UNIVERSITY OF CALIFORNIA
SANTA CRUZ

FLEXIBLE AND EFFICIENT ORDINAL REGRESSION WITH BAYESIAN NONPARAMETRICS

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

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Jizhou Kang

September 2024

The Dissertation of Jizhou Kang
is approved:

Professor Athanasios Kottas, Chair

Professor Juhee Lee

Professor Zehang Li

Professor Stephan Munch

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures **vi**

List of Tables **xv**

Abstract **xvii**

Dedication **xix**

Acknowledgments **xx**

1 Introduction **1**
1.1 Motivation and Objectives 1
1.2 Latent Variable Models for Ordinal Responses 5
1.3 Summary of contributions 9

2 A Structured Mixture Modeling Framework for Cross-sectional Ordinal Regression **12**
2.1 Introduction 12
2.2 General Methodology 16
2.2.1 From Building Blocks to General Model 16
2.2.2 Model Properties 20
2.2.3 Prior Specification 21
2.2.4 Posterior Inference 25
2.2.5 Assessing Model Flexibility 28
2.3 Specific Models for Ordinal Regression 34
2.3.1 The Common-weights Model 34
2.3.2 The Common-atoms Model 36
2.4 Data illustrations 37
2.4.1 Synthetic Data Examples 37
2.4.2 Credit ratings of U.S. firms 51
2.4.3 Retinopathy data 56
2.5 Discussion 59

3 A Nonparametric Modeling Approach for Ordinal Regression with Heterogeneous Responses 62
3.1 Introduction . . . 62
3.1.1 Background and Data . . . 62
3.1.2 Objectives and Outline . . . 65
3.2 Continuous Mixture Models . . . 67
3.2.1 Beta-Binomial and Logistic-Normal-Binomial . . . 67
3.2.2 Models for Ordinal Responses from Developmental Toxicity Study . . . 69
3.3 Discrete Mixture Models . . . 72
3.3.1 Models with Continuation-ratio Logits Kernel . . . 72
3.3.2 Models with Overdispersed Kernel . . . 77
3.4 Synthetic Data Examples . . . 80
3.4.1 First Synthetic Data Example . . . 81
3.4.2 Second Synthetic Data Example . . . 84
3.5 Real Data Illustrations . . . 87
3.6 Summary and Remarks . . . 95

4 A Flexible Modeling Framework for Longitudinal Ordinal Responses 97
4.1 Introduction . . . 97
4.2 The Modeling Approach for Binary Responses . . . 101
4.2.1 Model Specification . . . 102
4.2.2 Model Properties . . . 106
4.2.3 Prior Specification and Posterior Inference . . . 110
4.2.4 Connections with Existing Literature . . . 114
4.3 Data Illustrations with Binary Responses . . . 116
4.3.1 Synthetic data examples . . . 116
4.3.2 Real Application: *Studentlife* data . . . 126
4.4 Model for ordinal responses . . . 134
4.4.1 The extended model . . . 134
4.4.2 Data illustration . . . 137
4.5 Discussion . . . 138

5 A Case Study: Estimating Maturity of Sheepshead Minnows 141
5.1 Introduction . . . 141
5.2 Methodology . . . 147
5.3 Results . . . 152
5.3.1 Binary Responses . . . 152
5.3.2 Three-level Ordinal Responses . . . 157
5.4 Comments . . . 159

6 Conclusions 162

A Proofs **165**
A.1 Properties of Models for Cross-sectional Ordinal Regression 165
A.1.1 Proof of Proposition 2.2 165
A.1.2 Proof of Proposition 2.3 166
A.1.3 Proof of Lemma 2.1 167
A.1.4 Proof of Theorem 2.1 169
A.1.5 Proof of Proposition 2.4 173
A.2 Properties of Models for Heterogeneous Ordinal Responses 175
A.2.1 Proof of Proposition 3.1 175
A.2.2 Proof of Proposition 3.2 176
A.2.3 Proof of Proposition 3.3 177
A.3 Properties of Models for Longitudinal Ordinal Responses 178
A.3.1 Proof of Proposition 4.1 178
A.3.2 Proof of Proposition 4.2 179
A.3.3 Proof of Proposition 4.3 181

B Implementation Details **182**
B.1 MCMC of Models for Cross-sectional Ordinal Regression 182
B.1.1 The General Model 182
B.1.2 The Common-weights Model 185
B.1.3 The Common-atoms Model 186
B.2 MCMC of Models with Overdispersed Kernel 187
B.3 MCMC of Models for Longitudinal Ordinal Responses 190

List of Figures

2.1 Illustration of the continuation-ratio logits structure. 19

2.2 Illustration of how the two bounds can be used to set the monotonic pattern of the prior expected probability response curve. 23

2.3 Illustration of the prior specification strategy. In each panel, the red solid line is the prior expected probability response curve, the blue dashed lines and shaded region indicate the prior 95% interval estimate, and the green dotted lines show 5 prior realizations. . . 25

2.4 Synthetic data example. Posterior mean and 95% credible interval estimates for the marginal probability response curves under the common-weights (blue line and shaded region), common-atoms (orange line and shaded region), and general (red line and shaded region) models. In each panel, the green solid line is the true regression function. 39

2.5 Synthetic data example ($n = 800$). Box plots of the posterior samples for the six largest mixture weights, under the common-atoms and general models. 40

2.6 Second simulation example. Inference results for the marginal probability response curves. In each panel, the dashed line and shaded region correspond to the posterior mean and 95% credible interval estimates, whereas the (green) solid line denotes the true regression function. 42

2.7 Second simulation example. Box plots of the posterior samples for the three largest mixture weights under each of the nonparametric models. 44
2.8 Second simulation example. Posterior mean estimates of the three largest mixture weights and atoms. The red circle, blue plus, and green triangle correspond to the first, second, and third largest weights, respectively. 45
2.9 Second simulation example. Second simulation example. Inference results for the marginal probability response curves, under the informative prior specification. In each panel, the dashed line and shaded region correspond to the posterior mean and 95% credible interval estimates, whereas the (green) solid line denotes the true regression function. 46
2.10 Third simulation example. The true probability response surface $\pi_j(x_1, x_2)$, for $j = 1, 2, 3$ (from left to right). 47
2.11 Third simulation example. Posterior mean estimates of $\pi_j(x_1, x_2)$, for $j = 1, 2, 3$ (from left to right). 48
2.12 Third simulation example. Posterior mean of the three largest mixture weights for the common-atoms and general LSBP mixture models. 49
2.13 Credit ratings data. Posterior mean (lines) and 95% interval (shaded regions) estimates of probability response curves $\pi_j(x_s)$. Estimates for all five response categories are displayed in a single panel for each covariate. 53
2.14 Credit ratings data. Posterior mean estimates of probability response surfaces $\pi_j(x_2, x_3)$, for $j = 1, \ldots, 5$ (from left to right). . . 54
2.15 Credit ratings data. Posterior distributions of the probability of obtaining investment grade rating under the common-weights model. The red solid lines indicate the posterior mean. 55

2.16 Retinopathy data. Posterior densities for the probabilities and log odds ratios of the two retinopathy endpoints for smokers (in red) and non-smokers (in blue). The dashed line and shaded region correspond to the posterior mean and 95% credible interval, respectively 57
2.17 Retinopathy data. Posterior predictive distribution of the proportion for each ordinal response category. The dashed line indicates the observed proportion. 59

3.1 EG data. In each panel, a circle corresponds to a particular dam and the size of the circle is proportional to the number of implants. The coordinates of the circle are given by the toxin level and the proportion of the specific endpoint: non-viable fetuses among implants (left panel); malformations among live pups (middle panel); combined negative outcomes among implants (right panel). 65
3.2 EG data. Posterior mean (dotted line) and 95% interval estimate (dashed lines) for the dose response curves. The red solid line and shaded region is the posterior mean and 95% interval estimates obtained under a continuation-ratio logits model. 71
3.3 Connection between alternative encodings of the ordinal response. 75
3.4 First simulation example. Posterior mean and 95% interval estimates for the dose response curves under the mixture models with different kernel. In each panel, the posterior mean and interval estimates obtained under the model with and without overdispersed kernel are given by the blue dotted and dashed lines and the red dot-dashed line and shaded region, respectively. The green solid line is the true dose-response curve. In the top panel, a circle corresponds to a particular dam and the size of the circle is proportional to the number of implants. 83

3.5 First simulation example. Box plots of the posterior samples for the four largest mixture weights at the four observed dose levels and a new dose level, under each of the nonparametric models. 84
3.6 Second simulation example. Posterior mean and 95% interval estimates for the dose response curves under the mixture models with different kernel. In each panel, the posterior mean and interval estimates obtained under the model with and without overdispersed kernel are given by the blue dotted and dashed lines and the red dot-dashed line and shaded region, respectively. The green solid line is the true dose-response curve. In the top panel, a circle corresponds to a particular dam and the size of the circle is proportional to the number of implants. 86
3.7 Second simulation example. Box plot of the intracluster correlation posterior distributions at the observed toxin levels. In each panel, estimates under the "CW-Bin", "CW-LNB", "Gen-Bin" and "Gen-LNB" model are shown in red, blue, green, and purple, respectively. The orange dot marks the truth. 87
3.8 Second simulation example. Posterior distributions of the overdispersion parameters $\boldsymbol{\sigma^2}$ under the "CW-LNB" and the "Gen-LNB" model with prior $IG(3, 8/3)$ (in red) and $IG(2, 4/3)$ (in blue). . . 87
3.9 EG data. Posterior mean and 95% interval estimate for the dose response curves under the mixture models with different kernel. In each panel, the red solid line and shaded region is the posterior mean and interval estimates obtained under the model with continuation-ratio logits kernel, while the blue dotted and dashed lines are the estimates from the model with overdispersed kernel. 89
3.10 EG data. Box plot of the intracluster correlation posterior distributions at four observed toxin levels and for the new value of $x = 3.75$ g/kg. In each panel, estimates under "CW-Bin", "CW-LNB", "Gen-Bin" and "Gen-LNB" model are shown in red, blue, green, and purple, respectively. 90

3.11 EG data. Posterior distribution of the effective dose with 5% BMR (in red) and 10% BMR (in blue). The shaded region indicates the 95% credible interval. The corresponding benchmark dose is marked with "×". 91

3.12 EG data. Posterior mean ("∘") and 95% uncertainty bands (dashed lines) for the probability mass $\Pr(R \mid m = 12, G_{\mathbf{x}})$ (top panels) and conditional probability mass $\Pr(y \mid m = 12, R = 2, G_{\mathbf{x}})$ (bottom panels), at four observed toxin levels and for the new value of $x = 3.75$ g/kg. In each panel, estimates under "CW-Bin", "CW-LNB", "Gen-Bin" and "Gen-LNB" model are shown in red, blue, green, and purple, respectively. 92

3.13 EG data. Box plots of posterior predictive samples for the embryolethality, malformation, and combined risk endpoints at the observed toxin levels. The corresponding observed proportions are denoted by "∘". 93

4.1 Simulation study regarding the mean structure. Inference results for the probability response curve. In each panel, the dashed line and shaded region correspond to the posterior mean and 95% credible interval estimates, the (orange) dot is the original binary data, whereas the (green) cross denotes the true probability of generating that responses. 118

4.2 Simulation study regarding the mean structure. Prediction of the probability response curve for a new subject. In each panel, the dashed lines and shaded region shows the posterior mean and 95% interval estimates of probability response curve for a new subject. The solid lines are the posterior mean estimates of probability response curves for the in-sample subjects. The dotted line is the true probability function for generating binary responses. 119

4.3 Simulation study regarding the mean structure. Box and violin plots of the posterior samples of RMSE for different data generating process and sparsity level combinations. The red box corresponds to the proposed model while the blue box is for the simplified model. 120
4.4 Simulation study regarding the covariance structure. Inference results for the signal covariance kernels. In each panel, the dashed line and shaded region correspond to the posterior mean and 95% credible interval estimates, whereas the solid line denotes the true covariance kernel. 122
4.5 Simulation study regarding the covariance structure. Posterior interval estimate of correlation coefficients ("box") versus point estimate obtained from the true data generating process ("$\star$"). In each panel, the upper triangle and the lower triangle are for the Pearson and the rachoric correlation coefficient, respectively. . . . 123
4.6 Simulation study regarding the covariance structure. Histogram for the posterior samples of the 2-Wasserstein distance between the f.d.d.s. of the centralized signal process obtained from the proposed model (upper panel) and the simplified model (lower panel) to the truth. 124
4.7 Simulation study with irregular observing points. Visualization of the repeated measurements for each subject. The blue dot marks a positive response while the red cross represents a negative response. 125
4.8 Simulation study with irregular observing points. Posterior inference of a new subject's probability response curve. The dashed line and shaded region show the posterior mean and 95% interval estimates of probability response curve for a new subject. The dotted line is the true probability function for generating binary responses. As references, the solid lines are the posterior mean estimates of probability response curves for the in-sample subjects. 126
4.9 *Studentlife* data. Proportion of three types of response (positive, negative, and missing,) over time, for valence and arousal scores. . 129

4.10 *Studentlife* data. Empirical estimate of the correlation coefficients between binary responses within a week. In each panel, the upper triangle and the lower triangle are for the Pearson and the tetrachoric correlation coefficient, respectively. 129
4.11 *Studentlife* data. Posterior mean (dashed line) and 95% interval estimate (shaded region) of the probability response curve for an out-of-sample subject. The posterior mean estimates of probability response curves for in-sample subjects are given by the solid lines. The vertical shaded regions correspond to the four special time periods (see Section 4.3.2.1). 131
4.12 *Studentlife* data. Posterior density estimate of an out-of-sample subject's valence and arousal probability over the mood coordinate space on four specific days. In each panel, the crosses represent the posterior means of the in-sample subjects' valence and arousal probability mapped to the mood coordinate space. 132
4.13 *Studentlife* data. Posterior mean (solid line) and 95% interval estimate of the signal process covariance kernel. 132
4.14 Four levels arousal score data. Posterior mean (dashed line) and 95% interval estimate (shaded region) of probability response curve for an out-of-sample subject. The posterior mean estimates for the probability response curves of in-sample subjects are given by the solid lines. The vertical shaded regions correspond to the four special time periods (see Section 4.3.2.1). 139

5.1 Fish maturity data. Proportion of mature fish over time for each treatment group. 143
5.2 Fish maturity data. Empirical estimate of the correlation coefficients between binary responses over time. In each panel, the upper triangle and the lower triangle indicate the Pearson and the tetrachoric correlation coefficient, respectively, while the numbers on the diagonal are the variances. 144

5.3 Fish maturity data. Transition proportion matrix for fish in each treatment group. The i, j-th entry shows the observed proportion of transitions from stage i at time $t-1$ to stage j at time t. The treatment is specified by the title of each panel. 145
5.4 Fish maturity data. Posterior predictive mean (dashed line) and 95% interval estimate (shaded region) of the probability response curve for an out-of-sample subject. The posterior mean estimates of probability response curves for in-sample subjects are given by the solid lines. The corresponding treatment is specified as the title of each panel. 153
5.5 Fish maturity data. Posterior predictive distribution for $d^*(Z_g^*(\tau), Z_{g'}^*(\tau))$. In each panel, the black solid line is the kernel density estimation. The red dashed line indicates the mean, and the blue dotted lines mark the 95% interval. 155
5.6 Fish maturity data. Posterior distribution of $\varphi(Z_g^*(\tau))$ for fish of the specified treatment and population at the chosen time point. The red, blue, green, and purple histograms correspond to treatment $(\mathrm{PT} = 26, \mathrm{OT} = 26)$, $(\mathrm{PT} = 32, \mathrm{OT} = 26)$, $(\mathrm{PT} = 26, \mathrm{OT} = 32)$, and $(\mathrm{PT} = 32, \mathrm{OT} = 32)$, respectively. 156
5.7 Fish maturity data. Posterior empirical distribution for the time of maturity for fish of the specified treatment and population. The red, blue, green, and purple histograms correspond to treatment $(\mathrm{PT} = 26, \mathrm{OT} = 26)$, $(\mathrm{PT} = 32, \mathrm{OT} = 26)$, $(\mathrm{PT} = 26, \mathrm{OT} = 32)$, and $(\mathrm{PT} = 32, \mathrm{OT} = 32)$, respectively. 158
5.8 Fish maturity data with three-level ordinal responses. Posterior empirical distribution for the time of reaching maturity level 2 for fish of the specified treatment and population. The red, blue, green, and purple histograms correspond to treatment $(\mathrm{PT} = 26, \mathrm{OT} = 26)$, $(\mathrm{PT} = 32, \mathrm{OT} = 26)$, $(\mathrm{PT} = 26, \mathrm{OT} = 32)$, and $(\mathrm{PT} = 32, \mathrm{OT} = 32)$, respectively. 159

5.9 Fish maturity data with three-level ordinal responses. Posterior empirical distribution for the time of reaching maturity level 3 for fish of the specified treatment and population. The red, blue, green, and purple histograms correspond to treatment ($\mathrm{PT} = 26, \mathrm{OT} = 26$), ($\mathrm{PT} = 32, \mathrm{OT} = 26$), ($\mathrm{PT} = 26, \mathrm{OT} = 32$), and ($\mathrm{PT} = 32, \mathrm{OT} = 32$), respectively. 160

List of Tables

2.1 First simulation example. Summary of model comparison using the posterior predictive loss criterion. The values corresponding to the best model are given in bold. 41

2.2 Second simulation example. Summary of model comparison using the posterior predictive loss criterion. The values correspond to the best model are given in bold. 43

2.3 Third simulation example. Summary of model comparison results, using the RMSE $\bar{E}_j$, average 95% posterior credible interval length $\bar{L}_j$, and the coverage of the 95% posterior credible interval $\bar{R}_j$, for $j = 1, 2, 3$. The values that correspond to the best model are given in bold. 51

2.4 Credit ratings data. Summary of the posterior predictive loss criteria for model comparison. Each pair of numbers corresponds to $(G_j(\mathcal{M}), P_j(\mathcal{M}))$, $j = 1, \cdots, 5$. "Parametric" refers to the continuation-ratio logits model. The values for model with the smallest $G_j(\mathcal{M}) + P_j(\mathcal{M})$ are highlighted in bold. 52

3.1 EG data. BMD estimation under different models, based on posterior samples of ED. 92

3.2 EG data. Summary of comparison among the nonparametric models using the posterior predictive loss and interval score criteria. The values in bold correspond to the model favored by the particular criterion. 94

4.1 Simulation study with irregular observing points. Comparison between the proposed model and the generalized linear mixed effects model using two different criteria. The values in bold correspond to the model favored by the particular criterion. 126

4.2 *Studentlife* data. Summary of comparison between the proposed model and the generalized linear mixed effects model using two different criteria. The values in bold correspond to the model favored by the particular criterion. 133

5.1 Fish maturity data. Summary of comparison between two classes of models with different prior specifications using four different criteria. The values in bold correspond to the model favored by the particular criterion. 152

Abstract

Flexible and Efficient Ordinal Regression with Bayesian Nonparametrics

by

Jizhou Kang

Scientific discoveries are advanced by flexible and efficient statistical models. Grounded on Bayesian nonparametric modeling techniques, this thesis provides a toolbox for ordinal regression. The toolbox comprises models tailored for various settings, with shared characteristics of flexibility and efficiency. A key building block of the proposed models is a sequential mechanism to treat the ordinal response. Such mechanism implies a factorization of the response distribution that allows efficient, scalable computation through (partial) parallel sampling regarding the response categories. For problems under a cross-sectional setting, we develop nonparametric mixture models, leveraging the same sequential structure to define covariate-dependent mixture weights. Even though covariates are incorporated via linear functions, the mixture models admit flexible ordinal regression relationships, and they relax parametric assumptions for the response distribution. Moving towards modeling the dynamic evolution of ordinal responses from longitudinal studies, the critical insight is to treat the subjects measurements as stochastic process realizations at the corresponding time points. We propose a hierarchical framework that models the mean and covariance structure of the processes nonparametrically and simultaneously, a useful byproduct being a practical method for making predictions on any time scale. For all proposed models, we craft readily implementable Markov chain Monte Carlo algorithms that avoid specialized updates or tuning steps. A variety of synthetic and real data examples are used to illustrate the methods. In particular, the models for cross-sectional ordinal regression, along with their extensions, are examined in the context of risk assessment

for developmental toxicity studies. We also present a case study in evolutionary biology, in which our method for longitudinal ordinal responses is adapted to identify the impact of temperature on transgenerational responses, using repeated measurements on fish maturation data.

For my Families, Buddies, and Heroes, who are with me through every challenge and triumph.

Acknowledgments

This journey begins with an email from Thanasis five years ago, when he served as the director of our graduate program. He wrote "I'm contacting you regarding your application to our graduate program at UC Santa Cruz, which has generated a lot of enthusiasm." The wheel of fate rolls, and I was lucky to have Thanasis as my advisor. He is a wonderful advisor. He always has great ideas, encompassing not only research-related ideas that advance every piece of my work but also suggestions for teaching, extracurricular activities, and personal well-being, greatly enhancing my graduate school experience. This dissertation would have been impossible without his tireless editing and guidance. Throughout my time at UCSC, his consistent support and encouragement generate a lot more enthusiasm in my life, for which I cannot thank him enough.

My dissertation committee members deserve special thanks. Juhee Lee chaired my advance to candidacy committee. Her patience in addressing my questions has been invaluable, and I have repeatedly benefited from her expertise. I also appreciate her effort in organizing the journal club, where she has offered perspective and ideas that enriched my understanding on diverse topics. Zehang Li's statistical learning class is one of my favorites at UCSC. I aspire to approach my work with the same amount of professionalism, seriousness, and enthusiasm that he constantly embodies. Steve Munch provided numerous valuable insights and suggestions at my advancement exam, and it was through him that I obtained the fisheries data which plays an important role in this thesis. I am indebted to them for their valuable advice and insights, which have improved this thesis.

I am grateful to the other faculty members of the Statistics Department here at UCSC, who provided me with a solid statistical education. In particular, Herbie Lee offered generous helpful as my first-year advisor. He provided funding and

mentorship for a unique opportunity to design and develop my own online course. I am lucky to have taken classes from Bruno Sansó and Raquel Prado. Their profound expertise enriched my statistical skillsets. My achievement is rooted in the friendly and collaborative atmosphere created by all of these professors, for which I am very thankful.

I was lucky to have established connections with alumni of the department. Tony Pourmohamad is one of my mentors during my internship. I would consider myself fortunate if I am able to emulate his passion to work. I am grateful to him and the other mentor, Theo Koulis, for allowing me to experience the very best of collaborative research and real teamwork. It is always a pleasure to meet Matt Heiner at conferences. He generously offered encouragement and suggestions. Xiaotian Zheng and Chunyi Zhao provided constructive feedback and career advice. I thank them for being my role models and friends. Graduate school would not have been nearly as enjoyable without a group of supportive and loving fellow students, including Shuangjie Zhang and Zach Horton, who have continuously lifted my spirits.

Most profoundly, I extend my deepest gratitude to my family, especially my parents, for their selfless love. They have done so much to support me and my endeavors, enduring countless sleepless nights thinking about their child on the other side of the world, whom they could not see face-to-face for years. Their love and encouragement have created a sanctuary where I find trust, solace, and love, even when we are physically apart. My achievements are theirs, even though they expect nothing more than their child's happiness. I also want to express my heartfelt gratitude to my friends across the world, many of whom have been with me since childhood. Thank you for standing by my side despite the distance and time that separate us.

Chapter 1

Introduction

1.1 Motivation and Objectives

Analyzing subject responses, along with relevant predictors, constitutes a key challenge in statistics. In this work, we specifically target ordinal categorical responses. The setting of the study may be categorized as cross-sectional or longitudinal. We explore modeling approaches for either type of study.

Recent years have witnessed a rapid growth of ordinal data in a wide range of applications. Owing to its natural virtue in measuring unobservable features, such as degree of agreement, propensity in attitudes, and intensity of emotions, ordinal data is widely used in different fields. Examples include finance (agencies such as Standard and Poor's providing credit rating of companies ranging from "AAA" to "D"), biomedical sciences (the effect of a treatment categorized into "complete response" "partial response", "minimum response", and "no change"), social sciences (survey respondents giving their opinions on ordinal scale "agree", "neutral", and "disagree"), and environmental sciences (air quality rating such as "good", "fair", and "bad"). These ordinal variables, usually accompanied by relevant explanatory variables (covariates), form the ordinal regression problem.

The main objective is to examine the relationship between the ordinal response and the covariates, while appropriately accounting for the ordinal discrete nature of the responses.

Cross-sectional and longitudinal studies are both widely encountered in scientific disciplines due to their potential to address different questions of interest. The defining characteristic of a longitudinal study is that individuals are measured repeatedly through time, while in a cross-sectional study, a single outcome is measured for each individual. The major benefit of longitudinal studies is their ability to distinguish changes over time within individuals (aging effects) from differences among subjects with their baseline level (cohort effects). The benefit comes at a price of modeling challenges, especially in the direction of developing models that permit more general forms of dependence among the repeated measurements.

We identify flexibility and efficiency as the key objectives of the proposed methodologies. Validating whether the real data structure is compatible with the model assumptions can be demanding. Hence, flexible models that impose fewer restrictions on the data distribution are desirbale. Efficiency is another crucial consideration for practitioners. Models that demand fewer computational resources and less tuning sophistication are more appreciated in practice. Moreover, in applied fields, discoveries have been hindered by the available statistical tools, either because the tools are too restrictive, or because the computational cost is unaffordable.

Despite its importance, flexible and efficient modeling for ordinal regression is not sufficiently well developed in the literature. Existing approaches are either restrictive from a modeling perspective, making strong assumptions on the effect of covariates on the responses, or computationally demanding, requiring complex and/or inefficient algorithms to obtain inference.

Conventionally, methods for cross-sectional ordinal regression are based off of the generalized linear model (GLM) framework. The log odd ratios of response categories are linked to the linear predictor $\mathbf{x}^\top \boldsymbol{\beta}$. Different choices can be made on the form of odds ratios and link function, resulting in a variety of (parametric) modeling options (Agresti, 2010). Facilitated by augmented latent variables, Bayesian inference for such models is fairly efficient (Albert and Chib, 1993; Polson et al., 2013). However, parametric model formulations restrict flexibility, in terms of both the ordinal response distribution and the covariate-response relationship. To overcome such limitations, early work in the literature has explored semiparametric models (e.g. Basu and Chib, 2003; Choudhuri et al., 2007) and nonparametric models (e.g. Chib and Greenberg, 2010; Bao and Hanson, 2015). However, the computationally intensive inferential techniques hinder their popularity in practice. For example, posterior sampling under the model proposed in Chib and Greenberg (2010) requires a non-standard Metropolis-Hasting step with tailored proposal density.

In the context of longitudinal ordinal regression, models are developed under one of three broad approaches pertaining to marginal models, conditional models, or subject-specific models, postulating the generalized linear model framework. We refer to Molenberghs and Verbeke (2006) for a comprehensive review. These models typically assume a specific mechanism for the ordinal response evolution, and hence are restrictive in modeling dynamic ordinal regression relationships. Extensions have been developed in the Bayesian nonparametric literature, by incorporating a temporally dependent nonparametric prior on cross-sectional ordinal regression model (e.g. DeYoreo and Kottas, 2018b), through utilizing a smooth spline function for the regression relationship (e.g. Tang and Duan, 2012), or via embedding a nonparametric prior as the probability model for the random effects (e.g. Jara

et al., 2007). A question here is how to balance model flexibility and computation tractability.

Despite the general challenge of developing flexible and efficient models, a specific ordinal regression problem usually exhibits unique features that may bring further obstacles in modeling. In cross-sectional studies, exploratory data analysis usually suggests heterogeneity in terms of ordinal regression relationships. In modern longitudinal studies, it is common that the complete vector of repeated measurements is not collected on all subjects, leading to unbalanced longitudinal data. Additionally, overdispersed ordinal responses are widely encountered in certain applications, such as developmental toxicity studies. We seek to develop a general modeling framework that can be tailored for particular applications.

Such modeling objectives lead us to Bayesian nonparametrics, a rapidly growing field that offers a broad modeling framework for flexible and efficient statistical inference and prediction. An inherent virtue of Bayesian nonparametric models is their flexibility, which is characterized by large support on the space of relevant distributions and/or functions. Besides, efficient posterior simulation techniques for Bayesian nonparametric models have been developed. For reviews that cover theoretical, methodological and computational aspects of Bayesian nonparametric models, we refer to Müller et al. (2015) and Ghosal and van der Vaart (2017).

The goal of this thesis is to exploit the advantages of Bayesian nonparametric models to solve methodologically and practically relevant ordinal regression problems. We seek to develop a unified toolbox for ordinal regression under cross-sectional or longitudinal settings, emphasizing the two key aspects of flexibility and efficiency. The proposed models aim at relaxing restrictive assumptions on the (evolution of) ordinal response distribution and ordinal regression relationship, while involving efficient inference algorithms and interpretable expressions for key

quantities of interest. From the applications perspective, this work contributes to the analysis of ordinal data, a prevalent problem in several scientific fields.

1.2 Latent Variable Models for Ordinal Responses

In ordinal regression, the order of response categories should be taken into account. Motivated by exploiting the ordinality, latent variable models have been developed, postulating either a cumulative link structure or a sequential structure. In practice, the choice among the two alternatives is typically based on convention rather than a deliberate decision that takes the context and objectives of the specific problem into consideration. Here, we contrast these two structures in terms of handling order, with emphasis placed on the practical and methodological benefits of each approach.

Consider modeling a univariate ordinal response Y with C categories. We can equivalently encode the response as a vector of binary variables $\mathbf{Y} = (Y_1, \ldots, Y_C)$, such that $Y = j$ is equivalent to $Y_j = 1$ and $Y_k = 0$ for any $k \neq j$. It is typical to assume a multinomial response distribution, denoted by $\text{Mult}(1, \pi_1, \ldots, \pi_C)$. The task is to model the response category probabilities π_j, acknowledging the order of the categories.

The cumulative link model (McCullagh, 1980) assumes that the ordinal response arise from latent continuous random variable through discretization. In particular, let Z be a continuous variable with cumulative distribution function F. The observed ordinal variable is determined by $Y = j$ if and only if $Z \in (\varkappa_{j-1}, \varkappa_j]$, for $j = 1, \ldots, C$. Here $-\infty = \varkappa_0 < \varkappa_1 < \ldots < \varkappa_{C-1} < \varkappa_C = \infty$ are cut-off points on the latent scale, where, typically, $\varkappa_1 = 0$ for identifiability. The term "cumulative link model" is adopted because it is essentially the cumulative probabilities $\Pr(Y \leq j) = \pi_1 + \ldots + \pi_j$, $j = 1, \ldots, C$ that are linked to the latent

variable.

The sequential model (Tutz, 1991) starts with introducing $C-1$ latent continuous random variables, $(\mathcal{Z}_1, \ldots, \mathcal{Z}_{C-1}) \in \mathbb{R}^{C-1}$, where $\mathcal{Z}_j \overset{ind.}{\sim} F_j$, $j = 1, \ldots, C-1$. The ordinal response is then determined through a sequential dichotomization mechanism of the latent variables. Specifically, the mechanism allows the ordinal categories to be reached in consecutive order. To begin with, the ordinal response Y is allocated to the first category if $\mathcal{Z}_1 > 0$. If $\mathcal{Z}_1 \leq 0$, the process continues to assign $Y = 2$ if $\mathcal{Z}_2 > 0$. Given that category j is reached, the sign of $\mathcal{Z}_j$ is used to determined if the process stops or if it continues with higher categories. Note that because the distribution of the latent variables is allowed to vary with category and the split is binary, we can fix the cut-off for all latent variables at 0.

A key feature of the sequential model is that it facilitates a factorization of the multinomial distribution in terms of Binomial distributions. That is,

$$Mult(\mathbf{Y} \mid 1, \pi_1, \ldots, \pi_C) \;=\; Bin(Y_1 \mid m_1, p_1) \ldots Bin(Y_{C-1} \mid m_{C-1}, p_{C-1}),$$

where $m_1 = 1$, and $m_j = 1 - \sum_{k=1}^{j-1} Y_k$, for $j = 2, \ldots, C-1$. Here p_j denotes the conditional probability of response j, given that the response is j or higher, for $j = 1, \ldots, C-1$. It is linked with π_j through $p_j = \pi_j/(\pi_j + \ldots + \pi_C)$. This factorization is also referred to as the continuation-ratio parameterization of the multinomial distribution (Agresti, 2010).

Through formulated under different assumptions, the two approaches are compatible in modeling ordinal responses. In fact, there are scenarios where one can find equivalence between the two model formulations, i.e., one is a reparameterization of the other. Peyhardi et al. (2015) provides several examples, and in Section 2.2.5, we present the explicit reparameterization when the latent continuous variables are logistically distributed. In practice, the different methods may provide

similar results (Agresti, 2010). Nonetheless, given the context of the problem and/or the relevant scientific questions, one approach might be preferred over the other.

In general, we should prefer the model whose structure is better equipped to address the relevant scientific questions of the problem at hand. Consider, for instance, modeling the ordinal scale air quality index. It is plausible to assume that a continuous variable measuring air quality has a regression relationship with covariates. Apart from estimating the ordinal regression relationship, practitioners may also be interested in determining the responses' scale. Consequently, the cumulative link model is more appropriate, because the covariate effects are invariant to the choice of categories for the ordinal response, making it possible to compare models using different response scales. On the contrary, ordinal responses may arise from a sequential mechanism. For example, the disease severeness evolves from the mildest to the most severe. We can assume there is a sequential binary splitting process that determines the ordinal severeness level, from the mildest to the most severe, step by step. The covariate effects enter at every splitting point. Such a process facilitates direct inference for the covariate effects on the conditional probability of more severe disease, which is of particular interest in practice. Accordingly, the sequential model should be preferred.

When proposing models for ordinal data, the specific context of the problem may be unknown. Even if the context is known, both modeling assumptions could be equally applicable, or inapplicable, for the objectives. In either scenario, it is hard to select a model structure based on the essence of the problem. Therefore, we should also take the methodological objectives into account.

The cumulative link model essentially assumes the covariates affect a single latent continuous variable, and the cut-off points, which usually do not depend

on covariates, enter to discretize the latent variable. Because of this structured assumption on the covariate effects, it is more manageable to incorporate certain prior beliefs on the form of the covariate effects, such as monotonicity. Besides, cumulative link models have an immediate connection with regression models for continuous responses. Studying theoretical properties under the cumulative link modeling framework is thus more convenient, leveraging the existing results regarding continuous regression models.

When flexibility and efficiency is the major concern for the proposed method, the sequential model is arguably the better approach. The sequential model enables enhanced flexibility in terms of the ordinal regression relationship through allowing the covariates to affect all the $C-1$ latent continuous variables. In addition, by exploring the distributional assumptions for the latent variables, a wider scope of ordinal regression models can be developed. In contrast, to achieve a comparable level of flexibility postulating the cumulative link structure, we would have to embed covariates or dependence into the cut-off points. This is more challenging because the cut-off points must exhibit the order restriction. Critically, the sequential model boosts computation through the implied conditional independent structure of category-specific parameters, allowing parallel computing across response categories. Bearing in mind the key objectives of this thesis, we will adopt the sequential modeling structure as a key building block for our proposed Bayesian nonparametric modeling methods for ordinal regression.

We note here that the terminologies "cumulative link model" and "sequential model" represent classes of ordinal regression models. A specific model is determined in combination with a choice of link function and linear predictor. Besides, there are approaches other than using latent variables that are appropriate for ordinal regression. For instance, the adjacent categories model (Masters, 1982), which

is based on the odds ratio between two consecutive categories. A comprehensive review of approaches accounting for ordinality is presented in Tutz (2022).

1.3 Summary of contributions

This dissertation research focuses on exploiting the theoretical advantages of Bayesian nonparametric models to solve methodologically and practically relevant ordinal regression problems. The primary outcomes are Bayesian nonparametric models for cross-sectional or longitudinal ordinal regression, featuring flexibility and efficiency. We present thorough studies of the proposed models' properties. For all proposed models, we discuss approaches to prior specification and develop algorithms for conducting posterior inference. The methodologies are illustrated through a variety of data examples.

We begin in Chapter 2 with models for cross-sectional ordinal regression. We develop a nonparametric Bayesian modeling approach based on priors placed directly on the discrete distribution of the ordinal responses. The prior probability models are built from a structured mixture of multinomial distributions. We leverage the continuation-ratio logits representation to formulate the mixture kernel, with mixture weights defined through the logit stick-breaking process (Rigon and Durante, 2021) that incorporates the covariates through a linear function. The flexibility of the nonparametric mixture model is demonstrated by studying its Kullback-Leibler support. Moreover, we design an efficient and relatively easy to implement posterior simulation method, which also allows partial parallel sampling for category-specific parameters, prompting additional computational efficiency gains.

In Chapter 3, we explore modeling approaches for clustered ordinal responses, which arise, for instance, in developmental toxicology studies. Extra modeling

challenges emerge due to the extensive heterogeneity from various sources. Using data from a developmental toxicity experiment, we examine a spectrum of models, and demonstrate that flexibility is the key for reliable risk assessment. Notably, the nonparametric mixture models developed in Chapter 2 outshine traditional Bayesian hierarchical models in delivering coherent uncertainty quantification. The nonparametric models are then amplified with an overdispersed kernel, which offers enhanced control of variability. The models are illustrated and contrasted in drawing a series of inferences relevant to the toxicity study.

Another class of correlated ordinal responses arises from longitudinal studies, where the primary focus is on accommodating the temporal dependence. We present in Chapter 4, a flexible and efficient modeling approach for analyzing the dynamic evolution of the ordinal responses over time. This can be viewed as a longitudinal ordinal regression problem, where time is the only covariate. We tackle the problem from a functional data analysis perspective, treating the observations for each subject as realizations from subject-specific stochastic processes at the measured times. Leveraging the continuation-ratio logits representation, we model the discrete space processes through a sequence of continuous space processes. We utilize a hierarchical framework to model the mean and covariance kernel of the continuous space processes nonparametrically and simultaneously through a Gaussian process prior and an Inverse-Wishart process prior, respectively. The prior structure results in flexible inference for the evolution and correlation of ordinal responses, while allowing for borrowing of strength across all subjects.

The proposed method for longitudinal ordinal responses is particularly well-suited to problems in evolutionary biology. Chapter 5 delves into a detailed data illustration in this field. In particular, the data is taken from an experiment in which the maturity status of sheepshead minnows is recorded on ordinal scale

over several weeks. This component of the thesis is particularly compelling as it showcases the benefits of the flexible and efficient methodology in the context of a scientifically relevant problem.

The rest of the dissertation is organized as follows. Chapter 2 is devoted to nonparametric mixture models for cross-sectional ordinal regression. The utilization of these models, along with their extensions, in risk assessment for developmental toxicity studies is examined in Chapter 3. Turning to longitudinal settings, Chapter 4 presents a modeling approach for dynamic evolution of ordinal responses. It is followed by Chapter 5, in which a detailed analysis of a data set on sheepshead minnows maturation is provided to illustrate the practical benefits of the proposed model. We conclude with some future perspectives and remarks in Chapter 6. Technical details on proofs of theoretical results, and posterior simulation methods are provided in Appendices A and B, respectively.

Chapter 2

A Structured Mixture Modeling Framework for Cross-sectional Ordinal Regression

2.1 Introduction

Ordinal responses are widely encountered in many fields, including econometrics and the biomedical and social sciences, typically accompanied by covariate information. Hence, estimation and prediction of ordinal regression relationships remains a methodologically and practically relevant problem. The typical ordinal regression setting consists of a univariate ordinal response Y with C categories, and a covariate vector $\mathbf{x}$. The modeling challenge for the ordinal regression problem involves capturing general regression relationships in the response probabilities (especially for moderate to large C), while at the same time appropriately accounting for the ordinal nature of the response.

A commonly used approach involves cumulative link models (e.g., Agresti,

2013), under which the ordinal responses can be viewed as a discretized version of latent continuous responses, typically assumed normally distributed resulting in popular cumulative probit models. For Bayesian inference, such data augmentation facilitates posterior simulation (Albert and Chib, 1993). However, probit models preclude a flexible analysis of probability response curves, since covariate effects enter linearly and additively, and the normality assumption implies restrictions on the marginal response probabilities (e.g., Boes and Winkelmann, 2006). In general, parametric ordinal regression models sacrifice flexibility in the response distribution and/or the regression functions for the response probabilities.

To overcome such limitations, early work in the Bayesian nonparametrics literature has explored semiparametric models, focusing mostly on binary regression. Such methods relax parametric assumptions for the distribution of the latent variables (e.g., Basu and Chib, 2003) or for the regression function (e.g., Choudhuri et al., 2007). As a further extension, Chib and Greenberg (2010) modeled covariate effects additively by cubic splines, combined with a scale normal mixture for the latent responses, using the Dirichlet process (DP) prior (Ferguson, 1973) for the mixing distribution. More general DP mixture priors for the distribution of the latent continuous responses have been considered in Bao and Hanson (2015) and DeYoreo and Kottas (2018a). The latter involves a fully nonparametric Bayesian method under the density regression framework, modeling the joint distribution of covariates and latent responses with a DP mixture of multivariate normals. DeYoreo and Kottas (2020) provide a review of the joint response-covariate modeling approach with categorical variables. The density regression modeling framework is appealing with regard to the scope of ordinal regression inferences. However, it involves computationally intensive posterior simulation which does not scale with the number of covariates, and it is not suitable for

applications where the assumption of random covariates is not relevant.

The "logits regression family" (Agresti, 2013) offers an alternative approach to ordinal regression, based on direct modeling of the response distribution. Of particular interest to our methodology are continuation-ratio logits models. The continuation-ratio logits parameterization of the multinomial distribution implies a sequential mechanism, such that the ordinal response is determined through a sequence of binary outcomes. Starting from the lowest category, each binary outcome indicates whether the ordinal response belongs to that category or to one of the higher categories. The continuation-ratio logit for response category j is the logit of the conditional probability of response j, given that the response is j or higher. A key consequence is that, in a multinomial continuation-ratio logits regression model, the response distribution can be factorized into complete conditionals defined by Binomial logistic regression models.

To our knowledge, continuation-ratio logits have not been explored for general Bayesian nonparametric methods for ordinal regression. For nominal regression, Linderman et al. (2015) discussed a semiparametric model that, under the multinomial response distribution, replaces the linear covariate effects within the continuation-ratio logits by Gaussian process priors. More relevant to our methodology is the dependent DP mixture model in Kottas and Fronczyk (2013), based on a trinomial kernel that builds from the continuation-ratio logits formulation. This modeling approach was developed specifically in the context of developmental toxicity studies, rather than for general ordinal regression problems.

The continuation-ratio logits structure is attractive as a building block for general nonparametric Bayesian ordinal regression modeling, and this is the primary motivation for our methodology. We build the response distribution from a nonparametric mixture of multinomial distributions, mixing on the regression

coefficients under the continuation-ratio logits formulation for the mixture kernel. Model flexibility is enhanced through covariate dependent mixture weights, assigned a logit stick-breaking prior (Rigon and Durante, 2021). The stick-breaking structure, along with the logistic form for the underlying covariate dependent variables, yields a continuation-ratio logits regression representation also for the mixture weights. The similarity in the structure of the mixture kernel and the mixture weights is a distinguishing feature of the methodology, in terms of model properties and model implementation. We take advantage of this structure, as well as a latent variable model formulation, to explore the Kullback-Leibler support of the prior probability model.

Regarding model implementation, using the Pólya-Gamma data augmentation approach for logistic regression (Polson et al., 2013), we design an efficient Gibbs sampling algorithm for posterior inference. The posterior simulation method is ready to implement, in particular, it does not require specialized techniques or tuning of Metropolis-Hastings steps. Moreover, the product of Binomials formulation of the multinomial kernel yields a Gibbs sampler which, given all other model parameters, allows for separate updates for each set of mixture kernel parameters associated with each response category. Hence, the complexity of the inference procedure is not unduly increasing with the number of response categories.

The model yields flexible probability response curves expressed as weighted sums of parametric regression functions with local, covariate-dependent weights. As simplified versions of the general model structure, we explore mixture models that incorporate the covariates only in the kernel parameters or only in the weights. We study model properties and use synthetic and real data examples to compare the different model formulations.

Our objective is to develop a general toolbox for ordinal regression that allows flexibility in both the response distribution and the ordinal regression relationships. The toolbox comprises models of different complexity, all of which can be implemented with relatively straightforward posterior simulation methods. It also includes prior specification methods that range from a fairly non-informative choice to more informative options that enable incorporation of monotonicity trends for the probability response functions.

The rest of the chapter is organized as follows. In Section 2.2, we formulate the general modeling approach, and discuss prior specification, posterior inference, and model properties, including Kullback-Leibler support (with technical details given in Appendix A.1). Section 2.3 presents the two simplified mixture models. The methodology is illustrated in Section 2.4 with synthetic and real data examples. Section 2.5 concludes with discussion.

2.2 General Methodology

2.2.1 From Building Blocks to General Model

Consider an ordinal response Y with C categories, recorded along with a covariate vector $\mathbf{x} \in \mathbb{R}^p$. We can equivalently encode the response as a vector of binary variables $\mathbf{Y} = (Y_1, \ldots, Y_C)$, such that $Y = j$ is equivalent to $Y_j = 1$ and $Y_k = 0$ for any $k \neq j$.

The continuation-ratio logits regression model builds from the factorization of the multinomial distribution in terms of Binomial distributions,

$$Mult(\mathbf{Y} \mid 1, \pi_1, \ldots, \pi_C) \;=\; Bin(Y_1 \mid m_1, \varphi(\theta_1)) \ldots Bin(Y_{C-1} \mid m_{C-1}, \varphi(\theta_{C-1})), \tag{2.1}$$

where $m_1 = 1$, and $m_j = 1 - \sum_{k=1}^{j-1} Y_k$, for $j = 2, \ldots, C-1$, $\theta_j \equiv \theta_j(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}_j$, and $\varphi(\theta) = \exp(\theta)/\{1 + \exp(\theta)\}$ denotes the standard logistic function. The two parameterizations are linked through $\pi_1 = \varphi(\theta_1)$, $\pi_j = \varphi(\theta_j) \prod_{k=1}^{j-1} \{1 - \varphi(\theta_k)\}$, for $j = 2, \ldots, C-1$, and $\pi_C = \prod_{k=1}^{C-1} \{1 - \varphi(\theta_k)\}$. For notation simplicity, we use $K(\mathbf{Y} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{C-1})$, for the continuation-ratio logits representation of the multinomial distribution.

The parametric model is limited in the response distribution and the form of covariate effects. A strategy that surpasses these limitations and achieves flexible inference is to generalize the model via Bayesian nonparametric mixing. Using the kernel function in (2.1) in conjunction with a nonparametric prior for the covariate-dependent mixing distribution, we achieve the general nonparametric extension of the continuation-ratio logits model,

$$\mathbf{Y} \mid G_{\mathbf{x}} \sim \int K(\mathbf{Y} \mid \boldsymbol{\theta}) \, dG_{\mathbf{x}}(\boldsymbol{\theta}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \, K(\mathbf{Y} \mid \boldsymbol{\theta}_\ell(\mathbf{x})). \tag{2.2}$$

Here, the countable mixture form emerges under the nonparametric prior formulation for the mixing distribution that represents it as a discrete distribution, $G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \, \delta_{\boldsymbol{\theta}_\ell(\mathbf{x})}$, with covariate-dependent atoms, $\boldsymbol{\theta}_\ell(\mathbf{x})$, and weights, $\omega_\ell(\mathbf{x})$.

The prior formulation for $G_{\mathbf{x}}$ in (2.2) is generic. There are several options for building the model for the atoms and weights, a stick-breaking formulation for the latter being the more common strategy. The dependent DP (DDP) prior and related models (MacEachern, 2000; Quintana et al., 2022) has been explored in different applications, including simplified "common-weights" or "common-atoms" versions under which only the atoms or the weights, respectively, depend on the covariates. Other options include the kernel stick-breaking process (Dunson and Park, 2008), the probit stick-breaking process (Dunson and Rodríguez, 2011), and the logit stick-breaking process (Rigon and Durante, 2021).

As discussed below, for the ordinal regression problem with mixture kernel $K(\mathbf{Y} \mid \boldsymbol{\theta})$, the logit stick-breaking process (LSBP) prior offers a key advantage in model structure and in posterior simulation. Therefore, for the general model in (2.2), we assume the following LSBP prior for the covariate-dependent weights:

$$\omega_1(\mathbf{x}) = \varphi(\mathbf{x}^T\boldsymbol{\gamma}_1),\ \omega_\ell(\mathbf{x}) = \varphi(\mathbf{x}^T\boldsymbol{\gamma}_\ell)\prod_{h=1}^{\ell-1}(1-\varphi(\mathbf{x}^T\boldsymbol{\gamma}_h)),\ \ell \geq 2; \quad \boldsymbol{\gamma}_\ell \overset{i.i.d.}{\sim} N(\boldsymbol{\gamma}_0, \Gamma_0) \tag{2.3}$$

In addition, the atoms, $\boldsymbol{\theta}_\ell(\mathbf{x}) = (\theta_{1\ell}(\mathbf{x}), \ldots, \theta_{C-1,\ell}(\mathbf{x}))$, are built through a linear regression structure,

$$\theta_{j\ell}(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}_{j\ell} \mid \boldsymbol{\mu}_j, \Sigma_j \overset{ind.}{\sim} N(\mathbf{x}^T\boldsymbol{\mu}_j, \mathbf{x}^T\Sigma_j\mathbf{x}), \quad j = 1, \ldots, C-1,\ \ell \geq 1, \tag{2.4}$$

with the random variables that define the atoms assumed a priori independent of those that define the weights. The model is completed with the conjugate prior for the collection of hyperparameters $\boldsymbol{\psi} = \{\boldsymbol{\mu}_j, \Sigma_j\}_{j=1}^{C-1}$, that is,

$$\Sigma_j \overset{ind.}{\sim} IW(\nu_{0j}, \Lambda_{0j}^{-1}), \quad \boldsymbol{\mu}_j \mid \Sigma_j \overset{ind.}{\sim} N(\boldsymbol{\mu}_{0j}, \Sigma_j/\kappa_{0j}), \quad j = 1, \ldots, C-1. \tag{2.5}$$

In Section 2.2.3, we discuss prior specification for $\{\nu_{0j}, \Lambda_{0j}, \boldsymbol{\mu}_{0j}, \kappa_{0j}\}_{j=1}^{C-1}$, and for $\boldsymbol{\gamma}_0, \Gamma_0$.

To point to the benefit of working with the LSBP prior, we examine the continuation-ratio logits structure in (2.1). As illustrated in Figure 2.1, such structure implies a sequential mechanism in determining the ordinal response Y. At a generic step j, a Bernoulli variable $\mathcal{H}_j \sim Bern(\Delta_j)$ is generated to either set $Y = j$ if $\mathcal{H}_j = 1$, or to allocate Y to $\{k : k > j\}$ when $\mathcal{H}_j = 0$. The j-th step can only be reached if Y has not been assigned to $1, \ldots, j-1$. To bring in the covariate effects, we place a logit-normal prior on Δ_j, that is, $\Delta_j = \varphi(\mathbf{x}^T\boldsymbol{\beta}_j)$

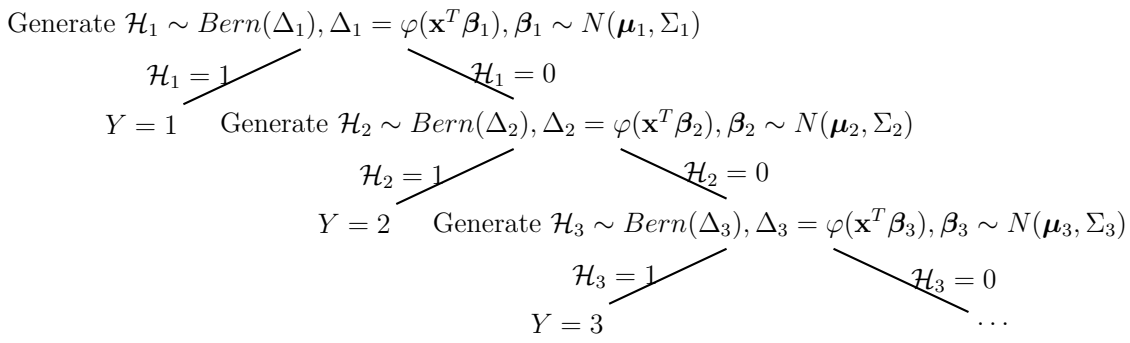


Figure 2.1: Illustration of the continuation-ratio logits structure.

and $\boldsymbol{\beta}_j \sim N(\boldsymbol{\mu}_j, \Sigma_j)$. This procedure provides a natural way of defining a stick-breaking process, engendering the LSBP as mentioned in Rigon and Durante (2021). Consider a configuration variable $\boldsymbol{\mathcal{L}}$, corresponding to $\mathbf{Y}$, that indicates the mixture component in (2.2) from which $\mathbf{Y}$ is generated. The same sequential generative process applies to $\boldsymbol{\mathcal{L}}$. At generic step ℓ, a Bernoulli variable $\mathcal{H}^*_\ell \sim Bern(\eta_\ell)$ is generated, serving the same role as $\mathcal{H}_j$ in determining whether $\boldsymbol{\mathcal{L}}$ locates at the current stage, or moves to later stages. Treating η_ℓ as the stick-breaking proportion, the covariate effects are incorporated through $\eta_\ell(\mathbf{x}) = \varphi(\mathbf{x}^T\boldsymbol{\gamma}_\ell)$. The resulting nonparametric model admits the countable mixture representation in (2.2), with weights and atoms depending on covariates in a similar fashion. We highlight this correspondence because it paves the way in developing tractable posterior inference strategies, which will be discussed in Section 2.2.4.

In this section, we consider properties under the general model formulation in (2.2) comprising the covariate-dependent weights and atoms in (2.3) and (2.4), respectively. In Section 2.3, we discuss the simpler common-weights and common-atoms models as a means to address the trade-off between the flexibility of model (2.2) and its potential computational cost. Our study of model properties and data illustrations explore such trade-off and suggest scenarios for which the simpler models may be suitable.

2.2.2 Model Properties

The covariate-response relationship can be studied through the marginal probability response curves $\Pr(\mathbf{Y} = j \mid G_{\mathbf{x}})$, for $j = 1, \ldots, C$. Given the ordinal nature of the response, also of interest are the conditional probability response curves, $\Pr(\mathbf{Y} = j \mid \mathbf{Y} \geq j, G_{\mathbf{x}})$. Here, we slightly abuse notation by writing $\mathbf{Y} = j$, while it is actually $\mathbf{Y} = \mathbf{1}_j$, the unit vector in $\mathbb{R}^C$ with the jth element equal to 1.

Based on the particular mixture of multinomial distributions for the general model in (2.2), the marginal probability response curve for $j = 1, \ldots, C$ can be expressed as

$$\Pr(\mathbf{Y} = j \mid G_{\mathbf{x}}) \;=\; \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \; \left\{ \varphi(\theta_{j\ell}(\mathbf{x})) \prod_{k=1}^{j-1} [1 - \varphi(\theta_{k\ell}(\mathbf{x}))] \right\}, \tag{2.6}$$

where the weights, $\omega_\ell(\mathbf{x})$, and atoms, $\theta_{j\ell}(\mathbf{x})$, are defined in (2.3) and (2.4), respectively, and we set $\varphi(\theta_{C\ell}(\mathbf{x})) \equiv 1$. Moreover, the conditional probability response curves are given by

$$\begin{aligned} \Pr(\mathbf{Y} = j \mid \mathbf{Y} \geq j, G_{\mathbf{x}}) \;&=\; \sum_{\ell=1}^{\infty} w_{j\ell}(\mathbf{x}) \, \varphi(\theta_{j\ell}(\mathbf{x})); \\ w_{j\ell}(\mathbf{x}) \;&=\; \frac{\omega_\ell(\mathbf{x}) \prod_{k=1}^{j-1} [1 - \varphi(\theta_{k\ell}(\mathbf{x}))]}{\sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \prod_{k=1}^{j-1} [1 - \varphi(\theta_{k\ell}(\mathbf{x}))]} \end{aligned} \tag{2.7}$$

Both the marginal and conditional probability response curves admit a weighted sum representation with component regression functions that correspond to the parametric continuation-ratio logits model. The covariate-dependent weights in (2.6) and (2.7) allow for local adjustment over the covariate space, thus enabling non-standard regression relationships and relaxing the restrictions on the covariate effects under the parametric model.

A useful observation is that the continuation-ratio logits model plays the role

of a parametric backbone for the nonparametric model, in the sense of prior expectation. More specifically, using (2.6), and the assumptions of the prior model in (2.2), (2.3) and (2.4),

$$\begin{aligned} \mathrm{E}(\Pr(\mathbf{Y} = j \mid G_{\mathbf{x}})) &= \sum_{\ell=1}^{\infty} \mathrm{E}(\omega_\ell(\mathbf{x}))\, \mathrm{E}\left\{\varphi(\theta_{j\ell}(\mathbf{x})) \textstyle\prod_{k=1}^{j-1}[1-\varphi(\theta_{k\ell}(\mathbf{x}))]\right\} \\ &= \mathrm{E}\left\{\varphi(\mathbf{x}^T\boldsymbol{\beta}_j) \textstyle\prod_{k=1}^{j-1}[1-\varphi(\mathbf{x}^T\boldsymbol{\beta}_k)]\right\}, \end{aligned} \tag{2.8}$$

where the last expectation is taken with respect to $\boldsymbol{\beta}_j \overset{ind.}{\sim} N(\boldsymbol{\mu}_j, \Sigma_j)$, $j = 1, \ldots, C-1$. Hence, the prior expectation for the marginal probability response curves under the nonparametric model reduces to the prior expectation under the parametric model. This property facilitates prior specification, as discussed in Section 2.2.3.

The general model can capture a spectrum of inferences, with the parameters $\boldsymbol{\gamma}_\ell$ controling the number of effective mixture components. Suppose the covariates take values in a bounded region. If $\boldsymbol{\gamma}_1$ results in $\varphi(\mathbf{x}^T\boldsymbol{\gamma}_1)$ effectively equal to one, then the nonparametric model collapses to its parametric backbone. On the other hand, if the first several $\boldsymbol{\gamma}_\ell$ are such that $\varphi(\mathbf{x}^T\boldsymbol{\gamma}_\ell)$ are relatively small, a larger number of effective components is favored, in the extreme utilizing a distinct multinomial component for each ordinal response. In practice, the strength of the nonparametric model lies between these two extremes.

2.2.3 Prior Specification

To implement the general model in (2.3), (2.4) and (2.5), we need to specify the parameters of the hyperpriors, that is, $(\boldsymbol{\gamma}_0, \Gamma_0)$ and $\{\nu_{0j}, \Lambda_{0j}, \boldsymbol{\mu}_{0j}, \kappa_{0j}\}_{j=1}^{C-1}$.

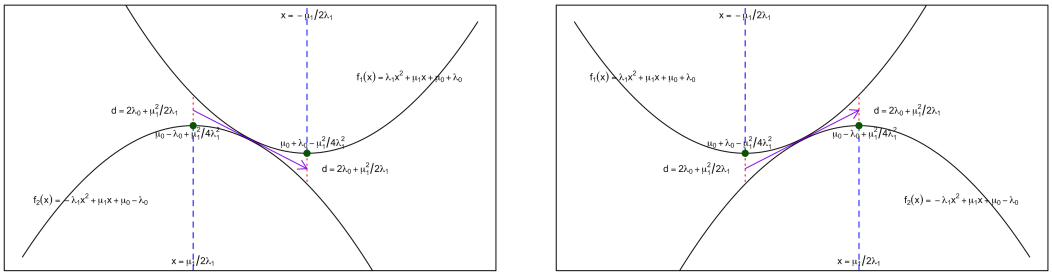
We set $\kappa_{0j} = \nu_{0j} = p + 2$ for all j, where p is the dimension of the covariate vector $\mathbf{x}$ (including the intercept). For the other prior hyperparameters, the proposed strategy is developed by first considering the prior expected probability

response curves to specify $\{\boldsymbol{\mu}_{0j}, \Lambda_{0j}\}_{j=1}^{C-1}$, and then using the prior expected weight placed on each mixing component to determine $\boldsymbol{\gamma}_0$ and Γ_0.

The weights and atoms of the mixture model have the same structure. Specifically, the weights are generated from a stick-breaking process with breaking proportion $\eta_\ell(\mathbf{x}) = \varphi(\mathbf{x}^T\boldsymbol{\gamma}_\ell)$, while the atoms can also be viewed as possessing a stick-breaking form with breaking proportion $\Delta_{j\ell}(\mathbf{x}) = \varphi(\mathbf{x}^T\boldsymbol{\beta}_{j\ell})$. Taking the prior into consideration, we have $\eta_\ell(\mathbf{x}) \sim LN(\mathbf{x}^T\boldsymbol{\gamma}_0, \mathbf{x}^T\Gamma_0\mathbf{x})$ and $\Delta_{j\ell}(\mathbf{x}) \sim LN(\mathbf{x}^T\boldsymbol{\mu}_{0j}, (\kappa_{0j}+1)/(\kappa_{0j}(\nu_{0j}-p-1))\mathbf{x}^T\Lambda_{0j}\mathbf{x})$, where $LN(\cdot,\cdot)$ denotes the logit-normal distribution. Therefore, a key quantity in prior specification is the expectation of a logit-normal distributed random variable, which does not have analytical form in general.

Nonetheless, if $Z \sim N(0, \sigma^2)$, then $\mathrm{E}(\varphi(Z)) = 0.5$, for any value of the variance σ^2 (Pirjol, 2013). This result motivates the default choice of hyperparameters we use in practice, that is, $\boldsymbol{\mu}_{0j} = \boldsymbol{\gamma}_0 = \mathbf{0}_p$, and $\Lambda_{0j} = \Gamma_0 = 10^2\,\mathbf{I}_p$. We refer to this specification as the "baseline" prior, which yields $\mathrm{E}(\Pr(\mathbf{Y} = j \mid G_\mathbf{x})) = 2^{-j}$, for $j = 1, \ldots, C-1$, and $\mathrm{E}(\Pr(\mathbf{Y} = C \mid G_\mathbf{x})) = 2^{-(C-1)}$, for all $\mathbf{x}$. The prior expectation of the weight associated with the ℓth mixing component is given by $2^{-\ell}$, for any ℓ.

In general, both the shape of the prior expected probability response curves and the prior expected weight placed on each mixing component depend on the expectation of the logit-normal distribution. Even though that expectation does not have analytical form, it can be readily obtained by simulation. Therefore, we can tune the prior hyperparameters and evaluate the prior expectation of $\eta_\ell(\mathbf{x})$ and $\Delta_{j\ell}(\mathbf{x})$. For instance, we can favor prior expected probability response curves possessing some specific pattern (such as monotonicity) and/or a certain number of mixture components. The following proposition, which can be obtained using



(a) Monotonically decreasing function. **(b)** Monotonically increasing function.

Figure 2.2: Illustration of how the two bounds can be used to set the monotonic pattern of the prior expected probability response curve.

results from Pirjol (2013), facilitates the tuning of prior hyperparameters.

Proposition 2.1. *If* $Z \sim N(\mu, \sigma^2)$, *then* $\varphi(\mu - \sigma^2/2) \leq E(\varphi(Z)) \leq \varphi(\mu + \sigma^2/2)$.

To illustrate the procedure in detail, we consider a special case where the covariates vector is $\mathbf{x} = (1, x)^T$ and the information is available for the first probability response curve. Suppose the prior hyperparameters are $\boldsymbol{\mu}_{01} = (\mu_{01,0}, \mu_{01,1})^T$ and $\Lambda_{01} = diag(\lambda_{01,0}, \lambda_{01,1})$. In such a case, the prior expected first probability response curve $\mathrm{E}(\Pr(\mathbf{Y} = 1 \mid G_{\mathbf{x}})) = \mathrm{E}(\varphi(\mathbf{x}^T\boldsymbol{\beta}_1))$, where $\mathbf{x}^T\boldsymbol{\beta}_1 \sim N(\mu_{01,0} + \mu_{01,1}x, (\kappa_{01} + 1)/(\kappa_{01}(\nu_{01} - p - 1))(\lambda_{01,0} + \lambda_{01,1}x^2))$. For notation simplicity, let us denote $\mu_s = \mu_{01,s}, \lambda_s = (\kappa_{01} + 1)/(2\kappa_{01}(\nu_{01} - p - 1))\lambda_{01,s}$, $s = 0, 1$. Then from Proposition 2.1, $\mathrm{E}(\Pr(\mathbf{Y} = 1 \mid G_{\mathbf{x}}))$ is bounded by

$$\varphi(-\lambda_1 x^2 + \mu_1 x + \mu_0 - \lambda_0) \leq \mathrm{E}(\Pr(\mathbf{Y} = 1 \mid G_{\mathbf{x}})) \leq \varphi(\lambda_1 x^2 + \mu_1 x + \mu_0 + \lambda_0)$$

Because the logistic function preserves monotonicity, it is helpful to study the relative position of the two parabolas inside. Indeed, we can choose the prior hyperparameters such that the two bounds squeeze a small region. The first prior expected probability response curve pinches through that region, possessing certain monotonicity, illustrated in Figure 2.2.

Specifically, suppose the prior guess for the first probability response curve is a decreasing function with respect to x. As shown in Figure 2.2a, we can put the range of x inside the two axes of symmetry. In addition, the quantity $d = 2\lambda_0 + \mu_1^2/2\lambda_1$ determines the maximum difference of the two bounds. The two vertices determine the prior mean at the minimum and maximum value of x. To summarize, the parameters $\mu_0, \mu_1, \lambda_0, \lambda_1$ can be specified by the equations

$$\begin{cases} \dfrac{\mu_1}{2\lambda_1} = a_1, \quad -\dfrac{\mu_1}{2\lambda_1} = -a_1 \\ 2\lambda_0 + \dfrac{\mu_1^2}{2\lambda_1} = a_2 \\ \mu_0 + \lambda_0 - \dfrac{\mu_1^2}{4\lambda_1} = -a_3 \\ \mu_0 - \lambda_0 + \dfrac{\mu_1^2}{4\lambda_1} = a_4 \end{cases} \iff \begin{cases} \mu_0 = \dfrac{a_4 - a_3}{2} \\ \mu_1 = -\dfrac{a_2 + a_3 + a_4}{2a_1} \\ \lambda_0 = \dfrac{a_2 - a_3 - a_4}{4} \\ \lambda_1 = \dfrac{a_2 + a_3 + a_4}{4a_1^2} \end{cases} \tag{2.9}$$

with positive numbers a_1, a_2, a_3, a_4 chosen based on the prior information. Note that λ_0 should be positive, so it imposes the constraint $a_2 > a_3 + a_4$ on the choice of these four numbers. Using (2.9), we can specific the prior hyperparameters $\mu_{01,0}, \mu_{01,1}, \lambda_{01,0}, \lambda_{01,1}$. The same strategy can be extended for the monotonically increasing case.

To specify $\boldsymbol{\mu}_{0j}$ and Λ_{0j} for $j > 1$, we can sequentially implement this strategy. Furthermore, if the dimension of covariates $p > 2$, it becomes more difficult to specify hyperparameters, but the same strategy can be applied by considering each covariate $x_s, s = 1, \ldots, p$ marginally while fixing $x_{s'}, s' \neq s$.

As a concrete example, consider an ordinal response with $C = 3$ categories, and a single covariate taking values in $(-10, 10)$. Suppose the prior information is that the first marginal probability response function is decreasing from 1 to 0, whereas the second is increasing from 0 to 1. For the first decreasing probability curve, we set $a_1 = a_2 = 10, a_3 = 6, a_4 = 2$ to specify $\boldsymbol{\mu}_{01}$ and Λ_{01}. As for the second

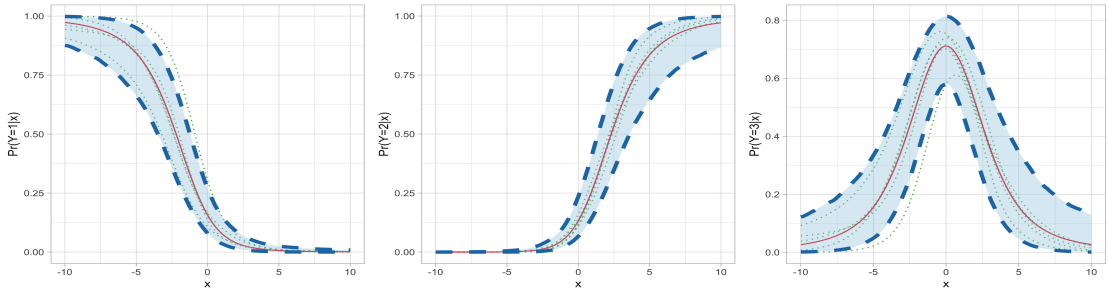


Figure 2.3: Illustration of the prior specification strategy. In each panel, the red solid line is the prior expected probability response curve, the blue dashed lines and shaded region indicate the prior 95% interval estimate, and the green dotted lines show 5 prior realizations.

probability curve, since $\mathrm{E}(\Pr(\mathbf{Y} = 2 \mid G_{\mathbf{x}})) = [1 - \mathrm{E}(\Pr(\mathbf{Y} = 1 \mid G_{\mathbf{x}}))]\mathrm{E}[\varphi(\mathbf{x}^T\boldsymbol{\beta}_2)]$ and utilizing the specified monotonicity for $\mathrm{E}(\Pr(\mathbf{Y} = 1 \mid G_{\mathbf{x}}))$, we focus on $\mathrm{E}[\varphi(\mathbf{x}^T\boldsymbol{\beta}_2)]$. To force a increasing trend, we further choose $\boldsymbol{\mu}_{02}$ and Λ_{02} by applying the strategy for the increasing case with same setting on a_1 to a_4. After solving the corresponding equations for hyperparameters, we obtain Figure 2.3, which shows point and interval estimates that reflect such prior information, with a fair amount of variability.

2.2.4 Posterior Inference

For Markov chain Monte Carlo (MCMC) posterior simulation, we work with a truncation approximation of the mixing distribution in the spirit of blocked Gibbs sampling for stick-breaking priors (Ishwaran and James, 2001). We favor the blocked Gibbs sampler as it results in practical model implementation and it allows for full posterior inference for general regression functionals. Hence, for posterior simulation, the mixing distribution $G_{\mathbf{x}}$ in (2.2) is replaced by $G^L_{\mathbf{x}} = \sum_{\ell=1}^{L} p_\ell(\mathbf{x})\, \delta_{\boldsymbol{\theta}_\ell(\mathbf{x})}$, with $\boldsymbol{\theta}_\ell(\mathbf{x})$ defined as before, and $p_\ell(\mathbf{x}) = \omega_\ell(\mathbf{x})$, for $\ell = 1, ..., L-1$, whereas $p_L(\mathbf{x}) = 1 - \sum_{\ell=1}^{L-1} p_\ell(\mathbf{x})$.

The truncation level L can be chosen to achieve any desired level of accuracy.

For normal mixtures with LSBP weights, Rigon and Durante (2021) show that, for fixed sample size and covariates, the L^1 distance between the prior predictive distribution of the sample under $G_{\mathbf{x}}$ and $G_{\mathbf{x}}^L$ decreases exponentially in L. The proof for this result (Theorem 1 in Rigon and Durante (2021)) applies to essentially any mixture kernel, and it thus also holds for the multinomial LSBP mixture model defined in (2.2), (2.3) and (2.4).

In practice, we can specify L using the prior expectation for the partial sum of weights. Under the prior in (2.3), $\mathrm{E}(\sum_{\ell=1}^{L} \omega_\ell(\mathbf{x})) = 1 - \{1 - \mathrm{E}(\varphi(\mathbf{x}^T\boldsymbol{\gamma}))\}^L$, where the expectation on the right-hand-side is with respect to $\boldsymbol{\gamma} \sim N(\boldsymbol{\gamma}_0, \Gamma_0)$. Hence, L can be selected by computing the expectation at a few representative values in the covariate space. Note that, when $\boldsymbol{\gamma}_0 = \mathbf{0}_p$, $\mathrm{E}(\varphi(\mathbf{x}^T\boldsymbol{\gamma})) = 0.5$, for any $\mathbf{x}$. We also recommend monitoring the posterior samples for $p_L(\mathbf{x})$ for different values $\mathbf{x}$ in the covariate space. Using a combination of such strategies, we worked with the (conservative) truncation level of $L = 50$ for the data examples of Section 2.4.

Denote by $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iC})$, where $Y_{ij} \in \{0, 1\}$ with $\sum_{j=1}^{C} Y_{ij} = 1$, the ith observed response, and by $\mathbf{x}_i$ the corresponding covariate vector, for $i = 1, \ldots, n$. We introduce latent configuration variables, $\{\mathcal{L}_i\}$, such that $\mathcal{L}_i = \ell$ if and only if $\mathbf{Y}_i$ is assigned to the ℓth mixture component. Then, the hierarchical model for the data can be expressed as

$$
\begin{aligned}
\mathbf{Y}_i \mid \{\boldsymbol{\beta}_{j\ell}\}, \mathcal{L}_i &\overset{ind.}{\sim} K(\mathbf{Y}_i \mid \boldsymbol{\theta}_{\mathcal{L}_i}) = \prod_{j=1}^{C-1} Bin(Y_{ij} \mid m_{ij}, \varphi(\mathbf{x}_i^T\boldsymbol{\beta}_{j\mathcal{L}_i})), \;\; i = 1, \ldots, n \\
\mathcal{L}_i \mid \{\boldsymbol{\gamma}_\ell\} &\overset{ind.}{\sim} \sum_{\ell=1}^{L} p_{i\ell}\, \delta_\ell(\mathcal{L}_i), \;\; i = 1, \ldots, n \\
\boldsymbol{\beta}_{j\ell} \mid (\boldsymbol{\mu}_j, \Sigma_j) &\overset{ind.}{\sim} N(\boldsymbol{\mu}_j, \Sigma_j), \;\; j = 1, \ldots, C-1, \;\; \ell = 1, \ldots, L \\
\boldsymbol{\gamma}_\ell &\overset{i.i.d.}{\sim} N(\boldsymbol{\gamma}_0, \Gamma_0), \;\; \ell = 1, \ldots, L-1 \\
(\boldsymbol{\mu}_j, \Sigma_j) &\overset{ind.}{\sim} N(\boldsymbol{\mu}_j \mid \boldsymbol{\mu}_{0j}, \Sigma_j/\kappa_{0j})\, IW(\Sigma_j \mid \nu_{0j}, \Lambda_{0j}^{-1}), \;\; j = 1, \ldots, C-1
\end{aligned}
\tag{2.10}
$$

where $m_{i1} = 1$, $m_{ij} = 1 - \sum_{k=1}^{j-1} Y_{ik}$, for $j = 2, \ldots, C-1$, $p_{i\ell} = \varphi(\mathbf{x}_i^T\boldsymbol{\gamma}_\ell) \prod_{h=1}^{\ell-1}(1 -$

$\varphi(\mathbf{x}_i^T\boldsymbol{\gamma}_h))$, for $\ell = 1, \ldots, L-1$, and $p_{iL} = \prod_{\ell=1}^{L-1}(1 - \varphi(\mathbf{x}_i^T\boldsymbol{\gamma}_\ell))$.

Akin to the ordinal response $\mathbf{Y}_i$ and its original form Y_i, we can view the latent configuration variable $\boldsymbol{\mathcal{L}}_i$ as the allocation of its multivariate form $\boldsymbol{\mathcal{L}}_i = (\mathcal{L}_{i1}, \ldots, \mathcal{L}_{iL}) \in \mathbb{R}^L$, with the connection defined as $\boldsymbol{\mathcal{L}}_i = \ell \iff \boldsymbol{\mathcal{L}}_i = \mathbf{1}_\ell$, the unit vector in $\mathbb{R}^L$ with the ℓth element equal to 1. An important observation is that the prior model for the $\boldsymbol{\mathcal{L}}_i$ in (2.10) can be equivalently defined through a continuation-ratio logits regression model for their multivariate images $\boldsymbol{\mathcal{L}}_i$. More specifically,

$$\boldsymbol{\mathcal{L}}_i \mid \{\boldsymbol{\gamma}_\ell\} \overset{ind.}{\sim} Bin(\mathcal{L}_{i1} \mid 1, \eta_1(\mathbf{x}_i)) \times Bin(\mathcal{L}_{i2} \mid 1 - \mathcal{L}_{i1}, \eta_2(\mathbf{x}_i)) \times \ldots$$
$$\times Bin\left(\mathcal{L}_{i,L-1} \mid 1 - \sum_{k=1}^{L-2} \mathcal{L}_{ik}, \eta_{L-1}(\mathbf{x}_i)\right)$$

where $\eta_\ell(\mathbf{x}_i) = \varphi(\mathbf{x}_i^T\boldsymbol{\gamma}_\ell)$, for $\ell = 1, \ldots, L-1$.

The form of the hierarchical model for the data, along with the observation above, elucidate the key model property discussed in Section 2.2.1. Under the (truncated) LSBP prior for the covariate-dependent weights, we achieve effectively the same structure for the weights and atoms of the general mixture model. In turn, this allows us to use the Pólya-Gamma data augmentation approach (Polson et al., 2013) to update both the atoms parameters as well as the ones for the weights. In particular, for each response $\mathbf{Y}_i$, we introduce two sets of Pólya-Gamma latent variables, such that conditionally conjugate updates emerge for the parameters defining both the weights and the atoms. Therefore, all model parameters can be updated via Gibbs sampling. Moreover, taking advantage of the continuation-ratio logits model structure for the mixture kernel, parallel computing for the different mixing components can be adopted, facilitating implementation in applications where the number of response categories is moderate to large. Details of the posterior simulation method are presented in Appendix B.1.

Using the posterior samples for model parameters, we can obtain full inference for any regression functional of interest. The MCMC posterior samples can also be used to estimate the posterior predictive distribution for new response $\mathbf{Y}_*$ given new covariate vector $\mathbf{x}_*$. Using superscript (t) to indicate the tth posterior sample for the model parameters, the tth posterior predictive sample is obtained by first sampling the corresponding configuration variable $\mathcal{L}_*^{(t)}$, such that $\mathcal{L}_*^{(t)} = \ell$ with probability $\varphi(\mathbf{x}_*^T \boldsymbol{\gamma}_\ell^{(t)}) \prod_{h=1}^{\ell-1} (1 - \varphi(\mathbf{x}_*^T \boldsymbol{\gamma}_h^{(t)}))$, for $\ell = 1, \ldots, L-1$ (and $\mathcal{L}_*^{(t)} = L$ with the remaining probability), and then sampling $\mathbf{Y}_*^{(t)}$ from $K(\cdot \mid \boldsymbol{\theta}_*^{(t)})$, with the jth element of $\boldsymbol{\theta}_*^{(t)}$ given by $\varphi(\mathbf{x}_*^T \boldsymbol{\beta}_{j\mathcal{L}_*^{(t)}}^{(t)})$, for $j = 1, \ldots, C-1$.

2.2.5 Assessing Model Flexibility

In Section 2.2.5.2, we study the Kullback-Leibler (KL) support of the proposed prior model, using results from Barrientos et al. (2012) for nonparametric mixtures for continuous responses. This study yields two results of independent interest: a formulation of the ordinal LSBP mixture model in terms of latent continuous responses (Section 2.2.5.1); and, a connection between the KL support of a prior for continuous responses and the induced prior for categorical outcomes arising from discretizing the continuous responses. Moreover, in Section 2.2.5.3, we contrast continuation-ratio logit and cumulative logit models. The purpose is to provide further motivation for the kernel choice of the LSBP mixture model. The proofs for all theoretical results are given in Appendix A.1.

2.2.5.1 Latent Variable Representation

Recall that the continuation-ratio logits structure implies a sequential mechanism involving binary steps to determine which of its C levels the ordinal response admits. The mechanism can also be represented through latent continuous vari-

ables, $\boldsymbol{\mathcal{Z}} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_{C-1})$, and a sequential, binary partition of $\mathbb{R}^{C-1}$, comprising sets

$$\mathcal{R}_1 = \mathbb{R}^+ \times \mathbb{R}^{C-2}; \;\; \mathcal{R}_j = (\mathbb{R}^-)^{j-1} \times \mathbb{R}^+ \times \mathbb{R}^{C-j-1}, \; j = 2, \ldots, C-1; \;\; \mathcal{R}_C = (\mathbb{R}^-)^{C-1}. \tag{2.11}$$

Hence, $\boldsymbol{\mathcal{Z}} \in \mathcal{R}_j$ if its first $j-1$ elements take negative values, and the j component is positive valued. Referring to the description of the continuation-ratio logits structure from Section 2.2.1, variable $\mathcal{Z}_j$ plays a similar role to Bernoulli variable $\mathcal{H}_j$, where now it is the sign of $\mathcal{Z}_j$ that specifies the ordinal response category, such that $\mathbf{Y} = j$ if-f $\mathcal{Z}_j > 0$, given that $\mathcal{Z}_k \leq 0$, for $k = 1, \ldots, j-1$. As stated in the following proposition, the multinomial model in (2.1) emerges for independent logistic variables $\mathcal{Z}_j$.

Proposition 2.2. *Consider ordinal response* $\mathbf{Y} = (Y_1, \ldots, Y_C)$*, where* $Y_j \in \{0, 1\}$ *with* $\sum_{j=1}^{C} Y_j = 1$*, and continuous random vector* $\boldsymbol{\mathcal{Z}} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_{C-1}) \in \mathbb{R}^{C-1}$*. Assume that:* $\mathbf{Y} \mid \boldsymbol{\mathcal{Z}} \sim \mathbf{1}(\mathbf{Y} = j \iff \boldsymbol{\mathcal{Z}} \in \mathcal{R}_j)$*, for* $j = 1, \ldots, C$*, with the* $\mathcal{R}_j$ *defined in (2.11); and,* $\boldsymbol{\mathcal{Z}} \mid \boldsymbol{\theta} \sim \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_j)$*, where* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{C-1})$*, and* $\mathfrak{L}(\cdot \mid \theta)$ *denotes the logistic distribution with mean* θ *and scale parameter* 1*. Then, marginalizing over* $\boldsymbol{\mathcal{Z}}$*,* $\mathbf{Y} \mid \boldsymbol{\theta}$ *follows the multinomial distribution with the continuation-ratio logits parameterization in (2.1).*

Proposition 2.2 formalizes the latent variable representation discussed in Tutz (1991), in particular, it provides the explicit connection between the values of $\mathbf{Y}$ and $\boldsymbol{\mathcal{Z}}$, and the complete distributional assumptions (including independence) for $\boldsymbol{\mathcal{Z}}$. This result further highlights the benefits of the binary choice, sequential structure. Because the order of the response variable is preserved in the sequential mechanism, order restrictions for the latent $\mathcal{Z}_j$ are not required. This is in contrast with cumulative link models where the cut-off variables that discretize the single latent

continuous response must be ordered. The proposition also suggests a direction for constructing more flexible models by relaxing the parametric assumption for the distribution of the $\mathcal{Z}_j$. Indeed, the next result shows that we can recover the ordinal regression model of Section 2.2.1 through a LSBP mixture model for $\boldsymbol{\mathcal{Z}}$ with a product logistic mixture kernel.

Proposition 2.3. *Consider ordinal response* $\mathbf{Y} = (Y_1, \ldots, Y_C)$*, where* $Y_j \in \{0, 1\}$ *with* $\sum_{j=1}^{C} Y_j = 1$*, and continuous random vector* $\boldsymbol{\mathcal{Z}} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_{C-1}) \in \mathbb{R}^{C-1}$*. Assume that:* $\mathbf{Y} \mid \boldsymbol{\mathcal{Z}} \sim \mathbf{1}(\mathbf{Y} = j \iff \boldsymbol{\mathcal{Z}} \in \mathcal{R}_j)$*, for* $j = 1, \ldots, C$*, with the* $\mathcal{R}_j$ *defined in (2.11); and,* $\boldsymbol{\mathcal{Z}} \mid G_{\mathbf{x}} \sim \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \left\{ \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_{j\ell}(\mathbf{x})) \right\}$*, where the* $\omega_\ell(\mathbf{x})$ *and* $\theta_{j\ell}(\mathbf{x})$ *are defined in (2.3) and (2.4), respectively. Then, marginalizing over* $\boldsymbol{\mathcal{Z}}$*,* $\mathbf{Y} \mid G_{\mathbf{x}}$ *follows the multinomial LSBP mixture model in (2.2).*

We note that Proposition 2.3 does not simplify posterior simulation; Gibbs sampling for the model augmented with latent $\boldsymbol{\mathcal{Z}}_i$ for each observed response $\mathbf{Y}_i$ would require imputing the $\boldsymbol{\mathcal{Z}}_i$, and it would still involve two sets of Pólya-Gamma latent variables. However, the latent variable formulation offers an alternative perspective to model structure, as well as a useful tool to study model properties, such as KL support discussed in the next section.

2.2.5.2 Kullback-Leibler Support of the LSBP Mixture Model

Consider a prior $\mathcal{F}$ on a space of densities $\mathfrak{F}$. Density $f^0 \in \mathfrak{F}$ is in the KL support of $\mathcal{F}$ if $\mathcal{F}(N_\epsilon(f^0)) > 0$, for any $\epsilon > 0$, where $N_\epsilon(f^0) = \{f : \int f^0(\mathbf{z}) \log(f^0(\mathbf{z})/ f(\mathbf{z}))\, \mathrm{d}\mathbf{z} < \epsilon\}$ is the (size ϵ) KL neighborhood of f^0. Keeping the focus on continuous distributions, the regression setting targets collections of densities $\{f^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, indexed by values in the covariate space $\mathcal{X}$. We defer technical details to Appendix A.1, but note that the extension of the KL support definition considers the KL divergence (in the standard definition above) at arbitrary, finite

sets of values in $\mathcal{X}$ (e.g., Barrientos et al., 2012).

Theorem 2.1 establishes the KL support of the multinomial LSBP mixture prior model defined in (2.2), (2.3) and (2.4). The theorem builds from results in Barrientos et al. (2012) who examined the KL support of stick-breaking process mixture models for covariate-dependent densities. It can be shown that the LSBP mixture with the product logistic kernel, given in Proposition 2.3 for the continuous random vector $\boldsymbol{\mathcal{Z}}$, satisfies the various conditions required for the KL results in Barrientos et al. (2012). Thus, the latent variable representation of the ordinal regression LSBP mixture model yields the key step towards establishing its KL support. The other step is provided by Lemma 2.1 which connects the KL support of priors for continuous distributions with the KL support of priors for distributions of discrete variables induced by discretizing the corresponding continuous variables.

For our purposes, the connection is achieved by starting with a generic prior $\mathcal{F}_{\mathbf{x}}$ for the covariate-dependent distribution of continuous random vector $\mathbf{Z} \in \mathbb{R}^{C-1}$. Then, prior $\mathcal{P}_{\mathbf{x}}$ for the distribution of ordinal response $\mathbf{Y} = (Y_1, \ldots, Y_C)$ is induced through $\mathbf{1}(\mathbf{Y} = j \iff \mathbf{Z} \in \mathcal{R}_j)$, for $j = 1, \ldots, C$, with the $\mathcal{R}_j$ defined in (2.11). Hence, prior $\mathcal{F}_{\mathbf{x}}$ for densities $f_{\mathbf{x}}$ gives rise to prior $\mathcal{P}_{\mathbf{x}}$ for ordinal probabilities $p_{\mathbf{x}}$ via the mapping

$$f_{\mathbf{x}} \mapsto p_{\mathbf{x}}(y) = \int_{\mathcal{R}_y} f_{\mathbf{x}}(\mathbf{z})\, \mathrm{d}\mathbf{z}, \quad \text{for } y = 1, \ldots, C. \tag{2.12}$$

The following lemma relates the KL support of priors $\mathcal{F}_{\mathbf{x}}$ and $\mathcal{P}_{\mathbf{x}}$.

Lemma 2.1. *Consider prior $\mathcal{F}_{\mathbf{x}}$ for densities $f_{\mathbf{x}}$, and the prior $\mathcal{P}_{\mathbf{x}}$ for ordinal probabilities $p_{\mathbf{x}}$ induced from (2.12). Assume that densities $\{f^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ are in the KL support of $\mathcal{F}_{\mathbf{x}}$, and consider probability mass functions $\{p^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, where $p^0_{\mathbf{x}}$ is defined from $f^0_{\mathbf{x}}$ according to (2.12). Then, the probability mass functions $\{p^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ are in the KL support of $\mathcal{P}_{\mathbf{x}}$.*

Key to Lemma 2.1 is an inequality that allows us to bound the sum in the KL divergence for probability mass functions by the integral in the KL divergence for densities, when the densities and mass functions are related as in (2.12). The result is not restricted to the specific partition in (2.11), and it thus offers general scope to study the KL support of priors for categorical distributions arising through discretization of latent continuous responses.

Finally, combining Lemma 2.1, Proposition 2.3, and results from Barrientos et al. (2012), we can derive the KL property for our model.

Theorem 2.1. *Denote by $\mathcal{P}_{\mathbf{x}}$ the LSBP mixture prior defined in (2.2), (2.3) and (2.4), and consider $\{p^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, a generic collection of covariate-dependent probabilities for an ordinal response with C categories. Assume that the probability of each response category is strictly positive. Then, the mass functions $\{p^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ are in the KL support of $\mathcal{P}_{\mathbf{x}}$.*

Full KL support is a key theoretical property of the prior model. For priors on spaces of continuous densities, it is typically the case that various regularity conditions are required for a generic density to be in the KL support of the prior. In our context, the underlying regularity conditions in the results from Barrientos et al. (2012) reduce to the condition that response probabilities are strictly positive. Finally, as discussed in Appendix A.1, KL support results can also be obtained for the simplified models of Section 2.3.

2.2.5.3 Continuation-ratio Logits vs Cumulative Logits

As discussed in the Introduction, cumulative link models provide a common approach to ordinal regression, with the inverse link typically specified through the distribution function of a continuous variable Z. Then, the ordinal response Y values can be developed through discretization of the latent continuous response

Z, in particular, $Y = j$ if and only if $Z \in (\varkappa_{j-1}, \varkappa_j]$, for $j = 1, \ldots, C$. Here $-\infty = \varkappa_0 < \varkappa_1 < \ldots < \varkappa_{C-1} < \varkappa_C = \infty$ are cut-off points, where, typically, $\varkappa_1 = 0$ for identifiability. Key examples are cumulative probit and cumulative logit models for which the continuous distribution is $N(Z \mid \vartheta, 1)$ and $\mathfrak{L}(Z \mid \vartheta)$, respectively. In the absence of covariates, the parameters of cumulative logit and continuation-ratio logits models can be related as shown in the following result.

Proposition 2.4. *Consider the two distinct model formulations for an ordinal response with C categories given by the cumulative logit model with parameters $(\vartheta, \varkappa_2, ..., \varkappa_{C-1})$, and the continuation-ratio logits model in (2.1) with parameters $(\theta_1, ..., \theta_{C-1})$. Then, the parameters of the two models are connected through $\vartheta = -\theta_1$ and the recursive expression $\varkappa_j = \log(e^{\varkappa_{j-1}} + e^{\varkappa_{j-1}+\theta_j} + e^{\theta_j - \theta_1})$, for $j = 2, \ldots, C-1$.*

Despite the one-to-one correspondence between the parameters of the two models, there is a key difference in regression modeling. Under continuation-ratio logits, covariate effects are modeled through $\theta_j = \mathbf{x}^T\boldsymbol{\beta}_j$, for $j = 1, ..., C-1$. Here, the order for the response outcomes is induced by the binary, sequential mechanism, and thus the regression model specification is not constrained by restrictions on its parameters. In contrast, under cumulative link models, the order for the response values requires ordered cut-off points, which makes it challenging to model them as functions of the covariates. Indeed, cumulative link models typically incorporate covariates only through the location parameter of the latent continuous response, e.g, the proportional odds regression model arises under the $\mathfrak{L}(Z \mid \mathbf{x}^T\boldsymbol{\beta})$ distribution.

On the other hand, sequential models, such as continuation-ratio logits models, are not invariant under reversal of the order of the response categories, while cumulative link models are (Peyhardi et al., 2015). Nonetheless, for several

applications, the order of the ordinal response categories is determined by the context of the problem and/or the relevant scientific questions. We provide an example in Section 2.4.3, where the order of the response (disease severeness) is encoded from the mildest to the most severe, because of primary interest is study of covariate effects on the progress from mild to severe levels.

We note that existing nonparametric Bayesian methods for ordinal regression (reviewed in the Introduction) build from cumulative link models. In particular, the fully nonparametric models in Bao and Hanson (2015) and DeYoreo and Kottas (2018a) can be expressed as mixtures of cumulative probit regressions, the former with mixture weights that do not depend on the covariates, the latter with covariate-dependent mixture weights. The earlier discussion suggests that the continuation-ratio logits formulation offers wider scope as a building block in nonparametric mixture modeling for ordinal regression.

2.3 Specific Models for Ordinal Regression

Here, we study simplified model versions, which are naturally suggested given the two building blocks of the general model. In particular, we discuss ordinal regression models that arise by retaining covariate dependence only in the atoms (Section 2.3.1) or only in the weights (Section 2.3.2). The different model versions are empirically compared in Section 2.4.

2.3.1 The Common-weights Model

As a first simplification, we can remove the covariate dependence from the mixture weights. That is, the ordinal regression mixture model is built from the common-weights mixing distribution $G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell \, \delta_{\boldsymbol{\theta}_\ell(\mathbf{x})}$, such that $\mathbf{Y} \mid G_{\mathbf{x}} \sim$

$\sum_{\ell=1}^{\infty} \omega_\ell \, K(\mathbf{Y} \mid \boldsymbol{\theta}_\ell(\mathbf{x}))$, where the covariate-dependent atoms are defined as in the general model in (2.4) and (2.5).

Regarding the prior model for the weights, one option would be to keep the LSBP structure, that is, reduce $\mathbf{x}^T\boldsymbol{\gamma}_\ell$ in (2.3) to scalar parameter γ_ℓ, with the γ_ℓ independent and identically normally distributed. We work instead with the DP prior for the weights: $\omega_1 = V_1$, and $\omega_\ell = V_\ell \prod_{h=1}^{\ell-1}(1 - V_h)$, for $\ell \geq 2$, where $V_\ell \mid \alpha \overset{i.i.d.}{\sim} \text{Beta}(1, \alpha)$.

Using the DP-induced prior for the weights allows connections with the well-established literature on DDP mixtures, including the early work with common-weights DDP priors, e.g., the ANOVA DDP (DeIorio et al., 2004) and the spatial DP (Gelfand et al., 2005). In particular, the common-weights model can be equivalently written as a DP mixture model:

$$\mathbf{Y} \mid F \sim \int K(\mathbf{Y} \mid \mathbf{x}^T\boldsymbol{\beta}_1, \ldots, \mathbf{x}^T\boldsymbol{\beta}_{C-1}) \, dF(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{C-1})$$

where F follows a DP prior with total mass parameter α, and centering distribution defined through $\boldsymbol{\beta}_j \mid (\boldsymbol{\mu}_j, \Sigma_j) \overset{ind.}{\sim} N(\boldsymbol{\mu}_j, \Sigma_j)$, for $j = 1, \ldots, C-1$. The model is completed with a Gamma(a_α, b_α) hyperprior for α, and the prior for the $(\boldsymbol{\mu}_j, \Sigma_j)$ in (2.5). For prior specification, we combine the approach for the atoms in the general model with techniques for specifying the prior for the total mass DP parameter. The posterior simulation method replaces the steps for updating the weights with the update for the DP weights under blocked Gibbs sampling. The details can be found in Appendix B.1.

With the expression for the weights appropriately adjusted, the common-weights model inherits the properties of the general model, discussed in Section 2.2.2. The prior expectation in (2.8) is not affected by the form of the weights. However, the probability response curves admit a potentially less flexible form than the

one in (2.6) under the general model. We still have a weighted combination of parametric regression functions, but now without the local adjustment afforded by covariate-dependent weights. The data analyses in Section 2.4 demonstrate the practical utility of the general model, but also include examples where the common-weights model yields practical, sufficiently flexible inference.

2.3.2 The Common-atoms Model

The alternative way to simplify the general model is to use mixing distribution $G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\, \delta_{\boldsymbol{\theta}_\ell}$, resulting in the common-atoms mixture model:

$$\mathbf{Y} \mid G_{\mathbf{x}} \sim \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\, K(\mathbf{Y} \mid \boldsymbol{\theta}_\ell)$$

where $\boldsymbol{\theta}_\ell = (\theta_{1\ell}, \ldots, \theta_{C-1,\ell})$. The covariate-dependent weights are defined using the LSBP prior in (2.3). The prior model for the atoms is built from $\theta_{j\ell} \mid \mu_j, \sigma_j^2 \overset{ind.}{\sim} N(\mu_j, \sigma_j^2)$, for $j = 1, \ldots, C-1$, and $\ell \geq 1$. The model is completed with the conjugate prior for the hyperparameters: $\sigma_j^2 \overset{ind.}{\sim} IG(a_{0j}, b_{0j})$, with $a_{0j} > 1$, and $\mu_j \mid \sigma_j^2 \overset{ind.}{\sim} N(\mu_{0j}, \sigma_j^2/\nu_{0j})$, for $j = 1, \ldots C-1$, where $IG(\cdot, \cdot)$ denotes the inverse-gamma distribution.

Model implementation builds from the general model, with appropriate adjustments for the atoms. Here, $\mathrm{E}(\Pr(\mathbf{Y} = j \mid G_{\mathbf{x}})) = \mathrm{E}\left\{\varphi(\theta_j) \prod_{k=1}^{j-1}(1 - \varphi(\theta_k))\right\}$, for $j = 1, \ldots, C$, where the expectation is taken with respect to $\theta_j \overset{ind.}{\sim} N(\mu_{0j}, (\nu_{0j} + 1)b_{0j}/\nu_{0j}(a_{0j} - 1))$ (obtained by marginalizing over the prior for (μ_j, σ_j^2)). Hence, the prior expected marginal probability response curves are constants over the covariate space. The prior specification strategy utilizes this property, by setting $\{\mu_{0j}, \nu_{0j}, a_{0j}, b_{0j}\}_{j=1}^{C-1}$ such that these constants correspond to prior information for the ordinal response probabilities. The key quantity is again the expectation of a logit-normal distributed random variable (discussed earlier in Section

2.2.3). The posterior sampling scheme is adapted from the general model, with the normal-inverse-Wishart update for the atoms parameters replaced by the univariate normal-inverse-Gamma analogue. Details are given in Appendix B.1.

The common-atoms mixture structure offers a parsimonious model formulation, especially for problems with a moderate to large number of response categories. On the other hand, the simplified model form involves a potential limitation. The marginal and conditional probability response curves have the form in (2.6) and (2.7), respectively, with $\theta_{j\ell}(\mathbf{x})$ replaced by $\theta_{j\ell}$. Hence, the covariates inform the shape of the regression curves only through the mixture weights. As a practical consequence, the common-atoms model typically activates a larger number of effective mixture components to estimate the regression relationship, and it thus encounters a higher risk of overfitting for problems with a moderate to large number of covariates. This point is illustrated with the data examples of Section 2.4.

2.4 Data illustrations

2.4.1 Synthetic Data Examples

We consider three simulation examples to demonstrate the modeling framework, including comparative study of the common-weights, common-atoms, and general models. The first example is designed to highlight the benefits of local, covariate-dependent weights in capturing non-standard shapes of probability response curves. The objective of the second example is to study how the different models handle the challenge of recovering standard regression relationships for which the non-parametric mixture model structure is not necessary. The third example compares the effectiveness of the different models in capturing non-linear and non-additive covariate effects, including comparison with the density regression model from

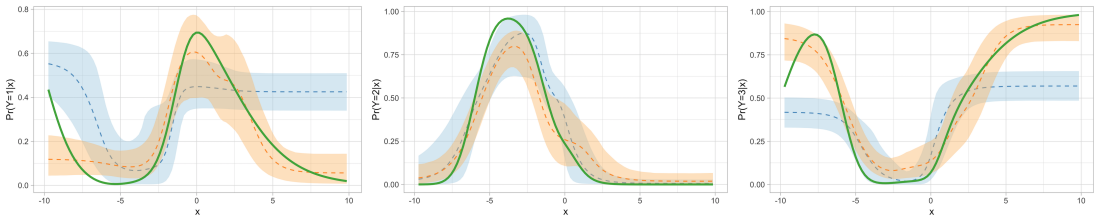
DeYoreo and Kottas (2018a).

2.4.1.1 First Synthetic Data Example

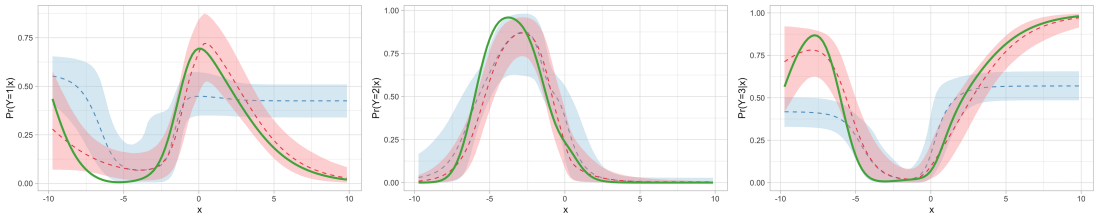
For the first experiment, to facilitate graphical illustrations, we consider an ordinal response with $C = 3$ categories, and one (continuous) covariate, where $x_i \overset{i.i.d.}{\sim} U(-10, 10)$, such that with the intercept, the covariate vector is $\mathbf{x}_i = (1, x_i)^T$. We generate the responses from a three component mixture of multinomial distributions, expressed in their continuation-ratio logits form. That is, $\mathbf{Y} \sim \sum_{k=1}^{3} w_k(\mathbf{x})\, K(\mathbf{Y} \mid \boldsymbol{\theta}_k(\mathbf{x}))$, where $\theta_{jk}(\mathbf{x}) = b_{jk0} + b_{jk1}x$, for $j = 1, 2$ and $k = 1, 2, 3$. The covariate dependence is introduced in the weights by computing $p_{j\mathbf{x}} = \Phi(a_{j0} + a_{j1}x)$, for $j = 1, 2$, where Φ is the $\mathrm{N}(0, 1)$ distribution function, and setting $(w_1(\mathbf{x}), w_2(\mathbf{x}), w_3(\mathbf{x})) = (p_{1\mathbf{x}}, (1 - p_{1\mathbf{x}})p_{2\mathbf{x}}, (1 - p_{1\mathbf{x}})(1 - p_{2\mathbf{x}}))$. The weights and atoms parameters are chosen such that the probability response curves have non-standard shapes (see Figure 2.4). We consider two sample sizes, $n = 200$ and $n = 800$.

The prior hyperparameters for the atoms are set according to the baseline choice. For the common-atoms and general models, we specify the LSBP prior hyperparameters $(\boldsymbol{\gamma}_0, \Gamma_0)$ to favor a priori more mixture components in the interval of covariate values $(-10, 0)$ where there is more variation in the regression functions. We note however that the prior specification is still fairly non-informative regarding the shape of the regression functions. In particular, under all three models, the prior mean estimates for the probability response curves are flat, and the prior 95% interval estimates span a substantial portion of the unit interval.

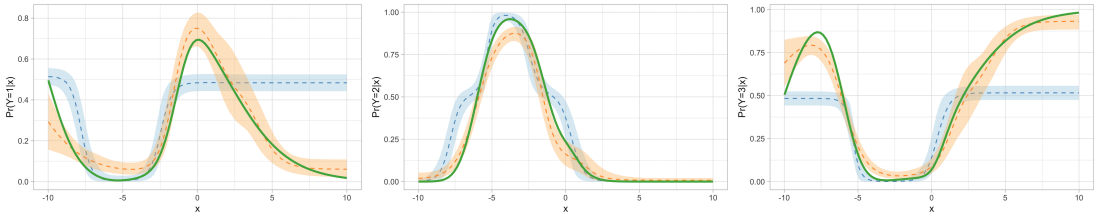
Inference results under the general and common-atoms models are contrasted with the common-weights model in Figure 2.4. As expected, the common-weights model does not recover well the non-standard regression functions for the first and



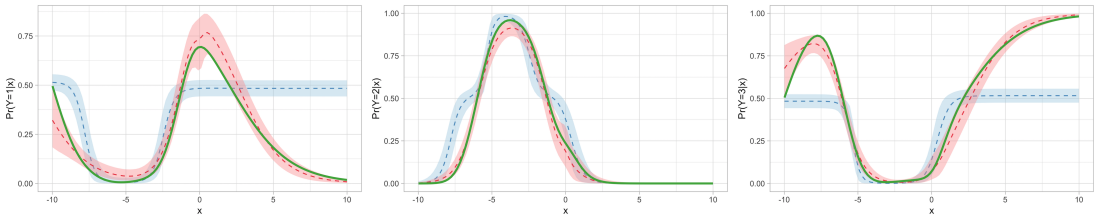
(a) Common-weights and common-atoms models ($n = 200$).



(b) Common-weights and general models ($n = 200$).



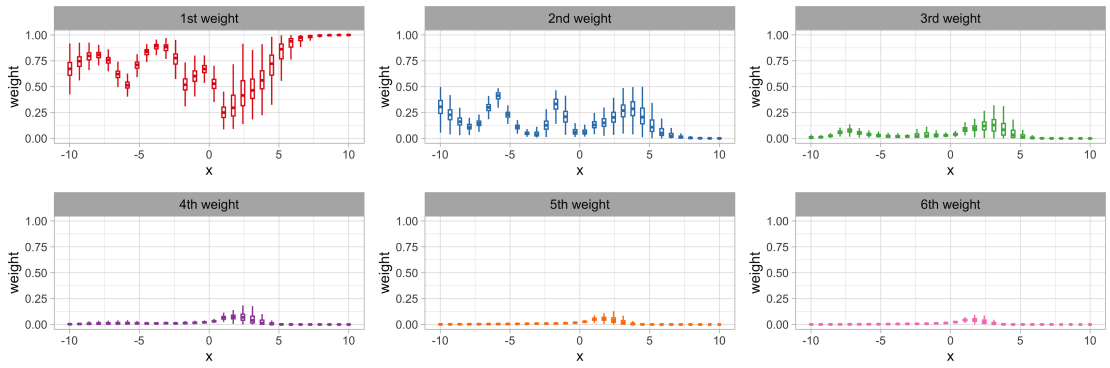
(c) Common-weights and common-atoms models ($n = 800$).



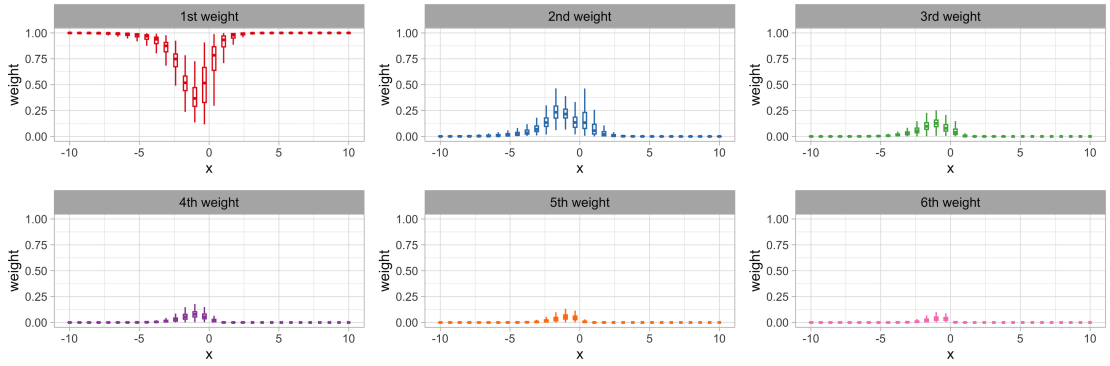
(d) Common-weights and general models ($n = 800$).

Figure 2.4: Synthetic data example. Posterior mean and 95% credible interval estimates for the marginal probability response curves under the common-weights (blue line and shaded region), common-atoms (orange line and shaded region), and general (red line and shaded region) models. In each panel, the green solid line is the true regression function.

third response categories. The two models that use covariate-dependent mixture weights perform notably better, with the general model resulting overall in more accurate estimation. Moreover, increasing the sample size results in more precise point estimates and more narrow posterior uncertainty bands.



(a) Common-atoms model.



(b) General model.

Figure 2.5: Synthetic data example ($n = 800$). Box plots of the posterior samples for the six largest mixture weights, under the common-atoms and general models.

Focusing on the models with covariate-dependent mixture weights (and the data set with $n = 800$), Figure 2.5 explores the posterior distribution of the six largest weights over the covariate space. For both models, it is essentially the first three largest weights that, given the data, define the probability vector of weights. However, we note the more local adjustment in the two largest weights under the common-atoms model, which becomes more pronounced in parts of the covariate space where the probability curves change more drastically. This is compatible with the common-atoms model's structure that seeks to fit the regression functions with atoms that do not change across the covariate space.

To further investigate how the proposed models behave in capturing non-

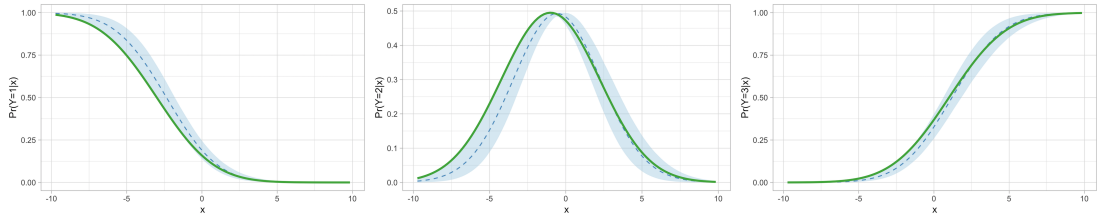
Table 2.1: First simulation example. Summary of model comparison using the posterior predictive loss criterion. The values corresponding to the best model are given in bold.

Model	$G_1(\mathcal{M})$	$P_1(\mathcal{M})$	$G_2(\mathcal{M})$	$P_2(\mathcal{M})$	$G_3(\mathcal{M})$	$P_3(\mathcal{M})$
Common-weights	132.59	141.27	72.76	83.54	136.40	141.91
Common-atoms	90.52	103.84	65.79	82.25	89.45	107.64
General	**89.96**	**96.25**	**64.14**	**72.56**	**88.24**	**94.39**

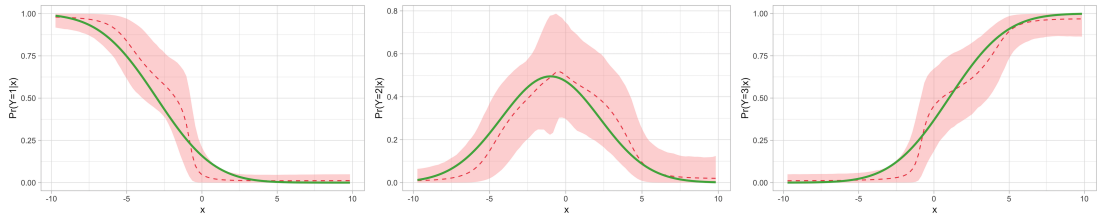
standard probability response curves, we conduct a formal model comparison using the posterior predictive loss criterion (Gelfand and Ghosh, 1998). The criterion contains a goodness-of-fit term and a penalty term. Since the response variable $\mathbf{Y}$ is multivariate, we consider the posterior predictive loss for every entry of it. Specifically, let $\mathbf{Y}_i^*$ denote the replicate response drawn from the posterior predictive distribution. Then, the goodness-of-fit term is defined as $G_j(\mathcal{M}) = \sum_{i=1}^n [\mathbf{Y}_{ij} - \mathrm{E}^{\mathcal{M}}(\mathbf{Y}_{ij}^* \mid \text{data})]^2$, whereas the penalty term is defined as $P_j(\mathcal{M}) = \sum_{i=1}^n \mathrm{Var}^{\mathcal{M}}(\mathbf{Y}_{ij}^* \mid \text{data})$, for $j = 1, \ldots, C$. The results are summarized in Table 2.1. The two models with covariate-dependent weights outperform the common-weights model. The common-atoms model and the general model are comparable in terms of goodness of fit. However, the common-atoms model activates more components to compensate for the constant atoms, resulting in a larger penalty.

2.4.1.2 Second Synthetic Data Example

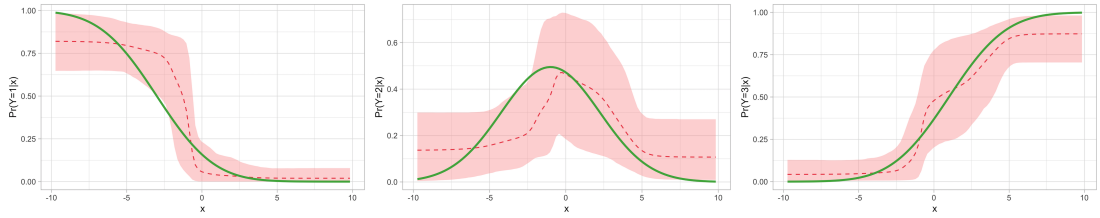
We generate $n = 100$ responses from a probit model, that is, we first sample normally distributed latent continuous variables $\tilde{y}_i$, and then discretize the $\tilde{y}_i$ with cut-off points to get the ordinal responses $\mathbf{Y}_i$, for $i = 1, \ldots, n$. The covariates, $(1, x_i)^T$, with the x_i sampled from the $Unif(-10, 10)$ distribution, enter through the mean of the normal distribution for the latent variables. The objective is to study how the different models handle the challenge of recovering standard



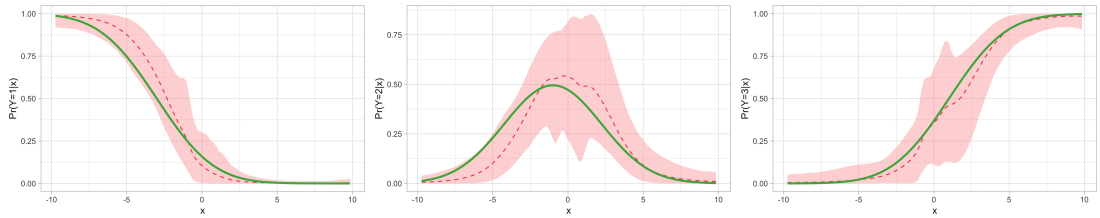
(a) Probit model.



(b) Common-weights model.



(c) Common-atoms model.



(d) General model.

Figure 2.6: Second simulation example. Inference results for the marginal probability response curves. In each panel, the dashed line and shaded region correspond to the posterior mean and 95% credible interval estimates, whereas the (green) solid line denotes the true regression function.

regression relationships for which the nonparametric mixture model structure is not necessary.

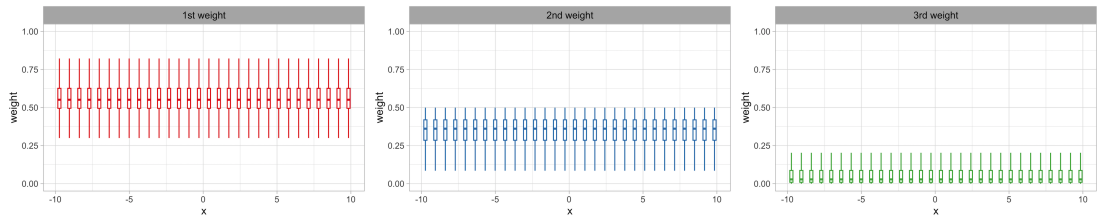
The nonparametric mixture models are applied to the data, using the (non-informative) baseline prior for their hyperparameters. Figure 2.6 plots posterior point and interval estimates for the marginal probability response curves, includ-

Table 2.2: Second simulation example. Summary of model comparison using the posterior predictive loss criterion. The values correspond to the best model are given in bold.

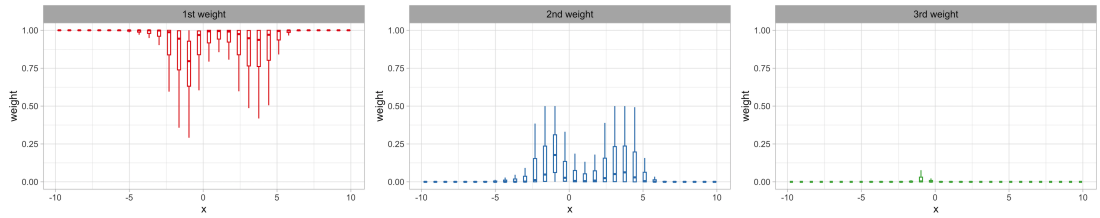
Model	$G_1(\mathcal{M})$	$P_1(\mathcal{M})$	$G_2(\mathcal{M})$	$P_2(\mathcal{M})$	$G_3(\mathcal{M})$	$P_3(\mathcal{M})$
Common-weights	**6.94**	7.94	12.95	13.59	**8.41**	9.00
Common-atoms	7.35	9.73	13.76	15.43	8.73	11.99
General	7.26	**7.38**	**12.94**	**12.78**	8.51	**8.84**

ing, as a reference point, estimates under the parametric probit model used to generate the data. As expected, the nonparametric models result in wider posterior uncertainty bands than the parametric model. In terms of recovering the underlying regression curves, the common-atoms model is less effective than the common-weights and the general model. As discussed in Section 2.3.2, this can be explained from the common-atoms model property that the regression curve shapes are adjusted essentially only through the mixture weights. The findings from the graphical comparison are supported by results from formal comparison, using the aforementioned posterior predictive loss criterion. The results, summarized in Table 2.2, suggests comparable performance for the common-weights and general models, whereas both outperform the common-atoms model.

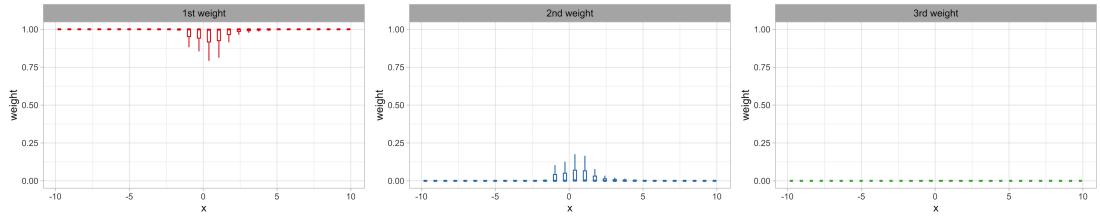
To further explore how the different nonparametric models utilize the mixture structure, Figure 2.7 shows the posterior distributions of the three largest mixture weights across covariate values. The general model is the most efficient in terms of the number of effective mixture components, using a second component (with small weight) only for covariates values around 0. This is to be expected, since it is those covariate values that result in practically relevant differences between the probit regression function (used to generate the data) and the logistic regression kernel. The common-atoms model activates effectively one extra component for covariate values where the regression functions are not flat. Compared to the general model, it places larger weights on the second component to account for the



(a) Common-weights model.



(b) Common-atoms model.

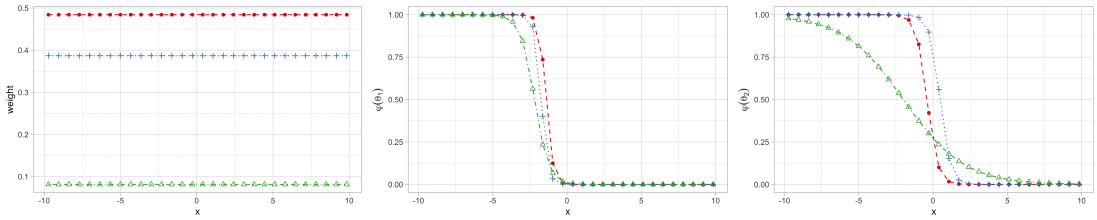


(c) General model.

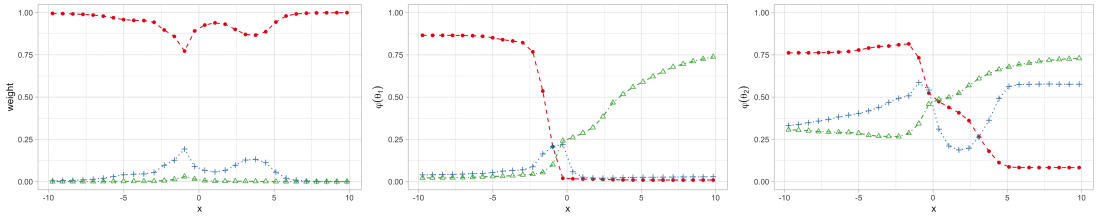
Figure 2.7: Second simulation example. Box plots of the posterior samples for the three largest mixture weights under each of the nonparametric models.

constant atoms. On the other hand, the mixture weights can not change with the covariates for the common-weights model. Hence, to recover the probit regression function, this model utilizes effectively three mixture components, with the second and third assigned larger (global) weight than the other two models.

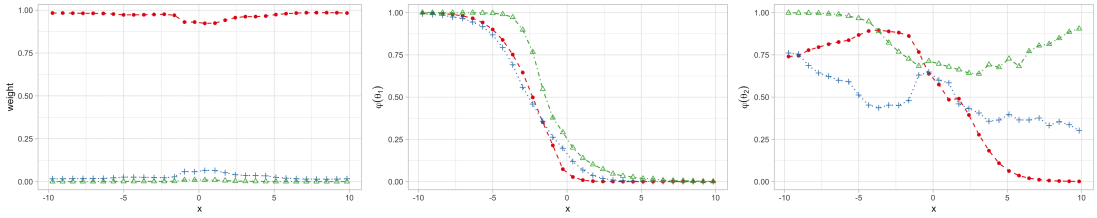
We also plot the posterior mean of the three largest weights and the corresponding atoms $\varphi(\theta_1)$ and $\varphi(\theta_2)$ in Figure 2.8. Combining with the posterior predictive loss criterion for each model, we can diagnose how the three models estimate the probability response curves. It appears that all three models are dominated by the mixing component with the largest weight, whose shape is similar to the truth. (The common-weights model favors two mixing components, but the two components are close to each other.) Regarding differences, the common-atoms



(a) Common-weigths model.



(b) Common-atoms model.

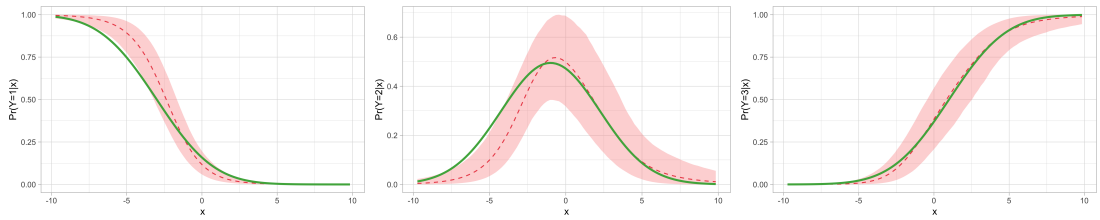


(c) General model.

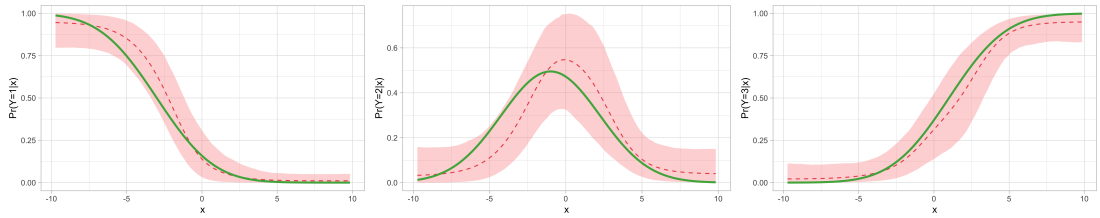
Figure 2.8: Second simulation example. Posterior mean estimates of the three largest mixture weights and atoms. The red circle, blue plus, and green triangle correspond to the first, second, and third largest weights, respectively.

model can only adjust the shape of the regression functions through the mixing weights. It thus uses more effective mixing components with shapes differing dramatically, yielding larger goodness-of-fit and penalty terms. The general model is overall the most effective in capturing the actual shape. It uses fewer and similar effective mixing components, leading to smaller penalty terms.

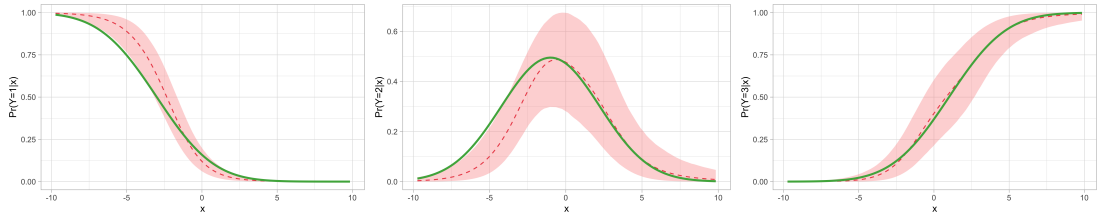
The sample size for this example was intentionally taken to be relatively small, in order to study sensitivity to the prior choice, as well as to demonstrate the practical utility of a more focused prior specification approach. If the monotonicity of two of the regression functions was in fact available as prior information, such information can be incorporated into the model, as discussed in Section 2.2.3.



(a) Common-weigths model.



(b) Common-atoms model.



(c) General model.

Figure 2.9: Second simulation example. Second simulation example. Inference results for the marginal probability response curves, under the informative prior specification. In each panel, the dashed line and shaded region correspond to the posterior mean and 95% credible interval estimates, whereas the (green) solid line denotes the true regression function.

Indeed, we consider a more information prior choice to reflect a decreasing shape for the first probability response function, and an increasing trend for the third response probability function. This set of informative prior hyperparameters leads to the posterior estimates shown in Figure 2.9. We note the more accurate posterior mean estimates and the reduction in the width of the posterior uncertainty bands, the improvement being particularly noteworthy for the common-atoms model.

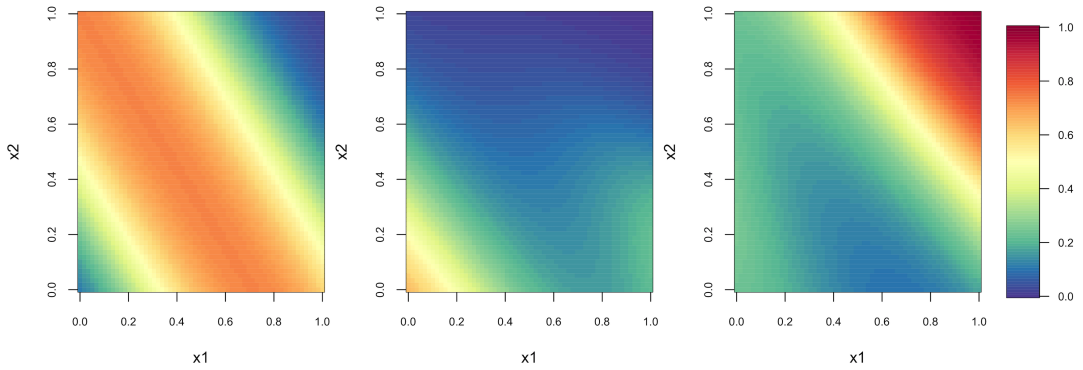
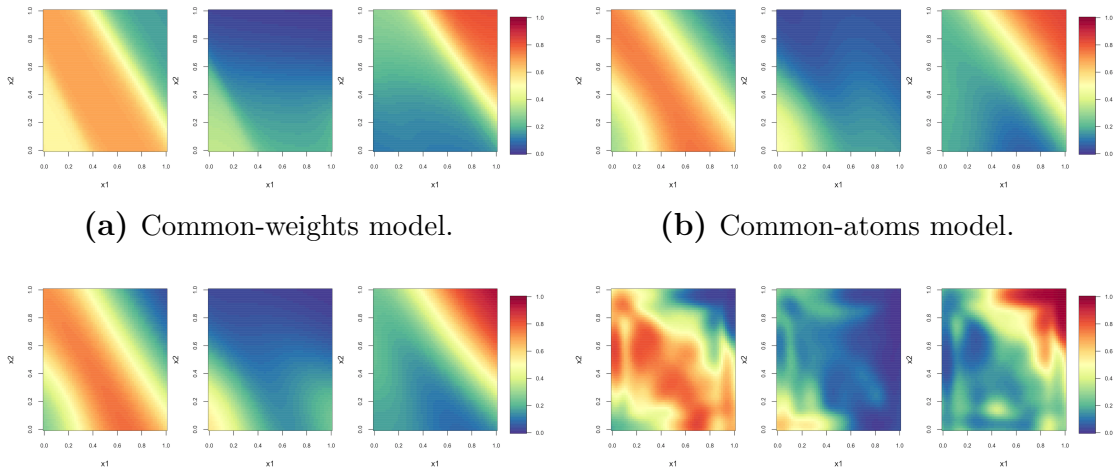


Figure 2.10: Third simulation example. The true probability response surface $\pi_j(x_1, x_2)$, for $j = 1, 2, 3$ (from left to right).

2.4.1.3 Third Synthetic Data Example

The purpose of the third simulation example is to investigate the effectiveness of the proposed models in capturing the joint effect of covariates. We consider two covariates and sample their values as $x_{is} \overset{i.i.d.}{\sim} Unif(0, 1)$, for $s = 1, 2$. The responses are sampled from the multinomial distribution with the continuation-ratio logits parameterization, where the $\theta_j(\mathbf{x})$, for $j = 1, 2$, are non-linear functions of the covariates. Specifically, we take $\theta_1(\mathbf{x}) = c_{11} + c_{12} \sin(a_{11}x_1 + a_{12}x_2)$, and $\theta_2(\mathbf{x}) = c_{21} + c_{22} \exp(a_{21}x_1 + a_{22}x_2)$. The covariate effects are non-linear and non-additive, resulting in non-standard probability response surfaces (displayed in Figure 2.10). We fit the general LSBP mixture model, as well as its two simplified versions. Note that covariates enter the mixture model structure linearly and additively, through the weights (common-atoms model), the atoms (common-weights model), or both (general model). It is therefore of interest to examine how the proposed models capture the non-standard probability response surfaces through the mixing of linear combinations of covariates. We take a fairly large sample size ($n = 3000$) to ensure the data is representative of the underlying data generating mechanism.

We set the prior hyperparameters for the atoms according to the baseline choice, while the hyperparameters for the weights are set to encourage a relatively large



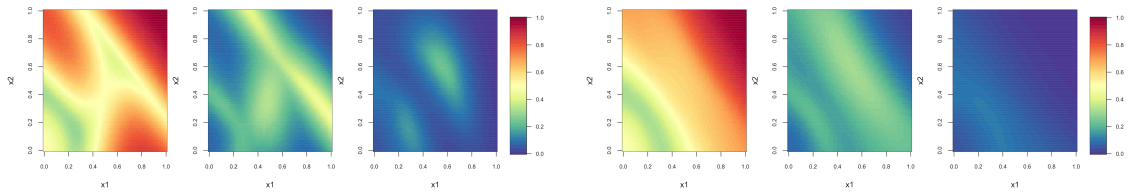
(a) Common-weights model. **(b)** Common-atoms model.

(c) General model. **(d)** Density regression model.

Figure 2.11: Third simulation example. Posterior mean estimates of $\pi_j(x_1, x_2)$, for $j = 1, 2, 3$ (from left to right).

number of effective mixture components. The truncation level is set as $L = 50$. Figure panels 2.11a, 2.11b, and 2.11c present the posterior mean estimates of the probability response surfaces under the common-weights model, the common-atoms model, and the general model, respectively. Although none of the models include non-linear or interaction terms for the covariates, the general model captures the non-linear joint effect particularly well, and the common-atoms model also demonstrates good estimation performance. These two models involve covariate-dependent weights, which allow for local adjustment in the regression surface estimates. As illustrated by this example, such local adjustment is beneficial, especially when the covariate effects are expected to be non-standard.

Focusing on the two LSBP mixture models with covariate-dependent weights, Figure 2.12 plots the posterior mean estimates for the three largest weights over a grid in the covariate space. The common-atoms model has more pronounced local changes, which is to be expected because it can adjust the shape of the regression surfaces only through the mixture weights. The general model exhibits more smooth estimated weight surfaces.



(a) Common-atoms model. **(b)** General model.

Figure 2.12: Third simulation example. Posterior mean of the three largest mixture weights for the common-atoms and general LSBP mixture models.

Focusing on the two LSBP mixture models with covariate-dependent weights, Figure 2.12 plots the posterior mean estimates for the three largest weights over a grid in the covariate space. The common-atoms model has more pronounced local changes, which is to be expected because it can adjust the shape of the regression surfaces only through the mixture weights. The general model exhibits more smooth estimated weight surfaces.

As discussed in the Introduction section of the main paper, the literature contains a relatively small collection of fully nonparametric Bayesian methods for ordinal regression. Here, we include comparison with the density regression model from DeYoreo and Kottas (2018a), for which the R code to implement the MCMC posterior simulation algorithm is available (from the online supplemental material of the journal article). Under the density regression modeling approach, the joint distribution of the two covariates and the latent continuous response is modeled with a DP mixture of trivariate normal densities, mixing on both the mean and the covariance matrix. The ordinal response probabilities emerge by discretizing the implied conditional distribution for the latent response given the covariates. The conditioning results in a model structure for the ordinal response distribution that can be interpreted as a mixture of probit regressions with covariate-dependent mixture weights.

The posterior mean estimates for the probability response surfaces are plotted

in Figure 2.11d. The density regression model captures the general trends, but it overestimates the local changes of the probability surfaces. This is likely due to both the different model structure (modeling the joint covariate-response distribution rather than the conditional response distribution) and to the underlying truth, which is much more structured than the density regression model. The results in Figure 2.11d are similar under different prior choices, in particular, under priors for the DP precision parameter that favor both large and fairly small number of distinct mixture components.

We further compare models based on their performance in estimating the probability response surfaces $\pi_j(x_1, x_2)$, for $j = 1, 2, 3$. We consider three metrics: the rooted mean square error (RMSE); the average length of the 95% posterior credible interval; and, the ratio at which the 95% credible interval covers the true probability. More specifically, for N grid points on the covariate space, the RMSE is calculated by

$$\bar{E}_j \;=\; N^{-1}\sqrt{\sum_{i=1}^{N}\{\pi_j^*(x_{1i}, x_{2i}) - \hat{\pi}_j(x_{1i}, x_{2i})\}^2}$$

where $\pi_j^*(x_{1i}, x_{2i})$ and $\hat{\pi}_j(x_{1i}, x_{2i})$ denote respectively the posterior mean estimate and the true value of the j-th category response probability at covariate values (x_{1i}, x_{2i}). In addition, the average 95% posterior credible interval length regarding the j-th category is obtained as $\bar{L}_j = N^{-1}\sum_{i=1}^{N}\{\pi_j^{\mathrm{U}}(x_{1i}, x_{2i}) - \pi_j^{\mathrm{L}}(x_{1i}, x_{2i})\}$, with $\pi_j^{\mathrm{U}}(x_{1i}, x_{2i})$ and $\pi_j^{\mathrm{L}}(x_{1i}, x_{2i})$ denoting the 97.5th and 2.5th percentiles of the posterior samples. Finally, the coverage percentage of the 95% posterior credible interval is calculated by

$$\bar{R}_j \;=\; N^{-1}\sum_{i=1}^{N}\mathbf{1}\{\pi_j^{\mathrm{L}}(x_{1i}, x_{2i}) \le \hat{\pi}_j(x_{1i}, x_{2i}) \le \pi_j^{\mathrm{U}}(x_{1i}, x_{2i})\}.$$

Table 2.3: Third simulation example. Summary of model comparison results, using the RMSE $\bar{E}_j$, average 95% posterior credible interval length $\bar{L}_j$, and the coverage of the 95% posterior credible interval $\bar{R}_j$, for $j = 1, 2, 3$. The values that correspond to the best model are given in bold.

Model	Level 1			Level 2			Level 3		
	E_1	L_1	R_1	E_2	L_2	R_2	E_3	L_3	R_3
Common-weights	4.08	0.09	0.59	2.24	0.03	0.93	2.59	0.09	0.74
Common-atoms	2.26	0.06	0.79	1.69	0.04	0.74	2.15	0.07	0.88
General	**1.63**	**0.04**	0.91	**0.83**	**0.02**	1.00	**1.31**	**0.04**	**0.92**
Density regression	3.82	0.07	**0.95**	2.09	0.06	**0.97**	3.58	0.10	**0.92**

Table 2.3 reports the metrics values for the four models. Among the LSBP mixture models, the general model performs better with respect to essentially all metrics, followed by the common-atoms model. These results reinforce the findings from the graphical comparison of the posterior mean estimates for the probability response surfaces. Also consistent with the graphical comparison in Figure 2.11, the general and common-atoms LSBP mixture models outperform the density regression model in terms of RMSE. This is also the case with respect to the average 95% posterior credible interval length. The density regression model achieves the best results for the coverage criterion, with the general LSBP mixture model a fairly close second. Overall, the general LSBP mixture model yields the best performance under the particular simulation scenario.

2.4.2 Credit ratings of U.S. firms

We consider data on Standard and Poor's (S&P) credit ratings for 921 U.S. firms in 2005 (Verbeek, 2008). The ordinal response is the firm's credit rating, originally recorded on a scale with seven categories. Since there were only 17 firms with rating of 1 or 7, and to facilitate illustration of inference results, we combine the responses in the first two and last two categories. We thus obtain an

Table 2.4: Credit ratings data. Summary of the posterior predictive loss criteria for model comparison. Each pair of numbers corresponds to $(G_j(\mathcal{M}), P_j(\mathcal{M}))$, $j = 1, \cdots, 5$. "Parametric" refers to the continuation-ratio logits model. The values for model with the smallest $G_j(\mathcal{M}) + P_j(\mathcal{M})$ are highlighted in bold.

	Parametric	Common-weights	Common-atoms	General
Level 1	$(92.65, 90.17)$	**(88.07,92.38)**	$(92.64, 104.97)$	$(86.61, 95.79)$
Level 2	$(158.71, 158.72)$	$(153.13, 158.92)$	$(156.04, 163.96)$	**(153.10,158.07)**
Level 3	$(150.18, 150.82)$	**(145.40,150.38)**	$(149.00, 152.03)$	$(148.11, 148.60)$
Level 4	$(95.95, 96.29)$	$(95.08, 97.23)$	$(97.41, 100.10)$	**(94.20,94.24)**
Level 5	**(17.80,17.46)**	$(17.85, 20.57)$	$(21.19, 31.04)$	$(17.74, 20.47)$

ordinal response scale ranging from 1 to 5, with higher ratings indicating higher creditworthiness. The data set includes five company characteristics that serve as covariates: book leverage (ratio of debt to assets), x_1; earnings before interest and taxes divided by total assets, x_2; standardized log-sales (proxy for firm size), x_3; retained earnings divided by total assets (proxy for historical profitability), x_4; and working capital divided by total assets (proxy for short-term liquidity), x_5.

The three nonparametric models were applied to the data, using the baseline choice for the atoms prior hyperparameters, and priors for the weights hyperparameters that favor a moderate to large number of distinct mixture components n^* (i.e., number of distinct $\mathcal{L}_i$ in the notation of Section 2.2.4). Given the number of covariates, one would expect that the common-atoms model requires larger n^*. Indeed, the posterior median for n^* is 8, 12, and 21 under the common-weights, general, and common-atoms model, respectively; in fact, the common-atoms model did not produce a posterior draw for n^* smaller than 10. The relative inefficiency of the common-atoms model is also reflected in its larger penalty term for the posterior predictive loss criterion. As detailed in Table 2.4, the model comparison using the posterior predictive loss criterion essentially does not distinguish between the general and common-weights models. Here, we discuss results under the common-weights model.

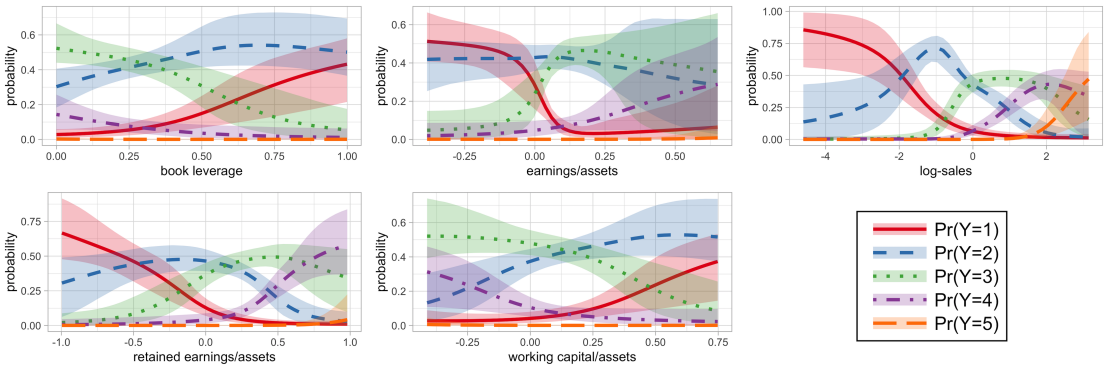


Figure 2.13: Credit ratings data. Posterior mean (lines) and 95% interval (shaded regions) estimates of probability response curves $\pi_j(x_s)$. Estimates for all five response categories are displayed in a single panel for each covariate.

We estimate first-order effects for each covariate x_s (denoted by $\pi_j(x_s)$, for $j = 1, \ldots, 5$), by computing posterior realizations for $\Pr(\mathbf{Y} = j \mid G_{\mathbf{x}})$ in (2.6) at a grid over the observed range for x_s, keeping the values of the other covariates fixed at their observed average. The resulting point and interval estimates are displayed in Figure 2.13. The estimates reveal some interesting relationships between the firm's characteristics and its credit rating. For instance, debt may help to fuel growth of the firm, while uncontrolled debt levels can lead to credit downgrades. Hence, an important question pertains to the relevant debt to assets ratio. The substantial increase in the probability of the lowest credit rating when book leverage gets larger than 0.4 (top left panel of Figure 2.13) suggests that the desirable ratio may not exceed 0.4. Moreover, there is a positive association between standardized log-sales (a proxy for firm size) and the firm's credit rating. The probability of the lowest credit rating decreases at a particular rate for low to moderate log-sales values, with the probability becoming exceedingly small for larger firms. The probabilities for ratings 2, 3 and 4 peak at increasingly larger log-sales values, and the probability of the highest rating is practically zero for low to moderate log-sales values and is increasing for the largest firms.

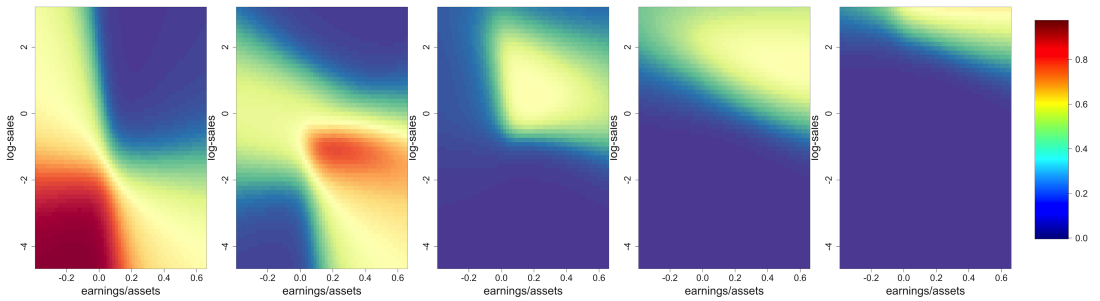


Figure 2.14: Credit ratings data. Posterior mean estimates of probability response surfaces $\pi_j(x_2, x_3)$, for $j = 1, \ldots, 5$ (from left to right).

Similarly to the first-order effects estimates, we can obtain inference for second-order probability response surfaces for any pair of covariates $(x_s, x_{s'})$, denoted by $\pi_j(x_s, x_{s'})$, for $j = 1, \ldots, 5$. As an illustration of the model's capacity to accommodate interaction effects among the covariates, Figure 2.14 plots posterior mean estimates for the second-order effects corresponding to earnings divided by total assets (x_2) and standardized log-sales (x_3).

Furthermore, it is also of interest to investigate the model performance on prediction. The credit rating of firms can be partitioned into two categories: investment grade (rating score is 3 or higher) and speculative grade. Because many bond portfolio managers are not allowed to invest in speculative grade bonds, firms with a speculative rating incur significant costs. It is helpful to check the models' implied posterior probability of obtaining an investment grade for a particular firm. We consider five prediction scenarios corresponding to the five covariates. In each scenario, we evaluate the change in the investment grade probability associated with one of the covariates changing from the 25th to the 75th percentile of the observed values, while holding all the other covariates at the average value of all observations. Figure 2.15 displays the posterior distribution of the probability of obtaining investment grade under the common-weights model.

Under the common-weights model, the probability moves along the expected

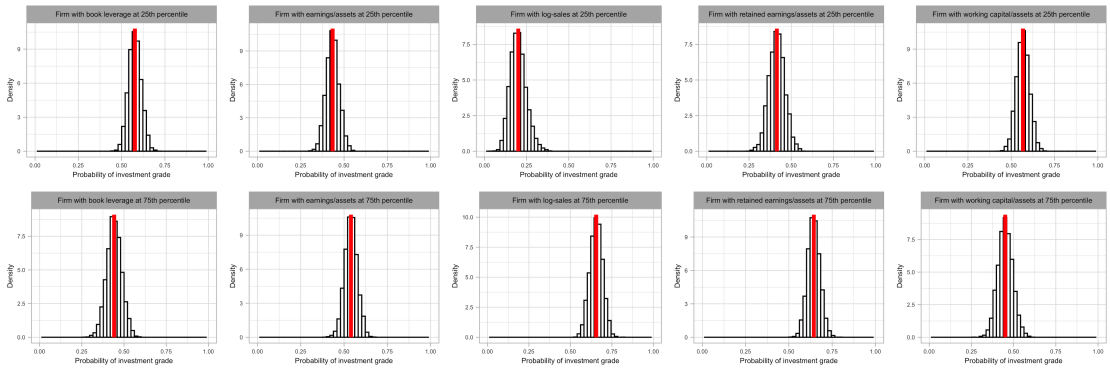


Figure 2.15: Credit ratings data. Posterior distributions of the probability of obtaining investment grade rating under the common-weights model. The red solid lines indicate the posterior mean.

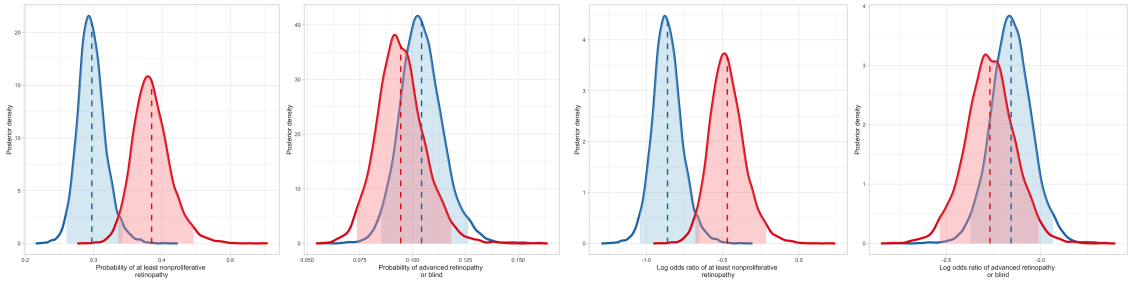
direction concerning all covariates, except for the working capital, which coincides with the discovery in Verbeek (2008). The results indicate higher leverage, meaning that a firm is financed relatively more with debt, reduces the expected credit rating. This is due to the fact that firms with high leverage face substantially higher debt financing costs. In addition, the larger firms, indicated by larger log-sales, have significantly better credit ratings than smaller firms, ceteris paribus. Higher earnings before interest and taxes and higher retained earnings also improve credit ratings. Furthermore, one would expect that maintaining a high level of working capital would enhance a company's credit rating since it reduces risk. However, a high level of working capital reduces profits, raising concern about the company's ability to cover interest payments. This argument suggests a concave relationship between working capital and credit rating, postulating that firms could have an optimal working capital ratio. Our result indicates that the optimal ratio lies between the first and third quartiles.

2.4.3 Retinopathy data

Problems from clinical research provide a broad application area for which the proposed modeling approach is particularly well-suited. For such problems, the severeness of a disease is often recorded in ordinal scale, and it is of interest to estimate effects of risk factors on disease status. This is a setting where it is natural to treat ordinal responses sequentially, from which conditional probability response relationships can be directly explored.

To illustrate the utility of our methodology in this context, we work with data set `retinopathy` from the R package "catdata" (Schauberger and Tutz, 2023). The data set is from a 6-year follow-up study of type 1 diabetic patients; it contains information about 613 patients' retinopathy status, recorded as: no retinopathy (ordinal level 1), nonproliferative retinopathy (ordinal level 2), and advanced retinopathy or blind (ordinal level 3). Also available is information on four risk factors: smoking status (smoker/non-smoker), diabetes duration (years), glycosylated hemoglobin (percent), and diastolic blood pressure (mmHg). The primary scientific question pertains to association between retinopathy and smoking status, adjusted for the other risk factors.

The standard proportional odds regression model does not appear suitable for this data set. Descriptive data analysis suggests that the odds of developing the retinopathy states are not proportional with respect to smoking; see Bender and Grouven (1998). This is therefore a useful illustrative case for the LSBP mixture model, which, in contrast to any particular parametric model, supports essentially any collection of covariate-dependent ordinal response probabilities. We apply the general model to the data with the ordinal scale for the responses, focusing on the endpoints of "at least nonproliferative retinopathy" ($\mathbf{Y} \geq 2$) and "advanced retinopathy or blind" ($\mathbf{Y} = 3$).



(a) $\tilde{\pi}_{1s}$ (left) and $\tilde{\pi}_{2s}$ (right). **(b)** r_{1s} (left) and r_{2s} (right).

Figure 2.16: Retinopathy data. Posterior densities for the probabilities and log odds ratios of the two retinopathy endpoints for smokers (in red) and non-smokers (in blue). The dashed line and shaded region correspond to the posterior mean and 95% credible interval, respectively

Since the main objective is to assess the relationship between smoking and retinopathy, we focus on inference results for the two retinopathy endpoint probabilities for smokers and non-smokers. Keeping the values for the other risk factors fixed at their observed average, Figure 2.16a displays the posterior densities for $\tilde{\pi}_{1s} = \Pr(\mathbf{Y} \geq 2 \mid G_{\mathbf{x}})$ and $\tilde{\pi}_{2s} = \Pr(\mathbf{Y} = 3 \mid G_{\mathbf{x}})$, where the subscript $s = 0, 1$ indicates non-smokers and smokers, respectively. These results point to an adverse effect of smoking on the development of at least nonproliferative retinopathy for diabetic patients, whereas there is no clear suggestion of an effect on the terminal endpoint (advanced retinopathy of blindness). Indeed, the posterior mean and 95% credible interval for $\tilde{\pi}_{11} - \tilde{\pi}_{10}$ are 0.088 and $(0.042, 0.144)$, whereas the corresponding estimates for $\tilde{\pi}_{21} - \tilde{\pi}_{20}$ are -0.010 and $(-0.031, 0.012)$.

Consider the smoker/non-smoker log odds ratios for the two retinopathy endpoints, that is, $r_{1s} = \log\{\tilde{\pi}_{1s}/(1 - \tilde{\pi}_{1s})\}$ and $r_{2s} = \log\{\tilde{\pi}_{2s}/(1 - \tilde{\pi}_{2s})\}$, for $s = 0, 1$. The proportional odds regression model assumes $\log\{\Pr(\mathbf{Y} \leq j)/\Pr(\mathbf{Y} > j)\} = \varkappa_j - \mathbf{x}^T\boldsymbol{\beta}$, for $j = 1, 2$, where the $\varkappa_j$ are the cut-off points in the notation of Section 2.2.5.3, and $\mathbf{x}$ comprises the four risk factors. Hence, assuming proportional odds specifically with regard to the risk factor of smoking imposes the constraint

$r_{11} - r_{10} = r_{21} - r_{20}$. As discussed in Bender and Grouven (1998), the proportional odds model identifies the diabetes duration, glycosylated hemoglobin, and blood pressure as significant risk factors, but estimates that the effect of smoking is negligible. Bender and Grouven (1998) question the proportional odds model assumption (the constraint above) based on descriptive data analysis, and estimation results from fitting separate binary logistic regressions to the two retinopathy endpoints.

The results from the LSBP mixture model highlight the benefits of flexible nonparametric Bayesian modeling. Using a probability model for the ordinal response, we can identify the disease endpoint for which smoking has an adverse effect (Figure 2.16a), as well as obtain clear evidence against the proportional odds structure with respect to the smoking risk factor (Figure 2.16b). And, to reiterate, such model-based inferences arise from a prior probability model that does not impose restrictions on the ordinal regression relationships, making it practically useful for applications where it is difficult to check whether the assumptions of a specific parametric model are compatible with the data generating mechanism.

As one illustration of predictive model assessment of the general LSBP mixture model, we examine the posterior predictive distribution of the test statistics r_j, $j = 1, 2, 3$, which represent the proportion of each ordinal response category among the n responses. That is, for each posterior sample of model parameters, we obtain posterior predictive samples for all in-sample subjects, and compute the proportion of each category. We therefore obtain the posterior predictive distribution of each test statistic, which can be compared with the observed proportion. The results, displayed in Figure 2.17, suggest that the model generates predictions which are compatible with the observed responses.

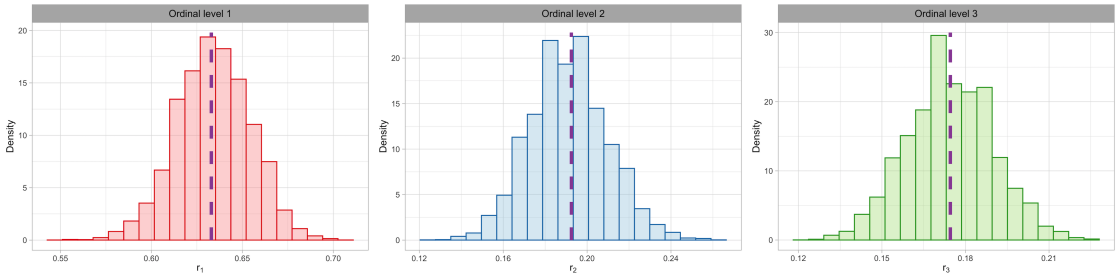


Figure 2.17: Retinopathy data. Posterior predictive distribution of the proportion for each ordinal response category. The dashed line indicates the observed proportion.

2.5 Discussion

We have developed Bayesian nonparametric mixture models for ordinal regression, modeling directly the discrete response distribution. The similarity between the logit stick-breaking prior and the continuation-ratio logits structure provides an elegant way of incorporating covariate effects in both the weights and the atoms of the mixture model, leading to the general model. To investigate the trade-off between model flexibility and complexity, we introduce two simpler models that retain covariate dependence only in the atoms (common-weights model) or only in the weights (common-atoms model). The methods yield a comprehensive toolbox that spans a wide range of flexibility in modeling ordinal regression relationships. Viewing the two simpler models as building blocks of the general model enables us to explore properties and develop inference algorithms under a unified framework. Full Kullback-Leibler support has been established as a key theoretical model property. The practical advantage of the proposed models lies in the convenience in prior specification and the computationally efficient posterior simulation method. With regard to the latter, the key feature is the combination of the continuation-ratio logits representation for the mixture kernel with the Pólya-Gamma data augmentation technique.

A practical consideration is which model to apply to a specific problem. The data examples of Section 2.4 were chosen to study different scenarios for suitability of the simplified models, as they pertain to the complexity of the probability response curves, the sample size, and the number of covariates. The common-weights model can not take advantage of the local adjustment offered by covariate-dependent weights, and this may be an issue for non-standard ordinal regression relationships. Among the two simpler model specifications, the common-atoms model is more suitable for complex covariate-response relationships. The caveat is that this model activates a large number of effective mixture components, thus increasing the computational cost and facing the potential risk of overfitting. Inheriting features from both of its building blocks, the general model offers the most versatile structure, especially for applications with sufficiently large amounts of data and non-standard regression relationships, as demonstrated by the synthetic data example of Section 2.4.1. Nonetheless, in applications with small to moderate sample sizes and moderate to large number of response categories, the two simpler models are useful options to consider.

The continuation-ratio logits structure boosts computation in two ways. First, it implies conditional independence for category-specific parameters, allowing partial parallel computing across response categories. In addition, the MCMC algorithm can be replaced by a mean-field variational inference approach. Taking advantage of the Pólya-Gamma technique, the variational strategy for our models can be framed within the well-established exponential family setting, for which there exists a closed-form coordinate ascent variational inference algorithm (Blei et al., 2017). Therefore, there is potential to scale up the proposed models to handle ordinal regression problems with large amounts of data.

The ordinal regression problem we have explored forms a building block for

more general model settings involving ordinal responses. In fact, Proposition 2.3 may widen the scope of the building block through alternative distributional assumptions for the latent variables. A feature of the modeling framework is its modularity. For example, the model structure can be embedded in a hierarchical framework to develop nonparametric inference for longitudinal ordinal regression. Repeated measurements of ordinal responses are typically measured with covariates over time. A possible way to approach such problems could be built upon models that allow the ordinal regression relationships at each particular time point to be estimated in a flexible fashion, combined with a hyper-model for evolving temporal dynamics. In addition, variable selection can be incorporated into the model through the priors for the parameters of the mixture kernel and weights, adapting techniques used for local mixtures of normal densities (e.g., Chung and Dunson, 2009; Heiner and Kottas, 2022). We will report on such extensions in future work.

Finally, the methodology can also be applied to problems where the components of the ordinal response $\mathbf{Y}$ are not necessarily binary. A specific application area involves developmental toxicity studies. Here, the covariate is the level of a particular toxin, and, for each pregnant laboratory animal exposed to a specific toxin level, the typical data structure involves responses recorded for its offspring on embryolethality, malformation, and normal offspring. The modeling methods can be elaborated to extend the dependent DP mixture model in Kottas and Fronczyk (2013) for developmental toxicology data analysis.

Chapter 3

A Nonparametric Modeling Approach for Ordinal Regression with Heterogeneous Responses

3.1 Introduction

3.1.1 Background and Data

Ordinal regression with responses being a sum of ordinal variables is a common occurrence in biomedical studies. In such a problem, a multivariate ordinal response $\mathbf{Y} = (Y_1, \ldots, Y_C)$ is recorded, along with a covariate $\mathbf{x}$. Here, each component of $\mathbf{Y}$ is an integer between 0 and m, and $\sum_{j=1}^{C} Y_j = m$. It is typically assumed that $\mathbf{Y} \sim \text{Mult}(m, \pi_1, \ldots, \pi_C)$. Contrasting with the ordinal response described in Chapter 2, we refer to variable of this type as the "extended" ordinal response. We can equivalently view $\mathbf{Y}$ as the sum of m ordinal variables, denoted as $\{\tilde{\mathbf{Y}}_q : q = 1, \ldots, m\}$, where $\tilde{\mathbf{Y}}_q$ represents a standard univariate ordinal response, encoded by binary variables. In this chapter, we will develop a modeling approach

that deals with overdispersed $\mathbf{Y}$. That is, responses which we might expect to be of multinomial form, but which exhibit a variance larger than that predicted by the multinomial model.

Segment II developmental toxicology studies provide an important area of application in which data of the aforementioned structure are prevailing. In these studies, at each experimental dose level, a number of pregnant laboratory animals (dams) are exposed to the toxin after implantation. Typically, the number of fetuses on ordered categories (e.g. prenatal death, malformation, and normal) are recorded as the response. The main objective is to examine the dose-response curve, which is defined by the (conditional) probability of an endpoint across the dose levels. Other inferential objectives involve solving the inverse problem, where interest lies in estimation of the dose level that induces a specified extra risk comparing to the control dose. Regarding the latter, coherent uncertainty quantification of the dose-response relationships is the key for ensuring accuracy. We refer to, for example, Kuk (2004) for a comprehensive discussion about developmental toxicity studies and the statistical issues therein.

In a standard Segment II developmental toxicology experiment, at each experimental toxin level, x_d, a number, n_d of pregnant laboratory animals (dams) are exposed to the toxin and the total number of implants, m_{di}, the number of non-viable fetuses (undeveloped embryos and/or prenatal deaths), R_{di}, and the number of live malformed (external, visceral or skeletal) pups, y_{di}, from each dam are recorded. We use $\mathbf{Y}_{di} = (R_{di}, y_{di}, m_{di} - R_{di} - y_{di})$ to denote the ordinal response, for the i-th animal at dose x_d. The data structure, $\{(x_d, \mathbf{Y}_{di}) : d = 1, \ldots, N; i = 1, \ldots, n_d\}$ falls in the extended ordinal regression setting, with replicated responses at each value of the single covariate (toxin level). Hereinafter, we refer to this particular data structure as the extended setting.

As an example, we consider the data from a study where ethylene glycol (EG), an organic solvent, is evaluated for toxic effects in pregnant rats. The study involves three active toxin levels at 1.25, 2.5, and 5 g/kg, and a control group, with the respective number of dams assigned to each group being 28, 29, 27 and 28. The number of implants ranges from 1 to 18 across all dams and all dose levels, with 25th, 50th, and 75th percentiles given by 12, 14, and 15, respectively. We work with the version of the data given in Table 1 of Fung et al. (1998).

The example data set is visualized in Figure 3.1. For each dam, we plot the observed proportions of embryolethality, malformation among live pups, and combined negative outcomes against the dose level. The color is used to facilitate identifying the same dam across panels. For the dose-response curves corresponding to these three endpoints, the empirical proportions suggest an overall increasing trend, although with no obvious parametric form for each dose-response curve. Moreover, vast variability is evident in the responses, of which the magnitude also differs across dose levels. Also noteworthy is a potentially different dose-response relationship for non-viable fetuses and malformed pups. A high dose usually exhibits an increase in the risk of embryolethality, while causing earlier mortality that prevents the pups surviving to be observed with malformations. Thus, it is biologically relevant to jointly model the distinct endpoints.

Because in Segment II toxicity experiments exposure occurs after implantation, we assume a distribution for the number of implants that does not depend on the toxin level. Through this chapter, we factorize the joint distribution as $p(m_{di}, R_{di}, y_{di} \mid x_d) = p(R_{di}, y_{di} \mid m_{di}, x_d)p(m_{di})$, and adopt a Poisson distribution with support shifted such that $m_{di} \geq 1$ for $p(m_{di})$. The focus here is on exploring modeling approaches for the toxin-dependent conditional distribution for the number of non-viable fetuses and malformations, (R_{di}, y_{di}), given m_{di}.

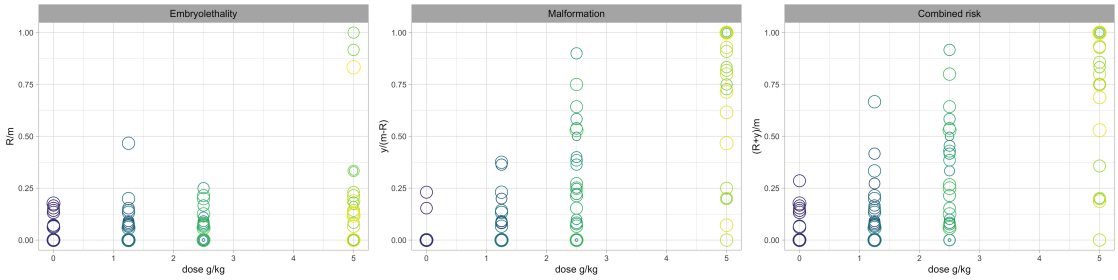


Figure 3.1: EG data. In each panel, a circle corresponds to a particular dam and the size of the circle is proportional to the number of implants. The coordinates of the circle are given by the toxin level and the proportion of the specific endpoint: non-viable fetuses among implants (left panel); malformations among live pups (middle panel); combined negative outcomes among implants (right panel).

3.1.2 Objectives and Outline

A gold mine of modeling challenges presented in the aforementioned data structure has captured attention in the statistical literature. To address the common occurrence of overdispersed responses in developmental toxicity studies, typically used approaches involve mixture models. We investigate a spectrum of mixture models, with mixing kernel that presumes a factorized multinomial structure. Starting with continuous mixtures, we examine two popular choices, namely the model based on Beta-Binomial (BB) distribution and Logistic-Normal-Binomial (LNB) distribution. We argue that these models preclude reliable risk assessment in this application, because of their parametric form in both the response distribution and the dose-response relationship.

Turning to discrete mixture models, the nonparametric mixture models proposed in Chapter 2 serve as a strong initial reference. Specifically, enhanced flexibility is achieved through a nonparametric mixture of continuation-ratio logits factorization of multinomial distributions, mixing through a dependent stick-breaking process prior placed on the probability parameters. Nonetheless, its kernel is restricted in terms of modeling extended ordinal response, providing

opportunities for novel models which enable more effective control of the ordinal responses' variability.

We consider a combination of these two types of mixture models. Specifically, we adopt a continuous mixture model as the kernel, which is then encapsulated in the discrete nonparametric mixing structure. The derived models inherit flexibility from the discrete mixture models, while potentially allowing an improvement in accounting overdispersion through the extra set of parameters introduced in the kernel. Regarding the choice of kernel, the LNB distribution is preferred from a computational efficiency consideration. Besides expanding a developed approach, motivation for examining the new model also originates from regulatory guidance (U.S. EPA, 1991), which requires considering an adequate set of models for developmental toxicity risk assessment.

Bayesian nonparametric methods have been explored as a powerful tool for analysis of development toxicology data. Focusing on studies that involve a discrete response, Dominici and Parmigiani (2001) proposed a product of Dirichlet process (DP) mxitures approach to deal with combined negative outcomes. Targeting the same type of responses, Fronczyk and Kottas (2014) built a nonparametric mixture model from a dependent Dirichlet process (DDP) prior, with the dependence of the mixing distributions governed by the dose level. Models that jointly consider various types of responses have also been explored, including, for binary and continuous responses (Hwang and Pennell, 2014), categorical and continuous responses (Fronczyk and Kottas, 2017), binary and continuous responses and litter size (Hwang and Pennell, 2018). The most relevant methodology is the one discussed in Kottas and Fronczyk (2013), which deals with ordinal responses as well. They use a product of Binomials as the kernel, to capture the nested structure of the responses, and a common-weights DDP prior for the dose-dependent mixing

distributions. We develop different (including more general) mixture models than the one in Kottas and Fronczyk (2013).

The rest of the chapter is organized as follows. In Section 3.2, we review the two typical continuous mixture models for accounting for overdispersion in ordinal responses, and demonstrate their limitations in uncertainty quantification with the EG data. The discrete nonparametric mixture models with either type of kernel are formulated and examined in depth in Section 3.3. Section 3.4 introduces two carefully designed simulation studies that reflect our main contributions. We compare the performance of the nonparametric mixture models through a series of risk assessments, conducted on the EG data. The main results are presented in Section 3.5. Finally, Section 3.6 concludes with a summary and discussion.

3.2 Continuous Mixture Models

This section focuses on providing an appropriate context for the models that will be examined later in this chapter. We start by reviewing properties of the classic Beta-Binomial and Logistic-Normal-Binomial distribution. Models built on them for ordinal responses in developmental toxicity study are also discussed.

3.2.1 Beta-Binomial and Logistic-Normal-Binomial

Both the BB and the LNB distribution can be viewed as a continuous mixture of the Binomial distribution. Consider modeling the number of positive responses, denoted by Y, among m trials. Then, the BB model assumes

$$Y \mid m, \theta, \lambda \sim BB(m, \theta, \lambda) := \int Bin(Y \mid m, \psi) Beta(\psi \mid \lambda\varphi(\theta), \lambda(1-\varphi(\theta)))d\psi.$$

On the other hand, the LNB model is formulated as

$$Y \mid m, \theta, \sigma^2 \sim LNB(m, \theta, \sigma^2) \coloneqq \int Bin(Y \mid m, \varphi(\psi)) N(\psi \mid \theta, \sigma^2) d\psi.$$

Here, $\varphi(x) = \exp(x)/(1 + \exp(x))$ denotes the standard logistic function. Under a regression setting, covariate effects can be incorporated into the model by setting $\theta = \theta(\mathbf{x})$.

For a deeper comprehension of these distributions, we consider the alternative encoding of Y with binary indicators $\{\tilde{Y}_q : q = 1, \ldots, m\}$, such that $Y = \sum_{q=1}^{m} \tilde{Y}_q$. Both the BB model and the LNB model postulate exchangeability, in lieu of independence, for $\tilde{Y}_q$, which induces marginal dependence among them. Capitalizing on overdispersion results for mixtures from exponential families (Shaked, 1980), we can show that the variance of Y under either of the models is larger than the variance of Y under a Binomial model, that is, the mixture models achieve overdispersion. The extent of overdispersion is controlled by the correlation between any pair of $\tilde{Y}_q$ and $\tilde{Y}_{q'}$, for $q, q' \in \{1, \ldots, m\}$.

Under the BB distribution, $\mathrm{E}(\tilde{Y}_q \mid \theta) = \varphi(\theta)$, which is the same as the mean under a Binomial distribution. For the correlation, $\mathrm{Corr}(\tilde{Y}_q, \tilde{Y}_{q'} \mid \lambda) = (1 + \lambda)^{-1}$. Therefore, λ controls the dependence among $\tilde{Y}_q$, hence the variance of Y, and is termed the overdispersion parameter.

Because the logit-normal integral in general does not have analytical form, neither $\mathrm{E}(\tilde{Y}_q \mid m, \theta, \sigma^2)$ nor $\mathrm{Corr}(\tilde{Y}_q, \tilde{Y}_{q'} \mid \theta, \sigma^2)$ are available in closed form for the LNB distribution. Nonetheless, we have the following approximation based on a second-order Taylor series expansion, which helps conceptualize the distribution.

Proposition 3.1. *Suppose* $\tilde{Y}_q \mid \psi \overset{i.i.d.}{\sim} Bern(\varphi(\psi))$, *for* $q = 1, \ldots, m$, *and* $\psi \mid$

$\theta, \sigma^2 \sim N(\theta, \sigma^2)$. *Then, marginalizing over* ψ,

$$\begin{aligned} E(\tilde{Y}_q \mid \theta, \sigma^2) &\approx \varphi(\theta) + \tfrac{\sigma^2}{2}\varphi''(\theta) \\ Corr(\tilde{Y}_q, \tilde{Y}_{q'} \mid \theta, \sigma^2) &\approx \tfrac{\sigma^2\varphi'(\theta)[4-\sigma^2(1-2\varphi(\theta))^2]}{4+\sigma^2(1-2\varphi(\theta))[2-4\varphi(\theta)-\sigma^2\varphi''(\theta)]}. \end{aligned} \tag{3.1}$$

The proof is shown in Appendix A.2. Proposition 3.1 reveals features of the LNB distribution, contrasting to the BB distribution, in two folds. Firstly, the LNB model introduces a fluctuation in the mean, with the magnitude managed by σ^2. Besides, both σ^2 and θ affect the correlation. To be aligned with the BB distribution, we term σ^2 the overdispersion parameter. Note however that the overdispersion parameter λ of the BB distribution affects only the variance of Y, while the LNB distribution σ^2 parameter influences both the mean and variance of Y.

In terms of Bayesian inference, the LNB model is more attractive. When the priors for hyperparameters are conditionally conjugate, leveraging the Pólya-Gamma data augmentation approach (Polson et al., 2013), we can obtain posterior samples of parameters through Gibbs sampling. On the contrary, Bayesian implementation of the BB model requires tuning Metropolis-Hasting samplers. For this reason, we choose the LNB distribution as the building block for the nonparametric mixture model with overdispersed kernel discussed in Section 3.3.2.

3.2.2 Models for Ordinal Responses from Developmental Toxicity Study

In developmental toxicity studies, the developing fetuses are at risk of fetal death due to the toxin insult. For those who survive the entire gestation period, malformation may be exhibited. The sequential nature of the response suggests factorizing the joint distribution as $p(R, y \mid m) = p(R \mid m)p(y \mid R, m)$. Let

$\mathbf{x} = (1, x)$, where x denotes the dose level. We omitted the subscript d and i for notation simplicity. The model that assumes each part of the factorization following a BB distribution is given by

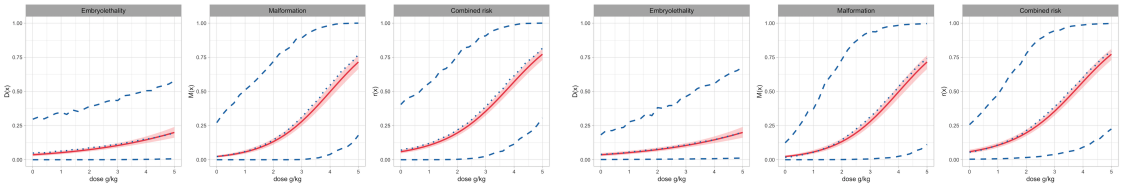
$$(R, y) \mid m, \theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \boldsymbol{\lambda} \sim BB(R \mid m, \theta_1(\mathbf{x}), \lambda_1) BB(y \mid m - R, \theta_2(\mathbf{x}), \lambda_2), \quad (3.2)$$

where $\theta_j(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}_j$, $j = 1, 2$, and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. Here, we assume a separate mixing distribution for each component, which is the key for effective interpretation and implementation of the model. Because of its induced interpretation for the response probabilities, the factorization is termed continuation-ratio in the literature. Accordingly, we term (3.2) the continuation-ratio Beta-Binomial ("CR-BB") model. Similarly, if the LNB distribution is used for each part of the factorization, the model is formulated as

$$(R, y) \mid m, \theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \boldsymbol{\sigma^2} \sim LNB(R \mid m, \theta_1(\mathbf{x}), \sigma_1^2) LNB(y \mid m - R, \theta_2(\mathbf{x}), \sigma_2^2), \quad (3.3)$$

where $\boldsymbol{\sigma^2} = (\sigma_1^2, \sigma_2^2)$, and it will be referred to as the continuation-ratio Logistic-Normal-Binomial ("CR-LNB") model.

To aid in exploring the relationship between the toxin level and the probability of the various endpoints, it is helpful to consider the underlying binary responses. In particular, for a generic dam with m implants exposed to toxin level x, we denote by $\tilde{\boldsymbol{R}} = \{\tilde{R}_q : q = 1, \cdots, m\}$ the non-viable fetus indicators, and $\tilde{\boldsymbol{y}} = \{\tilde{y}_l : l = 1, \cdots, m - \sum_{q=1}^{m} \tilde{R}_q\}$ the malformation indicators for the live pups, such that the extended ordinal response is $\mathbf{Y} = (R, y, m - R - y)$, where $R = \sum_{q=1}^{m} \tilde{R}_q$ and $y = \sum_{l=1}^{m-R} \tilde{y}_l$. The dose response curves are defined with the alternative encoding of the responses. Following the standard risk assessment methods in the literature (e.g. Krewski and Zhu, 1995), we consider the dose-response curves



(a) "CR-BB" model. **(b)** "CR-LNB" model.

Figure 3.2: EG data. Posterior mean (dotted line) and 95% interval estimate (dashed lines) for the dose response curves. The red solid line and shaded region is the posterior mean and 95% interval estimates obtained under a continuation-ratio logits model.

of embryolethality, malformation of viable fetus, and combined risk, implicitly conditioning on $m = 1$ and the model $\mathcal{M}$, defined respectively as $D(x) = \Pr(\tilde{R} = 1 \mid x)$, $M(x) = \Pr(\tilde{y} = 1 \mid \tilde{R} = 0, x)$, and $r(x) = \Pr(\tilde{R} = 1 \text{ or } \tilde{y} = 1 \mid x) = \Pr(\tilde{R} = 0 \text{ and } \tilde{y} = 1 \mid x) + \Pr(\tilde{R} = 1 \mid x)$.

For the EG data, we fit the "CR-BB" model and the "CR-LNB" model to obtain posterior inference for the dose-response curves. The resulting point and interval estimates are displayed in Figure 3.2, where, as a reference point, we also present the same inference under the continuation-ratio logits model. Without a mixing structure, the continuation-ratio logits model cannot account for overdispersion, leading to very narrow uncertainty bands. In contrast, the "CR-BB" and "CR-LNB" models provide overly wide interval estimates. This pattern emerges because the continuous mixture models pool the variability over the dose range, providing significant uncertainty even at dose levels with relatively small observed heterogeneity. Moreover, due to their parametric form, these models tend to overcompensate for the data heterogeneity by increasing the variability in the response distribution.

Such limitations of parametric continuous mixture models motivate us to consider discrete nonparametric mixture models, specifically the mixing structure induced by a dose-dependent stick-breaking process prior. By permitting clustered

mixing parameters, the discrete mixture models have the potential to manage the variability of response distribution more effectively. Next, we explore modeling approachess in this direction.

3.3 Discrete Mixture Models

3.3.1 Models with Continuation-ratio Logits Kernel

We consider a generalization of the continuation-ratio logits regression model via Bayesian nonparametric mixing. The model extension is achieved through a covariate-dependent nonparametric prior, $G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\, \delta_{(\theta_{1\ell}(\mathbf{x}),\theta_{2\ell}(\mathbf{x}))}$, leading to the general model

$$(R, y) \mid m, G_{\mathbf{x}} \sim \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\, Bin(R \mid m, \varphi(\theta_{1\ell}(\mathbf{x})))Bin(y \mid m - R, \varphi(\theta_{2\ell}(\mathbf{x}))). \quad (3.4)$$

As discussed in Chapter 2, the logit stick-breaking process (LSBP) prior has a structural similarity with the continuation-ratio logits, which offers key advantages in model properties and implementation. We assume the following LSBP prior for the covariate-dependent weights:

$$\omega_1(\mathbf{x}) = \varphi(\mathbf{x}^\top \boldsymbol{\gamma}_1),\ \omega_\ell(\mathbf{x}) = \varphi(\mathbf{x}^\top \boldsymbol{\gamma}_\ell) \prod_{h=1}^{\ell-1} (1 - \varphi(\mathbf{x}^\top \boldsymbol{\gamma}_h)),\ \ell \geq 2;\ \boldsymbol{\gamma}_\ell \overset{i.i.d.}{\sim} N(\boldsymbol{\gamma}_0, \Gamma_0) \quad (3.5)$$

In addition, the atoms are built through a linear regression structure,

$$\theta_{j\ell}(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}_{j\ell} \mid \boldsymbol{\mu}_j, \Sigma_j \overset{ind.}{\sim} N(\mathbf{x}^\top \boldsymbol{\mu}_j, \mathbf{x}^\top \Sigma_j \mathbf{x}), \quad j = 1, 2,\ \ \ell \geq 1, \quad (3.6)$$

with the random variables that define the atoms assumed a priori independent of those that define the weights. The model is completed with the conjugate prior

for the collection of hyperparameters $\boldsymbol{\psi} = \{\boldsymbol{\mu}_j, \Sigma_j : j = 1, 2\}$, that is,

$$\Sigma_j \overset{ind.}{\sim} IW(\nu_{0j}, \Lambda_{0j}^{-1}), \quad \boldsymbol{\mu}_j \mid \Sigma_j \overset{ind.}{\sim} N(\boldsymbol{\mu}_{0j}, \Sigma_j/\kappa_{0j}), \quad j = 1, 2. \tag{3.7}$$

We refer to the discrete mixture model in (3.4), with mixing weights and atoms specified respectively in (3.5) and (3.6), as the general mixture of product of Binomials kernel ("Gen-Bin") model.

We establish a useful connection of the nonparametric mixture model built for extended ordinal response, with a model built for the underlying response $\tilde{\boldsymbol{R}}$ and $\tilde{\boldsymbol{y}}$. Using the same nonparametric prior specified in (3.5) and (3.6), together with a product of Bernoullis kernel, the nonparametric mixture model for underlying binary response can be formulated as

$$(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{y}}) \mid m, G_{\mathbf{x}} \sim \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \prod_{q=1}^{m} Bern(\tilde{R}_q \mid \varphi(\theta_{1\ell}(\mathbf{x}))) \prod_{l=1}^{m-\sum_q \tilde{R}_q} Bern(\tilde{y}_l \mid \varphi(\theta_{2\ell}(\mathbf{x}))). \tag{3.8}$$

We can show that the mixture models (3.4) and (3.8) are equivalent in the sense that the moment generating function (MGF) of (R, y) under (3.4) is equal to the MGF of $(\sum \tilde{R}_q, \sum \tilde{y}_l)$ under (3.8). The result is formally stated in Proposition 3.2, with the proof presented in Appendix A.2.

Proposition 3.2. *Let $\mathcal{M}$ and $\tilde{\mathcal{M}}$ denote the mixture models (3.4) and (3.8), respectively. With the same m, and $G_{\mathbf{x}}$ formulated by (3.5) and (3.6),*

$$E_{\mathcal{M}}(e^{t_1 R + t_2 y} \mid m, G_{\mathbf{x}}) \;=\; E_{\tilde{\mathcal{M}}}(e^{t_1 \sum \tilde{R}_q + t_2 \sum \tilde{y}_l} \mid m, G_{\mathbf{x}}) \tag{3.9}$$

The subscript of the expectation refers to the distribution under which the expectation is taken.

Proposition 3.2 allows us to examine the dose-response curves for a dam with

a generic number of fetuses, which of course includes $m = 1$. Consequently, the expressions for the dose-response curves of embryolethality $D(x)$, malformation $M(x)$, and combined risk $r(x)$, under the proposed model, are given by

$$\begin{aligned} D(x) &= \Pr(\tilde{R} = 1 \mid G_{\mathbf{x}}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\, \varphi(\theta_{1\ell}(\mathbf{x})); \\ M(x) &= \Pr(\tilde{y} = 1 \mid \tilde{R} = 0, G_{\mathbf{x}}) = \sum_{\ell=1}^{\infty} \frac{\omega_\ell(\mathbf{x})[1 - \varphi(\theta_{1\ell}(\mathbf{x}))]}{\sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})[1 - \varphi(\theta_{1\ell}(\mathbf{x}))]}\, \varphi(\theta_{2\ell}(\mathbf{x})); \\ r(x) &= \Pr(\tilde{R} = 1 \text{ or } \tilde{y} = 1 \mid G_{\mathbf{x}}) = 1 - \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\, [1 - \varphi(\theta_{1\ell}(\mathbf{x}))][1 - \varphi(\theta_{2\ell}(\mathbf{x}))], \end{aligned} \tag{3.10}$$

with $\omega_\ell(\mathbf{x})$ and $\theta_{1\ell}(\mathbf{x})$, $\theta_{2\ell}(\mathbf{x})$ defined in (3.5) and (3.6), respectively. Note that all three dose-response curves admit a weighted sum representation with covariate-dependent weights, which enables local adjustment over the dose level, resulting in flexible estimation of the dose-response relationships.

Another equivalent encoding of the responses comes from the connection between the standard and extended ordinal response. Indeed, let $\{\tilde{\boldsymbol{Y}}_q : q = 1, \cdots, m\}$ be a collection of standard ordinal responses. That is, $\tilde{\boldsymbol{Y}}_q = (\tilde{Y}_{q1}, \tilde{Y}_{q2}, \tilde{Y}_{q3})$, where $\tilde{Y}_{qj}$ are binary, and only one $\tilde{Y}_{qj} = 1$, for $j = 1, 2, 3$. We can view $\tilde{\boldsymbol{Y}}_q$ as the ordinal response from an implant of the dam. They are linked with $\mathbf{Y} = (R, y, m-R-y)$ through $\mathbf{Y} = \sum_{q=1}^{m} \tilde{\boldsymbol{Y}}_q$. In addition, $\tilde{\boldsymbol{Y}}_q$ are connected with $\tilde{R}_q$ and $\tilde{y}_l$ through the sequential mechanism of the continuation-ratio logits structure, depicted in Figure 3.3. Introducing $\tilde{\boldsymbol{Y}}_q$ facilitates the study of overdispersion.

In the context of development toxicology studies, responses from the fetuses within the same dam are typically assumed to be positively correlated, resulting in overdispersion. Therefore, relevant modeling methods should promote positive intracluster correlations. Here, the cluster refers to the dam. Under the proposed model, the intracluster correlation at category j, $j = 1, 2, 3$, for any implants q

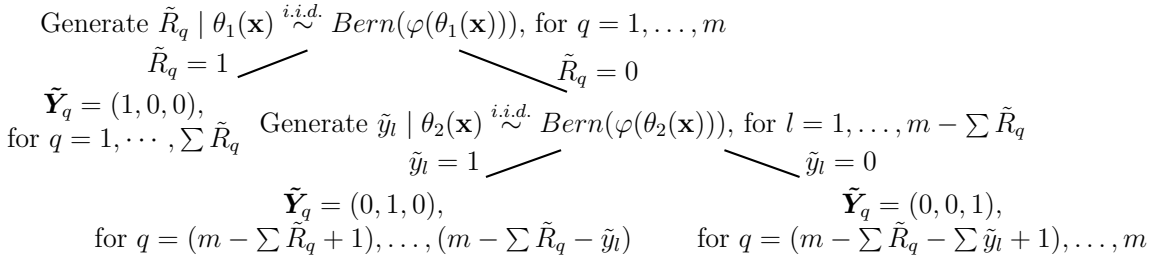


Figure 3.3: Connection between alternative encodings of the ordinal response.

and q' from the same dam, is given by

$$\text{Corr}(\tilde{Y}_{qj}, \tilde{Y}_{q'j} \mid G_{\mathbf{x}}) = \frac{\text{E}(\tilde{Y}_{qj}\tilde{Y}_{q'j} \mid G_{\mathbf{x}}) - \text{E}(\tilde{Y}_{qj} \mid G_{\mathbf{x}})\text{E}(\tilde{Y}_{q'j} \mid G_{\mathbf{x}})}{\{\text{Var}(\tilde{Y}_{qj} \mid G_{\mathbf{x}})\text{Var}(\tilde{Y}_{q'j} \mid G_{\mathbf{x}})\}^{1/2}}, \tag{3.11}$$

where $\text{E}(\tilde{Y}_{qj} \mid G_{\mathbf{x}}) = \text{E}(\tilde{Y}_{q'j} \mid G_{\mathbf{x}}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \{\varphi(\theta_{j\ell}(\mathbf{x})) \prod_{k=1}^{j-1}[1 - \varphi(\theta_{k\ell}(\mathbf{x}))]\}$, $\text{Var}(\tilde{Y}_{qj} \mid G_{\mathbf{x}}) = \text{Var}(\tilde{Y}_{q'j} \mid G_{\mathbf{x}}) = \text{E}(\tilde{Y}_{qj} \mid G_{\mathbf{x}}) - [\text{E}(\tilde{Y}_{qj} \mid G_{\mathbf{x}})]^2$, and $\text{E}(\tilde{Y}_{qj}\tilde{Y}_{q'j} \mid G_{\mathbf{x}}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \{\varphi(\theta_{j\ell}(\mathbf{x})) \prod_{k=1}^{j-1}[1 - \varphi(\theta_{k\ell}(\mathbf{x})]\}^2$, with $\varphi(\theta_{3\ell}(\mathbf{x})) \equiv 1$. Fronczyk and Kottas (2014) have shown that the intracluster correlation is positive under a common-weights DDP mixture of Binomial distributions. The required assumptions are that the variance, $\text{Var}(\tilde{Y}_{qj} \mid G_{\mathbf{x}})$, and correlation, $\text{Corr}(\tilde{Y}_{qj}, \tilde{Y}_{q'j} \mid G_{\mathbf{x}})$, are common within the cluster. These assumptions hold here, since any pair of $\tilde{Y}_{qj}$, $\tilde{Y}_{q'j}$ are associated with the same dose level x. As a result, the positive intracluster correlations result extends to our case, i.e., $\text{Corr}(\tilde{Y}_{qj}, \tilde{Y}_{q'j} \mid G_{\mathbf{x}}) > 0$, $\forall\, j$.

A practically relevant modeling aspect revolves around possible monotonicity restrictions for the dose-response functions. Developmental toxicity studies involve a small number of administered toxin levels. Hence, under nonparametric mixture models for the categorical responses, a monotonic trend in the prior expectation for the dose-response curves is desirable for effective interpolation and extrapolation inference. This is discussed in Kottas and Fronczyk (2013) and Fronczyk and Kottas (2014) under common-weights DDP mixture models, and is also relevant

in our model setting. Using the prior specification strategy of Section 2.2.3, we can incorporate a non-decreasing trend in the prior expected dose-response curves. We note however that prior (and thus posterior) realizations for the dose-response curves are not structurally restricted to be non-decreasing.

Two simplifications of the general model are discussed in Section 2.3, namely the common-weights model and the common-atoms model. Due to the monotonicity restriction of the prior expectation for the dose-response curves, the common-atoms model is not a practical option. This is because the common-atoms model adjusts the shape of dose-response curves only through the weights, resulting in prior expectations that are constant with respect to the toxin level covariate. Nonetheless, the common-weights model is worth exploring, because it bridges the general nonparametric mixture model proposed here with the model discussed in Kottas and Fronczyk (2013). The common-weights mixture with product of Binomial kernels ("CW-Bin") model is specified as

$$(R, y) \mid m, G_{\mathbf{x}} \sim \sum_{\ell=1}^{\infty} \omega_\ell \, Bin(R \mid m, \varphi(\theta_{1\ell}(\mathbf{x}))) Bin(y \mid m - R, \varphi(\theta_{2\ell}(\mathbf{x}))),$$

with $\omega_1 = V_1$, and $\omega_\ell = V_\ell \prod_{h=1}^{\ell-1}(1 - V_h)$, for $\ell \geq 2$, where $V_\ell \mid \alpha \overset{i.i.d.}{\sim} Beta(1, \alpha)$, and $\theta_{1\ell}(\mathbf{x})$, $\theta_{2\ell}(\mathbf{x})$ defined in (3.6). Kottas and Fronczyk (2013) adopt the same structure for the weights in their mixture model, while the atoms are chosen as Gaussian processes with the mean function postulating a linear regression form. They extend the common-weights model by incorporating a more flexible structure for the atoms. Note that their model still does not allow the dose-response curves for the embryolethality and combined negative outcome to have dose-dependent weights, which is an asset of our general model.

For Markov chain Monte Carlo (MCMC) posterior simulation, we notice that the blocked Gibbs sampler proposed in Chapter 2 is also applicable to conduct

posterior simulation with the "Gen-Bin" model and the "CW-Bin" model. Posterior realizations for the dose-response curves and intracluster correlations can be obtained by evaluating the corresponding expressions with MCMC posterior samples of model parameters. Moreover, for each endpoint, we can obtain the posterior distribution of a calibrated dose level for a specified probability, by (numerically) inverting the posterior realization of the corresponding dose-response curve. We illustrate the procedure with the EG data in Section 3.5.

The discrete mixing structure in conjunction with the restricted kernel implies *a priori* a trade-off between the variability of the response and the variability of the dose-response curve. Because overdispersion is not admitted in the kernel, the mixture model seeks to account for the vast variability in the response by activating more effective components. Contrarily, because of the discrete mixture structure, more effective components lead to less variability in the prior realizations of dose response curves, yielding overconfident prior intervals. Seeking coherent uncertainty quantification for both the response distribution and the dose-response curves, we consider building discrete mixture models with a kernel that allows higher level of dispersion.

3.3.2 Models with Overdispersed Kernel

Parallel to the development of the "Gen-Bin" model, we formulate the alternative modeling approach with overdispersed kernel starting from its parametric backbone in (3.3). Amplified with the general dose-dependent nonparametric prior we obtain

$$(R, y) \mid m, G_{\mathbf{x}}, \boldsymbol{\sigma}^{\mathbf{2}} \sim \sum_{\ell=1}^{\infty} \omega_{\ell}(\mathbf{x})\, LNB(R \mid m, \theta_{1\ell}(\mathbf{x}), \sigma_1^2) LNB(y \mid m-R, \theta_{2\ell}(\mathbf{x}), \sigma_2^2), \tag{3.12}$$

The prior on the weights $\omega_\ell(\mathbf{x})$ is specified as the same LSBP prior given in (3.5), while the atoms $\theta_{1\ell}(\mathbf{x})$, $\theta_{2\ell}(\mathbf{x})$ and their prior are specified as in (3.6) and (3.7). The model formulation is completed with $\sigma_j^2 \overset{i.i.d.}{\sim} IG(a_\sigma, b_\sigma)$, for $j = 1, 2$. This model formulation shall be referred to as the general mixture with product of LNB kernel ("Gen-LNB") model hereinafter.

The "Gen-LNB" model includes both the "CR-LNB" model and the "Gen-Bin" model as special (limiting) cases. If $\boldsymbol{\gamma}_1$ is such that $\varphi(\mathbf{x}^T\boldsymbol{\gamma}_1)$ is effectively equal to one, the nonparametric model collapses to the model in (3.3). If we let $\sigma_j^2 \to 0^+$ for $j = 1, 2$, the kernel collapses to the continuation-ratio logits model, resulting in the "Gen-Bin" model. The specific mixing structure allows smooth deviations from the Binomial, while keeping the extra level of flexibility, brought in by the discrete (infinite) mixture.

To investigate model properties, we build its connection with the nonparametric mixture model for the underlying $\tilde{\boldsymbol{R}}$ and $\tilde{\boldsymbol{y}}$. Specifically, consider the following model,

$$\begin{aligned}
&(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{y}}) \mid m, \psi_1, \psi_2 \sim \prod_{q=1}^{m} Bern(\tilde{R}_q \mid \varphi(\psi_1)) \prod_{l=1}^{m-\sum_q \tilde{R}_q} Bern(\tilde{y}_l \mid \varphi(\psi_2)), \\
&(\psi_1, \psi_2) \mid \theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \boldsymbol{\sigma}^{\mathbf{2}} \sim N(\psi_1 \mid \theta_1(\mathbf{x}), \sigma_1^2) N(\psi_2 \mid \theta_2(\mathbf{x}), \sigma_2^2), \\
&(\theta_1(\mathbf{x}), \theta_2(\mathbf{x})) \mid G_{\mathbf{x}} \sim G_{\mathbf{x}},\ G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \delta_{(\theta_{1\ell}(\mathbf{x}), \theta_{2\ell}(\mathbf{x}))},
\end{aligned} \tag{3.13}$$

with the same prior on $G_{\mathbf{x}}$ and $\boldsymbol{\sigma}^{\mathbf{2}}$ as for the "Gen-LNB" model. Then, the two model formulations are equivalent in terms of an equal MGF for the respective (R, y) and $(\sum \tilde{R}_q, \sum \tilde{y}_l)$.

Proposition 3.3. *With the same m and $G_{\mathbf{x}}$, equation (3.9) holds for $\mathcal{M}$ and $\tilde{\mathcal{M}}$, that is, the mixture models defined in (3.12) and (3.13), respectively.*

Proposition 3.3 allows us to implicitly condition on $m = 1$ when conduct-

ing inference for the dose-response curves. We denote the logit-normal integral $\int \varphi(\psi) N(\psi \mid \theta, \sigma^2) d\psi$ by $\varepsilon(\theta, \sigma^2)$. The expressions for dose-response curves at the aforementioned three endpoints are given by

$$
\begin{aligned}
D(x) &= \Pr(\tilde{R} = 1 \mid G_{\mathbf{x}}, \boldsymbol{\sigma^2}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\, \varepsilon(\theta_{1\ell}(\mathbf{x}), \sigma_1^2); \\
M(x) &= \Pr(\tilde{y} = 1 \mid \tilde{R} = 0, G_{\mathbf{x}}, \boldsymbol{\sigma^2}) = \sum_{\ell=1}^{\infty} \frac{\omega_\ell(\mathbf{x})[1 - \varepsilon(\theta_{1\ell}(\mathbf{x}), \sigma_1^2)]}{\sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})[1 - \varepsilon(\theta_{1\ell}(\mathbf{x}), \sigma_1^2)]}\, \varepsilon(\theta_{2\ell}(\mathbf{x}), \sigma_2^2); \\
r(x) &= \Pr(\tilde{R} = 1 \text{ or } \tilde{y} = 1 \mid G_{\mathbf{x}}, \boldsymbol{\sigma^2}) \\
&= 1 - \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\, [1 - \varepsilon(\theta_{1\ell}(\mathbf{x}), \sigma_1^2)][1 - \varepsilon(\theta_{2\ell}(\mathbf{x}), \sigma_2^2)].
\end{aligned}
$$

Flexible inference for the dose-response curves is again enabled with local-adjustable mixing weights.

The intracluster correlation under the general model with overdispersed kernel has a similar form as in (3.11), in which every component should include further conditioning on $\boldsymbol{\sigma^2}$. Specifically, $\mathrm{E}(\tilde{Y}_{qj} \mid G_{\mathbf{x}}, \boldsymbol{\sigma}^2) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \{\varepsilon(\theta_{j\ell}(\mathbf{x}), \sigma_j^2) \prod_{k=1}^{j-1} [1 - \varepsilon(\theta_{k\ell}(\mathbf{x}), \sigma_k^2)]\}$, $\mathrm{Var}(\tilde{Y}_{qj} \mid G_{\mathbf{x}}) = \mathrm{E}(\tilde{Y}_{qj} \mid G_{\mathbf{x}}) - [\mathrm{E}(\tilde{Y}_{qj} \mid G_{\mathbf{x}})]^2$, $\forall\, q \in \{1, \ldots, m\}$. Additionally, $\mathrm{E}(\tilde{Y}_{qj}\tilde{Y}_{q'j} \mid G_{\mathbf{x}}, \boldsymbol{\sigma^2}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \{\int \varphi^2(\psi_j) N(\psi_j \mid \theta_{j\ell}(\mathbf{x}), \sigma_j^2) d\psi_j\} \{\prod_{k=1}^{j-1} \int [1 - \varphi(\psi_k)]^2 N(\psi_k \mid \theta_{k\ell}(\mathbf{x}), \sigma_k^2) d\psi_k\}$. We set $\varphi(\theta_{3\ell}(\mathbf{x})) \equiv 1$ and $\sigma_3^2 = 0$. The positive intracluster correlation property can also be established in this context, as the model continues to assume a shared variance/correlation within the dam.

To obtain meaningful inference, it is important to use a proper, well-calibrated prior for $\boldsymbol{\sigma^2}$. This is a challenging task because of the lack of analytical form for the logit-normal integral. Nonetheless, we propose a general strategy for specifying the $IG(a_\sigma, b_\sigma)$ prior, based on the approximation in Proposition 3.1, and working with the mixture kernel, i.e., the LNB distribution. From the second line of (3.1), and noticing $\varphi(\theta) \in (0, 1)$, we can show that for small to moderate (but still providing enough variability) σ^2, the intracluster correlation is approximately $\sigma^2/4$.

Simple calculation yields that modeling by LNB in lieu of Binomial provides an extra $(m-1)\sigma^2/4$ folds of the variance In practice, we use a prior guess about the average variance deviation of R and y from the Binomial across the dose levels, and set the prior for $\boldsymbol{\sigma^2}$ accordingly, such that the overdispersion provided by the LNB kernel is enough to capture the extra variation. The other prior hyperparameters can be specified in the same fashion as the general model. Specifically, the prior specification strategy that ensures a monotonic trend in the prior expectation of dose-response curves can still be applied here.

Another appealing feature of the proposed model comes from the posterior simulation perspective. The mixing structure of the model is inherited from the "Gen-Bin" model, rendering the computational techniques developed for it readily adaptable here. We develop a blocked Gibbs sampler based on the MCMC algorithm in Appendix B.1, with modifications to account for the extra continuous mixing at the kernel. The detailed algorithm is presented in Appendix B.2. With the MCMC samples of model parameters, we can conduct any type of relevant inference, following the same procedure as the general model with original kernel.

To complete the spectrum of the proposed models, we also consider the simplification by removing the dose dependence in the mixing weights. That is, instead of determining weights through a LSBP prior, we use the stick-breaking formulation corresponding to the DP. We term this the common-weights mixture with product of LNB kernel ("CW-LNB") model.

3.4 Synthetic Data Examples

We conduct simulation experiments to demonstrate the practical benefits of using nonparametric mixture models in development toxicity study. Specifically, the first experiment is designed to highlight the benefits of local, dose-dependent

weights in capturing non-standard dose-response relationships. The objective of the second experiment is to illustrate the utility of the overdispersed kernel in capturing the vast heterogeneity of the data.

3.4.1 First Synthetic Data Example

For the first experiment, we consider four active dose level at 0.625, 1.25, 2.5, and 5 g/kg and a control group. We consider a total of $n = 100$ dams, evenly distributed across the dose levels. For each dam, the number of implants are generated from a Poisson distribution with mean 20. Conditioning on the number of implants, the responses are generated from a three component mixture of "CR-LNB" model, with dose-dependent model parameters. That is,

$$(R_{di}, y_{di}) \mid m_{di} \overset{ind.}{\sim} \sum_{k=1}^{3} w_k(\mathbf{x}_d)\, LNB(R_{di} \mid m_{di}, \theta_{1k}(\mathbf{x}_d), \sigma_1^2(\mathbf{x}_d)) \times LNB(y_{di} \mid m_{di} - R_{di}, \theta_{2k}(\mathbf{x}_d), \sigma_2^2(\mathbf{x}_d)),$$

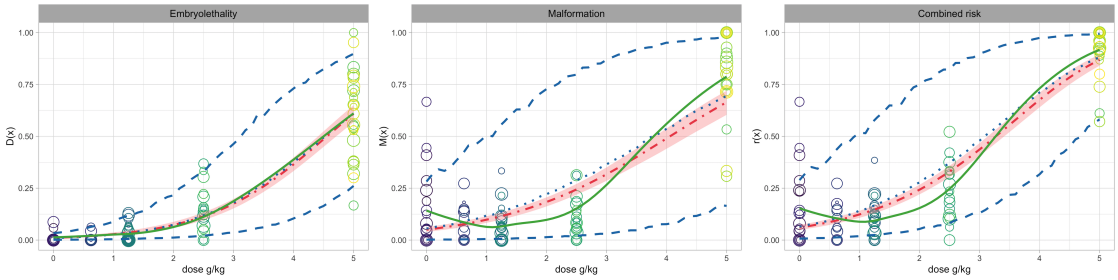
where $\mathbf{x}_d = (1, x_d)$, $\theta_{jk}(\mathbf{x}_d) = b_{jk0} + b_{jk1}x_d$, for $j = 1, 2$ and $k = 1, 2, 3$. The dose-dependent weights are induced by computing $p_j(\mathbf{x}_d) = \Phi(a_{j0} + a_{j1}x_d)$, for $j = 1, 2$, where $\Phi(\cdot)$ denotes the c.d.f. of the standard normal distribution, and setting $(w_1(\mathbf{x}_d), w_2(\mathbf{x}_d), w_3(\mathbf{x}_d)) = (p_1(\mathbf{x}_d), (1 - p_1(\mathbf{x}_d))p_2(\mathbf{x}_d), (1 - p_1(\mathbf{x}_d))(1 - p_2(\mathbf{x}_d)))$. Additionally, $\sigma_j^2(\mathbf{x}_d) = c_{j0} + c_{j1}x_d$, for $j = 1, 2$, and the parameters are chosen to ensure $\sigma^2(\mathbf{x}_d) > 0$.

We visualize the simulated data set in Figure 3.4a. For each dam, we plot the observed R_{di}/m_{di}, $y_{di}/(m_{di} - R_{di})$, and $(R_{di} + y_{di})/m_{di}$. In each panel, the solid line indicates the true dose-response curve. We intentionally set the true dose-response curves for the malformation and the combined negative outcome endpoints to exhibit a J-shape, which is referred to as the hormetic dose-response

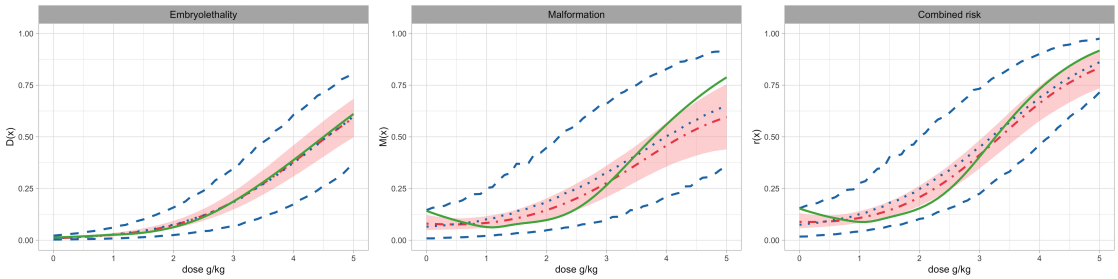
relationship in the toxicological sciences. Hormesis is a dose–response phenomenon characterized by beneficial effect to low exposures to toxins, and thus by opposite effects in small and large doses. The posterior mean and 95% interval estimates of dose-response curves under parametric models, i.e., the continuation-ratio logits model and the "CR-LNB" model, are also shown in Figure 3.4a. As expected, the standard models cannot capture the dip in the dose-response curves.

Figure 3.4 displays the posterior point and interval estimates of the dose-response curves under the nonparametric mixture models. All the models considered here capture the true dose-response curve for the embryolethality endpoint well, providing point estimates that are almost identical to the true monotonically increasing function. As for the dose-response curves corresponding to the malformation and combined risk endpoints, the two nonparametric mixture models without dose-dependent mixing weights ("CW-Bin" model and "CW-LNB" model) provide improved point and interval estimates comparing to their parametric backbones, but still fail in depicting the non-monotonic shape. On the contrary, the "Gen-Bin" model and the "Gen-LNB" model, permitting more efficient local adjustments, capture the dip in these dose-response curves. The comparison demonstrates the benefit of using the mixture model with dose-dependent weights, especially when the dose-response curves are expected to have non-standard shapes.

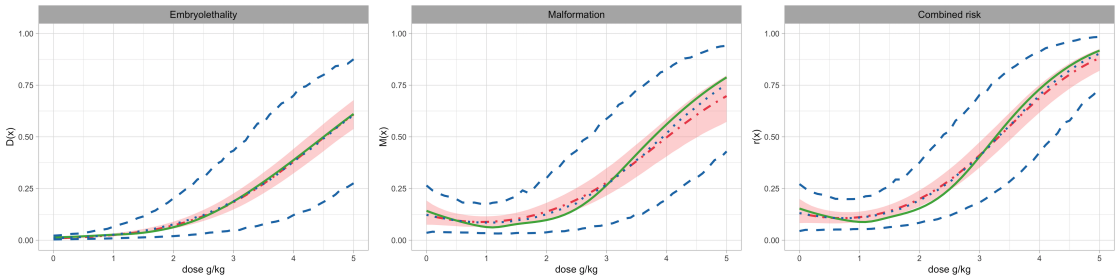
To further explore how the different nonparametric models utilize the mixture structure, Figure 3.5 shows the posterior distributions of the four largest mixture weights across dose levels. The models without overdispersed kernel generally activate more mixing components. This is to be expected, because these models rely on the mixing structure to account for overdispersion. The "Gen-Bin" model tends to use more mixing components at low dose region to help capture the dip of the dose-response curves. Equipped with overdispersed kernel, the "CW-LNB"



(a) Continuation-ratio logits model vs "CR-LNB" model.



(b) "CW-Bin" model vs "CW-LNB" model.



(c) "Gen-Bin" model vs "Gen-LNB" model.

Figure 3.4: First simulation example. Posterior mean and 95% interval estimates for the dose response curves under the mixture models with different kernel. In each panel, the posterior mean and interval estimates obtained under the model with and without overdispersed kernel are given by the blue dotted and dashed lines and the red dot-dashed line and shaded region, respectively. The green solid line is the true dose-response curve. In the top panel, a circle corresponds to a particular dam and the size of the circle is proportional to the number of implants.

model and the "Gen-LNB" model are more efficient in terms of the number of effective mixture components. Specifically, under the "Gen-LNB" model, we note the pronounced local adjustment of the mixing weights in the dose region where the dose-response curves change more drastically. On the contrary, the "CW-LNB"

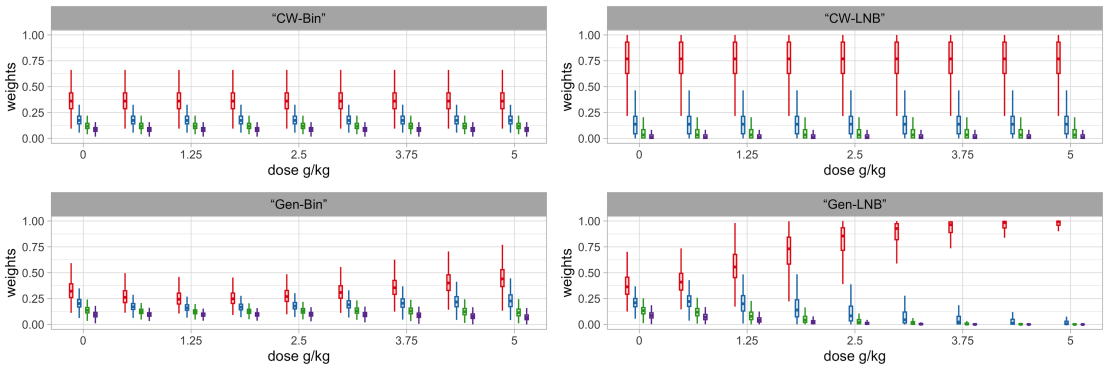


Figure 3.5: First simulation example. Box plots of the posterior samples for the four largest mixture weights at the four observed dose levels and a new dose level, under each of the nonparametric models.

model does not allow for local adjustable mixture weights, and thus can not capture as effectively the non-standard local behavior of the malformation and combined risk dose-response curves.

3.4.2 Second Synthetic Data Example

We consider active dose levels at 0.625, 1.25, 2.5, 3.75, and 5 g/kg, and a control group. A total of $n = 150$ dams are randomly assigned across the dose levels with uniform probability. We adopt the same process with the first simulation example to generate the number of implants for each dam. Then, the ordinal responses are obtained by sampling from

$$(R_{di}, y_{di}) \mid m_{di} \overset{ind.}{\sim} BB(R_{di} \mid m_{di}, \theta_1(\mathbf{x}_d), \lambda_1(\mathbf{x}_d))BB(y_{di} \mid m_{di} - R_{di}, \theta_2(\mathbf{x}_d), \lambda_2(\mathbf{x}_d)),$$

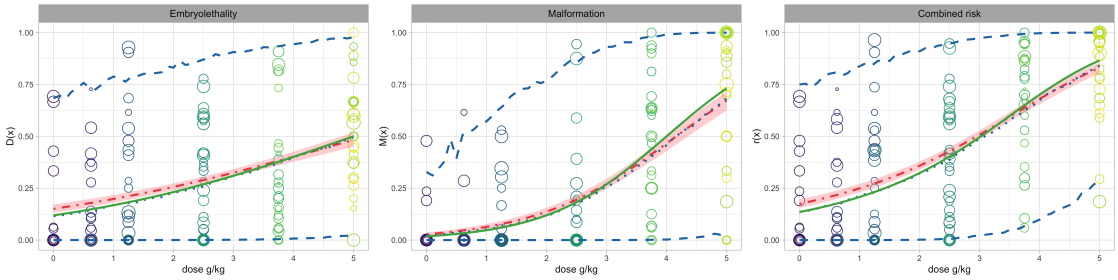
where $\theta_j(\mathbf{x}_d) = b_{j0} + b_{j1}x_d$, and $\lambda_j(\mathbf{x}_d) = c_{j0} + c_{j1}x_d > 0$, for $j = 1, 2$. The data are visualized in Figure 3.6a, including the posterior point and 95% interval estimates under the continuation-ratio logits model and the "CR-BB" model. Although the true dose-response curves here have relatively standard increasing

shape, the parametric models suffer in uncertainty quantification, due to the vast heterogeneity of the data. In particular, the "CR-BB" model is similar to the true data generating process, but the interval estimates obtained under it are too wide to be practically useful.

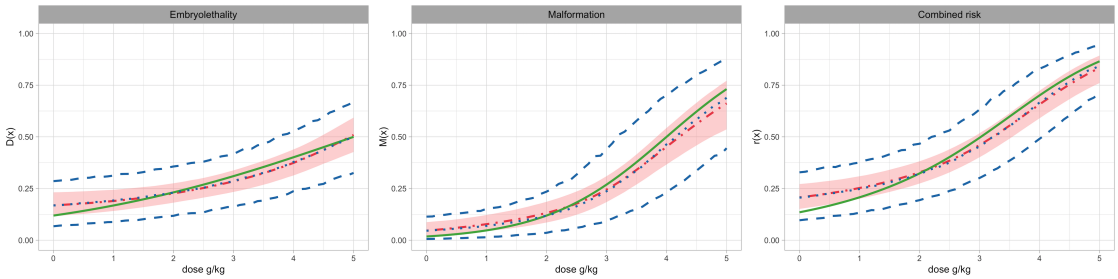
The nonparametric mixture models are applied to the data. Figure 3.6 plots posterior point and interval estimates for the dose-response curves. The nonparametric mixture models behave comparably in terms of recovering the underlying regression curves, evidenced by the similar posterior mean for the dose-response curves. The interval estimates under all the models capture the true dose-response curves. As expected, models with overdispersed kernel result in wider posterior uncertainty bands than the models with the continuation-ratio logits kernel. Observing the extensive dispersion in the data, a wider uncertainty band is arguably more reasonable.

To further investigate how the nonparametric mixture models behave in capturing the overdispersion of the data, we plot in Figure 3.7 posterior samples of the intracluster correlation. To facilitate comparison, we calculate the true intracluster correlations from the data generating process, and add them to the plot. In general, the models with overdispersed kernel perform better in capturing the truth. Without the help from overdispersed kernel, the "CW-Bin" model and the "Gen-Bin" model rely on the mixture structure to account for the extra dispersion, and are less effective in capturing the variability.

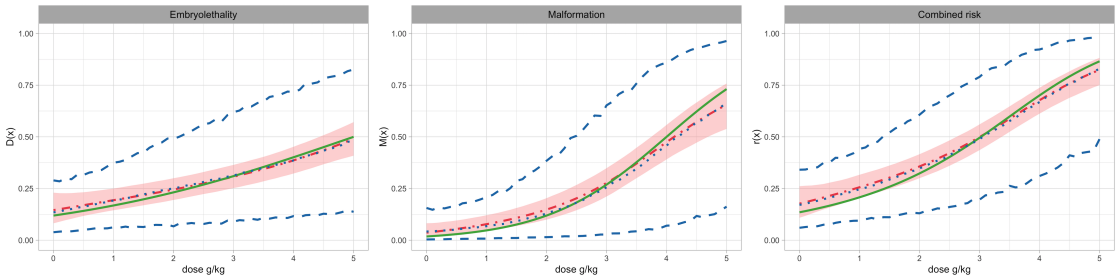
Finally, we conduct a sensitivity analysis regarding the prior of $\boldsymbol{\sigma^2}$ under the "CW-LNB" and "Gen-LNB" models. The results discussed above correspond to prior $\sigma_j^2 \sim IG(3, 8/3)$, for $j = 1, 2$, leading to an extra 33% variance on average a priori. We consider a more diffuse prior, namely $IG(2, 4/3)$, which induces the same level of extra variance on average. The posterior distributions of $\boldsymbol{\sigma^2}$ are



(a) Continuation-ratio logits model vs "CR-BB" model.



(b) "CW-Bin" model vs "CW-LNB" model.



(c) "Gen-Bin" model vs "Gen-LNB" model.

Figure 3.6: Second simulation example. Posterior mean and 95% interval estimates for the dose response curves under the mixture models with different kernel. In each panel, the posterior mean and interval estimates obtained under the model with and without overdispersed kernel are given by the blue dotted and dashed lines and the red dot-dashed line and shaded region, respectively. The green solid line is the true dose-response curve. In the top panel, a circle corresponds to a particular dam and the size of the circle is proportional to the number of implants.

shown in Figure 3.8. These distributions are comparable (and different from the prior distribution), suggesting effective learning of overdispersion from the data.

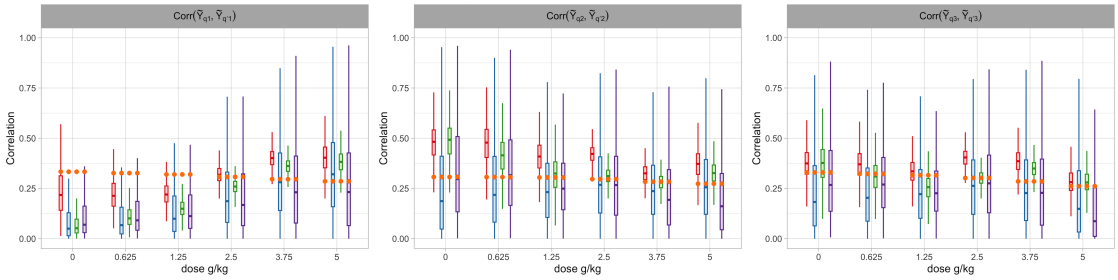


Figure 3.7: Second simulation example. Box plot of the intracluster correlation posterior distributions at the observed toxin levels. In each panel, estimates under the "CW-Bin", "CW-LNB", "Gen-Bin" and "Gen-LNB" model are shown in red, blue, green, and purple, respectively. The orange dot marks the truth.

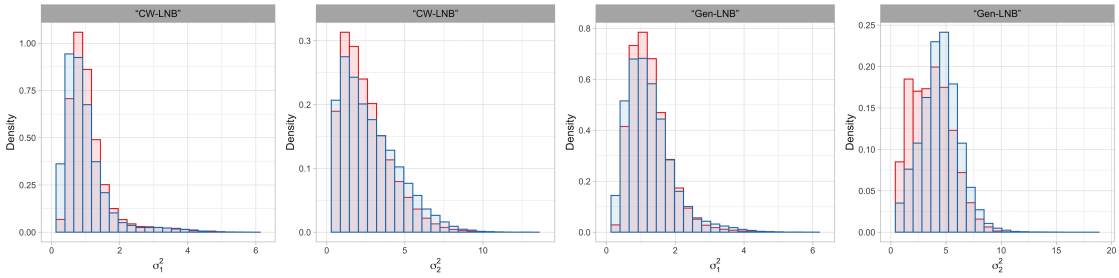


Figure 3.8: Second simulation example. Posterior distributions of the overdispersion parameters $\boldsymbol{\sigma^2}$ under the "CW-LNB" and the "Gen-LNB" model with prior $IG(3, 8/3)$ (in red) and $IG(2, 4/3)$ (in blue).

3.5 Real Data Illustrations

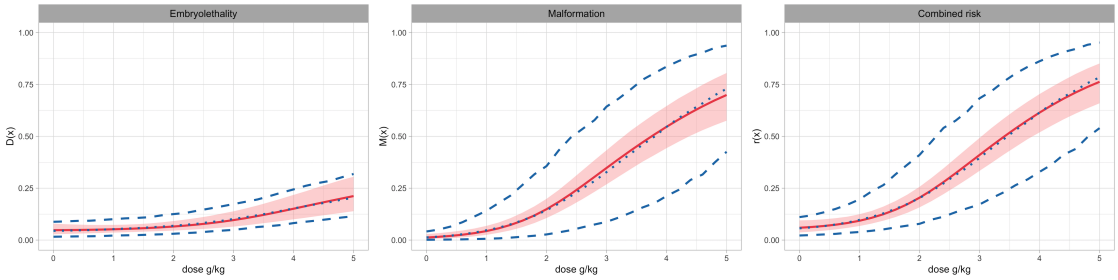
Working with the EG data, we illustrate the four nonparametric mixture models in addressing a spectrum of risk assessment tasks. We work with the (conservative) truncation level of $L = 50$ for the blocked Gibbs sampler. Posterior inference results are based on 5000 MCMC samples obtained every 2 iterations from a chain of 30000 iterations with a 20000 burn-in period.

We set the prior hyperparameters such that a monotonic increasing trend is incorporated in the prior expected dose-response curves. In line with this objective, the key is to specify $\boldsymbol{\mu}_{0j}$ and Λ_{0j}, $j = 1, 2$, by the strategy proposed in Section 2.2.3. The hyperparameters regarding the mixing weights under either the common-

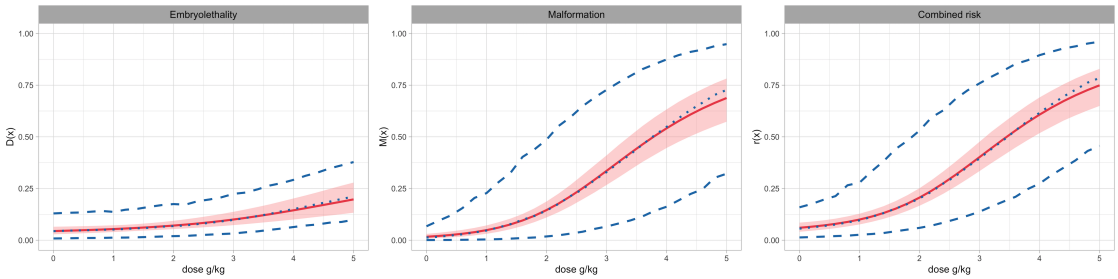
weights mixture or the general mixture are set such that they favor a priori a comparable number of distinct components. We use $IG(3, 1.2)$ as the prior for the overdispersion parameters in $\boldsymbol{\sigma^2}$, which provides approximately an extra 15% variance on average. For prior sensitivity analysis, we assume an alternative, more diffused prior, that is, $IG(2, 0.6)$. Despite the choice of prior and mixing structure, the posterior distributions of σ_1^2 and σ_2^2 are comparable.

Posterior estimates of dose-response curves under the discrete mixture models are displayed in Figure 3.9. The difference among the posterior point estimates of the dose-response curves is minor. The uncertainty bands provided by the models with overdispersed kernel are significantly wider, with the width changing across toxin levels. As shown in Figure 3.1, the variability of the responses increases with dose level. Uncertainty bands obtained under the "CW-Bin" and "Gen-Bin" model capture the trend in general, while they seem to underestimate the influence of the dose levels. Moreover, illustrated by a wider interval compared to that at 5 g/kg, the models with overdispersed kernel intensify the uncertainty at the region from 3 g/kg to 4 g/kg, where interpolation is actually needed. Without the help of overdispersed kernel, the models tend to be overconfident at this region. We also notice that comparing with the estimates under the continuous mixture model (Figure 3.2), the uncertainty bands obtained here are more plausible, indicating more effective control of variability under the discrete nonparametric mixture models.

We display posterior samples of the intracluster correlations at the observed toxin levels and a new level $x = 3.75$ g/kg, with the box plots in Figure 3.10. Despite the model and endpoint, the correlations depict an increasing trend with toxin levels, consistent with the observed data pattern. Moreover, the intracluster correlation at the new dose level is approximately the average of the correlations at the two



(a) "CW-Bin" model vs "CW-LNB" model.



(b) "Gen-Bin" model vs "Gen-LNB" model.

Figure 3.9: EG data. Posterior mean and 95% interval estimate for the dose response curves under the mixture models with different kernel. In each panel, the red solid line and shaded region is the posterior mean and interval estimates obtained under the model with continuation-ratio logits kernel, while the blue dotted and dashed lines are the estimates from the model with overdispersed kernel.

observed neighbors, indicating a smooth borrowing of strength across dose levels. As expected, the distribution of correlations from models with overdispersed kernel spread a wider range. Also noteworthy is that the magnitude of the intracluster correlation under the "CW-Bin" model is generally larger than the other models, which also means a larger variance for the response. However, as shown in Figure 3.9a, the posterior uncertainty for the dose-response curve under the "CW-Bin" model is shorter. This incongruity indicates a weak control of variability under the "CW-Bin" model, in which both the mixing structure and the kernel are restricted.

Estimating the effective dose (ED) and the benchmark dose (BMD) is crucial in developmental toxicity risk assessment. The procedure initiates with specifying the benchmark response level (BMR), denoted by α. After a dose-response model

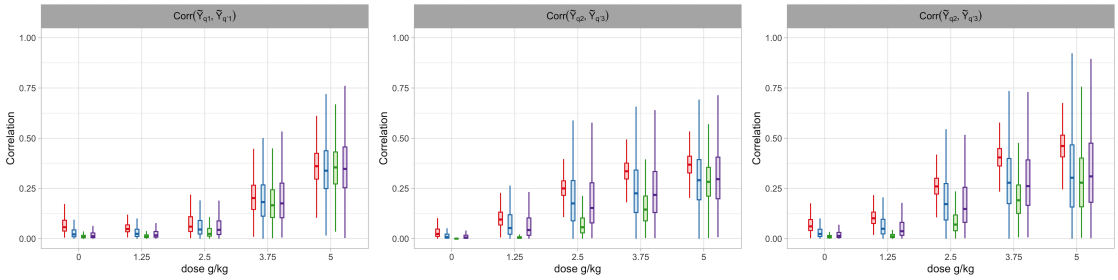
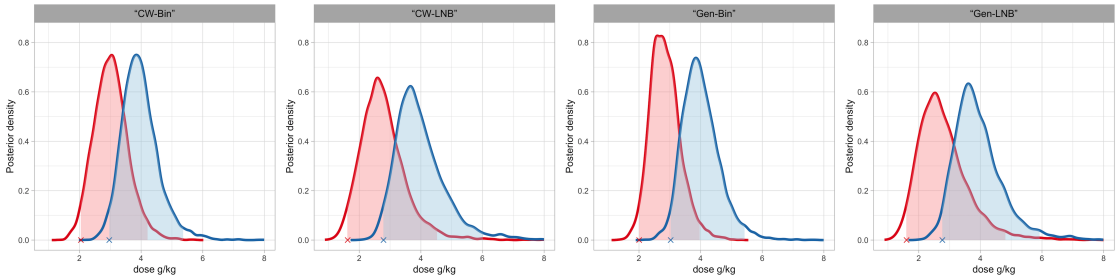


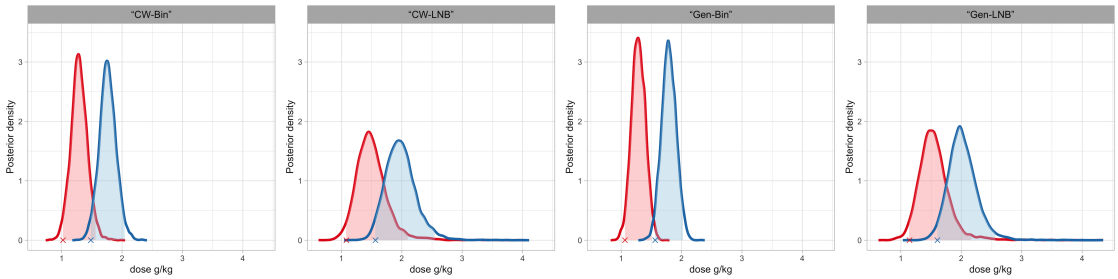
Figure 3.10: EG data. Box plot of the intracluster correlation posterior distributions at four observed toxin levels and for the new value of $x = 3.75$ g/kg. In each panel, estimates under "CW-Bin", "CW-LNB", "Gen-Bin" and "Gen-LNB" model are shown in red, blue, green, and purple, respectively.

is applied to the data, the effective dose ED_α is defined as the dose that induces an excess risk of α over control. As an example, for the embryolethality endpoint, the ED is found as the solution to the equation $D(\text{ED}_\alpha^D) - D(0)/(1 - D(0)) = \alpha$. With posterior samples of $D(x)$, we can numerically solve the equation, and obtain the posterior distribution of ED_α^D. Analogously, we obtain the posterior distribution of ED corresponds to the malformation and combined risk endpoints, using posterior realizations of $M(x)$ and $r(x)$, respectively. Then, BMD_α is defined as the left endpoint of the 95% credible interval of ED_α. Allen et al. (1994) found that BMD with $\alpha = 5\%$ is similar to the no observed adverse effect level (NOAEL). Additionally, agencies recommend reporting BMD at the level of 10% extra risk for dichotomous data (U.S. EPA, 2012). We focus on obtaining the posterior distribution of ED, and estimating BMD, at the three endpoints, for $\alpha = 5\%$ and $\alpha = 10\%$.

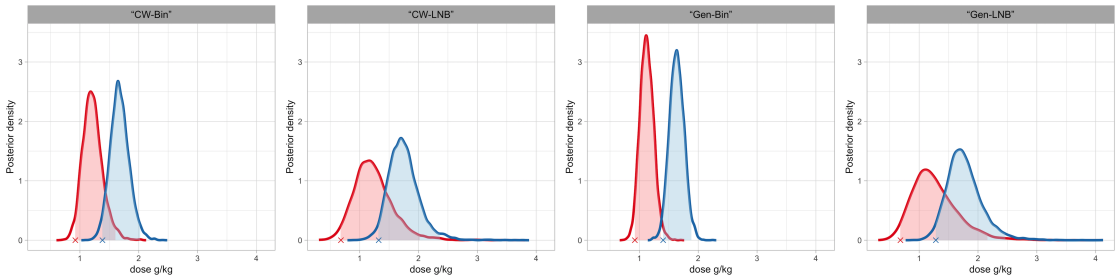
Figure 3.11 plots the posterior distribution of ED and the estimated BMD. We note that for embryolethality endpoint, the posterior samples of ED include extrapolation of toxin levels. The models with overdispersed kernel yield more dispersed distributions, while the models with the same kernel provide comparable results, despite the mixing structure. The estimated BMDs are summarized in Table



(a) Embryolethality endpoint.



(b) Malformation endpoint.



(c) Combined risk endpoint.

Figure 3.11: EG data. Posterior distribution of the effective dose with 5% BMR (in red) and 10% BMR (in blue). The shaded region indicates the 95% credible interval. The corresponding benchmark dose is marked with "×".

3.1. Results are generally robust across models. Therefore, it is manifested that the models themselves, in the absence of incorporating biochemical characteristics, are adequate for estimating BMD.

Figure 3.12 displays estimates for the probability mass functions corresponding to the number of non-viable fetuses given a specific number of implants $\Pr(R \mid m = 12, G_{\mathbf{x}})$, and the conditional probability mass functions of the number of malformations given a specified number of implants and the associated number

Table 3.1: EG data. BMD estimation under different models, based on posterior samples of ED.

Model	Embryolethality		Malformation		Combined risk	
	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 10\%$
"CW-Bin"	2.05	2.97	1.02	1.48	0.92	1.38
"CW-LNB"	1.62	2.79	1.08	1.56	0.68	1.32
"Gen-Bin"	2.00	3.03	1.06	1.56	0.92	1.40
"Gen-LNB"	1.64	2.77	1.14	1.60	0.68	1.28

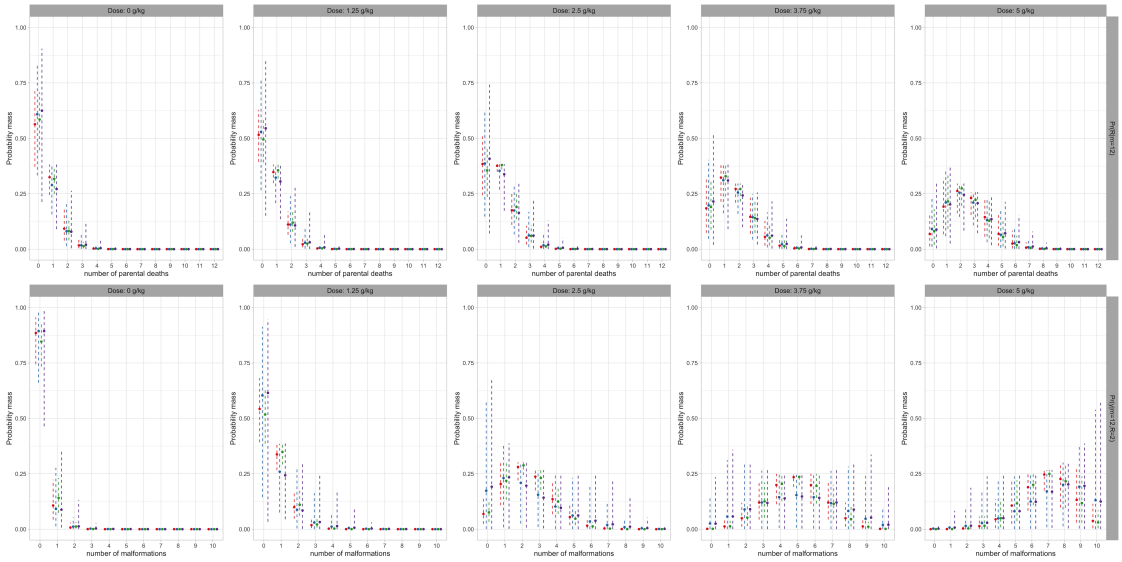
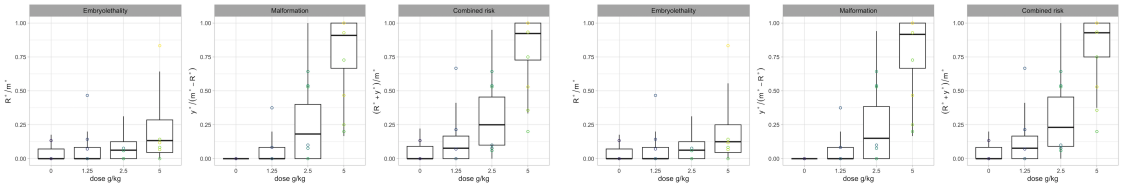


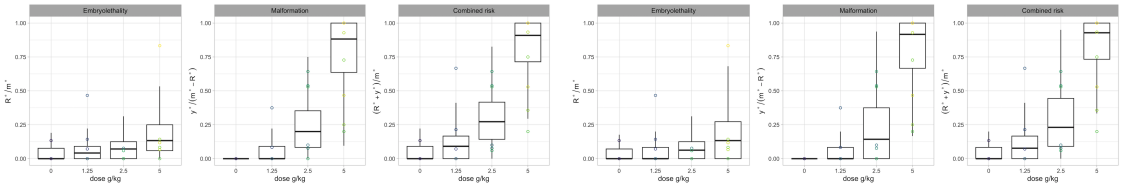
Figure 3.12: EG data. Posterior mean ("∘") and 95% uncertainty bands (dashed lines) for the probability mass $\Pr(R \mid m = 12, G_{\mathbf{x}})$ (top panels) and conditional probability mass $\Pr(y \mid m = 12, R = 2, G_{\mathbf{x}})$ (bottom panels), at four observed toxin levels and for the new value of $x = 3.75$ g/kg. In each panel, estimates under "CW-Bin", "CW-LNB", "Gen-Bin" and "Gen-LNB" model are shown in red, blue, green, and purple, respectively.

of non-viable fetuses $\Pr(y \mid m = 12, R = 2, G_{\mathbf{x}})$. All the models can uncover non-standard distributional shapes, especially for high toxin levels. Also noteworthy is the smooth evolution from right to left skewness in the conditional probability mass functions as the toxin level increases.

For the models considered in this section, we preform posterior predictive model checking based on cross-validation. Specifically, we use one randomly chosen



(a) "CW-Bin" model. **(b)** "CW-LNB" model.



(c) "Gen-Bin" model. **(d)** "Gen-LNB" model.

Figure 3.13: EG data. Box plots of posterior predictive samples for the embryolethality, malformation, and combined risk endpoints at the observed toxin levels. The corresponding observed proportions are denoted by "∘".

sample comprising data from 22 dams (approximately 20% of the data) spread roughly evenly across the dose levels as the test set, denoted by $\{(m'_{di}, R'_{di}, y'_{di}) : d = 1, \ldots, N, i = 1, \ldots, n'_d\}$. After fitting each model to the reduced data, we obtain, at each MCMC iteration, one set of posterior predictive sample at each observed dose level, denoted as m^*_d, R^*_d, and y^*_d. This is because the responses from the n_d dams at the d-th dose level share the same covariate $\mathbf{x}_d$. Based on the posterior predictive samples, Figure 3.13 displays box plots of the proportion of embryolethality R^*_d/m^*_d, malformation among live pups $y^*_d/(m^*_d - R^*_d)$, and combined negative outcomes $(R^*_d + y^*_d)/m^*_d$. The observed proportions from the test data points are marked by circles. None of these figures show evidence of ill-fitting.

We consider two types of model comparison, based on either the posterior predictive loss (PPL) criterion (Gelfand and Ghosh, 1998) or the interval score (IS) criterion (Gneiting and Raftery, 2007). The PPL criterion focuses on the first two moments of the predictive distribution, which may not be comprehensive

Table 3.2: EG data. Summary of comparison among the nonparametric models using the posterior predictive loss and interval score criteria. The values in bold correspond to the model favored by the particular criterion.

Endpoint	Criterion	"CW-Bin"	"CW-LNB"	"Gen-Bin"	"Gen-LNB"
Embryolethality	$G(\mathcal{M})$	0.72	0.72	**0.71**	0.72
	$P(\mathcal{M})$	0.56	0.53	**0.45**	0.58
	$S(\mathcal{M})$	20.73	**18.45**	20.46	18.73
Malformation	$G(\mathcal{M})$	1.34	1.39	**1.33**	1.36
	$P(\mathcal{M})$	1.18	1.10	**0.95**	1.17
	$S(\mathcal{M})$	16.07	14.97	16.81	**14.93**
Combined risk	$G(\mathcal{M})$	1.46	1.50	**1.43**	1.49
	$P(\mathcal{M})$	1.08	1.01	**0.89**	1.03
	$S(\mathcal{M})$	25.84	23.50	27.11	**20.91**

for risk assessment in developmental toxicity studies. As a complement, the IS criterion considers the quantiles of the predictive distribution, providing a more comprehensive perspective of the predictive distribution. We conduct model comparison based on these criteria applied to each of the endpoints. For instance, consider the embryolethality endpoint, we define the goodness-of-fit term as $G(\mathcal{M}) = \sum_{d=1}^{N}\sum_{i=1}^{n'_d}\{R'_{di}/m'_{di} - \mathrm{E}(R^*_d/m^*_d \mid \text{data})\}$, and the penalty term for model complexity as $P(\mathcal{M}) = \sum_{d=1}^{N} n'_d \mathrm{Var}(R^*_d/m^*_d \mid \text{data})$. The PPL criterion, comprised by these two terms, favors the model $\mathcal{M}$ that minimizes them. The IS criterion regarding the 95% credible interval is given by

$$S(\mathcal{M}) = \sum_{d=1}^{N}\sum_{i=1}^{n'_d}\{(u^e_d - l^e_d) + \frac{2}{\alpha}(l^e_d - \frac{R'_{di}}{m'_{di}})\mathbf{1}(\frac{R'_{di}}{m'_{di}} < l^e_d) + \frac{2}{\alpha}(\frac{R'_{di}}{m'_{di}} - u^e_d)\mathbf{1}(\frac{R'_{di}}{m'_{di}} > u^e_d)\},$$

where l^e_d and u^e_d denote the lower and upper limit of the posterior predictive 95% credible interval of the embryolethality endpoint at dose x_d, respectively, and $\alpha = 5\%$. The model with the smallest $S(\mathcal{M})$ is preferred. These terms are defined analogously for the other two endpoints, based on posterior predictive samples $y^*_d/(m^*_d - R^*_d)$ and $(R^*_d + y^*_d)/m^*_d$. We report the results in Table 3.2.

Based on the PPL criterion, the "Gen-Bin" model is preferred. The two models with overdispersed kernel have comparable goodness-of-fit, while as expected, have large penalty terms as well. Interestingly, the "CW-Bin" model also yields large penalty term. This is because the model imposes the most restricted structure, and thus tends to activate a larger number of effective components to capture the heterogeneity of the data. The IS criterion suggests an improvement from the use of mixture models with overdispersed kernel. In particular, for the malformation and combined risk endpoints, which exhibit vast variability, the "Gen-LNB" model that allows for the highest level of flexibility is preferred.

3.6 Summary and Remarks

We have explored a spectrum of modeling approaches that seek to introduce overdispersion through a mixing structure. We first illustrate that, the popular continuous mixture models fail in providing reliable uncertainty quantification for the dose-response curves. Contrarily, discrete mixture models, with mixing structure induced by dose-dependent stick-breaking process priors, offer a wealth of practical benefits. Notably, the enhanced flexibility offers rich inference for the response distributions and for the dose-response curves. Pursuing a more effective control of variability, we consider combining the two types of mixture models. Specifically, the general model is formulated with a continuous mixture model as the kernel in a discrete nonparametric mixing structure. We show that the derived models inherit the properties of their backbones, while ensuring efficient posterior inference. Data from a toxicity experiment involving an organic solvent were used to illustrate the discrete mixture models and to compare their performance with regard to a series of risk assessment tasks.

A crucial practical aspect entails selecting the appropriate model for a given

problem. The EG data example presented in Section 3.5 illuminates a plausible avenue. The "CW-Bin" model imposes restrictions in both the mixing kernel and the mixing weights, and may struggle with data with vast heterogeneity. If the risk assessment task only involves the first two moments of the predictive distribution, the "Gen-Bin" model may be more suitable. The key advantage of incorporating overdispersed kernel within a nonparametric mixture model lies in improved posterior predictive interval estimation. Among the two models with overdispersed kernel, the "Gen-LNB" model offers the most flexible structure for overdispersion, which is especially helpful if the data exhibit extensive variability. Overall, in order to obtain the best possible risk assessment in developmental toxicity studies, a comprehensive exploration of possible modeling options, as we conducted, is advocated by the regulatory agencies.

The modeling approaches examined in this chapter are directly applicable in other areas, which may involve more ordered categories and/or more covariates. For example, in pharmaceutical studies, participants are asked to report their responses to treatments in ordinal scale multiple times over a time period. The data comprises the frequency of each category, and may also include features of the participants. Additionally, if the time for each response is also available, models for longitudinal ordinal responses may be more appropriate, which is the focus of the next chapter.

Chapter 4

A Flexible Modeling Framework for Longitudinal Ordinal Responses

4.1 Introduction

Recent years have witnessed a rapid growth of longitudinal studies with binary and ordinal responses in several disciplines, including econometrics, and the health and social sciences. In such studies, of primary importance are the probability response curves, i.e., the probabilities of the response categories that evolve dynamically over time. This article aims at developing a hierarchical framework, customized to longitudinal settings, that allows flexible inference for the probability response curves. In addition, the defining characteristic of longitudinal data is that repeated measurements on the same subject induce dependence. Hence, a further objective is to flexibly model lead-lag correlations among repeated measurements.

The development of statistical methods for longitudinal binary and ordinal

data stems from models for longitudinal continuous responses, postulating the generalized linear model framework. Analogous to the continuous case, a specific model is formulated under one of three broad approaches pertaining to marginal models, conditional models, or subject-specific models. Marginal models provide alternative modeling options when likelihood-based approaches are difficult to implement. A conditional model describes the distribution of responses conditional on the covariates and also on part of the other components of the responses. In a subject-specific model, the effects of a subset of covariates are allowed to vary randomly from one individual to another. In the absence of predictor variables, functions of the observation time are usually used as covariates. We refer to Molenberghs and Verbeke (2006) for a comprehensive review. In Section 4.2.4, we elaborate on the connection of our proposed modeling approach with existing methods.

In this article, we introduce a novel viewpoint for longitudinal binary and ordinal data analysis. We begin with the model construction for longitudinal binary responses. The critical insight that distinguishes our methodology from the majority of the existing literature is functional data analysis. We treat the subjects' measurements as stochastic process realizations at the corresponding time points. The benefits are twofold. First, the models can incorporate unbalanced data from longitudinal studies in a unified scheme; directly inferring the stochastic process provides a well-defined probabilistic model for the missing values. Secondly, we can exploit the power of Bayesian hierarchical modeling for continuous functional data (e.g., Yang et al., 2016). To that end, we adopt the Binomial distribution with the logit link that connects binary responses to continuous signals, which, subject to additive measurement error, are then modeled as (conditionally) independent and identically distributed (i.i.d.) realizations from a Gaussian process (GP) with

random mean and covariance function. We place an Inverse-Wishart process (IWP) prior on the covariance function, and conditional on it, use a GP prior for the mean function. Therefore, the two essential ingredients in longitudinal modeling, the trend and the covariance structure, are modeled simultaneously and nonparametrically.

The hierarchical structure allows borrowing of strength across the subjects' trajectories. We apply a specific setting of hyperpriors for the GP and IWP priors, such that marginalizing over them, the latent continuous functions have a Student-t process (TP) prior. The TP enhances the flexibility of the GP (e.g., Shah et al., 2014). It retains attractive GP properties, such as analytic marginal and predictive distributions, and it yields predictive covariance that, unlike the GP, explicitly depends on the observed values. For inferential purposes, we represent the joint posterior distribution in multivariate form through evaluating the functions on the pooled grid, resulting in the common normal-inverse-Wishart conditional conjugacy. In conjunction with the Pólya-Gamma data augmentation technique (Polson et al., 2013), we develop a relatively simple and effective posterior simulation algorithm, circumventing the need for specialized techniques or tuning of Metropolis-Hastings steps.

To extend the model for ordinal responses, we utilize the continuation-ratio logits representation of the multinomial distribution. Such representation features an encoding of an ordinal response with C categories as a sequence of $C-1$ binary indicators, in which the j-th indicator signifies whether the ordinal response belongs to the j-th category or to one of the higher categories. We show that fitting a multinomial model for the ordinal responses is equivalent to fitting separately the aforementioned model on the binary indicators. Hence, we can conduct posterior simulation for each response category in a parallel fashion, leading to significant

computational efficiency gains in model implementation.

In modern longitudinal studies, it is common that the complete vector of repeated measurements is not collected on all subjects. As a specific example, in ecological momentary assessment (EMA) studies, emotions and behaviors are repeatedly measured for a cohort of participants, through wearable electronic devices (Ruwaard et al., 2018). For instance, in the *StudentLife* study (Wang et al., 2014), researchers monitored the students' mental status through pop-up questionnaires on their smartphones that prompted multiple times at pseudorandom intervals during the study period. Since the data collection process is based on the participants' conscious responding to prompted questions several times a day, non-response is inevitable. Missing values are typically considered to be a nuisance rather than a characteristic of EMA time series. Parametric and nonparametric Bayesian methods have been developed to handle longitudinal data with missingness; see Daniels and Xu (2020) for a review. The common issue is that one has to bear the drawbacks of making either structured or unstructured assumptions to manage missingness. The unstructured approach leads to flexibility, yet it may result in difficulties due to a large number of parameters relative to the sample size. Besides, the majority of the existing literature on longitudinal studies with missingness focuses on the scenario with continuous responses, and the extension to discrete responses is not trivial.

Accordingly, our contributions can be summarized as follows: (i) we model the mean and covariance jointly and nonparametrically, avoiding potential biases caused by a pre-specified model structure; (ii) we unify the toolbox for balanced and unbalanced longitudinal studies; (iii) the model encourages borrowing of strength, preserving systematic patterns that are common across all subject responses; (iv) we develop a computationally efficient posterior simulation method by taking

advantage of conditional conjugacy; (v) the model facilitates applications for ordinal responses with a moderate to large number of categories.

The rest of the chapter is organized as follows. Section 4.2 develops the methodology for binary responses, including model formulation, study of model properties, and the computational approach to inference and prediction. Section 4.3 illustrates the modeling approach through carefully designed simulation studies that reflect our main contributions and an EMA study that focuses on analyzing students' mental health through binary outcomes. The modeling extension for longitudinal ordinal responses is presented in Section 4.4, including an illustration involving an ordinal outcome from the same EMA study. Finally, Section 4.5 concludes with discussion.

4.2 The Modeling Approach for Binary Responses

Here, we develop the methodology for longitudinal binary responses. The data consist of repeated binary responses on n subjects, with the observation on subject i at time τ_{it} denoted by Y_{it}. The set of repeated outcomes for the i-th subject is collected into a T_i-dimensional vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT_i})^\top$. The hierarchical model construction is presented in Section 4.2.1. In Section 4.2.2, we discuss model properties related to our inference objectives. Bayesian inference and prediction is developed in Section 4.2.3. Finally, to place our contribution within the literature, we discuss in Section 4.2.4 the proposed model in the context of relevant Bayesian nonparametric approaches.

4.2.1 Model Specification

We examine the data from a functional data analysis perspective, treating each observed data vector $\mathbf{Y}_i$ as the evaluation of trajectory $Y_i(\tau)$ on grid $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iT_i})^\top$, for $i = 1, \ldots, n$. The n trajectories are assumed to be (conditionally) independent realizations from a continuous-time stochastic process. The prior probability model is built on the stochastic process. This approach avoids strong pre-determined assumptions on the transition mechanism within the sequence of subject-specific responses in $\mathbf{Y}_i$, while it is suitable to accommodate repeated measurements regardless of their observational pattern.

The functional data analysis view of longitudinal data dates back at least to Zhao et al. (2004), where it is suggested that functional data analysis tools, such as principal component analysis, can be used to capture periodic structure in longitudinal data. Indeed, Yao et al. (2005) study functional principal component analysis (FPCA) for sparse longitudinal data, a method that can provide effective recovery of the entire individual trajectories from fragmental data. FPCA has been applied in finance (Ingrassia and Costanzo, 2005), biomechanics (Donà et al., 2009), and demographic studies (Shamshoian et al., 2020). Its extension to examine sequences of discrete data is studied in Hall et al. (2008).

Our methodology builds from a GP-based hierarchical model for continuous functional data (Yang et al., 2016). Regarding mean-covariance estimation, the model in Yang et al. (2016) can be considered as a Bayesian counterpart of Yao et al. (2005). The hierarchical scheme enables a natural extension to studies with binary responses. We assume that, subject to measurement error, the i-th subject's responses, $Y_{it} \equiv Y_i(\tau_{it})$, depend on the i-th trajectory of the underlying process,

evaluated at times τ_{it}, through the following model

$$Y_i(\tau_{it}) \mid Z_i(\tau_{it}), \epsilon_{it} \overset{ind.}{\sim} Bin(1, \varphi(Z_i(\tau_{it}) + \epsilon_{it})), \quad t = 1, \ldots, T_i, \quad i = 1, \ldots, n,$$

where $\varphi(x) = \exp(x)/\{1 + \exp(x)\}$ denotes the expit function. The error terms are i.i.d. from a white noise process, that is, $\epsilon_{it} \mid \sigma_\epsilon^2 \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$, and independent of the process realizations $Z_i(\cdot)$. The main building block for the model construction is a hierarchical GP prior for the $Z_i(\cdot)$. In particular, given random mean function $\mu(\cdot)$ and covariance kernel $\Sigma(\cdot, \cdot)$, the $Z_i(\cdot)$ are i.i.d. GP realizations, denoted by $Z_i \mid \mu, \Sigma \overset{i.i.d.}{\sim} GP(\mu, \Sigma)$, for $i = 1, \ldots, n$. The hierarchical GP prior model is completed with nonparametric priors for the mean function and covariance kernel:

$$\mu \mid \Sigma \sim GP(\mu_0, \Sigma/\kappa), \quad \Sigma \sim IWP(\nu, \Psi_{\boldsymbol{\phi}}), \tag{4.1}$$

where $GP(\cdot, \cdot)$ and $IWP(\cdot, \cdot)$ denote the GP and IWP prior, respectively. The nonparametric prior reflects the intuition that parametric forms will generally not be sufficiently flexible for the mean and covariance functions.

We adopt an IWP prior for the covariance kernel, defined such that, on any finite grid $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)$ with $|\boldsymbol{\tau}|$ points, the projection $\Sigma(\boldsymbol{\tau}, \boldsymbol{\tau})$ follows an inverse-Wishart distribution with mean $\Psi_{\boldsymbol{\phi}}(\boldsymbol{\tau}, \boldsymbol{\tau})/(\nu - 2)$, denoted by $IW(\nu, \Psi_{\boldsymbol{\phi}}(\boldsymbol{\tau}, \boldsymbol{\tau}))$. Here, $\Psi_{\boldsymbol{\phi}}(\cdot, \cdot)$ is a non-negative definite function with parameters $\boldsymbol{\phi}$. Note that we use the parameterization from Dawid (1981) for the inverse-Wishart distribution, in particular, ν is the shape parameter and $\nu + |\boldsymbol{\tau}| - 1$ is the degrees of freedom parameter in the more common parameterization. Yang et al. (2016) validate that this parameterization defines an infinite dimensional probability measure whose finite dimensional projection on grid $\boldsymbol{\tau}$ coincides with the inverse-Wishart distribution $IW(\nu, \Psi_{\boldsymbol{\phi}}(\boldsymbol{\tau}, \boldsymbol{\tau}))$.

The model formulation is completed with prior specification for the hyperparameters. The error variance is assigned an inverse Gamma prior, $\sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon)$. We focus primarily on stationary specifications under the prior structure in (4.1). In particular, we work with mean function, $\mu_0(\tau) \equiv \mu_0$, and isotropic covariance function, $\Psi_{\boldsymbol{\phi}}$, within the Matérn class, a widely used class of covariance functions (Rasmussen and Williams, 2006). In general, the Matérn covariance function is specified by a scale parameter σ^2, a range parameter ρ, and a smoothness parameter ι. To encourage smoothness in the probability response curves, we set $\iota = 5/2$, such that the covariance kernel is given by

$$\Psi_{\boldsymbol{\phi}}(\tau, \tau') \;=\; \sigma^2 \left(1 + \frac{\sqrt{5}|\tau - \tau'|}{\rho} + \frac{5|\tau - \tau'|^2}{3\rho^2}\right) \exp\left(-\frac{\sqrt{5}|\tau - \tau'|}{\rho}\right),$$

where $\boldsymbol{\phi} = \{\sigma^2, \rho\}$. For hyperparameters μ_0, σ^2, ρ, we take the commonly used choice,

$$\mu_0 \sim N(a_\mu, b_\mu), \quad \sigma^2 \sim \text{Gamma}(a_\sigma, b_\sigma), \quad \rho \sim Unif(a_\rho, b_\rho).$$

Finally, we set $\kappa = (\nu - 3)^{-1}$, such that the continuous-time process for the $Z_i(\cdot)$ is a TP when μ and Σ are marginalized out (see Section 4.2.2 for details). As a consequence, parameter ν controls the tail heaviness of the marginal process, with smaller values of ν corresponding to heavier tails. We place a uniform prior on ν, $\nu \sim Unif(a_\nu, b_\nu)$, with $a_\nu > 3$ to ensure positive definiteness of Σ/κ.

As discussed in Diggle (1988), the correlation of repeated measurements on the same subject commonly has the following patterns. First, it should decrease with respect to the measurements' separation in time, while remaining positive to indicate the measurements are from the same subject. This feature is encapsulated by the form of the covariance kernel $\Psi_{\boldsymbol{\phi}}$. The IWP prior elicits realizations for which

this property holds a priori, while enabling a flexible estimate of the covariance structure with information from the data a posteriori. Second, measurements that are made arbitrarily close in time are subject to imperfect correlation, possibly caused by subsampling of each subject. This feature is represented by the error term in our model. Moreover, the motivation for adding the error term arises from the fact that measurement error is introduced in the estimation of a continuous-time function based on data collected at discrete time points.

Although the probability model is formulated through stochastic process realizations, posterior simulation is based on the corresponding finite dimensional distributions (f.d.d.s.). Consequently, to write the model for the data, we need to represent the likelihood and prior in multivariate forms through evaluating the functions on finite grids. Denoting $Y_i(\boldsymbol{\tau}_i)$ by $\mathbf{Y}_i$, $Z_i(\boldsymbol{\tau}_i)$ by $\mathbf{Z}_i$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT_i})^\top$, the model for the data can be written as

$$\begin{aligned} &\mathbf{Y}_i \mid \mathbf{Z}_i, \boldsymbol{\epsilon}_i \overset{ind.}{\sim} \prod_{t=1}^{T_i} Bin(1, \varphi(Z_{it} + \epsilon_{it})), \quad i = 1, \ldots, n, \\ &\mathbf{Z}_i \mid \mu(\boldsymbol{\tau}_i), \Sigma(\boldsymbol{\tau}_i, \boldsymbol{\tau}_i) \overset{ind.}{\sim} N(\mu(\boldsymbol{\tau}_i), \Sigma(\boldsymbol{\tau}_i, \boldsymbol{\tau}_i)), \quad \boldsymbol{\epsilon}_i \mid \sigma_\epsilon^2 \overset{ind.}{\sim} N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}). \end{aligned} \tag{4.2}$$

Notice that the grids $\{\boldsymbol{\tau}_i : i = 1, \ldots, n\}$ are not necessarily the same for all subjects. Therefore, the shared GP and IWP prior in (4.1) need to be evaluated on the pooled grid $\boldsymbol{\tau} = \cup_{i=1}^n \boldsymbol{\tau}_i$. If $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Psi}_{\boldsymbol{\phi}}$ denote $\mu(\boldsymbol{\tau})$, $\Sigma(\boldsymbol{\tau}, \boldsymbol{\tau})$, and $\Psi_{\boldsymbol{\phi}}(\boldsymbol{\tau}, \boldsymbol{\tau})$, respectively, then

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \mu_0, \nu \sim N(\mu_0 \mathbf{1}, (\nu - 3)\boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} \mid \nu, \boldsymbol{\phi} \sim IW(\nu, \boldsymbol{\Psi}_{\boldsymbol{\phi}}). \tag{4.3}$$

The hierarchical model formulation for the data in (4.2) and (4.3) forms the basis for the posterior simulation algorithm, which is discussed in detail in Section 4.2.3.

4.2.2 Model Properties

To fix ideas for the following discussion, we refer to $Z_i(\tau)$ as the signal process of the binary process $Y_i(\tau)$, and to $\mathcal{Z}_i(\tau) = Z_i(\tau) + \epsilon_i(\tau)$ as the latent process of $Y_i(\tau)$. Since the stochastic process is characterized by its f.d.d.s., we shall investigate the random vectors $\mathbf{Y}_{\boldsymbol{\tau}} = Y_i(\boldsymbol{\tau})$, $\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}} = \mathcal{Z}_i(\boldsymbol{\tau})$, and $\mathbf{Z}_{\boldsymbol{\tau}} = Z_i(\boldsymbol{\tau})$, for a generic grid vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)^\top$. We surpass the subject index i because the subject trajectories are identically distributed. Appendix A.3 includes proofs for the propositions included in this section.

Among the various inference goals in a study that involves longitudinal binary data, estimating the probability response curve and the covariance structure of the repeated measurements are the most important ones. In Proposition 4.1, we derive the probability response curves and covariance matrix of the binary vector $\mathbf{Y}_{\boldsymbol{\tau}}$, conditional on the signal vector $\mathbf{Z}_{\boldsymbol{\tau}}$ and error variance σ^2_ϵ. The probability response curve can be defined generically as $\mathbf{P}_{\mathbf{y}\boldsymbol{\tau}} = (\Pr(Y_{\tau_1} = y_{\tau_1} \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon), \ldots, \Pr(Y_{\tau_T} = y_{\tau_T} \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon))^\top$, where y_{τ_t} is either 0 or 1. Without loss of generality, we focus on $\mathbf{P}_{\mathbf{1}\boldsymbol{\tau}}$.

Proposition 4.1. *The probability response curve is given by* $\mathbf{P}_{\mathbf{1}\boldsymbol{\tau}} = E(\boldsymbol{\pi}(\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon)$*, where* $\boldsymbol{\pi}(\mathbf{x})$ *denotes the vector operator that applies the expit function to every entry of* $\mathbf{x}$*. Regarding the covariance matrix, for* $\tau \in \boldsymbol{\tau}$*,* $Var(Y_\tau \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon) = E(\varphi(\mathcal{Z}_\tau) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon) - E^2(\varphi(\mathcal{Z}_\tau) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon)$*, and for* $\tau, \tau' \in \boldsymbol{\tau}$*, with* $\tau' \neq \tau$*,* $Cov(Y_\tau, Y_{\tau'} \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon) = Cov(\varphi(\mathcal{Z}_\tau), \varphi(\mathcal{Z}_{\tau'}) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon)$*. The conditional expectations in all of the above expressions are with respect to distribution,* $\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}} \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon \sim N(\mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon \mathbf{I})$.

The practical utility of Proposition 4.1 lies on performing posterior inference for the probability response curve and the covariance structure of the binary process, conditioning on the signal process and the noise. With posterior samples of $\mathbf{Z}_{\boldsymbol{\tau}}$ and σ^2_ϵ, we can simulate $\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}$ from $N(\mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon \mathbf{I})$ and numerically compute the

corresponding moments in Proposition 4.1. The entries of $\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}$ are independent, given $\mathbf{Z}_{\boldsymbol{\tau}}$, and thus simulating $\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}$ is not computationally demanding, even when $|\boldsymbol{\tau}|$ is large.

We next establish a closer connection between the binary process and the signal process. Proposition 4.2 reveals that the evolution of the binary process over time can be (approximately) expressed as a function of the expectation of the signal process and the total variance. Moreover, the covariance of the binary process is approximately the covariance of the signal process scaled by a factor related to the expectation of the signal.

Proposition 4.2. *Consider the proposed model as described in (4.2) and denote* $\mu(\boldsymbol{\tau}) = \boldsymbol{\mu}$, *and* $\Sigma(\boldsymbol{\tau}, \boldsymbol{\tau}) = \boldsymbol{\Sigma}$. *Then,* $\forall \tau, \tau' \in \boldsymbol{\tau}$,

$$Pr(Y_\tau = 1 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\epsilon^2) \approx \varphi(E(Z_\tau \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})) + \frac{Var(Z_\tau \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sigma_\epsilon^2}{2} \varphi''(E(Z_\tau \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})),$$

$$Cov(Y_\tau, Y_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\epsilon^2) \approx \varphi'(E(Z_\tau \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}))\varphi'(E(Z_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}))\, Cov(Z_\tau, Z_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$- \frac{1}{4}[\, Var(Z_\tau \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sigma_\epsilon^2][\, Var(Z_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sigma_\epsilon^2]\varphi''(E(Z_\tau \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}))\varphi''(E(Z_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})).$$

Here, $\varphi'(x) = \frac{d\varphi(x)}{dx} = \varphi(x)[1-\varphi(x)]$ *and* $\varphi''(x) = \frac{d^2\varphi(x)}{dx^2} = \varphi(x)[1-\varphi(x)][1-2\varphi(x)]$.

Our inference results are based on exact expressions, such as the ones in Proposition 4.1. Nonetheless, the approximate expressions derived in Proposition 4.2 are practically useful to gain more insight on properties of the binary process, as well as for prior specification. Note that exploring properties of the binary process is not trivial due to the lack of general analytical forms for moments of logit-normal distributions. Hence, a connection with properties of the signal process is useful. For instance, if we specify the covariance for the signal process to decrease as a function of separation in time, an analogous structure will hold (approximately) for the binary process.

The previous discussion focuses on studying the f.d.d.s of the binary process given the signal process. Therefore, it is important to investigate the marginal f.d.d.s of the signal process. We show that, under the specification $\kappa = (\nu - 3)^{-1}$, the f.d.d.s. of the signal process correspond to a multivariate Student-t (MVT) distribution, and thus the signal process is a TP. We first state the definition of the MVT distribution and the TP (see, e.g., Shah et al., 2014). Notice that we use the covariance matrix as a parameter for the MVT distribution, instead of the more common parameterization based on a scale matrix.

Definition 4.1. *The random vector* $\mathbf{Z} \in \mathbb{R}^n$ *is MVT distributed, denoted* $\mathbf{Z} \sim MVT(\nu, \boldsymbol{\mu}, \boldsymbol{\Psi})$*, if it has density*

$$\frac{\Gamma(\frac{\nu+n}{2})}{[(\nu-2)\pi]^{n/2}\Gamma(\frac{\nu}{2})}|\boldsymbol{\Psi}|^{-1/2}\left(1+\frac{(\mathbf{Z}-\boldsymbol{\mu})^T\boldsymbol{\Psi}^{-1}(\mathbf{Z}-\boldsymbol{\mu})}{\nu-2}\right)^{-\frac{\nu+n}{2}}$$

where $\nu > 2$ *is the degrees of freedom parameter,* $\boldsymbol{\mu} \in \mathbb{R}^n$*, and* $\boldsymbol{\Psi}$ *is an* $n \times n$ *symmetric, positive definite matrix. Under this parameterization,* $E(\mathbf{Z}) = \boldsymbol{\mu}$ *and* $Cov(\mathbf{Z}) = \boldsymbol{\Psi}$*.*

Consider a process $Z(\tau)$ *formulated through mean function* $\mu(\tau)$*, a non-negative kernel function* $\Psi(\tau, \tau)$*, and parameter* $\nu > 2$*, such that its f.d.d.s correspond to the MVT distribution with mean vector and covariance matrix induced by* $\mu(\tau)$ *and* $\Psi(\tau, \tau)$*, respectively. Then,* $Z(\tau)$ *follows a TP, denoted by* $Z(\tau) \sim TP(\nu, \mu(\tau), \Psi(\tau, \tau))$*.*

Marginalizing over $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (4.2) and (4.3), the implied distribution for $\mathbf{Z}_{\boldsymbol{\tau}}$ is MVT, with degrees of freedom parameter ν (with $\nu > 3$ in our context), mean vector $\mu_0\mathbf{1}$, and covariance matrix $\boldsymbol{\Psi}_{\boldsymbol{\phi}} = \Psi_{\boldsymbol{\phi}}(\boldsymbol{\tau}, \boldsymbol{\tau})$. We thus obtain the following result for the signal process.

Proposition 4.3. *Under the model formulation in (4.2) and (4.3), the signal process follows marginally a TP, that is,* $Z \sim TP(\nu, \mu_0, \Psi_\phi)$.

Proposition 4.3 is beneficial in terms of both computation and interpretation. Without a constraint on κ, as in Yang et al. (2016), the marginal distribution of $\mathbf{Z}_{\boldsymbol{\tau}}$ does not have analytical form. Hence, for prediction at new time points, one has to sample from an IWP and a GP, which is computationally intensive, especially for a dense grid. In contrast, we can utilize the analytical form of the TP predictive distribution to develop a predictive inference scheme that resembles that of GP-based models (see Section 4.2.3). Moreover, the result highlights the model property that the degrees of freedom parameter ν controls how heavy tailed the process is. Smaller values of ν correspond to heavier tails. As ν gets larger, the tails resemble Gaussian tails. Additionally, ν controls the dependence between Z_τ and $Z_{\tau'}$, which are jointly MVT distributed, with smaller values indicating higher dependence. Such interpretation of parameter ν facilitates the choice of its hyperprior.

The local behavior of stochastic process realizations is crucial for interpolation. Under the longitudinal setting, continuous, or perhaps differentiable, signal process trajectories are typically anticipated. Evidently, the observed data can not visually inform the smoothness of signal process realizations. Rather, such smoothness should be captured in the prior specification that incorporates information about the data generating mechanism. For weakly stationary processes, mean square continuity is equivalent to the covariance function being continuous at the origin (Stein, 1999). And, the process is ι-times mean square differentiable if and only if the 2ι-times derivative of the covariance function at the origin exists and is finite. Under our model, the signal process follows a TP marginally. Its covariance structure is specified by the Matérn covariance function with smoothness parameter

ι. Referring to the behavior of the Matérn class of covariance functions at the origin, we obtain the following result for the mean square continuity and differentiability of the signal process.

Proposition 4.4. *Consider the proposed model with marginal signal process* $Z \sim TP(\nu, \mu_0, \Psi_\phi)$*, where* Ψ_ϕ *belongs to the Matérn family of covariance functions with smoothness parameter* ι*. Then, the signal process is mean square continuous and* $\lfloor \iota \rfloor$*-times mean square differentiable.*

The results in this section study several properties that are useful in model implementation. Indeed, the practical utility of such model properties with respect to prior specification and posterior inference is discussed in the next section.

4.2.3 Prior Specification and Posterior Inference

The model described in Section 4.2.1 contains parameters $\{\sigma_\epsilon^2, \mu_0, \sigma^2, \rho, \nu\}$ whose prior hyperparameters need to be specified. We develop a default specification strategy that relies on the model properties explored in Section 4.2.2.

First, we set the prior for μ_0 such that the prior expected probability response curve does not favor any category, and the corresponding prior uncertainty bands span a significant portion of the unit interval. For instance, this can be achieved with prior $\mu_0 \sim N(0, 100)$ which yields prior expected probability of positive response of about $1/2$ across τ. In general, we would not expect to have available prior information about the variance and correlation structure of the unobserved signal process, which are controlled by parameters σ^2 and ρ. However, Proposition 4.2 suggests an approximate relationship between the covariance structure of the binary process and the signal process, and we can thus specify the corresponding priors similarly to GP-based models. In particular, we select the uniform prior for the range parameter ρ such that the correlation between Z_τ and $Z_{\tau'}$ decreases to

0.05 when the difference between τ and τ' is within a pre-specified subset of the observation time window. For instance, for the data analysis in Section 4.3.2 where the total observation window comprises 72 days, we used a $Unif(3, 12)$ prior for ρ, which implies that the aforementioned correlation decreases to 0.05 when the time difference ranges from 7 to 31 days. The hyperprior for ν is $Unif(a_\nu, b_\nu)$. We specify $a_\nu > 3$ to reflect the constraint for $\Sigma/(\nu-3)$ to be a well-defined covariance matrix, and b_ν large enough such that the tail behavior of the marginal TP is hard to distinguish from that of a GP. For instance, a default choice is $a_\nu = 4$ and $b_\nu = 30$.

We follow Fong et al. (2010) to specify the prior for $\sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon)$. Integrating out σ_ϵ^2, the measurement error ϵ is marginally distributed as a univariate Student-t distribution with location parameter 0, scale parameter b_ϵ/a_ϵ, and degrees of freedom parameter $2a_\epsilon$. For a predetermined measurement error range $(-R, R)$ with degree of freedom υ, we can use the relationship $\pm t^{\upsilon}_{1-(1-q)/2}\sqrt{b_\epsilon/a_\epsilon} = \pm R$ to obtain $a_\epsilon = \upsilon/2$ and $b_\epsilon = R^2\upsilon/[2(t^{\upsilon}_{1-(1-q)/2})^2]$, where t^{υ}_q is the q-th percentile of a Student-t distribution with υ degrees of freedom.

Proceeding to posterior inference, we develop an MCMC algorithm based on (4.2) and (4.3). We introduce layers of latent variables, beginning with $\xi_{it} \sim PG(1, 0)$ for every observation Y_{it}, where $PG(a, b)$ denotes the Pólya-Gamma distribution with shape parameter a and tilting parameter b (Polson et al., 2013). Denote the collection of Pólya-Gamma variables for each subject by $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iT_i})^\top$. Also, introduce $\mathcal{Z}_{it} = Z_{it} + \epsilon_{it}$, and let $\boldsymbol{\mathcal{Z}}_i = (\mathcal{Z}_{i1}, \ldots, \mathcal{Z}_{iT_i})^\top$. Recall that $\boldsymbol{\tau} = \cup_{i=1}^n \boldsymbol{\tau}_i$ is the pooled grid. Denote the evaluations on the pooled grid by $\tilde{\mathbf{Z}}_i = Z_i(\boldsymbol{\tau})$ and let $\mathbf{Z}_i^* = \tilde{\mathbf{Z}}_i \setminus \mathbf{Z}_i$. That is, $\mathbf{Z}_i^* = Z_i(\boldsymbol{\tau}_i^*)$, where $\boldsymbol{\tau}_i^* = \boldsymbol{\tau} \setminus \boldsymbol{\tau}_i$ is the set of grid points at which the i-th trajectory misses observations. Then,

the model for the data $\{Y_{it} : t = 1, \ldots, T_i, i = 1, \ldots, n\}$ can be expressed as

$$
\begin{aligned}
&Y_{it} \mid \mathcal{Z}_{it}, \xi_{it} \overset{ind.}{\sim} \mathcal{B}(\mathcal{Z}_{it}, \xi_{it}),\ \xi_{it} \overset{i.i.d.}{\sim} PG(1, 0),\ t = 1, \ldots, T_i, \\
&\boldsymbol{\mathcal{Z}}_i \mid \mathbf{Z}_i, \sigma_\epsilon^2 \overset{ind.}{\sim} N(\mathbf{Z}_i, \sigma_\epsilon^2 \mathbf{I}_{T_i}),\ \tilde{\mathbf{Z}}_i = (\mathbf{Z}_i, \mathbf{Z}_i^*)^\top \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ i = 1, \ldots, n, \\
&\sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon),\ \boldsymbol{\mu} \mid \mu_0, \boldsymbol{\Sigma}, \nu \sim N(\mu_0 \mathbf{1}, (\nu - 3)\boldsymbol{\Sigma}),\ \mu_0 \sim N(a_\mu, b_\mu), \\
&\boldsymbol{\Sigma} \mid \nu, \boldsymbol{\Psi}_\phi \sim IW(\nu, \boldsymbol{\Psi}_\phi),\ \boldsymbol{\Psi}_\phi = \Psi_\phi(\boldsymbol{\tau}, \boldsymbol{\tau}),\ \boldsymbol{\phi} = \{\sigma^2, \rho\}, \\
&\sigma^2 \sim \text{Gamma}(a_\sigma, b_\sigma),\ \rho \sim Unif(a_\rho, b_\rho),\ \nu \sim Unif(a_\nu, b_\nu).
\end{aligned}
$$

Here, $\mathcal{B}(\mathcal{Z}_{it}, \xi_{it}) \propto \exp\{(Y_{it} - 0.5)\,\mathcal{Z}_{it} - 0.5\,\xi_{it}\mathcal{Z}_{it}^2\}$ denotes the probability mass function of Y_{it} conditional on both sets of latent variables, $\mathcal{Z}_{it}$ and ξ_{it}. Hence, the joint posterior density of all model parameters can be written as

$$
\begin{aligned}
&p(\{\boldsymbol{\mathcal{Z}}_i\}_{i=1}^n, \{\boldsymbol{\xi}_i\}_{i=1}^n, \{\tilde{\mathbf{Z}}_i\}_{i=1}^n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\epsilon^2, \mu_0, \sigma^2, \rho, \nu \mid \{\mathbf{Y}_i\}_{i=1}^n) \\
&\quad \propto \prod_{i=1}^n \{p(\mathbf{Y}_i \mid \boldsymbol{\mathcal{Z}}_i, \boldsymbol{\xi}_i) p(\boldsymbol{\xi}_i) p(\boldsymbol{\mathcal{Z}}_i \mid \mathbf{Z}_i, \sigma_\epsilon^2) p(\mathbf{Z}_i^* \mid \mathbf{Z}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\mathbf{Z}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\} \qquad (4.4) \\
&\quad \times p(\boldsymbol{\mu} \mid \mu_0, \boldsymbol{\Sigma}, \nu) p(\boldsymbol{\Sigma} \mid \sigma^2, \rho, \nu) p(\sigma_\epsilon^2) p(\mu_0) p(\sigma^2) p(\rho) p(\nu).
\end{aligned}
$$

The introduction of the latent variables enables a Gibbs sampling scheme with conditionally conjugate updates. Denote generically by $p(\boldsymbol{\theta} \mid -)$ the posterior full conditional for parameter $\boldsymbol{\theta}$. Notice that $p(\boldsymbol{\mathcal{Z}}_i, \boldsymbol{\xi}_i \mid -) \propto p(\mathbf{Y}_i \mid \boldsymbol{\mathcal{Z}}_i, \boldsymbol{\xi}_i) p(\boldsymbol{\xi}_i) p(\boldsymbol{\mathcal{Z}}_i \mid \mathbf{Z}_i, \sigma_\epsilon^2)$, which matches the Bayesian logistic regression structure in Polson et al. (2013). Therefore, $p(\boldsymbol{\mathcal{Z}}_i \mid -)$ and $p(\boldsymbol{\xi}_i \mid -)$ can be sampled directly. Factorizing the prior of $\tilde{\mathbf{Z}}_i$ as $p(\tilde{\mathbf{Z}}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\mathbf{Z}_i^* \mid \mathbf{Z}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\mathbf{Z}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, results in $p(\mathbf{Z}_i^*, \mathbf{Z}_i | -) \propto p(\mathbf{Z}_i^* \mid \mathbf{Z}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\mathbf{Z}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mathcal{Z}}_i \mid \mathbf{Z}_i, \sigma_\epsilon^2)$. This forms yields ready updates for $\mathbf{Z}_i^*$ and $\mathbf{Z}_i$ using GP-based predictive sampling. All other model parameters can be sampled using standard updates. The details of the MCMC algorithm are given in Appendix B.3.

We have linked the probability response curve and covariance structure of the

binary process $Y_i(\tau)$ to the corresponding signal process $Z_i(\tau)$. To estimate the signal process, we obtain posterior samples for $\mathbf{Z}_i^+ = Z_i(\boldsymbol{\tau}^+)$, where $\boldsymbol{\tau}^+ \supset \boldsymbol{\tau}$ is a finer grid than the pooled grid. Denote $\check{\boldsymbol{\tau}} = \boldsymbol{\tau}^+ \setminus \boldsymbol{\tau}$ as the time points where none of the subjects have observations, and let $\check{\mathbf{Z}}_i = Z_i(\check{\boldsymbol{\tau}})$. Using the marginal TP result from Proposition 4.3,

$$\begin{pmatrix} \tilde{\mathbf{Z}}_i \\ \check{\mathbf{Z}}_i \end{pmatrix} \sim MVT\left(\nu, \begin{pmatrix} \boldsymbol{\mu}_{0\boldsymbol{\tau}} \\ \boldsymbol{\mu}_{0\check{\boldsymbol{\tau}}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Psi}_{\boldsymbol{\tau},\boldsymbol{\tau}} & \boldsymbol{\Psi}_{\boldsymbol{\tau},\check{\boldsymbol{\tau}}} \\ \boldsymbol{\Psi}_{\check{\boldsymbol{\tau}},\boldsymbol{\tau}} & \boldsymbol{\Psi}_{\check{\boldsymbol{\tau}},\check{\boldsymbol{\tau}}} \end{pmatrix}\right),$$

where $\boldsymbol{\mu}_{0\cdot} = \mu_0 \mathbf{1}_{|\cdot|}$, and $\boldsymbol{\Psi}_{\cdot,\cdot}$ denotes the covariance function evaluation $\Psi_{\boldsymbol{\phi}}(\cdot,\cdot)$. Next, based on the conditionals of the MVT distribution (Shah et al., 2014),

$$\check{\mathbf{Z}}_i \mid \tilde{\mathbf{Z}}_i \sim MVT\left(\nu + |\boldsymbol{\tau}|, \check{\boldsymbol{\mu}}_{i\check{\boldsymbol{\tau}}}, \frac{\nu + S_{i\boldsymbol{\tau}} - 2}{\nu + |\boldsymbol{\tau}| - 2} \check{\boldsymbol{\Psi}}_{\check{\boldsymbol{\tau}},\check{\boldsymbol{\tau}}}\right), \tag{4.5}$$

with $\check{\boldsymbol{\mu}}_{i\check{\boldsymbol{\tau}}} = \boldsymbol{\Psi}_{\check{\boldsymbol{\tau}},\boldsymbol{\tau}} \boldsymbol{\Psi}_{\boldsymbol{\tau},\boldsymbol{\tau}}^{-1} (\tilde{\mathbf{Z}}_i - \boldsymbol{\mu}_{0\boldsymbol{\tau}}) + \boldsymbol{\mu}_{0\check{\boldsymbol{\tau}}}$, $S_{i\boldsymbol{\tau}} = (\tilde{\mathbf{Z}}_i - \boldsymbol{\mu}_{0\boldsymbol{\tau}})^\top \boldsymbol{\Psi}_{\boldsymbol{\tau},\boldsymbol{\tau}}^{-1} (\tilde{\mathbf{Z}}_i - \boldsymbol{\mu}_{0\boldsymbol{\tau}})$ and $\check{\boldsymbol{\Psi}}_{\check{\boldsymbol{\tau}},\check{\boldsymbol{\tau}}} = \boldsymbol{\Psi}_{\check{\boldsymbol{\tau}},\check{\boldsymbol{\tau}}} - \boldsymbol{\Psi}_{\check{\boldsymbol{\tau}},\boldsymbol{\tau}} \boldsymbol{\Psi}_{\boldsymbol{\tau},\boldsymbol{\tau}}^{-1} \boldsymbol{\Psi}_{\boldsymbol{\tau},\check{\boldsymbol{\tau}}}$. Using (4.5), given each posterior sample for $\tilde{\mathbf{Z}}_i$, μ_0, $\boldsymbol{\phi}$ and ν, we can complete the posterior realization for the signal process over the finer grid. As discussed in Section 4.2.2, we can then obtain full posterior inference for functionals of the binary process.

The predictive distribution of the signal process also illustrates the information borrowed across subjects. For the i-th subject, the grid, $\boldsymbol{\tau}^+$, where predictions are made can be partitioned as $\boldsymbol{\tau}_i \cup \boldsymbol{\tau}_i^* \cup \check{\boldsymbol{\tau}}$, where $\boldsymbol{\tau}_i^* = \boldsymbol{\tau} \setminus \boldsymbol{\tau}_i$ represents the grid points where subject i does not have observations, while at least one of the other subjects have observations. Then, we first predict $Z_i(\boldsymbol{\tau}_i^*)$ conditioning on $Z_i(\boldsymbol{\tau}_i)$ by the GP predictive distribution, and next predict $Z_i(\check{\boldsymbol{\tau}})$ conditioning on $Z_i(\boldsymbol{\tau}_i)$ and $Z_i(\boldsymbol{\tau}_i^*)$ by the TP predictive distribution. Comparing with the GP, (4.5) suggests the TP is scaling the predictive covariance by the factor $\frac{\nu + S_{i\boldsymbol{\tau}} - 2}{\nu + |\boldsymbol{\tau}| - 2}$. Note that $S_{i\boldsymbol{\tau}}$ is distributed as the sum of squares of $|\boldsymbol{\tau}|$ independent $MVT_1(\nu, 0, 1)$ random

variables and hence $\mathrm{E}(S_{i\boldsymbol{\tau}}) = |\boldsymbol{\tau}|$. Accordingly, if we have made good interpolation prediction, the predictive covariance for extrapolation of is expected to scale down and vice versa. Comparing with predicting both $Z_i(\boldsymbol{\tau}_i^*)$ and $Z_i(\check{\boldsymbol{\tau}})$ conditioning on $Z_i(\boldsymbol{\tau}_i)$ through the GP predictive distribution, our model allows using information across subjects to adjust the individual trajectory's credible interval.

Another crucial benefit of modeling the signal process as a TP emerges when we consider making predictions at $\check{\boldsymbol{\tau}}$, the grid points where none of the subjects have observations. Under the hierarchical GP prior in Yang et al. (2016), for which the marginal is not generally a TP, such predictions would require the conditional distribution $\boldsymbol{\Sigma}_{\check{\boldsymbol{\tau}},\check{\boldsymbol{\tau}}} \mid \boldsymbol{\Sigma}_{\boldsymbol{\tau},\boldsymbol{\tau}}$ from their joint inverse-Wishart distribution, which is not analytically available. We circumvent this issue by marginalizing out μ and Σ. The predictions are then based on the conditional $\check{\mathbf{Z}}_i \mid \tilde{\mathbf{Z}}_i$ from their joint multivariate t distribution, which is the MVT distribution in (4.5). Hence, for prediction on a grid denser than the pooled grid $\boldsymbol{\tau}$, the marginal TP specification for the signal process is a practically important model feature.

4.2.4 Connections with Existing Literature

Our methodology is broadly related with certain Bayesian nonparametric methods. The proposed model is related to a particular class of conditional models, known as transition models, which induce the aging effect by allowing past values to explicitly affect the present observation, usually through autoregressive dynamics. Di Lucca et al. (2013) studied a class of non-Gaussian autoregression models for continuous responses, which can be extended to handle binary longitudinal outcomes by treating them as a discretized version of the continuous outcomes. DeYoreo and Kottas (2018b) developed a nonparametric density regression model for ordinal regression relationships that evolve in discrete time. Compared with

the proposed methodology, these models are more flexible in terms of the binary response distribution. However, it is demanding to handle higher than first-order dynamics, and there is no natural way to treat missing data under a discrete time autoregressive framework, hindering applications for unbalanced longitudinal studies.

The proposed model is more closely related to subject-specific models, where the responses are assumed to be independent conditioning on subject-specific effects. The main approach has been to construct models for longitudinal binary responses building from the various Bayesian nonparametric models for longitudinal continuous data, developed under the mixed effects framework (e.g., Li et al., 2010; Ghosh and Hanson, 2010; Quintana et al., 2016). For instance, embedding a Dirichlet process mixture of normals prior as the probability model for the latent variables, Jara et al. (2007) and Tang and Duan (2012) consider binary responses, and Kunihama et al. (2019) handle mixed-scale data comprising continuous and binary responses. The proposed model differs in the way of treating subject-specific effects, and it arguably offers benefits in terms of computational efficiency.

There is a growing trend of adopting functional data analysis tools in longitudinal data modeling. These methods specify observations as linear combinations of functional principal components (FPCs), with the FPCs represented as expansions of a pre-specified basis. Bayesian methods include Jiang et al. (2020) for continuous responses, and Van Der Linde (2009) for binary and count responses. Challenges include inference which is sensitive to the basis choice, and a complex orthogonality constraint on the FPCs. Recently, Matuk et al. (2022) proposed an approach that can serve as foundation for generalized FPC analysis of sparse and irregular binary responses. Nonetheless, our model involves a more parsimonious formulation, including the structure with the GP and TP predictive distributions.

4.3 Data Illustrations with Binary Responses

4.3.1 Synthetic data examples

The principal goal of analyzing longitudinal data is to estimate the mean and covariance structure of the subject's repeated measurements. We conduct simulation studies to evaluate the proposed method on fulfilling this goal. In particular, Section 4.3.1.1 evaluates the proposed model's capacity to capture the fluctuation of the mean structure, and Section 4.3.1.2 explores its performance in estimating within subject covariance structure. In Section 4.3.1.3, we evaluate model performance on a scenario where the observations are made on irregular time points. Unless otherwise specified, the posterior analyses in this section are based on 5000 posterior samples collected every 4 iterations from a Markov chain of 30000 iterations, with the first 10000 samples being discarded.

4.3.1.1 Estimating Mean Structure

Consider a generic process of generating longitudinal binary responses,

$$\begin{aligned}
&\mathbf{Y}_i = Y_i(\boldsymbol{\tau}_i) \mid \mathcal{Z}_i(\boldsymbol{\tau}_i) \overset{ind.}{\sim} Bin(1, \eta(\mathcal{Z}_i(\boldsymbol{\tau}_i))), \quad \boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iT_i}), \quad i = 1, \ldots, n, \\
&\mathcal{Z}_i(\boldsymbol{\tau}_i) = \boldsymbol{\mathcal{Z}}_i = f(\boldsymbol{\tau}_i) + \boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i \quad \boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}),
\end{aligned} \tag{4.6}$$

where $\eta(\cdot)$ is a generic link function mapping $\mathbb{R}$ to $(0, 1)$, $f(\tau)$ is a signal function, and $\boldsymbol{\omega}_i$ is a realization from a mean zero continuous stochastic process that depicts the temporal covariance within subject. The objective is twofold. First, to estimate the subject's probability response curve, which is defined as the probability of obtaining positive response, as a function of time. Second, to estimate the true underlying signal function.

We consider three data generating processes. The specific choice of $\eta(\cdot)$, $f(\tau)$ and $\boldsymbol{\omega}_i$ for each generating process is summarized as follows:

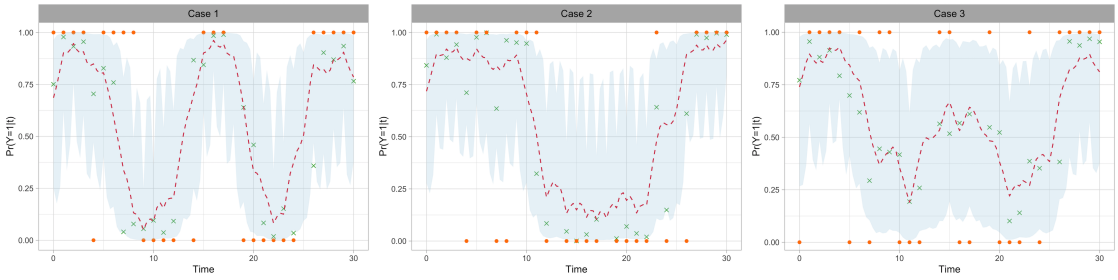
- Case 1: $\eta_1(\cdot) = \varphi(\cdot)$, where $\varphi(\cdot)$ is the expit function, $f_1(\tau) = 0.3 + 3\sin(0.5\tau) + \cos(\tau/3)$, and $\boldsymbol{\omega}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, K_1(\boldsymbol{\tau}, \boldsymbol{\tau}))$, with covariance kernel $K_1(\tau_t, \tau_{t'}) = \exp(-|\tau_t - \tau_{t'}|^2)$.

- Case 2: $\eta_2(\cdot) = \Phi(\cdot)$, where $\Phi(\cdot)$ denotes the CDF of standard normal distribution, $f_2(\tau) = 0.1 + 2\sin(0.25\tau) + \cos(0.25\tau)$, and $\boldsymbol{\omega}_i \overset{i.i.d.}{\sim} MVT(5, \mathbf{0}, K_2(\boldsymbol{\tau}, \boldsymbol{\tau}))$, with covariance kernel $K_2(\tau_t, \tau_{t'}) = \frac{1}{3}\exp(-|\tau_t - \tau_{t'}|^2)$.

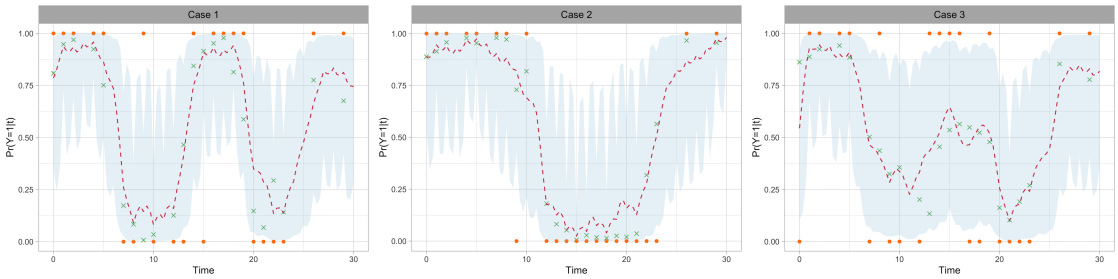
- Case 3: a mixture of Case 1 and Case 2, with equal probability of generating data from each model.

For $n = 30$ subjects, we simulate $T = 31$ binary observations at time $\tau = 0, \ldots, 30$, following the aforementioned data generating processes. To enforce an unbalanced study design, we randomly drop out a proportion of the simulated data. We term the drop out proportion sparsity level, for which we consider 10%, 25% and 50%.

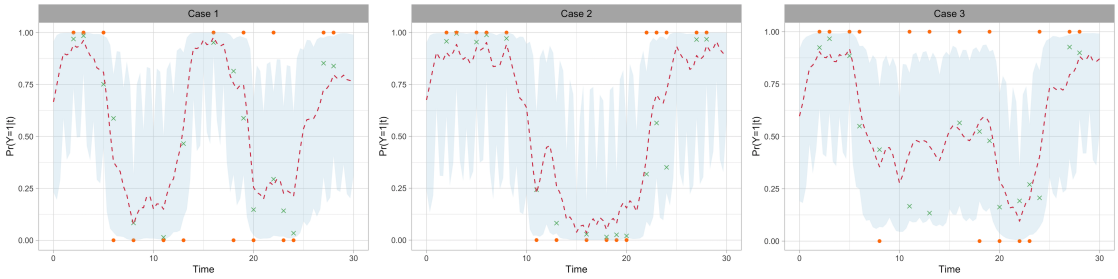
The proposed hierarchical model is applied to the data, with a weakly informative prior placed on the mean structure. We obtain posterior inference of the probability response curve and the signal process on a finer grid $\boldsymbol{\tau}^+ = (0, \frac{1}{3}, \frac{2}{3}, \ldots, 30)$. Figure 4.1 plots posterior point and interval estimates of the subject's probability response curve for a randomly selected one in each case. Despite the data generating process and the sparsity level, the model can recover the evolution of the underlying probability used in generating binary responses. We observe a shrink in the interval estimate at the set of grid points where at least one subject has observation, that is, $\boldsymbol{\tau}$. The increase in the credible interval width at $\check{\boldsymbol{\tau}}$ reflects the lack of information at those time grids.



(a) Sparsity level at 10%.



(b) Sparsity level at 25%.



(c) Sparsity level at 50%.

Figure 4.1: Simulation study regarding the mean structure. Inference results for the probability response curve. In each panel, the dashed line and shaded region correspond to the posterior mean and 95% credible interval estimates, the (orange) dot is the original binary data, whereas the (green) cross denotes the true probability of generating that responses.

We further investigate the model's ability in out-of-sample prediction, by estimating the probability response curve for a new subject from the same cohort. Figure 4.2 shows the posterior point and interval estimates of $\Pr(Y_*(\tau_{*t}) = 1)$, including, as a reference point, the posterior mean estimates of each subject's probability response curve $\Pr(Y_i(\tau_{it}) = 1)$, $i = 1, \ldots, n$. The true probability function that triggered the binary response, given as the signal transformed by the

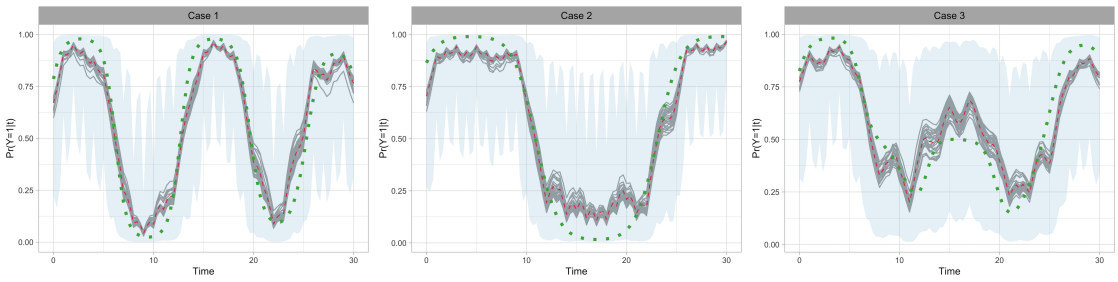


Figure 4.2: Simulation study regarding the mean structure. Prediction of the probability response curve for a new subject. In each panel, the dashed lines and shaded region shows the posterior mean and 95% interval estimates of probability response curve for a new subject. The solid lines are the posterior mean estimates of probability response curves for the in-sample subjects. The dotted line is the true probability function for generating binary responses.

link function, is also shown in the figure. It is obtained with the simulated data with 10% sparsity, while there is no major difference for the other two sparsity levels. The behavior of the probability response curve for the new subject is to be expected. It follows the overall trend depicted by the true underlying probability function, while suffers from a comparable level of measurement error with the observed subjects. Here, the point estimates exhibit local, non-smoothing behavior, which is due to the lack of repeat measurements. Actually, if we observe more responses per subject, the estimated probability response curve will become smoother.

It is also of interest to assess the model's ability in recovering the underlying continuous signal process, since the signal process describes the intrinsic behavior and is crucial to answer related scientific questions. In our proposed model, the signal process is modeled nonparametrically through a GP. To further emphasize the benefits of this model formulation, we compare the proposed model with its simplified backbone. The simpler model differs from the original one in modeling the mean function. Instead of modeling the mean function μ through a GP, we consider the parametric form $\mu(\tau) \equiv \mu_0$, with $\mu_0 \sim N(a_\mu, b_\mu)$. The model's ability in capturing the signal process is summarized by the rooted mean square error

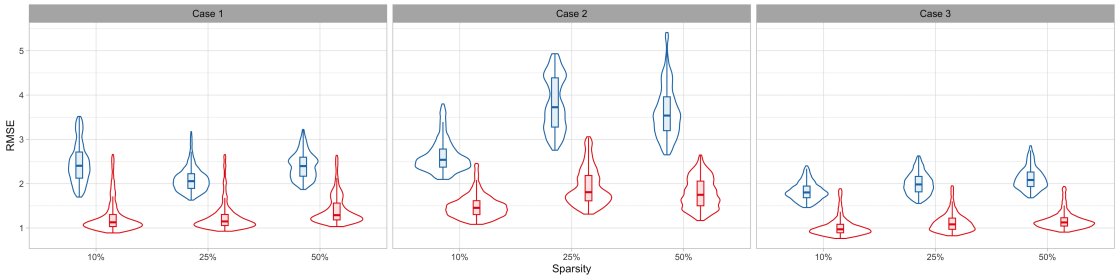


Figure 4.3: Simulation study regarding the mean structure. Box and violin plots of the posterior samples of RMSE for different data generating process and sparsity level combinations. The red box corresponds to the proposed model while the blue box is for the simplified model.

(RMSE), which is defined by $\text{RMSE}^{\mathcal{M}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\boldsymbol{\tau}^+|}\sum_{\tau\in\boldsymbol{\tau}^+}(\hat{Z}_i^{\mathcal{M}}(\tau) - f(\tau))^2}$. Here, $\hat{Z}_i^{\mathcal{M}}(\tau)$ denotes the estimated signal for subject i evaluated at time τ, under model $\mathcal{M}$, which can be obtained at every MCMC iteration. Figure 4.3 explores the posterior distribution of the RMSE under the proposed model and its simplified version, for different data generating process and sparsity level combinations. Despite the scenario, the proposed model shows a notably smaller RMSE. Contrasting the performance with the simpler model highlights the practical utility of including the GP prior layer for the mean function in terms of effective estimation of the underlying continuous signal process.

4.3.1.2 Estimating Covariance Structure

Since we emphasize the importance of modeling dependence in longitudinal data, we now explore how well our model works for estimating different covariance structure. Consider the data generating process in (4.6), with expit link function and signal $f(\tau) = 0.1 + 2\sin(0.5\tau) + \cos(0.5\tau)$. We examine a number of possible choices for generating $\boldsymbol{\omega}_i$, that imply covariance structures which would not be in the same form as the covariance kernel used in the proposed model. The primary interest is to exhibit the robustness of covariance kernel choice to different true

covariance structures. We let $T_i = T$ and $\tau_{it} = \tau_t$, namely that all subjects are observed over the same time grids. For $n = 100$ subjects, we generate sequences of length $T = 11$ at time $\tau = 0, \ldots, 10$. We study the following options of generating $\boldsymbol{\omega}_i$:

- Case 1: $\boldsymbol{\omega}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, K_1(\boldsymbol{\tau}, \boldsymbol{\tau}))$, with squared exponential kernel $K_1(\tau_t, \tau_{t'}) = \exp(-|\tau_t - \tau_{t'}|^2/(2 \cdot 3^2))$. Each realized trajectory is infinitely differentiable.
- Case 2: $\boldsymbol{\omega}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, K_2(\boldsymbol{\tau}, \boldsymbol{\tau}))$, with exponential kernel $K_2(\tau_t, \tau_{t'}) = \exp(-|\tau_t - \tau_{t'}|/5)$. Each realization is effectively from a continuous-time AR(1) GP.
- Case 3: $\boldsymbol{\omega}_i \overset{i.i.d.}{\sim} MVT(5, \mathbf{0}, K_3(\boldsymbol{\tau}, \boldsymbol{\tau}))$, with compound symmetry kernel $K_3(\tau_t, \tau_{t'}) = \mathbf{I}_{\{\tau_t = \tau_{t'}\}} + 0.4\mathbf{I}_{\{\tau_t \neq \tau_{t'}\}}$. The covariance between two observations remains a constant, despite their distance.
- Case 4: $\boldsymbol{\omega}_i \overset{i.i.d.}{\sim} MVT(5, \mathbf{0}, K_4(\boldsymbol{\tau}, \boldsymbol{\tau}))$, with kernel $K_4(\tau_t, \tau_{t'}) = 0.7K_2(\tau_t, \tau_{t'}) + 0.3K_3(\tau_t, \tau_{t'})$, a mixture of AR(1) and compound symmetry covariance structure.

In terms of longitudinal binary responses, the covariance structure can be elucidated in two senses, namely the covariance between the pair of binary data $(Y_i(\tau_t), Y_i(\tau_{t'}))$ and between the pair of signal $(Z_i(\tau_t), Z_i(\tau_{t'}))$. We consider the covariance structure of the signal process first. From Proposition 4.3, $\mathrm{Cov}(Z_i(\tau_t), Z_i(\tau_{t'})) = \Psi_{\boldsymbol{\phi}}(\tau_t, \tau_{t'})$, $\forall i$, where the covariance function $\Psi_{\boldsymbol{\phi}}$ is defined in (4.2.1). Hence, the signal covariance structure estimated from the model is also isotropic, facilitating a graphic comparison between the posterior estimate of $\Psi_{\boldsymbol{\phi}}(\tau_d)$ versus the true covariance kernel $K(\tau_d)$, where $\tau_d = |\tau_t - \tau_{t'}|$. The results are presented in Figure 4.4. As expected, the proposed model recovers the truth, despite the mis-specification of the covariance kernel. Comparing with the other

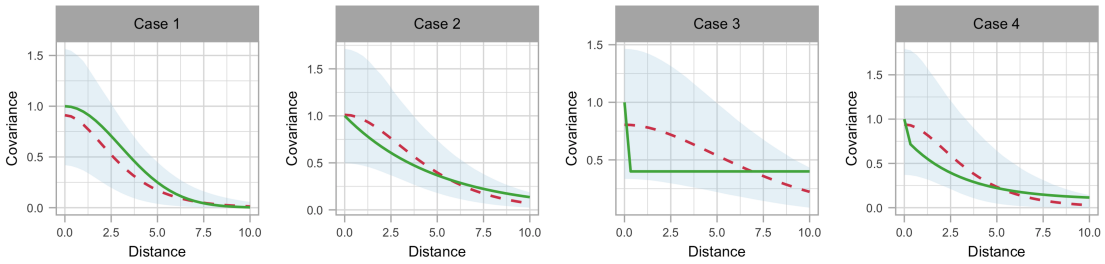
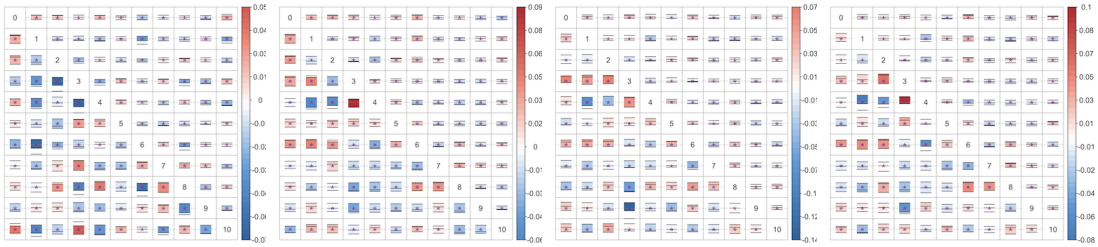


Figure 4.4: Simulation study regarding the covariance structure. Inference results for the signal covariance kernels. In each panel, the dashed line and shaded region correspond to the posterior mean and 95% credible interval estimates, whereas the solid line denotes the true covariance kernel.

three cases, the posterior point estimate of covariance kernel is less accurate in Case 3. This can be explained by noticing that the constant covariance in that case violates the model assumption. Nonetheless, the posterior interval still covers the truth.

As for the covariance between the pair of binary responses, we consider two measurements, the Pearson correlation coefficient and the tetrachoric correlation coefficient. For a review of the definitions and properties of these two correlation coefficients, we refer to Ekström (2011). At each MCMC iteration, we predict a new sequence of binary responses of length T, denoted as $\{Y_{i*}^{(s)}(\boldsymbol{\tau}) : s = 1, \ldots, S\}$. Correspondingly, we also obtain samples of binary sequences from the true data generating process, denoted by $\{\hat{Y}_{i*}^{(s)}(\boldsymbol{\tau}) : s = 1, \ldots, S\}$. Both sets of binary sequences form S/n datasets that mimic the original samples. From the datasets comprised by posterior predictive samples $Y_{i*}^{(s)}(\boldsymbol{\tau})$, we obtain interval estimates of the two correlation coefficients. In addition, for $\hat{Y}_{i*}^{(s)}(\boldsymbol{\tau})$ that are generated from the truth, we obtain point estimates, which can be viewed as the correlation coefficients from the data, accounting for the variation in the data generating process. Notice that marginally the binary process is not guaranteed to be isotropic. Hence, the correlation coefficients should be calculated for every possible pair of $(\tau_t, \tau_{t'}) \in \boldsymbol{\tau}$.



(a) Case 1. **(b)** Case 2. **(c)** Case 3. **(d)** Case 4.

Figure 4.5: Simulation study regarding the covariance structure. Posterior interval estimate of correlation coefficients ("box") versus point estimate obtained from the true data generating process ("$\star$"). In each panel, the upper triangle and the lower triangle are for the Pearson and the rachoric correlation coefficient, respectively.

The resulting point and interval estimates of both types of correlation coefficients are displayed in Figure 4.5. All the posterior interval estimates cover the truth, indicating that the proposed model effectively captures the binary covariance structure.

The simulation studies have illustrated the benefits of our approach, that is, avoiding possible bias in covariance structure estimation caused by mis-specification of the covariance kernel for the signal process. This model feature is driven by the IWP prior placed on the covariance function. To emphasize this point, we consider an alternative, simplified modeling approach, with $Z_i \overset{i.i.d.}{\sim} GP(\mu, \Psi_\phi)$, $\mu \sim GP(\mu_0, \Psi_\phi/\kappa)$. That is, instead of modeling the covariance function nonparametrically, we assume a covariance kernel of certain parametric form, specified by Ψ_ϕ. We consider the centralized signal process $\omega_i = Z_i - \mu$ evaluated at a finite grid $\boldsymbol{\tau}$, denoted as $\boldsymbol{\omega}_i$. Under the proposed model, $\boldsymbol{\omega}_i \overset{i.i.d.}{\sim} MVT(\nu, \mathbf{0}, \Psi_\phi(\boldsymbol{\tau}, \boldsymbol{\tau}))$, while under the simplified model, $\boldsymbol{\omega}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, (1 + \frac{1}{\kappa})\Psi_\phi(\boldsymbol{\tau}, \boldsymbol{\tau}))$. We know the true distribution of $\boldsymbol{\omega}_i$ from the data generating process. Therefore, we can compute the 2-Wasserstein distance between the model estimated distribution of $\boldsymbol{\omega}_i$ to the truth. The usage of 2-Wasserstein distance is motivated by its straightforward

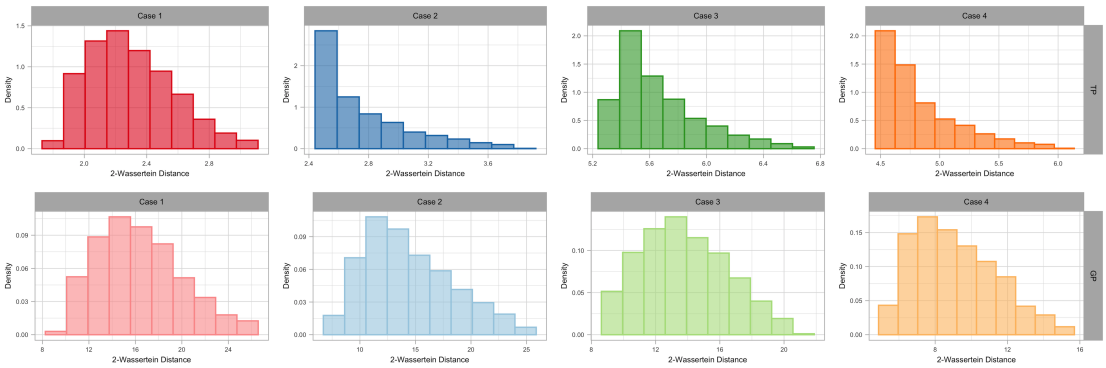


Figure 4.6: Simulation study regarding the covariance structure. Histogram for the posterior samples of the 2-Wasserstein distance between the f.d.d.s. of the centralized signal process obtained from the proposed model (upper panel) and the simplified model (lower panel) to the truth.

interpretation: a 2-Wasserstein distance of d means that coordinatewise standard deviations differ by at most d (Huggins et al., 2020, Thm. 3.4). Iterating over the posterior samples of model parameters, we obtain the distributions of 2-Wasserstein distance between the model estimated distribution of $\boldsymbol{\omega}_i$ and the truth, which is shown in Figure 4.6. Clearly, for the proposed model, the 2-Wasserstein distances are substantially small. Contrasting the performance highlights the practical benefits of modeling the covariance structure nonparametrically.

4.3.1.3 Model performance with irregular observing points

The simulation studies discussed above focus on longitudinal settings with observations made at integer time points, which is the typical scenario in longitudinal studies. To further illustrate the practical benefit of adopting the functional data analysis perspective, we consider a synthetic scenario in which observations are made irregularly. Specifically, the pooled grid $\boldsymbol{\tau}$ consists of 30 grid points that are uniformly sampled on the interval $(0, 30)$. We consider $n = 50$ subjects. For each of them, we first generate repeated measurements on the pooled grid, following the scheme described in Section 4.3.1.1 Case 1. The unbalanced setting is imposed

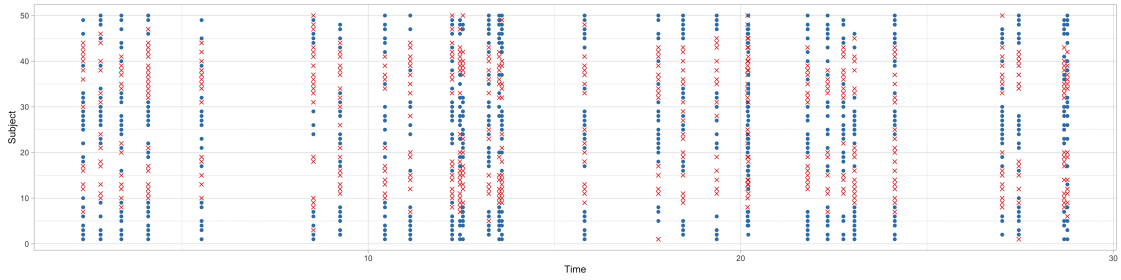


Figure 4.7: Simulation study with irregular observing points. Visualization of the repeated measurements for each subject. The blue dot marks a positive response while the red cross represents a negative response.

by randomly dropping out 30% of the simulated observations. The observed data are visualized in Figure 4.7, which shows heavily irregular pattern.

To assess the model's performance in out-of-sample prediction, we plot posterior point and interval estimates of a new subject's probability response curve in Figure 4.8, including the posterior mean estimate of each in-sample subjects' probability response curves. Similar to the scenarios discussed in Section 4.3.1.1, the predicted mean captures the true probability function well. Comparing to the cases with more regular observed time points, the shrinkage of the credible interval at observed points is less prominent. Nonetheless, the intervals are shorter at the region where observing points are more concentrated, which is to be expected.

Moreover, we compare our model with a traditional approach, which postulates a GLMM structure. Specifically, the model used for comparison is formulated as follows:

$$Y_{it} \mid \mathcal{Z}_{it} \overset{ind.}{\sim} Bin(1, \varphi(\mathcal{Z}_{it})),\ \mathcal{Z}_{it} = \tilde{\boldsymbol{\tau}}_{it}^{\top}\boldsymbol{\beta} + \sum_{k=1}^{K} S_{itk} b_k + \mu_i + \epsilon_{it},\ t = 1, \ldots, T_i,$$

for $i = 1, \ldots, n$. The components of this model are set similar to the modeling approach described in Section 4.3.2.3, except that here the cubic B-spline basis functions have 4 inner knots that separate the whole observing period into 5 equal

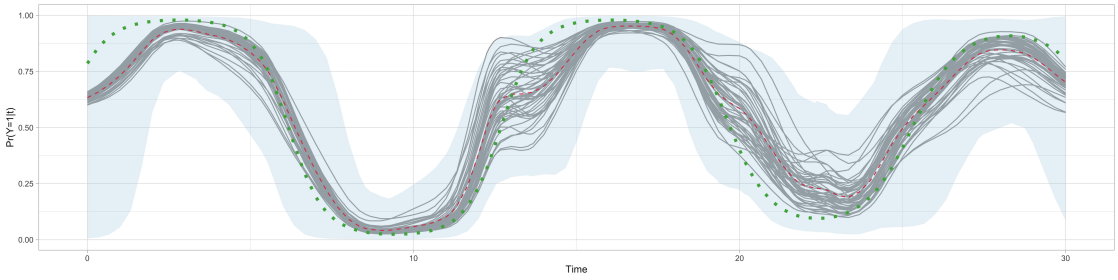


Figure 4.8: Simulation study with irregular observing points. Posterior inference of a new subject's probability response curve. The dashed line and shaded region show the posterior mean and 95% interval estimates of probability response curve for a new subject. The dotted line is the true probability function for generating binary responses. As references, the solid lines are the posterior mean estimates of probability response curves for the in-sample subjects.

Table 4.1: Simulation study with irregular observing points. Comparison between the proposed model and the generalized linear mixed effects model using two different criteria. The values in bold correspond to the model favored by the particular criterion.

Model	Posterior predictive loss			CRPS
	$G(\mathcal{M})$	$P(\mathcal{M})$	$G(\mathcal{M}) + P(\mathcal{M})$	
Proposed	**125.78**	**152.33**	**278.11**	**0.12**
GLMM	150.55	154.16	304.71	0.14

length intervals. We perform model comparison using the posterior predictive loss criterion and CRPS, with the results summarized in Table 4.1. Our model is favored by both criteria. The key distinction between the two models is that we adopt a flexible, functional data analysis modeling approach, which appears to be beneficial, especially when the observing time points are highly irregular.

4.3.2 Real Application: *Studentlife* data

4.3.2.1 Data for Analysis

Studentlife (Wang et al., 2014) is a study that integrates automatic sensing data and an EMA component to probe students' mental health status and to study its

relationship with students' academic performance and behavior trends. The data were collected by a smartphone app carried by 48 students over a 10-week term at Dartmouth College. The dataset is available from the R package "studentlife" (Fryer et al., 2022).

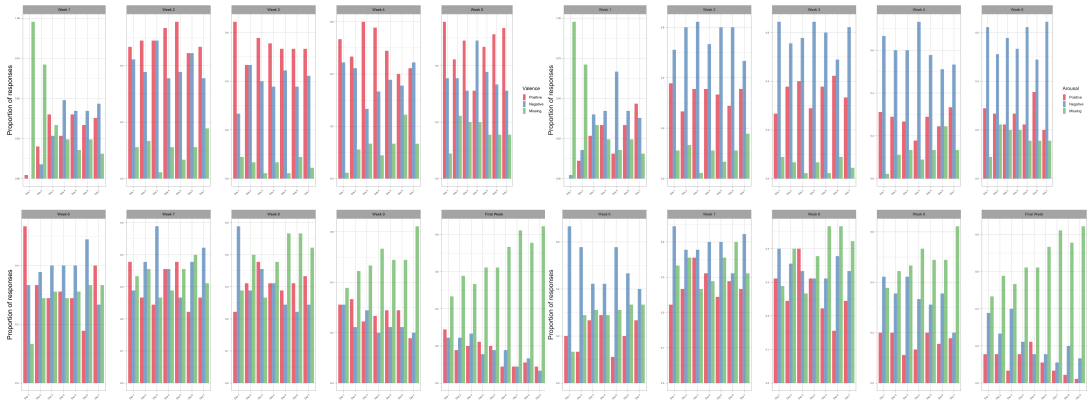
We focus on a subset of the data that corresponds to assessing the students' emotional status. In the *Studentlife* study, the assessment of emotion is conducted by the Photographic Affect Meter (PAM), a tool for measuring affect in which users select from a wide variety of photos the one which best suits their current mood (Pollak et al., 2011). The PAM survey is deployed to the mobile app and prompts everyday during the study period. The participants either respond to the survey, or ignore it, introducing missingness. The outcome of the survey contains two attributes, the PAM valence and the PAM arousal. They are scores of -2 to 2 (excluding 0) that measure the subject's extent of displeasure to pleasure or state of activation ranging from low to high, respectively. We dichotomize the valence and arousal scores by their sign, representing the positive values by 1. In this section, we focus on analyzing the change of binary valence and arousal responses to evaluate students' affects as the term progresses.

The data were collected during the spring 2013 term at Dartmouth college. We set the study period according to the official academic calendar, from the first day of classes (March 25, 2013) to the end of the final exam period (June 4, 2013), resulting in a total of 72 days. We exclude subjects with less than 12 responses, resulting in 45 students. The longitudinal recordings of valence or arousal of the i-th student are denoted by $Y_i(\boldsymbol{\tau}_i)$, for $i = 1, \ldots, 45$, where the student-specific grid points are a subset of $\boldsymbol{\tau} = (0, 1, \ldots, 71)^\top$, representing the days on which the measurements are recorded. Several special events occurred during the study period, and we are particularly interested in investigating the change of students'

affects on the time intervals around these events. Specifically, the events and corresponding periods are: (i) Days following the Boston marathon bombing (April 15, 2013 to April 17, 2013); (ii) The Green Key (a spring festival at Dartmouth) period (May 17, 2013 to May 18, 2013); (iii) The Memorial Day long weekend (May 25, 2013 to May 27, 2013); (iv) The final examination period (May 31, 2013 to June 3, 2013).

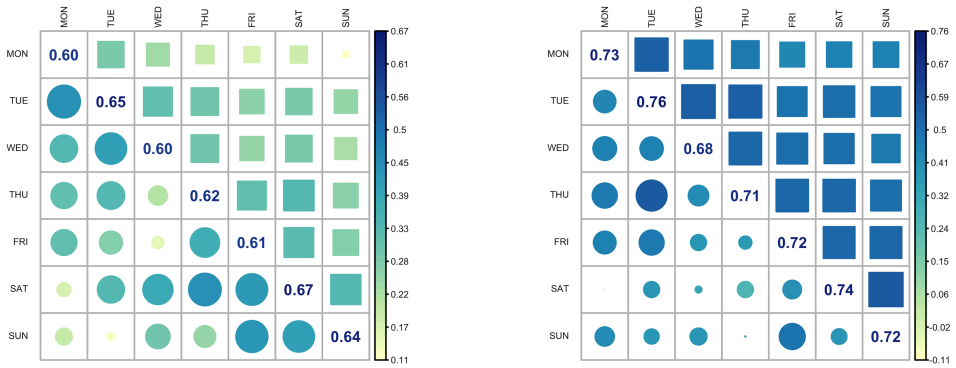
We retrieve the data for the specific responses and study period from the R package "studentlife" that contains the database for the entire study. Over all observations, the percentage of missing values is 31.1%. We treat the missingness as missing-at-random. The main reson is that the responses are from an ecological momentary assessment (EMA) study (mentioned explicitly in the original publication (Wang et al., 2014) and the description of the corresponding R package). To reduce the potential bias caused by nonrandom missing responses, at the design stage, EMA studies place a premium on obtaining high levels of subject compliance with the assessment protocol (Shiffman et al., 2008). As a result, one can assume the occurrence of missing values is driven by a completely random process (Ruwaard et al., 2018), and therefore ignorable (see e.g. Hedeker et al., 2009; Shiffman et al., 2009).

For an example of empirical evaluation for the specific data, we plot the proportion of the three types of responses (positive, negative, and missing) over time, for valence and arousal scores, aggregated over the subjects. The corresponding plot is displayed in Figure 4.9. The plots show no strong pattern of missingness over time, apart from that more missing responses appear at the beginning and toward the end of the study. Combining with the feature from the design of EMA studies, the missing-at-random assumption is arguably plausible for our illustrative data analyses.



(a) Valence score **(b)** Arousal score

Figure 4.9: *Studentlife* data. Proportion of three types of response (positive, negative, and missing,) over time, for valence and arousal scores.



(a) Valence. **(b)** Arousal.

Figure 4.10: *Studentlife* data. Empirical estimate of the correlation coefficients between binary responses within a week. In each panel, the upper triangle and the lower triangle are for the Pearson and the tetrachoric correlation coefficient, respectively.

We further explore the correlations between the binary responses within a week. We split the whole observation sequence into batches representing a week, and empirically calculate the Pearson and the tetrachoric correlation coefficient for each pair of time and distance combinations. Figure 4.10 presents the results. It suggests that the correlation of the students' response to valence and arousal decreases slowly in time.

4.3.2.2 Analysis and Results

We fit the proposed model for the binary valence and arousal responses separately. We specify the prior for the model parameters by the procedure mentioned in Section 4.2.3. We suggest the default hyperprior for μ_0 and ν as $\mu_0 \sim N(0, 100)$ and $\nu \sim Unif(4, 30)$. σ^2 and ρ control the covariance structure. Their prior hyperparameters can be determined by exploring the covariance structure of the data. On the other hand, the hyperprior for σ_ϵ^2 depends on the belief about the range and the degree of freedom of the measurement error. In general, the measurement error reflect the remaining variability of the underlying continuous process, whose major change has been captured by the signal process. Consequently, we assume the measurement error range should be small, and we pick a moderate value for the error degree of freedom. Posterior inference results are based on 5000 MCMC samples obtained every 4 iterations from a chain of 50000 iterations with a 30000 burn-in period (which is conservative).

We first examine in Figure 4.11 the probability response curves, defined as the probability of obtaining positive valence or arousal as a function of time. For the valence, the happiness level drops as the term begins and increases when the term ends. The Boston marathon bombing may have had a minor effect on the valence. We observe local peaks around the Green Key festival and the Memorial Day holiday. As the students finish their exams, there is a trend toward happiness. As for arousal, it is relatively stable at the beginning of the term, and fluctuates as the term progresses. There is a drop in activation level after the Boston marathon bombing and during the final exam period, while the activation level reaches a local maximum at around the Green Key festival and the Memorial Day holiday.

Moreover, we assess the student's emotional status on specific days. According to Russell (1980), various states of emotional status can be represented by points

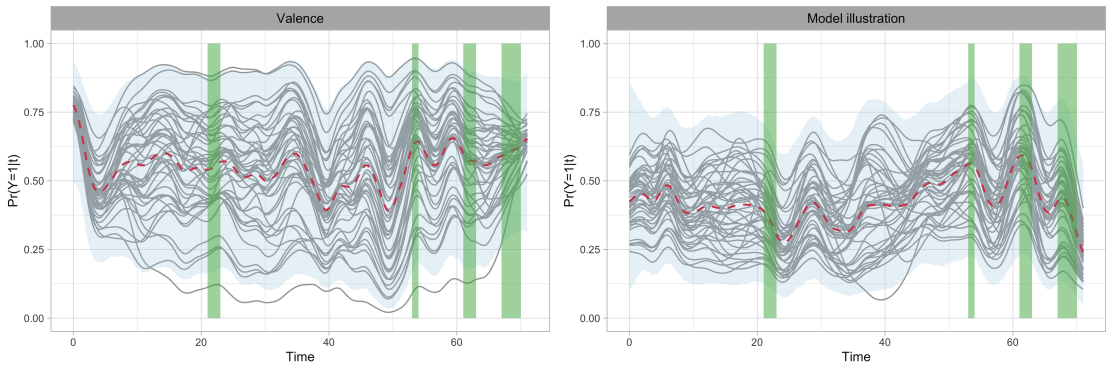
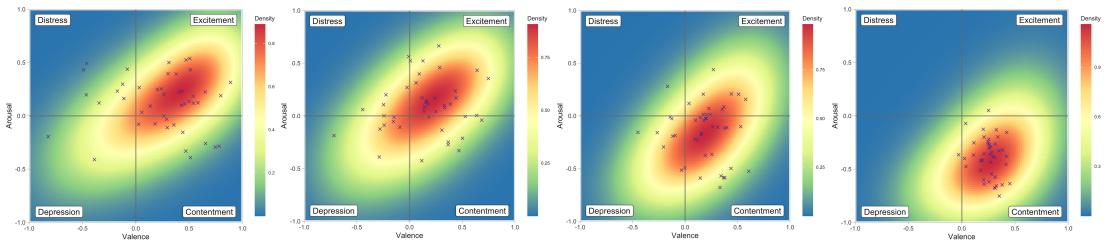


Figure 4.11: *Studentlife* data. Posterior mean (dashed line) and 95% interval estimate (shaded region) of the probability response curve for an out-of-sample subject. The posterior mean estimates of probability response curves for in-sample subjects are given by the solid lines. The vertical shaded regions correspond to the four special time periods (see Section 4.3.2.1).

located at the two dimensional mood coordinate space spanned by valence for the horizontal dimension and arousal for the vertical dimension. Moods such as excitement, distress, depression, and contentment, are represented by points in the quadrants of the space. For each observation, we can map the corresponding pairs of probabilities for positive valence and arousal onto the unit square in the mood space. In Figure 4.12, the density heatmap is obtained by the posterior samples of positive probabilities for a new student of the same cohort, while the posterior means of the in-sample positive probabilities are marked by crosses. Panels (a) and (b) suggest the students are mostly excited at the festival and holiday. Moving from panel (c) to panel (d), we observe that the happiness level increases and the activation level decreases towards the end of the exam period.

We also obtain the posterior point and 95% interval estimate for the covariance kernel of the signal process, which is displayed in Figure 4.13. It is noteworthy that there is a similar decreasing trend for the two distinct binary responses of valence and arousal. The practical range, defined as the distance at which the correlation is 0.05, has an estimated mean of 20.99 for valence and 22.97 for arousal.



(a) Green Key **(b)** Memorial Day **(c)** Final exams begin **(d)** Final exams end

Figure 4.12: *Studentlife* data. Posterior density estimate of an out-of-sample subject's valence and arousal probability over the mood coordinate space on four specific days. In each panel, the crosses represent the posterior means of the in-sample subjects' valence and arousal probability mapped to the mood coordinate space.

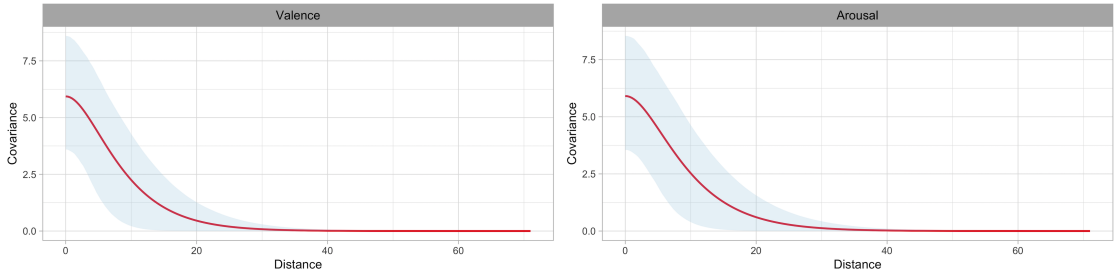


Figure 4.13: *Studentlife* data. Posterior mean (solid line) and 95% interval estimate of the signal process covariance kernel.

4.3.2.3 Performance comparisons

For comparison with a traditional approach, we consider an analysis of the data under the GLMM setting. In particular, we assume, for $i = 1, \ldots, n$,

$$Y_{it} \mid \mathcal{Z}_{it} \overset{ind.}{\sim} Bin(1, \varphi(\mathcal{Z}_{it})), \;\; \mathcal{Z}_{it} = \tilde{\boldsymbol{\tau}}_{it}^{\top}\boldsymbol{\beta} + \sum_{k=1}^{K} S_{itk} b_k + \mu_i + \epsilon_{it}, \;\; t = 1, \ldots, T_i,$$

where $\tilde{\boldsymbol{\tau}}_{it} = (1, \tau_{it})^{\top}$, $\boldsymbol{\beta}$ is the vector of fixed effects, and $\epsilon_{it} \overset{i.i.d.}{\sim} N(0, \sigma_{\epsilon}^2)$ is the measurement error. To allow flexibility in modeling the time effect, we consider cubic B-spline basis functions with $K = 9$ knots that separate naturally the observed interval by week; S_{itk} is the k-th basis associated with time, with parameter

Table 4.2: *Studentlife* data. Summary of comparison between the proposed model and the generalized linear mixed effects model using two different criteria. The values in bold correspond to the model favored by the particular criterion.

Response	Model	Posterior predictive loss			CRPS
		$G(\mathcal{M})$	$P(\mathcal{M})$	$G(\mathcal{M}) + P(\mathcal{M})$	
Valence	Proposed	**428.09**	**475.31**	**903.40**	**0.19**
	GLMM	456.09	475.83	931.92	0.20
Arousal	Proposed	**457.62**	496.63	**954.25**	**0.20**
	GLMM	476.17	**492.28**	968.45	0.21

$b_k \overset{i.i.d.}{\sim} N(0, \sigma_b^2)$. Finally, $\mu_i \overset{i.i.d.}{\sim} N(0, \sigma_\mu^2)$ are subject-specific random effects. The model is implemented using the integrated nested Laplace approximation (INLA) approach (Rue et al., 2009) with the "INLA" package in R (Rue et al., 2017). We used the default choices provided by the R package for the prior on $\boldsymbol{\beta}$ (a flat prior), and for the values of the variance terms, σ_ϵ^2, σ_b^2, and σ_μ^2.

We perform model comparison using two different metrics: the posterior predictive loss criterion which combines a goodness-of-fit term, $G(\mathcal{M})$, and a penalty term, $P(\mathcal{M})$, for model complexity (Gelfand and Ghosh, 1998); and, the continuous ranked probability score (CRPS), defined in terms of predictive cumulative distribution functions (Gneiting and Raftery, 2007). Both criteria can be calculated from the posterior samples for model parameters, and both favor the model with a smaller value. Table 4.2 summarizes the results. For the valence response, both criteria favor the proposed model. As for the arousal response, the proposed model provides a more accurate fit to the data, while being penalized more than the GLMM with respect to model complexity. Nonetheless, our model is favored in terms of total posterior predictive loss, as well as by the CRPS criterion.

4.4 Model for ordinal responses

4.4.1 The extended model

We extend the model developed in Section 2.2.1 to handle ordinal responses. Suppose the observation on subject i at time τ_{it}, denoted by Y_{it}, takes C possible categories. We can equivalently encode the response as a vector with binary entries $\mathbf{Y}_{it} = (Y_{i1t} \ldots, Y_{iCt})$, such that $Y_{it} = j$ is equivalent to $Y_{ijt} = 1$ and $Y_{ikt} = 0$ for any $k \neq j$. We assume a multinomial response distribution for $\mathbf{Y}_{it}$, factorized in terms of binomial distributions,

$$Mult(\mathbf{Y}_{it} \mid m_{it}, \omega_{i1t}, \ldots, \omega_{iCt}) = \prod_{j=1}^{C-1} Bin(Y_{ijt} \mid m_{ijt}, \varphi(Z_{ijt} + \epsilon_{ijt})) \tag{4.7}$$

where $m_{it} = \sum_{j=1}^{C} Y_{ijt} \equiv 1$, $m_{i1t} = m_{it}$, and $m_{ijt} = m_{it} - \sum_{k=1}^{j-1} Y_{ikt}$. This factorization bridges the gap between binary and ordinal responses. Similar to the model for binary responses, we adopt a functional data analysis perspective on $\{Z_{ijt}\}$, modeling them separately through the hierarchical framework developed in Section 2.2.1. That is, $Z_{ij}(\tau) \mid \mu_j, \Sigma_j \overset{i.i.d.}{\sim} GP(\mu_j, \Sigma_j)$, for $i = 1, \ldots, n$, and $\mu_j | \Sigma_j \overset{ind.}{\sim} GP(\mu_{0j}, (\nu_j - 3)\Sigma_j)$, $\Sigma_j \overset{ind.}{\sim} IWP(\nu_j, \Psi_{\phi_j})$, where $\boldsymbol{\phi}_j = \{\sigma_j^2, \rho_j\}$, for $j = 1, \ldots, C-1$. The error terms are modeled as $\epsilon_{ijt} \mid \sigma_{\epsilon j}^2 \overset{ind.}{\sim} N(0, \sigma_{\epsilon j}^2)$. Hence, the hierarchical model for the data can be expressed as

$$\begin{aligned}
&\mathbf{Y}_i | \{\mathbf{Z}_{ij}\}, \{\boldsymbol{\epsilon}_{ij}\} \overset{ind.}{\sim} \prod_{t=1}^{T_i} \prod_{j=1}^{C-1} Bin(Y_{ijt} \mid m_{ijt}, \varphi(Z_{ijt} + \epsilon_{ijt})), \quad i = 1, \ldots, n, \\
&\mathbf{Z}_{ij} \mid \mu_j(\boldsymbol{\tau}_i), \Sigma_j(\boldsymbol{\tau}_i, \boldsymbol{\tau}_i) \overset{ind.}{\sim} N(\mu_j(\boldsymbol{\tau}_i), \Sigma_j(\boldsymbol{\tau}_i, \boldsymbol{\tau}_i)), \quad \boldsymbol{\epsilon}_{ij} \mid \sigma_{\epsilon j}^2 \overset{ind.}{\sim} N(\mathbf{0}, \sigma_{\epsilon j}^2 \mathbf{I}), \\
&\boldsymbol{\mu}_j \mid \mu_{0j}, \boldsymbol{\Sigma}_j, \nu_j \overset{ind.}{\sim} N(\mu_{0j}\mathbf{1}, (\nu_j - 3)\boldsymbol{\Sigma}_j); \; \boldsymbol{\Sigma}_j \mid \nu_j, \boldsymbol{\Psi}_j \overset{ind.}{\sim} IW(\nu_j, \boldsymbol{\Psi}_j),
\end{aligned} \tag{4.8}$$

for $j = 1, \ldots, C-1$, where $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{iT_i})^\top$, $\mathbf{Z}_{ij} = (Z_{ij1}, \ldots, Z_{ijT_i})^\top$, $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijT_i})^\top$, and the collection of the functional evaluations on the pooled grid

$\boldsymbol{\tau}$ are denoted by the corresponding bold letter.

The structure in (4.7) is referred to as the continuation-ratio logits representation of the multinomial distribution (Tutz, 1991). In the context of Bayesian nonparametric modeling, it has been used as the kernel of nonparametric mixture models for cross-sectional ordinal regression (Kang and Kottas, 2022).

Examining model properties reveals the practical utility of the continuation-ratio logits structure. The factorization in (4.7) allows us to examine the probability response curves and the within subject covariance structure in the same fashion as for binary responses. Specifically, the continuation-ratio logit for response category j is the logit of the conditional probability of response j, given that the response is j or higher. As a consequence, for any finite grid $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)^\top$, the probability response curves are given by

$$\begin{aligned} \mathbf{P}_{\mathbf{j}\boldsymbol{\tau}} &= (\Pr(Y_{\tau_1} = j \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon), \ldots, \Pr(Y_{\tau_T} = j \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma^2_\epsilon))^\top \\ &= \mathrm{E}\left(\boldsymbol{\pi}_{j\boldsymbol{\tau}} \mid \mathbf{Z}_{j\boldsymbol{\tau}}, \sigma^2_{\epsilon j}\right) \prod_{k=1}^{j-1} \mathrm{E}\left((1 - \boldsymbol{\pi}_{k\boldsymbol{\tau}}) \mid \mathbf{Z}_{k\boldsymbol{\tau}}, \sigma^2_{\epsilon k}\right), \end{aligned} \tag{4.9}$$

where $\boldsymbol{\pi}_{j\boldsymbol{\tau}} = (\varphi(\boldsymbol{\mathcal{Z}}_{j1}), \ldots, \varphi(\boldsymbol{\mathcal{Z}}_{jT}))^\top$ and $\boldsymbol{\mathcal{Z}}_{j\boldsymbol{\tau}} \mid \mathbf{Z}_{j\boldsymbol{\tau}}, \sigma^2_{\epsilon j} \sim N(\mathbf{Z}_{j\boldsymbol{\tau}}, \sigma^2_{\epsilon j}\mathbf{I}_T)$, for $j = 1, \ldots, C$. To avoid redundant expressions, we include the term $\boldsymbol{\pi}_{C\boldsymbol{\tau}}$ and set it always equal to 1. As for the covariance structure, we study the joint probability of the repeated measurements on the same subject at time τ and τ' taking category j and j'. Exploiting the conditional independence structure across the categories,

$$\begin{aligned} &\Pr(Y_\tau = j, Y_{\tau'} = j' \mid \{\mathbf{Z}_{j\boldsymbol{\tau}}\}, \{\sigma^2_{\epsilon j}\}) \\ &= \begin{cases} \mathrm{E}(\pi_{j\tau}\pi_{j\tau'} \mid \mathbf{Z}_{j\boldsymbol{\tau}}, \sigma^2_{\epsilon j}) \prod_{k \neq j} \mathrm{E}[(1 - \pi_{k\tau})(1 - \pi_{k\tau'}) \mid \mathbf{Z}_{k\boldsymbol{\tau}}, \sigma^2_{\epsilon k}] & j = j' \\ \mathrm{E}[\pi_{j\tau}(1 - \pi_{j\tau'}) \mid \mathbf{Z}_{j\boldsymbol{\tau}}, \sigma^2_{\epsilon j}] \, \mathrm{E}[(1 - \pi_{j'\tau})\pi_{j'\tau'} \mid \mathbf{Z}_{j'\boldsymbol{\tau}}, \sigma^2_{\epsilon j'}] & \\ \quad \times \prod_{k \neq j, j'} \mathrm{E}[(1 - \pi_{k\tau})(1 - \pi_{k\tau'}) \mid \mathbf{Z}_{k\boldsymbol{\tau}}, \sigma^2_{\epsilon k}] & j \neq j' \end{cases} . \end{aligned} \tag{4.10}$$

Hence, we can explore the covariance of the two ordinal responses $\mathbf{Y}_\tau, \mathbf{Y}_{\tau'}$ by studying the pairwise covariance for each entry.

The continuation-ratio logits structure is also key to efficient model implementation. It implies a sequential mechanism, such that the ordinal response is determined through a sequence of binary outcomes. Starting from the lowest category, each binary outcome indicates whether the ordinal response belongs to that category or to one of the higher categories. This mechanism inspires a novel perspective on the model implementation. That is, we can re-organize the original data set containing longitudinal ordinal responses to create $C-1$ data sets with longitudinal binary outcomes. Then, fitting model (4.8) to the original data set is equivalent to fitting the model of Section 2.2.1 separately on the $C-1$ re-organized data sets. The procedure is elaborated below.

Denote the set of all possible subject and time indices by $\boldsymbol{\mathcal{I}}_1$, that is, $\boldsymbol{\mathcal{I}}_1 = \{(i,t) : i = 1,\ldots,n, t = 1,\ldots,T_i\}$. To build the first re-organized data set with binary outcomes, we create binary indicators $Y_{it}^{(1)}$, such that $Y_{it}^{(1)} = 1$ if $Y_{i1t} = 1$ and $Y_{it}^{(1)} = 0$ if $Y_{i1t} = 0$. The first data set is then $\boldsymbol{\mathcal{D}}_1 = \{Y_{it}^{(1)} : (i,t) \in \boldsymbol{\mathcal{I}}_1\}$. Moving to the second data set, we first filter out the observations that are already categorized into the smallest scale, and denote the remaining indices set by $\boldsymbol{\mathcal{I}}_2 = \boldsymbol{\mathcal{I}}_1 \setminus \{(i,t) : Y_{i1t} = 1\}$. This is the set of indices with original ordinal responses belonging to categories higher than or equal to the second smallest scale. Then, we create new binary indicators $Y_{it}^{(2)}$, such that $Y_{it}^{(2)} = 1$ if $Y_{i2t} = 1$, and $Y_{it}^{(2)} = 0$ if $Y_{i2t} = 0$. The second data set is obtained as $\boldsymbol{\mathcal{D}}_2 = \{Y_{it}^{(2)} : (i,t) \in \boldsymbol{\mathcal{I}}_2\}$. The process is continued until we obtain the $(C-1)$-th data set, $\boldsymbol{\mathcal{D}}_{C-1} = \{Y_{it}^{(C-1)} : (i,t) \in \boldsymbol{\mathcal{I}}_{C-1}\}$, where $\boldsymbol{\mathcal{I}}_{C-1}$ is the indices set such that the original ordinal responses belong to either category $C-1$ or C. Notice that every re-organized data set $\boldsymbol{\mathcal{D}}_j$, for $j = 1,\ldots,C-1$, contains longitudinal binary outcomes for which the model

of Section 2.2.1 is directly applicable. Provided the priors placed on each ordinal response category's parameters are independent, it is straightforward to verify that fitting separately the model for binary responses to the re-organized data sets $\{\boldsymbol{\mathcal{D}}_j : j = 1, \ldots, C-1\}$ is equivalent to fitting model (4.8) to the original data set. We formalize the conclusion in the following proposition.

Proposition 4.5. *Fitting the ordinal responses model in (4.8) is equivalent to fitting the model for binary responses separately,* $C-1$ *times to the data sets* $\{\boldsymbol{\mathcal{D}}_j : j = 1, \ldots, C-1\}$.

Based on Proposition 4.5, the posterior simulation algorithm for the ordinal responses model can be parallelized and implemented on separate cores. In applications where the number of response categories is moderate to large, such a parallel computing scheme is especially beneficial. Also, since the binary responses model serves as the backbone for modeling ordinal responses, the prior specification strategy and the posterior simulation method described in Section 4.2.3 can be readily extended to model (4.8). Finally, from (4.9) and (4.10), it is clear that the posterior samples obtained from the $C-1$ separate models suffice to obtain full posterior inference for the ordinal response process.

4.4.2 Data illustration

As an illustration example, we consider the PAM arousal score on the original scale, which is obtained from the same EMA study discussed in Section 4.3.2. PAM arousal is a -2 to 2 (excluding 0) score. We examine the same cohort of students on the same study period as described in Section 4.3.2. Over all observations, the distribution of arousal scores involves 16.6% for level -2, 27.7% for level -1, 12.6% for level 1, and 12% for level 2, while 31.1% of the observations are missing.

To implement model (4.8), we follow the procedure outlined above Proposition 4.5. We re-organize the original data into separate data sets $\{\boldsymbol{\mathcal{D}_j} : j = 1, \ldots, 3\}$, each of them containing the binary responses indicating whether the arousal scores are at level j or a higher level. Then, the proposed model is fitted to the three data sets in parallel.

The primary inference focus is on the change of arousal scores as the term progresses, which is depicted by the probability response curve of each response level. We display posterior point and interval estimates of $\mathbf{P_{j\tau}}$ (defined in (4.9)) in Figure 4.14. The probability of the highest arousal level drops dramatically as the term begins, indicating that the excitement of a new quarter may vanish within a week. The Boston marathon bombing slightly triggers higher probability for moderately low to low arousal level. There is a drop of the probability for moderately high to high arousal level after the Green Key festival and the Memorial Day holiday. The exams may have a significant impact on the arousal level. We observe peaks of arousal at the beginning of the final exam period, and also the middle of the term, which corresponding to the midterm exam period. Since the students are taking different courses, the midterm exam times vary, resulting in some curves with lead or lag peaks compared to the majority. This pattern is not clear in the analysis of binary arousal scores. Hence, examining the finer ordinal scale enables us to discover subtle changes of the students activation states.

4.5 Discussion

We have developed a novel Bayesian hierarchical model for analyzing longitudinal binary data. We approach the problem from a functional data analysis perspective, resulting in a method that is suitable for either regularly or irregularly spaced longitudinal data. The modeling approach achieves flexibility and

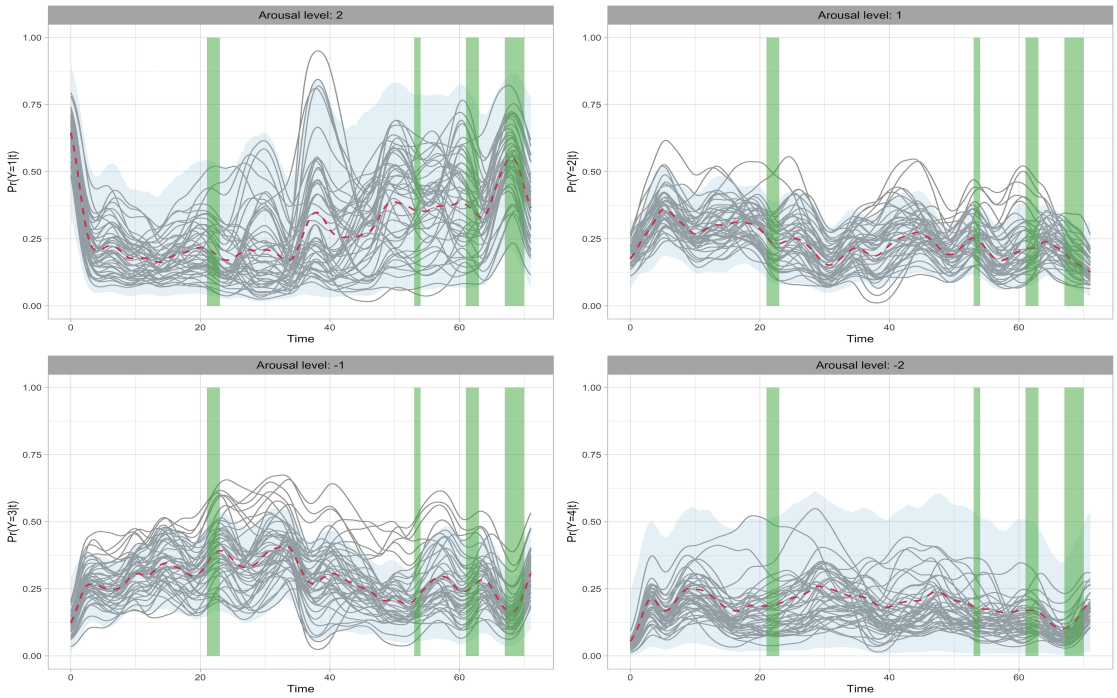


Figure 4.14: Four levels arousal score data. Posterior mean (dashed line) and 95% interval estimate (shaded region) of probability response curve for an out-of-sample subject. The posterior mean estimates for the probability response curves of in-sample subjects are given by the solid lines. The vertical shaded regions correspond to the four special time periods (see Section 4.3.2.1).

computational efficiency in full posterior inference. With regard to the former, the key model feature is the joint and nonparametric modeling of the mean and covariance structure. As illustrated by the data application, our approach enables interpretable inference with coherent uncertainty quantification, and provides improvement over the GLMM approach. The model formulation enables a natural extension to incorporate ordinal responses, which is accomplished by leveraging the continuation-ratio logits representation of the multinomial distribution. This representation leads to a factorization of the multinomial model into separate binomial models, on which the modeling approach for binary responses can be applied. The computational benefit is retained, since we can utilize parallel computing across response categories.

The proposed methodology for modeling longitudinal binary and ordinal responses can be elaborated in different directions. We have focused on stationary specifications for the hierarchical GP prior. Nonstationary model components can be incorporated through a time-varying mean function μ_0 and/or a nonstationary covariance kernel Ψ_ϕ. Moreover, longitudinal studies typically have predetermined covariates associated with each subject, or time-varying covariates corresponding to each observation. The predetermined covariates can be incorporated in the model through the prior placed on the mean function of the signal process. Using the functional linear model may be a possible strategy for the more challenging task of accounting for time-varying covariates.

In the EMA study example discussed in this chapter, the two response attributes can be modeled jointly. Although such an approach may encourage borrowing of information and improve uncertainty quantification, it also introduces a challenge, which is to account for the correlation between the bivariate responses. A potential solution is to induce correlation through shared hyperparameters. Nonetheless, we do not pursue this here, but rather choose to emphasize the practical utility of our flexible modeling framework through this real application.

Chapter 5

A Case Study: Estimating Maturity of Sheepshead Minnows

5.1 Introduction

We present a case study using data from a real longitudinal study in population biology. The purpose is to illustrate, with a concrete example, the benefit of a flexible and efficient modeling approach in answering relevant questions. The methodology presented here is also an extension to the model for longitudinal ordinal responses, discussed in Chapter 4, in terms of incorporating predetermined (categorical) covariates.

The specific data we study here is about the maturity of male sheepshead minnows in three states, Connecticut (CT), Maryland (MD), and South Carolina (SC) (data obtained courtesy of Dr. Steve Munch, NOAA, SWFSC, FED). The response variable is the discrete ordinal color stage, indicating maturity status ranging from juvenile to adult. For each fish, its maturity status is measured at eight equally spaced time points. Corresponding to the responses, three categorical

experiment conditions (parent temperature, offspring temperature, and exposure day) are treated as covariates. Of direct interest is estimating differences in trends in maturity across the treatment combinations.

Because the experiment conditions involve treatment of two consecutive generations, we are also interested in using the data set to test a theory about transgenerational plasticity (TGP). In our usage, TGP is manifest as a significant interaction between parent and offspring environments affecting offspring phenotype. The existing theory predicts that the magnitude of TGP in response to temperature on the Atlantic coast of the US, if exists, should decrease with increasing latitude. To provide evidence supporting the theory, we will estimate the TGPs for each population, and comparing the magnitude of them.

Before delving into analysis, we start with descriptive and exploratory analysis of the data. The data consist of complete records for $n = 319$ (63 from CT, 99 from MD, and 157 from SC) fish. The categorical treatments, parent temperature (PT, 26 or 32), offspring temperature (OT, 26 or 32), and exposure day (ED, 7, 30, or 45), naturally divide fish into 12 groups, such that the fish within a group shares the same treatment combination. The number of fish in each group ranges from 8 to 39. We use the tuple $(\mathrm{PT}, \mathrm{OT}, \mathrm{ED})$ to refer to the treatments hereinafter.

Male sheepshead minnows go through a sequence of five color stages as they approach maturity (Lee et al., 2017), resulting in the five-level ordinal response. To facilitate simple and interpretable model output, we start with a binary version of the response, representing either immature (first three ordinal levels) or mature (fourth and fifth ordinal levels). This is also aligned with the goal of examining TGP, where the primary focus is on differentiating between immature and mature. Additionally, because we are interested in discerning the effect of treatments on the transition from less mature to more mature categories, we also assess the finer

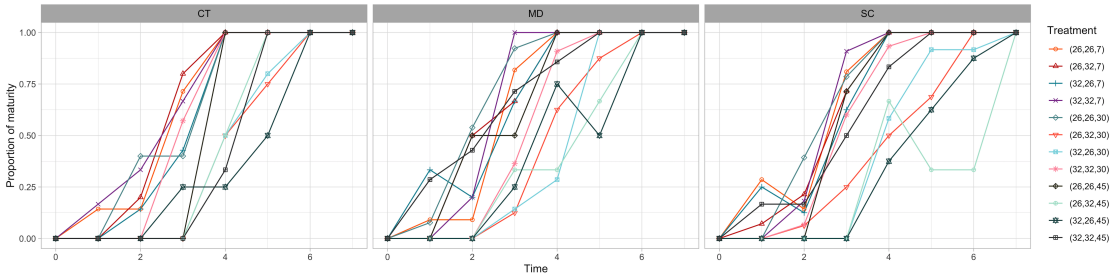


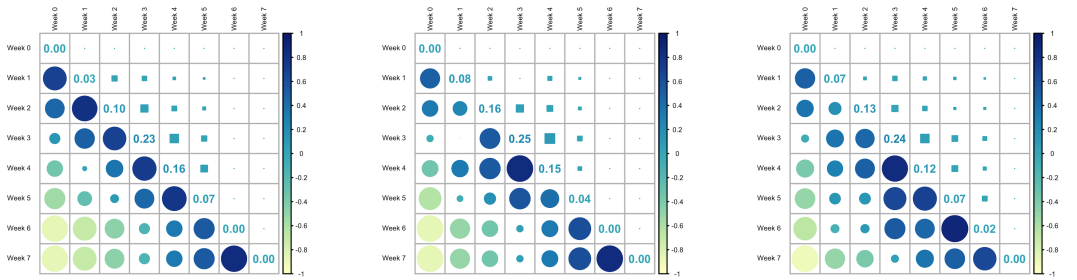
Figure 5.1: Fish maturity data. Proportion of mature fish over time for each treatment group.

scale ordinal responses.

For the binary version of the responses, we plot the proportion of mature fish for each population and treatment group in Figure 5.1. While an overall increasing pattern is observed, individual trajectories differ in terms of slope, monotonicity, and inflection points. It is also suspectable to assume the effect of covariates are additive, as evidenced by the intricate appearance of the curves. Therefore, a flexible model that permits nonstandard evolution of binary responses and imposes no specific pattern among the treatment groups may be more suitable here.

We further explore the correlations between the binary responses. Over the observed time grid, we empirically calculate the Pearson and the tetrachoric correlation coefficient for each pair of time and distance combinations for each population. Figure 5.2 displays the result. The plots all suggest a fast decreasing of correlation with time, through no obvious choice for modeling the temporal correlation. Shown on the diagonals of each panel in Figure 5.2 are the variances of the binary responses, which suggest nonstationarity.

Moreover, we seek to analyze how treatment affects transitions from lower to higher categories. As an empirical study, we calculate the transition proportion matrices based on the observed responses in each treatment group. Displayed in Figure 5.3a are the empirical proportion matrices under different treatment

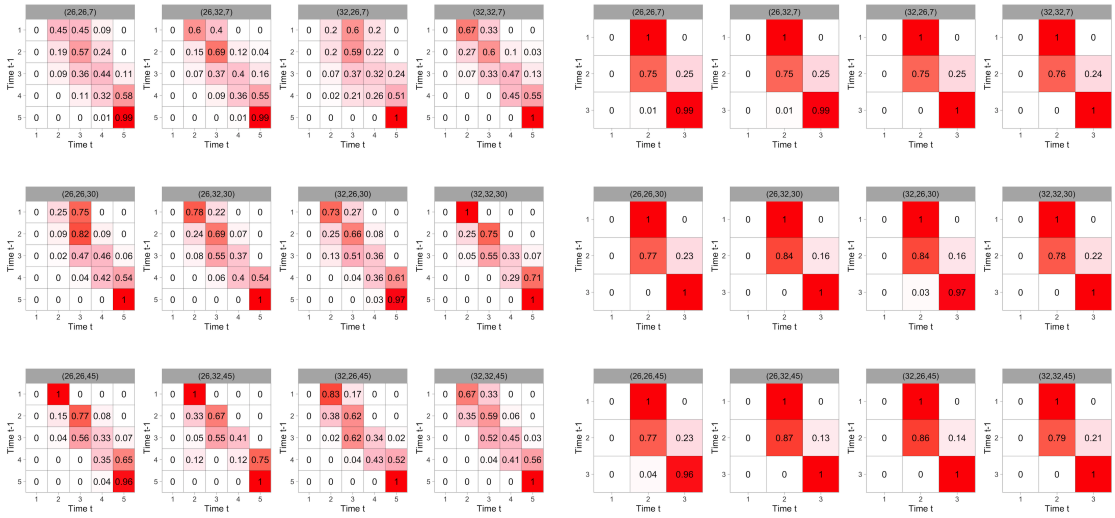


(a) CT Population. **(b)** MD Population. **(c)** SC Population.

Figure 5.2: Fish maturity data. Empirical estimate of the correlation coefficients between binary responses over time. In each panel, the upper triangle and the lower triangle indicate the Pearson and the tetrachoric correlation coefficient, respectively, while the numbers on the diagonal are the variances.

conditions, calculated with the original five-level ordinal maturity status. 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 juvenile (first ordinal level), adolescent (second, third, and fourth ordinal levels), and adult (last ordinal level). With the three-level ordinal responses, the transition proportion matrices based on the observed responses in each treatment group are shown in Figure 5.3b. From a biological point of view, fish are supposedly more mature at time t than at time $t-1$. This is also the pattern we observed from both figures. Though, with the three-level ordinal responses, it is more prominent that the treatment effect differs by category. Hence, a sequential model for ordinal data, which emphasizes on contrasting a category with categories from higher levels, is more suitable in the context of the problem.

The preliminary analyses suggest that it is a challenging problem. From a methodological perspective, the key is to balance between model flexibility and interpretability. We need a flexible model to accommodate the nonstandard evolution of responses over time and the complexities of treatment effects. Meanwhile,



(a) Original ordinal scale. **(b)** Collapsed ordinal scale.

Figure 5.3: Fish maturity data. Transition proportion matrix for fish in each treatment group. The i, j-th entry shows the observed proportion of transitions from stage i at time $t-1$ to stage j at time t. The treatment is specified by the title of each panel.

the rather sophisticated model structure should not preclude direct answers to relevant scientific questions. As a related concern, it can be challenging to identify key quantities of interest in complex models. For instance, while the effect of a covariate may be represented by the corresponding regression coefficient in a linear regression model, such a straightforward interpretation may not apply to more intricate models. We have to explore various options. Besides, the data exhibits an imbalance in subjects across treatment groups, with some groups having fewer subjects. In the proposed model, it is crucial to promote the pooling of information across subjects in disparate groups.

The objectives of this chapter are twofold. Firstly, we address the scientific relevant questions, specifically, to quantify the variations in maturity trends across treatment combinations, and to establish evidence indicating a decrease in the magnitude of TGP with rising latitude. Secondly, we extend the methodology intro-

duced in Chapter 4 to incorporate predetermined categorical covariates. Because the categorical covariates naturally partition subjects into groups, the proposed solution is to adopt the flexible model for temporal evolving binary or ordinal responses as the building block. The blocks for each group are then interconnected through priors placed on the model parameters.

Statistical tools have been utilized in studying the effect of some covariates to fish maturity. The majority of them focus on time-varying covariates, such as the length of the fish, and use parametric logistic regression or some variant (see e.g. Bobko and Berkeley, 2004). Under the linear mixed effects framework, Munch et al. (2021) examined the treatment effects to fish maturity, using age at maturation as the response variable. Besides, a substantial body of literature is dedicated to longitudinal binary/ordinal regression, aimed at various applications across a wide spectrum (see e.g. Barcella et al., 2018, for a Bayesian nonparametric model with application in a clinical study). A more relevant work is DeYoreo and Kottas (2018c). Motivated also by an application in fisheries research that involves longitudinal ordinal responses, their model builds on a dependent Dirichlet process prior for time-dependent mixing distributions to capture dynamically evolving relationships between age, length and maturity. Because their approach accounts for the joint stochastic mechanism of covariates and responses, it is not directly applicable in our scenario, where the covariates are deterministic.

From an alternative perspective, models can be developed by postulating a Markov chain structure. Specifically, these models consists of flexible, though parsimonious, formulation for the marginal probability distribution and the transition probability matrix (see e.g. Bartolucci et al., 2009). However, these models usually focus on estimating probability response curves at discrete time grid. In this application, inferring probability response curves on a continuous scale is more

beneficial for our objectives.

The reminder of the chapter is organized as follows. We formulate the model for analyzing the fish maturity data set in Section 5.2, with emphasis placed on exploring potential modeling options. These modeling options are formally compared using the real data, and the best model are used to conduct inference to answer relevant scientific questions. These analyses are presented in Section 5.3. Finally, we conclude in Section 5.4 with some comments.

5.2 Methodology

We start from formulating the model with binary responses. Let n denote the total number of fish. Based on their treatment conditions, the fish can be grouped into $G = 12$ groups. Let $\mathbf{Y}_{gi}$ denote the observed binary maturity status sequence at grid $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_T)^\top$ of the i-th subject in g-th group for $g = 1, \cdots, G$, and $i = 1, \cdots, n_g$, where $\sum_{g=1}^{G} n_g = n$.

At the observed data level, we assume the binomial model, that is,

$$Y_{git} \mid Z_{git}, \boldsymbol{\epsilon}_{git} \overset{ind.}{\sim} Bin(1, \varphi(Z_{git} + \epsilon_{git})),\ t = 1, \ldots, T,\ i = 1, \ldots, n_g,\ g = 1, \ldots, G.$$

Here $\varphi(\cdot)$ denotes the standard logistic function. The error term $\epsilon_{git} \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$. We assume Z_{git} is the evaluation of a continuous signal process $Z_{gi}(\tau)$ at time t. We adopt a functional data analysis perspective and focus on modeling the signal process $Z_{gi}(\tau)$, which is the key in our model formulation.

A typical flexible model for the continuous signal process $Z_{gi}(\tau)$ is Gaussian process (GP). Specifically, we assume

$$Z_{gi}(\tau) | \mu_g(\tau), \Sigma_g(\tau, \tau) \overset{i.i.d.}{\sim} GP(\mu_g(\tau), \Sigma_g(\tau, \tau)), \quad i = 1, \ldots, n_g,\ g = 1, \ldots, G.$$

That is, we assume that the signal processes for the subjects within each group are independent realizations from a GP with the group-specific mean function $\mu_g(\cdot)$, and covariance kernel $\Sigma_g(\cdot,\cdot)$.

The hierarchical nonparametric prior for the mean and covariance function of a GP, developed in Section 4.2, is adopt here as the joint prior for μ_g and Σ_g, i.e.,

$$\begin{aligned} \mu_g(\tau)|\Sigma_g(\tau,\tau),\mu_{0g}(\tau),\nu_g &\overset{ind.}{\sim} GP(\mu_{0g}(\tau),(\nu_g-3)\Sigma_g(\tau,\tau)), \\ \Sigma_g(\tau,\tau)|\nu_g,\Psi_{\sigma_g^2,\rho_g}(\tau,\tau) &\overset{ind.}{\sim} IWP(\nu_g,\Psi_{\sigma_g^2,\rho_g}(\tau,\tau)). \end{aligned} \tag{5.1}$$

We further assume $\mu_{0g}(\tau) \equiv \mu_{0g}$, that is, a constant mean function over time. To encourage smooth estimation of the signal function $Z_{gi}(\tau)$, we specify the covariance kernel of the Inverse-Wishart process (IWP) as Matérn covariance kernel with smoothness 5/2,

$$\Psi_{\sigma_g^2,\rho_g}(\tau,\tau') = \sigma_g^2(1+\frac{\sqrt{5}|\tau-\tau'|}{\rho_g}+\frac{5|\tau-\tau'|^2}{3\rho_g^2})\exp(-\frac{\sqrt{5}|\tau-\tau'|}{\rho_g}).$$

To emphasize the hyperparameters of the prior and to simplify the notation, we denote the aforementioned joint prior for the mean and covariance function as $JP(\mu_{0g},\sigma_g^2,\rho_g,\nu_{0g})$.

This prior specification yields twofold advantages. First, despite potential misspecification of a specific mean and covariance structure, the flexibility inherent in the nonparametric prior enables it to capture both the trend and covariance of the process. Second, we can show marginally $Z_{gi}(\tau)$s are realizations from student-t process (TP), i.e,

$$Z_{gi}(\tau) \mid \mu_{0g},\nu_g,\Psi_{\sigma_g^2,\rho_g}(\tau,\tau) \overset{i.i.d.}{\sim} TP(\nu_g,\mu_{0g}(\tau),\Psi_{\sigma_g^2,\rho_g}(\tau,\tau)). \tag{5.2}$$

Consequently, we can obtain interpolation of $Z_{gi}(\tau)$ on a more denser time grid in

an efficient way.

Our model formulation facilitate direct answers to the relevant scientific questions. The model assumes that the observed binary maturity status is related to a continuous signal process, which is a realization from a TP. Specifically, we use $\tilde{Z}_g$ to denote $TP(\nu_g, \mu_{0g}(\tau), \Psi_{\sigma_g^2,\rho_g}(\tau,\tau))$. For a generic new subject in group g, from (5.2), its signal process, denoted by $Z_g^*(\tau)$, satisfies $Z_g^*(\tau) \mid \tilde{Z}_g \sim \tilde{Z}_g$, and $\varphi(Z_g^*(\tau))$ is the probability of maturity. We term $\varphi(Z_g^*(\tau))$ the maturity profile of treatment corresponding to group g, because it characterizes the evolution dynamics of maturity. We rely on $\varphi(Z_g^*(\tau))$ to assess the differences in trends in maturity across treatment conditions.

We complete the model formulation with a discussion on specifying priors for the hyperparameters $\{\mu_{0g}, \sigma_g^2, \rho_g, \nu_g : g = 1, \ldots, G\}$. Here, we focus on the induced dependence between groups. Note in one of the extreme case, if we assume $\mu_{0g} \overset{ind.}{\sim} \pi(\mu_{0g})$, $\sigma_g^2 \overset{ind.}{\sim} \pi(\sigma_g^2)$, $\rho_g \overset{ind.}{\sim} \pi(\rho_g)$, and $\nu_g \overset{ind.}{\sim} \pi(\nu_g)$, then effectively we are fitting models separately to each group of data, with no borrowing of information across groups. It should not be encouraged in this case. Another extreme is to assume $\mu_{0g} \equiv \mu_0 \sim \pi(\mu_0)$, $\sigma_g^2 \equiv \sigma^2 \sim \pi(\sigma^2)$, $\rho_g \equiv \rho \sim \pi(\rho)$, and $\nu_g \equiv \nu \sim \pi(\nu)$. That is, despite the corresponding group, the individual signal processes $Z_{gi}(\tau)$ are all realizations from the same TP. It is not a good choice because the treatment effects in maturity is ignored. Therefore, a valid option should lie in between these two extremes. We explore various options.

Prior for μ_{0g}:

Option (a): consider the vector of indicators for each group, denoted by $\mathbf{x}_g = (1, 1_{(g=1)}, \ldots, 1_{(g=G)})^\top$. We assume $\mu_{0g} = \mathbf{x}_g^\top \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_G)$. Additionally, $\alpha_0 \sim N(a_0, b_0)$, α_1 is fixed at 0 for identifiability, and $\alpha_g \overset{i.i.d.}{\sim} Lap(0, \sigma_\alpha^2/\lambda_\alpha)$, for $g = 2, \ldots, G$.

Option (b): consider the vector of indicators for each treatment condition, denoted by $\mathbf{x}_g = (1, 1_{(PT=32)}, 1_{(OT=32)}, 1_{(ED=30)}, 1_{(ED=45)})$. We assume $\mu_{0g} = \mathbf{x}_g^\top \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_4)$, and $\alpha_g \overset{i.i.d.}{\sim} Lap(0, \sigma_\alpha^2/\lambda_\alpha)$, for $g = 0, \ldots, 4$.

Prior for σ_g^2:

Option (a): we assume the scale parameter σ_g^2 of groups that have the same exposure days are conditionally i.i.d., i.e., $\sigma_g^2 \mid \theta_s \overset{i.i.d.}{\sim} Gamma(a_\sigma, a_\sigma \theta_s^{-1})$, and $\theta_s \overset{i.i.d.}{\sim} IG(a_\theta, b_\theta)$, for $s = 1, 2, 3$.

Option (b): we assume that conditioning on a common parameter θ, the scale parameters σ_g^2 are i.i.d., i.e., $\sigma_g^2 \mid \theta \sim Gamma(a_\sigma, a_\sigma \theta^{-1})$, and $\theta \sim IG(a_\theta, b_\theta)$.

Prior for ρ_g:

Option (a): we assume the smoothness parameters for each group are i.i.d., i.e., $\rho_g \overset{i.i.d.}{\sim} Unif(a_\rho, b_\rho)$.

Option (b): we assume a common smoothness parameter shared by groups, i.e., $\rho_g \equiv \rho \sim Unif(a_\rho, b_\rho)$.

Prior for ν_g:

Option (a): we assume the degrees of freedom parameters for each group are i.i.d., i.e., $\nu_g \overset{i.i.d.}{\sim} Unif(a_\nu, b_\nu)$.

Option (b): we assume a common degrees of freedom parameter shared by groups, i.e., $\nu_g \equiv \nu \sim Unif(a_\nu, b_\nu)$.

Note that for each parameter, option (a) yields a weaker dependence comparing to option (b). We consider these options for computation efficiency. For all the scenarios, posterior inference can be conducted through a Gibbs sampler modified from the algorithm presented in Section 4.2.3.

We perform a formal model comparison to choose the best one among these options. We start from priors that induce less dependence among the groups, and moving towards options that induce more dependence and better performance under the comparison criteria. For the class of models that assume the continuous signal process follows a TP, we consider the following specific options for the priors:

- Model $\mathcal{M}_1$: option (b) for μ_{0g}, while option (a) for σ_g^2, ρ_g and ν_g;
- Model $\mathcal{M}_2$: option (a) for μ_{0g}, σ_g^2, and ν_g, while option (b) for ρ_g;
- Model $\mathcal{M}_3$: option (a) for μ_{0g} and σ_g^2, while option (b) for ρ_g and ν_g;
- Model $\mathcal{M}_4$: option (a) for μ_{0g}, while option (b) for σ_g^2, ρ_g and ν_g.

As an alternative modeling approach, we consider a simplified version of the proposed model. That is, instead of placing a IWP prior on the covariance kernel, we assume it has the deterministic structure of $\Psi_{\sigma_g^2,\rho_g}$. The structure for the mean function is unchanged. This alternative approach models the signal process as realizations from a GP. Under this class of models, we consider two specific choices which differ by the priors on the hyperparameters:

- Model $\mathcal{M}_5$: option (a) for μ_{0g}, while option (b) for σ_g^2 and ρ_g;
- Model $\mathcal{M}_6$: option (a) for μ_{0g} and ρ_g, while option (b) for σ_g^2.

We consider three model comparison criteria, namely the posterior predictive loss (PPL), the Watanabe-Akaike information criterion (WAIC), and the Log Pseudo Marginal Likelihood (LPML) with conditional predictive ordinate on deviance scale. Note that the PPL and WAIC contain a penalty term for model complexity, and LPML implicitly punishes complex model because it is based on leave-one-out cross-validation posterior predictive probability of the data. For all

Table 5.1: Fish maturity data. Summary of comparison between two classes of models with different prior specifications using four different criteria. The values in bold correspond to the model favored by the particular criterion.

	Model	PPL		WAIC	LPML
		$G(\mathcal{M})$	$P(\mathcal{M})$		
TP models	$\mathcal{M}_1$	146.26	234.54	1058.43	1130.85
	$\mathcal{M}_2$	121.54	197.52	896.01	1042.71
	$\mathcal{M}_3$	109.56	**158.44**	790.46	995.01
	$\mathcal{M}_4$	**104.93**	163.85	**773.62**	**961.79**
GP models	$\mathcal{M}_5$	177.15	335.04	1337.66	1381.95
	$\mathcal{M}_6$	159.14	244.93	1121.11	1145.70

three criteria, we prefer the model with smaller value. The result, presented in Table 5.1, favors the model $\mathcal{M}_4$ in the class of TP models. The results presented hereinafter are based on this selected model.

To handle the three-level ordinal responses, we leverage the continuation-ratio logits structure to extend the selected model for binary responses. In particular, we use $\varphi(Z^*_{gj}(\tau))$ to model the conditional probability of observing maturity level j given that the maturity level is higher than $j-1$, for $j = 1, 2$. We assume a full factorization across response categories, such that the model can be implemented efficiently through parallelization. We refer to Section 4.4.1 for the technical details.

5.3 Results

5.3.1 Binary Responses

We first examine in Figure 5.4 the probability response curves, defined as the probability of maturity as a function of time. In all the groups, individual probability response curve exhibits a smooth increasing pattern. The width of uncertainty bands, as expected, is related to the number of subjects in the corresponding group.

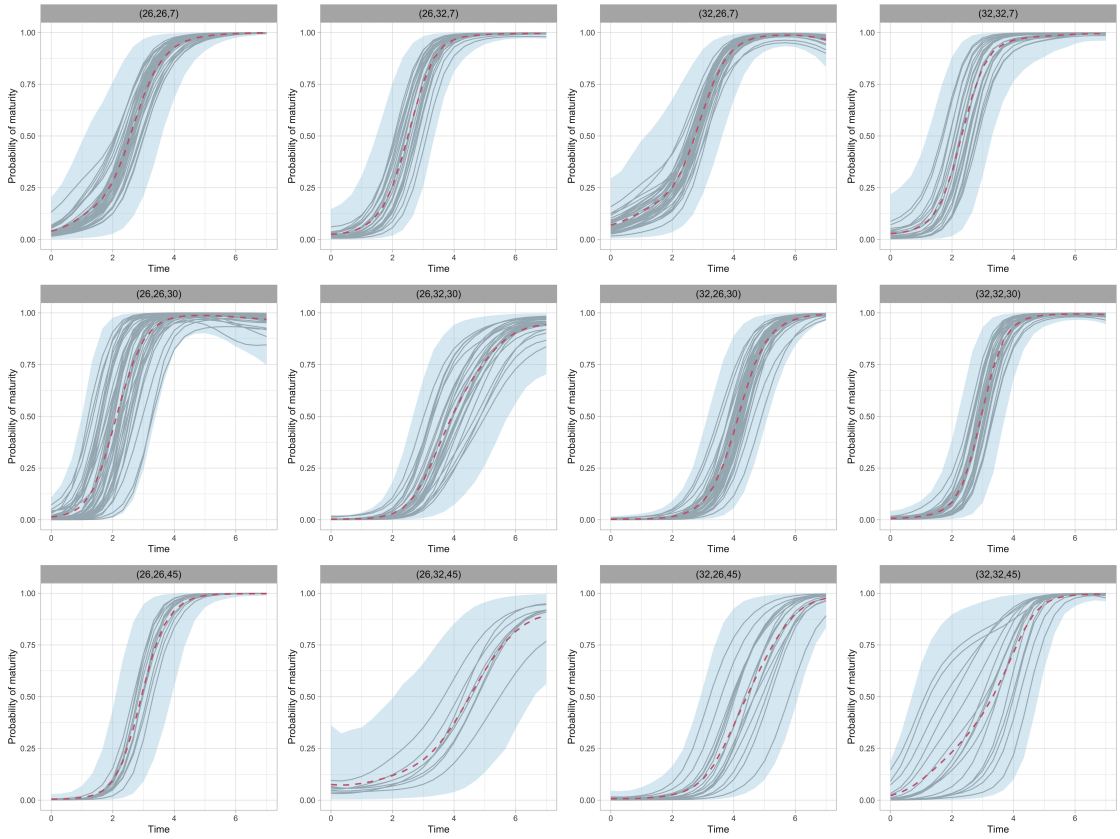


Figure 5.4: Fish maturity data. Posterior predictive mean (dashed line) and 95% interval estimate (shaded region) of the probability response curve for an out-of-sample subject. The posterior mean estimates of probability response curves for in-sample subjects are given by the solid lines. The corresponding treatment is specified as the title of each panel.

Nonetheless, since the model permits borrowing of information across the groups, even in groups with fewer subjects, the uncertainty bands are still plausible. The maturity rate differs by treatment groups. We will elaborate the differences more because it quantifies the effect of the treatment on the trends in maturity.

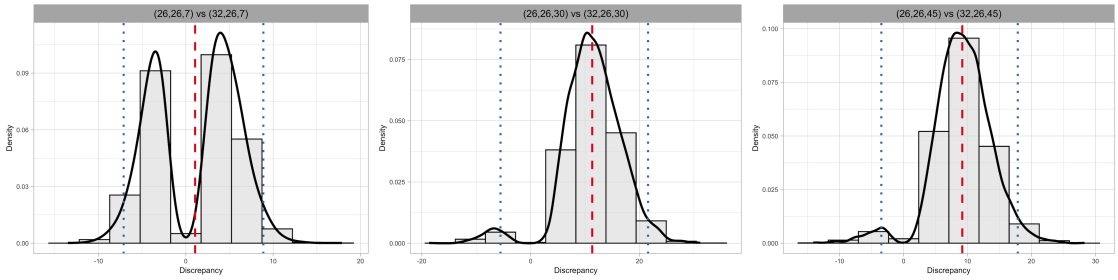
The maturity profile for group g is represented by $\varphi(Z_g(\tau))$. Since the standard logistic function $\varphi(\cdot)$ is monotonic increasing, we propose to use a measurement of discrepancy between $Z_g(\tau)$ and $Z_{g'}(\tau)$ to quantify the relative effect of the treatment for group g to the treatment for group g' in maturity. We care about

both the magnitude of the relative effect and the direction of it. Let $\boldsymbol{\tau}^*$ denote a time grid on $[0, T]$ with $|\boldsymbol{\tau}^*|$ intervals. The proposed measurement of discrepancy is defined as

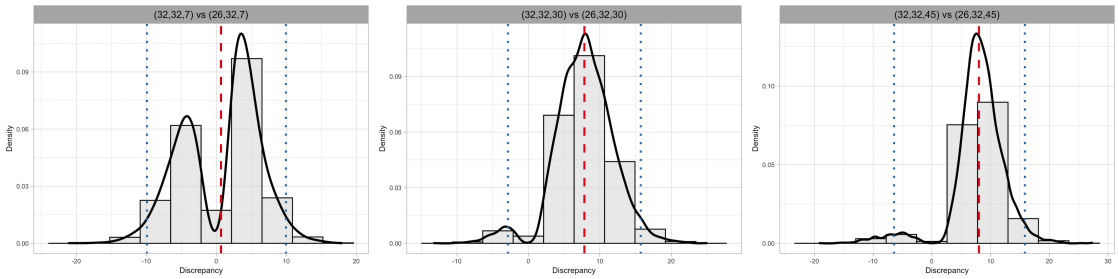
$$d^*(Z_g(\tau), Z_{g'}(\tau)) = \text{sign}(Z_g(\tau), Z_{g'}(\tau)) \times d(Z_g(\tau), Z_{g'}(\tau)),$$

where $d(Z_g(\tau), Z_{g'}(\tau)) = (\frac{T}{|\boldsymbol{\tau}^*|} \sum_{\tau \in \boldsymbol{\tau}^*} \|Z_g(\tau) - Z_{g'(\tau)}\|^2)^{1/2}$ measures the magnitude of the discrepancy. The sign function takes value 1 if $Z_g(\tau)$ reaches the predetermined threshold ($\varphi^{-1}(0.9)$) earlier than $Z_{g'}(\tau)$, and is -1 otherwise. Introducing the sign function enables us to distinguish the relative effects of treatments on fish maturation, such that the treatment leading to an earlier maturity has a positive relative effect, and vice versa.

In Figure 5.5, we show selected pairwise comparison of relative effect on maturity trend. Salinas and Munch (2012) showed that sheepshead minnows from Florida exhibit thermal TGP such that the fastest growing offspring at a given temperature are those whose parents were held at the same temperature. Inspired by their conclusion, we explore pairwise comparison for groups with the same offspring temperature, controlling the exposure day. We contrast the group with the same parent temperature to the group with the different temperature, such that a positive discrepancy meaning the group with the same parent and offspring temperature mature faster. For 30 or 45 exposure days, displayed in the middle and right panel of Figure 5.5, the distributions of discrepancy based on posterior predictive sample of $Z^*_g(\tau)$ and $Z_{g'}(\tau)$ are mostly positive, suggesting that the group with the same parent and offspring temperature mature faster than their counterpart. On the other hand, when exposure day is 7, the distributions of discrepancy exhibit bimodality, with a positive mode and a negative mode. It indicates that the thermal TGP is not significant for brief exposure periods.



(a) Pairwise discrepancy between groups with offspring temperature 26.

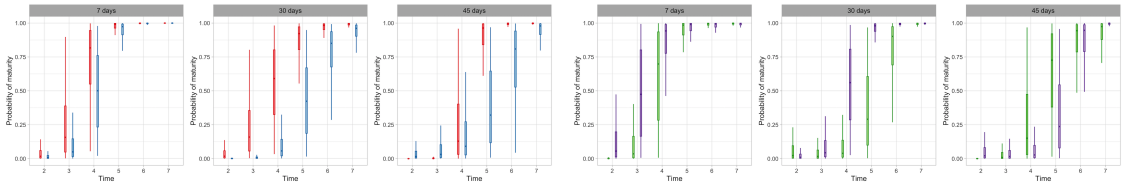


(b) Pairwise discrepancy between groups with offspring temperature 32.

Figure 5.5: Fish maturity data. Posterior predictive distribution for $d^*(Z_g^*(\tau), Z_{g'}^*(\tau))$. In each panel, the black solid line is the kernel density estimation. The red dashed line indicates the mean, and the blue dotted lines mark the 95% interval.

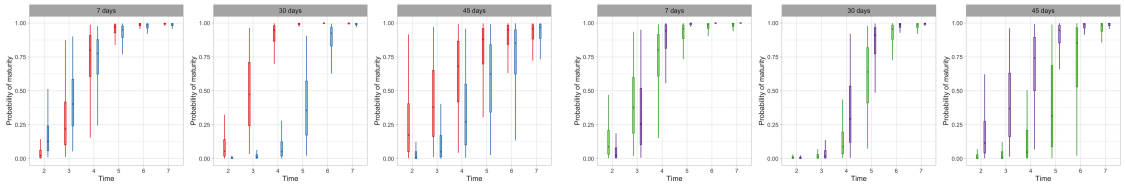
Next, we explore the relationship between the latitude corresponding to each population and the magnitude of TGP. To distinguish between populations, we fit the proposed model separately to the three populations, and obtain posterior samples of $Z_{gl}^*(\tau)$, where $l = 1, 2, 3$, representing the three population. Munch et al. (2021) defines the TGP as the regression coefficient corresponding to the interaction term of parent and offspring temperature in a linear mixed effects model. With a more advanced model, there is no clear definition of TGP. We examine two types of inferences.

Firstly, we contrast $\varphi(Z_{gl}^*(\tau))$ and $\varphi(Z_{g'l}(\tau))$ at observed time grid. We focus on comparing groups with different parent temperature, controlling the offspring temperature, exposure days, and population. The results, displayed in Figure 5.6, indicate that TGP is also manifested as the fastest growing offspring at a



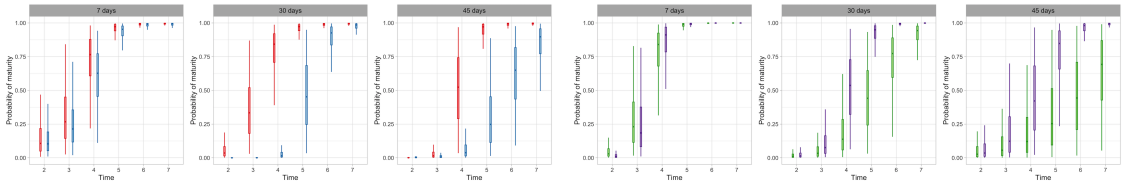
(a) CT, with OT = 26.

(b) CT, with OT = 32.



(c) MD, with OT = 26.

(d) MD, with OT = 32.



(e) SC, with OT = 26.

(f) SC, with OT = 32.

Figure 5.6: Fish maturity data. Posterior distribution of $\varphi(Z_g^*(\tau))$ for fish of the specified treatment and population at the chosen time point. The red, blue, green, and purple histograms correspond to treatment $(\text{PT} = 26, \text{OT} = 26)$, $(\text{PT} = 32, \text{OT} = 26)$, $(\text{PT} = 26, \text{OT} = 32)$, and $(\text{PT} = 32, \text{OT} = 32)$, respectively.

given temperature are those with the same parent temperature. The magnitude of TGP, in general, is not significant with 7 exposure days, but significant with longer exposure period, and also appear to be decrease with increasing latitude.

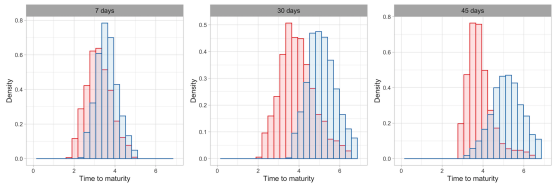
In Munch et al. (2021), they implemented a linear mixed effects model with the time to maturity as the response variable. Because maturity status is only observed on a discrete time grid, they use linear interpolation to obtain a continuous estimate of maturation time. Inspired by this idea, and to exploit the advantage of our model in terms of posterior prediction at any time grid, we consider solving the inverse problem, i.e., finding the smallest time $t \in \boldsymbol{\tau}^*$ such that probability of maturity $\varphi(Z_{gl}^*(\tau))$ exceeds a prespecified threshold. (We use 0.9 in the analysis

presented later). This time, denoted by τ^*_{gl}, is an implicit signal for the maturity age under our model. Holding the same population, exposure days, and offspring temperature, we compare posterior distributions of τ^*_{gl} for groups with the same or the different parent temperature. The results are presented in Figure 5.7.

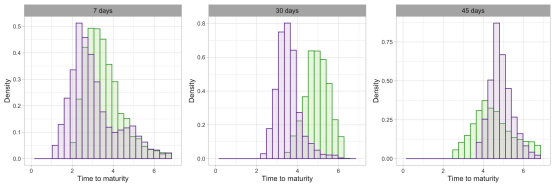
These histograms confirm that, for fish in CT, MD, and SC, we also observe thermal TGP in such a manner that the fastest growing offspring at a given temperature are those whose parents were held at the same temperature. The magnitude of the TGP can be viewed as the separation of the two histograms in the same panel. In general, for longer exposure period, the separation seems to be significant, and its magnitude decreases with increasing latitude. These results provide graphical evidence to support the theory about TGP and latitude.

5.3.2 Three-level Ordinal Responses

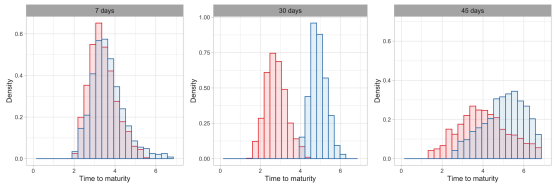
Moreover, we analyze the version of data with the three-level ordinal response. Here, with a finer ordinal response scale, the primary goal is to assess the TGP in relation to the different stages in the transition from immaturity to maturity. To evaluate TGP during the transition from juvenile to adolescent, we approach this as an inverse problem. Specifically, leveraging the posterior samples of the probability response curves estimated on a high-resolution time grid, we determine the earliest time point (referred to as"time to adolescen") at which the probability of reaching adolescence is 0.99. This high threshold was set because some treatment groups began with a high probability of being classified as adolescent. Even with this relatively extreme threshold probability, the TGP is nearly negligible, as evidenced by the distribution of "time to adolescent", depicted in Figure 5.8. Overall, the two histograms in each panel display the expected pattern; however, their separation is not significant.



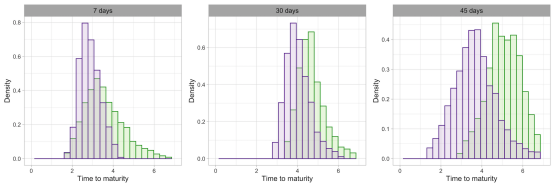
(a) CT, with OT = 26.



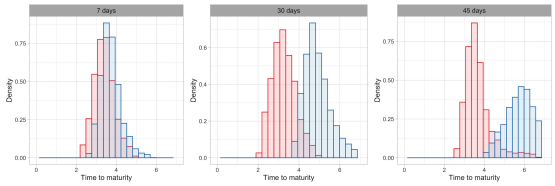
(b) CT, with OT = 32.



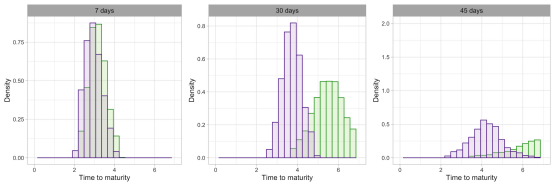
(c) MD, with OT = 26.



(d) MD, with OT = 32.



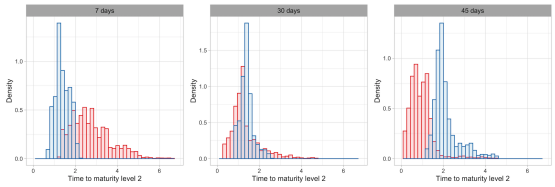
(e) SC, with OT = 26.



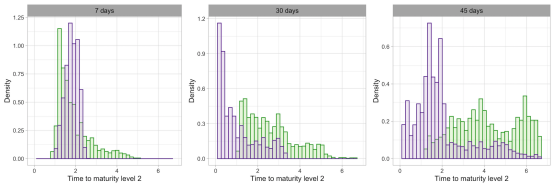
(f) SC, with OT = 32.

Figure 5.7: Fish maturity data. Posterior empirical distribution for the time of maturity for fish of the specified treatment and population. The red, blue, green, and purple histograms correspond to treatment $(\text{PT} = 26, \text{OT} = 26)$, $(\text{PT} = 32, \text{OT} = 26)$, $(\text{PT} = 26, \text{OT} = 32)$, and $(\text{PT} = 32, \text{OT} = 32)$, respectively.

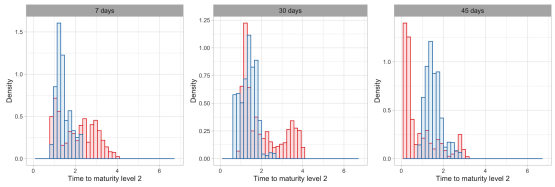
Additionally, we perform a similar analysis to evaluate TGP during the transition from adolescence to adult. We define "time to adul" as the earliest time point at which the probability of being classified as an adult reaches 0.5. The distributions of "time to adul" for different treatment groups and locations are presented in Figure 5.9. Here, even with a more conservative threshold, thermal TGP appears to be prominent. At each location, TGP is more pronounced with longer exposure durations, while it is not significant with a 7-day exposure. Furthermore, across different locations, the magnitude of TGP, as manifested by the separation of the two histograms in each panel, tends to decrease with increasing latitude. In conclusion, these results illustrate that thermal TGP influences the maturation of



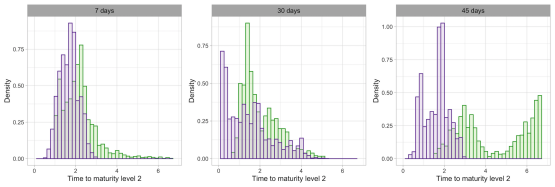
(a) CT, with OT = 26.



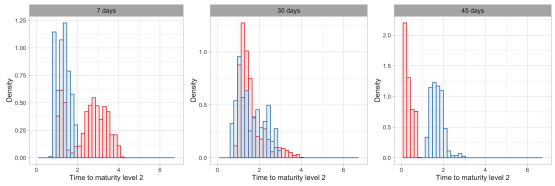
(b) CT, with OT = 32.



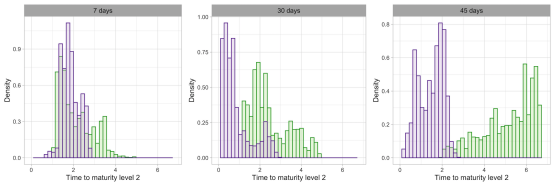
(c) MD, with OT = 26.



(d) MD, with OT = 32.



(e) SC, with OT = 26.



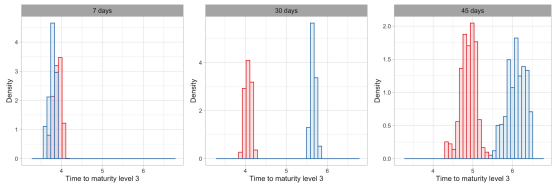
(f) SC, with OT = 32.

Figure 5.8: Fish maturity data with three-level ordinal responses. Posterior empirical distribution for the time of reaching maturity level 2 for fish of the specified treatment and population. The red, blue, green, and purple histograms correspond to treatment $(\text{PT} = 26, \text{OT} = 26)$, $(\text{PT} = 32, \text{OT} = 26)$, $(\text{PT} = 26, \text{OT} = 32)$, and $(\text{PT} = 32, \text{OT} = 32)$, respectively.

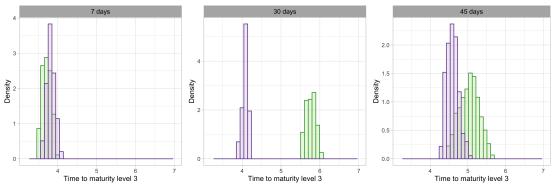
fish from adolescence to adult, but it is less impactful during the transition from juvenile to adolescence. Reaching this conclusion highlight the practical utility of analyzing the three-level ordinal responses with a flexible sequential model.

5.4 Comments

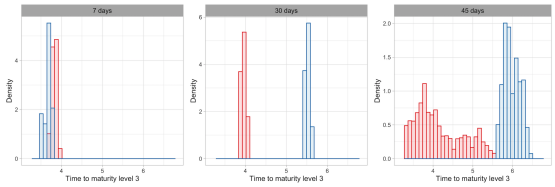
With a case study from fishery science, we illustrate the methodology for dynamic evolution of binary and ordinal responses, developed in Chapter 4, as a powerful building block for model that takes into account the effect of predetermined



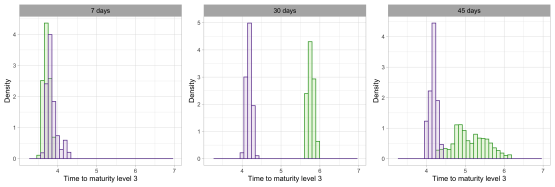
(a) CT, with OT = 26.



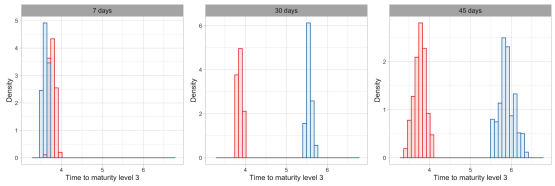
(b) CT, with OT = 32.



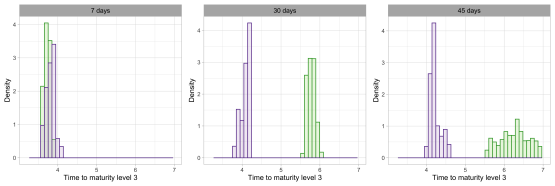
(c) MD, with OT = 26.



(d) MD, with OT = 32.



(e) SC, with OT = 26.



(f) SC, with OT = 32.

Figure 5.9: Fish maturity data with three-level ordinal responses. Posterior empirical distribution for the time of reaching maturity level 3 for fish of the specified treatment and population. The red, blue, green, and purple histograms correspond to treatment $(\text{PT} = 26, \text{OT} = 26)$, $(\text{PT} = 32, \text{OT} = 26)$, $(\text{PT} = 26, \text{OT} = 32)$, and $(\text{PT} = 32, \text{OT} = 32)$, respectively.

experiment conditions. However, we do not enforce nonstationarity in the mean and covariance structure of the signal process, which is arguably a convincing assumption, both from the exploratory data analysis and from a biology perspective. The primary concern here is implementation efficiency. Moreover, lacking sufficient prior information to account for nonstationarity in a structural way, we are concerned that the potential benefits may not be enough to compensate the extra costs on implementation. Nonetheless, we perform the first ever formal statistical analysis on this data, and the inferences we obtained generally agree with what is expected to be true biologically and is considered to be plausible.

The flexible nature of the model offers a model-based approach to contrast the effect of treatment conditions on maturity, which is not restricted to any particular form. This approach is more flexible comparing to classic approaches, such as the linear mixed effects model, under which the effect of a treatment usually has a parametric form. However, the flexibility induces a challenge, which is to find appropriate metrics to quantify the effects. In Section 5.3, we explore various options, which are by no means an exhaustive set. The metrics are also specific to the questions of interest for this specification, and should differ case-by-case. Although one may question the general applicability of the proposed flexible model comparing to the traditional parametric approaches, we see this as an asset of flexible models. For complex real problems, it is expected that statisticians collaborating with domain experts to determine the inferences to be derived from the model.

Chapter 6

Conclusions

Exploiting the theoretical advantages of Bayesian nonparametric models, we have developed a suite of statistical models to solve methodologically and practically relevant ordinal regression problems. Leveraging the sequential treatment of ordinal responses, induced by the continuation-ratio factorization, we circumvent the limitations of parametric models by adopting well-crafted nonparametric priors, leading to flexibility in ordinal response distribution, covariate-response relationship, and dependence among clustered responses. The proposed models seek to strike a good balance between model complexity and implementation difficulty. With regard to the latter, the key feature is the efficient Gibbs sampling algorithms for all proposed models.

To conclude this dissertation, we discuss some implications and possible extensions of the presented work. The feature of the general model in Section 2.2 of having effectively the same structure for the mixture weights and atoms is both theoretically appealing and practically powerful. It emphasizes the structural similarity between nonparametric priors for discrete distributions and models for categorical data, linking these two fields that have been parallelly explored in the literature. We are excited about the great potential of leveraging this similarity to

boost new nonparametric mixture models for categorical data analysis.

The models discussed in Chapter 2 and Chapter 4 shed light on extending one another. Seeking more flexible modeling of ordinal regression relationships in a cross-sectional setting, in lieu of a linear function, we can incorporate covariate effects through a Gaussian process on the covariate space. A practical concern is that the resulting model may break the balance between flexibility and efficiency. This can be addressed by using a more regularized weights structure, such as the geometric weights prior (Mena et al., 2011). Besides, an alternative path for developing longitudinal ordinal regression models is built on flexible models for the cross-sectional setting, extending them with a hyper-model for evolving temporal dynamics.

In the big data era, many fields have witnessed technological advancements resulting in snowballing of large-scale data. Although we have proposed efficient posterior simulation methods scalable with ordinal levels, it remains a challenging task to scale up inference in the presence of massive data sets within the MCMC realm. The most popular alternative to MCMC involves variational inference (VI) algorithms. A key advantage of the proposed methods is that they fall within the category of conditionally conjugate models, for which there exists a closed-form coordinate ascent VI algorithm (Blei et al., 2017). We notice the VI algorithms for some special cases of our models in Linderman et al. (2015) and Rigon and Durante (2021). Moreover, deriving VI algorithms for the proposed models will facilitate embedding them in a more complex ordinal regression settings.

This dissertation provides a general toolbox for ordinal regression, comprising various models tailored for specific problem settings, united under a sequential treatment for the ordinal responses. We augment model flexibility by nonparametric priors, employing considerable effort to guarantee a parsimonious model

formulation and efficient model implementation. It was a primary objective outlined in the introduction that the methodology developed herein would lead to new avenues of exploration in other scientific fields. Our impetus toward structured model specification and implementable computation techniques, coupled with the development of publicly available software (e.g. R packages), will contribute to achieving this goal.

Appendix A

Proofs

A.1 Properties of Models for Cross-sectional Ordinal Regression

A.1.1 Proof of Proposition 2.2

Proof. Under the augmented model, we have

$$\Pr(\mathbf{Y} = j, \boldsymbol{\mathcal{Z}} \mid \boldsymbol{\theta}) = \Pr(\mathbf{Y} = j \mid \boldsymbol{\mathcal{Z}})\, f(\boldsymbol{\mathcal{Z}} \mid \boldsymbol{\theta}) = \mathbf{1}(\boldsymbol{\mathcal{Z}} \in \mathcal{R}_j) \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_j).$$

Integrating out $\boldsymbol{\mathcal{Z}}$, we obtain

$$\begin{aligned}
\Pr(\mathbf{Y} = j \mid \boldsymbol{\theta}) &= \int \mathbf{1}(\boldsymbol{\mathcal{Z}} \in \mathcal{R}_j) \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_j)\, \mathrm{d}\boldsymbol{\mathcal{Z}} = \int_{\mathcal{R}_j} \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_j)\, \mathrm{d}\boldsymbol{\mathcal{Z}} \\
&= \left(\int_0^{\infty} \mathfrak{L}(\mathcal{Z}_j \mid \theta_j)\, \mathrm{d}\mathcal{Z}_j \right) \prod_{k=1}^{j-1} \left(\int_{-\infty}^{0} \mathfrak{L}(\mathcal{Z}_k \mid \theta_k)\, \mathrm{d}\mathcal{Z}_k \right) \\
&= \varphi(\theta_j) \prod_{k=1}^{j-1} \{1 - \varphi(\theta_k)\},
\end{aligned}$$

for $j = 2, \ldots, C-1$. Similarly, $\Pr(\mathbf{Y} = 1 \mid \boldsymbol{\theta}) = \varphi(\theta_1)$, and $\Pr(\mathbf{Y} = C \mid \boldsymbol{\theta}) = \prod_{k=1}^{C-1}\{1 - \varphi(\theta_k)\}$. Therefore, $\mathbf{Y} \mid \boldsymbol{\theta} \sim K(\mathbf{Y} \mid \boldsymbol{\theta})$, i.e., the ordinal response distribution is the multinomial with the continuation-ratio logits parameterization.

□

A.1.2 Proof of Proposition 2.3

Proof. In this scenario, under the augmented model,

$$\Pr(\mathbf{Y} = j, \boldsymbol{\mathcal{Z}} \mid G_{\mathbf{x}}) = \mathbf{1}(\boldsymbol{\mathcal{Z}} \in \mathcal{R}_j) \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_{j\ell}(\mathbf{x})).$$

Integrating out $\boldsymbol{\mathcal{Z}}$, the probability for the j-th response category becomes

$$\begin{aligned}
\Pr(\mathbf{Y} = j \mid G_{\mathbf{x}}) &= \int_{\mathcal{R}_j} \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_{j\ell}(\mathbf{x})) \, \mathrm{d}\boldsymbol{\mathcal{Z}} \\
&= \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \int_{\mathcal{R}_j} \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_{j\ell}(\mathbf{x})) \, \mathrm{d}\boldsymbol{\mathcal{Z}} \\
&= \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \left\{ \left(\int_0^{+\infty} \mathfrak{L}(\mathcal{Z}_j \mid \theta_{j\ell}(\mathbf{x})) \, \mathrm{d}\mathcal{Z}_j \right) \prod_{k=1}^{j-1} \left(\int_{-\infty}^{0} \mathfrak{L}(\mathcal{Z}_k \mid \theta_{k\ell}(\mathbf{x})) \, \mathrm{d}\mathcal{Z}_k \right) \right\} \\
&= \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \left\{ \varphi(\theta_{j\ell}(\mathbf{x})) \prod_{k=1}^{j-1} [1 - \varphi(\theta_{k\ell}(\mathbf{x}))] \right\},
\end{aligned}$$

for $j = 2, \ldots, C-1$. The function under integration and countable summation in the first line takes non-negative values, and we can thus switch the order of the two operations. Similarly, we obtain $\Pr(\mathbf{Y} = 1 \mid G_{\mathbf{x}}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \, \varphi(\theta_{1\ell}(\mathbf{x}))$, and $\Pr(\mathbf{Y} = C \mid G_{\mathbf{x}}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \left\{ \prod_{k=1}^{C-1} [1 - \varphi(\theta_{k\ell}(\mathbf{x}))] \right\}$. Hence, $\mathbf{Y} \mid G_{\mathbf{x}} \sim \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \, K(\mathbf{Y} \mid \boldsymbol{\theta}_\ell(\mathbf{x}))$, i.e., the multinomial LSBP mixture model. □

A.1.3 Proof of Lemma 2.1

Proof. Consider the set of probability densities $\{f^0_{\mathbf{x}}(\mathbf{z}) : \mathbf{x} \in \mathcal{X}\}$ for $\mathbf{Z} \in \mathbb{R}^{C-1}$. Let $\mathfrak{F}$ be the set of all distributions defined on $\mathbb{R}^{C-1}$. A prior $\mathcal{F}_{\mathbf{x}} = \{F_{\mathbf{x}}(w, B) : \mathbf{x} \in \mathcal{X}\}$ on $\mathfrak{F}^{\mathcal{X}}$ is a stochastic process on an appropriate probability space $(\mathcal{W}, \mathscr{F}, \Pi)$, such that for every $\mathbf{x} \in \mathcal{X}$ and almost every $w \in \mathcal{W}$, $F_{\mathbf{x}}(w, \cdot) \in \mathfrak{F}$. The set of densities $\{f^0_{\mathbf{x}}(\mathbf{z}) : \mathbf{x} \in \mathcal{X}\}$ having KL property relative to $\mathcal{F}_{\mathbf{x}}$ refers to

$$\Pi\left\{w \in \mathcal{W} : \int f^0_{\mathbf{x}_t}(\mathbf{z}) \log(f^0_{\mathbf{x}_t}(\mathbf{z})/f_{\mathbf{x}_t}(\mathbf{z}))\, \mathrm{d}\mathbf{z} < \epsilon, \quad t = 1, \ldots, T\right\} > 0, \tag{A.1}$$

for any $\epsilon > 0$, $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathcal{X}$, $T \in \mathbb{N}^+$.

Now consider the set of probability masses $\{p_{\mathbf{x}}(y) : \mathbf{x} \in \mathcal{X}\}$ for ordinal variable Y with C categories. Denote the set of all distributions on $\{1, \ldots, C\}$ by $\mathfrak{P}$. Let $\{\mathcal{R}_1, \ldots, \mathcal{R}_C\}$ be a partition of $\mathbb{R}^{C-1}$. To connect with the distribution of continuous variable, we consider the mapping from $\mathfrak{F}^{\mathcal{X}}$ to $\mathfrak{P}^{\mathcal{X}}$, given by

$$f_{\mathbf{x}} \mapsto p_{\mathbf{x}}(y) = \int_{\mathcal{R}_y} f_{\mathbf{x}}(\mathbf{z})\, \mathrm{d}\mathbf{z}, \quad y = 1, \ldots, C. \tag{A.2}$$

This mapping induces a prior on $\mathfrak{P}^{\mathcal{X}}$ from $\mathcal{F}_{\mathbf{x}}$. The prior, denoted by $\mathcal{P}_{\mathbf{x}}$, is a $\mathfrak{P}$-valued stochastic process on probability space $(\mathcal{W}, \mathscr{P}, \Pi)$. Additionally, let $p^0_{\mathbf{x}}(y)$ denote the discrete distribution induced by $f^0_{\mathbf{x}}(\mathbf{z})$. Following the definition of KL property for continuous distributions, we say $\{p^0_{\mathbf{x}}(y) : \mathbf{x} \in \mathcal{X}\}$ possesses the KL property with respect to $\mathcal{P}_{\mathbf{x}}$ if

$$\Pi\left\{w \in \mathcal{W} : \sum_{y=1}^{C} p^0_{\mathbf{x}_t}(y) \log(p^0_{\mathbf{x}_t}(y)/p_{\mathbf{x}_t}(y)) < \epsilon, \quad t = 1, \ldots, T\right\} > 0 \tag{A.3}$$

for any $\epsilon > 0$, $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathcal{X}$, $T \in \mathbb{N}^+$.

To prove the lemma, it is adequate to show that every w that satisfies the

criterion in (A.1) also satisfies the criterion in (A.3).

The proof relies on the following inequality of KL divergence,

$$\int_{\mathcal{A}} g_1(\mathbf{u}) \log\left(\frac{g_1(\mathbf{u})}{g_2(\mathbf{u})}\right) \mathrm{d}\mathbf{u} \geq \int_{\mathcal{A}} g_1(\mathbf{u})\,\mathrm{d}\mathbf{u} \times \log\left(\frac{\int_{\mathcal{A}} g_1(\mathbf{u})\,\mathrm{d}\mathbf{u}}{\int_{\mathcal{A}} g_2(\mathbf{u})\,\mathrm{d}\mathbf{u}}\right), \tag{A.4}$$

where $g_r(\mathbf{u})$ is a density of $\mathbf{u} \in \mathbb{R}^s$, $r = 1, 2$ and $\mathcal{A}$ is a generic subset of $\mathbb{R}^s$. To show this inequality, let $\mathcal{H}_r = \int_{\mathcal{A}} g_r(\mathbf{u})\,\mathrm{d}\mathbf{u}$, such that $h_r(\mathbf{u}) = g_r(\mathbf{u})/\mathcal{H}_r$, $r = 1, 2$, are densities on $\mathcal{A}$. The left-hand-side of (A.4) can be expressed as $\mathcal{H}_1 \int_{\mathcal{A}} h_1(\mathbf{u}) \log(\frac{\mathcal{H}_1 h_1(\mathbf{u})}{\mathcal{H}_2 h_2(\mathbf{u})})\,\mathrm{d}\mathbf{u} = \mathcal{H}_1 \log(\frac{\mathcal{H}_1}{\mathcal{H}_2}) + \mathcal{H}_1 \int_{\mathcal{A}} h_1(\mathbf{u}) \log(\frac{h_1(\mathbf{u})}{h_2(\mathbf{u})})\,\mathrm{d}\mathbf{u} \geq \mathcal{H}_1 \log(\frac{\mathcal{H}_1}{\mathcal{H}_2})$, because $\int_{\mathcal{A}} h_1(\mathbf{u}) \log(\frac{h_1(\mathbf{u})}{h_2(\mathbf{u})})\,\mathrm{d}\mathbf{u}$ is the KL divergence between densities h_1 and h_2.

For any set $\mathcal{R}_y$ in the partition of $\mathbb{R}^{C-1}$, from (A.4) we obtain

$$\begin{aligned}\int_{\mathcal{R}_y} f^0_{\mathbf{x}_t}(\mathbf{z}) \log(f^0_{\mathbf{x}_t}(\mathbf{z})/f_{\mathbf{x}_t}(\mathbf{z}))\,\mathrm{d}\mathbf{z} &\geq \int_{\mathcal{R}_y} f^0_{\mathbf{x}_t}(\mathbf{z})\,\mathrm{d}\mathbf{z} \times \log\left(\frac{\int_{\mathcal{R}_y} f^0_{\mathbf{x}_t}(\mathbf{z})\,\mathrm{d}\mathbf{z}}{\int_{\mathcal{R}_y} f_{\mathbf{x}_t}(\mathbf{z})\,\mathrm{d}\mathbf{z}}\right) \\ &= p^0_{\mathbf{x}_t}(y) \log\left(\frac{p^0_{\mathbf{x}_t}(y)}{p_{\mathbf{x}_t}(y)}\right),\end{aligned}$$

for $y = 1, \ldots, C$. Consider any $w \in \mathcal{W}$ satisfying $\int f^0_{\mathbf{x}_t}(\mathbf{z}) \log(f^0_{\mathbf{x}_t}(\mathbf{z})/f_{\mathbf{x}_t}(\mathbf{z}))\,\mathrm{d}\mathbf{z} < \epsilon$, for $t = 1, \ldots, T$. Then, we have

$$\begin{aligned}\epsilon &> \int f^0_{\mathbf{x}_t}(\mathbf{z}) \log(f^0_{\mathbf{x}_t}(\mathbf{z})/f_{\mathbf{x}_t}(\mathbf{z}))\,\mathrm{d}\mathbf{z} = \sum_{y=1}^{C} \int_{\mathcal{R}_y} f^0_{\mathbf{x}_t}(\mathbf{z}) \log(f^0_{\mathbf{x}_t}(\mathbf{z})/f_{\mathbf{x}_t}(\mathbf{z}))\,\mathrm{d}\mathbf{z} \\ &\geq \sum_{y=1}^{C} p^0_{\mathbf{x}_t}(y) \log\left(\frac{p^0_{\mathbf{x}_t}(y)}{p_{\mathbf{x}_t}(y)}\right).\end{aligned}$$

We have thus obtained that w also satisfies $\sum_{y=1}^{C} p^0_{\mathbf{x}_t}(y) \log(p^0_{\mathbf{x}_t}(y)/p_{\mathbf{x}_t}(y)) < \epsilon$, for $t = 1, \ldots, T$, which completes the argument for the proof. □

A.1.4 Proof of Theorem 2.1

Proof. Based on Proposition 2.3, the multinomial LSBP mixture model can be formulated in terms of latent continuous responses as follows: $\mathbf{Y} \mid \boldsymbol{\mathcal{Z}} \sim \mathbf{1}(\mathbf{Y} = j \iff \boldsymbol{\mathcal{Z}} \in \mathcal{R}_j)$, for $j = 1, \ldots, C$, and

$$\begin{aligned}
&\boldsymbol{\mathcal{Z}} \mid G_{\mathbf{x}} \sim \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_{j\ell}(\mathbf{x})) = \int \prod_{j=1}^{C-1} \mathfrak{L}(\mathcal{Z}_j \mid \theta_j(\mathbf{x}))\, \mathrm{d}G_{\mathbf{x}}(\boldsymbol{\theta}), \\
&\omega_1(\mathbf{x}) = \varphi(\mathbf{x}^T \boldsymbol{\gamma}_1), \;\; \omega_\ell(\mathbf{x}) = \varphi(\mathbf{x}^T \boldsymbol{\gamma}_\ell) \prod_{h=1}^{\ell-1} (1 - \varphi(\mathbf{x}^T \boldsymbol{\gamma}_h)), \;\; \ell \geq 2, \\
&\theta_{j\ell}(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}_{j\ell} \mid \boldsymbol{\mu}_j, \Sigma_j \overset{ind.}{\sim} N(\mathbf{x}^T \boldsymbol{\mu}_j, \mathbf{x}^T \Sigma_j \mathbf{x}), \quad j = 1, \ldots, C-1, \;\; \ell \geq 1.
\end{aligned} \tag{A.5}$$

where $\{\mathcal{R}_j : j = 1, \ldots, C\}$ is the partition of $\mathbb{R}^{C-1}$ defined in equation (2.11).

Let $\mathcal{F}_{\mathbf{x}}$ be the above LSBP mixture prior for continuous random vector $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{C-1}$. The original multinomial LSBP mixture prior on ordinal distributions is denoted by $\mathcal{P}_{\mathbf{x}}$. Consider the set of probability masses $\{p^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ on $\{1, \ldots, C\}$. From Lemma 2.1, to show that $\{p^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ has the KL property with respect to $\mathcal{P}_{\mathbf{x}}$, we can utilize the result regarding the KL support of $\mathcal{F}_{\mathbf{x}}$.

Suppose the probability densities $\{f^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ on $\mathbb{R}^{C-1}$ satisfy the regularity conditions (v) to (viii) in Barrientos et al. (2012, Theorem 5). We prove that $\{f^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ possesses the KL property relative to $\mathcal{F}_{\mathbf{x}}$. The proof consists of two parts. We first show that $\mathcal{F}_{\mathbf{x}}$ falls in the scheme of dependent stick-breaking process (DSBP) priors (Barrientos et al., 2012, Definition 4). Then, we confirm that the mixture kernel of the model for $\boldsymbol{\mathcal{Z}}$ satisfies the conditions in Barrientos et al. (2012, Theorem 10).

For the LSBP prior discussed in this paper, we introduce the marginal distributions $\mathscr{V}^{V_\ell}_{\mathcal{X}}$, $G^0_{\mathcal{X}}$, and the copula functions $\mathscr{C}^{V_\ell}_{\mathcal{X}}$, $\mathscr{C}^{\theta}_{\mathcal{X}}$, $\ell = 1, 2, \ldots$, defined as

follows:

$$\mathscr{V}_{\mathcal{X}}^{V_\ell} = \{N(\mathbf{x}^T\boldsymbol{\gamma}_0, \mathbf{x}^T\Gamma_0\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}, \quad G_{\mathcal{X}}^0 = \{\prod_{j=1}^{C-1} N(\mathbf{x}^T\boldsymbol{\beta}_j, \mathbf{x}^T\Sigma_j\mathbf{x}) : \mathbf{x} \in \mathcal{X}\},$$

$$\mathscr{C}_{\mathcal{X}}^{V_\ell} = \{C_{\mathbf{x}_1,\ldots,\mathbf{x}_d}(u_1,\ldots,u_d) = \Phi_{S(\mathbf{x}_1,\ldots,\mathbf{x}_d)}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d)), \mathbf{x}_1,\ldots,\mathbf{x}_d \in \mathcal{X}\},$$

$$\mathscr{C}_{\mathcal{X}}^{\theta} = \{C_{\mathbf{x}_1,\ldots,\mathbf{x}_d}(u_1,\ldots,u_d) = \prod_{j=1}^{C-1} \Phi_{R_j(\mathbf{x}_1,\ldots,\mathbf{x}_d)}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d)), \mathbf{x}_1,\ldots,\mathbf{x}_d \in \mathcal{X}\},$$

where Φ is the c.d.f. of the standard normal distribution. In addition, $\Phi_{S(\mathbf{x}_1,\ldots,\mathbf{x}_d)}$ and $\Phi_{R_j(\mathbf{x}_1,\ldots,\mathbf{x}_d)}$ denote the c.d.f. of a $d-$variate normal distribution with mean 0, variance 1, and correlation matrix $S(\mathbf{x}_1,\ldots,\mathbf{x}_d)$, $R_j(\mathbf{x}_1,\ldots,\mathbf{x}_d)$, whose (s,t)-entry is given respectively by

$$S(\mathbf{x}_1,\ldots,\mathbf{x}_d)_{(s,t)} = \frac{\mathbf{x}_s^T\Gamma_0\mathbf{x}_t}{\sqrt{\mathbf{x}_s^T\Gamma_0\mathbf{x}_s}\sqrt{\mathbf{x}_t^T\Gamma_0\mathbf{x}_t}}, \quad R_j(\mathbf{x}_1,\ldots,\mathbf{x}_d)_{(s,t)} = \frac{\mathbf{x}_s^T\Sigma_j\mathbf{x}_t}{\sqrt{\mathbf{x}_s^T\Sigma_j\mathbf{x}_s}\sqrt{\mathbf{x}_t^T\Sigma_j\mathbf{x}_t}}$$

Let $\mathscr{T}$ denote the transformation induced by the standard logistic function, i.e. $x \mapsto \varphi(x)$, which is strictly increasing, and define $\mathscr{V}_{\mathcal{X}}^{\eta_\ell} := \mathscr{T} \circ \mathscr{V}_{\mathcal{X}}^{V_\ell}$, $\mathscr{C}_{\mathcal{X}}^{\eta_\ell} := \mathscr{C}_{\mathcal{X}}^{V_\ell}$. Consequently, it is easy to check that the LSBP prior fits in the definition of DSBP prior given in Barrientos et al. (2012). Specifically, the LSBP prior can be written as $DSBP(\mathscr{C}_{\mathcal{X},\mathbb{N}}^{\eta}, \mathscr{C}_{\mathcal{X}}^{\theta}, \mathscr{V}_{\mathcal{X},\mathbb{N}}^{\eta}, G_{\mathcal{X}}^0)$.

Now, consider the mixing kernel of (A.5), given by

$$\chi(\mathbf{z} \mid \boldsymbol{\theta}) = \prod_{j=1}^{C-1} \frac{e^{-(z_j-\theta_j)}}{(1+e^{-(z_j-\theta_j)})^2}.$$

We check that the kernel satisfies conditions (i) $-$ (iv) of Barrientos et al. (2012, Theorem 5). Letting $\chi_j(z) = \frac{e^{-(z-\theta_j)}}{(1+e^{-(z-\theta_j)})^2}$, we have

$$\chi_j'(z) = -\frac{e^{-(z-\theta_j)}(1-e^{-(z-\theta_j)})}{(1+e^{-(z-\theta_j)})^3}. \tag{A.6}$$

Obviously, $\chi(\mathbf{z} \mid \boldsymbol{\theta})$ is continuous and strictly positive on $\mathbb{R}^{C-1}$. It is bounded above by $\frac{1}{4^{C-1}}$ because from (A.6), each $\chi_j(z)$ takes its maximum $\frac{1}{4}$ at $z = \theta_j$. Condition (i) is satisfied. In addition, since $\chi_j(z)$ decreases as $z > \theta_j$, we can choose $l_1 = \sqrt{\sum_{j=1}^{C-1} \theta_j^2}$ such that $\chi(\mathbf{z} \mid \boldsymbol{\theta})$ decreases as $\mathbf{z}$ outside the ball $\{\mathbf{z} : \|\mathbf{z}\| < l_1\}$. Condition (ii) is also satisfied. As for condition (iii), it is satisfied because $\chi_j'(z)/\chi(z) \to -1$ as $z \to \infty$, and $\chi_j'(z)/\chi_j(z) \to 1$ as $z \to -\infty$. Finally, condition (iv) is obviously satisfied. Moreover, $\mathscr{C}_{\mathcal{X}}^{\eta_\ell}$ and $\mathscr{C}_{\mathcal{X}}^{\theta}$ are copulas with positive density on the appropriate unitary hyper-cubes. Based on Barrientos et al. (2012, Theorem 10), $\{f_{\mathbf{x}}^0 : \mathbf{x} \in \mathcal{X}\}$ possess the KL property relative to $\mathcal{F}_{\mathbf{x}}$. Consequently, by Lemma 2.1, the induced distributions on discrete space $\{p_{\mathbf{x}}^0 : \mathbf{x} \in \mathcal{X}\}$ possess the KL property relative to $\mathcal{P}_{\mathbf{x}}$.

To complete the proof, we finally consider the required regularity conditions such that $p_{\mathbf{x}}^0$ is in the KL support of $\mathcal{P}_{\mathbf{x}}$. We notice that the regularity conditions for continuous distributions are not necessary in our ordinal regression setting. In fact, the only condition we need for $p_{\mathbf{x}}^0$ is that it comprises strictly positive probabilities for all response categories. To obtain this result, we first show that there exists a specific $f_{\mathbf{x}}^0$ that enables the connection with $p_{\mathbf{x}}^0$ given in (A.2). Then, we show that such $f_{\mathbf{x}}^0$ satisfies the regularity conditions (v) to (viii) of Barrientos et al. (2012, Theorem 5).

For a generic probability mass $p_{\mathbf{x}}^0$ on $\{1, \ldots, C\}$, without loss of generality, we assume $p_{\mathbf{x}}^0$ is positive hereinafter. We define

$$f_{\mathbf{x}}^0(\mathbf{z}) = \sum_{y=1}^{C} \frac{p_{\mathbf{x}}^0(y) \mathbf{1}_{\mathcal{R}_y}(\mathbf{z}) \phi(\mathbf{z})}{\int_{\mathcal{R}_y} \phi(\mathbf{z}) \, \mathrm{d}\mathbf{z}}, \tag{A.7}$$

where $\phi(\mathbf{z})$ denotes the p.d.f. of the standard $C-1$ dimensional normal distribution. It is straightforward to check that $\int_{\mathcal{R}_y} f_{\mathbf{x}}^0(\mathbf{z}) \, \mathrm{d}\mathbf{z} = p_{\mathbf{x}}^0(y)$, for $y = 1, \ldots, C$. Hence, the relationship defined in (A.2) is satisfied.

We now show that this specific $f^0_{\mathbf{x}}(\mathbf{z})$ satisfies the regularity conditions (v) to (viii) of Barrientos et al. (2012, Theorem 5). To proceed, we notice that because $\phi(\mathbf{z})$ is symmetric around the origin, we have

$$\frac{1}{2^{C-1}} \leq \int_{\mathcal{R}_y} \phi(\mathbf{z})\, \mathrm{d}\mathbf{z} \leq \frac{1}{2}, \quad y = 1, \ldots, C \tag{A.8}$$

where $\{\mathcal{R}_y : y = 1, \ldots, C\}$ is the partition defined in equation (2.11). Hence, condition (v) follows directly from (A.8). For condition (vi), because $\log(f^0_{\mathbf{x}}(\mathbf{z}))$ is also bounded, we have

$$\int f^0_{\mathbf{x}}(\mathbf{z}) \log(f^0_{\mathbf{x}}(\mathbf{z}))\, \mathrm{d}\mathbf{z} = \sum_{y=1}^{c} \frac{p^0_{\mathbf{x}}(y)}{\int_{\mathcal{R}_y} \phi(\mathbf{z})\, \mathrm{d}\mathbf{z}} \int_{\mathcal{R}_y} \log(f^0_{\mathbf{x}}(\mathbf{z}))\phi(\mathbf{z})\, \mathrm{d}\mathbf{z} < \infty.$$

To show that condition (vii) holds, let $\mathcal{B}$ denote the set of boundaries for the partition $\{\mathcal{R}_y : y = 1, \ldots, C\}$. The set $\mathcal{B}$ has measure 0, and outside $\mathcal{B}$ the function $\log(f^0_{\mathbf{x}}(\mathbf{z})/h_\delta(\mathbf{z}))$ is bounded, where $h_\delta(\mathbf{z}) = \inf_{\|\mathbf{z}'-\mathbf{z}\|<\delta} f^0_{\mathbf{x}}(\mathbf{z}')$. Therefore, we can obtain

$$\begin{aligned}\int f^0_{\mathbf{x}} \log(f^0_{\mathbf{x}}(\mathbf{z})/h_\delta(\mathbf{z}))\, \mathrm{d}\mathbf{z} &= \int_{\mathcal{B}} f^0_{\mathbf{x}} \log(f^0_{\mathbf{x}}(\mathbf{z})/h_\delta(\mathbf{z}))\, \mathrm{d}\mathbf{z} + \int_{\mathcal{B}^C} f^0_{\mathbf{x}} \log(f^0_{\mathbf{x}}(\mathbf{z})/h_\delta(\mathbf{z}))\, \mathrm{d}\mathbf{z} \\ &= \sum_{y=1}^{C} \frac{p^0_{\mathbf{x}}(y)}{\int_{\mathcal{R}_y} \phi(\mathbf{z})\, \mathrm{d}\mathbf{z}} \int_{\mathcal{B}^C \cap \mathcal{R}_y} \log(f^0_{\mathbf{x}}(\mathbf{z})/h_\delta(\mathbf{z}))\phi(\mathbf{z})\, \mathrm{d}\mathbf{z} < \infty.\end{aligned}$$

Finally, condition (viii) can be verified using the fact that the tails of $\log(\chi(\mathbf{z}))$ behave like $|\mathbf{z}|$.

Hence, we have proved that the set of generic, positive probability mass functions $\{p^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ on $\{1, \ldots, C\}$ possesses the KL property with respect to the proposed nonparametric prior model.

Similarly, we can establish the KL property for the nonparametric prior induced by the two simpler models. First, it is straightforward to check that the common-

weights and the common-atoms nonparametric priors fall in the wDDP (Barrientos et al., 2012, Definition 2) and θDSBP (Barrientos et al., 2012, Definition 4) framework, respectively. In addition, the mixing kernel is the same as in Theorem 2.1, and therefore the relevant regularity conditions are satisfied as shown earlier. Hence, by Theorem 5 and Theorem 10 of Barrientos et al. (2012), and Lemma 2.1, we obtain the following corollary.

Corollary A.1. *Denote by $w\mathcal{P}_{\mathbf{x}}$ the common-weights LSBP mixture prior discussed in Section 2.3.1, and $\theta\mathcal{P}_{\mathbf{x}}$ the common-atoms LSBP mixture prior proposed in Section 2.3.2. Consider $\{p^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, a generic collection of covariate-dependent probabilities for an ordinal response with C categories. Assume that the probability of each response category is strictly positive. Then, the mass functions $\{p^0_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ are in the KL support of $w\mathcal{P}_{\mathbf{x}}$ and $\theta\mathcal{P}_{\mathbf{x}}$.*

□

A.1.5 Proof of Proposition 2.4

Proof. Denote by $\Pr(Y = j \mid \mathcal{M}_1)$ and $\Pr(Y = j \mid \mathcal{M}_2)$, for $j = 1, ..., C$, the j-th category response probability under the continuation-ratio logits model and the cumulative logit model, respectively. The parameter vector for the former model is $(\theta_1, ..., \theta_{C-1})$, whereas for the latter it is $(\vartheta, \varkappa_2, ..., \varkappa_{C-1})$. Recall that, for identifiability, $\varkappa_1 = 0$.

For the first response probability, we have $\Pr(Y = 1 \mid \mathcal{M}_1) = e^{\theta_1}/(1 + e^{\theta_1})$, and $\Pr(Y = 1 \mid \mathcal{M}_2) = \Pr(Z \leq 0 \mid \mathcal{M}_2) = e^{-\vartheta}/(1 + e^{-\vartheta})$, from which we obtain $\vartheta = -\theta_1$. The equality for these two parameters holds true for any value of C, but this step also establishes the result for the simplest case of $C = 2$

For $C = 3$, from the argument above, we have $\vartheta = -\theta_1$. Now, under the

continuation-ratio logits model

$$\Pr(Y = 3 \mid \mathcal{M}_1) = \frac{1}{1 + e^{\theta_1}} \frac{1}{1 + e^{\theta_2}}, \tag{A.9}$$

while under the cumulative logit model

$$\Pr(Y = 3 \mid \mathcal{M}_2) = \Pr(Z > \varkappa_2 \mid \mathcal{M}_2) = \frac{1}{1 + e^{\varkappa_2 - \vartheta}} = \frac{1}{1 + e^{\varkappa_2 + \theta_1}}. \tag{A.10}$$

Setting (A.9) and (A.10) equal, and solving for $\varkappa_2$, we have

$$\varkappa_2 = \log(1 + e^{\theta_2} + e^{\theta_2 - \theta_1}) = \log(e^{\varkappa_1} + e^{\varkappa_1 + \theta_2} + e^{\theta_2 - \theta_1}),$$

which establishes the result for $C = 3$.

To prove the proposition by induction, assume the correspondence between the parameters of the two models holds true for an ordinal response with $C - 1$ categories, that is, $\vartheta = -\theta_1$ and $\varkappa_j = \log(e^{\varkappa_{j-1}} + e^{\varkappa_{j-1} + \theta_j} + e^{\theta_j - \theta_1})$, for $j = 2, \ldots, C - 2$.

Now, assume that the ordinal response has C categories. To complete the argument, we work with the probability for category C. For the continuation-ratio logits model:

$$\Pr(Y = C \mid \mathcal{M}_1) = \frac{1}{1 + e^{\theta_{C-1}}} \prod_{k=1}^{C-2} \frac{1}{1 + e^{\theta_k}} = \frac{1}{1 + e^{\theta_{C-1}}} \frac{1}{1 + e^{\varkappa_{C-2} + \theta_1}}$$

where the equality $\prod_{k=1}^{C-2} 1/(1 + e^{\theta_k}) = 1/(1 + e^{\varkappa_{C-2} + \theta_1})$ is obtained by starting from the right-hand side and using recursively (for $j = C - 2, C - 3, ..., 2$) the induction assumption $\varkappa_j = \log(e^{\varkappa_{j-1}} + e^{\varkappa_{j-1} + \theta_j} + e^{\theta_j - \theta_1})$. On the other hand,

$$\Pr(Y = C \mid \mathcal{M}_2) = \Pr(Z > \varkappa_{C-1} \mid \mathcal{M}_2) = \frac{1}{1 + e^{\varkappa_{C-1} - \vartheta}} = \frac{1}{1 + e^{\varkappa_{C-1} + \theta_1}}.$$

Setting the two probabilities equal to each other, we can solve for $\varkappa_{C-1}$, resulting in

$$\varkappa_{C-1} = \log(e^{\varkappa_{C-2}} + e^{\varkappa_{C-2}+\theta_{C-1}} + e^{\theta_{C-1}-\theta_1}),$$

thus completing the induction argument.

Note that we can write $\varkappa_j = \log(A\, e^{\varkappa_{j-1}} + B)$, where $A = 1 + e^{\theta_j} > 1$, and $B = e^{\theta_j - \theta_1} > 0$, from which we can easily confirm the order restriction for the cut-off points, $\varkappa_j > \varkappa_{j-1}$, for $j = 2, \ldots, C-1$. □

A.2 Properties of Models for Heterogeneous Ordinal Responses

A.2.1 Proof of Proposition 3.1

Proof. By definition, we have

$$\mathrm{Corr}(\tilde{Y}_q, \tilde{Y}_{q'} \mid \theta, \sigma^2) = \frac{\mathrm{E}(\tilde{Y}_q \tilde{Y}_{q'} \mid \theta, \sigma^2) - \mathrm{E}(\tilde{Y}_q \mid \theta, \sigma^2)\mathrm{E}(\tilde{Y}_{q'} \mid \theta, \sigma^2)}{\sqrt{\mathrm{Var}(\tilde{Y}_q \mid \theta, \sigma^2)\mathrm{Var}(\tilde{Y}_{q'} \mid \theta, \sigma^2)}}.$$

Using law of total expectation/variance,

$$\begin{aligned}
\mathrm{E}(\tilde{Y}_q \tilde{Y}_{q'} \mid \theta, \sigma^2) &= \mathrm{E}(\varphi^2(\psi) \mid \theta, \sigma^2); \\
\mathrm{E}(\tilde{Y}_q \mid \theta, \sigma^2) = \mathrm{E}(\tilde{Y}_{q'} \mid \theta, \sigma^2) &= \mathrm{E}(\varphi(\psi) \mid \theta, \sigma^2); \\
\mathrm{Var}(\tilde{Y}_q \mid \theta, \sigma^2) = \mathrm{Var}(\tilde{Y}_{q'} \mid \theta, \sigma^2) &= \mathrm{E}(\varphi(\psi) \mid \theta, \sigma^2) - \{\mathrm{E}(\varphi(\psi) \mid \theta, \sigma^2)\}^2,
\end{aligned}$$

where the expectation is taken with respect to $\psi \sim N(\theta, \sigma^2)$.

Write $\psi = \theta + \zeta$, where $\zeta \sim N(0, \sigma^2)$. By Taylor expansion around the mean,

$$\varphi(\psi) \approx \varphi(\theta) + \zeta\varphi'(\theta) + \frac{\zeta^2}{2}\varphi''(\theta).$$

Then taking expectation yields $\mathrm{E}(\varphi(\psi) \mid \theta, \sigma^2) \approx \varphi(\theta) + \frac{\sigma^2}{2}\varphi''(\theta)$. Using the same technique,

$$\varphi^2(\psi) \approx \varphi^2(\theta) + 2\zeta\varphi(\theta)\varphi'(\theta) + \zeta^2\{(\varphi'(\theta))^2 + \varphi(\theta)\varphi''(\theta)\}.$$

Taking expectation yields $\mathrm{E}(\tilde{Y}_q\tilde{Y}_{q'} \mid \theta, \sigma^2) \approx \varphi^2(\theta) + \sigma^2\{(\varphi'(\theta))^2 + \varphi(\theta)\varphi''(\theta)\}$.

We further notice that $\varphi'(\theta) = \varphi(\theta)(1-\varphi(\theta))$, and $\varphi'(\theta) = \varphi(\theta)(1-\varphi(\theta))(1-2\varphi(\theta))$. The final results emerge after simple algebra. □

A.2.2 Proof of Proposition 3.2

Proof. Starting with the moment generating function for $\sum_{q=1}^{m}\tilde{R}_q$ and $\sum_{l=1}^{m-\sum_q \tilde{R}_q}\tilde{y}_l$ under model $\tilde{\mathcal{M}}$, we have

$$\begin{aligned}
&E_{\tilde{\mathcal{M}}}(e^{t_1\sum\tilde{R}_q + t_2\sum\tilde{y}_l} \mid m, G_{\mathbf{x}}) \\
&= \sum_{\ell=1}^{\infty}\omega_\ell(\mathbf{x})\left\{\prod_{q=1}^{m}\sum_{\tilde{R}_q=0,1} e^{t_1\tilde{R}_q}Bern(\tilde{R}_q \mid \varphi(\theta_{1\ell}(\mathbf{x})))\right\}\left\{\prod_{l=1}^{m-\sum_q\tilde{R}_q}\sum_{\tilde{y}_l=0,1} e^{t_2\tilde{y}_l}Bern(\tilde{y}_l \mid \varphi(\theta_{2\ell}(\mathbf{x})))\right\} \\
&= \sum_{\ell=1}^{\infty}\omega_\ell(\mathbf{x})\left\{\prod_{q=1}^{m}\sum_{\tilde{R}_q=0,1} e^{t_1\tilde{R}_q}Bern(\tilde{R}_q \mid \varphi(\theta_{1\ell}(\mathbf{x})))\right\}\left\{1+\varphi(\theta_{2\ell}(\mathbf{x}))(e^{t_2}-1)\right\}^{m-\sum\tilde{R}_q} \\
&= \sum_{\ell=1}^{\infty}\omega_\ell(\mathbf{x})\left[\prod_{q=1}^{m}\sum_{\tilde{R}_q=0,1}\left\{e^{t_1\tilde{R}_q}Bern(\tilde{R}_q \mid \varphi(\theta_{1\ell}(\mathbf{x})))\{1+\varphi(\theta_{2\ell}(\mathbf{x}))(e^{t_2}-1)\}^{1-\tilde{R}_q}\right\}\right] \\
&= \sum_{\ell=1}^{\infty}\omega_\ell(\mathbf{x})\left[(1-\varphi(\theta_{1\ell}(\mathbf{x})))\{1+\varphi(\theta_{2\ell}(\mathbf{x}))(e^{t_2}-1)\} + e^{t_1}\varphi(\theta_{1\ell}(\mathbf{x}))\right]^m \\
&= \sum_{\ell=1}^{\infty}\omega_\ell(\mathbf{x})\sum_{R=0}^{m}\binom{m}{R}\left\{e^{t_1}\varphi(\theta_{1\ell}(\mathbf{x}))\right\}^R\left[(1-\varphi(\theta_{1\ell}(\mathbf{x})))\{1+\varphi(\theta_{2\ell}(\mathbf{x}))(e^{t_2}-1)\}\right]^{m-R} \\
&= \sum_{\ell=1}^{\infty}\omega_\ell(\mathbf{x})\left\{\sum_{R=0}^{m}e^{t_1R}Bin(R \mid m, \varphi(\theta_{1\ell}(\mathbf{x})))\right\}\left[1+\varphi(\theta_{2\ell}(\mathbf{x}))(e^{t_2}-1)\right]^{m-R} \\
&= \sum_{\ell=1}^{\infty}\omega_\ell(\mathbf{x})\left\{\sum_{R=0}^{m}e^{t_1R}Bin(R \mid m, \varphi(\theta_{1\ell}(\mathbf{x})))\right\}\left\{\sum_{y=0}^{m-R}e^{t_2y}Bin(y \mid m-R, \varphi(\theta_{2\ell}(\mathbf{x})))\right\} \\
&= E_{\mathcal{M}}(e^{t_1R+t_2y} \mid m, G_{\mathbf{x}}),
\end{aligned}$$

which completes the argument for the proof. □

A.2.3 Proof of Proposition 3.3

Proof. We first show the equality holds on the parametric backbones of the corresponding nonparametric mixture models. Let $\mathcal{M}^*$ and $\tilde{\mathcal{M}}^*$ denote the kernel of the mixture model $\mathcal{M}$ and $\tilde{\mathcal{M}}$, respectively. Following the strategy in proving Proposition 3.2, we start from the MGF for $\sum_{q=1}^{m} \tilde{R}_q$ and $\sum_{l=1}^{m-\sum_q \tilde{R}_q} \tilde{y}_l$,

$$
\begin{aligned}
&E_{\tilde{\mathcal{M}}^*}(e^{t_1 \sum \tilde{R}_q + t_2 \sum \tilde{y}_l} \mid m, \theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \boldsymbol{\sigma^2}) \\
&\quad = \left\{ \int \prod_{q=1}^{m} \sum_{\tilde{R}_q=0,1} e^{t_1 \tilde{R}_q} Bern(\tilde{R}_q \mid \varphi(\psi_1)) N(\psi_1 \mid \theta_1(\mathbf{x}), \sigma_1^2) d\psi_1 \right\} \\
&\qquad \times \left\{ \int \prod_{l=1}^{m-\sum_q \tilde{R}_q} \sum_{\tilde{y}_l=0,1} e^{t_2 \tilde{y}_l} Bern(\tilde{y}_l \mid \varphi(\psi_2)) N(\psi_2 \mid \theta_2(\mathbf{x}), \sigma_2^2) d\psi_2 \right\} \\
&\quad = \int \int \left\{ \prod_{q=1}^{m} \sum_{\tilde{R}_q=0,1} e^{t_1 \tilde{R}_q} Bern(\tilde{R}_q \mid \varphi(\psi_1)) \right\} \left\{ \prod_{l=1}^{m-\sum_q \tilde{R}_q} \sum_{\tilde{y}_l=0,1} e^{t_2 \tilde{y}_l} Bern(\tilde{y}_l \mid \varphi(\psi_2)) \right\} \\
&\qquad \times N(\psi_1 \mid \theta_1(\mathbf{x}), \sigma_1^2) N(\psi_2 \mid \theta_2(\mathbf{x}), \sigma_2^2) d\psi_1 d\psi_2 \\
&\quad = \int \int \left\{ \sum_{R=0}^{m} e^{t_1 R} Bin(R \mid m, \varphi(\psi_1)) \right\} \left\{ \sum_{y=0}^{m-R} e^{t_2 y} Bin(y \mid m-R, \varphi(\psi_2)) \right\} \\
&\qquad \times N(\psi_1 \mid \theta_1(\mathbf{x}), \sigma_1^2) N(\psi_2 \mid \theta_2(\mathbf{x}), \sigma_2^2) d\psi_1 d\psi_2 \\
&\quad = \int \left\{ \sum_{R=0}^{m} e^{t_1 R} Bin(R \mid m, \varphi(\psi_1)) \right\} N(\psi_1 \mid \theta_1(\mathbf{x}), \sigma_1^2) d\psi_1 \\
&\qquad \times \int \left\{ \sum_{y=0}^{m-R} e^{t_2 y} Bin(y \mid m-R, \varphi(\psi_2)) \right\} N(\psi_2 \mid \theta_2(\mathbf{x}), \sigma_2^2) d\psi_2 \\
&\quad = E_{\mathcal{M}^*}(e^{t_1 R + t_2 y} \mid m, \theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \boldsymbol{\sigma^2}).
\end{aligned}
$$

Turning to the nonparametric mixture model, applying the equality for parametric model on every mixing component, we have

$$
\begin{aligned}
&E_{\tilde{\mathcal{M}}}(e^{t_1 \sum \tilde{R}_q + t_2 \sum \tilde{y}_l} \mid m, G_{\mathbf{x}}, \boldsymbol{\sigma}^{\mathbf{2}}) \\
&\quad = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \Bigg\{ \int \prod_{q=1}^{m} \sum_{\tilde{R}_q=0,1} e^{t_1 \tilde{R}_q} Bern(\tilde{R}_q \mid \varphi(\psi_1)) N(\psi_1 \mid \theta_{1\ell}(\mathbf{x}), \sigma_1^2) d\psi_1 \Bigg\} \\
&\qquad \times \Bigg\{ \int \prod_{l=1}^{m-\sum_q \tilde{R}_q} \sum_{\tilde{y}_l=0,1} e^{t_2 \tilde{y}_l} Bern(\tilde{y}_l \mid \varphi(\psi_2)) N(\psi_2 \mid \theta_{2\ell}(\mathbf{x}), \sigma_2^2) d\psi_2 \Bigg\} \\
&\quad = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \Bigg[\int \Bigg\{ \sum_{R=0}^{m} e^{t_1 R} Bin(R \mid m, \varphi(\psi_1)) \Bigg\} N(\psi_1 \mid \theta_{1\ell}(\mathbf{x}), \sigma_1^2) d\psi_1 \Bigg] \\
&\qquad \times \Bigg[\int \Bigg\{ \sum_{y=0}^{m-R} e^{t_2 y} Bin(y \mid m-R, \varphi(\psi_2)) \Bigg\} N(\psi_2 \mid \theta_{2\ell}(\mathbf{x}), \sigma_2^2) d\psi_2 \Bigg] \\
&\quad = E_{\mathcal{M}}(e^{t_1 R + t_2 y} \mid m, G_{\mathbf{x}}, \boldsymbol{\sigma}^{\mathbf{2}}).
\end{aligned}
$$

Therefore, we obtaine the desired equation. □

A.3 Properties of Models for Longitudinal Ordinal Responses

A.3.1 Proof of Proposition 4.1

Proof. For the probability response curve $\mathbf{P}_{\mathbf{1}\boldsymbol{\tau}}$, we have

$$
\begin{aligned}
\mathbf{P}_{\mathbf{1}\boldsymbol{\tau}} &= \int (\Pr(Y_{\tau_1} = 1 \mid \boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}, \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2), \ldots, \Pr(Y_{\tau_T} = 1 \mid \boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}, \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2))^\top p(\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}} \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2)\, d\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}} \\
&= \int \boldsymbol{\pi}(\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}) N(\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}} \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2 \mathbf{I})\, d\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}} = \mathrm{E}(\boldsymbol{\pi}(\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2).
\end{aligned}
$$

Then, to find the diagonal and off-diagonal elements for the covariance matrix of $\mathbf{Y}_{\boldsymbol{\tau}}$, we use the law of total variance/covariance. For the diagonal elements, we

can write

$$\begin{aligned}\mathrm{Var}(Y_\tau \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2) &= \mathrm{Var}[\mathrm{E}(Y_\tau \mid \mathcal{Z}_{\boldsymbol{\tau}}) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2] + \mathrm{E}[\mathrm{Var}(Y_\tau \mid \mathcal{Z}_{\boldsymbol{\tau}}) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2] \\ &= \mathrm{Var}[\varphi(\mathcal{Z}_\tau) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2] + \mathrm{E}[\varphi(\mathcal{Z}_\tau)(1-\varphi(\mathcal{Z}_\tau)) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2] \\ &= \mathrm{E}[\varphi(\mathcal{Z}_\tau) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2] - \mathrm{E}^2[\varphi(\mathcal{Z}_\tau) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2].\end{aligned}$$

Similarly, for the off-diagonal entries, we obtain

$$\begin{aligned}\mathrm{Cov}(Y_\tau, Y_{\tau'} \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2) &= \mathrm{Cov}[\mathrm{E}(Y_\tau \mid \boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}), \mathrm{E}(Y_{\tau'} \mid \boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2] + \mathrm{E}[\mathrm{Cov}(Y_\tau, Y_{\tau'} \mid \boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}}) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2] \\ &= \mathrm{Cov}[\varphi(\mathcal{Z}_\tau), \varphi(\mathcal{Z}_{\tau'}) \mid \mathbf{Z}_{\boldsymbol{\tau}}, \sigma_\epsilon^2].\end{aligned}$$

□

A.3.2 Proof of Proposition 4.2

Proof. To establish the result, we first prove the following lemma.

Lemma A.1. *Consider the bivariate vector* $\mathbf{Z} = (Z_1, Z_2)^\top$ *that follows* $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *where* $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ *and* $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \gamma\sigma_1\sigma_2 \\ \gamma\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. *Then we have,*

$$\begin{aligned}&E(\varphi(Z_i)) \approx \varphi(\mu_i) + \frac{\sigma_i^2}{2}\varphi''(\mu_i), \quad i = 1, 2, \\ &E(\varphi(Z_1)\varphi(Z_2)) \approx \varphi(\mu_1)\varphi(\mu_2) + \frac{1}{2}[\sigma_1^2\varphi''(\mu_1)\varphi(\mu_2) + 2\gamma\sigma_1\sigma_2\varphi'(\mu_1)\varphi'(\mu_2) + \sigma_2^2\varphi(\mu_1)\varphi''(\mu_2)].\end{aligned}$$

Proof. To show the result, we write $\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\zeta}$, where $\boldsymbol{\zeta} \sim N(0, \boldsymbol{\Sigma})$. By Taylor expansion around the mean,

$$\varphi(Z_i) \approx \varphi(\mu_i) + \zeta_i\varphi'(\mu_i) + \frac{\zeta_i^2}{2}\varphi''(\mu_i).$$

Then taking expectation yields $\mathrm{E}(\varphi(Z_i)) \approx \varphi(\mu_i) + \frac{\sigma_i^2}{2}\varphi''(\mu_i)$, $i = 1, 2$.

As for $E(\varphi(Z_1)\varphi(Z_2))$, consider the function $f(\mathbf{Z}) = \varphi(Z_1)\varphi(Z_2)$, using the bivariate version of Taylor expansion,

$$f(\mathbf{Z}) \approx f(\boldsymbol{\mu}) + \bigtriangledown f(\boldsymbol{\mu})^\top \boldsymbol{\zeta} + \frac{1}{2}\boldsymbol{\zeta}^\top \bigtriangledown^2 f(\boldsymbol{\mu})\boldsymbol{\zeta}.$$

Similarly, taking expectation with respect to $\boldsymbol{\zeta}$ we can obtain the result. □

Turning to the proof of Proposition 4.2, we notice that $\mathbf{Z}_{\boldsymbol{\tau}} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Marginalizing out $\mathbf{Z}_{\boldsymbol{\tau}}$, we have $\boldsymbol{\mathcal{Z}}_{\boldsymbol{\tau}} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\epsilon^2 \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \sigma_\epsilon^2\mathbf{I})$. Therefore, for any $\tau, \tau' \in \boldsymbol{\tau}$, we have

$$\begin{pmatrix} \mathcal{Z}_\tau \\ \mathcal{Z}_{\tau'} \end{pmatrix} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\epsilon^2 \sim N(\begin{pmatrix} \mu_\tau \\ \mu_{\tau'} \end{pmatrix}, \begin{pmatrix} \Sigma_{\tau,\tau} + \sigma_\epsilon^2 & \Sigma_{\tau,\tau'} \\ \Sigma_{\tau',\tau} & \Sigma_{\tau',\tau'} + \sigma_\epsilon^2 \end{pmatrix})$$

To establish the connection with the mean and covariance of the signal process, we write

$$\begin{pmatrix} \mu_\tau \\ \mu_{\tau'} \end{pmatrix} = \begin{pmatrix} \mathrm{E}(Z_\tau \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \mathrm{E}(Z_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{pmatrix}$$

$$\begin{pmatrix} \Sigma_{\tau,\tau} + \sigma_\epsilon^2 & \Sigma_{\tau,\tau'} \\ \Sigma_{\tau',\tau} & \Sigma_{\tau',\tau'} + \sigma_\epsilon^2 \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(Z_\tau \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sigma_\epsilon^2 & \mathrm{Cov}(Z_\tau, Z_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \mathrm{Cov}(Z_\tau, Z_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) & \mathrm{Var}(Z_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sigma_\epsilon^2 \end{pmatrix}$$

Similar to the proof of Proposition 4.1, we can show

$$\begin{aligned} &\Pr(Y_t = 1 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\epsilon^2) = \mathrm{E}(\varphi(\mathcal{Z}_\tau) \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\epsilon^2) \\ &\mathrm{Cov}(\mathbf{Y}_\tau, \mathbf{Y}_{\tau'} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\epsilon^2) = \mathrm{Cov}[\varphi(\mathcal{Z}_\tau), \varphi(\mathcal{Z}_{\tau'}) \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\epsilon^2] \end{aligned}$$

Applying Lemma A.1, the desired outcome emerges as a direct consequence of algebraic simplification. □

A.3.3 Proof of Proposition 4.3

Proof. The result is proved by considering the corresponding f.d.d.s. on any finite grids $\boldsymbol{\tau}$. Let the bold letter denote the corresponding process evaluated at $\boldsymbol{\tau}$. From the model assumption mentioned in (4.2) and (4.3), we have

$$\mathbf{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \boldsymbol{\mu}|\boldsymbol{\Sigma} \sim N(\mu_0\mathbf{1}, (\nu-3)\boldsymbol{\Sigma}),\ \boldsymbol{\Sigma} \sim IW(\nu, \boldsymbol{\Psi}).$$

To obtain the marginal distribution of $\mathbf{Z}$, we have

$$p(\mathbf{Z}) = \int\int p(\mathbf{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}) p(\boldsymbol{\Sigma})\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}.$$

Marginalizing over the mean vector $\boldsymbol{\mu}$, we obtain $\mathbf{Z} \mid \boldsymbol{\Sigma} \sim N(\mu_0\mathbf{1}, (\nu-2)\boldsymbol{\Sigma})$. Based on that,

$$\begin{aligned} p(\mathbf{Z}) &= \int p(\mathbf{Z} \mid \boldsymbol{\Sigma}) p(\boldsymbol{\Sigma})\, d\boldsymbol{\Sigma} \\ &\propto \int \frac{\exp\{-\frac{1}{2}\mathrm{Tr}[(\boldsymbol{\Psi}_\phi + \frac{(\mathbf{Z}-\mu_0\mathbf{1})(\mathbf{Z}-\mu_0\mathbf{1})^\top}{\nu-2})\boldsymbol{\Sigma}^{-1}]\}}{|\boldsymbol{\Sigma}|^{(\nu+|\boldsymbol{\tau}|+1)/2}}\, d\boldsymbol{\Sigma} \\ &\propto [1 + \frac{(\mathbf{Z}-\mu_0\mathbf{1})^\top \boldsymbol{\Psi}_\phi^{-1}(\mathbf{Z}-\mu_0\mathbf{1})}{\nu-2}]^{-(\nu+|\boldsymbol{\tau}|)/2}, \end{aligned}$$

which can be recognized as the kernel of a MVT distribution. Therefore, the result holds.

□

Appendix B

Implementation Details

B.1 MCMC of Models for Cross-sectional Ordinal Regression

B.1.1 The General Model

The development of the posterior simulation method for the general model (2.10) relies heavily on effectively the same structure for the weights and atoms of the mixture model. The Pólya-Gamma data augmentation approach is used to update parameters defining both the weights and atoms, leading to conditionally conjugate update for all parameters. Denote the Pólya-Gamma distribution with shape parameter b and tilting parameter c by $PG(b, c)$. Specifically, for each $\mathbf{Y}_i$, we introduce two groups of Pólya-Gamma latent variables $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{i,L-1})$ and $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{i,C-1})$, where $\xi_{i\ell} \overset{i.i.d.}{\sim} PG(1, 0)$ and $\zeta_{ij} \overset{ind.}{\sim} PG(m_{ij}, 0)$. Proceeding to the joint posterior, the contribution from $\mathbf{Y}_i$ is given by

$$f(\mathbf{Y}_i \mid \{\boldsymbol{\beta}_{j\ell}\}, \mathcal{L}_i, \boldsymbol{\zeta}_i) \propto \prod_{j=1}^{C-1} \exp\{\frac{\zeta_{ij}}{2}(\mathbf{x}_i^T \boldsymbol{\beta}_{j\mathcal{L}_i} - \upsilon_{ij}/\zeta_{ij})^2\},$$

where $\upsilon_{ij} = Y_{ij} - \frac{m_{ij}}{2}$. Likewise, let $\iota_{i\ell} = \mathcal{L}_{i\ell} - \frac{1}{2}$, we can write the contribution from $\boldsymbol{\mathcal{L}}_i$ as

$$f(\boldsymbol{\mathcal{L}}_i \mid \{\boldsymbol{\gamma}_\ell\}, \boldsymbol{\xi}_i) \propto \prod_{\ell=1}^{L-1} \exp\{\frac{\xi_{i\ell}}{2}(\mathbf{x}_i^T \boldsymbol{\gamma}_\ell - \iota_{i\ell}/\xi_{i\ell})^2\}.$$

These expressions admit closed-form full conditional distributions for $\{\boldsymbol{\beta}_{j\ell}\}$ and $\{\boldsymbol{\gamma}_\ell\}$.

We outline the MCMC sampling algorithm for the full augmented model. This process can be achieved entirely with Gibbs updates, by iterating the following steps. For notation simplicity, we let $(\phi \mid -)$ denote the posterior full conditional distribution for parameter ϕ.

Step 1: update parameters in the atoms. In this step, we update two sets of parameters, $\{\boldsymbol{\beta}_{j\ell} : j = 1, \ldots, C-1, \ell = 1, \ldots, L\}$ and $\{\zeta_{ij} : i = 1, \ldots, n, j = 1, \ldots, C-1\}$. Denote the set of distinct values of the configuration variables by $\{\mathcal{L}_r^* : r = 1, \ldots, n^*\}$. Following Polson et al. (2013), it can be done by $(\boldsymbol{\beta}_{j\ell} \mid -) \sim N(\tilde{\boldsymbol{\mu}}_{j\ell}, \tilde{\Sigma}_{j\ell})$ and $(\zeta_{ij} \mid -) \sim PG(m_{ij}, \mathbf{x}_i^T \boldsymbol{\beta}_{jL_i})$, where

- if $\ell \notin \{\mathcal{L}_r^* : r = 1, \ldots, n^*\}$: $\quad \tilde{\boldsymbol{\mu}}_{j\ell} = \boldsymbol{\mu}_j$, $\quad \tilde{\Sigma}_{j\ell} = \Sigma_j$;
- if $\ell \in \{\mathcal{L}_r^* : r = 1, \ldots, n^*\}$: $\quad \tilde{\boldsymbol{\mu}}_{j\ell} = \tilde{\Sigma}_{j\ell}(X_\ell^T \boldsymbol{\upsilon}_\ell + \Sigma_j^{-1} \boldsymbol{\mu}_j)$, $\quad \tilde{\Sigma}_{j\ell} = (X_\ell^T \Omega_\ell X_\ell + \Sigma_j^{-1})^{-1}$.

Here X_ℓ is the matrix whose column vectors are given by $\{\mathbf{x}_i : \mathcal{L}_i = \ell\}$, Ω_ℓ is the diagonal matrix with diagonal elements $\{\zeta_{ij} : \mathcal{L}_i = \ell\}$, and $\boldsymbol{\upsilon}_\ell$ is the vector of $\{\upsilon_{ij} : \mathcal{L}_i = \ell\}$. Notice that updating $\{\boldsymbol{\beta}_{j\ell}\}$ can be run in parallel across categories $j = 1, \ldots, C-1$.

Step 2: update parameters in the weights. Similarly, we update $\{\boldsymbol{\gamma}_\ell : \ell = 1, \ldots, L-1\}$ and $\{\xi_{i\ell} : i = 1, \ldots, n, \ell = 1, \ldots, L-1\}$ from $(\boldsymbol{\gamma}_\ell \mid -) \sim$

$N(\tilde{\boldsymbol{\gamma}}_\ell, \tilde{\Gamma}_\ell)$ and $(\xi_{i\ell} \mid -) \sim PG(1, \mathbf{x}_i^T \boldsymbol{\gamma}_\ell)$, where $\tilde{\boldsymbol{\gamma}}_\ell = \tilde{\Gamma}_\ell(X_\ell^T \boldsymbol{\iota}_\ell + \Gamma_0^{-1}\boldsymbol{\gamma}_0)$ and $\tilde{\Gamma}_\ell = (X_\ell^T \Xi_\ell X_\ell + \Gamma_0^{-1})^{-1}$. We denote the diagonal matrix formed by $\{\xi_{i\ell} : \mathcal{L}_i = \ell\}$ as Ξ_ℓ, and the vector of$\{\iota_{i\ell} : \mathcal{L}_i = \ell\}$ as $\boldsymbol{\iota}_\ell$.

Step 3: update configuration variables. Update $\mathcal{L}_i$, for $i = 1, \ldots, n$ from

$$P(\mathcal{L}_i = \ell \mid -) = \frac{p_{i\ell} \prod_{j=1}^{C-1} Bin(Y_{ij} \mid m_{ij}, \varphi(\boldsymbol{x}_i^T \boldsymbol{\beta}_{j\ell}))}{\sum_{\ell=1}^{L} p_{i\ell} \prod_{j=1}^{C-1} Bin(Y_{ij} \mid m_{ij}, \varphi(\boldsymbol{x}_i^T \boldsymbol{\beta}_{j\ell}))}$$

where $\{p_{i\ell} : \ell = 1, \ldots, L\}$ are calculated as $p_{i1} = \varphi(\mathbf{x}_i^T \boldsymbol{\gamma}_1)$, $p_{i\ell} = \varphi(\mathbf{x}_i^T \boldsymbol{\gamma}_\ell) \prod_{h=1}^{\ell-1}(1 - \varphi(\mathbf{x}_i^T \boldsymbol{\gamma}_h))$, $\ell = 2, \ldots, L-1$, and $p_{iL} = \prod_{\ell=1}^{L-1} (1 - \varphi(\mathbf{x}_i^T \boldsymbol{\gamma}_\ell))$.

Step 4: update hyperparameters. By conjugacy, updating hyperparameters is standard. We update $\{\boldsymbol{\mu}_j\}$ and $\{\Sigma_j\}$ by $(\boldsymbol{\mu}_j \mid -) \sim N(\boldsymbol{\mu}_j^*, \Sigma_j/\kappa_j^*)$ and $(\Sigma_j \mid -) \sim IW(\nu_j^*, (\Lambda_j^*)^{-1})$, with the parameters given by

$$\begin{aligned}
&\boldsymbol{\mu}_j^* = \frac{\kappa_{0j}}{\kappa_{0j} + n^*}\boldsymbol{\mu}_{0j} + \frac{n^*}{\kappa_{0j} + n^*}\bar{\boldsymbol{\beta}}_j, \;\; \kappa_j^* = n^* + \kappa_{0j}, \;\; \nu_j^* = n^* + \nu_{0j} \\
&\bar{\boldsymbol{\beta}}_j = \frac{1}{n^*}\sum_{r=1}^{n^*} \boldsymbol{\beta}_{j\mathcal{L}_r^*}, \;\; S_j = \sum_{r=1}^{n^*} (\boldsymbol{\beta}_{j\mathcal{L}_r^*} - \bar{\boldsymbol{\beta}}_j)(\boldsymbol{\beta}_{j\mathcal{L}_r^*} - \bar{\boldsymbol{\beta}}_j)^T, \\
&\Lambda_j^* = \Lambda_{0j} + S_j + \frac{n^*\kappa_{0j}}{n^* + \kappa_{0j}}(\bar{\boldsymbol{\beta}}_j - \boldsymbol{\mu}_{0j})(\bar{\boldsymbol{\beta}}_j - \boldsymbol{\mu}_{0j})^T.
\end{aligned}$$

We refer to the above process as the "general process". From the connection discussed in Section 2.3, the Gibbs sampler for the two simpler models are straightforwardly adapted from the general process.

B.1.2 The Common-weights Model

In the scenario that a common-weights model is adopted, the mixing weights and the configuration variables are determined by

$$\mathcal{L}_i \mid \boldsymbol{\omega} \sim \sum_{\ell=1}^{L} \omega_l \delta_\ell(\mathcal{L}_i),\ \boldsymbol{\omega} \mid \alpha \sim f(\boldsymbol{\omega} \mid \alpha),\ \alpha \sim Gamma(a_\alpha, b_\alpha),$$

where $f(\boldsymbol{\omega} \mid \alpha)$ stands for a special case of the generalized Dirichlet distribution

$$f(\boldsymbol{\omega} \mid \alpha) = \alpha^{L-1}\omega_L^{\alpha-1}(1-\omega_1)^{-1}(1-(\omega_1+\omega_2))^{-1}\ldots(1-\sum_{\ell=1}^{L-2}\omega_\ell)^{-1},$$

while the atoms are the same as in the general model. Hence, we only need to introduce the group of Pólya-Gamma latent variables $\{\boldsymbol{\zeta}_i : i = 1, \ldots, n\}$, which enable the same conjugate update in sampling atoms related parameters. We keep **Step 1** and **Step 4** in the general process, whereas the other two steps are replaced by:

Step 2*: update parameters in the weights. The parameters to be updated in this step involve $\{\omega_\ell : \ell = 1, \ldots, L-1\}$ and α. From Ishwaran and James (2001), it can be done by sample $V_\ell^* \overset{ind.}{\sim} Beta(1 + M_\ell, \alpha + \sum_{h=\ell+1}^{L} M_h)$ for $\ell = 1, \ldots, L-1$. Then let $\omega_1 = V_1^*$, $\omega_\ell = V_\ell^* \prod_{h=1}^{\ell-1}(1-V_h^*)$, $\ell = 2, \ldots, L-1$ and $\omega_L = 1 - \sum_{\ell=1}^{L-1}\omega_\ell$. In addition, a new sample of α is obtained from $(\alpha \mid -) \sim Gamma(a_\alpha + L - 1, b_\alpha - \sum_{\ell=1}^{L-1}\log(1-V_\ell^*))$.

Step 3*: update configuration variables. Update $\mathcal{L}_i$, $i = 1, \ldots, n$, from

$$P(\mathcal{L}_i = \ell \mid -) = \frac{\omega_\ell \prod_{j=1}^{C-1} Bin(Y_{ij} \mid m_{ij}, \varphi(\boldsymbol{x}_i^T \boldsymbol{\beta}_{j\ell}))}{\sum_{\ell=1}^{L} \omega_\ell \prod_{j=1}^{C-1} Bin(Y_{ij} \mid m_{ij}, \varphi(\boldsymbol{x}_i^T \boldsymbol{\beta}_{j\ell}))}.$$

B.1.3 The Common-atoms Model

If one choose to fit the common-atoms model, the linear regression terms in the atoms are simplified by $\theta_{j\ell}$ with prior $\theta_{j\ell} \overset{ind.}{\sim} N(\mu_j, \sigma_j^2)$, $j = 1, \ldots, C-1$ and $\ell = 1, \ldots, L$. We replace **Step 1** and **Step 4** of the general process with the following alternatives, while the other steps remain the same.

Step 1*: update parameters in the atoms. The two sets of parameters $\{\theta_{j\ell} : j = 1, \ldots, C-1, \ell = 1, \ldots, L\}$ and $\{\zeta_{ij} : i = 1, \ldots, n, j = 1, \ldots, C-1\}$ are now updated by $(\theta_{j\ell} \mid -) \sim N(\tilde{\mu}_{j\ell}, \tilde{\sigma}^2_{j\ell})$ and $(\zeta_{ij} \mid -) \sim PG(m_{ij}, \theta_{j\mathcal{L}_i})$, where

- if $\ell \notin \{\mathcal{L}_r^* : r = 1, \ldots, n^*\}$: $\tilde{\mu}_{j\ell} = \mu_j$, $\tilde{\sigma}^2_{j\ell} = \sigma_j^2$,
- if $\ell \in \{\mathcal{L}_r^* : r = 1, \ldots, n^*\}$: $\tilde{\mu}_{j\ell} = \tilde{\sigma}_j^2(\sum_{\{i:\mathcal{L}_i=\ell\}} \upsilon_{ij} + \mu_j/\sigma_j^2)$, $\tilde{\sigma}^2_{j\ell} = \sigma_j^2/(\sigma_j^2 \sum_{\{i:\mathcal{L}_i=\ell\}} \zeta_{ij} + 1)$.

Step 4*: update hyperparameters. That is, we update $\{\mu_j : j = 1, \ldots, C-1\}$ and $\{\sigma_j^2 : j = 1, \ldots, C-1\}$ by $(\mu_j \mid -) \sim N(\mu_j^*, \sigma_j^2/\nu_j^*)$ and $(\Sigma_j \mid -) \sim IW(\nu_j^*, (\Lambda_j^*)^{-1})$, where

$$\mu_j^* = \frac{\nu_{0j}\mu_{0j} + n^*\bar{\theta}_j}{\nu_{0j} + n^*}, \;\; \nu_j^* = n^* + \nu_{0j}, \;\; a_j^* = a_j + n^*/2 \;\; \bar{\theta}_j = \frac{1}{n^*}\sum_{r=1}^{n^*} \theta_{jr},$$
$$b_j^* = b_j + \frac{1}{2}\sum_{r=1}^{n^*}(\theta_{jr} - \bar{\theta}_j)^2 + \frac{n^*\nu_{0j}}{n^* + \nu_{0j}}\frac{(\bar{\theta}_j - \mu_{0j})^2}{2}.$$

Finally, for notation consistency, we should also replace the terms $\mathbf{x}_i^T\boldsymbol{\beta}_{j\ell}$ with $\theta_{j\ell}$ in **Step 3**, while keeping the same updating mechanism.

With the posterior samples for model parameters drawn by the MCMC mechanism described above, we can obtain full inference for any regression functional of interest. For instance, for any $j = 1, \ldots, C$, posterior realizations for the marginal

probability response curve, $\Pr(\mathbf{Y} = j \mid G_{\mathbf{x}})$, can be computed over a grid in $\mathbf{x}$ via

$$\sum_{\ell=1}^{L} \left\{ \varphi(\mathbf{x}^T \boldsymbol{\gamma}_\ell^{(t)}) \prod_{h=1}^{\ell-1} (1 - \varphi(\mathbf{x}^T \boldsymbol{\gamma}_h^{(t)})) \right\} \left\{ \varphi(\mathbf{x}^T \boldsymbol{\beta}_{j\ell}^{(t)}) \prod_{k=1}^{j-1} [1 - \varphi(\mathbf{x}^T \boldsymbol{\beta}_{k\ell}^{(t)})] \right\}$$

where $\varphi(\mathbf{x}^T \boldsymbol{\gamma}_L^{(t)}) = \varphi(\mathbf{x}^T \boldsymbol{\beta}_{C\ell}^{(t)}) \equiv 1$, and the superscript (t) indicates the tth posterior sample for the model parameters.

B.2 MCMC of Models with Overdispersed Kernel

In this section, we design the posterior simulation steps for the "Gen-LNB" model, and discuss its modification to accommodate for the "CW-Bin" model. Consider the data $\{(x_d, \mathbf{Y}_{di}) : d = 1, \ldots, N; i = 1, \ldots, n_d\}$. For the continuous mixing structure in the kernel, and the enveloping discrete mixing structure, we introduce latent variables $\{\boldsymbol{\psi}_{di} = (\psi_{di1}, \psi_{di2})\}$, and configuration variables $\{\mathcal{L}_{di}\}$. The model is then formulated hierarchically as

$$\begin{aligned}
(R_{di}, y_{di}) \mid \boldsymbol{\psi}_{di} &\overset{ind.}{\sim} Bin(R_{di} \mid m_{di}, \varphi(\psi_{di1})) Bin(y_{di} \mid m_{di} - R_{di}, \varphi(\psi_{di2})) \\
\boldsymbol{\psi}_{di} \mid \{\boldsymbol{\beta}_{j\ell}\}, \mathcal{L}_{di}, \boldsymbol{\sigma^2} &\overset{ind.}{\sim} \prod_{j=1}^{2} N(\psi_{dij} \mid \mathbf{x}_d^T \boldsymbol{\beta}_{j\mathcal{L}_{di}}, \sigma_j^2), \;\; d = 1, \ldots, N \;\; i = 1, \ldots, n_d \\
\mathcal{L}_{di} \mid \{\boldsymbol{\gamma}_\ell\} &\overset{ind.}{\sim} \sum_{\ell=1}^{L} p_{d\ell}\, \delta_\ell(\mathcal{L}_i), \;\; d = 1, \ldots, N \;\; i = 1, \ldots, n_d \\
\boldsymbol{\beta}_{j\ell} \mid (\boldsymbol{\mu}_j, \Sigma_j) &\overset{ind.}{\sim} N(\boldsymbol{\mu}_j, \Sigma_j), \;\; j = 1, 2, \;\; \ell = 1, \ldots, L \\
\boldsymbol{\gamma}_\ell &\overset{i.i.d.}{\sim} N(\boldsymbol{\gamma}_0, \Gamma_0), \;\; \ell = 1, \ldots, L-1 \\
(\boldsymbol{\mu}_j, \Sigma_j) &\overset{ind.}{\sim} N(\boldsymbol{\mu}_j \mid \boldsymbol{\mu}_{0j}, \Sigma_j/\kappa_{0j})\, IW(\Sigma_j \mid \nu_{0j}, \Lambda_{0j}^{-1}), \;\; j = 1, 2 \\
\sigma_j^2 &\overset{i.i.d.}{\sim} IG(a_\sigma, b_\sigma), \;\; j = 1, 2
\end{aligned}$$

where $p_{d\ell} = \varphi(\mathbf{x}_d^T \boldsymbol{\gamma}_\ell) \prod_{h=1}^{\ell-1}(1-\varphi(\mathbf{x}_d^T \boldsymbol{\gamma}_h))$, for $\ell = 1, \ldots, L-1$, and $p_{dL} = \prod_{\ell=1}^{L-1}(1-\varphi(\mathbf{x}_d^T \boldsymbol{\gamma}_\ell))$.

The hierarchical model formulation reminds us the Pólya-Gamma data augmentation approach (Polson et al., 2013), which is the key to conditionally conjugate updates for all parameters. For each response (R_{di}, y_{di}), we introduce two groups of Pólya-Gamma latent variables $\boldsymbol{\xi}_{di} = (\xi_{di1}, \ldots, \xi_{di,L-1})$ and $\boldsymbol{\zeta}_{di} = (\zeta_{di1}, \zeta_{di2})$, where $\xi_{di\ell} \overset{i.i.d.}{\sim} PG(1,0)$ and $\zeta_{di1} \sim PG(m_{di}, 0)$, $\zeta_{di2} \sim PG(m_{di} - R_{di}, 0)$, independently. Here $PG(b,c)$ denotes the Pólya-Gamma distribution with shape parameter b and tilting parameter c.

We outline the MCMC sampling algorithm for the full augmented model. This process can be achieved entirely with Gibbs updates, by iterating the following steps. For notation simplicity, we let $(\phi \mid -)$ denote the posterior full conditional distribution for parameter ϕ.

Step 1: update parameters in the atoms. In this step, we update two sets of parameters, $\{\boldsymbol{\psi}_{di} : d = 1, \ldots, N, i = 1, \ldots, n_d\}$, $\{\boldsymbol{\zeta}_{di} : d = 1, \ldots, N, i = 1, \ldots, n_d\}$ and $\{\boldsymbol{\beta}_{j\ell} : j = 1, 2, \ell = 1, \ldots, L\}$. Denote the set of distinct values of the configuration variables by $\{\mathcal{L}_r^* : r = 1, \ldots, n^*\}$. Following Polson et al. (2013), it can be done by $(\psi_{dij} \mid -) \sim N(\phi_{dij}, \tau_{dij}^2)$, with $\phi_{di1} = \frac{\mathbf{x}^\top \boldsymbol{\beta}_{1\mathcal{L}_{di}} + \sigma_1^2(R_{di} - \frac{m_{di}}{2})}{1+\sigma_1^2 \zeta_{di1}}$, $\phi_{di2} = \frac{\mathbf{x}^\top \boldsymbol{\beta}_{2\mathcal{L}_{di}} + \sigma_2^2(y_{di} - \frac{m_{do} - R_{di}}{2})}{1+\sigma_2^2 \zeta_{di2}}$, $\tau_{dij}^2 = \frac{\sigma_j^2}{1+\sigma_j^2 \zeta_{dij}}$, $j = 1, 2$. It is then followed by updating $(\zeta_{di1} \mid -) \sim PG(m_{di}, \psi_{di1})$ and $(\zeta_{di2} \mid -) \sim PG(m_{di} - R_{di}, \psi_{di2})$. Finally, we update $\boldsymbol{\beta}_{j\ell}$ by sampling from $(\boldsymbol{\beta}_{j\ell} \mid -) \sim$

$N(\tilde{\boldsymbol{\mu}}_{j\ell}, \tilde{\Sigma}_{j\ell})$, where

$$\begin{aligned}
&\bullet \text{ if } \ell \notin \{\mathcal{L}_r^* : r = 1, \ldots, n^*\}: \quad \tilde{\boldsymbol{\mu}}_{j\ell} = \boldsymbol{\mu}_j, \\
&\qquad\qquad\qquad\qquad\qquad\quad \tilde{\Sigma}_{j\ell} = \Sigma_j; \\
&\bullet \text{ if } \ell \in \{\mathcal{L}_r^* : r = 1, \ldots, n^*\}: \quad \tilde{\boldsymbol{\mu}}_{j\ell} = \tilde{\Sigma}_{j\ell}(\tfrac{X_\ell^T \boldsymbol{\psi}_\ell}{\sigma_j^2} + \Sigma_j^{-1}\boldsymbol{\mu}_j), \\
&\qquad\qquad\qquad\qquad\qquad\quad \tilde{\Sigma}_{j\ell} = (\tfrac{X_\ell^T X_\ell}{\sigma_j^2} + \Sigma_j^{-1})^{-1}.
\end{aligned}$$

Here X_ℓ is the matrix whose column vectors are given by $\{\mathbf{x}_d : \mathcal{L}_{di} = \ell\}$, and $\boldsymbol{\psi}_\ell$ is the vector of $\{\psi_{dij} : \mathcal{L}_{di} = \ell\}$. Notice that updating $\{\boldsymbol{\beta}_{j\ell}\}$ can be run in parallel across categories $j = 1, \ldots, C-1$.

Step 2: update parameters in the weights. Similarly, we update $\{\boldsymbol{\gamma}_\ell : \ell = 1, \ldots, L-1\}$ and $\{\xi_{i\ell} : i = 1, \ldots, n, \ell = 1, \ldots, L-1\}$ from $(\boldsymbol{\gamma}_\ell \mid -) \sim N(\tilde{\boldsymbol{\gamma}}_\ell, \tilde{\Gamma}_\ell)$ and $(\xi_{i\ell} \mid -) \sim PG(1, \mathbf{x}_i^T\boldsymbol{\gamma}_\ell)$, where $\tilde{\boldsymbol{\gamma}}_\ell = \tilde{\Gamma}_\ell(X_\ell^T\boldsymbol{\iota}_\ell + \Gamma_0^{-1}\boldsymbol{\gamma}_0)$ and $\tilde{\Gamma}_\ell = (X_\ell^T\Xi_\ell X_\ell + \Gamma_0^{-1})^{-1}$. We denote the diagonal matrix formed by $\{\xi_{i\ell} : \mathcal{L}_i = \ell\}$ as Ξ_ℓ, and the vector of$\{\iota_{i\ell} : \mathcal{L}_i = \ell\}$ as $\boldsymbol{\iota}_\ell$.

Step 3: update configuration variables. Update $\mathcal{L}_i$, for $i = 1, \ldots, n$ from

$$P(\mathcal{L}_i = \ell \mid -) = \frac{p_{i\ell}\prod_{j=1}^{C-1} Bin(Y_{ij} \mid m_{ij}, \varphi(\boldsymbol{x}_i^T\boldsymbol{\beta}_{j\ell}))}{\sum_{\ell=1}^{L} p_{i\ell}\prod_{j=1}^{C-1} Bin(Y_{ij} \mid m_{ij}, \varphi(\boldsymbol{x}_i^T\boldsymbol{\beta}_{j\ell}))}$$

where $\{p_{i\ell} : \ell = 1, \ldots, L\}$ are calculated as $p_{i1} = \varphi(\mathbf{x}_i^T\boldsymbol{\gamma}_1)$, $p_{i\ell} = \varphi(\mathbf{x}_i^T\boldsymbol{\gamma}_\ell)\prod_{h=1}^{\ell-1}(1-\varphi(\mathbf{x}_i^T\boldsymbol{\gamma}_h))$, $\ell = 2, \ldots, L-1$, and $p_{iL} = \prod_{\ell=1}^{L-1}(1-\varphi(\mathbf{x}_i^T\boldsymbol{\gamma}_\ell))$.

Step 4: update overdispersion parameters. The posterior full conditional distribution of σ_j^2 is given by $(\sigma_j^2 \mid -) \sim IG(a_j^*, b_j^*)$, where

$$a_j^* = a_\sigma + \frac{n_j^*}{2}, \;\; b_j^* = b_\sigma + \frac{\sum_{d=1}^{N}\sum_{i=1}^{n_d}(\psi_{dij} - \mathbf{x}_d^\top\boldsymbol{\beta}_{j\mathcal{L}_{di}})}{2},$$

where $n_1^* = \sum_{d=1}^{N} n_d$ and $n_2^* = \sum_{d=1}^{N}\sum_{i=1}^{n_d}\mathbf{1}(m_{di} - R_{di} > 0)$.

Step 5: update hyperparameters. By conjugacy, updating hyperparameters is standard. We update $\{\boldsymbol{\mu}_j\}$ and $\{\Sigma_j\}$ by $(\boldsymbol{\mu}_j \mid -) \sim N(\boldsymbol{\mu}_j^*, \Sigma_j/\kappa_j^*)$ and $(\Sigma_j \mid -) \sim IW(\nu_j^*, (\Lambda_j^*)^{-1})$, with the parameters given by

$$\boldsymbol{\mu}_j^* = \frac{\kappa_{0j}}{\kappa_{0j}+n^*}\boldsymbol{\mu}_{0j} + \frac{n^*}{\kappa_{0j}+n^*}\bar{\boldsymbol{\beta}}_j, \;\; \kappa_j^* = n^* + \kappa_{0j}, \;\; \nu_j^* = n^* + \nu_{0j}$$
$$\bar{\boldsymbol{\beta}}_j = \frac{1}{n^*}\sum_{r=1}^{n^*} \boldsymbol{\beta}_{j\mathcal{L}_r^*}, \;\; S_j = \sum_{r=1}^{n^*}(\boldsymbol{\beta}_{j\mathcal{L}_r^*} - \bar{\boldsymbol{\beta}}_j)(\boldsymbol{\beta}_{j\mathcal{L}_r^*} - \bar{\boldsymbol{\beta}}_j)^T,$$
$$\Lambda_j^* = \Lambda_{0j} + S_j + \frac{n^*\kappa_{0j}}{n^*+\kappa_{0j}}(\bar{\boldsymbol{\beta}}_j - \boldsymbol{\mu}_{0j})(\bar{\boldsymbol{\beta}}_j - \boldsymbol{\mu}_{0j})^T.$$

The posterior sampling algorithm for the "CW-LNB" model is adapt from this sampling scheme, with **Step 2** and **Step 3** been replaced by **Step** 2^* and **Step** 3^* described in Appendix B.1, while keeping the other steps unchanged.

B.3 MCMC of Models for Longitudinal Ordinal Responses

Based on the joint posterior distributions derived from (4.4), we design the MCMC sampling algorithm for the proposed model with binary responses. This process can be achieved entirely with Gibbs updates, by iterating the following steps. For notation simplicity, we let $(\phi \mid -)$ denote the posterior full conditional distribution for parameter ϕ.

Step 1: For $i = 1, \ldots, n$ update $\boldsymbol{\mathcal{Z}}_i$ from $N(\mathbf{m}_i, \boldsymbol{\mathcal{V}}_i)$, where $\boldsymbol{\mathcal{V}}_i = (\Omega_i + (1/\sigma_\epsilon^2)\mathbf{I})^{-1}$, and $\mathbf{m}_i = \boldsymbol{\mathcal{V}}_i(\boldsymbol{\lambda}_i + (1/\sigma_\epsilon^2)\mathbf{Z}_i)$. Here Ω_i denote the diagonal matrix of $\boldsymbol{\xi}_i$, and $\boldsymbol{\lambda}_i = (Y_{i1} - 1/2, \ldots, Y_{iT_i} - 1/2)^\top$.

Step 2: Update the Pólya-Gamma random variables ξ_{it} by sample from $PG(1, \mathcal{Z}_{it})$, for $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$.

Step 3: Update σ_ϵ^2 by sample from $IG(a_\epsilon + \sum_{i=1}^n T_i/2, b_\epsilon + \sum_{i=1}^n (\boldsymbol{\mathcal{Z}}_i - \mathbf{Z}_i)^\top (\boldsymbol{\mathcal{Z}}_i - \mathbf{Z}_i)/2)$.

Step 4: Update $\tilde{\mathbf{Z}}_i$ for $i = 1, \ldots, n$,

- In the case that all the subjects having observations on a common grid, $\mathbf{Z}_i^*$ vanishes and $\tilde{\mathbf{Z}}_i = \mathbf{Z}_i$. It has full conditional distribution $\mathbf{Z}_i \mid - \sim N(\tilde{\boldsymbol{\mu}}_i, \tilde{\mathbf{V}}_i)$, where $\tilde{\mathbf{V}}_i = ((1/\sigma_\epsilon^2)\mathbf{I} + \boldsymbol{\Sigma}^{-1})^{-1}$, and $\tilde{\boldsymbol{\mu}}_i = \tilde{\mathbf{V}}_i((1/\sigma_\epsilon^2)\boldsymbol{\mathcal{Z}}_{\boldsymbol{i}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$.
- In the case that the repeated measurements for the subjects are collected on uncommon grids, we first update $\mathbf{Z}_i^*$ from $N(\boldsymbol{\mu}_i^*, \boldsymbol{V}_i^*)$, where

$$\boldsymbol{\mu}_i^* = \mu(\boldsymbol{\tau}_i^*) + \Sigma(\boldsymbol{\tau}_i^*, \boldsymbol{\tau}_i)\Sigma(\boldsymbol{\tau}_i, \boldsymbol{\tau}_i)^{-1}(\mathbf{Z}_i - \mu(\boldsymbol{\tau}_i)) = \mathbf{B}_i\mathbf{Z}_i - \mathbf{u}_i,$$
$$\boldsymbol{V}_i^* = \Sigma(\boldsymbol{\tau}_i^*, \boldsymbol{\tau}_i^*) - \Sigma(\boldsymbol{\tau}_i^*, \boldsymbol{\tau}_i)\Sigma(\boldsymbol{\tau}_i, \boldsymbol{\tau}_i)^{-1}\Sigma(\boldsymbol{\tau}_i, \boldsymbol{\tau}_i^*),$$

  with $\mathbf{B}_i = \Sigma(\boldsymbol{\tau}_i^*, \boldsymbol{\tau}_i)\Sigma(\boldsymbol{\tau}_i, \boldsymbol{\tau}_i)^{-1}$ and $\mathbf{u}_i = \mathbf{B}_i\mu(\boldsymbol{\tau}_i) - \mu(\boldsymbol{\tau}_i^*)$.

  Then, to update $\mathbf{Z}_i$, we sample from $N(\tilde{\boldsymbol{\mu}}_i, \tilde{\mathbf{V}}_i)$, where

$$\tilde{\mathbf{V}}_i = [(1/\sigma_\epsilon^2)\mathbf{I} + \Sigma(\boldsymbol{\tau}_{\boldsymbol{i}}, \boldsymbol{\tau}_{\boldsymbol{i}})^{-1} + \mathbf{B}_i^T(\boldsymbol{V}_i^*)^{-1}\mathbf{B}_i]^{-1},$$
$$\tilde{\boldsymbol{\mu}}_i = \tilde{\mathbf{V}}_i[(1/\sigma_\epsilon^2)\boldsymbol{\mathcal{Z}}_i + \Sigma(\boldsymbol{\tau}_{\boldsymbol{i}}, \boldsymbol{\tau}_{\boldsymbol{i}})^{-1}\mu(\tau_i) + \mathbf{B}_i^T(\boldsymbol{V}_i^*)^{-1}(\mathbf{u}_i + \mathbf{Z}_i^*)].$$

Step 5: Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ jointly by sample from $N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}/\kappa^*)$ and $IW(\nu^*, \boldsymbol{\Psi}^*)$, respectively, with

$$\boldsymbol{\mu}^* = \frac{\kappa}{\kappa + n}\boldsymbol{\mu}_0 + \frac{n}{\kappa + n}\tilde{\mathbf{Z}}^m, \quad \kappa^* = n + \kappa, \quad \nu^* = n + \nu$$
$$\boldsymbol{\Psi}^* = \boldsymbol{\Psi} + S + \frac{n\kappa}{n + \kappa}(\tilde{\mathbf{Z}}^m - \boldsymbol{\mu}_0)(\tilde{\mathbf{Z}}^m - \boldsymbol{\mu}_0)^T, \quad S = \sum_{i=1}^n (\tilde{\mathbf{Z}}_i - \tilde{\mathbf{Z}}^m)(\tilde{\mathbf{Z}}_i - \tilde{\mathbf{Z}}^m)^{top},$$

where $\tilde{\mathbf{Z}}^m$ denote the mean of $\{\tilde{\mathbf{Z}}_i\}_{i=1}^n$.

Step 6: Update μ_0 from $N(a_\mu^*, b_\mu^*)$, where $b_\mu^* = [\mathbf{1}^\top[(\nu-3)\boldsymbol{\Sigma}]^{-1}\mathbf{1} + \frac{1}{b_\mu}]^{-1}$, and $a_\mu^* = b_\mu^*[\mathbf{1}^\top[(\nu-3)\boldsymbol{\Sigma}]^{-1}\boldsymbol{\mu} + \frac{a_\mu}{b_\mu}]$.

Step 7: Update σ^2 from Gamma$(a_\sigma + \frac{(\nu+|\boldsymbol{\tau}|-1)|\boldsymbol{\tau}|}{2}, b_\sigma + \frac{1}{2}tr(\boldsymbol{\Psi}_\rho\boldsymbol{\Sigma}^{-1}))$. Here $\boldsymbol{\Psi}_\rho$ denotes the correlation matrix $\boldsymbol{\Psi}_\phi/\sigma^2$.

Step 8: Using the Griddy-Gibbs sampler by Ritter and Tanner (1992), update ρ from

$$P(\rho = c_l \mid -) = \frac{|\boldsymbol{\Psi}_{c_l}|^{(\nu+|\boldsymbol{\tau}|-1)/2}\exp(-\frac{1}{2}tr(\boldsymbol{\Psi}_{c_l}\boldsymbol{\Sigma}^{-1}))}{\sum_{l=1}^{G}|\boldsymbol{\Psi}_{c_l}|^{(\nu+|\boldsymbol{\tau}|-1)/2}\exp(-\frac{1}{2}tr(\boldsymbol{\Psi}_{c_l}\boldsymbol{\Sigma}^{-1}))},$$

where $c_1, \ldots, c_G$ are grid points on a plausible region of ρ and $\boldsymbol{\Psi}_{c_l}$ denotes the correlation matrix when ρ taking the value c_l.

Step 9: Using the Griddy-Gibbs sampler, update ν from

$$P(\nu = c_l \mid -) = \frac{N(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, (c_l-3)\boldsymbol{\Sigma})IW(\boldsymbol{\Sigma} \mid c_l + |\boldsymbol{\tau}| - 1, \boldsymbol{\Psi}_\phi)}{\sum_{l=1}^{G}N(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, (c_l-3)\boldsymbol{\Sigma})IW(\boldsymbol{\Sigma} \mid c_l + |\boldsymbol{\tau}| - 1, \boldsymbol{\Psi}_\phi)}.$$

where $c_1, \ldots, c_G$ are grid points on a plausible region of ν.

Bibliography

Agresti, A. (2010), *Analysis of Ordinal Categorical Data*, Hoboken, NJ, USA: Wiley, second edition.

— (2013), *Categorical Data Analysis*, Hoboken, NJ, USA: Wiley, third edition.

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

Allen, B. C., Kavlock, R. J., Kimmel, C. A., and Faustman, E. M. (1994), "Dose-response assessment for developmental toxicity. II. Comparison of generic benchmark dose estimates with no observed adverse effect levels." *Fundamental and applied toxicology*, 23, 487–495.

Bao, J. and Hanson, T. (2015), "Bayesian Nonparametric Multivariate Ordinal Regression," *Canadian Journal of Statistics*, 43, 337–357.

Barcella, W., De Iorio, M., and Malone-Lee, J. (2018), "Modelling Correlated Binary Variables: An Application to Lower Urinary Tract Symptoms," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 67, 1083–1100.

Barrientos, A. F., Jara, A., and Quintana, F. A. (2012), "On the Support of MacEachern's Dependent Dirichlet Processes and Extensions," *Bayesian Analysis*, 7, 277–310.

Bartolucci, F., Lupparelli, M., and Montanari, G. E. (2009), "Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes," *Annals of Applied Statistics*, 3, 611 – 636.

Basu, S. and Chib, S. (2003), "Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models," *Journal of the American Statistical Association*, 98, 224–235.

Bender, R. and Grouven, U. (1998), "Using Binary Logistic Regression Models for Ordinal Data with Non-proportional Odds," *Journal of Clinical Epidemiology*, 51, 809–816.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877.

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

Boes, S. and Winkelmann, R. (2006), "Ordered Response Models," *Allgemeines Statistisches Archiv*, 90, 167–181.

Chib, S. and Greenberg, E. (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. B. (2009), "Nonparametric Bayes conditional distribution modeling with variable selection," *Journal of the American Statistical Association*, 104, 1646–1660.

Daniels, M. J. and Xu, D. (2020), "Bayesian Methods for Longitudinal Data with Missingness," in Lesaffre, E., Baio, G., and Boulanger, B. (editors), *Bayesian Methods in Pharmaceutical Research*, Chapman and Hall/CRC, 185–205.

Dawid, A. P. (1981), "Some Matrix-variate Distribution Theory: Notational Considerations and A Bayesian Application," *Biometrika*, 68, 265–274.

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

DeYoreo, M. and Kottas, A. (2018a), "Bayesian Nonparametric Modeling for Multivariate Ordinal Regression," *Journal of Computational and Graphical Statistics*, 27, 71–84.

— (2018b), "Modeling for Dynamic Ordinal Regression Relationships: An Application to Estimating Maturity of Rockfish in California," *Journal of the American Statistical Association*, 113, 68–80.

— (2018c), "Modeling for Dynamic Ordinal Regression Relationships: An Application to Estimating Maturity of Rockfish in California," *Journal of the American Statistical Association*, 113, 68–80.

— (2020), "Bayesian Nonparametric Density Regression for Ordinal Responses," in Fan, Y., Nott, D., Smith, M. S., and Dortet-Bernadet, J.-L. (editors), *Flexible Bayesian Regression Modelling*, Academic Press, 65–90.

Di Lucca, M. A., Guglielmi, A., Müller, P., and Quintana, F. A. (2013), "A Simple Class of Bayesian Nonparametric Autoregression Models," *Bayesian Analysis*, 8, 63 – 88.

Diggle, P. J. (1988), "An Approach to the Analysis of Repeated Measurements," *Biometrics*, 44, 959–971.

Dominici, F. and Parmigiani, G. (2001), "Bayesian semiparametric analysis of developmental toxicology data," *Biometrics*, 57, 150–157.

Donà, G., Preatoni, E., Cobelli, C., Rodano, R., and Harrison, A. J. (2009), "Application of Functional Principal Component Analysis in Race Walking: An Emerging Methodology," *Sports Biomechanics*, 8, 284–301.

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

Dunson, D. B. and Rodríguez, A. (2011), "Nonparametric Bayesian Models through Probit Stick-breaking Processes," *Bayesian Analysis*, 6, 145–177.

Ekström, J. (2011), "The Phi-coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate," Technical report, Department of Statistics, UCLA.

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

Fong, Y., Rue, H., and Wakefield, J. (2010), "Bayesian Inference for Generalized Linear Mixed Models," *Biostatistics*, 11, 397–412.

Fronczyk, K. and Kottas, A. (2014), "A Bayesian Nonparametric Modeling Framework for Developmental Toxicity Studies (with discussion)," *Journal of the American Statistical Association*, 109, 873–893.

— (2017), "Risk Assessment for Toxicity Experiments with Discrete and Continuous Outcomes: A Bayesian Nonparametric Approach," *Journal of Agricultural, Biological and Environmental Statistics*, 22, 585–601.

Fryer, D., Nguyen, H., and Orban, P. (2022), *Studentlife: Tidy Handling and Navigation of the Student-Life Dataset*. R package version 1.1.0.

Fung, K. Y., Marro, L., and Krewski, D. (1998), "A Comparison of Methods for Estimating the Benchmark Dose Based on Overdispersed Data from Developmental Toxicity Studies," *Risk Analysis*, 18, 329–342.

Gelfand, A. E. and Ghosh, S. K. (1998), "Model choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11.

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

Ghosal, S. and van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference*, Cambridge, UK: Cambridge University Press.

Ghosh, P. and Hanson, T. (2010), "A Semiparametric Bayesian Approach to Multivariate Longitudinal Data," *Australian & New Zealand Journal of Statistics*, 52, 275–288.

Gneiting, T. and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American statistical Association*, 102, 359–378.

Hall, P., Müller, H.-G., and Yao, F. (2008), "Modelling Sparse Generalized Longitudinal Observations with Latent Gaussian Processes," *Journal of the Royal Statistical Society: Series B*, 70, 703–723.

Hedeker, D., Mermelstein, R. J., Berbaum, M. L., and Campbell, R. T. (2009), "Modeling mood variation associated with smoking: an application of a heterogeneous mixed-effects model for analysis of ecological momentary assessment (EMA) data," *Addiction*, 104, 297–307.

Heiner, M. and Kottas, A. (2022), "Bayesian nonparametric density autoregression with lag selection," *Bayesian Analysis*, 17, 1245–1273.

Huggins, J., Kasprzak, M., Campbell, T., and Broderick, T. (2020), "Validated Variational Inference via Practical Posterior Error Bounds," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Hwang, B. S. and Pennell, M. L. (2014), "Semiparametric Bayesian joint modeling of a binary and continuous outcome with applications in toxicological risk assessment," *Statistics in Medicine*, 33, 1162–1175.

— (2018), "Semiparametric Bayesian joint modeling of clustered binary and continuous outcomes with informative cluster size in developmental toxicity assessment," *Environmetrics*, 29, e2526.

Ingrassia, S. and Costanzo, G. D. (2005), "Functional Principal Component Analysis of Financial Time Series," in Vichi, M., Monari, P., Mignani, S., and Montanari, A. (editors), *New Developments in Classification and Data Analysis*, Berlin, Heidelberg: Springer Berlin Heidelberg.

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

Jara, A., José García-Zattera, M., and Lesaffre, E. (2007), "A Dirichlet Process Mixture Model for the Analysis of Correlated Binary Responses," *Computational Statistics & Data Analysis*, 51, 5402–5415.

Jiang, L., Zhong, Y., Elrod, C., Natarajan, L., Knight, R., and Thompson, W. K. (2020), "BayesTime: Bayesian Functional Principal Components for Sparse Longitudinal Data," arXiv:2012.00579.

Kang, J. and Kottas, A. (2022), "Structured Mixture of Continuation-ratio Logits Models for Ordinal Regression," arXiv:2211.04034.

Kottas, A. and Fronczyk, K. (2013), "Flexible Bayesian modelling for clustered categorical responses in developmental toxicology," in Damien, P., Dellaportas, P., Polson, N. G., and Stephens, D. A. (editors), *Bayesian Theory and Applications*, Oxford, UK: Oxford University Press, 70–83.

Krewski, D. and Zhu, Y. (1995), "A Simple Data Transformation for Estimating Benchmark Doses in Developmental Toxicity Experiments," *Risk Analysis*, 15, 29–39.

Kuk, A. Y. C. (2004), "A Litter-Based Approach to Risk Assessment in Developmental Toxicity Studies via a Power Family of Completely Monotone Functions," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 53, 369–386.

Kunihama, T., Halpern, C. T., and Herring, A. H. (2019), "Non-parametric Bayes models for Mixed Scale Longitudinal Surveys," *Journal of the Royal Statistical Society: Series C*, 68, 1091–1109.

Lee, W.-S., Mangel, M., and Munch, S. B. (2017), "Developmental order of a secondary sexual trait reflects gonadal development in male sheepshead minnows (Cyprinodon variegatus)," *Evolutionary Ecology Research*, 18, 531–538.

Li, Y., Lin, X., and Müller, P. (2010), "Bayesian Inference in Semiparametric Mixed Models for Longitudinal Data," *Biometrics*, 66, 70–78.

Linderman, S., Johnson, M. J., and Adams, R. P. (2015), "Dependent Multinomial Models Made Easy: Stick-Breaking with the Pólya-gamma Augmentation," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, Cambridge, MA, USA: MIT Press.

MacEachern, S. N. (2000), "Dependent Dirichlet processes," Technical report, Department of Statistics, The Ohio State University.

Masters, G. N. (1982), "A Rasch model for partial credit scoring," *Psychometrika*, 47, 149–174.

Matuk, J., Herring, A. H., and Dunson, D. B. (2022), "Bayesian Functional Principal Component Analysis using Relaxed Mutually Orthogonal Processes," arXiv:2205.12361.

McCullagh, P. (1980), "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, 42, 109–127.

Mena, R. H., Ruggiero, M., and Walker, S. G. (2011), "Geometric stick-breaking processes for continuous-time Bayesian nonparametric modeling," *Journal of Statistical Planning and Inference*, 141, 3217–3230.

Molenberghs, G. and Verbeke, G. (2006), *Models for Discrete Longitudinal Data*, Springer Series in Statistics, Springer-Verlag.

Munch, S. B., Lee, W. S., Walsh, M., Hurst, T., Wasserman, B. A., Mangel, M., and Salinas, S. (2021), "A latitudinal gradient in thermal transgenerational plasticity and a test of theory," *Proceedings of the Royal Society B: Biological Sciences*, 288, 20210797.

Müller, P., Quintana, F., Jara, A., and Hanson, T. (2015), *Bayesian Nonparametric Data Analysis*, New York, NY: Springer.

Peyhardi, J., Trottier, C., and Guédon, Y. (2015), "A new specification of generalized linear models for categorical responses," *Biometrika*, 102, 889–906.

Pirjol, D. (2013), "The Logistic-normal Integral and Its Generalizations," *Journal of Computational and Applied Mathematics*, 237, 460–469.

Pollak, J. P., Adams, P., and Gay, G. (2011), "PAM: A Photographic Affect Meter for Frequent, in Situ Measurement of Affect," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models Using Pólya-Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349.

Quintana, F. A., Johnson, W. O., Waetjen, L. E., and Gold, E. B. (2016), "Bayesian Nonparametric Longitudinal Data Analysis," *Journal of the American Statistical Association*, 111, 1168–1181.

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022), "The Dependent Dirichlet Process and Related Models," *Statistical Science*, 37, 24–41.

Rasmussen, C. E. and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, The MIT Press.

Rigon, T. and Durante, D. (2021), "Tractable Bayesian Density Regression via Logit Stick-breaking Priors," *Journal of Statistical Planning and Inference*, 211, 131–142.

Ritter, C. and Tanner, M. A. (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler," *Journal of the American Statistical Association*, 87, 861–868.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion)." *Journal of the Royal Statistical Society B*, 71, 319–392.

Rue, H., Riebler, A. I., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017), "Bayesian computing with INLA: A review," *Annual Reviews of Statistics and Its Applications*, 395–421.

Russell, J. A. (1980), "A Circumplex Model of Affect." *Journal of Personality and Social Psychology*, 39, 1161–1178.

Ruwaard, J., Kooistra, L., and Thong, M. (2018), *Ecological Momentary Assessment in Mental Health Research: A Practical Introduction, with Examples in R*, APH Mental Health, 1st (build 2018-11-26) edition.

Salinas, S. and Munch, S. B. (2012), "Thermal legacies: transgenerational effects of temperature on growth in a vertebrate," *Ecology Letters*, 15, 159–163.

Schauberger, G. and Tutz, G. (2023), *catdata: Categorical Data*. R package version 1.2.3.

Shah, A., Wilson, A., and Ghahramani, Z. (2014), "Student-t Processes as Alternatives to Gaussian Processes," in Kaski, S. and Corander, J. (editors), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, Reykjavik, Iceland: PMLR.

Shaked, M. (1980), "On Mixtures from Exponential Families," *Journal of the Royal Statistical Society - Series B*, 42, 192–198.

Shamshoian, J., Şentürk, D., Jeste, S., and Telesca, D. (2020), "Bayesian Analysis of Longitudinal and Multidimensional Functional Data," *Biostatistics*, 23, 558–573.

Shiffman, S., Gwaltney, C. J., Balabanis, M. H., Liu, K. S., Paty, J. A., Kassel, J. D., Hickcox, M., and Gnys, M. (2009), "Immediate antecedents of cigarette smoking: an analysis from ecological momentary assessment." *Journal of Abnormal Psychology*, 111, 531–545.

Shiffman, S., Stone, A. A., and Hufford, M. R. (2008), "Ecological momentary assessment," *Annual Review of Clinical Psychology*, 4, 1–32.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Series in Statistics, Springer.

Tang, N.-S. and Duan, X.-D. (2012), "A Semiparametric Bayesian Approach to Generalized Partial Linear Mixed Models for Longitudinal Data," *Computational Statistics and Data Analysis*, 56, 4348–4365.

Tutz, G. (1991), "Sequential Models in Categorical Regression," *Computational Statistics and Data Analysis*, 11, 275–295.

— (2022), "Ordinal regression: A review and a taxonomy of models," *WIREs Computational Statistics*, 14, e1545.

U.S. EPA (1991), *Guidelines for Developmental Toxicity Risk Assessment*, U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC.

— (2012), *Benchmark Dose Technical Guidance*, U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC.

Van Der Linde, A. (2009), "A Bayesian latent variable approach to functional principal components analysis with binary and count data," *AStA Advances in Statistical Analysis*, 93, 307–333.

Verbeek, M. (2008), *A Guide to Modern Econometrics*, Hoboken, NJ, USA: Wiley, 3rd ed. edition.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A. T. (2014), "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, New York, NY, USA: Association for Computing Machinery.

Yang, J., Zhu, H., Choi, T., and Cox, D. D. (2016), "Smoothing and Mean–Covariance Estimation of Functional Data with a Bayesian Hierarchical Model," *Bayesian Analysis*, 11, 649 – 670.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590.

Zhao, X., Marron, J. S., and Wells, M. T. (2004), "The Functional Data Analysis View of Longitudinal Data," *Statistica Sinica*, 14, 789–808.