CTFC method showed good performance on predictor selection and out-of-sample prediction under each simulation scenario. At sizes $N \in \{300, 400, 500\}$, the CTFC was comparable to the Lasso in terms of MSPE, and it was much better than the Lasso when $N \in \{700, 1200\}$. Figure C.2 summarizes the comparative MSPE performance across training sample sizes.

4.4 Chiari Malformation I Data

As a real example we consider data from a family-based study of Chiari malformation type I (CMI). CMI is a hereditary disorder resulting in extension (herniation) of the cerebellar tonsils into the foramen magnum. CMI is sometimes asymptomatic, but can also present many neurological symptoms, including hydrocephalus, muscle weakness, insomnia, or depression. The goal of the original research was to identify genetic loci associated with a positive diagnosis of Chiari Malformation I. In our approach, we model a continuous phenotype, the average tonsillar herniation, that has some relationship with this binary diagnosis.

Full details for this family-based study are described in Markunas et al. (2013), but we summarize the most relevant aspects here. Families with at least two afflicted individuals were enrolled via self-referral. CMI status was determined by an MRI measurement for some individuals ($N = 126$), but in a preponderance of cases ($N = 241$), CMI status was derived from examination of medical records or from the individual's physician. Where MRI info was available, if the MRI indicated cerebellar tonsillar herniation of 3 mm or more for both cerebral tonsils or herniation of 5 mm or more for either tonsil, the subject was classified as affected. Each of the $N = 367$ individuals had some genotype information determined for a large ($N = 214436$) set of SNPs. These individuals were genotyped for a larger set of SNPs ($N = 592532$), but after quality control this set was reduced, and not all remaining SNPs were available for all individuals. We determined the largest overlapping set of available SNPs and MRI information, and after filtering in this manner we had 105 observations, including information for 210164 SNPs. These 105 individuals represented 47 families; of these 47 families, 16 were represented by only a single family member. The final data set also included four nongenetic factors: the age of the individual at the time of the MRI, the individual's sex, whether the individual was the proband for his/her family, and whether the family had any clinical features associated with connective tissue disease (CTD). We discretized the age at MRI into one of four classes using the SUGS method (Wang and Dunson 2011). Families with formally diagnosed hereditary CTDs had been excluded from the study, but Markunas et al. included identification of families where at least one member exhibited conditions associated with CTDs. The goal of this identification was to drive a stratified analysis to account for potentially different CMI disease pathways.

Because all non-missing observations for average tonsillar herniation were non-negative, we needed to transform them to a scale more appropriate for normal kernels. In addition, 15 of the 105 observations had average tonsillar herniation of zero. We used as our response $log([\text{avg. hern}] + 0.01)$ to address the zero observations and to model on a scale consistent with normal kernels. While this is related to the criteria used to determine affected status in the original work, the two measurements are not directly comparable.

Family-based studies (Spielman and Ewens 1996) were developed to combat the spurious associations that result from population heterogeneity. In some cases, there may be no association between a response of interest and some observed predictor, but the structure of the population can induce correlation where none in fact exists. The use of family-based controls eliminates mismatching of subjects with ethnically different controls and thus avoids inducing marginal correlation. One aspect of the family-based study that we do not directly address is ascertainment bias, which can arise when families are recruited based upon the presence of the disease of interest. For a hereditary disorder, the increased probability of at least one member of the family suffering the affliction means that some families are more likely to be part of the enrollment set; the increased proportion of affected individuals can introduce bias into estimates of association with different predictors. Pfeiffer et al. (2001) demonstrate the resulting bias in logistic regressions under these circumstances, and present an approach to this problem in family-based studies of disease incidence, using a conditional likelihood to address the ascertainment effect. Zhang et al. (2009) also addresses logistic regression for disease incidence and provides a semiparametric Bayes approach for situations involving very low-dimensional predictors.

4.4.1 Cross-Validation

To assess performance of the CTFC method in this case, we conducted a leave-one-out cross-validation process. In parallel, we excluded one observation at a time from the full set ($N = 105$) and conducted the complete predictor selection and model estimation using the remaining observations. Once the model was estimated using this "training" set, we examined the prediction error. In many cases, the pattern of missingness resulted in singleton "families" with no remaining family members to provide additional prediction information. In those cases where there were remaining family members, we developed predictions for the left-out j^{th} family member based on the fixed SNPs for that family member as well as the conditional expectation for the left-out random effect, $b_{ij}|b_{i\setminus j}$. The resulting prediction for the left-out individual was based upon the component-specific intercepts $\theta_{h_1 \cdots h_p}$, weighted according to the predictor-driven loadings $\prod_{j=1}^{p} \pi_{h_j}^{(j)}(x_{ij})$ for distinct $h_1, \ldots, h_p$ as specified in (4.2). The 105 different training samples were well-aligned in their identification of important predictors. Figure C.3 illustrates the results of the first pass selection, showing this general agreement on important predictors. We compared the mean squared prediction error (MSPE) with random forests (Breiman 2001) and the coverage probabilities with quantile regression random forests (Meinshausen 2006). While neither random forests nor quantile regression random forests are designed explicitly to address random effects, the number of singleton families in this analysis may have made this less of a factor. Over the left out observations, CTFC had an average MSPE of 3.09 and RF had an average MSPE of 2.92. The CTFC method's 95% coverage intervals included the left-out observation 94% of the time, in stark contrast to the QRF's average rate of 82%. This sharp discrepancy in coverage is likely due to the nontrivial form for the conditional density. Even though the random forest family of methods performs well for MSPE, they appear to fall short in assessment of higher moments for the conditional distributions.

4.4.2 Full Data Set

Using the full dataset of subjects with both MRI information and genotype information, we ran the variable selection and model estimation steps previously outlined. We ran the second stage of predictor selection for 1000 iterations over the 32 predictors selected in the first stage. The small sample size led to some numerical instabilities for certain combinations of predictors, and the second stage of predictor selection quickly found only a few stable configurations. This identified four genetic factors and one physical factor, the age at MRI, as predictors important for the prediction of the transformed response; the descriptions of these factors are shown in Table C.1. We ran the Gibbs sampler for 20,000 burn-in iterations and derived conditional density estimates for the phenotype from a retained sample of 20,000. Figure C.4 shows selected conditional density estimates for particular predictor vectors. In each case, the levels of the three final predictors were fixed at two copies of the major allele for RS11877713, zero/one copy for RS16954106, and two copies for RS10981955. We varied RS6894946 between zero/one and two copies of its major allele and varied the age at MRI across the four discretized levels. The conditional densities indicate some interaction between the two predictors. While the change from zero/one copy to two copies of the major allele for RS6894946 generally promotes the leftmost peak, the effect is most pronounced at the second level of the age at MRI variable. In that case, the credible bands indicate more support for this interaction. The credible bands are wide in each case, which is not surprising given the small sample size. This also illustrates the complex nature of this phenotype and indicates that there are likely additional explanatory predictors.

The conclusions from this approach to the data do not widely concur with the Markunas et al. study. This could be due to several factors. Many fewer subjects had available MRI data, giving us a much smaller sample size. We were able to use only 105 of the 367 original observations, representing 47 of the original 66 families; Markunas et al. had complete information on affliction status and predictors for all 367 individuals. Furthermore, after filtering for available MRI and predictor data, several of the families were represented by a single individual, hampering our ability to distinguish between intra-family variation and residual variation. In addition, we are using a continuous measurement of a physical condition rather than a binary disease status as our response, and used one of many possible transformations of this measurement in our model. The binary measurement used by Markunas et al. is related to but not equivalent to the continuous measurements of tonsillar herniation that we used as a response. The demarcation of “normal” and “afflicted” is based upon much clinical experience, but is necessarily subjective; our use of the continuous underlying quantity is free of this subjectivity. We also examined a set of 210160 SNPs (after filtering for missingness) without regard for spacing of those SNPs. In contrast, Markunas et al. used a set of 12056 SNPs derived from an initial clustering of SNPs and thinning to get an acceptable spacing between SNPs. The SNPs we identified may be in linkage disequilibrium with those identified by Markunas et al.. Finally, we made specific assumptions about intra-family correlation structure that the original study did not use.

4.5 Discussion

We have presented the CTFC, the conditional tensor factorization method for correlated data. This is a general method for the analysis of correlated data in the presence of high-dimensional categorical predictor sets. In simulation studies and in an analysis of a real data example relating a continuous phenotype to a large number of SNPs and other categorical predictors, we have seen good performance. When appropriate parallelization steps are taken, the method scales well to high-dimensional predictor sets. Most importantly, the models produced by the CTFC provide insight beyond the conditional expectation for different combinations of predictors, and capture the complexity in higher moments of the phenotype's conditional distribution. This ability to identify important predictors and to capture complex conditional densities in the presence of correlated observations makes the CTFC an attractive method for the modeling of quantitative phenotypes. The CTFC method gives a differential expression of the predictors across different combinations of those predictors. That is, the influence of one level of a predictor may produce a straightforward Gaussian conditional distribution when combined with a specific combination of other predictors, but changing the level of that predictor results in a more complex conditional density. What we provide is a targeted method for the commonly encountered special case, modeling a *continuous* response conditional on *categorical* predictors. In addition, we provide an approach for potentially correlated observations. In cases where the typical LMM is too inflexible to describe apparent variation, our nonparametric approach gives an attractive alternative.

Chapter 5

Discussion and Future Directions

In the first dissertation paper we presented a general approach for the introduction of one-dimensional marginal prior information in different applications of NP Bayes techniques, specifically the DPM model and the canonical Dirichlet prior. The multivariate unordered categorical data example suggests one future direction for this work. The original source for that data (the North Carolina PUMS) also includes cross-tabulation information for many of the quantities involved. Since these tables were formed from a much larger dataset (the data for the entire state of North Carolina), they could provide very reliable prior information about the correlation structure between the two variables they summarize. While simply incorporating a two-way table is not, in theory, a significant departure from the work already done, it may expose further difficulties with our assumptions about the form of the joint distribution of the marginals, $p_0(\boldsymbol{\theta})$, induced under the base nonparametric prior. In the work already presented, we made simplifying assumptions about that induced prior. Those assumptions may have been warranted, but as the marginal prior information becomes more complex, the corresponding induced $p_0(\boldsymbol{\theta})$ may not be well-approximated. In addition, we considered only two nonparametric priors in our analysis. Different NP Bayes approaches, such as that developed in Dunson and Xing (2009) to address multivariate unordered categorical data, might also benefit from our treatment of marginal prior information.

In the second and third papers we addressed the problem of density regression in the presence of multiple interacting categorical predictors, with and without correlated responses. While the wide array of genetic epidemiology and other data sets involving large numbers of SNPs makes this special case important, the technique will be of greater utility when we extend it to predictors and responses of mixed type. The speed of computation, particularly in the mixed model setting, remains an issue. While we can achive faster real time results through parallelization, further refinements of our technique could reduce the total computation time needed. Even though the discussions around "Big Data" contain a great deal of hyperbole, there are certainly difficult high-dimensional questions to answer, and techniques like ours, successfully adapted, could be important tools to separate the signal from the noise.

Appendix A: Chapter 2

A.1 Proofs

Proof of Theorem 1. Let $\mathcal{A}$ be the Borel sets of a Hausdorff space $\mathcal{F}$ and let θ be a measurable map from $(\mathcal{F}, \mathcal{A})$ to the measurable space $(\Theta, \mathcal{B})$. Let P_0 be the probability measure over $\mathcal{B}$ defined by $P_0(B) = \pi(\theta^{-1}B)\ \forall B \in \mathcal{B}$. The results of Hoffmann-Jørgensen (1971) give the existence of a regular conditional probability function $\Lambda_0(A|\theta)$ such that $\Lambda_0(A|\cdot)$ is $\mathcal{B}$-measurable for each $A \in \mathcal{A}$, $\Lambda_0(\cdot|\theta)$ is a probability distribution over $\mathcal{A}$ for each $\theta \in \Theta$, and that $\Lambda_0(A|\theta(f))$ is a version of the conditional probability of A given $\mathcal{B}$, in that

$$\mathrm{E}_0[1(\theta(f)) \in B) \times \pi_0(A|\theta(f))] \equiv \int_B \Lambda_0(A|\theta)\ P_0(d\theta) = \pi_0(A \cap \theta^{-1}B)\ \forall B \in \mathcal{B},$$

where θ represents either the function mapping $\mathcal{F}$ to Θ or a point in Θ, depending on the context.

Let P_1 be a probability measure on $\mathcal{B}$ such that $P_1 \ll P_0$. Define $\pi_1 : \mathcal{A} \to [0,1]$ by

$$\pi_1(A) = \int \Lambda_0(A|\theta) P_1(d\theta).$$

Then clearly $0 = \pi_1(\emptyset) \leq \pi_1(A) \leq \pi_1(\mathcal{F}) = 1$ for all $A \in \mathcal{A}$. Additionally, for a countable disjoint collection of sets $\{A_1, A_2, \ldots\} \subset \mathcal{A}$ with $A = \cup A_i$, we have

$$\begin{aligned}
\pi_1(A) &= \int \Lambda_0(A|\theta) P_1(d\theta) \\
&= \int \sum_{i=1}^{\infty} \Lambda_0(A_i|\theta) P_1(d\theta) \\
&= \sum_{i=1}^{\infty} \int \Lambda_0(A_i|\theta) P_1(d\theta) \\
&= \sum_{i=1}^{\infty} \pi_1(A_i),
\end{aligned}$$

where the second-to-last line follows from the monotone convergence theorem. Therefore, $\pi_1(A)$ is a probability measure on $(\mathcal{F}, \mathcal{A})$. To compute the marginal distribution of π_1, let $B \in \mathcal{B}$ and $h(\theta) = dP_1/dP_0$. Then

$$\begin{aligned}
\pi_1(\theta^{-1}B) &= \int \Lambda_0(\theta^{-1}B|\theta) P_1(d\theta) \\
&= \int \Lambda_0(\theta^{-1}B|\theta) h(\theta) P_0(d\theta).
\end{aligned}$$

The Radon-Nikodym derivative $h(\theta)$ is positive and measurable, and so we can express $h(\theta)$ as the limit of simple functions, $h(\theta) = \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} 1(\theta \in B_{n,k})$. By the monotone convergence theorem we have

$$\begin{aligned}
\pi_1(\theta^{-1}B) &= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} \int \Lambda_0(\theta^{-1}B|\theta) 1(\theta \in B_{n,k}) P_0(d\theta) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} \Lambda_0(\theta^{-1}(B \cap B_{n,k})) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} P_0(B \cap B_{n,k}) = \int_B h(\theta) P_0(d\theta) = \int_B \frac{dP_1}{dP_0}(\theta) P_0(d\theta) = P_1(B).
\end{aligned}$$

Finally, the Radon-Nikodym derivative of π_1 with respect to π_0 can be found via a similar calculation: For any $A \in \mathcal{A}$,

$$\begin{aligned}
\pi_1(A) &= \int \Lambda_0(A|\theta) P_1(d\theta) \\
&= \int \Lambda_0(A|\theta) h(\theta) P_0(d\theta) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} \int \pi_0(A|\theta) 1(\theta \in B_{n,k}) P_0(d\theta) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} h_{n,k} \pi_0(A \cap \theta^{-1} B_{n,k}) \\
&= \int_A \left(\lim_{n\to\infty} \sum_{k=1}^{n} h_k 1(\theta(f) \in B_{n,k}) \right) \pi_0(df) \\
&= \int_A h(\theta(f)) \pi_0(df) = \int_A \frac{dP_1}{dP_0}(\theta(f)) \pi_0(f).
\end{aligned}$$

Proof of Lemma 2: Let $\mathcal{A}$ be the Borel sets of a Hausdorff space $\mathcal{F}$. For $k \in \{0, 1\}$ let π_k be a probability measure on $(\mathcal{F}, \mathcal{A})$ and let P_k be the measure on $(\Theta, \mathcal{B})$ induced by the measurable map $\theta : \mathcal{F} \to \Theta$. Recall that if $\pi_1 \not\ll \pi_0$, then the KL divergence $D(\pi_1||\pi_0)$ is infinite. On the other hand, we will show that if $\pi_1 \ll \pi_0$ and $P_1 \ll P_0$, then the KL divergence $D(\pi_1||\pi_0)$ of π_0 from π_1 can be expressed in terms of marginal and conditional densities with respect to a common dominating measure, and that if P_1 and P_0 are fixed, the divergence is minimized by matching the conditional distributions of π_0 and π_1.

Let μ be a dominating measure for π_0 and π_1, and let ν be a dominating measure for P_1 and P_0. The results of Hoffmann-Jørgensen (1971) give the existence of a regular conditional probability function $\tilde{\Lambda}_0(\cdot|\cdot) : \mathcal{A} \times \Theta \to [0, 1]$ with the properties described in the proof of Theorem 1. Now for each $A \in \mathcal{A}$ and $\theta \in \Theta$, define $\Lambda_0(A|\theta) = \tilde{\Lambda}_0(A|\theta) \times 1(\pi_0(A) > 0)$. It is easy to check that this is measurable in θ for each $A \in \mathcal{A}$, is a version of the conditional probability of A given θ and is dominated by π_0, and therefore by μ, for each $\theta \in \Theta$. Therefore, the measures $\{\Lambda_0(\cdot|\theta) : \theta \in \Theta\}$ form a dominated class with densities $\{\lambda_0(\cdot|\theta) :$

$\theta \in \Theta\}$ with respect to μ. By Tonelli's theorem we can write

$$\mathrm{Pr}_0(\{f, \theta\} \in A \times B) \equiv \pi_0(A \cap \theta^{-1}B) = \int_B \int_A \lambda_0(f|\theta)p_0(\theta)\ \mu(df) \times \nu(d\theta),$$

and so π_0 has a density $\lambda_0(f|\theta)p_0(\theta)$ with respect to the product measure $\mu \times \nu$. The same construction can be made for π_1, giving the existence of a conditional probability density $\lambda_1(f|\theta)$ for which

$$\mathrm{Pr}_1(\{f, \theta\} \in A \times B) \equiv \pi_1(A \cap \theta^{-1}B) = \int_B \int_A \lambda_1(f|\theta)p_1(\theta)\ \mu(df) \times \nu(d\theta).$$

Letting $B = \{\theta : p_0(\theta) > 0\}$, the KL divergence is

$$\begin{aligned}
D(\pi_1||\pi_0) &= \int_\Theta \int_\mathcal{F} \log \tfrac{\lambda_1(f|\theta)p_1(\theta)}{\lambda_0(f|\theta)p_0(\theta)} \lambda_1(f|\theta)p_1(\theta)\ \mu(df) \times \nu(d\theta) \\
&= \int_B \int_\mathcal{F} \log \tfrac{\lambda_1(f|\theta)p_1(\theta)}{\lambda_0(f|\theta)p_0(\theta)} \lambda_1(f|\theta)p_1(\theta)\ \mu(df) \times \nu(d\theta) \\
&= \int_B \int_\mathcal{F} \log \tfrac{\lambda_1(f|\theta)}{\lambda_0(f|\theta)} \lambda_1(f|\theta)p_1(\theta)\ \mu(df) \times \nu(d\theta) + \int_B \log \tfrac{p_1(\theta)}{p_0(\theta)} p_1(\theta)\ \nu(d\theta) \\
&= \int_\Theta D(\Lambda_1(\cdot|\theta)||\Lambda_0(\cdot|\theta))\ P_1(d\theta) + D(P_1||P_0), \qquad \text{(A.1)}
\end{aligned}$$

where the last line follows from the assumption that $P_1 \ll P_0$ and so $P_1(B) = P_0(B) = 1$. Since the integrand in (A.1) is always greater than or equal to zero, we have $D(\pi_1||\pi_0) \geq D(P_1||P_0)$ with equality when $\Lambda_1(\cdot|\theta) = \Lambda_0(\cdot|\theta)$ for θ-a.e. P_1.

A.2 Figures

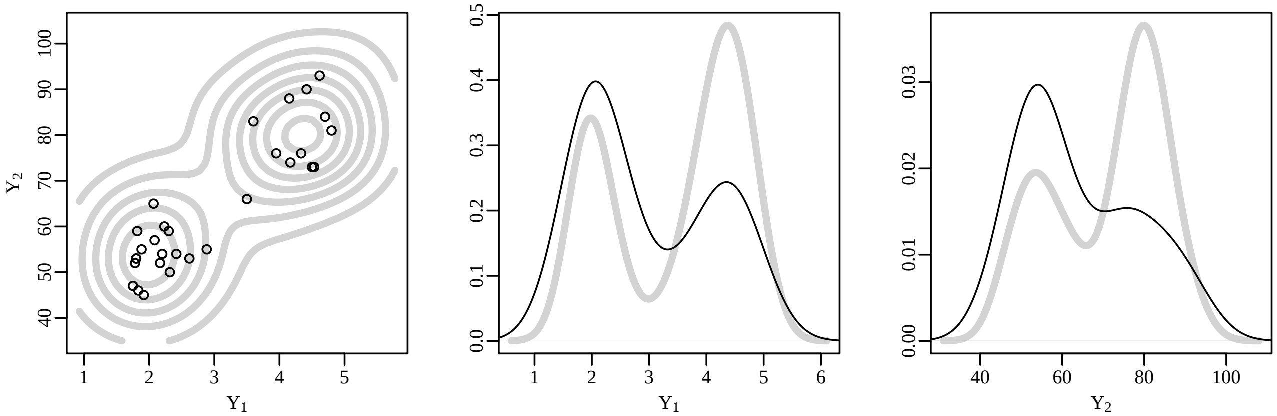

Figure A.1: Population and sample: The left-most panel shows the contours of the population density and a scatterplot of the $n = 30$ randomly sampled observations. The center and right panels show marginal densities for the population (light gray) and sample (black).

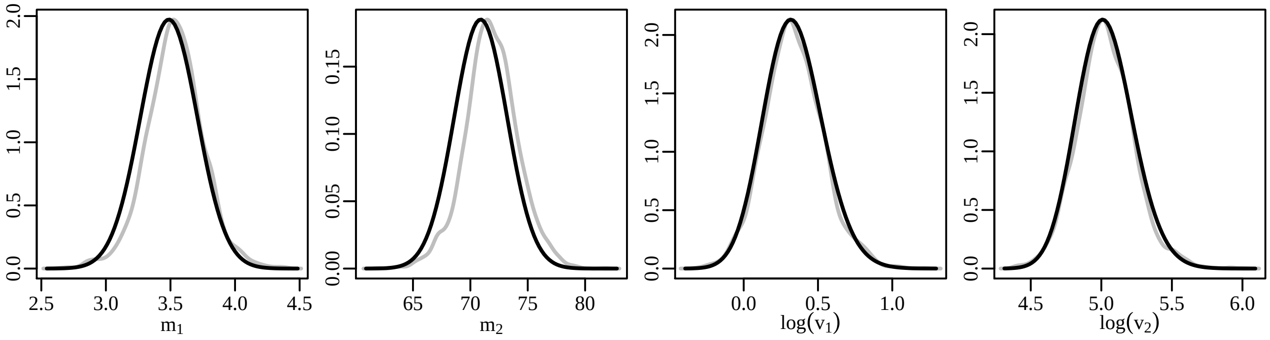

Figure A.2: p_1 priors (black) and kernel density estimates of priors induced by π_0^I (grey).

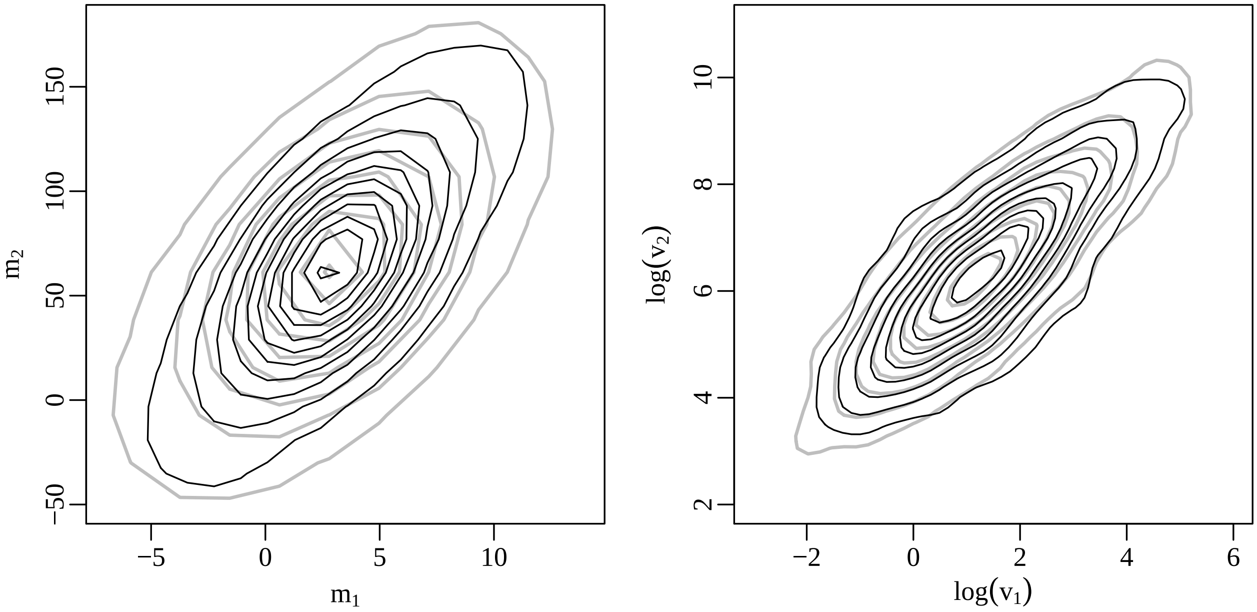

Figure A.3: Comparison of approximated p_0 (grey) and p_0 induced by π_0^N (black).

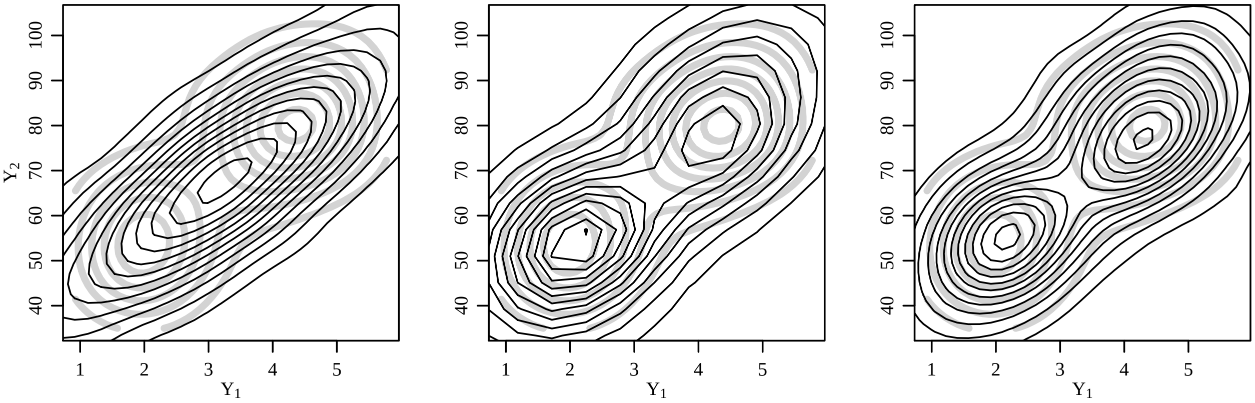

Figure A.4: Contour plots of the posterior predictive density in black and the population density in gray, under π_0^I, π_0^N and π_1 from left to right.

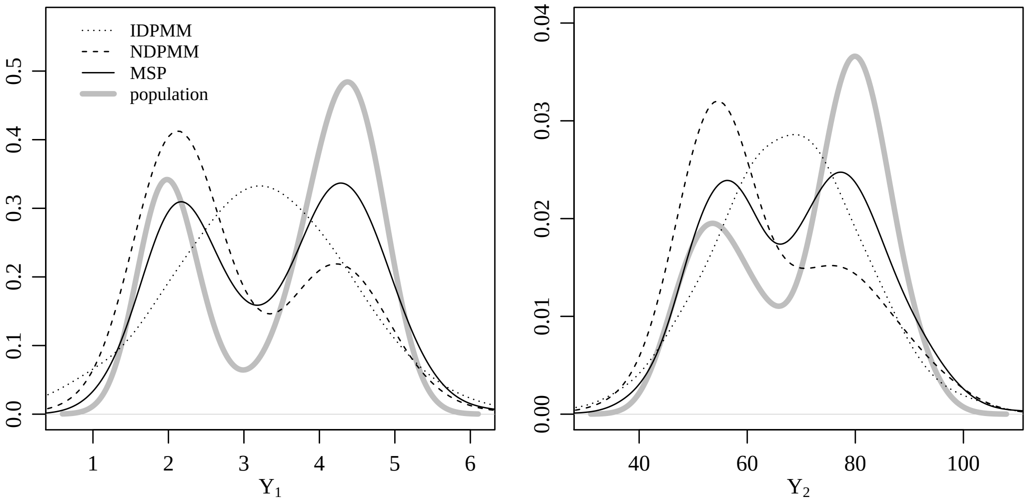

Figure A.5: Marginal population densities and estimates from the three priors: informative DPMM (IDPMM), noninformative DPMM (NDPMM) and marginally specified prior (MSP).

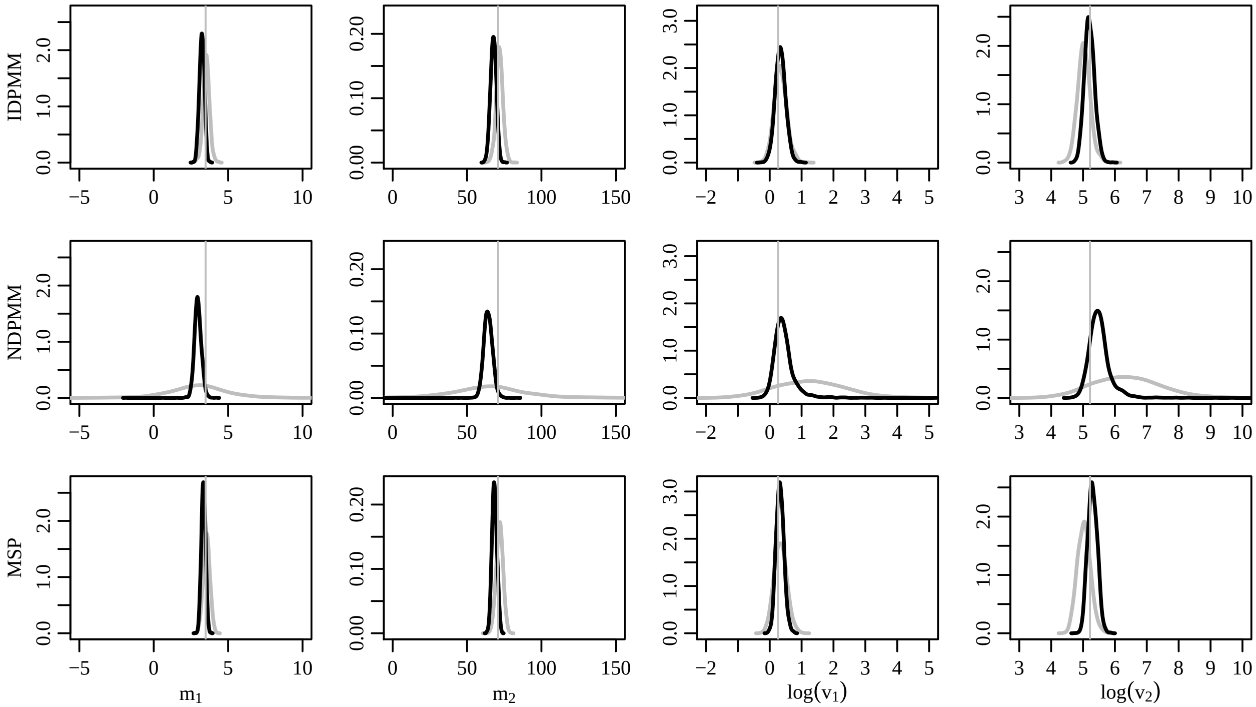

Figure A.6: Priors (gray) and posteriors (black) for the marginal means and log variances. Gray vertical lines indicate the corresponding population values derived from the full ($n = 272$) data set.

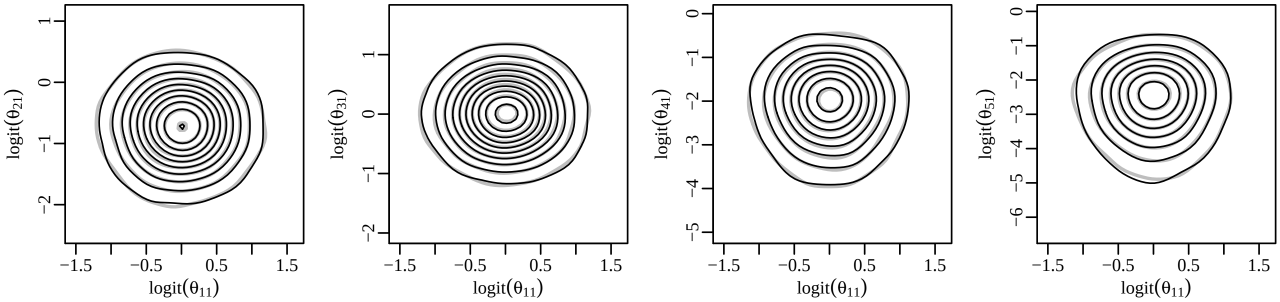

Figure A.7: Comparison of approximated p_0 (grey) and p_0 induced by π_0^N (black) for a subset of the margins. To facilicate comparison, a logit transform was used.

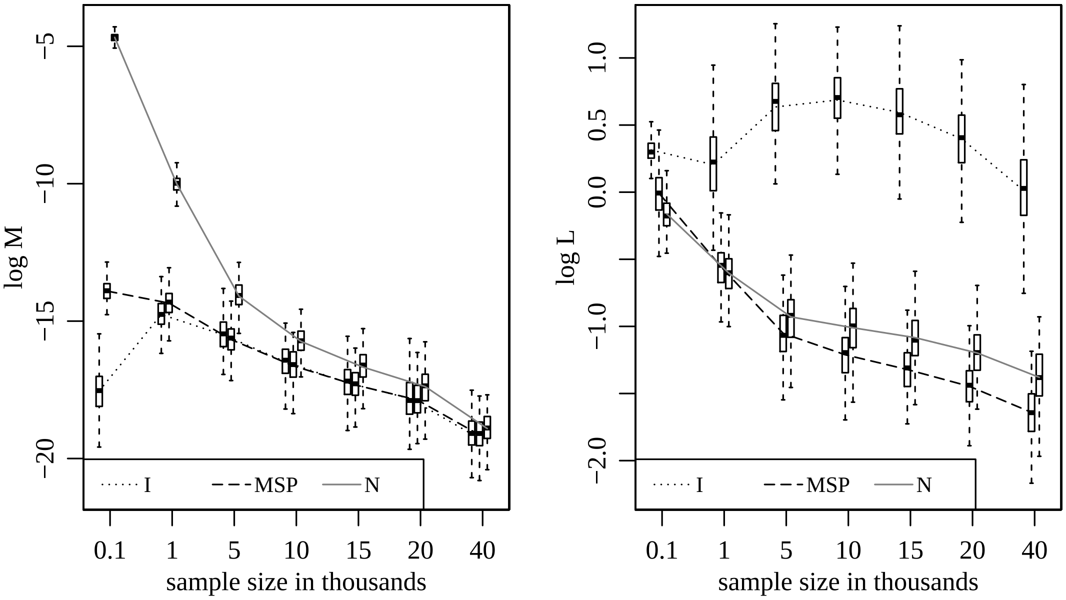

Figure A.8: Comparison of M and L metrics on the log scale for π_0^I (I), π_0^N (N) and π_1 (MSP) at various sample sizes.

Appendix B: Chapter 3

B.1 Tables

Table B.1: Details for SNPs included in the final CTF model for the molecular epidemiology data.

Gene	SNP	Chromosome position
IGFBP5	rs11575170	217256085
TGFBR3	rs17880594	92118885
CHC1L	rs9331997	47986441
XPA	rs3176745	99478631

Table B.2: Single nucleotide polymorphisms with binary-coded response profiles matching rs17880594.

		Chromosome
Gene	SNP	Position
TGFBR3	rs17880594	92118885
BCL2	rs4987854	58944606

Table B.3: Single nucleotide polymorphisms with the same binary-coded response profile as rs9331997.

		Chromosome			Chromosome
Gene	SNP	Position	Gene	SNP	Position
CHC1L	rs9331997	47986441	PTCH2	rs11573584	45066714
FGF6	rs17177088	4426536	PTCH2	rs11573575	45069786
NF1	rs17880186	26506162	CAPN9	rs28359586	228950268
LIG3	rs3136010	30349600	BCL2	rs4987713	59126421
FGFR1	rs17175968	38398906	FGF1	rs17223786	141959671
GSR	rs8190916	30693097	CAPN1	rs17881440	64706952
GSR	rs8190961	30684329	DNCH1	rs17512481	101553661
RFC4	rs3917098	188005373	DNCH1	rs17540908	101530971
BRCA2	rs4987117	31812236	RAD23B	rs11573610	109085600
TNFRSF14	rs11573983	2482568	RAD23B	rs11573639	109096805
FANCC	rs4647349	97119778	FGF10	rs17234471	44348233
FANCC	rs4647527	96917421	FGF10	rs17227997	44363031
FANCC	rs4647530	96915299	CDC27	rs11570580	42551656
PPIA	rs17860078	44807042	CDC27	rs11570463	42613866
WT1	rs5030148	32411407	ELA2	rs17216572	802482
CDC42	rs16831112	22290768	ELA2	rs17216558	802398
CDC42	rs16826272	22259266	ELA1	rs17860348	50012447
DMC1	rs11570424	37264433	CASP6	rs5030524	110842527
NQO2	rs28383640	2962424	FANCA	rs17225754	88408115
PGR	rs11571218	100439035	FANCA	rs17226218	88385110
PGR	rs11571191	100465781	MCM4	rs17287656	49044945
PGR	rs11571143	100504864	MCM4	rs17334528	49050745
RAD21	rs16889009	117946263	MCM4	rs17334423	49043956
POLB	rs3136810	42347715	MCM4	rs17334388	49041325
RAF1	rs5746260	12598664	MCM4	rs17334444	49045259
RAF1	rs5746186	12634639	MCM4	rs17287775	49052710
RAF1	rs5746244	12601660	MCM4	rs17334570	49053493
PLA2G4A	rs12720501	185097363	MCM4	rs17287649	49044042
TGFBR3	rs17883484	92143941	CHEK2	rs17880159	27414052
TGFBR3	rs17884942	91997887	CHEK2	rs17884403	27413867

Table B.4: Single nucleotide polymorphisms with the same binary-coded response profile as rs9331997 (cont'd).

		Chromosome			Chromosome
Gene	SNP	Position	Gene	SNP	Position
POLG	rs1801377	87661134	PRKDC	rs8178198	48902462
IGF2R	rs8191880	160412860	PRKDC	rs8178023	49011100
IGF2R	rs8191759	160370739	PRKDC	rs8178203	48895779
NFKB1	rs4648034	103723297	PRKDC	rs8178019	49014249
NFKB1	rs4647967	103648667	PRKDC	rs8178098	48962528
MMP14	rs17886822	22385392	PRKDC	rs8178254	48852731
CAPN6	rs17885539	110381073	E2F1	rs3213151	31735677
MSH6	rs3136295	47874069	PLA2G5	rs11573266	20285946
UHRF1	rs17883957	4897362	PLA2G5	rs11573257	20281406
NEIL1	rs5745918	73432177	POLK	rs5744688	74918394
FGF20	rs17515275	16895733	POLK	rs5744654	74908042
HK2	rs28362992	74948254	MCM3AP	rs17176254	46523745
PRKDC	rs8178199	48902372	NBS1	rs13312875	91058912
PRKDC	rs8178100	48962068	NBS1	rs11782136	91030314
PRKDC	rs8178031	49008531	NBS1	rs1805832	91058243
PRKDC	rs8178205	48895013	ACTB	rs13447431	5531830
PRKDC	rs8178180	48907242	E2F4	rs3730392	65783637
PRKDC	rs8178194	48903699	MNAT1	rs4151296	60416414
PRKDC	rs8178081	48974070	CDC16	rs17338382	114054745
PRKDC	rs8178208	48892675	CDC16	rs17338089	114034538
PRKDC	rs8178133	48935851	UMPS	rs17843831	125940890
PRKDC	rs8178189	48905990	UMPS	rs17843817	125938826
PRKDC	rs8178186	48906417	MMP2	rs17859847	54072820
PRKDC	rs8178020	49013821	CDC14B	rs16905626	98366751
PRKDC	rs8178110	48956030	CDC14B	rs16911213	98353465
PRKDC	rs8178154	48932763	CDC14B	rs16910936	98295236
PRKDC	rs8178256	48852275	MYBPC3	rs11570089	47318048
PRKDC	rs8178220	48875683	MYBPC3	rs11570052	47327990
PRKDC	rs8178069	48978232	RAD9A	rs17887226	66918883
PRKDC	rs8178109	48956139	CDC34	rs16989739	482530

Table B.5: Single nucleotide polymorphisms with the same binary-coded response profile as rs9331997 (cont'd).

		Chromosome			Chromosome
Gene	SNP	Position	Gene	SNP	Position
CDC34	rs16990717	494574	CDC7	rs13447522	91754636
CDC34	rs16990514	484034	CDC7	rs13447481	91744826
MLH1	rs4647267	37032389	E2F3	rs4134948	20592183
GCSH	rs8177861	79682996	E2F3	rs4134988	20601214
CCNA2	rs3217755	122965533	PPARD	rs9658160	35499882
FGF13	rs17497235	137622887	DNAJC3	rs17882245	95127641
FGF13	rs17538934	137578160	CKN1	rs1479646	60250260
FGF13	rs17510123	137612210	CKN1	rs4647114	60232029
PAWR	rs8176872	78538735	HGF	rs5745640	81224063
CDC25C	rs11567965	137692492	E2F2	rs3218185	23715878
CDC25C	rs11567960	137694074	WRN	rs11574310	31094575
IGFBP4	rs10305281	35854525	TAF11	rs4646912	34964538
MLL	rs9332811	117867064	REV3L	rs17510914	111801601
MLL	rs9332780	117851777	REV3L	rs17510485	111874469
CCNB2	rs28383493	57184700	REV3L	rs17540138	111743699
RAD17	rs17236198	68703997	EDNRA	rs10305874	148636071
FGFR3	rs3135891	1776372	EDNRA	rs10305882	148641981
FGFR3	rs3135837	1766073	EDNRA	rs10305891	148653972
NEIL2	rs8191662	11680702	EDNRA	rs10305873	148628144
EGFR	rs17336905	55200059	EDNRA	rs10305877	148636750
EGFR	rs17337037	55207181	XRCC3	rs3212073	103240944
EGFR	rs17289260	55120684	CCNI	rs4252906	78195123
EGFR	rs17290538	55226523	CCNI	rs4252822	78206864
EGF	rs11569137	111150901	CCNI	rs4252941	78189857
MCM2	rs17538530	128810869	JUNB	rs17887128	12763196
BNIP1	rs5745129	172511519	CTNND1	rs11570177	57302097
ESR1	rs9340797	152204717	CTNND1	rs11570227	57340387
ESR1	rs9340862	152252435	NR3C1	rs10482677	142661078
ESR1	rs9340996	152384481	NR3C1	rs10482624	142748328
ESR1	rs9340992	152384082			

Table B.6: Comparison of mean square prediction error (MSPE) and coverage proportion (COV) for different methods applied to molecular epidemiology data.

Metric	CTF	RF	QRF	BART
MSPE	0.263	0.353	-	0.425
95% Coverage	0.961	-	0.928	0.817

B.2 Figures

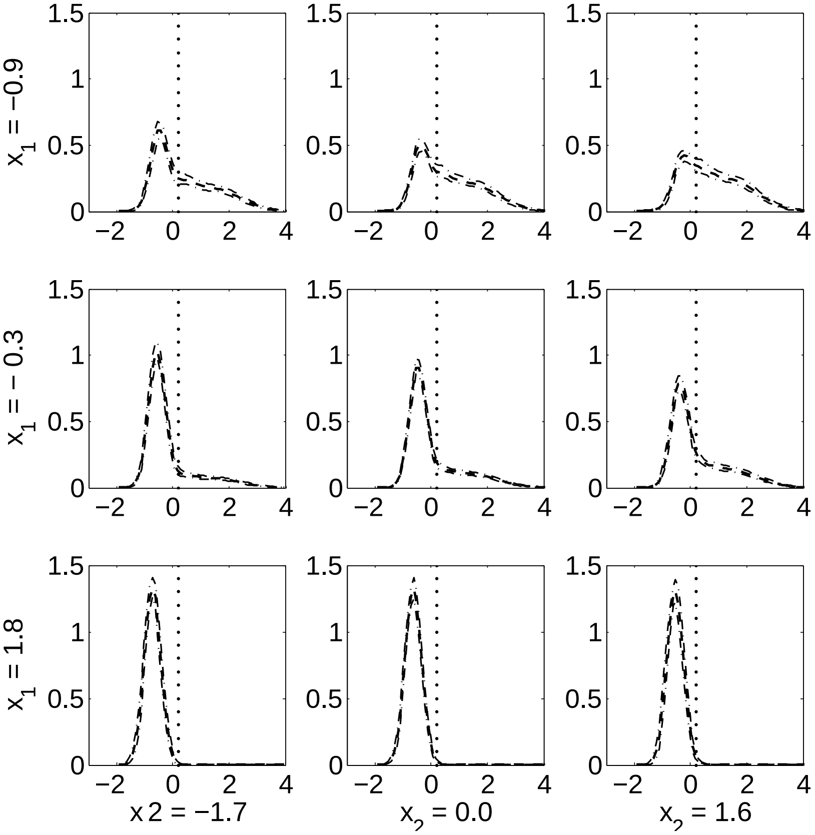

Figure B.1: Conditional densities with similar means but predictor-dependent higher moments; (Chung and Dunson 2009)

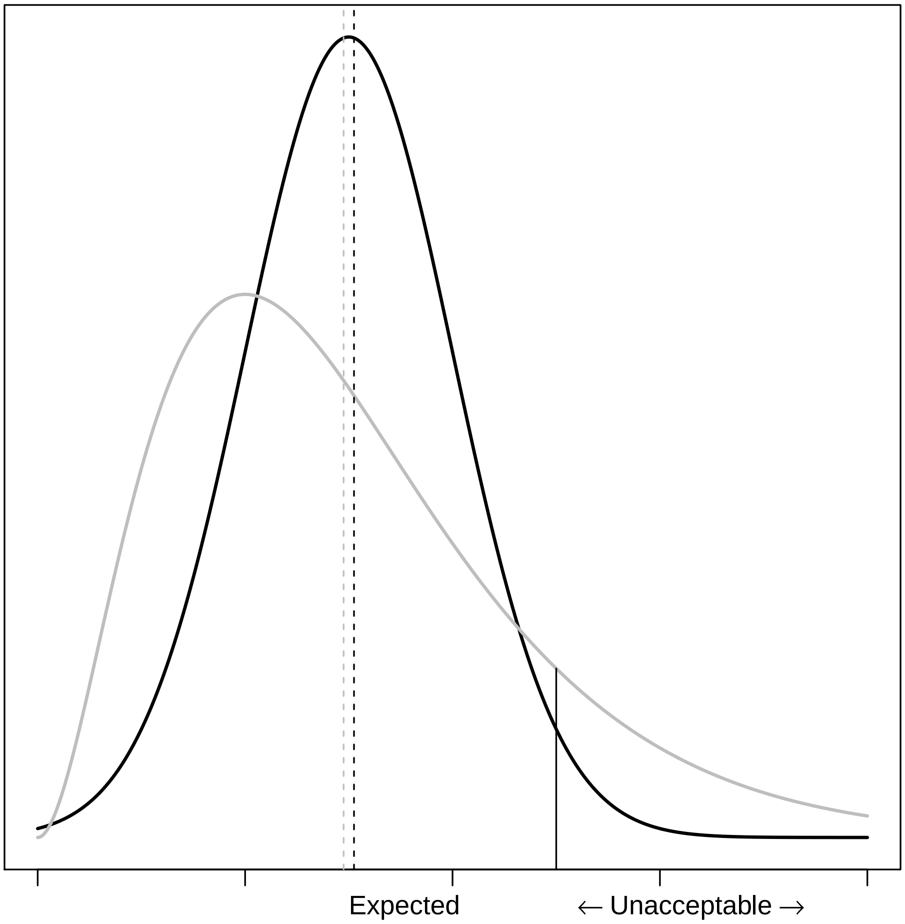

Figure B.2: Hypothetical scenario of conditional distributions with the same expected value but different tail probabilities

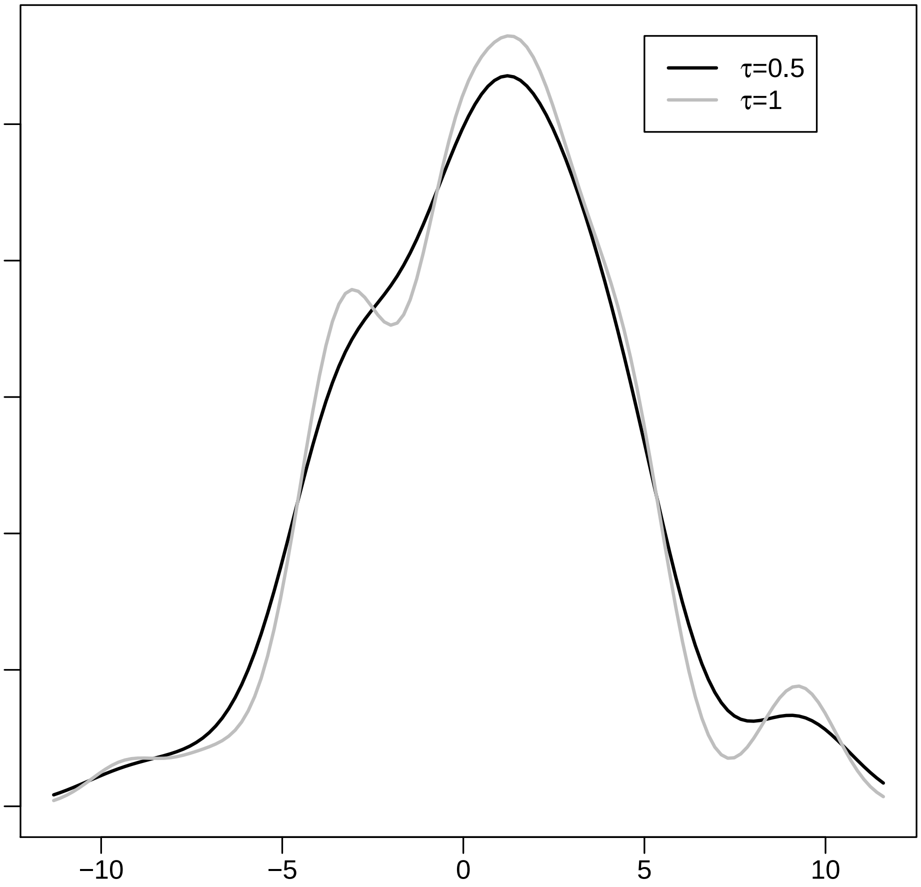

Figure B.3: Simulation study density; three underlying predictors interact to produce a complex population density. Separate lines indicate different assumed residual precisions.

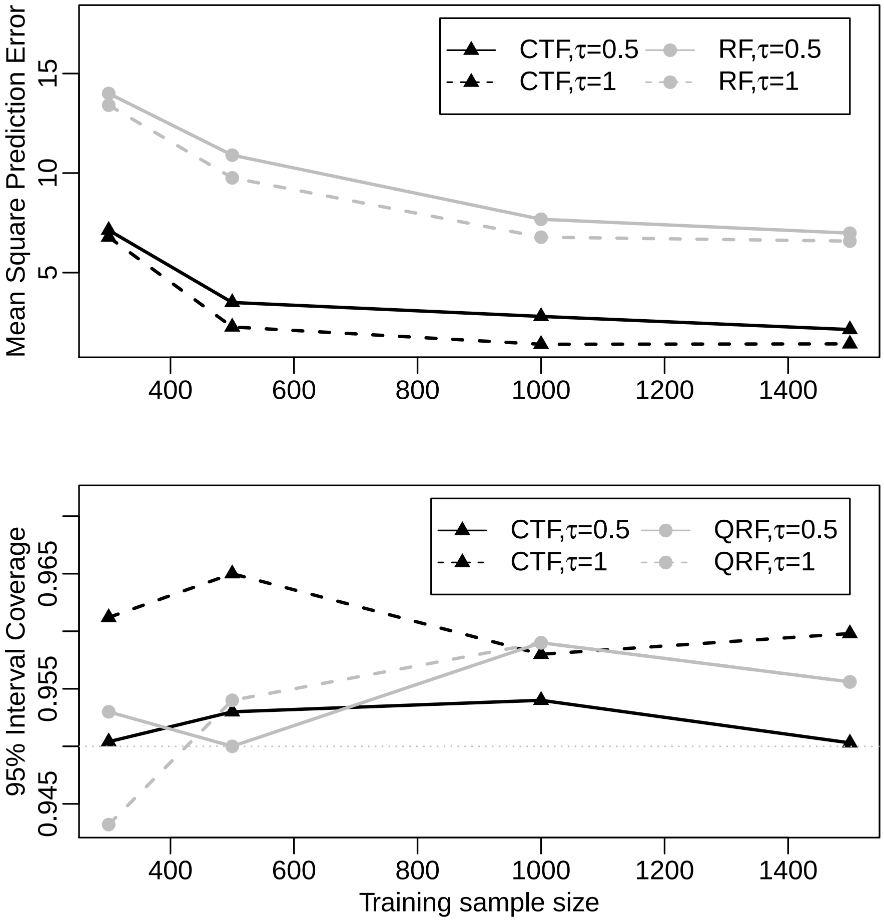

Figure B.4: Simulation study results, demonstrating the improved performance of the CTF in mean square prediction error (MSPE) relative to random forests and comparable performance in coverage (COV) relative to quantile regression random forests.

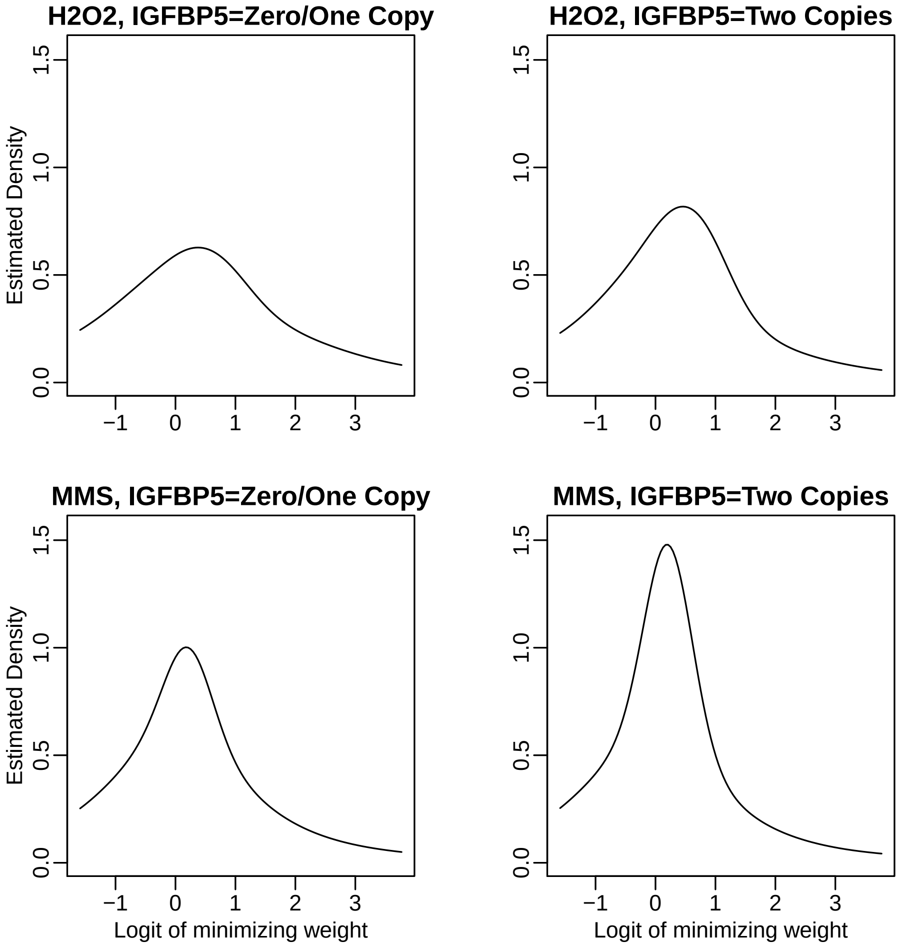

Figure B.5: Selected conditional densities for CTF estimated model of molecular epidemiology data, varying the exposure and the number of copies of the dominant allele at the IGFBP5 SNP. All other SNPs are held at the “Zero/One Copy” level.

Appendix C: Chapter 4

C.1 Tables

Table C.1: Predictors selected in analysis of full data set, presented in inclusion probability order.

Predictor
Chromosome 5, SNP RS6894946
Age at MRI (discretized)
Chromosome 18, SNP RS11877713
Chromosome 15, SNP RS16954106
Chromosome 9, SNP RS10981955

C.2 Figures

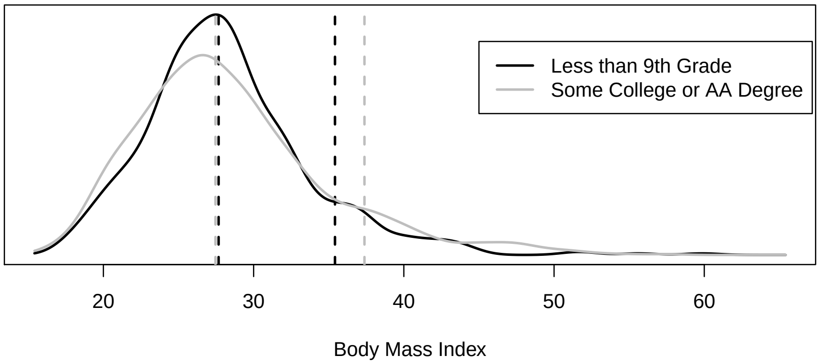

Figure C.1: Body Mass Index (BMI) data from NHANES 2001-2002. The black curve indicates the empirical BMI density for adults with less than a 9^{th} grade education; the grey curve is for adults with some college or an associate's degree. Vertical lines indicate 50^{th} and 90^{th} percentiles for the separate populations; the 50^{th} percentile for the two populations is very similar.

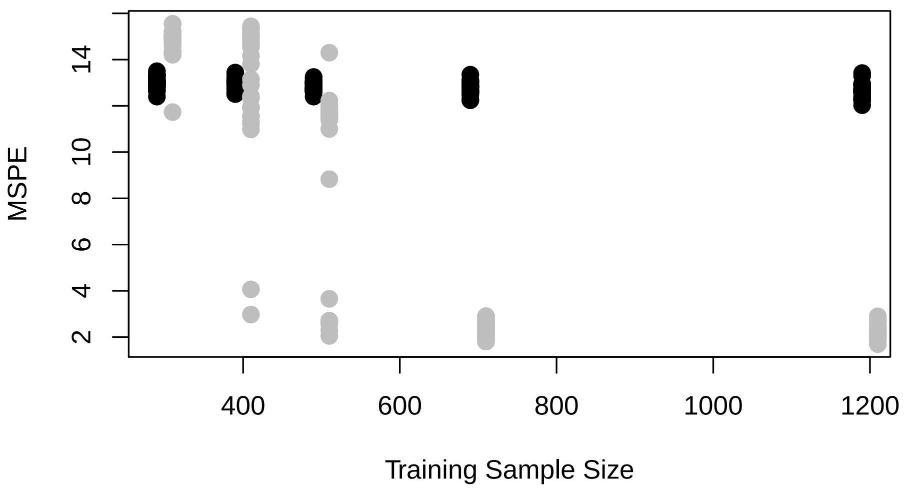

Figure C.2: Summary of simulation study results. Grey dots indicate mean square prediction error (MSPE) for the CTFC method. Black dots indicate mean square prediction error for the Lasso.

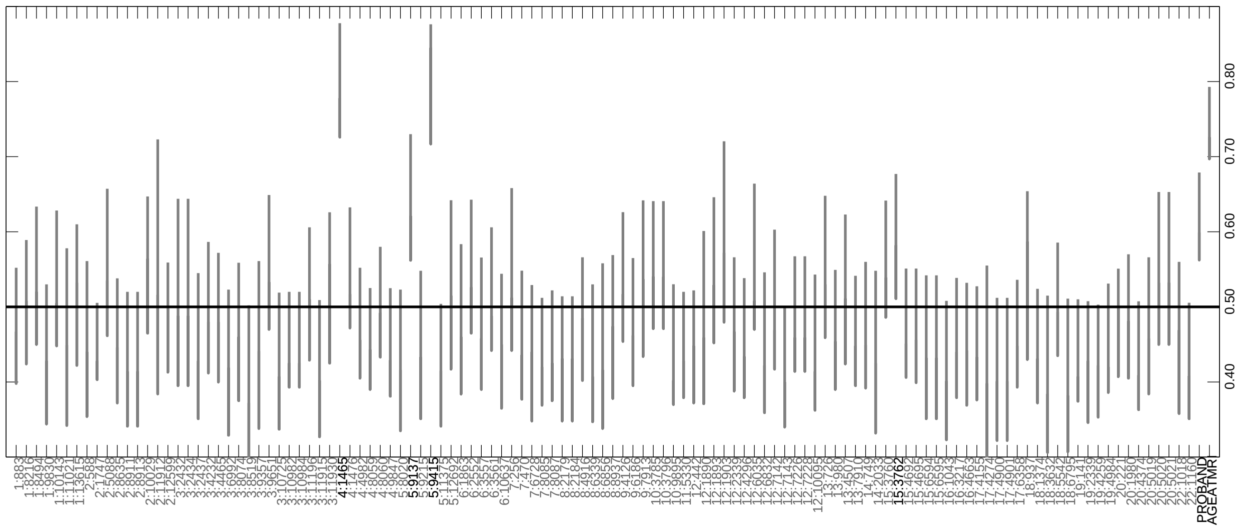

Figure C.3: First-pass inclusion probabilities for predictors selected in at least one leave-out scenario. Heavy gray vertical lines indicate the range of inclusion probabilities across leave-out sets. Predictors selected in all leave-out scenarios are indicated with darker labels on the horizontal axis.

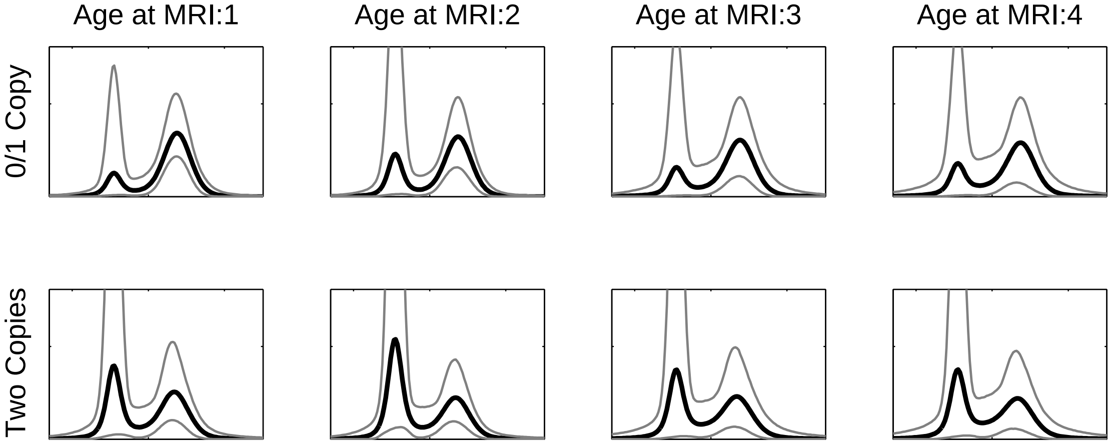

Figure C.4: Conditional densities with varying levels of RS6894946 and age at MRI. Rows indicate different levels of RS6894946 (zero/one copy or two copies of the major allele), and columns indicate different levels of the age at MRI variable. Heavy black lines indicate posterior mean, light grey lines indicate 95% credible intervals.

Bibliography

Antoniak, C. A. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

Besag, J., Green, P., Higdon, D., and Mengersen, K. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–41, Feb 1995. ISSN 0883-4237. doi: 10.1214/ss/1177010123.

Bhattacharya, A. and Dunson, D. B. Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107(497):362–377, Mar 2012. ISSN 0162-1459. doi: 10.1080/01621459.2011.646934.

Bishop, C. and Svensén, M. Bayesian hierarchical mixtures of experts. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, pages 57–64, 2003.

Blackwell, D. and MacQueen, J. Ferguson distributions via Polya urn schemes. *Annals of Statistics*, 1(2):353–355, 1973. ISSN 0090-5364. doi: 10.1214/aos/1176342372.

Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Bush, C. and MacEachern, S. A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, JUN 1996. ISSN 0006-3444. doi: 10.1093/biomet/83.2.275.

Bush, C., Lee, J., and MacEachern, S. Minimally informative prior distributions for non-parametric Bayesian analysis. *Journal of the Royal Statistical Society B*, 72(2):253–268, 2010.

Chen, X., Liu, C.-T., Zhang, M., and Zhang, H. A forest-based approach to identifying gene and gene-gene interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19199–19203, DEC 4 2007. ISSN 0027-8424. doi: 10.1073/pnas.0709868104.

Chipman, H., George, E., and McCulloch, R. Bayesian ensemble learning. In *Advances in Neural Information Processing Systems*, pages 265–272. MIT Press, 2006.

Chipman, H., George, E., and McCulloch, R. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298, 2010.

Chu, W. and Ghahramani, Z. Probabilistic models for incomplete multi-dimensional arrays. *Proceedings of the* 12th *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

Chung, Y. and Dunson, D. Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488):1646–1660, 2009.

Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, JUN 2009. ISSN 1471-0056. doi: 10.1038/nrg2579.

Cowell, R., Dawid, A., Lauritzen, S., and Spiegelhalter, D. *Probabilistic Networks and Expert Systems*. Springer, 1999.

Csiszar, I. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975. ISSN 0091-1798. doi: 10.1214/aop/1176996454.

Dawid, A. and Lauritzen, S. Hyper Markov laws in the statistical-analysis of decomposable graphical models. *Annals of Statistics*, 21(3):1272–1317, SEP 1993. ISSN 0090-5364. doi: 10.1214/aos/1176349260.

De Iorio, M., Muller, P., Rosner, G., and MacEachern, S. An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215, 2004.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Dobra, A. and Lenkoski, A. Copula Gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics*, 5(2A):969–993, Jun 2011. ISSN 1932-6157. doi: 10.1214/10-AOAS397.

Dobra, A. and Massam, H. The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Statistical Methodology*, 7(3):240 – 253, 2010. ISSN 1572-3127. doi: 10.1016/j.stamet.2009.04.002. URL `http://www.sciencedirect.com/science/article/pii/S1572312709000215`.

Dunson, D. and Park, J. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

Dunson, D. and Xing, C. Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association (Theory and Methods)*, 104(487):1042–1051, 2009.

Eilers, P. and de Menezes, R. Quantile smoothing of array CGH data. *Bioinformatics*, 21 (7):1146–1153, APR 1 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti148.

Escobar, M. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

Escobar, M. and West, M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

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

Ferguson, T. S. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4):615–629, 1974.

Ferguson, T. S. Bayesian density estimation by mixtures of normal distributions. In Chernoff, H., Rizvi, M. H., Rustagi, J. S., and Siegmund, D., editors, *Recent Advances in Statistics: Papers in honor of Herman Chernoff on his sixtieth birthday*, pages 287–302. Academic Press, New York, 1983.

Fienberg, S. E. and Rinaldo, A. Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference*, 137(11):3430–3445, Nov 2007. ISSN 0378-3758. doi: 10.1016/j.jspi.2007.03.022.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

George, E. and McCulloch, R. Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

Goodman, L. How to ransack social mobility tables and other kinds of cross-classification tables. *American Journal of Sociology*, pages 1–40, 1969.

Griffin, J. and Steel, M. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.

Hannah, L., Blei, D., and Powell, W. B. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12:1923–1953, 2011.

Hayes, B. J., Visscher, P. M., and Goddard, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(1):47–60, FEB 2009. ISSN 0016-6723. doi: 10.1017/S0016672308009981.

Ho, J. W. K., Stefani, M., dos Remedios, C. G., and Charleston, M. A. A model selection approach to discover age-dependent gene expression patterns using quantile regression models. *BMC Genomics*, 10, SEP 2009. ISSN 1471-2164. doi: 10.1186/1471-2164-10-S3-S16. 8th International Conference on Bioinformatics, Biopolis, SINGAPORE, SEP 07-11, 2009.

Hoerl, A. and Kennard, R. Ridge regression - biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–&, 1970. ISSN 0040-1706. doi: 10.2307/1267351.

Hoff, P. Hierarchical multilinear models for multiway data. *Computational Statistics and Data Analysis*, 55:530–543, 2011.

Hoffmann-Jørgensen, J. Existence of conditional probabilities. *Math. Scand.*, 28:257–264 (1972), 1971. ISSN 0025-5521.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. *PLOS Genetics*, 4(7), JUL 2008. ISSN 1553-7390. doi: 10.1371/journal.pgen.1000130.

Hoh, J., Wille, A., Zee, R., Cheng, S., Reynolds, R., Lindpaintner, K., and Ott, J. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *ANNALS OF HUMAN GENETICS*, 64(5):413–417, SEP 2000. ISSN 0003-4800. doi: 10.1046/j.1469-1809.2000.6450413.x.

Ishwaran, H. and James, L. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association (Theory and Methods)*, 96(453):161–173, 2001.

Jara, A. and Hanson, T. A class of mixtures of dependent tail-free processes. *Biometrika*, 98(3):553–566, 2011.

Jordan, M. and Jacobs, R. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

Kalli, M., Griffin, J., and Walker, S. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

Kass, R. E. and Wasserman, L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90 (431):928–934, 1995. ISSN 0162-1459.

Kleinman, K. and Ibrahim, J. A semiparametric Bayesian approach to the random effects model. *Biometrics*, 54(3):921–938, SEP 1998a. ISSN 0006-341X. doi: 10.2307/2533846.

Kleinman, K. and Ibrahim, J. A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine*, 17(22):2579–2596, NOV 30 1998b. ISSN 0277-6715. doi: 10.1002/(SICI)1097-0258(19981130)17:22<2579::AID-SIM948>3.0.CO;2-P.

Koenker, R. and Bassett, G. regression quantiles. *Econometrica*, 46(1):33–50, 1978. ISSN 0012-9682. doi: 10.2307/1913643.

Langseth, H., Nielsen, T., Rumí, R., and Salmerón, A. Inference in hybrid Bayesian networks with mixtures of truncated basis functions. In *Sixth European Workshop on Probabilistic Graphical Models*, pages 171–178, 2012.

Lauritzen, S. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, DEC 1992. ISSN 0162-1459. doi: 10.2307/2290647.

Lijoi, A. and Regazzini, E. Means of a Dirichlet process and multiple hypergeometric functions. *The Annals of Probability*, 32(2):1469–1495, 2004.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. FaST linear mixed models for genome–wide association studies. *Nature Methods*, 8(10):833–835, 2011.

Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., and Heckerman, D. Improved linear mixed models for genome–wide association studies. *Nature Methods*, 9 (6):525–526, 2012.

Lo, A. On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12(1):351–357, 1984. ISSN 0090-5364. doi: 10.1214/aos/1176346412.

Lou, X.-Y., Chen, G.-B., Yan, L., Ma, J. Z., Zhu, J., Elston, R. C., and Li, M. D. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *American Journal of Human Genetics*, 80(6):1125–1137, JUN 2007. ISSN 0002-9297. doi: 10.1086/518312.

MacEachern, S. N. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics - Simulations*, 23:727–741, 1994.

MacEachern, S. N. and Müller, P. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.

Marchini, J., Donnelly, P., and Cardon, L. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417, APR 2005. ISSN 1061-4036. doi: 10.1038/ng1537.

Markunas, C. A., Soldano, K., Dunlap, K., Cope, H., Asiimwe, E., Stajich, J., Enterline, D., Grant, G., Fuchs, H., Gregory, S. G., and Ashley-Koch, A. E. Stratified Whole Genome Linkage Analysis of Chiari Type I Malformation Implicates Known Klippel-Feil Syndrome Genes as Putative Disease Candidates. *PLOS ONE*, 8(4), APR 19 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0061521.

Massam, H., Liu, J., and Dobra, A. A conjugate prior for discrete hierarchical log-linear models. *Annals of Statistics*, 37(6A):3431–3467, DEC 2009. ISSN 0090-5364. doi: 10.1214/08-AOS669.

McAuliffe, J., Blei, D., and Jordan, M. Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing*, 16(1):5–14, 2006.

Meinshausen, N. Quantile regression forests. *Journal of Machine Learning Research*, 7: 983–999, 2006.

Moala, F. and O'Hagan, A. Elicitation of multivariate prior distributions: A nonparametric Bayesian approach. *Journal of Statistical Planning and Inference*, 140(7):1635–1655, 2010.

Moral, S., Rumí, R., and Salmerón, A. Mixtures of truncated exponentials in hybrid Bayesian networks. In *Proceedings of the Sixth European Conference on Symbolic and Quantitative Approaches to Reasoning under Uncertainty (ECSQARU 2001*, pages 156–167, 2001.

Muller, P., A., E., and West, M. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79, 1996.

Neal, R. Regression and classification using gaussian process priors. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics*, volume 6, pages 475–501. Oxford University Press, 1999.

Olive, P., Wlodek, D., and Banath, J. DNA double-strand breaks measured in individual cells subjected to gel electrophoresis. *Cancer Research*, 51(17):4671–4676, 1991.

Papaspiliopoulos, O. and Roberts, G. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

Pearl, J. *Probabilistic reasoning in intelligent systems : networks of plausible inference.* Morgan Kaufmann, 1999.

Petrone, S., Guindani, M., and Gelfand, A. Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society B*, 71(4):755–782, 2009.

Pfeiffer, R. M., Gail, M. H., and Pee, D. Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika*, 88:933–948, 2001.

Pitman, J. Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, probability and game theory*, volume 30 of *IMS Lecture Notes Monogr. Ser.*, pages 245–267. Inst. Math. Statist., Hayward, CA, 1996. doi: 10.1214/lnms/1215453576. URL `http://dx.doi.org/10.1214/lnms/1215453576`.

Rakitsch, B., Lippert, C., Stegle, O., and Borgwardt, K. A Lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214, JAN 15 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts669.

Reich, B. and Fuentes, M. A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Annals of Applied Statistics*, 1(1):249–264, 2007.

Reich, B., Bondell, H., and Li, L. Sufficient dimension reduction via Bayesian mixture modeling. *Biometrics*, 67:886–895, 2011.

Rodriguez, A., Dunson, D., and Taylor, J. Bayesian hierarchically weighted finite mixture models for samples of distributions. *Biostatistics*, 10(1):155–171, 2009.

Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Shahbaba, B. and Neal, R. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.

Spielman, R. and Ewens, W. The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics*, 59(5):983–989, NOV 1996. ISSN 0002-9297.

Sudderth, E. and Jordan, M. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Proceedings of Neural Information Processing Systems*, 2008.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological*, 58(1):267–288, 1996. ISSN 0035-9246.

Tokdar, S., Zhuy, Y., and Ghosh, J. Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Analysis*, 5(2):319–344, 2010.

Tucker, L. Some mathematical notes on 3-mode factor analysis. *Psychometrika*, 31(3):279, 1966. ISSN 0033-3123. doi: 10.1007/BF02289464.

Walker, S. Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36:45–54, 2007.

Wang, L. and Dunson, D. B. Fast Bayesian Inference in Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, MAR 2011. ISSN 1061-8600. doi: 10.1198/jcgs.2010.07081.

Waterhouse, S., MacKay, D., and Robinson, T. Bayesian methods for mixtures of experts. In *Advances in Neural Information Processing Systems*, pages 351–357. MIT Press, 1996.

Xu, Z., Yan, F., and Qi, Y. Infinite Tucker decomposition: Nonparametric Bayesian models for multiway data analysis. *Proceedings of the* 29th *International Conference on Machine Learning*, 2012.

Yamato, H. Characteristic functions of means of distributions chosen from a Dirichlet process. *Ann. Probab.*, 12(1):262–267, 1984. ISSN 0091-1798.

Yang, Y. and Dunson, D. Bayesian conditional tensor factorizations for high-dimensional classification. Submitted to JRSS-B, 2012. URL `http://isds.duke.edu/research/papers/2012-12`.

Yau, C., Papaspiliopoulos, O., Roberts, G., and Holmes, C. Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society (B)*, 73(1):37–57, 2011.

Yi, N. Statistical analysis of genetic interactions. *Genetics Research*, 92(5-6):443–459, OCT-DEC 2010. ISSN 0016-6723. doi: 10.1017/S0016672310000595.

Zhang, L., Mukherjee, B., Hu, B., Moreno, V., and Cooney, K. A. Semiparametric Bayesian modeling of random genetic effects in family-based association studies. *Statistics in Medicine*, 28:113–139, 2009.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., and Buckler, E. S. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–U118, APR 2010. ISSN 1061-4036. doi: 10.1038/ng.546.

Zou, F., Huang, H., Lee, S., and Hoeschele, I. Nonparametric Bayesian variable selection with applications to multiple quantitative trait loci mapping with epistasis and gene-environment interaction. *Genetics*, 186(1):385–394, 2010.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B-Methodological*, 67(Part 2):301–320, 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x.