UNIVERSITY OF CALIFORNIA
SANTA CRUZ

A BAYESIAN FRAMEWORK FOR FULLY NONPARAMETRIC ORDINAL REGRESSION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS AND APPLIED MATHEMATICS

by

Maria Noel DeYoreo

September 2014

The Dissertation of Maria Noel DeYoreo
is approved:

Professor Athanasios Kottas, Chair

Professor Marc Mangel

Professor Stephan Munch

Professor Raquel Prado

Professor Bruno Sansó

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

A Bayesian Framework for Fully Nonparametric Ordinal Regression

by

Maria Noel DeYoreo

Traditional approaches to ordinal regression rely on strong parametric assumptions for the regression function and/or the underlying response distribution. While they simplify inference, restrictions such as normality and linearity are inappropriate for most settings, and the need for flexible, nonlinear models which relax common distributional assumptions is clear. Through the use of Bayesian nonparametric modeling techniques, nonstandard features of regression relationships may be obtained if the data suggest them to be present. We introduce a general framework for multivariate ordinal regression, which is not restricted by linearity or additivity assumptions in the covariate effects. In particular, we assume the ordinal responses arise from latent continuous random variables through discretization, and model the latent response-covariate distribution using a Dirichlet process mixture of multivariate normals. We begin with the binary regression setting, both due to its prominent role in the literature and because it requires more specialized model development under our framework. In particular, we use a square-root-free Cholesky decomposition of the normal kernel covariance matrix, which facilitates model identifiability while allowing for appropriate dependence structure. Moreover, this model structure has the computational advantage of simplifying the implementation of Markov Chain Monte Carlo posterior simulation. Next, we develop modeling and inference methods for ordinal regression, including the underdeveloped setting that involves multivariate ordinal responses. Standard

parametric models for ordinal regression suffer from computational challenges arising from identifiability constraints and parameter estimation, whereas due to the flexible nature of the nonparametric model, we overcome these difficulties. The modeling approach is further developed to handle ordinal regressions which are indexed in discrete-time, through use of a dependent Dirichlet process prior, which estimates the unique regression relationship at each time point in a flexible way while incorporating dependence across time. We consider several examples involving synthetic data to study the scope of the proposed methodology with respect to inference and prediction under both standard and more complex scenarios for the underlying data generating mechanism. Moreover, a variety of real data examples are used to illustrate our methods. As this methodology is especially well-suited to problems in ecology and population dynamics, we target applications in these areas. In particular, our methods are used to provide a detailed analysis of a data set on rockfish maturity and body characteristics collected across different years.

Acknowledgments

Many people have contributed in one way or another, either directly or indirectly, to the development of this thesis. First and foremost, I want to thank my advisor, Athanasios Kottas, for his constant encouragement throughout my PhD studies, enthusiastic attitude towards research and advising, and attention to detail, which greatly improved every piece of writing I produced during my studies. I aspire to approach my work with the same amount of care, seriousness, and enthusiasm, and one day, to mentor others, instilling them with the knowledge, curiosity, and confidence to make their own contributions and discover their unique interests within the field of statistics. I know that a large part of my success in graduate school is attributable to Thanasis, who was always more than willing to nominate or recommend me for any award, job, or research opportunity, and for this I am forever grateful.

I have loved being a part of the Applied Math and Statistics department here at UCSC, and that is in large part due to the wonderful group of professors here. I would like to thank them all for their dedication to teaching and mentoring, and for making the department a friendly, positive place to spend the past several years. In particular, those on my dissertation committee deserve special thanks. Bruno Sanso has been a wonderful department chair throughout my time here, welcoming from the start, and affording me the opportunity to teach a summer course, as well as written many letters of recommendation on my behalf. Raquel Prado and Marc Mangel have also provided letters of recommendation for me during my job search, for which I am very thankful. Raquel's time series class was one of my favorites at UCSC, and she chaired my advancement to candidacy committee,

which led to valuable advice on the direction of my future research. Marc is a dedicated teacher and mentor, helped open up many collaborations that I benefited from, and was always willing to include me in his research group meetings and activities. Steve Munch was a valuable member of my advancement committee, and it was through him that I obtained the fisheries data which plays a large role in the final research project of this thesis.

Graduate school would not have been nearly as enjoyable without my fellow graduate students in AMS. They have helped make our windowless office space an enjoyable place to be. I look forward to seeing where the future takes them, and running into them at conferences. Santa Cruz Masters has been such a wonderful team to be a part of, and I want to thank the coaches and my lanemates for keeping me sane during graduate school; I looked forward to the lunchtime swim workouts each day, and will greatly miss being a part of this team.

Finally, all I have ever achieved throughout my life has been possible due to the support of my family. My parents, Jim and Kelly, have always encouraged me to pursue my goals, and have sacrificed so much to help me achieve those goals, while teaching me the value of a well-rounded and balanced life. They are great role models. The love, friendship, and support of my sister Sarah and brother Patrick is so important, and I am incredibly fortunate to have them as my younger siblings. My husband Kamron has been with me day to day through this entire process, beginning in college at UCSB, and throughout graduate school, and I couldn't have done it all without him.

Chapter 1

Introduction

1.1 Motivation and Background

A fundamental problem in statistics lies in quantifying the relationship between a response variable and a set of related variables (covariates) thought to affect the response in some way. The response variable may either be continuous or discrete, and we work within the latter setting, focusing our attention specifically on situations involving one or more ordered categorical responses.

Binary and ordinal responses measured along with covariates are present in applications from fields including the social sciences, economics, and the biological sciences. Social science data is frequently qualitative, often with some ordering. Sample surveys usually result in correlated ordinal data, since respondents assign ratings on ordinal scales to a set of questions, and the responses given by a single rater are correlated. Ordinal data is also encountered in econometrics, since rating agencies, such as Standard and Poor's and Moody's, use an ordinal grading scale. Binary responses arise anytime a yes/no, positive/negative, or

0/1 type of observation is made, for instance representing the presence/absence of a disease or characteristic in a biomedical study.

Motivated by the prevalence of problems requiring the use of binary and ordinal regression, and the unique challenges that stem from the discrete and ordered nature of the response, we aim to develop new methods which have utility in several of the aforementioned settings. From a modeling perspective, interest centers on determining the regression relationship between the responses and covariates, while appropriately accounting for the dependence between variables, as well as the ordinal form of the responses.

Traditional approaches to binary and ordinal regression are based off of linear methods, in which the response is assumed to depend explicitly on a linear combination of the covariates. Specifically, in the binary case, the probability of a positive response is assumed to be equal to a transformation of the linear predictor $\boldsymbol{x}^T\boldsymbol{\beta}$ defined by a particular distribution function, resulting in a generalized linear model (GLM) (McCullagh and Nelder, 1983). Alternatively, the model can be augmented with latent variables (Albert and Chib, 1993), so that the binary or ordinal response Y arises as a discretized version of a latent continuous response Z. Here, it is typically assumed that $Z = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$, with ϵ following some parametric distribution, such as the normal or logistic distribution. When multiple ordinal responses occur for each observation, a multivariate distribution must be introduced for the vector of latent responses $\boldsymbol{Z}$.

A Bayesian version of the GLM allows for uncertainty in the regression coefficients $\boldsymbol{\beta}$, and the resulting predictive inference, however there are clearly limitations inherent in the approach, regardless of the mode of inference. The large body of literature on Bayesian

regression has aimed at developing more flexible regression functions, and quantifying uncertainty in the resulting estimates. A natural way to overcome the restrictive assumptions of parametric linear regression models is to adopt a Bayesian nonparametric approach.

Early work involving Bayesian nonparametric priors has focused on providing more general inference for either the regression function or the error distribution. Semiparametric approaches to binary and ordinal regression move away from the GLM framework, targeting either linearity or the link function. This has been studied through the use of basis function representations for the regression relationship (e.g., Denison et al., 2002), by assigning a nonparametric prior to the distribution function that defines the link (e.g., Newton et al., 1996), or with the addition of a nonparametric random effects term to the linear predictor (e.g., Follmann and Lamberdt, 1989). In contrast to these approaches, our aim is to be simultaneously flexible about the regression relationship, as well as the response distribution.

We work within a Bayesian nonparametric framework, under which priors are placed on the distribution that generates the data, rather than on parameters of a particular assumed distribution as in parametric modeling. Instead of modeling directly the distribution for $\boldsymbol{Z}$ conditional on a fixed covariate vector $\boldsymbol{x}$, we use a version of implied conditional regression, estimating the joint density $f(\boldsymbol{z}, \boldsymbol{x})$, and the marginal covariate density $f(\boldsymbol{x})$, to obtain general inference for the conditional latent response density $f(\boldsymbol{z} \mid \boldsymbol{x})$. Flexible modeling for $f(\boldsymbol{z}, \boldsymbol{x})$ results in a flexible form for the conditional latent response distribution, and hence, also for the implied regression relationship. This approach to curve fitting regression is discussed further in Section 1.2.

The contributions of this thesis are primarily in ordinal regression, as we aim to

develop a set of tools which are powerful for certain regression problems involving binary or ordinal responses, and illustrate the power of our methods using substantial applications from fields we target. The starting point for the methodology lies in Bayesian nonparametric mixture modeling, an overview of which is given in Section 1.2. The main contributions of this work as well as a discussion of the modeling limitations are provided in Section 1.3.

1.2 Bayesian Nonparametric Mixture Modeling

Our goal is to model flexibly the joint latent response-covariate density, which, for a single binary or ordinal response Y, is $f(z, \boldsymbol{x})$. We refer here to a univariate response, but the methods can be extended to multiple responses $\boldsymbol{Y}$, a setting we explore in Chapter 3. A natural extension of the assumption that a single parametric distribution generates the data involves multiple distributions from the same family assumed to generate the data, each with a given probability. Finite mixture modeling achieves increased flexibility in that it can accommodate multimodal or skewed densities, however the number of components in the mixture must be specified, or better yet, treated as random, requiring advanced computational techniques. An arguably better alternative is to place a nonparametric prior on the discrete random mixing distribution.

The mixture model we utilize is a location-scale mixture of multivariate normal distributions for $f(z, \boldsymbol{x})$, which can be expressed as $\int \mathrm{N}(z, \boldsymbol{x}; \boldsymbol{\mu}, \Sigma) dG(\boldsymbol{\mu}, \Sigma)$. Note that a finite mixture model with K components is obtained when $G = \sum_{k=1}^{K} p_k \delta_{(\boldsymbol{\mu}_k, \Sigma_k)}$, for a probability vector $(p_1, \ldots, p_K)$. Alternatively, a discrete nonparametric prior can be placed on G, resulting in a countable mixture, which has clear advantages in terms of selection

of the number of mixture components based on the data, and the ability to accommodate highly nonstandard distributions. In addition, inference can be simpler than under a finite mixture model when the number of components K is large, or treated as random.

We choose to work with the well-studied Dirichlet process (DP) prior (Ferguson, 1973) for the random distribution G, due to its simplicity, analytical tractability, availability of computational techniques for inference, and proven success in many applications. This prior choice results in a DP mixture model (Antoniak, 1974) for the distribution of $(Z, \boldsymbol{X})$. We will denote the DP prior distribution for G by $\mathrm{DP}(\alpha, G_0)$, which is defined in terms of a centering distribution G_0, such that $\mathrm{E}(G) = G_0$, and a precision parameter $\alpha > 0$, which controls how close realizations of G are to G_0. The DP generates almost surely discrete distributions (Ferguson, 1973; Blackwell, 1973). More specifically, using its constructive definition (Sethuraman, 1994), a realization G from a $\mathrm{DP}(\alpha, G_0)$ is almost surely of the form $G = \sum_{l=1}^{\infty} p_l \delta_{\boldsymbol{\theta}_l}$. The locations $\boldsymbol{\theta}_l$ are independent realizations from the centering distribution G_0, and the weights are determined through stick-breaking from beta distributed random variables. In particular, let $v_l \overset{iid}{\sim} \mathrm{beta}(1, \alpha)$, $l = 1, 2, \ldots$, independently of $\{\boldsymbol{\theta}_l\}$, and define $p_1 = v_1$, and for $l = 2, 3, \ldots$, $p_l = v_l \prod_{r=1}^{l-1}(1 - v_r)$. The discreteness of the DP allows for ties in the mixing parameters. The data is therefore clustered into a typically small number of groups relative to the sample size n, with the distribution of the number of distinct groups (clusters) depending on α, where larger α values favor more clusters.

Assuming a continuous response Y and continuous covariates $\boldsymbol{X}$, the implied conditional regression approach of obtaining inference for the conditional density $f(y \mid \boldsymbol{x})$ through estimation for the joint density $f(y, \boldsymbol{x})$ and the marginal $f(\boldsymbol{x})$ is not new. It dates

back to Nadaraya (1964) and Watson (1964), who proposed kernel density estimators for $f(y, \boldsymbol{x})$ and $f(\boldsymbol{x})$. Alternatively, ideas from Bayesian nonparametric density estimation can be adopted, using multivariate normal DP mixtures for the joint response-covariate distribution (as in Müller et al., 1996).

The DP mixture curve fitting approach has not, until recently, been widely utilized. This may be attributed to the fact that estimation techniques used in DP mixture density estimation are not sufficient for general inference in DP mixture conditional regression. In particular, the regression estimates proposed by Müller et al. (1996), and in the more recent work of Rodriguez et al. (2009), provide only an approximate point estimate for the conditional response density $f(y \mid \boldsymbol{x}; G)$. To obtain full inference for the conditional density, and hence the regression function, the posterior distribution for G must be sampled (e.g., Taddy and Kottas, 2010).

In contrast to MCMC techniques which involve marginalizing G over its DP prior (e.g., Escobar and West, 1995; Bush and MacEachern, 1996; Neal, 2000), we use a posterior simulation technique in which the infinite dimensional G is truncated to a finite level N via the constructive definition (Ishwaran and James, 2001). In particular, $G \approx G_N = \sum_{l=1}^{N} p_l \delta_{\boldsymbol{\theta}_l}$, where the atoms $\boldsymbol{\theta}_l = (\boldsymbol{\mu}_l, \Sigma_l)$, remain i.i.d. from G_0, and the weights arise through stick-breaking, such that $v_1, \ldots, v_{N-1} \overset{iid}{\sim} \text{beta}(1, \alpha)$ as in the countable representation, but $v_N = 1$, so that $p_N = 1 - \sum_{l=1}^{N-1} p_l$. Throughout this document, all inferences involving G are based on G_N.

As a consequence of posterior sampling for G, posterior realizations are available for the joint response-covariate density $f(y, \boldsymbol{x}; G)$, the marginal covariate density $f(\boldsymbol{x}; G)$,

and hence the conditional response density $f(y \mid \boldsymbol{x}; G)$ at any fixed $(y, \boldsymbol{x})$. This yields full inference for the implied conditional regression relationship. Partition the mean vector of the normal kernel such that $\boldsymbol{\mu} = (\mu^y, \mu^x)$, where μ^y denotes the mean of Y, and μ^x denotes the mean of $\boldsymbol{X}$, and partition the covariance matrix such that $\Sigma^{yy} = \text{var}(Y)$, $\Sigma^{xx} = \text{cov}(\boldsymbol{X})$, and $\Sigma^{yx} = \text{cov}(Y, \boldsymbol{X})$. The conditional mean $\text{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}; G)$ has the form of a weighted sum of conditional expectations $\sum_{l=1}^{N} w_l(\boldsymbol{x})(\mu_l^y + \Sigma_l^{yx}(\Sigma_l^{xx})^{-1}(\boldsymbol{x} - \mu_l^x))$, with weights $w_l(\boldsymbol{x}) = p_l \text{N}(\boldsymbol{x}; \mu_l^x, \Sigma_l^{xx}) / \sum_{r=1}^{N} p_r \text{N}(\boldsymbol{x}; \mu_r^x, \Sigma_r^{xx})$, which are covariate-dependent. Evaluating this expression over a grid in $\boldsymbol{x}$, provides posterior realizations for the conditional mean regression function. Complete and proper inference under the flexible DP mixture curve fitting approach to regression provides a powerful framework for nonparametric estimation of the conditional response density and the regression function.

An important extension of the DP mixture framework which we utilize in Chapter 4 involves modeling a collection of random distributions which are related in some way, such as over space, time, or covariates. In particular, we develop a dependent Dirichlet process (DDP) prior for data which is indexed in discrete-time. The general DDP formulation (MacEachern, 2000) introduces time-dependence into the weights and atoms of G_t, the random distribution at time t, so that $G_t = \sum_{l=1}^{\infty} p_{l,t} \delta_{\boldsymbol{\theta}_{l,t}}$, where $p_{1,t} = v_{1,t}$, and for $l = 2, 3, \ldots$, $p_{l,t} = v_{l,t} \prod_{r=1}^{l-1} v_{r,t}$. The stick-breaking proportions $\boldsymbol{v_l} = \{v_{l,t} : t = 1, 2, \ldots\}$ and locations $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{l,t} : t = 1, 2, \ldots)$ arise as i.i.d. realizations from stochastic processes indexed in time, and each $v_{l,t}$ must be marginally beta distributed with first parameter equal to 1, so that for any fixed t, G_t is marginally distributed as a DP. Technical details for the DDP, as well as DP mixture models briefly introduced here, will be developed throughout

the thesis in the specific settings in which they are utilized.

1.3 Research Objectives

Our main contribution in this thesis is to develop modeling strategies for binary and ordinal regression that are more flexible from an inferential perspective than existing approaches. Our methods for binary regression can be compared with other recent approaches, in that they share commonalities with the few existing fully nonparametric regression models which treat the covariates as random. However, as the response variables are made more complex, becoming ordinal and finally multivariate ordinal, there are very few methods available for obtaining flexible inference. Fully nonparametric regression models in the multivariate ordinal setting in particular are essentially nonexistent. We also propose an original construction of the DDP in order to model ordinal regression relationships which evolve in discrete-time.

The primary motivations in taking a Bayesian nonparametric curve fitting approach to regression include: the desire to let the data drive the form of the regression relationships, the ability to properly quantify uncertainty in regression functions and other inferences, the implicit modeling for the covariate distribution and accommodation of dependence in covariates, and the ease of implementation relative to the significant attributes afforded by Bayesian nonparametrics.

The proposed models are widely applicable, although there are some limitations of the joint modeling approach to regression. The main restriction concerns dimensionality, as these methods are meant for problems involving small to moderate numbers of covariates.

A very large number of covariates does not present a problem methodologically, but rather, practically. This is a consequence of the need to estimate the entire joint distribution, and to obtain the posterior of the conditional response distribution over a grid of $\boldsymbol{x}$ values. Inference for these conditional relationships can be made more feasible if interest centers only on regressions over a subset of the covariates, or for just a few $\boldsymbol{x}$ values. Another consequence of the joint modeling framework is that covariates are treated as random, which is an attribute in many settings. However, if the covariates are fixed prior to sampling or can not be viewed as random, as in controlled experiments involving pre-determined treatment and control groups, then methods which attempt to model the covariates along with the response are not appropriate.

A further remark involves the types of covariates that may be handled. The modeling builds from a multivariate normal kernel, and thus can accommodate continuous covariates, possibly after transformation. Discrete covariates which have some ordering can also be handled using a latent variable approach, which we apply in the data example of Section 3.3.4, as well as the main data analysis of Chapter 4. However, for nominal categorical covariates we must move beyond mixtures of multivariate normals. In that case, a mixed continuous-discrete kernel can be used, in which the multivariate normal is retained for the latent responses and continuous covariates, and the discrete covariates are incorporated through a discrete kernel component, possibly conditional on the continuous covariates and/or latent responses. Nominal categorical covariates were not explored in this thesis, however the core of the methodology is already in place, requiring relatively minor modification to be applicable to problems involving categorical covariates.

The practical utility of the proposed ordinal regression modeling framework will be demonstrated through a variety of data examples. The proposed methods are particularly well-suited to problems in population dynamics and evolutionary biology, thus many of our data illustrations fall within these areas. In particular, the model for dynamic ordinal regressions is used to provide a detailed analysis of a data set on rockfish maturation and body measurements, recorded across different years.

We begin in Chapter 2 with the special case of binary regression. This is an important problem which should be distinguished from ordinal regression, because it requires a different model and computational techniques for inference; this is primarily due to differences arising from identifiability considerations. Ordinal regression with one or more responses is the focus in Chapter 3, in which theoretical properties of the model are studied, and it is shown to be more flexible and arguably simpler to implement than standard parametric models. A new DDP prior is developed in Chapter 4 for modeling dynamic ordinal regressions, and Chapter 5 provides a detailed analysis of a data set on rockfish maturity and body characteristics collected over time, as well as an illustration using Citigroup stock data. We conclude with some final remarks in Chapter 6. Technical details on proofs of theoretical results, posterior simulation methods, and properties of the DDP model are provided in Appendices A, B, and C, respectively.

Chapter 2

A Fully Nonparametric Modeling Approach to Binary Regression

2.1 Introduction

Standard approaches to binary regression in both classical and Bayesian settings involve potentially restrictive distributional assumptions as well as those of linearity in relating the response to covariates. The simplest, yet commonly used approach to regression with a dichotomous response assumes the probability of positive response is related to a linear combination of the covariates (the linear predictor) via a link function, or transformation by a CDF. That is, given covariates $\boldsymbol{x}$ and regression coefficients $\boldsymbol{\beta}$, it is assumed that $\Pr(Y = 1) = F(\boldsymbol{x}^T\boldsymbol{\beta})$. Common choices for F include the logistic and normal distributions, each resulting in a small range of monotonic, symmetric trends that may be present for the probability response curve, regardless of what the data suggest. Inference

under a GLM framework can be performed using either a classical or Bayesian approach to inference, which are essentially the same in that they are equally restrictive.

There has been substantial effort devoted to relaxing the functional form of the linear predictor, through the use of basis functions, including spline based approaches (e.g., Denison et al., 2002), and generalized additive models (Hastie and Tibshirani, 1990), applied in a Bayesian context by Wood and Kohn (1998). Under these approaches, the linear predictor is modified by applying a smoothing function to each covariate separately and assuming the transformed covariates are additive in their effects. However, the underlying distributional assumption is still present through the link function.

The motivation for Bayesian nonparametric methodology lies in the notion that the model should support a wide range of distributional shapes and regression relationships. In an attempt to create more flexible models to handle asymmetry, which the standard links can not, as well as overdispersion, which arises when the data exhibit more variability than expected under the model for the observations, several Bayesian semiparametric approaches to binary regression have been developed. Early work has targeted either the link, treating it as a random function with a nonparametric prior (Newton et al., 1996; Basu and Mukhopadhyay, 2000), or linearity, for instance, by viewing the intercept of the linear predictor as a random effects term having a nonparametric prior (Follmann and Lamberdt, 1989; Mukhopadyay and Gelfand, 1997; Walker and Mallick, 1997). More recently, Choudhuri et al. (2007) relaxed the linearity assumption by placing a Gaussian process prior on the argument of the inverse link. Trippa and Muliere (2009) assumed each binary response to arise from a random colored tesselation, and placed a DP prior (Ferguson, 1973) on the

space of colored tesselations. Shahbaba and Neal (2009), Dunson and Bhattacharya (2010), and Hannah et al. (2011), have proposed nonparametric solutions to the regression problem with categorical responses, building off the curve fitting approach to regression of Müller et al. (1996), as described in Section 1.2.

The focus of this chapter is on building a Bayesian nonparametric model for binary responses, measured along with covariates. The idea of inducing a regression model through the joint response-covariate distribution is attractive, since in many settings the covariates are not fixed prior to sampling. We target problems of this type, developing a flexible model for fully nonparametric binary regression. The foundation of the proposed methodology is different from the existing nonparametric modeling approaches. The key distinguishing feature of the proposed model involves the introduction of latent continuous responses, in similar spirit to parametric probit models; see, for instance, Albert and Chib (1993).

In Section 2.2, we formulate the mixture model for binary regression. We discuss identifiability for the parameters of the mixture kernel distribution, as well as prior specification approaches, and give details for posterior inference. In Section 2.3, the methodology is applied to simulated data, and problems from environmetrics and evolutionary biology are studied, using data sets from the literature for illustration. Section 2.4 contains further discussion to place our contribution within the existing literature. Technical details on the identifiability result, prior specification and posterior simulation, and the expressions for the model comparison criterion used in Section 2.3 are provided in the Appendices.

2.2 Methodology

2.2.1 The Modeling Approach

Let $\{(y_i, \boldsymbol{x}_i) : i = 1, ..., n\}$ denote the data, where each observation consists of a binary response y_i along with a vector of covariates, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$. The continuous auxiliary variables, z_i, determine the observed binary responses y_i by their sign, such that $y_i = 1$ if and only if $z_i > 0$. Instead of seeking a nonparametric model directly for the regression function, we estimate the joint distribution of latent responses and covariates, $f(z, \boldsymbol{x})$, which induces a flexible model for the regression relationship, $\Pr(Y = 1 \mid \boldsymbol{x})$.

Focusing on p continuous covariates, $\boldsymbol{X}$, and a single binary response Y, with corresponding latent continuous response Z, a normal distribution is a natural choice for the kernel in a mixture representation for $f(z, \boldsymbol{x})$. The Dirichlet process is then used as a prior for the random mixing distribution G, to create a mixture model of the form: $f(z, \boldsymbol{x}; G) = \int \mathrm{N}_{p+1}(z, \boldsymbol{x}; \boldsymbol{\mu}, \Sigma)\mathrm{d}G(\boldsymbol{\mu}, \Sigma)$, $G \mid \alpha, \boldsymbol{\psi} \sim \mathrm{DP}(\alpha, G_0(\cdot; \boldsymbol{\psi}))$, where α is the Dirichlet process precision parameter, and $\boldsymbol{\psi}$ the parameters of the Dirichlet process centering distribution. Applying the constructive definition of the DP with $\boldsymbol{\theta_l} = (\boldsymbol{\mu}_l, \Sigma_l)$, the model admits a representation as a countable mixture of multivariate normals, $f(z, \boldsymbol{x}; G) = \sum_{l=1}^{\infty} p_l \mathrm{N}_{p+1}(z, \boldsymbol{x}; \boldsymbol{\mu}_l, \Sigma_l)$.

For the normal kernel distribution, let $\boldsymbol{\mu} = (\mu^z, \mu^x)$, where μ^z denotes the mean of Z, and μ^x denotes the mean of $\boldsymbol{X}$, and partition the covariance matrix such that $\Sigma^{zz} = \mathrm{var}(Z)$, $\Sigma^{xx} = \mathrm{cov}(\boldsymbol{X})$, a $p \times p$ matrix, and $\Sigma^{zx} = \mathrm{cov}(Z, \boldsymbol{X})$, a row vector of length p. Then, integrating over the latent response z, the induced model for the observables assumes

the form

$$f(y, \boldsymbol{x}; G) = \sum_{l=1}^{\infty} p_l \mathrm{N}_p(\boldsymbol{x}; \mu_l^x, \Sigma_l^{xx}) \mathrm{Bern}\left(y; \Phi\left(\frac{\mu_l^z + \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(\boldsymbol{x} - \mu_l^x)}{(\Sigma_l^{zz} - \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(\Sigma_l^{zx})^T)^{1/2}}\right)\right), \quad (2.1)$$

where $\Phi(\cdot)$ denotes the standard normal distribution function.

Flexible inference for the binary regression functional can be obtained through $\Pr(Y = 1 \mid \boldsymbol{x}; G) = \Pr(Y = 1, \boldsymbol{x}; G)/f(\boldsymbol{x}; G)$. Marginalizing over z in $f(z, \boldsymbol{x}; G)$, the marginal distribution for $\boldsymbol{x}$ is $f(\boldsymbol{x}; G) = \sum_{l=1}^{\infty} p_l \mathrm{N}_p(\boldsymbol{x}; \mu_l^x, \Sigma_l^{xx})$. Hence, the implied conditional regression function can be expressed as a weighted sum of the form $\sum_{l=1}^{\infty} w_l(\boldsymbol{x})\pi_l(\boldsymbol{x})$, with covariate-dependent weights $w_l(\boldsymbol{x}) = p_l \mathrm{N}_p(\boldsymbol{x}; \mu_l^x, \Sigma_l^{xx}) / \sum_{j=1}^{\infty} p_j \mathrm{N}_p(\boldsymbol{x}; \mu_j^x, \Sigma_j^{xx})$, and probabilities

$$\pi_l(\boldsymbol{x}) = \Phi\left(\frac{\mu_l^z + \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(\boldsymbol{x} - \mu_l^x)}{(\Sigma_l^{zz} - \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(\Sigma_l^{zx})^T)^{1/2}}\right), \quad (2.2)$$

which have the probit form with component-specific intercept and slope parameters.

The dependence structure of the mixture kernel in $f(z, \boldsymbol{x}; G)$ is key to obtaining general inference for the implied binary regression function. However, is it sensible to leave all elements of the kernel covariance matrix Σ unrestricted? In the case of a single mixture component, which arises in the limit as $\alpha \to 0^+$, the regression function $\Pr(y = 1 \mid \boldsymbol{x}; G)$ has the form a single normal cumulative distribution function, as given in (2.2). This function takes the same value for any x when μ^z and Σ^{zx} are scaled by a positive constant c, and Σ^{zz} by c^2, indicating that different combinations of μ and Σ result in the same probability of positive response. Hence, there is an identification problem if μ and Σ are unrestricted. This limiting case of our model is a parametric probit model, albeit with random covariates. In this setting, if identification constraints are not imposed, then prior distributions become increasingly important yet difficult to specify, and the use of noninformative priors can be

problematic and create computational difficulties (Hobert and Casella, 1996; McCulloch et al., 2000; Koop, 2003). In addition, empirical evidence based on simulated data suggests that, without parameter restrictions, the correlations implied by the covariance matrices Σ_l are not representative of the correlations that generated the data, and undesirable behavior is present in the uncertainty bands of the binary regression functional at the extreme regions of the covariate space. For these reasons, and the fundamental belief that within a particular cluster or mixture component the corresponding parameters should be identifiable, we now focus on restricting the kernel of the mixture.

Here, we employ the standard definition of likelihood identifiability, such that a parameter θ for a family of distributions $\{f(x \mid \theta) : \theta \in \Theta\}$ is identifiable if distinct values of θ correspond to distinct probability density functions, that is, if $\theta \neq \theta'$, then $f(x \mid \theta)$ is not the same function of x as $f(x \mid \theta')$. Under our setting, the focus is on the kernel of the mixture model for the observed data, $f(y, \boldsymbol{x}; G)$, which has the form

$$k(y, \boldsymbol{x}; \boldsymbol{\eta}) = \mathrm{N}_p(\boldsymbol{x}; \mu^x, \Sigma^{xx})\mathrm{Bern}\left(y; \Phi\left(\frac{\mu^z + \Sigma^{zx}(\Sigma^{xx})^{-1}(\boldsymbol{x} - \mu^x)}{(\Sigma^{zz} - \Sigma^{zx}(\Sigma^{xx})^{-1}(\Sigma^{zx})^T)^{1/2}}\right)\right), \tag{2.3}$$

with $\boldsymbol{\eta} = (\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zz}, \Sigma^{zx})$. Note that if z and $\boldsymbol{x}$ are independent in the mixture kernel, the probability in the Bernoulli response becomes $\Phi(\mu^z/(\Sigma^{zz})^{1/2})$; hence, a restriction – for instance, on Σ^{zz} – is required for identifiability. This is in fact the only restriction necessary to obtain an identifiable kernel, and we thus retain the ability to estimate Σ^{zx}, which is significant in capturing the dependence of Y on $\boldsymbol{X}$ under the mixture distribution. The specific result is given in the following lemma whose proof can be found in Appendix A.1.1.

LEMMA 1. *The parameters* $(\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zx})$ *are identifiable in the model for observed data which has the form in (2.3), provided* Σ^{zz} *is fixed to a constant.*

While intuitively straightforward, fixing Σ^{zz} to a constant is challenging operationally. The usual conditionally conjugate inverse-Wishart choice for $G_0(\Sigma)$ does not offer the solution, due to the single degree of freedom parameter in the inverse-Wishart distribution, which does not allow for one element of Σ to be fixed while freely estimating the rest of the matrix. This problem is overcome by aid of a square-root-free Cholesky decomposition of Σ. This decomposition is useful for modeling longitudinal data (Daniels and Pourahmadi, 2002), as well as specifying conditional independence assumptions for the elements of a normal random vector (Webb and Forster, 2008). Let β be a unit lower triangular matrix, and let Δ be a diagonal matrix with positive elements, $(\delta_1, ..., \delta_{p+1})$, such that $\Delta = \beta\Sigma\beta^T$. Hence, $\Sigma = \beta^{-1}\Delta(\beta^{-1})^T$, where β^{-1} is also lower triangular with all its diagonal elements equal to 1, and $\det(\Sigma) = \prod_{i=1}^{p+1} \delta_i$. Moreover, $\delta_1 = \Sigma^{zz}$, and thus the identifiability restriction can be implemented by setting the first element of Δ equal to a constant value; $\delta_1 = 1$ is used from this point forward. Instead of mixing directly on Σ, the mixing takes place on β and the p free elements of Δ, denoted by $(\delta_2, ..., \delta_{p+1})$. Hence, the mixture model for the joint density of the latent response and covariates is now written as:

$$f(z, \boldsymbol{x}; G) = \sum_{l=1}^{\infty} p_l \mathrm{N}_{p+1}(z, \boldsymbol{x}; \boldsymbol{\mu}_l, {\beta_l}^{-1}\Delta_l(\beta_l^{-1})^T). \tag{2.4}$$

While this decomposition of Σ allows for the necessary flexibility in viewing only part of the covariance matrix as random, its real utility lies in the existence of a conditionally conjugate centering distribution G_0, which enables development of an efficient

Gibbs sampler for posterior simulation. In particular, a multivariate normal G_0 component for the vector $\tilde{\boldsymbol{\beta}}$, which is composed of the $p(p+1)/2 = q$ free elements of β, and independent inverse-gamma components for $\delta_2, ..., \delta_{p+1}$ result in full conditional distributions which are multivariate normal and inverse-gamma, respectively. Therefore, G_0 comprises independent components for $\boldsymbol{\mu}, \tilde{\boldsymbol{\beta}}$, and $\delta_2, ..., \delta_{p+1}$, such that it has the form $\mathrm{N}_{p+1}(\boldsymbol{\mu}; \boldsymbol{m}, V)\mathrm{N}_q(\tilde{\boldsymbol{\beta}}; \boldsymbol{\theta}, C) \prod_{i=2}^{p+1} \mathrm{IG}(\delta_i; \nu_i, s_i)$.

2.2.2 Posterior Inference for Binary Regression

In order to simulate from the full posterior distribution, we utilize the blocked Gibbs sampler, which is based on a finite truncation approximation to G (Ishwaran and Zarepour, 2000; Ishwaran and James, 2001), as described in Section 1.2. Introducing configuration variables $\boldsymbol{L} = (L_1, ..., L_n)$, each taking values in $\{1, ..., N\}$, and letting $\boldsymbol{W}_l = (\boldsymbol{\mu}_l, \tilde{\boldsymbol{\beta}}_l, \boldsymbol{\delta}_l)$, the hierarchical version of the Dirichlet process mixture model for the data given the latent continuous responses, $\boldsymbol{z} = (z_1, ..., z_n)$, becomes

$$
\begin{aligned}
y_i \mid z_i &\overset{ind.}{\sim} 1_{(y_i=1)}1_{(z_i>0)} + 1_{(y_i=0)}1_{(z_i\leq 0)}, \quad i = 1, ..., n \\
(z_i, \boldsymbol{x}_i) \mid \boldsymbol{W}, L_i &\overset{ind.}{\sim} \mathrm{N}_{p+1}(z_i, \boldsymbol{x}_i; \boldsymbol{\mu}_{L_i}, \beta_{L_i}^{-1}\Delta_{L_i}(\beta_{L_i}^{-1})^T), \quad i = 1, ..., n \\
L_i \mid \boldsymbol{p} &\overset{iid}{\sim} \sum_{l=1}^{N} p_l \delta_l(L_i), \quad i = 1, ..., n \\
\boldsymbol{W}_l \mid \boldsymbol{\psi} &\overset{iid}{\sim} \mathrm{N}_{p+1}(\boldsymbol{\mu}_l; m, V)\mathrm{N}_q(\tilde{\boldsymbol{\beta}}_l; \boldsymbol{\theta}, C) \prod_{i=2}^{p+1} \mathrm{IG}(\delta_{i,l}; \nu_i, s_i) \quad l = 1, ..., N
\end{aligned}
$$

and the prior implied for $\boldsymbol{p}$ by the stick-breaking construction defined through beta$(1, \alpha)$ random variables corresponds to a generalized Dirichlet (GD) distribution (Connor and

Mosimann, 1969), with density $f(\boldsymbol{p} \mid \alpha) = \alpha^{N-1} p_N^{\alpha-1}(1-p_1)^{-1}(1-(p_1+p_2))^{-1} \times \cdots \times (1-\sum_{l=1}^{N-2} p_l)^{-1}$ (Ishwaran and James, 2001). The full Bayesian model is completed with a gamma(a_α, b_α) prior for α, with mean a_α/b_α, and with conditionally conjugate hyperpriors for $\boldsymbol{\psi} = (\boldsymbol{m}, V, \boldsymbol{\theta}, C, s_2, \ldots, s_{p+1})$, specifically: $\boldsymbol{m} \sim \mathrm{N}_{p+1}(\boldsymbol{a}_{\boldsymbol{m}}, B_{\boldsymbol{m}})$, $V \sim \mathrm{IW}_{p+1}(a_V, B_V)$, $\boldsymbol{\theta} \sim \mathrm{N}_q(\boldsymbol{a}_{\boldsymbol{\theta}}, B_{\boldsymbol{\theta}})$, $C \sim \mathrm{IW}_q(a_C, B_C)$, and $s_i \overset{ind.}{\sim} \text{gamma}(a_{s_i}, b_{s_i})$, for $i = 2, ..., p+1$. Here, $S \sim \mathrm{IW}_k(a, B)$ indicates that the $k \times k$ positive definite matrix S follows an inverse-Wishart distribution with density proportional to $\mid S \mid^{-(a+k+1)/2} \exp\{-0.5\mathrm{tr}(BS^{-1})\}$. The notation $\delta_{i,l}$ is used for element i of the vector $\boldsymbol{\delta}_l$ corresponding to the diagonal of Δ_l. Moreover, where convenient, we use the Σ notation for the structured covariance matrix, where the elements of Σ are computed through $\Sigma = \beta^{-1}\Delta(\beta^{-1})^T$.

A key feature of the modeling approach is that simulation from the full posterior distribution, $p(\boldsymbol{W}, \boldsymbol{L}, \boldsymbol{p}, \boldsymbol{\psi}, \alpha, \boldsymbol{z} \mid \text{data})$, is possible via Gibbs sampling. We next discuss posterior simulation details focusing on a result that enables Gibbs sampling updates for the parameters that define the covariance matrices of the normal mixture components.

The updates for $\boldsymbol{p}$ and α are generic for any choice of mixture kernel (Ishwaran and Zarepour, 2000). In particular, the implied prior $f(\boldsymbol{p} \mid \alpha)$ is GD with parameter vectors $(1, 1, \ldots, 1)$ and $(\alpha, \alpha, \ldots, \alpha)$, and the remaining term in the full conditional for $\boldsymbol{p}$ is $\prod_{i=1}^n \sum_{l=1}^N p_l \delta_l(L_i) = \prod_{l=1}^N p_l^{M_l}$, which is GD with parameter vectors $(M_1+1, M_2+1, \ldots, M_{N-1}+1)$ and $(N-1+\sum_{l=1}^N M_l, \ldots, 2+M_{N-1}+M_N, 1+M_N)$, where $M_l = \mid \{i : L_i = l\} \mid$ is the size of mixture component l. Hence, the full conditional for $\boldsymbol{p}$ is GD with parameter vectors $(M_1+1, M_2+1, \ldots, M_{N-1}+1)$ and $(\alpha+\sum_{l=2}^N M_l, \ldots, \alpha+M_{N-1}+M_N, \alpha+M_N)$, which can be sample constructively through latent beta$(1+M_l, \alpha+\sum_{r=l+1}^N M_r)$ random

variables, for $l = 1, \ldots, N-1$. Combining the gamma(a_α, b_α) prior for α with $f(\boldsymbol{p} \mid \alpha)$, yields a full conditional for α which is gamma$(a_\alpha + N - 1, b_\alpha - log(p_N))$.

Each L_i, $i = 1, ..., n$, is sampled from a discrete distribution on $\{1, ..., N\}$, with probabilities proportional to $p_l \mathrm{N}_{p+1}(z_i, \boldsymbol{x}_i; \boldsymbol{\mu}_l, \Sigma_l)$, for $l = 1, ..., N$. The full conditional distributions for the components of $\boldsymbol{\psi}$ are easily found using standard conjugate updating. The full conditional distribution for each z_i is a truncated version of the normal distribution $\mathrm{N}(\mu^z_{L_i} + \Sigma^{zx}_{L_i}(\Sigma^{xx}_{L_i})^{-1}(\boldsymbol{x}_i - \mu^x_{L_i}), 1 - \Sigma^{zx}_{L_i}(\Sigma^{xx}_{L_i})^{-1}(\Sigma^{zx}_{L_i})^T)$, with the restriction $z_i > 0$ if $y_i = 1$, and $z_i \leq 0$ if $y_i = 0$.

Letting $\{L^*_j, j = 1, ..., n^*\}$ be the vector of distinct values of $\boldsymbol{L}$, the full conditional for $\boldsymbol{W}_l$ is proportional to $G_0(\boldsymbol{W}_l \mid \boldsymbol{\psi}) \prod_{j=1}^{n^*} \prod_{\{i:L_i = L^*_j\}} \mathrm{N}_{p+1}(z_i, \boldsymbol{x}_i; \boldsymbol{\mu}_{L^*_j}, \beta^{-1}_{L^*_j} \Delta_{L^*_j} (\beta^{-1}_{L^*_j})^T)$. If $l \notin \{L^*_j : j = 1, ..., n^*\}$, then $\boldsymbol{W}_l \sim G_0(\cdot \mid \boldsymbol{\psi})$. If $l \in \{L^*_j : j = 1, ..., n^*\}$, then the full conditional distribution for each element of $\boldsymbol{W}_l = (\boldsymbol{\mu}_l, \tilde{\boldsymbol{\beta}}_l, \delta_{2,l}, ..., \delta_{p+1,l})$ arises from the product of a normal likelihood component, based on $\{(z_i, \boldsymbol{x}_i) : L_i = L^*_j\}$, and the base distribution G_0. Therefore, when $l = L^*_j$, for $j = 1, ..., n^*$, the full conditional for $\boldsymbol{\mu}_l$ is multivariate normal with mean vector $(V^{-1} + M_l \Sigma_l^{-1})^{-1}(V^{-1}\boldsymbol{m} + \Sigma_l^{-1} \sum_{\{i:L_i = l\}} (z_i, \boldsymbol{x}_i)^T)$ and covariance matrix $(V^{-1} + M_l \Sigma_l^{-1})^{-1}$.

Lemma 2, whose proof can be found in Appendix B.1, provides the result for the posterior full conditional distributions of the $\tilde{\boldsymbol{\beta}}_l$ and the $\delta_{i,l}$, for $l = 1, \ldots, N$, and $i = 2, ..., p+1$. Before stating the lemma, we fix the required notation. As discussed earlier, vector $\tilde{\boldsymbol{\beta}}$ consists of the lower triangle of free elements of matrix β. For instance, if $p = 2$, the mixture kernel is a trivariate normal, and the free elements of β are $(\beta_{2,1}, \beta_{3,1}, \beta_{3,2})$, corresponding to $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3)$. The matrix Δ contains vector $\boldsymbol{\delta}$ on its diagonal. Let $r = p+1$

represent the dimension of the mixture kernel. Let $\boldsymbol{d}_i$ be a vector of length $r(r-1)/2 = q$, containing $r-1$ nonzero terms, occurring in elements $k(k+1)/2$ for $k = 1, ..., r-1$. Let T_i be a block diagonal matrix of dimension $q \times q$ with $r-1$ blocks, which can be constructed from square matrices $T_i^1, ..., T_i^{r-1}$ of dimensions $1, ..., r-1$. Matrix T_i^j occurs in rows and columns $j(j-1)/2+1$ to $j(j+1)/2$ of T_i.

LEMMA 2. *Consider the following Bayesian probability model:*

$$(y_{i,1}, ..., y_{i,r}) \mid \boldsymbol{\mu}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\delta} \overset{ind.}{\sim} \mathrm{N}_r(\boldsymbol{\mu}, \beta^{-1}\Delta(\beta^{-1})^T), \quad i = 1, ..., n,$$

with a multivariate normal prior for $\boldsymbol{\mu}$, independent inverse-gamma priors on the diagonal elements of Δ, $\delta_k \sim \mathrm{IG}(\nu_k, s_k)$, $k = 1, ..., r$, and a multivariate normal prior on the vector comprising the lower triangular elements of β, $\tilde{\boldsymbol{\beta}} \sim \mathrm{N}_q(\boldsymbol{\theta}, D)$. Then, the posterior full conditional distribution for δ_k, $k = 1, ..., r$, is an inverse-gamma distribution with shape parameter $\nu_k + 0.5n$ and scale parameter $s_k + 0.5\sum_{i=1}^n \{(y_{i,k} - \mu_k) + \sum_{j<k} \beta_{kj}(y_{i,j} - \mu_j)\}^2$. In addition, the posterior full conditional for $\tilde{\boldsymbol{\beta}}$ is multivariate normal with mean vector $(D^{-1} + \sum_{i=1}^n T_i)^{-1}(D^{-1}\boldsymbol{\theta} + \sum_{i=1}^n T_i \boldsymbol{d}_i)$ and covariance matrix $(D^{-1} + \sum_{i=1}^n T_i)^{-1}$. Here, the non-zero elements of $\boldsymbol{d}_i$ are $-(y_{i,2} - \mu_2)/(y_{i,1} - \mu_1), ..., -(y_{i,r} - \mu_r)/(y_{i,r-1} - \mu_{r-1})$, and the (m,n)-th element of matrix T_i^j, for $j = 1, ..., r-1$, is given by $T_{i,mn}^j = (y_{i,m} - \mu_m)(y_{i,n} - \mu_n)/\delta_{j+1}$, for $m = 1, ..., j, n = 1, ..., j$.

This lemma provides the information necessary to obtain the remaining full conditional distributions, which are available in closed-form. Let $\boldsymbol{y}_i^* = (z_i, \boldsymbol{x}_i)$ denote the augmented latent response-covariate vector, such that $y_{i,1}^* = z_i$ and $y_{i,j+1}^* = x_{ij}$, for $j = 1, ..., p$. Then, when

$l = L_j^*$, for $j = 1, ..., n^*$, the full conditional distribution for $\delta_{k,l}$ is inverse-gamma with shape parameter $\nu_k + 0.5M_l$ and scale parameter $s_k + 0.5 \sum_{\{i:L_i=L_j^*\}} \{(y_{i,k}^* - \mu_{k,l}) + \sum_{j<k} \beta_{kj,l}(y_{i,j}^* - \mu_{j,l})\}^2$. The full conditional for $\tilde{\boldsymbol{\beta}}_l$ is multivariate normal with covariance matrix $(C^{-1} + \sum_{\{i:L_i=L_j^*\}} T_i)^{-1}$, and mean vector $(C^{-1} + \sum_{\{i:L_i=L_j^*\}} T_i)^{-1}(C^{-1}\boldsymbol{\theta} + \sum_{\{i:L_i=L_j^*\}} T_i \boldsymbol{d}_i)$. The p non-zero terms in the vector $\boldsymbol{d}_i$ are $-(y_{i,2}^* - \mu_{2,l})/(y_{i,1}^* - \mu_{1,l}), ..., -(y_{i,p+1}^* - \mu_{p+1,l})/(y_{i,p}^* - \mu_{p,l})$, and for $j = 1, ..., p$, the matrix T_i^j contains elements $T_{i,mn}^j = (y_{i,m}^* - \mu_{m,l})(y_{i,n}^* - \mu_{n,l})/\delta_{j+1,l}$, $m = 1, ..., j, n = 1, ..., j$.

The mixing distribution G, approximated by $(\boldsymbol{p}, \boldsymbol{W})$, is imputed as a component of the posterior simulation algorithm, enabling full inference for any functional of $f(y, \boldsymbol{x}; G)$. The binary regression functional is the main quantity of interest, and is estimated as $\Pr(Y = 1 \mid \boldsymbol{x}; G) = \Pr(Y = 1, \boldsymbol{x}; G)/f(\boldsymbol{x}; G)$, where $f(\boldsymbol{x}; G) = \sum_{l=1}^N p_l \mathrm{N}_p(\boldsymbol{x}; \mu_l^x, \Sigma_l^{xx})$, and $\Pr(Y = 1, \boldsymbol{x}; G) = \sum_{l=1}^N p_l \mathrm{N}_p(\boldsymbol{x}; \mu_l^x, \Sigma_l^{xx})\pi_l(\boldsymbol{x})$, with $\pi_l(\boldsymbol{x})$ given in (2.2). Therefore, full inference for $\Pr(Y = 1 \mid \boldsymbol{x}; G)$ can be readily obtained for any covariate value $\boldsymbol{x}$, providing a point estimate along with uncertainty quantification for the binary regression function. Inference can also be obtained for the covariate distribution, $f(\boldsymbol{x}; G)$, as well as the covariate distribution conditional on a particular value of y, $f(\boldsymbol{x} \mid y; G)$, which we refer to as inverse inferences, discussed further in the context of the data example of Section 2.3.2.

2.2.3 Prior Specification Strategies

We discuss two approaches to hyperprior specification by considering the limiting case of the model as $\alpha \to 0^+$, which corresponds to a single mixture component (Taddy and Kottas, 2010). Both approaches use an approximate range and center of $\boldsymbol{x}$, say $\boldsymbol{r}^x$ and $\boldsymbol{c}^x$, both vectors of length p, with the objective being to center and scale the mixture kernel

appropriately using only a small amount of prior information. Under the assumption of a single mixture component, the marginal moments are given by $\mathrm{E}((Z, \boldsymbol{X})^T) = \boldsymbol{a_m}$, and $\mathrm{cov}((Z, \boldsymbol{X})^T) = \mathrm{E}(\Sigma) + B_{\boldsymbol{m}} + (a_V - p - 2)^{-1} B_V$. We therefore set $\boldsymbol{a_m} = (0, \boldsymbol{c}^x)$, and let $B_{\boldsymbol{m}} = 0.5\mathrm{diag}(1, (r_1^x/4)^2, ..., (r_p^x/4)^2)$, using c_j^x and $(r_j^x/4)^2$ as proxies for the marginal mean and variance of x_j, for $j = 1, ..., p$. We set $a_V = p + 3$, which yields a dispersed prior for V albeit with finite prior expectation, and determine B_V such that $(a_V - p - 2)^{-1} B_V = B_{\boldsymbol{m}}$. Next, we must determine values for the prior hyperparameters associated with $\tilde{\boldsymbol{\beta}}$ and the $\boldsymbol{\delta}$, and this is where the two approaches differ.

The first approach uses prior simulation to induce approximately uniform$(-1, 1)$ priors on all correlations of the mixture kernel covariance matrix, while appropriately centering the variances. Note that the number of correlations grows at a rate of $O(p^2)$, making this approach practically feasible only for a small number of covariates. In particular, with a single covariate the kernel covariance matrix comprises correlation, $\rho = -\tilde{\beta}(\tilde{\beta}^2 + \delta)^{-1/2}$, and variance, $\sigma^2 = \tilde{\beta}^2 + \delta$. Here, $\tilde{\beta}$ and δ are scalar parameters with G_0 components $\mathrm{N}(\theta, c)$ and $\mathrm{IG}(\nu, s)$, respectively, and the hyperpriors are: $\theta \sim \mathrm{N}(a_\theta, b_\theta)$, $c \sim \mathrm{IG}(a_c, b_c)$, and $s \sim \mathrm{gamma}(a_s, b_s)$. We set $\mathrm{E}(\tilde{\beta}) = a_\theta = 0$, and build the specification for the other hyperparameters from $\mathrm{E}(\sigma^2) = b_\theta + b_s^{-1}(\nu - 1)^{-1} a_s + (a_c - 1)^{-1} b_c$. We first fix the shape parameters ν, a_c and a_s to values that yield relatively large prior dispersion, for instance, $\nu = a_c = 2$ results in infinite prior variance for the inverse-gamma distributions. Next, using $(r^x/4)^2$ as a proxy for $\mathrm{E}(\sigma^2)$, we find constants k_1, k_2, k_3, where $k_1 + k_2 + k_3 = 1$, such that $k_1(r^x/4)^2 \approx b_\theta$, $k_2(r^x/4)^2 \approx b_s^{-1}(\nu - 1)^{-1} a_s$, and $k_3(r^x/4)^2 \approx (a_c - 1)^{-1} b_c$, while at the same time the induced prior on ρ is approximately uniform on $(-1, 1)$. Finally, with

k_1, k_2, k_3 specified, b_θ, b_s, and b_c can be determined accordingly.

While this approach is attractive when a relatively noninformative prior is desired, it is difficult to implement with a moderate to large number of covariates. An alternative strategy arises from studying the distribution which is implied for (β, Δ) if Σ is inverse-Wishart distributed. Using properties of partitioned Wishart and inverse-Wishart matrices (e.g., Box and Tiao, 1973; Eaton, 2007, Ch. 8), it can be shown that $\Sigma \sim \text{IW}_{p+1}(v, T)$ implies inverse-gamma distributions for the δ_i, $i = 2, \ldots, p$, and a normal distribution for $\tilde{\boldsymbol{\beta}}$ given the $\{\delta_i\}$. It is customary to specify noninformative priors on the inverse-Wishart scale, usually fixing the degrees of freedom parameter to a small value, and the inverse scale parameter to be a diagonal matrix. Here, we use the smallest possible integer value for v that ensures a finite expectation for the $\text{IW}_{p+1}(v, T)$ distribution, that is, $v = p + 3$, and set $\text{E}(\Sigma) = T = \text{diag}(T_1, \ldots, T_{p+1}) = \text{diag}(1, (r_1^x/4)^2, ..., (r_p^x/4)^2)$. Then, as shown in Appendix A.2, the distributions implied on δ_i, for $i = 2, \ldots, p+1$, are $\text{IG}(0.5(v + i - (p+1)), 0.5T_i)$. Hence, we let $\nu_i = 0.5(v + i - (p+1))$, and $\text{E}(s_i) = 0.5T_i$; for the data examples of Section 2.3, we worked with exponential priors for the s_i resulting in $b_{s_i} = 2/T_i$. Moreover, the $\text{IW}_{p+1}(v, T)$ distribution implies a normal distribution for the i-th row of matrix β, given δ_i; see Appendix A.2. This can be translated into a distribution for $\tilde{\boldsymbol{\beta}}$ conditionally on the δ_i, specifically, a normal distribution with zero mean vector and covariance matrix $\text{BD}(S_1, \ldots, S_p)$, which denotes a block diagonal matrix with elements $S_i = \delta_{i+1}\text{diag}(T_1^{-1}, \ldots, T_i^{-1})$, for $i = 1, ..., p$. Now, after marginalizing out $\boldsymbol{\theta}$, the G_0 prior component for $\tilde{\boldsymbol{\beta}}$ becomes $\text{N}_q(\boldsymbol{a_\theta}, B_{\boldsymbol{\theta}} + C)$. We therefore specify $\boldsymbol{a_\theta}$ to be equal to the zero mean vector, and since we have a further prior on C, and S_i is a function of δ_{i+1}, we set

$B_{\boldsymbol{\theta}} + \mathrm{E}(C) = \mathrm{BD}(\hat{S}_1, \ldots, \hat{S}_p)$, where $\hat{S}_i$ is a proxy for S_i obtained by replacing δ_{i+1} with its marginal prior mean. Finally, $B_{\boldsymbol{\theta}}$ and $\mathrm{E}(C)$ can be specified to be equal to each other or assigned different portions of $\mathrm{BD}(\hat{S}_1, \ldots, \hat{S}_p)$.

The prior for α can be studied separately, as the value of α controls the number of distinct components in the mixture, or the number of clusters (Antoniak, 1974; Escobar and West, 1995; Liu, 1996). If one has some information or beliefs regarding the number of clusters in the data, then the relations $\mathrm{E}(n^* \mid \alpha) \approx \alpha \log((\alpha + n)/\alpha)$, and $\mathrm{Var}(n^* \mid \alpha) \approx \alpha\{\log((\alpha + n)/\alpha) - 1\}$ can be used to specify a prior for α based on the expectation and uncertainty in the number of mixture components. In many applications, it may be best to use a prior on α which favors a small number of clusters.

2.3 Data Illustrations

2.3.1 Simulated Data

In order to study the performance of the model in the simplest regression setting with a single covariate, latent response-covariate data $\{(z_i, x_i) : i = 1, ..., 150\}$ was simulated from a mixture of three bivariate normal distributions. The binary observations were then determined directly from the signs of the latent responses, producing an observed data set $\{(y_i, x_i) : i = 1, ..., 150\}$. As a tool for model validation, the inference from this binary regression model can be compared to that obtained from a model which views $\{(z_i, x_i) : i = 1, ..., 150\}$ as observed data, effectively making the regression problem one of density estimation, using the technique of Müller et al. (1996). In particular, the inference obtained for $\Pr(Y = 1 \mid x; G)$ under the binary regression model can be compared to that for

$\Pr(Z > 0 \mid x; G)$ under the density estimation model. The inference from the model which treats z as observed is viewed as the best that can possibly be achieved under the binary regression model. Under the simulation approach to prior specification, values of $k_1 = k_3 = 0.175$ and $k_2 = 0.65$ correspond to roughly uniform$(-1, 1)$ priors for ρ, and result in $\mathrm{E}(\sigma^2) \approx (r^x/4)^2$.

The inference for $\Pr(Y = 1 \mid x; G)$ under the binary regression model may be compared with that for $\Pr(Z > 0 \mid x; G)$ from the density estimation model, and both may be evaluated in terms of their similarities to the true $\Pr(Z > 0 \mid x)$. Median and 95% interval estimates for the two functions are displayed along with the truth in Figure 2.1. The point estimate for $\Pr(Z > 0 \mid x; G)$ is slightly closer to the truth at the peak, as well as at the extreme values of the covariate space. While there are subtle differences in the estimation for $\Pr(Y = 1 \mid x; G)$ to that for $\Pr(Z > 0 \mid x; G)$, the fact that the proposed model is performing almost as well as the model for density estimation is encouraging. The interval estimates are widest at the endpoints under both models, as is to be expected since there is less data at these regions.

The three bivariate normal components which generated the data had very different correlation structure. One mixture component generated positively correlated data, one generated negatively correlated data, and the other generated uncorrelated data. It would clearly be restrictive to force $\mathrm{cov}(Z, X)$ to be zero in each component, and the proposed model avoids doing so while maintaining identifiability.

The latent unobserved responses are a key component of the modeling strategy. Since the variance of Z within each component must be fixed for identifiability, and the

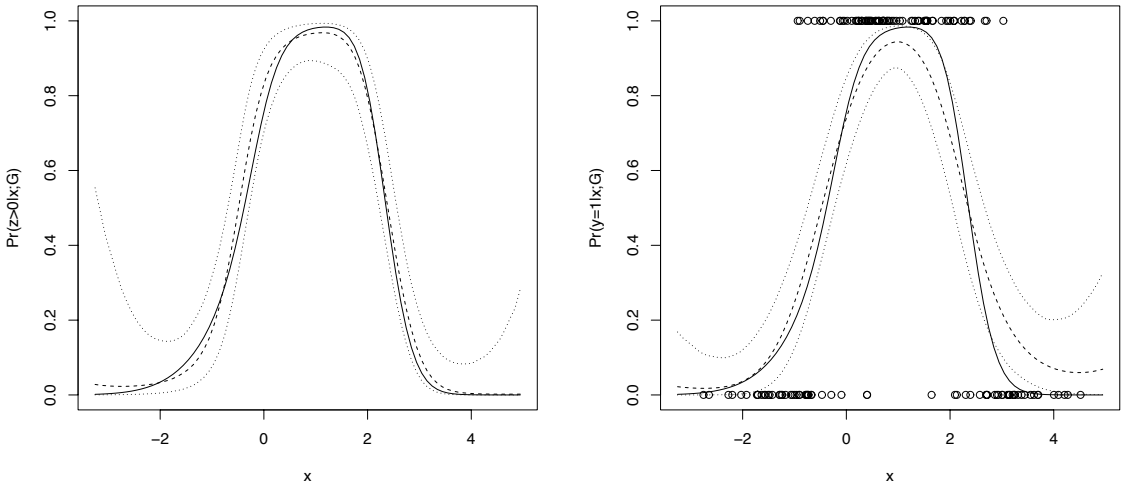

Figure 2.1: Simulation example. Left: Posterior median (dashed line) and 95% uncertainty bands (dotted lines) for $\Pr(Z > 0 \mid x; G)$ are compared with the true data generating $\Pr(Z > 0 \mid x)$ (solid line). Right: Posterior median (dashed line) and 95% uncertainty bands (dotted lines) for $\Pr(Y = 1 \mid x; G)$ are compared with the true data generating $\Pr(Z > 0 \mid x)$ (solid line).

cut-off point set to 0, any inference obtained for Z must be thought of as relative. Hence, while the scale and location of $f(z \mid x; G)$ is not directly interpretable, the inference should meaningful when contrasted against estimates involving different covariate values. Because $\Pr(Y = 1 \mid x; G) = \Pr(Z > 0 \mid x; G)$, the observed binary responses fully inform about the latent response CDF at 0, but not at other values. Within each normal kernel, it can be shown that if $\Pr(Y = 1 \mid x_1) < \Pr(Y = 1 \mid x_2)$, this implies that $\mathrm{E}(Z \mid X = x_1) < \mathrm{E}(Z \mid X = x_2)$. Therefore, although there is some information about the latent responses which is lost in observing only the binary responses, we may be able to compare the estimated latent response distributions across covariate values, particularly their locations. As an illustration of the types of estimated conditional distributions which arise from the model, we show posterior inference for $f(z \mid x; G)$ for three values of x from both the model which

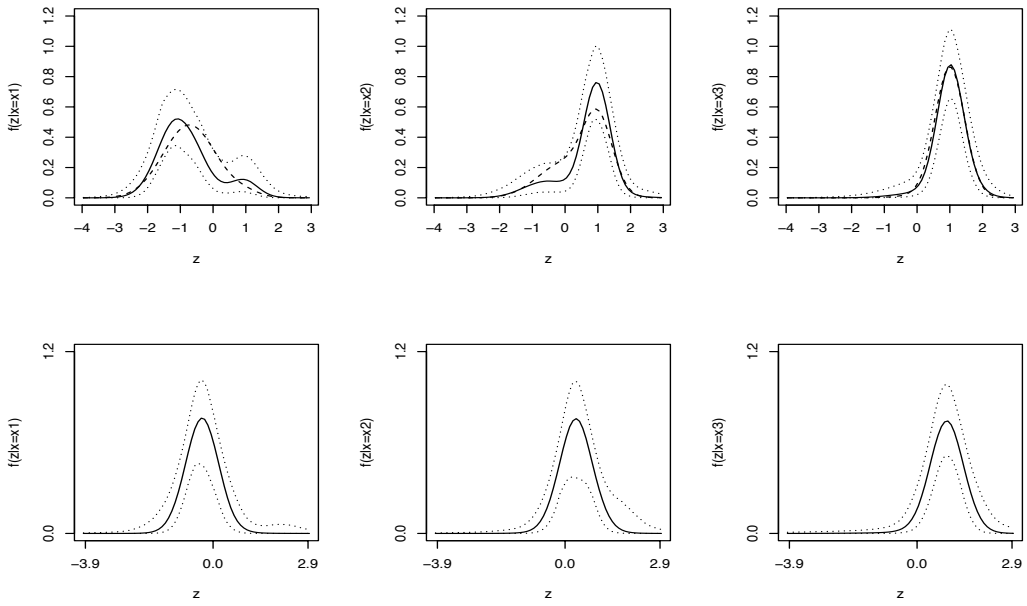

Figure 2.2: Simulation example. The top row shows point (solid) and 95% interval estimates (dotted) for $f(z \mid x; G)$ for three different values of x obtained from the model based on observed data (z_i, x_i), as well as the true densities (dashed). The bottom row shows the corresponding inference obtained from the binary regression model.

actually observes the latent responses, and the binary regression model which sees only the binary responses, in Figure 2.2. These values were chosen from different regions of the covariate space; in particular, x_1 represents the average of the minimum and the mean covariate value, $x_2 = 0$, and x_3 represents the mean. Even though the binary responses tell relatively little about the latent responses, the model is estimating the relative locations very well, and the similarity with the estimates from the model which sees the latent responses is apparent.

2.3.2 Ozone Data

Ozone is a gas which has detrimental consequences when it occurs near the Earth's surface. Ground-level ozone is a harmful pollutant, making up most of the smog which is visible in the sky over large cities. Because of the effects ozone has on the environment and our health, its concentration is monitored by environmental agencies. Rather than recording the actual concentration, presence or absence of an exceedance over a given ozone concentration threshold may be measured, and the number of ozone exceedances in a particular area is of interest.

We work with data set `ozone` from the "ElemStatLearn" R package. The data set includes measurements of ozone concentration in parts per billion, wind speed in miles per hour, temperature in degrees Fahrenheit, and radiation in langleys, recorded over 111 days from May to September of 1973 in New York. To construct a binary ozone exceedance response, we define an exceedance as an ozone concentration which is larger than 70 parts per billion. Therefore, we can model the probability of an ozone exceedance as a function of wind speed, temperature, and radiation, using the Dirichlet process mixture binary regression model. In addition, the modeling approach is evidently appropriate here, since it is natural to estimate conditional relationships between the four environmental variables through modeling the stochastic mechanism for their joint distribution. We are not suggesting dichotomizing a continuous response in practice, but use this example to illustrate a practically relevant setting in which a binary response may arise as a discretized version of a continuous response. Moreover, the existence of the continuous ozone concentrations enables comparison of inferences from the binary regression model with a model based on

the actual continuous responses.

Prior specification was performed using the first approach discussed in Section 2.2.3 that favors uniform priors for the correlations of the kernel covariance matrix. Although the corresponding priors were not all close to the uniform on $(-1, 1)$ under the inverse-Wishart prior specification approach, both methods resulted in prior mean estimates for $\Pr(Y = 1 \mid x_j)$, $j = 1, 2, 3$, that were, for each of the three random covariates, constant around 0.5, with 90% interval bands that essentially span the unit interval. All posterior inference results discussed below were robust to the prior choice.

The marginal binary response curves for the probability of exceedance as a function of wind speed, temperature, and radiation, are shown in the top row of Figure 2.3. There is a decreasing trend in probability as wind speed increases, with the probability being essentially 0 when wind speed is greater than 15 mph. The opposite trend is observed with temperature, as the probability of exceedance is near 0 when temperature is less than 75 degrees, and above 0.8 when temperature exceeds 90 degrees. A non-monotonic unimodal response curve is obtained as a function of radiation, with peak probability occurring at moderate values of radiation, and declining with higher and lower values. Bivariate surfaces indicating probability of exceedance as a function of temperature and wind speed, as well as radiation and wind speed, are shown in Figure 2.4.

For this illustrative data example, the continuous ozone concentration responses are also available. We can therefore compare the binary regression model inferences for $\Pr(Y = 1 \mid x_j)$ with the ones for $\Pr(Z > 70 \mid x_j)$, under the corresponding density estimation model – a Dirichlet process mixture based on a four-dimensional normal kernel with

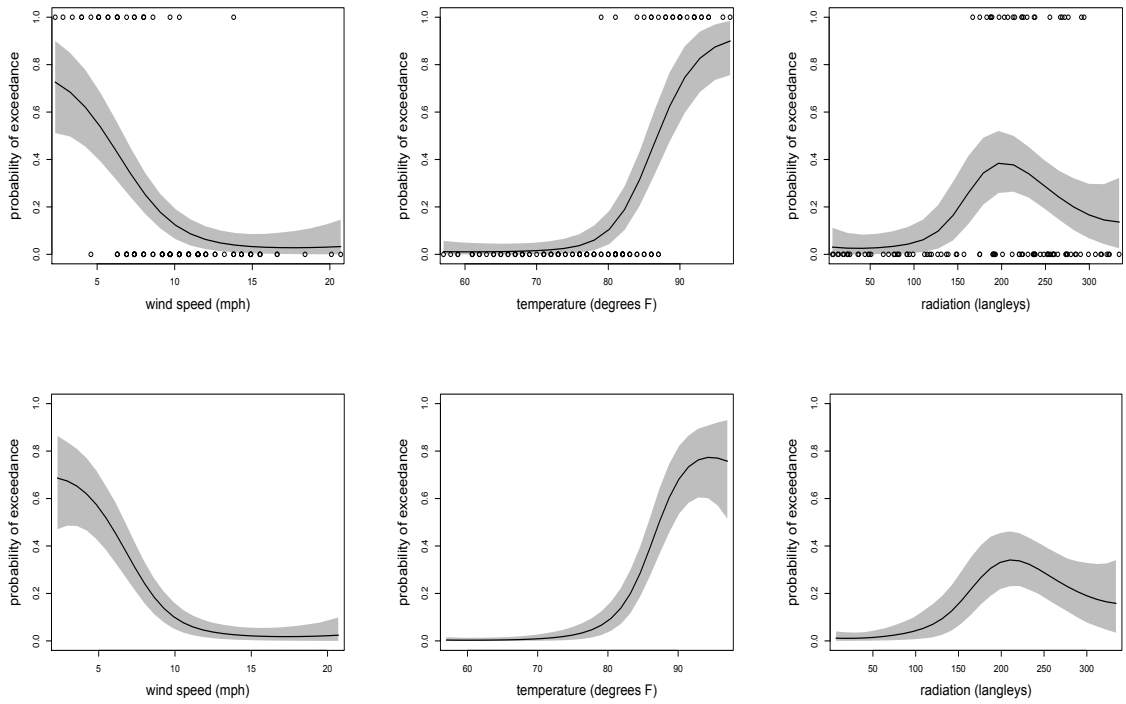

Figure 2.3: Ozone data. Posterior mean (solid line) and 90% uncertainty bands (gray shaded regions) for probability of exceedance versus wind speed (left panels), temperature (middle panels), and radiation (right panels). The top row plots results under the binary regression model, including the binary response data in each panel. The bottom row shows results under the model for density estimation, applied to $\{(z_i, \boldsymbol{x}_i)\}$.

unrestricted covariance matrix – applied to the original data set $\{(z_i, \boldsymbol{x}_i) : i = 1, ..., 111\}$. Results are shown in the bottom row of Figure 2.3, based on a prior choice for the density estimation model that induces prior estimates for the $\Pr(Z > 70 \mid x_j)$ curves that are similarly diffuse to the ones for $\Pr(Y = 1 \mid x_j)$. Save for some differences in the uncertainty bands, the density estimation model reveals similar trends for the regression functions to the ones uncovered by the binary regression model.

As another appealing consequence of estimating the joint response-covariate distribution, we can obtain inference for the distribution of covariates conditional on a particular

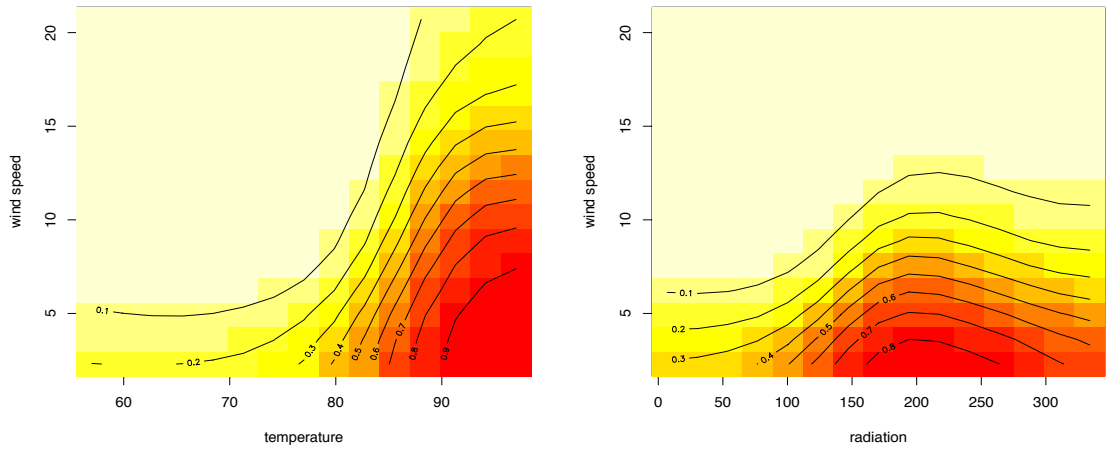

Figure 2.4: Ozone data. Posterior mean surface for probability of exceedance versus temperature and wind speed (left panel), and radiation and wind speed (right panel). Probabilities ranging from 0 to 1 are indicated by a spectrum of colors from white to red.

value of y. These inverse inferences may be of interest in many settings, as they indicate how the covariate distribution differs given a positive versus a negative binary response. Such inferences are not possible under a model directly for the conditional response distribution (with the implicit assumption of fixed covariates). Figure 2.5 shows estimates for the density of each covariate conditional on the binary exceedance response, $f(x_j \mid y = 1)$ and $f(x_j \mid y = 0)$, for $j = 1, 2, 3$. Note that when an exceedance occurs, temperature is generally higher and wind speed lower. In addition, the conditional densities associated with an exceedance have smaller dispersion than those associated with a non-exceedance, indicating that a smaller range of covariate values are supported when an exceedance occurs.

Recall from Section 2.2 that if we make the simplifying assumption $\Sigma^{zx} = 0$ for the covariance matrix of the kernel in $f(z, \boldsymbol{x}; G)$, we obtain a kernel for $f(y, \boldsymbol{x}; G)$ that comprises independent components $\mathrm{N}_p(\boldsymbol{x}; \mu^x, \Sigma^{xx})$ and $\mathrm{Bern}(y; \Phi(\mu^z))$. The implied condi-

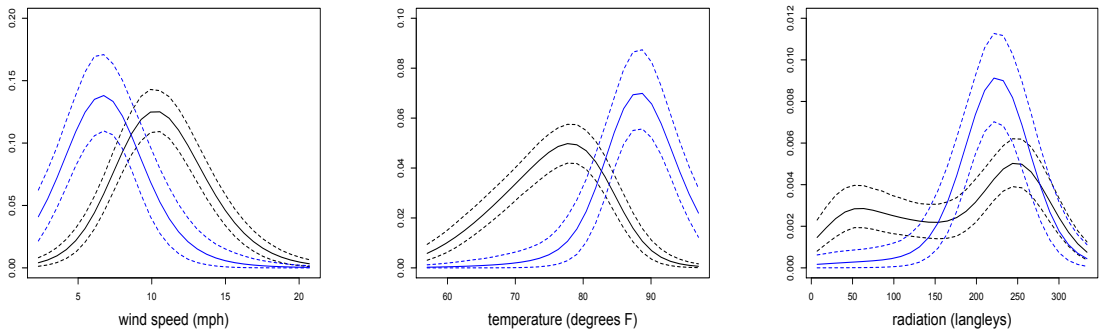

Figure 2.5: Ozone data. Posterior mean estimates (solid lines) and 90% uncertainty bands (dashed lines) for the density of wind speed (left panel), temperature (middle panel), and radiation (right panel), given an ozone concentration exceedance (blue) and non-exceedance (black).

tional regression function is again a weighted sum of probabilities with the same covariate-dependent weights as the proposed model, but probabilities which are not functions of $\boldsymbol{x}$; the probability $\pi_l(\boldsymbol{x})$ in expression (2.2) reduces to $\pi_l = \Phi(\mu_l^z)$. Mixtures of this product-kernel form have been previously proposed in the literature; see, for instance, Dunson and Bhattacharya (2010).

We fitted the simpler product-kernel model to the ozone data, using hyperpriors that induce similarly diffuse prior estimates for the regression functions with the general binary regression model. Differences in the response probabilities produced by the product-kernel mixture model – not shown here – tend to occur at peaks or low points of the curves in Figure 2.3. In general, the product-kernel model underestimates the probability surface or curve when it takes a high value, and overestimates regions of low probability. In addition, the uncertainty bands from the product-kernel model are generally wider than those produced by the proposed model.

For a more formal comparison, we use the posterior predictive loss criterion of

Gelfand and Ghosh (1998). The criterion favors the model m that minimizes the predictive loss measure $D_k(m) = P(m) + \{k/(k+1)\}G(m)$, with penalty term $P(m) = \sum_{i=1}^{n} \text{var}^{(m)}(Y_{new,i} \mid \text{data})$, and goodness of fit term $G(m) = \sum_{i=1}^{n}\{y_i - \text{E}^{(m)}(Y_{new,i} \mid \text{data})\}^2$. Here, $\text{E}^{(m)}(Y_{new,i} \mid \text{data})$ is the mean under model m of the posterior predictive distribution for replicated response $Y_{new,i}$ with corresponding covariate value $\boldsymbol{x}_i$. The variance is similarly defined. Details involving expressions contributing to $D_k(m)$ for each model are given in Appendix B.2, but note that computations are based on the conditional posterior predictive distribution of Y given $\boldsymbol{x}$. The penalty term under the product-kernel model is 10.17, while it is 7.95 under the proposed model, and the goodness of fit terms are 4.17 and 4.08, respectively. Hence, regardless of the choice for constant k, the criterion favors the general Dirichlet process binary regression model.

2.3.3 Estimating Natural Selection Functions in Song Sparrows

In addition to enabling more general modeling of binary regression relationships, the latent variables may be practically relevant in specific applications. Often, we may only observe whether or not some event occurred, although there exists an underlying continuous response which drives the binary observation. The ozone data was used to illustrate an environmental application for which the latent continuous responses are actually present. In applications in biology, the latent response may represent maturity, which is recorded on a discretized scale, or an unobservable trait or measure of health. In general, the continuous responses may be latent either because they are actually unobservable, or as consequence of recording taking place on a discretized scale. As an example of the former scenario, consider a binary response which represents survival. While we only observe survival on a

binary scale, it is meaningful to conceptualize an underlying process which drives survival. Quantifying the probability of survival as a function of phenotypic traits is of great interest in evolutionary biology (Lande and Arnold, 1983; Schluter, 1988; Janzen and Stern, 1998). Survival can be thought of as a measure of fitness, and the fitness surface describes the relationship between phenotypic traits and fitness. The proposed methodology is particularly well-suited for this area of application, as it allows flexible inference for the shape of the fitness surface and for the distribution of population traits under a joint modeling framework that incorporates the scientifically relevant latent fitness responses.

As an illustration, we consider a standard data set from the relevant literature that records overwinter mortality along with six morphological traits in a population of 145 female song sparrows (Schluter, 1988). The traits measured consist of weight, wing length, tarsus length, beak length, beak depth, and beak width. Our initial analysis included four traits – weight, wing length, tarsus length, and beak length – as beak width and depth are highly discretized, correlated with beak length, and did not appear to be associated with a trend in survival. This analysis revealed tarsus length and beak length to be the main targets of selection, which is consistent with the findings of Schluter and Smith (1986). A key objective in this example is to obtain inferences for functionals used to assess the strength and form of natural selection acting on phenotypic traits, and we thus focus on the two traits associated with survival.

The model was applied with standardized covariates tarsus length (X_1) and beak length (X_2), measured in millimeters, using the second approach to prior specification involving the inverse-Wishart distribution. The estimated selection curves are shown in Figure

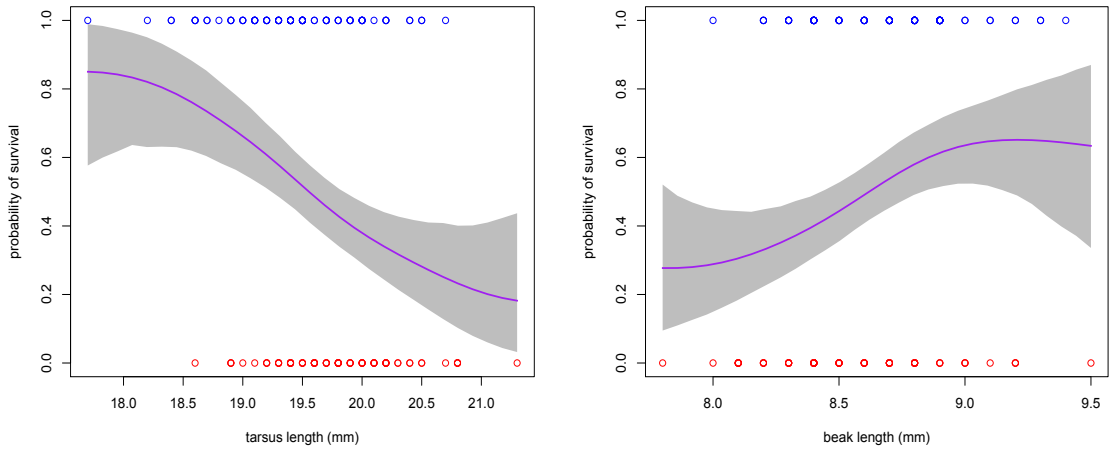

Figure 2.6: Song sparrows data. Posterior mean (purple) and 90% uncertainty bands (gray shaded regions) for the probability of survival as a function of tarsus length and beak length.

2.6, revealing a strong decreasing trend in fitness over tarsus length, in which a sparrow with tarsus length 20.55 millimeters has a 10% lower probability of surviving overwinter than a sparrow with tarsus length just 0.5 mm shorter. The opposite trend in fitness is present over beak length, as longer beaks are associated with higher probabilities of survival. The posterior median estimate for the probability of survival as a function of both traits (Figure 2.7, left panel) confirms that the combination of long beaks and short tarsi is optimal for fitness; importantly, it also indicates that a short tarsus provides the more significant contribution to higher probability of survival. The corresponding posterior interquartile range estimate (Figure 2.7, right panel) depicts more uncertainty in the survival probability surface for sparrows having both a short beak and short tarsus, and those with both a long beak and long tarsus.

For each of the two traits, we estimated the standardized directional selection differential, $\bar{x}_j^* - \bar{x}_j$, $j = 1, 2$, which provides a measure of selection intensity representing

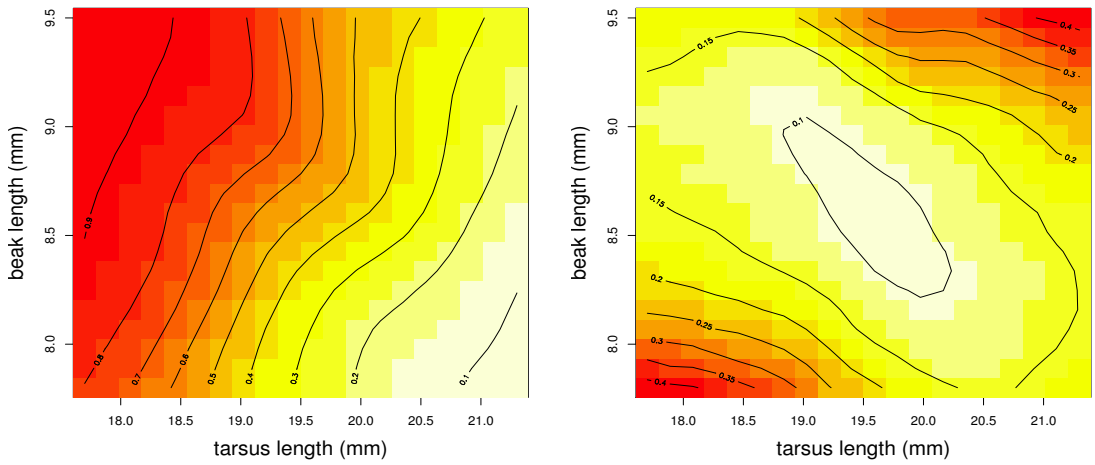

Figure 2.7: Song sparrows data. Posterior median surface (left panel) and interquartile range surface (right panel) for the probability of survival as a function of tarsus length and beak length.

the change in mean value of a phenotype produced by selection (Lande and Arnold, 1983). Here, $\bar{x}_j = \int x_j f(x_j) \mathrm{d}x_j$ is the mean value of phenotypic trait x_j before selection, and $\bar{x}_j^* = \int x_j f(x_j \mid Y = 1) \mathrm{d}x_j = \{\Pr(Y = 1)\}^{-1} \int x_j \Pr(Y = 1, x_j) \mathrm{d}x_j$ is the mean value after selection; the marginal probability $\Pr(Y = 1)$ is referred to as mean absolute fitness. Under our model, $\bar{x}_j = \sum_{l=1}^{N} p_l \mu_l^{x_j}$, mean absolute fitness is given by $\sum_{l=1}^{N} p_l \Phi(\mu_l^z)$, and $\int x_j \Pr(Y = 1, x_j; G_N) \mathrm{d}x_j$ is approximated with a Riemann sum. The posterior mean estimate for the standardized selection differential for tarsus length was -0.31, with a 90% posterior credible interval of $(-0.46, -0.18)$. For beak length, the posterior mean and 90% credible interval for the standardized selection differential were 0.22 and $(0.09, 0.36)$. Note that these intervals do not contain zero. Combined with the estimated regression curves, these results give strong evidence that directional selection is acting on tarsus length and beak length, favoring sparrows with long beaks and short tarsi.

The average gradient of the selection surface, weighted by the phenotype distribution, is given under our model by the vector

$$\left(\int \frac{\partial \Pr(Y = 1 \mid \boldsymbol{x}; G)}{\partial x_1} f(\boldsymbol{x}; G) \mathrm{d}\boldsymbol{x}, \int \frac{\partial \Pr(Y = 1 \mid \boldsymbol{x}; G)}{\partial x_2} f(\boldsymbol{x}; G) \mathrm{d}\boldsymbol{x} \right)^T.$$

Under a linear regression structure with a multivariate normal distribution for the phenotypic traits, the selection gradient is equivalent to the vector of linear regression slopes (Lande and Arnold, 1983). Janzen and Stern (1998) do not incorporate in their approach a distributional assumption for $f(\boldsymbol{x})$, and approximate the j-th selection gradient by $n^{-1} \sum_{i=1}^{n} \partial \Pr(Y = 1 \mid \boldsymbol{x}) / \partial x_j \mid_{\boldsymbol{x}=\boldsymbol{x}_i}$. Our joint mixture modeling approach avoids the assumption of normality for the phenotypic distribution, as well as the need to estimate the integral by assuming the sample represents the population distribution. The integrand of the i-th component of the selection gradient vector can be written as $\{\partial \Pr(Y = 1, \boldsymbol{x}; G)/\partial x_i\} - \{\Pr(Y = 1 \mid \boldsymbol{x}; G) \partial f(\boldsymbol{x}; G)/\partial x_i\}$, for $i = 1, 2$. We omit the specific expressions for each of these two terms, but note that both are analytically available as a consequence of the mixture of normals representation for $f(z, \boldsymbol{x}; G)$. Finally, the average gradient of the relative selection surface, also referred to as the directional selection gradient by Lande and Arnold (1983), is obtained by dividing each element of the selection gradient vector by mean absolute fitness. We obtained posterior mean estimates of -0.27 and 0.18, with corresponding 90% credible intervals of $(-0.40, -0.14)$ and $(0.06, 0.31)$, for the directional selection gradient associated with tarsus length and beak length, respectively.

The presence of stabilizing or disruptive selection can be explored by considering the change in the phenotypic variance-covariance matrix due to selection, that is, the change from the pre-selection covariance matrix P, with elements $\int (x_1 - \bar{x}_1, x_2 -$

$\bar{x}_2)^T(x_1 - \bar{x}_1, x_2 - \bar{x}_2)f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, to the post-selection covariance matrix P^*, with elements $\int(x_1 - \bar{x}_1^*, x_2 - \bar{x}_2^*)^T(x_1 - \bar{x}_1^*, x_2 - \bar{x}_2^*)f(\boldsymbol{x} \mid Y = 1)\mathrm{d}\boldsymbol{x}$. The stabilizing selection differential matrix is given by $P^* - P + (\bar{x}_1^* - \bar{x}_1, \bar{x}_2^* - \bar{x}_2)^T(\bar{x}_1^* - \bar{x}_1, \bar{x}_2^* - \bar{x}_2)$ (Lande and Arnold, 1983), where negative values for a particular trait indicate the presence of stabilizing selection, while positive values indicate disruptive selection. The posterior mean for the matrix element corresponding to tarsus length is 0.038, that for beak length is -0.020, and the off-diagonal element has a posterior mean of -0.018. The 90% posterior credible intervals for each element of the matrix all include zero, indicating lack of significant evidence for stabilizing or disruptive selection acting on either trait.

One way to check if a kernel with independent components for $\boldsymbol{x}$ and y would be adequate is to study in posterior predictive space the correlations between the latent response and the two traits. Denoting by Θ the vector comprising all model parameters, the joint posterior predictive distribution is given by $p(z, \boldsymbol{x} \mid \text{data}) = \int \sum_{l=1}^{N} p_l \mathrm{N}_3(z, \boldsymbol{x}; \boldsymbol{\mu}_l, \Sigma_l)p(\Theta \mid \text{data})\,\mathrm{d}\Theta$, which requires sampling one of $(\Sigma_1, \ldots, \Sigma_N)$ with probabilities $p_1, \ldots, p_N$ for each set of posterior samples. The correlations resulting from these posterior predictive draws for the kernel covariance matrix are plotted in Figure 2.8. These results suggest that it would be restrictive to force uncorrelated mixture kernel components, since the distribution of correlations associated with (Z, X_1) is right-skewed and centered on negative values, while that for (Z, X_2) is mainly focused on positive values and left-skewed, a pattern which is consistent with the shape of the estimated binary regression curves.

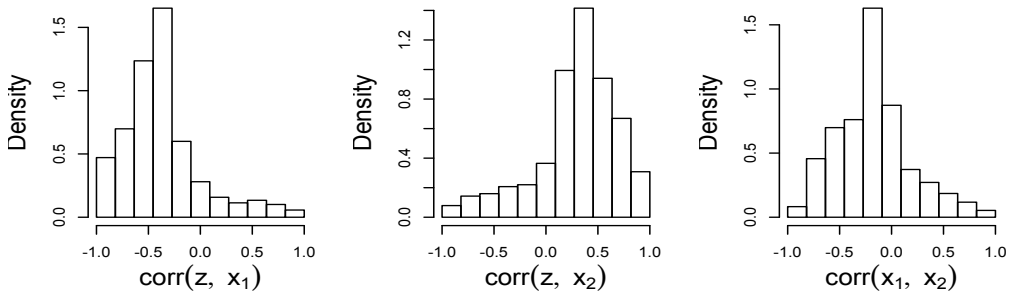

Figure 2.8: Song sparrows data. Posterior predictive samples for $\text{corr}(z, x_1)$ (left panel), $\text{corr}(z, x_2)$ (middle panel), and $\text{corr}(x_1, x_2)$ (right panel).

2.4 Discussion

We have presented a flexible method for estimating the regression relationship between binary responses and continuous covariates, which is built from a DP mixture model for the latent response-covariate distribution. Identifiability was established for the parameters of the mixture kernel, in which the covariance matrix was reparameterized in such a way that allows for viewing only part of the matrix as random, while retaining the desirable features of conjugacy. Full conditional distributions were derived for the random elements of the covariance matrix, providing the key component of an efficient Markov chain Monte Carlo algorithm for posterior simulation. Two strategies for prior specification were discussed. The methodology was illustrated with simulated data and two examples that were chosen to indicate the practical utility of the modeling approach for problems in the environmental sciences and in population biology.

We discussed the special case of the model arising from $\Sigma^{zx} = 0$ in the mixture kernel, which has been previously proposed with the further restriction that Σ^{xx} is diagonal

(Dunson and Bhattacharya, 2010). There, the simplicity of independence among covariates within mixture components was viewed as appealing, and the response was modeled as independent of the covariates within the kernel, resulting in what was termed a product-kernel. In a related approach, Shahbaba and Neal (2009) also build a model for the joint distribution $f(y, \boldsymbol{x})$, but do so by separately estimating $f(\boldsymbol{x})$ and $\Pr(y \mid \boldsymbol{x})$, where the latter is assumed to be a multinomial logit model within a mixture component. Due to the difficulties arising from estimation of full covariance matrices unless the inflexible inverse-Wishart is used as a prior, they too assume $x_1, ..., x_p$ to be independent within each component. This idea was generalized by Hannah et al. (2010) to allow any standard generalized linear model to take the place of the multinomial logit model.

The independence assumptions discussed above are, in general, restrictive. The proposed justification is that because independence is imposed only within each component, dependence arises when more than one component is contained in the mixture. Therefore, the ability of product-kernel models to approximate the regression relationship and the covariate distribution is enhanced through the mixture. However, in order to correctly capture the covariate distribution and the dependence of Y on $\boldsymbol{X}$ in complex problems, there is need for models which allow for dependence within clusters. Dunson and Bhattacharya (2010) note that if interest centers on quantifying dependence, then there is no need to introduce a response, and the method for joint modeling can still be used in this case. If estimation of dependence is in fact the goal, this is clearly more adequately achieved when random variables are allowed to depend on one another through more than just clustering. In this work, the introduction of latent variables and reparameterization of the covariance

matrix allow these assumptions to be relaxed.

The proposed modeling approach relies on the choice of the multivariate normal distribution for the mixture kernel. This choice can accommodate essentially any type of continuous covariate, possibly through use of appropriate transformation. As will become clear in later chapters, it can also handle ordinal categorical covariates $\boldsymbol{X}$ by incorporating in the model associated continuous variables, $\boldsymbol{X}_c$, such that $\boldsymbol{X}$ arises from $\boldsymbol{X}_c$ through discretization. In particular, although in this case inferences were not affected, beak length in the data example of Section 2.3.3 was recorded only to the nearest tenth, and it could therefore be treated as a discrete covariate. In a data example of Section 3.3.4, as well as the main data analysis involving fish maturity of Chapter 4, we will exploit the ordinal nature of the discrete covariates in this way. Extensions of the modeling approach to incorporate ordinal responses follow naturally. This will be the focus of the next chapter.

Chapter 3

A General Framework for Multivariate Ordinal Regression

3.1 Introduction

The idea that categorical random variables arise as discretized versions of underlying latent continuous random variables becomes even more natural when taken to the ordinal setting, as the categorical response levels have a natural ordering. Like the probit GLM for binary responses, the parametric probit model for a multi-category response assumes that $\Pr(Y_i \leq j) = \Phi(\gamma_j - \boldsymbol{x}_i^T\boldsymbol{\beta})$, for $j = 1, \ldots, C$, and cut-offs $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_{C-1} < \gamma_C = \infty$, with $\gamma_1 = 0$ for identifiability. In terms of latent responses (introduced by Albert and Chib, 1993), the model assumes $Y_i = j$ if and only if $Z_i \in (\gamma_{j-1}, \gamma_j]$, for $j = 1, \ldots, C$, and $Z_i \sim \mathrm{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1)$, to give $\Pr(Y_i = j) = \int_{\gamma_{j-1}}^{\gamma_j} \mathrm{N}(z_i; \boldsymbol{x}_i^T\boldsymbol{\beta}, 1)dz_i$.

The multivariate binary probit model (Ashford and Sowden, 1970) generalizes

the binary probit model to accommodate correlated binary responses using a multivariate normal distribution for the underlying latent variables. In this setting, $\boldsymbol{Z}_i$ is a vector, and maximum likelihood estimation is intractable when more than just a few responses are present. To obtain an identifiable model, restrictions must be imposed on the covariance matrix of the multivariate normal distribution for $\boldsymbol{Z}_i$. One way to handle this is to restrict the covariance matrix to be a correlation matrix, which complicates Bayesian inference since there does not exist a conjugate prior for correlation matrices. Chib and Greenberg (1998) discuss inference under this challenging model, using a random walk Metropolis algorithm to sample the correlation matrix, however the matrix generated is not guaranteed to be positive definite. Imai and van Dyk (2005) and Liu (2001) use parameter expansion with data augmentation as developed by Liu and Wu (1999) to expand the parameter space so that unrestricted covariance matrices may be sampled, and a one-to-one mapping is used to imply a set of draws for correlation matrices. Talhouk et al. (2012) worked with a sparse correlation matrix arising from conditional independence assumptions, and used a parameter expansion strategy to expand the correlation matrix into a covariance matrix, and update the covariance matrix in a Gibbs sampling step.

To avoid the issue of constrained covariance matrices in the multivariate ordinal probit model, Webb and Forster (2008) reparameterized Σ in such a way that it is simple to fix its diagonal elements without sacrificing closed-form full conditional distributions. Lawrence et al. (2008) used a parameter expansion technique, in which the parameter space includes unrestricted covariance matrices, which are then normalized to correlation matrices. The multivariate ordinal probit model brings in an additional level of complexity since it

requires estimation for the cut-offs in addition to the challenges arising from correlation matrices. The cut-offs are often highly correlated with the latent responses, suffering from problems of posterior degeneracy.

The assumption of normality on the latent variables is restrictive, in particular for data which contains a large proportion of observations at high or low ordinal levels, and relatively few observations at moderate levels. As a consequence of the unimodal, bell-shape of the normal distribution, the effect of each covariate on the probability response curves is somewhat restrictive. In particular, it is easy to see that, for an ordinal response Y, $\Pr(Y = 1 \mid x)$ and $\Pr(Y = C \mid x)$ are monotonically increasing or decreasing as a function of covariate x, and they must have the opposite type of monotonicity. The direction of monotonicity changes exactly once in moving from category 1 to C (referred to as the single-crossing property). In addition, the relative effect of covariates k and l, or the ratio of $\partial \Pr(Y = m \mid \boldsymbol{x})/\partial x_k$ to $\partial \Pr(Y = m \mid \boldsymbol{x})/\partial x_l$, is equal to β_k/β_l, which does not depend on m or $\boldsymbol{x}$. That is, the relative effect of one covariate to another on the probability of response is the same for every ordinal level and any covariate value. See Boes and Winkelmann (2006) for a discussion of some of these properties.

As we saw in the last chapter, even with only one probability curve to be estimated, there is a large set of literature devoted to modeling this function. Bayesian inference which relies on alternative latent-response distributions is somewhat limited, particularly in the multivariate setting. For a univariate ordinal response, Chib and Greenberg (2010) assume that the latent response arises from scale mixtures of normals, and the covariate effects to be additive upon transformation by cubic splines. This allows nonlinearities to be

obtained in the marginal regression curves, however the assumption of additive covariate effects is restrictive. Moreover, there are aspects of the spline-based approach such as prior specification and choice of location and number of knots that make implementing the model non-trivial. Gill and Casella (2009) extend the parametric ordinal probit model by introducing subject-specific random effects terms, and modeling them with a Dirichlet process (DP) prior.

Chen and Dey (2000) modeled the latent variables with scale mixtures of normal distributions, with means linear on the covariates. In the context of multivariate ordinal data without covariates, Kottas et al. (2005) modeled the distribution of the latent variables with a Dirichlet process (DP) mixture of multivariate normals, which is sufficiently flexible to represent essentially any pattern in a contingency table while using fixed cut-offs. This represents a significant advantage in using a nonparametric model, because in the parametric models discussed, the estimation of cut-offs represented a computational burden, requiring nonstandard inferential techniques such as hybrid MCMC samplers (Johnson and Albert, 1999) and reparamaterization to achieve transformed cut-offs which do not have an order restriction (Chen and Dey, 2000).

Finally, related work includes Shahbaba and Neal (2009), Dunson and Bhattacharya (2010), Hannah et al. (2011), and Papageorgiou et al. (2014), as they develop nonparametric models for joint response-covariate distributions. Shahbaba and Neal (2009) considered classification of a univariate response using a multinomial logit kernel, and this was extended by Hannah et al. (2011) to accommodate alternative response types with mixtures of generalized linear models. Dunson and Bhattacharya (2010) studied DP mixtures

of independent kernels, and Papageorgiou et al. (2014) build a model for spatially-indexed data of mixed type (count, categorical, and continuous). These models would not be appropriate for ordinal data, or, particularly in the first three cases, when inferences are to be made on the association or correlation between variables.

The preceding discussion should indicate that there are significant challenges involved in fitting parametric multivariate probit models, and a large amount of research is dedicated to providing new inferential techniques in this setting. While there is clearly interest and utility in this model, the assumption of normality on the latent variables is highly restrictive. There are few existing nonparametric approaches to ordinal regression, and they are virtually nonexistent in the multivariate case. Semiparametric models for binary regression are more common, since in this case there is a single regression function to be modeled. When taken to the setting involving a single ordinal response with $C \geq 3$ classifications, it becomes much harder to incorporate flexible priors for each of the $C - 1$ probability response curves. Semiparametric prior specifications appear daunting in the multivariate ordinal regression setting where, in addition to flexible regression relationships, it is desirable to achieve general dependence structure between the ordinal responses.

In this chapter, we introduce a Bayesian nonparametric regression model for univariate and multivariate ordinal responses. The covariates remain treated as random, and the way in which they affect the response is driven by the data, as we do not assume a linear relationship between the latent responses and covariates as in standard probit regression and its variations, or any independence assumptions in the covariate effects. An appealing

aspect of the nonparametric modeling approach taken is that the cut-offs may be fixed. In Section 3.2 we formulate the model and show that, with fixed cut-offs, it can approximate arbitrarily well any set of probabilities on the ordinal outcomes. To do so, we establish the Kullback-Leibler (KL) property of our prior, proving the induced prior on the space of mixed ordinal-continuous distributions assigns positive probability to all KL neighborhoods of all densities in this space. A variety of data illustrations are provided in Section 3.3. One of these involves analysis of multirater agreement data, in which the association between the ordinal variables is a key inferential objective. This association is described by the correlations between the latent variables in the standard ordinal probit model, termed "polychoric correlations" in the social sciences (e.g., Olsson, 1979). Section 3.4 concludes the chapter.

3.2 Modeling Strategy, Properties, and Inference

3.2.1 Model Formulation

Suppose that k ordinal categorical variables are recorded for each of n individuals, along with p continuous covariates, so that for individual i we observe a response vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ik})$ and a covariate vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, with $y_{ij} \in \{1, \ldots, C_j\}$, and $C_j > 2$. Introduce latent continuous random variables $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,k})$, $i = 1, \ldots, n$, such that $y_{ij} = l$ if and only if $\gamma_{j,l-1} < z_{ij} \leq \gamma_{j,l}$, for $j = 1, \ldots, k$, and $l = 1, \ldots, C_j$. For example, in a biomedical application, y_{i1} and y_{i2} could represent severity of two different symptoms of patient i, recorded on a categorical scale ranging from "no problem" to "severe", along with covariate information weight, age, and blood pressure. The assumption

that the ordinal responses represent discretized versions of latent continuous responses is realistic for many settings, such as the one considered here. Note also that the assumption of random covariates is appropriate in this application, and that the medical measurements are all related and arise in the form of a data vector. This motivates our focus on building a model for the joint density $f(\boldsymbol{z}, \boldsymbol{x})$, which is a multivariate density of dimension $k + p$, which in turn implies a model for the conditional response distribution $f(\boldsymbol{y} \mid \boldsymbol{x})$.

To model $f(\boldsymbol{z}, \boldsymbol{x})$ in a flexible way while retaining interpretability and computational feasibility, we use a DP mixture (Ferguson, 1973; Antoniak, 1974) of multivariate normals model, mixing on the mean vector and covariance matrix. That is, we assume $(\boldsymbol{z}_i, \boldsymbol{x}_i) \mid G \overset{iid}{\sim} \int \mathrm{N}(\cdot; \boldsymbol{\mu}, \Sigma) dG(\boldsymbol{\mu}, \Sigma)$, and place a DP prior on the random mixing distribution G. The hierarchical model is formulated by introducing a latent mixing parameter $\boldsymbol{\theta_i} = (\boldsymbol{\mu}_i, \Sigma_i)$ for each data vector, to give

$$\begin{aligned} (\boldsymbol{z}_i, \boldsymbol{x}_i) \mid \boldsymbol{\theta}_i &\overset{ind.}{\sim} \mathrm{N}(\boldsymbol{\mu}_i, \Sigma_i), \quad i = 1, \ldots, n \\ \boldsymbol{\theta}_i \mid G &\overset{iid}{\sim} G, \quad i = 1, \ldots, n \\ G \mid \alpha, \boldsymbol{m}, V, S &\sim \mathrm{DP}(\alpha, G_0) \end{aligned} \tag{3.1}$$

with $G_0(\boldsymbol{\mu}, \Sigma; \boldsymbol{\psi}) = \mathrm{N}(\boldsymbol{\mu}; \boldsymbol{m}, V)\mathrm{IW}(\Sigma; \nu, S)$. The model is completed with hyperpriors on $\boldsymbol{\psi}$, and a prior on α:

$$\boldsymbol{m} \sim \mathrm{N}(\boldsymbol{a_m}, B_{\boldsymbol{m}}), \quad V \sim \mathrm{IW}(a_V, B_V), \quad S \sim \mathrm{W}(a_S, B_S), \quad \alpha \sim \mathrm{gamma}(a_\alpha, b_\alpha), \tag{3.2}$$

where, for $(k+p)\times(k+p)$ matrices S and V, $\mathrm{W}(a_S, B_S)$ denotes a Wishart distribution with mean $a_S B_S$, and $\mathrm{IW}(a_V, B_V)$ denotes an inverse-Wishart distribution with mean $(a_V - (k + p) - 1)^{-1} B_V$.

From the constructive definition for G, the prior model for $f(\boldsymbol{z}, \boldsymbol{x})$ has an almost sure representation as a countable mixture of multivariate normals, and the proposed model can therefore be seen to be a nonparametric extension of the multivariate probit model, albeit with random covariates. This implies a countable mixture of normals for $f(\boldsymbol{z} \mid \boldsymbol{x}; G)$, from which the latent $\boldsymbol{z}$ may be integrated out to reveal the induced model for the regression relationships. In general, for a multivariate response $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ with an associated covariate vector $\boldsymbol{X}$, the probability that $\boldsymbol{Y}$ takes on the values $\boldsymbol{l} = (l_1, \ldots, l_k)$, where $l_j \in \{1, \ldots, C_j\}$, $j = 1, \ldots, k$ takes the form

$$\Pr(\boldsymbol{Y} = \boldsymbol{l} \mid \boldsymbol{x}; G) = \sum_{r=1}^{\infty} w_r(\boldsymbol{x}) \int_{\gamma_{k,l_k-1}}^{\gamma_{k,l_k}} \cdots \int_{\gamma_{1,l_1-1}}^{\gamma_{1,l_1}} \mathrm{N}(\boldsymbol{z}; m_r(\boldsymbol{x}), S_r) d\boldsymbol{z} \tag{3.3}$$

with covariate-dependent weights $w_r(\boldsymbol{x}) \propto p_r \mathrm{N}(\boldsymbol{x}; \mu_r^x, \Sigma_r^{xx})$ and mean vectors $m_r(\boldsymbol{x}) = \mu_r^z + \Sigma_r^{zx}(\Sigma_r^{xx})^{-1}(\boldsymbol{x} - \mu_r^x)$, and covariance matrices $S_r = \Sigma_r^{zz} - \Sigma_r^{zx}(\Sigma_r^{xx})^{-1}\Sigma_r^{xz}$. Here, $(\boldsymbol{\mu}_r, \Sigma_r)$ are the atoms in the DP prior constructive definition, where $\boldsymbol{\mu}_r$ is partitioned into μ_r^z and μ_r^x according to random vectors $\boldsymbol{Z}$ and $\boldsymbol{X}$, and $(\Sigma_r^{zz}, \Sigma_r^{xx}, \Sigma_r^{zx}, \Sigma_r^{xz})$ are the components of the corresponding partition of covariance matrix Σ_r.

To illustrate, consider a bivariate response $\boldsymbol{Y} = (Y_1, Y_2)$, with covariates $\boldsymbol{X}$. The probability assigned to the event $(Y_1 = l_1) \cap (Y_2 = l_2)$ is obtained using (3.3), which involves evaluating bivariate normal CDFs. However, one may be interested in the marginal relationship between individual components of $\boldsymbol{Y}$ and the covariates. Referring to the example given at the start of this section, we may obtain the probability that both symptoms are severe as a function of $\boldsymbol{X}$, but also how the first varies as a function of $\boldsymbol{X}$. The marginal

inference, $\Pr(Y_1 = l_1 \mid \boldsymbol{x}; G)$, is given by the expression

$$\sum_{r=1}^{\infty} w_r(\boldsymbol{x}) \left\{ \Phi\left(\frac{\gamma_{1,l_1} - m_r(\boldsymbol{x})}{s_r} \right) - \Phi\left(\frac{\gamma_{1,l_1-1} - m_r(\boldsymbol{x})}{s_r} \right) \right\} \tag{3.4}$$

where $m_r(\boldsymbol{x})$ and s_r are the conditional mean and variance for z_1 conditional on $\boldsymbol{x}$ implied by the joint distribution $\mathrm{N}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\mu}_r, \Sigma_r)$. Expression (3.4) provides also the form of the ordinal regression curves in the case of a single response.

It can be seen that the expressions for the regression relationships have the form of countable mixtures, with component-specific kernels which take the form of parametric probit regressions, and weights which are covariate-dependent. This allows one to obtain nonlinear, nonstandard relationships, by favoring a set of parametric models with varying probabilities depending on the location in the covariate space. Many of the limitations of standard parametric models including relative covariate effects which are constant in terms of the covariate and the ordinal level, monotonicity, and the single-crossing property of the response curves are thereby overcome.

3.2.2 Model Properties

In (3.1), Σ was left an unrestricted covariance matrix, and given an inverse-Wishart base distribution in G_0. In a parametric probit model, one way to ensure identifiability, as an alternative to working with correlation matrices, is to fix $\gamma_{j,2}$ (in addition to $\gamma_{j,1}$), for $j = 1, \ldots, k$. As shown by the following result, this extends to the random covariate setting. Under the mixture setting, model identifiability refers to identifiability of the mixture kernel parameters in the induced model for $(\boldsymbol{Y}, \boldsymbol{X})$, so that within a cluster or mixture component, the parameters are identifiable. Lemma 3 establishes that there is no

issue with identifiability in letting Σ be a general covariance matrix, as a consequence of assuming fixed cut-offs $(\gamma_{j,1}, \ldots, \gamma_{j,C_j-1})$, for $j = 1, \ldots, k$.

Lemma 3. *The parameters* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *are identifiable in the kernel of the mixture model for* $(\boldsymbol{Y}, \boldsymbol{X})$ *as long as* $C_j > 2$ *for all* $j = 1, \ldots, k$.

Refer to Appendix A.1.2 for a proof of this result. If $C_j = 2$ for some j, additional restrictions are needed for identifiability, and these are discussed later in Section 3.2.5.

Identifiability is a basic model property, and is achieved here by fixing the cut-offs. However, this may appear to be a significant restriction with respect to the resulting inferences, as under a parametric probit model, fixing all cut-offs would prohibit the model from being able to adequately assign probabilities to the regions determined by the cut-offs. We therefore seek to determine if the nonparametric model with fixed cut-offs is sufficiently flexible to accommodate any distribution in the class being considered. Kottas et al. (2005) provide an informal argument (also utilized by Savitsky and Dalal, 2014) that the normal DP mixture model for multivariate ordinal responses without covariates can approximate arbitrarily well any probability distribution for a contingency table. The basis for this argument is that, in the limit, one mixture component can be placed within each set of cut-offs corresponding to a specific ordinal vector, with the mixture weights assigned accordingly to each cell.

Here, we provide a more formal proof of the full support of our model for ordinal-continuous data. A prior model has large support if it can generate densities which are arbitrarily close to any true data-generating density. In addition to being a desirable prop-

erty on its own, the ramifications of large support are significant, as it is a key condition which is used in the study of posterior consistency (e.g., Ghosh and Ramamoorthi, 2003). Using the KL divergence as a measure of distance, a particular density $f_0(\boldsymbol{w})$ is said to be in the KL support of the prior $\mathcal{P}$, if $\mathcal{P}\{K_\epsilon(f_0(\boldsymbol{w}))\} > 0$ for every $\epsilon > 0$, where $K_\epsilon(f_0(\boldsymbol{w})) = \{f : \int f_0(\boldsymbol{w}) log(f_0(\boldsymbol{w})/f(\boldsymbol{w}))d\boldsymbol{w} < \epsilon\}$. The KL property is said to be satisfied if any true density $f_0(\boldsymbol{w})$ is in the KL support of the prior.

It has been established that the DP location mixture of multivariate normal kernels prior satisfies the KL property (Wu and Ghosal, 2008). That is, if the mixing distribution G is given a DP prior on the space of probability measures on $\boldsymbol{\mu}$, and a normal kernel is chosen so that $f(\boldsymbol{w}; G, \Sigma) = \int \mathrm{N}(\boldsymbol{w}; \boldsymbol{\mu}, \Sigma)dG(\boldsymbol{\mu})$, with Σ a diagonal matrix, then the prior induced on the space of densities assigns positive probability to all KL neighborhoods of all densities. Letting this induced prior be denoted by $\mathcal{P}$, and modeling the joint distribution of $(\boldsymbol{X}, \boldsymbol{Z})$ with a DP location mixture of normals, the KL property yields:

$$\mathcal{P}\left(\{f : \int f_0(\boldsymbol{x}, \boldsymbol{z}) \log(f_0(\boldsymbol{x}, \boldsymbol{z})/f(\boldsymbol{x}, \boldsymbol{z}))d\boldsymbol{x}d\boldsymbol{z} < \epsilon\}\right) > 0 \tag{3.5}$$

for all $\epsilon > 0$ and all densities $f_0(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{D}$, where $\mathcal{D}$ denotes the space of densities on $\mathbb{R}^{p+k}$.

To establish the KL property of the prior on mixed continuous-ordinal distributions $(\boldsymbol{X}, \boldsymbol{Y})$ induced from multivariate continuous distributions $(\boldsymbol{X}, \boldsymbol{Z})$, we must assume there exists a true $p_0(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}^*$, with $\mathcal{D}^*$ the space of distributions on $\mathbb{R}^p \times \{1, \ldots, C_1\} \times \cdots \times \{1, \ldots, C_k\}$. Let $f_0(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{D}$ be a density function such that

$$p_0(\boldsymbol{x}, l_1, \ldots, l_k) = \int_{\gamma_{k,l_k-1}}^{\gamma_{k,l_k}} \cdots \int_{\gamma_{1,l_1-1}}^{\gamma_{1,l_1}} f_0(\boldsymbol{x}, z_1, \ldots, z_k)dz_1 \cdots dz_k, \tag{3.6}$$

for $l_j \in \{1, \ldots, C_j\}$. That is, $f_0(\boldsymbol{x}, \boldsymbol{z})$ is an underlying density on the latent continuous scale which induces the corresponding true distribution on the ordinal variables. Note that

at least one $f_0 \in \mathcal{D}$ does exist for each $p_0 \in \mathcal{D}^*$, with one such f_0 described in Appendix A.3. The next theorem establishes that, as a consequence of the KL property of the DP mixture of normals (3.5), the prior assigns positive probability to all KL neighborhoods of all $p_0(\boldsymbol{x}, \boldsymbol{y})$, as well as all KL neighborhoods of all conditional distributions $p_0(\boldsymbol{y} \mid \boldsymbol{x})$.

Lemma 4. *Assume the true distribution of a mixed continuous-ordinal random variable is $p_0(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{D}^*$, and let $f_0(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{D}$ be the corresponding continuous density function, which satisfies $\mathcal{P}\{K_\epsilon(f_0(\boldsymbol{x}, \boldsymbol{z}))\} > 0$. Then $\mathcal{P}\{K_\epsilon(p_0(\boldsymbol{x}, \boldsymbol{y}))\} > 0$, and $\mathcal{P}\{K_\epsilon(p_0(\boldsymbol{y} \mid \boldsymbol{x}))\} > 0$.*

Lemma 4, which is proved in Appendix A.3, establishes full support for a model arising from a DP location mixture of multivariate normal kernels, a simpler version of our model. Combined together, the properties of identifiability and full support reflect a major advantage of the proposed model. That is, it can approximate arbitrarily well any distribution on $(\boldsymbol{Y}, \boldsymbol{X})$, as well as any conditional distribution for $(\boldsymbol{Y} \mid \boldsymbol{X})$, while at the same time avoiding the need to impute cut-offs or work with correlation matrices, both of which are major challenges in fitting multivariate probit models.

The cut-offs can be fixed to arbitrary increasing values, which we recommend to be equally spaced and centered at zero. As confirmed empirically with simulated data (refer to Section 3.3 for details), the choice of cut-offs does not affect inferences for the ordinal regression relationships, only the center and scale of the latent variables, which must be interpreted relative to the cut-offs.

3.2.3 Prior Specification

To implement the model, we need to specify the parameters of the hyperpriors in (3.2). A default specification strategy similar to that in Section 2.2.3 is developed by considering the limiting case of the model as $\alpha \to 0^+$, which results in a single normal distribution for $(\boldsymbol{Z}, \boldsymbol{X})$. This limiting model is essentially the multivariate probit model, with the addition of random covariates. The only covariate information we use here is an approximate center (such as the midpoint of the data) and range of each covariate, denoted by $\boldsymbol{c}^x = (c_1^x, \ldots, c_p^x)$ and $\boldsymbol{r}^x = (r_1^x, \ldots, r_p^x)$. Then c_m^x and $(r_m^x/4)^2$ are used as proxies for the marginal mean and variance of X_m. We also seek to center and scale the latent variables appropriately, using the cut-offs. Since Y_j is supported on $\{1, \ldots, C_j\}$, latent continuous variable Z_j must be supported on values slightly below $\gamma_{j,1}$, up to slightly above γ_{j,C_j-1}. Let $r_j^z = (\gamma_{j,C_j-1} - \gamma_{j,1})$, and use $r_j^z/4$ as a proxy for the standard deviation of Z_j.

In the limit, with $(\boldsymbol{Z}, \boldsymbol{X}) \mid \boldsymbol{\mu}, \Sigma \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, we find $\mathrm{E}(\boldsymbol{Z}, \boldsymbol{X}) = \boldsymbol{a_m}$, and $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{X}) = a_S B_S (\nu - d - 1)^{-1} + B_V (a_V - d - 1)^{-1} + B_{\boldsymbol{m}}$, with $d = p + k$. Then, assuming each set of cut-offs $(\gamma_{j,0}, \ldots, \gamma_{j,C_j})$ are centered at 0, fix $a_m = (0, \ldots, 0, \boldsymbol{c^x})$. Letting $D = \mathrm{diag}\{(r_1^z/4)^2, \ldots, (r_k^z/4)^2, (r_1^x/4)^2, \ldots, (r_p^x/4)^2\}$, each of the three terms in $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{X})$ can be assigned an equal proportion of the total covariance, and set to $(1/3)D$, or to $(1/2)D$ to inflate the variance slightly. For dispersed but proper priors with finite expectation, ν, a_V, and a_S can be fixed to $d + 2$. Fixing these parameters allows for B_S and B_V to be determined accordingly, completing the default specification strategy for the hyperpriors of $\boldsymbol{m}$, V, and S.

Although we have developed an approach to prior specification which utilizes the

model for $(\boldsymbol{Z}, \boldsymbol{X})$, the focus of this work is in modeling regression functions, so we should also consider the priors which are induced for the regression relationships. In the strategy outlined above, the form of $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{X})$ was diagonal, so that in the prior, we favor independence between $\boldsymbol{Z}$ and $\boldsymbol{X}$. In the expressions for the regression functions in (3.3) and (3.4), it is easy to see that if $\Sigma_l^{zx} = 0$ for all l, then $m_l(\boldsymbol{x}) = \mu_l^z$, and the probabilities (given by the differences in CDFs) no longer depend on $\boldsymbol{x}$. This leads to regression curves which are flat in the prior mean, and not increasing or decreasing over the covariate space, and this method can therefore be considered if noninformative priors are desired, or when it is unknown how the ordinal responses vary over $\boldsymbol{X}$.

3.2.4 Posterior Inference

The hierarchical model involving the truncation approximation to G is expressed as:

$$
\begin{aligned}
y_{ij} = l \quad \text{iff} \quad \gamma_{j,l-1} < z_{ij} \leq \gamma_{j,l}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k \\
(\boldsymbol{z}_i, \boldsymbol{x}_i) \mid \{\boldsymbol{\mu}_l, \Sigma_l\}, L_i \overset{ind.}{\sim} \mathrm{N}(\boldsymbol{\mu}_{L_i}, \Sigma_{L_i}), \quad i = 1, \ldots, n \\
L_i \mid \boldsymbol{p} \overset{iid}{\sim} \sum_{l=1}^{N} p_l \delta_l(L_i), \quad i = 1, ..., n \\
\boldsymbol{p} \mid \alpha \sim \mathrm{GD}((1, 1, \ldots, 1), (\alpha, \alpha, \ldots, \alpha)) \\
(\boldsymbol{\mu}_l, \Sigma_l) \mid \boldsymbol{\psi} \overset{iid}{\sim} \mathrm{N}(\boldsymbol{\mu}_l; \boldsymbol{m}, V)\mathrm{IW}(\Sigma_l; \nu, S), \quad l = 1, \ldots, N
\end{aligned}
$$

and the full model is completed with conditionally conjugate priors on $\boldsymbol{\psi}$ and α as given in (3.2).

All full conditional distributions are available in closed-form, allowing a Gibbs

sampler to be used for sampling from the full posterior distribution $p(\boldsymbol{\mu}, \Sigma, \boldsymbol{L}, \boldsymbol{p}, \alpha, \boldsymbol{\psi}, \boldsymbol{z} \mid$ data). The full conditional distribution for each $\boldsymbol{\mu}_l$ is normal, that for Σ_l is inverse-Wishart, and each L_i is drawn from the discrete distribution on $\{1, \ldots, N\}$. Each latent z_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$, has a truncated normal full conditional distribution, supported on the interval $(\gamma_{j,y_{i,j}-1}, \gamma_{j,y_{i,j}}]$.

The regression functional $\Pr(\boldsymbol{Y} = \boldsymbol{l} \mid \boldsymbol{x}; G)$ (estimated by a truncated version of (3.3)) can be computed over a grid in $\boldsymbol{x}$ at every MCMC iteration. This yields an entire set of samples for this probability at any covariate value $\boldsymbol{x}$. Note that $\boldsymbol{x}$ may include just a portion of the covariate vector or a single covariate, but in full generality this probability could be estimated at any fixed covariate vector $\boldsymbol{x}$. As indicated in (3.4), in the multivariate response setting with $k > 1$, we may want to show inference for just individual components of $\boldsymbol{Y}$ over the covariate space.

In some applications, in addition to modeling how $\boldsymbol{Y}$ varies across $\boldsymbol{X}$, we may also be interested in how the distribution of $\boldsymbol{X}$ changes at different ordinal values of $\boldsymbol{Y}$. As a feature of the joint-modeling approach which treats $\boldsymbol{X}$ as random, we can obtain inference for $f(\boldsymbol{x} \mid \boldsymbol{y}; G_N)$, which can be evaluated at fixed ordinal levels $\boldsymbol{y}$. We refer to these inferences as inverse relationships, which will be obtained for a data example in the next section.

While these functionals involving the mixing distribution are of primary interest, particularly the regression functionals, the association between the ordinal variables in a multivariate ordinal setting may also be a key target of inference. In the social sciences, the correlations between pairs of latent responses, $\text{corr}(Z_r, Z_s)$, are termed polychoric

correlations (Olsson, 1979) when a single multivariate normal distribution is used for the underlying latent response distribution. Here, there are N multivariate normals present, each having a particular probability given by the probability vector $\boldsymbol{p}$, hence we can sample a single $\text{corr}(Z_r, Z_s)$ at each MCMC iteration, providing posterior predictive distributions for polychoric correlations, which can be used to assess overall agreement between pairs of response variables. As an alternative and possibly more informative measure of association, we can obtain inference for probability of agreement over each covariate, or probability of agreement at each ordinal level. These inferences can be used to determine where in the covariate space response variables tend to agree, as well as the ordinal levels which are associated with more agreement. In the social sciences it is common to assess agreement among multiple raters or judges who are assigning a grade to the same item. We illustrate our methods on a data set of this type, referred to as multirater agreement data, in which both estimating regression relationships and modeling agreement are major objectives.

3.2.5 Accommodating Binary Responses

All discussion up to this point has focused on multivariate ordinal responses, with $C_j > 2$ for all j. However, if one or more responses is binary, then the kernel of the model proposed for ordinal responses with unrestricted $\boldsymbol{\mu}$ and Σ is not identifiable. In the univariate probit model studied in Chapter 2, identifiability was facilitated by fixing Σ^{zz}, using the computationally convenient square-root free Cholesky decomposition of Σ which uses the relationship $\Sigma = \beta^{-1}\Delta\beta^{-T}$, with β a unit lower triangular matrix, and Δ diagonal.

When multiple ordinal responses exist and one or more is binary, it follows from the univariate case that we can not hope to estimate all elements of Σ, in particular the

covariance elements corresponding to the binary responses. Identifiability in this setting can be accomplished by fixing the diagonal elements of Σ^{zz} which represent the variances of the latent binary responses. The covariance elements $\Sigma^{z_i z_j}$, $i \neq j$, all remain free, which is important since the association between the responses may be of interest.

The decomposition of Σ used in the univariate binary case may be useful in the multivariate setting as well. The key result here is that if $(W_1, \ldots, W_n) \sim \mathrm{N}(\boldsymbol{\mu}, \beta^{-1}\Delta\beta^{-T})$, then $\mathrm{Var}(W_i \mid W_1, \ldots, W_{i-1}) = \delta_i$, for $i = 2, \ldots, n$. This was used by Webb and Forster (2008) for modeling multivariate binary data. Therefore, if $(\boldsymbol{Z}, \boldsymbol{X}) \sim \mathrm{N}(\boldsymbol{\mu}, \beta^{-1}\Delta\beta^{-T})$, with $(Z_1, \ldots, Z_r)$ binary, and $(Z_{r+1}, \ldots, Z_k)$ ordinal, then fixing δ_1 fixes $\mathrm{var}(Z_1)$, fixing δ_2 fixes $\mathrm{var}(Z_2 \mid Z_1)$, and so on. The scale of the latent binary responses may therefore be constrained by fixing δ_1, the variance of the first latent binary response, Z_1, and the conditional variances $(\delta_2, \ldots, \delta_r)$ of the remaining latent binary responses $(Z_2, \ldots, Z_r)$. The conditional variances $(\delta_{r+1}, \ldots, \delta_{k+p})$ are not restricted, since they correspond to the scale of latent ordinal responses or covariates, which are identifiable under our model with fixed cut-offs.

3.3 Data Examples

3.3.1 Simulated Data

The model was extensively tested on simulated data. We describe here some observations and results from a series of simulations, in which the primary goal was to assess how well the model can estimate the challenging regression functionals, which exhibit highly nonlinear trends. We also explored effects of sample size, choice of cut-offs, and number of

response categories, by modifying the simulation setting in various ways.

Bivariate continuous data $\{(z_i, x_i), i = 1, \ldots, n\}$ was simulated according to a mixture of 4 bivariate normals. A set of cut-off points dividing $\mathbb{R}$ into 3 regions was specified, so that an ordinal response y_i is implied by each continuous z_i, producing data observations $\{(y_i, x_i), i = 1, \ldots, n\}$, to which our model is applied. Two samples of sizes $n = 200$ and $n = 800$ were produced from this simulation setting, to give data $\{(y_i, x_i), i = 1, \ldots, n\}$, with $y_i \in \{1, 2, 3\}$. The effect of the sample size was observed in the uncertainty bands for the regression functions, which were reduced in width and made smoother with the larger sample size. The regression estimates capture the truth well in both cases, but are smoother and more accurate with more data, as expected. The cut-offs were specified as $\gamma_1 = -5$ and $\gamma_2 = 5$. We stated previously that the cut-offs may be fixed to arbitrary increasing values, and that the choice has no impact on inference involving the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$, only between $\boldsymbol{Z}$ and $\boldsymbol{X}$. To test this point, the model is fit to the same data but with cut-offs of $\gamma_1 = -20$ and $\gamma_2 = 20$. The regression functions, i.e. $\Pr(Y = j \mid x; G)$, $j = 1, 2, 3$, are unaffected by the change in cut-offs, as expected. Rather, the scale of the estimated distribution for Z is increased, since $\mid Z \mid > 20$ is needed for observations corresponding to $Y \neq 2$. The model was also applied using the more challenging cut-offs $\gamma_1 = 0$ and $\gamma_2 = 0.1$, which correspond to a narrow interval for Z producing $Y = 2$. These cut-offs force the model to generate components with small variance, lying in the interval $(0, 0.1)$, and it succeeds, producing ordinal regression functions unchanged from the previous set. Figure 3.1 shows inference for the regression functions from the simulation with $n = 200$ using the more challenging cut-off points of $\boldsymbol{\gamma} = (-\infty, 0, 0.1, \infty)$, as well as the simulation

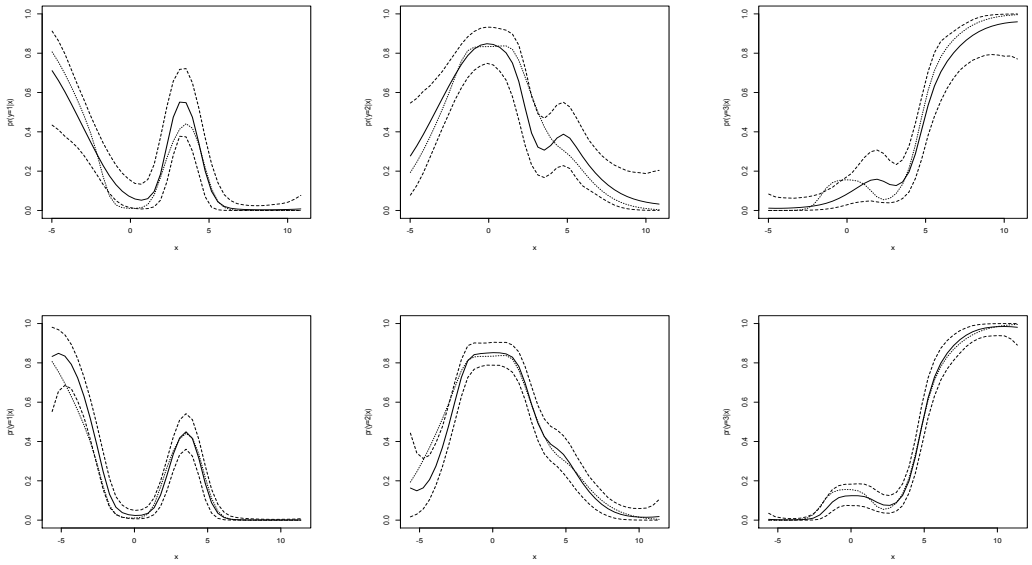

Figure 3.1: Simulated data. Posterior mean (solid) and 95% interval estimates (dashed) for $\Pr(Y = j \mid x; G_t)$, for $j = 1, 2, 3$ by column. Top row shows inference from a simulation with $n = 200$, and bottom row corresponds to a larger sample size $n = 800$. The truth is shown as a dotted line.

with $n = 800$ using the cut-offs points of $\boldsymbol{\gamma} = (-\infty, -5, 5, \infty)$.

Finally, the simulation setting was modified to produce an ordinal response with 5 categories for Y, by partitioning $\mathbb{R}$ into 5 regions. The cut-off points were chosen to create a range of shapes for the regression functions, one having a standard monotonic trend, and some with very nonlinear trends, including unimodality and bimodality. While these highly nonlinear curves may rarely be present in practice, they certainly test the model's ability to capture the truth, however challenging it may be. With only 200 samples spread across 5 ordinal levels, the model captures the regression relationships very well, as can be seen in Figure 3.2.

In the data illustrations that follow, the default prior specification strategy outlined in Section 3.2.3 was used. The posterior distributions for each component of $\boldsymbol{m}$ always

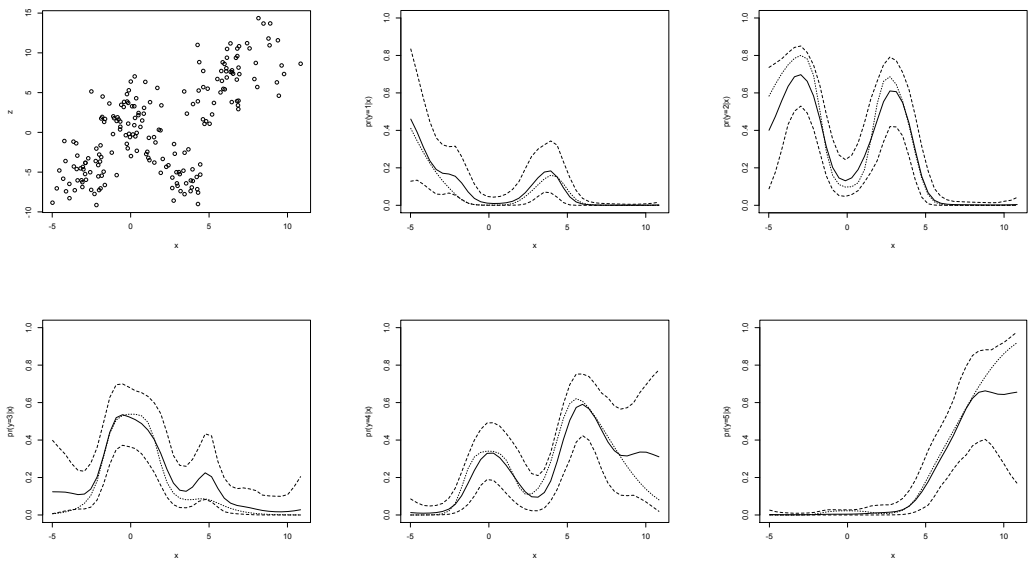

Figure 3.2: Simulated data. Posterior mean (solid) and 95% interval estimates (dashed) for $\Pr(Y = j \mid x; G_t)$, for $j = 1, \ldots, 5$. The truth is shown as a dotted line. The simulated continuous $\{(z_i, x_i)\}$ which generated data observations $\{(y_i, x_i)\}$ is shown on the top left.

appeared very peaked compared to the prior, indicating the prior on $\boldsymbol{m}$ to be sufficiently diffuse. Sensitivity to the prior for V and S was tested by comparing posterior distributions under the strategy outline above to those obtained from more diffuse priors on V and S, by letting the diagonal matrices B_V and B_S have larger elements, also increasing the expectation of V and S element-wise. Some sensitivity to the priors was found in terms of the learning for the hyperparameters V and S, however this was not reflected in the posterior inferences for the regression functions, which displayed little to no change when the priors were altered. The prior for α was also studied, and we noticed a moderate amount of learning taking place for α for larger data sets, and a small amount for smaller data sets, which is consistent with what is known about α in DP mixture models. The priors for α were in all cases chosen to favor reasonably large values relative to the size of the data set,

placing positive probability on a wide range of values for α to be relatively diffuse on the number of components in the mixture.

3.3.2 Ozone Data

One area of application where the proposed model for ordinal data is particularly well-suited falls in the environmental sciences, where observations of some key environmental variable are recorded on the ordinal scale, however there is an underlying continuous random variable which is not observed. If environmental covariates are present, these should in most cases be viewed as random, and modeled jointly along with the variable viewed as the latent response.

To illustrate the application of our methods in a setting of this type, we turn again to the ozone concentration data from Section 2.3.2. We define an ordinal response containing three categories, representing the concentration level. Ozone concentration greater than 100 ppb is defined as a "high" level, and assigned the ordinal level 3. This can be considered an extreme level of ozone concentration, as only about 6% of observations are this high. Concentration falling between 50 and 100 is considered "medium", corresponding to level 2. Approximately 22% of observations fall in this range. Finally, anything less than 50 ppb is probably not of a high enough level to be of concern, and these concentrations are assigned level 1.

In this example continuous ozone concentration is contained in the data, and we create a discretized response indicating ozone concentration level to illustrate the proposed method. However, in other real settings where data on ozone is recorded, it may only be available in an ordinal form, such as whether or not it exceeded a certain threshold,

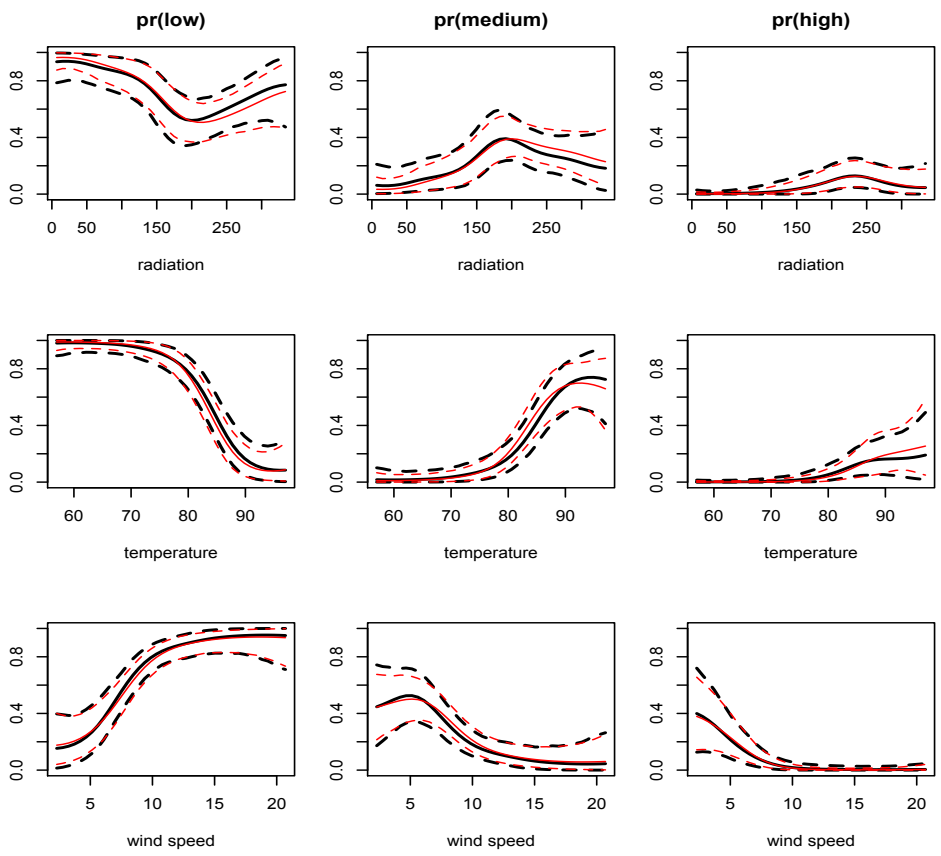

Figure 3.3: Ozone data. Mean (solid) and 95% interval estimates (dashed) for $\Pr(Y = j \mid x_l; G)$ (thick black) compared to $\Pr(\gamma_{j-1} < Z \leq \gamma_j \mid x_l; G)$ (red), for $j = 1, 2, 3$ and $l = 1, 2, 3$, giving the probability of ozone concentration being low, medium, and high over covariates radiation, temperature, and wind speed.

or whether it was "high", "medium", or "low". This idea can be generalized to other environmental characteristics or outcomes, which may only be available on an ordinal scale, but in reality are continuous.

The ordinal regression model was applied to the ozone data, with ozone concentration level response and the three environmental variables as covariates. Two different sets of cut-offs were specified, those being $(-\infty, -3, 3, \infty)$ and $(-\infty, -1, 100, \infty)$, and these had no effect on the inferences reported below. To assess and validate the inferences given

by the model, we can compare the results from the proposed model which only sees the discretized ozone concentration, to one which observes the actual continuous ozone concentration. Specifically, we model the continuous data vector $(Z, \boldsymbol{X})$ with a DP mixture of multivariate normals, using the curve-fitting approach to regression of Müller et al. (1996) and extended by Taddy and Kottas (2010). We compare the univariate regression curves $\Pr(Y = j \mid x_l; G)$ to $\Pr(\gamma_{j-1} < Z \leq \gamma_j \mid x_l; G)$, $j = 1, 2, 3$, and $l = 1, 2, 3$, the latter being from the model which observes Z rather than Y, essentially giving us a benchmark to compare our model to which represents the best possible inference which could be obtained if no loss in information occurred by observing Y rather than Z. Figure 3.3 compares mean and 95% interval estimates for the regression curves produced by the model for ordinal data to the comparable inferences from the model which observes $(z_1, \ldots, z_n)$. The similarity between the two sets of inferences is clear. There are some regions where the interval bands from the ordinal regression model are slightly wider, particularly for the covariate radiation, which is the most nonlinear, however these differences are very subtle, and not even present in some of the inferences. The trends in ozone concentration classification probabilities conditional on temperature as well as wind speed are somewhat standard, exhibiting monotonic relationships, while the regression curves associated with radiation are more nonlinear. The ability to capture such a wide range of patterns is a feature of the flexible nonparametric model for the latent response-covariate distribution.

Figure 3.4 displays bivariate surfaces which illustrate the posterior mean estimates for probability of each level as a function of radiation and temperature. Interaction effects are implicit in the joint response-covariate framework, without the need to account for

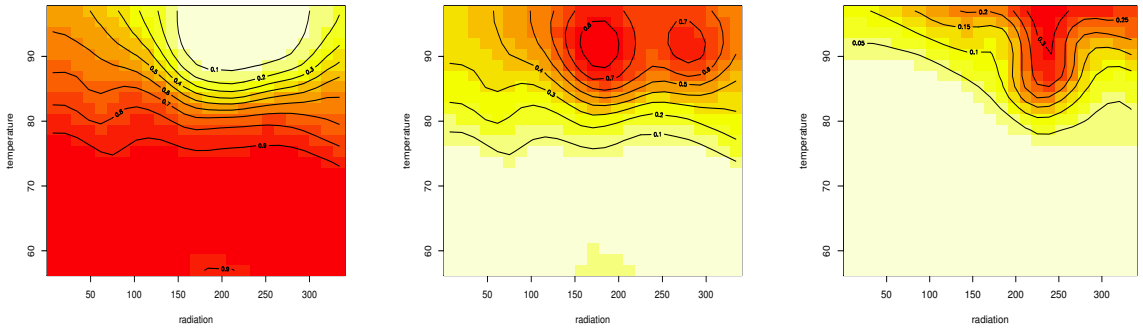

Figure 3.4: Ozone data. Posterior mean estimates for $\Pr(Y = j \mid x_1, x_2; G)$ for $j = 1, 2, 3$, corresponding to low (left), medium (middle) and high (right). White indicates a posterior mean probability 0, and red indicates probability 1.

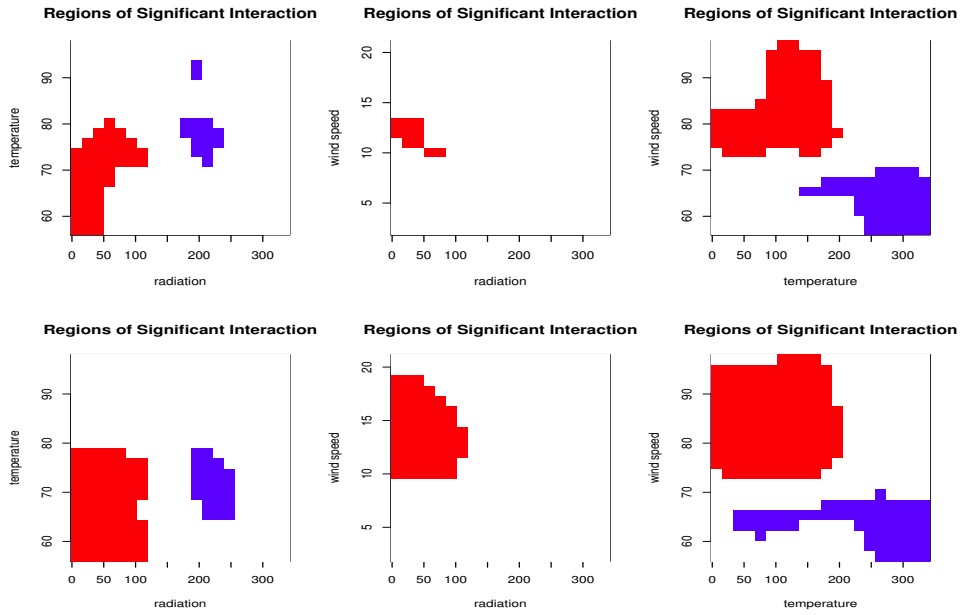

Figure 3.5: Ozone data. Regions of significant positive interaction (red) and negative interaction (blue) are compared under the model for ordinal data (top row) and a model applied directly to the latent response-covariate data (bottom row).

them with additional terms. To study how the covariates interact in terms of their effect on ozone concentration, we turn to a standard decomposition from sensitivity analysis which decomposes a deterministic function into main effects and interactions. The deterministic function we work with is $\mathrm{E}(Z \mid \boldsymbol{X})$, which, for two covariates $\boldsymbol{X} = (X_1, X_2)$, is equal to $\mathrm{E}(\mathrm{E}(Z \mid \boldsymbol{X}))+f_1(X_1)+f_2(X_2)+f_{12}(X_1, X_2)$, with $f_i(X_i) = \mathrm{E}(\mathrm{E}(Z \mid \boldsymbol{X}) \mid X_i)-\mathrm{E}(\mathrm{E}(Z \mid \boldsymbol{X}))$ the main effect of X_i, and $f_{12}(X_1, X_2)$ the interaction between X_1 and X_2 (Oakley and O'hagan, 2004). This gives $f_{12}(X_1, X_2) = \mathrm{E}(Z \mid \boldsymbol{X}) + \mathrm{E}(Z) - \mathrm{E}(Z \mid X_1) - \mathrm{E}(Z \mid X_2)$. Each of these terms is easily available from the mixture of multivariate normals representation for (Z, X_1, X_2). For any fixed covariate vector (x_1, x_2), a set of posterior samples is available. Note that all terms in $f_{12}(X_1, X_2)$ are computed using the same joint distribution $f(z, \boldsymbol{x}; G)$ at a particular MCMC sample, and thus the sign of $f_{12}(x_1, x_2)$ is meaningful. The magnitude of the interaction at a particular covariate vector, $f_{12}(x_1^0, x_2^0)$, is also comparable to the magnitude at a different location, $f_{12}(x_1^1, x_2^1)$, within a given MCMC sample.

We obtain posterior samples for the pairwise interactions over the covariate space, and focus on the regions of significant interaction. That is, we indicate the regions of significant positive interaction, for which the 2.5 percentile for f_{12} is above 0, and the regions of significant negative interaction, for which the 97.5 percentile is below 0. We also compare this to the inference obtained from the model applied to data $\{(z_i, \boldsymbol{x}_i)\}$, which should certainly be able to infer the way in which the elements of $\boldsymbol{X}$ interact to affect Z, and in this case there is not potential for complications arising from the changing scale and location of Z across MCMC samples. The inferences under the ordinal regression model are shown in the top row of Figure 3.5, and the inferences from the continuous response model

are shown in the bottom row. Red indicates regions of significant positive interaction, and blue indicates significant negative interaction. The two sets of figures largely agree, with the main difference being that the ordinal model is slightly more conservative in that it assigns significance to smaller regions.

3.3.3 Credit Ratings of US Companies

We now consider an example involving Standard and Poor's (S&P) credit ratings for 921 US firms in 2005. This example is taken from Verbeek (2008), in which an ordered logit model was applied to the data, and was also used by Chib and Greenberg (2010) to illustrate a flexible modeling approach involving cubic splines and DP mixture errors. For each firm, a credit rating on a seven-point ordinal scale is available, along with five characteristics, which provide $X_1, \ldots, X_5$. As consistent with the analysis of Chib and Greenberg (2010), we combined the first two categories as well as the last two categories, to produce an ordinal response with 5 levels, where higher ratings indicate more creditworthiness. The covariates in this application are book leverage X_1 (ratio of debt to assets), earnings before interest and taxes divided by total assets X_2, standardized log sales X_3 (proxy for firm size), retained earnings divided by total assets X_4 (proxy for historical profitability), and working capital divided by total assets X_5 (proxy for short-term liquidity).

The posterior mean estimates for the marginal probability curves, $\Pr(Y = j \mid x_k; G)$, for $j = 1, \ldots, 5$ and $k = 1, \ldots, 5$, are shown in Figure 3.6. Each panel displays the set of regression functions associated with a single covariate. These inferences display some differences from the results obtained by Chib and Greenberg (2010), which could be due to the additivity assumption of the covariate effects in the regression function under

their model. Empirical regression functions computed by calculating proportions of observations assigned to each class over a grid in each covariate give convincing evidence that the regression relationships estimated by our model fit the data quite well.

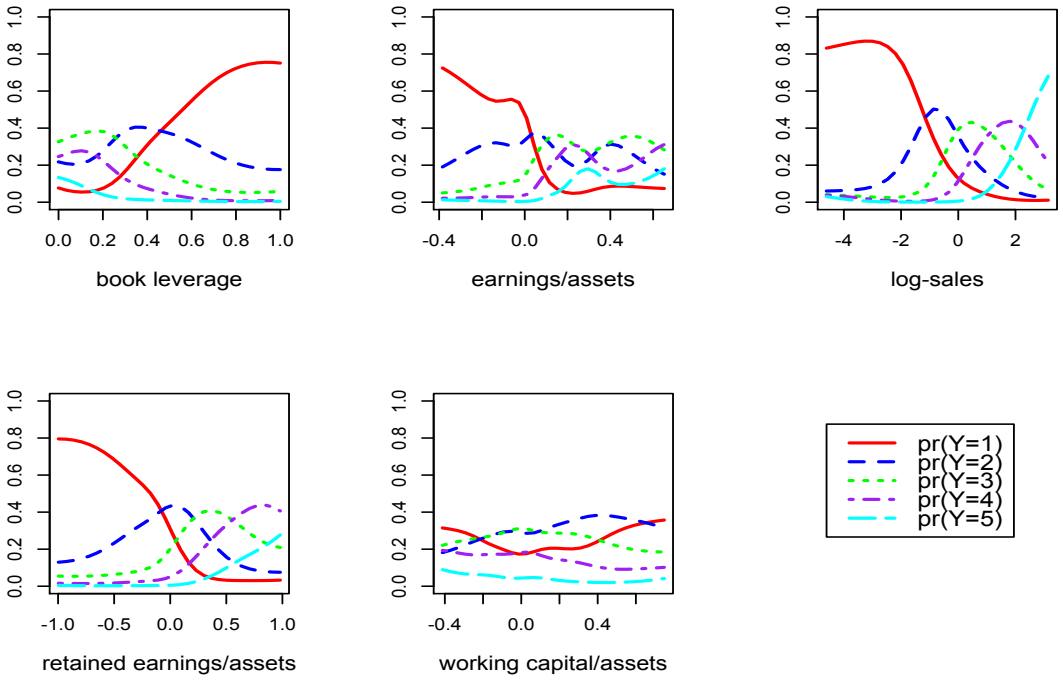

Figure 3.6: Credit rating data. Posterior mean estimates for $\Pr(Y = j \mid x_k; G)$, for each covariate $k = 1, \ldots, 5$. All 5 ordinal response curves corresponding to the 5 ordinal levels are displayed in a single panel corresponding to a common covariate.

The most nonstandard trends appear to be present over X_2, which is earnings before interest and taxes divided by total assets. The covariate X_3, which represents firm size, has some interesting as well as sensible probability trends associated with it. The probability of rating level 1 is somewhat constant for low values of X_3, and then is decreasing to 0, indicating that small firms have a similar probability of receiving the lowest rating, and at some point, the larger the firm, the closer to 0 this probability becomes. The probability

curves for levels $2, 3$, and 4 are all quadratic shaped, with peaks occurring at larger values for higher ratings. Finally, the probability of receiving the highest rating is increasing as a function of X_3. In summary, the size of a firm is positively related to credit rating, as expected.

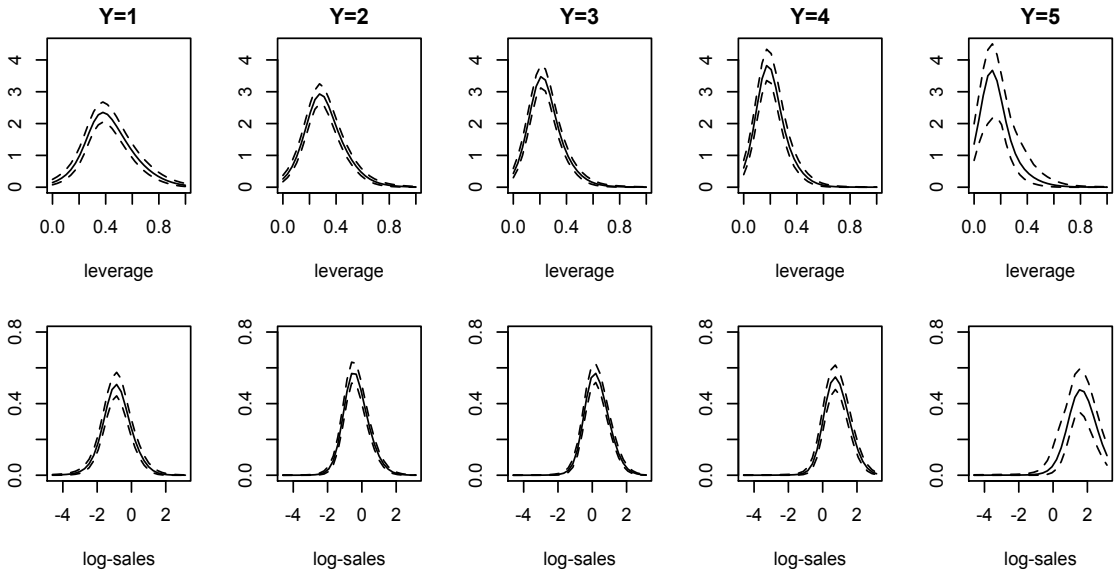

Figure 3.7: Credit rating data. Posterior mean (solid) and 95% interval estimates (dashed) for distributions of covariates book leverage (first row) and standardized log-sales (second row), conditional on ordinal credit rating, arranged by column, i.e. column 1 corresponds to $f(x \mid Y = 1; G)$.

One distinguishing feature of our approach is that we model the joint distribution of the responses and covariates, viewing the covariates as random. This allows our model to accommodate interactions between covariates, which is not done in the cubic spline model, since this assumes the transformed covariates are additive in their effects on the latent response. The assumption of random covariates is appropriate here, as the covariates are characteristics of companies which are not fixed prior to sampling, and their distribution is unknown. In addition to inference for the regression relationships, we may obtain inference

for the covariate distribution, or for any covariate conditional on a specific ordinal rating. These we refer to as inverse relationships, as introduced in Section 3.2.4. It may be of interest to investors and econometricians to know, for example, approximately how large is a company's leverage, given that it has a rating of 2? Is the distribution of leverage much different from that of a company which has a rating of 3? The distributions of the covariates book leverage (X_1) and standardized log-sales (X_3), are shown in Figure 3.7 for each of the 5 ordinal ratings. In the first row, we show $f(x_1 \mid Y = j; G)$, for $j = 1, \ldots, 5$. In general, the distribution of book leverage is centered on decreasing values as rating increases, since higher ratings are associated with lower leverage, and the distribution becomes more and more peaked, having a smaller standard deviation. The interval bands are slightly wider for the distribution associated with $y = 1$ than for $y = 2$, 3, or 4, and they are much wider for $y = 5$, since there are very few observations in this category. The distribution of log-sales (Figure 3.7, second row) has a mode which occurs at increasing values as Y increases, indicating that if one firm has a higher rating than another, it likely also has higher sales.

3.3.4 Standard and Poor Grades of Countries

As a second example from econometrics, we turn our attention to a data set taken from Simonoff (2003), concerned with modeling S&P ratings of $n = 31$ countries as a function of debt service ratio and income, the latter given as an ordinal variable with levels of low, medium, and high. Ratings range from 1 to 7, with 1 indicating the best rating of AAA, and 7 the worst of CCC. This data set concerns a very small sample with a fairly large number of categories, and it will be interesting to see how the model performs in this setting.

Since the covariate income is discrete, we can not simply treat it as part of the continuous covariate vector $\boldsymbol{X}$ in our model. While income is recorded on an ordinal scale, it is truly continuous, and therefore it makes sense to model income W as arising from a latent continuous random variable (this method of modeling ordinal covariates was also used by Ronning and Kukuk, 1996). Therefore, let $\boldsymbol{Z} = (Z_1, Z_2)$, and assume Y arises through Z_1 just as W arises from Z_2. Let the continuous covariate X represent debt service ratio. While rating is viewed as the response, we are once again interested in inverse relationships, such as the distribution of debt service ratio or (discrete) income as a function of a given S&P rating.

Figure 3.8 shows the posterior mean and 95% interval bands for the 7 probability response curves as a function of debt service ratio. We see both monotonic trends (for response categories 1, 2, and 7), as well as nonlinear ones, most notably for categories 4 and 5. The interval bands are wider than in the other examples, given the very small sample size.

Posterior mean and 95% interval bands for the probability of each ordinal rating as a function of income, $\Pr(Y = j \mid W = w; G)$ for $w = 1, 2, 3$ (low, medium, and high) are shown in Figure 3.9. The expression for $\Pr(Y = j \mid W = w; G)$ is obtained from $\Pr(Y = j, W = w; G)/\Pr(W = w; G)$, where the numerator contains a double integral of a bivariate normal density function, thereby requiring evaluation of bivariate normal CDFs. One trend we observe is that the probability of receiving a top rating of 1, 2, or 3 is highest for high-income countries, the probability of receiving a moderate rating of 4 or 5 is highest for medium-income countries, and the probability of receiving a poor rating is highest for

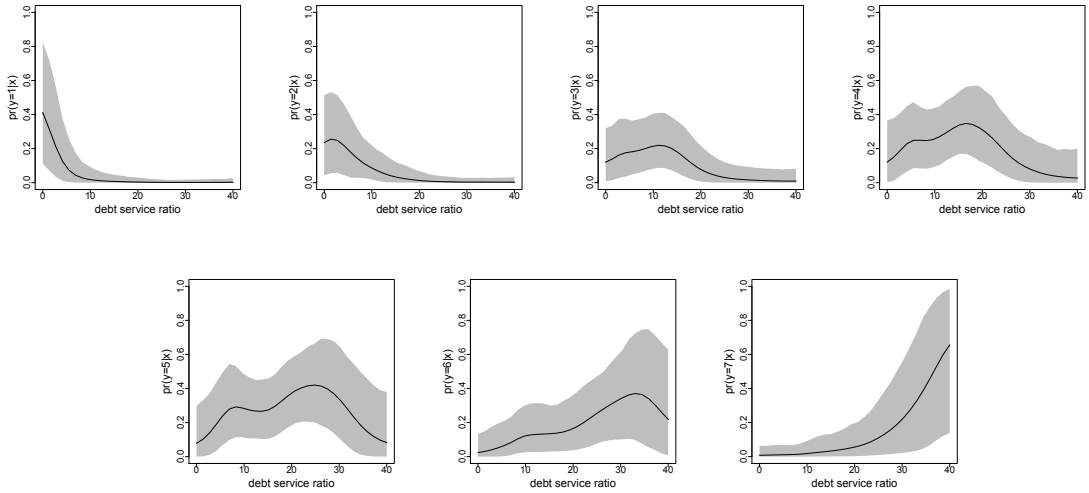

Figure 3.8: S&P ratings of countries data. Posterior mean (solid) and 95% interval estimates (gray shaded regions) for probability curves associated with each rating as a function of debt service ratio.

low-income countries. The uncertainty bands suggest large uncertainty in the probabilities associated with the top 3 ratings for high-income countries, as well as the probabilities associated with the worst 3 ratings for low-income countries. It appears highly unlikely for a country to receive one of the top 2 ratings unless it is high-income, however there does appear to be some positive probability of a middle-income country receiving one of the lowest ratings.

The latent continuous responses represent latent continuous credit rating in this application. The method for posterior simulation involves sampling $z_{i,1}$, for $i = 1, \ldots, 31$, which represent the country-specific latent ratings. Although we may observe the same ordinal rating for multiple countries, the latent continuous ratings may have slightly different distributions. The two countries with AA (level 2) ratings are Canada and Australia. These countries both have income classified as high, however Canada has no debt, and Australia

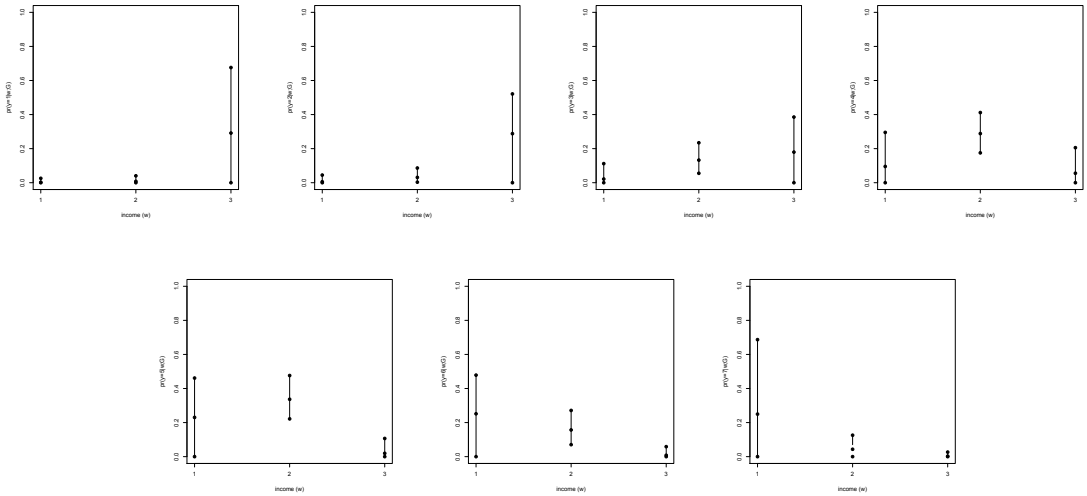

Figure 3.9: S&P ratings of countries data. Posterior mean and 95% interval bands for each category as a function of discrete income levels of low (1), medium (2), and high (3).

has a debt service ratio which is around 10. This value is not particularly high, but since higher debt service ratio seems to be associated with poorer ratings, we would expect that Canada would be closer to receiving a better rating of AAA than Australia. Posterior distributions for latent continuous ratings for Canada and Australia are shown in Figure 3.10 on the left. Australia is represented by the solid line, and Canada by the dashed line. The gray lines represent the cut-off points for the ordinal ratings. It is therefore the case that Canada's ordinal rating of AA is closer to a AAA than is Australia's AA, as expected.

Now consider the four countries with A (level 3) ratings: Chile, Czech Republic, Hungary, and Slovenia. Of these countries, all are classified as middle income except for Slovenia, which has high income. The debt service ratios range from 8.9 (Czech Republic) to 15.8 (Chile), with Slovenia at 9.1, and Hungary at 12.9. The latent response distributions are shown on the right of Figure 3.10. We see that Slovenia's latent rating distribution is centered on values very close to the cut-off for a higher grade of AA. The other three

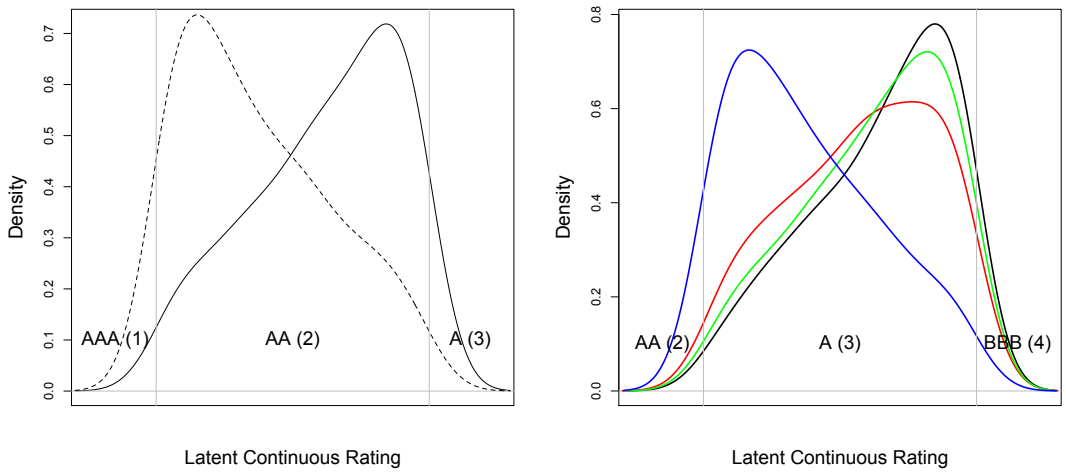

Figure 3.10: S&P ratings of countries data. Left: Posterior distributions for latent continuous ratings for Australia (solid) and Canada (dashed), the two countries with AA rating. Right: Inference corresponding to the four countries with A rating.

distributions are very similar, but there are subtle differences. Chile appears closest to receiving a BBB, which makes sense given its higher debt service ratio. The Czech Republic has a latent continuous distribution centered on slightly lower (better) values than does Hungary, which is expected given its low debt service ratio. One interesting feature to note is that differences in income seem to have a large effect on the distributions, while differences in debt service ratio do not appear to have a great impact.

3.3.5 Analysis of Multirater Agreement Data

A variety of methods exist for analyzing ordinal data collected from multiple raters when the goal is to measure agreemen, ranging from the commonly used κ statistic (Cohen, 1960) and its extensions (Fleiss, 1971), which are indices calculated from observed and expected agreement of raters, to model based approaches involving log-linear models

(Tanner and Young, 1985). We do not attempt to review all of the approaches to modeling multirater agreement data, rather our focus lies in the use of fully model-based approaches for analysis of ordinal data collected from multiple raters along with covariate information. The proposed multivariate ordinal regression model is powerful in this setting, offering flexibility and novelty in terms of the modeling framework and available inferences. We focus on a scenario involving a set of expert graders who evaluate student essays, rating them on an ordinal scale. We contrast our approach to the parametric model of Johnson and Albert (1999), from where this data example is taken, and the nonparametric approach of Savitsky and Dalal (2014), both containing fully model based Bayesian approaches to inference, and similar in spirit to ours, utilizing latent responses.

Multirater agreement data arises when k raters assign ordinal scores to n individuals, so that $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ik})$, for $i = 1, \ldots, n$. The raters typically use the same classification levels, and therefore each $y_{ij} \in \{1, \ldots, C\}$. This data could be summarized in a contingency table, however, we are concerned with problems in which possibly relevant covariate information is also available for each individual. We assume that each judge assigns an ordinal rating to individual i, which represents a discretized version of a continuous rating, so that z_{ij} determines y_{ij}. That is, z_{ij} is the continuous latent score observed by judge j on individual i.

This is in contrast to the formulation of Johnson and Albert (1999), in which the assumption is made that all judges actually agree on the intrinsic worth of each item, so that $z_{ij} = W_i + \epsilon_{ij}$, where W_i represents the true latent score, and ϵ_{ij} is the error observed by judge j. Then, W_i is assumed linearly related to the covariates, being normal with

mean containing the term $\boldsymbol{x}_i^T\beta$. The inflexibility of the latent response distribution is clear, and since the distribution of z_{ij} does not have a judge-specific mean, random cut-offs are necessary to allow the scores to vary among judges.

Savitsky and Dalal (2014) note the inflexibility of the latent response distribution assumed by Johnson and Albert (1999), modeling the latent random vector of a single rater with a DP mixture of independent normals, as the normal kernel distributions have diagonal covariance matrices (a product-kernel model). The dependence is therefore introduced over the latent scores of a single rater, but the data vectors arising from each rater are assumed independent. It is therefore not so clear how to extract inference for inter-rater agreement in this model, which is one of the major objectives in modeling data of this type in the social sciences.

There is no notion in our model of an intrinsic true score for an individual, since we assume each rater has his or her own beliefs which determine the score assigned to a particular individual. An overall score for an individual could be obtained by somehow averaging over the latent scores assigned by each rater, however, if extracting a true underlying score which it is believed all raters agree on is the goal, our method is probably not the best choice. Rather, we focus on modeling relationships between the ordinal scores and covariates, as well as inferring the association between the ordinal variables, over both the covariate space and the ordinal levels. Our method is novel in its use for modeling multi-rater agreement data in many ways, most notably in the nonparametric approach which can accommodate complex dependence among raters, and the assumption of random covariates.

We now apply our method to a problem involving three expert graders who eval-

uate $n = 198$ student essays, each assigned a rating on an ordinal scale of 1 through 10 (these represent raters 2, 3, and 4 from the data given in Chapter 5 of Johnson and Albert, 1999). The covariates average word length and total number of words are used as the $p = 2$ covariates, which may or may not explain to some extent the ratings given by a particular judge. The traditional measure of agreement between raters l and m in the social sciences, $\rho_{lm} = \text{corr}(Z_l, Z_m)$, the polychoric correlation, can be assessed through the posterior predictive distribution for this correlation from the DP mixture model, by sampling one of the $(\rho_{lm,1}, \ldots, \rho_{lm,N})$ with probabilities $(p_1, \ldots, p_N)$ at each MCMC iteration, producing samples for $\rho_{0,lm}$. All three distributions favor most heavily positive correlations (raters 1 and 3 appear to agree most strongly), but place substantial probability on negative correlations, suggesting there is some disagreement present between all pairs of raters. We can determine where raters l and m tend to agree or disagree by finding the latent continuous ratings which are assigned to observations for which $\text{corr}(Z_l, Z_m)$ tends to be of a certain strength and direction. That is, we can look at $E(z_{il} \mid \text{data})$ and $E(z_{im} \mid \text{data})$ arranged by $\text{E}(\text{corr}(z_{i,l}, z_{i,m}) \mid \text{data})$, as $\{\Sigma_l : l = 1, \ldots, N\}$, and L_i imply a particular $\text{corr}(z_{i,l}, z_{i,m})$. This shows, for instance, that raters 1 and 2 strongly agree on very low ratings, and they disagree when rater 2 gives low ratings and rater 1 gives high ratings. It is also the case for the other pairs of raters that they strongly agree mainly at low scores.

There are a variety of regression relationships present in this example. These can be used to assess how ratings tend to vary across covariates, as well as how raters behave in comparison to one another. Defining a high rating as 8 or larger, and a low rating as 3 or lower, we show inference for probability of high and low ratings as functions of average

word length and total number of words for each rater in Figure 3.11. There appears to be a strong trend in rating as a function of number of words for each rater, with rater 2 in particular giving higher ratings for essays with more words. The regression curves for high ratings associated with rater 2 are somewhat separated from raters 1 and 3, and rater 2 clearly assigns more high ratings than other raters.

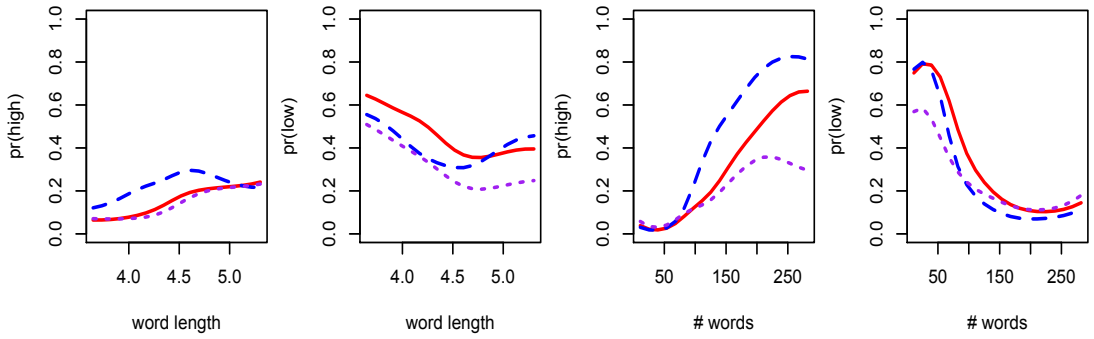

Figure 3.11: Multirater data. Posterior mean estimates for probability of high and low rating as a function of average word length (2 left plots) and number of words (2 right plots), for raters 1 (solid red), 2 (dashed blue), and 3 (dotted purple).

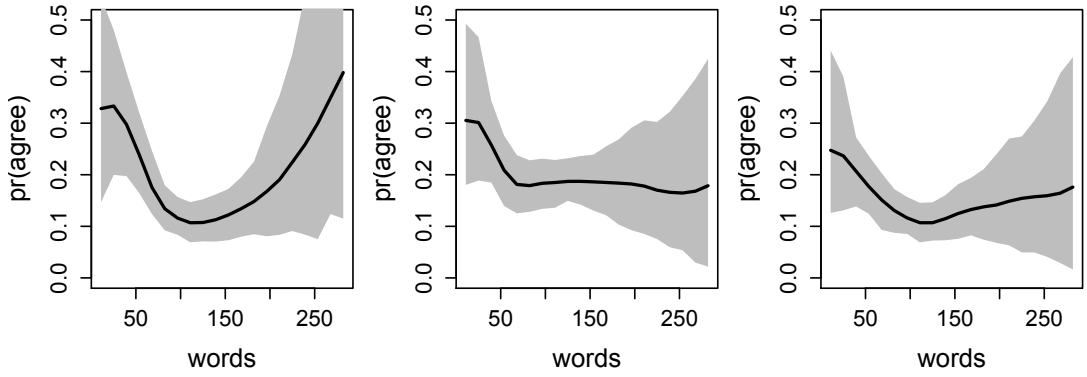

Figure 3.12: Multirater data. Posterior mean (solid) and 95% interval estimates (gray) for probability of agreement for raters 1 and 2 (left), 1 and 3 (middle), and 2 and 3 (right), over covariate number of words.

To determine the regions of the covariate space in which raters tend to agree or disagree, we show probability of perfect agreement for pairs of raters in Figure 3.12, over

	$Y_1 = \text{H}$	$Y_1 = \text{L}$	$Y_2 = \text{H}$	$Y_2 = \text{L}$	$Y_3 = \text{H}$	$Y_3 = \text{L}$
$Y_1 = \text{H}$			0.54	0.15	0.46	0.06
$Y_1 = \text{L}$			0.13	0.48	0.04	0.51
$Y_2 = \text{H}$	0.36	0.18			0.27	0.14
$Y_2 = \text{L}$	0.08	0.56			0.07	0.41
$Y_3 = \text{H}$	0.63	0.11	0.55	0.18		
$Y_3 = \text{L}$	0.04	0.78	0.15	0.54		

Table 3.1: Multirater data. Agreement and disagreement probabilities for pairs of raters, with disagreement highlighted in gray. The row labels indicate the event being conditioned on. H refers to high ratings of $\{8, 9, 10\}$, and L refers to low ratings of $\{1, 2, 3\}$

the covariate number of words. This suggests that raters 1 and 2 agree most strongly on grades for essays with few or many words. The other two pairs of raters tend to agree most for essays with few words, and the trends in agreement probabilities are more constant for these pairs of raters.

Finally, to assess the strength of agreement between raters on high and low scores, we show posterior means for the probability that one rater gives a particular high/low rating, conditional on the rating given by another rater. Posterior means for $\Pr(Y_l \mid Y_m; G)$ for $l, m \in \{1, 2, 3\}$, and Y_l and Y_m taking values of $\{8, 9, 10\}$ (high) or $\{1, 2, 3\}$ (low), are given in Table 3.1. Each row represents an event being conditioned on, while each column is the event a probability is being assigned to. For example, row 1, column 3, gives $\Pr(Y_2 \in \{8, 9, 10\} \mid Y_1 \in \{8, 9, 10\}; G)$. The cells corresponding to disagreement are highlighted with gray. The first two rows give probabilities conditioned on rater 1, and indicate that rater 2 has substantially more disagreement with rater 1 than does rater 3.

The last two rows suggest that rater 2 disagrees more with rater 3 than rater 1 disagrees with rater 3. Finally, comparing the grades given by raters 1 and 3 to those of rater 2 in the middle two rows, we see slightly more disagreement between raters 1 and 2 than raters 1 and 3; however not to a large degree.

3.4 Summary and Remarks

We have presented a fully nonparametric approach to modeling multivariate ordinal data with covariates, which represents a significant contribution to the existing methods for ordinal regression. The power of the framework lies in the flexible model for the latent responses and covariates, which allows the data to drive the way in which the covariates affect the response, while naturally accounting for dependence and interactions.

We showed that our model can accommodate any distribution for mixed ordinal-continuous data, by providing a proof of the KL property of the prior. While our objective was to show the power of the model to approximate any data-generating distribution while at the same time assuming fixed cut-offs, which is provided by the KL property. Flexibility is achieved while assuming fixed cut-offs, without restrictions on the covariance matrix of the normal kernel. This is a very appealing feature of the model, since it avoids the need to estimate cut-offs or work with correlation matrices, which are the most challenging aspects of fitting multivariate probit and related models.

The version of the multivariate probit model we introduced in Section 3.1, and the setting under which we work to build our model, is one in which there exists a single vector of covariates $\boldsymbol{X}$ for each response vector $\boldsymbol{Y}$. That is, the covariates are not specific to the

particular response variable, but rather $(\boldsymbol{Y}, \boldsymbol{X})$ arise as a truly multivariate vector. However, there is another version of the model in which p_j distinct covariates $(X_{j,1}, \ldots, X_{j,p_j})$ exist for each response variable Y_j. This regression setting was described for multivariate continuous responses by Tiao and Zellner (1964), and this is the version of the multivariate binary probit model developed in the work of Chib and Greenberg (1998).

Scenarios which make use of distinct covariates for each response fall broadly into two categories. The first consists of problems in which only a portion of the covariate vector is thought to affect a particular response, but there may be some overlap in the subset of covariates which generate the responses. Chib and Greenberg (1998) considered a voting behavior problem of this kind in which the first of two responses was assumed to be generated by a subset of the covariates associated with the second response. Although treated with a distinct covariate model, this setting could really be accommodated by modeling all covariates $\boldsymbol{X}$ jointly with $\boldsymbol{Y}$, and conditioning on the relevant subset of $\boldsymbol{X}$ in the regression inferences.

The other type of example which is often approached using a distinct covariate regression model really does not fall into the multivariate regression setting. For instance, Tiao and Zellner (1964) mention an example in which each response Y_j corresponds to a particular company j, and therefore $\boldsymbol{X}_j$ is company-specific. Chib and Greenberg (1998) illustrate their model with the famous Six Cities data, in which $\boldsymbol{Y} = (Y_1, \ldots, Y_4)$ represents wheezing status at ages 7 through 10. In these settings, we argue that the responses are actually univariate ordinal, in which hierarchical structure exists, creating dependence in the company-specific or age-specific distributions. This motivates our work on modeling

for dynamic ordinal regression relationships, which builds on the work presented here, so that at any particular time point a unique regression relationship is estimated in a flexible fashion, while dependence is incorporated across time.

In the next chapter, we will extend the model for ordinal regression to handle ordinal regression problems which also contain an aspect of time. That is, the observations are made up of ordinal responses along with covariates, as well as a time index. Many of the properties and results developed in this chapter carry over to the time-dependent setting, since at any given point in time, the ordinal regression model already developed holds.

Chapter 4

Modeling for Dynamic Ordinal Regression Relationships

The motivation for this chapter stems from ordinal regression problems, in which the observations also contain some measure of time. The goal then becomes flexible modeling for dynamically evolving mixed ordinal-continuous distributions. The model for ordinal regression developed in the previous chapter was seen to be powerful for modeling a single multivariate distribution which implies a set of ordinal regressions, and we now require a model for a time series of multivariate distributions. We build on previous work through use of the dependent Dirichlet process (DDP), which we first review in Section 4.1, and develop a new method for incorporating dependence in the weights of the DP in Section 4.2. Two versions of the DDP model for density estimation are presented in Section 4.3. The general DDP model for regression with time-dependent weights and atoms that we choose to work with due to its superior forecasting ability is applied to carefully chosen simulated

data scenarios in Section 4.4. In the following chapter, we consider two real data examples, focusing primarily on a case study involving data on rockfish sampled along the coast of California, which contains temporal structure, since the date of sampling is available in addition to the ordinal response on maturity, along with covariates length and age.

4.1 Dependent Dirichlet Process Priors

Consider data indexed in discrete-time, giving rise to a set of distributions, which are related but not identical. The goal then becomes to model each distribution in a flexible way, while realizing that the set of distributions are correlated, and therefore this dependence must be accounted for in an appropriate fashion. To build on previous knowledge and results, we would like to retain the well-studied DP mixture model marginally at each time $t \in \mathcal{T}$, with $\mathcal{T} = \{1, 2, \dots\}$. We thus must extend the DP prior to model $G_{\mathcal{T}} = \{G_t : t \in \mathcal{T}\}$, a set of dependent distributions such that each G_t follows a DP marginally. The constructive definition of the DP which expresses a realization G from a $\text{DP}(\alpha, G_0)$ as a countable mixture of point masses such that $G = \sum_{l=1}^{\infty} p_l \delta_{\boldsymbol{\theta}_l}$ can be extended to model $G_{\mathcal{T}}$ by introducing dependence in the weights or the atoms.

The general formulation of the DDP introduced by MacEachern (2000) expresses the atoms $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{l,t} : t \in S\}$, $l = 1, 2, \dots$ as i.i.d. sample paths from a stochastic process over S, and the latent beta random variables which drive the weights, $\boldsymbol{v}_l = \{v_{l,t} : t \in S\}$, $l = 1, 2, \dots$, as i.i.d. realizations from a stochastic process with $\text{beta}(1, \alpha_t)$ marginal distributions. The data could be indexed in time, space, or by a covariate, and S represents the corresponding index set, often being $\mathbb{R}$ or $\mathbb{Z}^+$, the latter holding in our case. The

DDP model for data indexed in discrete-time expresses G_t as $\sum_{l=1}^{\infty} p_{l,t}\delta_{\boldsymbol{\theta}_{l,t}}$, for $t \in \mathcal{T}$. The locations $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{l,t} : t \in \mathcal{T}\}$ are i.i.d. for $l = 1, 2, \ldots$, from a time series model for the kernel parameters. The stick-breaking weights $\boldsymbol{p}_l = \{p_{l,t} : t \in \mathcal{T}\}$, $l = 1, 2, \ldots$, arise through a latent time series with beta$(1, \alpha_t)$ marginal distributions, independently of $\{\boldsymbol{\theta}_l\}$.

The general DDP can be simplified by introducing dependence only in the weights, such that the atoms are not time dependent, or alternatively, the atoms can be time dependent while the weights remain independent of time. We refer to these as common weights (or single-p) and common atoms models, respectively. The most natural of the two simplifications is to assume that the locations are constant over time, and introduce dependence in the weights through dependent beta random variables, so that $G_t = \sum_{l=1}^{\infty} p_{l,t}\delta_{\boldsymbol{\theta}_l}$ with $\boldsymbol{\theta}_l \overset{iid}{\sim} G_0$ and $p_{1,t} = v_{1,t}$, $p_{l,t} = v_{l,t}\prod_{r=1}^{l-1}(1 - v_{r,t})$, for $l = 2, 3, \ldots$, with each $\{v_{l,t} : t \in \mathcal{T}\}$ a realization from a time series model with beta$(1, \alpha)$ marginals. Equivalently, we can write $p_{1,t} = 1 - \beta_{1,t}$, $p_{l,t} = (1 - \beta_{l,t})\prod_{r=1}^{l-1}\beta_{r,t}$, for $l = 2, 3, \ldots$, with each $\{\beta_{l,t} : t \in \mathcal{T}\}$ a realization from a time series model with beta$(\alpha, 1)$ marginals, which we will utilize. This simplification is natural for time series data, as it assumes a set of atoms common to each distribution, which are favored to different degrees at each time-point. In addition, introducing dependence in the atoms is not always straightforward, particularly if $\boldsymbol{\theta}_l$ is of large dimension.

There have been many variations of the DDP model proposed in the literature since it was introduced. The common weights version was originally discussed (MacEachern, 2000), in which a Gaussian process was used to generate dependent locations, with the autocorrelation function controlling the degree to which distributions which are "close"

are similar, and how quickly this similarity decays. De Iorio et al. (2004) consider also a common weights model, in which the index of dependence is a covariate, a key application of DDP models. In the order-based DDP of Griffin and Steel (2006), covariates are used to sort the weights. Covariate dependence is incorporated in the weights in the kernel and probit stick-breaking models of Dunson and Park (2008) and Chung and Dunson (2011), respectively, however these are not DDP models, as they do not retain the DP marginally. Gelfand et al. (2005) developed a DP mixture model for spatial data, using a spatial Gaussian process to induce dependence in distributions indexed in space. For data indexed in discrete-time, as in our setting, Rodriguez and ter Horst (2008) apply a common weights model, with atoms arising from a dynamic linear model. Taddy (2010) assumes the alternative simplification of the DDP, having common atoms, and models each independent time series of stick-breaking proportions $\{v_{l,t} : t \in \mathcal{T}\}$ using an autoregressive beta stick-breaking process McKenzie (1985), which generates positively correlated beta random variables. Nieto-Barajas et al. (2012) also use the common atoms simplification of the DDP, modeling a time series of random distributions by linking the beta random variables through latent binomially distributed random variables.

4.2 A Dependent Nonparametric Prior

To generate a correlated series $(\beta_{l,1}, \ldots, \beta_{l,T})$, we define a stochastic process

$$\mathcal{B} = \left\{ \beta_t = \exp\left(-\frac{\zeta^2 + \eta_t^2}{2\alpha} \right) : t \in \mathcal{T} \right\}, \tag{4.1}$$

which is built from a standard normal random variable ζ and a stochastic process $\eta_{\mathcal{T}}$ which has standard normal marginal distributions. This transformation leads to marginal

distributions $\beta_t \sim \text{beta}(\alpha, 1)$ for any t. To see this, take two standard normal random variables Y_1 and Y_2. Assuming $Y_1, Y_2 \overset{ind.}{\sim} \text{N}(0,1)$ implies $U = Y_1^2 + Y_2^2 \sim \chi^2_2$, or $U \sim \exp(0.5)$. Now, if $W = 0.5U$, then $W \sim \exp(1)$. Finally, the transformation $B = \exp(-W/\alpha)$ yields $f(b) = \alpha b^{\alpha-1}$, that is $B \sim \text{beta}(\alpha, 1)$. Therefore, to have $\beta_t \sim \text{beta}(\alpha, 1)$, we need $\zeta \sim \text{N}(0,1)$ and $\eta_t \sim \text{N}(0,1)$.

Because we work with distributions indexed in discrete-time, we assume $\eta_{\mathcal{T}}$ to be a first-order autoregressive (AR) process. The requirement of standard normal marginal distributions on $\eta_{\mathcal{T}}$ leads to a restriction on the variance of the AR(1) model, such that $\eta_{l,t} \sim \text{N}(\phi\eta_{l,t-1}, 1-\phi^2)$, $t = 2, \ldots, T$. Thus $|\phi| < 1$, which implies stationarity for the stochastic process for $\eta_{\mathcal{T}}$. The correlation in $(\beta_{l,t}, \beta_{l,t+k})$ is driven by the autocorrelation present in $\eta_{\mathcal{T}}$, and this induces dependence in the weights $(p_{l,t}, p_{l,t+k})$, which leads to dependent distributions (G_t, G_{t+k}).

While the AR(1) process is the simplest and most natural choice for a discrete-time index, higher-order AR processes may be preferred, though we note that the $\eta_{\mathcal{T}}$ process is just contributing to the dependence in the stick-breaking random variables, which in turn drives the dependence in the weights, and hence the distributions. For other settings involving distributions indexed by a covariate or spatial location, alternative stochastic processes can be chosen appropriately. For instance, a Gaussian process is a natural choice for spatially-referenced data.

4.2.1 Properties of the DDP Prior Model

Autocovariance of the Stochastic Process $\mathcal{B}$

Let $\rho(k) = \text{corr}(\eta_t, \eta_{t+k})$, which is equal to ϕ^k under the assumption for of an AR(1) process for η. The autocovariance function associated with $\mathcal{B}$ is

$$\text{cov}(\beta_t, \beta_{t+k}) = \frac{\alpha^{3/2}(1-\rho^2(k))^{1/2}}{(2+\alpha)^{1/2}\left((1-\rho^2(k)+\alpha)^2-\alpha^2\rho^2(k)\right)^{1/2}} - \frac{\alpha^2}{(\alpha+1)^2}, \tag{4.2}$$

as described in Appendix C.1.

This covariance function can be converted to a correlation function by dividing (4.2) by $\text{var}(\beta_t) = \alpha/((\alpha+1)^2(\alpha+2))$, giving

$$\text{corr}(\beta_t, \beta_{t+k}) = \frac{\alpha^{1/2}(1-\rho^2(k))^{1/2}(\alpha+1)^2(\alpha+2)^{1/2}}{\left((1-\rho^2(k)+\alpha)^2-\alpha^2\rho^2(k)\right)^{1/2}} - \alpha(\alpha+2). \tag{4.3}$$

This autocorrelation function is shown for k ranging from 1 to 50, at various values of α assuming an AR(1) process for $\eta_{\mathcal{T}}$ in Figure 4.1. Smaller values for α lead to smaller correlations for any fixed ϕ at a particular lag, and ϕ controls the strength of correlation, with large ϕ producing large correlations which decay slowly. The parameters ϕ and α in combination can lead to a wide range of correlations, however it is clear from the figures that $\alpha \geq 1$ implies a lower bound near 0.5 on the correlation for any lag k.

We now turn to the limiting behavior of the autocorrelation function. Using (4.3), the $\lim_{\alpha\to 0^+} \text{corr}(\beta_t, \beta_{t+k})$ is easily found to be 0. At the other extreme, consider the limiting behavior as $\alpha \to \infty$, which is tending towards 0.5 as $\rho(k) \to 0^+$, and 1 as $\rho(k) \to 1^-$, taking values between 0.5 and 1 for intermediate values of $\rho(k)$. Assuming $\rho(k) = \phi^k$, gives $\lim_{\phi\to 0^+} \text{corr}(\beta_t, \beta_{t+k}) = \alpha^{1/2}(\alpha+1)(\alpha+2)^{1/2} - \alpha(\alpha+2)$. This tends to 0.5 quickly as $\alpha \to \infty$. At the other extreme value of ϕ, we have $\lim_{\phi\to 1^-} \text{corr}(\beta_t, \beta_{t+k}) = 1$.

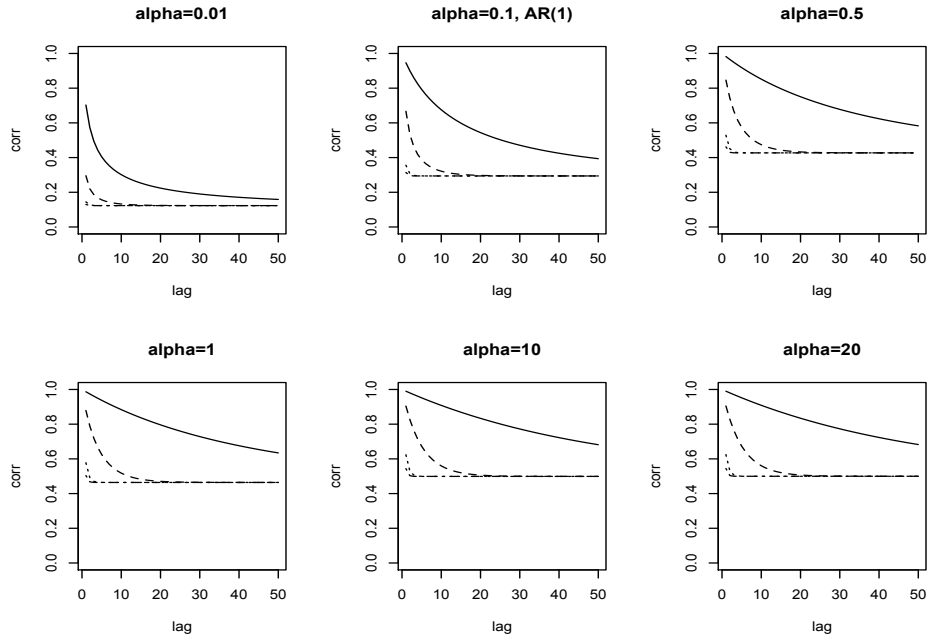

Figure 4.1: Autocorrelation function for $\mathcal{B}$ assuming $\eta_{\mathcal{T}} \sim$AR(1), with values of ϕ of 0.99 (solid), 0.9 (dashed), 0.5 (dotted), and 0.3 (dashed/dotted).

Note that the covariance expressions are functions of $\rho^2(k)$ but not $\rho(k)$, which is to be expected since η_t^2 enters the expression for β_t, and thus $-\rho(k)$ and $\rho(k)$ have the same effect in the correlation. The same is true of the correlation in the DP weights, which is derived below. We believe that $\phi \in (0, 1)$ is natural since we are trying to build a stochastic process for distributions correlated in time through a transformation of an AR process, which we expect should be positively correlated. However, all that is strictly required to preserve the DP marginals is $\mid \phi \mid < 1$.

Autocovariance of the DP Weights

Assume that the random distribution at time t is realized in the form $G_t(\cdot) = \sum_{l=1}^{\infty} p_{l,t}\delta_{\boldsymbol{\theta}_l}(\cdot)$, where $p_{l,t} = (1 - \beta_{l,t}) \prod_{k=1}^{l-1} \beta_{k,t}$, and $\boldsymbol{\beta}_l = (\beta_{l,1}, \ldots, \beta_{l,T})$ is generated by

$\mathcal{B}$, from an underlying AR(1) process for $\eta_{\mathcal{T}}$ with coefficient ϕ. Since each element of $\boldsymbol{\beta}_t = (\beta_{1,t}, \beta_{2,t}, \dots)$ is i.i.d. beta$(\alpha, 1)$, marginally, G_t is distributed as DP(α, G_0). The expression for $\text{cov}(p_{l,t}, p_{l,t+1})$ is

$$\left\{\frac{\alpha^{3/2}(1-\phi^2)^{1/2}}{(2+\alpha)^{1/2}((1-\phi^2+\alpha)^2-\alpha^2\phi^2)^{1/2}}\right\}^{l-1}\left\{1-\frac{2\alpha}{\alpha+1}+\frac{\alpha^{3/2}(1-\phi^2)^{1/2}}{(2+\alpha)^{1/2}((1-\phi^2+\alpha)^2-\alpha^2\phi^2)^{1/2}}\right\}-\frac{\alpha^{2l}}{(1+\alpha)^{2l+2}}, \quad (4.4)$$

as derived in Appendix C.2.

To obtain $\text{corr}(p_{l,t}, p_{l,t+1})$, we need $\text{var}(p_{l,t})$ which is the same for every t, and since $\beta_{l,t}$ is independent of $\beta_{k,t}$, for $k \neq l$, we use the fact that for independent random variables $X_1, \dots, X_n$, $\text{var}(\prod_{i=1}^n X_i) = \prod_{i=1}^n(\text{var}(X_i) + \text{E}^2(X_i)) - \prod_{i=1}^n \text{E}^2(X_i)$. Note that $\text{var}(\beta_{l,t}) = \text{var}(1-\beta_{l,t}) = \alpha/((1+\alpha)^2(\alpha+2))$, $\text{E}(\beta_{l,t}) = \alpha/(\alpha+1)$, and $\text{E}(1-\beta_{l,t}) = 1/(\alpha+1)$, to get $\text{var}(w_{l,t}) = \text{var}((1-\beta_{l,t})\prod_{k=1}^{l-1}\beta_{k,t}) =$

$$\left(\frac{\alpha}{(1+\alpha)^2(2+\alpha)}+\frac{1}{(1+\alpha)^2}\right)\left(\frac{\alpha}{(1+\alpha)^2(2+\alpha)}+\frac{\alpha^2}{(1+\alpha)^2}\right)^{l-1}-\frac{\alpha^{2l}}{(1+\alpha)^{2l+2}}. \quad (4.5)$$

The ratio of (4.4) to (4.5) provides $\text{corr}(p_{l,t}, p_{l,t+1})$, which is plotted for various values of α and ϕ, over weight index l in Figure 4.2. A number of features of this plot can be noted. First, as expected, larger values of ϕ near 1 lead to larger correlations for weights of any lag. Second, the decay in correlations with weight index is faster for small α and small ϕ.

Numerical exploration suggests that as $\alpha \to 0^+$, $\text{corr}(p_{1,t}, p_{1,t+1}) \to 1$ for any value of ϕ. However, the correlation in the second weight, $\text{corr}(p_{2,t}, p_{2,t+1})$, tends towards 0. For very large α, the correlation is constant over weight index l, fixed at some value which is an increasing function of ϕ, always larger than ϕ, and which does not go below 0.5. For large α, as $\phi \to 0^+$, the correlation tends to 0.5, and as $\phi \to 1^-$, the correlation tends to 1.

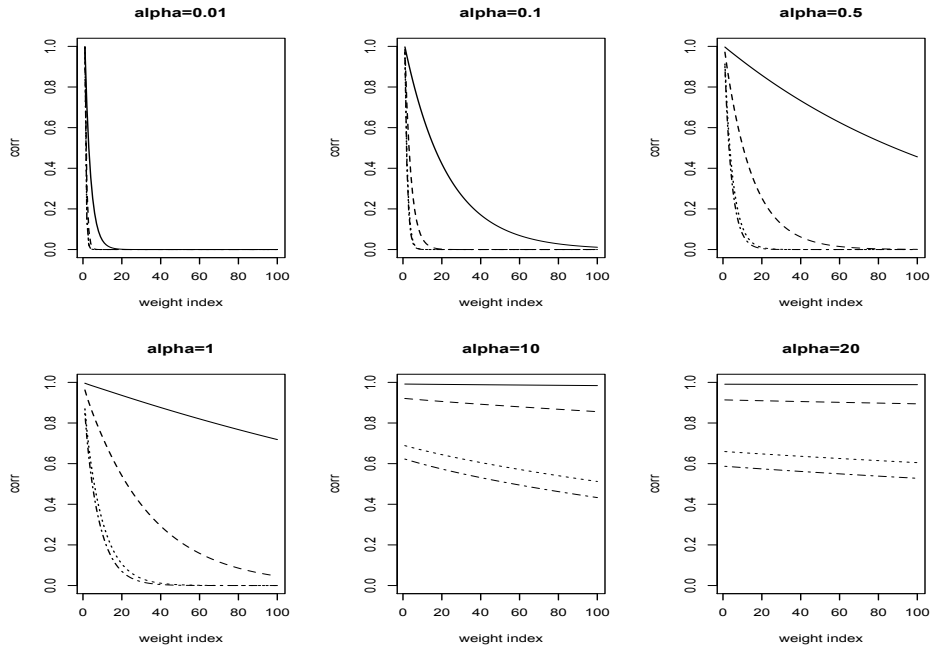

Figure 4.2: Each panel shows $\text{corr}(p_{l,t}, p_{l,t+1})$ for fixed α for four values of ϕ ranging from 0.99 (solid) to 0.3 (dashed/dotted) over weight index l running from 1 to 100.

Therefore, in the limit, as $\alpha \to \infty$, $\text{corr}(p_{1,t}, p_{1,t+1})$ is contained in $(0.5, 1)$.

While we focused only on consecutive weights, note that $\text{corr}(p_{l,t}, p_{l,t+k})$ has the same expression as $\text{corr}(p_{l,t}, p_{l,t+k})$, with ϕ replaced by ϕ^k. Therefore, the limits in the correlations of weights k time lags apart remain the same for any k as ϕ tends to 0 and 1, or as α tends to 0. For large α, $\text{corr}_\phi(p_{1,t}, p_{1,t+k}) = \text{corr}_{\phi^k}(p_{1,t}, p_{1,t+1})$, and therefore the lag k correlations also are between $(0.5, 1)$, but decay with k.

Autocovariance of Consecutive Distributions

We obtain $\text{corr}(G_t(A), G_{t+1}(A))$ for two consecutive distributions G_t and G_{t+1}, and a measurable subset $A \subset \mathbb{R}$ (details given in Appendix C.3). This has the expression:

$$\text{corr}(G_t(A), G_{t+1}(A)) = \frac{(\alpha+1)(\sum_{l=1}^{\infty}\sigma_{l,l}G_0(A) + \sum_{l\neq m}\sigma_{l,m}G_0^2(A) + \sum_{l=1}^{\infty}\frac{\alpha^{2(l-1)}}{(1+\alpha)^{2l}}G_0(A)(1-G_0(A)))}{G_0(A)(1-G_0(A))}$$

$$= \sum_{l=1}^{\infty}\frac{\alpha^{2(l-1)}}{(1+\alpha)^{2l-1}} + \frac{(\alpha+1)(\sum_{l=1}^{\infty}\sigma_{l,l} + \sum_{l\neq m}\sigma_{l,m}G_0(A))}{1-G_0(A)}, \tag{4.6}$$

where $\sigma_{l,l}$ is defined in equation (4.4), and $\sigma_{l,m} = \text{cov}(p_{l,t}, p_{l,t+1})$. Letting $a = \min(l,m)$ and $b = \max(l,m)$, the expression for $\sigma_{l,m}$ is

$$\text{E}\left(\beta_{j,t}\beta_{j,t+1}\right)^{a-1}\left(\frac{\alpha}{1+\alpha}\right)^{b-a-1}\left\{\frac{\alpha}{1+\alpha} - \left(\frac{\alpha}{1+\alpha}\right)^2 - \text{E}\left(\beta_{j,t}\beta_{j,t+1}\right) + \text{E}\left(\beta_{j,t}\beta_{j,t+1}\right)\frac{\alpha}{1+\alpha}\right\} - \frac{\alpha^{l+m}}{(1+\alpha)^{l+m+2}} \tag{4.7}$$

where $\text{E}(\beta_{j,t}\beta_{j,t+1}) = \frac{\alpha^{3/2}(1-\rho^2(1))^{1/2}}{(2+\alpha)^{1/2}((1-\rho^2(1)+\alpha)^2-\alpha^2\rho^2(1))^{1/2}}$, as described in Appendix C.1.

Stationarity

We find that the resulting stochastic process $\mathcal{B}$ which produces the stick-breaking proportions is strongly stationary, using the fact that it is a transformation of a strongly stationary process $\eta_{\mathcal{T}}$. See Appendix C.4 for a discussion of how this can be shown.

Although the process generating the stick-breaking weights is stationary, the random process $\boldsymbol{G}$ generates non-stationary time series. Assume observations Y_t arise from G_t. Marginalizing over $\boldsymbol{G}$ gives $\text{E}(Y_t) = \text{E}(G_0)$ and $\text{var}(Y_t) = \text{var}(G_0)$, the mean and variance of the centering/base distribution G_0. However, given G_t, $\text{E}(Y_t \mid G_t) = \sum_{l=1}^{\infty} p_{l,t}\theta_l$, and $\text{var}(Y_t \mid G_t) = \sum_{l=1}^{\infty} p_{l,t}\theta_l^2 - (\sum_{l=1}^{\infty} p_{l,t}\theta_l)^2$. For two time points t and $t+k$,

$$\text{cov}(Y_t, Y_{t+k} \mid G_t, G_{t+k}) = \sum_{j=1}^{\infty}\sum_{l=1}^{\infty} p_{l,t}p_{j,t+k}l\theta_l\theta_j - \left(\sum_{l=1}^{\infty} p_{l,t}\theta_l\right)\left(\sum_{j=1}^{\infty} p_{j,t+k}\theta_j\right),$$

indicating the random process $\boldsymbol{G}$ generates non-stationary time series with non-constant variance.

4.3 DDP Mixture Modeling for Density Estimation

4.3.1 Common Atoms DDP Hierarchical Model

Assume, at time t, there exist n_t observations, denoted by $\{\boldsymbol{y}_{t,i} : i = 1, \ldots, n_t\}$. To model the time-evolving distributions in a flexible way, while retaining some common behavior from one time to the next, we apply a common atoms DDP mixture model with truncation. This expresses the distribution for data at time t as $f(\boldsymbol{y}; G_t) = \int k(\boldsymbol{y}; \boldsymbol{\theta}) dG_t(\boldsymbol{\theta}) \approx \sum_{l=1}^{N} p_{l,t} k(\boldsymbol{y}; \boldsymbol{\theta}_l)$, with $k(\cdot; \boldsymbol{\theta})$ a kernel distribution having parameters $\boldsymbol{\theta}$.

Inference under this model requires updating the probabilities $\{p_{l,t}\}$, or equivalently the $\{\beta_{l,t}\}$, which are defined through $\{\eta_{l,t}\}$, $\{\zeta_l\}$, and α. A number of updating schemes can be considered. The distribution $f(\beta_2, \ldots, \beta_T)$ can not be built directly using $\prod_{t=2}^{T} f(\beta_t \mid \beta_{t-1})$, as there does not appear to be available a useful expression for $f(\beta_t \mid \beta_{t-1})$. In designing an MCMC algorithm to learn the beta random variables $\{\beta_{l,t}\}$, one choice is to work with $\{\beta_{l,t}\}$ and $\{\eta_{l,t}\}$. In this case, we require $f(\beta_t \mid \eta_t, \alpha)$, which has the form $f(\beta_t \mid \eta_t, \alpha) = \left\{(2^{1/2}\alpha\beta_t^{\alpha-1} \exp(\eta_t^2/2))/(\pi^{1/2}(-2\alpha \log \beta_t - \eta_t^2)^{1/2})\right\} 1_{(0, exp\{-\eta_t^2/(2\alpha)\})}(\beta_t)$. Alternatively, we can work with $\{\zeta_l\}$ and $\{\eta_{l,t}\}$, which, along with α, provide $\{\beta_{l,t}\}$, and hence $\{p_{l,t}\}$. Assuming a normal kernel to model continuous observation vectors $\boldsymbol{y}_{t,i}$, and introducing configuration variables $\boldsymbol{L}$ such that $L_{t,i} = l$ if-f observation $\boldsymbol{y}_{t,i}$ is assigned to component l, for $l = 1, \ldots, N$, the hierarchical model is:

$$\{\boldsymbol{y}_{t,i}\} \mid \{\boldsymbol{\mu}_l\}, \{\Sigma_l\}, \{L_{t,i}\} \sim \prod_{t=1}^{T} \prod_{i=1}^{n_t} \mathrm{N}(\boldsymbol{\mu}_{L_{t,i}}, \Sigma_{L_{t,i}})$$

$$\{L_{t,i}\} \mid \{\beta_{l,t}\} \sim \prod_{t=1}^{T} \prod_{i=1}^{n_t} \{[\sum_{l=1}^{N-1} (1-\beta_{l,t}) \prod_{r=1}^{l-1} \beta_{r,t} \delta_l(L_{t,i})] + \prod_{r=1}^{N-1} \beta_{r,t} \delta_N(L_{t,i})\}$$

$$\zeta_l \overset{iid}{\sim} \mathrm{N}(0,1), \quad l = 1, \ldots, N-1$$

$$\eta_{l,1} \overset{iid}{\sim} \mathrm{N}(0,1), \quad l = 1, \ldots, N-1$$

$$\eta_{l,t} \mid \eta_{l,t-1}, \phi \sim \mathrm{N}(\phi \eta_{l,t-1}, 1-\phi^2), \quad l = 1, \ldots, N-1,\ t = 2, \ldots, T$$

$$(\boldsymbol{\mu}_l, \Sigma_l) \mid \boldsymbol{\psi} \overset{iid}{\sim} \mathrm{N}(\boldsymbol{\mu}; \boldsymbol{m}, V)\mathrm{IW}(\Sigma; \nu, D), \quad l = 1, \ldots, N, \tag{4.8}$$

with priors on α, $\boldsymbol{\psi}$, and ϕ. The parameters $\{\zeta_l\}$ and $\{\eta_{l,t}\}$ can be updated individually with slice samplers, which involves drawing alternatively from uniform random variables and truncated normal random variables. The parameters α and ϕ, given priors $\mathrm{IG}(a_\alpha, b_\alpha)$, and Uniform on $(0,1)$ or $(-1,1)$, respectively, can be sampled using a Metropolis-Hastings algorithm. The configuration variables $\{L_{t,i}\}$ and kernel parameters $\{\boldsymbol{\mu}_l, \Sigma_l\}$ are common to DP mixture models, and follow standard updates. Finally, the parameters $\boldsymbol{\psi} = (\boldsymbol{m}, V, D)$ have closed-form full conditional distributions, given normal, inverse-Wishart, and Wishart priors. The full conditionals and posterior simulation details are described in Appendix B.3.1.

Functionals for Density Estimation and Forecasting

First note that, for any time $t = 1, \ldots, T$, and letting Θ denote all model parameters, the posterior predictive distribution for $\boldsymbol{y}_{0t}$, a new observation at time t, is given by

$$p(\boldsymbol{y}_{0t} \mid \text{data}) = \int p(\boldsymbol{y}_{0t} \mid \Theta) p(\Theta \mid \text{data}) d\Theta$$

$$= \int \int p(\boldsymbol{y}_{0t} \mid \{\boldsymbol{\mu}_l, \Sigma_l\}, L_{t0}) p(L_{t0} \mid \Theta) p(\Theta \mid \text{data}) dL_{t0} d\Theta$$

$$= \int \sum_{l=1}^{N} p_{l,t} \mathrm{N}(\boldsymbol{y}_{0t}; \boldsymbol{\mu}_l, \Sigma_l) p(\Theta \mid \text{data}) d\Theta$$

$$= \mathrm{E}(f(\boldsymbol{y}_{0t}; G_t) \mid \text{data}),$$

which can be evaluated for any $\boldsymbol{y}_{0t}$ using Monte Carlo integration.

Now consider forecasting to the next time not contained in the data. This requires samples for $\boldsymbol{p}_{T+1} = \{p_{l,T+1}\}$, for which we need samples for $\boldsymbol{\eta}_{T+1} = \{\eta_{l,T+1}\}$, in addition to the posterior samples we already have. Then $p(\boldsymbol{y}_{0,T+1} \mid \text{data}) =$

$$\int \int \sum_{l=1}^{N} p_{l,T+1} \mathrm{N}(\boldsymbol{y}_{0,T+1}; \boldsymbol{\mu}_l, \Sigma_l) \prod_{l=1}^{N-1} \mathrm{N}(\eta_{l,T+1}; \phi \eta_{l,T}, 1 - \phi^2) p(\Theta \mid \text{data}) d\boldsymbol{\eta}_{T+1} d\Theta \qquad (4.9)$$

Each MCMC posterior sample for Θ can be used to draw a sample for $\boldsymbol{\eta}_{T+1}$, and then to calculate $\sum_{l=1}^{N} p_{l,T+1} \mathrm{N}(\boldsymbol{y}_{0,T+1}; \boldsymbol{\mu}_l, \Sigma_l)$ for any $\boldsymbol{y}_{0,T+1}$, providing full inference for the forecast at any $\boldsymbol{y}_{0,T+1}$.

Forecasting to later times, such as $T + 2$, requires samples for $\boldsymbol{\eta}_{T+2}$, and hence $\boldsymbol{\eta}_{T+1}$ as well. The expression for $p(\boldsymbol{y}_{0,T+2} \mid \text{data})$ follows from the one-step ahead forecast, and inference can be obtained in a similar fashion.

Accommodating Missing Data

It is likely that in applications, there may be one or more years for which there is no data, and such is the case with the data we will later study. We consider the situation in which data is completely missing at some time or a set of time points. If a small number of years are missing, we essentially have regularly spaced data, but there is a gap which we must deal with since our model is for regularly spaced data.

Let r represent the year(s) for which there is no data. The posterior distribution for model parameters, conditioning on only the observed data, and integrating over the missing data, leaves a joint posterior proportional to

$$\{\prod_{t\neq r}\prod_{i=1}^{n_t}\mathrm{N}(\boldsymbol{y}_{t,i};\boldsymbol{\mu}_{L_{t,i}},\Sigma_{L_{t,i}})\}\{\prod_{t\neq r}\prod_{i=1}^{n_t}\sum_{l=1}^{N}p_{l,t}\delta_l(L_{t,i})\}\{\prod_{l=1}^{N-1}\mathrm{N}(\zeta_l;0,1)\}$$
$$\{\prod_{l=1}^{N-1}\mathrm{N}(\eta_{l,1};0,1)\}\{\prod_{t=2}^{T}\prod_{l=1}^{N-1}\mathrm{N}(\eta_{l,t};\phi\eta_{l,t-1},1-\phi^2)\}\{\prod_{l=1}^{N}\mathrm{N}(\boldsymbol{\mu}_l;\boldsymbol{m},V)\mathrm{IW}(\Sigma_l;\nu,D)\}p(\alpha,\boldsymbol{\psi},\phi). \tag{4.10}$$

There are a few changes in the posterior full conditionals, in particular the full conditionals for $\eta_{l,r}$ are now missing the contributions from the second line of the model for $L_{r,i}$, having only the contributions from the AR model:

$$p(\eta_{l,r}\mid\ldots,\text{data})\propto\mathrm{N}\left(\eta_{l,r};\frac{\phi(\eta_{l,r-1}+\eta_{l,r+1})}{1+\phi^2},\frac{1-\phi^2}{1+\phi^2}\right).$$

The full conditionals for $\{\zeta_l\}$ and α reflect the lack of data at r as well. Since posterior samples are obtained for all $\eta_{l,t}$, even at $t=r$, posterior inference is still available for $f(\boldsymbol{y}_{0,r};G_r)$ at missing time points r.

Data Examples

Various simplifications of the model were tested for purposes of code debugging, and the full model was also applied to simulated data which corresponded to a similar scenario to that assumed in the hierarchical model. In all cases, the posterior density estimates were observed to be capturing the truth well, with more uncertainty associated with smaller sample sizes. As $\alpha\to 0^+$, a single normal distribution holds, being the same at any time t. Prior specification for the DP mixture hyperpriors may therefore follow a

standard default specification strategy in which we use a rough measure of the center and range of the data to scale the mixture component appropriately.

Simulated data from a mixture of normals with time-varying locations At each time point assume a mixture of three normals with constant weights and locations which depend on time. In particular, assume a 3 component mixture with weights $(0.3, 0.2, 0.5)$, locations $\mu_{l,t}$, and variances $(6, 2, 3)$ not dependent on time. The locations $\mu_{l,t}$, $l = 1, 2, 3$, are assumed to follow an AR(1) model with AR coefficients 1, 0.7, and -0.5, and innovation variances equal to 1. This is a mixture autoregressive model, which is a special case of a Markov switching AR model (Früwirth-Schnatter, 2006) with transition matrix having all rows equal to the weight distribution $(0.3, 0.2, 0.5)$, and is closer to a common weights DDP model, the alternative simplification of the DDP, than our model.

A histogram of the data at each time point is shown in Figure 4.3, along with the posterior mean and 95% intervals estimates. The model appears to be doing a very good job of estimating these densities, even though the data was simulated from a model with time-varying locations and constant weights, a scenario which is not contained in our model.

Simulated data from a skew-normal distribution In this example, data was simulated from a skew-normal (SN) distribution with time-varying parameters. The SN distribution has density function $f(y) = 2\phi\{(y-\xi)/\omega\}\Phi\{\alpha(y-\xi)/\omega\}$, with ξ a location parameter, α a shape parameter, and ω a scale parameter. The parameters of the SN distribution at time t are ξ_t, ω_t, and α_t, for $t = 1, \ldots, T = 18$. We assume $\alpha_t = 10t/(T-1)+5-10T/(T-1)$,

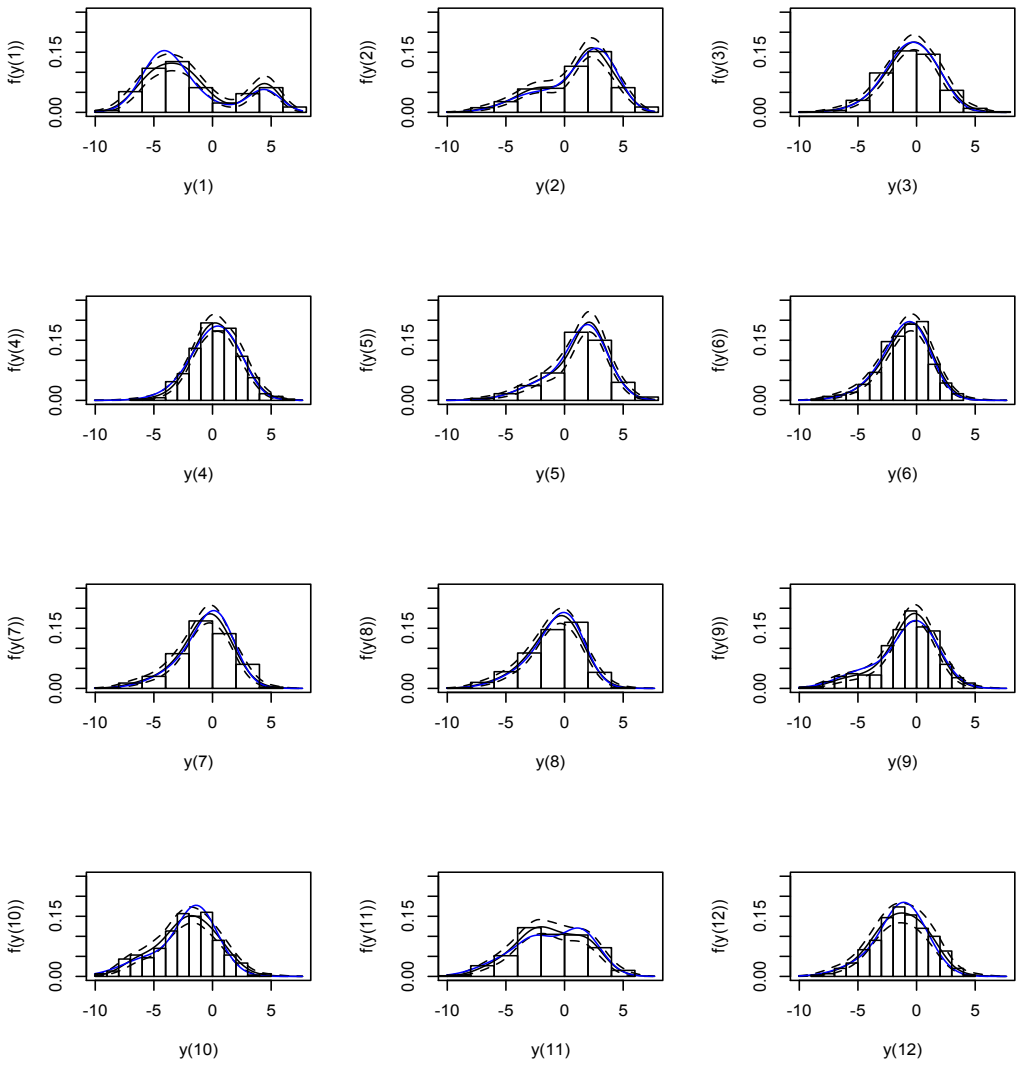

Figure 4.3: Constant weights simulation. Posterior mean and 95% interval estimates for $f(y_{0t}; G)$ (black) compared to a histogram of the data and the true densities (blue).

that is α has a linear trend with skewness ranging from -5 at time $t = 1$ to 5 at $t = T$, meaning the distributions will be left-skewed early on, and right-skewed later on. The scale parameter is given a quadratic trend $\omega_t = 0.2t^2 - 0.2tT + 20$, which has larger values for time points near the beginning and end of the time series, and smaller values for time points near the middle. Finally, the location parameter ξ is given a periodic trend, using the model $\xi_t = 10\cos(\pi t/4) + \epsilon_t$ with $\epsilon_t \sim \mathrm{N}(0, 1)$.

This produces a time series of 18 densities which begin left-skewed with large variance, become symmetric with smaller variance, and finally become right-skewed with increasing variance. The model is first fit to the entire set of data (results not shown because they are indistinguishable from those shown next). Next, the observations at time $t = 13$ are removed, and the model is applied using the missing data techniques previously described. Density estimates for all years are shown in Figure 4.4, including year 13, for which there is no data (the figure still shows the data which was present at this year before becoming missing). Notice that there is more uncertainty present at year 13 than in other years, as there should be, however the point estimate is capturing the location of the missing data fairly well, and may indicate some slight right skewness, which is in fact present in the model which generated the data at this time point, with skewness parameter $\alpha_{13} = 2.06$.

Forecasting Results The forecast distribution can be obtained using (4.9). The results obtained in various settings combined with further exploration suggests that this simplified DDP prior which incorporates dependence only in the weights is not sufficient for forecasting in all settings. Due to the nature of this model, if the shape or location of the modes of the distributions are changing over time, some normal components will be present at a particular

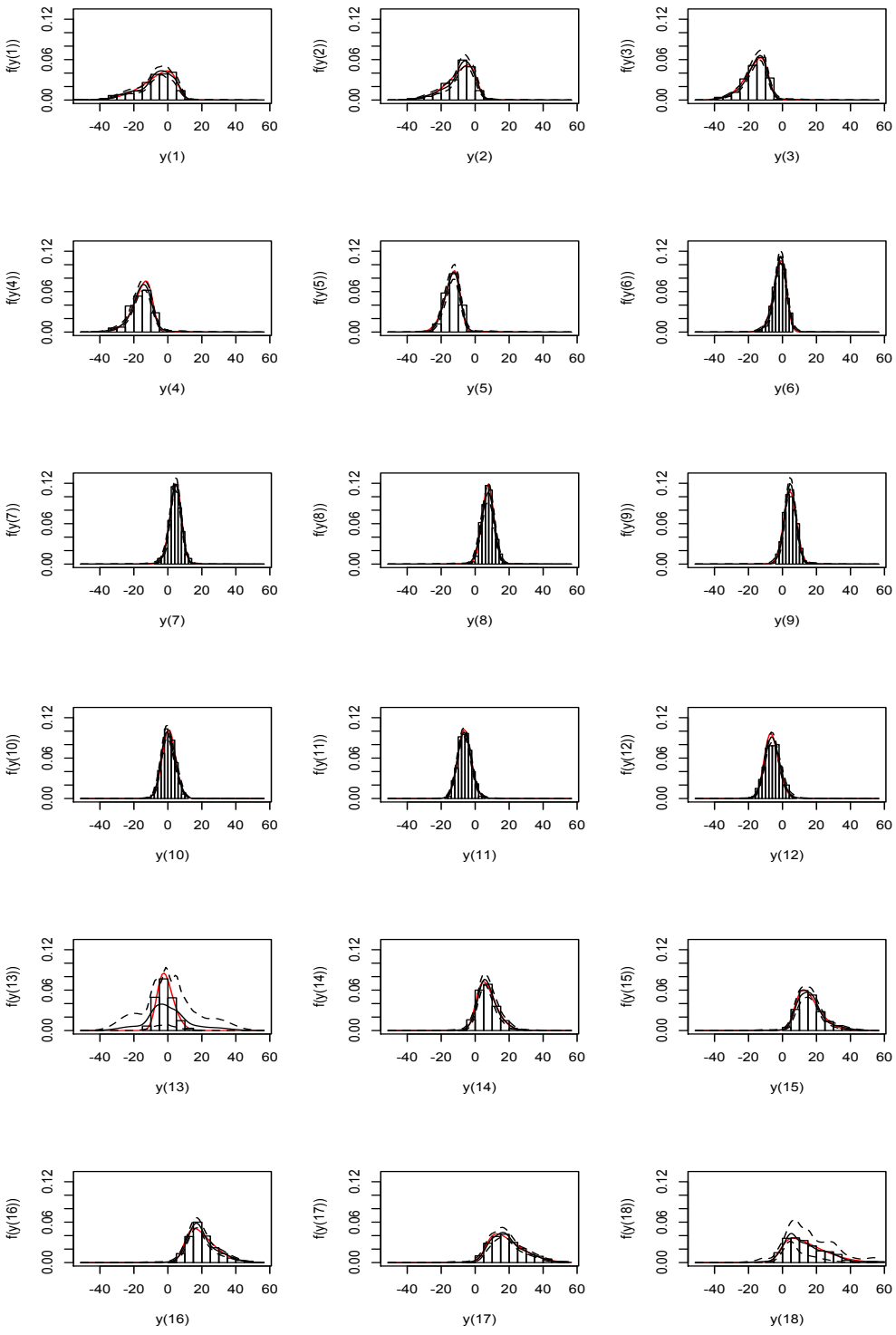

Figure 4.4: SN with missing year simulation. Posterior mean and 95% interval estimates for $f(y_{0t}; G)$ (black) compared to a histogram of the data and data-generating densities (red).

time point that are not needed, hence they receive a weight close to zero. For example, in the SN simulation, the distributions are changing drastically from the first year to the last year in terms of their support. The model is able to accommodate this heterogeneity in shape and location of distributions by using enough components to produce any of these distributions, and at a particular time adjusting the weights according to which components are favored.

However, when taken to the next time point for which there is no data, new weights are generated at each location assuming $\eta_{l,T+1} \sim \mathrm{N}(\phi\eta_{l,T}, 1-\phi^2)$, and $p_{l,T+1}$ is calculated using the ζ_l, which are common to each time, and the time-specific draws for $\eta_{l,T+1}$. The newly generated weights therefore should be similar to those at $p_{l,T}$, however the degree of similarity depends on ϕ (with ϕ values closer to 1 producing more similarity), and the value of $\eta_{l,T}$. What enters into the calculation of the weights is not $\eta_{l,t}$, but $\mid \eta_{l,t} \mid$, and $\mid \eta_{l,t+1} \mid$ is on average larger than $\mid \eta_{l,t} \mid$ when $\mid \eta_{l,t} \mid$ is near zero, and on average smaller than $\mid \eta_{l,t} \mid$ when $\mid \eta_{l,t} \mid$ is large, being exactly equal to $\mid \eta_{l,t} \mid$ in expectation for any value of ϕ when $\mid \eta_{l,t} \mid = \sqrt{2/\pi}$. This leads to very large proportions $1-\beta_{l,T}$ at time T being smaller at $T+1$, and very small proportions $1-\beta_{l,T}$ being larger. In general, this leads to weights at time $T+1$ which are somewhat similar to those at time T, but the very small weights at time T are often larger at $T+1$, and the very large weights at time T are often smaller at $T+1$. Because the components of the mixture are placed all over the support of the entire set of T densities, this behavior may produce forecasted distributions which are not as similar to the last time points as they probably should be.

This behavior is reflected in the forecasts produced for the simulated data. In

the first example involving a mixture of normals with time-varying weights and constant locations, the forecast was similar to that at time T, with the very peaked modes at T being slightly less so at $T+1$, and having wider uncertainty bands. In the SN simulation, the forecasted distribution is fairly symmetric, and places positive probability over the entire region supported by the set of distributions. This behavior is not what would be expected given the distributions at the last few time points, which do not place much weight on negative values, and are right-skewed. However, the combination of the process which generates $\{p_{l,T+1}\}$ and the nature of the common atoms DDP model leads to weights which are larger than zero placed at components centered on negative values, creating a fairly symmetric distribution covering the entire region between approximately -30 and 40. When the model is applied to only the last 5 years of data, a right-skewed distribution similar to that present at T is produced for the forecast. The forecasted distribution from the common weights simulation has a mean very similar to that at the previous time-point, and makes sense given the time series of densities. Forecasting is successful in these two settings because the support of the distributions is not changing to the degree that the SN densities were.

4.3.2 Extension to a More General DDP Model

The common atoms DDP performs well in density estimation, however it does not always forecast densities which are similar to only the most recent time points, often producing predictions which are somewhat of an average of the entire time-series of densities. In light of these results, we seek a variation of the current model with an eye towards forecasting. There are a number of extensions to the model which can be considered.

Staying in the realm of a common atoms DDP, the time series model for the $\{\eta_{l,t}\}$ random variables, which is now assumed an AR(1), could be replaced with any time series model having standard normal marginal distributions. The most natural extension would be to use a higher-order AR model, however, this does not appear to solve the issues present in forecasting when the distributions change support significantly over time. The obvious way to control this, it seems, is to make the atoms of the DP also time-dependent, in addition to the weights, so that $G_t = \sum_{l=1}^{\infty} p_{l,t}\delta_{\boldsymbol{\theta}_{l,t}}$. The simplest extension that should achieve the necessary flexibility in forecasting is to make the means of the atoms time-dependent, and keep the covariance parameters constant in time. A vector autoregressive model of order 1, VAR(1), assumed for $\boldsymbol{\mu}_{l,1}, \ldots, \boldsymbol{\mu}_{l,T}$, has the form $\boldsymbol{\mu}_{l,t} \mid \boldsymbol{\mu}_{l,t-1}, \Theta, \boldsymbol{m}, V \sim \mathrm{N}(\boldsymbol{m} + \Theta\boldsymbol{\mu}_{l,t-1}, V)$, where Θ and V are, in general, full matrices. This gives rise to the stationary distribution having mean $(I-\Theta)^{-1}\boldsymbol{m}$, and covariance matrix C where C solves the equation $C - \Theta C \Theta^T = V$. The analytic solution to this equation is given by $\mathrm{vec}(C) = (I - \Theta \otimes \Theta)^{-1}\mathrm{vec}(V)$, where the operator $\mathrm{vec}(A)$ stacks the columns of the matrix A. Since the stationary distribution for the VAR(1) is not convenient to work with, we assume a prior $\boldsymbol{\mu}_{l,1} \sim \mathrm{N}(\boldsymbol{m_0}, V_0)$, $l = 1, \ldots, N$. All but the last line of the hierarchical model in (4.8) still holds, but the rest of the model is completed as follows:

$$
\begin{aligned}
\boldsymbol{\mu}_{l,1} \mid \boldsymbol{m_0}, V_0 &\sim \mathrm{N}(\boldsymbol{m_0}, V_0), \quad l = 1, \ldots, N \\
\boldsymbol{\mu}_{l,t} \mid \boldsymbol{\mu}_{l,t-1}, \Theta, \boldsymbol{m}, V &\sim \mathrm{N}(\boldsymbol{m} + \Theta\boldsymbol{\mu}_{l,t-1}, V), \quad l = 1, \ldots, N,\ t = 2, \ldots, T \\
\Sigma_l \mid \nu, D &\overset{iid}{\sim} \mathrm{IW}(\Sigma_l; \nu, D), \quad l = 1, \ldots, N
\end{aligned} \tag{4.11}
$$

with priors on α, ϕ, Θ, and $\boldsymbol{\psi} = (\boldsymbol{m}, V, D)$.

The full conditionals for $\{\eta_{l,t}\}$, $\{\zeta_l\}$, ϕ, α and D are the same in this model. As before, $L_{t,i}$ is drawn from the discrete distribution on $\{1,\ldots,N\}$, but now with probabilities proportional to $p_{l,t}\mathrm{N}(\boldsymbol{y}_{t,i};\boldsymbol{\mu}_{l,t},\Sigma_l)$ for $l = 1,\ldots,N$, i.e., $\boldsymbol{\mu}_{l,t}$ replaces $\boldsymbol{\mu}_l$. The update for Σ_l is also changed to reflect the time-dependence in the atoms, being $\mathrm{IW}(\nu + M_l, D + \sum_{\{(t,i):L_{ti}=l\}}(\boldsymbol{y}_{t,i} - \boldsymbol{\mu}_{l,t})(\boldsymbol{y}_{t,i} - \boldsymbol{\mu}_{l,t})^T)$, where $M_l = \mid \{t,i\} : L_{t,i} = l \mid$. The updates for the remaining parameters in this model are provided in Appendix B.3.2.

The matrix Θ, if left a full matrix, seems difficult to work with. A VAR(1) process is stable if the polynomial $\mid I - \Theta u \mid$ has no roots within or on the complex unit circle. A VAR process is stationary if its mean and covariance functions are time-invariant, and stability is a sufficient condition for stationarity of a VAR. Due to the complex form for the stable region of Θ, common prior choices are noninformative, and do not have support on only the stable region.

The simplifying assumption that Θ is diagonal with elements $\theta_1,\ldots,\theta_d$ implies that each element of $\boldsymbol{\mu}_{l,t}$ has a mean which depends on only the corresponding element of $\boldsymbol{\mu}_{l,t-1}$ and not the other elements, which seems reasonable for most applications. In this case, individual Uniform priors on $(0,1)$ or $(-1,1)$ can be used for each θ_i, and the full conditional for $\theta_1,\ldots,\theta_d$ is $p(\theta_1,\ldots,\theta_d \mid \ldots, \text{data}) \propto \prod_{l=1}^{L}\prod_{t=2}^{T}\mathrm{N}(\boldsymbol{\mu}_{l,t};\boldsymbol{m} + \Theta\boldsymbol{\mu}_{l,t-1}, V)\prod_{i=1}^{d}1_{(a,1)}(\theta_i)$, where a is either 0 or -1. The individual θ_i elements can be updated with a Metropolis-Hastings algorithm, or the parameters $(\theta_1,\ldots,\theta_d)$ can be updated jointly, using a proposal distribution which is a multivariate normal distribution on the logit scale.

A further possible assumption is that V is diagonal, in which case $\boldsymbol{\mu}_{l,1}$ can be started from the stationary distribution, in which case the VAR(1) is just a set of d AR(1)

models: $\mu_{l,1,i} \mid m_i, V_i, \theta_i \sim \mathrm{N}(\frac{m_i}{1-\theta_i}, \frac{V_i}{1-\theta_i^2})$ and $\mu_{l,t,i} \mid \mu_{l,t-1,i}, \theta_i, m_i, V_i \sim \mathrm{N}(m_i + \theta_i \mu_{l,t-1,i}, V_i)$, for $i = 1, \ldots, d$. However, the V matrix being diagonal implies independence in the elements of $\boldsymbol{\mu}_{l,t}$, which is probably more restrictive than is realistic. We therefore advocate for a full covariance matrix V.

For prior specification of the hyperparameters, we use the limiting result that as $\alpha \to 0^+$ and $\boldsymbol{\theta} \to \mathbf{0}$, the model for $\boldsymbol{Y}_t$, the random variable at time t, is $\mathrm{N}(\boldsymbol{\mu}_t, \Sigma)$, with $\boldsymbol{\mu}_t \sim \mathrm{N}(\boldsymbol{m}, V)$. Therefore, the marginal prior moments $\mathrm{E}(\boldsymbol{Y}_t)$ and $\mathrm{Cov}(\boldsymbol{Y}_t)$ are the same for any $t = 2, \ldots, T$, and these expressions can be set to a global (over all t) mean and covariance estimate based on the range and center of the data, as was done in the previous version of the model. It remains to specify only $\boldsymbol{m}_0$ and V_0, the mean and covariance for the initial distributions $\boldsymbol{\mu}_{l,1}$. We propose a fairly conservative specification, noting that in the limit, $\mathrm{E}(\boldsymbol{Y}_1) = \boldsymbol{m}_0$, and $\mathrm{Cov}(\boldsymbol{Y}_1) = a_D B_D (\nu - d - 1)^{-1} + V_0$. Therefore, $\boldsymbol{m}_0$ can be set to the mean or midrange of the data at $t = 1$, and V_0 can be set to $\mathrm{diag}((r_1^1/4)^2, \ldots, (r_d^1/4)^2) - a_D B_D (\nu - d - 1)^{-1}$, where $\boldsymbol{r}^1$ indicates the range vector at $t = 1$.

Implementation in the Skew-Normal Setting

For univariate data $y_{t,i}$, the stationary distribution for the AR(1) process is easily available. We use this and assume $\mu_{l,1} \mid m, v, \theta \sim \mathrm{N}(m/(1-\theta), v/(1-\theta^2))$, where variance v replaces the covariance matrix V. In the example which follows, the choice of Uniform priors on $(-1, 1)$ or $(0, 1)$ for θ and ϕ did not affect the results, as these parameters each had posterior distributions which were concentrated far away from 0, centered on values of 0.86 and 0.78.

This model produces a series of density estimates which look almost identical to those produced from the simpler common atoms model. Forecasting involves generating new means at time $T+1$ in addition to new weights, which can be accomplished by drawing $\mu_{l,T+1} \sim \mathrm{N}(m + \theta\mu_{l,T}, v)$ using the posterior samples for $\theta, m, \mu_{l,T}$, and v. The forecasting distribution is displayed in Figure 4.5 along with the densities from the last 5 years of data. Looking closely at the means $\mu_{l,t}$ and weights $w_{l,t}$ produced by this model at T versus $T+1$ suggests a larger degree of similarity than in the model with common atoms (Figure 4.6), which may be surprising since, intuitively, restricting the locations to be the same at each time should lead to more similarities across time. Figure 4.6 plots the posterior means for $\{\mu_l\}$ on the x-axis, and the corresponding posterior means for $\{p_{l,T}\}$ (black) and forecasted weights $\{p_{l,T+1}\}$ (blue) on the y-axis. It can be seen from this plot that for the common weights model (left), the small weights at time T associated with $\mu_{l,T} < 0$ are estimated to be quite a bit larger at $T+1$, and some of the larger weights take on smaller values at $T+1$, producing a fairly symmetric forecasting distribution with left tail extending much farther into negative values than is suggested by the trend in the data. The right figure gives the equivalent inference from the more general model, in which $\mu_{l,T+1} \neq \mu_{l,T}$. It is clear that more similarity is present for $f(y_{0,T+1}; G_{T+1})$ and $f(y_{0,T}; G_T)$ under the general DDP model. Note that the same number of effective components are present at time T and $T+1$ under this model, and their locations and weights match up fairly well. The main differences are the slightly smaller weights at $T+1$ for the smallest and largest (effective) μ values near 7 and 30, and the slightly larger weights near the center, which produces a fairly symmetric forecasting distribution.

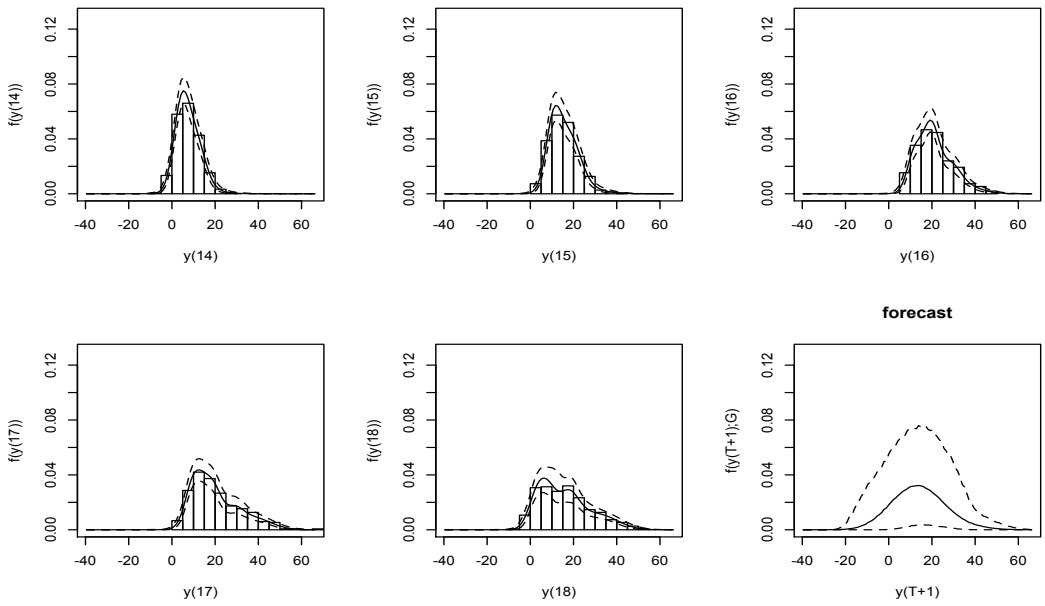

Figure 4.5: SN simulation. Inference from general DDP model. Posterior mean and 95% interval estimates for $f(y_{0,T+1}; G)$, the forecasting distribution, compared with the last 5 years of data.

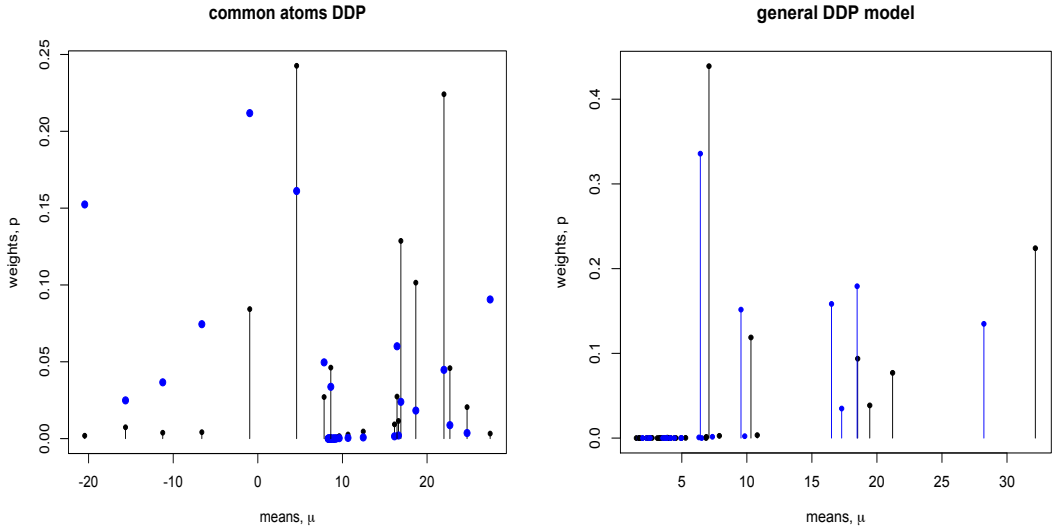

Figure 4.6: SN simulation. Inference from the common location model (left) versus general DDP model (right). The posterior mean weights are plotted on the x-axis and the posterior mean weights $p_{l,T}$ (black) and mean forecasting weights $p_{l,T+1}$ (blue) on the y-axis.

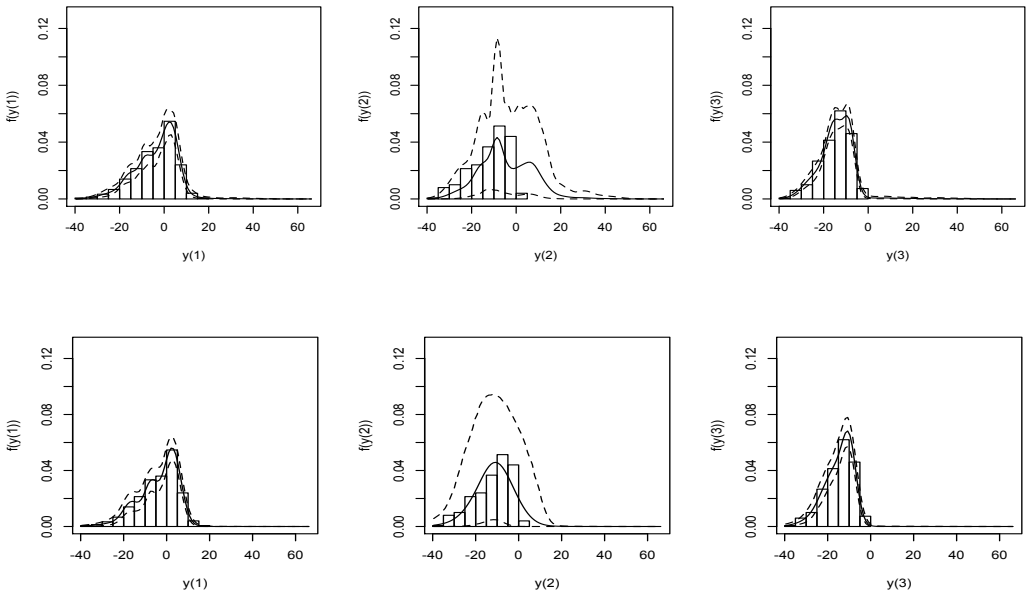

Figure 4.7: SN simulation with data at $t = 2$ missing. Mean and 95% interval estimates from the common atoms model (top) versus general DDP model (bottom) for the densities at $t = 1, 2, 3$, in which the data at $t = 2$ was missing.

Missing data can be taken into account using this model in the same way as discussed for the simpler model. Note that, since the normal locations now also depend on time, the update for $\boldsymbol{\mu}_{l,r}$ is the same update that occurs for $\boldsymbol{\mu}_{l,t}$ when $M_{l,t} = 0$. The model is applied to the SN example with year 13 missing, which produced the inferences under the common atoms model shown in Figure 4.4. The estimates for the missing year distribution, $f(y_{0,13}; G)$, are much smoother under the general model, and the tails of the distribution do not extend out as far as under the common atoms DDP.

The models are also applied with year 2 missing, which is a more interesting scenario since the data is more significantly skewed at $t = 1$ and $t = 3$. The distributions for the first 3 years, including the missing year 2, are shown in Figure 4.7. The first row corresponds to the simpler model. The general model is definitely performing better, with

a point estimate that approximates the distribution of missing data fairly well, exhibiting a small amount of left-skewness.

4.4 Modeling Dynamic Regressions

4.4.1 Continuous Response Variable

We now extend the model for density estimation of univariate random variables $\{Y_t\}$ to a model for regression of continuous random variables $\{Z_t\}$ on covariates $\{\boldsymbol{X}_t\}$. We focus our attention on the more general model with weights and atoms which are indexed by time. The DDP mixture model is applied to estimate $f_t(z, \boldsymbol{x})$, using $f_t(z, \boldsymbol{x}; G_t)$, $t = 1, \ldots, T$, and the output of the MCMC can be used to obtain inference for the conditional distribution $f_t(z \mid \boldsymbol{x}; G_t) = f_t(z, \boldsymbol{x}; G_t)/f_t(\boldsymbol{x}; G_t)$. The model was first tested on data which was produced by simulating from the model, and in all cases was able to identify very well the locations and weights (of the non-negligible components) of the data-generating mixture. The results did suggest some difficulty in learning $\boldsymbol{\theta}$, ϕ, and $\boldsymbol{\psi}$, which were estimated very well in some simulations and not as well in others. In the cases with poorer estimation, the posterior distributions for $\boldsymbol{\theta}$ and ϕ were certainly identifying the correct region of the intervals $(0, 1)$ or $(-1, 1)$, however the estimation was better for the means and weights, even though $\boldsymbol{\theta}$, ϕ, and $\boldsymbol{\psi}$ appear in the time series models for these parameters.

Skew-Normal Conditional Densities

Bivariate data $(z, x)_{t,i}$, with $T = 12$ and $n_t = 200$ for all t was simulated in the following way. First, $x_{t,i} \sim \mathrm{N}(5t, 30^2)$, so that the series of distributions are normal with

increasing means. Then, $z_{t,i}$ was simulated from a SN distribution with time and covariate-dependent locations ξ_t, and time-dependent skewness and scale parameters α_t and ω_t. In particular, $z_{t,i} \mid x_{t,i} \sim \text{SN}(\xi_{t,i}, \omega_t, \alpha_t)$, with $\xi_{t,i} = a_t + b_t x_{t,i}^2$, where a_t is a linear function of t such that $a_1 = -80$ and $a_T = 80$, and b_t begins negative at -0.03, and ends at 0.03, being nondecreasing. The other parameters ω_t and α_t have quadratic and linear trends as in the SN simulation for univariate density estimation.

The posterior mean estimates for the bivariate surfaces $f_t(z, x; G_t)$ are shown with the data overlaid in Figure 4.8. As implied by these figures, the marginals $f_t(x; G_t)$ and $f_t(z; G_t)$ are also well-captured, being symmetric for X_t and moving from left-skewed to right-skewed for Z_t. The expectation of Z_t over the covariate X_t is shown at each time in Figure 4.9, displaying quadratic trends of various forms at early and later time points.

Forecasts for $f_{T+1}(x; G_{T+1})$, $f_{T+1}(z; G_{T+1})$, and $\text{E}_{T+1}(Z \mid X; G_{T+1})$ are displayed in Figure 4.10. Since X_t is normal with mean having a parametric trend and constant variance, the true $f_{T+1}(x)$ is available to compare with the forecast. This is shown in red, and the distribution $f_T(x)$ is also included in blue. The marginal distribution for Z_{T+1} is not available analytically, but samples can be obtained by drawing an x_{T+1} from the normal distribution at time $T + 1$, and sampling $z_{T+1} \mid x_{T+1}$ from a SN distribution with location ξ_{T+1} a function of x_{T+1}, and parameters α_{T+1} and ω_{T+1} calculated from their parametric functions. The estimate $f_{T+1}(z; G_{T+1})$ is right-skewed, and closely approximates $f_T(z; G_T)$, but does not quite capture the true distribution for Z at a future time point in which all parameters continue to evolve parametrically. The estimates for $\text{E}_{T+1}(Z \mid X; G_{T+1})$ are compared with $\text{E}_T(Z \mid X; G_{T+1})$ (blue) and $\text{E}_{T+1}(Z \mid X; G_{T+1})$ (red).

Conditional densities $f_t(z \mid x; G_t)$ are estimated at fixed covariate values of $x = -20$, 0, and 40, and at time points $t = 2$, 4, 8, and 10, and compared with the actual SN densities $f_t(z \mid x)$ in Figure 4.11. The forecast distribution $f_{T+1}(z \mid x; G_{T+1})$ is also obtained at each covariate value.

Mixture of Normals with Time-dependent Weights and Means

Assume data is generated from a mixture of three normal distributions with weights and means that depend on time, varying parametrically. In particular, let $p_{1,t} = 0.4 \mid \sin(t/3) \mid$, $p_{2,t} = 0.6 \mid \sin(t/7) \mid$, for $t = 1, \ldots, T-1$, $p_{1,T} = 0$, $p_{2,T} = 1$, and let $p_{3,t} = 1 - p_{1,t} - p_{2,t}$ for each t. The mean vectors $\boldsymbol{\mu}_{l,t}$ are given parametric trends: $\boldsymbol{\mu}_{1,t} = (15, 28 + t)$, $\boldsymbol{\mu}_{2,t} = (10 + 0.7t, 35 + t)$, and $\boldsymbol{\mu}_{3,t} = (10 + 0.7t, 20 + t)$. Bivariate data vectors $(z_{t,i}, x_{t,i})$ are generated according to the mixture $\sum_{l=1}^{3} p_{l,t} \mathrm{N}(\boldsymbol{\mu}_{l,t}, W_l)$, for $t = 1, \ldots, 9$ and $i = 1, \ldots, 100$, with W_1 uncorrelated, W_2 negatively correlated, and W_3 positively correlated. This leads to a diverse set of conditional expectations $\mathrm{E}_t(Z \mid X)$, containing both linear and nonlinear trends.

The model captures well the densities $f_t(x)$, which range from unimodal to multimodal, shown in Figure 4.12. A range of trends are present in the conditional expectations $\mathrm{E}_t(Z \mid X)$, including linear (both increasing and decreasing) and quadratic trends, with a dipping behavior in the center of the peak of the quadratic function at moderate to later times. The mean and 95% interval estimates from the model are compared with the truth and the data in Figure 4.13. In all cases, the point estimates contain the truth, and the model appears to be doing particularly well in capturing the complex forms present near $t = 7$, both in terms of the point estimates and the interval bands.

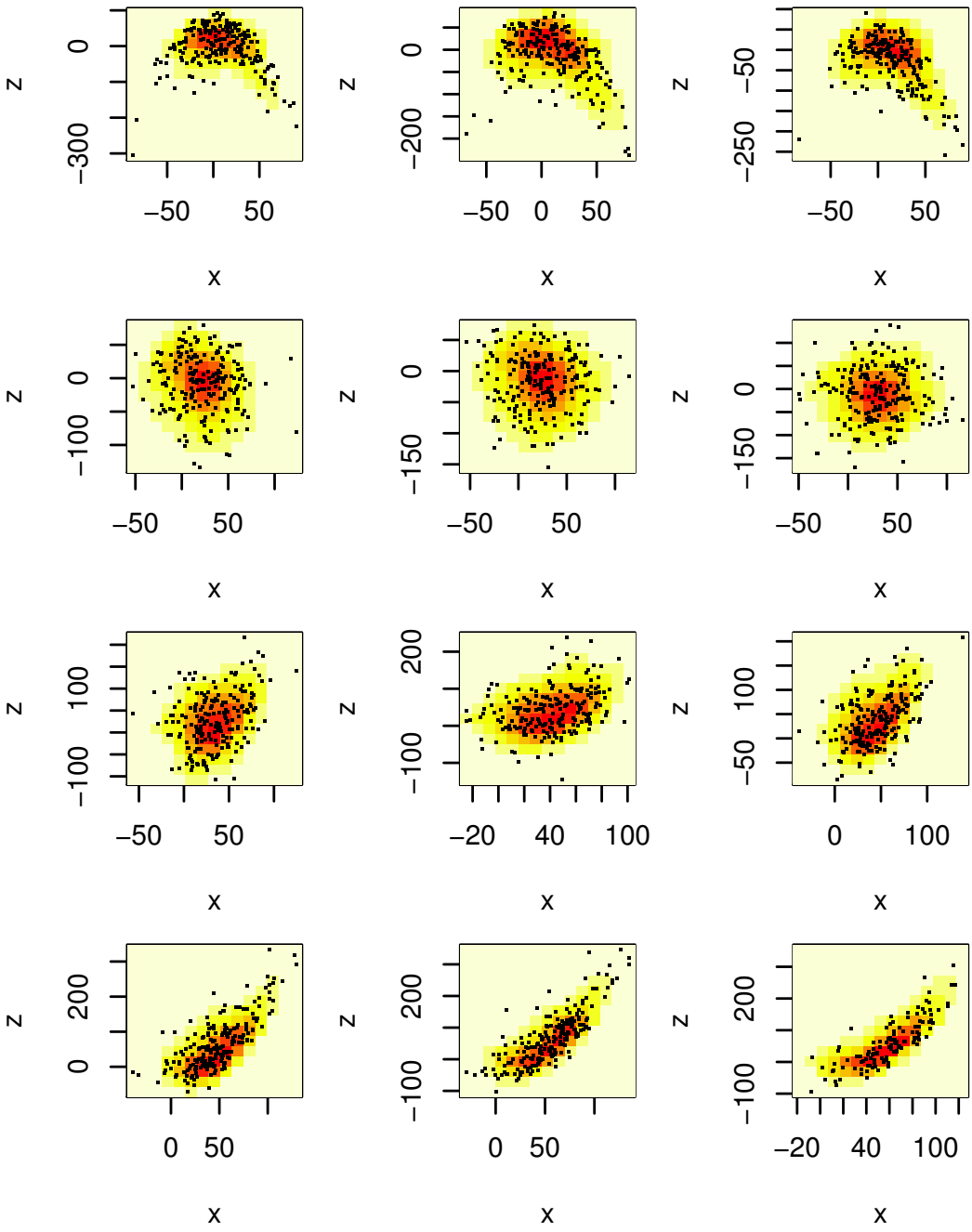

Figure 4.8: SN regression example. Mean estimates for $f_t(z, x; G_t)$, $t = 1, \ldots, 12$, with the data overlaid as small points.

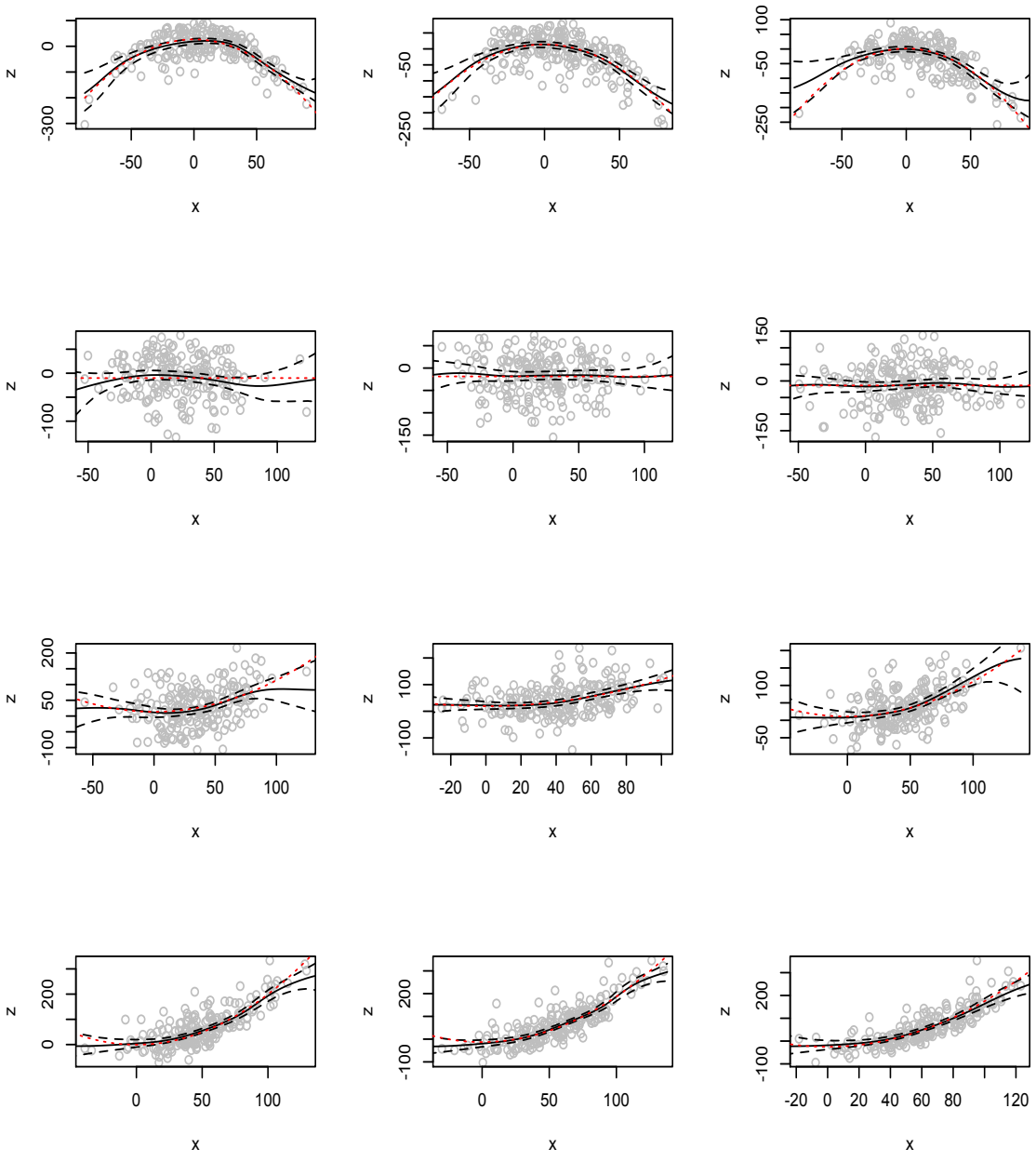

Figure 4.9: SN regression example. Mean and 95% interval estimates for $\mathrm{E}_t(Z \mid X; G_t)$. The truth is shown in red.

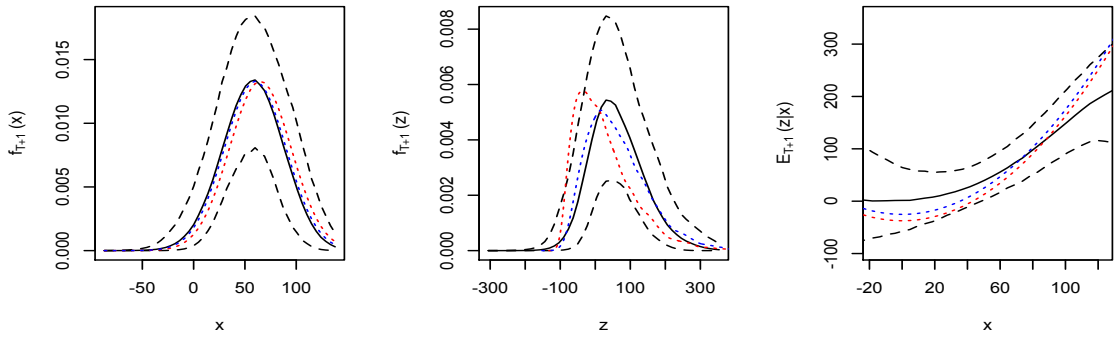

Figure 4.10: SN regression example. Mean and 95% interval estimates for forecasting distributions $f_{T+1}(x; G_{T+1})$ (truth in red, $f_T(x)$ in blue), $f_{T+1}(z; G_{T+1})$ (truth in red, $f_T(z)$ in blue), and $\mathrm{E}_{T+1}(Z \mid X; G_{T+1})$ (truth in red, $\mathrm{E}_T(Z \mid X)$ in blue).

4.4.2 Dynamic Ordinal Regressions

To accommodate ordinal responses, we add an additional level to the hierarchical model in which ordinal responses Y arise from latent continuous responses Z, such that $Y_{t,i} = j$ iff $Z_{t,i} \in (\gamma_{j-1}, \gamma_j]$, for $j = 1, \ldots, C$, with $\gamma_0 = -\infty$ and $\gamma_C = \infty$. Then, the DDP mixture model is applied to estimate $f_t(z, \boldsymbol{x})$. Now, $z_{t,i}$ is unobserved and must be updated in the MCMC with a truncated normal distribution lying on the interval $(\gamma_{y_{t,i}-1}, \gamma_{y_{t,i}}]$, with mean $\mu^z_{L_{t,i},t} + \Sigma^{zx}_{L_{t,i}}(\Sigma^{xx}_{L_{t,i}})^{-1}(\boldsymbol{x}_{t,i} - \mu^x_{L_{t,i},t})$ and variance $\Sigma^{zz}_{L_{t,i}} - \Sigma^{zx}_{L_{t,i}}(\Sigma^{xx}_{L_{t,i}})^{-1}\Sigma^{xz}_{L_{t,i}}$. The ordinal regression functions have the form

$$\Pr_t(Y = j \mid \boldsymbol{x}; G_t) = \sum_{r=1}^{N} \pi_{r,t}(\boldsymbol{x}) \left\{ \Phi\left(\frac{\gamma_j - m_{r,t}(\boldsymbol{x})}{s_r}\right) - \Phi\left(\frac{\gamma_{j-1} - m_{r,t}(\boldsymbol{x})}{s_r}\right) \right\} \tag{4.12}$$

with $\pi_{r,t}(\boldsymbol{x}) \propto p_{r,t}\mathrm{N}(\boldsymbol{x}; \mu^x_{r,t}, \Sigma^{xx}_r)$, $m_{r,t}(\boldsymbol{x}) = \mu^z_{r,t} + \Sigma^{zx}_r(\Sigma^{xx}_r)^{-1}(\boldsymbol{x} - \mu^x_{r,t})$ and $s_r^2 = \Sigma^{zz}_r - \Sigma^{zx}_r(\Sigma^{xx}_r)^{-1}\Sigma^{xz}_r$.

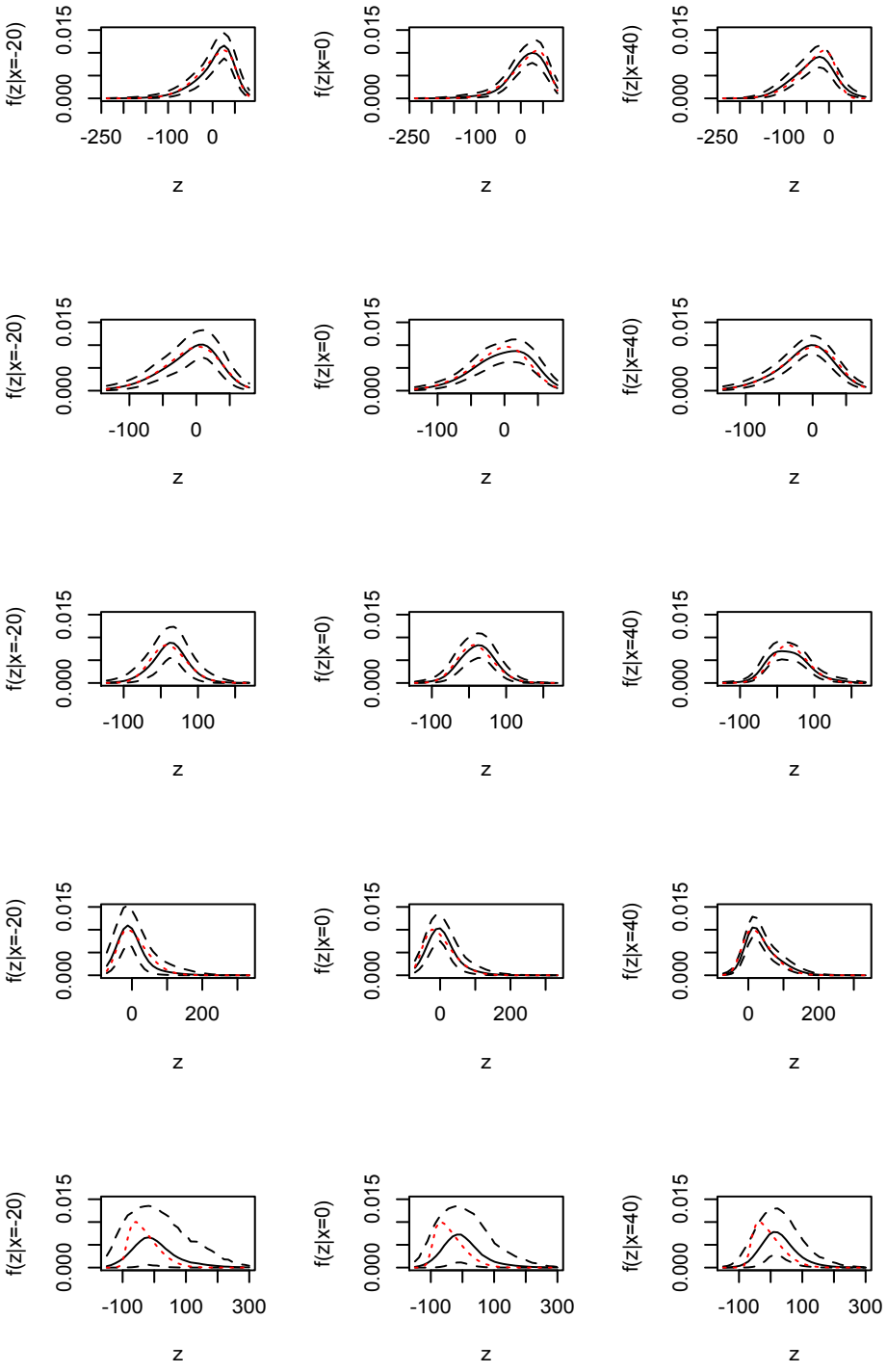

Figure 4.11: SN regression example. Inference for $f_t(z \mid x; G_t)$ for $x = -20$, 0, and 40 (by column) and at $t = 2$, 4, 8, 10, and $T + 1 = 13$ (by row). Mean (solid) and 95% interval estimates (dashed) are compared with the truth (red) in each case.

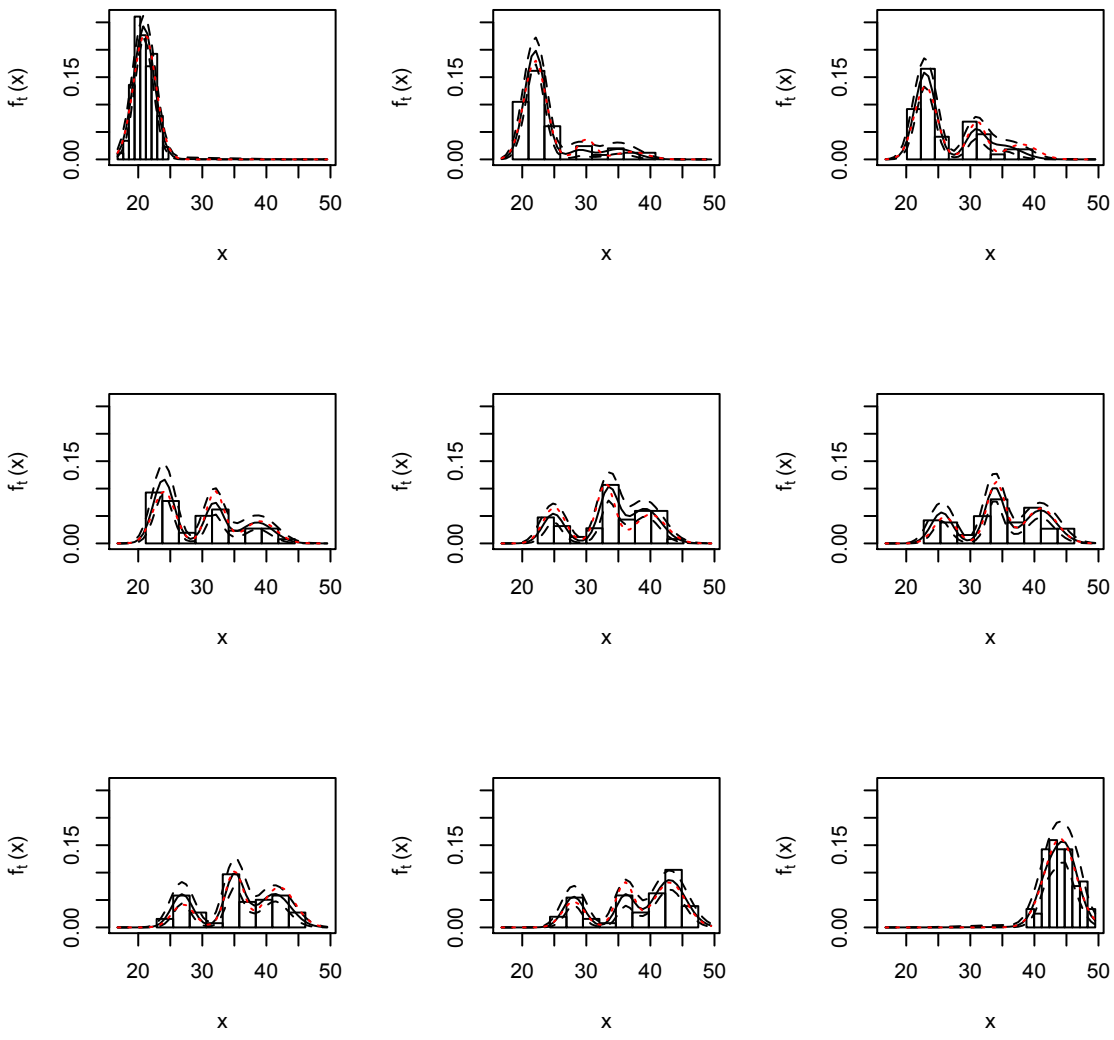

Figure 4.12: Mixture of normals regression example. Mean (solid) and 95% interval estimates (dashed) for the marginal distributions $f_t(x; G_t)$. The data is given as a histogram and the density which generated it shown in red.

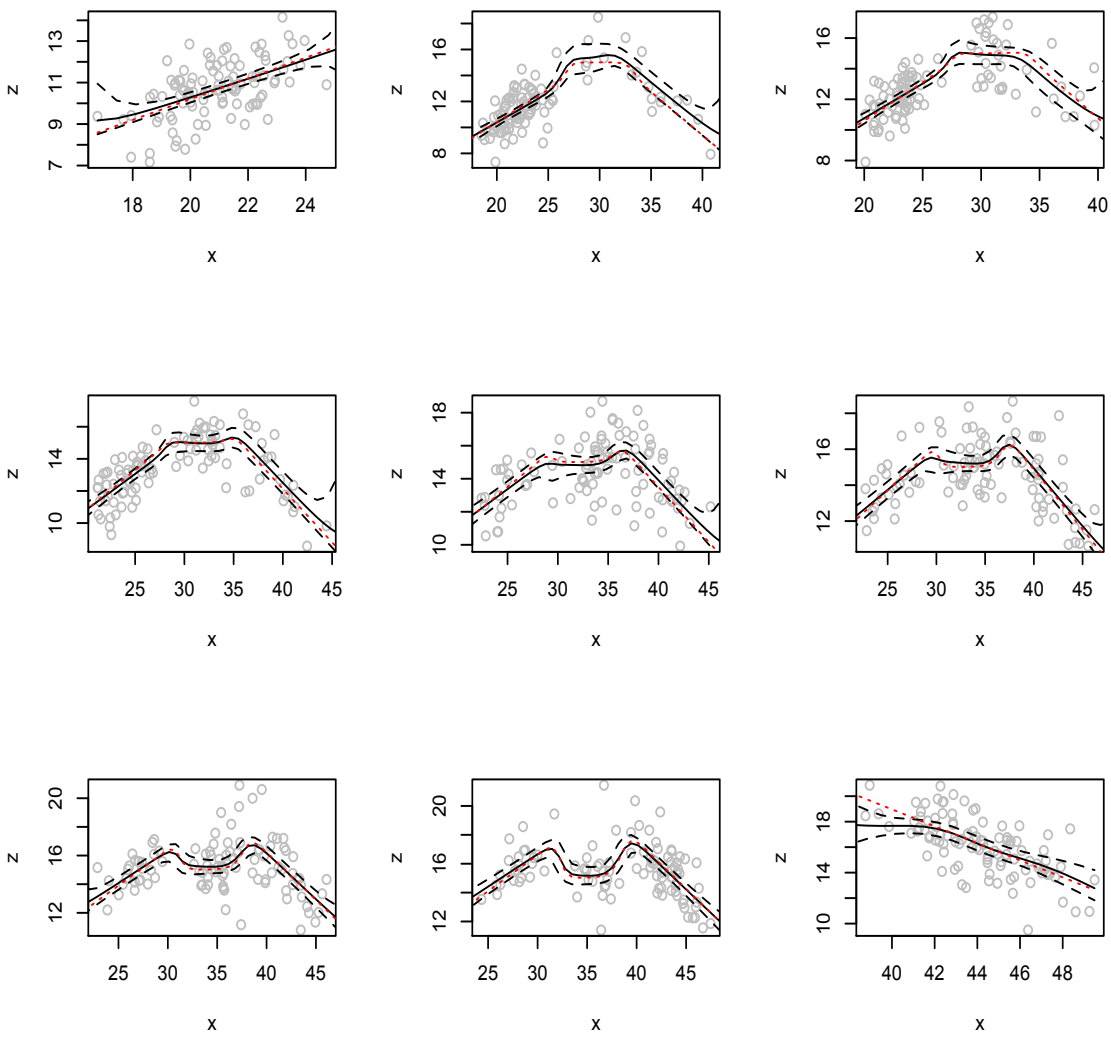

Figure 4.13: Mixture of normals regression example. Mean (solid) and 95% interval estimates (dashed) for the conditional expectations $\mathrm{E}_t(Z \mid X; G_t)$. The data is also shown along with the true functionals in red.

Ordinal Responses from a Conditional Skew-Normal Distribution

Ordinal responses $\{y_{t,i}\}$ with $C = 3$ categories are generated from the SN responses $\{z_{t,i}\}$ (obtained in the first example of Section 4.4.1), using cut-offs of $\gamma_1 = 0$ and $\gamma_2 = 80$. Note that the ordinal regression estimates from the model fit to data (Y, X) can be compared to the equivalent functional from the model which is applied to (Z, X), which can be used as a benchmark, representing the best possible inference that may be obtained under the ordinal model. This expression has the same form, since $\Pr(Y = j \mid x) = \Pr(\gamma_{j-1} < Z \leq \gamma_j \mid x)$. For this simulation, we also know the truth, which is $\Pr_t(\gamma_{j-1} < Z \leq \gamma_j \mid x) = F_{\text{SN}}(\gamma_j; \xi_t, \omega_t, \alpha_t) - F_{SN}(\gamma_{j-1}; \xi_t, \omega_t, \alpha_t)$, where F_{SN} denotes the CDF of the SN distribution.

The true data-generating probability functions are compared with the mean and 95% interval estimates from the model fit to $\{(z, x)_{t,i}\}$ and the model for ordinal regression. Figure 4.14 illustrates these regression curves at $t = 2$, 8, and 12, time points which are representative of the different types of behavior observed in these trends, as well as the forecasts for time 13. The point estimates for $\Pr_t(\gamma_{j-1} < Z \leq \gamma_j \mid x)$ are extremely similar to the estimates for the ordinal regression curves $\Pr_t(Y = j \mid x)$, and any differences that occur are very subtle. The interval bands from the ordinal regression model are slightly wider in some places than those from the model for density estimation. Overall, the similarity in the ordinal regressions and their proxies from the model which observes z is clear, and the DDP regression models seem to be capturing the truth well.

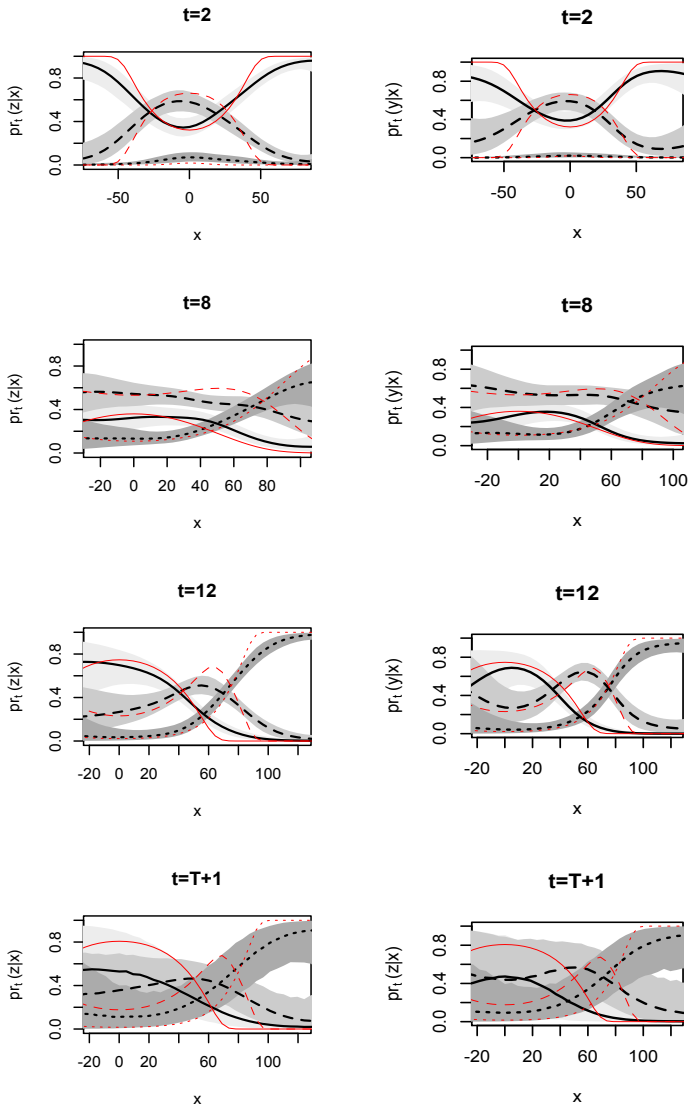

Figure 4.14: Ordinal SN conditional densities regression example. Posterior mean (black) and 95% interval estimates (gray shaded regions) for $\Pr_t(\gamma_{j-1} < Z \leq \gamma_j \mid x)$ (left column) and $\Pr_t(Y = j \mid x)$ (right column), for $j = 1, 2, 3$, compared to the truth (red). Category 1 is indicated by a solid line, category 2 by a dashed line, and category 3 by a dotted line.

Ordinal Responses Generated by Parametric Mixtures of Normals

The second simulation setting in Section 4.4.1 was extended to generate ordinal responses $Y \in \{1, 2, 3\}$ using cut-off points of $\gamma_1 = 12$ and $\gamma_2 = 14$. The DDP models are fit to the data (Y, X) and (Z, X), producing inferences which are shown for $t = 1$, 2, 7, and 9 in Figure 4.15 and compared to the truth. The curves are more linear at $t = 1$, favoring category 1 with decreasing probability as X increases, and at $t = 9$, favoring category 3 with decreasing probability. The intermediate time points are of similar forms, having quadratic shapes, favoring category 1 near the boundaries in X and category 3 for moderate X. The model which observes Z is doing better in some places, such as when $t = 9$, however the model which sees only Y is performing very well in capturing the nonlinear shapes which are present, for example at $t = 7$. The model is even estimating the slight dip in the peak of the quadratic trend for $\Pr_t(Y = 3 \mid x)$ which is present at $t = 7$. The interval estimates produced when the discretized ordinal variable is treated as the response are wider in essentially all places as compared to the corresponding inference from the model which sees the continuous version.

4.5 Concluding Remarks

In this chapter, we set out to develop a DDP prior model for ordinal regressions indexed in time. We began with a simplification of the DDP, in which only the probabilities are time-dependent. This model appeared to be very successful in density estimation, however, further exploration suggested that its ability to forecast was not so strong, as a consequence of the common atoms restriction, which may lead to atoms being present in

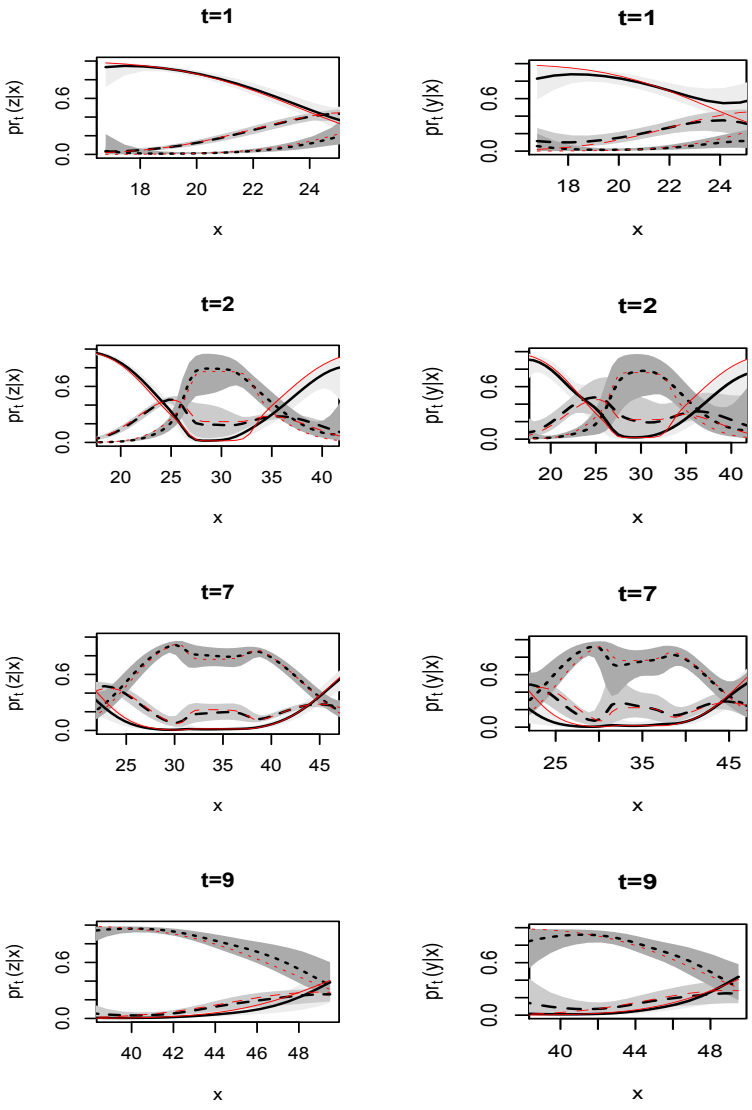

Figure 4.15: Discretized mixture of normals regression example. Posterior mean (black) and 95% interval estimates (gray shaded regions) for $\Pr_t(\gamma_{j-1} < Z \leq \gamma_j \mid x)$ (left column) and $\Pr_t(Y = j \mid x)$ (right column), for $j = 1, 2, 3$, compared to the truth (red). Category 1 is indicated by a solid line, category 2 by a dashed line, and category 3 by a dotted line.

regions where there is no data at a particular time. Because forecasting is an important aspect of modeling for time series, and in particular, we require estimation for densities at years of the time series for which there is no data, we move to a more general model, in which the atoms are also time-dependent. While we did not consider the alternative DDP simplification, one future area of research which could be interesting would involve formal study and comparison of the common weights and common atoms DDP models. Under what settings does the common weights model yield superior inference, and vice versa?

The simulation settings of this chapter were all developed for a purpose, beginning with those used to test the models for density estimation. For example, the simulation involving time-dependent atoms and constant probabilities revealed that the common atoms model was able to capture densities which were generated from a scenario closer to the common weights model. The simulation involving SN densities represented an interesting but also realistic scenario, in which fairly standard densities exist at each time, smoothly evolving from left to right-skewed, and in terms of scale and location. The simulation studies of Section 4.3.1 were then extended to test the model for regression, and finally for ordinal regression.

Having tested the model extensively on simulated data, we are confident in its power to uncover the truth in complex scenarios. Next, we move in Chapter 5 to real data illustrations, focusing first on a case study involving data on rockfish, which contains temporal structure. Chapter 5 also contains an example, involving Citigroup stock data, to indicate other settings in which these methods may be utilized.

Chapter 5

Applications of Temporal Ordinal Regression Methods

5.1 Estimating Maturity of Rockfish

The problem of modeling maturity as a function of age or length is extremely important in fisheries science, as estimates of age at maturity play a large role in population model estimates of sustainable harvest rates (Clark, 1991; Hannah et al., 2009). Virtually all methods for studying maturity as a function of age or length use logistic regression or some variant. Hannah et al. (2009) use logistic regression with a single covariate to study the proportion of fish mature by length and age separately for female Cabezon and yelloweye rockfish. Bobko and Berkeley (2004) also collapsed maturity into just two levels and applied logistic regression with length as a covariate, combining data on female black rockfish collected from November through March of three consecutive years. To obtain age

at 50% maturity, they used their estimate for length at 50% maturity and solved for the corresponding age given by the von Bertalanffy growth curve. Morgan and Hoenig (1997) discuss how estimates of proportion mature at a particular age are often flawed because the length-distribution at a given age is not taken into account, that is it is often assume that maturity is independent of length after conditioning on age. Their method to account for length weights observations of a given age differently depending on length class (an interval of lengths), and they apply a probit regression to model maturity over age.

Our modeling framework can be used to study time-evolving relationships between maturation, length, and age. It is well-suited to this problem, as these three variables constitute a random vector, and although maturity is recorded on an ordinal scale, it is truly continuous. The ability of our approach to treat the covariates and response jointly as random variables, handle multiple ordinal maturation categories, provide flexible inference for a variety of functionals involving maturity and covariate measurements, and incorporate time-dependence distinguishes it from the standard approaches. These features are viewed as strengths and attributes, but also imply that we must be somewhat careful in interpreting results, and realize that although the benefits are numerous, there are some limitations inherent in this approach.

One possible price to be paid for the flexibility afforded by this approach is that we can not incorporate monotonicity restrictions without substantially altering the modeling framework. From a biological point of view, we expect the larger or older the fish, the smaller the probability it is immature, hence monotonically decreasing (increasing) probability curves associated with the level immature (mature). Our approach is in contrast to

more standard techniques such as logistic regression, which force these probability curves to be increasing or decreasing. We must also be careful with extrapolation, as the flexibility of the model may lead to unrealistic behavior outside of the range of the data. In previous simulations we observed that the model could uncover smooth, monotonic trends if they were present, and we will see for this example, that it is retaining reasonable levels of monotonicity, even though there is nothing to force or encourage this behavior. The model is capturing the trends suggested by the data, which often fit with what we expect or believe to be true biologically, but not always.

5.1.1 Description of the Data and Preliminary Results

Before delving into analysis, we discuss the data and the assumptions we make about it. The data on female Chilipepper rockfish that we will analyze consists of date of sampling, age recorded in years, length in millimeters, and maturity recorded on an ordinal scale from 1 to 6, representing immature (1), early and late vitellogenesis $(2, 3)$, eyed larvae (4), and post-spawning $(5, 6)$ (data obtained courtesy of Steve Munch, NOAA, SWFSC, FED). Because we are not necessarily interested in differentiating between every one of these maturity levels, and to make the model output simple and more interpretable, we collapse maturity into 3 ordinal levels, representing immature (1), pre-spawning mature $(2, 3, 4)$, and post-spawning mature $(5, 6)$.

Many observations have age missing or maturity recorded as unknown. Exploratory analysis suggests there to be no systematic pattern in missingness, for example the length distribution using only the complete data looks identical to the length distribution using the data which has missing values for age and/or maturity. Further discussion with fisheries

research scientists at NOAA having expertise in aging of rockfish and data collection (Don Pearson, NOAA, SWFSC, FED) revealed that the reason for missing age in a sample is that otoliths were not collected or have not yet been aged. Maturity may be recorded as unknown because it can be difficult to distinguish between stages, and samplers are told to record unknown unless they are reasonably sure of the stage. Therefore, there is no systematic reason that age or maturity is not present, and it is reasonable to assume that the data are missing at random, or that the probability an observation is missing does not depend on the missing values, allowing us to ignore the missing-data mechanism, and base inferences only on the complete data.

The months of December-February are the interesting ones biologically, since this is the time that Chilipepper rockfish spawn, and the various levels of maturity are all present during these months. Considering complete observations of maturity, length, and age from these winter months, with year as an index of dependence, observations occur in years 1993-2007, with no observations in 2003, 2005, or 2006. The approach to handling data for which entire years are missing (as described in Section 4.3.1) must be applied. If age is treated as a continuous covariate, so that $\boldsymbol{X}$ represents (length, age), we run into problems, as the model places one mixture component centered at each discrete value of age, with a variance component which approaches zero; that is the model places a point mass at each value of age. Computational problems arise very quickly in the MCMC from attempting to invert singular matrices. Clearly, treating age as continuous is not appropriate here, as there are only around 25 distinct values of age in over $2,200$ observations.

Age is in fact an ordinal random variable, such that a recorded age j implies the

fish was between j and $j+1$ years of age. This relationship between discrete recorded age and continuous age is obtained by the following reasoning. Chilipepper rockfish are winter spawning, so they are assumed to be born on January 1. The annuli (rings) of the otiliths (ear stones) are counted in order to determine age, and these also form sometime around January. Thus, for each ring, there has been one year of growth. While it may seem strange at first that there is not a single 0 recorded for age, this is because these "young of the year" are very small and extremely unlikely to be caught by a fishery except for a few unusual species, of which Chilipepper is not one (personal communication with Alec MacCall, NOAA, SWFSC, FED).

We therefore treat age much in the same way as maturity. Let U represent observed ordinal age, let U^* represent underlying continuous age, and assume, for $j = 1, 2, \ldots,$ that $U = j$ iff $U^* \in (j, j+1]$. We can equivalently say that $U = j$ iff $\log(U^*) \in (\log(j), \log(j+1)]$, for $j = 1, 2, \ldots,$ and $U = 0$ iff $\log(U^*) \in (-\infty, 0]$, so that the support of the latent continuous random variable corresponding to age is $\mathbb{R}$. Letting W be the latent continuous random vector which determines U through discretization, we assume $u_{t,i} = j$ iff $w_{t,i} \in (\log(j), \log(j+1)]$, for $j = 0, 1, \ldots,$ so that W is interpretable as log-age on a continuous scale. Assuming Z represents maturity on a continuous scale, and X represents length, the DDP mixture model is then applied to the trivariate continuous vector $\{(z, w, x)_{t,i}\}$, $t = 1, \ldots, 15$, $i = 1, \ldots, n_{15}$.

The distributions for length display a range of unimodal and skewed, as well as nonstandard shapes, and are shown for 6 years in Figure 5.1. The 95% posterior interval estimates reflect the different sample sizes in these years; in 1994 there are 271 observations,

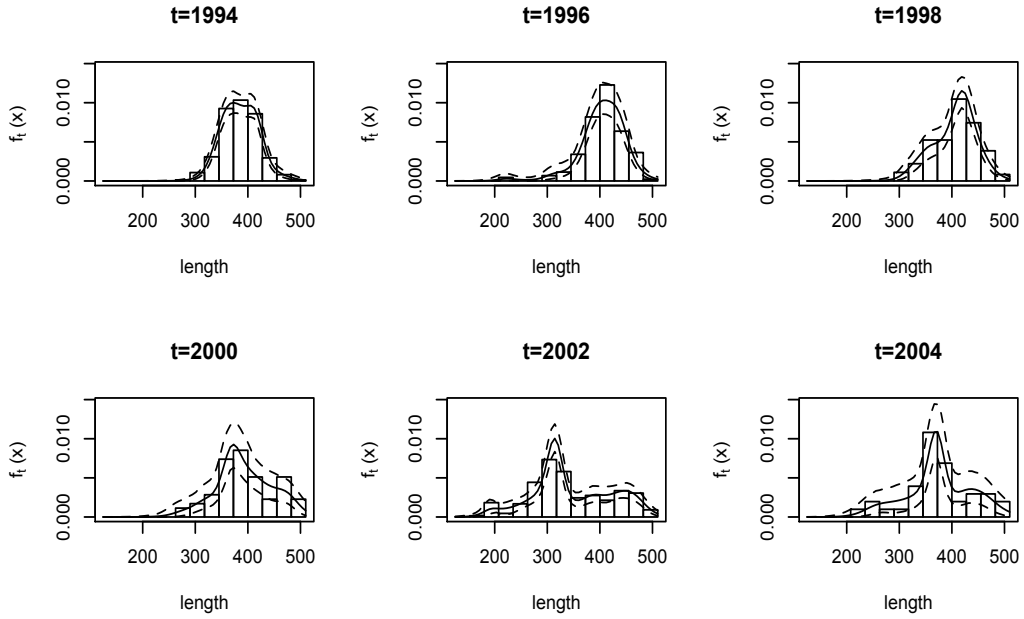

Figure 5.1: Fish maturity example. Posterior mean and 95% interval estimates for the distribution of length in millimeters across 6 years, with the data shown as a histogram.

whereas in 2000 and 2004 there are only 64 and 37 observations, respectively. One attractive feature of our model is that inference is obtained over age on a continuous scale. The distribution of log-age is given by $f_t(w; G_t)$, which at a particular value w_0, can be divided by $\exp(w_0)$ to obtain the corresponding density estimate at age $u_0^* = \exp(w_0)$. These densities are shown in Figure 5.2 for the same years as length. The year 2002 favors the particular age of 3, and 2004 favors age 5.

The posterior mean surface for the bivariate distributions of age and length are shown in Figure 5.3 for all time points. An ellipse with a slight "banana" shape appears at each year, though some nonstandard features and differences across years are present. Picture a line or curve going through the center of these distributions, representing $\mathrm{E}_t(X \mid U^* = u^*; G_t)$, which is obtained at a particular u^* by evaluating $\mathrm{E}_t(X \mid W = w; G_t) =$

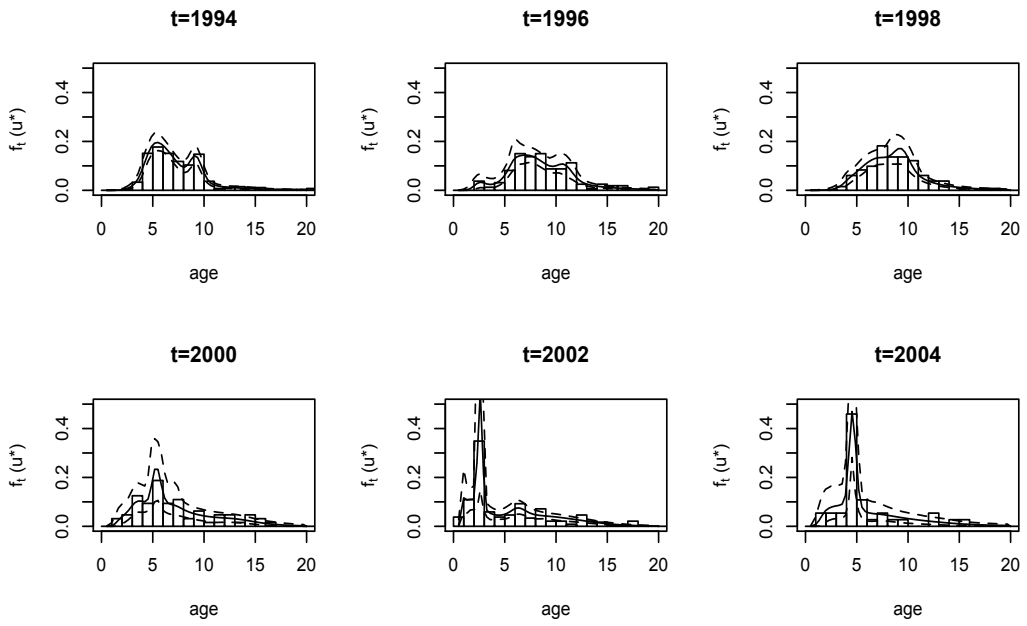

Figure 5.2: Fish maturity example. Posterior mean and 95% interval estimates for the distribution of age on a continuous scale across 6 years, with the data shown as a histogram.

$\sum_{l=1}^{N} \pi_{l,t}(\mu_l^x + \Sigma_l^{xw}(\Sigma_l^{ww})^{-1}(w - \mu_l^w))$ with $\pi_{l,t} \propto p_{l,t}\mathrm{N}(w; \mu_l^w, \Sigma_l^{ww})$ at $w = log(u^*)$, for which we show posterior mean and 95% interval bands for 3 years in Figure 5.4. These are analogous to the von Bertalanffy growth curves used to obtain length-at-age (shown in red). While the von Bertalanffy growth equation is a particular function of age and three parameters (estimated here using nonlinear least squares), our model is simply estimating the joint distribution of length and age, which implies the form of length as a function of age. These curves happen to be very similar to the von Bertalanffy growth curves, with slight differences, for instance in 2002. Note that our approach yields uncertainty quantification in the growth curves, while the usual technique of obtaining point estimates of parameters and plugging in these estimates to obtain a fitted growth curve does not allow for this. This seems an important distinction, as the attainment of unique growth curves by group

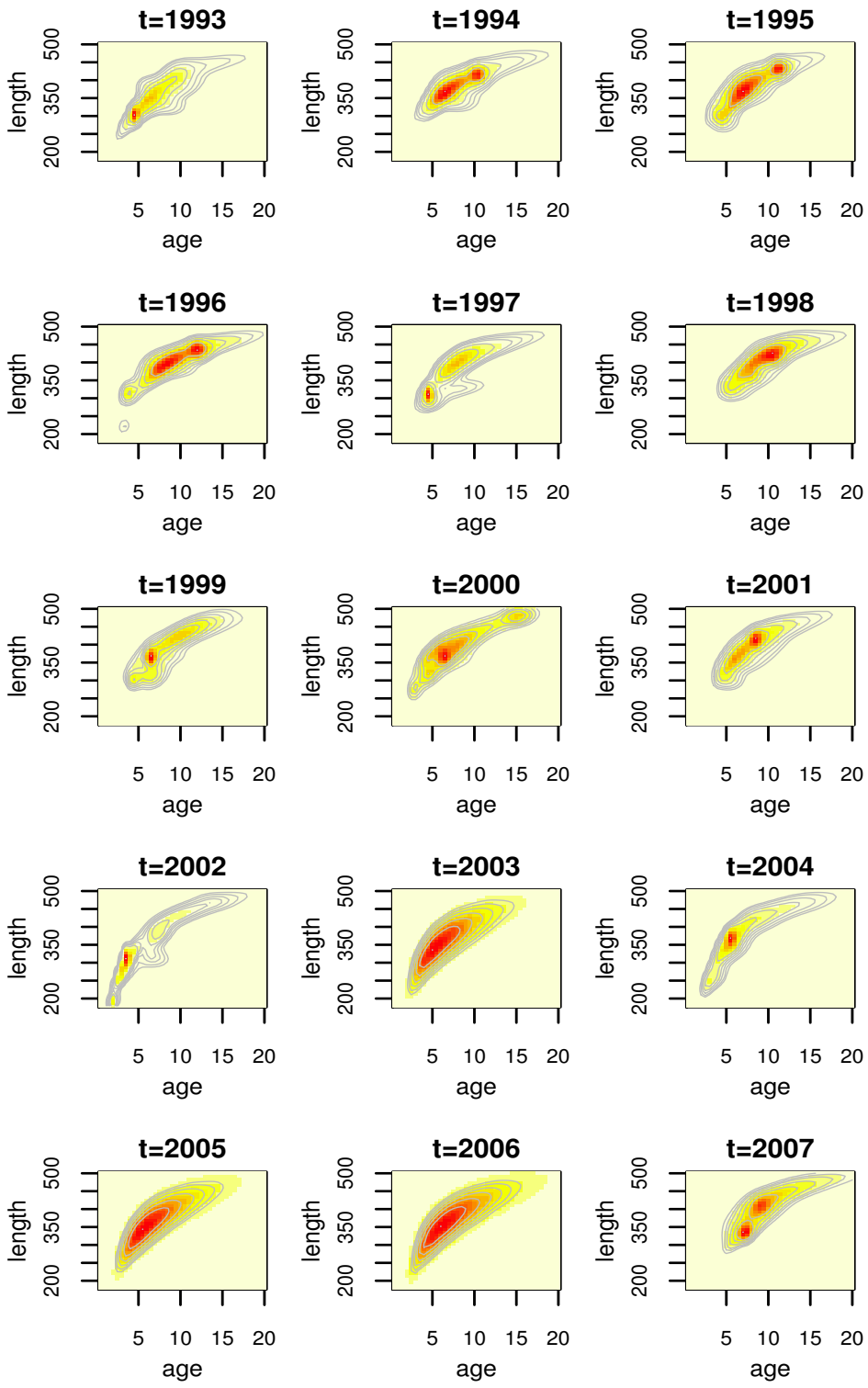

Figure 5.3: Fish maturity example. Posterior mean estimates for the distribution of age and length (mm) across all years.

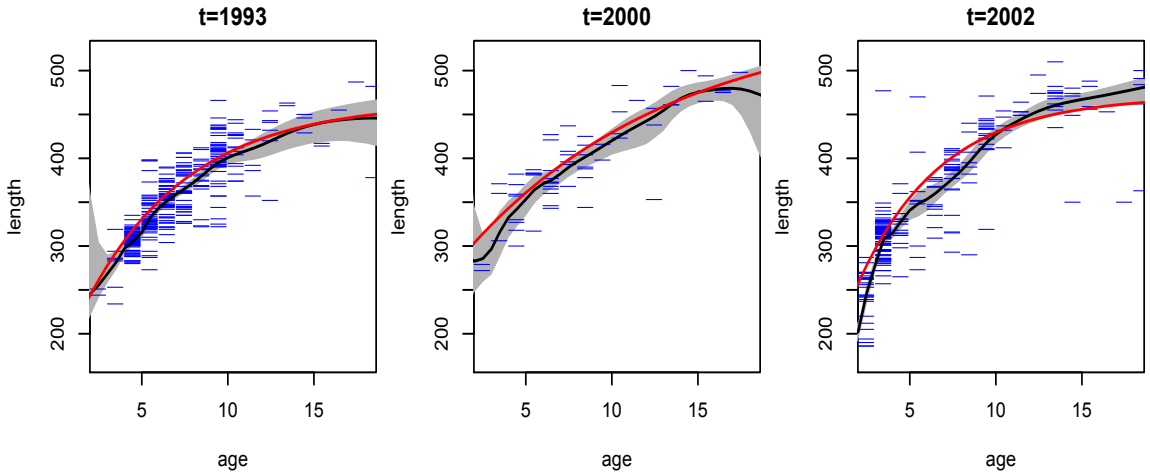

Figure 5.4: Fish maturity example. Posterior mean and 95% interval estimates for $\mathrm{E}_t(X \mid U^* = u^*; G_t)$, evaluated for a grid of values u^*, or the expected value of length over age, compared the the von Bertalanffy growth curves (red) with the data overlaid.

(i.e. by location or cohort) are often used to suggest that the groups differ in some way, however this type of analysis should appropriately take into account the uncertainty in the estimated curves.

Referring again to Figure 5.3, note that there is actually no data in 2003, 2005, or 2006, and the model places more smooth, standard shapes for the distributions in those years, as it should. The distribution in 2002 appears to extend down farther to smaller ages and lengths; in fact this year is unique in that all of the age 1 fish are present in this year (9 of them). In addition, 26 out of 34 of the age 2 fish and 83 out of 118 of the age 3 fish are contained in this year. That is, this year contains a very large proportion of the young fish which are present in the data.

The last year 2007, in addition to containing few observations, is peculiar. There are no fish that are younger than age 6 in this year. Of the 5 age 6 fish, 3 are immature, and of the 7 age 7 fish, all are immature. From this point on, the proportion of immature

fish decreases quickly. In all years combined, less than 10% of age 6 as well as age 7 fish are immature. This year is clearly an anomaly. Is there really some change point at which fish are suddenly maturing much later? Or is this an error on the part of the person who determined and recorded age and maturity at this time? As there are no observations in 2005 or 2006, and a small number of observations in 2007 which seem to not agree with the other years of data, we now report inferences only up to 2004.

5.1.2 Results for Functionals Involving Maturity

Inference for the maturation probability curves is shown over length and age in Figures 5.5 and 5.6. The probability that a fish is immature (solid black line) is generally decreasing over length, reaching a value near 0 at around 350 mm in most years. There is a large change in this probability over length in 2002 and 2004 as compared to other years, as these years suggest a probability close to 1 for very small fish near 200 to 250 mm. Turning to age, the probability of immaturity is also decreasing with age, also showing differences in 2002 and 2004 in comparison to other years. There is not an indication of a general trend in the probabilities associated with levels 2 or 3. Years 1995-1997 and 1999 display similar behavior, with a peak in probability of post-spawning for moderate length values near 350 mm, and ages 6-7, favoring pre-spawning fish at other lengths and ages. The last four years 2001-2004 suggest the probability of pre-spawning mature to be increasing with length up to a point and then leveling off, while post-spawning is favored most for large fish. Post-spawning appears to have a lower probability than pre-spawning mature for any age at all years, with the exception of 1998, for which the probability associated with post-spawning is very high for older fish.

The Pacific States Marine Fisheries Commission states that all Chilipepper rockfish are mature at around 4-5 years, and at size 304 to 330 mm. A stock assessment produced by the Pacific Fishery Management Council (Field, 2009) fitted a logistic regression to model maturity over length, from which it appears that 90% of fish are mature around 300-350 mm. As our model does not enforce monotonicity on the probability of maturity across age, we obtain posterior distributions for the first age not less than 2 (since biologically all fish under 2 should be immature) at which the probability of maturity exceeds 90%, given that it exceeds 90% at some point. That is, for each posterior sample we evaluate $\Pr_t(Y > 1 \mid u^*; G_t)$ over a grid in u^* beginning at 2 and find the smallest value of u^* at which this probability exceeds 90%. Note that there were very few posterior samples for which this probability did not exceed 90% for any age, namely just 4 samples in 1993 and 8 in 2003. The estimates for age at 90% maturity are shown in Figure 5.7. The most noticeable feature is the very narrow interval bands in 2002. Recall that this year contained an abnormally large number of young fish. In this year, over half of fish age 2 (meaning age 2-3) are immature, and over 90% of age 3 (meaning age 3-4) fish are mature, so we would expect the age at 90% maturity to be above 3 but less than 4, which it is. The early and later years seem to be suggesting that fish are maturing later than in the intermediate years, the trend having somewhat of a bathtub shape, with the exception of 1998 which is placed on somewhat higher ages. A similar analysis is performed for length, and suggests a similar trend over time as the age analysis.

Due to the monotonicity in the maturity probability curve in standard approaches, and the fact that age and length are not viewed as random variables, the point at which

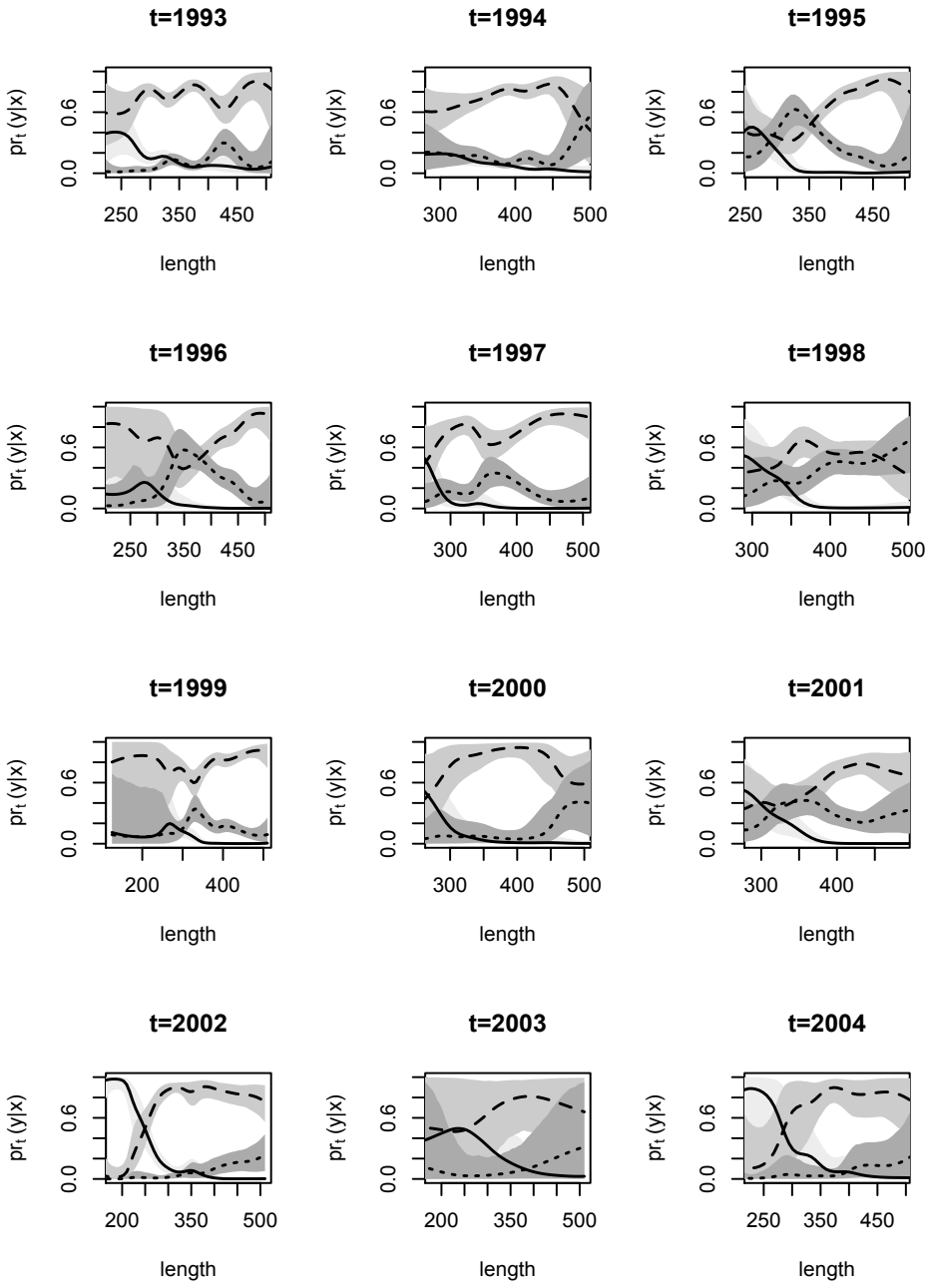

Figure 5.5: Fish maturity example. Posterior mean (black lines) and 95% interval estimates (gray shaded regions) for the marginal ordinal probability curves associated with length. Category 1 (immature) given by solid line, category 2 (mature) given by dashed, and category 3 (post-spawning) shown as a dotted line.

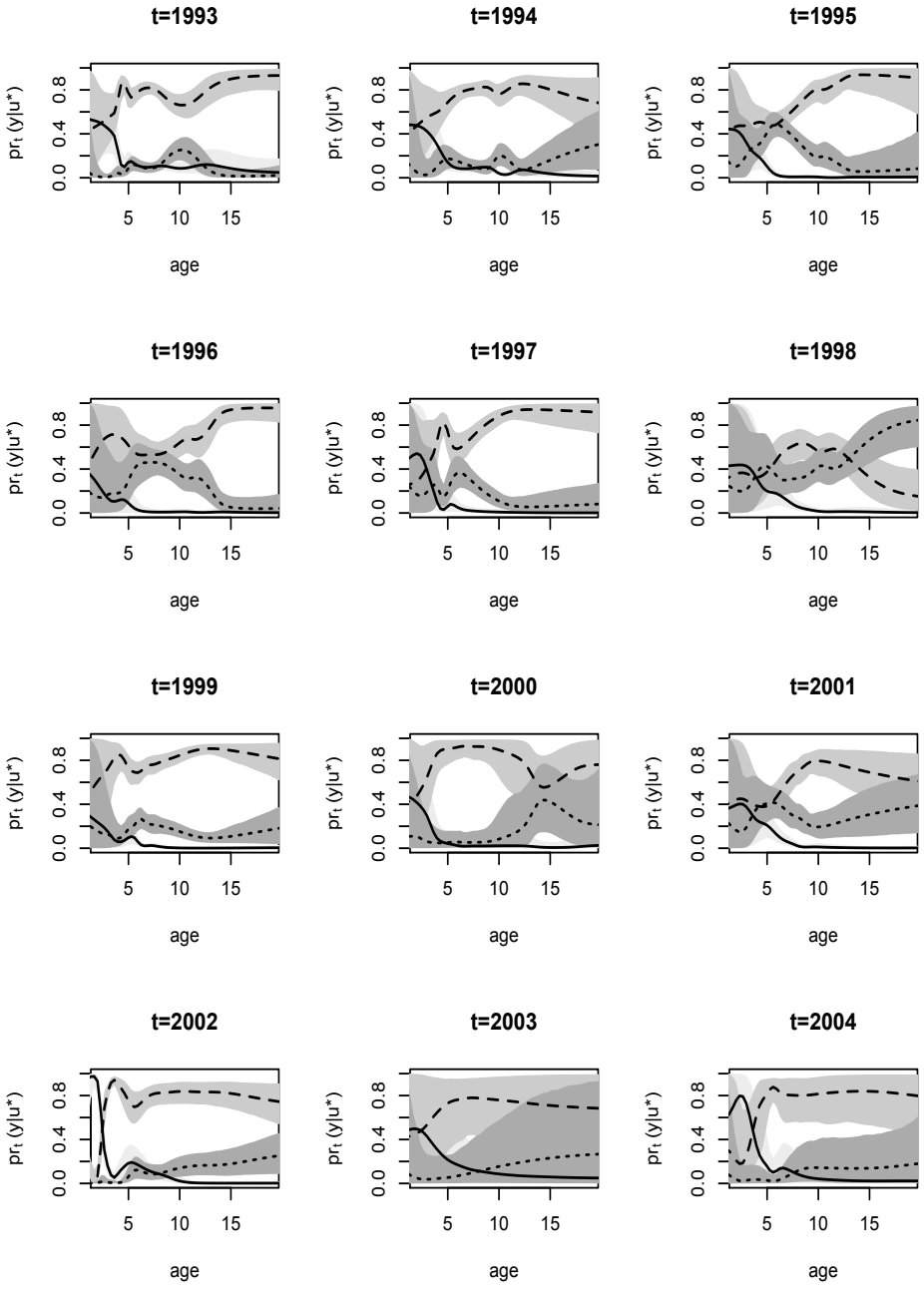

Figure 5.6: Fish maturity example. Posterior mean (black lines) and 95% interval estimates (gray shaded regions) for the marginal ordinal probability curves associated with age. Category 1 (immature) given by solid line, category 2 (mature) given by dashed, and category 3 (post-spawning) shown as a dotted line.

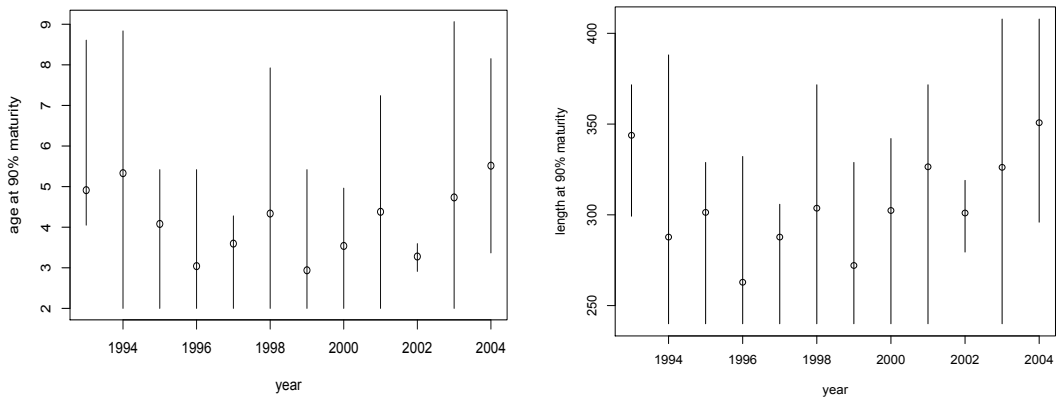

Figure 5.7: Fish maturity example. Posterior mean and 90% intervals for the smallest value of age above 2 years at which probability of maturity first exceeds 90% (left), and similar inference for length (right). Refer to Section 5.1.2 for details.

maturity exceeds a certain probability is a reasonable quantity to obtain in order to study the age or length at which most fish are mature. However, since we are treating age and length as random variables, we can actually obtain their distribution at a given maturity level. These are inverse inferences, in which we study $f_t(x \mid Y = 1; G_t)$ as opposed to $\Pr_t(Y = 1 \mid x; G_t)$, for instance. It is most informative to look at age and length for immature fish, as this makes it clear at which age or length there is essentially no probability assigned to the immature category. The posterior mean for $f_t(u^*, x \mid Y = 1; G_t)$ is shown in Figure 5.8. These can be compared to the bivariate distributions for age and length which closely resemble $f_t(u^*, x \mid Y > 1; G_t)$, since most fish are mature.

5.1.3 Model Checking

We now seek to validate our model, by studying how well it fits the observed data, as well as how it performs in terms of prediction. Two methods of model checking

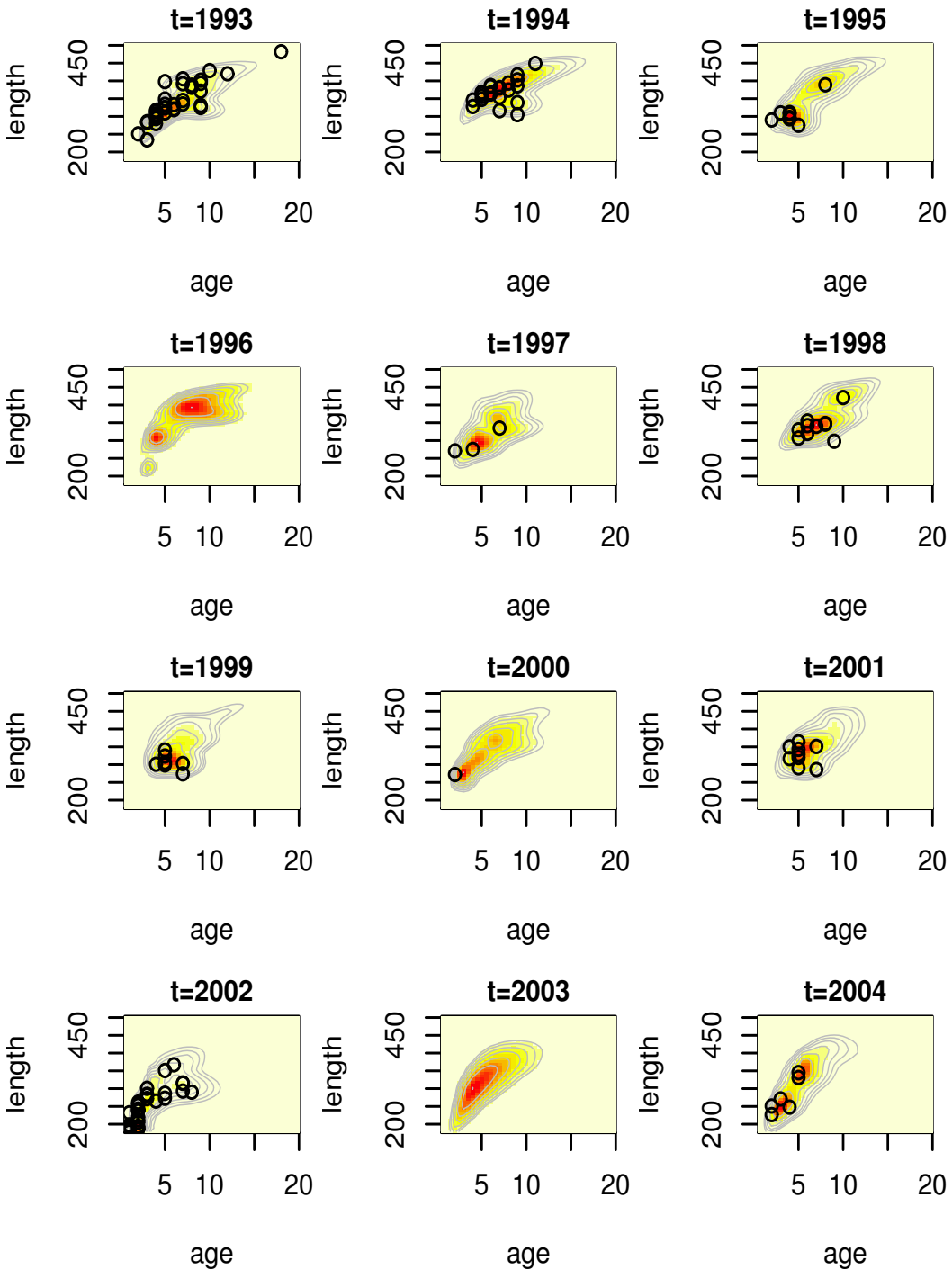

Figure 5.8: Fish maturity example. Posterior mean estimate for the distribution of age and length for immature fish over time, with the age and length of immature fish overlaid.

are performed, one which involves applying the model to a reduced data set and obtaining Bayesian residuals for the observations that were left out. The second approach involves posterior predictive checking, in which we generate replicate data sets from the predictive distribution and compare these to the real data, using various test quantities.

To analyze residuals with cross-validation, we randomly select 20% of the observations in each year and refit the model, leaving out these observations. The inferences obtained for the ordinal regressions and density estimates (not shown) appear unchanged from those based on the complete data. We now obtain residuals for each observation $(\tilde{y}, \tilde{u}, \tilde{x})$ which was left out. The residuals we consider are of the form $\tilde{y}_{t,i} - \mathrm{E}_t(Y \mid U = \tilde{u}_{t,i}, X = \tilde{x}_{t,i}; G_t)$, so that there is one residual for each MCMC posterior sample. Note that the residuals are bounded by -2 and 2, and the closer to 0 the better, although the residual will often not be right near 0, as, for $\tilde{y} = 1$ or $\tilde{y} = 3$, this can only happen if the model assigns a probability of exactly 1 to the particular category. The residuals for $\tilde{y} = 1$ will always be negative, while those for $\tilde{y} = 3$ will be positive, and since there are relatively few 1s, we do not expect to see very many negative residuals. The posterior mean and 95% interval estimates for the residuals are shown for each time, ordered by covariate (age is first ordered followed by length), in Figure 5.9. The reason why there appear to be two or three distinct ranges in residuals at each time is due to the discrete nature of the responses. That is, for two covariate values close together, the model assigns approximately the same expectation, but if one of those observations is 1 and the other is 2, the residuals will differ by approximately 1. There does not appear to be any sort of trend in residuals across covariate values, meaning we are not systematically under or overestimating maturity for fish of a

particular covariate size. All years look roughly similar in terms of the types of residuals we see, with any noticeable differences due to the differences in number of observations left out or number of residuals (determined by sample size in each particular year).

Next, we study how well the model fits the observed data by creating replicate data sets from the posterior predictive distribution. Using the output from the model applied to the full data set, we simulate replicate data sets, $(y, u, x)^{rep}_{t,i}$, $t = 1, \ldots, 12$, $i = 1, \ldots, n_t$, for each MCMC iteration. We then choose some test quantity $T(\{y, u, x\}_t)$, $t = 1, \ldots, 12$, and for each replicate data set, determine the value of the test quantity and compare the distribution of test quantities with the value computed from the real data set. To obtain Figure 5.10 we computed, for each replicate sample, the proportion of age 6 fish that were of maturity levels 1 and 2. Boxplots of these proportions are shown, with the true proportions in the real data set indicated as blue points. The width of each box is proportional to the number of age 6 fish in that year. Figure 5.11 refers to fish of at least age 7 and longer than 400 mm. Finally, we compute the sample correlation of length and age for fish of maturity level 2 in Figure 5.12. These figures all suggest that the model is predicting data which is very similar to the observed data in terms of inferences we really care about.

For comparison, we also fitted the simpler common atoms DDP model to the data, and computed the predictive criterion of Gelfand and Ghosh (1998), used earlier in Section 2.3.2, which is composed of a sum of squares goodness of fit term, and sum of predictive variances penalty term. The goodness of fit term at time t is given by $\sum_{i=1}^{n_t}(y_{t,i} - \mathrm{E}_t(Y \mid U = u_{t,i}, X = x_{t,i}; \text{data}))^2$, and the penalty term is $\sum_{i=1}^{n_t}(\mathrm{E}_t(Y^2 \mid U = u_{t,i}, X = x_{t,i}; \text{data}) -$

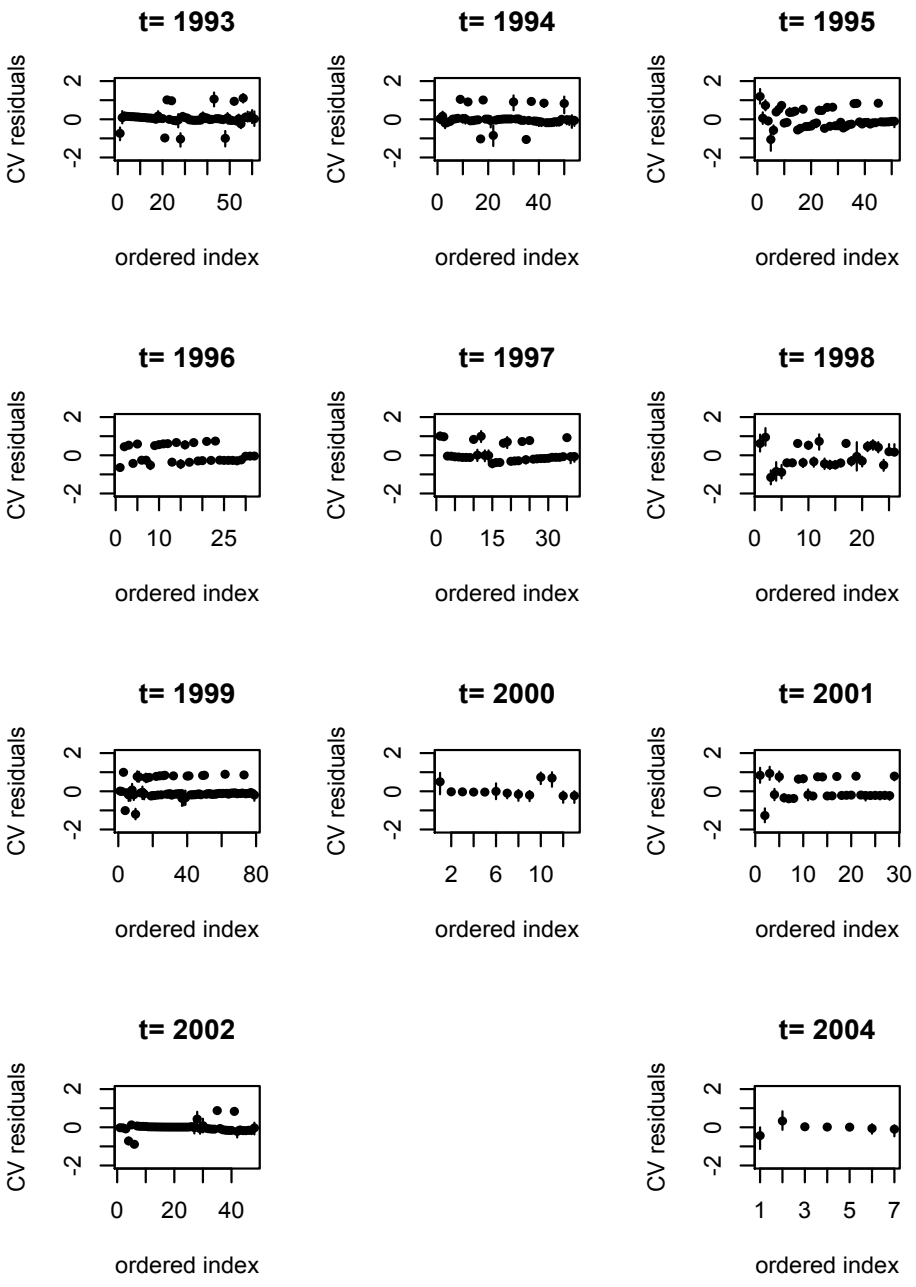

Figure 5.9: Fish maturity example. Posterior mean and 95% intervals for the cross-validation residuals, ordered by covariate values.

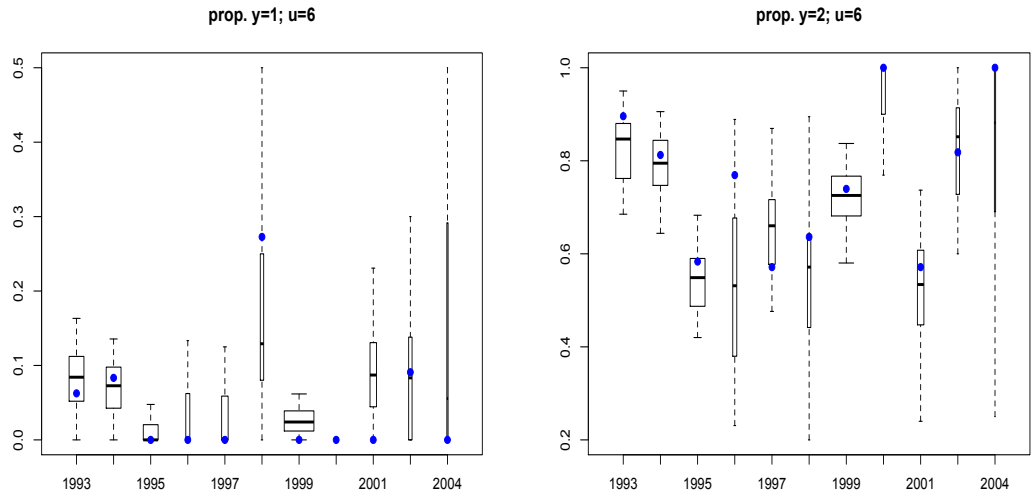

Figure 5.10: Fish maturity example. Distributions of the proportion of age 4 fish that were of maturity level 1 (left) and 2 (right) in the replicated data sets are shown as boxplots, with width proportional to the number of age 6 fish in each year. The true proportion in the real data set is given as a blue circle.

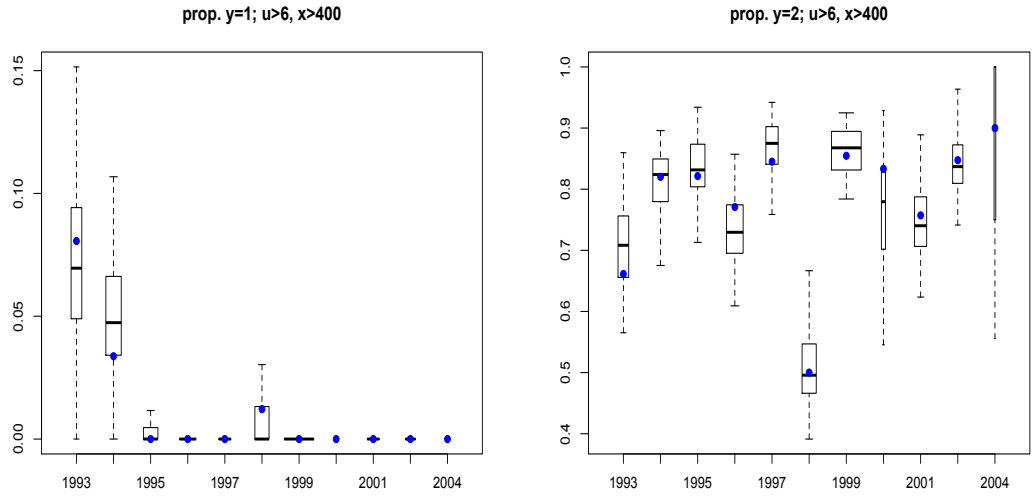

Figure 5.11: Fish maturity example. Distributions of the proportion of fish age 7 and above and length larger than 400 mm that were of maturity level 1 (left) and 2 (right) in the replicated data sets are shown as boxplots, with width proportional to the number fish of this age and length in each year. The true proportion in the real data set is given as a blue circle.

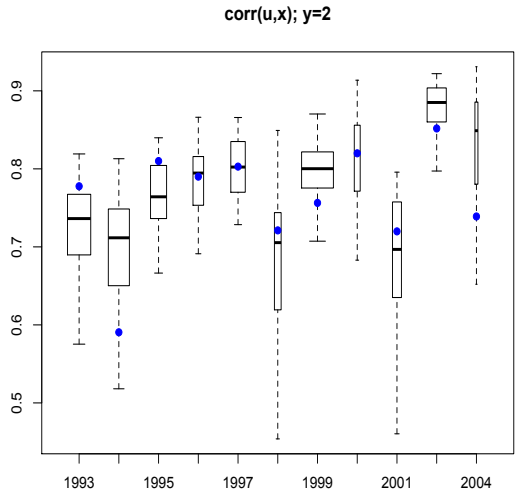

Figure 5.12: Fish maturity example. Distributions of the sample correlation of fish that were of maturity level 2 in the replicated data sets are shown as boxplots, with width proportional to the number of level 2 fish in each year. The sample correlation present in the real data set is given as a blue circle.

$\mathrm{E}_t^2(Y \mid U = u_{t,i}, X = x_{t,i}; \text{data}))$. Letting Θ denote all model parameters,

$$\mathrm{E}_t(Y \mid U = u_{t,i}, X = x_{t,i}; \text{data}) = \frac{\int\int y p_t(y, u_i, x_i \mid \Theta) p(\Theta \mid \text{data}) d\Theta dy}{\int p_t(u_i, x_i \mid \Theta) p(\Theta \mid \text{data}) d\Theta},$$

where $p_t(u_i, x_i \mid \Theta)$ is $\sum_{l=1}^{N} p_{l,t} \mathrm{N}(x_i; \mu_{l,t}^x, \Sigma_l^{xx}) \int_{\log(u_i)}^{\log(u_i+1)} \mathrm{N}(w; \mathrm{E}_{l,t}(W \mid X = x_i), \mathrm{Var}_l(W \mid X = x_i)) dw$, and $\mathrm{E}_{l,t}(W \mid X = x_i) \equiv \mathrm{E}_l(W \mid X = x_i)$ under the common atoms model. The term $p_t(y, u_i, x_i \mid \Theta)$ has a similar form, but with a double integral over z and w, and $\int y p_t(y, u_i, x_i \mid \Theta) dy = \sum_{j=1}^{3} j p_t(y, u_i, x_i \mid \Theta)$. The numerator and denominator can therefore each be evaluated via Monte Carlo integration of these expressions. The penalty term requires $\mathrm{E}_t(Y^2 \mid U = u_{t,i}, X = x_{t,i}; \text{data})$, for which the numerator requires Monte Carlo integration of $\sum_{j=1}^{3} j^2 p_t(y, u_i, x_i \mid \Theta)$. The goodness of fit and penalty term are computed for each year, and we find that the goodness of fit term is lower under the general model at every time point, although by a small amount. The penalty terms show larger

differences, and are lower under the general model for all years except 1996, 2005, and 2007. This suggests that the general model is providing a better fit to the data with less uncertainty, except in the few cases in which the posterior predictive variance is larger under the more complex model. Overall, the general model is preferred to the common atoms model using this criterion, which confirms our earlier conclusions based on simulated data.

5.2 Modeling Stock Price Changes over Time

Before 1997, all stocks traded on the New York Stock Exchange were priced in eighths. In 1997, they moved to pricing in sixteenths, and are now using a decimal system. There has been some discussion by the Securities and Exchange Commission (SEC) of bringing back the fractional pricing on stocks (for a recent report by the SEC on the effects of decimalization, see `http://www.sec.gov/news/studies/2012/decimalization-072012.pdf`), and corporate bonds still trade in eighths. In analyzing price changes of stocks which are traded in fractions, it is not adequate to treat the price changes as continuous, particularly if the range of changes is not too large. Müller and Czado (2009) and others referenced therein argue that the possible returns occur in clusters which are well separated due to the small range of price changes, and therefore the data must be modeled using a discrete-response model. Their ordinal-response stochastic volatility model assumes a latent continuous time series in which each latent response is normal with mean linear on the covariates and log-variances (log-volatilities) forming an AR(1) process. Most analyses of stock returns or price changes use time series models in which one observation exists at each point in time,

however our model requires replication, being appropriate only for settings in which multiple observations occur at each point in time.

The focus of this example is to model the price changes of Citigroup stock. Citigroup is an American financial services corporation. It has the largest financial services network in the world and is the third largest bank holding company in the US. We focus on the time period of 12 years from 1980 to 1991, considering each year to be one period in time, and weekly price changes to be observations within each time period, resulting in approximately 52 observations per year. The log-volume of trades in each week is used as a covariate. This data is publicly available, obtained here from Yahoo Finance. A quick look at the data shows weekly price changes occur in multiples of \$0.125, as expected. For ease of interpretation of results, we collapse levels of return into 7 ordered categories, representing "large negative" (1), "moderate negative" (2), "small negative" (3), "insignificant change" (4), "small positive" (5), "moderate positive" (6), and "large positive" (7). We define price change (in absolute value) of greater than \$2.50 as large, greater than \$1.25 but not greater than \$2.50 as moderate, greater than \$0.25 but not greater than \$1.25 as small, and less than or equal to \$0.25 as insignificant. This results in proportions of $(0.08, 0.11, 0.20, 0.20, 0.19, 0.12, 0.12)$ assigned to each level, so that a large negative change is the least likely outcome, and small or no change are equally most likely.

The recession of the early 1980s had a severe effect on financial institutions, and in 1987 the stock market crashed in what is known as Black Monday. On average, the returns from 1985 and 1991 are focused on relatively higher values, and those from 1987 on lower values. In most years, the distribution of ordinal returns is roughly symmetric, favoring

moderate values, however some years show different patterns, favoring extreme values (as in 1987) or a particular value (5 is favored in 1985) with higher probability. The years 1988, 1990, and 1991 show more uniform behavior across values. Exploratory analysis of return versus volume suggests that return does not necessarily increase or decrease with volume, but that the potential for extreme losses or gains may increase with volume. This has been suggested by others, and is not surprising since positive or negative news about a company generally leads to more trading action.

The distributions of log-volume suggest an increasing trend in volume over time, with some right-skewness for some years, and differences in standard deviations or peaks (being more peaked at later time points, and less in 1981-1985). The posterior mean for each ordinal return level over log-volume is shown in Figure 5.13 over time. Blue indicates a positive return, black a negative, and red insignificant change. The dotted lines represent extreme changes, dashed lines represent moderate changes, and solid lines represent small changes. In most years, there appears to be a decreasing trend over log-volume for levels 3, 4, 5, those indicating small changes (solid lines), and an increasing trend for level 1 or 7 or both (dotted lines). The trends in 1981-1985 are fairly similar, having high probability associated with little or no change, except for large volume, for which the probability of a large positive change increases and becomes large. In 1986, a large positive return is very likely when log-volume is moderate to large, a similar pattern to that observed in 1991. In 1987, the year of the stock market crash, we see that for the first time, the probability of a large negative return is most likely for low to moderate volume. In fact, the three most likely outcomes when volume is low to moderate involve negative return. Interestingly, for

large volume the probability of a large positive return becomes most likely.

For the extreme levels 1 and 7, corresponding to large losses and gains, we show the posterior mean along with 95% interval bands (Figures 5.14 and 5.15). It is clear that the probability of a large negative return is significantly higher overall in 1987, and fairly high in 1990 compared to other years, and that the uncertainty is larger in 1987 and later than in earlier years. The probability associated with a large positive gain is generally increasing with volume, and takes on fairly high values for any volume in 1991.

Inference for the probability allocated to each ordinal level over time, $\Pr_t(Y = j; G_t)$, is shown in Figure 5.16. These are consistent with the earlier description of the distributions of returns in the data, favoring small changes in price early on. The probabilities are fairly uniform across values in 1986 but favor more large positive returns, and a significant increase in probability of large negative returns is seen in 1987. In 1988-1990 the probabilities are again fairly uniform across values, and in 1991 the large positive return has the highest probability.

Qualitative conclusions may be drawn from this analysis. It appears that extreme changes in price are more likely when volume is high. The years 1980-1985 were not associated with large gains or losses, and were therefore not risky years for investors, whereas the years 1986-1991 were much more volatile, producing large positive losses or gains with higher probability. The year 1987 was not a good year for this stock, it contained many extreme losses, but also extreme gains, hence it was very volatile.

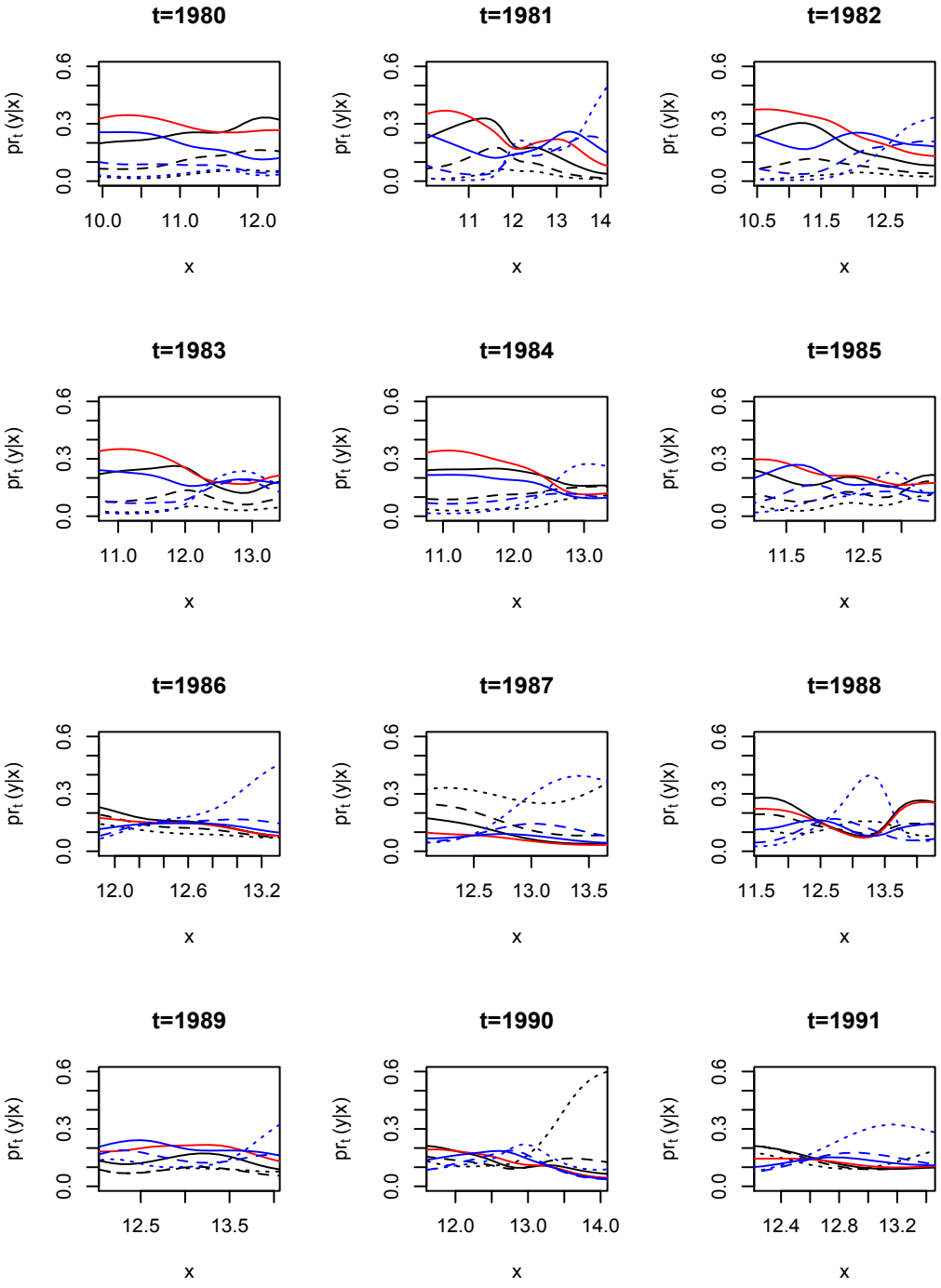

Figure 5.13: Citigroup example. Posterior mean estimates for $\Pr_t(y = j \mid x; G_t)$, for $j = 1, \ldots, 7$. Blue indicates a positive return, black a negative, and red little/no change. The dotted lines represent extreme changes, dashed represents moderate changes, and solid represents small changes.

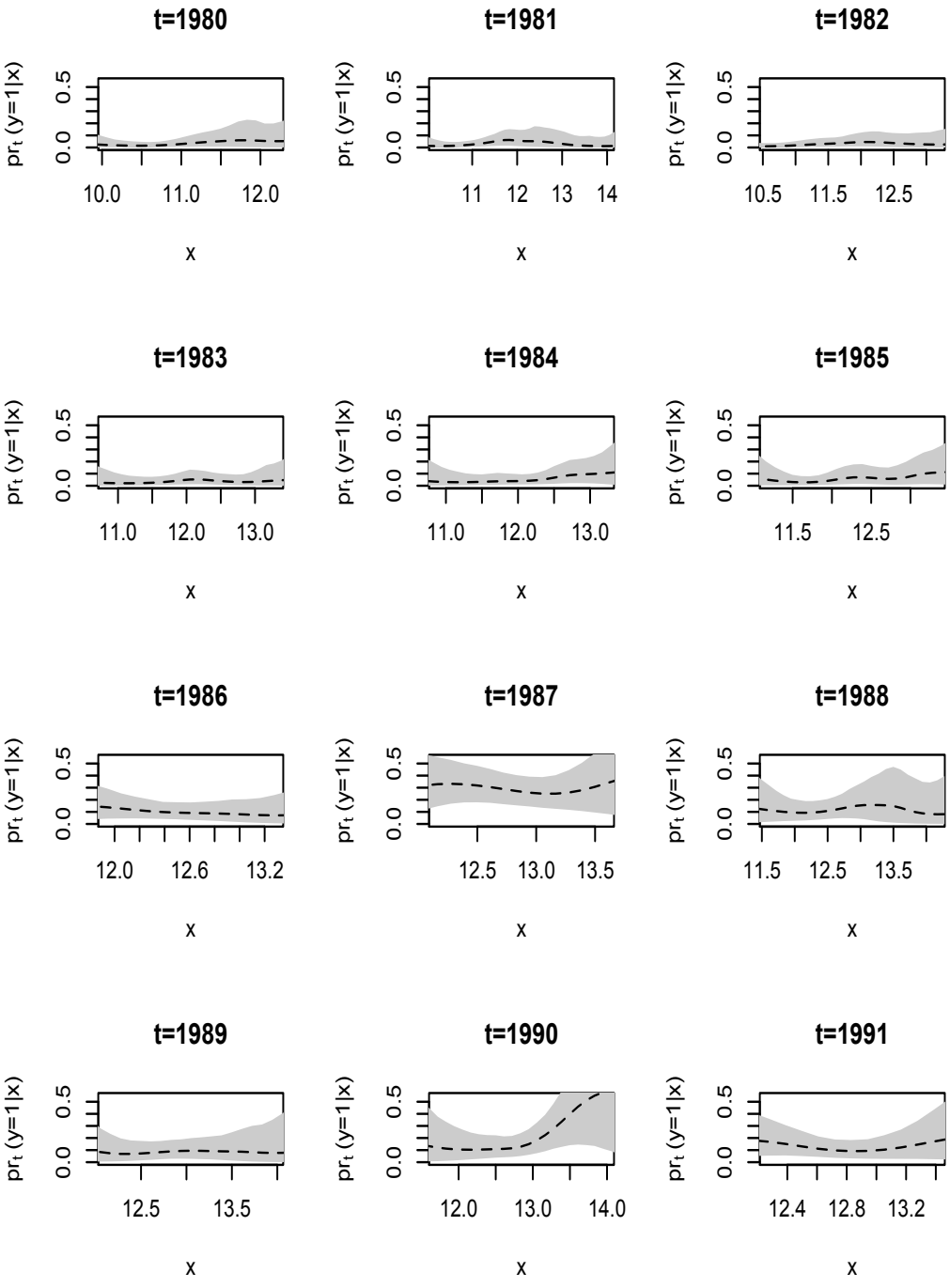

Figure 5.14: Citigroup example. Posterior mean and 95% interval estimates for $\Pr_t(y = 1 \mid x; G_t)$, or the probability of a large negative return.

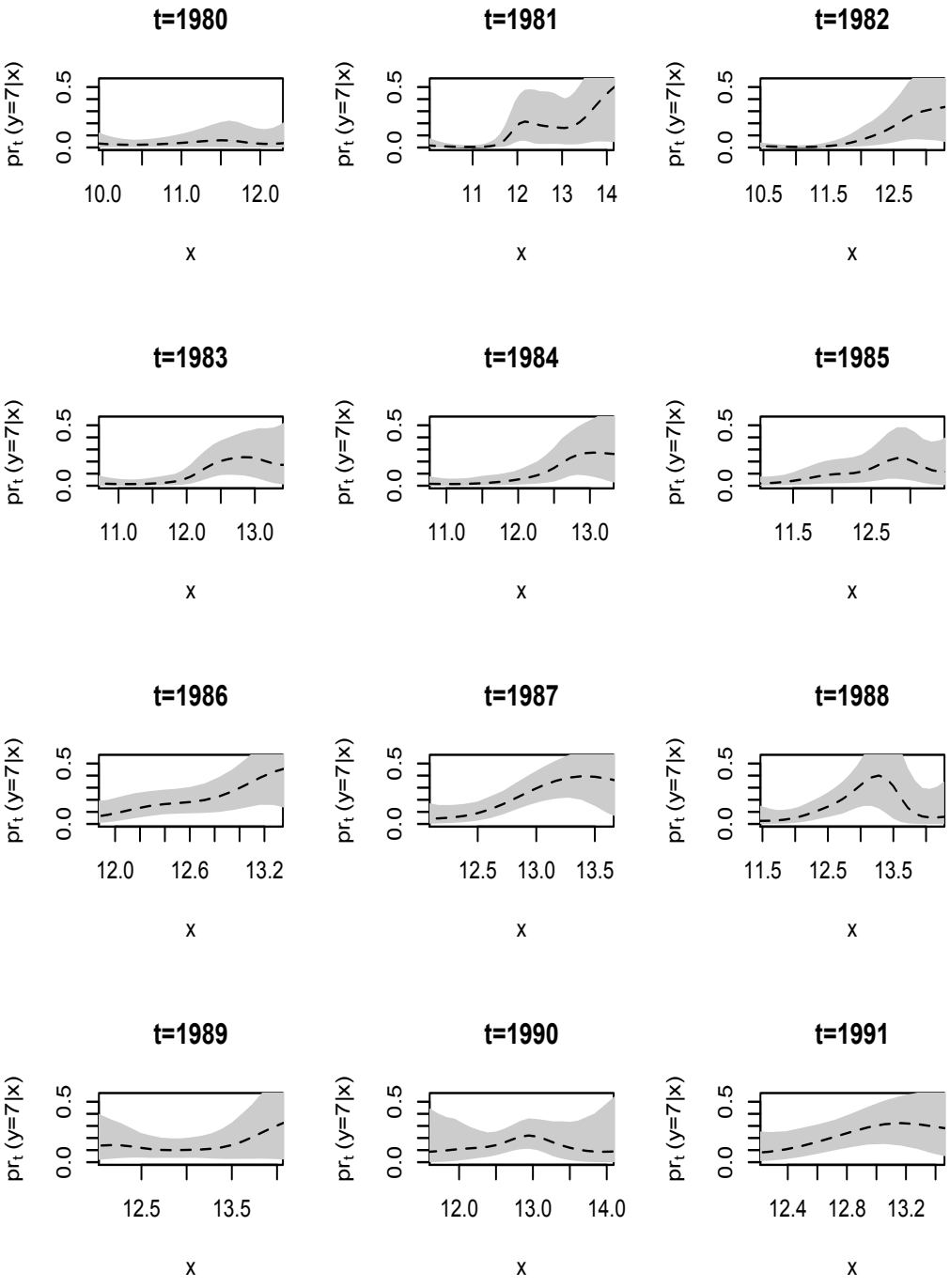

Figure 5.15: Citigroup example. Posterior mean and 95% interval estimates for $\Pr_t(y = 7 \mid x; G_t)$, or the probability of a large positive return.

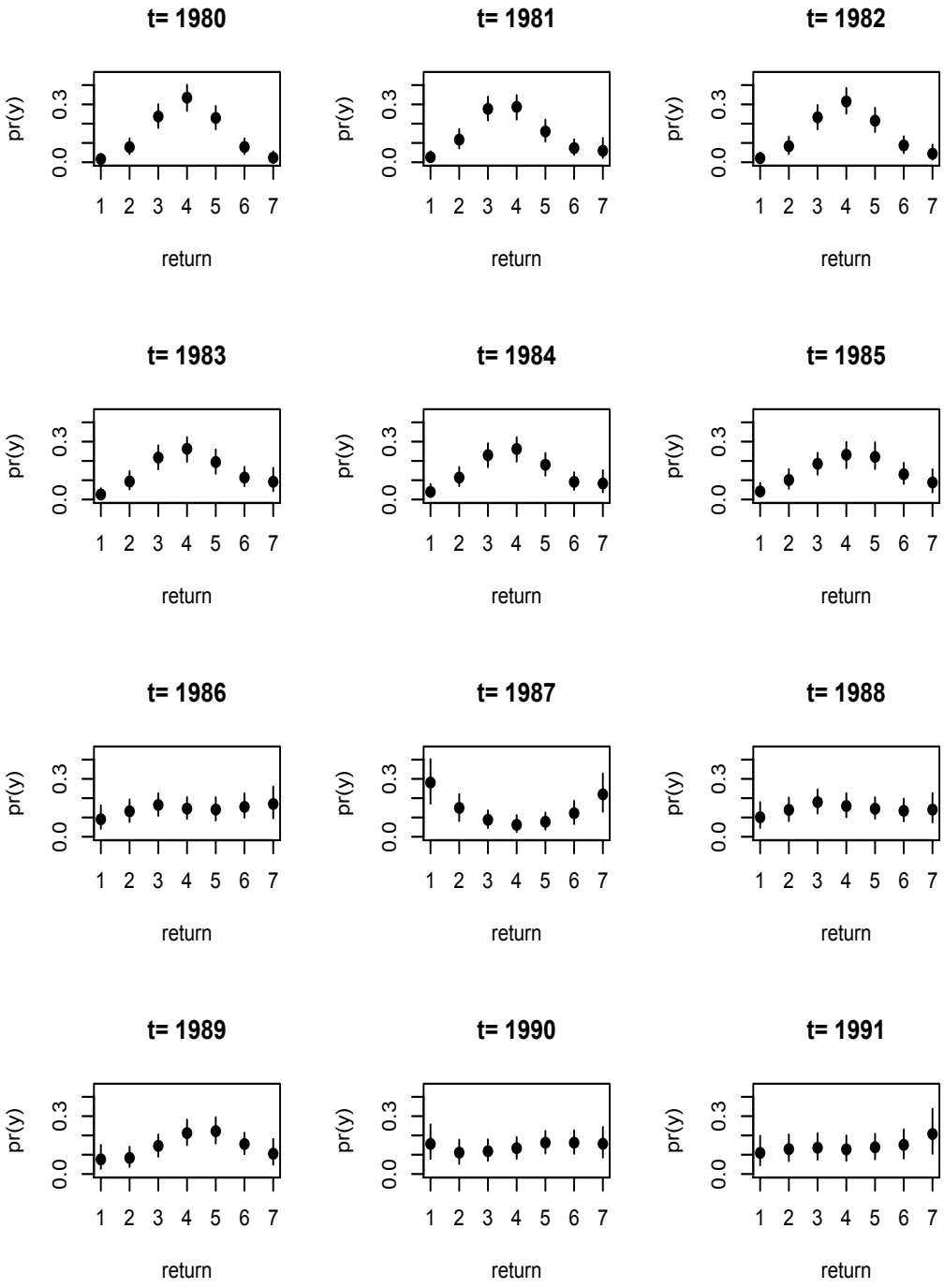

Figure 5.16: Citigroup example. Posterior mean and 95% interval estimates for $\Pr_t(y = j \mid G_t)$, for $j = 1, \ldots, 7$.

5.3 Comments

The methods for dynamic ordinal regression, developed in Chapter 4 and elaborated in this chapter through specific data examples, are widely applicable to modeling mixed ordinal-continuous distributions indexed in discrete-time. The ordinal regression model at any particular point in time is flexible, as the DP mixture representation is retained for the latent response-covariate distribution, as studied in Chapter 3. This means that our approach can provide flexible inference for multiple functionals of the response-covariate distribution. However, we can not force relationships such as monotonicity in the regression functions without substantial changes in the modeling framework from the start. We see this as an attribute in most settings. Nevertheless, in situations in which it is believed that monotonicity exists, we must realize that the data will dominate the model output, and may not produce strictly monotonic relationships.

In the fish maturity example considered in Section 5.1, it is scientifically accepted that monotonicity exists in the relationship between maturity and age or length. Although our model does not force this, the inferences generally agree with what is expected to be true biologically. Our model is also extremely relevant to this setting, as the covariates age and length are treated as random, and the ordinal nature of age is accounted for by the model using variables which represent underlying continuous age. The set of inferences that are provided under this framework, including estimates for length as a function of maturity, which is comparable to a growth curve, make this modeling approach powerful for the particular application considered, as well as related problems.

While year of sampling was considered to be the index of dependence in this

analysis, an alternative is to consider cohort as an index of dependence. All fish born in the same year, or the same age in a given year, represent one cohort. Grouping fish by cohort rather than year of record should lead to more homogeneity within a group, however there are also some possible issues since fish will generally be younger as cohort index increases. This is a consequence of having a particular set of years for which data is collected, i.e., the cohort of fish born in 2006 can not be older than 4 if data collection stopped in 2009. Due to complications such as these, combined with exploration of the data and the relationships implied within each cohort, we decided to treat year of data collection as the index of dependence, but cohort indexing could be explored further. There are also other possible changes that may be made in this analysis, such as collapsing maturity in a different way, or considering each of the 6 levels of maturity to be distinct.

We also provided an illustration using Citigroup stock data in Section 5.2, to indicate the potential for utility of our methods for problems in econometrics. There are a large number of econometric time series models which are built to handle features such as stochastic volatility, in which the variance of a process is non-constant over time. While the dynamic ordinal regressions model is not a time series model, it can be used when multiple observations exist at each point in time, accommodating nonstandard features across time, such as differences in variance, as well as within a particular point in time.

Chapter 6

Conclusions

To conclude this dissertation, we discuss some extensions of the proposed methods. We began the thesis by stating that our focus was on regression involving ordered categorical responses. While this is an accurate claim, we note that it is possible to extend the methods to settings containing multivariate mixed ordinal-continuous responses. The models for ordinal regression were developed to handle ordinal responses and continuous covariates, as well as ordinal covariates, through use of a multivariate normal mixture kernel. It can therefore also handle mixed ordinal-continuous responses, with continuous and/or ordinal covariates. For any combination of ordinal and continuous variables, the mixture of multivariate normals model can be applied to the joint distribution of the latent continuous variables and the truly continuous variables. The distinction between the responses and covariates comes into play later, when producing conditional inferences.

A number of applications which were considered here mainly as illustrations could be expanded. One is the fish maturity example, in which certain aspects of the analysis

could be performed differently, as already discussed in Section 5.3. While time was the focus of dependence here, similar applications may instead contain a spatial index. To incorporate dependence across spatial locations, or sites, a spatial model could be introduced in place of the time series model for the atoms of the DP (Gelfand et al., 2005), and the weights could either be assumed constant in space, or dependence introduced by a spatial model (Duan et al., 2007). At any particular location, the ordinal regression model induced through a dependent mixture of multivariate normal kernels retains the same form as under the time-dependent setting. To account for space when there exist few spatial locations, such as "northern" and "southern" port complexes in the fish maturity data, a dependent DP prior may be assigned to the finite collection of mixing distributions, say G_n and G_s in the setting mentioned, in similar spirit to the methods of De Iorio et al. (2004) or Teh et al. (2006).

The model for ordinal regression was applied to a data set of multirater agreement in Section 3.3.5. As most of the existing methods for quantifying agreement among raters are restrictive in terms of incorporating covariate information and dependence between raters, the nonparametric multivariate ordinal regression model has potential for utility in this setting. However, more thought is needed in terms of determining an effective way of measuring agreement between raters. In addition, if obtaining an overall underlying score for each subject is a key component of inference, the model as it stands must be tailored to provide estimation for the intrinsic subject-specific score.

We mention throughout the thesis that our methods accommodate interactions between covariates in a natural way, through the joint modeling framework. A question

remains on how to quantify those interactions. In the ozone data example of Section 3.3.2, we proposed using standard decompositions from sensitivity analysis to determine the pairwise interactions. It remains to be determined whether this technique, when applied to the deterministic function which is an expectation of the latent response, allows for estimation of the magnitude of the interactions, or only the sign. Further study in the ordinal response setting, as well as with a continuous response, may indicate if this decomposition is an appropriate method for studying interactions among covariates when Bayesian nonparametric curve fitting models are applied.

This dissertation provides a collection of flexible modeling and inference methods for a variety of problems in ordinal regression. We take a Bayesian nonparametric approach, using existing prior models and developing a new DDP prior for distributions indexed in discrete-time. Our impetus for performing Bayesian nonparametric curve fitting regression lies in the significant attributes afforded by this approach in terms of flexibility, uncertainty quantification, the treatment of covariates, and accommodation of dependence. The methodology contained in this thesis has a very wide scope, being applicable to a number of data analysis problems, as indicated by the examples drawn from the social sciences, econometrics, and especially, the biological and environmental sciences.

Bibliography

Albert, J. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Antoniak, C. (1974), "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.

Ashford, J. and Sowden, R. (1970), "Multi-variate Probit Analysis," *Biometrics*, 26, 535–546.

Basu, S. and Mukhopadhyay, S. (2000), "Bayesian Analysis of Binary Regression Using Symmetric and Asymmetric Links," *The Indian Journal of Statistics Series B*, 62, 372–387.

Blackwell, D. (1973), "Discreteness of Ferguson Selections," *The Annals of Statistics*, 1, 356–358.

Bobko, S. and Berkeley, S. (2004), "Maturity, ovarian cycle, fecundity, and age-specific parturition of black rockfish (Sebastes melanops)," *Fisheries Bulletin*, 102, 418–429.

Boes, S. and Winkelmann, R. (2006), "Ordered Response Models," *Advances in Statistical Analysis*, 90, 165–179.

Box and Tiao (1973), *Bayesian Inference in Statistical Analysis*, Reading, Massachusetts: Addison-Wesley.

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

Canale, A. and Dunson, D. (2011), "Bayesian Kernel Mixtures for Counts," *Journal of the American Statistical Association*, 106, 1528–1539.

Chen, M. and Dey, D. (2000), "Bayesian Analysis for Correlated Ordinal Data Models," in *Generalized Linear Models: A Bayesian Perspective*, eds. Dey, D., Ghosh, S., and Mallick, B., New York: Marcel Dekker, pp. 135–162.

Chib, S. and Greenberg, E. (1998), "Analysis of multivariate probit models," *Biometrika*, 85, 347–361.

— (2010), "Additive cubic spline regression with Dirichlet process mixture errors," *Journal of Econometrics*, 156, 322–336.

Choudhuri, N., Ghosal, S., and Roy, A. (2007), "Nonparametric binary regression using a Gaussian process prior," *Statistical Methodology*, 4, 227–243.

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

Clark, W. (1991), "Groundfish exploitation rates based on life history parameters," *Canadian Journal of Fisheries and Aquatic Sciences*, 48, 734–750.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20, 37–46.

Connor, R. and Mosimann, J. (1969), "Concepts of Independence for proportions with a generalization of the Dirichlet distribution," *Journal of the American Statistical Association*, 64, 194–206.

Daniels, M. and Pourahmadi, M. (2002), "Bayesian analysis of covariance matrices and dynamic models for longitudinal data," *Biometrika*, 89, 553–566.

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

Denison, D., Holmes, C., Mallick, B., and Smith, A. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, England: Wiley.

Duan, J., Guindani, M., and Gelfand, A. (2007), "Generalized Spatial Dirichlet Process Models," *Biometrika*, 94, 805–825.

Dunson, D. and Bhattacharya, A. (2010), "Nonparametric Bayes regression and classification through mixtures of product kernels," *Bayesian Statistics*, 9, 145–164.

Dunson, D. and Park, J. (2008), "Kernel Stick-breaking Processes," *Biometrika*, 95, 307–323.

Eaton, M. (2007), *Multivariate Statistics: A Vector Space Approach*, Beachwood, Ohio: Institute of Mathematical Statistics.

Escobar, M. and West, M. (1995), "Bayesian Density Estimation and Inference using Mixtures," *Journal of the American Statistical Association*, 90, 577–568.

Ferguson, T. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

Field, J. (2009), "Status of the Chilipepper rockfish, Sebastes Goodei, in 2007," Tech. rep., Southwest Fisheries Science Center.

Fleiss, J. (1971), "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, 76, 378–382.

Follmann, D. and Lamberdt, E. (1989), "Generalizing logistic regression by nonparametric modelling," *Journal of the American Statistical Association*, 84, 295–300.

Früwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer.

Gelfand, A. and Ghosh, S. (1998), "Model choice: A minimum posterior predictive loss approach," *Biometrika*, 1, 1–11.

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

Ghosh, J. and Ramamoorthi, R. (2003), *Bayesian Nonparametrics*, New York: Springer.

Gill, J. and Casella, G. (2009), "Nonparametric Priors for Ordinal Bayesian Social Science Models: Specification and Estimation," *Journal of the American Statistical Association*, 104, 453–464.

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

Hannah, L., Blei, D., and Powell, W. (2011), "Dirichlet Process Mixtures of Generalized Linear Models," *Journal of Machine Learning Research*, 1, 1–33.

Hannah, R., Blume, M., and Thompson, J. (2009), "Length and age at maturity of female yelloweye rockfish (Sebastes rubberimus) and cabezon (Scorpaenichthys marmoratus) from Oregon waters based on histological evaluation of maturity," Tech. rep., Oregon Department of Fish and Wildlife.

Hastie, T. and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 1, 297–318.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

Hobert, J. and Casella, G. (1996), "The effect of improper priors on Gibbs sampling in hierarchical linear mixed models," *Journal of the American Statistical Association*, 61, 1461–1473.

Imai, K. and van Dyk, D. (2005), "A Bayesian Analysis of the multivariate probit model using marginal data augmentation," *Journal of Econometrics*, 124, 311–334.

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

Ishwaran, H. and Zarepour, M. (2000), "Markov Chain Monte Carlo in approximate Dirichlet and Beta two-parameter process Hierarchical Models," *Biometrika*, 87, 371–390.

Janzen, F. and Stern, H. (1998), "Logistic Regression for Empirical Studies of Multivariate Selection," *Evolution*, 52, 1564–1571.

Johnson, V. and Albert, J. (1999), *Ordinal Data Modeling*, New York: Springer.

Koop, G. (2003), *Bayesian Econometrics*, Chichester: John Wiley and Sons.

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

Lande, R. and Arnold, S. (1983), "The Measurement of Selection on Correlated Characters," *Evolution*, 37, 1210–1226.

Lawrence, E., Bingham, D., Liu, C., and Nair, V. (2008), "Bayesian Inference for Multivariate Ordinal Data Using Parameter Expansion," *Technometrics*, 50, 182–191.

Liu, C. (2001), "Bayesian analysis of multivariate probit models - discussion on the art of data augmentation by Van Dyk and Meng," *Journal of Computational and Graphical Statistics*, 10, 75–81.

Liu, J. and Wu, Y. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Assocation*, 94, 1264–1274.

MacEachern, S. (2000), "Dependent Dirichlet processes," Tech. rep., The Ohio State University Department of Statistics.

McCullagh, P. and Nelder, J. (1983), *Generalized Linear Models*, London: Chapman and Hall.

McCulloch, P., Polson, N., and Rossi, P. (2000), "A Bayesian analysis of the multinomial probit model with fully identified parameters," *Journal of Econometrics*, 99, 173–193.

McKenzie, E. (1985), "An Autoregressive Process for Beta Random Variables," *Management Science*, 31, 988–997.

Morgan, M. and Hoenig, J. (1997), "Estimating Maturity-at-Age from Length Stratified Sampling," *Journal of Northwest Atlantic Fisheries Science*, 21, 51–63.

Mukhopadyay, S. and Gelfand, A. (1997), "Dirichlet Process Mixed Generalized Linear Models," *Journal of the American Statistical Association*, 92, 633–639.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian Curve Fitting Using Multivariate Normal Mixtures," *Biometrika*, 83, 67–79.

Nadaraya, E. (1964), "On Estimating Regression," *Theory of Probability and its Applications*, 9, 141–142.

Neal, R. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Newton, M., Czado, C., and Chappell, R. (1996), "Bayesian Inference for Semiparametric Binary Regression," *Journal of the American Statistical Association*, 91, 142–153.

Nieto-Barajas, L., Müller, P., Ji, Y., Lu, Y., and Mills, G. (2012), "A Time-Series DDP for Functional Proteomics Profiles," *Biometrics*, 68, 859–868.

Oakley, J. and O'hagan, A. (2004), "Probabilistic sensitivity analysis of complex models: a Bayesian approach," *Journal of the Royal Statistical Society, Series B*, 66, 751–769.

Olsson, U. (1979), "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient," *Psychometrika*, 44, 443–460.

Papageorgiou, G., Richardson, S., and Best, N. (2014), "Bayesian nonparametric models for spatially indexed data of mixed type," Pre-print available at `http://arxiv.org/abs/1408.1368`.

Rodriguez, A., Dunson, D., and Gelfand, A. (2009), "Bayesian Nonparametric Functional Analysis through Bayesian Density Estimation," *Biometrika*, 96, 149–162.

Rodriguez, A. and ter Horst, E. (2008), "Bayesian Dynamic Density Estimation," *Bayesian Analysis*, 3, 339–366.

Ronning, G. and Kukuk, M. (1996), "Efficient Estimation of Ordered Probit Models," *Journal of the American Statistical Association*, 91, 1120–1129.

Savitsky, T. and Dalal, S. (2014), "Bayesian non-parametric analysis of multirater ordinal data, with application to prioritizing research goals for prevention of suicide," *Journal of the Royal Statistical Society: Series C*, 63, 539–557.

Schluter, D. (1988), "Estimating the Form of Natural Selection on a Quantitative Trait," *Evolution*, 42, 849–861.

Schluter, D. and Smith, J. (1986), "Natural Selection on Beak and Body Size in the Song Sparrow," *International Journal of Organic Evolution*, 40, 221–231.

Schwartz, L. (1965), "On Bayes Procedures," *Z. Wahrsch. Verw. Gebiete*, 4, 10–26.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

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

Simonoff, J. (2003), *Analyzing Categorical Data*, New York: Springer-Verlag.

Taddy, M. (2010), "Autoregressive Mixture Models for Dynamic Spatial Poisson Processes: Application to Tracking the Intensity of Violent Crime," *Journal of the American Statistical Association*, 105, 1403–1417.

Taddy, M. and Kottas, A. (2010), "A Bayesian Nonparametric Approach to Inference for Quantile Regression," *Journal of Business and Economic Statistics*, 28, 357–369.

Talhouk, A., Doucet, A., and Murphy, K. (2012), "Efficient Bayesian Inference for Multivariate Probit Models with Sparse Inverse Correlation Matrices," *Journal of Computational and Graphical Statistics*, 21, 739–757.

Tanner, M. and Young, M. (1985), "Modeling Ordinal Scale Disagreement," *Psychological Bulletin*, 98, 408–415.

Teh, Y., Jordan, M., and Blei, D. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581.

Tiao, G. and Zellner, A. (1964), "On the Bayesian Estimation of Multivariate Regression," *Journal of the Royal Statistical Society, Series B*, 26, 277–285.

Trippa, L. and Muliere, P. (2009), "Bayesian Nonparametric Binary Regression via Random Tesselations," *Statistics and Probability Letters*, 79, 2273–2282.

Verbeek, M. (2008), *A Guide to Modern Econometrics*, John Wiley and Sons, 3rd ed.

Walker, S. and Mallick, B. (1997), "Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing," *Journal of the Royal Statistical Society B*, 59, 845–860.

Watson, G. (1964), "Smooth Regression Analysis," *Sankhya: The Indian Journal of Statistics, Series A*, 26, 359–372.

Webb, E. and Forster, J. (2008), "Bayesian Model Determination for Multivariate Ordinal and Binary Data," *Computational Statistics and Data Analysis*, 52, 2632–2649.

Wood, S. and Kohn, R. (1998), "A Bayesian Approach to Robust Binary Nonparametric Regression," *Journal of the American Statistical Association*, 93, 203–213.

Wu, Y. and Ghosal, S. (2008), "Kullback Leibler Property of Kernel Mixture Priors in Bayesian Density Estimation," *Electronic Journal of Statistics*, 2, 298–331.

Appendix A

Theoretical Results

A.1 Identifiability Results

A.1.1 Proof of Lemma 1

Recall the kernel distribution in (2.3) for which we wish to prove that parameters $(\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zx})$ are identifiable, fixing $\Sigma^{zz} = 1$. Assume that

$$k(y, \boldsymbol{x}; \mu_1^x, \mu_1^z, \Sigma_1^{xx}, \Sigma_1^{zx}) = k(y, \boldsymbol{x}; \mu_2^x, \mu_2^z, \Sigma_2^{xx}, \Sigma_2^{zx}). \tag{A.1}$$

If this implies $(\mu_1^x, \mu_1^z, \Sigma_1^{xx}, \Sigma_1^{zx}) = (\mu_2^x, \mu_2^z, \Sigma_2^{xx}, \Sigma_2^{zx})$, then $(\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zx})$ are identifiable.

From (A.1), it must be the case that $\mathrm{N}_p(\boldsymbol{x}; \mu_1^x, \Sigma_1^{xx}) = \mathrm{N}_p(\boldsymbol{x}; \mu_2^x, \Sigma_2^{xx})$. This follows from summing each side of (A.1) over the two possible values of y. Because the mean vector and covariance matrix are identifiable for the multivariate normal likelihood, it can be concluded that $\mu_1^x = \mu_2^x$, and $\Sigma_1^{xx} = \Sigma_2^{xx}$. Now, after this simplification, each side of the equality in (A.1) consists of a Bernoulli distribution for $k(y \mid \boldsymbol{x})$, and since y is either 0 or 1, the corresponding Bernoulli probabilities must be equal. Since Φ is a monotonically

increasing function of its argument, the arguments of Φ are equal, that is,

$$\frac{\mu_1^z + \Sigma_1^{zx}(\Sigma^{xx})^{-1}(\boldsymbol{x} - \mu^x)}{(1 - \Sigma_1^{zx}(\Sigma^{xx})^{-1}(\Sigma_1^{zx})^T)^{1/2}} = \frac{\mu_2^z + \Sigma_2^{zx}(\Sigma^{xx})^{-1}(\boldsymbol{x} - \mu^x)}{(1 - \Sigma_2^{zx}(\Sigma^{xx})^{-1}(\Sigma_2^{zx})^T)^{1/2}}.$$

This can be written in the form $\boldsymbol{a}^T\boldsymbol{x} + b = 0$, and in order for this to be true for all $\boldsymbol{x}$, each element of vector $\boldsymbol{a}$ must be 0, and scalar b must be 0. The two equations $\boldsymbol{a} = \mathbf{0}$ and $b = 0$ require

$$\frac{\Sigma_1^{zx}}{(1 - \Sigma_1^{zx}(\Sigma^{xx})^{-1}(\Sigma_1^{zx})^T)^{1/2}} = \frac{\Sigma_2^{zx}}{(1 - \Sigma_2^{zx}(\Sigma^{xx})^{-1}(\Sigma_2^{zx})^T)^{1/2}} \tag{A.2}$$

$$\frac{\mu_1^z - \Sigma_1^{zx}(\Sigma^{xx})^{-1}\mu^x}{(1 - \Sigma_1^{zx}(\Sigma^{xx})^{-1}(\Sigma_1^{zx})^T)^{1/2}} = \frac{\mu_2^z - \Sigma_2^{zx}(\Sigma^{xx})^{-1}\mu^x}{(1 - \Sigma_2^{zx}(\Sigma^{xx})^{-1}(\Sigma_2^{zx})^T)^{1/2}} \tag{A.3}$$

Using (A.2), (A.3) can be replaced by $\mu_1^z\Sigma_2^{zx} = \mu_2^z\Sigma_1^{zx}$. Writing these two equations component-wise, and letting Σ_{ji}^{zx} denote element i of the vector Σ_j^{zx}, results in two systems of p equations:

$$\frac{(\Sigma_{1i}^{zx})^2}{1 - \Sigma_1^{zx}(\Sigma^{xx})^{-1}(\Sigma_1^{zx})^T} = \frac{(\Sigma_{2i}^{zx})^2}{1 - \Sigma_2^{zx}(\Sigma^{xx})^{-1}(\Sigma_2^{zx})^T}, \quad i = 1, ..., p \tag{A.4}$$

$$\mu_1^z\Sigma_{2i}^{zx} = \mu_2^z\Sigma_{1i}^{zx}, \quad i = 1, ..., p \tag{A.5}$$

When $p = 1$ such that Σ^{zx} is a scalar, (A.4) becomes $\mid \Sigma_1^{zx} \mid = \mid \Sigma_2^{zx} \mid$, which has only the solution $\Sigma_1^{zx} = \Sigma_2^{zx}$, since $\Sigma_1^{zx} = -\Sigma_2^{zx}$ would violate (A.2). Then from (A.5) we conclude $\mu_1^z = \mu_2^z$.

In general, with p covariates, (A.4) can be written as

$$(\Sigma_{1i}^{zx})^2 - (\Sigma_{1i}^{zx})^2 \sum_{k=1}^{p}\sum_{j=1}^{p} \Sigma_{2j}^{zx}\Sigma_{2k}^{zx}(\Sigma^{xx})_{jk}^{-1} = (\Sigma_{2i}^{zx})^2 - (\Sigma_{2i}^{zx})^2 \sum_{k=1}^{p}\sum_{j=1}^{p} \Sigma_{1j}^{zx}\Sigma_{1k}^{zx}(\Sigma^{xx})_{jk}^{-1}, \quad i = 1, ..., p$$

Because (A.5) implies $\Sigma_{1l}^{zx}\Sigma_{2m}^{zx} = \Sigma_{1m}^{zx}\Sigma_{2l}^{zx}$ for any $l, m = 1, ..., p$, the equation reduces to $(\Sigma_{1i}^{zx})^2 = (\Sigma_{2i}^{zx})^2$. The constraint $\Sigma_{1l}^{zx}\Sigma_{2m}^{zx} = \Sigma_{1m}^{zx}\Sigma_{2l}^{zx}$ leaves only $\Sigma_1^{zx} = -\Sigma_2^{zx}$ and $\Sigma_1^{zx} = \Sigma_2^{zx}$

as possible solutions. The first can be eliminated as well, since this contradicts (A.2). This leaves as the only feasible solution $\Sigma_1^{zx} = \Sigma_2^{zx}$, which implies $\mu_1^z = \mu_2^z$ from (A.5).

It has been shown that if $k(y, \boldsymbol{x}; \mu_1^x, \mu_1^z, \Sigma_1^{xx}, \Sigma_1^{zx}) = k(y, \boldsymbol{x}; \mu_2^x, \mu_2^z, \Sigma_2^{xx}, \Sigma_2^{zx})$, then this implies $(\mu_1^x, \mu_1^z, \Sigma_1^{xx}, \Sigma_1^{zx}) = (\mu_2^x, \mu_2^z, \Sigma_2^{xx}, \Sigma_2^{zx})$. Therefore, applying directly the definition, the parameters $(\mu^x, \mu^z, \Sigma^{xx}, \Sigma^{zx})$ are identifiable in the kernel of the mixture.

A.1.2 Proof of Lemma 3

Here we show that the parameters in the kernel of the induced model for mixed ordinal-continuous $(\boldsymbol{y}, \boldsymbol{x})$ with fixed cut-offs are identifiable, as stated in Section 3.2.2. We use the definition of likelihood identifiability, setting

$$k(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1) = k(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2), \tag{A.6}$$

for arbitrary $(\boldsymbol{y}, \boldsymbol{x})$ such that $y_i \in \{1, \ldots, C_i\}$ and $x \in \mathbb{R}^p$. If this implies $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\Sigma_1 = \Sigma_2$, then the kernel is identifiable.

For observed $\boldsymbol{y}$ such that $y_j \in \{1, \ldots, C_j\}$ with $C_j > 2$, for all $j = 1, \ldots, k$, $k(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ is expressed as $\int_{\gamma_{k,y_k-1}}^{\gamma_{k,y_k}} \cdots \int_{\gamma_{1,y_1-1}}^{\gamma_{1,y_1}} \mathrm{N}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\mu}, \Sigma) dz_1 \ldots dz_k$. As a consequence of (A.6), we have that $\mathrm{N}(\boldsymbol{x}; \boldsymbol{\mu}_1^x, \Sigma_1^{xx}) = \mathrm{N}(\boldsymbol{x}; \boldsymbol{\mu}_2^x, \Sigma_2^{xx})$, for all $\boldsymbol{x} \in \mathbb{R}^p$, and therefore $\boldsymbol{\mu}_1^x = \boldsymbol{\mu}_2^x$, and $\Sigma_1^{xx} = \Sigma_2^{xx}$.

It also must be the case that for each $j = 1, \ldots, k$, $k(y_j \mid \boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1) = k(y_j \mid \boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)$, for any $y_j \in \{1, \ldots, C_j\}$. That is,

$$\Phi\left(\frac{\gamma_{j,l} - \mu_1^{z_j} - \Sigma_1^{z_j x}(\Sigma^{xx})^{-1}(\boldsymbol{x} - \mu^x)}{(\Sigma_1^{z_j z_j} - \Sigma_1^{z_j x}(\Sigma^{xx})^{-1}\Sigma_1^{x z_j})^{1/2}}\right) = \Phi\left(\frac{\gamma_{j,l} - \mu_2^{z_j} - \Sigma_2^{z_j x}(\Sigma^{xx})^{-1}(\boldsymbol{x} - \mu^x)}{(\Sigma_2^{z_j z_j} - \Sigma_2^{z_j x}(\Sigma^{xx})^{-1}\Sigma_2^{x z_j})^{1/2}}\right), \tag{A.7}$$

for $l = 1, \ldots, C_j - 1$.

For (A.7) to be true for any $\boldsymbol{x}$, it must be that

$$\frac{\Sigma_1^{z_j x}}{(\Sigma_1^{z_j z_j} - \Sigma_1^{z_j x}(\Sigma^{xx})^{-1}\Sigma_1^{x z_j})^{1/2}} = \frac{\Sigma_2^{z_j x}}{(\Sigma_2^{z_j z_j} - \Sigma_2^{z_j x}(\Sigma^{xx})^{-1}\Sigma_2^{x z_j})^{1/2}}, \tag{A.8}$$

and

$$\frac{\gamma_{j,l} - \mu_1^{z_j}}{(\Sigma_1^{z_j z_j} - \Sigma_1^{z_j x}(\Sigma^{xx})^{-1}\Sigma_1^{x z_j})^{1/2}} = \frac{\gamma_{j,l} - \mu_2^{z_j}}{(\Sigma_2^{z_j z_j} - \Sigma_2^{z_j x}(\Sigma^{xx})^{-1}\Sigma_2^{x z_j})^{1/2}}, \tag{A.9}$$

for $l = 1, \ldots, C_j - 1$. Using (A.8), (A.9) becomes $(\gamma_{j,l} - \mu_1^{z_j})\Sigma_2^{z_j x} = (\gamma_{j,l} - \mu_2^{z_j})\Sigma_1^{z_j x}$, and working with 2 of these $C_j - 1$ equations, the system has the solution $\Sigma_1^{z_j x} = \Sigma_2^{z_j x}$. Then from (A.9), $\mu_1^{z_j} = \mu_2^{z_j}$, and from (A.8), $\Sigma_1^{z_j z_j} = \Sigma_2^{z_j z_j}$.

Notice that we required 2 of the $C_j - 1$ equations of the form $(A.9)$ to arrive at this solution. Therefore, if $C_j = 2$ for some j, we are unable to identify all free parameters $\Sigma^{z_j z_j}, \mu^{z_j}, \Sigma^{z_j x}$, which we have identified if $C_j > 2$. In this case, fix $\Sigma^{z_j z_j}$, and then μ^{z_j} and $\Sigma^{z_j x}$ are identifiable, as in Appendix A.1.1. Although we do not require free cut-offs here due to the flexibility provided by the mixture, if $C_j > 3$, the cut-offs $\gamma_{j,3}, \ldots, \gamma_{j,C_j-1}$ are also identifiable, if treated as parameters.

Finally, (A.6) implies $k(y_j, y_{j'}; \boldsymbol{\mu}_1, \Sigma_1) = k(y_j, y_{j'}; \boldsymbol{\mu}_2, \Sigma_2)$, for $j, j' \in \{1, \ldots, k\}$, $j \neq j'$. Because identifiability has already been established for all parameters except $\Sigma^{z_j z_{j'}}$, this implies that, for any $y_j \in \{1, \ldots, C_j\}$ and $y_{j'} \in \{1, \ldots, C_{j'}\}$,

$$\int_{\gamma_{j',y_{j'}-1}}^{\gamma_{j',y_{j'}}} \int_{\gamma_{j,y_j-1}}^{\gamma_{j,y_j}} \mathrm{N}\left((z_j, z_{j'})^T; (\mu^{z_j}, \mu^{z_{j'}})^T, \begin{pmatrix} \Sigma^{z_j z_j} & \Sigma_1^{z_i z_{j'}} \\ \Sigma_1^{z_i z_{j'}} & \Sigma^{z_{j'} z_{j'}} \end{pmatrix} \right) dz_j dz_{j'} =$$

$$\int_{\gamma_{j',y_{j'}-1}}^{\gamma_{j',y_{j'}}} \int_{\gamma_{j,y_j-1}}^{\gamma_{j,y_j}} \mathrm{N}\left((z_j, z_{j'})^T; (\mu^{z_j}, \mu^{z_{j'}})^T, \begin{pmatrix} \Sigma^{z_j z_j} & \Sigma_2^{z_j z_{j'}} \\ \Sigma_2^{z_j z_{j'}} & \Sigma^{z_{j'} z_{j'}} \end{pmatrix} \right) dz_j dz_{j'}. \tag{A.10}$$

We use the result that $\int_{-\infty}^{b}\int_{-\infty}^{a} \mathrm{N}((w_1,w_2)^T;(0,0)^T,V)dw_1dw_2$ is monotonically increasing in V_{12}, for constants a and b. This can be shown with the following:

$$\begin{aligned}
&\frac{\partial}{\partial V_{12}}\int_{-\infty}^{b}\int_{-\infty}^{a} \mathrm{N}((w_1,w_2)^T;(0,0)^T,V)dw_1dw_2 \\
&= \int_{-\infty}^{b}\int_{-\infty}^{a} \frac{\partial}{\partial V_{12}}\mathrm{N}((w_1,w_2)^T;(0,0)^T,V)dw_1dw_2 \\
&= \int_{-\infty}^{b}\int_{-\infty}^{a} \frac{\partial^2}{\partial w_1 \partial w_2}\mathrm{N}((w_1,w_2)^T;(0,0)^T,V)dw_1dw_2 \\
&= \frac{\partial^2}{\partial w_1 \partial w_2}\int_{-\infty}^{b}\int_{-\infty}^{a} \mathrm{N}((w_1,w_2)^T;(0,0)^T,V)dw_1dw_2 \\
&= N((a,b)^T;(0,0)^T,V) > 0.
\end{aligned}$$

This result implies that $k(Y_j = 1, Y_{j'} = 1; \boldsymbol{\mu}, \Sigma)$ is monotonically increasing in $\Sigma^{z_j z_{j'}}$ and therefore, $\Sigma_1^{z_j z_{j'}} = \Sigma_2^{z_j z_{j'}}$.

A.2 Distributions Implied by the Inverse-Wishart

Assume $\Sigma \sim \mathrm{IW}_r(v,T)$, with $r = p+1$, and partition Σ into blocks, Σ_{11}, Σ_{12}, Σ_{21}, and Σ_{22}, of dimensions $q \times q$, $q \times (r-q)$, $(r-q) \times q$, and $(r-q) \times (r-q)$, respectively. Moreover, consider the corresponding partition for matrix T. Then, applying propositions 8.7 and 8.8 of Eaton (2007), we obtain:

(a) $\Sigma_{11} \sim \mathrm{IW}_q(v-(r-q), T_{11})$.

(b) $\Sigma_{22\cdot 1} \sim \mathrm{IW}_{r-q}(v, T_{22\cdot 1})$, where $\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ and $T_{22\cdot 1} = T_{22} - T_{21}T_{11}^{-1}T_{12}$.

(c) $\Sigma_{11}^{-1}\Sigma_{12} \mid \Sigma_{22\cdot 1}^{-1} \sim \mathrm{MN}_{q,r-q}(T_{11}^{-1}T_{12}, T_{11}^{-1}, \Sigma_{22\cdot 1})$. Here, MN denotes the matrix normal distribution such that, conditionally on $\Sigma_{22\cdot 1}$, $\mathrm{vec}(\Sigma_{11}^{-1}\Sigma_{12}) \sim \mathrm{N}_{q(r-q)}(\mathrm{vec}(T_{11}^{-1}T_{12}), T_{11}^{-1} \otimes \Sigma_{22\cdot 1})$.

We now assume T is diagonal, with elements $(T_1, \ldots, T_{p+1})$, as this is the case relevant to our prior specification approach. Let $T^i = \text{diag}(T_1, \ldots, T_i)$. Applying result (b) with $q = p$, we obtain $\delta_{p+1} \sim \text{IG}(0.5v, 0.5T_{p+1})$. This uses the fact that $\Sigma_{22\cdot 1} = \delta_{p+1}$ as a consequence of the (β, Δ) parameterization, and the simplification of $T_{22\cdot 1}$ to $T_{22} = T_{p+1}$ when T is diagonal. Applying result (a) with $q = p$, we obtain the marginal distribution of the upper left p dimensional block of the covariance matrix Σ, which is $\Sigma_{1:p,1:p} \sim \text{IW}_p(v - 1, T^p)$. Next, using result (b) for matrix $\Sigma_{1:p,1:p}$ with $q = p - 1$, we have $\delta_p \sim \text{IG}(0.5(v-1), 0.5T_p)$, since $(\Sigma_{1:p,1:p})_{22\cdot 1} = \delta_p$. Analogously, applying results (a) and (b) in succession, we obtain $\delta_i \sim \text{IG}(0.5(v + i - (p+1)), 0.5T_i)$, for $i = 2, \ldots, p+1$.

For each $i = 2, \ldots, p+1$, result (a) yields an $\text{IW}_i(v + i - (p+1), T^i)$ distribution for $\Sigma_{1:i,1:i}$, that is, for the upper left block of Σ of dimension i. Then, applying result (c) to $\Sigma_{1:i,1:i}$ with $q = i - 1$, we obtain $(-\beta_{i,1}, \ldots, -\beta_{i,i-1})^T \mid \delta_i \sim \text{N}_{i-1}((0, \ldots, 0)^T, \delta_i (T^{i-1})^{-1})$, for $i = 2, \ldots, p+1$. This uses the fact that $(T^i)_{12} = (0, \ldots, 0)^T$, $\text{vec}((\Sigma_{1:i,1:i})_{11}^{-1}(\Sigma_{1:i,1:i})_{12}) = (-\beta_{i,1}, \ldots, -\beta_{i,i-1})^T$, and $(\Sigma_{1:i,1:i})_{22\cdot 1} = \delta_i$.

A.3 Proof of Lemma 4

We first show that there exists at least one $f_0(\boldsymbol{x}, \boldsymbol{z})$ for any $p_0(\boldsymbol{x}, \boldsymbol{y})$, as defined in (3.6). This is related to the example given by Canale and Dunson (2011) for modeling count data, in which $f_0(z)$ with univariate $z \in \mathbb{R}$ induces probability mass function $p_0(y)$. Let $\boldsymbol{s} = (s_1, \ldots, s_k)$, with each $s_j \in \{1, \ldots, C_j\}$, and define

$$f_0(\boldsymbol{x}, \boldsymbol{z}) = \sum_{\boldsymbol{s}} \frac{p_0(\boldsymbol{x}, s_1, \ldots, s_k) \prod_{j=1}^k 1_{(\gamma^*_{j,s_j-1}, \gamma^*_{j,s_j}]}(z_j)}{\prod_{j=1}^k (\gamma^*_{j,s_j} - \gamma^*_{j,s_j-1})}$$

where $\gamma^*_{j,l} = \gamma_{j,l}$ if $l \in \{1, \ldots, C_j - 1\}$, $\gamma^*_{j,0} = b_j$, and $\gamma^*_{j,C_j} = d_j$, with $-\infty < b_j < \gamma_{j,1}$ and $\gamma_{j,C_j-1} < d_j < \infty$, for $j = 1, \ldots, k$. Then this $f_0(\boldsymbol{x}, \boldsymbol{z})$ satisfies the relationship given in (3.6), inducing $p_0(\boldsymbol{x}, \boldsymbol{l})$ upon integration.

Now we prove the lemma. Let $\mathrm{KL}(f_0, f)$ be the KL distance between f_0 and f. The chain rule for relative entropy states that

$$\mathrm{KL}(f_0(\boldsymbol{x}, \boldsymbol{z}), f(\boldsymbol{x}, \boldsymbol{z})) = \mathrm{KL}(f_0(\boldsymbol{x}), f(\boldsymbol{x})) + \mathrm{KL}(f_0(\boldsymbol{z} \mid \boldsymbol{x}), f(\boldsymbol{z} \mid \boldsymbol{x})), \tag{A.11}$$

and therefore

$$\int f_0(\boldsymbol{x}, \boldsymbol{z}) \log \left(f_0(\boldsymbol{x}, \boldsymbol{z}) / f(\boldsymbol{x}, \boldsymbol{z}) \right) d\boldsymbol{z} d\boldsymbol{x} \geq \int f_0(\boldsymbol{x}) \log \left(f_0(\boldsymbol{x}) / f(\boldsymbol{x}) \right) d\boldsymbol{x},$$

so that if $f(\boldsymbol{x}, \boldsymbol{z}) \in K_\epsilon(f_0(\boldsymbol{x}, \boldsymbol{z}))$, then $f(\boldsymbol{x}) \in K_\epsilon(f_0(\boldsymbol{x}))$. That is,

$$K_\epsilon(f_0(\boldsymbol{x}, \boldsymbol{z})) \subseteq K_\epsilon(f_0(\boldsymbol{x})) = \{f(\boldsymbol{x}, \boldsymbol{z}) : \mathrm{KL}(f_0(\boldsymbol{x}), f(\boldsymbol{x})) < \epsilon\}.$$

Using the KL property of the prior model for $(\boldsymbol{x}, \boldsymbol{z})$, $\mathcal{P}\{K_\epsilon(f_0(\boldsymbol{x}))\} \geq \mathcal{P}\{K_\epsilon(f_0(\boldsymbol{x}, \boldsymbol{z}))\} > 0$, so that the prior $\mathcal{P}$ assigns positive probability to all KL neighborhoods of the true marginal covariate distribution $f_0(\boldsymbol{x})$.

Now, take $f \in K_{\epsilon/2}(f_0(\boldsymbol{x}, \boldsymbol{z}))$. By the chain rule, $\mathrm{KL}(f_0(\boldsymbol{x}), f(\boldsymbol{x})) < \epsilon/2$, and $\mathrm{KL}(f_0(\boldsymbol{z} \mid \boldsymbol{x}), f(\boldsymbol{z} \mid \boldsymbol{x})) < \epsilon/2$. We now use the result that for two distributions $g_1(\boldsymbol{t})$ and $g_2(\boldsymbol{t})$, with $\boldsymbol{t} = (t_1, \ldots, t_s)$,

$$\int_{A_s} \cdots \int_{A_1} g_1(\boldsymbol{t}) \log \left(\frac{g_1(\boldsymbol{t})}{g_2(\boldsymbol{t})} \right) d\boldsymbol{t} \geq \int_{A_s} \cdots \int_{A_1} g_1(\boldsymbol{t}) d\boldsymbol{t} \times \log \left(\frac{\int_{A_s} \cdots \int_{A_1} g_1(\boldsymbol{t}) d\boldsymbol{t}}{\int_{A_s} \cdots \int_{A_1} g_2(\boldsymbol{t}) d\boldsymbol{t}} \right), \tag{A.12}$$

Applying this result with $g_1(\boldsymbol{t}) = f_0(\boldsymbol{z} \mid \boldsymbol{x})$, $g_2(\boldsymbol{t}) = f(\boldsymbol{z} \mid \boldsymbol{x})$, and $A_j = (\gamma_{j,l_j-1}, \gamma_{j,l_j})$, for $j = 1, \ldots, k$, then the right hand side of the equation becomes $p_0(\boldsymbol{l} \mid \boldsymbol{x}) \log(p_0(\boldsymbol{l} \mid \boldsymbol{x}) / p^*(\boldsymbol{l} \mid \boldsymbol{x}))$, with $p^*(\boldsymbol{l} \mid \boldsymbol{x}) = \int_{\gamma_{k,l_k-1}}^{\gamma_{k,l_k}} \cdots \int_{\gamma_{1,l_1-1}}^{\gamma_{1,l_1}} f(\boldsymbol{z} \mid \boldsymbol{x}) d\boldsymbol{x}$. Now, summing each side of (A.12) over

$l_j = 1, \ldots, C_j$, and $j = 1, \ldots, k$, and multiplying by $\int_{\mathbb{R}} f_0(\boldsymbol{x})d\boldsymbol{x}$, we have

$$\int_{\mathbb{R}} f_0(\boldsymbol{x}) \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_0(\boldsymbol{z} \mid \boldsymbol{x}) \log\left(\frac{f_0(\boldsymbol{z} \mid \boldsymbol{x})}{f(\boldsymbol{z} \mid \boldsymbol{x})}\right) d\boldsymbol{z}d\boldsymbol{x} \geq \int_{\mathbb{R}} f_0(\boldsymbol{x}) \sum_{l_k=1}^{C_k} \cdots \sum_{l_1=1}^{C_1} p_0(\boldsymbol{l} \mid \boldsymbol{x}) \log\left(\frac{p_0(\boldsymbol{l} \mid \boldsymbol{x})}{p^*(\boldsymbol{l} \mid \boldsymbol{x})}\right) d\boldsymbol{x}.$$

The left side of the equation is by definition $\mathrm{KL}(f_0(\boldsymbol{z} \mid \boldsymbol{x}), f(\boldsymbol{z} \mid \boldsymbol{x}))$, which is less than $\epsilon/2$, and the right side is $\mathrm{KL}(p_0(\boldsymbol{y} \mid \boldsymbol{x}), p^*(\boldsymbol{y} \mid \boldsymbol{x}))$, which also must be less than $\epsilon/2$. This implies $\mathrm{KL}(p_0(\boldsymbol{x}, \boldsymbol{y}), p^*(\boldsymbol{x}, \boldsymbol{y})) < \epsilon$, by (A.11). Defining $K_\epsilon(p_0(\boldsymbol{x}, \boldsymbol{y})) = \{f(\boldsymbol{x}, \boldsymbol{z}) : \mathrm{KL}(p_0(\boldsymbol{x}, \boldsymbol{y}), p^*(\boldsymbol{x}, \boldsymbol{y})) < \epsilon\}$, we have $K_{\epsilon/2}(f_0(\boldsymbol{x}, \boldsymbol{z})) \subseteq K_{\epsilon/2}(p_0(\boldsymbol{y} \mid \boldsymbol{x}))$ and $K_{\epsilon/2}(f_0(\boldsymbol{x}, \boldsymbol{z})) \subseteq K_\epsilon(p_0(\boldsymbol{x}, \boldsymbol{y}))$, implying $\mathcal{P}\{K_\epsilon(p_0(\boldsymbol{x}, \boldsymbol{y}))\} \geq \mathcal{P}\{K_{\epsilon/2}(f_0(\boldsymbol{x}, \boldsymbol{z}))\} > 0$ and $\mathcal{P}\{K_{\epsilon/2}(p_0(\boldsymbol{y} \mid \boldsymbol{x}))\} \geq \mathcal{P}\{K_{\epsilon/2}(f_0(\boldsymbol{x}, \boldsymbol{z}))\} > 0$.

Appendix B

Posterior Simulation Details

B.1 Proof of Lemma 2

Here, we provide details for Lemma 2 of Section 2.2.2. Consider $\boldsymbol{y} = (y_1, ..., y_r) \mid \boldsymbol{\mu}, \beta, \Delta \sim \mathrm{N}_r(\boldsymbol{\mu}, \beta^{-1}\Delta(\beta^{-1})^T)$, such that the likelihood for β is proportional to $\exp\{-(\boldsymbol{y} - \boldsymbol{\mu})^T\beta^T\Delta^{-1}\beta(\boldsymbol{y} - \boldsymbol{\mu})\}$. First, focus on determining the likelihood for $\tilde{\boldsymbol{\beta}}$, a vector of length $q = r(r-1)/2$. Write $\beta(\boldsymbol{y} - \boldsymbol{\mu})$ as $M(1, \tilde{\boldsymbol{\beta}}^T)^T$, for a matrix M, of dimension $r \times (q+1)$ which has row i containing i nonzero elements, the first being $(y_i - \mu_i)$, occurring in column 1, and the rest being $(y_1 - \mu_1), ..., (y_{i-1} - \mu_{i-1})$, occurring in columns $2 + (i-1)(i-2)/2$ to $i + (i-1)(i-2)/2$. Then, the likelihood for $\tilde{\boldsymbol{\beta}}$ can be written proportional to $\exp\{-(1, \tilde{\boldsymbol{\beta}}^T)M^T\Delta^{-1}M(1, \tilde{\boldsymbol{\beta}}^T)^T\}$. Let $C = M^T\Delta^{-1}M$. If there exists a symmetric, positive definite matrix T and vector $\boldsymbol{d}$ for which $(1, \tilde{\boldsymbol{\beta}}^T)C(1, \tilde{\boldsymbol{\beta}}^T)^T = \tilde{\boldsymbol{\beta}}^T T\tilde{\boldsymbol{\beta}} - 2\tilde{\boldsymbol{\beta}}^T T\boldsymbol{d} + R$, where R is a constant that does not depend on $\tilde{\boldsymbol{\beta}}$, then the likelihood for $\tilde{\boldsymbol{\beta}}$ corresponds to a normal distribution with mean vector $\boldsymbol{d}$ and covariance matrix T^{-1}. The left side of the above equation is $C_{11} + 2\sum_{j=2}^{q+1}\tilde{\beta}_{j-1}C_{1j} + \sum_{j=2}^{q+1}\sum_{i=2}^{q+1}\tilde{\beta}_{j-1}\tilde{\beta}_{i-1}C_{ij}$, and the last of these terms is

just $\tilde{\boldsymbol{\beta}}^T C_{q\times q}\tilde{\boldsymbol{\beta}}$, where $C_{q\times q}$ denotes the $q\times q$ submatrix of C obtained by deleting the first row and column of C. Therefore, with $T = C_{q\times q}$, we seek $\boldsymbol{d}$ such that $-\tilde{\boldsymbol{\beta}}^T T\boldsymbol{d} = \sum_{j=2}^{q+1}\tilde{\beta}_{j-1}C_{1j}$. Equating the coefficient associated with $\tilde{\beta}_i$, $i = 1, ..., q$, on each side of the equation results in a system of q equations:

$$-\sum_{j=1}^{q} d_j T_{i-1,j} = C_{1i}, \quad i = 2, ..., q+1. \tag{B.1}$$

As explained in Section 2.2, T is a block diagonal matrix which can be constructed from square matrices $T^1, ..., T^{r-1}$, of dimensions $1, ..., r-1$, where

$$T^j_{mn} = (y_m - \mu_m)(y_n - \mu_n)/\delta_{j+1}, \quad m = 1, ..., j, \quad n = 1, ..., j. \tag{B.2}$$

The symmetry of T follows from the symmetry of C, but it remains to be shown that T is positive definite. For a non-zero vector $\boldsymbol{v}$, we must have $\boldsymbol{v}^T T\boldsymbol{v} > 0$. When $r = 2$, $\boldsymbol{v}^T T\boldsymbol{v}$ becomes $v_1^2(y_1-\mu_1)^2/\delta_2$. When $r = 3$, $\boldsymbol{v}^T T\boldsymbol{v}$ is the sum of the result for $r = 2$ and the term $(v_2(y_1-\mu_1)+v_3(y_2-\mu_2))^2/\delta_3$. For $r = 4$, the term $(v_4(y_1-\mu_1)+v_5(y_2-\mu_2)+v_6(y_3-\mu_3))^2/\delta_4$ is added to the result for $r = 3$. In general, a term of the form $(v_{q-r+2}(y_1 - \mu_1) + ... + v_q(y_{r-1} - \mu_{r-1}))^2/\delta_r$ is added in going from $r-1$ to r dimensions. Clearly, T is positive semidefinite. However, to have $\boldsymbol{v}^T T\boldsymbol{v} > 0$, and all elements of T strictly positive, it must be the case that $y_i \neq \mu_i$, for $i = 1, ..., r-1$, which holds true with probability 1, since $\boldsymbol{\mu}$ is a continuous random vector.

We now derive the form of the mean vector $\boldsymbol{d}$. Because T is sparse, the system of q equations (B.1) can be divided into $r-1$ sets of equations, where set j consists of j equations with j unknowns, $d_{1+j(j-1)/2}, ..., d_{j(j+1)/2}$. Let the index $1+j(j-1)/2$ be denoted by (1) and let the index $j(j+1)/2$ be denoted by (j). Set the first $j-1$ of these elements

equal to 0, so that $d_{1+j(j-1)/2} = ... = d_{j(j+1)/2-1} = 0$. Then the j equations become

$$-d_{(j)}T_{(1),(j)} = C_{1,(1)+1}, ..., -d_{(j)}T_{(j),(j)} = C_{1,(j)+1}. \tag{B.3}$$

The solution $d_{(j)} = -(y_{j+1} - \mu_{j+1})/(y_j - \mu_j)$ satisfies these j equalities (B.3), since the elements $C_{1,(1)+1}, ..., C_{1,(j)+1}$ are $(y_1-\mu_1)(y_{j+1}-\mu_{j+1})/\delta_{j+1}, ..., (y_j-\mu_j)(y_{j+1}-\mu_{j+1})/\delta_{j+1}$, and the elements $T_{(1),(j)}, ..., T_{(j),(j)}$ are $(y_1 - \mu_1)(y_j - \mu_j)/\delta_{j+1}, ..., (y_j - \mu_j)(y_j - \mu_j)/\delta_{j+1}$, as given in (B.2), so that

$$-C_{1,(1)+1}/T_{(1),(j)} = ... = -C_{1,(j)+1}/T_{(j),(j)} = -(y_{j+1} - \mu_{j+1})/(y_j - \mu_j).$$

With n data vectors, $(y_{i,1}, ..., y_{i,r})$, for $i = 1, ..., n$, the likelihood for $\tilde{\boldsymbol{\beta}}$ is proportional to a normal with mean $(\sum_{i=1}^n T_i)^{-1}(\sum_{i=1}^n T_i \boldsymbol{d}_i)$, and covariance matrix $(\sum_{i=1}^n T_i)^{-1}$, where T_i and $\boldsymbol{d}_i$ are computed using the i-th observation. When combined with a normal prior for $\tilde{\boldsymbol{\beta}}$, the full conditional is also normal.

Next, consider the likelihood for the δ_k, which up to the proportionality constant is given by $\prod_{k=1}^r \delta_k^{-1/2} \exp\{-\text{tr}(\beta^T \Delta^{-1} \beta (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T)/2\}$. By properties of trace, $\text{tr}(\beta^T \Delta^{-1} \beta (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T) = \text{tr}(\beta (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T \beta^T \Delta^{-1})$. Let $A = \beta (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T \beta^T$. Since Δ is diagonal with $\boldsymbol{\delta}$ on the diagonal, the likelihood for each δ_k is proportional to $\delta_k^{-1/2} \exp\{-A_{kk}/(2\delta_k)\}$. The diagonal elements of A are the squares of $\beta(\boldsymbol{y} - \boldsymbol{\mu})$, which are $A_{kk} = \{(y_k - \mu_k) + \sum_{j<k} \beta_{kj}(y_j - \mu_j)\}^2$. Then, with n data vectors, $(y_{i,1}, ..., y_{i,r})$, $i = 1, ..., n$, the likelihood for δ_k, $k = 1, ..., r$, is proportional to an inverse-gamma with shape parameter $(n/2) - 1$ and scale parameter $0.5 \sum_{i=1}^n \{(y_{i,k} - \mu_k) + \sum_{j<k} \beta_{kj}(y_{i,j} - \mu_j)\}^2$. When combined with an inverse-gamma prior, this results in a posterior full conditional distribution which is inverse-gamma.

B.2 Model Comparison Criterion

The predictive loss measure used for model comparison in Section 2.3.2 requires for each model m the posterior predictive mean, $\mathrm{E}^{(m)}(y_{new,i} \mid \text{data})$, and posterior predictive variance, $\mathrm{var}^{(m)}(y_{new,i} \mid \text{data})$, for replicated response $y_{new,i}$ with associated covariate vector $\boldsymbol{x}_i$.

Denote generically by Θ the full parameter vector for either the product-kernel model or for the more general binary regression model developed in Section 2. For the former model, $\mathrm{E}(y \mid \boldsymbol{x}_i; \text{data}) = \{p(\boldsymbol{x}_i \mid \text{data})\}^{-1} \int \sum_{l=1}^{N} p_l \mathrm{N}_p(\boldsymbol{x}_i; \mu_l^x, \Sigma_l^{xx}) \Phi(\mu_l^z)\, p(\Theta \mid \text{data}) \mathrm{d}\Theta$, with $p(\boldsymbol{x}_i \mid \text{data}) = \int \sum_{l=1}^{N} p_l \mathrm{N}_p(\boldsymbol{x}_i; \mu_l^x, \Sigma_l^{xx})\, p(\Theta \mid \text{data}) \mathrm{d}\Theta$, and $\mathrm{E}(y^2 \mid \boldsymbol{x}_i; \text{data})$ also has the same form. Under the proposed model, $\mathrm{E}(y \mid \boldsymbol{x}_i; \text{data})$ is given by

$$\{p(\boldsymbol{x}_i \mid \text{data})\}^{-1} \int \sum_{l=1}^{N} p_l \mathrm{N}_p(\boldsymbol{x}_i; \mu_l^x, \Sigma_l^{xx}) \Phi\left(\frac{\mu_l^z + \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(\boldsymbol{x}_i - \mu_l^x)}{(\Sigma_l^{zz} - \Sigma_l^{zx}(\Sigma_l^{xx})^{-1}(\Sigma_l^{zx})^T)^{1/2}} \right) p(\Theta \mid \text{data}) \mathrm{d}\Theta$$

where $p(\boldsymbol{x}_i \mid \text{data}) = \int \sum_{l=1}^{N} p_l \mathrm{N}_p(\boldsymbol{x}_i; \mu_l^x, \Sigma_l^{xx})\, p(\Theta \mid \text{data}) \mathrm{d}\Theta$, and $\mathrm{E}(y \mid \boldsymbol{x}_i; \text{data}) = \mathrm{E}(y^2 \mid \boldsymbol{x}_i; \text{data})$. Hence, under both models, straightforward Monte Carlo integration using the posterior samples for model parameters yields estimates for the required posterior predictive means and variances.

B.3 MCMC Details for DDP Model

B.3.1 Updates for Constant Atoms DDP Model

Here we give details for posterior full conditionals required to implement the model in Section 4.3.1.

Updating the Weights

The full conditional for $(\{\zeta_l\}, \{\eta_{l,t}\})$ is given by $p(\{\zeta_l\}, \{\eta_{l,t}\} \mid \ldots, \text{data}) \propto$

$$\prod_{l=1}^{N-1} \mathrm{N}(\zeta_l; 0, 1)\mathrm{N}(\eta_{l,1}; 0, 1) \prod_{t=2}^{T} \prod_{l=1}^{N-1} \mathrm{N}(\eta_{l,t}; \phi\eta_{l,t-1}, 1-\phi^2) \prod_{t=1}^{T} \prod_{i=1}^{n_t} \sum_{l=1}^{N} p_{l,t}\delta_l(L_{t,i}).$$

Write $\prod_{i=1}^{n_t} \sum_{l=1}^{N} p_{l,t}\delta_l(L_{t,i}) = \prod_{l=1}^{N} p_{l,t}^{M_{l,t}}$, where $M_{l,t} = \mid \{(t,i) : L_{t,i} = l\} \mid$, i.e., the number of observations at time t assigned to component l. Filling in the form for $\{p_{l,t}\}$ gives

$$\prod_{i=1}^{n_t} \sum_{l=1}^{N} p_{l,t}\delta_l(L_{t,i}) = \left(1 - \exp\left(-\frac{\zeta_1^2 + \eta_{1,t}^2}{2\alpha}\right)\right)^{M_{1,t}} \exp\left(-\frac{M_{N,t}\sum_{l=1}^{N-1}(\zeta_l^2 + \eta_{l,t}^2)}{2\alpha}\right)$$
$$\prod_{l=2}^{N-1} \left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}} \exp\left(-\frac{M_{l,t}\sum_{r=1}^{l-1}(\zeta_r^2 + \eta_{r,t}^2)}{2\alpha}\right).$$

The full conditional for each ζ_l, $l = 1, \ldots, N-1$, is therefore

$$p(\zeta_l \mid \ldots, \text{data}) \propto \exp\left(-\frac{\zeta_l^2}{2}\right) \exp\left(-\frac{\sum_{t=1}^{T}\sum_{r=l+1}^{N} M_{r,t}}{2\alpha}\right) \prod_{t=1}^{T} \left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}}$$

giving

$$p(\zeta_l \mid \ldots, \text{data}) \propto \mathrm{N}(\zeta_l; 0, (1 + \alpha^{-1}\sum_{t=1}^{T}\sum_{r=l+1}^{N} M_{r,t})^{-1}) \prod_{t=1}^{T} \left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}}$$

We use a slice sampler to update ζ_l, with the following steps:

- Draw $u_t \sim \text{uniform}\left(0, \left(1 - \exp\left(-\frac{\zeta_l^2+\eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}}\right)$, for $t = 1, \ldots, T$.
- Draw $z_l \sim \mathrm{N}(0, (1 + \alpha^{-1}\sum_{t=1}^{T}\sum_{r=l+1}^{N} M_{r,t})^{-1})$, restricted to the lie in the interval $\{\zeta_l : u_t < \left(1 - \exp\left(-\frac{\zeta_l^2+\eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}}, t = 1, \ldots, T\}$. Solving for ζ_l in each of these T equations gives $\zeta_l^2 > -\eta_{l,t}^2 - 2\alpha\log(1 - u_t^{1/M_{l,t}})$, for $t = 1, \ldots, T$. Therefore, if $-\eta_{l,t}^2 - 2\alpha\log(1 - u_t^{1/M_{l,t}}) < 0$ for all t, then ζ_l has no restrictions, and is therefore sampled from a normal distribution. Otherwise, if $-\eta_{l,t}^2 - 2\alpha\log(1 - u_t^{1/M_{l,t}}) > 0$ for some

t, then $|\ \zeta_l\ | > \max_t\{(-\eta_{l,t}^2 - 2\alpha\log(1-u_t^{1/M_{l,t}}))^{1/2}\}$. This then requires sampling ζ_l from a normal distribution, restricted to the intervals $(-\infty, -\max_t\{(-\eta_{l,t}^2 - 2\alpha\log(1-u_t^{1/M_{l,t}}))^{1/2}\})$, and $(\max_t\{(-\eta_{l,t}^2 - 2\alpha\log(1-u_t^{1/M_{l,t}}))^{1/2}\}, \infty)$.

In the second step above, we may have to sample from a normal distribution, restricted to two disjoint intervals. The resulting distribution is therefore a mixture of two truncated normals, with probabilities determined by the (normalized) probability the normal assigns to each interval. These truncated normals both have mean 0 and variance $(1 + \sum_{t=1}^{T}\sum_{r=l+1}^{N} M_{r,t}/\alpha)^{-1}$, and each mixture component has equal probability.

The full conditional for each $\eta_{l,t}$, $l = 1, \ldots, N-1$, $t = 2, \ldots, T-1$, is proportional to

$$\mathrm{N}(\eta_{l,t}; 0, \frac{\alpha}{\sum_{r=l+1}^{N} M_{r,t}})\mathrm{N}(\eta_{l,t}; \phi\eta_{l,t-1}, 1-\phi^2)\mathrm{N}(\eta_{l,t+1}; \phi\eta_{l,t}, 1-\phi^2)\left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}}$$

$$\propto \mathrm{N}\left(\eta_{l,t}; \frac{\phi\alpha(\eta_{l,t-1} + \eta_{l,t+1})}{\phi^2(\alpha - \sum_{r=l+1}^{N} M_{r,t}) + \alpha + \sum_{r=l+1}^{N} M_{r,t}}, \frac{\alpha(1-\phi^2)}{\phi^2(\alpha - \sum_{r=l+1}^{N} M_{r,t}) + \alpha + \sum_{r=l+1}^{N} M_{r,t}}\right)$$
$$\left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}}$$

Each $\eta_{l,t}$, $l = 1, \ldots, N-1$, and $t = 2, \ldots, T-1$, can therefore be sampled with a slice sampler:

- Draw $u \sim \mathrm{Unif}\left(0, \left(1 - \exp\left(-\frac{\zeta_l^2+\eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}}\right)$.
- Draw $\eta_{l,t} \sim \mathrm{N}\left(\eta_{l,t}; \frac{\phi\alpha(\eta_{l,t-1}+\eta_{l,t+1})}{\phi^2(\alpha-\sum_{r=l+1}^{N} M_{r,t})+\alpha+\sum_{r=l+1}^{N} M_{r,t}}, \frac{\alpha(1-\phi^2)}{\phi^2(\alpha-\sum_{r=l+1}^{N} M_{r,t})+\alpha+\sum_{r=l+1}^{N} M_{r,t}}\right)$, restricted to $\left\{\eta_{l,t} : \left(1 - \exp\left(-\frac{\zeta_l^2+\eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}} > u\right\}$, giving $\eta_{l,t}^2 > -2\alpha\log(1-u^{1/M_{l,t}}) - z_l^2$.

In the second step above, we will again either sample from a single normal or a mixture of truncated normals, where each normal has the same mean and variance, but the truncation intervals differ. Since the mean of this normal is not zero, the weights assigned to each truncated normal are not the same. The unnormalized weight assigned to the normal which places positive probability on $((-2\alpha \log(1 - u^{1/M_{l,t}}) - \zeta_l^2)^{1/2}, \infty)$ is given by $1 - F((-2\alpha \log(1 - u^{1/M_{l,t}}) - \zeta_l^2)^{1/2})$, where F is the CDF of the normal for $\eta_{l,t}$ given in the second step. The unnormalized weight given to the component which places positive probability on $(-\infty, -(-2\alpha \log(1-u^{1/M_{l,t}})-\zeta_l^2)^{1/2})$ is given by $F(-(-2\alpha \log(1-u^{1/M_{l,t}})-\zeta_l^2)^{1/2})$.

The full conditionals for $\eta_{l,1}$ and $\eta_{l,T}$ are slightly different. The full conditional for $\eta_{l,1}$ is

$$p(\eta_{l,1} \mid \ldots, \text{data}) \propto \mathrm{N}(\eta_{l,1}; 0, \frac{\alpha}{\sum_{r=l+1}^{N} M_{r,1}}) \mathrm{N}(\eta_{l,1}; 0, 1) \mathrm{N}(\eta_{l,2}; \phi\eta_{l,1}, 1-\phi^2) \left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,1}^2}{2\alpha}\right)\right)^{M_{l,1}},$$

$$\propto \mathrm{N}\left(\eta_{l,1}; \frac{\phi\alpha\eta_{l,2}}{\alpha + \sum_{r=l+1}^{N} M_{r,1} - \phi^2 \sum_{r=l+1}^{N} M_{r,1}}, \frac{\alpha(1-\phi^2)}{\alpha + \sum_{r=l+1}^{N} M_{r,1} - \phi^2 \sum_{r=l+1}^{N} M_{r,1}}\right) \left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,1}^2}{2\alpha}\right)\right)^{M_{l,1}}.$$

For $\eta_{l,T}$, we have:

$$p(\eta_{l,T} \mid \ldots, \text{data}) \propto \mathrm{N}(\eta_{l,T}; 0, \frac{\alpha}{\sum_{r=l+1}^{N} M_{r,T}}) \mathrm{N}(\eta_{l,T}; \phi\eta_{l,T-1}, 1-\phi^2) \left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,T}^2}{2\alpha}\right)\right)^{M_{l,T}},$$

which is proportional to

$$\mathrm{N}\left(\eta_{l,T}; \frac{\phi\alpha\eta_{l,T-1}}{\alpha + \sum_{r=l+1}^{N} M_{r,T} - \phi^2 \sum_{r=l+1}^{N} M_{r,T}}, \frac{\alpha(1-\phi^2)}{\alpha + \sum_{r=l+1}^{N} M_{r,T} - \phi^2 \sum_{r=l+1}^{N} M_{r,T}}\right) \left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,T}^2}{2\alpha}\right)\right)^{M_{l,T}}$$

The slice samplers for $\eta_{l,1}$ and $\eta_{l,T}$ are therefore implemented in the same way as for $\eta_{l,t}$, except the normals which are sampled from have different means and variances.

Updating α

The full conditional for α is

$$p(\alpha \mid \ldots, \text{data}) \propto p(\alpha) \exp\left(-\frac{\sum_{t=1}^{T} M_{N,t} \sum_{l=1}^{N-1}(\zeta_l^2 + \eta_{l,t}^2)}{2\alpha}\right) \exp\left(-\frac{\sum_{t=1}^{T}\sum_{l=2}^{N-1} M_{l,t} \sum_{r=1}^{l-1}(\zeta_r^2 + \eta_{r,t}^2)}{2\alpha}\right)$$
$$\prod_{t=1}^{T}\prod_{l=1}^{N-1}\left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}}$$

Therefore, with $p(\alpha) = \text{IG}(a_\alpha, b_\alpha)$, we have

$$p(\alpha \mid \ldots, \text{data}) = \text{IG}\left(\alpha; a_\alpha, b_\alpha + \frac{1}{2}\sum_{t=1}^{T}\left(M_{N,t}\sum_{l=1}^{N-1}(\zeta_l^2 + \eta_{l,t}^2) + \sum_{l=2}^{N-1} M_{l,t}\sum_{r=1}^{l-1}(\zeta_r^2 + \eta_{r,t}^2)\right)\right)$$
$$\prod_{t=1}^{T}\prod_{l=1}^{N-1}\left(1 - \exp\left(-\frac{\zeta_l^2 + \eta_{l,t}^2}{2\alpha}\right)\right)^{M_{l,t}}$$

The parameter α can be sampled using a Metropolis-Hastings algorithm. In particular we work with $\log(\alpha)$, and use a normal proposal distribution centered at the log of the current value of α.

Updating ϕ

The full conditional for AR parameter ϕ is

$$p(\phi \mid \ldots, \text{data}) \propto p(\phi) \prod_{t=2}^{T}\prod_{l=1}^{N-1} \text{N}(\eta_{l,t}; \phi\eta_{l,t-1}, 1-\phi^2)$$
$$\propto (1-\phi^2)^{-(N-1)(T-1)/2} \exp\left(-\sum_{t=2}^{T}\sum_{l=1}^{N-1}\frac{1}{2(1-\phi^2)}(\eta_{l,t} - \phi\eta_{l,t-1})^2\right) p(\phi)$$

We assume $p(\phi) = \text{uniform}(0,1)$ or $p(\phi) = \text{uniform}(-1,1)$, and apply a M-H algorithm to sample $\log\left(\frac{\phi}{1-\phi}\right)$ or $\log\left(\frac{\phi+1}{1-\phi}\right)$, respectively, using a normal proposal distribution.

Updating Remaining Parameters

The configuration variables $L_{t,i}$, $t = 1, \ldots, T$, $i = 1, \ldots, n_t$, have full conditionals given by

$$p(L_{t,i} \mid \ldots, \text{data}) \propto \mathrm{N}(\boldsymbol{y}_{t,i} \mid \boldsymbol{\mu}_{L_{t,i}}, \Sigma_{L_{t,i}})(\sum_{l=1}^{N-1} p_{l,t} \prod_{r=1}^{l-1} p_{r,t}\delta_l(L_{t,i}) + p_{N,t}\delta_N(L_{t,i}))$$

where $p_{1,t} = 1-\beta_{1,t}$, $p_{l,t} = (1-\beta_{l,t}) \prod_{r=1}^{l-1} \beta_{r,t}$, for $l = 2, \ldots, N-1$, and $p_{N,t} = 1-\sum_{l=1}^{N-1} p_{l,t}$. Therefore, $L_{t,i}$ is drawn from the discrete distribution on $\{1, \ldots, N\}$, with probabilities $p_{l,ti} \propto p_{l,t}\mathrm{N}(\boldsymbol{y}_{t,i} \mid \boldsymbol{\mu}_l, \Sigma_l)$ for $l = 1, \ldots, N$.

The parameters $(\{\boldsymbol{\mu}_l\}, \{\Sigma_l\})$ of the normal kernel have full conditionals:

$$p(\boldsymbol{\mu}_l, \Sigma_l \mid \ldots, \text{data}) \propto \mathrm{N}(\boldsymbol{\mu}_l; \boldsymbol{m}, V)\mathrm{IW}(\Sigma_l; \nu, D) \prod_{k=1}^{n^*} \prod_{\{(t,i):L_{ti}=L_k^*\}} N(\boldsymbol{y}_{t,i}; \boldsymbol{\mu}_{L_k^*}, \Sigma_{L_k^*})$$

where L_k^*, $k = 1, \ldots, n^*$ are the distinct values of $\boldsymbol{L}$. Therefore, for $l \notin \{L_k^*\}$, we simulate $\boldsymbol{\mu}_l \sim \mathrm{N}(\boldsymbol{m}, V)$, and $\Sigma_l \sim \mathrm{IW}(\nu, D)$. And, for $l \in \{L_k^*\}$, with $M_k^* = \mid \{(t,i) : L_{ti} = L_k^*\} \mid$, $\boldsymbol{\mu}_l \sim N(a_{\boldsymbol{\mu}}, B_{\boldsymbol{\mu}})$, with

$$B_{\boldsymbol{\mu}} = (V^{-1} + M_k^*(\Sigma_{L_k^*})^{-1})^{-1}, \quad a_{\boldsymbol{\mu}} = B_{\boldsymbol{\mu}}(\boldsymbol{m}V^{-1} + \sum_{\{(t,i):L_{ti}=L_k^*\}} \boldsymbol{y}_{t,i}(\Sigma_{L_k^*})^{-1}),$$

and $\Sigma_l \sim \mathrm{IW}(\nu + M_k^*, D + \sum_{\{(t,i):L_{ti}=L_k^*\}}(\boldsymbol{y}_{t,i} - \boldsymbol{\mu}_l)(\boldsymbol{y}_{t,i} - \boldsymbol{\mu}_l)^T)$.

The updates for $\boldsymbol{\psi} = (\boldsymbol{m}, V, D)$ are straightforward. Assuming a prior $\boldsymbol{m} \sim \mathrm{N}(\boldsymbol{a_m}, B_{\boldsymbol{m}})$, then the posterior full conditional for $\boldsymbol{m}$ is $p(\boldsymbol{m} \mid \ldots, \text{data}) \propto \mathrm{N}(\boldsymbol{a_m}^*, B_{\boldsymbol{m}}^*)$, with $B_{\boldsymbol{m}}^* = (B_{\boldsymbol{m}}^{-1} + NV^{-1})^{-1}$ and $\boldsymbol{a_m}^* = B_{\boldsymbol{m}}^*(\boldsymbol{a_m}B_{\boldsymbol{m}}^{-1} + V^{-1}\sum_{l=1}^{N} \boldsymbol{\mu}_l)$.

Assuming a prior $V \sim \mathrm{IW}(a_V, B_V)$, we sample $(V \mid \ldots, \text{data}) \sim \mathrm{IW}(a_V + N, B_V + \sum_{l=1}^{L}(\boldsymbol{\mu}_l - \boldsymbol{m})(\boldsymbol{\mu}_l - \boldsymbol{m})^T)$.

Finally, letting $D \sim \mathrm{W}(a_D, B_D)$, we sample $(D \mid \ldots, \text{data}) \sim \mathrm{W}(a_D + \nu N, (B_D^{-1} + \sum_{l=1}^{N}(\Sigma_l)^{-1})^{-1})$.

B.3.2 Updates for General DDP Model

Here, posterior full conditionals required for the model of Section 4.3.2 are given.

The parameter $\boldsymbol{m}$, when given a prior $\mathrm{N}(\boldsymbol{a_m}, B_{\boldsymbol{m}})$ has full conditional

$$p(\boldsymbol{m} \mid \ldots, \text{data}) \propto \mathrm{N}(\boldsymbol{m}; \boldsymbol{a_m}, B_{\boldsymbol{m}}) \prod_{l=1}^{N} \prod_{t=2}^{T} \mathrm{N}(\boldsymbol{\mu}_{l,t}; \boldsymbol{m} + \Theta \boldsymbol{\mu}_{l,t-1}, V)$$

$$\propto \mathrm{N}(\boldsymbol{m}; \boldsymbol{a_m}, B_{\boldsymbol{m}}) \prod_{l=1}^{N} \prod_{t=2}^{T} \mathrm{N}(\boldsymbol{m}; \boldsymbol{\mu}_{l,t} - \Theta \boldsymbol{\mu}_{l,t-1}, V)$$

so that $\boldsymbol{m}$ can be updated with a $\mathrm{N}(\boldsymbol{a_m}^*, B^*_{\boldsymbol{m}})$, with

$$B^*_{\boldsymbol{m}} = (B^{-1}_{\boldsymbol{m}} + N(T-1)V^{-1})^{-1}$$

and

$$a^*_{\boldsymbol{m}} = B^*_{\boldsymbol{m}}(B^{-1}_{\boldsymbol{m}} \boldsymbol{a_m} + V^{-1} \sum_{l=1}^{N} \sum_{t=2}^{T} (\boldsymbol{\mu}_{l,t} - \Theta \boldsymbol{\mu}_{l,t-1}))$$

The parameter V, when given a prior that is $\mathrm{IW}(a_V, B_V)$, has full conditional $p(V \mid \ldots, \text{data}) \propto \mathrm{IW}(a_V + N(T-1), B_V + \sum_{l=1}^{N} \sum_{t=2}^{T} (\boldsymbol{\mu}_{l,t} - \boldsymbol{m} - \Theta \boldsymbol{\mu}_{l,t-1})(\boldsymbol{\mu}_{l,t} - \boldsymbol{m} - \Theta \boldsymbol{\mu}_{l,t-1})^T)$.

The updates for $\boldsymbol{\mu}_{l,t}$ are $\mathrm{N}(\boldsymbol{m}^*, V^*)$, with $\boldsymbol{m}^*$ and V^* given by:

- For $t = 2, \ldots, T-1$, if $M_{l,t} = 0$, then the update for $\boldsymbol{\mu}_{l,t}$ has $V^* = (V^{-1} + (\Theta^{-1} V \Theta^{-T})^{-1})^{-1}$ and $\boldsymbol{m}^* = V^*(V^{-1}(\boldsymbol{m} + \Theta \boldsymbol{\mu}_{l,t-1}) + (\Theta^{-1} V \Theta^{-T})^{-1} \Theta^{-1}(\boldsymbol{\mu}_{l,t+1} - \boldsymbol{m}))$

- For $t = 2, \ldots, T-1$, if $M_{l,t} \neq 0$, then the update for $\boldsymbol{\mu}_{l,t}$ has $V^* = (V^{-1} + (\Theta^{-1} V \Theta^{-T})^{-1} + M_{l,t} \Sigma_l^{-1})^{-1}$ and $\boldsymbol{m}^* = V^*(V^{-1}(\boldsymbol{m} + \Theta \boldsymbol{\mu}_{l,t-1}) + (\Theta^{-1} V \Theta^{-T})^{-1} \Theta^{-1}(\boldsymbol{\mu}_{l,t+1} - \boldsymbol{m}) + \Sigma_l^{-1} \sum_{\{i: L_{t,i} = l\}} \boldsymbol{y}_{t,i})$

- for $t = 1$, if $M_{l,1} = 0$, then the update for $\boldsymbol{\mu}_{l,1}$ has $V^* = ((\Theta^{-1} V \Theta^{-T})^{-1} + V_0^{-1})^{-1}$, and $\boldsymbol{m}^* = V^*((\Theta^{-1} V \Theta^{-T})^{-1} \Theta^{-1}(\boldsymbol{\mu}_{l,2} - \boldsymbol{m}) + V_0^{-1} \boldsymbol{m_0})$, unless $d = 1$, in which case

the updates are $V^* = ((1-\theta^2)V^{-1} + \theta^2 V^{-1})^{-1} = V$ and $m^* = V^*((1-\theta^2)V^{-1}(1-\theta)^{-1}m + \theta V^{-1}(\mu_{l,2} - m))$

- for $t = 1$, if $M_{l,1} \neq 0$, then the update for $\boldsymbol{\mu}_{l,1}$ has $V^* = (M_{l,1}\Sigma_l^{-1} + (\Theta^{-1}V\Theta^{-T})^{-1} + V_0^{-1})^{-1}$ and $\boldsymbol{m}^* = V^*(\Sigma_l^{-1}\sum_{\{i:L_{1,i}=l\}} \boldsymbol{y}_{1,i} + (\Theta^{-1}V\Theta^{-T})^{-1}\Theta^{-1}(\boldsymbol{\mu}_{l,2} - \boldsymbol{m}) + V_0^{-1}\boldsymbol{m_0})$, and for $d = 1$, $V^* = ((1-\theta^2)V^{-1} + \theta^2 V^{-1} + M_{l,1}\Sigma_l^{-1})^{-1}$ and $m^* = V^*((1-\theta^2)V^{-1}(1-\theta)^{-1}m + \theta V^{-1}(\mu_{l,2} - m) + \Sigma_l^{-1}\sum_{\{i:L_{1,i}=l\}} y_{1,i})$
- for $t = T$, if $M_{l,T} = 0$, then the update for $\boldsymbol{\mu}_{l,T}$ has $V^* = V$, and $\boldsymbol{m}^* = \boldsymbol{m} + \Theta\boldsymbol{\mu}_{l,T-1}$
- for $t = T$, if $M_{l,T} \neq 0$, then the update for $\boldsymbol{\mu}_{l,T}$ has $V^* = (M_{l,T}\Sigma_l^{-1} + V^{-1})^{-1}$ and $\boldsymbol{m}^* = V^*(\Sigma_l^{-1}\sum_{\{i:L_{T,i}=l\}} \boldsymbol{y}_{t,i} + V^{-1}(\boldsymbol{m} + \Theta\boldsymbol{\mu}_{l,T-1}))$

Appendix C

Properties of the DDP Prior Model

Details are provided for the properties of the dependent stick-breaking process given in Section 4.2.1.

C.1 Autocovariance of $\mathcal{B}$

The last term in expression 4.2 is obtained by noting that $\mathrm{E}(\beta_t) = \alpha/(\alpha+1)$, since the process is stationary with $\beta_t \sim \mathrm{beta}(\alpha, 1)$ at any time t.

The first term of (4.2) is $\mathrm{E}(\beta_t\beta_{t+k}) = \mathrm{E}(e^{-\zeta^2/\alpha})\mathrm{E}(e^{-(\eta_t^2-\eta_{t+k}^2)/(2\alpha)}) =$

$$\frac{\alpha^{3/2}(1-\rho(k)^2)^{1/2}}{(2+\alpha)^{1/2}((1-\rho(k)^2+\alpha)^2-\alpha^2\rho(k)^2)^{1/2}}. \tag{C.1}$$

Since $\zeta^2 \sim \chi_1^2$, we have that $\mathrm{E}(e^{-\zeta^2/\alpha}) = \alpha^{1/2}/(2+\alpha)^{1/2}$ through the moment generating function of $X = \zeta^2$ which gives $\mathrm{E}(e^{tX}) = (1-2t)^{-1/2}$, and we evaluate this for $t = -1/\alpha$. For the term $\mathrm{E}(e^{-(\eta_t^2+\eta_{t+k}^2)/(2\alpha)})$, we use the fact that $(\eta_t, \eta_{t+k}) \sim N(0, C(k))$, with $C(k)$ a covariance matrix with diagonal elements equal to 1 and off-diagonal elements equal to

$\rho(k)$. Integration results in

$$\mathrm{E}\left(-\frac{\eta_t^2+\eta_{t+k}^2}{2\alpha}\right)=\frac{\alpha(1-\rho(k)^2)^{1/2}}{((1-\rho(k)^2+\alpha)^2-\alpha^2\rho(k)^2)^{1/2}}.$$

Combining these terms yields expression (C.1) and (4.2) follows.

C.2 Autocovariance of Consecutive Weights

The expression for $\mathrm{E}(p_{l,t}p_{l,t+1})$ is:

$$\mathrm{E}(p_{l,t}p_{l,t+1})=\mathrm{E}\left\{(1-\beta_{l,t})(1-\beta_{l,t+1})\prod_{k=1}^{l-1}\beta_{k,t}\beta_{k,t+1}\right\}$$

$$=\mathrm{E}\left\{\prod_{k=1}^{l-1}\beta_{k,t}\beta_{k,t+1}\right\}-\mathrm{E}\left\{\beta_{l,t}\prod_{k=1}^{l-1}\beta_{k,t}\beta_{k,t+1}\right\}-\mathrm{E}\left\{\beta_{l,t+1}\prod_{k=1}^{l-1}\beta_{k,t}\beta_{k,t+1}\right\}+\mathrm{E}\left\{\beta_{l,t}\beta_{l,t+1}\prod_{k=1}^{l-1}\beta_{k,t}\beta_{k,t+1}\right\}.$$

Using the fact that each times series $\boldsymbol{\beta}_l$ is independent of $\boldsymbol{\beta}_m$ (i.e., independence exists across the index l, and only stick-breaking proportions $\beta_{l,t},\beta_{l,t'}$ for $t,t'\in\{1,\ldots,T\}$ are correlated), the expression above simplifies to

$$\prod_{k=1}^{l-1}\mathrm{E}\left\{\beta_{k,t}\beta_{k,t+1}\right\}\left\{1-\mathrm{E}(\beta_{l,t})-\mathrm{E}(\beta_{l,t+1})+\mathrm{E}(\beta_{l,t}\beta_{l,t+1})\right\},$$

with $\mathrm{E}(\beta_{l,t})=\mathrm{E}(\beta_{l,t+1})=\alpha/(\alpha+1)$, and $\mathrm{E}(\beta_{l,t}\beta_{l,t+1})$ is given in (C.1).

To get $\mathrm{corr}(p_{l,t},p_{l,t+1})$, we also need $\mathrm{E}(p_{l,t})=\mathrm{E}\{(1-\beta_{l,t})\prod_{k=1}^{l-1}\beta_{k,t}\}$, and since $\beta_{1,t},\beta_{2,t},\ldots$ are independent, and each are marginally beta(α,1), $\mathrm{E}\{(1-\beta_{l,t})\prod_{k=1}^{l-1}\beta_{k,t}\}=\alpha^l/(1+\alpha)^{l+1}$. This gives the expression for $\mathrm{cov}(p_{l,t},p_{l,t+1})$ as given in (4.4).

C.3 Autocorrelation of Consecutive Distributions

Since G_t is marginally distributed as a $\mathrm{DP}(\alpha, G_0)$, its variance is $\mathrm{var}(G_t(A)) = G_0(A)(1 - G_0(A))/(\alpha + 1)$. By definition, $\mathrm{cov}(G_t(A), G_{t+1}(A))$

$$= \mathrm{E}(\mathrm{cov}(G_t(A), G_{t+1}(A) \mid \boldsymbol{\theta})) + \mathrm{cov}(\mathrm{E}(G_t(A) \mid \boldsymbol{\theta}), \mathrm{E}(G_{t+1}(A) \mid \boldsymbol{\theta})),$$

since $G_t(A) = \sum_{l=1}^{\infty} p_{l,t}\delta_{\boldsymbol{\theta}_l}(A)$, where $\delta_{\boldsymbol{\theta}_l}(A)$ depends on $\boldsymbol{\theta}_l$, which is itself a random variable, and for fixed α, $p_{l,t}$ does not depend on further parameters. The first term in the expression above is then $\mathrm{E}(\mathrm{cov}(G_t(A), G_{t+1}(A) \mid \boldsymbol{\theta})) =$

$$\sum_{l=1}^{\infty} \mathrm{cov}(p_{l,t}, p_{l,t+1})\mathrm{E}(\delta^2_{\boldsymbol{\theta}_l}(A) \mid \boldsymbol{\theta}) + \sum_{l \neq m} \mathrm{cov}(p_{l,t}, p_{m,t+1})\mathrm{E}(\delta_{\boldsymbol{\theta}_l}(A)\delta_{\boldsymbol{\theta}_m}(A) \mid \boldsymbol{\theta})$$

$$= \sum_{l=1}^{\infty} \mathrm{cov}(p_{l,t}, p_{l,t+1})G_0(A) + \sum_{l \neq m} \mathrm{cov}(p_{l,t}, p_{m,t+1})G_0^2(A)$$

using the fact that $\mathrm{E}(\delta_{\boldsymbol{\theta}_l}(A)) = G_0(A)$ and $\mathrm{var}(\delta_{\boldsymbol{\theta}_l}(A)) = G_0(A)(1 - G_0(A))$.

The second term is

$$\mathrm{cov}\left(\mathrm{E}(G_t(A) \mid \boldsymbol{\theta}), \mathrm{E}(G_{t+1}(A) \mid \boldsymbol{\theta})\right) = \mathrm{cov}\left(\sum_{l=1}^{\infty} \mathrm{E}(p_{l,t})\delta_{\boldsymbol{\theta}_l}(A), \sum_{m=1}^{\infty} \mathrm{E}(p_{m,t+1})\delta_{\boldsymbol{\theta}_m}(A) \mid \boldsymbol{\theta}\right)$$

$$= \mathrm{cov}\left(\sum_{l=1}^{\infty} \frac{\alpha^{l-1}}{(1+\alpha)^l}\delta_{\boldsymbol{\theta}_l}(A), \sum_{m=1}^{\infty} \frac{\alpha^{m-1}}{(1+\alpha)^m}\delta_{\boldsymbol{\theta}_m}(A) \mid \boldsymbol{\theta}\right)$$

using the fact that $\mathrm{E}(p_{l,t}) = \alpha^{l-1}/(1+\alpha)^l$. Then, because $\boldsymbol{\theta}_l$ and $\boldsymbol{\theta}_m$ are independent for $m \neq l$, this becomes

$$= \sum_{l=1}^{\infty} \mathrm{var}\left(\frac{\alpha^{l-1}}{(1+\alpha)^l}\delta_{\boldsymbol{\theta}_l}(A) \mid \boldsymbol{\theta}\right)$$

$$= \sum_{l=1}^{\infty} \frac{\alpha^{2(l-1)}}{(1+\alpha)^{2l}} G_0(A)(1 - G_0(A)).$$

Finally, letting $\sigma_{l,m} = \mathrm{Cov}(p_{l,t}, p_{m,t+1})$, and putting everything together, we have (4.6). The form of $\sigma_{l,m} = \mathrm{cov}(p_{l,t}, p_{m,t+1})$ can be worked out in a similar way to $\sigma_{l,l}$.

C.4 Stationarity of the $\mathcal{B}$ Process

Since $\eta_{\mathcal{T}}$ is a stationary process, $f_{\boldsymbol{\eta_t}}(\eta_{t_1}, \ldots, \eta_{t_k}) = f_{\boldsymbol{\eta_{t+s}}}(\eta_{t_1+s}, \ldots, \eta_{t_k+s})$. Applying a transformation to $\boldsymbol{\eta_t}$ and $\boldsymbol{\eta_{t+s}}$ to obtain the distribution of $\boldsymbol{\beta_t}$ and $\boldsymbol{\beta_{t+s}}$, we find that $f(\beta_{t_1}, \ldots, \beta_{t_k})$ is the same function of $\boldsymbol{\beta_t}$ as $f(\beta_{t_1+s}, \ldots, \beta_{t_k+s})$ is of $\boldsymbol{\beta_{t+s}}$. That is, these density functions are the same. Therefore, the stochastic process $\mathcal{B}$ is strongly stationary, as a consequence of the strong stationarity of the $\eta_{\mathcal{T}}$ process.