

RICE UNIVERSITY

**Bayesian Semiparametric and Flexible Models for Analyzing
Biomedical Data**

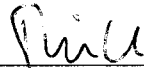
by

Luis G. León Novelo

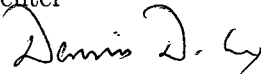
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

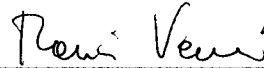
APPROVED, THESIS COMMITTEE:



Peter Müller, Professor, Director
Biostatistics, M. D. Anderson Cancer
Center



Dennis Cox, Professor, Chair
Statistics



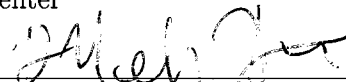
Marina Vannucci, Professor
Statistics



Fred Oswald, Associate Professor
Psychology



Kim-Anh Do, Professor
Biostatistics, M. D. Anderson Cancer
Center



Bekele Nebiyou, Professor
Biostatistics, M. D. Anderson Cancer
Center

HOUSTON, TEXAS

AUGUST, 2009

UMI Number: 3421146

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

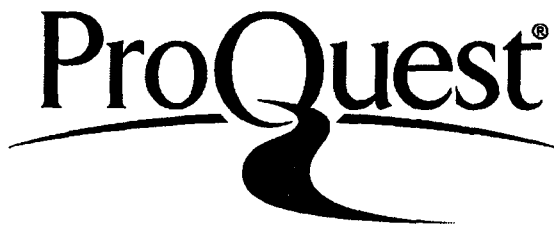
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3421146

Copyright 2010 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Bayesian Semiparametric and Flexible Models for Analyzing Biomedical Data

by

Luis G. León Novelo

In this thesis I develop novel Bayesian inference approaches for some typical data analysis problems as they arise with biomedical data. The common theme is the use of flexible and semi-parametric Bayesian models and computation intensive simulation-based implementations. In chapter 2, I propose a new approach for inference with multivariate ordinal data. The application concerns the assessment of toxicities in a phase III clinical trial. The method generalizes the ordinal probit model. It is based on flexible mixture models. In chapter 3, I develop a semi-parametric Bayesian approach for bio-panning phage display experiments. The nature of the model is a mixed effects model for repeated count measurements of peptides. I develop a non-parametric Bayesian random effects distribution and show how it can be used for the desired inference about organ-specific binding. In chapter 4, I introduce a variation of the product partition model with a non-exchangeable prior structure. The model is applied to estimate the success rates in a phase II clinical of patients with sarcoma. Each patient presents one subtype of the disease and subtypes are grouped by good, intermediate and poor prognosis. The prior model respects the varying prognosis across disease subtypes. Two subtypes with equal prognoses are more likely a priori to have similar success rates than two subtypes with different prognoses.

Acknowledgements

I would like to show my appreciation and gratitude to Peter Müller for guiding me and supporting me through my years at Rice. I am also grateful to Kim-Anh Do, Nebiyou Bekele and Fernando Quintana for letting me learn from them while working on this thesis. Thank you to my committee members for contributing to make this project what it is today. I am also grateful to Professors Raúl Rueda and Eduardo Gutiérrez for starting me on this academic journey.

Thank you to my mother Enna, grandmothers Enna and Josefa, grandfather Enrique, siblings Enrique, Alejandro, Oscar and Enna for all their support and love along the way.

Contents

Abstract	ii
Acknowledgements	iii
List of Figures	vii
List of Tables	xi
1 Introduction	1
1.1 Bayesian Inference	2
1.2 Non-parametric Bayesian Inference	2
1.3 Overview	4
1.4 Contributions of this Thesis	6
2 Assessing Toxicities in a Clinical Trial: Bayesian Inference for Multivariate Ordinal Data	7
2.1 Overview	7
2.2 Background	9
2.2.1 Clinical Trials	10
2.2.2 Binary Data Model	12
2.2.3 Multiresponse Categorical Data Model	14
2.3 A Phase III Clinical Trial	16
2.4 A Hierarchical Model for Ordinal Data Nested within Categories . . .	18

2.5	Priors, Posterior and Simulation Scheme	21
2.6	Applications	25
2.6.1	A Simulated Dataset	25
2.6.2	A Phase III Clinical Trial of Retinoid Isotretinoin	27
2.7	Discussion	30
3	Dirichlet Process Mixture Models for Discrete Human Data in Phage	
	Display Model	37
3.1	Overview	38
3.2	Background	41
3.2.1	Dirichlet Distribution	41
3.2.2	Dirichlet Process and Dirichlet Process Mixture Model	42
3.2.3	Gibbs Sampling Scheme for DPM	46
3.2.4	Example of Application of DPM in Density Estimation	52
3.3	Data	57
3.4	Model	60
3.4.1	A Semi-parametric Mixture of Poisson Model	60
3.4.2	Posterior Simulation	63
3.5	Selecting Significant Tripeptide-Tissue Pairs	66
3.6	A Simulation Study	69
3.7	Results	72
3.8	Discussion	78
4	Borrowing Strength with Hierarchical Models over Non-Exchangeable	
	Subpopulations	81
4.1	Outline	81
4.2	Data	83
4.3	Non-Exchangeable Product Partition Model	84
4.3.1	Model Definition	84

	vi
4.3.2 Some Properties of the NEPPM	87
4.4 Comparison and Operating Characteristics	89
4.4.1 Competing Models	89
4.4.2 Average Sample Size and Stopping Probabilities	95
4.5 Results	99
4.6 Discussion	99
Bibliography	108

List of Figures

2.1	Distribution of the latent random variable $v_i \sim N(x_i^t \beta, 1)$ according to Albert and Chib's model considering four categories, $K = 3$. The cutpoints θ_2 and θ_3 are random. The observed value z_i is indicator of the interval where v_i falls.	15
2.2	Illustration of the distribution of the latent variable v_{i1} in the model described in (2.2)-(2.4) when the covariate takes values -1 and 1 . Here we consider: $J = 1$ type of toxicity, no patient-specific random effect r_i , $G = 2$ components in the mixture of normals and three, $K = 2$, possible ordinal outcomes. In both mixtures, the darkly shaded, lightly shaded and white areas correspond to the probabilities π_{1k} of the ordinal outcome taking the values $0, 1$ and 2 , respectively.	22
2.3	Boxplots of the simulated posterior samples of the ordinal probit model parameters (β_j 's). Boxes corresponding to samples of β_j 's with 0.01-quantile greater than zero are shaded.	29
2.4	Estimated mixture of normal distribution of the latent variable v_{ij} conditioned on $r_i = 0$ for $j = 1, \dots, 7$. Shaded curve corresponds to placebo ($x = -1$) and dashed to isotretinoin ($x = 1$). Darkly and lightly shaded areas represent, respectively, the probability of no toxicity (π_{j0}) and toxicity at grade 1 (π_{j1}) under placebo.	30

3.1	Base measure $F_0=Ga(9, 2)$, posterior base measure $F_1 := E(F x^n)$, empirical cdf and set of simulated measures sampled from the posterior distribution of the DP($\alpha = 5, F_0$) when a sample $x^n: 1,2,2,3,3$ and 3 of this DP was observed.	44
3.2	Histogram of the house prices in Boston. The continuous line is the non-parametric estimation of the corresponding density function using the MDP model in (3.25)	56
3.3	Observed sequence of tripeptide-tissue pair counts across the three stages. Each line represents the three observed counts of a pair. The three panels depicts the pairs with: (a) non decreasing, (b) oscillating and (c) non increasing counts.	59
3.4	Histograms of a simulated sample of future observations of β and δ . They estimate the true probability density function (continuous line) of the distribution that generated the data	71
3.5	Thirty pairs with highest estimated values of \bar{m}_i in (3.36). Lower section presents counts for the three stages: circle, triangle and cross. The middle section shows the value of \bar{m}_i . The upper part depicts the posterior probability of increasing means across the three stages, p_i	74
3.6	Respectively, (a) and (b) nondecreasing and oscillating observed tripeptide-tissue pair counts across the three stages. In red the 219 selected pairs using the optimal rule (3.35) with a threshold value of 1.	75
3.7	Histograms of simulated samples of the final distribution of the parameters β_i and δ_i for three specific pairs.	76

- 4.1 Comparison of the estimated success rates under the four models under S0. The horizontal axis shows the $n = 12$ success rates under the simulation truth S_0 . The upper panel shows the absolute bias. The lower panel shows mean squared error. Both are arranged by true p_i . The five lines and labels correspond to the models NEPPM with $\gamma = 1$ [1], NEPPM with $\gamma = 1/2$ [2], parametric Hierarchical [3], HLRM [4] and separate [5] models. 93
- 4.2 Scenario S0. Coverage probabilities of the central 95% credible intervals under the NEPPM with $\gamma = 1$ vs. the HLRM. From left to right, the first three success rates correspond to poor prognosis ($x_i = -1$), the following six to intermediate ($x_i = 0$) and the last three to good prognosis ($x_i = 1$). The character sizes are proportional to the sample size N_i 94
- 4.3 Scenarios S1 through S4. Panels (a) through (d) summarize simulation under scenarios S1 through S4. Comparison of the estimated success rates under the NEPPM (star “*”) vs. the HLRM (circle “o”). The horizontal axis shows the $n = 12$ true success rates under the assumed scenario. The upper, medium and lower panels show absolute value of the bias, mean square error and coverage probability of the central 95% credible interval, respectively. All are arranged by subtype. Under S1 (panel (a)), from left to right, the first three success rates correspond to poor prognosis ($x_i = -1$), the following six to intermediate ($x_i = 0$) and the last three to good prognosis ($x_i = 1$). The point sizes are proportional to the sample size N_i 96

- 4.4 S1 through S4. Panels (a) through (d) show the average number of patients (\bar{N}_i) and early stopping probabilities (\bar{p}_i) under scenarios S1 through S4. Both summaries are with respect to repeat experimentation. Summaries are arranged by the simulation truth p_i , shown on the horizontal axis. In each panel, the upper part shows the average number of patients \bar{N}_i that enter into the study for each arm. The lower part shows the early stopping probability (\bar{p}_i). The stars (“*”) show summaries under the NEPPM. The circles (“o”) show summaries for the HLRM. The character size is proportional to N_i , the maximum sample size for each subtype. 98
- 4.5 Central 95% credible intervals for the success rates of the treatment in the sarcoma subtypes in the study when applying the proposed NEPPM with parameter $\gamma = 1$. Right square bracket marks the 5% percentile. The upper two and the rest CI’s correspond, respectively, to the two and ten sarcoma subtypes with good and intermediate prognoses. 100

List of Tables

2.1	Toxicity frequency for randomized eligible patients by study arms. In the placebo (Isotretinoin) group, 171 (427) out of 577 (589) patients exhibited some type of toxicity. In bold the proportion of patients in the study arm belonging to the cell.	33
2.2	Simulated data set, marginal posterior cell probabilities with central 95% credible intervals. In bold are the true cell probabilities.	34
2.3	Marginal posterior cell probabilities (central 95% credible intervals) of toxicity. The table only reports grades up to $k = 3$. The marginal probabilities for <i>grade G4 abnormal vision</i> are 0 (with 95% C.I. (0, 0.002)) under placebo, and 0.001 (with 95% C.I., (0, 0.002)) under isotretinoin.	35
2.4	Probability of different toxicities (at any grade; rows) conditional on the same patient having experienced other toxicities (at any grade; columns). The two numbers x/y in each cell report probabilities under placebo/isotretinoin. For comparison the diagonal reports in bold the marginal probabilities of exhibiting each type of toxicity.	36
3.1	Considered tripeptide-tissue pairs by using the rule (3.35) with a threshold value of 7. The expected values of FDR and FNR are 0.337 and 0.113 respectively. Pairs considered: 30.	77
4.1	Reported number of successes/trials for each one of the sarcoma subtypes.	84
4.2	Simulation truth p_i under five alternative scenarios	91

Chapter 1

Introduction

1.1 Bayesian Inference

Throughout this thesis I use the Bayesian paradigm for statistical inference. Bayesian inference is characterized by a joint probability model on all unknown quantities, including observable data y and parameters θ . Classical inference in contrast uses only probability models for y indexed by θ . Under the Bayesian paradigm, all relevant information after seeing the data is contained in the posterior distribution $p(\theta | y)$. The main challenges are the construction of appropriate prior probability models $p(\theta)$, and the often computationally intensive assessment of relevant summaries of the high dimensional posterior distribution $p(\theta | y)$.

Over the last two decades a barrage of new methods commonly known as Markov Chain Monte Carlo (MCMC) have been proposed to deal with the latter problem. Most Bayesian inference can be represented as posterior expectation of appropriate functions of the parameters. The main idea of MCMC is to approximate posterior expectations by ergodic averages over Markov chain simulations that are set up to have $p(\theta | y)$ as asymptotic distribution. These developments are well summarized in, among many other references, Cappé and Robert (2002) and Lopes and Gamerman (2006) .

1.2 Non-parametric Bayesian Inference

The second big challenge concerns the choice of the prior probability model. Conventional parametric priors are families of prior probability models $p(\theta | \eta)$ indexed by a *finite* dimensional parameter η . Typical examples of these types of priors are normal models, Beta distributions, etc. In many applications this assumption of finite dimensions turns out to be too restrictive. A typical situation is the specification of random effects distributions. Assuming a parametric random effects model implies a very homogeneous population of experimental units (patients, peptides, etc). It does not properly reflect the population heterogeneity that is typical for many biomedical

problems. Patients come with a wide variety of treatment histories, different genotypes, family histories, etc. Peptides are related to different biologic functions, have different interactions, etc. One approach to address this problem is the use of more flexible prior probability models. In particular, *non-parametric* Bayesian models have been used to generalize parametric models. A technical definition of a non-parametric Bayesian model $p(\eta)$ is a probability model that allows η to be infinite dimensional, like a random probability measure. Let $N(x; m, s)$ denote a normal kernel with moments (m, s) . For example,

$$p(\theta) = \int N(\theta; m, s) dG(m)$$

is a mixture of normals indexed by the mixing measure G (and a scale s). Here G is a random probability measure. The model is completed with a hyperprior $p(G)$ on G . Since the probability measure G is infinite dimensional, this is formally a non-parametric Bayesian model. Often also the resulting model for θ is referred to as “non-parametric”. Perhaps the most popular non-parametric Bayesian models are based on the Dirichlet process (DP) prior $p(G) = \text{DP}(\alpha, G_0)$. The DP is defined in Ferguson (1973) and Antoniak (1974). Good recent reviews of such models appear in Quintana and Müller (2004), and Walker, Damien, Laud and Smith (1999, JRSSB). The term *non-parametric* Bayesian inference could be considered a misnomer, since the defining property is the exact opposite, an infinite dimensional parameter space. But the terminology is traditional, and simply motivated by the fact that inference closely resembles traditional classical non-parametric inference, such as kernel density estimation. From a data analysis perspective, the infinite dimensional nature of the parameter space is not critical, and I refer to any similarly flexible probability model as *non-parametric*. In particular, inference based on mixture model generalization of underlying parametric models are usually considered “non-parametric” inference.

1.3 Overview

In chapter 2, I consider modeling and inference for ordinal outcomes nested within categorical responses. I propose a mixture of normal distributions for latent variables associated with the ordinal data. This mixture model allows us to fix without loss of generality the cutpoint parameters that link the latent variable with the observed ordinal outcome. Moreover, the mixture model is shown to be more flexible in estimating cell probabilities when compared to the traditional Bayesian ordinal probit regression model with random cutpoint parameters. I extend the model to account for possible dependence among the outcomes in different categories. I apply the model to a randomized phase III study to compare treatments on the basis of toxicities recorded by type of toxicity and grade within type. The data include the different (categorical) toxicity types exhibited in each patient. Each type of toxicity has an (ordinal) grade associated to it. The dependence among the different types of toxicity exhibited by the same patient is modeled by introducing patient-specific random effects.

In chapter 3, I discuss inference for a human phage display experiment with three stages. The data are tripeptide counts by tissue and stage. The primary aim of the experiment is to identify ligands that bind with high affinity to a given tissue. I formalize the research question as inference about the monotonicity of mean counts over stages. The inference goal is then to identify a list of peptide-tissue pairs with significant increase over stages. I develop a semi-parametric model as a mixture of Poisson distributions with a Dirichlet process prior on the mixing measure. The posterior distribution under this model allows the desired inference about the monotonicity of mean counts. However, the desired inference summary as a list peptide-tissue pairs with significant increase involves a massive multiplicity problem. I consider two alternative approaches to address this multiplicity issue. First I propose an approach based on the control of the posterior expected false discovery rate. I notice that the implied solution ignores the relative size of the increase. This motivates a second

approach based on a utility function that includes explicit weights for the size of the increase.

In chapter 4, I introduce a non-parametric Bayesian model for phase II clinical trial with patients presenting different subtypes of the disease under study. The subtypes are not a priori exchangeable. The lack of a priori exchangeability hinders the straightforward use of traditional hierarchical models to implement borrowing of strength across disease subtypes. We introduce instead a random partition model for the set of disease subtypes. All subtypes within the same cluster share a common success probability. The random partition model is a variation of the product partition model that allows us to model a non-exchangeable prior structure. This model is the categorical covariate version of the more general non exchangeable product partition model proposed in Müller et al. (2009). In particular the data arises from a phase II clinical trial of patients with sarcoma, a rare type of cancer affecting connective or supportive tissues and soft tissue (e.g., cartilage and fat). Each patient presents one subtype of the disease and subtypes are grouped by good, intermediate and poor prognosis. The prior model should respect the varying prognosis across disease subtypes. Two subtypes with equal prognosis should be more likely a priori to co-cluster than any two subtypes with different prognosis. The practical motivation for the proposed approach is that the number of accrued patients within each disease subtype is too small to assess the success rates with the desired precision if we were to analyze the data for each subtype separately. It would be practically impossible to carry out a clinical study of possible new therapies. Like a hierarchical model, the proposed clustering approach considers all observations, across all disease subtypes, to estimate individual success rates. But in contrast with the standard hierarchical models, the model considers disease subtypes *a priori* non-exchangeable. This implies that when assessing the success rate for a particular type our model borrows more information from the outcome of the patients sharing same prognosis than from the others.

1.4 Contributions of this Thesis

The main contributions in this thesis to existing methods are the following. The ordinal data model greatly generalizes conventional ordinal probit models. It builds on earlier work by Zhou (2005), who used a construction with nested categorical and ordinal models. The new approach proposed in this thesis is more parsimonious. For the specific application the critical advantage is the use of patients as experimental units (rather than adverse events). This enables us to correctly model dependence across adverse events relate to the same patient.

The semi-parametric model for the biopanning phage data is the first non-parametric approach for such data in the literature. Besides the actual model, another specific methodological innovations is the use of decision theoretic rules to identify organ specific peptide binding.

The non-exchangeable probability partition model for the phase II clinical trial is the first application of the categorical covariate version of these models in these kind of studies in the literature. I explore the characteristics of this model and compare it, via simulation, with models that would be naturally used to model this data.

Chapter 2

Assessing Toxicities in a Clinical Trial: Bayesian Inference for Multivariate Ordinal Data

2.1 Overview

We address modeling and inference for data that include ordinal outcomes nested within categories. The data format can alternatively be seen as multivariate ordinal data with each dimension of the multivariate outcome corresponding to one level of a categorical variable. The motivating application is to model adverse event (toxicity) data in clinical trials. Toxicity type and severity are usually recorded as categorical and ordinal outcome, respectively. In a randomized phase III study, in addition to the efficacy of the study agent, investigators and regulators are also interested in learning about the toxicity profile of the study agent. Traditionally, simple descriptive statistics such as cross-tabulations have been provided. However, this purely descriptive approach fails to offer an in-depth understanding of how the treatment affects both the toxicity type and the severity associated with a specific type of toxicity.

The multinomial probit (MNP) model (Aitchison and Bennett, 1970) and the

multinomial logit model are popular model choices for implementing regression for categorical outcomes. However, the computational burden associated with implementing full posterior inference hinders the routine application of these models in applied work. In recent years, there have been some advances using classical and Bayesian approaches. In particular, the method of simulated moments by McFadden (1989), and Gibbs sampling with data augmentation as discussed in Albert and Chib (1993) and McCulloch and Rossi (1994), have made the required computations in the multinomial probit model more practical.

For inference with ordinal data, many authors have proposed methods in a classical (McCullagh, 1980) and Bayesian (Albert and Chib, 1993; Doss, 1994; Cowles, 1996) framework. A natural way to model ordinal data is to introduce an underlying continuous latent variable. The ordinal outcome is linked with the latent variable through a set of cutpoints. The probability of an ordinal outcome is represented by the probability that this latent continuous variable falls within a given interval defined by the cutpoints. The ordinal probit model is characterized by the assumption that this latent variable follows a normal distribution.

Albert and Chib (1993) proposed Bayesian inference for the ordinal probit regression parameters. The model includes a diffuse prior on the cutpoint parameters. Cowles (1996) proposed improved posterior simulation with a hybrid Gibbs/Metropolis-Hastings sampling scheme which updates the cutpoint parameters jointly with the other parameters. This approach reduces the high auto-correlation and achieves practical convergence within a reasonable number of iterations of the MCMC simulation.

We propose a mixture model which can model ordinal data without the need to estimate cutpoint parameters. We show that in the proposed mixture model, the cutpoints can be fixed without loss of generality. While standard ordinal models assume that the regression lines which characterize the ordinal outcomes are parallel (thus leading to the proportional odds assumption when using a logistic link), our model is flexible in the sense that it is able to fit data when this parallel regressions

assumption is violated. This is especially attractive when modeling multiple ordinal variables. We avoid the need to check that each variable meets the parallel regressions assumption. A similar model with a mixture of normals distribution for the latent probit score is introduced in Kottas *et al.* (2005). They use a non parametric mixture. Our model differs by using a finite mixture, introducing a regression on covariates and using patient-specific random effects. Besides these innovations, the most important contribution of this work is the application to inference for adverse event rates.

This chapter builds on earlier work by Zhou (2005), who used a construction with nested categorical and ordinal models. The rest of the chapter is organized as follows. In Section 2.2 we briefly describe the clinical trials that a drug has to pass successfully in order to reach the market. In this section we also present the model by Albert and Chib (1993) for the Bayesian analysis of binary data and its extension to ordinal data. In Section 2.3, we introduce a phase III clinical trial. In Section 2.4, we present a joint multinomial and ordinal probit model to estimate the cell probabilities of multiple categorical outcomes with different ordinal levels nested in each categorical outcome. The prior specifications and posterior inference are discussed in Section 2.5. We illustrate properties of the model by applying the model to a simulated dataset and data from a phase III clinical trial. The results are presented in Section 2.6. A summary and discussion of possible extensions are presented in Section 2.7.

2.2 Background

In this section, we present the bio-medical and statistical background of the work presented in this chapter. First, we explain what a clinical trials is. Later, we briefly introduce the models proposed by Albert and Chib (1993) for the Bayesian analysis of binary and ordinal data. Both models exploit the idea of data augmentation by introducing a latent variable. The same idea is used in the statistical model we propose in this chapter.

2.2.1 Clinical Trials

Clinical trials are medical studies statistically designed to assess the safety and efficacy of a new treatment –drug, device, psychological therapy, etc– for and with humans. From now on, we will focus on clinical trials testing new drugs and use the word drug and treatment interchangeably. Clinical trials are also performed in order to “extend the label” of an already marketed drug, that is, to prove more benefits to its consumers or to expand the range of diseases for which the drug has a positive effect. A clinical trial is the last step of a long chain in the development of a new treatment; potential drugs have to be discovered, purified, characterized, and tested in labs before proceeding with a clinical trial.

Clinical trials involve human beings as experimental units and, thereby, a number of ethical considerations must be taken into account. Among other controls, at each participating institution an Institutional Review Board must approve and supervise these studies. In order to include a patient in the study, the researcher must get informed consent from the patient, that is, the participants must be aware of the risks involved in the trial. Besides, the participant has the right to withdraw from the trial at any moment. The eligibility of patients must be restricted. For example, pregnant women or patients with better medical options than the tested treatment are usually not allowed to enroll.

Clinical trials for drugs are classified according to their specific objective in four different kinds: phases I, II, III and IV. A new drug must successfully pass the first three phase trials before reaching the market. Phase IV trials are post-marketing studies. Next we briefly describe the different types of clinical trials.

Phase I trials have the objective of finding the acceptable or safe dosage of the new drug. A small number of, generally healthy, patients (20 to 50) are enrolled and followed very closely. The participants do not expect any health benefit. The study involves escalation of the doses before an intolerable level of toxicity is identified.

Phase II trials have the purpose of showing the efficacy of the new drug. The benefits of the drug are compared against a control - either patients under placebo or under best standard of care. The study may include patients treated at control. But many phase II trials include no such control group, and the comparison is only implicit by setting the standards for the judgment. These studies recruit between 20 and 300 patients who, in contrast with phase I clinical trials, can expect health benefits. A statistical comparison between the group of patients receiving the new drug and the control is performed. Often the development of a new drug fails for lack of evidence of beneficial effects in phase II. If the drug is still promising, its safety and efficacy are tested again in a trial involving a large number of patients in a subsequent phase III study.

Phase III trials study definitively assesses the safety and efficacy of the drug. They involve a large number of participants (between 300 and 3,000 or more depending on the disease). As phase II clinical trials, they involve a comparison between a group under the new drug and a control group. Phase III trials usually involve randomization to treatment versus control.

Phase IV trials are also known as “post marketing surveillance trials”. Their objective is to expand the label of the drug. That is, to prove additional benefits to its consumers or to expand the range of diseases for which the drug has a positive effect. Moreover, phase IV trials give information about long term side effects not detected, due to its duration, in the phase III clinical trial. They also may be useful to analyze possible interactions with other drugs.

Phase II and III clinical trials compare the results between the treatment and control groups. In order to establish causal conclusions between the new drug and the results observed in the patients a random assignment of each patient to either a treatment or a control group is important. This randomization has ethical implications related to the decision of stopping a trial. To stop a trial a balance between

individual and *collective* interests (Pocock, 1992; Palmer and Rosenberger, 1999) is necessary. The individual interest suggests to put the patient on what the physician considers the “best treatment” available. If we only followed this goal, then many trials would not proceed. The collective interest requires to treat some individuals not under this “best treatment.” This is done in order to gather data to prove that there is a superior treatment. The disagreement among physicians about the “best treatment” makes randomization ethical (Freedman, 1987). The objective of clinical trials is to create consensus about the best treatment among physicians. Finally, it is worth to mention that adaptive random allocation designs for phase II and III clinical trials are being considered. The idea of these designs is that the probability of assigning a patient to a group changes each time the information is updated and increases the probability of a new patient entering the study to be assigned to what at that point of the study is thought to be the “best treatment”.

2.2.2 Binary Data Model

Suppose that we have a vector of binary outcomes $\mathbf{z} = (z_1, \dots, z_n)$ with $z_i \in \{0, 1\}$. Assume that each observation is associated with a vector of measurements (covariates) x_i of dimension p . We are interested in estimating the probability of observing the response k for a future observation with covariate vector x_f , $k = 0$ or 1 .

A common approach to this problem is the use of generalized linear models. In particular, the probit and the logit models. They assume that $z_i | \pi_i$ are independent and Bernoulli(π_i) distributed. The value of π_i is related to the covariate x_i by means of a known link function ψ^{-1} mapping the interval $(0, 1)$ into the real line. The variable x_i is linearly related with $\psi^{-1}(\pi_i)$, that is $\psi^{-1}(\pi_i) = x_i^T \beta$ where β is a p -dimensional unknown parameter vector. Commonly ψ is chosen to be a cdf. When ψ is the normal cdf we have a probit model and when it is a logistic cdf we are working with a logistic model. An important property of the probit model is that we can choose prior distributions for β that makes posterior inference tractable. See

discussion below. The logistic model allows a nice interpretation of the parameters.

Posterior simulation in the described model is rather complicated. To facilitate posterior simulation, Albert and Chib (1993) propose an equivalent augmented model. They introduce a vector of continuous latent random variables $\mathbf{v} = (v_1, \dots, v_n)$, define the binary response as $z_i = 1(v_i > 0)$ and assume a linear relationship between z_i and the vector x_i , that is:

$$z_i = x_i^t \beta + \epsilon_i \quad \text{for } i = 1, \dots, n.$$

The distribution of the random errors ϵ_i plays an important role in their model. In particular when the errors are standard normal (logistic) distributed we are working with a probit (logistic) model. The variance is set to 1 for identifiability reasons. In the augmented probit model, the joint pdf of the data and the parameters is given by:

$$p(\beta, \mathbf{v}, \mathbf{z}) \propto p(\beta) \times \prod_{i=1}^n \{1(v_i > 0)1(z_i = 1) + 1(v_i < 0)1(z_i = 0)\} \times N(v_i | x_i^t \beta, 1),$$

where $N(y | m, s)$ denotes the normal cdf with mean m and variance s . This implies the complete conditional posterior distribution

$$p(\beta | \mathbf{v}, \mathbf{z}) \propto p(\beta) \times \prod_{i=1}^n N(v_i | x_i^t \beta, 1). \quad (2.1)$$

And $p(v_i | \mathbf{z}, \beta)$ is a truncated (at zero) normal distribution. It is truncated from the right when $z_i = 0$ and truncated from the left when $z_i = 1$. The expression (2.1) is the posterior distribution of β when considering the Bayesian linear regression model $z_i = x_i^t \beta + \epsilon_i$ with $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1)$. Denote by $N_p(m, \Sigma)$ the p -dimensional normal distribution with mean vector m and covariance matrix Σ . Considering a prior for β , $p(\beta) = N_p(m_\beta, \Sigma_\beta)$, yields the posterior distribution $p(\beta | \mathbf{z}) = N_p(m_1, \Sigma_1)$ where $\Sigma_1 = (\Sigma_\beta^{-1} + X^t X)^{-1}$ and $m_1 = \Sigma_1(\Sigma_\beta^{-1} m_\beta + X^t \mathbf{z})$, where X is the design matrix with i -th row equal to x_i . A Gibbs sampler defined by iterative draws from the two conditional posterior distributions above is used to generate a Monte Carlo posterior

sample of β : β^1, \dots, β^M . These values lead to Monte Carlo estimation of π for a future observation:

$$Pr[z_f = 1 \mid x_f, \mathbf{z}] \approx \frac{1}{M} \sum_{m=1}^M \Phi^{-1}(x_f^T \beta^m),$$

where Φ denotes the standard normal cdf.

2.2.3 Multiresponse Categorical Data Model

Albert and Chib (1993) generalized the model given above to consider ordinal responses $z_i \in \{0, \dots, K-1\}$ with $K > 2$. The response z_i is equal to k if the latent variable z_i falls in the random interval $[\theta_k, \theta_{k+1})$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{K-1})$ is an imputed latent random vector of cutpoints, $\theta_0 = -\infty$ and $\theta_K = \infty$. For identifiability reasons, θ_1 is fixed to 0. See Figure 2.1 for a graphical representation in an example. The joint distribution of the parameters and the data is:

$$p(\beta, \mathbf{v}, \mathbf{z}, \boldsymbol{\theta}) \propto p(\beta, \boldsymbol{\theta}) \times \prod_{i=1}^n \{1(\theta_{z_i} \leq v_i < \theta_{z_i+1})\} \times N(v_i \mid x_i^t \beta, 1),$$

The complete conditional posterior distribution for β is exactly the same as in the binary data model given above; $p(v_i \mid \mathbf{z}, \beta)$ is a truncated normal distribution taking values in the interval $[\theta_{z_i}, \theta_{z_i+1})$. Assume a noninformative prior for θ_k , $p(\theta_k) \propto 1$. The complete conditional distribution of θ_k is uniform in the interval $[\max_i \{v_i : z_i = k\}, \min_i \{v_i : z_i = k+1\})$.

Straightforward implementation of the Gibbs sampling scheme using these complete conditional distributions yields a poorly mixing Markov chain. The cutpoints and the latent variables move too slowly. Cowles (1996) algorithm accelerates convergence by replacing alternate sampling from $p(\boldsymbol{\theta} \mid \mathbf{z}, \mathbf{v}, \beta)$ and $p(\mathbf{v} \mid \mathbf{z}, \beta, \boldsymbol{\theta})$, by instead sampling from the joint distribution of the latent variables and the cutpoints conditional in the data and the rest of the parameters,

$$p(\mathbf{v}, \boldsymbol{\theta} \mid \beta, \mathbf{z}) \propto p(\mathbf{v} \mid \mathbf{z}, \beta, \boldsymbol{\theta}) \times p(\boldsymbol{\theta} \mid \mathbf{z}, \beta).$$

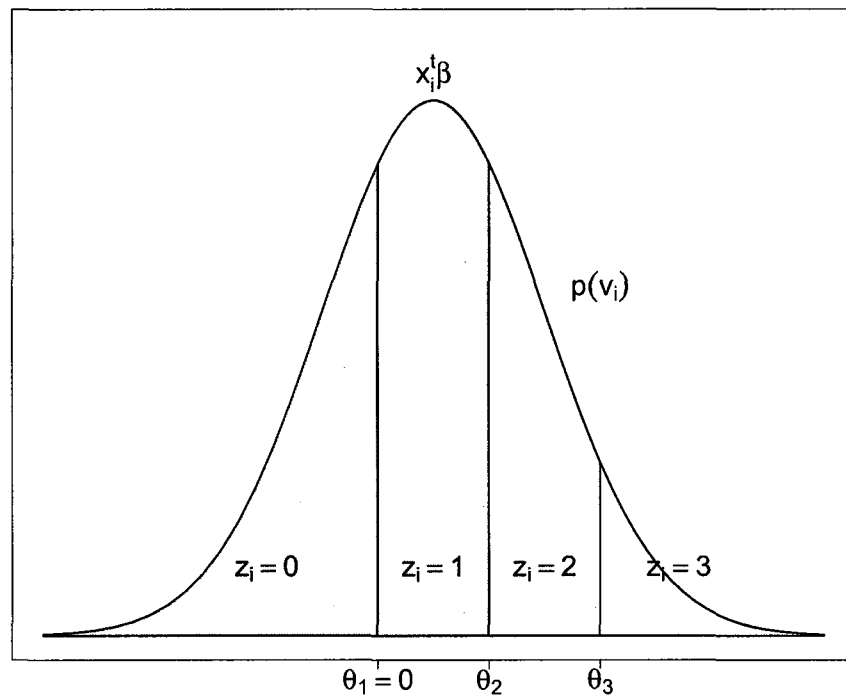


Figure 2.1: Distribution of the latent random variable $v_i \sim N(x_i^t \beta, 1)$ according to Albert and Chib's model considering four categories, $K = 3$. The cutpoints θ_2 and θ_3 are random. The observed value z_i is indicator of the interval where v_i falls.

In other words, the modification of the algorithm consists on sampling from $p(\theta \mid \mathbf{z}, \beta)$ instead of from $p(\theta \mid \mathbf{z}, \mathbf{v}, \beta)$. A Metropolis-Hasting algorithm (see, Cowles, 1996; Johnson and Albert 1999, p. 135) is used to simulate from the former distribution.

Once a posterior sample $(\beta^1, \boldsymbol{\theta}^1), \dots, (\beta^M, \boldsymbol{\theta}^M)$ from $p(\beta, \boldsymbol{\theta} \mid \mathbf{z})$ is obtained the Monte Carlo estimate of cell probability $\pi_k \equiv Pr[z_f = k]$ for a future observation with covariate x_f is

$$\pi_k \approx \frac{1}{M} = \sum_{m=1}^M \{ \Phi^{-1}(x_f^T \beta^m - \theta_{k+1}^m) - \Phi^{-1}(x_f^T \beta^m - \theta_k^m) \} \quad \text{for } k = 0, \dots, K - 1.$$

Finally, we note that Albert and Chib's model for more than two categories requires the "parallel regression assumption," the probit version of the proportional odds assumption of the logistic model. That is,

$$\begin{aligned} Pr[z \leq k \mid x] &= Pr[x^t \beta + \epsilon \leq \theta_{k+1}] \\ &= Pr[\epsilon \leq \theta_{k+1} - x^t \beta] = \Phi(\theta_{k+1} - x^t \beta) \end{aligned}$$

and, therefore,

$$f_k(x) := \Phi^{-1}(Pr[z \leq k \mid x]) = \theta_{k+1} - x^t \beta.$$

In other words, the assumption is that the functions f_k for $k = 0, \dots, K - 1$ are parallel lines.

2.3 A Phase III Clinical Trial

Studies have suggested that retinoid chemoprevention may help control second primary tumors, disease recurrence, and mortality for stage I non-small-cell lung cancer (NSCLC) patients. A National Cancer Institute (NCI) intergroup phase III trial of 1166 patients with pathologic stage I NSCLC was conducted to validate the efficacy of isotretinoin, a retinoid hypothesized to have chemopreventive properties. Patients were randomly assigned to receive either placebo or isotretinoin (30 mg/day) for 3 years in a double-blinded study. Patients were stratified at randomization by tumor

stage, histology, and smoking status. A total of 589 patients received isotretinoin while the remaining patients received placebo.

One of the objectives of the phase III study is to assess the treatment effect on different types of toxicity and the grade associated with each of them. In this chapter we focus on inference related to this objective only. The treatment-related toxicities include: cheilitis, conjunctivitis, arthralgia, hypertriglyceridemia, headache, and abnormal vision. Cheilitis is dryness, usually associated with an uncomfortable sensation of the lips with scaling and cracking and accompanied by a characteristic burning sensation. Conjunctivitis is one of the most common nontraumatic eye complaints, involving the inflammation of conjunctiva. Arthralgia is pain in the joints. Hypertriglyceridemia is an excess of triglycerides in the blood. With the exception of hypertriglyceridemia, toxicity was graded by use of the Common Toxicity Criteria, a toxicity scale used by the NCI for Adverse Events. Triglyceride toxicity was graded as follows: grade 1 toxicity was defined as more than 2.5 times but less than or equal to five times the normal level; grade 2 toxicity was defined as more than five times but less than or equal to 10 times the normal level; and grade 3 toxicity was defined as more than 10 times the normal triglyceride level or if a patient experienced complications (e.g., pancreatitis) at any grade of triglyceride toxicity. If patients experienced multiple incidents of the same toxicity, only one incident at the highest grade was counted. When multiple different toxicities are reported for the same patient, the corresponding observed toxicity levels are expected to correlate. We will introduce the desired correlation with patient-specific random effects in the model. Reported toxicity rates are per patient, i.e., toxicity rates are probabilities that a patient reports a certain type of toxicity and grade. A summary of the data is shown in Table 2.1.

2.4 A Hierarchical Model for Ordinal Data Nested within Categories

For each recorded adverse event, the data report one variable. This variable is an ordinal outcome, z_{ij} , which reports the grade at which the j^{th} categorical outcome was observed on the i^{th} individual, $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, J\}$, where, respectively, n and J denote the total number of patients and the total number of different toxicity types recorded in the study. The variable z_{ij} takes values $k = 0, 1, \dots, K_j$. The observation $z_{ij} = k$ indicates that the i^{th} patient exhibited the toxicity of type j at grade k . The additional grade $k = 0$ is used to indicate that toxicity j was not recorded for patient i .

Let X be a $(n \times H)$ matrix of possible regressors, with the i^{th} row, x_i , recording H covariates for the i^{th} patient, $i = 1, \dots, n$. For inference on a dose effect one could use $x_{i1} = 0$ when the drug is not present, $x_{i1} = 1$ for the lowest dose of the drug, $x_{i1} = 2$ for the second lowest dose and so on. In our specific example, we have a dichotomous covariate. We use $x_{i1} = 1$ when patient i is treated with isotretinoin, and $x_{i1} = -1$ for placebo. Considering just the treatment effect we have $H = 1$ and the i^{th} row of the covariate matrix, X , is just x_{i1} . In general, the covariates could be occasion-specific and indexed by patient i and toxicity j . We only use patient-specific covariates in the application and proceed therefore for simplicity with patient-specific covariates only.

We set up an ordinal probit regression for z_{ij} on covariates x_i . The cell probability $Pr(z_{ij} = k)$ is represented as the probability that a continuous latent variable v_{ij} falls into the interval $(\theta_{kj}, \theta_{k+1,j})$. A patient specific random effect r_i induces correlation across all toxicity observations for the same patient. Multiple cutpoints are required for the K_j ordinal outcomes:

$$z_{ij} = k \quad \text{if } \theta_{kj} \leq v_{ij} < \theta_{k+1,j} \quad \text{for } k = 0, 1, \dots, K_j. \quad (2.2)$$

$$v_{ij} = x_i^T \beta_j + r_i + \xi_{ij}, \quad (2.3)$$

where,

$$r_i \sim N(0, \sigma_r^2), \quad \xi_{ij} \sim \sum_{g=1}^G p_{jg} N(\mu_{jg}, \sigma_\xi^2) \quad (2.4)$$

and

$$(p_{j1}, p_{j2}, \dots, p_{jG}) \sim \text{Dirichlet}(\alpha_{j1}, \alpha_{j2}, \dots, \alpha_{jG}), \quad j = 1, \dots, J. \quad (2.5)$$

Here, β_j parameterize the ordinal probit model for z_{ij} . Notice that β_j does not include an intercept parameter. An intercept is already implicitly included in μ_{jg} . Consider the implied model for each v_{ij} after marginalizing with respect to r_i . The marginal distribution for each v_{ij} , $j = 1, \dots, J$, is a mixture of normal distributions sharing the scale parameter ($\sqrt{\sigma_r^2 + \sigma_\xi^2}$), with distinct location parameters ($x_i^T \beta_j + \mu_{jg}$), fixed number of components (G) and weights (p_{jg}). It can be shown that without loss of generality we can fix the cutpoints θ_{jk} when working with the mixture of normal model in (2.4) instead of a single normal. See also the discussion below. While we assume that G is fixed, for $j = 1, \dots, J$, in (2.4), using different G_j for each toxicity is possible without additional complications.

The mixture model can alternatively be written as a hierarchical model by introducing a latent indicator variable w_{ij} . Specifically, conditional on $w_{ij} = g$, β_j and r_i , the latent variable v_{ij} follows a normal distribution: $v_{ij} \mid w_{ij} = g, \beta_j, r_i, \mu_{jg} \sim N(x_i^T \beta_j + r_i + \mu_{jg}, \sigma_\xi^2)$. The prior probability for $w_{ij} = g$ is $Pr(w_{ij} = g) = p_{jg}$. Let $\Phi(\cdot)$ denote the standard normal cdf. Marginalizing with respect to both r_i and the latent variable v_{ij} , we have:

$$Pr(z_{ij} = k \mid w_{ij} = g, \beta_j, \mu_{jg}) = \Phi\left(\frac{\theta_{k+1,j} - x_i^T \beta_j - \mu_{jg}}{\sqrt{\sigma_r^2 + \sigma_\xi^2}}\right) - \Phi\left(\frac{\theta_{kj} - x_i^T \beta_j - \mu_{jg}}{\sqrt{\sigma_r^2 + \sigma_\xi^2}}\right) \quad (2.6)$$

$$Pr(w_{ij} = g) = p_{jg}. \quad (2.7)$$

For each category j the probability of a response at level k is

$$\pi_{jk} \equiv Pr(z_{ij} = k) = \sum_{g=1}^G Pr(z_{ij} = k \mid w_{ij} = g) p_{jg} \quad (2.8)$$

For reasons of identifiability we fix the variance of the normal kernels in (2.4). We recommend $\sigma_\xi^2 = 1$ and $\sigma_r^2 = 4$. See the argument below.

For later reference we state the joint probability model. Let $\mathbf{z} = (z_{ij} : i = 1, \dots, n, j = 1, \dots, J)$ denote the data. Let f_{β_j} denote the prior distribution for β_j . Assume that μ_{jg} for $j = 1, \dots, J$ and $g = 1, \dots, G$ are *a priori* independent, that for each j , given the imputed hyperparameter ϕ_j , $\mu_{j1}, \dots, \mu_{jG}$ is a random sample from $f_\mu(\cdot | \phi_j)$ and that $\phi_1, \dots, \phi_J \stackrel{\text{iid}}{\sim} f_\phi$. Let $\beta = (\beta_1, \dots, \beta_J)$ denote all probit regression coefficients, and similarly for $\boldsymbol{\mu}$, \mathbf{r} , \mathbf{v} , \mathbf{w} and $\boldsymbol{\phi}$. Let $N(x|m, s)$ indicate a normal pdf with mean m and variance s evaluated at x . The joint distribution of \mathbf{z} and all the model parameters is

$$\begin{aligned}
p(\mathbf{z}, \mathbf{v}, \mathbf{r}, \boldsymbol{\beta}, \boldsymbol{\mu}, \mathbf{w}, \boldsymbol{\phi}) = & \\
& \prod_{i=1}^n \prod_{j=1}^J 1[\theta_{z_{ij},j} < v_{ij} < \theta_{z_{ij}+1,j}] \prod_{i=1}^n \prod_{j=1}^J N(v_{ij} | x_i^T \beta_j + r_i + \mu_{j,w_{ij}}, \sigma_\xi^2) \\
\times & \prod_{j=1}^J f_{\beta_j}(\beta_j) \prod_{i=1}^n N(r_i | 0, \sigma_r^2) \\
\times & \prod_{j=1}^J \left\{ \prod_{g=1}^G p_{jg}^{\#\{i:w_{ij}=g\}} \prod_{g=1}^G p_{jg}^{\alpha_{jg}-1} \right\} \prod_{j=1}^J \left\{ f_\phi(\phi_j) \prod_{g=1}^G f_\mu(\mu_{jg} | \phi_j) \right\}
\end{aligned} \tag{2.9}$$

To show that the cutpoints θ_{jk} in (2.2) can be fixed, consider the following simplified version of the right side in (2.4). Let $v \sim \sum_{g=1}^G p_g N(\mu_g, \sigma^2)$ and define $\pi_k \equiv Pr(z = k) = Pr(\theta_k \leq v < \theta_{k+1})$, $k = 0, 1, \dots, K$. We show by a constructive argument that an appropriate choice of $(G, p_g, \mu_g, g = 1, \dots, G)$ can approximate an arbitrary set of desired cell probabilities $(\pi_0^*, \pi_1^*, \dots, \pi_K^*)$. In particular, the parallel regression assumptions of the probit model is not required. A similar argument was used in Kottas *et al.* (2005) for infinite Dirichlet process mixtures of normal distributions. Consider a mixture of normal distributions with $G \geq K$ components. Place one component of the mixture into each interval $[\theta_k, \theta_{k+1})$ by choosing $\mu_k = \frac{1}{2}(\theta_k + \theta_{k+1})$, and set $p_k = \pi_k^*$. Specify σ such that $1 - \epsilon$ of the probabilities of each kernel is between the adjacent cutpoints. This trivially achieves $|\pi_k - \pi_k^*| < \epsilon$ for $k = 0, 1, \dots, K$. Therefore, the cutpoints θ_{jk} in (2.2) can be fixed without loss of generality.

We recommend as a default choice of cutpoint parameters θ_k : $\theta_0 = -\infty$, $\theta_{K+1} =$

∞ , $\{\theta_1, \dots, \theta_K\} = \{0, \pm 4, \pm 8, \dots\}$. For reasons of identifiability, we suggest fixing $\sigma = 1$. This choice implicitly restricts cell probabilities π_2, \dots, π_{K-1} to be at most 0.95. The first and last cell probabilities, π_1 and π_K , are unrestricted. This is important in the context of the later application to adverse event rates, when the first level of the ordinal outcome corresponds to no toxicity, which is often greater than 0.95. If larger cell probabilities are desired for intermediate outcomes, the widths between θ_k and θ_{k+1} can be increased or decreased accordingly. Figure 2.2 illustrates the model for one toxicity with three possible outcomes, i.e., $z_{ij} \in \{0, 1, 2\}$ and $K_j = 2$. The figure shows how the proposed model with $G = 2$ fits the cell probabilities π_{1k} . The figure shows the mixture of normal distribution of the latent random variable, v_{i1} , under two values of the covariate $x_i = -1, 1$. In both mixtures, the darkly shaded, lightly shaded and white areas correspond to the probability of the ordinal outcome taking the values 0, 1 and 2, respectively. Notice that this particular set of cell probabilities does not satisfy the parallel regression assumption. Therefore, these probabilities cannot be represented by a model with a unimodal distributed latent random variable and random cutpoints. Finally, we fix σ_r^2 at $\sigma_r^2 = 4$, implying non-negligible prior probability for random effects r_i to be in a range that covers several cutpoints θ_k .

2.5 Priors, Posterior and Simulation Scheme

We use conjugate priors for the probit regression parameters, centering the prior to represent the prior judgment about the marginal prevalence of the outcomes and the effects of the covariates. We use

$$f_{\beta_j}(\beta_j) \equiv N(m_{\beta_j}, \sigma_{\beta}). \quad (2.10)$$

As default choice for G , the size of the mixture model in (2.4), we suggest $G = K - 1$. Our recommendation is based on empirical evidence. On one hand, small values of G create faster mixing Markov chains, but may not be sufficient to fit the data.

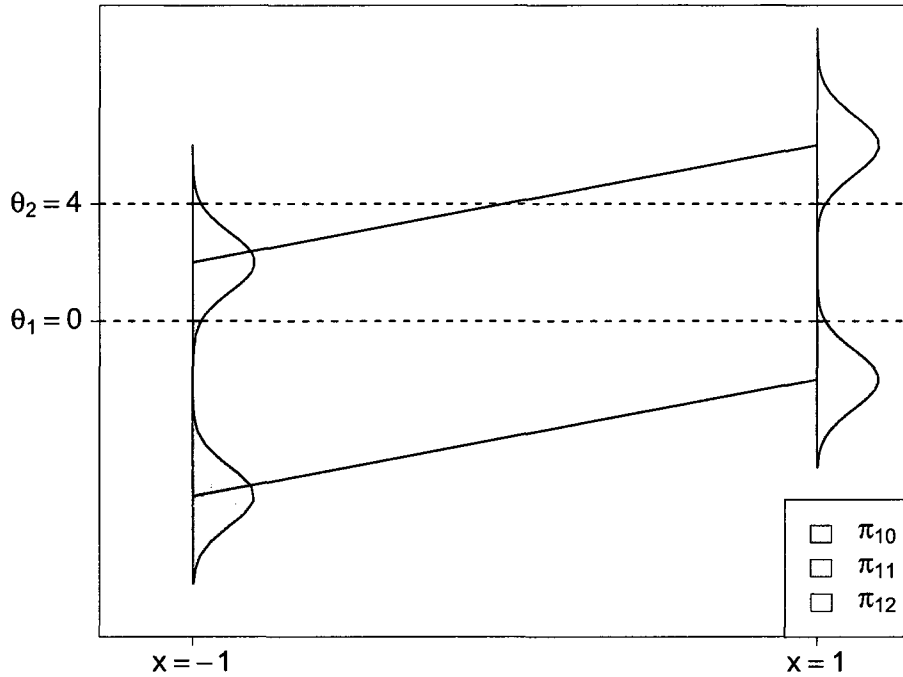


Figure 2.2: Illustration of the distribution of the latent variable v_{i1} in the model described in (2.2)-(2.4) when the covariate takes values -1 and 1 . Here we consider: $J = 1$ type of toxicity, no patient-specific random effect r_i , $G = 2$ components in the mixture of normals and three, $K = 2$, possible ordinal outcomes. In both mixtures, the darkly shaded, lightly shaded and white areas correspond to the probabilities π_{1k} of the ordinal outcome taking the values $0, 1$ and 2 , respectively.

On the other, large values of G may overparameterize the model leading to poorly mixing Markov chains. Alternatively, using reversible jump MCMC (Green, 1995), G could be included in the parameter vector and estimated as part of the inference. But since the parameters of interest are the cell probabilities π_{jk} , and inference on mixture-specific parameters is not of interest we prefer the approach with fixed large values of G . Formally implied inference on the parameters of the mixture model, including μ_{jg} and p_{jg} , should not be interpreted. Problems related to label switching (arbitrary permutation of the terms in the mixture) and node duplication (replicating

essentially identical terms) make the posterior distribution on μ_{jg} and p_{jg} meaningless to interpret.

We assume that w_{ij} takes on discrete values $1, 2, \dots, G$ with prior probability p_{j1}, \dots, p_{jG} , respectively. For the location parameter (μ_{jg}) in the components of the mixture of normal model (2.4), we use independent normal priors $f_{\mu}(\mu_{jg}|\phi_j) \equiv N(\phi_j, \sigma_{\mu}^2)$ with a conjugate hyperprior $f_{\phi_j}(\phi) \equiv N(0, \sigma_{\phi}^2)$.

Keeping in mind the default suggestion for the cutpoints θ_k we recommend $\sigma_{\mu} \approx 1/J \sum_j K_j$, i.e., half the span from the first to last cutpoint, averaging across all toxicities.

An investigator might be interested in assessing how different dose levels affect toxicity grade. Our model may be used in this context. For a cytotoxic agent, it is usually assumed that a higher dose incurs worse toxicity. The parameter β_j , the dose effect on the toxicity grade, may be restricted to be positive when there are only two dose levels and the lower dose group is the reference group. When there are M dose groups, β_j becomes an $(M-1)$ -dimensional vector. In this case one could enforce monotonicity of toxicity with increasing dose by introducing an order constraint on β_j as follows: assuming that the lowest dose is the reference group and the highest dose group is group M , monotonicity can be represented as $\beta_{(M-1)j} > \beta_{(M-2)j} > \dots > \beta_{1j}$. This assumption guarantees *a priori* that a higher dose incurs worse toxicity.

All full conditional posterior distributions are derived from the joint probability model (2.9). Since we adopt conjugate priors for all parameters, all full conditional posterior distributions have tractable closed forms, allowing straightforward implementation of Markov Chain Monte Carlo (MCMC) posterior simulation using a Gibbs sampler. We start with initial values for the latent variables \mathbf{v} and \mathbf{w} . The values for \mathbf{v} need to comply with (2.2). One iteration of the Gibbs sampler is described by the following transition probabilities:

- (a) For each $j = 1, \dots, J$, we first marginalize with respect to ϕ_j (recall that ϕ_j is the prior mean for μ_{jg}). Then, sample $(\beta_j, \mu_{j1}, \dots, \mu_{jG})$ from the joint complete

conditional posterior distribution. That is, from the posterior distribution for the linear regression:

$$v_{ij} - r_i = x_i^T \beta_j + \sum_{g=1}^G \mathbf{1}(w_{ij} = g) \mu_{jg} + \epsilon_i, \quad \text{for } i = 1, \dots, n,$$

with response $y_i^* = v_{ij} - r_i$. The residuals are $\epsilon_1, \dots, \epsilon_n \stackrel{\text{iid}}{\sim} N(0, \sigma_\xi^2)$ and the prior is the multivariate normal distribution resulting from the product of (2.10) and the prior distribution of $(\mu_{j1}, \dots, \mu_{jG})$ when marginalizing with respect to ϕ . The latter is a multivariate normal distribution of dimension G with mean zero and covariance matrix $\sigma_\mu^2 I_G + \mathbf{1}_G \mathbf{1}_G^t \sigma_\phi^2$, where I_G is the identity matrix of dimension G and $\mathbf{1}_G$ is the column vector of length G with all entries equal to 1.

- (b) For each $i = 1, \dots, n$, we sample r_i from the complete conditional posterior distribution, or, equivalently, from the posterior distribution for a linear regression with response $y_j^{**} = v_{ij} - (x_i^T \beta_j + \mu_{t,w_{ij}})$,

$$y_j^{**} = r_i + \epsilon_j, \quad \text{for } j = 1, \dots, J,$$

with $\epsilon_1, \dots, \epsilon_J \stackrel{\text{iid}}{\sim} N(0, \sigma_\xi^2)$ and prior $r_i \sim N(0, \sigma_r^2)$.

- (c) We update β_j with a random walk Metropolis-Hasting transition probability. We generate $\tilde{\beta} \sim N(\beta_j, c^2)$ where $c > 0$ and compute the posterior distribution of β_j (marginalizing with respect to \mathbf{w}):

$$p(\beta_j \mid \boldsymbol{\mu}, \mathbf{r}, \mathbf{p}, \mathbf{z}) \propto f_{\beta_j}(\beta_j) \prod_{i=1}^n \left[\Phi_{x_i^T \beta_j, \sigma_\xi^2} \left(\theta_{z_{ij}+1,j} - r_i - \sum_{g=1}^G p_{jg} \mu_{jg} \right) - \Phi_{x_i^T \beta_j, \sigma_\xi^2} \left(\theta_{z_{ij},j} - r_i - \sum_{g=1}^G p_{jg} \mu_{jg} \right) \right]$$

where $\Phi_{\mu\sigma^2}(x)$ represents the normal cdf with mean μ and variance σ^2 evaluated at x . Let $A = p(\tilde{\beta} \mid \boldsymbol{\mu}, \mathbf{r}, \mathbf{p}, \mathbf{z}) / p(\beta_j \mid \boldsymbol{\mu}, \mathbf{r}, \mathbf{p}, \mathbf{z})$. With probability $\min\{1, A\}$ we replace β_j by $\tilde{\beta}$.

- (d) The latent indicator variables w_{ij} in equation (2.7) are sampled from the complete conditional posterior distribution (marginalized with respect to z_{ij}):

$$Pr(w_{ij} = g \mid \beta_j, r_i, \mu_{jg}, p_{jg}) \propto p_{jg} [\Phi_{\mu_{jg}, \sigma_\xi^2}(\theta_{z_{ij}+1, j} - x_i^T \beta_j - r_i) - \Phi_{\mu_{jg}, \sigma_\xi^2}(\theta_{z_{ij}, j} - x_i^T \beta_j - r_i)].$$

- (e) For each j , we update (p_{j1}, \dots, p_{jG}) from

$$f(p_{j1}, \dots, p_{jG} \mid w_{1j}, \dots, w_{nj}) = \text{Dirichlet}(\alpha'_{j1}, \dots, \alpha'_{jG}),$$

$$\text{with } \alpha'_{jg} = \alpha_{jg} + \#\{i : w_{ij} = g\}.$$

- (f) The latent variables v_{ij} are updated by draws from the truncated normal distribution

$$f(v_{ij} \mid \beta_j, r_i, w_{ij} = g, \mu_{jg}, z_{ij} = k) \propto N(v_{ij} \mid x_i^T \beta_j + r_i + \mu_{jg}, \sigma_\xi^2) I(\theta_k \leq z_{ij} < \theta_{k+1, j}).$$

Finally, for each toxicity type and grade, we evaluate the posterior probability of toxicity for a future patient. For each type of toxicity, using the equations (2.6) and (2.8), the posterior probability $\pi_{jk} \equiv Pr(z_{fj} = k \mid \mathbf{z})$, $f = n + 1$, that a future patient with covariate vector x_f exhibits the toxicity j at level k is estimated from the Gibbs sampler outputs. Let η^m denote the imputed value of the generic parameter η after m iterations of the Gibbs sampler. We report

$$\begin{aligned} \pi_{jk} &= \sum_{g=1}^G Pr(z_{fj} = k \mid w_{fj} = g, \mathbf{z}) Pr(w_{fj} = g \mid \mathbf{z}) \\ &\approx \frac{1}{M} \sum_{m=1}^M \sum_{g=1}^G \left\{ \Phi \left(\frac{\theta_{k+1, j} - x_f^T \beta_j^m - \mu_{jg}^m}{\sqrt{\sigma_\tau^2 + \sigma_\xi^2}} \right) - \Phi \left(\frac{\theta_{kj} - x_f^T \beta_j^m - \mu_{jg}^m}{\sqrt{\sigma_\tau^2 + \sigma_\xi^2}} \right) \right\} p_{jg}^m, \end{aligned}$$

where M is the total number of MCMC iterations retained after an initial burn-in.

2.6 Applications

2.6.1 A Simulated Dataset

We use a simulated dataset to validate the model. A total of $n = 1000$ subjects were assigned into two groups, A and B , of equal size. For each subject i , there were four

($J = 4$) ordinal outcomes labeled $z_i = (z_{i1}, \dots, z_{i4})$ with four ($K = 3$) possible values, i.e., $z_{ij} \in \{0, 1, 2, 3\}$. The observations z_i were generated according to the model defined by (2.2)-(2.4) but with r_i generated from a mixture of normal distributions, $r_i \sim 0.75 N(0, 1) + 0.25 N(4, 1)$. In (2.2) we fixed the values of the cutpoints, θ_{jk} , as $(\theta_{0k}, \theta_{1k}, \dots, \theta_{4k}) = (-\infty, -3, 1, 2, \infty)$ for all j . We deliberately chose cutpoints different from the default cutpoints that are used in the analysis model. In (2.3) we set $x_i = -1$ for group A and $x_i = 1$ for group B . The slope parameters in the probit regression were set to $(\beta_1, \dots, \beta_4) = (0, 0.5, 1, 1.5)^T$. In (2.4) we set the variance $\sigma_\xi^2 = 1$, the number of components in the normal mixture $G = 3$, the normal means $(\mu_{j1}, \mu_{j2}, \mu_{j3}) = (-5, 0, 3)$ and the mixture weights $(p_{j1}, p_{j2}, p_{j3}) = (0.9, 0.05, 0.05)$ for all j . The true cell probabilities are reported in Table 2.2.

We fit model (2.2)-(2.5) with priors for the parameters as described in Section 4. More specifically, we assume that the probit parameters in (2.3), β_j , follow the normal distribution specified in (2.10) with $m_{\beta_j} = 0$ and $\sigma_{\beta_j}^2 = 1$. In (2.4) we assume $G = 2$, and weights distributed according to (2.5) with $\alpha_{j1} = \alpha_{j2} = 1$ for all j . The normal means are assigned conjugate priors, $\mu_{jg} \sim N(\phi, \sigma_\mu^2 = 16)$ and $\phi \sim N(0, \sigma_\phi^2 = 10^4)$. We use the default choices $\sigma_\xi^2 = 1$, $\sigma_r^2 = 4$ and $\sigma_\mu^2 = 16$. The values of the cutpoints in (2.2), θ_{jk} , were set to $(\theta_{j0}, \dots, \theta_{j4}) = (-\infty, -4, 0, 4, \infty)$ for all j .

We simulated a total of 110,000 iterations of the posterior MCMC scheme. After an initial burn-in of 10,000 iteration, the imputed parameters were saved after each 10^{th} iteration, yielding to a posterior Monte Carlo sample size of 10,000. The marginal posterior probabilities for each combination of toxicity type and grade were estimated and compared with the true cell probabilities in Table 2.2. The model reports reasonable estimates of the cell probabilities; in 26 out of 32 cells the true cell probability is within the reported 95% central credible interval.

For comparison we also implemented an ordinal probit regression with random cutpoints, but a (single) normal distribution for the latent probit scores (Albert and Chib, 1993). For a fair comparison we included patient-specific random effects as

in the proposed model. Thus the alternative model is (2.2)-(2.5) with $G = 1$ and random cutpoints θ_2 through θ_{K_j} ($\theta_1 = 0$ is fixed). We used the posterior MCMC implementation of Cowles (1996). We summarized the marginal posterior distributions for the cell probabilities as in Table 2.2 (not shown). We find that for only 10 out of the 32 cell probabilities the central 95% posterior credible intervals contain the simulation truth. To assess the efficiency of the posterior MCMC for the proposed model versus the ordinal probit regression we recorded serial autocorrelations of the Markov chain simulations. We find comparable values (not shown). In summary, we conclude that the proposed model and the conventional ordinal probit regression lead to comparable results with slightly more flexibility of the proposed model. We caution against over-interpreting the comparison. The simulated data set is relatively large with $n = 1000$, and the reported comparison is based on only one simulated data set.

Finally we investigated robustness of posterior inference with respect to choices of the prior hyperparameters. To explore prior sensitivity we considered several alternative choices. Shifting the values of all α_{jg} to either 1/3 or 3 did not change inference appreciably. Similarly, using a non informative prior, $p(\beta) \propto 1$, for the regression parameter did not substantially affect the inference. Overall, we found that the posterior estimates are quite robust with respect to prior specification.

2.6.2 A Phase III Clinical Trial of Retinoid Isotretinoin

We applied the proposed model for inference in the phase III clinical trial introduced in Section 2. As in the simulation study, we chose $N(0, 1)$ priors for the ordinal probit parameters. The size of the mixture was fixed at $G = 2$ with equal *a priori* weights by setting $\alpha_{jg} = 1$. A vague hyperprior centered at 0 with the variance of $\sigma_\phi^2 = 10^4$ was imposed on ϕ . The cutpoints were chosen following the default choices. The variances were set to the default choices $\sigma_\xi^2 = 1$, $\sigma_r^2 = 4$ and $\sigma_\mu^2 = 16$.

Saving every 10^{th} iteration after a 10,000 iteration burn-in, a Monte Carlo poste-

rior sample of size 10,000 was saved to estimate cell probabilities. Table 2.3 displays the estimated cell probabilities together with central 95% credible intervals. Note the near zero probabilities for some of the higher toxicity grades. The proposed model is appropriate to handle sparse tables. The estimates formally confirm and quantify what is expected from inspection of the data. There were more incidences of cheilitis and conjunctivitis observed in the isotretinoin group than in the placebo group. Elevated triglyceride levels were found more frequently in the isotretinoin group than in the placebo group. In contrast, more patients experienced headache in the placebo group. Posterior inference confirms that adverse event rates under treatment and placebo differ significantly. Figure 2.3 summarizes posterior inference for the probit regression parameters. It is 99% certain that the treatment (isotretinoin) had an undesired (i.e., $\beta_j > 0$) effect on cheilitis, conjunctivitis and hyper-triglyceride.

The posterior distribution allows us to report coherent probabilities for any event of interest. In particular, we can report inference on joint and conditional probabilities of adverse events across different toxicities. For example, Table 4 reports conditional probabilities for each adverse event (at any grade) given an adverse event in another toxicity for the same patient. For comparison the table also reports the marginal probabilities (in the diagonal). The considerable variation of probabilities in each row confirms that the toxicities exhibited by the same patient are not independent. The inclusion of subject specific random effects r_i was critical in fitting this data. For example, the first row reports that the probability that a patient exhibits abnormal vision is low marginally, but considerably increased when the patient has experienced fatigue or headache.

Figure 2.4 shows the estimated distribution for the underlying latent variable, v_{ij} , in equation (2.3) conditioned on $r_i = 0$. Let $\bar{\beta}_j = E(\beta_j | \mathbf{z})$, $\bar{p}_{jg} = E(p_{jg} | \mathbf{z})$ and $\bar{\mu}_{jg} = E(\mu_{jg} | \mathbf{z})$. The figure shows $x_i^T \bar{\beta}_j + \sum_g \bar{p}_{jg} N(\bar{\mu}_{jg}, 1)$, for $x = -1, 1$ and $j = 1 \dots, 7$.

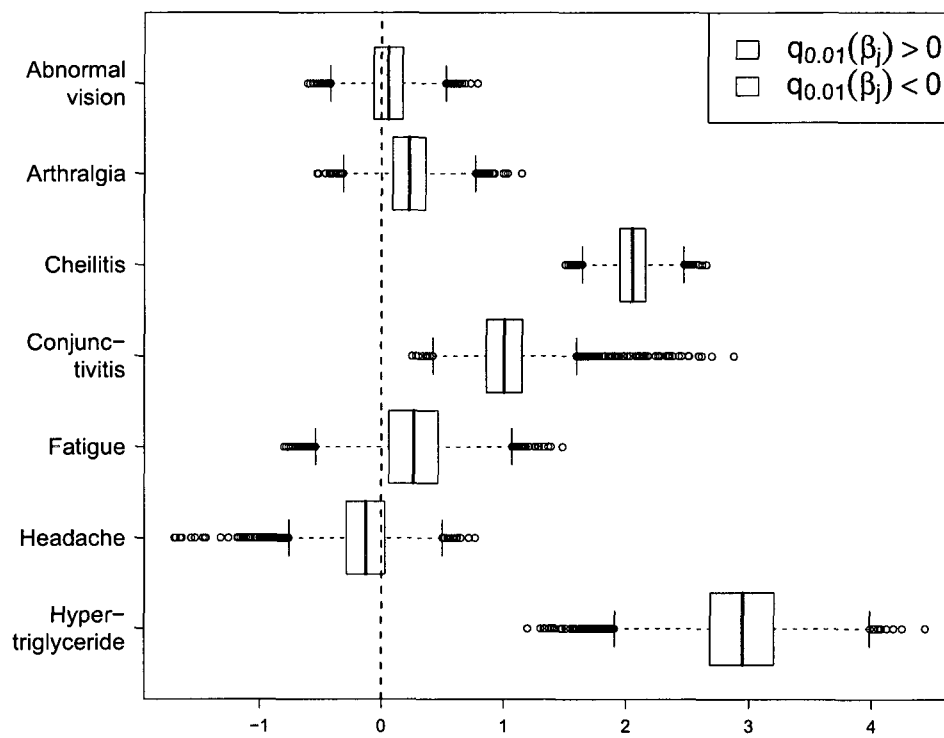


Figure 2.3: Boxplots of the simulated posterior samples of the ordinal probit model parameters (β_j 's). Boxes corresponding to samples of β_j 's with 0.01-quantile greater than zero are shaded.

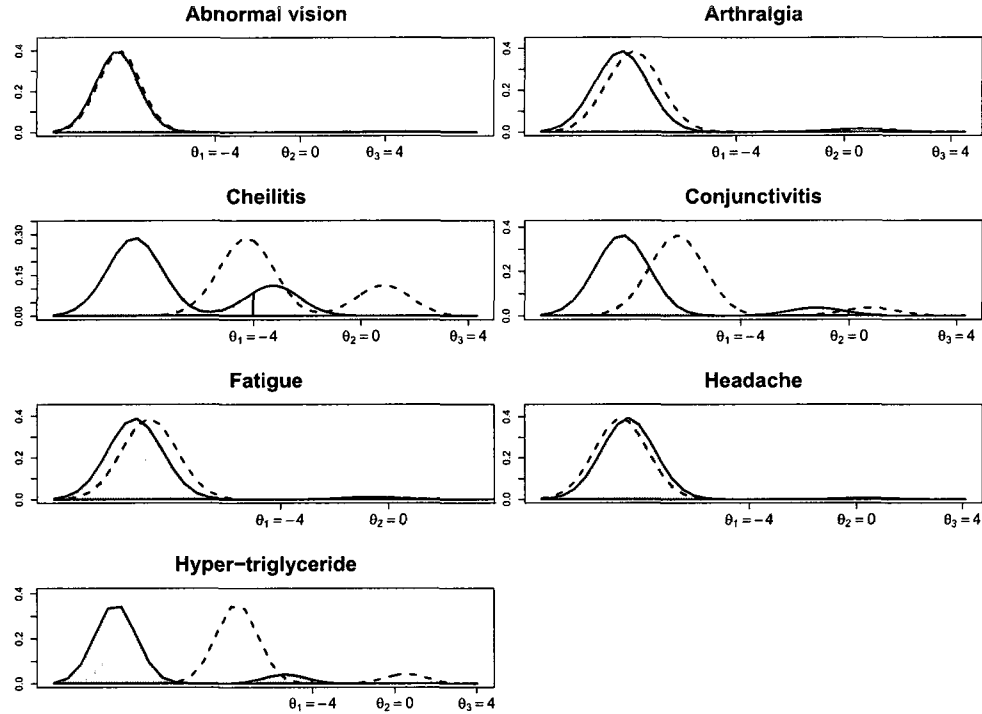


Figure 2.4: Estimated mixture of normal distribution of the latent variable v_{ij} conditioned on $r_i = 0$ for $j = 1, \dots, 7$. Shaded curve corresponds to placebo ($x = -1$) and dashed to isotretinoin ($x = 1$). Darkly and lightly shaded areas represent, respectively, the probability of no toxicity (π_{j0}) and toxicity at grade 1 (π_{j1}) under placebo.

2.7 Discussion

We have proposed a Bayesian hierarchical model to analyze ordinal data nested within categories. Our model characterizes the ordinal/categorical data structure by a variation of the ordinal probit model. We provide posterior summaries to assess treatment effects. In the phase III clinical trial example, traditional analysis might simply group the toxicities levels into two: no toxicity (0) and some toxicity (1+2+3+4), and then, apply a Chi-squared test or Fisher's exact test to compare the two treatment groups

for a particular toxicity type instead. The proposed approach provides alternative model-based posterior inference. In accordance with the natural structure of the data, our model treats toxicity grade as an ordinal data. The proposed model accounts for the (high) dependence across different toxicities within the same patient. The proposed model allows for extensions to more complicated designs by appropriate changes in the linear model (2.3). For example, one could accommodate repeated observations of adverse events by replacing the latent variable v_{ij} in (2.3) by v_{ijh} for the h^{th} repeated observation of type of toxicity j for patient i , and defining a new set of random effects R_{ij} .

This model also has interesting applications in other areas such as health outcomes research and clinical trial design. For example, some studies have shown that even when treatments are known to be effective, many patients who could benefit from them are not getting these treatments. Beta blocker medication, given after heart attacks, can reduce mortality; blood-thinning medication can prevent stroke; and thrombolytic therapy given immediately after a heart attack can reduce the damage from the attack. The outcome instrument has focused on assessing the overall level of functioning after receiving the treatment conditional on patients' prognostic characteristics. The overall level of functioning is a quantified variable on an ordinal scale. Therefore, by assessing the ordinal outcomes within each category, health outcome researchers will be able to identify and address the barriers to better care and, eventually, translate these findings into practical strategies to improve care.

One critical issue is the choice of the size of the mixture in modeling ordinal outcomes. We suggested as a rule of thumb to set the size of mixture, G , equal to the number ordinal levels minus two ($K - 1$). Alternatively, one could treat G as an unknown parameter and use reversible jump MCMC.

In summary, we have introduced an approach for flexible, model-based inference for the adverse events reported in a Phase III clinical trial. The model includes dependence across adverse events for the same patient. The computational effort of

implementation is comparable to a traditional ordinal probit regression.

Table 2.1: Toxicity frequency for randomized eligible patients by study arms. In the placebo (Isotretinoin) group, 171 (427) out of 577 (589) patients exhibited some type of toxicity. In bold the proportion of patients in the study arm belonging to the cell.

Placebo						
Toxic effect	No Tox	G1	G2	G3	G4	
Abnormal vision	565(0.979)	9(0.016)	0(0)	2(0.003)	1(0.002)	
Arthralgia	548(0.95)	19(0.033)	10(0.017)	0(0)	-	
Cheilitis	493(0.854)	76(0.132)	8(0.014)	0(0)	-	
Conjunctivitis	530(0.919)	43(0.075)	3(0.005)	1(0.002)	-	
Fatigue	558(0.967)	12(0.021)	5(0.009)	2(0.003)	-	
Headache	554(0.96)	16(0.028)	3(0.005)	4(0.007)	-	
Hyper-triglyceride	551(0.955)	22(0.038)	4(0.007)	0(0)	-	
Isotretinoin						
Abnormal vision	579(0.983)	8(0.014)	1(0.002)	1(0.002)	0(0)	
Arthralgia	544(0.924)	30(0.051)	10(0.017)	5(0.008)	-	
Cheilitis	212(0.36)	245(0.416)	122(0.207)	10(0.017)	-	
Conjunctivitis	449(0.762)	98(0.166)	31(0.053)	11(0.019)	-	
Fatigue	572(0.971)	14(0.024)	3(0.005)	0(0)	-	
Headache	580(0.985)	9(0.015)	0(0)	0(0)	-	
Hyper-triglyceride	514(0.873)	64(0.109)	10(0.017)	1(0.002)	-	

Table 2.2: Simulated data set, marginal posterior cell probabilities with central 95% credible intervals. In bold are the true cell probabilities.

Group A				
Ordinal				
level	0	1	2	3
T1	0.619(0.64) (0.576, 0.658)	0.299(0.274) (0.264, 0.337)	0.023(0.027) (0.017, 0.031)	0.059(0.059) (0.045, 0.074)
T2	0.65(0.683) (0.609, 0.689)	0.272(0.246) (0.237, 0.309)	0.028(0.021) (0.02, 0.036)	0.05(0.05) (0.037, 0.064)
T3	0.662(0.72) (0.62, 0.703)	0.253(0.22) (0.216, 0.29)	0.029(0.017) (0.019, 0.041)	0.056(0.042) (0.042, 0.072)
T4	0.736(0.757) (0.697, 0.774)	0.183(0.193) (0.15, 0.218)	0.036(0.016) (0.024, 0.049)	0.045(0.034) (0.032, 0.061)
Group B				
T1	0.646(0.64) (0.604, 0.685)	0.274(0.274) (0.239, 0.311)	0.022(0.027) (0.016, 0.03)	0.057(0.059) (0.044, 0.073)
T2	0.581(0.586) (0.538, 0.623)	0.335(0.306) (0.297, 0.372)	0.03(0.036) (0.023, 0.038)	0.054(0.071) (0.041, 0.069)
T3	0.52(0.517) (0.475, 0.564)	0.379(0.348) (0.341, 0.417)	0.036(0.049) (0.027, 0.045)	0.065(0.086) (0.05, 0.082)
T4	0.405(0.432) (0.361, 0.45)	0.469(0.4) (0.434, 0.502)	0.056(0.061) (0.045, 0.069)	0.069(0.107) (0.055, 0.086)

Table 2.3: Marginal posterior cell probabilities (central 95% credible intervals) of toxicity. The table only reports grades up to $k = 3$. The marginal probabilities for *grade G4 abnormal vision* are 0 (with 95% C.I. (0, 0.002)) under placebo, and 0.001 (with 95% C.I., (0, 0.002)) under isotretinoin.

Toxic effect	No Tox	G1	G2	G3
Placebo				
Abn. vision	0.974 (0.962, 0.985)	0.02 (0.011, 0.033)	0.002 (0.001, 0.005)	0.003 (0.001, 0.006)
Arthralgia	0.937 (0.918, 0.954)	0.042 (0.029, 0.057)	0.019 (0.013, 0.026)	0.002 (0.001, 0.005)
Cheilitis	0.808 (0.774, 0.84)	0.172 (0.145, 0.201)	0.02 (0.012, 0.03)	0 (0, 0)
Conjunctivitis	0.896 (0.872, 0.918)	0.077 (0.063, 0.093)	0.025 (0.015, 0.036)	0.001 (0, 0.002)
Fatigue	0.967 (0.952, 0.978)	0.024 (0.015, 0.035)	0.009 (0.005, 0.015)	0.001 (0, 0.001)
Headache	0.962 (0.945, 0.976)	0.029 (0.017, 0.044)	0.008 (0.004, 0.013)	0.001 (0, 0.003)
H.-triglyceride	0.971 (0.955, 0.983)	0.028 (0.017, 0.042)	0.001 (0, 0.003)	0 (0, 0)
Isotretinoin				
Abn. vision	0.972 (0.958, 0.983)	0.023 (0.013, 0.035)	0.002 (0, 0.004)	0.003 (0.001, 0.006)
Arthralgia	0.922 (0.9, 0.941)	0.054 (0.039, 0.073)	0.021 (0.014, 0.028)	0.003 (0.002, 0.006)
Cheilitis	0.398 (0.355, 0.443)	0.403 (0.365, 0.439)	0.177 (0.153, 0.203)	0.022 (0.015, 0.031)
Conjunctivitis	0.777 (0.739, 0.814)	0.163 (0.133, 0.196)	0.052 (0.04, 0.066)	0.007 (0.004, 0.011)
Fatigue	0.959 (0.943, 0.973)	0.029 (0.019, 0.042)	0.011 (0.006, 0.017)	0.001 (0, 0.002)
Headache	0.968 (0.953, 0.98)	0.024 (0.015, 0.036)	0.007 (0.004, 0.012)	0.001 (0, 0.002)
H.-triglyceride	0.851 (0.82, 0.879)	0.086 (0.064, 0.11)	0.056 (0.038, 0.075)	0.007 (0.002, 0.013)

Table 2.4: Probability of different toxicities (at any grade; rows) conditional on the same patient having experienced other toxicities (at any grade; columns). The two numbers x/y in each cell report probabilities under placebo/isotretinoin. For comparison the diagonal reports in bold the marginal probabilities of exhibiting each type of toxicity.

	Abnormal vision (AV)	Arthralgia (AR)	Cheilitis (CH)	Conjunctivitis (CO)	Fatigue (FA)	Headache (HE)	Hyper- triglyceride (HT)
AV	0.026/0.028	0.175/0.186	0.071/0.043	0.108/0.099	0.171/0.209	0.249/0.266	0.078/0.11
AR	0.428/0.518	0.063/0.078	0.12/0.107	0.162/0.199	0.227/0.316	0.326/0.391	0.134/0.201
CH	0.533/0.923	0.366/0.821	0.192/0.602	0.302/0.824	0.348/0.774	0.442/0.823	0.334/0.729
CO	0.438/0.78	0.267/0.566	0.163/0.305	0.104/0.223	0.256/0.511	0.344/0.603	0.176/0.411
FA	0.223/0.306	0.121/0.166	0.061/0.053	0.082/0.095	0.033/0.041	0.168/0.238	0.061/0.105
HE	0.366/0.304	0.195/0.161	0.087/0.044	0.125/0.087	0.189/0.186	0.038/0.032	0.095/0.098
HT	0.089/0.581	0.062/0.382	0.051/0.181	0.049/0.275	0.053/0.381	0.074/0.456	0.029/0.149

Chapter 3

Dirichlet Process Mixture Models for Discrete Human Data in Phage Display Model

3.1 Overview

We develop semi-parametric Bayesian inference for data obtained from a human phage display experiment. The experiment is carried out to learn about preferential binding of proteins in certain organs. The long-term goal is to exploit such knowledge to develop targeted therapies that could deliver a drug to specific tissues and limit side effects such as toxicity (Kolonin et al., 2006; Arap et al., 2006). A phage library is a collection of millions of phages, each displaying different peptide sequences. In a bio-panning experiment the phage display library is exposed to a target. Phages with proteins that do not bind to the target are washed away, leaving only those with proteins that are binding specifically to the target. A critical limitation of the described experiment is the lack of any amplification. Suppose that the peptide A binds to the tissue T with high affinity. In addition, suppose that we are using a library that contains a small amount of peptide A among a large number of different peptides. We may observe only a small count of the peptide A in the tissue T and therefore, we may not detect this binding behavior. To mitigate this limitation Kolonin et al. (2006) proposed to perform multistage phage display experiment, that is, to perform successive stages of panning (usually three or four) to enrich peptides that bind to the targets. This procedure allows for the counts of peptides like A referred above to increase in every stage and, therefore, it increases the chance of detecting their binding behavior.

Kolonin et al. (2006) use a Bayesian beta-binomial model to make a list of the peptides with relatively large increases from the first to the third stages. The outcome of this first step is a list of peptide/tissue pairs with the highest posterior probability of increasing frequencies over the three stages. In a next step, for each peptide A tissue T in this list, they compare the count of peptide A in tissue T in the last (third) stage, versus the count of peptide A in the unselected library (a representative draw from the peptide library injected into the first stage of the phage display experiment). They consider a two-by-two table with counts in tissue T in the first row and counts

in the unselected library in the second row. The two columns are A versus not A. They carry out a test of independence in these two-by-two contingency tables, using Fisher's exact test. Finally, in a third step they consider a similar two-by-two table, but now with the second row reporting counts for all tissues (after stage 3). Again, a test for independence is carried out to test for preferential binding of A to tissue T. A peptide A passing all three filters is reported as binding with high affinity to tissue T.

Ji et al. (2007) proposed a Bayesian hierarchical model as a way of accounting for correlation between measurements and reducing the number of parameters. They used a False Discovery Rate (FDR) criterion (see, Newton, 2004) to report high-binding peptides for a 3-stage phage display experiment with mouse data. Later, in Section 3, we argue that this parametric hierarchical model is inappropriate for the human data described below. Taking advantage of the large sample size of our data set, we propose instead a semiparametric Bayesian model that avoids some of the limitations of the fully parametric model described by Ji et al. Also, we propose an alternative criterion to select high-binding peptides based on a decision theoretic framework.

The main contributions of this paper is the use of a non-parametric prior to avoid the limitations of specific parametric assumptions, and the use of a decision theoretic framework to address the multiplicity issues arising in the selection of a list of tripeptides-tissue pairs that are reported for significant affinity.

The remainder of this chapter is organized as follows. The proposed model is a Dirichlet Process Mixture (DPM) model, Section 3.2 reviews the basic properties of the Dirichlet Process and DPM models, describes the standard MCMC algorithm to simulate posterior samples from this latter model and gives an example of an application of DPM model in density estimation. Section 3.3 presents a detailed description of the multistage human data. In section 3.4 We propose a Bayesian semi-parametric mixture Poisson model and describe the Markov Chain Monte Carlo(MCMC) simula-

tion scheme for obtaining random samples of the posterior distribution of the imputed parameters of the model. In Section 3.5, we discuss the criterion to select the peptides that bind with high affinity to certain tissues. We perform a simulation study to assess further the properties of our model and our peptide selecting criterion in Section 3.6. In Section 3.7, we show the results of applying our model to the human three-stage phage display experiment. Finally, we provide some concluding remarks in Section 3.8.

3.2 Background

The probabilistic model in Section 3.4 is a Dirichlet Process Mixture (DPM) model. The objectives of this section is, first, to review the DPM and, second, to show posterior simulation and posterior predictive simulation for a future observation. With the first objective in mind, we review the Dirichlet distribution, the Dirichlet Process and its basic properties. To explain the main ideas of posterior and posterior predictive simulation we review the Gibbs sampling scheme proposed by MacEachern (1994). Finally, we give an example of an application in density estimation of the DPM.

3.2.1 Dirichlet Distribution

The Dirichlet distribution is the multivariate extension of the beta distribution and is the conjugate prior model for the parameters of a multinomial model.

Let $Ga(\alpha, \beta)$ denote the gamma distribution with shape parameter $\alpha \geq 0$ and scale $\beta > 0$. Define $Ga(\alpha = 0, \beta)$ as a point mass at zero. For $\alpha > 0$ the gamma pdf is

$$Ga(z \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \exp(-z/\beta) z^{\alpha-1} \mathbf{1}_{(0, \infty)}(z), \quad (3.1)$$

where $\mathbf{1}_S(z)$ denotes the indicator function of the set S .

The Dirichlet distribution is defined with all its parameters positive (see Wilks, 1962). Ferguson (1973) extends it to allow some, but not all, parameters to be equal to zero. Let Z_1, \dots, Z_k be independent random $Ga(\alpha_j, 1)$ variables respectively, with $\alpha_j \geq 0$ for every j and $\alpha_j > 0$ for some j , $j = 1, \dots, k$. The Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_k$, here denoted by $\text{Dirichlet}(\alpha_1, \dots, \alpha_k)$, is defined as the distribution of the vector (Y_1, \dots, Y_k) , where

$$Y_j = Z_j / \sum_{i=1}^k Z_i \quad \text{for } j = 1, \dots, k. \quad (3.2)$$

When using the notation $\text{Dirichlet}(\alpha_1, \dots, \alpha_k)$ it is assumed that $\alpha_j \geq 0$ for all j , and $\alpha_j > 0$ for some j .

When $\alpha_j = 0$, the corresponding Y_j is a point mass at 0, i.e. $Y_j = 0$ almost surely. When $\alpha_j > 0$ for all j , the pdf in \mathfrak{R}^{k-1} of (Y_1, \dots, Y_{k-1}) is

$$f(y_1, \dots, y_{k-1}) = \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} \prod_{j=1}^{k-1} y_j^{\alpha_j - 1} \left(1 - \sum_{i=1}^{k-1} y_i\right)^{\alpha_k - 1} 1_S(y_1, \dots, y_{k-1}), \quad (3.3)$$

where S is the set,

$$S = \{(y_1, \dots, y_{k-1}) \in \mathfrak{R}^{k-1} \mid y_j \geq 0, \sum_{j=1}^{k-1} y_j \leq 1\}.$$

When $k = 2$, (3.3) reduces to the Beta distribution, denoted by $\text{Beta}(\alpha_1, \alpha_2)$.

3.2.2 Dirichlet Process and Dirichlet Process Mixture Model

Consider the complete separable metric space (Ferguson, 1974), \mathcal{X} with its Borel σ -algebra \mathcal{A} . Let

$$\mathcal{F} = \{F : F \text{ is a probability measure on } \mathcal{X}\}.$$

Let \mathcal{A} denote some suitable σ -algebra of subsets of \mathcal{F} , for example the Borel sets generated by the topology of weak convergence. We say that \mathcal{P} is a random probability measure (RPM) if \mathcal{P} is a probability measure in $(\mathcal{F}, \mathcal{A})$. If F denotes a random probability chosen according to \mathcal{P} , then $F(B)$ for B measurable set in \mathcal{X} is a random variable. Let x_1, \dots, x_n be a random sample from F . Consider a Bayesian model, with a prior distribution on $F \sim \mathcal{P}$. Inference about F is based on the posterior distribution $F \mid x_1, \dots, x_n$, using the information available in the sample.

With this framework in mind Ferguson (1973) stated two “desirable” characteristic that the RPM, \mathcal{P} , the prior for F , should satisfy: (i) Its support should be large and (ii) the posterior distribution should be mathematically convenient. In the same paper, he introduces and proves the existence of the Dirichlet Process (DP) RPM. A measure F is generated by a DP if for any partition of the sample space, $\{B_1, \dots, B_k\}$,

the vector of random probabilities $(F(B_1), \dots, F(B_k))$ follows a Dirichlet distribution as defined in the previous subsection:

$$(F(B_1), \dots, F(B_k)) \sim \text{Dirichlet}(\alpha F_0(B_1), \dots, \alpha F_0(B_k)).$$

We denote this by $F \sim \text{DP}(\alpha, F_0)$. The DP prior is indexed with two parameters: the precision parameter α and the base measure F_0 . The base measure F_0 defines the expectation $E(F(B)) = F_0(B)$, and α is the precision parameter that defines the variance. For details of the role of these parameters see Walker et al. (1999).

We say that the collection of random elements X_1, \dots, X_n is a **random sample of size n of the Dirichlet Process DP** if given, $F \sim \text{DP}$, $X_1, \dots, X_n \mid F$ is a random sample from F . One of the main characteristics of the DP is that it is conjugate, implying a simple posterior updating rule. Let δ_x denote a point mass at x . If

$$X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F \quad \text{and} \quad F \sim \text{DP}(\alpha, F_0),$$

then the posterior distribution of F is, again, a DP given by

$$F \mid X_1, \dots, X_n \sim \text{DP}(\alpha + n, F_1), \tag{3.4}$$

with $F_1 := E(F \mid X_1, \dots, X_n) \propto F_0 + \sum_{i=1}^n \delta_{X_i}$ a compromise between the empirical distribution and the base measure F_0 (See Figure 3.1).

Sethuraman (1994) shows that any $F \sim \text{DP}(\alpha, F_0)$ can be represented as

$$F(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\mu_h}, \quad \mu_h \stackrel{i.i.d.}{\sim} F_0 \tag{3.5}$$

where

$$w_h = U_h \prod_{j < h} (1 - U_j) \quad \text{with } U_h \stackrel{i.i.d.}{\sim} \text{Beta}(1, \alpha).$$

That is, a realization of a DP is a discrete random measure with countable support. Its support is a sequence of values independently sampled from F_0 with respective jump sizes generated by a “stick breaking” procedure. In particular F is a discrete measure, that is, the DP generates, almost surely, discrete random measures. Figure

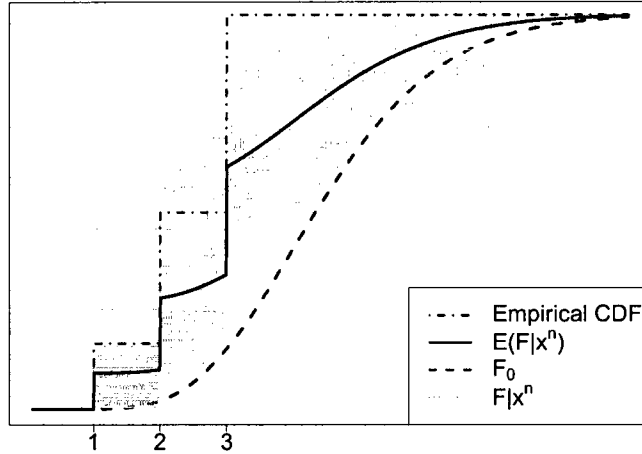


Figure 3.1: Base measure $F_0 = Ga(9, 2)$, posterior base measure $F_1 := E(F | x^n)$, empirical cdf and set of simulated measures sampled from the posterior distribution of the DP($\alpha = 5, F_0$) when a sample $x^n: 1, 2, 2, 3, 3$ and 3 of this DP was observed.

3.1 shows the base measure $F_0 = Ga(9, 2)$, the posterior base measure F_1 , the empirical cdf and a set of simulated measures sampled from the posterior distribution of the DP($\alpha = 5, F_0$) via (3.5) when the sample of size $n = 6$ of this DP: 1, 2, 2, 3, 3 and 3 was observed.

The DP allows analytic evaluation of the marginal distribution $p(X_1, \dots, X_n)$, integrating with respect to $F \sim DP(\alpha, F_0)$. The marginal distribution is also known as the Pólya urn (Blackwell and MacQueen, 1973). We will exploit the Pólya urn later in this section to simulate from a DP. This representation, in particular, implies that if X_1, \dots, X_n is a random sample of this sample: X_1^*, \dots, X_K^* are a random sample from F_0 . This implies that

$$Pr[X_1 \in \cdot] = F_0(\cdot). \quad (3.6)$$

The predictive distribution (3.4) can be shown to be

$$Pr[X_i \in \cdot | X_1, \dots, X_{i-1}] = \frac{\alpha}{i} F_0(\cdot) + \frac{1}{i} \sum_{j=1}^{i-1} \delta_{X_j}(\cdot), \quad \text{for } i = 2, \dots, n. \quad (3.7)$$

Multiplying $p(X_1)$ and $p(X_i | X_1, \dots, X_{i-1})$, $i = 2, \dots, n$, defines the joint marginal distribution $p(X_1, \dots, X_n)$. When F_0 has finite support and $\alpha = 1$, the two expressions above imply that sampling (X_1, \dots, X_n) can be interpreted as drawing from an urn whose initial proportion of balls of color x is $F_0(\{x\})$. We draw the i -th ball from the urn, its color X_i is registered, the ball is put back into the urn and a new ball of the same color as the one just extracted is added to the urn.

Mostly due to its computational advantages and the easy interpretation of the parameters, the DP is the most popular nonparametric Bayesian prior. Nevertheless, an almost surely discrete prior is not desirable in many applications. A simple solution is to consider a convolution of $G \sim DP$ with a continuous kernel. Such a (hierarchical) model is known as Dirichlet process mixtures (DPM) (MacEachern, 1994, Escobar and West, 1995):

$$\begin{aligned} x_i | \theta_i &\sim F(x_i | \theta_i) \text{ and independent for } i = 1, \dots, n \\ \theta_1, \dots, \theta_n &| G \stackrel{i.i.d.}{\sim} G \\ G &\sim DP(\alpha, G_0), \end{aligned} \quad (3.8)$$

i.e. $x_i \sim \int F(x_i | \theta) dG(\theta)$ and $G \sim DP(\alpha, G_0)$.

The model above is easily extended to consider non identically distributed samples such as the regression model: $x_i | \theta_i \sim N(z_i^T \theta_i, \sigma)$ where z_i is a vector of covariates of the same dimension as θ_i and σ the common precision. In general, suppose that the distributions of each x_i are (possibly different) known distributions F_i , indexed with θ_i , and possibly additional parameters σ that are common across all i , i.e.,

$$X_i | \theta_i, \sigma \sim F_i(X_i | \theta_i, \sigma).$$

Moreover, we can include a hyperprior distribution for the parameter α and, besides, consider that the base measure G_0 depends on parameters γ with prior distribution

$p(\gamma)$. The general DPM becomes

$$\begin{aligned}
I. & \quad x_i \mid \theta_i, \sigma \sim F_i(x_i \mid \theta_i, \sigma) \text{ and are independent, for } i = 1, \dots, n, \\
II. & \quad \theta_1, \dots, \theta_n \mid G^{iid} \text{ and } \sigma \sim p(\sigma), \\
III. & \quad G \mid \alpha, \gamma \sim DP(\alpha, G_0(\cdot, \mid \gamma)) \\
IV. & \quad \alpha \sim p(\alpha) \text{ and } \gamma \sim p(\gamma).
\end{aligned} \tag{3.9}$$

The levels I-III (with α and γ fixed) correspond to the model in Lo(1984), while levels II-IV correspond, essentially, to the Antoniak (1974) model. Escobar and West (1995) join both models including all the levels (I-IV). F_i determines if the distribution of x_i is continuous or discrete. In the following subsection we describe the Gibbs sampling scheme for posterior simulation of $\theta_1, \dots, \theta_n$.

3.2.3 Gibbs Sampling Scheme for DPM

This subsection summarizes posterior MCMC simulation in DP mixture models as described in West et al. (1994). We describe posterior MCMC in the mixture model defined in (3.9).

Complete conditional posterior distribution of θ_i : We use transition probabilities defined by sampling θ_i from its complete conditional posterior given the currently imputed values of the other θ_j 's, all other parameters and the observed data. For the moment, we assume that the parameters α and σ are fixed and suppress them in the notation. We will include them later.

Assume $\theta_i \sim G$ and $G \sim DP(\alpha, G_0)$. Let $\theta^{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n)$. The updating rule (3.4) of the DP implies, $G \mid \theta^{-i} \sim DP(\alpha + n - 1, \alpha G_0 + \sum_{j \neq i} \delta_{\theta_j})$. We can marginalize G , using the Pólya urn, and find

$$Pr[\theta_i \in \cdot \mid \theta^{-i}] = \frac{\alpha}{\alpha + n - 1} G_0(\cdot) + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\theta_j}(\cdot). \tag{3.10}$$

That is, we can directly simulate from $\theta_i \mid \theta^{-i}$, without any need to generate the RPM G .

Expression (3.10) tells us that the probability that θ_i differs from θ_j , for $j \neq i$, is $\alpha/(\alpha + n - 1)$. Moreover, denoting with g_0 the pdf corresponding to G_0 ,

$$p(\theta_i | \theta^{-i}) = \frac{\alpha}{\alpha + n - 1} g_0(\theta_i) + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\theta_j}(\theta_i). \quad (3.11)$$

The previous equation, conditional independence of x_1, \dots, x_n given $\theta_1, \dots, \theta_n$ and Bayes theorem imply

$$\begin{aligned} p(\theta_i | \theta^{-i}, x_1, \dots, x_n) &\propto p(x_1, x_2, \dots, x_n | \theta_i, \theta^{-i}) p(\theta_i | \theta^{-i}) \\ &\propto \prod_{j=1}^n p(x_j | \theta_j) \left\{ \alpha g_0(\theta_i) + \sum_{j \neq i} \delta_{\theta_j}(\theta_i) \right\} \\ &\propto \alpha p(x_i | \theta_i) g_0(\theta_i) + \sum_{j \neq i} p(x_i | \theta_j) \delta_{\theta_j}(\theta_i). \end{aligned} \quad (3.12)$$

Let $p(x_i) = \int p(x_i | \theta) g_0(\theta) d\theta$ denote the marginal of x_i under g_0 . Then

$$p(x_i | \theta_i) g_0(\theta_i) = \frac{p(x_i | \theta_i) g_0(\theta_i)}{p(x_i)} p(x_i) = p(\theta_i | x_i) p(x_i),$$

we get,

$$p(\theta_i | \theta^{-i}, x_1, x_2, \dots, x_n) \propto q_{i0} g_{i0}(\theta_i) + \sum_{j \neq i} q_{ij} \delta_{\theta_j}(\theta_i), \quad (3.13)$$

where $q_{i0} = \alpha p(x_i)$ is the product of α and the marginal distribution $p(x_i)$ of x_i ; $q_{ij} = p(x_i | \theta_j)$; and $g_{i0}(\theta_i) \propto p(x_i | \theta_i) g_0(\theta_i)$, the posterior distribution of θ_i given x_i under a Bayes model with prior $\theta_i \sim G_0$ and sampling model $p(x_i | \theta_i)$.

We observe that a random sample of size n : $\theta_1, \theta_2, \dots, \theta_n$ of a DP can be equivalently represented by the triplet (K, θ^*, φ) where K is the number of distinct values among the θ_i , $\theta^* = \{\theta_1^*, \theta_2^*, \dots, \theta_K^*\}$ are these distinct values, and φ are indicators $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_n)$, with $\varphi_i = k$ if $\theta_i = \theta_k^*$. We refer to (K, θ^*, φ) as a ‘‘configuration.’’

A configuration (West, 1990, MacEachern, 1994) classifies the data $x^n := \{x_1, \dots, x_n\}$ into K different clusters with $n_k = \#\{i | \varphi_i = k\}$ observations that share the common parameter θ_k^* . We use S_k to denote the k -th cluster of observation indices, $S_k = \{i | \varphi_i = k\}$. In other words, a configuration is simply a partition of $\{1, \dots, n\}$

into clusters defined by the unique values θ_j^* , together with a list of these unique values. In the Gibbs sampling scheme we will update, in the first step, the configuration given the previous one, and, in the second step, $\theta_1^*, \theta_2^*, \dots, \theta_K^*$ given φ and K .

Moreover, since the θ_k^* 's are a random sample from the base measure G_0 , The unique values θ_k^* 's are conditionally independent given φ , with posterior densities:

$$\begin{aligned} p(\theta_k^* | x^n, \varphi, K) &\propto p(x^n | \theta_k^*, \varphi, K) p(\theta_k^* | \varphi, K) \\ &\propto \left\{ \prod_{i \in S_k} f_i(x_i | \theta_k^*) \right\} g_0(\theta_k^*). \end{aligned} \quad (3.14)$$

Denoting by K^{-i}, n_k^{-i} and S_k^{-i} for $k = 1, \dots, K^{-i}$ and $\theta^{*-i} := (\theta_1^{*-i}, \dots, \theta_{K^{-i}}^{*-i})$ the configuration corresponding to the random sample θ^{-i} , the conditional prior (3.10) is equivalent to

$$Pr[\theta_i \in \cdot | \theta^{-i}] = \frac{\alpha}{\alpha + n - 1} G_0(\cdot) + \frac{1}{\alpha + n - 1} \sum_{k=1}^{K^{-i}} n_k^{-i} \delta_{\theta_k^{*-i}}(\cdot). \quad (3.15)$$

In words, θ_i is different from the other parameters and drawn from G_0 with probability proportional to α , and otherwise equal to the k -th already observed value, θ_k^{*-i} , with probability proportional to the number of times this value has been observed in the sample θ^{-i} , i.e., $\propto n_k^{-i}$.

The extension of the expression (3.15) from n to $n + 1$ yields to the predictive distribution of a new value θ_i with $i = n + 1$. This distribution is identical to the expected value of G given θ^*, φ, K . This is easily seen by

$$p(\theta_{n+1} | \theta^*, K, \varphi) = \int p(\theta_{n+1} | G) dp(G | \theta^*, K, \varphi) = \int G(\theta_{n+1}) dp(G | \theta^*, K, \varphi) = \bar{G}.$$

Thus, once we have it, the posterior sample of the parameters can be used to estimate G . The predictive distribution (on the random effects) is

$$Pr[\theta_{n+1} \in \cdot | \theta^*, \varphi, K] = E(G | \theta^*, \varphi, K) = \frac{\alpha}{\alpha + n} G_0(\cdot) + \frac{1}{\alpha + n} \sum_{k=1}^K n_k \delta_{\theta_k^*}(\cdot). \quad (3.16)$$

Therefore, the posterior distribution of a future observation x_{n+1} given a configuration is

$$Pr[x_{n+1} \in \cdot | \theta^*, \varphi, K] = \frac{\alpha}{\alpha + n} F_{n+1}(\cdot | \theta_{n+1}) + \frac{1}{\alpha + n} \sum_{k=1}^K n_k F_{n+1}(\cdot | \theta_k^*), \quad (3.17)$$

where θ_{n+1} is a new sample from G_0 and F_{n+1} is the cdf of x_{n+1} . The Gibbs sampling scheme to simulate parameters from the posterior distribution is based on the above discussion and described next.

Gibbs Sampling Scheme: MacEachern (1994) introduced the following Gibbs sampling scheme to simulate random samples $\theta_1, \dots, \theta_n$ from the posterior under model (3.9). The posterior is easiest described in terms of the configuration parameters θ^* , K and φ . We are still assuming that σ , γ and α are known, the simulation of these parameters can be added, as we will see later, straightforward to the Gibbs sampling scheme, so that, we keep these variables out of the notation. Call (3.13),

$$p(\theta_i | x^n, \theta^{-i}, \varphi^{-i}, K^{-i}) = q_{i0}g_{i0}(\theta_i) + \sum_{k=1}^{K^{-i}} q_{ik}\delta_{\theta_k^{-i}}(\theta_i). \quad (3.18)$$

Under model (3.9) the weights q_{ik} are given by,

$$q_{ik} = \begin{cases} c\alpha h_i(x_i), & k = 0, \\ cn_k^{-i} f_i(x_i | \theta_k^*), & 1 \leq k \leq K^{-i}. \end{cases} \quad (3.19)$$

Here f_i is the pdf corresponding to F_i , g_{i0} is the posterior pdf of θ_i , obtained by updating g_0 with the likelihood $f_i(x_i | \theta_i)$, that is,

$$g_{i0}(\theta_i) \propto f_i(x_i | \theta_i)g_0(\theta_i), \quad (3.20)$$

whose normalization constant, $h_i(x_i)$, is the marginal density of x_i ,

$$h_i(x_i) = \int f_i(x_i | \theta_i)g_0(\theta_i) d\theta_i,$$

and c is a normalization constant (across $k = 0, \dots, K^{-i}$).

Equation (3.18) implies the posterior distribution for the indicator variables in the configuration,

$$Pr(\varphi_i = k | x^n, \theta^{-i}, \varphi^{-i}, K^{-i}) = q_{ik}. \quad (3.21)$$

We can simulate samples of θ^* , K and φ by iterating over the following transition probabilities:

- (a) Given the previous values for θ^* , K and φ , we generate a new configuration by sampling new values for the indicator variables using (3.21), replacing $\theta_1^*, \theta_2^*, \dots$. When sampling $\varphi_i = 0$ we associate the observation i , x_i , with a new draw from g_{i0} given in (3.20) and update the configuration accordingly.

Note that using a base measure G_0 conjugate to f_i is mathematically convenient. It makes sampling from g_{i0} and computation of h_i easier. For computational reasons conjugate families should be considered when appropriate. In the model introduced in Section 3.4 this is the case.

Repeated use of step (a) defines a Markov chain with limiting distribution equal to the posterior distribution $p(\theta^*, K, \varphi \mid x^n)$. But, this Markov chain mixes poorly. Since it shifts one value of θ_i at a time, rarely does a value of θ_k^* in a cluster change. This change needs the chain to pass through a middle state of low probability in which all indices in the cluster S_k are moved to other clusters. Convergence is accelerated by avoiding this phenomenon. Once the values of θ^* , K and φ have been obtained in step (a), the vector θ^* is resampled conditional on the values of K and φ . That is, a second step is added to the Gibbs sampling scheme (after each step (a)),

- (b) Given K and φ , we draw a new set of parameters θ^* by sampling new values from the posterior distribution (3.14) .

Sequential iterations over (a) and (b) defines a Markov chain that converges to the posterior distribution $p(\theta^*, K, \varphi \mid x^n)$. For details about this convergence see MacEachern and Müller (1998) . The posterior distribution of any function of the parameters can be estimated based on the posterior sample. For example the average of (3.17), with respect to the simulated values of $\theta_1^*, \dots, \theta_K^*$, and θ_{n+1} estimates the predictive distribution $p(x_{n+1} \mid x^n) = \overline{G}$. This average is used for the density estimation. See the example given in the next subsection.

Now we extend the Gibbs sampling scheme to include the common parameter σ . Let $p(\sigma)$ denote the prior for σ . We draw this parameter from the appropriate

posterior conditional distribution,

$$\begin{aligned}
 p(\sigma \mid x^n, \theta^*, \varphi, K) &\propto p(x^n \mid \theta^*, \varphi, K, \sigma) p(\sigma \mid \theta^*, \varphi, K) \\
 &\propto p(\sigma) \prod_{i=1}^n f_i(x_i \mid \theta_i, \sigma) \\
 &\propto p(\sigma) \prod_{k=1}^K \prod_{k \in S_k} f_i(x_i \mid \theta_k^*, \sigma).
 \end{aligned} \tag{3.22}$$

Therefore, the Gibbs sampling scheme given by (a) and (b) is extended, when necessary, by adding step (c):

- (c) Generate σ conditional on the imputed values of the parameters θ^* , φ and K using the expression (3.22).

The distributions given in (3.14) and (3.21) used in steps (a) and (b) are conditional on $\{\alpha, \gamma\}$. The Gibbs sampling scheme can be extended to include $\{\alpha, \gamma\}$ by adding an extra step: sampling from the appropriate posterior distribution, $p(\alpha, \gamma \mid x^n, \theta, \sigma) = p(\alpha, \gamma \mid x^n, \theta^*, \varphi, K, \sigma)$. This can be done (West, 1992 and Escobar and West, 1995) in the following way:

Suppose that α and γ are *a priori* independent with densities

$$p(\alpha, \gamma) = p(\alpha)p(\gamma).$$

Model (3.9) implies that α and γ given the parameters θ^* , K , φ , and σ , remain independent. Therefore, α and γ can be considered separately. Due to the nature of the DP (see Antoniak, 1974), only the value of K matters in the posterior distribution of α , i.e.,

$$p(\alpha \mid x^n, \theta^*, \varphi, K, \sigma) = p(\alpha \mid K).$$

West (1992) proposed a model augmentation with a latent Beta random variable for the parameter α . If the prior distribution $p(\alpha)$ is $Ga(a, b)$, then the posterior distribution $p(\alpha \mid K)$ is a mixture of gamma distributions. This allows easy simulation from $p(\alpha \mid K)$. A new step is added to the Gibbs sampling scheme,

- (d) Update the total mass parameter α :

1. Draw a latent variable $\eta \sim \text{Beta}(\alpha + 1, n)$.
2. Sample α from the mixtures of gammas:

$$\pi_\eta \text{Ga}(\alpha \mid a + K, b - \log \eta) + (1 - \pi_\eta) \text{Ga}(\alpha \mid a + K - 1, b - \log \eta), \quad (3.23)$$

where,

$$\frac{\pi_\eta}{1 - \pi_\eta} = \frac{\alpha K - 1}{n(b - \log \eta)}$$

Finally, we define a transition probability to update γ . Recall that γ is introduced into the model through G_0 . We get, from (3.14),

$$p(\gamma \mid x^n, \theta^*, \varphi, K, \sigma) = p(\gamma \mid \theta^*, K) \propto p(\gamma) \prod_{k=1}^K g_0(\theta_k^* \mid \gamma). \quad (3.24)$$

For computational reasons, in applications it is convenient to assume a prior distribution $p(\gamma)$ that is conjugate to g_0 . We need to add a new step to our Gibbs sampling scheme:

- (e) Conditional on K and φ we update γ using (3.24).

As technical note, in their examples Escobar and West (1995) noticed that the MCMC scheme stated above can get “trapped” in local modes of the posterior distribution. To “free” it, instead of running a much longer chain, they suggest to reinitialize every certain number of Gibbs sampling iterations, for example, 10,000 by the configuration with $K = n$ and new values for $\theta_1^*, \dots, \theta_n^*$ resampled from g_{i0} in (3.20) without changing the current values of the remaining parameters.

3.2.4 Example of Application of DPM in Density Estimation

In this section we present an application of a Dirichlet Process Mixture (DPM) model to non-parametric density estimation.

We use the Boston housing-price data from Harrison and Rubinfeld (1978). This data appears online as Boston data in the statlib index. The file contains measurements of 13 characteristics of houses in Boston and their price (named MEDV in

the file). We apply the DPM model to estimate the distribution of the house prices (disregarding the rest of the information in the file).

A particular case of model (3.9) is given in Escobar and West (1995), and implemented in the R-package “DPpackage” (Alejandro Jara, 2007):

$$\begin{aligned}
& I. \quad y_i \mid \mu_i, \lambda_i \sim N(\mu_i, \lambda_i) \text{ for } i = 1, \dots, n, \\
& II. \quad (\mu_1, \lambda_1), \dots, (\mu_n, \lambda_n) \mid G \stackrel{\text{iid}}{\sim} G \\
& III. \quad G \mid \alpha, k_0, \mu_1, \psi_1 \sim \text{DP}(\alpha, G_0(\cdot \mid k_0, \mu_1, \psi_1)), \\
& \quad \quad G_0(\mu_i, \lambda_i \mid k_0, \mu_1, \psi_1) = N(\mu_i \mid m_1, k_0 \lambda_i) \text{ Ga}(\lambda \mid \eta_1, \psi_1) \\
& IV. \quad \alpha \sim \text{Ga}(a_\alpha, b_\alpha), m_1 \sim N(m_2, s_2), \\
& \quad \quad \psi_1 \sim \text{Ga}(\eta_2, \psi_2) \text{ and } k_0 \sim \text{Ga}(a_{k_0}/2, b_{k_0}/2),
\end{aligned} \tag{3.25}$$

where $N(m, \tau)$ denotes the normal distribution with mean m and precision τ , and $\text{Ga}(\alpha, \beta)$ the gamma distribution with mean α/β ; $\eta_1, a_\alpha, b_\alpha, s_2, \eta_2, \psi_2, a_k, b_k > 0$ and $m_2 \in \mathfrak{R}$.

Notice that the variables $y_i, (\mu_i, \lambda_i)$ and (m_1, k_0, ψ_1) of the model in (3.25) play the role of the variables x_i, θ_i and γ , respectively, in the general DPM (3.9). Let $y^n = \{y_1, \dots, y_n\}$ represent the prices of $n = 506$ houses. Besides, we denote with $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_K^*)$ and $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_K^*)$ the vectors of size K containing the unique values of the μ_i and the λ_i respectively. We use the notation of the previous section for the variables related with the configurations. When stating a complete conditional posterior distribution in this section, we make explicit only the relevant quantities (for example, since the complete conditional posterior distribution of k_0 depends only on $m_1, \boldsymbol{\mu}^*$ and $\boldsymbol{\lambda}^*$, we just write $k_0 \mid m_1, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*$).

We now describe steps (a)-(e) of the MCMC sampling scheme given in the previous section for this specific example. There is no common parameter σ so that step (c) is not necessary. Let $t(y \mid m, s, \nu)$ denote a student t distribution with location m , scale s and degrees of freedom ν .

- (a) Given the currently imputed values of $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$, K and φ , generate a new configuration by simulating $\varphi_1, \dots, \varphi_n$ from the complete conditional posterior distribution

bution,

$$P(\varphi_i = k \mid y_i, \boldsymbol{\mu}^{*-i}, \boldsymbol{\lambda}^{*-i}, \varphi^{-i}, K^{-i}) = q_{ik}, \quad \text{for } k = 0, \dots, K^{-i}$$

where,

$$\begin{aligned} q_{i0} &= c\alpha \text{ Student} \left(y_i \mid m_1, \frac{\eta_1}{\psi_1} \frac{k_0}{k_0 + 1}, 2\eta_1 \right), \\ q_{ik} &= cn_k^{-i} N(y_i \mid \mu_k^*, \lambda_k^*) \quad \text{for } k = 1, \dots, K^{-i}, \end{aligned}$$

and c is a normalization constant such that $q_0 + \dots + q_{K^{-i}} = 1$. Whenever we sample $\varphi_i = 0$, we generate a new observation (μ^*, λ^*) from

$$\begin{aligned} g_{i0}(\mu^*, \lambda^*) &= N \left(\mu^* \mid \frac{k_0 m + y_i}{k_0 + 1}, (k_0 + 1) \lambda^* \right) \\ &\quad \times Ga \left(\lambda^* \mid \eta_1 + n_k/2, \psi_1 + \frac{1}{2} \frac{k_0}{k_0 + 1} (y_i - m_1)^2 \right) \end{aligned}$$

and update, accordingly, the new configuration by $K = K + 1$ and $\varphi_i = K + 1$.

- (b) Given K and φ , generate a new set of parameters (μ_k^*, λ_k^*) for $k = 1, \dots, K$ from the distribution

$$\begin{aligned} p(\mu_k^*, \lambda_k^* \mid y^n, \varphi, K, k_0, \psi_1) &= \\ &N \left(\mu^* \mid \frac{k_0 + n_k \bar{y}_k m}{k_0 + n_k}, (k_0 + n_k) \lambda^* \right) \\ &\times Ga \left(\lambda_k^* \mid \eta_1 + n_k/2, \psi_1 + \frac{1}{2} \left[\sum_{i \in S_k} (y_i - \bar{y}_k)^2 + \frac{n_k}{k_0 + n_k} (m_1 - \bar{y}_k)^2 \right] \right), \end{aligned}$$

where \bar{y}_k is the mean of the observations in S_k , that is $\bar{y}_k = \sum_{i \in S_k} y_i / n_k$.

- (d) Updating the total mass parameter α :

1. Let α denote the currently imputed parameter value. Generate a latent random variable $\eta \sim \text{Beta}(\alpha + 1, n)$.
2. Sample the new value of α from

$$\begin{aligned} p(\alpha \mid K, \eta) &= \pi_\eta Ga(\alpha \mid a_\alpha + K, b_\alpha - \log \eta) \\ &\quad + (1 - \pi_\eta) Ga(\alpha \mid a_\alpha + K - 1, b_\alpha - \log \eta), \end{aligned}$$

where,

$$\frac{\pi_\eta}{1 - \pi_\eta} = \frac{a_\alpha + K - 1}{n(b_\alpha - \log \eta)}.$$

e Hyperparameters of G_0 , γ : simulate from the complete conditional posterior distributions

$$p(m_1 | \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*, r) = N \left(m_1 \mid \frac{m_2 s_2 + k_0 \sum_{k=1}^K \mu_k^* \lambda_k^*}{s_2 + k_0 \sum_{k=1}^K \lambda_k}, s_2 + k_0 \sum_{k=1}^K \lambda_k \right),$$

$$p(\psi_1 | K, \boldsymbol{\lambda}^*) = Ga \left(\psi_1 \mid \eta_2 + K \eta_1, \psi_2 + \sum_{k=1}^K \lambda_k^* \right),$$

and,

$$p(k_0 | \boldsymbol{\tau}^*, m_1, r) = Ga \left(k_0 \mid \frac{K + \alpha_{k_0}}{2}, \frac{\beta_{k_0} + \sum \tau_k^* (\mu_k^* - m_1)^2}{2} \right).$$

Finally, we estimate the density of the house prices at a point y_f based on the (approximated) sample of size M from the joint posterior distribution of the parameters: $(\boldsymbol{\mu}^{*m}, \boldsymbol{\lambda}^{*m}, \varphi^m, K^m, \alpha^m, k_0^m, m_1^m, \eta^m)$, for $m = 1, \dots, M$, via (3.17) by,

$$p(y_f | y^n) = \frac{1}{M} \sum_{m=1}^M \left\{ \frac{\alpha^m}{\alpha^m + n} N(y_f | \mu_{n+1}^{*m}, \lambda_{n+1}^{*m}) + \frac{1}{\alpha^m + n} \sum_{k=1}^{K^m} n_k^m N(y_f | \mu_j^{*m}, \lambda_j^{*m}) \right\},$$

where $(\mu_{n+1}^{*m}, \lambda_{n+1}^{*m})$ is a new sample from g_{i0} given in (3.2.4) whose parameters are specified by the corresponding parameters of the m -th simulated observation: k_0^m, m_1^m, η^m .

We use the same values for the hyperparameters given in the set of hyperparameters “prior4” in the example in the help of the function “DPdensity” in the R package “DPpackage”. In the R example, the density of the velocities, relative to our own galaxy, of 82 galaxies from six well-separated conic sections of the space is estimated. The values are $a_\alpha = 2, b_\alpha = 1, m_2 = 0, s_2 = 10^{-5}, \alpha_{k_0} = 1, \beta_{k_0} = 100, \eta_1 = \eta_2 = 3$, and $\psi_2 = 0.5$. After a 1000 iteration burn-in period we generated $M = 10,000$ observations from the posterior distribution of the parameters by taking an observation every ten Gibbs iterations. All this using the R function “DPdensity”. The estimated density is shown in Figure 3.2.

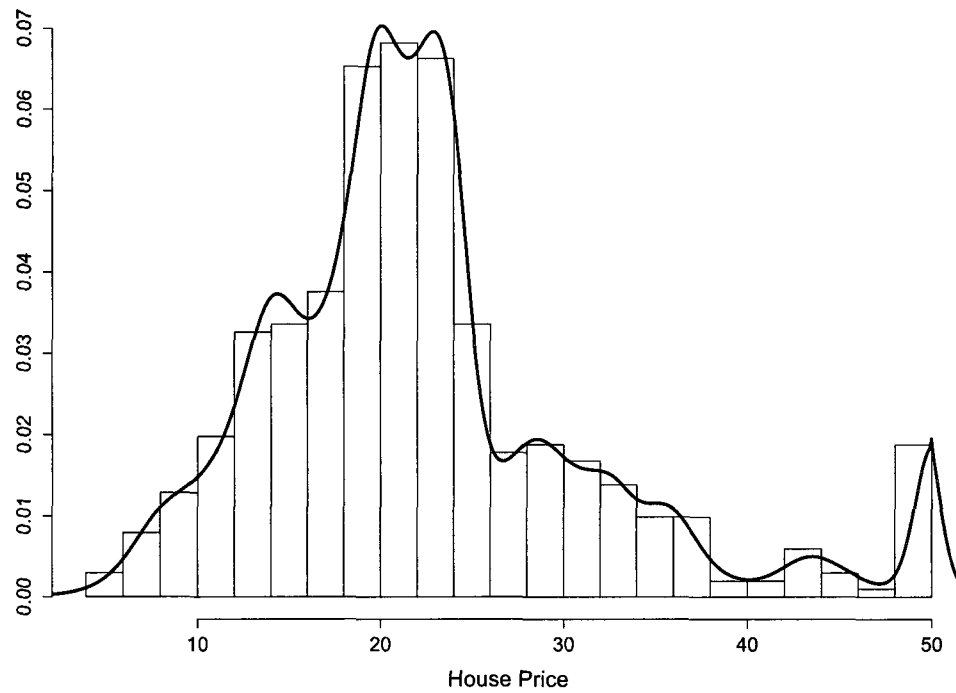


Figure 3.2: Histogram of the house prices in Boston. The continuous line is the nonparametric estimation of the corresponding density function using the MDP model in (3.25)

3.3 Data

We now return to the biopanning experiment described in Section 3.1. The data are from a biopanning experiment carried out at M. D. Anderson Cancer Center. The data come from three consecutive human subjects who met the formal criteria for brain-based determination of death (Wijdicks, 2001). See Arap et al. (2002) for details on patient selection and clinical procedure following clinical ethics criteria..

The purpose of the experiment is to identify peptides with increasing counts over the consecutive stages. At each stage we record counts for peptide/tissue pairs. Peptides are denoted as CX_7C (here C =cysteine, X = any amino acid, represented by a letter). Tissues are Bone-Marrow, Fat, Muscle, Prostate and Skin.

At each stage a phage display peptide library was injected into a new patient, and 15 minutes later biopsies were collected from each of the target tissues and the peptide counts were recorded. For the second and third stage the injected phage display peptide library was the already enriched phage display library from the previous stage.

The original data are counts for all unique 7-mers X_7 . However, we summarize the data using all implied 3-mers. For example, the 7-mer AGAGADR corresponds to the four unique tripeptides AGA,GAG,DAG and ADR. Note that we do not distinguish between a tripeptide and its mirror (e.g., DAG and GAD are counted as the same) and each tripeptide contained in a 7-mer is counted only once (e.g., the count on AGA is incremented only once, although it is contained twice in the 7-mer). So, an observed 7-mer AGAGADR contributes a count for the four tripeptides AGA,GAG,DAG and ADR. The main reason for recording 3-mers are problems related to sparse counts that would result from recording the 20^7 possible 7-mers. In contrast there are only 4200 (20^3 , minus duplicate mirrors) tripeptides. It is believed that the 4200 distinct 3-mers are still a sufficiently rich class to differentiate between binding sites. See, for example, Arap et al. (2002), Ji et al. (2006) and Kolonin et al. (2006) who also use tripeptides. Finally, the data corresponding to the third stage contains two seven-

peptides, 'XRGFRAA' in muscle and 'KTXXARX' in skin where one or more amino-acids, denoted by an 'X', are not identified. We consider the four tripeptides identified in the first one and discard counts for the second seven-peptide.

The data reports counts for 4200 tripeptides and 5 tissues over 3 consecutive stages. For the analysis, we excluded tripeptide-tissue pairs for which the sum of their counts over the three stages was below 5, leaving $n = 2763$ distinct pairs. Figure 3.3 shows the parallel coordinates plot of raw data for these tripeptides-tissue pairs. The desired inference is to identify tripeptide-tissue pairs with an increasing pattern across the three stages, i.e., to mark lines in the figure that show a clear increasing trend from first to third stage. Some lines can be clearly classified as increasing, without reference to any probability model. But for many lines the classification is not obvious. The purpose of the proposed model-based approach is to define where to draw the line to define a significant increase.

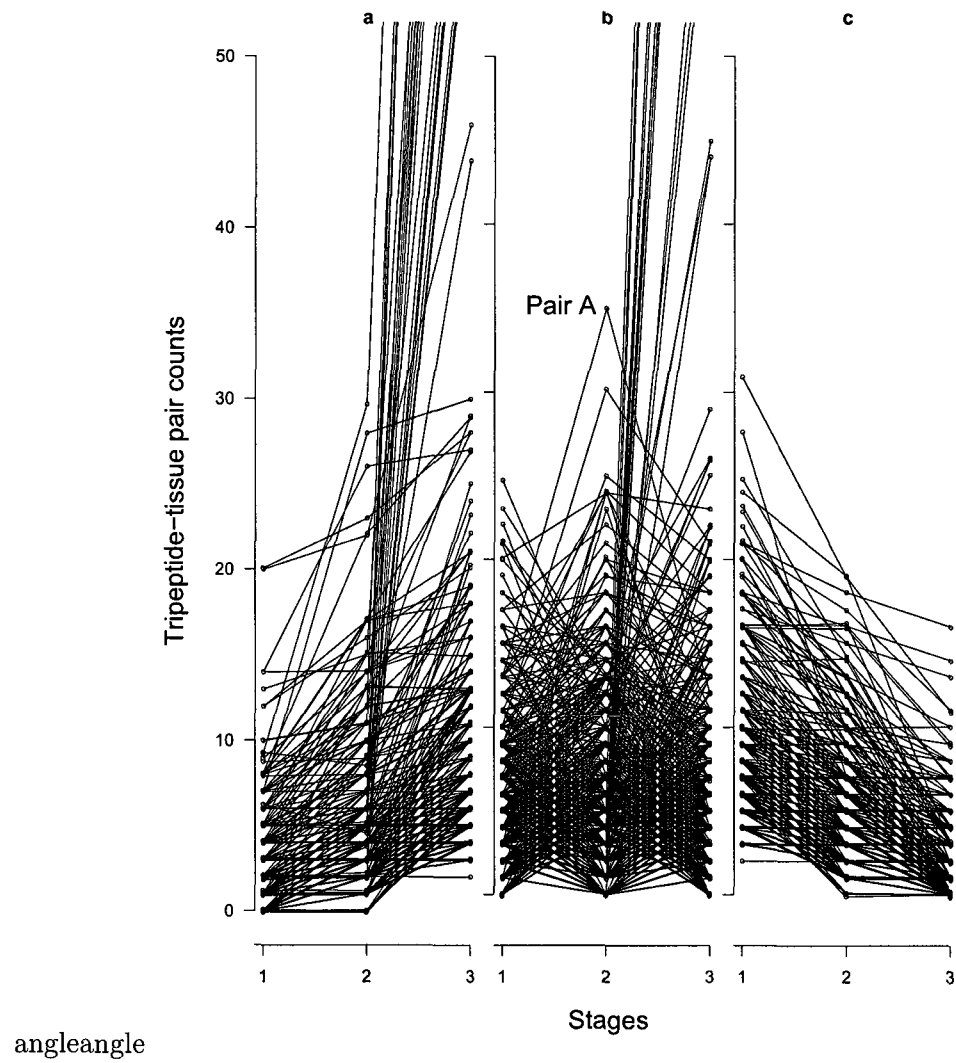


Figure 3.3: Observed sequence of tripeptide-tissue pair counts across the three stages. Each line represents the three observed counts of a pair. The three panels depicts the pairs with: (a) non decreasing, (b) oscillating and (c) non increasing counts.

3.4 Model

3.4.1 A Semi-parametric Mixture of Poisson Model

Our approach is model-based. we will use $N_i = (N_{i1}, N_{i2}, N_{i3})$ to denote the observed counts for tripeptide-tissue pair i across the three stages, for pairs $i = 1, \dots, n$. We cast the desired selection of tissue-specific tripeptides as inference about the increasing trend of mean counts in a probability model for the observed data N_i . Ji et al. (2007) have proposed a model-based approach based on a mixture of normal distributions in the parameters of the assumed Poisson distribution of the observations. The strength and attraction of their model is its parsimonious nature. For the relatively small mouse data set, this characteristic is important. Nevertheless, this model has a limitation. The human data analyzed here is ten-fold larger. This allows us to consider a more elaborate model. The model in Ji et al. assumes a linear relation on the log-Poisson scale. For example, consider the pairs that are reported as oscillating in Figure 3.3 (b). Although the data shows a marked difference in slopes from stages 1 to 2 versus from stages 2 to 3, the model assumes one common slope. This is a concern when the imputed overall slope is positive, e.g., the pair marked by A in the Figure 3.3 (b). Outliers like pair A in Figure 3.3 (b) can inappropriately drive the inference. Taking advantage of the larger sample size, the semiparametric nature of the model we propose can mitigate this problem. Finally, in Ji et al. binding tripeptides are reported in terms of statistical significance, formalized as the posterior probability of the overall slope being greater than zero. We will propose an approach that also takes into account the size of the overall slope and is more suitable to incorporate biological significance.

In summary, the choice of an appropriate probability model is driven by the following considerations. First, we wish to limit the impact of specific parametric modeling choices on the inference about monotonicity of the mean counts. The large number of recorded pairs allows us to use a semi-parametric approach that reduces dependence

on a specific parametric model. Second, we will later build on the probability model to define a formal decision problem for the selection of a final list of tripeptide-tissue pairs. For this, and for a simulation study to validate the model, we have to rely on efficient and fast computations.

These two competing desiderata lead us to consider a semi-parametric mixture model. We will use a mixture of parametric models, with a nonparametric prior on the mixing measure. Under the Bayesian paradigm, nonparametric priors refer to probability models on probability distributions. The non-parametric prior on the mixing measure greatly generalizes the underlying parametric model, in much the same way as a mixture of independent normal kernels can approximate arbitrary multivariate distributions in kernel density estimates. Similar semi-parametric mixture models have successfully been applied for Bayesian inference in a variety of other applications, including, for example, Müller and Rosner (1997), Mukhopadhyay and Gelfand (1997), and Kleinman and Ibrahim (1998). The special case of binary outcomes has been discussed, among many others, by Basu and Mukhopadhyay (2000).

We start with a sampling model for N_i conditional on assumed mean counts across stages for the peptide-tissue combination i . Conditional on the mean counts we assume independent Poisson sampling. In anticipation of the final inference goal we parameterize the mean counts as $(\mu_i, \mu_i\beta_i, \mu_i\delta_i)$, allowing us to describe increasing mean counts by the simple event $1 < \beta_i < \delta_i$. We write $\text{Poi}(x | m)$ to indicate a Poisson distributed random variable x with mean m .

$$p(N_{i1}, N_{i2}, N_{i3} | \mu_i, \beta_i, \delta_i) = \text{Poi}(N_{i1} | \mu_i) \text{Poi}(N_{i2} | \mu_i\beta_i) \text{Poi}(N_{i3} | \mu_i\delta_i) \quad (3.26)$$

for $i = 1, \dots, n$. The sampling model includes different poison-slopes for each stage, in contrast to the model proposed by Ji et al. (2007). The parameter μ_i can be thought as the expected count mean of the pair i across the three stages if we were not enriching the tripeptide library at every stage. We extend (3.26) to the desired semi-parametric mixture model by assuming a non-parametric prior for a random effects distribution for $\theta_i \equiv (\beta_i, \delta_i)$. Let $G(b, d)$ denote a bivariate random probability

measure. We discuss the probability model for G below. A parametric random effects distribution for μ_i keeps computation simple.

$$(\beta_i, \delta_i | G) \sim G, \text{ i.i.d. , and } \mu_i \sim Ga(s_\mu, s_\mu \cdot t_\mu) \quad (3.27)$$

The parametrization of the Gamma distribution is chosen such that $E(X) = a/b$ for $X \sim Ga(a, b)$. Choosing a prior for the random probability measure G requires a nonparametric prior. The most commonly used model is the Dirichlet process (DP) prior (Ferguson, 1973; Antoniak, 1974). We write $G \sim DP(\alpha, G_0)$ for a DP prior on the random probability measure G . The DP prior is indexed with two parameters, a total mass parameter α and a base measure G_0 . The total mass parameter is a precision parameter, and the base measure defines the prior expectation, $E(G) = G_0$. See, for example MacEachern and Müller (1998), Walk et al. (1999), and Müller et al. (2004) for recent reviews of the DP prior, including posterior inference for DP mixtures similar to the model used here. We assume

$$G \sim DP(\alpha, G_0) \text{ with } G_0(b, d) = Ga(b | s_\beta, s_\beta t_\beta) \cdot Ga(d | s_\delta, s_\delta t_\delta). \quad (3.28)$$

The base measure in the DP prior are independent gamma distributions. The model is completed with a prior on the hyperparameters

$$\alpha \sim Ga(a_\alpha, b_\alpha), t_\beta \sim Ga(t_\beta | a_{t_\beta}, b_{t_\beta}), \quad (3.29)$$

$$t_\delta \sim Ga(t_\delta | a_{t_\delta}, b_{t_\delta}), \text{ and, } t_\mu \sim Ga(t_\mu | a_{t_\mu}, b_{t_\mu}), \quad (3.30)$$

3.4.2 Posterior Simulation

The model defined in (3.26)-(3.30) is a DPM. It includes a conjugate Poisson sample model and gamma baseline distribution of the DP, and conditionally conjugate prior specifications for other parameters. This greatly facilitates posterior simulation by MCMC simulation.

The implementation of posterior MCMC follows the standard posterior simulation method for DPM models given in, for example, Neal (2000) and MacEachern and Müller (1998). This method was described in the Section 3.2. In particular, we use the model augmentation with a latent Beta random variable proposed by West (1992) (step (d) of the Section 3.2), to implement inference for the total mass parameter α .

The remaining parameters μ_i, t_μ, t_β and t_δ have closed form conditional posterior distributions conditional on currently imputed values for all other parameters and latent variables. Let \mathbf{N} denote the data. When stating a complete conditional posterior distribution on this section, we make explicit only the relevant quantities (for example, since $t_\mu | \mathbf{N}, \mu_1, \dots, \mu_n, t_\beta, t_\delta$ does not depend on \mathbf{N}, t_β or t_δ we just write $t_\mu | \mu_1, \dots, \mu_n$).

We apply the Gibbs sampling scheme consisting of steps (a)-(e) in Section 3.2 to simulate from the posterior distribution of the model proposed model. We notice that the parameters $x_i, F_i(\cdot | \theta_i)$, and γ of the general DPM model (3.9) are equivalent to $N_i, (3.26), (t_\beta, t_\delta)$ in our specific proposed model. There is no common parameter σ , for all N_i 's and, then, step (c) is not necessary. The rest of the parameters in (3.9) and its correspondents in our specific application have the same names; an extra step (f) is included in the Gibbs sampling scheme to account for the parameters μ_i ; and, the following configuration – a set of variables $(K, (\beta^*, \delta^*), \varphi)$ equivalent to $(\beta_1, \delta_1), \dots, (\beta_n, \delta_n)$, see Section 3.2– is required:

- $(\beta^*, \delta^*) = (\beta_1^*, \delta_1^*), \dots, (\beta_K^*, \delta_K^*)$ represents the K unique values of $(\beta_1, \delta_1), \dots, (\beta_n, \delta_n)$.

- $\varphi = (\varphi_1, \dots, \varphi_n)$ is the vector of indicator variables indexing the value of $(\beta_1^*, \delta_1^*), \dots, (\beta_K^*, \delta_K^*)$ related with (β_i, δ_i) for $i = 1, \dots, n$. That is, $\varphi_i = k$, if $(\beta_i, \delta_i) = (\beta_k^*, \delta_k^*)$.

This configuration determines the following set of variables:

- $S_k := \{i : \varphi_i = k\}$ is the set of sample indices of (β_i, δ_i) related to (β_k^*, δ_k^*) for $k = 1, \dots, K$.
- $n_k := \#S_k$.

We will denote as K^{-i}, n_k^{-i} and S_k^{-i} for $k = 1, \dots, K^{-i}$, $(\beta^*, \delta^*)^{-i} = (\beta_1^*, \delta_1^*)^{-i}, \dots, (\beta_{K^{-i}}^*, \delta_{K^{-i}}^*)^{-i}$ to the configuration corresponding to $(\beta, \delta)^{-i} := (\beta_1, \delta_1), \dots, (\beta_{i-1}, \delta_{i-1}), (\beta_{i+1}, \delta_{i+1}), \dots, (\beta_n, \delta_n)$. We now describe the corresponding steps (a)-(e) of Section 3.2 and extra step (f) for the proposed model.

- (a) Given the current imputed values $(K, (\beta^*, \delta^*), \varphi)$, generate a new configuration by simulating $\varphi_1, \dots, \varphi_n$ from the complete conditional posterior distribution,

$$P(\varphi_i = k | N_i, (\beta, \delta)^{-i}, \varphi^{-i}, K^{-i}) = q_{ik}, \quad \text{for } k = 0, \dots, K^{-i}$$

where,

$$q_{i0} = c\alpha \left[\frac{(s_\beta t_\beta)^{s_\beta}}{(\mu_i + s_\beta t_\beta)^{s_\beta + N_{i2}}} \frac{\Gamma(s_\beta + N_{i2})}{\Gamma(s_\beta)} \right] \left[\frac{(s_\delta t_\delta)^{s_\delta}}{(\mu_i + s_\delta t_\delta)^{s_\beta + N_{i3}}} \frac{\Gamma(s_\delta + N_{i3})}{\Gamma(s_\delta)} \right],$$

$$q_{ik} = cn_k^{-i} \beta_k^* N_{i2} \delta_k^* N_{i3} \exp\{-\mu_i(\beta_k^* + \delta_k^*)\}, \quad \text{for } j = 1, \dots, K^{-i},$$

and c is a normalization constant such that $q_{i0} + \dots + q_{iK^{-i}} = 1$. Whenever we sample $\varphi_i = 0$, we generate a new observation (β_i, δ_i) from

$$Ga(\beta_i | s_\beta + N_{i2}, s_\beta t_\beta + \mu_i) \times Ga(\delta_i | s_\delta + N_{i2}, s_\delta t_\delta + \mu_i),$$

and update, accordingly, the new configuration by $K = K + 1$ and $\varphi_i = n + 1$.

(b) Given K and φ , generate a new set of parameters (β^*, δ^*) from the distributions

$$p(\beta_k^* | \mathbf{N}, \varphi, K) = Ga \left(\beta_k^* \middle| s_\beta + \sum_{i \in S_k} N_{i2}, s_\beta t_\beta + \sum_{i \in S_k} \mu_i \right), \quad \text{for } k = 1, \dots, K,$$

and

$$p(\delta_k^* | \mathbf{N}, \varphi, K) = Ga \left(\delta_k^* \middle| s_\delta + \sum_{i \in S_k} N_{i3}, s_\delta t_\delta + \sum_{i \in S_k} \mu_i \right), \quad \text{for } k = 1, \dots, K.$$

(d) Update the total mass parameter α :

(1) let α denote the currently imputed parameter value. Generate $\eta \sim Beta(\alpha + 1, n)$ and,

(2) sample the new value of α from

$$p(\alpha | K, \eta) = \pi_\eta Ga(\alpha | a_\alpha + K, b_\alpha - \log \eta) \\ + (1 - \pi_\eta) Ga(\alpha | a_\alpha + K - 1, b_\alpha - \log \eta),$$

where,

$$\frac{\pi_\eta}{1 - \pi_\eta} = \frac{a_\alpha + K - 1}{n(b_\alpha - \log \eta)}$$

(e) Update hyperparameters of G_0 . Simulate from the complete conditional posterior distributions,

$$p(t_\beta | K, \beta_1^*, \dots, \beta_K^*) = Ga \left(t_\beta \middle| a_{t_\beta} + K s_\beta, b_{t_\beta} + s_\beta \sum_{k=1}^K \beta_k^* \right). \\ p(t_\delta | K, \delta_1^*, \dots, \delta_K^*) = Ga \left(t_\delta \middle| a_{t_\delta} + K s_\delta, b_{t_\delta} + s_\delta \sum_{k=1}^K \delta_k^* \right).$$

(f) Finally, update μ_i and the hyperparameter t_μ :

$$p(\mu_i | \mathbf{N}, t_\mu, \beta_i, \delta_i) = Ga(\mu_i | N_{i1} + N_{i2} + N_{i3} + s_\mu, 1 + \beta_i + \delta_i + s_\mu t_\mu). \\ p(t_\mu | \mu_1, \dots, \mu_n) = Ga \left(t_\mu \middle| a_{t_\mu} + n s_\mu, b_{t_\mu} + s_\mu \sum_{i=1}^n \mu_i \right).$$

In order to avoid the chain to get “trapped” we reinitialize the configuration every 10,000 Gibbs sampling iterations with $K = n$ getting new values for $(\beta_1^*, \delta_1^*), \dots, (\beta_n^*, \delta_n^*)$ by resampling from (3.4.2) without changing the values of the remaining parameters.

3.5 Selecting Significant Tripeptide-Tissue Pairs

Posterior MCMC allows us to carry out essentially all posterior inference of interest. In particular, let \mathbf{N} denote the observed data and $p_i = Pr(\delta_i > \beta_i > 1 \mid \mathbf{N})$ denote the posterior probability for increasing mean counts for peptide-tissue pair i . The posterior Monte Carlo sample allows easy evaluation of p_i as empirical average of $I(\delta_i > \beta_i > 1)$ over all imputed values (δ_i, β_i) in the Monte Carlo posterior sample. Let $d_i \in \{0, 1\}$ denote an indicator for reporting significant affinity for the peptide-tissue pair i , i.e., increasing mean counts. A reasonable decision rule is to report all pairs with marginal posterior probability beyond a threshold:

$$d_i^* = I(p_i > t). \quad (3.31)$$

The rule d^* can be justified in terms of the False Discovery Rate (FDR) concept (Newton, 2004) or, alternatively, as an optimal Bayes rule. To define an optimal rule we need to augment the probability model to a decision problem by introducing a utility function. Let θ and y generically denote all unknown parameters and all observable data. A utility function $u(d, \theta, y)$ formalizes relative preferences for decision d under hypothetical outcomes y and under an assumed truth θ . For example, a utility function could be

$$u(d, \theta, y) = \sum_i d_i I(\delta_i > \beta_i > 1) + k \sum_i (1 - d_i)(1 - I(\delta_i > \beta_i > 1)), \quad (3.32)$$

i.e., a linear combination of the number of true positive selections d_i and true negatives. For a given probability model, data and utility function, the optimal Bayes rule is defined as the rule that maximizes u in expectation over all not observed variable, and conditional on all observed variables. In our case,

$$d^B = \arg \max_{\delta} E(u(d, \theta, y) \mid y).$$

It can be shown that d^* arises as Bayes rule under several utility functions that trade off false positive and false negative counts.

A shortcoming of the rule d^* is that it implicitly weights all true positives and true negatives equally (and then $t = 1/2$). But not all true positives are equally desirable to the investigators in a phage display experiment.

Besides, the FDR criterion is a Bayes rule of the form of (3.31) when considering a particular utility function of the form (3.32). Under this criterion, the selected pairs conform the largest list such that the expected FDR (proportion of false positives in this list) is bounded by a quantity $0 < \alpha < 1$. The same list is obtained when choosing k in (3.32) such that the set of p_i 's greater than $t := k/(k+1)$ and $I_k := \{i : p_i > t\}$, satisfies:

$$S := \sum_{i \in I_k} \alpha - (1 - p_i) \geq 0 \quad \text{and} \quad S + \alpha - (1 - p_j) < 0, \quad \text{for any } j \notin I_k. \quad (3.33)$$

Ji et al. (2007) selected the pairs with increasing means across the three stages according to the FDR-based criterion (3.31) where p_i is the probability of (common) positive slope (across the stages 1 and 2 and 2 and 3) according to their model. It can be seen from the form of (3.32) that this criterion does not take into account the size of the increase.

In contrast, we will use a utility function that gives weights to the pairs proportional to the relative increment from the first to the third stages (i.e. δ_i) if the means of the three phases increase, i.e., if $\delta_i > \beta_i > 1$.

We choose the utility function of Müller et al. (2006) given by

$$U(d, w) = \sum_{i=1}^n d_i w_i - k \sum_{i=1}^n (1 - d_i) w_i - cD, \quad (3.34)$$

where $w = (w_1, \dots, w_n)$ is a vector of weights, in our application we consider

$$w_i = \delta_i I\{\delta_i > \beta_i > 1\};$$

D is the number of pairs that we declare that have increasing means across the three stages of the experiment, i.e., $D = \sum_{i=1}^n d_i$; $c > 0$ represents the cost of declaring that a pair has increasing means; and $k > 0$ is such that $k w_i$ is the cost of not declaring that the pair i has increasing means when indeed it really does.

Let $\bar{m}_i = E(w_i \mid \mathbf{N})$, when maximizing the utility function, straightforward algebra shows that the optimal rule is

$$d_i^B = I(\bar{m}_i \geq c/(k+1)). \quad (3.35)$$

Once we have a simulated sample $(\beta_i^1, \delta_i^1), \dots, (\beta_i^M, \delta_i^M)$ of (β_i, δ_i) given the data. We can estimate \bar{m}_i with

$$\bar{m}_i \approx \frac{1}{M} \sum_{k=1}^M \delta_i^k I\{\delta_i^k > \beta_i^k > 1\}, \quad (3.36)$$

and then make the decision d_i^B .

3.6 A Simulation Study

We carried out a simulation study to validate the proposed approach. We generated $n = 2000$ observations of the model described in (3.26) through (3.30), except that the random probability measure G is replaced by a Gamma distribution with fixed parameters

$$\mu_k \sim Ga(\mu | s_\mu^f, s_\mu^f t_\mu^f), \text{ i.i.d. and } (\beta_k, \delta_k) \sim Ga(\beta | s_\beta^f, s_\beta^f t_\beta^f) Ga(\delta | s_\delta^f, s_\delta^f t_\delta^f) \quad (3.37)$$

We set the hyper-parameters such that the expected value of μ_i and its variance are small and, besides, β_i and δ_i have both mean 8 and variances 30 and 120 respectively. The idea behind is that μ_i is interpreted as the mean of the counts through the three stages of the pair i if there were no enrichment. Since, initially, the library contains a small amount of the particular tripeptide related with the pair i among the large number of different tripeptides, we expect μ_i to be small. The parameters β_i and δ_i represent the folds of the expected count values from the first stage to the second and third stages respectively due to the library enrichment. We allow these last parameters to have large variances. The Gamma parameters were set to $s_\mu^f = 3.6, t_\mu^f = 5/6, s_\beta^f = 13/6, t_\beta^f = 1/8, s_\delta^f = 0.53$ and $t_\delta^f = 0.125$.

The hyper-parameters of the model described in (3.26) through (3.30) were chosen taking into account the same considerations and set to $s_\mu = 1, a_{t_\mu} = 3, b_{t_\mu} = 2, s_\beta = s_\delta = 1, a_{t_\beta} = a_{t_\delta} = 2.5, b_{t_\beta} = b_{t_\delta} = 9, a_\alpha = b_\alpha = 1$.

Saving every 10^{th} iteration after a 10,000 iteration burn-in, a Monte Carlo posterior sample of size $M = 5,000$ was saved. We performed convergence assessment for the proposed parameter values to ensure that the MCMC algorithm converges well. On the basis of the convergence criteria in Cowless and Taylor (1996), we found that the Markov chains mixed very well and converged rapidly. In 723 out of the 2,000 simulated cases it turned out that $\delta_i > \beta_i > 1$.

Using the FDR criterion described by equations (3.31) and (3.33), we selected the pairs such that, under the assumptions of our model, the expected false discovery

rate was 0.2. This implies that the expected False Negative Rate (FNR, proportion of false negatives, relative to the total number of unselected pairs) was 0.117. We declared that 715 pairs had increasing means across the three stages. Of them 578 really did. The observed FDR and FNR were 0.192 and 0.113 respectively.

Selecting the pairs according to the utility function (3.34), with $c/(k+1) = 8$, we declared that 722 pairs had increasing means. Of them, 575 actually exhibited this pattern. The expected values of the FDR and FNR were 0.206 and 0.117 respectively. The observed values of these quantities were 0.204 and 0.116 respectively. The number of pairs chosen by both methods was 688.

Our model is a particular case of a Dirichlet Process Mixture model. MacEachern and Müller (1998) mention that if the data follows a distribution according to a Dirichlet Process Mixture model, the predictive final distribution of a future observation matches with the posterior expected value of the density that generated the data. Thus, the true distribution of the data can be estimated from a (simulated) sample of future observations. This is, averages over the values of the expression (3.17) evaluated at the values of the posterior simulated configurations $((\beta^*, \delta^*, \varphi, k))$. We employed this method to simulate a sample of a future observations of β and δ . In Figure 3.4, we compare the histograms of this simulated sample with the true distribution.

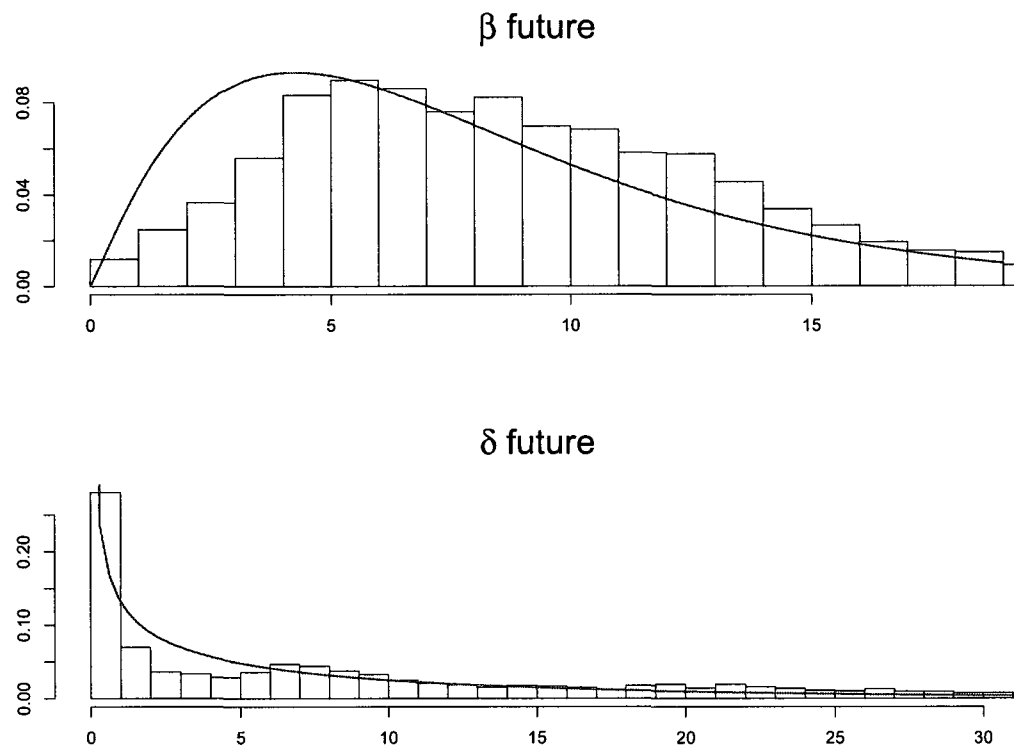


Figure 3.4: Histograms of a simulated sample of future observations of β and δ . They estimate the true probability density function (continuous line) of the distribution that generated the data

3.7 Results

In this section we present analysis and results by applying the proposed method to the phage display data described in section 3.3.

The parameter values in our proposed priors are elicited by consulting with the investigators. The values of the hyper-parameters are the same as the ones used in the simulation study. The parameter μ_i is interpreted as the expected counts if there were no enrichment of the library of tripeptides at every stage and this is the case for the first stage. We assume that most of the phage counts are small in the initial state. Therefore, we set the expected value for the first stage counts μ_i to 0.1 and its variance to 0.03. We do not assume any knowledge of the means increment between the first and the second stage (in terms of β_i) and between the first and the third stage (in terms of δ_i). We center these values around 6 allowing for a large variance equal to 180.

We obtained a Monte Carlo posterior sample of size $M = 5,000$ storing the values of the imputed parameters every ten iterations after a burn-in of 10,000 iterations. Analogous to the simulation study, we performed convergence assessment for the proposed parameter values to ensure that the MCMC algorithm converged well. We found that the Markov chains mixed very well and converges rapidly.

Table 3.1 shows the 30 pairs with highest values of \bar{m}_i , i.e the pairs chosen according to the optimal rule (3.35) with a threshold value of seven. Figure 3.5 depicts these pairs. We notice that there are some pairs, such as, the tripeptide ARF in the tissue fat, that present a small posterior probability of increasing means, p_i , that would not be selected according to the decision rule (3.31) but that are selected when using the decision rule (3.35).

Using the optimal rule (3.35), with a threshold value of one, 219 tripeptide-tissue pairs are selected. Figure 3.6 highlights these pairs. This criterion selects the pairs clearly having high increasing counts across the three stages. Over some pairs with nondecreasing counts, in agreement with its related utility function (3.34), the cri-

terion chooses pairs with an oscillating count pattern but with a substantially large increment from the first to the second stages or from the second to the third stages. Besides, it picks pairs that have small counts over the three stages but with a large increment in the second or/and third stage count in comparison to the previous count. On the contrary, there are some other pairs that despite having relatively large and nondecreasing counts over the three stages are not selected; the selecting method is detecting that this event happens because the corresponding tripeptide has a large base-line count to start with and not necessarily a strongly binding behavior to the respective tissue. The expected values of the FDR and FNR are 0.534 and 0.089 respectively. Due to space limitation these pairs are not listed in this manuscript.

In Figure 3.7, we plotted the histograms of the simulated samples of the posterior distributions of β_i and δ_i for three specific pairs. We notice that the bimodal behavior of the first two pairs considered in this plot cannot be detected by the parametric version of our model. The third pair is an example in which there is no statistical evidence of an increasing pattern (since $\beta_i < 1$). Nevertheless, the decision rule (3.35) tends to pick this pair due to its large count observed at the third stage.

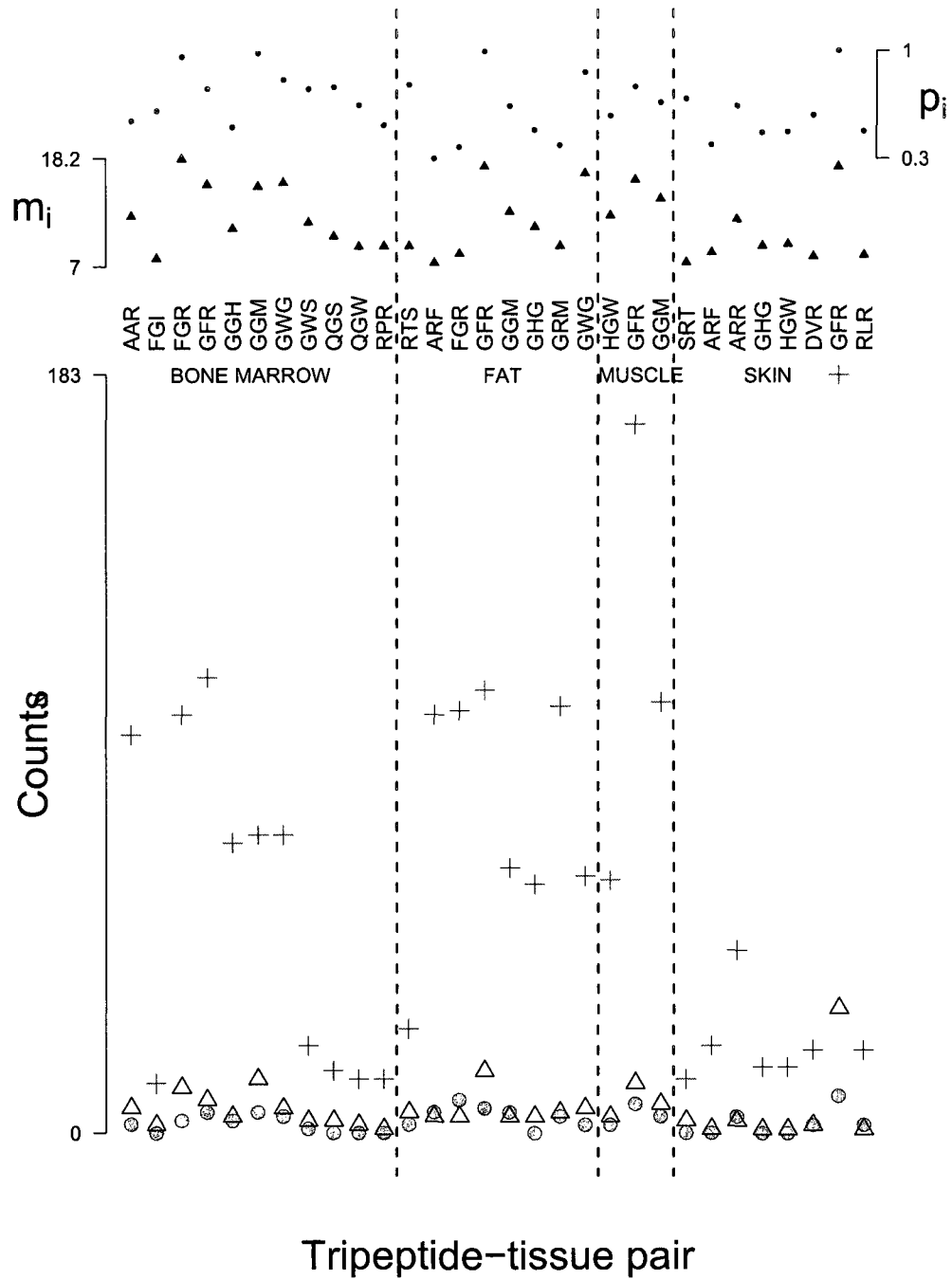


Figure 3.5: Thirty pairs with highest estimated values of \bar{m}_i in (3.36). Lower section presents counts for the three stages: circle, triangle and cross. The middle section shows the value of \bar{m}_i . The upper part depicts the posterior probability of increasing means across the three stages, p_i .

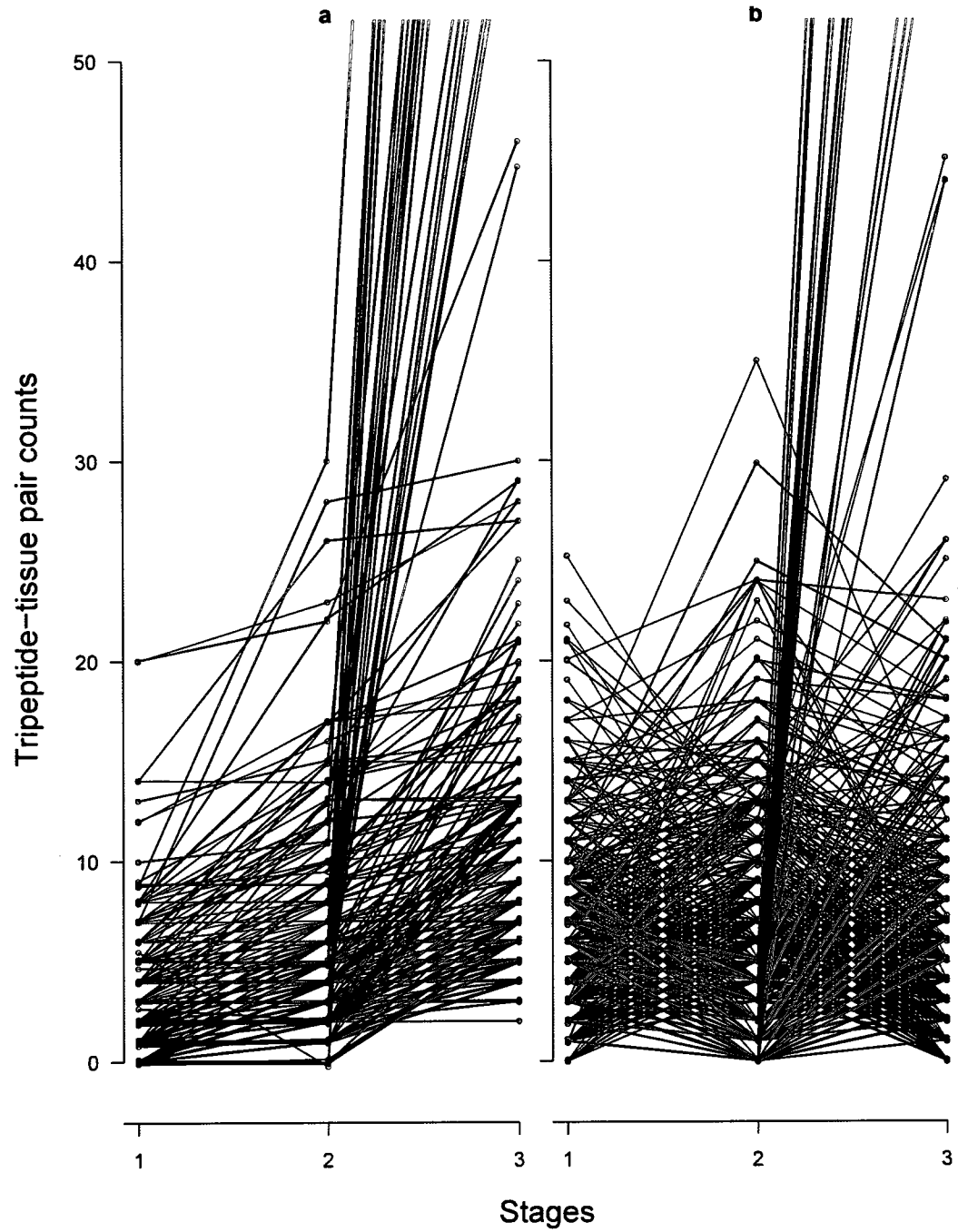


Figure 3.6: Respectively, (a) and (b) nondecreasing and oscillating observed tripeptide-tissue pair counts across the three stages. In red the 219 selected pairs using the optimal rule (3.35) with a threshold value of 1.

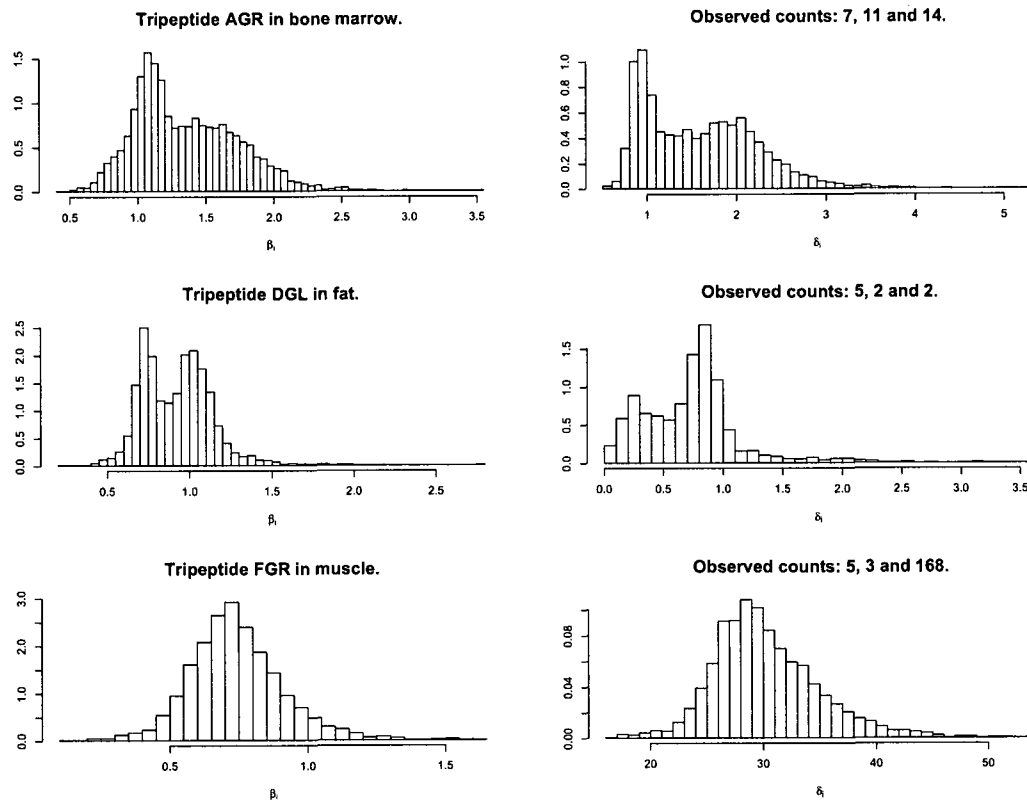


Figure 3.7: Histograms of simulated samples of the final distribution of the parameters β_i and δ_i for three specific pairs.

Table 3.1: Considered tripeptide-tissue pairs by using the rule (3.35) with a threshold value of 7. The expected values of FDR and FNR are 0.337 and 0.113 respectively.

Pairs considered: 30.

Tissue	Tripeptide	Stage 1	Stage 2	Stage 3	\bar{m}_i	p_i
BM	AAR	2	6	96	12.19	0.56
BM	FGI	0	2	12	7.90	0.63
BM	FGR	3	11	101	18.17	0.96
BM	GFR	5	8	110	15.51	0.76
BM	GGH	3	4	70	10.97	0.53
BM	GGM	5	13	72	15.32	0.98
BM	GWG	4	6	72	15.69	0.82
BM	GWS	1	3	21	11.59	0.76
BM	QGS	0	3	15	10.15	0.77
BM	QGW	0	2	13	9.10	0.66
BM	RPR	0	1	13	9.14	0.54
BM	RTS	2	5	25	9.14	0.79
F	ARF	5	4	101	7.41	0.33
F	FGR	8	4	102	8.39	0.41
F	GFR	6	15	107	17.36	0.99
F	GGM	5	4	64	12.70	0.66
F	GHG	0	4	60	11.13	0.51
F	GRM	4	5	103	9.18	0.42
F	GWG	2	6	62	16.66	0.87
F	HGW	2	4	61	12.31	0.60
MU	GFR	7	12	171	15.97	0.78
MU	GGM	4	7	104	14.02	0.68
MU	SRT	0	3	13	7.46	0.70
PR	ARF	0	1	21	8.50	0.42
PR	ARR	4	3	44	11.92	0.66
PR	GHG	0	1	16	9.20	0.49
PR	HGW	0	1	16	9.38	0.50
SK	DVR	2	2	20	8.10	0.60
SK	GFR	9	30	183	17.34	1.00
SK	RLR	2	1	20	8.24	0.50

3.8 Discussion

We have proposed semi-parametric model-based statistical inference for high dimension count data arising from phage experiments with parallel biopanning. The probability model is extended to a decision problem by adding a utility function for the choice of reported peptide/tissue pairs.

Previously, Ji et al. (2007) introduced a model for the analysis of phage experiment data based on mouse data. Analogous to their model, ours takes into account the correlation that exists between the different stages and detects the tripeptides that tend to bind with a specific tissue. Since there is just one observation across the three stages for every tripeptide-tissue pair, we need to impose a hierarchical structure that allows us to borrow information of all the pairs in order to make statistical inferences of the behavior of one in particular. A visual inspection of the human data, such as the one presented on Figure 3.3, shows the existence of pairs with oscillating counts and the presence of outliers. This indicates that the log-linearity of the means assumption on the model proposed by Ji et al. may not be appropriate in this case. In order to avoid an inference that can be misled by outliers, we require a more robust model against them. These two phenomena are taken into account by our model, the first by the specific structure of it and, the second, taking advantage of our larger data set, by its nonparametric nature. In addition, the model we proposed has an easy interpretation of the parameters at the upper level of the hierarchical model: μ_i is the mean counts of the tripeptide-tissue pair i if there were no enrichment of the tripeptide library at every stage, while β_i and δ_i are the folds of these counts at the second and third stages respectively due to this enrichment. Moreover, this parametrization allows an easy description of the phenomenon we are interested in- the increasing mean counts through the three stages- in terms of the parameters. If the biologists are interested in studying other phenomena involving the means, once we have simulated the posterior sample of our parameters, it is easy to compute the posterior probability of the events equivalent to these phenomena.

The proposed model includes random baseline means μ_i to allow for variation of the amount of tripeptides in the original phage display peptide library. It is important to include such random effects. Large counts at the third stage can be explained by either preferential binding of a certain tripeptide, or simply random high initial count in the original library. The inclusion of the random baseline means allows us to differentiate these two cases.

In selecting the tripeptide-tissue pairs with increasing means across the three stages, we face a massive multiplicity problem. Ji et al. approach this problem through the application of a simple Bayesian FDR, providing the investigator with an estimate of the error rate in the selection procedure. Their approach can be embedded in a decision theoretic framework where the utility function takes into consideration statistical significance, that is, the proportion of false positives and false negatives; but it does not take into account the biological significance such as the size of the mean increase through the three stages. Our decision theoretic approach specifically targets this issue through the definition of its corresponding utility function, and the estimation of the error rate in the selection procedure, such as the FDR and false negative rate, can also be computed using this methodology. Other utility functions can be considered.

The nonparametric nature of our model is an improvement in sophistication over its parametric version. Nevertheless, the MCMC simulation scheme of the posterior sample is straightforward. The posterior distribution of any parameter given the rest can be derived directly and has an analytical expression due to the use of conjugate families. The computing time necessary to generate the MCMC sample is relatively short.

The simulation study we performed was satisfactory. If the data is generated according to the parametric version of our model, our nonparametric model is able to detect the event of interest with high accuracy; the observed FDR and false negatives rate are close to their respective expected values.

When selecting 219 pairs, our model and selecting criterion were able to choose the pairs with high increasing counts as well as the ones with oscillating and/or small counts over the three stages but exhibiting large counts on the second and/or the third stages in relation to the previous stage count. These pairs are candidates for a deeper study of their binding behavior.

In consulting with the biology investigators, they would expect a small count if there were no tripeptide enrichment of the library from the previous stage. In agreement with its interpretation given above, we initially assume that the parameter μ_i is small by choosing a prior distribution with small mean and variance. It is necessary to assume a prior small variance since we have just one three-count observation for every tripeptide-tissue pair. The other two parameters β_i and δ_i can be assumed to have diffuse priors.

Our model is not considering that different tissues can have different binding behaviors. For example, there may be a tissue that absorbs more tripeptides, or present different count variances. Although this problem is ameliorated considering the baseline mean count parameter μ_i , the model could be improved by priori incorporating a correlation structure of the pairs sharing the same tissue.

Chapter 4

Borrowing Strength with Hierarchical Models over Non-Exchangeable Subpopulations

4.1 Outline

We address statistical inference in a phase II trial of sarcoma. Sarcoma is a rare cancer affecting connective and soft tissues (e.g., cartilage and fat). Sarcoma is a very heterogeneous disease with many different subtypes with widely different prognosis. In the proposed design we pay particular attention to the heterogeneous nature of the disease and the fact that different disease subtypes are related but can not be considered *a priori* exchangeable. We classify different subtypes by the overall prognosis as poor, intermediate or good. The objective of the study is to assess the efficacy of a new drug in patients with different subtypes of sarcoma. Let p_i , $i = 1, \dots, n$, denote the probability of response (defined below) for a patient with disease subtype i . One possible approach is to analyze the different subtypes as separate studies. But due to the rare nature of some of the subtypes the enrollment in $n = 12$ separate studies would be way too slow. This leads us to consider borrowing of strength

across related subtypes. This could be done with a hierarchical model consisting of a submodel for each disease, plus a prior distribution for p_i that assumes that all p_i arise from some underlying common distribution $p(p_i | \eta)$. The p_i are interpreted as subtype-specific effects, and η could characterize an overall success probability. For example, one could assume $\text{logit } p_i \sim N(\mu, \tau)$ and complete the model with a prior $p(\mu, \tau)$. A major limitation of this approach is that it assumes that disease subtypes are a priori exchangeable. Formally the prior model $\text{logit } p_i \sim N(\mu, \tau)$ is invariant with respect to arbitrary permutations of the indices $i = 1, \dots, n$. This is not appropriate since different subtypes are not exchangeable. Disease subtypes with poor prognosis are known to be different from those with good prognosis. For data analysis conditional on a large enough data set this might be no problem, as the likelihood would asymptotically dominate the prior. But for design, when inference is initially based on very small sample sizes, such details of the prior model can matter. An easy fix is to replace the exchangeable model with a partially exchangeable model using a regression. Let $x_i \in \{-1, 0, 1\}$ denote an indicator for the subtype prognosis. We could assume $\text{logit } p_i \sim N(\mu_i, \tau)$ with $\mu_i = \beta_0 + \beta_1 x_i$. The problem with this approach is that disease subtypes are grouped by overall prognosis and this grouping is frozen. However, while overall prognosis for a subtype is important, it is not obvious that it determines the most appropriate grouping. One of the eligibility criteria of the study is failure of prior therapy, i.e., in a sense all patients enter the study with a poor prognosis.

We propose a novel approach that can be characterized as intermediate between an exchangeable hierarchical model and a regression. We treat the appropriate grouping as a random quantity, ρ , and define a probability model for this random partition ρ . The model is indexed with the covariates x_i . Thus the model includes a priori a preference for grouping by x_i , but allows for alternative grouping as the data dictates. In other words, we propose a semi-parametric model that respects the non-exchangeable nature present in the data.

The rest of this chapter is organized as follows. Section 4.2 describes the motivating phase II sarcoma dataset. The non-exchangeable model is formalized and discussed in detail in Section 4.3. Section 4.4 reports a comparison of the proposed model versus several alternatives. In Section 4.5 the model is applied to the sarcoma data. Section 4.6 concludes with a discussion.

4.2 Data

A single arm Phase II clinical trial for sarcoma is carried out at M.D. Anderson Cancer Center. The objective of the study is to assess the efficacy of Irinotecan on patients with different sarcoma subtypes. Irinotecan is a water soluble and commercially available chemical agent that is thought to reduce the sarcoma cancerigenic tumors. Its Dose-Limiting Toxicity (DLT) has been measured through a Phase I clinical trial (Masuda et al, 2000). The DLT is defined as diarrhea and bone marrow suppression. The patients in this trial received the drug in cycles of 21 days, where the drug was administered during the first two weeks and nothing during the third one. A week with drug treatment consisted of $16 \text{ mg}/\text{m}^2$ of daily injections of Irinotecan for five consecutive days and two days of resting.

Treatment efficacy is measured as tumor shrinkage. More specifically, tumor sizes at the end of the second cycle and, if necessary, at the end of the fourth cycle were compared with the size at the beginning of the study. If complete (total disappearance of tumor) or partial responses (at least 30% shrinkage) is observed at the end of the second cycle, the treatment is considered a success for this patient. In contrast, when reporting progressive disease (20% or more increase), the trial is declared a failure. If none of the above criteria is met, the disease is considered stabilized. In this case, the tumor is measured again after the fourth cycle, and the treatment is declared a failure only when progressive disease is reported.

This study is still ongoing. So far, a total of 179 patients have met the eligibility

criteria and have been recruited in the protocol. Of them, 164 participants exhibited one out of eight sarcoma subtypes related with an intermediate prognosis and only 15 exhibited one out of two subtypes of the disease related to good prognosis. No patient with a sarcoma subtype related to poor prognosis has been reported yet. Table 4.1 shows the available data.

Table 4.1: Reported number of successes/trials for each one of the sarcoma subtypes.

Intermediate Prognosis	
subtype	successes/trials
Leiomyosarcoma	6/28
Liposarcoma	7/29
MFH	3/29
Osteosarcoma	5/26
Synovial	3/20
Angiosarcoma	2/15
MPNST	1/5
Fibrosarcoma	1/12
Good Prognosis	
Ewing's	0/13
Rhabdo	0/2

4.3 Non-Exchangeable Product Partition Model

4.3.1 Model Definition

Let y_i and N_i respectively denote the number of successes (i.e., patients with positive outcome) and the total number of patients presenting with sarcoma subtype i , $i =$

$1, \dots, n$, $n = 12$. Denote also by $\mathbf{p} = (p_1, \dots, p_n)^t$ the vector of success rates. Let $\rho = (S_1, \dots, S_K)$ denote a partition of $\{1, \dots, n\}$ into K clusters S_k , $k = 1, \dots, K$, i.e., $\{1, \dots, n\} = \cup_{k=1}^K S_k$ and $S_k \cap S_{k'} = \emptyset$ for $k \neq k'$. Let $x^n = \{x_1, \dots, x_n\}$ the set of values of the ordinal covariates. We assume that all disease subtypes in the same cluster have similar success rates. More specifically, the logits of the success rate with indices in the cluster k are drawn from the same normal distribution with mean θ_k^* . The partition ρ is equivalently characterized by cluster membership indicators φ_i , for $i = 1, \dots, n$ with $\varphi_i = k$ if $i \in S_k$. We use ρ or (φ, K) interchangeably.

We assume the following model (NEPPM):

$$\begin{aligned} y_i | p_i &\sim \text{Bin}(p_i, N_i) \\ \text{logit}(p_i) &\equiv \theta_i \stackrel{\text{iid}}{\sim} N(\theta_k^*, \tau_p), \quad \text{for } i \in S_k \\ \theta_k^* &\stackrel{\text{iid}}{\sim} N(0, \tau_\theta), \text{ for } k = 1, \dots, K \quad \text{and} \quad \rho \sim p(\rho | x^n), \end{aligned} \quad (4.1)$$

The definition of $p(\rho | x^n)$ will be discussed below. Here, $N(m, s)$ denotes the normal distribution with mean m and precision s . We fix $\tau_p = 18$ such that the ratio of θ_{i_1} and θ_{i_2} for any two i_1, i_2 in the same cluster is between $1/2.5$ and 2.5 with probability 0.95 . We fix $\tau_\theta = 1/4$. This value allows for sufficiently spread out values of θ_k^* , and thus p_i s. The model defines a compromise between an exchangeable hierarchical model, separate models and a partially exchangeable hierarchical regression model.

To define the model $p(\rho)$ we resort to the product partition models (PPMs) (Hartigan, 1990; Barry and Hartigan, 1993)) The idea is to construct a probability distribution $p(\rho_n)$ on the space of partitions of $\{1, \dots, n\}$, by introducing a cohesion function $c(A) \geq 0$ for every $A \subset \{1, \dots, n\}$, measuring how tightly grouped the elements in A are thought to be. A product partition model is defined as

$$p(\rho_n = (S_1, \dots, S_k)) = \frac{1}{g_n} \prod_{k=1}^K c(S_k), \quad p(y^n | \rho_n) = \prod_{k=1}^K p(y_i : i \in S_k), \quad (4.2)$$

Model (4.2) is easily seen to be conjugate. A remarkable connection between PPMs and the Dirichlet process (DP) (Ferguson, 1973) is pointed out, for example, in Quintana and Iglesias (2003) and Dahl (2003)). The DP is discrete with probability 1 and

the distribution on partitions induced by ties corresponds to a PPM with cohesions $c(S_k) = \alpha \times (\#S_k - 1)!$. Here $\#A$ denotes the cardinality of a set A and α is the total mass parameter of the DP prior.

A feature of the PPM induced by the DP is that it is *a priori* exchangeable with respect to the experimental units, which makes it inadequate for our application.

We now define a non-exchangeable PPM (NEPPM). In particular, applied to the clustering of disease sub-types $\{1, \dots, n\}$ we define a model that increases the prior probability of any two subtypes i and i' with equal prognoses $x_i = x_{i'}$ to cluster together.

In general, we define a probability model for random partitions of experimental units $\{1, \dots, n\}$ with categorical covariates $x_i \in \{1, \dots, Q\}$ such that clusters with homogeneous covariates are encouraged *a priori*.

We define

$$Pr(\rho_n = (S_1, \dots, S_K)) \propto \prod_{k=1}^K c(S_k), \quad \text{where } c(S_k) = c_D(S_k)d(S_k), \quad (4.3)$$

$c_D(S_k) = \alpha(\#S_k - 1)!$ is the cohesion induced by the DP and

$$d(S_k) = \left(\frac{\prod_{q=1}^Q m_{kq}!}{(\#S_k + Q - 1)!} \right)^\gamma. \quad (4.4)$$

Here Q is the number of different categories, m_{kq} the number of experimental units of category q in the cluster S_k and γ is a nonnegative constant, common to all cohesions, that gives strength to the cohesion of homogeneous clusters. We refer to $d(S_k)$ as a “similarity function.” It serves the purpose of increasing the probability of forming clusters with more homogeneous covariate values x_i , $i \in S_k$. The higher the value of γ , the stronger the prior emphasis on homogeneous clusters. More homogeneous clusters S_k have larger $d(S_k)$. In our specific application, we have $Q = 3$ prognoses, and m_{kq} for $q = -1, 0, 1$, are, respectively, the numbers of sarcoma subtypes with poor, intermediate and good prognosis in the cluster S_k . As desired, the resulting prior probability model is non-exchangeable.

Let $f_k(\rho_n)$ denote the predictive probability (PP) function, i.e., the conditional probability of a hypothetical new $(n+1)$ -st unit being allocated to cluster k , conditional on ρ_n . Let $K(n)$ denote the number of clusters in ρ_n . We find

$$f_k(\rho_n) = \frac{c(S_k \cup \{n+1\})}{c(S_k)} \propto \begin{cases} \#S_k \left(\frac{m_{k\ell}+1}{\#S_k+Q} \right)^\gamma & \text{for } 1 \leq k \leq K(n), \\ \alpha/Q^\gamma & \text{for } k = K(n) + 1. \end{cases} \quad (4.5)$$

where $x_{n+1} = \ell$ is the category of the new experimental unit and $c(\emptyset) := 1$. In the context of our application x_i is the prognosis for subtype i . Posterior simulation follows a simple modification of standard Gibbs-sampling schemes for DP or PPM models, as described in, e.g., MacEachern and Müller (1998) or Quintana (2006). See further details in the Appendix.

4.3.2 Some Properties of the NEPPM

The proposed model reduces to the DP Pólya urn when all x_i are equal, i.e., $Q = 1$. As a consequence the NEPPM reduces to a DPM. That is, if $Q = 1$, then, $m_{k1} = \#S_k$ and $d(S_k)$ reduces to

$$d(S_k) \propto \frac{\Gamma(\#S_k + \beta)}{\Gamma(\#S_k + \beta)} = 1. \quad (4.6)$$

Similarly, when $\gamma = 0$ the similarity function drops out of the model and the NEPPM reduces to the DPM. On the other hand, the model can easily be extended for more complicated covariates x_i . For multiple categorical covariates one could introduce several similarity functions and use the product to modify the cohesion functions in (4.4).

For $\gamma = 1$ the model reduces to a special case of the model introduced in Müller et al. (2009), using the default similarity function for categorical covariates. In general, they suggest to use a similarity function defined with an auxiliary probability model as $d(S_k) \equiv g(x_k^*) =$

$$g(x_k^*) = \int \prod_{i \in S_k} q(x_i | \xi_k) q(\xi_k) d\xi_k,$$

where x_k^* is the set of the values of the covariate in the cluster k , ξ_k is a latent variable, and $q(\cdot | \xi_k)$ and $q(\cdot)$ are auxiliary probability models. For categorical covariates $x_i \in \{1, \dots, Q\}$ they introduce a vector $\xi_k = (\xi_{k1}, \dots, \xi_{kQ})$ and use $q(x_i = q | \xi_{kq}) = \xi_{kq}$ and $q(\xi)$ as a Dirichlet distribution with parameters β_1, \dots, β_Q . We get,

$$g(x_k^*) = \frac{\Gamma(\sum_q \beta_q) \prod_q \Gamma(\beta_q + m_{kq})}{\prod_q \Gamma(\beta_q) \Gamma(\#S_k + \sum_q \beta_q)}. \quad (4.7)$$

Assuming that the parameters of this Dirichlet distribution are all equal to, say, β , the expression above reduces to:

$$g(x_k^*) = \frac{\Gamma(Q\beta) \prod_q \Gamma(\beta + m_{kq})}{\Gamma(\beta)^Q \Gamma(\#S_k + Q\beta)} \quad (4.8)$$

and, when $\beta = 1$,

$$g(x_k^*) \propto \frac{\prod_q \Gamma(m_{kq} + 1)}{\Gamma(\#S_k + Q)} = \frac{\prod_q m_{kq}!}{(\#S_k + Q - 1)!}$$

This is exactly the coefficient $d(S_k)$ in (4.4) of the definition of our NEPPM with $\gamma = 1$.

The proposed model $p(\rho_n | x^n)$ defines a sequence of probability models across sample sizes n . The question arises whether the model is coherent across n . Ideally the model for n should arise from marginalization of the model under $n + 1$. Below we show that, in general, this is not true. Let $\rho_n = (S_1, \dots, S_K)$ denote a partition of the set $\{1, \dots, n\}$. Recall the model, using the notation in (4.8)

$$p(\rho_n = (S_1, \dots, S_K) | x^n) = \frac{1}{g_n(x^n)} \prod_{k=1}^K c_D(S_k) g(x_k^*), \quad (4.9)$$

The normalizing constant, $g_n(x^n)$, is equal to $\sum_{\rho_n} \prod_{k=1}^K c_D(S_k) g(x_k^*)$.

Let $(\rho_n, \varphi_{n+1} = \ell)$ denote the partition of $\{1, \dots, n + 1\}$ after adding the index $n + 1$ to the ℓ -th cluster of ρ_n , for $\ell = 1, \dots, K + 1$. Then,

$$\begin{aligned} p(\varphi_{n+1} = \ell | \rho_n, x^n, x_{n+1}) &= \frac{Pr(\rho^n, \varphi_{n+1} = \ell | x^n, x_{n+1} = q)}{Pr(\rho^n | x^n)} \\ &= \frac{g_n(x^n)}{g_{n+1}(x^{n+1})} \frac{c(S_{\ell+})}{c(S_{\ell})} \times \begin{cases} \frac{m_{\ell q} + \beta}{\#S_{\ell} + \beta Q}, & 1 \leq \ell \leq K \\ \alpha/Q, & \ell = K + 1. \end{cases} \end{aligned}$$

Notice that in general,

$$Pr(\rho^n | x^n) \neq \sum_{k=1}^{K(n)+1} Pr(\rho^n, \varphi_{n+1} = l | x^n, x_{n+1} = q) f_k(\rho_n)$$

where f_k was defined in (4.5). That is, $Pr(\rho_n | x^n)$ is not obtained when marginalizing $Pr(\rho^n, \varphi_{n+1} | x^n, x_{n+1} = q)$ with respect to φ_{n+1} .

4.4 Comparison and Operating Characteristics

In this section we compare via simulation the performance of the proposed model vs. different alternative models. The comparison is in terms of bias, mean square error and coverage probability (CP). Later, in Subsection 4.4.2 we continue the comparison for the best two models by focusing on performance summaries that are relevant for the clinical trial design. Specifically, we will consider the probability (under repeat experimentation) of correctly identifying disease subtypes for which the treatment is not effective. Early identification of such disease subtypes is important to avoid exposure of patients to unnecessary risks. In the implementation of the proposed NEPPM model we assume that the parameter α in the definition of c_D in (4.5) is $Ga(5,0.5)$ distributed. Here $Ga(a,b)$ denotes the gamma distribution with mean a/b .

4.4.1 Competing Models

We compare the proposed NEPPM (4.1) with the following alternative models. The first model entirely abandons borrowing strength across subtypes. The second model borrows strength, but assumes a priori exchangeable subtypes. The third model respects the lack of exchangeability across sub-types, but goes to the extreme of grouping the subtypes by the covariate, in our case prognosis, and fixing this grouping for the rest of the analysis.

Separate Models: Assume a separate model for each disease subtype. There is no borrowing of strength or pooling of information across subtypes. That is, for $i = 1, \dots, n$,

$$\begin{aligned} y_i | p_i &\sim \text{Bin}(p_i, N_i) \\ \text{logit}(p_i) &\sim N(\mu_i, \tau_i) \\ \mu_i &\sim N(0, 1/4) \text{ and } \tau_i \sim \text{Ga}(1.25, 5). \end{aligned}$$

The hyperprior means match the parameters in the NEPPM model (4.1).

Parametric Hierarchical Model: The model borrows strength across subtypes, but treats a priori all subtypes symmetrically. No prognosis information is considered.

$$\begin{aligned} y_i | \mathbf{p} &\stackrel{\text{iid}}{\sim} \text{Bin}(p_i, N_i) \\ \text{logit}(p_i) &\stackrel{\text{iid}}{\sim} N(\mu, \tau) \\ \mu &\sim N(0, 1/4) \text{ and } \tau \sim \text{Ga}(1.25, 5). \end{aligned}$$

The use of hierarchical models with a priori exchangeable subpopulations is a standard approach for many biomedical inference problems that require borrowing of strength across subpopulations.

Hierarchical with Logistic Regression Model (HLRM): The HLRM assumes partial exchangeability by fixing the success rates as a logistic regression on x_i , the overall prognosis of disease subtype i . In other words, ρ is fixed as the grouping determined by the prognosis covariate. There is no learning about the partition.

$$\begin{aligned} y_i | \mathbf{p} &\stackrel{\text{iid}}{\sim} \text{Bin}(p_i, N_i) \\ \text{logit}(p_i) &\sim N(\beta_0 + \beta_1 x_i, \tau) \\ (\beta_0, \beta_1)^t &\sim \text{MVN}(0, I_2/4), \end{aligned} \tag{4.10}$$

where the precision τ is fixed to $\tau = 18$ like in (4.1).

Table 4.2: Simulation truth p_i under five alternative scenarios

scenario	good	intermediate	poor
N_i	6, 6, 8	40, 40, 30, 20, 20,10	7, 7, 6
S0	.59,.55,.47	.45,.4,.36,.34,.3,.26	.23,.19,.17
S1	.28,.26,.24	.18,.16,.15,.17,.2,.19	.12,.13,.10
S2	.28,.26,.18	.24,.16,.15,.17,.10,.19	.12,.20,.10
S3	.40,.45,.35	.15,.10,.15,.12,.10,.10	.45,.35,.40
S4	.30,.40,.15	.20,.10,.30,.20,.15,.12	.15,.40,.30

The covariate x_i is equal to $-1, 0$ or 1 when i indexes a sarcoma subtype with poor, intermediate or good prognosis, respectively; $MVN(\mu, \Lambda)$ is the multivariate normal distribution with mean μ and precision matrix Λ ; and I_2 denotes the 2×2 identity matrix. The precision matrix for the vector of regression parameters is chosen to match the hyperparameter means in (4.1).

In the comparison, we consider $n = 12$ different experimental units (sarcoma subtypes) with a categorical covariable x_i with values -1 (poor), 0 (intermediate) and 1 (good). Each simulated trial realization consists of n independent observations $y_i \sim \text{Bin}(p_i, N_i)$ with success rates fixed at an assumed simulation truth, and fixed sample size N_i . We use $N_i = 6, 6, 8, 40, 40, 30, 20, 20, 10, 7, 7$ and 6 , respectively. The first three subtypes have poor overall prognosis, $x_i = -1, i = 1, \dots, 3$. The last three subtypes have good prognosis, $x_i = 1, i = 10, \dots, 12$. The remaining six subtypes have $x_i = 0$. The sample sizes N_i are chosen to match the expected accrual under the 12 sarcoma subtypes in the motivating phase II sarcoma trial. For the simulation truth on p_i we consider five scenarios, S0 through S4, summarized in Table 4.2.

Scenarios S0 and S1 favor the HLRM model. The grouping by prognosis is perfect and the monotonicity assumption implicit in the HLRM is satisfied. The remaining scenarios represent varying levels of mismatch between prognosis and true success

probabilities and violations of the monotonicity assumption of the HLRM model. In S2, monotonicity is overall satisfied but the grouping by covariates is not perfect. In S3 the grouping is perfect but monotonicity violated. Finally, in S4 both, grouping by covariates and monotonicity are violated.

We generated $M = 200$ repeat simulations of the entire trial under these five scenarios (the match of $M = 200$ with $\sum N_i$ is coincidence). For each simulation $m = 1, \dots, 200$, and for each $i = 1, \dots, n$, we estimated p_i by the posterior mean \bar{p}_i^m . We evaluated bias and mean square error by

$$bias(\bar{p}_i) \approx \left| \frac{1}{M} \sum_{m=1}^M \bar{p}_i^m - p_i \right| \text{ and } mse(\bar{p}_i) \approx \frac{1}{M} \sum_{m=1}^M (\bar{p}_i^m - p_i)^2,$$

i.e., we use Monte Carlo averages to evaluate the (frequentist) means with respect to repeat experimentation. For scenario S0, we estimated the three models described earlier in this section plus two versions of the proposed NEPPM, once with $\gamma = 1/2$ and once with $\gamma = 1$ in (4.4). The comparison of all these models is summarized in Figure 4.1. In terms of MSE, the two version of the NEPPM perform similarly. The HLRM (4.10) produces the estimators with the lowest values of the MSE. This is to be expected since scenario S0 strongly favors HLRM. The NEPPM and HLRM perform better than the remaining models because they are the only ones that borrow strength across sarcoma subtypes and acknowledge the similarity of the success rates corresponding to the same prognosis. Figure 4.2 compares the CP of the central 95% credible interval (CI) under the HLRM vs. the NEPPM with $\gamma = 1$. The HLRM has low CP for the subtypes with lowest and largest intermediate prognosis success rates. This is due to the fixed grouping of all intermediate prognosis subtypes, leading to excessive shrinkage for the subtypes with $x_i = 0$ with lowest and highest true p_i . The HLRM model does not allow any change or weighting of the grouping.

Figure 4.3 compare NEPPM vs. HLRM under scenarios S1-S4. In scenario S1 the HLRM performs overall better than the NEPPM. Scenario S2 is the same as S1 but swapping in the simulation truth one intermediate (p_8) for a poor prognosis success

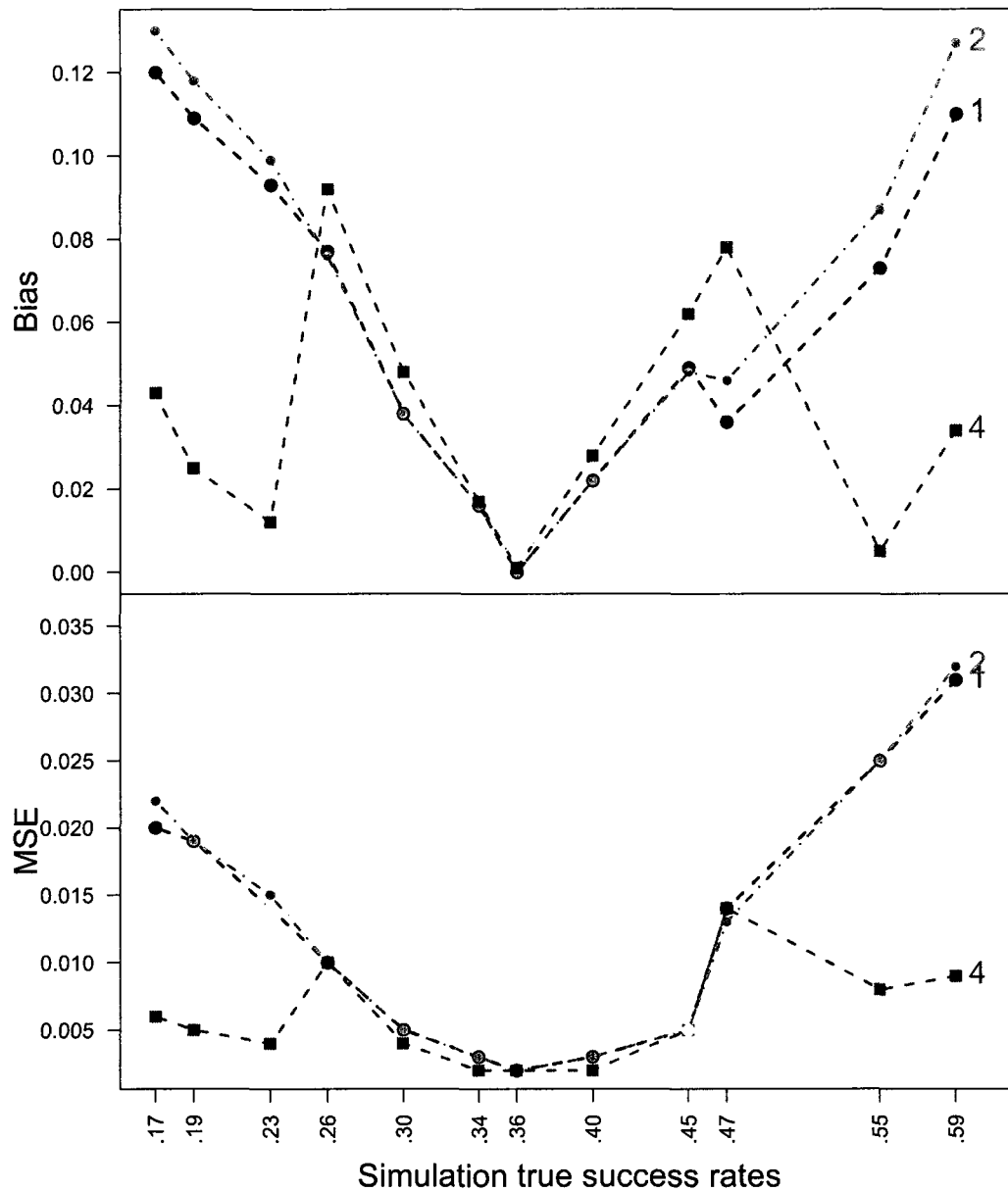


Figure 4.1: Comparison of the estimated success rates under the four models under S_0 . The horizontal axis shows the $n = 12$ success rates under the simulation truth S_0 . The upper panel shows the absolute bias. The lower panel shows mean squared error. Both are arranged by true p_i . The five lines and labels correspond to the models NEPPM with $\gamma = 1$ [1], NEPPM with $\gamma = 1/2$ [2], parametric Hierarchical [3], HLRM [4] and separate [5] models.

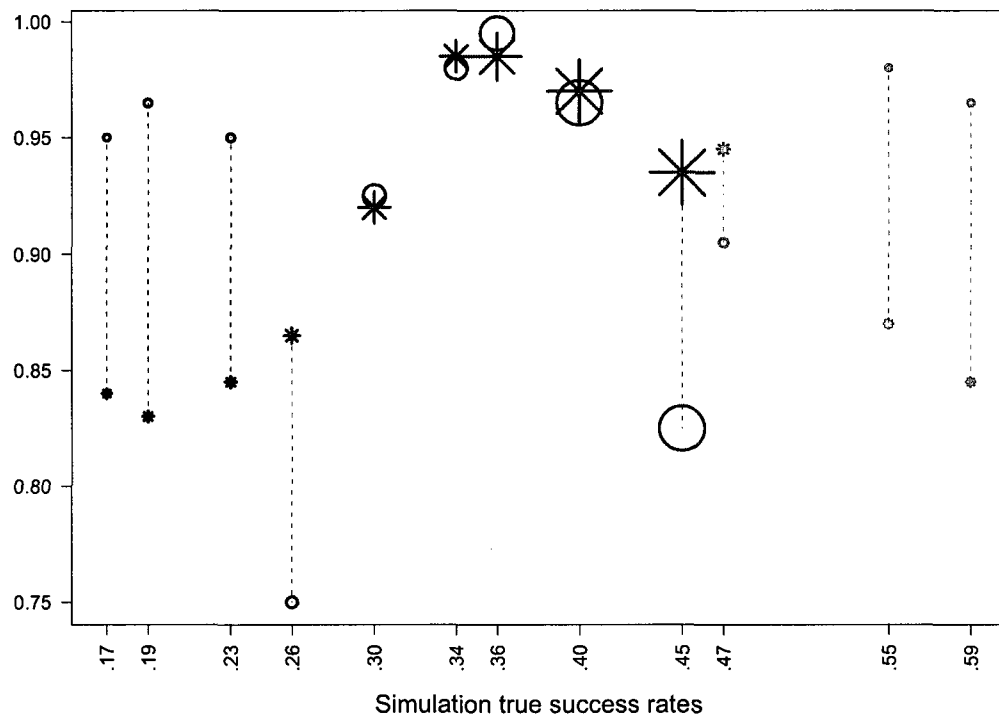


Figure 4.2: Scenario S0. Coverage probabilities of the central 95% credible intervals under the NEPPM with $\gamma = 1$ vs. the HLRM. From left to right, the first three success rates correspond to poor prognosis ($x_i = -1$), the following six to intermediate ($x_i = 0$) and the last three to good prognosis ($x_i = 1$). The character sizes are proportional to the sample size N_i .

rate (p_{11}) and another intermediate (p_7) for a good prognosis success rate (p_3). Figure 4.3 (b) shows that the HLRM fails to accurately estimate the swapped probabilities. The fixed grouping (by covariate) is inappropriate for these subtypes, leading to poor estimates of the corresponding success rates. Scenario S3 favors the NEPPM. The grouping is exact but the monotonicity assumed by the HLRM is violated. Figure 4.3 (c) shows how the NEPPM outperforms the HLRM in this scenario. The success rate estimates under the NEPPM have the favorable bias and MSE. The HLRM performs poorly in terms of CP even for the intermediate prognosis success rates in subtypes with large sample sizes N_i . In S4 the grouping by prognosis is inappropriate and the monotonicity assumption of the HLRM is violated. As Figure 4.3 (d) shows, the NEPPM performs better than the HLRM in this scenario.

4.4.2 Average Sample Size and Stopping Probabilities

We continue the comparison of NEPPM versus HLRM. We now focus on summaries that are relevant for early stopping for futility. We will accrue patients by cohorts of 10. We will stop recruiting patients for sarcoma subtype i and cancel the i -th study arm if

$$Pr[p_i > 0.175 \mid \text{data so far}] < 0.10. \quad (4.11)$$

Already accrued data for canceled study arms will continued to be used in the inference for other sarcoma subtypes, i.e., it remains part of the data set.

For both models we continue to use the same hyperparameters as in the previous section. The (maximum) sample sizes N_i , $i = 1, \dots, n$, remain as before. The best model should accrue the smallest number of patients in study arms for which the treatment is inefficient, that is when the success rate p_i is less than 0.175.

Results are summarized in Figure 4.4. Each panel reports the average number of patients in each study arm (\bar{N}_i), and for each study arm, the probability of early stopping (\bar{p}_i). Like the reported bias and MSE in the earlier comparison, both summaries, \bar{N}_i and \bar{p}_i , are with respect to repeat experimentation, i.e., they are an expectation

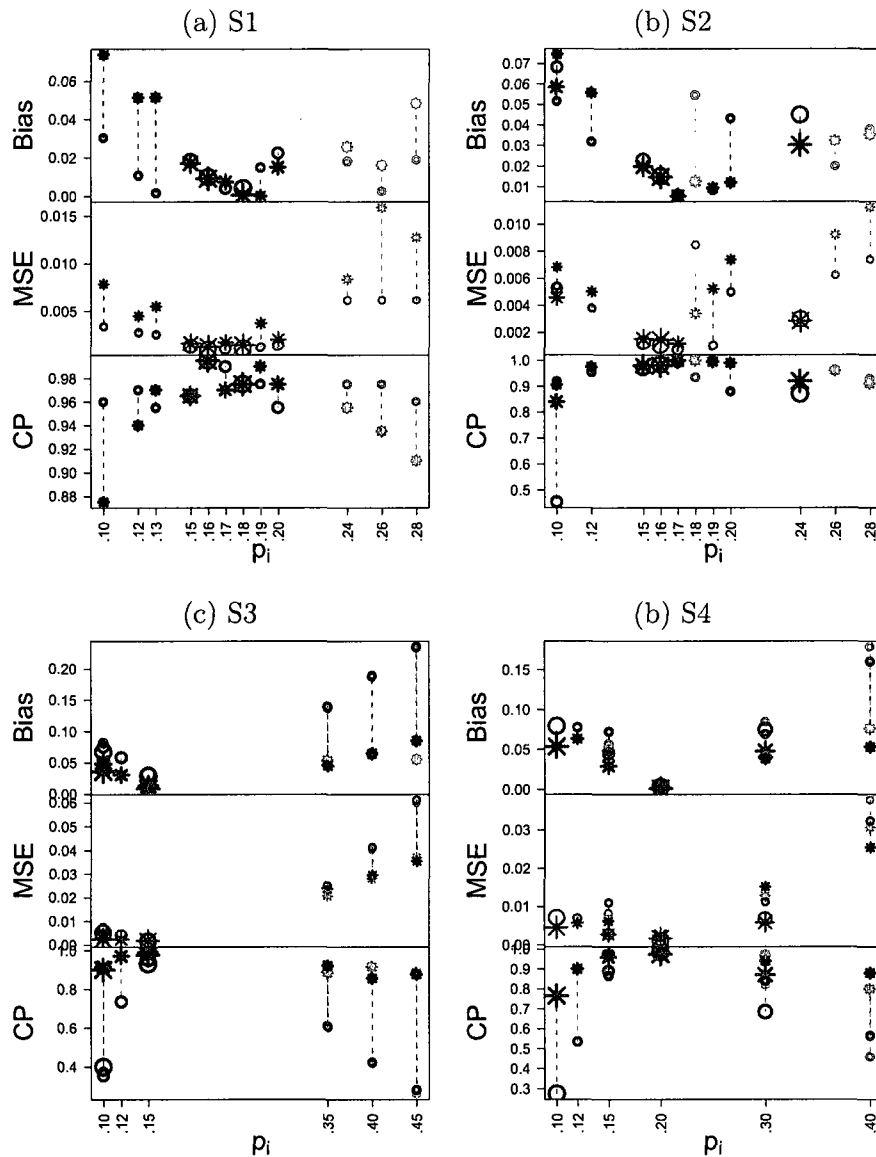


Figure 4.3: Scenarios S1 through S4. Panels (a) through (d) summarize simulation under scenarios S1 through S4. Comparison of the estimated success rates under the NEPPM (star “*”) vs. the HLRM (circle “o”). The horizontal axis shows the $n = 12$ true success rates under the assumed scenario. The upper, medium and lower panels show absolute value of the bias, mean square error and coverage probability of the central 95% credible interval, respectively. All are arranged by subtype. Under S1 (panel (a)), from left to right, the first three success rates correspond to poor prognosis ($x_i = -1$), the following six to intermediate ($x_i = 0$) and the last three to good prognosis ($x_i = 1$). The point sizes are proportional to the sample size N_i .

and a probability with respect to repeated simulations.

Recall that scenario S1 favors the HLRM. In particular, the success rates are ordered according to prognoses as assumed by the HLRM. Not surprisingly, the HLRM outperforms NEPPM for the subtypes with few patients. The implicit monotonicity assumption and the fixed grouping (which happens to match the simulation truth) greatly improve inference for these subtypes. The early stopping probability for subtypes with poor prognosis is (correctly) high and is low for subtypes with good prognosis. In contrast, under S2 there is no clear winner. See Figure 4.4 (b). The HLRM is overall more aggressive in the sense that it leads to higher early stopping probabilities. But it does so even when the treatment is effective (i.e., $p_i > 0.175$). We observe similar summaries under S3. Figure 4.4 (c) indicates that the HLRM stops earlier, across study arms, including those with $p_i \geq 0.175$. Finally, Figure 4.4 (d) shows comparable average sample sizes under S4 for both models.

In summary, we gain precision when incorporating the information of the covariates in the model and the covariates have predictive power. Among the 5 competing models considered in Section 4.1, the partially exchangeable HLRM and the NEPPM with random partitions show the best overall performance. Comparing these two models directly, when the assumed true success rates are monotone increasing with respect to prognosis, then the HLRM is optimal in terms of bias, MSE and coverage probability. However, if the assumed simulation truth does not match the grouping by prognosis x_i or the monotonicity is violated, then the NEPPM performs better. By its nature, the HLRM introduces strong prior beliefs about the similarity of the success rates. The model groups subtypes by x_i , and allows no modification of this fixed clustering. Inference is precise when these beliefs happen to be right. In contrast, the NEPPM introduces similar beliefs, but allows for uncertainty. The model allows the data to speak and correct the clustering in case the prior beliefs were inaccurate. In terms of early stopping probabilities, the HLRM tends to stop study arms earlier than the NEPPM. In all but S1 where HLRM wins, there is not a clear winner

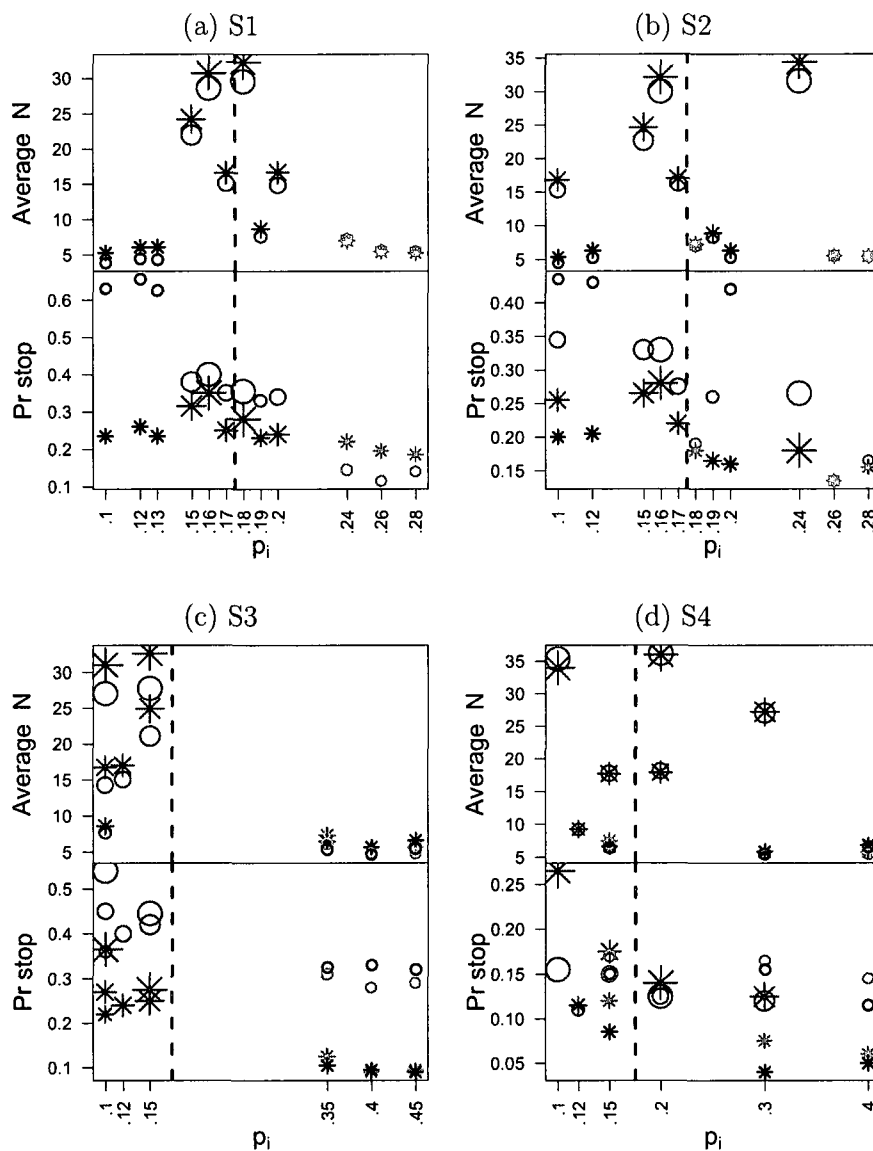


Figure 4.4: S1 through S4. Panels (a) through (d) show the average number of patients (\bar{N}_i) and early stopping probabilities (\bar{p}_i) under scenarios S1 through S4. Both summaries are with respect to repeat experimentation. Summaries are arranged by the simulation truth p_i , shown on the horizontal axis. In each panel, the upper part shows the average number of patients \bar{N}_i that enter into the study for each arm. The lower part shows the early stopping probability (\bar{p}_i). The stars (“*”) show summaries under the NEPPM. The circles (“o”) show summaries for the HLRM. The character size is proportional to N_i , the maximum sample size for each subtype.

model.

4.5 Results

We implemented inference for the data described in Section 4.2 using the proposed non-exchangeable partition model 4.1 with $\gamma = 1$. The parameter α of the Dirichlet process associated with the probability in the space of partitions of the space of indices was assumed to be $Ga(5, 0.5)$ distributed.

Saving every 10th iteration after a 10,000 iteration burn-in, a Monte Carlo posterior sample of size 10,000 was saved to estimate the success rates. The Markov chain mixed well. Both the central 95% credible intervals and the 5% percentile of the success rate p_i for each sarcoma subtype in the study are shown in Figure 4.5. Only for Angiosarcoma, Synovial and MFH we find posterior probabilities greater 95% that p_i is greater than 0.1.

4.6 Discussion

We proposed an approach to borrow strength across non-exchangeable subpopulations. The usual approach to borrow strength across subpopulations is through hierarchical models, perhaps one of the most successful Bayesian approaches in biomedical data analysis. In a hierarchical model, the estimation of any subpopulation-specific effect borrows the same amount of strength from all observation in other subpopulations. In a partially exchangeable hierarchical regression model inference borrows strength across all subpopulations that are grouped together in some fixed a priori grouping by covariates. In contrast, the proposed model introduces the non-exchangeability only stochastically, with random partitions, through the prior distribution and the estimation of subpopulation-specific effects p_i borrows more strength from the subset of observations, that according to our prior beliefs, are exchangeable

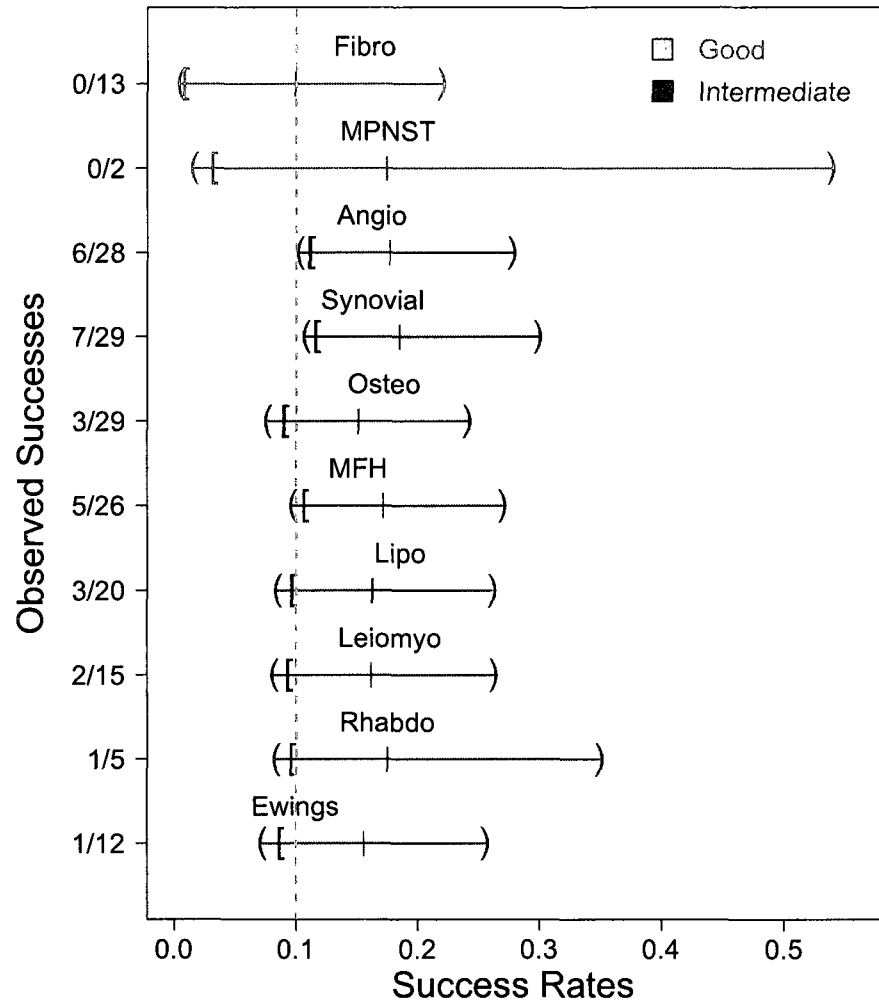


Figure 4.5: Central 95% credible intervals for the success rates of the treatment in the sarcoma subtypes in the study when applying the proposed NEPPM with parameter $\gamma = 1$. Right square bracket marks the 5% percentile. The upper two and the rest CI's correspond, respectively, to the two and ten sarcoma subtypes with good and intermediate prognoses.

with the observation i . The proposed model is robust in the sense that it allows the data to speak and correct prior assumptions when the prior beliefs happen not to be confirmed by the data. However, the precision of estimates under the proposed model can be lower than under a partially exchangeable hierarchical regression because inference has to account for the uncertainty in the grouping. This was confirmed in the simulation study. The proposed model performed better when the prior beliefs were not matched by the simulation truth.

The proposed model is useful when sample sizes are small in some subpopulations. With large samples in all subpopulations, a separate model for each subpopulation could be considered. However, small samples are usually the case in early phase clinical trials. In particular due to the rare nature of the sarcomas studied in the motivating phase II clinical trial borrowing strength is essential to obtain precise estimations.

The proposed model included a novel distribution over random partitions that gives increased *a priori* weights to homogeneous partitions. The additional computational effort compared to a conventional Pólya urn (induced by a Dirichlet process mixture) is minimal. Depending on the application, in principle any sampling model and any distribution for the cluster-specific effects can be considered. The proposed model is a particular case of the more general PPMx model introduced in Müller et al. (2009). Their model uses covariate information to change the prior probability of clustering. They consider continuous, ordinal and nominal covariates.

In summary, we have proposed an easy to implement model to borrow strength across non-exchangeable subpopulations. The approach compromises between borrowing too much strength (such as the hierarchical regression model) and no borrowing at all (such as separate models for each subpopulation). The proposed model is suitable to estimate the success rates in the motivating sarcoma trial presented here by borrowing strength across the different subtypes of the disease. It considers a pair of sarcoma subtypes to be *a priori* likely to be exchangeable when they are related

with the same prognosis.

Appendix

Gibbs Sampling Scheme

Any set of imputed parameters $\theta_1, \dots, \theta_n$ corresponds to a partition $\rho = (S_1, \dots, S_k)$ of the indices $\{1, \dots, n\}$. Let $\theta_1^*, \dots, \theta_K^*$ the unique values in the sample in order of appearance and define $S_k = \{i : \theta_i = \theta_k^*\}$.

Consider the Dirichlet process mixture model (Antoniak, 1974) in (3.9). Let $\theta^{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n)$ denote all values but θ_i . Let K^{-i} be the number of unique values in θ^{-i} , let $\theta_1^{*-i}, \dots, \theta_{K^{-i}}^{*-i}$ be these different values in order of appearance and ρ^{-i} be the partition implied by θ^{-i} . Use equation (4.5) with the $n-1$ observations and let the observation i be the future observation (exchangeability allows us to permute the indices) to get:

$$p[\theta_i | \theta^{-i}] = f_{K^{-i}+1}(\rho^{-i})g_0(\theta_i) + \sum_{k=1}^{K^{-i}} f_k(\rho^{-i})\delta_{\theta_k^{*-i}}(\theta_i),$$

where δ_x is a pointmass at x and g_0 is the pdf corresponding to G_0 . Therefore,

$$\begin{aligned} p[\theta_i | \theta^{-i}, y^n] &\propto p[y^n | \theta_i, \theta^{-i}]p[\theta_i | \theta^{-i}] \\ &\propto \left\{ \prod_{j=1}^n p[y_j | \theta_j] \right\} \left\{ f_{K^{-i}+1}(\rho^{-i})G_0(\theta_i) + \sum_{k=1}^{K^{-i}} f_k(\rho^{-i})\delta_{\theta_k^{*-i}}(\theta_i) \right\} \\ &\propto p[y_i | \theta_i]f_{K^{-i}+1}(\rho^{-i})G_0(\theta_i) + \sum_{k=1}^{K^{-i}} f_k(\rho^{-i})p[y_i | \theta_i]\delta_{\theta_k^{*-i}}(\theta_i) \end{aligned}$$

Using (4.5) we get:

$$p[\theta_i | \theta^{-i}, y^n] \propto \frac{\alpha}{Q^\gamma} p[y_i | \theta_i] G_0(\theta_i) + \sum_{k=1}^{K^{-i}} \#S_k \left(\frac{m_{kl} + 1}{\#S_k + Q} \right)^\gamma p[y_i | \theta_k^{*-i}] \delta_{\theta_k^{*-i}}(\theta_i),$$

where l is the value of the categorical covariate (in our particular, the prognosis) corresponding to the i^{th} experimental unit (patient).

Observing that

$$p(y_i | \theta_i)G_0(\theta_i) = \frac{p(y_i | \theta_i)G_0(\theta_i)}{p(y_i)}p(y_i) = p(\theta_i | y_i) \int p(y_i | \theta) d\theta,$$

we get that with probability $q_{i0} := \alpha/Q^\gamma \times \int p(y_i | \theta) d\theta$, θ_i is a new observation from $p(y_i | \theta_i)$.

Consider the DPM Gibbs sampling scheme (MacEachern and Müller, 1998 and Neal, 2000), described in Subsection 3.2.3, with a Dirichlet process with total mass parameter α and base measure G_0 . In this scheme, θ_i is set equal to θ_k^{*-i} with probability $q_{ik}^D := \#S_k p(\theta_i|y_i)$. In contrast, we will set θ_i equal to θ_k^{*-i} with probability

$$q_{ik} = q_{ik}^D \times \left(\frac{m_{kl} + 1}{\#S_k + Q} \right)^\gamma.$$

The expression above has an important practical implication. Assume we have code for posterior simulation under the DP prior with parameters α and G_0 . Only a slight modification in the predictive probability function, i.e., in $Pr[\theta_i = \theta_k^* | \boldsymbol{\theta}^{-i}, y^n]$, of the Dirichlet process is necessary to implement posterior simulation under the proposed

nonexchangeable product partition model.

In the implementation of the NEPPM in (4.1) the normal prior for θ_k^* in model plays the role of G_0 in the algorithm described above. Besides, we assume a gamma distribution for the DP total mass parameter α . To do so, we use the model augmentation with a latent Beta random variable proposed by West (1992) (step (d) of the Section 3.2), to implement posterior inference for α .

Conclusion

In this thesis I have presented three research projects using non-parametric Bayesian model-based statistical inference for biomedical data. The common themes of these analyses are flexible statistical models and a refinement of previously published parametric analyses for similar datasets. Some limitations remain.

In the second chapter, I introduced a model to analyze adverse event data gathered in a phase III clinical trial. Each data record consists of the observed grades of seven different adverse events exhibited by a patient (including grade 0 for adverse events that were not recorded for the patient). To my knowledge, this approach is the first model-based inference that accounts for and assesses the very plausible correlation of the grades of the adverse events exhibited by the same patient. Besides, the proposed model is more flexible than standard models for ordinal data in the sense that it is able to fit cell probabilities (i.e. the probability that a patient exhibits certain adverse event at a determined level) that do not necessarily satisfy the parallel regression assumption. The parallel regression assumption is the probit version of the proportional odds assumption for the logit model. The data structure in this problem is very common in other applications, for example in a survey where each question has an ordinal outcome and it is expected that the answers given by the same respondent are correlated. The implementation of the model presents one difficulty. The latent variable that determines the ordinal outcome is distributed according to a mixture of normal distributions. Determining the correct number of components, G , in this mixture is not trivial. We empirically explore different values of G and make a

recommendation. Another approach is to make G random by using, for example, reversible jump. This would greatly complicate inference. Since the parameter G is not of primary interest we did not pursue this direction. The model has two more limitations. First, we showed that for any given set of toxicity probabilities under placebo ($x = -1$), there exists a mixture of normal densities that can represent these probabilities as the area under the curve in the (fixed) intervals $(\theta_k, \theta_{k+1}]$. The model assumes that by a simple shift of this mixture on the horizontal axis we can represent the toxicity probabilities under treatment ($x = 1$). Second, a univariate parameter, the patient-specific random effect, models the dependency of the different toxic grades reported by the same patient. This implies, in particular that toxic grades are positively correlated. For the particular application this is appropriate. Nevertheless, in general, a scalar patient-specific random effect may not be sufficient to model the correlation structure across adverse events. A possible solution is to consider a multivariate patient-specific random effect. Modifications of the model to analyze data with different dependence structures could be studied. An example is mentioned in the discussion section of the second chapter. It considers the analysis of repeated toxicity measures of the adverse events in the same patient.

In chapter 3, I introduced a semi-parametric model to analyze the outcome of multistage phage display experiments with humans. The aim of the phage display experiment is to identify peptides that bind with high affinity to specific tissues. That is, the objective is a list of peptides binding to specific tissues. The particular experiment considered has three stages. For every peptide-tissue pair, we only have one observation: the triplet of counts in the three stages. The hierarchical structure of the model allows borrowing strength across all pairs. The binding behavior of a peptide is reflected by the event of having increasing mean counts across the three stages. The model allows a biological interpretation of the parameters and an easy representation of the event of increasing mean counts. Besides, the model is mathematically tractable. The MCMC simulation is straightforward due to the use of

conjugate families. A simple inspection of the data suggests that there are outliers. The semi-parametric nature of the model allows for possible heterogeneity of the data and robustifies inference. A limitation of the model is that it assumes all pairs to be *a priori* exchangeable. Instead, a more realistic model should assume that the peptide A in tissue T is more likely to behave similar to the same peptide A in other tissue or similar to any other peptide in the same tissue T.

In the same chapter, we propose an inference from a decision theoretic point of view. At the moment of selecting the list of pairs to report to the biologist collaborator for further research and interpretation, we face a massive multiplicity problem. For each peptide/tissue pair we are testing the alternative hypothesis of increasing means. The most commonly used multiplicity adjustment is Bonferroni's Criterion. This procedure works well for testing few hypotheses, but it is too conservative when the number of hypotheses is very large. A more recent method to address massive multiplicity problems is by controlling the (frequentist) False Discovery Rate (FDR). FDR is the expected proportion of false positives in the list of reported comparisons ("discoveries"). We control the related posterior FDR, i.e., the posterior expected proportion of false discoveries. By controlling I mean establishing an upper bound on the posterior expected FDR while maximizing the number of reported pairs in the list. This Bayesian procedure can be characterized as a Bayes rule under a utility function that considers statistical significance. But the utility function ignores biological significance. That is, it only takes into account whether or not the pair has (significantly) increasing counts, but ignores the size of this increase. This observation leads us to consider an alternative FDR that does account for biologic significance, simply by modifying the underlying utility function.

In chapter 4 I address inference for non-exchangeable experimental units. The motivating application is to a clinical trial for rare sarcomas, including patients from $n = 12$ different disease subtypes with very slow accrual for some of the subtypes. A practical clinical trial design requires borrowing of strength across the disease

subtypes to reach any meaningful conclusion. The disease subtypes are the non-exchangeable experimental units. In general, I consider problems when experimental units are grouped according to a categorical covariate, with small sample sizes in some groups and when the experimental units may not be exchangeable across the values of the covariate. In the motivating application the covariate is the overall prognosis for each disease subtype. Borrowing strength across the categories is necessary to increase the precision of inference in the small-size categories, in our case the disease subtypes with few patients. A standard procedure to borrow strength across subpopulations is a hierarchical model when the experimental units are exchangeable, or a hierarchical regression under partial exchangeability. The hierarchical regression groups all experimental units with the same categorical covariate together and borrows strength across them. When the outcome is binary the model becomes a hierarchical logistic regression model (HLRM). Under the hierarchical regression model the grouping is fixed. Inappropriate grouping induced by the covariate can lead to poor inference.

I propose a semi-parametric model that is more robust against inappropriate grouping by considering random grouping, or partitions. The model introduces the covariate through the prior on the random partition. Inference for the parameter corresponding to the i -th subpopulation borrows more strength from observations in subpopulations that are with high probability grouped together with i . The data can correct a prior guess on the grouping when the prior (grouping) beliefs are inaccurate. In the same chapter I compare the proposed model with some parametric approaches. I compare the performance in terms of bias, mean square error and coverage probabilities under different scenarios. I show that the proposed model is more robust against inappropriate grouping than standard approaches. Unfortunately, we could not show that the model is any better than the HLRM in terms of stopping probabilities and average sample sizes when applied in a clinical trial design. That is, the proposed model does not detect inefficient treatments any faster than the HLRM. The model

is applied to the motivating sarcoma phase II trial. The rare nature of the disease makes borrowing strength essential.

The proposed model is based on a product partition model for random partitions. The model assigns high prior probability to partitions with homogeneous clusters. That is, clusters with few different values of the covariate associated to the indices in the cluster. The model is a particular case of the more general PPMx model introduced in Müller et al. (2009). Their model considers continuous, ordinal and categorical covariates. In the same paper their model is applied to examples with ordinal and continuous covariate. To my knowledge, this is the first application of the PPMx model with categorical covariate. The computation burden associated with the implementation of this model is only slightly greater than the computational effort involved in inference for random partitions induced by the Dirichlet process mixture model (Pólya urn).

In summary, the three flexible and non-parametric Bayesian models proposed in this thesis are an improvement over previously used parametric models to analyze similar data-sets. All models can be extended and applied to other problems with appropriate modifications. They all have limitations and improvements are possible.

Bibliography

- [1] John Aitchison and Jo A. Bennett. Polychotomous quantal response by maximum indicant. *Biometrika*, 57:253–262, 1970.
- [2] James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.
- [3] Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.
- [4] Wadih Arap, Mikhail G. Kolonin, Martin Trepel, Johanna Lahdenranta, Marina Card-Vila, Paul J. Giordano, Ricardo J. Mintz, Peter U. Ardel, Virginia J. Yao, Claudia I. Vidal, Limor Chen, Anne Flamm, Heli Valtanen, Lisa M. Weavind, Marshall E. Hicks, Raphael E. Pollock, Gregory H. Botz, Corazon D. Bucana, Erkki Koivunen, Dolores Cahill, Patricia Troncoso, Keith A. Baggerly, Rebecca D. Pentz, Kim-Anh Do, Christopher J. Logothetis, and Renata Pasqualini. Steps toward mapping the human vasculature by phage display. *Nature Medicine*, 8(2):121 – 1271, 2002.
- [5] Daniel Barry and J. A. Hartigan. A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88:309–319, 1993.
- [6] Sanjib Basu and Saurabh Mukhopadhyay. Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhyā*, 62:372–387, 2000.

- [7] David Blackwell and James B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.
- [8] Olivier Cappé and Christian P. Robert. Markov chain Monte Carlo: 10 years and still running! In Adrian E. Raftery, Martin Abba Tanner, and Martin Wells, editors, *Statistics in the 21st century*, pages 302–311. Chapman & Hall Ltd, 2002.
- [9] Mary Kathryn Cowles. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6:101–111, 1996.
- [10] Mary Kathryn Cowles and Bradley P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904, 1996.
- [11] David B. Dahl. An improved merge-split sampler for conjugate Dirichlet Process mixture models. Technical Report 1086, Department of Statistics, University of Wisconsin, 2003.
- [12] Hani Doss. Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*, 22:1763–1786, 1994.
- [13] Michael Escobar and Michael West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.
- [14] Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, April 1973.
- [15] Thomas S. Ferguson. Prior distributions on spaces of probability measures. *Annals of Statistics*, 2(4):615–629, August 1974.

- [16] Dani Gamerman and Hedibert F. Lopes. Markov chain Monte Carlo: Stochastic simulation for bayesian inference: second edition. In *Texts in Statistical Science Series*. Chapman & Hall Ltd, 2006.
- [17] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.
- [18] David Jr. Harrison and Daniel L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, March 1978.
- [19] J. A. Hartigan. Partition models. *Communications in Statistics: Theory and Methods*, 19:2745–2756, 1990.
- [20] Alejandro Jara. Applied bayesian non- and semi-parametric inference using DP-package. *Rnews (in press)*, 2007.
- [21] Yuan Ji, Guosheng Yin, Kam-Wah Tsui, Mikhail G. Kolonin, Jessica Sun, Wadih Arap, Pasqualini Renata, and Kim-Anh Do. Bayesian mixture models for complex high dimensional count data in phage display experiments. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 56(2):139–152, 2007.
- [22] Valen E. Johnson and James H. Albert. *Ordinal Data Modeling*. Springer-Verlag Inc, 1999.
- [23] Ken P. Kleinman and Joseph .G. Ibrahim. A semi-parametric bayesian approach to the random effects model. *Biometrics*, 54:921–938, 1998.
- [24] Mikhail G. Kolonin, Jessica Sun, Kim-Anh Do, Claudia I. Vidal, Yuan Ji, Keith A. Baggerly, Renata Pasqualini, and Wadih Arap. Synchronous selection of homing peptides for multiple tissues by in vivo phage display. *FASEB J.*, 20:979–981, 2006.

- [25] Athanasios Kottas, Peter Müller, and Fernando Quintana. Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14(3):610–625, 2005.
- [26] Albert Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12:351–357, 1984.
- [27] Steven N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, 23:727–741, 1994.
- [28] Steven N. MacEachern and Peter Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.
- [29] Noriyuki Masuda, Shunichi Negoro, Shinzoh Kudoh, Takahiko Sugiura, Kazuhiko Nakagawa, Hideo Saka, Minoru Takada, Hisanobu Niitani, and Masahiro Fukuoka. Phase i and pharmacologic study of docetaxel and irinotecan in advanced nonsmall-cell lung cancer. *Journal of Clinical Oncology*, 18:2996–3003, 2000.
- [30] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B: Methodological*, 42:109–142, 1980.
- [31] Robert McCulloch and Peter E. Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64:207–240, 1994.
- [32] Daniel McFadden. A method of simulated moments for estimation of discrete response models without numerical integration (STMA V31 2344). *Econometrica*, 57:995–1026, 1989.
- [33] Saurabh Mukhopadhyay and Alan E. Gelfand. Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association*, 92:633–639, 1997.

- [34] Peter Müller, Giovanni Parmigiani, and Kenneth Rice. FDR and Bayesian multiple comparisons rules. Working Paper 115. <http://www.bepress.com/jhubiostat/paper115>, 2006.
- [35] Peter Müller, Fernando A. Quintana, and Rosner Gary L. Bayesian clustering with regression. Submitted, 2009.
- [36] Peter Müller, Fernando A. Quintana, and Gary L. Rosner. Hierarchical meta-analysis over related non-parametric Bayesian models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 66:735–749, 2004.
- [37] Peter Müller and Gary Rosner. A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, 92:1279–1292, 1997.
- [38] Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.
- [39] Michael A. Newton. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics (Oxford)*, 5(2):155–176, 2004.
- [40] Christopher R. Palmer and William F. Rosenberger. Ethics and practice: alternative designs for phase III randomized clinical trials. *Controlled clinical trials*, 20:172–186, 1999.
- [41] Stuart J. Pocock. When to stop a clinical trial. *British medical journal*, 305:235–240, 1992.
- [42] Fernando A. Quintana and Pilar L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 65(2):557–574, 2003.

- [43] Fernando A. Quintana and Peter Müller. Nonparametric Bayesian assessment of the order of dependence for binary sequences. *Journal of Computational and Graphical Statistics*, 13:213–231, 2004.
- [44] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.
- [45] Stephen G. Walker, Paul Damien, Purushottam W. Laud, and Adrian F. M. Smith. Bayesian nonparametric inference for random distributions and related functions (Disc: P510-527). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61:485–509, 1999.
- [46] Michael West. Bayesian kernel density estimation. Technical Report ISDS Discussion Paper 90-A02, Duke University, 1990.
- [47] Michael West. Hyperparameter estimation in Dirichlet process mixture models. Technical Report 92-A03, Institute of Statistics and Decision Sciences. Duke University, 1992.
- [48] Mike West, Peter Müller, and Michael D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. In P. R. Freeman and A. F. M. Smith, editors, *Aspects of Uncertainty. A Tribute to D. V. Lindley*, pages 363–386. John Wiley & Sons, 1994.
- [49] Eelco F.M. Wijdicks. The diagnosis of brain death. *N. Engl. J. Med.*, 344:1215–1221, 2001.
- [50] Samuel S. Wilks. *Mathematical Statistics*. Wiley, New York, NY, 1962.
- [51] Xian Zhou. *Bayesian Inference of Ordinal Data*. PhD thesis, Rice University, 2005.