Some Recent Advances in Non- and Semiparametric Bayesian Modeling with Copulas, Mixtures, and Latent Variables

by

Jared S. Murray

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
David B. Dunson

_____
Fan Li

_____
Vincent Joseph Hotz

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2013

Abstract

Some Recent Advances in Non- and Semiparametric Bayesian Modeling with Copulas, Mixtures, and Latent Variables

by

Jared S. Murray

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
David B. Dunson

_____
Fan Li

_____
Vincent Joseph Hotz

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2013

Abstract

This thesis develops flexible non- and semiparametric Bayesian models for mixed continuous, ordered and unordered categorical data. These methods have a range of possible applications; the applications considered in this thesis are drawn primarily from the social sciences, where multivariate, heterogeneous datasets with complex dependence and missing observations are the norm.

The first contribution is an extension of the Gaussian factor model to Gaussian copula factor models, which accommodate continuous and ordinal data with unspecified marginal distributions. I describe how this model is the most natural extension of the Gaussian factor model, preserving its essential dependence structure and the interpretability of factor loadings and the latent variables. I adopt an approximate likelihood for posterior inference and prove that, if the Gaussian copula model is true, the approximate posterior distribution of the copula correlation matrix asymptotically converges to the correct parameter under nearly any marginal distributions. I demonstrate with simulations that this method is both robust and efficient, and illustrate its use in an application from political science.

The second contribution is a novel nonparametric hierarchical mixture model for continuous, ordered and unordered categorical data. The model includes a hierarchical prior used to couple component indices of two separate models, which are also linked by local multivariate regressions. This structure effectively overcomes the limitations of existing mixture models for mixed data, namely the overly strong

local independence assumptions. In the proposed model local independence is replaced by local conditional independence, so that the induced model is able to more readily adapt to structure in the data. I demonstrate the utility of this model as a default engine for multiple imputation of mixed data in a large repeated-sampling study using data from the Survey of Income and Participation. I show that it improves substantially on its most popular competitor, multiple imputation by chained equations (MICE), while enjoying certain theoretical properties that MICE lacks.

The third contribution is a latent variable model for density regression. Most existing density regression models are quite flexible but somewhat cumbersome to specify and fit, particularly when the regressors are a combination of continuous and categorical variables. The majority of these methods rely on extensions of infinite discrete mixture models to incorporate covariate dependence in mixture weights, atoms or both. I take a fundamentally different approach, introducing a continuous latent variable which depends on covariates through a parametric regression. In turn, the observed response depends on the latent variable through an unknown function. I demonstrate that a spline prior for the unknown function is quite effective relative to Dirichlet Process mixture models in density estimation settings (i.e., without covariates) even though these Dirichlet process mixtures have better theoretical properties asymptotically. The spline formulation enjoys a number of computational advantages over more flexible priors on functions. Finally, I demonstrate the utility of this model in regression applications using a dataset on U.S. wages from the Census Bureau, where I estimate the return to schooling as a smooth function of the quantile index.

To my family, and to Carolyn, the love of my life, so she can finally see what it's all been for.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

Φ	Normal cumulative distribution function
$\overset{iid}{\sim}$	Independent and identically distributed

Abbreviations

DP	Dirichlet Process
MICE	Multiple Imputation by Chained Equations
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
PMM	Predictive Mean Matching
GLM	Generalized Linear Model
GDP	Generalized Double Pareto (distribution)
PX	Parameter Expansion/Expanded
GCFM	Gaussian Copula Factor Model
HPD	Highest Posterior Density (interval)
SIPP	Survey of Income and Program Participation
HCMM-LD	Hierarchically Coupled Mixture Model with Local Dependence
MPMN	Mixture of Product Multinomials
ITF	Infinite Tensor Factorization

Acknowledgements

I would like to thank everyone in the Department of Statistical Science, which is a collegial, nurturing environment unlike any I have ever been in before. Many thanks go to my advisor, Jerry Reiter, for his support and guidance in my research endeavors and professional development, and for somehow always finding the time to read drafts, troubleshoot, or brainstorm. He is the model of a mentor. Thanks to Joe Lucas for taking me on in my early years, for showing me the ropes and encouraging me. I also want to thank David Dunson for thoughtful discussion and comments, and class projects that somehow become peer reviewed papers.

I owe a tremendous debt of gratitude to my fellow students at DSS, past and present. Special thanks are due to my sometime officemates and good friends Chris Challis and David McClure, who brought small but ferocious dinosaurs and a peculiarly sunny disposition (respectively) to wherever we found ourselves. Thanks also to Richard Hahn, the preeminent "helpful discussant" and an invaluable source for post-Ph.D. advice.

I am immensely grateful to my family for their support; my parents, my sister, and the niece or nephew I haven't met yet. And also to my old friends, too many to name, who I speak with too infrequently but think of often and fondly, especially in trying times. Finally, my deepest and most heartfelt thanks to my future wife Carolyn, who has been everything anyone could hope for in a companion, and so much more.

1

Introduction

Many scientific problems require flexible multivariate probability models. For example, a survey or test might ask several questions designed to measure some latent trait or ability. When datasets have partially or completely missing observations, confidence or credible intervals should be widened to appropriately reflect uncertainty. This can be achieved by integrating over the distribution of missing values given the observed data. Similarly, the dependence structure of a collection of variables in a finite population can be estimated by "filling in" the unsampled units with draws from the predictive distribution, given the sampled observations.

Flexible modeling is challenging, and multivariate modeling especially so - as the dimension increases the information contained in a finite sample decreases rapidly. But fully parametric models make restrictive assumptions about the joint distribution and its dependence structure that are often violated in real data. Hence there is a need for models which are flexible but carefully structured. In this thesis I present three such models: (1) A semiparametric Gaussian copula factor model, which assumes parametric dependence structure with nonparametric marginal distributions, (2) A hierarchically coupled mixture model with local dependence, carefully spec-

ified to capture complex dependence in mixed continuous, ordered and unordered categorical data, and (3) A latent variable density regression model, which provides a particularly tractable alternative to dependent mixture models. These methods are motivated by applications in the social sciences, but are appropriate in a wide range of applied settings.

This thesis is organized in four chapters: In Chapter 2 I extend Gaussian factor models to the case of mixed continuous, ordinal and count variables. Gaussian factor models provide a lower-dimensional representation of the covariance matrix of the data. When the outcomes are multivariate normal this decomposition of the covariance matrix summarizes all the dependence in the data. This is generally not the case for mixed data. I describe the Gaussian copula factor model, provide new theoretical results on the consistency of estimation under an approximate likelihood, and introduce new default priors for factor analysis with non-Gaussian variables. I develop efficient parameter-expanded inference and apply the model to a dataset from political science. This chapter closely follows my paper with David Dunson, Larry Carin, and Joe Lucas which appeared as Murray et al. (2013).

In Chapter 3 I present more flexible joint models for mixed data. The Gaussian copula factor model is semiparametric in that the marginal distributions need not be specified, but the use of a Gaussian copula is itself a substantial parametric assumption. Further, the Gaussian copula factor model is not trivially extended to unordered categorical variables. To address these limitations I introduce a novel hierarchical mixture model which is carefully structured to capture dependence within and between continuous and categorical data. I illustrate with a simulation study on genuine survey data, showing that this model can be superior to the most popular competitor in multiple imputation tasks.

In Chapter 4 I introduce a density regression model for a continuous dependent variable based on a model with continuous latent variables. Unlike most existing

density regression methods I do not rely on discrete mixture models with dependent atoms and/or weights. This drastically reduces the number of parameters in the model without sacrificing much flexibility, and the continuous nature of the model enables the use of sophisticated Monte Carlo techniques for posterior inference. I demonstrate with simulations and an application to U.S. wage data, and discuss extensions to particular covariate spaces and to multivariate models.

In the remainder of this chapter I provide some necessary background material on existing joint modeling techniques, on Bayesian nonparametric modeling with mixture models, and on multiple imputation.

1.1 A Review of Joint Models for Mixed Data

1.1.1 Factor models

Factor models induce dependence in multivariate data y_i by integrating with respect to latent variables following some distribution. The simplest example is the linear factor model, which takes the form

$$y_i = \Lambda\eta_i + \epsilon_i \tag{1.1}$$

where y_i is a $p \times 1$ vector of observed variables, Λ is a $p \times k$ matrix of factor loadings ($k < p$), η_i is a length k vector of latent variables and the vector ϵ_i contains uncorrelated errors or disturbances. I assume $E(y_i) = 0$ without any loss of generality. Marginalizing over the latent factors yields $Cov(y_i) = \Lambda\Lambda' + \Sigma$. If $k > p - 1$ any covariance matrix may be represented in this fashion, but typically $k < p$ and in modern high-dimensional applications k may be orders of magnitude less than p, providing regularized estimates of high-dimensional covariance matrices. Factor models date at to at least Spearman (1904) and have a rich and complex history with key advances taking place in different fields. Bartholomew et al. (2011) (chap. 1) give a comprehensive history of these developments.

The Gaussian factor model further assumes that $\eta_i \sim N(0, I)$ and $\epsilon_i \sim N(0, \Sigma)$ with $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. To extend this model to mixed continuous and ordered categorical data, Muthen (1984) proposed relating the observed discrete responses to latent continuous variables by thresholding:

$$z_i = \mu + \Lambda\eta_i + \epsilon_i$$

$$y_{ij} = \begin{cases} z_{ij}, & y_{ij} \text{ is continuous} \\ \sum_{c=1}^{C_j} c\mathbf{1}(\gamma_{jc-1} < z_{ij} \leq \gamma_{jc}), & y_{ij} \text{ is discrete} \end{cases}$$

where C_j is the number of levels for j when it is categorical and each γ_{jc} is a cutpoint with $\gamma_{j0} = -\infty$, $\gamma_{j1} = 0$ and $\gamma_{jc_j} = \infty$. Muthen (1984) used a two-stage procedure for maximum likelihood estimation. The seminal reference for full posterior inference via MCMC in multivariate probit models is Albert and Chib (1993), with further improvements by Cowles (1996) who gave a blocked Metropolis Hastings update for the threshold parameters. In Chapter 2 I describe a novel parameter expansion MCMC scheme that provides still further improvements.

Moustaki and Knott (2000) considered a broader class of models, known as generalized latent trait models, where each manifest variable follows a distribution in the exponential family (see also Sammel et al. (1997); Dunson (2003, 2000)). The factor structure enters through the linear predictor:

$$E(y_{ij} \mid \eta_i) = g_j^{-1}(\mu_j + \lambda_j\eta_i), \tag{1.2}$$

where g_j is the link function for variable j, with the Gaussian factor model recovered by taking g as the identity and the thresholded formulation of Muthen (1984) corresponds to the probit link. In general marginalizing out η_i will induce marginal means and variances that depend on λ_j, complicating the interpretation of both the factor loadings and scores compared to the Gaussian factor model.

One may generalize the Gaussian factor model in another direction by assuming that

$$z_i = \Lambda\eta_i + \epsilon_i, \quad y_{ij} = f_j(z_{ij}) \tag{1.3}$$

where each f_j is a monotone function. This is the semiparametric Gaussian copula factor model I introduce in Chapter 2. This formulation recovers the model proposed by Muthen (1984) when the f_j are step functions for discrete manifest variables and the identity for continuous manifest variables. Allowing the f_j to range over all monotone functions makes this model substantially more general, however. I discuss this and related semiparametric approaches to factor models in detail in Chapter 2.

1.1.2 Mixture and latent class models

Mixture models decompose the density of a random variable Y into a convex combination of k normalized kernels $f(y;\theta_h)$, i.e.

$$g(y;\theta) = \sum_{h=1}^{k} \pi_h f(y;\theta_h) \tag{1.4}$$

where $0 \leq \pi \leq 1$ and $\sum_{h=1}^{k} \pi_h = 1$. A recent comprehensive overview of of mixture modeling appears in McLachlan and Peel (2000); see also Marin et al. (2005) for a review focused specifically on Bayesian inference. Mixture models may be rewritten hierarchically by introducing component assignment indicators

$$\Pr(H_i = h) = \pi_h, \ 1 \leq h \leq k \tag{1.5}$$

$$Y_i \sim F_{\theta_h}. \tag{1.6}$$

This representation is useful for computing maximum likelihood estimates via expectation-maximization or implementing posterior inference via Gibbs sampling in a fully Bayesian approach.

With suitable choices for $f(\cdot;\theta)$ and enough components k a mixture model can approximate a wide variety of true densities. Not all specifications will perform

equally well. For example, if the true density has heavy tails and the kernel does not the approximation tends to be poor. If the density is rough or multimodal and k is small the mixture model will also tend to be a poor fit. Hence appropriate values for k will depend on the kernel, and vice-versa. Fully Bayesian approaches to mixture modeling typically fix a kernel f and either place a prior on k, treating it as a parameter to be estimated (e.g. Richardson and Green (1997)) or choose a large value for k and use a prior on π that favors giving most components low mixture weights (Ishwaran and James, 2001; Ishwaran and Zarepour, 2002). The latter approach can often be cast as an approximation to an infinite dimensional mixture model ($k = \infty$). I will discuss such models in the next section and in later chapters.

For multivariate data it can be difficult to select an appropriate kernel. When $y \in \mathbb{R}^p$ the multivariate normal with component-specific mean and covariance parameters is a popular and convenient approach. When y includes mixed continuous, categorical and other data types, there is no obvious choice for f. One candidate is a product kernel, i.e. if y is p dimensional, $f(y;\theta) = \prod_{j=1}^{p} f_j(y_j;\theta^{(j)})$. The equivalent augmented model is

$$\Pr(H_i = h) = \pi_h,\ 1 \leq h \leq k \tag{1.7}$$

$$Y_{ij} \sim F_j(\theta_h),\ \text{independently.} \tag{1.8}$$

With a product kernel the outcomes are mutually independent given the component index, and dependence is induced by marginalizing out the component index (similar to the factor model, where the latent variables were continuous). Mixtures with product kernels are also known as latent class models. Often in latent class modeling the components correspond to substantively meaningful latent populations. However, in this thesis I instead use mixtures as flexible and parsimonious models for multivariate distributions. In this context forcing all the dependence in a multi-

variate distribution to be represented through a single latent cluster index is overly restrictive. In Chapter 3 I develop a new class of models that retains much of the computational simplicity of product kernel mixtures but relaxes their strong conditional independence assumptions.

1.2 Nonparametric Bayes and the Dirichlet Process

Nonparametric Bayesian models are usually defined as those which model infinite dimensional quantities, such as probability densities, stochastic processes, or functions. A distinguishing characteristic of nonparametric Bayesian models is that they tend to have a data-adaptive quality, in that model complexity is allowed to grow with the sample size "automatically" (that is, in a manner determined by the prior). This adaptive modeling approach is also useful in contexts which involve very high but finite dimensional objects such as large probability mass functions for contingency tables or huge covariance matrices. Here I provide a brief introduction to the Dirichlet process, which is most relevant to the methods in this thesis and has inspired much of the recent developments in nonparametric Bayesian modeling more generally.

1.2.1 The Dirichlet Process

The Dirichlet process is a distribution over probability measures parameterized by H, the base probability measure, and α, the concentration. We write $P \sim DP(\alpha, H)$ to indicate that P follows a DP. Ferguson (1974) characterizes the DP as $P \sim DP(\alpha, H)$ if and only if for any measurable partition $\{B_1, \ldots B_k\}$ of the sample space the random vector $(P(B_1), P(B_2), \ldots P(B_k)) \sim Dir(\alpha H(B_1), \alpha H(B_2), \ldots, \alpha H(B_k))$. Hence the prior mean of $P(A)$ for any measureable set A is simply $P(A)$ with variance decreasing as α increases.

A constructive definition of the Dirichlet process was provided by Sethuraman

(1994) who showed that $P \sim DP(\alpha, H)$ if and only if P can be constructed as

$$V_h \overset{iid}{\sim} Beta(1, \alpha), \quad \theta_h \overset{iid}{\sim} H \tag{1.9}$$

$$\pi_h = V_h \prod_{l \leq h} (1 - V_h) \tag{1.10}$$

$$P(B) = \sum_{h=1}^{\infty} \pi_h \mathbf{1}(\theta_h \in B) \tag{1.11}$$

The stick breaking representation shows clearly that draws from the DP are discrete almost surely. Further, the Ferguson (1974) characterization shows that distributions drawn from a DP are rough in the sense that the prior distribution of the probability mass assigned to two adjacent subsets has a weak negative correlation. As a model for many types of data these are undesirable properties since we expect the data to come from reasonably smooth and/or continuous distributions. In these cases the DP mixture model

$$P \sim DP(\alpha, H)$$

$$g(x) = \int_B f(x; \theta)\, dP(\theta) \tag{1.12}$$

is preferred. By (1.9)-(1.11), (1.12) can also be written as

$$g(x) = \sum_{h=1}^{\infty} \pi_h f(x; \theta_h) \tag{1.13}$$

where π_h and θ_h are as defined in (1.10) and (1.9) respectively. Hence DP mixtures are a generalization of finite mixture models to an infinite number of components.

1.2.2 Posterior inference in DP mixtures

MCMC samplers for DP mixtures can be exact or approximate, and may either represent or marginalize over the DP distributed measure P. Marginal samplers are based on the Polya urn scheme described in Blackwell and MacQueen (1973): If

$\theta_1, \theta_2, \ldots \theta_n, \theta_{n+1} \overset{iid}{\sim} P$, $P \sim DP(\alpha, H)$ then the predictive distribution of θ_{n+1} is

$$(\theta_{n+1} \mid \theta_1, \theta_2, \ldots \theta_n) \sim \frac{\alpha}{\alpha + n} H + \sum_{i=1}^{n} \frac{1}{\alpha + n} \delta_{\theta_i} \tag{1.14}$$

$$= \frac{\alpha}{\alpha + n} H + \sum_{h=1}^{k} \frac{n_k}{\alpha + n} \delta_{\tilde{\theta}_h} \tag{1.15}$$

where $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ are the k distinct values of θ_i for $1 \leq i \leq n$ and $n_h = \sum_{i=1}^{n} \mathbf{1}\Big(\theta_i = \tilde{\theta}_h\Big)$. Inference based on representation (1.14) is described in Escobar and West (1995). Bush (1996) proposed a more efficient sampler based on the representation in (1.15). Neal (2000) reviews these methods and their extension to nonconjugate priors on θ, proposing what is probably the most popular marginal sampler for Dirichlet process mixture models, his Algorithm 8 (although very recently Favaro and Teh (2013) proposed further refinements with improved computational efficiency).

Alternatively, the measure P may be instantiated during MCMC sampling. Ishwaran and James (2001) proposed truncating the stick breaking representation in (1.9) by setting $V_K \equiv 1$ for some large K. They show that this approximation can be quite accurate. Exact samplers include Papaspiliopoulos and Roberts (2008)'s retrospective sampling approach and the slice sampler of Walker (2007). Papaspiliopoulos (2008) connected the two approaches and proposed a blocked variant. It relies on the data-augmented likelihood for a single observation y_i

$$L(\theta, \pi;\ y_i, u_i, H_i) = \mathbf{1}(0 \leq u_i \leq \pi_{H_i}) f(y_i; \theta_{H_i}), \quad H_i \in \mathbb{N}. \tag{1.16}$$

Integrating out u_i leaves $\pi_{H_i} f(y_i; \theta_{H_i})$, which yields (1.13) on summing out H_i. The condition $\mathbf{1}(0 \leq u_i \leq \pi_{H_i})$ ensures that the conditional likelihood for H_i is nonnegative only for finitely many $h \in \mathbb{N}$, so it may be trivially updated even though there are in principle an infinite number of clusters. Since full conditional distribution for the stick breaking variables V_h is slightly awkward, Papaspiliopoulos (2008) update

the V_h are marginally over u from standard Beta full conditionals as in Ishwaran and James (2001). Only those V_h with $h \leq \max\{H_i : 1 \leq i \leq n\}$ need to be updated as the remainder are draws from the prior, instantiated as needed during the update for u_i.

1.3 Multiple Imputation

Multiple imputation (MI) is a principled method to adjust inference to account for uncertainty due to missing data. It was introduced in a series of papers in the late 1970's and early 1980's by Don Rubin and collaborators, and outlined in detail in Rubin (1987). The basic principle is to account for uncertainty due to missing data by averaging over the posterior predictive distribution of the missing data.

Let Q be the estimand of interest, and partition the dataset as $Y = (Y_{mis}, Y_{obs})$ where Y_{mis} is the missing data and Y_{obs} is the observed data. Let $\hat{Q}$ be an estimator of Q with associated variance estimate U. The Bayesian justification for multiple imputation begins with the identities

$$P(Q \mid Y_{obs}) = \int P(Q \mid Y_{mis}, Y_{obs}) P(Y_{mis} \mid Y_{obs})\, dY_{obs} \tag{1.17}$$

$$E(Q \mid Y_{obs}) = E(E(Q \mid Y_{mis}, Y_{obs}) \mid Y_{obs}) \tag{1.18}$$

$$\begin{aligned} Var(Q \mid Y_{obs}) = {} & Var(E(Q \mid Y_{mis}, Y_{obs}) \mid Y_{obs}) \\ & + E(Var(Q \mid Y_{mis}, Y_{obs}) \mid Y_{obs}) \end{aligned} \tag{1.19}$$

all of which follow from basic probability. MI essentially constructs Monte Carlo estimates of (1.18) and (1.19). Assume $Y_{mis}^{(m)}$ are samples from $P(Y_{mis} \mid Y_{obs})$, and

define $\hat{Q}_m$ as $\hat{Q}$ computed over the sample $(Y_{obs}, Y_{mis}^{(m)})$ with $\hat{U}_m$ defined similarly. Let

$$\bar{Q}_M = \sum_{m=1}^{M} \hat{Q}_m / M \tag{1.20}$$

$$\bar{U}_M = \sum_{m=1}^{M} \hat{U}_m / M \tag{1.21}$$

$$B_M = \sum_{m=1}^{M} (\hat{Q}_m - \bar{Q}_M)^2 / (M-1) \tag{1.22}$$

$$T_M = \bar{U}_M + \left(\frac{M+1}{M}\right) B_M. \tag{1.23}$$

When $M = \infty$, $(Q - \bar{Q}_\infty) \sim N(0, T_\infty)$. In practice small M (3-10) is usually sufficient, in which case the normal distribution is replaced by a t distribution with degrees of freedom $(M-1)(1 + \bar{U}_M/((1+1/M)B_M))^2$. For smaller sample sizes a more accurate adjusted degrees of freedom was given by Barnard and Rubin (1999). For univariate estimands confidence intervals may be constructed directly, and inverted to obtain p-values. For multicomponent estimands further approximations are necessary; see e.g. Chapter 3 in Rubin (1987) and Chapter 4, Section 3 of Schafer (1997) or the recent review in Reiter and Raghunathan (2007).

From the imputer's perspective the challenge lies in specifying a model for $P(Y_{mis} \mid Y_{obs})$. A natural approach is to assume a probability model indexed by Θ so that $P(Y_{mis} \mid Y_{obs}) = \int P(Y_{mis}, \Theta \mid Y_{obs}) d\theta$. Samples from $P(Y_{mis} \mid Y_{obs})$ are collected by sampling (approximately) from the joint posterior distribution $P(Y_{mis}, \Theta \mid Y_{obs})$ and discarding the samples of Θ. This is often a trivial modification of MCMC methods for complete-data problems, where the missing data is sampled from its full conditional $P(Y_{mis} \mid \Theta, Y_{obs})$ which often reduces to $P(Y_{mis} \mid \Theta)$.

In general the imputation model need not correspond to the ultimate model used by the analyst. Discrepancy between analysis and imputation models is called

uncongeniality (Meng, 1994). As long as the imputation model does not omit features that are important in the analysis model the resulting inferences tend to be valid. In fact, when information is used in the imputation model that is not available to the analyst, MI can give superefficient inference relative to the analysis that ignores this information. When the analyst is not the imputer, as in the traditional applications of MI by federal agencies, the imputation procedure must preserve as many features of the data as possible. Since these will not all be known in advance, it is appealing to use imputation procedures which can adapt to complex data.

1.3.1 Proper MI and probability models

Rubin (1987) notes that MI procedures derived by sampling from the posterior distribution under a Bayesian model tend to be proper in the sense of being approximately unbiased and giving confidence intervals with coverage rates at least as large as the nominal rate. Intuitively a good imputation model not only matches the data well but also "injects" enough variance into the samples of $(Y_{mis} \mid Y_{obs})$ to sufficiently inflate the pooled confidence intervals. This is why methods that fix an estimate $\hat{\Theta}$ of Θ and impute from $P(Y_{mis} \mid \hat{\Theta}, Y_{obs})$ are usually improper (Rubin, 1987). A simple example of this is the Census Bureau's hot deck, which imputes by 1) stratifying on some fully observed categorical variables and 2) drawing a value for missing observations randomly from the observed data in the same cell. The hot deck fixes Θ at the empirical distribution function within each cell, and the imputations have variance that is too low, even in large samples where the empirical distribution is a good approximation to the truth. When some cells have few cases the performance degrades further, which limits the number of conditioning variables that can be included.

The appeal of the hot deck as a nonparametric MI engine is due largely to its simplicity; it only requires the choice of conditioning variables. Specifying a complete probability model is challenging, particularly in the presence of mixed data. This has

lead to the popularity of multiple imputations by chained equations (MICE), which avoids using joint probability models by relying on a series of univariate conditional models instead (Raghunathan et al., 2001; Van Buuren and Oudshoorn, 1999). MICE imputes the data by iteratively sampling missing observations from the conditional models, like in a Gibbs sampler. The models may be specified explicitly, e.g. logistic regressions for categorical data, or implicitly with techniques like predictive mean matching (PMM). PMM is a generalization of the hot deck which samples donor values for a missing observation by computing the predictions from a linear regression model and imputing an observed value with a similar predicted mean.

MICE is currently the most popular approach for multiply imputing mixed data. However, it has its drawbacks. First, the series of conditional models need not (and often will not) correspond to any proper joint model. This undermines the theoretical foundation for MI, and it is unclear when to expect proper imputations from this method. Perhaps more importantly, specifying each conditional model is labor intensive, and in practice many analysts simply use default choices, such as univariate GLMs with only main effects. These specifications are restrictive, even if they happen to correspond to a proper joint distribution. For example, with multivariate categorical data imputing from a series of univariate conditional logistic regressions with main effects only corresponds to a loglinear model with only two-way interactions. Similarly, imputing continuous data from a series of linear regressions with main effects only and normal errors corresponds to a multivariate normal joint distribution. In Chapter 3 I develop an alternative proper joint model for mixed data and demonstrate its utility as a default imputation engine.

2

Bayesian Semiparametric Gaussian Copula Factor Models for Mixed Data

2.1 Introduction

Factor analysis and its generalizations are powerful tools for analyzing and exploring multivariate data, routinely used in applications as diverse as social science, genomics and finance. The typical Gaussian factor model is given by

$$y_i = \Lambda \eta_i + \epsilon_i \tag{2.1}$$

where y_i is a $p \times 1$ vector of observed variables, Λ is a $p \times k$ matrix of factor loadings ($k < p$), $\eta_i \sim N(0, I)$ is a $k \times 1$ vector of latent variables or factor scores, and $\epsilon_i \sim N(0, \Sigma)$ are idiosyncratic noise with $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Marginalizing out the latent variables, $y_i \sim N(0, \Lambda\Lambda' + \Sigma)$, so that the covariance in y_i is explained by the (lower-dimensional) latent factors. The model in (2.1) may be generalized to incorporate covariates at the level of the observed or latent variables, or to allow dependence between the latent factors. For exposition we focus on this simple case.

This model has been extended to data with mixed measurement scales, often by linking observed categorical variables to underlying Gaussian variables which

follow a latent factor model (e.g. Muthén (1983)). An alternative is to include shared latent factors in separate generalized linear models for each observed variable (Sammel et al., 1997; Moustaki and Knott, 2000; Dunson, 2003, 2000). Unlike in the Gaussian factor model, the latent variables typically impact both dependence and the form of the marginal distributions. For example, suppose $y_i = (y_{i1}, y_{i2})'$ are bivariate counts assigned Poisson log-linear models: $\log \mathrm{E}(y_{ij} \mid \eta_i) = \mu_j + \lambda \eta_i$. Then λ governs both the dependence between $y_{i1},\ y_{i2}$ and the overdispersion in each marginal distribution. This confounding can lead to substantial artifacts and misleading inferences. In addition, computation in such models is difficult, and requiring marginal distributions in the exponential family can be restrictive.

There is a growing literature on semiparametric latent factor models to address the latter problem. A number of authors have proposed mixtures of factor models (Ghahramani and Beal, 2000). Song et al. (2010) instead allow flexible error distributions in Eq. (2.1). Yang and Dunson (2010) proposed a broad class of semiparametric structural equation models that allow an unknown distribution for η_i. When building such flexible mixture models there is a sacrifice to be made in terms of interpretation, parsimony and computation, and subtle confounding effects remain. It would be appealing to retain the simplicity, interpretability and computational scalability of Gaussian factor models while allowing the marginal distributions to be unknown and free of the dependence structure.

To accomplish these ambitious goals we propose a semiparametric Bayesian Gaussian copula model utilizing the extended rank likelihood of Hoff (2007) for the marginal distributions. This approximation avoids a full model specification and is in some sense not fully Bayesian, but in practice we expect that this rank-based likelihood discards only a mild amount of information while providing robust inference. An additional contribution of this chapter is to provide new theoretical and empirical justification for this approach.

We proceed as follows: In Section 2.2, we propose the Gaussian copula factor model for mixed scale data and discuss its relationship to existing models. In Section 2.3 we develop a Bayesian approach to inference, specifying prior distributions and outlining a straightforward and efficient Gibbs sampler for posterior computation. Section 2.4 contains a simulation study, and Section 2.5 illustrates the utility of this method in a political science application. Section 2.6 concludes with a discussion.

2.2 The Gaussian copula factor model

A p-dimensional copula $\mathbb{C}$ is a distribution function on $[0,1]^p$ where each univariate marginal distribution is uniform on $[0,1]$. Any joint distribution F can be completely specified by its marginal distributions and a copula; that is, there exists a copula $\mathbb{C}$ such that

$$F(y_1,\ldots,y_p) = \mathbb{C}(F_1(y_1),\ldots,F_p(y_p)) \tag{2.2}$$

where F_j are the univariate marginal distributions of F (Sklar, 1959). If all F_j are continuous then $\mathbb{C}$ is unique, otherwise it is uniquely determined on $\text{Ran}(F_1)\times\cdots\times \text{Ran}(F_p)$ where $\text{Ran}(F_j)$ is the range of F_j. The copula of a multivariate distribution encodes its dependence structure, and is invariant to strictly increasing transformations of its univariate margins. Here we consider the Gaussian copula:

$$\mathbb{C}(u_1,\ldots u_p) = \Phi_p(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_p) \mid C), \quad (u_1,\ldots,u_p)\in[0,1]^p \tag{2.3}$$

where $\Phi_p(\cdot|C)$ is the p-dimensional Gaussian cdf with correlation matrix C and Φ is the univariate standard normal cdf. Combining (2.2) and (2.3) we have

$$F(y_1,\ldots,y_p) = \Phi_p(\Phi^{-1}(F_1(y_1)),\ldots,\Phi^{-1}(F_p(y_p)) \mid C). \tag{2.4}$$

From (2.4) a number of properties are clear: The joint marginal distribution of any subset of y has a Gaussian copula with correlation matrix given by the appropriate

submatrix of C, and $y_j \perp\!\!\!\perp y_{j'}$ if and only if $c_{jj'} = 0$. When F_j, $F_{j'}$ are continuous, $c_{jj'} = Corr(\Phi^{-1}(F_j(y_j), \Phi^{-1}(F_{j'}(y_{j'}))$ which is an upper bound on $Corr(y_j, y_{j'})$ (attained when the margins are Gaussian) (Klaassen and Wellner, 1997). The rank correlation coefficients Kendall's tau and Spearman's rho are monotonic functions of $c_{jj'}$ (Hult and Lindskog, 2002). For variables taking finitely many values $c_{jj'}$ gives the polychoric correlation coefficient (Muthén, 1983).

If the margins are all continuous, zeros in $R = C^{-1}$ imply conditional independence, as in the multivariate Gaussian distribution. However this is generally not the case when some variables are discrete. Even in the simple case where $p = 3$, Y_3 is discrete and $c_{13}c_{23} \neq 0$, if $r_{12} = 0$ then Y_1 and Y_2 are in fact *dependent* conditional on Y_3 (a similar result holds when conditioning on several continuous variables and a discrete variable as well - details available in supplementary materials). Results like these suggest that sparsity priors for R in Gaussian copula models (e.g. Pitt et al. (2006); Dobra and Lenkoski (2011)) are perhaps not always well-motivated when discrete variables are present, and should be interpreted only with great care.

A Gaussian copula model can be expressed in terms of latent Gaussian variables z. Let $F_j^{-1}(t) = \inf\{t : F_j(y) \geq t, y \in \mathbb{R}\}$ be the usual pseudo-inverse of F_j and suppose Ω is a covariance matrix with C as its correlation matrix. If $z \sim N(0, \Omega)$ and $y_j = F_j^{-1}(\Phi(z_j/\sqrt{\omega_{jj}}))$ for $1 \leq j \leq p$ then $F(y)$ has a Gaussian copula with correlation matrix C and univariate marginals F_j. We utilize this representation to generalize the Gaussian factor model to Gaussian copula factor models by assigning z a latent factor model:

$$\eta_i \sim N(0, I), \quad z_i|\eta_i \sim N(\Lambda\eta_i,\ I) \tag{2.5}$$

$$y_{ij} = F_j^{-1}\left(\Phi\left(\frac{z_{ij}}{\sqrt{1 + \sum_{h=1}^{k} \lambda_{jh}^2}}\right)\right). \tag{2.6}$$

Inference takes place on the scaled loadings

$$\tilde{\lambda}_{jh} = \frac{\lambda_{jh}}{\sqrt{1 + \sum_{h=1}^{k} \lambda_{jh}^2}} \tag{2.7}$$

so that $c_{jj'} = \sum_{h=1}^{k} \tilde{\lambda}_{jh}\tilde{\lambda}_{j'h}$. Rescaling is important as the factor loadings are not otherwise comparable across the different variables - even though Λ is technically identified it is not easily interpreted. We also consider the *uniqueness* of variable j, given by

$$u_j = 1 - \sum_{h=1}^{k} \tilde{\lambda}_{jh}^2 = \frac{1}{1 + \sum_{h=1}^{k} \lambda_{jh}^2} \tag{2.8}$$

In the Gaussian factor model $u_j = \sigma_j^2/(\sigma_j^2 + \sum_{h=1}^{k} \lambda_{jh}^2)$, the proportion of variance unexplained by the latent factors. In the Gaussian *copula* factor model this exact interpretation does not hold, but u_j still represents a measure of dependence on common factors.

2.2.1 Relationship to existing factor models

The Gaussian factor model and probit factor models are both special cases of the Gaussian copula factor model. Probit factor models for binary or ordered categorical data parameterize each margin by a collection of "cutpoints" $\gamma_{j0}, \ldots \gamma_{jc_j}$ (taking $\gamma_{j0} = -\infty$ and $\gamma_{jc_j} = \infty$ without loss of generality) so that $F_j(c) = \Phi\left(\gamma_{jc}(1 + \sum_{h=1}^{k} \lambda_{jh}^2)^{-1/2}\right)$. Then F_j has the pseudoinverse

$$F^{-1}(u_{ij}) = \sum_{c=1}^{c_j} c\mathbf{1}\left(\Phi\left(\frac{\gamma_{jc-1}}{\sqrt{1 + \sum_{h=1}^{k} \lambda_{jh}^2}}\right) < u_{ij} \leq \Phi\left(\frac{\gamma_{jc}}{\sqrt{1 + \sum_{h=1}^{k} \lambda_{jh}^2}}\right)\right). \tag{2.9}$$

Plugging this into (2.6) and simplifying gives $y_{ij} = \sum_{c=1}^{C_j} c\mathbf{1}(\gamma_{jc-1} < z_{ij} \leq \gamma_{jc})$ where $z_i \sim N(0, \Lambda\Lambda' + I)$, which is the data augmented representation of an ordinal probit

factor model. Naturally the connection extends to mixed Gaussian/probit margins as well.

Other factor models that contain Gaussian/probit models as special cases include semiparametric factor models, which assume non-Gaussian latent variables η_i or errors ϵ_i, retaining the linear model formulation (2.1) so that marginally $Cov(y_i) = \Lambda Cov(\eta_i)\Lambda' + \Sigma$. But $F(y_i)$ no longer has a Gaussian copula, and since the joint distribution is no longer elliptically symmetric the covariance matrix is unlikely to be an adequate measure of dependence. Further, the dependence and marginal distributions are confounded since the implied correlation matrix depends on the parameters of the marginal distributions.

The Gaussian copula factor model overcomes these shortcomings. In the Gaussian copula factor model $\tilde{\Lambda}$ governs the dependence separately from the marginal distributions, representing a factor-analytic decomposition for the scale-free copula parameter C rather than $Cov(y_i)$. The Gaussian copula model is also invariant to strictly monotone transformations of univariate margins. Therefore it is consistent with the common assumption that there exist monotonic functions $h_1, \ldots, h_p$ such that $(h_1(y_1), \ldots h_p(y_p))'$ follows a Gaussian factor model, while existing semiparametric approaches are not. Researchers using our method are not required to consider numerous univariate transformations to achieve "approximate normality".

2.2.2 Marginal Distributions

One way to deal with marginal distributions in a copula model is to specify a parametric family for each margin and infer the parameters simultaneously with C (see e.g. Pitt et al. (2006) for a Bayesian approach). This is computationally expensive for even moderate p, and there is often no obvious choice of parametric family for every margin. Since our goal is not to learn the whole joint distribution but rather to characterize its dependence structure, we would prefer to treat the marginal dis-

tributions as nuisance parameters.

A popular semiparametric method for continuous observations is a two-stage approach in which an estimator $\hat{F}_j$ is used to compute "pseudodata" $\hat{z}_{ij} = \Phi^{-1}(\hat{F}_j(y_{ij}))$, which are treated as fixed to infer the copula parameters. A natural candidate is $\hat{F}_j(t) = \frac{n}{n+1}\sum_{i=1}^{n}\frac{1}{n}\mathbf{1}(y_{ij} \leq t)$, the (scaled) empirical marginal cdf. Klaassen and Wellner (1997) considered such estimators in the Gaussian copula and Genest et al. (1995) developed them in the general case. However, this method cannot handle discrete margins. To accommodate mixed discrete and continuous data Hoff (2007) proposed an approximation to the full likelihood called the extended rank likelihood, derived as follows: Since the transformation $y_{ij} = F_j^{-1}(\Phi(z_{ij}))$ is nondecreasing, when we observe $y_j = (y_{1j}, \ldots, y_{nj})$ we also observe a partial ordering on $z_j = (z_{1j}, \ldots z_{nj})$. To be precise we have that

$$z_j \in D(y_j) \equiv \{z_j \in \mathbb{R}^n : y_{ij} < y_{i'j} \Rightarrow z_{ij} < z_{i'j}\} \tag{2.10}$$

The set $D(y_j)$ is just the set of possible $z_j = (z_{1j}, \ldots, z_{nj})$ which are consistent with the ordering of the observed data on the j^{th} variable. Let $D(Y) = \{Z \in \mathbb{R}^{p\times n} : z_j \in D(y_j) \quad \forall\ 1 \leq j \leq p\}$. Then we have

$$P(Y|C, F_1, \ldots F_p) = P(Y, Z \in D(Y)|C, F_1, \ldots F_p) \tag{2.11}$$

$$= P(Z \in D(Y)|C) \times P(Y|Z \in D(Y), C, F_1, \ldots F_p) \tag{2.12}$$

where (2.12) holds because given C the event $Z \in D(Y)$ does not depend on the marginal distributions. Hoff (2007) proposes dropping the second term in (2.12) and using $P(Z \in D(Y)|C)$ as the likelihood. Intuitively we would expect the first term to include most of the information about C. Simulations in Section 2.4 provide further evidence of this. Hoff (2007) shows that when the margins are all continuous the marginal ranks satisfy certain relaxed notions of sufficiency for C, although these fail when some margins are discrete. Unfortunately theoretical results for applications involving mixed data have been lacking.

To address this we give a new proof of strong posterior consistency for C under the extended rank likelihood with nearly any mixture of discrete and continuous margins (barring pathological cases which preclude identification of C). Posterior consistency will generally fail under Gaussian/probit models when any margin is misspecified. A similar result for continuous data and a univariate rank likelihood was obtained by Gu and Ghosal (2009). We replace Y with $Y^{(n)}$ for notational clarity below.

Theorem 1. *Let Π be a prior distribution on the space of all positive semidefinite correlation matrices $\mathcal{C}$ with corresponding density $\pi(C)$ with respect to a measure ν. Suppose $\pi(C) > 0$ almost everywhere with respect to ν and that $F_1, \ldots, F_p$, are the true marginal cdfs. Then for C_0 a.e. $[\nu]$ and any neighborhood $\mathcal{A}$ of C_0 we have that*

$$\lim_{n\to\infty} \Pi(C \in \mathcal{A} \mid Z^{(n)} \in D(Y^{(n)})) = 1 \text{ a.s. } [G^{\infty}_{C_0,F_1,\ldots,F_p}] \tag{2.13}$$

where $G^{\infty}_{C_0,F_1,\ldots,F_p}$ is the distribution of $\{y_i\}_{i=1}^{\infty}$ under $C_0, F_1, \ldots F_p$.

The proof is in Appendix A.1. We assumed a prior $\pi(C)$ having full support on $\mathcal{C}$. Under factor-analytic priors fixing $k < p$ restricts the support of π, and posterior consistency will only hold if C_0 has a factor analytic decomposition in k or fewer factors. But by setting k large (or inferring it) it is straightforward to define factor-analytic priors which have full-support on $\mathcal{C}$ (further discussion in Section 2.6). In practice, many correlation matrices which do not exactly have a k-factor decomposition are still well-approximated by a k-factor model. Finally, the result also applies to posterior consistency for $\tilde{\Lambda}$ if k is chosen correctly, given compatible identifying restrictions.

The efficiency of semiparametric estimators such as ours is also an important issue. Hoff et al. (2011) give some preliminary results which suggest that pseudo-MLE's based on the rank likelihood for continuous margins may be asymptotically relatively efficient. However, it is unclear whether even these results apply to the

case of mixed continuous and discrete margins, which is our primary focus. Simulations of the efficiency of posterior means under the extended rank likelihood versus correctly specified parametric models appear in Section 2.4.1. These results give an indication of the worst-case scenario in terms of efficiency lost in using the likelihood approximation, and are quite favorable in general.

2.3 Prior Specification and Posterior Inference

2.3.1 Prior Specification

Since the factor model is invariant under rotation or scaling of the loadings and scores, we assume that sufficient identifying conditions are imposed (by introducing sign constraints and fixed zeros in Λ), or that inference is on C which does not suffer from this indeterminacy. For brevity we also assume here that k is known and fixed. Suggestions for incorporating uncertainty in k are given in the Discussion section.

A common prior for the unrestricted factor loadings in Gaussian, probit or mixed factor models is $\lambda_{jh} \sim N(0, 1/b)$. However, these priors have some troubling properties outside the Gaussian factor model: When $\sigma_j \equiv 1$ as in probit or mixed Gaussian/probit factor models – or in our copula model – the implied prior on u_j is

$$\pi(u_j) = \frac{(b/2)^{-k/2}}{\Gamma(k/2)} \left(\frac{1}{u_j^2}\right) \left(\frac{1-u_j}{u_j}\right)^{k/2-1} \exp\left[-\frac{b}{2}\left(\frac{1-u_j}{u_j}\right)\right]. \tag{2.14}$$

Figure 2.1 shows that these priors are quite informative on the uniquenesses, especially as k increases. When k is small they are particularly informative on the scaled loadings, shrinking $\tilde{\lambda}_{jh}$ toward *large* values, rather than toward zero. This effect becomes worse as the prior variance increases. The problem is that the normal prior puts insufficient mass near zero. Coupled with the normalization this results in a "smearing" of mass across the columns of $\tilde{\Lambda}$, deflating u_j, inducing spurious correlations, and giving inappropriately high probability to values of the

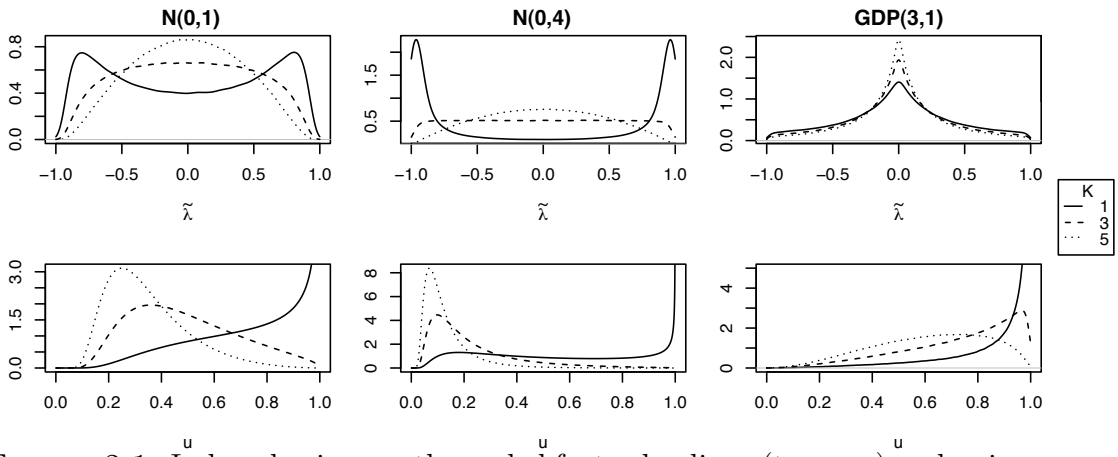

FIGURE 2.1: Induced priors on the scaled factor loadings (top row) and uniquenesses (bottom row) implied by different priors as K varies

scaled loadings near ± 1. Therefore the normal prior is a very poor default choice in these models.

To address these shortcomings we consider shrinkage priors on λ_{jh} that place significant mass at or near zero. Such parsimony is also desirable for more interpretable results. Shrinkage priors have been thoroughly explored in the regression context (see e.g. Polson and Scott (2010) and references therein). In that context heavy-tailed distributions are desirable. While somewhat heavy tails are appealing here (so that $\pi(\tilde{\lambda}_{jh})$ decays slowly to zero as $|\tilde{\lambda}_{jh}| \to 1$), *extremely* heavy tails are inappropriate. Very heavy tails imply that with significant probability a single unscaled loading (say λ_{jm}) in a row j will be much larger than the others so that $\tilde{\lambda}_{jh} \approx \lambda_{jh}/|\lambda_{jm}|$ for $1 \leq h \leq k$. The resulting joint prior on $\tilde{\lambda}_j = (\tilde{\lambda}_{j1}, \ldots, \tilde{\lambda}_{jk})$ will assign undesirably high probability to vectors with one entry near ± 1 and the rest near 0, yielding correlations which are approximately 0 or ± 1. Applying these priors in this new setting requires extra care.

As a default choice we recommend the generalized double Pareto (GDP) prior of

(Armagan et al., 2011) which has the density

$$\pi(\lambda_{jh}) = \frac{\alpha}{2\beta}\left(1 + \frac{|\lambda_{jh}|}{\beta}\right)^{-(\alpha+1)} \tag{2.15}$$

which we will refer to as $\lambda_{jh} \sim GDP(\alpha, \beta)$. The GDP is a flexible generalization of the Laplace distribution with a sharper peak at zero and heavier tails. It has the following scale-mixture representation: $\lambda_{jh}|\psi_{jh} \sim N(0, \psi_{jh})$, $\psi_{jh}|\xi_{jh} \sim Exp(\xi_{jh}^2/2)$, and $\xi_{jh} \sim Ga(\alpha, \beta)$ which leads to conditional conjugacy and a simple Gibbs sampler. The GDP's tail behavior is determined by α, and β is a scale parameter. Armagan et al. (2011) handle the hyperparameters by either fixing them both at 1 or assigning them a hyperprior. Here taking $\alpha = 3$, $\beta = 1$ is a good default choice: The $GDP(3, 1)$ distribution has mean 0 and variance 1, and $Pr(|\lambda_{jh}| < 2) \approx 0.96$. Critically, taking $\alpha = 3$ leads to tails of $\pi(\lambda_{jh})$ light enough to induce a sensible prior on $\tilde{\lambda}_{jk}$.

Figure 2.1 shows draws from the implied prior on u_j and $\tilde{\lambda}_{jh}$ under the $GDP(3, 1)$ prior, which are much more reasonable than the current default Normal priors. Note that as K increases, the prior puts increasing mass near zero without changing a great deal in the tails. This is reasonable since we expect each variable to load highly on only a few factors, and is not a feature of the light-tailed normal priors. The prior on the uniquenesses remains relatively flat under the GDP prior, whereas the normal prior increasingly favors lower values and less parsimony.

2.3.2 Parameter-Expanded Gibbs Sampling

For efficient MCMC inference we introduce a parameter-expanded (PX) version of the original model. The PX approach (Meng and Van Dyk, 1999; Liu and Wu, 1999) adds redundant (non-identified) parameters to reduce serial dependence in MCMC and improve convergence and mixing behavior. Naive Gibbs sampling in our model suffers from high autocorrelation due to strong dependence between Z and Λ. We modify (2.6) by adding scale parameters $V = \text{diag}(v_1^2, \ldots, v_p^2)$ to weaken this

dependence:

$$w_i \sim N\left(V^{1/2}\Lambda\eta_i, V\right) \tag{2.16}$$

$$y_{ij} = F_j^{-1}\left(\Phi\left(\frac{w_{ij}}{v_j\sqrt{1+\sum_{j=1}^{K}\lambda_{jh}^2}}\right)\right) \tag{2.17}$$

Since w_{ij}/v_j and z_{ij} are equal in distribution (2.17) is observationally equivalent to the original model. We assume that V is independent of the inferential parameters *a priori* so that $\pi(\Lambda, H, V|Y) = \pi(\Lambda, H|Y)\pi(V)$ (where H' is the $n \times k$ matrix with entries η_{ik}) and the marginal posterior distribution of the inferential parameters is unchanged.

We choose the conjugate PX prior $1/v_j^2 \sim Ga(n_0/2, n_0/2)$ (independently). The greatest benefits from PX are realized when the PX prior is most diffuse, which would imply sending $n_0 \to 0$ and an improper PX prior. The resulting posterior for (Λ, H, V) is also improper, but we can prove that the samples of (Λ, H) from the corresponding Gibbs sampler still have the desired stationary distribution $\pi(\Lambda, H|Y)$ (Appendix A.2). The PX Gibbs sampler is implemented as follows:

PX parameters: Draw $1/v_j^2 \sim Ga(n/2, s_j/2)$ where $\Psi_j = \text{diag}(\psi_{j1}^2/2, \ldots, \psi_{jh_j}^2/2)$ and $s_j = z_j(I - H_j'(\Psi_j^{-1} + H_j H_j')^{-1}H_j)z_j'$.

Factor Loadings: We assume a lower triangular loadings matrix with a positive diagonal; the extension to other constraints is straightforward. Let $k_j = \max(k, j)$ and H_j' be the $n \times k_j$ matrix with entries η_{ik} for $1 \le k \le k_j$ and $1 \le i \le n$. Update nonzero elements in row j of Λ as $\lambda_j' \sim N(\hat{\lambda}_j'/v_j,\ (\Psi_j^{-1} + H_j H_j')^{-1})$ where $\hat{\lambda_j}' = (\Psi_j^{-1} + H_j H_j')^{-1}H_j z_j'$ and λ_{jj} is restricted to be positive if $j \le k$.

Hyperparameters: Update $(1/\psi_{jh}|-) \sim InvGauss(|\xi_{jh}/\lambda_{jh}|,\ \xi_{jh}^2)$ and $(\xi_{jh}|-) \sim Ga(\alpha+1,\ \beta+|\lambda_{jh}|)$ where $InvGauss(a, b)$ is the inverse-Gaussian distribution with mean a and scale b.

Factor scores: Draw η_i from $(\eta_i|-) \sim N\left([\Lambda'\Lambda + I]^{-1}\Lambda' z_i, [\Lambda'\Lambda + I]^{-1}\right)$.

Augmented Data: Update Z elementwise from

$$(z_{ij}|-) \sim TN\left(\sum_{h=1}^{k} \lambda_{jh}\eta_{ki}, 1, z_{ij}^{l}, z_{ij}^{u}\right) \tag{2.18}$$

where $TN(m, v, a, b)$ denotes the univariate normal distribution with mean m and variance v truncated to (a, b), $z_{ij}^{l} = \max\{z_{i'j} : y_{i'j} < y_{ij}\}$ and $z_{ij}^{u} = \min\{z_{i'j} : y_{i'j} > y_{ij}\}$. If y_{ij} is missing then $(z_{ij}|-) \sim N(\sum_{h=1}^{k} \lambda_{jh}\eta_{ki}, 1)$. Note that (2.18) doesn't require a matrix inversion since $(z_{ij} \perp\!\!\!\perp z_{ij'} \mid \Lambda, \eta_i, Y)$ for $j \neq j'$, a unique property of our factor analytic representation and a significant computational benefit as p grows.

The PX-Gibbs sampler has mixing behavior at least as good as Gibbs sampling under the original model (which fixes $V = I$) (Liu and Wu, 1999; Meng and Van Dyk, 1999), and the additional computation is negligible. The PX-Gibbs sampler often increases the smallest effective sample size (associated with the largest loadings) by an order of magnitude or more in both real and synthetic data. The improved mixing is also vital for the multimodal posteriors sometimes induced by shrinkage priors. To our knowledge this is the first application of PX to factor analysis of mixed data, but PX has previously been applied to Gibbs sampling in Gaussian factor models by Ghosh and Dunson (2009) who introduce scale parameters for η_i to reduce dependence between H and Λ. Since MCMC in our model suffers primarily from dependence between Z and Λ our approach is more appropriate. Hoff (2007) and Dobra and Lenkoski (2011) also use priors on unidentified covariance matrices to induce a prior on correlation matrices in Gaussian copula models. But the motivation there is simply to derive tractable MCMC updates and dependence between the priors on C and V precludes our strategy of choosing an optimal PX prior, limiting the benefits of PX.

2.3.3 Posterior Inference

Given MCMC samples we can address a number of inferential problems. The posterior distribution of the factor scores η_i provide a measure of the latent variables for each data point, describing a projection of the observed data into the latent factor space, and the factors themselves are characterized by the variables which load highly on them. Even if the factors are not directly interpretable this is a very useful exploratory technique for mixed data which is robust to outliers and handles missing data automatically (unlike common alternatives such as principal component analysis).

We can also do inference on conditional or marginal dependence relationships in y_i. Here there is no need for identifying constraints in Λ which simplifies model specification. Tests of independence like $H_0 : c_{jj'} \leq \epsilon$ versus $H_1 : c_{jj'} > \epsilon$ are simple to construct from MCMC output. When the variables are continuous the conditional dependence relationships are encoded in $R = C^{-1}$ which we can compute as

$$R = (\tilde{\Lambda}\tilde{\Lambda}' + U)^{-1} = U^{-1} - U^{-1}\tilde{\Lambda}[I + \tilde{\Lambda}'U^{-1}\tilde{\Lambda}]^{-1}\tilde{\Lambda}'U^{-1} \tag{2.19}$$

Eq. (2.19) requires calculating only k-dimensional inverses, rather than p-dimensional inverses, a significant benefit of our factor-analytic representation.

As discussed in Section 2.2 the presence of discrete variables complicates inference on conditional dependence. Additionally, two discrete variables may be effectively marginally independent even if $|c_{jj'}| > 0$ simply by virtue of their levels of discretization. For these reasons, and for more readily interpretable results, it can be valuable to consider aspects of the posterior predictive distribution $\pi(y^*|Y)$. Under our semiparametric model this distribution is somewhat ill-defined, but we can sample from an approximation to $\pi(y^*|Y)$ by drawing $\tilde{\Lambda}$ via the PX-Gibbs sampler, drawing $z^* \sim N(0, \tilde{\Lambda}\tilde{\Lambda}' + U)$ and setting $y_j^* = \hat{F}_j^{-1}(\Phi(z_j^*))$ where $\hat{F}_j$ are estimators of each marginal distribution. This disregards some uncertainty when making predic-

tions; Hoff (2007) provides an alternative based on the values of $z_{ij}^l,\ z_{ij}^u$ from (2.18) but (in keeping with his observations) we find both approaches to perform similarly.

Posterior predictive sampling of conditional distributions is detailed in Section 2 of the supplement. Importantly, the factor-analytic representation of C allows us to directly sample any conditional distributions of interest (rather than using rejection sampling from the joint posterior predictive) by reducing the problem of sampling a truncated multivariate normal distribution to that of sampling independent truncated univariate normals.

2.4 Simulation Study

When fitting models in the following simulations we used the $GDP(3,1)$ prior for λ_{jk} and take $1/\sigma^2 \sim Gamma(2,2)$ for the Gaussian factor model and uniform priors on the cutpoints in the probit model. The cutpoints in the probit model were updated using independence Metropolis-Hastings steps with a proposal derived from the empirical cdfs. All models were fit using our R package.

2.4.1 Relative efficiency

First we examine finite-sample relative efficiency of the extended rank likelihood in the "worst-casse scenario" (for our method). We compare the posterior mean correlation matrix under the Gaussian copula factor model with the extended rank likelihood to that under 1) a Gaussian factor model, when the factor model is true and 2) a probit factor model, when the probit model is true. Both are special cases of the Gaussian copula factor model so we can directly compare the parameters.

The true (unscaled) factor loadings were sampled iid $GDP(3,1)$. For the probit case each margin had five levels with probabilities sampled $Dirichlet(1/2,\ldots,1/2)$. We fix k at the truth; additional simulations suggest that the relative performance is similar under misspecified k. We performed 100 replicates for various $p/k/n$ com-

binations in Figure 2.2. Each model was fit using 100,000 MCMC iterations after a 10,000 burn-in, keeping every 20th sample. MCMC diagnostics for a sample of the fitted models indicated no convergence issues. We assess the performance of each method by computing a range of loss functions: Average and maximum absolute bias ($\frac{2}{p(p-1)} \sum_{i<j<p} |\hat{c}_{ij} - c_{ij}|$ and $\max_{i<j<p} |\hat{c}_{ij} - c_{ij}|$ respectively), root squared error: $\left[2 \sum_{i<j<p} (\hat{c}_{ij} - c_{ij})^2\right]^{1/2}$ and Stein's loss: $\text{tr}(\hat{C}C^{-1}) + \log\det(\hat{C}C^{-1}) - p$. Stein's loss is (up to a constant) the KL divergence from the Gaussian copula density with correlation matrix C to the Gaussian copula density with correlation matrix $\hat{C}$ and is therefore natural to consider here.

Figure 2.2 shows that the two methods are more or less indistinguishable in the probit case. Our method slightly outperforms the probit model in many cases because we do not have to specify a prior for the cutpoints. There are also some computational benefits here since in the copula model we avoid Metropolis-Hastings steps for the marginal distributions. In the continuous case our model also does well, although the Gaussian model is somtimes substantially more efficient under Stein's loss. But as p grows our model is increasingly competitive.

2.4.2 Misspecification Bias and Consistency

The previous simulations suggest that the loss in efficiency in worst-case scenarios is quite often minimal. To illustrate the practical benefit of our model (and the impact of our consistency result) in a realistic scenario we simulated data from a one-factor Gaussian copula factor model using the marginal distributions from the dataset we analyze in Section 2.5 (shown in Fig. 2.3). For simplicity we take $\tilde{\Lambda} = \tilde{\lambda}1$ and consider $\tilde{\lambda} = 0.7$ and 0.8 (although we did not constrain the loadings to be equal when fitting models to the simulated data). Results in Section 2.5 suggest that these are plausible values.

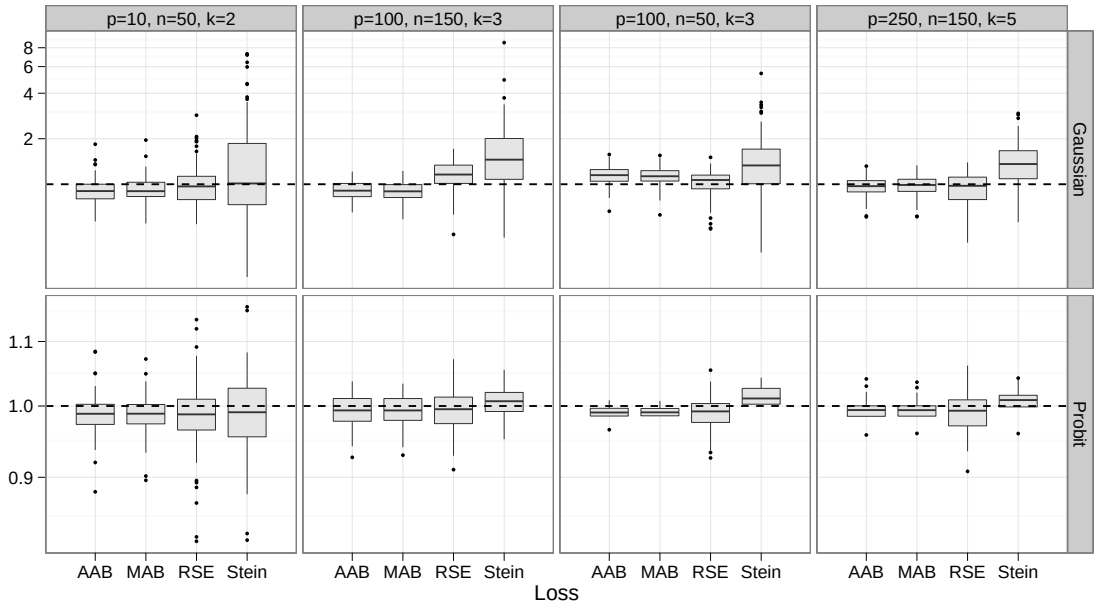

FIGURE 2.2: Efficiency (ratio of the loss under our model to that under the Gaussian/probit model) of the posterior mean under a range of loss functions

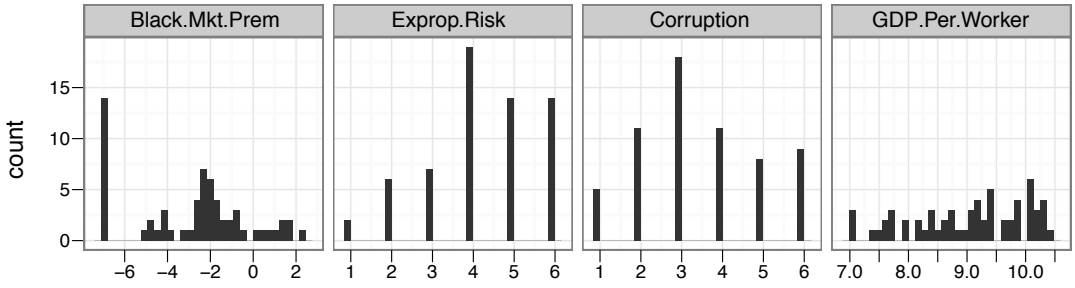

FIGURE 2.3: Distributions of 4 variables from the political risk dataset in Quinn (2004). The fifth, Ind.Jud, is binary with 34/62 ones.

Figure 2.4 shows that factor loadings for the two continuous variables (black market premium and GDP per worker) are underestimated by the Gaussian/probit model. When all the variables are dependent there is a "ripple" effect, so that even factor loadings for discrete variables are subject to some bias. We should expect this behavior in general – the copula correlations bound the observed Pearson correlations from above (in absolute value), with the bounds obtained only under Gaussian margins. The difference between the Pearson and copula correlation parameters, and

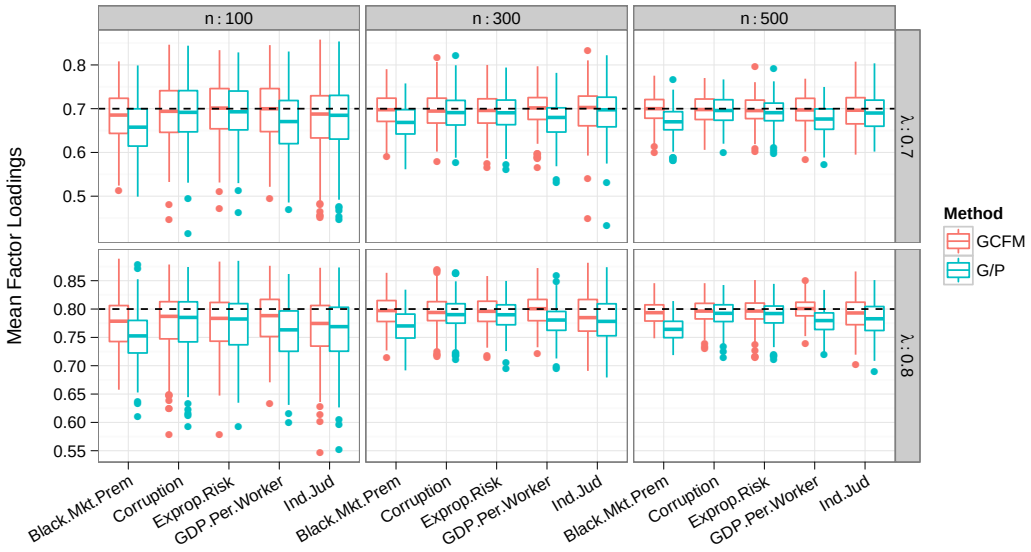

FIGURE 2.4: Posterior mean factor loadings using 100 simulated datasets generated with the margins in Section 5 using our model (GCFM) versus a mixed Gaussian/probit model (G/P).

hence the asymptotic bias, depends entirely on the form of the marginal distributions. This makes proper choices of transformations critical in the parametric model. Our model relieves this concern entirely. Although the magnitude of these effects is relatively mild here there is little reason to suspect this is true in general, especially in more complex models with multiple factors and a larger number of observed variables.

2.5 Application: Political-Economic Risk

Quinn (2004) considers measuring political-economic risk, a latent quantity, using five proxy variables and a Gaussian/probit factor model. Political economic risk is defined as the risk of a state "manipulat[ing] economic rules to the advantage of itself and its constituents" (North and Weingast, 1989, pp. 808). The dataset includes five indicators recorded for 62 countries: independent judiciary, black market

premium, lack of appropriation risk, corruption, and gross domestic product per worker (GDPW) (Fig. 2.3). Additional background on political-economic risk and on the variables in this dataset is provided by Quinn (2004), and the data are available in the R package MCMCpack. Quinn (2004) transforms the positive continuous variables GDPW and black market premium by $\log(x)$ and $\log(x + 0.001)$ (resp.). The disproportionate number of zeros in black market premium (14/62 observations) leaves a large spike in the left tail and the normality assumption is obviously invalid. Since Quinn (2004) has already implicitly assumed a Gaussian copula, our model is a natural alternative to the misspecified Gaussian/probit model used there.

To explore sensitivity to prior distributions we fit the copula model under several priors: $GDP(3, 1)$, $N(0, 1)$ and the $N(0, 4)$ priors used by Quinn (2004). We use 100,000 MCMC iterations and save every 10^{th} sample after a burn-in of 10,000 iterations. Standard MCMC diagnostics gave no indication of lack of convergence. Figure 2.5 shows posterior means and credible intervals for the scaled loadings under each prior. Note that the $N(0, 4)$ prior, intended to be noninformative, is actually very informative on the scaled loadings (Fig 2.1). It pulls the scaled loadings toward ± 1, with most pronounced influence on the binary variable Ind.Jud and the other categorical variables. The GDP prior instead shrinks toward zero as we would expect.

We also compare our model to the Gaussian/probit model in Quinn (2004), but using the $GDP(3, 1)$ prior in both cases. Figure 2.6 shows posterior predictive means and credible intervals for Kendall's tau, as well as the observed values and bootstrapped confidence intervals. Our model fits well, considering the limited sample size, and fits almost uniformly better than the Gaussian-probit model. Other posterior predictive checks on rank correlation measures and in subsets of the data show similar results.

Incorrectly assuming a normal distribution for log black market premium is espe-

cially damaging. The copula correlation between GDPW and black market premium (on which the data are most informative) is underestimated in the Gaussian/probit model: mean -0.33 and 95% HPD interval $(-0.46, -0.22)$ as opposed to -0.56 and $(-0.73, -0.40)$ under our model. This is also evident in the posterior predictive samples of Kendall's τ in Fig. 2.6. Figure 2.7 shows density estimates of draws from the bivariate posterior predictive of black market premium and GDPW. The Gaussian/probit model is clearly not a good fit, assigning very little mass to the bottom-right corner (which contains almost 25% of the data). The Gaussian copula factor model assigns appropriately high density to this region. Estimates of the latent variables are impacted as well: Figure 2.8 plots the mean factor scores from each model (after shifting and scaling to a common range) for low-risk countries. The seven countries with the lowest risk have identical covariate values except on GDPW. Our model infers mean scores that are sorted by GDPW (higher GDPW yielding a lower score). The Gaussian/probit model instead assigns these countries almost identical scores.

2.6 Discussion

We have developed a new semiparametric approach to the factor analysis of mixed data that is both robust and efficient. We propose new default prior distributions for factor loadings that are more suited to routine use of this model (and similar models, such as probit factor models). We also induce attractive new priors on correlation matrices in Gaussian copula models; these are both more flexible and parsimonious than the inverse Wishart prior used by Hoff (2007), and much more efficient computationally than the graphical model priors of Dobra and Lenkoski (2011). They admit optimal parameter expansion schemes which are easy to implement. Additionally, they are readily extended to informative specifications, to include covariates or to more complex latent variable models.

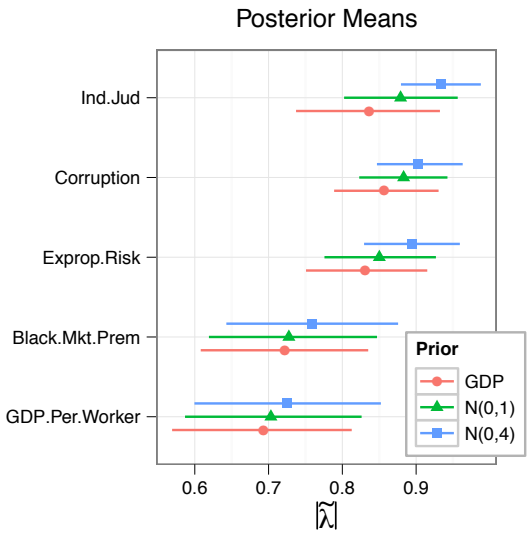

FIGURE 2.5: Posterior means/90% HPD intervals for scaled factor loadings under the different priors. Differences due to priors are larger for discrete variables, and largest for Ind.Jud (which is binary)

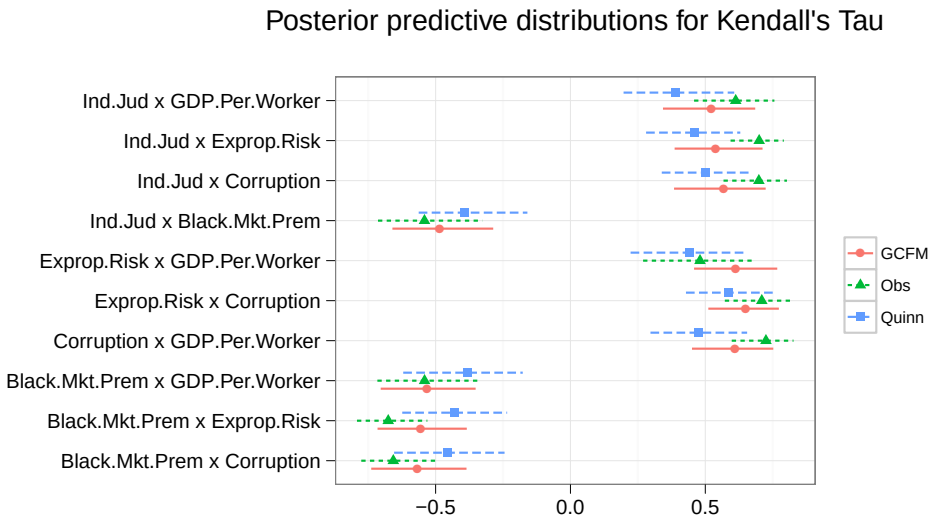

FIGURE 2.6: Posterior predictive mean and 95% HPD intervals of Kendall's τ under our model (Cop) and the Gaussian-probit model (Mix) as well as the observed values and bootstrapped 95% confidence intervals.

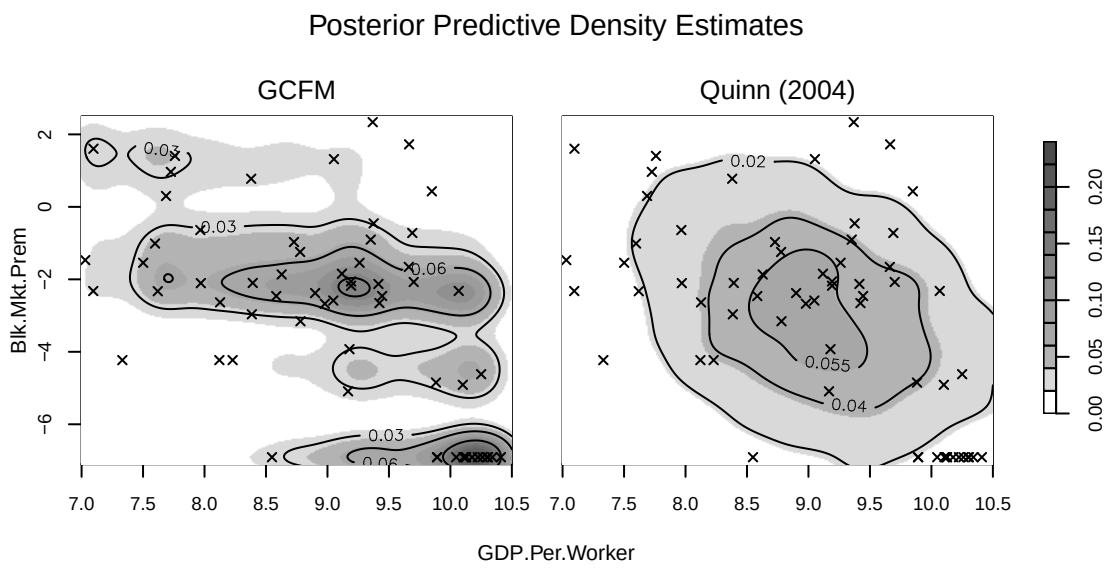

FIGURE 2.7: Posterior predictive distributions of log GDP and log black market premium, with observed data scatterplots. Note the cluster of points in the bottom-right corner; even though they represent over 20% of the sample the predictive density from the model in Quinn (2004) assigns very little mass to this area.

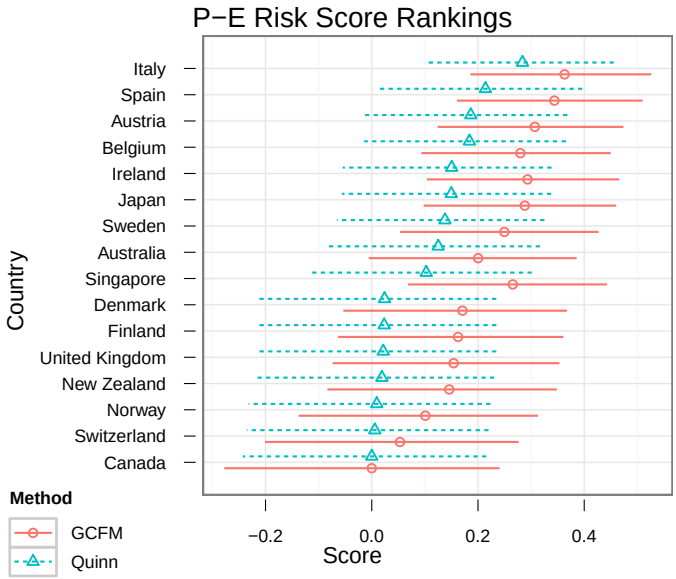

FIGURE 2.8: Comparison of the political-economic risk ranking obtained via our model and the mixed-data factor analysis of Quinn (2004). Points are posterior means and lines represent marginal 90% credible intervals.

We have not considered the issue of uncertainty in the number of factors, but it is straightforward to do so by adapting existing methods for Gaussian factor models. In addition to posterior predictive checks, these include stochastic search (Carvalho et al., 2008), reversible jump MCMC (Lopes and West, 2004), Bayes factors (Ghosh and Dunson, 2009; Lopes and West, 2004) and nonparametric priors (Paisley and Carin, 2009; Bhattacharya and Dunson, 2011). The latter are especially promising when interest lies in C since they preserve the computational advantages of factor-analytic priors while providing full support on correlation matrices (which fails for fixed $k < p$). Particularly when the plausible range of k is quite small, posterior predictive checks can be effective.

3

Nonparametric Bayes for Multiple Imputation of Mixed Data

3.1 Introduction

Large-scale surveys capture a variety of heterogeneous data on respondents. These multivariate responses frequently include a mix of continuous and ordered or unordered categorical variables. Many of these variables are also subject to missingness. A popular approach to handling missing data is multiple imputation (MI) (Rubin, 1976, 1987). MI operates by repeatedly sampling the missing data from its predictive distribution under an appropriate probability model. Analysts can account for uncertainty in a principled manner by pooling estimates and confidence intervals computed in each completed dataset.

Routine implementation of proper MI in heterogenous datasets is hampered by a lack of sufficiently flexible and tractable joint models for mixed data. This has been a key driver of the popularity of methods that generate imputations from a series of univariate conditional models. While these methods can perform well in practice – often outperforming existing joint models – they have theoretical and practical

drawbacks. This motivates us to develop next-generation joint modeling techniques.

In this chapter we introduce a flexible nonparametric Bayesian joint model for mixed continuous, ordered and unordered categorical data suitable for use as a default imputation method, and demonstrate its superior performance over conditional imputation methods in a genuine dataset under repeated sampling. Nonparametric Bayesian methods are attractive for MI due to their ability to incorporate increasingly complex features of the data as the sample size increases. This allows the imputer to potentially preserve structure in the data that he or she may not have anticipated but may be important to analysts. This data-adaptive behavior is not automatic, however; in nonparametric models aspects of the prior can dominate even in very large samples. We provide a carefully constructed hierarchical model and prior distributions suitable for a wide range of applications. Its modular structure allows extension to further incorporate domain knowledge or special features of particular datasets.

The chapter proceeds as follows: In Section 3.2 we describe some common challenges when imputing mixed data in large surveys. We discuss how these manifest in our motivating application, imputing missing data in the Survey of Income and Program Participation (SIPP). In Section 3.3 we introduce the model and prior specification, and discuss the relationship between our proposal and related methods. In Section 3.4 we present extensive simulations using SIPP data demonstrating the improved performance of MI using our model versus a popular competitor. Finally, in Section 3.5 we discuss extensions and future work.

3.2 Motivation

The Survey of Income and Program Participation (SIPP) is a large panel survey conducted by the U.S. Census Bureau. SIPP records a large number of variables on each individual in participating households, including demographic and household

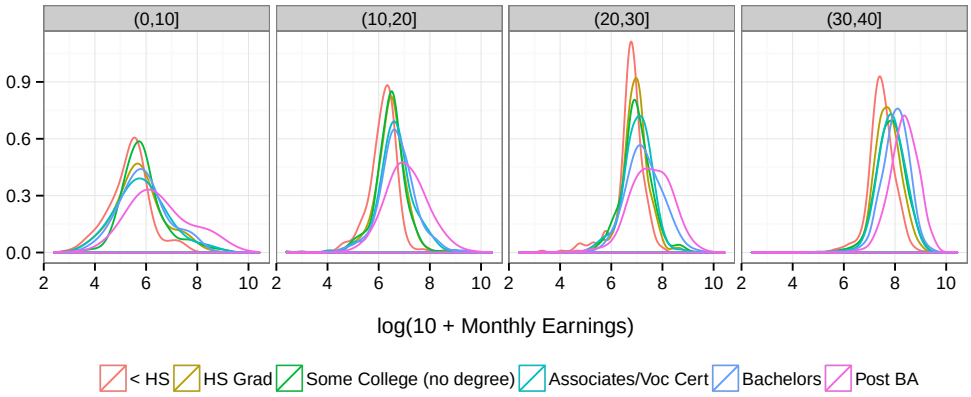

FIGURE 3.1: Log monthly earnings by usual hours worked and education level for those reporting at most 40 hours.

characteristics, labor force participation, and taxes, assets, liabilities and sources of income (including transfer programs).

SIPP is characteristic of large official surveys in that it includes many categorical variables and a smaller number of continuous variables, with complicated dependence and nonstandard distributions. For example, Figure 3.1 plots (log) total earnings by usual hours worked and education level. The distribution of income is complex and varies across levels of the discrete variables: When usual hours ≤ 10, the earnings distribution is right skewed, but as the number of hours increases it eventually becomes slightly skewed left. In the first three panels, increasing education level is associated with increased dispersion in the distribution of log earnings. Compare this with the last panel, where increased education beyond high school is primarily associated with a location shift in earnings. There is also evidence of higher order dependence in the marginal distributions of the categorical variables. Table 3.2 shows analysis of deviance tables for 1, 2, and 3 way loglinear models fit to subsets of the categorical variables, chosen more or less at random. All indicate some evidence of interactions.

Datasets like SIPP are difficult to impute with joint models since there are few

Table 3.1: Analysis of deviance tables for loglinear models fit to three subsets the SIPP data

Race, gender, education level, hourly

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1 way	108	10191.25			
2 way	69	211.50	39	9979.76	$< 10^{-6}$
3 way	20	39.92	49	171.58	$< 10^{-6}$

Marital status, usual hours, gender, no. own children

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1 way	132	6350.60			
2 way	91	1383.23	41	4967.36	$< 10^{-6}$
3 way	30	168.83	61	1214.40	$< 10^{-6}$

Occupation, gender, education level, hourly, union

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1 way	1073	37323.52			
2 way	879	2261.08	194	35062.43	$< 10^{-6}$
3 way	467	545.25	412	1715.83	$< 10^{-6}$

joint models for such heterogenous data. A common approach is to use a general location model (Olkin and Tate (1961); Little and Schluchter (1985) and (Schafer, 1997, Ch. 9)). For continuous variables Y and discrete variables X, the general location model assumes that $(Y \mid X = x) \sim N(\mu_x, \Sigma_x)$ and $X \sim \pi$ with $\pi \sim Dir(\alpha)$ (see also Liu and Rubin (1998) who generalize the $(Y \mid X)$ model to the class of elliptically symmetric distributions). Estimation under this model is infeasible unless each cell contains a large number of observations, so further constraints are necessary. Typical restrictions include assuming $\Sigma_x \equiv \Sigma$ for all x, $\mu_x = D(x)B$ for a matrix of regression coefficients B and design vector $D(x)$, and loglinear constraints on π to include interactions only up to a certain order. The general location model is already somewhat limited by its assumption of conditional multivariate normality, and imposing a common covariance structure and limited interactions makes it quite restrictive.

Given the difficulty of specifying proper joint probability models for mixed out-

comes, many have advocated instead specifying a sequence of univariate models for each variable subject to missingness conditional on all the others. This is known as the "chained equations" or "fully-conditional" approach (Raghunathan et al., 2001; Van Buuren and Oudshoorn, 1999). Such a collection of conditional models may not correspond to a proper joint model, which somewhat undermines the theoretical justifications behind MI.

Perhaps more important than theoretical concerns are the practical challenges in implementing the chained equations approach. It requires specifying each univariate model, which is labor-intensive and challenging with even a moderate number of variables. For example, categorical variables are most often implemented via a conditional multinomial logit model, which requires a distinct coefficient vector for each level of each variable. Even in fairly large datasets these coefficients may be poorly estimated, especially as we consider interaction terms. In practice most if not all variables are imputed using default choices such as logistic regression or predictive mean matching using main effects of all the other variables. In modest-sized datasets (e.g., a few hundred cases) with low fractions of missing information the bias induced by the inevitable model misspecification can be small relative to the combined imputation/complete data variance, but this is much less likely in larger datasets like federal surveys.

Mixture models are a flexible alternative to traditional parametric joint models for multiple imputation. Mixtures can capture a range of distributional shapes and complex dependence. In some specific settings these models have proven useful for MI (e.g. Böhning et al. (2007); Elliott and Stettler (2007); Vermunt and Ginkel (2008); Gebregziabher and DeSantis (2010); Si and Reiter (2013); Kim et al. (2013); Manrique-Vallier and Reiter (2012a,b)). However, to our knowledge there have been no applications of mixture modeling to imputing mixed continuous and categorical data. In the next section we introduce a Bayesian nonparametric mixture model

which generalizes many of these existing mixture models (and the general location model) to multivariate continuous, ordered and unordered categorical data.

3.3 A Bayesian Nonparametric Model for MI of Mixed Data

Let $X_i = (X_{i1}, \ldots, X_{ip})'$ be a vector of p ordered and/or unordered categorical variables for respondent i, with X_{ij} taking values in $1, 2, \ldots, d_j$. Let $Y_i = (Y_{i1}, \ldots, Y_{iq})'$ be a vector of continuous variables taking values in $\mathbb{R}^q$. In Section 3.5 we discuss extending this model to include other variable types, such as counts. Note that the use of X and Y is for convenience, not to suggest "predictors" and "responses".

Our approach, which we call a hierarchically coupled mixture model with local dependence (HCMM-LD), begins with separate mixture models for X and Y and combines them through dependent cluster assignment and local regressions. We define $H_{ix} \in \mathbb{N}$ and $H_{iy} \in \mathbb{N}$ to be component indices for the X and $Y \mid X$ models, respectively, with $Z_i \in \mathbb{N}$ a third top-level component index.

3.3.1 Data model

Given the component indices we assume that

$$\Pr(X_i = x \mid H_{ix} = h_x, \{\psi_h\}_{h=1}^{\infty}) = \prod_{j=1}^{p} \psi_{h_x x_j}^{(j)} \tag{3.1}$$

$$(Y_i \mid X_i = x, H_{iy} = h_y, \{B_h, \Sigma_h\}_{h=1}^{\infty}) \sim N(D(x)B_{h_y}, \Sigma_{h_y}). \tag{3.2}$$

The parameters for each X component are given a Dirichlet prior

$$\{\psi_{h_x}^{(j)} : h_x \in \mathbb{N}\} \overset{iid}{\sim} Dir(\gamma_j) \text{ for } 1 \leq j \leq p. \tag{3.3}$$

Reasonable default choices for γ_j include $(1, 1, \ldots, 1)$ or $(1/d_j, 1/d_j, \ldots, 1/d_j)$. We prefer the latter, but find our results are usually insensitive to this choice. Alternatively prior information about the marginal distribution for X_j could be included here.

The parameters for each Y component are given hierarchical priors:

$$\{(B_{h_y}, \Sigma_{h_y})\} \overset{iid}{\sim} MatN(B, I, T_B) \times IW(d, \Sigma) \tag{3.4}$$

$$(B, \Sigma) \sim MatN(B_0, I, \sigma_0^2 I) \times W(c, \Sigma_0) \tag{3.5}$$

where $MatN(\mu, \Phi, \Sigma)$ is the matrix normal distribution, i.e. the distribution of the $p^* \times q$ dimensional matrix $\Phi^{1/2} S \Sigma^{1/2}$ when S is $p^* \times q$ with $s_{ij} \overset{iid}{\sim} N(0,1)$. We assume that $T_B = \text{diag}(1/\tau_1, \ldots, 1/\tau_q)$ and assign $\tau_1, \ldots \tau_q$ independent $G(0.5, 0.5)$ priors. In our applications we find the posterior predictive distributions to be largely insensitive to this choice, but this may not always be true. In particular, if the additive mean model is nearly correct for one or more elements of Y, then some of these variance components will tend toward zero. In this case prior specification will become more important, and different parametrizations or more sophisticated MCMC techniques may be warranted (see e.g. Gelman (2006)). It may also be advantageous to allow these variance components to vary by j, the corresponding entry in X, or to further model them if $D(X)$ includes interactions.

To complete the hyperprior, note that $E(\Sigma_h) = \frac{c}{d-q-1}\Sigma_0$. Absent good *a priori* information about the scale of Y, a weakly data dependent prior can be derived by scaling the data and taking $c = q+1$, $d = q+2$ and $\Sigma_0 = \frac{1}{q+1} I$. In sufficiently large samples inferences are insensitive to the choice of (B_0, σ_0^2); we use $(0, 2)$.

Component index model

The component indices are given the hierarchical prior

$$\Pr(H_{ix} = h_x, H_{iy} = h_y \mid Z_i = z) = \phi^{(x)}_{zh_x} \phi^{(y)}_{zh_y} \tag{3.6}$$

$$\Pr(Z_i = z \mid \lambda) = \lambda_z. \tag{3.7}$$

Each $\phi_z^{(s)}$ for $s \in \{x, y\}$ and $z \in \mathbb{N}$ is constructed independently as a stick breaking process, as is λ:

$$\phi_{zh}^{(s)} = V_{zh}^{(s)} \prod_{l<h} (1 - V_{zl}^{(s)}), \ \{V_{zh}^{(s)} : z \in \mathbb{N}, h \in \mathbb{N}\} \overset{iid}{\sim} Beta(1, \beta_s) \tag{3.8}$$

$$\lambda_h = W_h \prod_{l<h} (1 - W_l), \ \{W_h : h = 1, 2, \dots\} \overset{iid}{\sim} Beta(1, \alpha). \tag{3.9}$$

Banerjee et al. (2013) establish that this is a well-defined prior, which they call an *infinite tensor factorization (ITF)* prior. The parameters α, β_x and β_y are each assigned gamma hyperpriors with shape and rate parameters equal to 0.5.

Marginalizing Z gives $\Pr(H = (h_x, h_y)) = \sum_{z=0}^{\infty} \lambda_z \phi_{zh_x}^{(x)} \phi_{zh_y}^{(y)}$, inducing dependence between H_x and H_y. If we imagine the infinite probability mass function for (H_x, H_y) arranged in a matrix, most of the mass is near the top-left corner and along the first row and column. We discuss the benefits of this choice in the next section.

3.3.2 Properties of the HCMM-LD

A key feature of the HCMM-LD is its *local dependence*. First, we allow the relationships within Y as well as between X and Y to vary across Y components. This makes hierarchical priors on (B_{h_y}, Σ_{h_y}) essential. A simpler, standard prior choice for $\{(B_{h_y}, \Sigma_{h_y})\}$ would have them iid from some prior. Since many components will have few or no observations with any particular X_j value, the corresponding coefficient would be drawn from the prior if we do not allow borrowing of information about B_{h_y} across components. The hierarchical prior still allows larger components to adapt to local changes in the impact of X on Y. Similarly, if the Σ_{h_y} do not have a hierarchical prior then many components will have covariance essentially drawn from the prior, leading to imputations for Y that can vary wildly. Assuming the relationships between the elements of Y are locally linear but varying across clusters allows the model to capture globally nonlinear features in the distribution of Y, while the

hierarchical prior ensures that small components have reasonable parameter values.

Another form of local dependence results from (3.6) and (3.7), the hierarchical model for the component indices. It is straightforward to derive the distribution of (Y, X) marginalizing over the lower-level cluster indices H_x and H_y:

$$f(X, Y \mid Z = z) = \left(\sum_{h_y=1}^{\infty} \phi_{zh_y}^{(y)} N(Y; D(X)B_{h_y}, \Sigma_{h_y}) \right) \left(\sum_{h_x=1}^{\infty} \phi_{zh_x}^{(x)} \prod_{j=1}^{p} \psi_{h_x X_j}^{(j)} \right). \quad (3.10)$$

Therefore, within top-level clusters we have local dependence within X (which is, to our knowledge, novel for mixture models with multivariate categorical data) as well as between X and Y and within Y. Note that $f(X, Y \mid Z = z)$ and $f(X, Y \mid Z = z')$ will differ only in their respective lower-level stick breaking weights $(\phi_z^{(y)}, \phi_z^{(x)})$ and $(\phi_{z'}^{(y)}, \phi_{z'}^{(x)})$, providing shrinkage somewhat analagous to the hierarchical prior on the Y component parameters.

Equation (3.10) shows how the HCMM-LD may be cast as a "mixture of mixture models" by marginalizing over Z. For any z, $(X \mid Z = z)$ follows a Dirichlet process mixture of product multinomials (MPMN) model as described by Dunson and Xing (2009). The distribution of $(Y \mid X, Z = z)$ is very close to the ANOVA-DDP model introduced by De Iorio et al. (2004), except we use a normal location-scale mixture rather than a location mixture (and hierarchical priors). Again, these distributions are not independent across z due to the sharing of atoms corresponding to lower-level components.

The HCMM-LD has interesting limiting cases. If $\beta_y \to 0$ then $f(Y \mid X, Z = z)$ has a single cluster and is therefore equivalent to a multivariate ANOVA model. The resulting model for (X, Y) is essentially a general location model, except the Dirichlet marginal model for X is replaced with the MPMN model. There are reasons to prefer the latter: The MPMN model can represent any probability distribution whereas the Dirichlet prior requires restrictive assumptions about the depth of interactions in X,

and the latent class structure of the MPMN model also makes posterior inference and simulating from posterior predictive distributions straightforward. When $\alpha \to 0$ there is a single top-level cluster index, in which case X is once again the MPMN and the model for $(Y \mid X)$ is the location-scale ANOVA-DDP. This simpler model is somewhat limited compared to the HCMM-LD; for example, $E(Y \mid X = x)$ in this reduced case is $E(Y \mid X) = \sum_{h_y=1}^{\infty} \phi_{h_y}^{(y)} D(X)\beta_h \equiv D(X)\tilde{\beta}_h$. This is a well-known limitation of "single-p" dependent Dirichlet processess (MacEachern, 2000). Compare this to the conditional distribution for Y under the full HCMM-LD:

$$f(Y \mid X = x) = \sum_{h=1}^{\infty} \frac{w_h(x)}{\sum_{l_y=1}^{\infty} w_l(x)} N(Y_i; D(x)B_h, \Sigma_h) \tag{3.11}$$

where $w_h(x) = \sum_{h_x,z} \lambda_z \phi_{zh_x}^{(x)} \phi_{zh_y}^{(y)} \prod_{j=1}^{p} \psi_{h_x x_j}^{(j)}$. This form of the weights depends on X, allowing for unanticipated interactions in the distribution of $Y \mid X$. Further, because it does not separate into the product of separate terms for each X_j (due to local dependence) it allows for meaningful interactions in the conditional distribution of H_y given X, which is a unique feature of the HCMM-LD compared to existing joint mixture models.

3.3.3 Related work

Dunson and Xing (2009) proposed a tensor factorization model for multivariate categorical data which assumes that the joint probability mass function of the observations is decomposed as in (3.1), with a stick-breaking prior on H_x, and showed that any probability tensor can be decomposed in this fashion. Si and Reiter (2013) illustrated the utility of this model for multiple imputation of categorical data, and Manrique-Vallier and Reiter (2012a,b) showed how to incorporate structural zeros in the contingency table. Vermunt and Ginkel (2008) proposed a similar latent class mixture for imputing multivariate categorical data, as did Gebregziabher and De-

Santis (2010).

A generalization of model in Dunson and Xing (2009) to include additional univariate kernels of different types was developed by Dunson and Bhattacharya (2010). However, the authors note that when the number of variables grows the number of clusters must also grow to accommodate the dependence in the joint distribution. This is due to the local independence assumptions of the latent class or product kernel formulation, which forces all the dependence to be represented through a single cluster index. The HCMM-LD is able to avoid such strong local independence assumptions.

A number of authors have proposed other joint mixture models that do include some limited local dependence, typically as a way to induce priors on conditional distributions of interest. For mixed data, a common approach is to decompose the joint kernel into a conditional kernel for the outcome of interest and a marginal product kernel for predictors. Shahbaba and Neal (2009) and Molitor et al. (2010) proposed DP mixtures with kernels of this form for categorical outcomes, and Hannah et al. (2011) provided extensions to more general cases and asymptotic theory. This formulation incorporates local dependence between predictors and the response but not within the predictors, and the assumption of local independence between the preditors can lead to the same proliferation of clusters as in the latent class model (Hannah et al. (2011) includes some discussion of this phenomenon).

Dunson and Bhattacharya (2010) suggest in their discussion that one might overcome the proliferation of clusters by instead coupling a series of mixture models through *dependent* cluster assignment. This was later implemented by Banerjee et al. (2013) who introduce the ITF prior and use it to couple univariate mixture models for each variable. Compared to the HCMM-LD this model allows only weak local dependence (through shared lower-level components within top-level components). The local dependence features of the HCMM-LD allow us to simultaneously

avoid the proliferation of clusters in the latent class model and center our model on a reasonable alternative (the MPMN model for X and a multivariate regression or ANOVA-DDP for $Y \mid X$).

Wade et al. (2011) proposed the enriched DP, which is somewhat similar to the ITF in that it induces dependent cluster assignment. The enriched DP separates a joint distribution into a conditional and a marginal, and assigns each each a DP prior (where the base measure for the conditional varies across the marginal). However, the enriched DP lacks the symmetry of the ITF, which makes it difficult to understand the induced joint distribution. This is unappealing, since our partition into "X" and "Y" is essentially arbitrary. The enriched DP is also unable to collapse to the ANOVA-DDP conditional and MPMN marginal model, which is a desirable feature of the HCMM-LD.

Finally, Canale and Dunson (2011) model mixed ordered data (count, ordinal and continuous) by thresholding latent variables which follow a DP mixture (see also Kottas et al. (2005) for the ordinal case). In the HCMM-LD it would be straightforward to include ordinal and count variables via data augmentation in a similar fashion, although there is a nontrivial increase in computation when doing so. While this modeling framework could in principle be extended to include unordered categorical data via latent "utilities" (similar to the multinomial probit data augmentation in Albert and Chib (1993)) the resulting model would be quite difficult to specify, and latent mixture model would be very high dimensional. The HCMM-LD strikes a balance between flexibility and tractability that is often advantageous in applications.

3.4 SIPP simulations

To evaluate the performance of HCMM-LD in multiple imputation, we conducted a repeated sampling simulation study on a population taken from the first wave of the 2008 SIPP panel. We define the population as individuals who reported

Table 3.2: Variables in the SIPP simulation study

Variable	Levels
Monthly earnings from employment	Continuous
Age	Continuous
Gender	2
Race	5
Marital Status	6
Born in the US	2
Number of own children in the home	4 (0,1,2, or 3+)
Education level	6
Occupation	23
Worker Class	3 (Private, Nonprofit, Government)
Union	2
Hourly	2
Usual Hours worked	9 (0-80 in increments of 10 hours, 80+)

positive income from work during the reference period in Wave 1, and exclude records with missing entries (less than 1%, not counting initial nonresponders). Our final population consists of 30,507 respondents. We created 500 datasets with missingness by taking simple random samples of size $6,000$ and introducing approximately 35% missingness completely at random while ensuring that each dataset had about 500 complete cases.

With guidance from Census staff we selected the 13 variables (2 continuous and 11 categorical) listed in Table 3.2. We chose a modest number of variables to make a large simulation study more efficient while keeping the problem challenging. For example, the contingency table formed by the 11 categorical variables has over 7 million cells and is therefore very sparse. Further discussion of the scalability of the HCMM-LD is deferred to Section 3.5.

We compare the HCMM-LD to multiple imputations via chained equations as implemented in the R package MICE (van Buuren and Groothuis-Oudshoorn, 2011). Our goal is to compare default procedures so we did not alter any of the options to MICE except that we generate $M = 10$ imputed datasets instead of the default

$M = 5$. The default procedure imputes continuous variables via predictive mean matching (Little, 1988) and uses logistic regressions to impute discrete variables. Each conditional model includes a main effect for every other variable.

For the HCMM-LD we use the default priors described in Section 3.3 and center and scale the continuous variables. We included main effects for each of the variables in Table 3.2 in the design vectors $D(X_i)$. We perform $200,000$ MCMC iterations, using the sampler described in Appendix B, discarding the first $100,000$ iterations and keeping the imputations from every $10,000^{th}$ iteration thereafter. This is very conservative; examination of a handful of datasets suggests that these numbers could be reduced by at least half without impacting the results. In practice the imputer should carefully examine MCMC diagnostics of relevant identified parameters, such as marginal means, quantiles, and variances or covariances of the imputed data. We ran the simulations in a heterogenous cluster environment so the run times varied, but a mid-range desktop machine completed 10,000 iterations of the MCMC sampler in about 45-55 minutes.

After imputing the data, we compute a number of estimates in each completed dataset and combine them using "Rubin's rules" (Rubin, 1987). Complete data estimates were computed using the survey package in R (R Development Core Team, 2011; Lumley, 2004), including a finite population correction.

3.4.1 Results

Cell Means

We begin by examining cell means of log monthly earnings, conditioning on various subsets of the other variables. In each case we restrict to cells with expected counts of at least 30. It would also be possible to consider untransformed incomes, but due to the skewness of the income distribution we combine imputations on the log scale where normal approximations are more likely to hold. In applications we can

transform to the original scale after pooling, e.g. using the delta method.

Figure 3.2 shows the coverage rates and average width of 95% pooled confidence intervals for the cell means of log monthly earnings by age (discretized into 10 year intervals except for < 18, $18 - 25$, and 65+), sex and presence of own children. The HCMM-LD imputations are clearly superior: About half of MICE's intervals have coverage under 75%, with several more under 25% and some approaching 0%. In contrast, the worst coverage rate using the HCMM-LD is just under 75% with the majority near or greater than the nominal 95% rate. Of the estimates where the HCMM-LD imputations undercover, MICE imputations undercover to a much greater degree. The widths of the confidence intervals are comparable, and there are a number of instances where the HCMM-LD's coverage rates are better with shorter intervals. This suggests that MICE's lack of coverage is due to bias, which is confirmed by Figure 3.3. Overall the range of bias under our method is substantially smaller when standardizing the bias by the standard deviation of the pooled estimate (left) or the true cell mean (right). For MI this is desirable, since we seek to generate valid complete data inferences across a wide range of estimands.

A significant factor driving this difference is the complex relationship between age and income. Earnings tend to be lowest in the young (SIPP recordings earnings information on respondents 15 or older), increasing during working years and falling off again as those who can afford to retire do so. Additionally, the variance in earnings is low in the younger cohort, roughly stable through the working years, and increasing near and after retirement age. The bias also appears to be a function of sample size; for example, in 18-24 year olds MICE's standardized bias is 4.2 SDs for men with children (N=449) and 3.0 for women with children (N=564), compared to 2.5 (0.4) for men (women) without their own children in the home (N=1,362 and 1,204, respectively). Interactions also appear to be at play; the effect of having their own child in the home varies across the respondent's age, due in part to its high

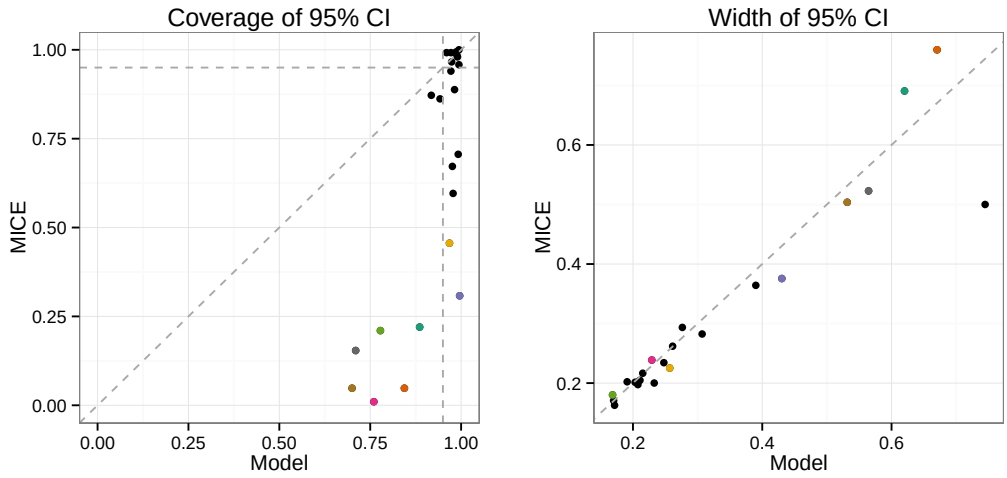

FIGURE 3.2: (Left) Coverage rate of pooled nominal 95% CI for mean log monthly earnings by age, gender, and own children in the home (Yes/No) (Right) Average CI width of 95% CI.

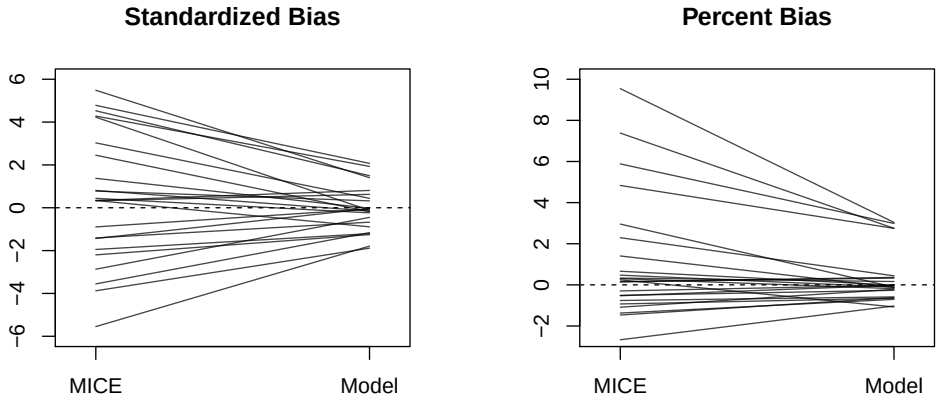

FIGURE 3.3: Standardized (left) and percent (right) bias of pooled estimates of population means, by age (18 and under, 19-25,25-65 in ten year increments, and over 65) and presence of own child under 18 in the household. Each line represents a cell mean, with the left and right endpoints at the bias under MICE and HCMM-LD, respectively.

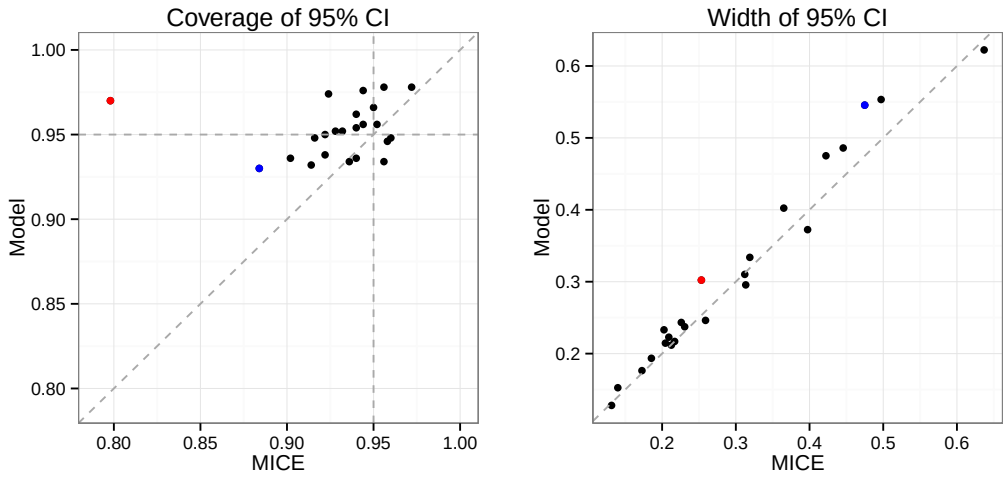

FIGURE 3.4: (Left) Coverage rate of pooled nominal 95% CI for mean log monthly earnings by occupation. (Right) Average width of 95% CI.

correlation with the age of the children, and across the genders as well. For example, the population difference in log wages for 18-24 year old women is -0.159, versus -0.076 for 35-44 year old women. In men the population differences are -0.064 for 18-24 year olds and 0.232 for 35-44 year olds.

Figures 3.4 and 3.5 show the coverage rates and average widths of the pooled confidence intervals for cell means by occupation and by occupation and education level (respectively). For occupation alone coverage under the HCMM-LD within 93-98% for each cell. MICE does well in most cases too, although coverage dips to about 80% in one case. For occupation and education, nearly all the HCMM-LD CIs have coverage over or just under 95%, with a single exception at 85%. Most of the MICE intervals have the advertised coverage, but there are a few that dip below 90% and one that falls to about 65%. In both cases the overall range of standardized or percent bias is lower under the model-based imputations than under MICE.

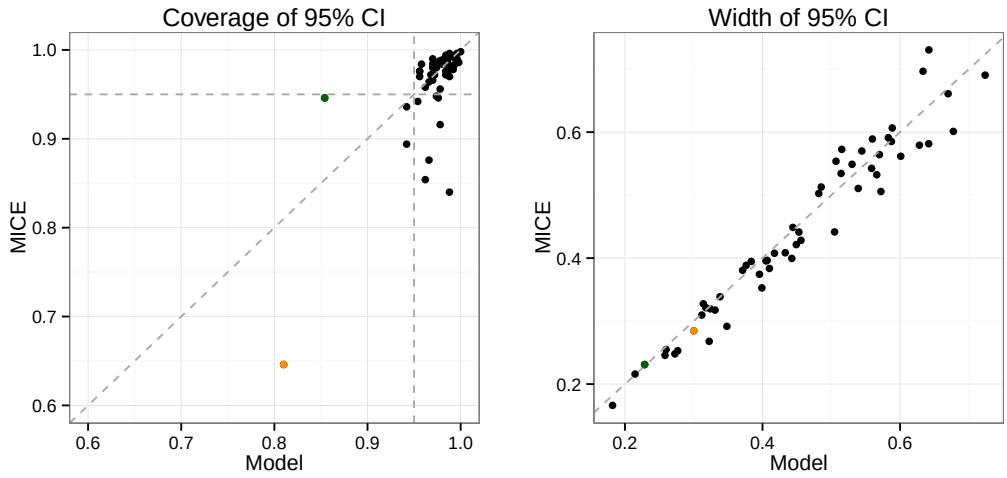

FIGURE 3.5: (Left) Coverage rate of pooled nominal 95% CI for mean log monthly earnings by occupation and education level. (Right) Average width of 95% CI.

Regression Coefficients

Next we consider linear regressions of log earnings on age, gender, usual hours worked (recoded as < 30, 30-60, and 60+), and indicators for married with spouse present and own child under 18 in the household. To begin we fit a model including an age squared term as well as two- and three-way interactions between sex, own child, and marital status. Figure 3.6 shows pooled estimates of the coefficients and the average width of their confidence intervals. The HCMM-LD imputations are clearly superior. Including the squared term in age is challenging for both methods, since it tends to give high leverage to points at low and high age values and neither method has been modified to anticipate the nonlinear relationship. However, the HCMM-LD still offers over 50% coverage rates for the age coefficients and the intercept, whereas the coverage rate under MICE drops to zero.

Figure 3.7 shows results from the same 3-way regression model without the age squared term. Coverage is generally improved for both methods but the HCMM-LD tends to have better coverage rates, particularly for the two way interactions.

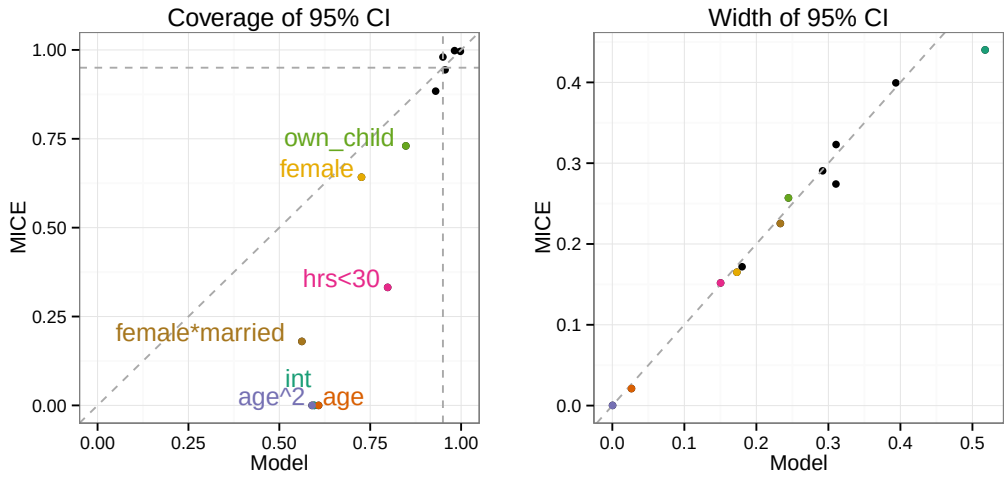

FIGURE 3.6: (Left) Coverage rate of pooled nominal 95% CI for the regression with three-way interaction and age squared. (Right) Average width of 95% CI.

Under both methods the interactions are pulled toward zero, but more so with MICE compared to the model. Finally, as a point of reference we also consider the regression with main effects only (Fig. 3.8). We expected MICE to do well here, since this is a submodel of the actual regression used in its predictive mean matching imputation. However, there is one case where MICE's coverage dropped to about 80%, compared to 95% under the HCMM-LD. This is the coefficient for hours worked > 60, and the lack of coverage appears to be due to the relatively small sample size of this group (807 in the population) and large effect (about 0.25 in this particular model), which combine to make predictive mean matching less effective. The average pooled estimate from MICE was 0.17 compared to 0.25 from the model, and the average width of the CI was similar between the two. Figure 3.9 shows that adding the age squared term to this model has similar effects as in the three way interaction model, pulling down coverage for the intercept, age and age squared terms to zero for MICE about about 60% for the model. The model based imputations also have better coverage for the remaining coefficients in this case.

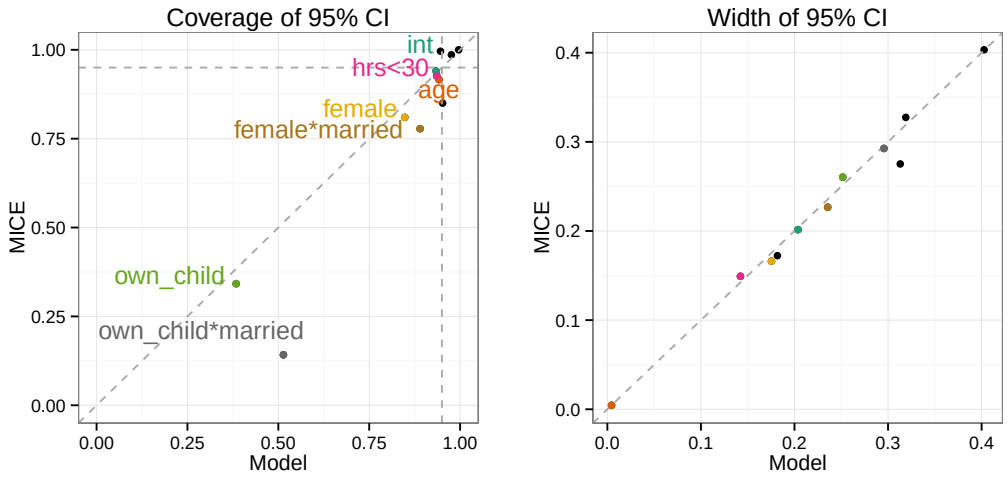

FIGURE 3.7: (Left) Coverage rate of pooled nominal 95% CI for regression with three-way interaction *without* age squared. (Right) Average width of 95% CI.

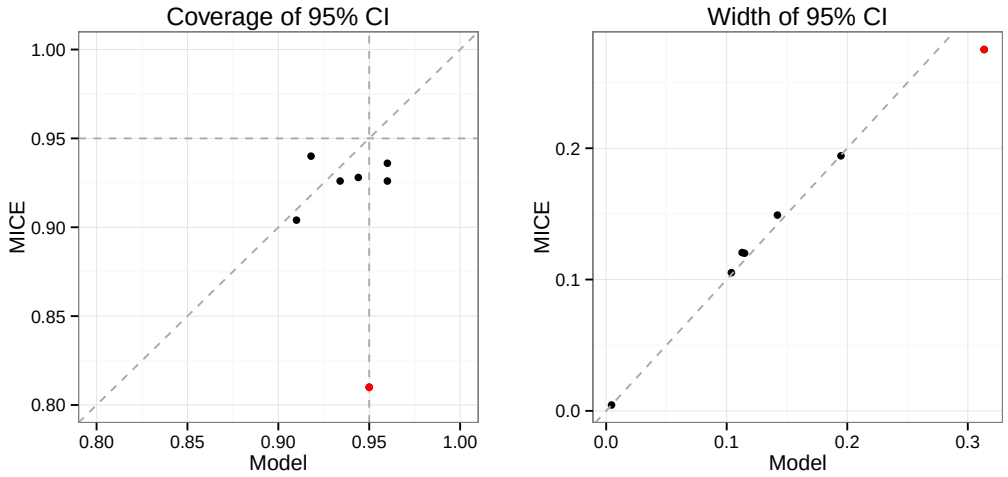

FIGURE 3.8: (Left) Coverage rate of pooled nominal 95% CI for the main effects regression.(Right) Average width of 95% CI.

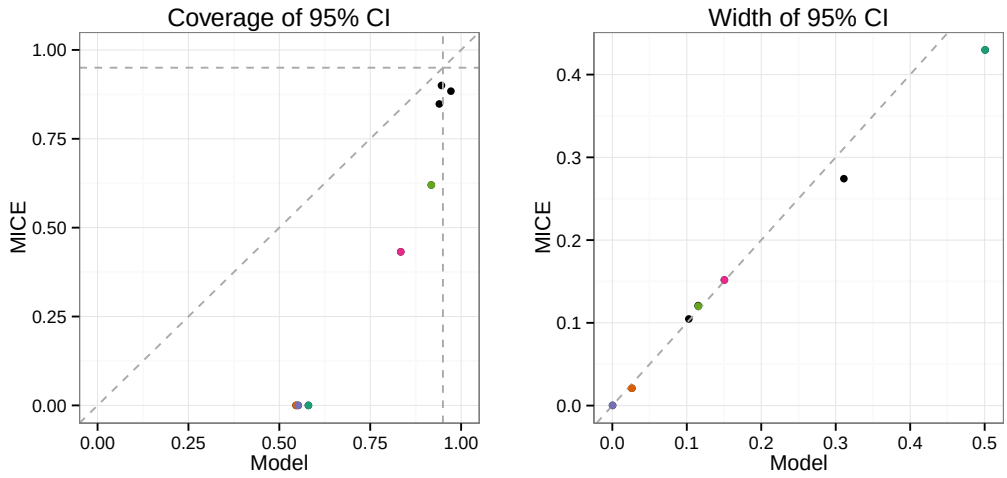

FIGURE 3.9: (Left) Coverage rate of pooled nominal 95% CI for the main effects regression plus an age squared term.(Right) Average width of 95% CI.

Conditional Frequencies

We also examine the quality of categorical imputations by estimating cell frequencies of categorical variables. In all cases we restrict to cases where $E(N_c) \times p \geq 10$ and $E(N_c) \times (1-p) \geq 10$, where p is the true proportion and N_c is the cell size in our simple random samples, to make the normal approximation somewhat more plausible. Figure 3.10 displays results from estimating the proportion of respondents with their own child under 18 in the home by sex, race and age. The HCMM-LD based imputations perform much better than MICE, for which some coverage rates drop all the way to zero. Coverage rates for the HCMM-LD never drop below 60% and are better than MICE in every case but one. Figure 3.11 shows that MICE has very good or very poor coverage in large cells, consistent with the lack of coverage arising from misspecification bias. The HCMM-LD tends to have somewhat lower coverage in the larger cells than in the smaller cells, but not nearly to the extent of MICE. This is probably due to finite-sample bias; larger cells are more sensitive to finite sample bias since the complete data standard errors are smaller. Figure 3.12

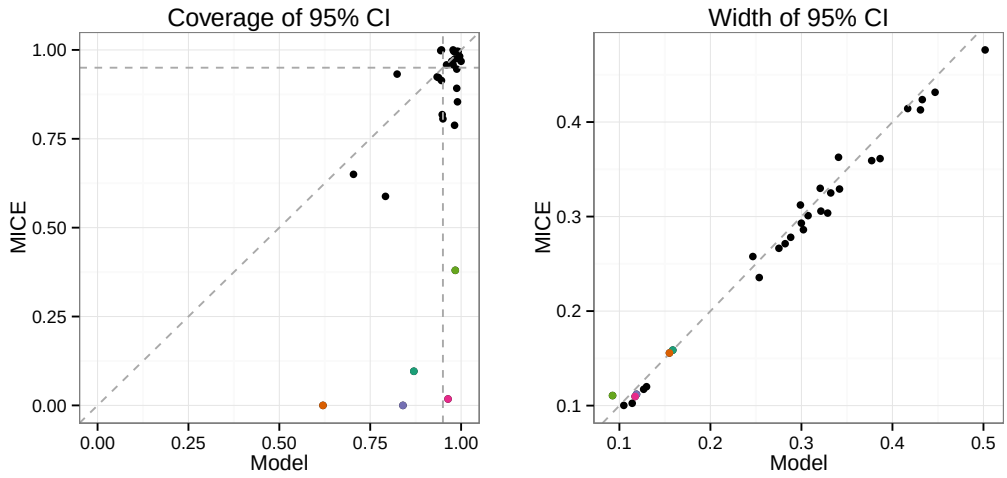

FIGURE 3.10: (Left) Coverage rate of pooled nominal 95% CI for proportion with own child < 18 in the household by age, race and sex. (Right) Average width of 95% CI.

shows results from estimating the proportion of usual hours stratifying on marital status, gender and number of own children in the home (0, 1, 2 or 3+). Coverage is often similar between the two methods, but occasionally much better under the model based imputations.

3.5 Conclusion and future work

We have introduced a Bayesian nonparametric joint model for mixed continuous and categorical data and demonstrated with repeated sampling simulations using SIPP data that, as a default imputation engine, it can substantially improve on the most popular competing method across a range of estimands. This runs counter to the prevailing wisdom that joint models are not competitive with imputation by chained equations (comments to this effect appear in Gelman (2004); van Buuren (2007); Stuart et al. (2009); He et al. (2010); Drechsler (2010), among others). Our comparison has been between default implementations of each method. Either could be modified to incorporate dataset-specific prior knowledge – and this is good practice

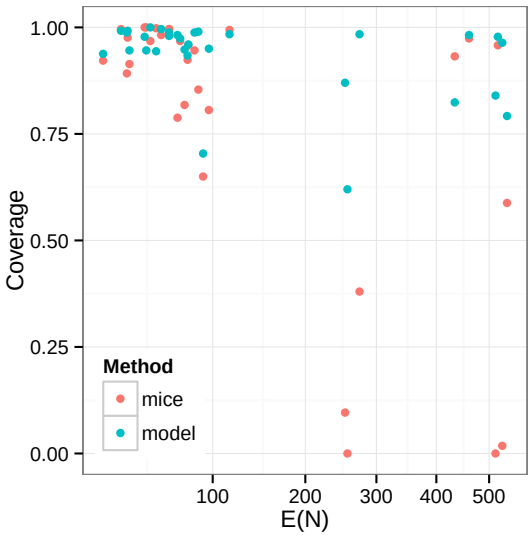

FIGURE 3.11: Coverage by expected cell size for proportion with own child < 18 in the household by age, race and gender.

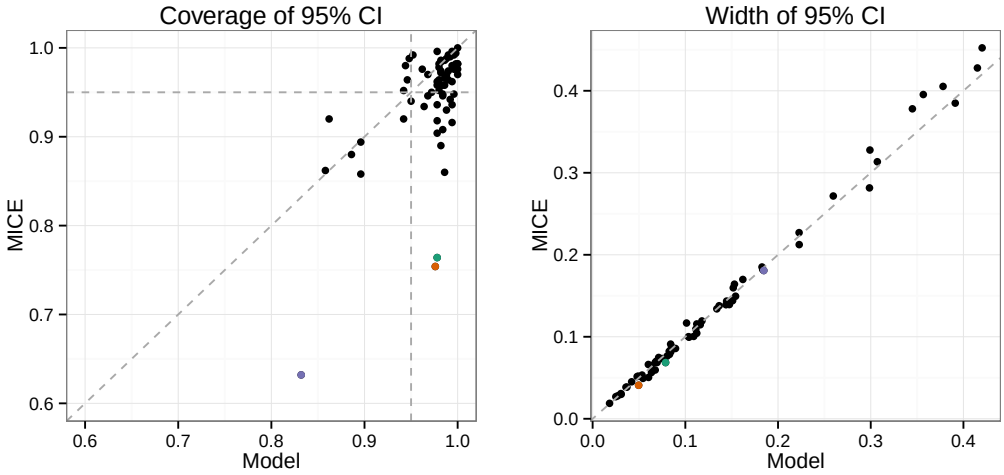

FIGURE 3.12: (Left) Coverage rate of pooled nominal 95% CI for cell frequencies of usual hours worked by marital status, gender and own child. (Right) Average width of 95% CI.

– but default procedures are valuable when this information is lacking or difficult to translate into a statistical model. They are also useful testbeds for examining an imputation method's ability to capture unanticipated structure in data.

The examples in Section 3.4 are representative of many other estimands we examined. For many estimands the difference between MICE and HCMM-LD is modest, but for others the improvement under the model-based approach can be dramatic. Rarely is the performance significantly worse under the HCMM-LD than under MICE. We suspect that as the sample size grows large the performance gap will increase, as the differences appear to be driven mostly by misspecification bias. In large samples the complete data standard errors will be smaller, making pooled confidence intervals more sensitive to any bias introduced by the imputation procedure. The HCMM-LD has the potential to increase in complexity and capture additional features of the data, unlike MICE.

Although we have not performed a systematic evaluation of the properties of the HCMM-LD in large sample, high-dimensional settings, we are optimistic about its performance. Computationally, fitting the HCMM-LD reduces to fitting a series of mixture and regression models. Computation time scales roughly linearly with sample size, the primary bottleneck being the computation of likelihoods when resampling cluster indices (although of course larger samples will also require longer MCMC runs). These steps could be further optimized in our existing code. Increasing the dimension of X is clearly feasible; for example, in an MPMN model Si and Reiter (2013) considered some simulations with some 50 categorical variables. Increasing q, the dimension of Y, is more of a strain since aspects of posterior inference require $O(q^3)$ operations. However, fitting large mixtures of multivariate normals is a well-studied problem and a number of specialized, efficient algorithms exist, for example utilizing parallel architectures to compute likelihoods. Alternative priors for component parameters can provide computational benefits and additional regulariza-

tion in higher dimensions, for example, by assuming a factor-analytic decomposition for the covariance matrices. Any of these methods are straightforward to adapt to the HCMM-LD.

There are a number of interesting directions to extend this work. We have focused on jointly modeling all the variables, but in practice it would be appealing to avoid modeling completely observed variables (especially design variables). The modular nature of the HCMM-LD makes it conceptually straightforward to incorporate fully observed covariates, and we are currently evaluating the best ways to do so. We would like to incorporate other types of variables, such as counts or durations. Finally, structural zeros in contingency tables (from impossible combinations or skip patterns), semicontinuous variables and range or inequality restrictions are all common complications in MI. Linear restrictions in MI of continuous variables are considered by Kim et al. (2013) who utilize truncated mixture models. Manrique-Vallier and Reiter (2012a,b) also used truncation to include structural zeros in contingency tables. Adapting these approaches to mixed data, where constraints on continuous variables depend on the values of categorical variables and vice-versa is an active area of research.

4

Density Estimation and Regression Using Spline Transformation Priors

4.1 Introduction

Discrete mixture models have been the workhorse of nonparametric Bayesian density estimation, due largely to the tractability of the Dirichlet process. In more recent years considerable effort has been devoted to adapting these methods for density *regression*, which models a collection of densities $\{f(y \mid x) : x \in \mathcal{X}\}$. There are a range of applications where mean regression models are inadequate; for example in regression models for income it is usually the median or some other quantile that is of interest. Often we want to consider multiple quantiles simultaneously, or some other feature of the distribution (tail probabilities, skewness, and so on). These problems are most naturally approached by allowing the entire distribution to depend on covariates.

In this chapter we develop a flexible but tractable model for density estimation and regression problems based on continuous latent variable models. The chapter proceeds as follows. In Section 4.2 we introduce the model for a single density,

including prior specification, and illustrate its use in simulations. In Section 4.3 we extend this model to density regression, and illustrate with an example using data from the 1990 and 2000 U.S. Census to examine changes in wage structure. Finally, in Section 4.4 we discuss a number of possible extensions to this work.

4.2 Transformation priors for a single density

We begin with the class of transformation models suggested by Kundu and Dunson (2011):

$$y_i = g(u_i) + \epsilon_i \tag{4.1}$$

where $u_i \sim U(0,1)$, $\epsilon_i \sim N(0, \sigma^2)$ are independent. The u_i are latent variables, inducing a model for the density of y_i through marginalization:

$$f(y \mid \sigma) = \int_0^1 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-g(u))^2} d\,u. \tag{4.2}$$

In the limit as $\sigma_y \to 0$, $Y \stackrel{d}{=} g(U)$ where $U \sim U(0,1)$. Pati et al. (2011) provide a number of theoretical tools to assess the prior support of this model. Intuitively, if g is the pseduo-inverse of a distribution function F, then $Y \sim F$ and therefore this class of models is quite flexible. It also generalizes the normal location mixture model: If g is a step function with $g(u) = \mu_h$ for $u \in [\nu_h, \nu_{h+1})$ for an increasing sequence ν on $[0,1)$ such that $\nu_1 = 0$ and $\sum_{h=0}^{\infty}(\nu_{h+1} - \nu_h) = 1$, then we have

$$f(y \mid \sigma_y) = \int_0^1 \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2\sigma_y^2}(y-\mu_h)^2} \mathbf{1}(u \in [\nu_h, \nu_{h+1}))d\,u \tag{4.3}$$

$$= \sum_{h=1}^{\infty}(\nu_{h+1} - \nu_h)\frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2\sigma_y^2}(y-\mu_h)^2}. \tag{4.4}$$

The representation in (4.4) is intimately related to the augmented model used in slice sampling mixture models (Walker, 2007; Kalli et al., 2011) where the prior on mixture weights induces a prior on the sequence ν in an obvious way.

Smooth functions g yield an *uncountably* infinite location mixture of normals. Other location models could be used in place of the normal by assuming another distribution for ϵ_i. Kundu and Dunson (2011) propose a Gaussian process prior on g and suggest a squared exponential covariance kernel. This is a flexible choice, but it has some drawbacks. Motivated by computational considerations, Kundu and Dunson (2011) discretize the support of u_i, i.e. $\Pr(u_i = u_h) = 1/k$ for some $0 < u_1 < u_2 < \cdots < u_k < 1$. This turns the model into a finite location mixture of normals, with a prior that encourages smoothness on the locations and a uniform mixing distribution:

$$f(y \mid \sigma_y) = \int_0^1 \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2\sigma_y^2}(y-g(u))^2} \, d\, u \tag{4.5}$$

$$= \sum_{h=1}^{k} \frac{1}{k} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2\sigma_y^2}(y-g(u_k))^2}. \tag{4.6}$$

The approximation to (4.1) is clearly improving with $k \to \infty$, but so is the computation required to update each u_i in a Gibbs sampler, and a very fine grid can lead to numerical problems in updating the Gaussian process.

We propose instead modeling g as a linear combination of polynomial splines with a fixed number of varying knots, which we will call the transformation spline density model (TSDM). With Gaussian priors on the coefficients of the spline basis expansion this model is also a Gaussian process with a particular covariance function (Rasmussen and Williams, 2005, Chapter 6), although we will consider other priors as well. While this generally comprises a smaller class of possible functions g, it remains flexible while allowing for efficient posterior inference without discretization.

Specifically, we assume:

$$g(u) = \sum_{h=1}^{k} \beta_h b_h(u), \tag{4.7}$$

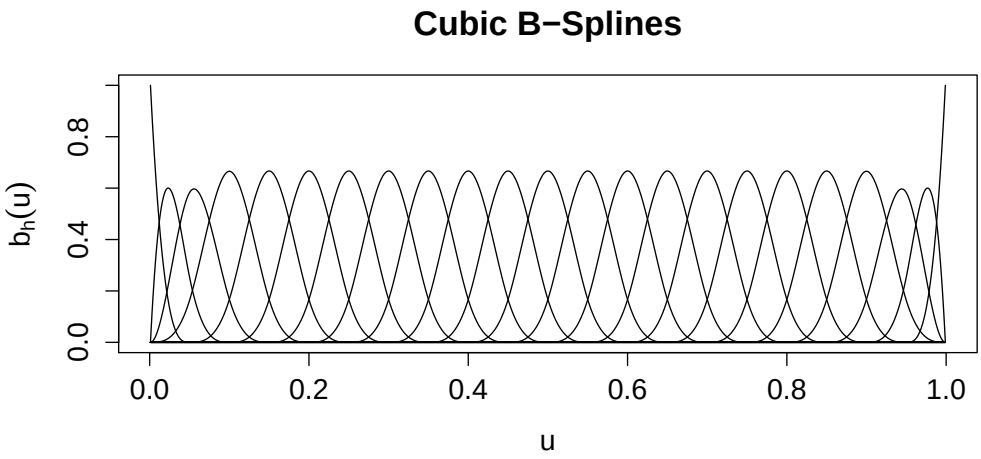

FIGURE 4.1: Cubic B-splines with 19 evenly-spaced interior knots

where β is a vector of coefficients and each $b_h(u)$ is a fixed basis function. Polynomial splines are a natural choice; here we focus on cubic splines, which ensures that g has continuous second derivatives. For highly irregular densities a lower order spline might be reasonable and is a simple modification of the methods outlined here.

Cubic splines are given by

$$b_1(u) = 1,\ b_2 = u,\ b_3(u) = u^2,\ b_4(u) = u^3,\ b_h(u) = (u - \xi_{h-4})_+^3 \tag{4.8}$$

for $5 \leq h \leq k$, where $(x)_+ = x\mathbf{1}(x > 0)$ and $0 < \xi_1 < \xi_2 < ... < \xi_k < 1$ are knot locations. Rather than working directly with cubic splines, we will focus on the equivalent cubic B-spline basis. B-splines have a more complicated form but are easily computed via a recursive algorithm that is both efficient and numerically stable (De Boor, 1978). Figure 4.1 shows the cubic B-spline basis functions with 19 evenly spaced interior knots. Each basis function is nonzero only on a small section of $[0, 1]$, so recomputing $g(u)$ can be done quickly during Monte Carlo simulations. The B-spline formulation also makes it straightforward to specify priors on β that encourage smoothness in g, which are detailed in Section 4.2.1.

One appealing feature of the TSDM is that all its partial derivatives are available

analytically. This allows the TSDM to be fit using sophisticated MCMC techniques that use the gradients of the log posterior; in this chapter we use Hamiltonian Monte Carlo (Neal, 2011), specifically the adaptive variant implemented in Stan (Stan Development Team, 2013). Some of the derivatives have complicated expressions, but these can be efficiently computed with existing B-spline software. Derivatives for β are obviously simple to compute. The partial derivative of g with respect to u is a quadratic B-spline which can be evaluated via the same recursive algorithm used for g itself (De Boor, 1978). Less well-known is that when the knots are distinct their partial derivative curves also have a B-spline representation (Piegl and Tiller, 1998). For the examples in this chapter we used the Stan package's automatic differentiation capabilities, but incorporating code to evaluate the gradient directly should dramatically increase its computational efficiency and is a subject of future work.

4.2.1 Prior specification

It will be convenient to rewrite (4.1) as

$$y_i = g_0(u_i) + g(u_i) + \epsilon_i \tag{4.9}$$

for some additional function g_0 which is either known or modeled with a few parameters. A natural choice for g_0 is

$$g_0(u) = m_0 + s_0 Q(u) \tag{4.10}$$

for some appropriately standardized quantile function Q, so that m_0 and s_0 control the location and scale of the centering distribution Q^{-1}. In this chapter we use a logistic approximation to the normal quantile function, i.e. $g_0(u) = s_0 \operatorname{logit}(u)/1.7$, and we fix $m_0 = 0$ and center the data prior to analysis. Now g represents the deviation from the centering distribution, so a sensible prior mean is constant at zero, which is achieved by taking $E(\beta) = 0$ *a priori*. Theoretical arguments for

centering in transformation models appear in Pati et al. (2011). Practical benefits include substantially more efficient posterior sampling and the ability to specify the tails of f through Q (note that the range of g is finite almost surely because the basis functions are bounded).

A common choice for $p(\beta)$ in the literature is the improper first-order random walk prior (Lang and Brezger, 2004):

$$p(\beta_1) \propto 1 \tag{4.11}$$

$$\beta_h \sim N(\beta_{h-1}, \tau_\beta) \text{ for } 2 \leq h \leq k. \tag{4.12}$$

This prior penalizes differences in adjacent coefficients, which is equivalent to applying regularization to the derivative of g (Lang and Brezger, 2004). In the current setting where u_i is unobserved it is best to avoid improper priors. Fortunately it is straightforward to be weakly informative: If our prior guess g_0 is good, then β should be concentrated around 0, and in any event the range of g (and therefore β) should not exceed the range of the data much, if at all. Hence a reasonable default choice for $p(\beta_1)$ is $N(0, (0.25r)^2)$ where r is the (observed or expected) range of y.

For a distribution Q^{-1} that is symmetric about 0, another weakly informative choice can be derived as follows. Assume k is odd and the knots are equally spaced. Then $g(0.5) = \frac{1}{6}\beta_{(k-1)/2} + \frac{2}{3}\beta_{(k+1)/2} + \frac{1}{6}\beta_{(k+3)/2}$. Define the alternative random walk prior:

$$\beta_{(k+1)/2} \sim N(0, \tau_{0.5}) \tag{4.13}$$

$$\beta_h \sim N(\beta_{h+1}, \tau_\beta) \text{ for } 1 \leq h \leq (k-1)/2 \tag{4.14}$$

$$\beta_h \sim N(\beta_{h-1}, \tau_\beta) \text{ for } (k+3)/2 \leq h \leq k. \tag{4.15}$$

It follows that $g(0.5) \sim N(0, \frac{1}{18}\tau_\beta + \tau_{0.5})$. Typically $\tau_\beta << \tau_{0.5}$, so $\tau_{0.5}$ may be chosen based on the range of the data, or to reflect prior beliefs about the median. This prior is appealing in that the variance of the B-spline coefficients is higher in the tails (near $h = 1$ and $h = k + 4$), and is symmetric: $Var(\beta_{2k-1-s}) = Var(\beta_{2k-1+s})$.

In our examples we have found little difference between the two random-walk priors, but in other applications one or the other may be more useful.

We let $1/\tau_\beta \sim Gamma(w/2, wv/2)$, where v characterizes the prior expected difference between adjacent coefficients and w controls the prior variance; $w = 1$ and $v = 0.1$ are reasonable defaults for most problems. The normal random walk can be extended to a t_ν-random walk by introducing local scale parameters so that $(\beta_h \mid \gamma_h) \sim N(\beta_{h-1}, \gamma_h \tau_\beta)$ with $1/\gamma_h \sim Gamma(\nu/2, \nu/2)$. Stronger shrinkage priors can be induced with different priors on the local scale parameters, as in e.g. Scheipl and Kneib (2009). The examples in this chapter use a t_3 random walk prior.

Our strategy for the knots is to choose a relatively small k and allow their locations to vary. The examples in this chapter all use 9 interior knots. This sacrifices some ability to recover very fine structure in g, but that is less of a problem here than in traditional spline models, where u_i would be observed, since it is far less likely that we will have enough data to infer that structure anyway. The prior distribution for the knot locations is induced by assigning the consecutive differences $(\xi_2 - \xi_1, \xi_3 - \xi_2, \ldots, \xi_k - \xi_{k-1})$ a joint uniform prior, which is centered on evenly spaced knots. Larger prior support could be obtained by giving k a prior distribution and updating it with a reversible jump Metroplis step; these methods are well-established for usual spline models (e.g. Denison et al. (1998)) but introduce a level of computational complexity that is undesirable. We will see in simulations and in an application that the fixed-k model works quite well in practice. Finally, we assume σ follows a standard half-normal distribution to reflect a prior belief that values near 0 are likely. In general σ tends to be well-determined by the data.

4.2.2 Simulation study

We compared the TSDM to a standard DP mixture of normals (DPMN) for a few interesting distributions: A log Gamma distribution with shape 1 and scale 2, a t

distribution with 3 degrees of freedom, and two of the normal mixtures from Marron and Wand (1992). The first, which we refer to as MW-4 (following the numbering scheme in the original paper) is $(2/3)N(0,1)+(1/3)N(0,1/10)$ and the second (MW-8) is $(3/4)N(0,1) + (1/4)N(1,1/3)$. These densities are shown in Figure 4.2, along with the posterior means for each model on one of the simulated datasets. We evaluated the performance of each method by computing the pointwise maximum between the posterior mean density and the truth as well as the L_1 distance between the posterior mean density and the truth.

For the DPMN we used standard hierarchical priors on the component means and variances, and Neal (2000)'s algorithm 8 with $M = 1$ as implemented in DPpackage (Jara et al., 2011). Priors for the TSDM are as described above, using the asymmetric t_3 random walk for β. Four separate chains were run for 2,000 iterations each, with the first half discarded as burn-in and keeping every second sample thereafter for a final MCMC sample size of 2,000. For the DPMN we ran the MCMC for 5,000 burn-in iterations and 20,000 regular iterations, saving every 10th sample for a final MCMC sample size of 2,000. For each distribution we sampled 50 datasets of size $n = 100$ and estimated the posterior mean of the density function.

The results are summarized in Figure 4.3. The most striking difference is on MW-4, where the TSDM uniformly outperformed under L_1 loss and also did significantly better under the pointwise maximum. This is despite the fact that the true density is in fact a mixture of normal distributions. There are two reasons why the DPMN struggles: First, there are two relatively balanced clusters, which is somewhat in conflict with the DP prior. Second, the hierarchical conjugate prior shrinks the component variances together, which is evidently problematic here. The estimates shown in Figure 4.2 are typical of the estimates across the datasets; the DPMN falls short of the peak and lacks the sharp inflection point. The TSDM is able to adapt to the peak, but to do so the bandwidth parameter must be small and so the density

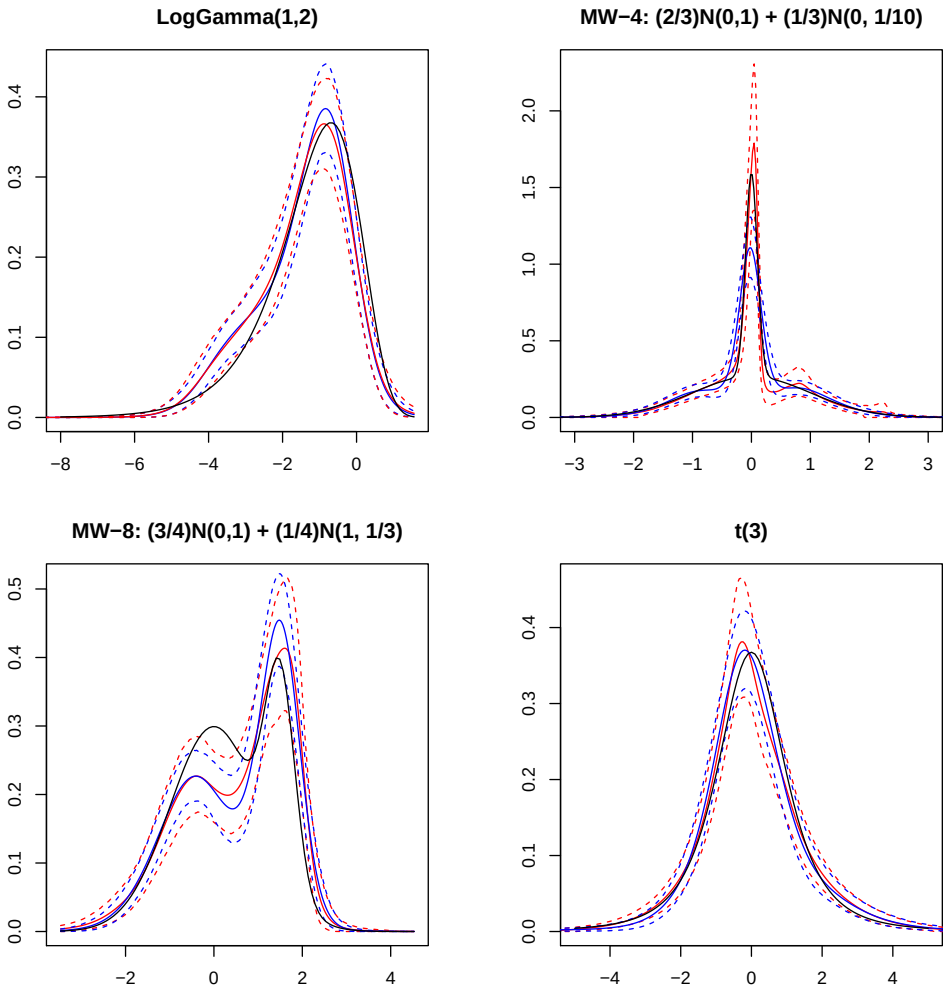

FIGURE 4.2: The true densities of the simulations in Section 4.2.2, overlaid with a randomly-chosen posterior mean and 90% credible interval under the DPMN (blue) and TSDM (red).

estimate tends to be a little rougher in the tails. Allowing the bandwidth to depend on u_i could help with this problem (see Section 4.4 for further discussion on this point).

On the MW-8, t, and log Gamma densities the performance is very similar. The DPMN performs slightly better for MW-8 and the TSDM does slightly better for the log Gamma under both losses, which is expected since MW-8 is in fact a mixture of normals. The two methods perform similarly on the t distribution, although the

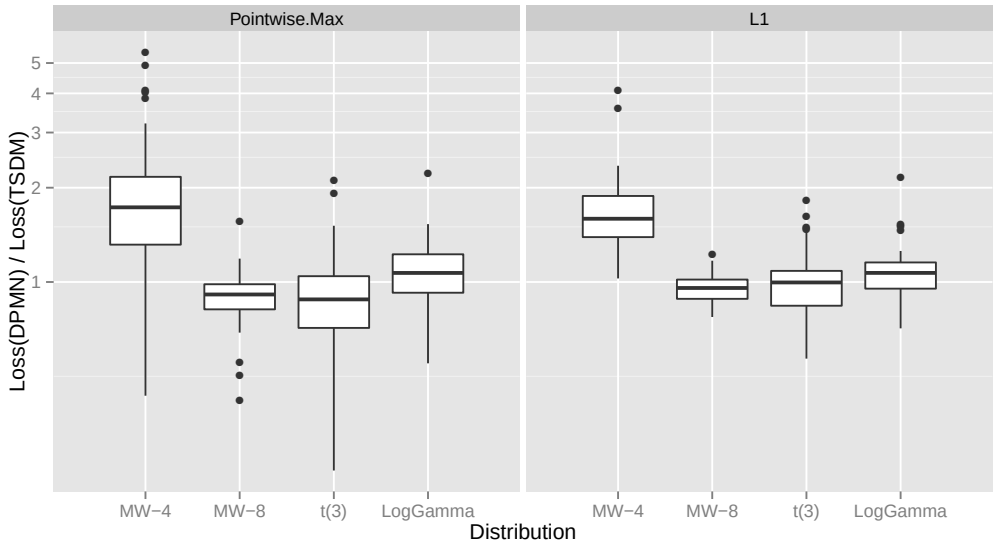

FIGURE 4.3: Relative pointwise maximum and L_1 distances between the posterior mean density estimate and the truth for the 50 simulated datasets.

DPMN seems to have an edge under the pointwise maximum loss. These differences are minor relative to the sampling variability, and the two methods generally give similar estimates as is evident from Figure 4.2. Even though there is substantial asymptotic theory behind the DPMN for density estimation it is clear that the TSDM is competitive in finite samples.

4.3 Density Regression

We turn our attention to the problem of density regression. Modifying (4.1) to be a density regression model has been proposed by Kundu and Dunson (2011) and Bhattacharya et al. (2012). The former induce a regression model by including u_i in separate functions for the outcome and covariates:

$$y_i = g_y(u_i) + \epsilon_i^y \tag{4.16}$$

$$x_{ij} = g_j(u_i) + \epsilon_{ij}^x \tag{4.17}$$

where dependence is induced by the common latent variable in a fashion reminiscent of the linear factor model. In fact, with $k = 1$ the Gaussian factor model is a special

case of (4.17), recovered when $g_y(u) = a_y\Phi^{-1}(u)$ and $g_j = a_j\Phi^{-1}(u)$ for each j. Joint modeling has some distinct drawbacks, which I discuss further in Section 4.3.1, and this model is somewhat difficult to extend to nominal covariates. Kundu and Dunson (2011) also discretize the support of u_i, which in this case turns the model into a location mixture of normals with smoothly varying means and a diagonal covariance matrix, similar to the single density case. This is undesirable in many applications.

Bhattacharya et al. (2012) instead assume that

$$y_i = g(u_i, x_i) + \epsilon_i \tag{4.18}$$

with $g(u_i, x_i)$ drawn from a Gaussian process. This avoids unnecessary joint modeling, and is substantially more flexible than (4.16)-(4.17). However, the authors focus on theoretical developments and do not describe posterior inference. One practical concern is the specification of the covariance function with mixed continuous and categorical covariates; another is computation. In particular it will be difficult to efficiently perform MCMC inference in this model without discretizing u.

We take an altogether different approach here, which we will call the transformation spline density regression model (TSDRM). First, (4.9) is modified slightly to incorporate covariates into the centering distribution, so that

$$y_i = m_0(x_i) + s_0Q(u_i) + g(u_i) + \epsilon_i, \tag{4.19}$$

which centers the model on a homoskedastic regression. It would be straightforward to also let s_0 to vary with x, centering the model on a heteroskedastic regression, although we do not pursue this here.

Rather than replacing $g(u_i)$ with $g(u_i, x_i)$ we allow the *prior* for u_i to depend on x_i. Specifically we assume u_i follows a beta regression with mean parameter $\mu(x) \in (0, 1)$ and dispersion parameter $\phi(x) \in \mathbb{R}^+$, where $a(x) = \mu(x)\phi(x)$ and

$b(x) = (1 - \mu(x))\phi(x)$. The complete data model is

$$(y_i \mid x_i, u_i) \sim N(m_0(x_i) + s_0 Q(u_i) + g(u_i), \sigma^2) \tag{4.20}$$

$$(u_i \mid x_i) \sim Beta\left(\mu(x)\phi(x), (1 - \mu(x))\phi(x)\right). \tag{4.21}$$

As in the single density case we obtain the induced distribution for $(y \mid x)$ by integrating out u_i.

The priors on σ and g are the same as the single density model. For the beta regression we assume

$$\text{logit}(\mu(x)) = \eta_\mu(x) \tag{4.22}$$

$$\log(\phi(x)) = -\eta_\phi(x). \tag{4.23}$$

The TSDRM is evidently somewhat less flexible than using the more general form $g(u_i, x_i)$ as in (4.18). The influence of all the covariates on the distribution of y (apart from its mean and potentially its scale) must be filtered through a single latent variable, and the distribution of that latent variable is restricted to belong to a mean-dispersion beta regression. However, it is quite useful to have a relatively simple model for general density regression problems that can then be extended as needed. We discuss some of the possible extensions in Section 4.4, while in Section 4.3.2 we show that this model has excellent fit to a real dataset.

4.3.1 Related work

The literature on nonparametric Bayesian models for dependent probability measures is large and growing rapidly. There have been numerous proposals to model a conditional density flexibly by fitting the joint distribution of the outcome and covariates (West and Escobar, 1993; Muller, 1996; Rodriguez et al., 2009; Shahbaba and Neal, 2009; Taddy, 2010; Molitor et al., 2010; Dunson and Bhattacharya, 2010; Hannah et al., 2011). This approach has some drawbacks. It is obviously unappealing to model a covariate such as time which is inherently not random, and essentially

without measurement error. It is also wasteful in terms of statistical efficiency. Most importantly, the Bayesian machinery doesn't "know" that only one conditional is of interest.

Consider the common situation where there are strong relationships among the covariates but relatively weak dependence of the response on covariates. Intuitively, the likelihood is based on a joint model, so the posterior distribution represents our best guess at the joint distribution. This need not (and usually will not) correspond to our best guess at the conditional distribution, even if the implied conditional model has the desired form. With the implicit penalty on model complexity in the Bayesian approach the posterior will tend to capture the strongest relationships and smooth away the weaker ones, regardless of whether we might find them interesting (see Hahn et al. (2013) for detailed discussion of this phenomenon in an important special case).

A similar situation arises in classical nonparametric estimation of conditional densities, where estimators are often constructed from estimates of the joint and marginal distributions, i.e. as $\hat{p}_{xy}(x,y)/\hat{p_x}(x)$ where $\hat{p}_{xy}(x,y),\ \hat{p_x}(x)$ are kernel density estimates. The solution in that context is to choose smoothing parameters for $\hat{p}_{xy}$ and $\hat{p_x}$ using a loss function that explicitly penalizes errors in the implied *conditional* distribution (e.g. (Bashtannyk and Hyndman, 2001; Hall et al., 2004)). There is no equivalent strategy in the fully Bayesian approach other than constructing the likelihood function correctly, based on the conditional distribution of interest.

There have been a number of proposals for conditional density estimation in the nonparametric Bayes literature that avoid the pitfalls of joint modeling. Most take the form of infinite mixture models with dependent weights and/or atoms. MacEachern (1999, 2000) developed the dependent Dirichlet process (DDP): For $\{\theta_h\} \overset{iid}{\sim} P_{0x}$ (where P_0 is some stochastic process over $\mathcal{X}$), $P_x = \sum_{h=1}^{\infty} \pi_h(x)\delta_{\theta_h}$. A

smooth density regression model is obtained via $f(y \mid x) = \int \mathcal{K}(y; x, \theta) dP_x(\theta)$ for some kernel $\mathcal{K}$. If $\pi_h(x) = V_h(x) \prod_{l<h}(1 - V_l(x))$ with $V_h(x) \sim Beta(1, \alpha)$ for all $x \in X$, then at each covariate value $P_x \sim DP(\alpha, P_{0x})$. Griffin and Steel (2006) proposed a DDP for continuous covariates which achieves dependent weights by reordering a common collection of stick breaking weights V_h with the order depending on x. Chung and Dunson (2011) proposed the local DP by introducing components and weights which are assigned to locations in covariate space. The distributions P_x are constructed via the weights in a neighborhood of x, encouraging P_x and $P_{x'}$ to be similar when $x,\ x'$ are close.

In general, constructing a prior on the weights such that $P_x \sim DP(\alpha, P_{0x})$ can be quite challenging. "Fixed-π" DDP models instead take $V_h(x) \equiv V_h$ with $V_h \sim Beta(1, \alpha)$, alleviating this difficulty while still allowing dependence in the atoms. The ANOVA-DDP described in Chapter 3 is one example of a fixed-π DDP; others include the spatial DDP (Gelfand et al., 2005), the restricted DDP (Dunson and Peddada, 2008) and the dynamic DDP model in Caron et al. (2008). I describe some of the limitations of fixed-π models in Chapter 3; essentially, they tend to have trouble adapting to local structure.

Several authors have proposed density regression models outside the DDP framework: Dunson et al. (2007) proposed a model that expresses P_x as a convex combination of basis distributions. Noting that this model has some limitations, including an unsatisfying sample dependence in the prior, Dunson and Park (2008) proposed the kernel stick breaking process (KSBP). In their formulation P_x is again expressed as a convex combination of independent measures G_h with DP priors, but the weights are given by

$$w_h(x) = V_h K(x, x_h) \prod_{l<h} (1 - V_l K(x, x_l)) \tag{4.24}$$

where $V_h \overset{iid}{\sim} Beta(1, \alpha)$ and $K(x, x')$ is a normalized kernel. The locations x_h are assigned a further prior, and the effective number of G_h is controlled by α. The probit stick breaking process instead replaces the Beta stick breaking weights with probit transformations of stochastic processes:

$$w_h(x) = \Phi(\mu_h(x)) \prod_{l<h} (1 - \Phi(\mu_l(x))) \tag{4.25}$$

where Φ is the normal cdf (Chung and Dunson, 2009; Rodríguez and Dunson, 2011). Karabatsos and Walker (2012) proposed yet another dependent mixture model defined by making the mixture weights a regression on covariates.

All of the preceding models induce dependence through discrete mixture weights, components, or both. Such models constitute the majority of density regression models proposed to date. A notable exception is the logistic Gaussian process model introduced by Tokdar et al. (2010). Inference in this model is somewhat complex, however, and the subspace projection technique does not appear to lend itself to mixed categorical and continuous regressors. Jara and Hanson (2011) also consider density regression models using transformed Gaussian processses. For high-dimensional problems with continuous regressors, Shen and Ghosal (2013) recently proposed a model which also uses B-splines, but in a fundamentally different way than the TSDRM. There is no latent variable; tensor products of B-splines instead play a role analagous to covariate dependent mixture probabilities. This model does not naturally handle categorical regressors, however. There is a distinct lack of simple, flexible density regression models for mixed covariates; this is partly addressed by the TSDRM.

4.3.2 Application: Estimating the Return to Education

We illustrate the TSDRM on a subset of public use microdata from the U.S. Census Bureau, which was originally compiled by Angrist et al. (2006) to illustrate a quantile

regression model. For the 1980, 1990 and 2000 samples they extracted all U.S. born white and black men aged 40-49 with positive annual earnings and positive hours worked in the year prior to the Census. Records with imputed values were excluded, and wages were adjusted to 1989 dollars. Angrist et al. (2006) and its supplementary materials contain complete details of how the data were obtained and cleaned. The response is log monthly wages, and the covariates include years of education, experience (defined as age-education-6) and race. The resulting dataset has over 200,000 observations; for our purposes we randomly sampled 1,000 records from each of the 1990 and 2000 files, and fit the model to the 1990 and 2000 datasets separately. The 1990 sample is self-weighting, but the 2000 file includes person-level weights; we drew our sample with probabilities proportional to these, so that our subsample is approximately a simple random sample from the target population, instead of a sample from the microdata file.

The covariate vector x includes education, experience, experience squared, and an indicator for race (1 if black, 0 otherwise). The continuous covariates are standardized to have zero mean and unit variance. The regression functions and priors are

$$m_0(x) = \alpha_0' x,\ \alpha_0 \sim N(0, 10I) \tag{4.26}$$

$$\eta_\mu(x) = \alpha_{\mu 0} + \alpha_\mu' x,\ (\alpha_{\mu 0}, \alpha_\mu')' \sim N(0, 2.5I) \tag{4.27}$$

$$\eta_\phi(x) = \alpha_{\phi 0} + \alpha_\phi' x,\ (\alpha_{\mu 0}, \alpha_\phi')' \sim N(0, 2.5I). \tag{4.28}$$

The remaining parameters are assigned the priors described in previous sections. We fit the model using Stan 2.0.1, running 2 chains for 20,000 iterations with half used as burn-in. The performance of the sampler was improved substantially by randomly perturbing the adapted step size by setting `stepsize_jitter=0.5`. Standard MCMC diagnostics on the quantiles and estimated density values were all excellent.

Before describing some results, we assess model fit using posterior predictive

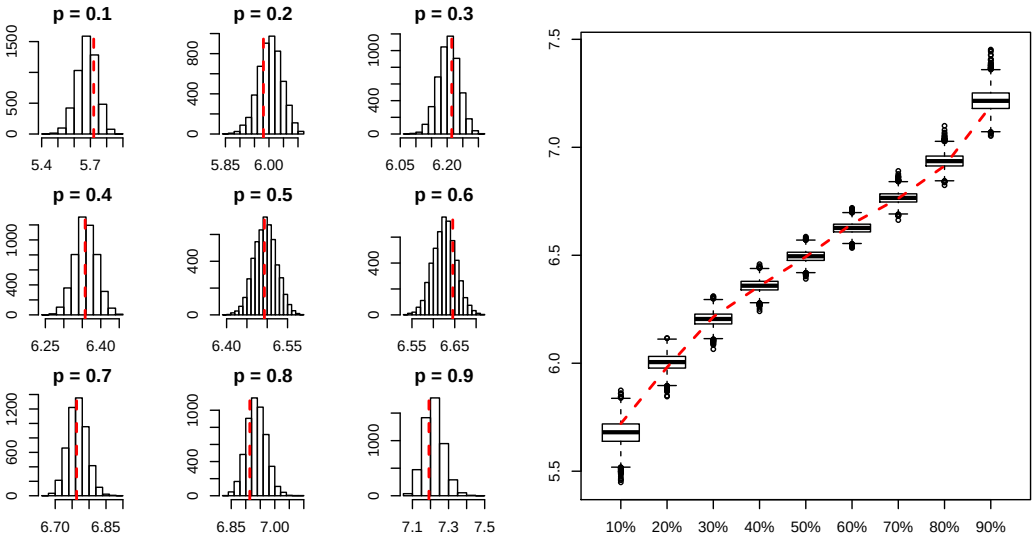

FIGURE 4.4: Posterior predictive samples of marginal sample quantiles for the 1990 dataset. Dashed lines are the observed values.

checks. For each sample of the parameters we generated a new vector of responses by sampling $\tilde{y}_i$ from the predictive distribution at x_i for $1 \leq \ i \leq n$ and computing relevant summary statistics on the new dataset. Figures 4.4 and 4.5 display posterior predictive distributions of marginal sample deciles for the 1990 and 2000 datasets, respectively. Overall the fit is quite good in both cases.

To assess the fit of conditional distributions, we computed the quantile regression coefficients at $p = .1, .25, .5, .75, .9$ in each replicated dataset. Since the sample is relatively small, stratifying on particular values of the covariates would give highly variable estimates of conditional quantiles and make posterior predictive checks unreliable. The quantile regression coefficients should be more stable and provide at least some capacity to detect model misfit in the conditional distributions.

Figures 4.6 and 4.7 show the results for 1990 and 2000, respectively. Again the results are quite good overall. The modes of the posterior predictive distributions are all close to the observed values. There is substantial variation for most of the coefficients, particularly the coefficient for race. This is unsurprising, since the two

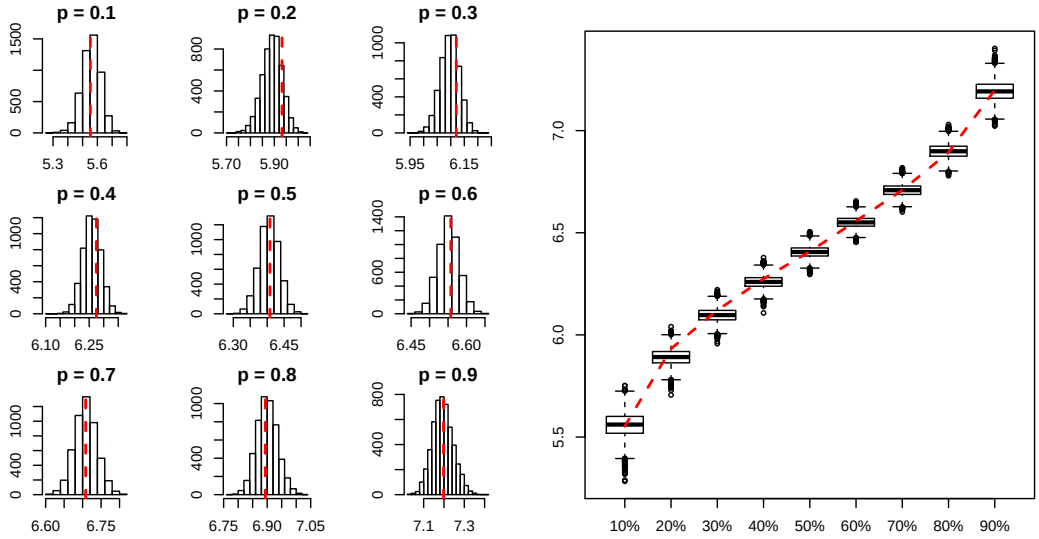

FIGURE 4.5: Posterior predictive samples of marginal sample quantiles for the 2000 dataset. Dashed lines are the observed values.

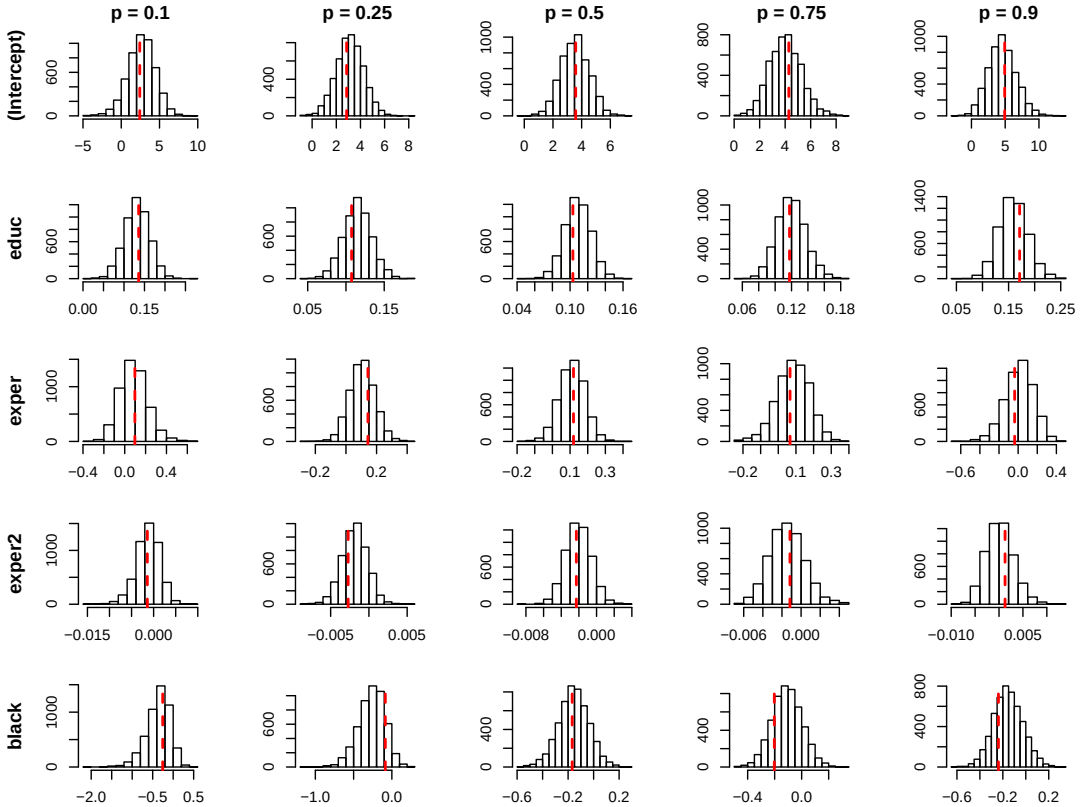

FIGURE 4.6: Posterior predictive samples of quantile regression coefficients for the 1990 dataset. Dashed lines are the observed values.

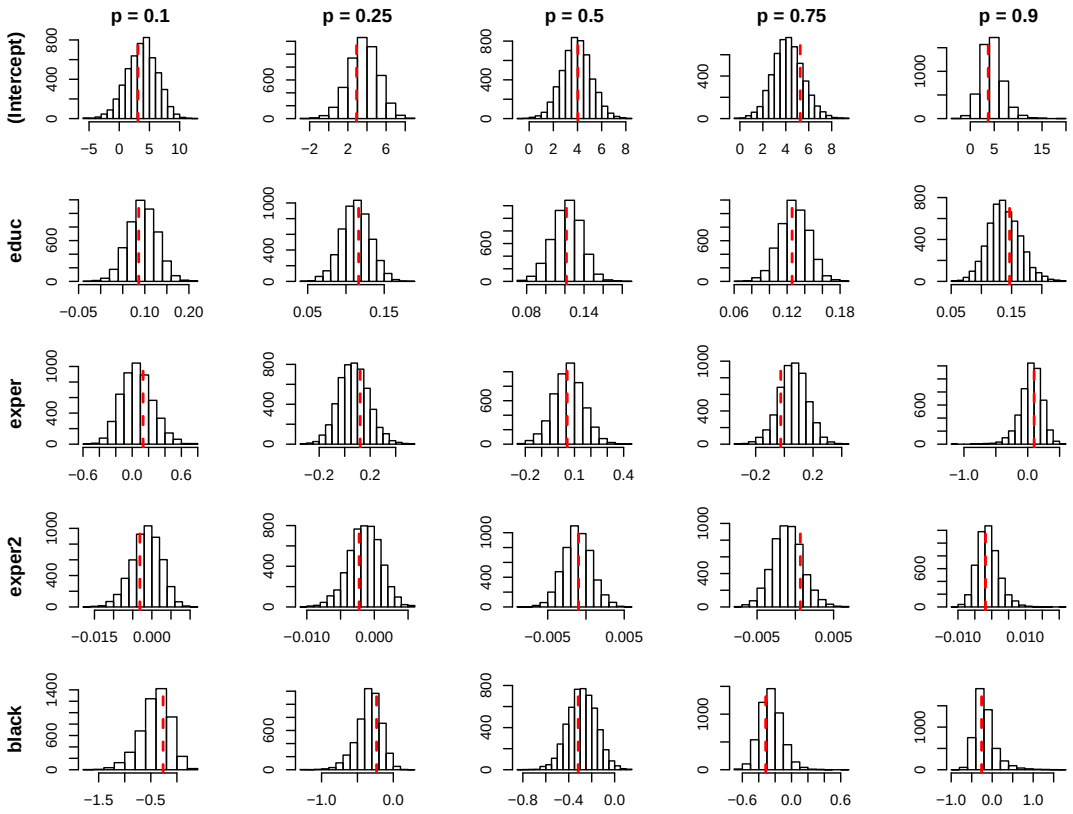

FIGURE 4.7: Posterior predictive samples of quantile regression coefficients for the 2000 dataset. Dashed lines are the observed values.

datasets are both approximately 93% white, but even for this coefficient the posterior predictive checks do not suggest serious discrepancies. Formal model assessment and choice is a topic for future work, however the posterior predictive checks are encouraging.

Angrist et al. (2006) used this data to estimate the return to education as a function of the quantile index; we will obtain a smooth estimate of this function using the TSDRM. Let $Q_x(p)$ be the quantile function of monthly wages at $X = x$. We are interested in

$$h(p; x_1, x_2) = 100 \times \frac{Q_{x_2}(p) - Q_{x_1}(p)}{Q_{x_1}(p)}, \tag{4.29}$$

i.e. the predicted percentage change in monthly wage when changing the covariate

vector from $X = x_1$ to $X = x_2$. Angrist et al. (2006) fit a linear quantile regression to estimate the effect of each additional year of education, so that (4.29) simplifies to

$$h(p; x_1, x_2) = 100 \times \left(e^{\gamma_{educ}^{(p)}} - 1 \right) \tag{4.30}$$

for any values of the other variables, where $\gamma_{educ}^{(p)}$ is the education coefficient at quantile index p. This is appealing, as the coefficients have a direct interpretation as a marginal effect. But it is also somewhat restrictive, as it supposes homogenous effects in the population at each quantile index. The TSDRM instead assumes homogenous effects on the latent variable u_i, due to the use of the log and logit link functions and linear regressions for $\eta_\mu(x),\ \eta_\phi(x)$. However, those restrictions do not propagate to y_i and in general the effect of a covariate will depend on the value of all the others. We will first discuss inference for h at a completely specified covariate vector, and then introduce a method for estimating marginal effects.

With T samples from the posterior distribution of the parameters we can construct posterior samples of (4.29) as follows: For $x = x_1$ and $x = x_2$, repeat the following for $1 \leq t \leq T$:

1. Compute the conditional cdf at $X = x$ over a fine grid $\tilde{y}_1, \ldots, \tilde{y}_R$ using the formula

$$F_x^{(t)}(y) = \int_0^1 \Phi\left(\frac{y - (g_0^{(t)}(u) + g^{(t)}(u))}{\sigma^{(t)}} \right) p\left(u; a^{(t)}(x), b^{(t)}(x)\right) du, \tag{4.31}$$

   where $p(u; a, b)$ is the pdf of a $Beta(a, b)$ distribution. The integral can be done numerically, but in some datasets this can become unstable due to occasional extreme samples for g or the beta parameters. However,

$$F_x^{(t)}(y) = E_{(u|a^{(t)}(x), b^{(t)}(x))}\left(\Phi\left(\frac{y - (g_0^{(t)}(u) + g^{(t)}(u))}{\sigma^{(t)}} \right) \right) \tag{4.32}$$

so that Monte Carlo integration can be used instead. In this case one set of samples from $p(u \mid a(x), b(x))$ can be used to compute $F_x^{(t)}(y)$ over the whole grid of y values. The Monte Carlo error only needs to be small relative to the overall error from posterior sampling, so this method can be quite efficient.

2. Invert $F_x^{(t)}(y)$ to obtain the quantile function $Q_x^{(t)}(p)$. The simplest way to do this is via linear interpolation based on the $(F_x^{(t)}(\tilde{y}_r), \tilde{y}_r)$ pairs computed in step 1. If some other interpolant $\hat{F}_x^{(t)}$ is used instead then we just solve $\hat{F}_x^{(t)}(y) - \tilde{p}_r = 0$ for a grid $0 < \tilde{p}_1 \leq \tilde{p}_2 \leq \cdots \leq \tilde{p}_S < 1$, which can be done very efficiently since $\hat{F}_x^{(t)}$ is monotone and smooth.

3. Compute

$$h^{(t)}(\tilde{p}_r; x_1, x_2) = 100 \times \frac{Q_{x_1}^{(t)}(\tilde{p}_r) - Q_{x_2}^{(t)}(\tilde{p}_r)}{Q_{x_1}^{(t)}(\tilde{p}_r)} \tag{4.33}$$

over a fine grid.

Averaging $h^{(t)}(\tilde{p}; x_1, x_2)$ over $1 \leq t \leq T$ gives an estimate of the posterior mean of h, and we can construct pointwise credible bands as well. The same procedure can be used for any other functionals of interest.

The marginal effect of a single variable can be estimated with a slight modification of step 1. Let $x_{i,c}$ be the variable of interest for observation i and $x_{i,-c}$ be the remaining variables. Then if the two levels of X_c to be compared are s and $s+\delta$, let $x_{1i} = (s, x_{i,-c})$ and $x_{2i} = (s+\delta, x_{i,-c})$. For each of these covariate vectors, compute

$$F_s^{(t)}(y) = \sum_{i=1}^{n} \frac{1}{n} F_{x_{1i}}^{(t)}(y) \tag{4.34}$$

$$F_{s+\delta}^{(t)}(y) = \sum_{i=1}^{n} \frac{1}{n} F_{x_{2i}}^{(t)}(y), \tag{4.35}$$

and then invert them to obtain quantile functions as in step 2, and compute h as in step 3. That is, in step 1 we estimate the cdf of Y given X_c but marginalized over the other covariates with respect to their empirical distribution. Similar methods could be used to estimate conditional effects when X_c is bivariate to assess interaction effects. Hill (2011) suggested a similar strategy for computing average treatment effects in nonlinear mean regression models.Another distribution for X_{-c} could be used instead of the empirical distribution, and in principle it would be possible to do a fully Bayesian analysis of the joint distribution to get an estimate for the marginal effect accounting for uncertainty in the marginal distribution of X_c. We do not pursue these methods here, however.

Marginal effect estimates for $X_{educ} = 16$ versus $X_{educ} = 12$ in 1990 and 2000 are shown in Figure 4.8. The posterior mean and 50% and 90% credible intervals are given, as well as the point estimates from OLS and quantile regressions (the quantile regressions were computed at 20 regularly spaced points from 0.1 to 0.9). There is clear evidence that the effect varies across the income distribution; in 1990 it is largest at the lower and upper quantiles, while in 2000 it is low at the low quantiles and high at the upper quantiles, with a sharp twist in the middle. This is consistent with the findings in Angrist et al. (2006), who confirmed this pattern with additional data from the Current Population Survey.

This application illustrates a key feature of the TSDRM: It is able to estimate smooth conditional quantile functions (and other functionals, like h), unlike quantile regression methods, which will tend give rough estimates that are also sensitive to the chosen quantile indices. The model is straightforward to specify, requiring only the choice of three regression functions, all of which have meaningful scales. Inference is somewhat time consuming, taking several hours for each of the two datasets, but once Stan can take advantage of efficient routines to compute splines and their derivatives this should be improved substantially.

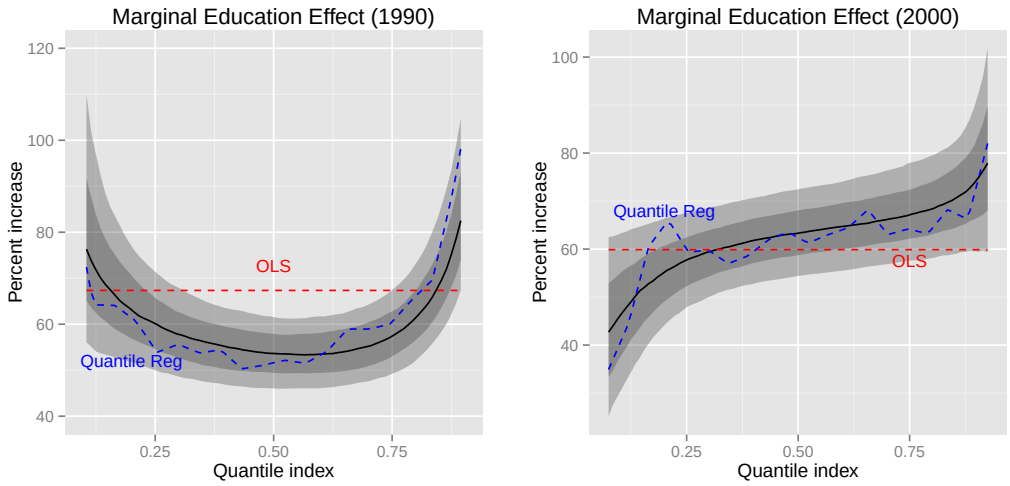

FIGURE 4.8: Marginal percentage difference in monthly wages for 16 years of education versus 12. The straight dashed line is the OLS estimate, the curved dashed line is the quantile regression estimate, and the solid black like is the posterior mean under the TSDRM. The dark and light gray bands are pointwise 50% and 90% credible intervals, respectively.

4.4 Conclusion and future work

We have presented a new model for density estimation and regression. There are a number of interesting directions to extend this work. One possibility is to use more informative priors, for example requiring g to be monotone. This is straightforward to do using our spline-based prior, either by requiring β to be nondecreasing, or by using spline bases that are designed specifically for monotone function estimation. Another interesting extension is allowing the bandwidth σ to vary with u_i; the resulting model generalizes discrete location-scale mixtures of normals in the same fashion that the TSDM generalizes location mixtures. We can also extend the model to multivariate data by sharing the latent variables in separate functions, i.e., $y_{ij} = g_j(u_i) + \epsilon_{ij}$, as in Kundu and Dunson (2011). When multivariate density regression is the ultimate goal, it might be reasonable to allow further dependence in ϵ_i as well.

In some density regression problems the simple beta regression prior on u_i may

be inadequate. One way to allow for further heterogeneity would be to include random effects into the mean and dispersion models. For example, if these random effects were drawn from a Dirichlet process or another discrete prior then $p(u_i \mid x_i)$ could take on a wider variety of different shapes. A more flexible option would to replace $g(u)$ with $g(u, x)$, but preserve the tractability of the spline model by using a functional ANOVA-type decomposition (Wahba 1990), i.e.

$$g(u, x) = \sum_{j=1}^{p} g_j(u_i, x_j). \tag{4.36}$$

with g_j modeled via tensor splines, or hierarchically if x_j is discrete. For particular covariate spaces other priors might make more sense. For example, if X is comprised entirely of categorical variables then we might assume

$$g(u, x) = \sum_{h=1}^{k} \beta_{xh} b_h(u). \tag{4.37}$$

With a p dimensional vector of covariates the spline coefficients β form a $(p+1)$-way tensor which can be further modeled via a low-rank approximation; Hoff (2010) used a similar model for cell means in crossclassified data. The continuous mixture formulation, and particularly the spline representation, should open up new avenues for flexible, tractable modeling of conditional densities.

Appendix A

Proofs for Chapter 2

In this appendix I give a detailed proof of posterior consistency in the semiparametric Gaussian copula model using the extended rank likelihood (Theorem 1 in Chapter 2). In section A.2 I detail the marginal and blocked PX Gibbs samplers. I show their equivalence as the PX prior becomes increasingly diffuse (and improper in the limit) which proves that the PX Gibbs sampler in Chapter 2 has the correct target distribution.

A.1 Proof of Theorem 1

Proof. We require a variant of Doob's theorem, presented in Gu and Ghosal (2009):

Theorem 2. *Let X_i be observations whose distributions depend on a parameter θ, both taking values in Polish spaces. Assume $\theta \sim \Pi$ and $X_i|\theta \sim P_\theta$. Let $\mathcal{X}_N$ be the σ-field generated by $X_1, \ldots, X_N$ and $\mathcal{X}_\infty = \sigma\langle\bigcup_{i=1}^{\infty} \mathcal{X}_i\rangle$. If there exists a $\mathcal{X}_\infty$ measurable function f such that for $(\omega, \theta) \in \Omega^\infty \times \Theta$, $\theta = f(\omega)$ a.e. $[P_\theta^\infty \times \Pi]$ then the posterior is strongly consistent at θ for almost every θ $[\Pi]$.*

Therefore we must establish the existence of a consistent estimator of C which is

measurable with respect to the σ-field generated by the sequence $\{D(Y^{(m)})\}_{m=1}^{\infty}$ (a coarsening of the σ-field generated by $\{Y^{(m)}\}_{m=1}^{\infty}$). Let $R_{nij} = \sum_{h=1}^{n} \mathbf{1}([y_{hj} \leq y_{ij}]) = n\hat{F}_j(y_{ij})$. Let $\mathbf{R}_{ni}(Y^{(n)})$ be the p-vector with entry j given by R_{nij} and let $\mathbf{R}_n(Y^{(n)}) = \{\mathbf{R}_{ni}\}_{i=1}^{n}$. Observe that the information contained in the extended rank likelihood (namely the boundary conditions in the definition of the set $D(Y^{(n)})$) is equivalent to the information contained in $R_n(Y^{(n)})$. Hence a function that is measurable with respect to $\mathcal{R}_n$, the σ-field generated by $\{R_m(Y^{(m)})\}_{m=1}^{n}$, is also measurable with respect to the σ-field generated by $\{D(Y^{(m)}\}_{m=1}^{N}$ and we may work exclusively with the former.

Let $\hat{U}_{nij} = \frac{R_{nij}}{n+1}$ and $\hat{U}_{ni} = (\hat{U}_{ni1}, \ldots \hat{U}_{nip})'$. Then $\hat{U}_{nij} \xrightarrow{as} U_{ij}$ where $U_{ij} = F_j(y_{ij})$ by the SLLN, so $\hat{U}_{ni} \xrightarrow{as} U_i$ and therefore U_i is $\mathcal{R}_\infty = \sigma\langle \bigcup_{i=1}^{\infty} \mathcal{R}_i \rangle$ measurable. Note that if F_j is discrete U_{ij} is merely a relabeling of y_{ij} (each category/integer is "labeled" with its marginal cumulative probability). So U_i is a sample from a Gaussian copula model with correlation matrix C_0 where the continuous margins are all $U[0,1]$ and the discrete marginal distributions are completely specified. The problem of estimating C from U_i reduces to estimating ordinary and polychoric/polyserial correlations with fixed marginals and it is straightforward to verify that the distribution of U_i is a regular parametric family admitting a consistent estimator of C, say $h_N(U_1, \ldots, U_N)$. Therefore there exists a sequence of $\mathcal{R}_\infty$ measurable functions $h_N(U_1, \ldots, U_N) \to h(U_1, U_2, ...) = C_0$ almost surely and

$$C_0 = h(U_1, U_2, ...) = h^*(\{\mathbf{R}_{Ni} : N \geq 1, 1 \leq i \leq N\}) \text{ a.s. } [G^{\infty}_{C_0, F_1, \ldots, F_p}] \qquad \text{(A.1)}$$

where (A.1) holds because a null set under the measure induced by $R_n(Y^{(n)})$ is also null under $G^{\infty}_{C_0, F_1, \ldots, F_p}$. □

A.2 Validity of the PX Sampler

Let Θ be the inferential parameters and let $s_j = z_j(I - H_j'(\Psi_j^{-1} + H_j H_j')^{-1} H_j) z_j'$ Our working prior for $(v_1, \ldots v_p)$ is $\prod_{j=1}^{P} IG(v_j^2; n_0/2, n_0/2)$. To verify that samples of Θ from the PX-Gibbs sampler have stationary distribution $\pi(\Theta|Y)$ we need to show that as $n_0 \to 0$ the transition kernels under the marginal sampling scheme (alternately drawing from $\pi(W|\Theta, Y)$ and $\pi(\Theta|W)$) and the blocked sampling scheme (alternately drawing from $\pi(W|\Theta, V, Y)$ and $\pi(\Theta, V|W)$) converge (Meng and Van Dyk, 1999). The t^{th} updates under the two schemes are as follows:

Scheme 1: Draw $1/v_{0j}^2 \sim Ga(n_0/2, n_0/2)$ and $1/v_{1j}^2 \sim Ga\left(\frac{n_0+n}{2}, \frac{n_0+v_{0j}^2 s_j}{2}\right)$. Set $r = v_{j0}/v_{j1}$ and draw $\lambda_j \sim N(r\hat{\lambda_j}', (\Psi_j^{-1} + HH')^{-1})$

Scheme 2: Draw $1/v_{tj}^2 \sim Ga\left(\frac{n_0+n}{2}, \frac{n_0+v_{(t-1)j} s_j}{2}\right)$. Set $r = v_{(t-1)j}/v_{tj}$ and draw $\lambda_j \sim N(r\hat{\lambda_j}', (\Psi_j^{-1} + HH')^{-1})$

Updates for the rest of Θ under both schemes are the same as in Section 2.3.2. As $n_0 \to 0$ under Scheme 1 the distribution of $1/v_{0j}^2$ approaches a point mass at 1 and Scheme 1 converges to Scheme 2 with $n_0 = 0$.

Appendix B

Posterior Inference in the HCMM-LD

Posterior inference is available through Gibbs sampling, which we outline below. Banerjee et al. (2013) describe an exact partially collapsed Gibbs sampler for ITF mixtures which could be adopted directly. However, we present a truncation approximation here that is simpler to implement and approaches the infinite dimensional model in the limit, building on Ishwaran and James (2001)'s blocked Gibbs sampler for truncated stick breaking priors. For integers k_0, k_x and k_y we set $W_{k_0} = 1$ and, for each $1 \leq z \leq k_0$, $V^{(x)}_{zk_x} = 1$ and $V^{(y)}_{zk_y} = 1$. The experiments in Section 3.4 use $k_0 = 15$, $k_y = 60$ and $k_x = 90$.

A reasonable way to choose the truncation levels is to initially choose a fairly high number (something on the order of $\sqrt{n}$ seems to work well), initialize the MCMC algorithm with a large number of small clusters and monitor the number of occupied components during a burn-in phase. If the number of occupied components is close to the truncation level, it is probably too small and should be increased and the burn-in repeated. Otherwise the MCMC can proceed, unless the difference is substantial and computation time is a concern, in which case the truncation level can be decreased

and the burn-in repeated. This is essentially how we arrived at the truncation levels in Chapter 3, although we chose somewhat higher truncation levels to allow for potential differences in the 500 simulated datasets.

The MCMC algorithm is as follows:

- Z: For each observation, sample Z_i from

$$\Pr(Z_i = z \mid H_i = (h_x, h_y), \Psi, \lambda) \propto \lambda_z \phi^{(x)}_{zh_x} \phi^{(y)}_{zh_y} \tag{B.1}$$

  for $1 \leq z \leq k_0$

- X_{mis}: For each observation i sample each missing entry of X_i from its full conditional distribution,

$$\Pr(X_{ij} = x_j \mid -) \propto \psi^{(j)}_{h_x x_j} N(Y_i;\ D_i(x_j) B_{h_y}, \Sigma_{h_y}), \tag{B.2}$$

  where $D_i(x_j)$ is the design vector obtained by setting $X_{ij} = x_j$ and holding the other elements of X_i at their current values. If the number of categorical variables subject to missingness is relatively small, it may be feasible to update all the missing entries in X_i in a block. This will lead to a better mixing chain when there are strong dependencies in the distribution of X. In practice we find this simpler update to work well. It is much more efficient computationally, since the sample space for the blocked update gets large rapidly as the number of missing variables increases.

- H_x For each observation update H_{xi} from

$$\Pr(H_{xi} = h_x \mid Z_i = z, -) \propto \phi^{(x)}_{zh_x} \prod_{j=1}^{p} \psi^{(j)}_{h_x X_{ij}} \tag{B.3}$$

- (Y^{mis}, H_y): Update the cluster index for Y and the missing entries in a block

by first sampling H_{yi} marginally over Y_i^{mis} according to

$$\Pr(H_{yi} = h_y \mid Z_i = z, Y_i^{obs}, X_i, \phi_z^{(y)}, \{B_h, \Sigma_h\}) \propto \phi_{zh_y}^{(y)} N(Y_i^{obs};\ D_i B_{h_y}^*, \Sigma_{h_y}^*) \tag{B.4}$$

where $B_{h_y}^*$ is obtained by droppping the columns of B_{h_y} correspoinding to missing observations in Y_i and $\Sigma_{h_y}^*$ is the relevent submatrix of Σ_{h_y}. Given the new component index, sample the missing entries of Y from

$$(Y_i^{mis} \mid -) \sim N(\tilde{\mu} + D_i \tilde{B}_{h_y}, \tilde{\Sigma}_{h_y}) \tag{B.5}$$

where $N(Y_i^{mis};\ \tilde{\mu} + D_i \tilde{B}_{h_y}, \tilde{\Sigma}_{h_y})$ is the standard conditional distribution of $(Y_i^{mis} \mid Y_i^{obs}, -)$. The block update is critical when $\tilde{B}_{h_y}$ and/or $\tilde{\Sigma}_{h_y}$ vary substantially across clusters, as is often the case.

- Cluster parameters: For each $1 \le z \le k_0$, $1 \le h_x \le k_x$ and $1 \le j \le p$ sample

$$(\psi_{h_x}^{(j)} \mid -) \sim Dir\left(\gamma_{j1} + n_{h_x 1}, \gamma_{j2} + n_{h_x 2}, \ldots, \gamma_{jd_j} + n_{h_x d_j},\right), \tag{B.6}$$

where $n_{h_x c} = \sum_{i=1}^n \mathbf{1}(H_{xi} = h_x, X_{ij} = c)$.

For each $1 \le h_y \le k_y$ and $1 \le r \le q$ sample

$$(B_{h_y r} \mid -) \sim N\left(V(\tau_q B_q + D'_{h_y} \tilde{y}_{hr} / \tilde{\sigma}^2_{h_y r}), V\right) \tag{B.7}$$

where $V = (\tau_q I + D'_{h_y} D_{h_y} / \tilde{\sigma}^2_{h_y r})^{-1}$ and D_{h_y} is the matrix obtained by stacking the vectors $\{D_i(X_i) : H_{yi} = h_y\}$ and $\tilde{y}_{hr}$ is the vector obtained by concatenating $\{y_{ir} - \tilde{\mu}_{iq} : H_{yi} = h_y\}$. The parameters $\tilde{\mu}_{iq}, \tilde{\sigma}_{h_y r}$ are the mean and variance of the conditional normal distribution $p(Y_{iq} \mid Y_{i/q}, H_{yi})$.

Finally, for each $1 \le h_y \le k_y$ sample

$$\Sigma_{h_y} \sim IW(d + \sum_{i=1}^n \mathbf{1}(H_{xi} = h_x), \Sigma + S_{h_y}), \tag{B.8}$$

where $S_{h_y} = \sum_{i:H_{iy}=h_y} (Y_i - D_i B_{h_y})(Y_i - D_i B_{h_y})'$

- Hyperparameters: For each entry of B sample

$$(B_{jr} \mid -) \sim N\left((k_y\tau_r + 1/\sigma_0^2)^{-1}k_x\tau_r \sum_{h=1}^{k_y} B_{hjr}, (k_y\tau_r + 1/\sigma_0^2)^{-1}\right). \tag{B.9}$$

For $1 \leq r \leq q$ sample

$$(\tau_r \mid -) \sim G\left(\frac{a_\tau + k_y}{2}, \frac{b_\tau + \sum_{h=1}^{k_y}(B_{hr} - B_r)'(B_{hr} - B_r)}{2}\right). \tag{B.10}$$

- Mixing proportions: Sample λ by drawing

$$(W_h \mid -) \sim Beta\left(1 + m_h, \alpha + n - \sum_{l=1}^{h} m_l\right), \tag{B.11}$$

where $m_h = \sum_{i=1}^{n} \mathbf{1}(Z_i = h)$, and set $\lambda = V_{0h}\prod_{l<h}(1 - V_{0l})$. For $1 \leq z \leq k_0$ sample

$$(V_{zh}^{(x)} \mid -) \sim Beta\left(1 + r_{zh}^{(x)}, \beta_x + m_z - \sum_{l=1}^{h} r_{zl}^{(x)}\right) \tag{B.12}$$

$$(V_{zh}^{(y)} \mid -) \sim Beta\left(1 + r_{zh}^{(y)}, \beta_y + m_z - \sum_{l=1}^{h} r_{zl}^{(y)}\right) \tag{B.13}$$

where $r_{zh}^{(x)} = \sum_{i=1}^{n} \mathbf{1}(Z_i = z)\mathbf{1}(H_{xi} = h)$ and $r_{zh}^{(y)}$ is defined similarly. Set $\phi_{zh}^{(x)} = V_{zh}^{(x)}\prod_{l<h}(1 - V_{zl}^{(x)})$, $\phi_{zh}^{(y)} = V_{zh}^{(y)}\prod_{l<h}(1 - V_{zl}^{(y)})$

- Concentration parameters: Let $\mathcal{Z}_{occ}$ be the set of occupied top-level clusters. Sample the concentration parameters from their gamma full conditionals:

$$\alpha \sim G(a_0 + k_0 - 1, b_0 - \log(\lambda_{k_0})) \tag{B.14}$$

$$\beta_x \sim G\left(a_x + k_x, b_x - \sum_{z \in \mathcal{Z}_{occ}} \log\left(\phi_{zk_x}^{(x)}\right)\right) \tag{B.15}$$

$$\beta_y \sim G\left(a_y + k_y, b_y - \sum_{z \in \mathcal{Z}_{occ}} \log\left(\phi^{(y)}_{zk_y}\right)\right) \tag{B.16}$$

Bibliography

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669.

Angrist, J., Chernozhukov, V., and FernándezVal, I. (2006), "Quantile regression under misspecification, with an application to the US wage structure," *Econometrica*.

Armagan, A., Dunson, D., and Lee, J. (2011), "Generalized double Pareto shrinkage," *arXiv preprint arXiv:1104.0861*.

Banerjee, A., Murray, J., and Dunson, D. B. (2013), "Bayesian learning of joint distributions of objects," in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Scottsdale, AZ, USA.

Barnard, J. and Rubin, D. B. (1999), "Miscellanea. Small-sample degrees of freedom with multiple imputation," *Biometrika*, 86, 948–955.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011), *Latent variable models and factor analysis: A unified approach*, vol. 899, Wiley. com.

Bashtannyk, D. M. and Hyndman, R. J. (2001), "Bandwidth selection for kernel conditional density estimation," *Computational Statistics & Data Analysis*, 36, 279–298.

Bhattacharya, A. and Dunson, D. B. (2011), "Sparse Bayesian infinite factor models." *Biometrika*, 98, 291–306.

Bhattacharya, A., Pati, D., and Dunson, D. B. (2012), "Latent factor density regression models," .

Blackwell, D. and MacQueen, J. (1973), "Ferguson Distributions Via Polya Urn Schemes," *The annals of statistics*.

Böhning, D., Seidel, W., Alfó, M., Garel, B., Patilea, V., Walther, G., Di Zio, M., Guarnera, U., and Luzi, O. (2007), "Imputation through finite Gaussian mixture models," *Computational Statistics & Data Analysis*, 51, 5305–5316.

Bush, C. (1996), "A semiparametric Bayesian model for randomised block designs," *Biometrika*, 83, 275–285.

Canale, A. and Dunson, D. B. (2011), "Bayesian multivariate mixed-scale density estimation," .

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2008), "Bayesian Inference for Linear Dynamic Models With Dirichlet Process Mixtures," *IEEE Transactions on Signal Processing*, 56, 71–84.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008), "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456.

Chung, Y. and Dunson, D. B. (2009), "Nonparametric Bayes Conditional Distribution Modeling With Variable Selection." *Journal of the American Statistical Association*, 104, 1646–1660.

Chung, Y. and Dunson, D. B. (2011), "The local Dirichlet process." *Annals of the Institute of Statistical Mathematics*, 63, 59–80.

Cowles, M. K. (1996), "Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models," *Statistics and Computing*, 6, 101–111.

De Boor, C. (1978), *A practical guide to splines*, vol. 27, Springer-Verlag New York.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205–215.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian curve fitting," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 333–350.

Dobra, A. and Lenkoski, A. (2011), "Copula Gaussian graphical models and their application to modeling functional disability data," *The Annals of Applied Statistics*, 5, 969–993.

Drechsler, J. (2010), "Multiple imputation of missing values in the wave 2007 of the IAB Establishment Panel," *IAB Discussion Paper*.

Dunson, D. (2000), "Bayesian latent variable models for clustered mixed outcomes," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62, 355–366.

Dunson, D. B. (2003), "Dynamic Latent Trait Models for Multidimensional Longitudinal Data," *Journal of the American Statistical Association*, 98, 555–563.

Dunson, D. B. and Bhattacharya, A. (2010), "Nonparametric Bayes regression and classification through mixtures of product kernels," *Bayesian Stats*.

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

Dunson, D. B. and Peddada, S. D. (2008), "Bayesian nonparametric inference on stochastic ordering," *Biometrika*, 95, 859–874.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007), "Bayesian density regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 163–183.

Dunson, D. D. B. and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, 104, 1042–1051.

Elliott, M. R. and Stettler, N. (2007), "Using a mixture model for multiple imputation in the presence of outliers: the 'Healthy for life' project," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56, 63–78.

Escobar, M. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the american statistical association*, pp. 577–588.

Favaro, S. and Teh, Y. W. (2013), "MCMC for Normalized Random Measure Mixture Models," *Statistical Science*, 28, 335–359.

Ferguson, T. (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*.

Gebregziabher, M. and DeSantis, S. M. (2010), "Latent class based multiple imputation approach for missing categorical data," *Journal of Statistical Planning and Inference*, 140, 3252–3262.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021–1035.

Gelman, A. (2004), "Parameterization and Bayesian Modeling," *Journal of the American Statistical Association*, 99, 537–545.

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 1, 515–534.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995), "A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions," *Biometrika*, 82, 543.

Ghahramani, Z. and Beal, M. (2000), "Variational inference for Bayesian mixtures of factor analysers," *Advances in neural information processing systems*, 12, 449–455.

Ghosh, J. and Dunson, D. B. (2009), "Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis," *Journal of Computational and Graphical Statistics*, 18, 306–320.

Griffin, J. E. and Steel, M. F. J. (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194.

Gu, J. and Ghosal, S. (2009), "Bayesian ROC curve estimation under binormality using a rank likelihood," *Journal of Statistical Planning and Inference*, 139, 2076–2083.

Hahn, P. R., Carvalho, C. M., and Mukherjee, S. (2013), "Partial Factor Modeling: Predictor-Dependent Shrinkage for Linear Regression," *Journal of the American Statistical Association*, 108, 999–1008.

Hall, P., Racine, J., and Li, Q. (2004), "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, 99, 1015–1026.

Hannah, L. A., Blei, D. M., and Powell, W. B. (2011), "Dirichlet Process Mixtures of Generalized Linear Models," *The Journal of Machine Learning Research*, 999999, 1923–1923–1953–1953.

He, Y., Zaslavsky, A. M., Landrum, M. B., Harrington, D. P., and Catalano, P. (2010), "Multiple imputation in a large-scale complex survey: a practical guide." *Statistical methods in medical research*, 19, 653–70.

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240.

Hoff, P. (2007), "Extending the rank likelihood for semiparametric copula estimation," *Annals of Applied Statistics*, 1, 265–283.

Hoff, P. (2010), "Hierarchical multilinear models for multiway data," .

Hoff, P., Niu, X., and Wellner, J. (2011), "Information bounds for Gaussian copulas," *Arxiv preprint arXiv:1110.3572*, pp. 1–24.

Hult, H. and Lindskog, F. (2002), "Multivariate extremes, aggregation and dependence in elliptical distributions," *Advances in Applied Probability*, 34, 587–608.

Ishwaran, H. and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Ishwaran, H. and Zarepour, M. (2002), "Dirichlet prior sieves in finite normal mixtures," *Statistica Sinica*, 12, 941–963.

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

Jara, A., Hanson, T., Quintana, F. A., Müller, P., and Rosner, G. L. (2011), "DPpackage: Bayesian Semi- and Nonparametric Modeling in R," *Journal of Statistical Software*, 40, 1–30.

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

Karabatsos, G. and Walker, S. G. (2012), "Adaptive-modal Bayesian nonparametric regression," *Electronic Journal of Statistics*, 6, 2038–2068.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2013), "Multiple Imputation of Missing or Faulty Values Under Linear Constraints," .

Klaassen, C. and Wellner, J. (1997), "Efficient estimation in the bivariate normal copula model: normal margins are least favourable," *Bernoulli*, pp. 55–77.

Kottas, A., Müller, P., and Quintana, F. (2005), "Nonparametric Bayesian Modeling for Multivariate Ordinal Data," *Journal of Computational and Graphical Statistics*, 14, 610–625.

Kundu, S. and Dunson, D. B. (2011), "Latent Factor Models for Density Estimation," .

Lang, S. and Brezger, A. (2004), "Bayesian P-Splines," *Journal of Computational and Graphical Statistics*, 13, 183–212.

Little, R. J. A. (1988), "Missing-Data Adjustments in Large Surveys," *Journal of Business & Economic Statistics*, 6, 287–296.

Little, R. J. A. and Schluchter, M. D. (1985), "Maximum likelihood estimation for mixed continuous and categorical data with missing values," *Biometrika*, 72, 497–512.

Liu, C. and Rubin, D. (1998), "Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data," *Biometrika*, 85, 673–688.

Liu, J. S. and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264.

Lopes, H. and West, M. (2004), "Bayesian model assessment in factor analysis," *Statistica Sinica*, 14, 41–68.

Lumley, T. (2004), "Analysis of Complex Survey Samples," *Journal of Statistical Software*, 09.

MacEachern, S. N. (1999), "Dependent nonparametric processes," in *Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55, Department of Statistics, The Ohio State University, American Statistical Association.

MacEachern, S. N. (2000), "Dependent Dirichlet Processes," .

Manrique-Vallier, D. and Reiter, J. P. (2012a), "Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros," .

Manrique-Vallier, D. and Reiter, J. P. (2012b), "Bayesian Estimation of Discrete Truncated Latent Structure Models," .

Marin, J., Mengersen, K., and Robert, C. (2005), "Bayesian modelling and inference on mixtures of distributions," *Handbook of statistics*, 25, 459–507.

Marron, J. S. and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models (Wiley Series in Probability and Statistics)*, Wiley-Interscience, 1 edn.

Meng, X. and Van Dyk, D. (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301.

Meng, X.-L. (1994), "Multiple-Imputation Inferences with Uncongenial Sources of Input," *Statistical Science*, 9, 538–558.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010), "Bayesian profile regression with an application to the National survey of children's health," *Biostatistics*, 11, 484–98.

Moustaki, I. and Knott, M. (2000), "Generalized latent trait models," *Psychometrika*, 65, 391–411.

Muller, P. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Murray, J., Dunson, D., Carin, L., and Lucas, J. E. (2013), "Bayesian Gaussian copula factor models for mixed data," *Journal of the ....*

Muthén, B. (1983), "Latent Variable Structural Equation Modeling with Categorical Data," *Journal of Econometrics*, 22, 43–65.

Muthen, B. (1984), "A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators," *Psychometrika*, 49, 115–132.

Neal, R. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, 9, 249–265.

Neal, R. (2011), "MCMC for Using Hamiltonian Dynamics," in *Handbook of Markov Chain Monte Carlo*, chap. 5.

North, D. C. and Weingast, B. R. (1989), "Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England," *The journal of economic history*, 49, 803–832.

Olkin, I. and Tate, R. F. (1961), "Multivariate Correlation Models with Mixed Discrete and Continuous Variables," *The Annals of Mathematical Statistics*, 32, 448–465.

Paisley, J. and Carin, L. (2009), "Nonparametric factor analysis with beta process priors," *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pp. 1–8.

Papaspiliopoulos, O. (2008), "A note on posterior sampling from Dirichlet mixture models," .

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

Pati, D., Bhattacharya, A., and Dunson, D. B. (2011), "Posterior convergence rates in non-linear latent variable models," p. 31.

Piegl, L. a. and Tiller, W. (1998), "Computing the derivative of NURBS with respect to a knot," *Computer Aided Geometric Design*, 15, 925–934.

Pitt, M., Chan, D., and Kohn, R. (2006), "Efficient Bayesian inference for Gaussian copula regression models," *Biometrika*, 93, 537–554.

Polson, N. and Scott, J. (2010), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," *Bayesian Statistics*, 9.

Quinn, K. M. (2004), "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses," *Political Analysis*, 12, 338–353.

R Development Core Team, R. (2011), "R: A Language and Environment for Statistical Computing," .

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey methodology*, 27, 85–96.

Rasmussen, C. E. and Williams, C. K. I. (2005), *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)*, The MIT Press.

Reiter, J. P. and Raghunathan, T. E. (2007), "The Multiple Adaptations of Multiple Imputation," *Journal of the American Statistical Association*, 102, 1462–1471.

Richardson, S. and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)," *Journal of the Royal Statistical Society: Series B (Methodological)*, 59, 731–792.

Rodríguez, A. and Dunson, D. B. (2011), "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian Analysis*, 6, 145–177.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2009), "Bayesian nonparametric functional data analysis through density estimation," *Biometrika*, 96, 149–162.

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley.

Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63, 581–592.

Sammel, M., Ryan, L., and Legler, J. (1997), "Latent Variable Models for Mixed Discrete and Continuous Outcomes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 667–678.

Schafer, J. (1997), *Analysis of incomplete multivariate data*, CRC press.

Scheipl, F. and Kneib, T. (2009), "Locally adaptive Bayesian P-splines with a Normal-Exponential-Gamma prior," *Computational Statistics & Data Analysis*, 53, 3533–3552.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*.

Shahbaba, B. and Neal, R. (2009), "Nonlinear Models Using Dirichlet Process Mixtures," *The Journal of Machine Learning Research*, 10, 1829–1850.

Shen, W. and Ghosal, S. (2013), "Adaptive Bayesian density regression for high-dimensional data," .

Si, Y. and Reiter, J. P. (2013), "Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys," *Journal of Educational and Behavioral Statistics*, pp. 1076998613480394–.

Sklar, A. (1959), "Fonctions de répartition à n dimensions et leurs marges," *Publ. Inst. Statist. Univ. Paris*, 8, 229–231.

Song, X.-Y., Pan, J.-H., Kwok, T., Vandenput, L., Ohlsson, C., and Leung, P.-C. (2010), "A semiparametric Bayesian approach for structural equation models." *Biometrical journal. Biometrische Zeitschrift*, 52, 314–32.

Spearman, C. (1904), "" General Intelligence," Objectively Determined and Measured," *The American Journal of Psychology*.

Stan Development Team (2013), "Stan: A C++ Library for Probability and Sampling, Version 2.0.1," .

Stuart, E. A., Azur, M., Frangakis, C., and Leaf, P. (2009), "Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative." *American journal of epidemiology*, 169, 1133–9.

Taddy, M. A. (2010), "Inverse Regression for Analysis of Sentiment in Text," .

Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010), "Bayesian density regression with logistic Gaussian process and subspace projection," *Bayesian Analysis*, 5, 319–344.

van Buuren, S. (2007), "Multiple imputation of discrete and continuous data by fully conditional specification." *Statistical methods in medical research*, 16, 219–42.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011), "mice: Multivariate Imputation by Chained Equations in R," .

Van Buuren, S. and Oudshoorn, K. (1999), "Flexible multivariate imputation by MICE," .

Vermunt, J. and Ginkel, J. V. (2008), "Multiple imputation of incomplete categorical data using latent class analysis," *Sociological ...*, pp. ???–???

Wade, S., Mongelluzzo, S., and Petrone, S. (2011), "An enriched conjugate prior for Bayesian nonparametric inference," *Bayesian Analysis*, 6, 359–385.

Walker, S. G. (2007), "Sampling the Dirichlet Mixture Model with Slices," *Communications in Statistics - Simulation and Computation*, 36, 45–54.

West, M. and Escobar, M. D. (1993), *Hierarchical priors and mixture models, with application in regression and density estimation*, Institute of Statistics and Decision Sciences, Duke University.

Yang, M. and Dunson, D. B. (2010), "Bayesian Semiparametric Structural Equation Models with Latent Variables," *Psychometrika*, 75, 675–693.

Biography

Jared Scott Murray was born in Concord, NH on October 22, 1984, and was raised in the nearby town of Epsom. He attended the University of New Hampshire where he earned a B.S. in Interdisciplinary Mathematics (Statistics) in May 2009. He matriculated at Duke University that fall, and completed his M.S. in statistics in December of 2011. He graduated from Duke with his Ph.D. in 2013, under the direction of Jerome Reiter.