2011

BAYESIAN SEMIPARAMETRIC GENERALIZATIONS OF LINEAR MODELS USING POLYA TREES

Angela Schoergendorfer
University of Kentucky, angelasch@gmail.com

ABSTRACT OF DISSERTATION

Angela Schoergendorfer

The Graduate School

University of Kentucky

2011

BAYESIAN SEMIPARAMETRIC GENERALIZATIONS OF LINEAR MODELS USING POLYA TREES

ABSTRACT OF DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements of the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By

Angela Schoergendorfer

Lexington, Kentucky

Co-Directors: Dr. Adam Branscum, Professor of Statistics,

and Dr. Kert Viele, Professor of Statistics

Lexington, Kentucky

2011

ABSTRACT OF DISSERTATION

BAYESIAN SEMIPARAMETRIC GENERALIZATIONS OF LINEAR MODELS USING POLYA TREES

In a Bayesian framework, prior distributions on a space of nonparametric continuous distributions may be defined using Polya trees. This dissertation addresses statistical problems for which the Polya tree idea can be utilized to provide efficient and practical methodological solutions.

One problem considered is the estimation of risks, odds ratios, or other similar measures that are derived by specifying a threshold for an observed continuous variable. It has been previously shown that fitting a linear model to the continuous outcome under the assumption of a logistic error distribution leads to more efficient odds ratio estimates. We will show that deviations from the assumption of logistic error can result in great bias in odds ratio estimates. A one-step approximation to the Savage-Dickey ratio will be presented as a Bayesian test for distributional assumptions in the traditional logistic regression model. The approximation utilizes least-squares estimates in the place of a full Bayesian Markov Chain simulation, and the equivalence of inferences based on the two implementations will be shown. A framework for flexible, semiparametric estimation of risks in the case that the assumption of logistic error is rejected will be proposed.

A second application deals with regression scenarios in which residuals are correlated and their distribution evolves over an ordinal covariate such as time. In the context of prediction, such complex error distributions need to be modeled carefully

and flexibly. The proposed model introduces dependent, but separate Polya tree priors for each time point, thus pooling information across time points to model gradual changes in distributional shapes. Theoretical properties of the proposed model will be outlined, and its potential predictive advantages in simulated scenarios and real data will be demonstrated.

KEYWORDS: Polya trees, risk estimation, logistic regression, Bayesian nonparametrics, longitudinal data.

Angela Schoergendorfer

Student's signature

June 17, 2001

Date

BAYESIAN SEMIPARAMETRIC GENERALIZATIONS OF LINEAR MODELS USING POLYA TREES

By

Angela Schoergendorfer

Dr. Adam Branscum

Co-Director of Dissertation

Dr. Kert Viele

Co-Director of Dissertation

Dr. Arne Bathke

Director of Graduate Studies

June 17, 2011

Date

RULES FOR THE USE OF DISSERTATIONS

Unpublished dissertations submitted for the Master's and Doctor's degrees and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgments.

Extensive copying or publication of the dissertation in whole or in part requires also the consent of the Dean of the Graduate School of the University of Kentucky.

A library which borrows this dissertation for use by its patrons is expected to secure the signature of each user.

Name and Address Date

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

DISSERTATION

Angela Schoergendorfer

The Graduate School

University of Kentucky

2011

BAYESIAN SEMIPARAMETRIC GENERALIZATIONS OF LINEAR MODELS USING POLYA TREES

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements of the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By

Angela Schoergendorfer

Lexington, Kentucky

Co-Directors: Dr. Adam Branscum, Professor of Statistics,

and Dr. Kert Viele, Professor of Statistics

Lexington, Kentucky

2011

To Svea and Bruno Schörgendorfer

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor Dr. Adam Branscum for his support, advice and patience during the process of writing this dissertation. I would also like to thank Dr. Timothy Hanson for his ideas and feedback that contributed to the methods presented in this work. I extend my thanks to the members of my dissertation committee for their encouragement and feedback, and the Department of Statistics, specifically Dr. Arnold Stromberg and Dr. Arne Bathke, for all the support I have received during my time at the University of Kentucky.

I have received an unmeasurable amount of moral support from Kyle Mullikin, Ingrid, Helene, and Doris Schörgendorfer, Franz Grieger, Sandra Vlasich, and Dustin and Melissa Lueker. My gratitude goes out to them.

TABLE OF CONTENTS

Acknowledgments **iii**

List of Tables **vi**

List of Figures **vii**

1 Introduction **1**
- 1.1 Priors on spaces of distributions . . . 2
  - 1.1.1 Dirichlet processes . . . 3
  - 1.1.2 Polya trees . . . 4
  - 1.1.3 Fitting Polya tree models . . . 11
- 1.2 Bayes factors . . . 16
- 1.3 Dissertation outline . . . 18

2 Bayesian Semiparametric Risk Regression with Measurement Data **19**
- 2.1 Introduction . . . 19
- 2.2 Background . . . 22
- 2.3 Impact of model misspecification . . . 25
- 2.4 Testing for goodness of fit in logistic regression . . . 29
  - 2.4.1 Empirical Bayes test . . . 31
  - 2.4.2 Theoretical results . . . 32
- 2.5 Estimation . . . 37
- 2.6 Simulation study . . . 39
  - 2.6.1 Full Bayesian approach . . . 39
  - 2.6.2 Simulation data . . . 41
  - 2.6.3 Test performance . . . 41
  - 2.6.4 Estimation . . . 51
- 2.7 Examples . . . 52
  - 2.7.1 Risk factors for obesity in the Health and Retirement Study . . . 53
  - 2.7.2 Risk of diabetes in NHANES . . . 57
- 2.8 Summary . . . 62

3 A Dependent Polya Tree Model for Regression Analysis **64**
- 3.1 Introduction . . . 64
- 3.2 Background . . . 65
- 3.3 Dependent Polya tree model . . . 67
  - 3.3.1 Theoretical structure . . . 69
- 3.4 Simulation study . . . 71
  - 3.4.1 Error models . . . 72
  - 3.4.2 Simulation settings . . . 73
  - 3.4.3 Results . . . 75
- 3.5 Example . . . 83
  - 3.5.1 Growth data . . . 83
- 3.6 Summary . . . 89

4 Summary and Outlook **91**

A Appendix **93**
A.1 Notation 93

Bibliography **94**

Vita **105**

LIST OF TABLES

1.1 Cutoff values for Bayes factors 17

2.1 Simulation results without covariates (1) 43
2.2 Simulation results without covariates (2) 44
2.3 Median of $\log_{10}(BF)$ without covariates 47
2.4 Simulation results with covariates (1) 48
2.5 Simulation results with covariates (2) 49
2.6 Median of $\log_{10}(BF)$ with covariates 50
2.7 Results for risk estimation with true logistic error 52
2.8 Results for risk estimation with true exponential error 53
2.9 Risk factors included in the analysis of the Health and Retirement data 54

3.1 LPML values for model comparison for three sample sizes n 78
3.2 Age groups defined for the IgG data set 84
3.3 LPML values from the DFPT model for the IgG data set 87
3.4 Parameter estimates for two models from the DFPT model 87
3.5 LPML values from various models for the IgG data set presented by Jara and Hanson [in press]) 89

A.1 Polya tree notation 93
A.2 Distributions 93

LIST OF FIGURES

1.1 Schematic of the construction of a Polya tree 6
1.2 Possible posterior density estimates from a finite PT distribution for various levels J 9
1.3 Posterior samples from a mixture of Polya trees 14
1.4 Estimated distributions from a mixture of Polya trees 15

2.1 True risk and estimated risk from ordinary logistic regression and AFT model for three non-logistic error distributions 27
2.2 True log odds ratios for three non-logistic error distributions 28
2.3 Risk estimation for any arbitrary shape of the residual distribution . 38
2.4 Histogram of residuals for the Health and Retirement data 55
2.5 Estimated risk of obesity 56
2.6 Estimated odds ratios of obesity for risk factors exercise, smoking and alcohol 58
2.7 Histogram of residuals for the NHANES data set 59
2.8 Odds ratio estimates for gender estimated 60
2.9 Estimated odds ratios of diabetes 61

3.1 Predictive error densities for a skew-normal distribution 79
3.2 Predictive error densities for a distribution changing from normal to right skewed 80
3.3 Predictive error densities for a distribution changing from normal to bimodal 81
3.4 Predictive error densities for a distribution changing from left to right skewed 82
3.5 Distribution of ages 85
3.6 Predictive densities and estimated median function 88

Chapter 1 Introduction

The strict parametric assumptions of standard model theory, while simplifying computations for estimation and inference, in practice are rarely met by real data. At least for large sample sizes, deviations from parametric assumptions may not affect estimation of the mean structure dramatically, but in the context of prediction of individual observations distributional misspecifications may have a great effect and lead to inappropriate inferences. For example, we will show that common violations of parametric assumptions, such as skewness, in the context of risk estimation can lead to dramatic biases. Nonparametric methods, on the other hand, make no assumptions about the general shape of distributions and are therefore more flexible in accommodating patterns observed in the data.

Gelfand [1999] describes the objective of semiparametric modeling as enriching the class of standard parametric models by specifying at least portions of the model nonparametrically, while retaining the main linear structure. This dissertation presents two semiparametric generalizations of linear models using nonparametric Bayesian methods. In the models presented here, the residual error distribution will be modeled nonparametrically, while the remaining parametric formulation of the model is maintained. This results in a median, rather than a mean, regression model.

The remainder of this chapter presents an overview of methods that will be employed in the method development in this dissertation. Section 1.1 discusses nonparametric methods that have been developed for the Bayesian framework. Specifically, the concept of the Polya tree prior, which is a generalization of the Dirichlet process, and computational aspects of Polya tree models are explained. Furthermore, approaches to model selection in the Bayesian setting, such as Bayes factors and log-pseudo marginal likelihood, are presented in Section 1.2, as these metrics will be employed in model comparisons. Section 1.3 gives an outline of the remainder of the dissertation.

1.1 Priors on spaces of distributions

Parametric statistical models specify a probability model f_θ that is completely known up to a parameter vector θ, where $\theta \in \Theta \subset \mathbb{R}^p$. In parametric Bayesian statistics, uncertainty about θ is addressed by assigning it a prior distribution $p_\theta(\theta)$, which quantifies how likely or unlikely sets $A \in \Theta$ are to contain the "true" value of θ. Nonparametric statistical models add flexibility by allowing the entire function $f \in \mathcal{F}$ to be arbitrary. Here f is the parameter and Bayesian nonparametrics attempts to put a prior $\mathcal{P}(.)$ on the space of probability distributions $\mathcal{F}$.

First developments in Bayesian nonparametric methods were presented by Freedman [1963] and Fabius [1964]. After further theoretical developments in the 1960's and 1970's (see, e.g., Kraft [1964], Kraft and van Eeden [1964], Ferguson [1973, 1974], Antoniak [1974]), applications of Bayesian nonparametric methods became widespread in the 1990's, following developments in computational sampling methods such as the Gibbs sampler [Gelfand and Smith, 1990, Casella and George, 1992] and the Metropolis-Hastings algorithm [Tierney, 1994], which allowed flexible posterior simulation for complex models. Possibly the most popular method for nonparametric Bayesian modeling has been the Dirichlet process (DP). Polya trees, a generalization of DPs, have been slightly less common in applications. Other nonparametric priors include Pitman-Yor processes [Pitman and Yor, 1997], gamma processes [Kalbfleisch, 1978], extended gamma processes [Dykstra and Laud, 1981], and beta processes [Hjort, 1990]. For an overview of Bayesian nonparametric methods, see Gelfand [1999] and Walker et al. [1999]; for an overview of their applications to common inference problems, see Müller and Quintana [2004].

In the following sections, Dirichlet processes and Polya trees are explained in detail. Polya trees are the distribution of choice for method development in this dissertation, and Dirichlet process models will be used as an alternative model in an application presented in Chapter 3.

1.1.1 Dirichlet processes

The Dirichlet process (DP) and its properties were introduced by Ferguson [1973]:

Definition 1.1. *Let $\alpha > 0$ be a scalar and G_0 a probability measure. A random probability measure G on the space Ω is said to have a Dirichlet process prior with parameter αG_0, written $G \sim DP(\alpha G_0)$, if for any finite measurable partition $(A_1, \ldots, A_k)$ of Ω, the random vector $\big(G(A_1), \ldots, G(A_k)\big)$ has a Dirichlet distribution with parameter $\big(\alpha G_0(A_1), \ldots, \alpha G_0(A_k)\big)$.*

G_0 is the base measure, or centering distribution, of the Dirichlet process, and the weight parameter α gauges the variability of G around G_0. For fixed sample size, with increasing values of α, G is forced to follow the shape of G_0 more closely. Sampling in DP models is facilitated by the fact that DPs are conjugate priors: if $G \sim DP(\alpha G_0)$, the posterior distribution $G|y$ upon observing data $y = (y_1, \ldots, y_n)$ is $DP\big((\alpha + n)G_0^*\big)$, where $G_0^* = \alpha(\alpha + n)^{-1} G_0 + (\alpha + n)^{-1} \sum_{i=1}^{n} \delta(y_i)$ and $\delta(y_i)$ denotes the measure giving mass one to the point y_i.

A Dirichlet process gives probability 1 to the set of discrete distributions. To obtain continuous distributions and avoid issues that could arise from misspecification of the base measure, G_0 may be defined as coming from a parametric family of distributions $\{G_\theta\}_\theta$. By defining a prior $p_\theta(\theta)$ and $G|\alpha, G_\theta \sim DP(\alpha G_\theta)$, a mixture of DPs is obtained, where marginally $G \sim \int DP(\alpha G_\theta) p_\theta(d\theta)$ [Antoniak, 1974].

A more popular alternative to a mixture of DPs is a Dirichlet process mixture (DPM) model. The nonparametric distribution G is then defined as coming from a mixture of parametric distributions, where the mixing distribution is a DP: $G(\cdot) \sim \int G_\theta(\cdot) dF(\theta)$, where the kernel G_θ is a parametric distribution and $F|\alpha, F_0 \sim DP(\alpha F_0)$ [Hanson et al., 2005]. This construction results in a continuous G with probability 1 as long as the kernel function G_θ is continuous.

Escobar [1994] and Escobar and West [1995] develop a Gibbs sampler algorithm

for posterior computation for DPMs without explicitly drawing posterior iterates of G. Further computational developments for DPMs were presented, e.g., by Bush and MacEachern [1996], MacEachern and Müller [1998], and Neal [2000].

Inferences about G and functionals thereof may be of interest in certain applications. Explicit sampling is simplified by an alternative, constructive representation of the Dirichlet process, which was introduced by Sethuraman [1994]. Let $G = \sum_{i=1}^{\infty} p_i \delta(\theta_i)$, where the vectors $\theta = (\theta_1, \theta_2, \ldots)$ and $p = (p_1, p_2, \ldots)$ are independent, the distribution of the θ_i's is that of an independent, identically distributed sample from G_0, and $p_i = v_i \prod_{j=1}^{i-1}(1 - v_j)$, $v_i \overset{iid}{\sim} \text{beta}(1, \alpha)$. Then $G(\cdot) \sim DP(\alpha G_0)$. Gelfand and Kottas [2002] use this representation to develop a computational approach that samples from the posterior distribution of G and therefore allows for inferences about G.

Dirichlet processes and mixtures have been employed in a variety of data analysis problem, for example in semiparametric median regression models [Kottas and Gelfand, 2011], to model random effects distributions [Bush and MacEachern, 1996, Kleinman and Ibrahim, 1998], survival analysis [Kuo and Mallick, 1997, Pennell and Dunson, 2006], and to evaluate goodness of fit of parametric distributions [Carota and Parmigiani, 1998, Viele, 2007].

1.1.2 Polya trees

Polya trees (PT) were introduced by Ferguson [1974], and Lavine [1992, 1994] as well as Mauldin et al. [1992] gave an overview of their definition and properties. To define a Polya tree, let $e_j(k)$ be the j-fold binary representation of the number $k-1$. Let Ω be a separable measureable space, and define a separating binary tree of partitions of Ω such that for every level $j = 1, 2, \ldots$ of the tree, the collection $\{B(j,k) : k = 1, \ldots, 2^j\}$ partitions Ω such that $\Omega = B(1,1) \cup B(1,2), B(1,1) \cap B(1,2) = \varnothing$, and for all $j = 1, 2, \ldots, B(j,k) = B(j+1, 2k-1) \cup B(j+1, 2k)$, and $B(j+1, 2k-1) \cap B(j+1, 2k) = \varnothing$. Further, let $\Pi = \{B(j,k) : j = 1, 2, \ldots; k = 1, \ldots, 2^j\}$, i.e., the set of partitioning

sets.

Definition 1.2. *A random probability measure G on Ω is said to have a Polya tree distribution, or a Polya tree prior, with parameter $(\Pi, \mathcal{A})$, written $G \sim PT(\Pi, \mathcal{A})$, if there exist nonnegative numbers $\mathcal{A} = \{\alpha_{j,k} : j = 1, 2, \ldots; k = 1, \ldots, 2^j\}$ and random variables $\mathcal{Y} = \{Y_{e_j(k)} : j = 1, 2, \ldots; k = 1, \ldots, 2^j\}$ such that the following hold:*

1. *all random pairs $(Y_{e_j(2k-1)}, Y_{e_j(2k)})$ in $\mathcal{Y}$ are independent;*
2. *for every $j = 1, 2, \ldots, k = 1, \ldots, 2^{j-1}$, $Y_{e_j(2k-1)} \sim beta(\alpha_{j,2k-1}, \alpha_{j,2k})$, and $Y_{e_j(2k)} = 1 - Y_{e_j(2k-1)}$;*
3. *for every $j = 1, 2, \ldots$ and every $k = 1, \ldots, 2^j$, $G\big(B(j,k)\big) = \prod_{i=1}^{j} Y_{e_j(\lceil k2^{i-j} \rceil)}$.*

Polya trees fall into the more broad category of tail-free processes [Freedman, 1963]. A tail-free process is defined analogously to Definition 1.2, with the generalization that there is no specific distribution imposed on $Y_{e_j(k)}$ [Ferguson, 1974]. A Dirichlet process is a special case of a Polya tree that is attained if for every j and k, $\alpha_{j,k} = \alpha_{j+1,2k-1} + \alpha_{j+1,2k}$.

Figure 1.1 visualizes the idea of the construction of a Polya tree for the sample space $\Omega = (0, 1]$. The PT prior is defined by a sequence of binary partitions on the sample space and conditional branch probabilities $Y_{e_j(k)}$. At each level j, the probability of any set $B(j,k)$ is defined as the product of all conditional branch probabilities along the path leading from the top node of the tree to that set. By defining a distribution on the branch probabilities, a distribution on $G\big(B(j,k)\big)$ is induced.

In applications, the partitions in Π are induced by "centering" G on a fixed distribution G_0. To do this, the sets $B(j,k)$ are defined as the intervals $\big(G_0^{-1}((k-1)/2^j), G_0^{-1}(k/2^j)\big]$, for $j = 1, 2, \ldots; k = 1, \ldots, 2^j$. Partitions induced in this way will be denoted by Π_0. We further choose $\alpha_{j,2k-1} = \alpha_{j,2k}, \forall j = 1, 2, \ldots; k = 1, \ldots, 2^j$.

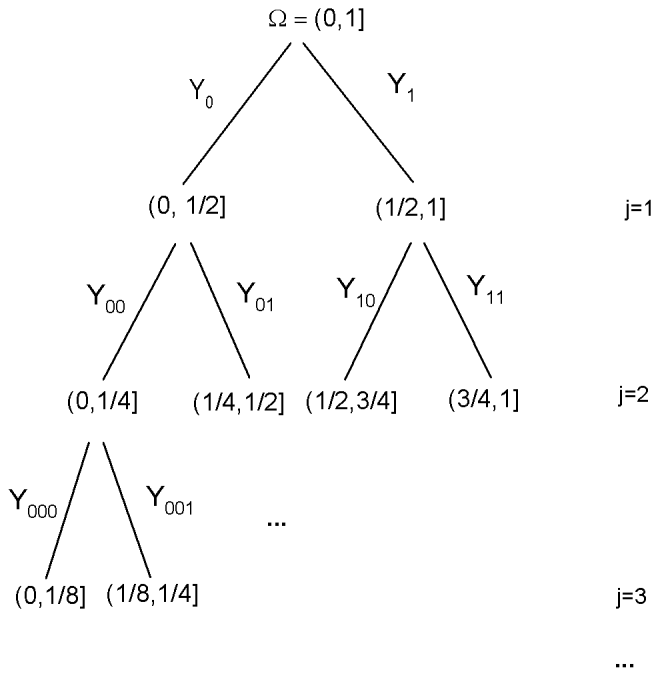


Figure 1.1: Schematic of the construction of a Polya tree

With these selections, the prior distributions on $Y_{e_j(k)}$ are symmetric around 1/2 and the prior mean of $G\big(B(j,k)\big)$ is $E\big[G\big(B(j,k)\big)\big] = \prod_{i=1}^{j} E[Y_{e_j(\lceil k2^{i-j}\rceil)}] = 1/2^j = G_0\big(B(j,k)\big)$ by the independence of the $Y_{e_j(k)}$'s, where the expectation is with respect to the PT distribution. Therefore, we write $E[G] = G_0$.

Lavine [1992] outlines three aspects of Polya trees that are affected by the choice of $\mathcal{A}$. First, the $\alpha_{j,k}$'s control the rate at which the updated predictive distribution changes from the prior distribution to the distribution of the sample. For large $\alpha_{j,k}$'s the predictive distribution is close to G_0, while for small $\alpha_{j,k}$'s its shape is mainly determined by the empirical distribution function of the data. Second, $\alpha_{j,k}$ affects the smoothness of G. For instance, choosing $\alpha_{j,k} = j^2$ yields a prior on the space of absolutely continuous distributions with probability one [Kraft, 1964, Ferguson, 1974, p. 621]. Finally, the $\alpha_{j,k}$'s impact the extent to which random G can vary around its prior mean G_0. In particular, larger $\alpha_{j,k}$'s allow for less variability of G about its mean. A common choice is $\alpha_{j,k} = c\rho(j)$, where $c > 0$ is fixed and $\rho(j)$ is an increasing, positive function, as used, e.g., in Berger and Guglielmi [2001], Walker and Mallick [1999], among many others. Alternatively, a prior distribution on c could be introduced. For Polya trees centered around a distribution G_0 with $\alpha_{j,k} = c\rho(j)$, we will use the notation $PT(c, \rho(.), G_0)$.

Polya trees are conjugate priors [Ferguson, 1974], which follows from the conjugacy of the beta priors defined on the branch probabilities $Y_{e_j(k)}$. Specifically, if $G \sim PT(\Pi_0, \mathcal{A})$ and $y = (y_1, \ldots, y_n)$, where $y_i|G \overset{iid}{\sim} G$, then upon observing data y, the posterior $G|y \sim PT(\Pi_0, \mathcal{A}|y) = PT(\Pi_0, \mathcal{A}^*)$ with $\mathcal{A}^* = \{\alpha^*_{j,k} = \alpha_{j,k} + n(j,k,y)\}$, where $n(j,k,y)$ is the number of observations in y that fall into set $B(j,k)$.

A simple PT is characterized by an infinite number of parameters, the branch probabilities. In practice, fitting PT models is done computationally by either marginalization or truncation to a finite tree. A finite Polya tree is a PT truncated at a fixed level J. The resulting prior is no longer nonparametric in the sense that it has an

infinite number of parameters, but rather richly parametric (i.e., it has a large, but finite, number of parameters).

Let $\Pi^J = \{\{B(j,k)\} : j = 1, \ldots, J; k = 1, \ldots, 2^j\}$. A finite Polya tree is defined as follows:

Definition 1.3. *A random probability measure G on Ω is said to have a finite Polya tree prior with parameter $(\Pi^J, \mathcal{A}^J)$, written $G \sim FPT(\Pi^J, \mathcal{A}^J)$, if there exist non-negative numbers $\mathcal{A}^J = \{\alpha_{j,k} : j = 1, \ldots, J; k = 1, \ldots, 2^j\}$ and random variables $\mathcal{Y} = \{Y_{e_j(k)} : j = 1, \ldots, J; k = 1, \ldots, 2^j\}$ such that the following hold:*

1. *all random pairs $(Y_{e_j(2k-1)}, Y_{e_j(2k)})$ in $\mathcal{Y}$ are independent;*
2. *for every $j = 1, \ldots, J, k = 1, \ldots, 2^{j-1}$, $Y_{e_j(2k-1)} \sim beta(\alpha_{j,2k-1}, \alpha_{j,2k})$, and $Y_{e_j(2k)} = 1 - Y_{e_j(2k-1)}$;*
3. *for every $j = 1, \ldots, J$ and every $k = 1, \ldots, 2^j$, $G\big(B(j,k)\big) = \prod_{i=1}^{j} Y_{e_j(\lceil k2^{i-j} \rceil)}$.*
4. *On sets $B(J,k)$, G follows G_0.*

The predictive Polya tree density for a future observation y_{n+1} that is obtained upon observing data y is

$$g(y_{n+1}|y) = g_0(y_{n+1}) 2^J \prod_{j=1}^{J} \frac{cj^2 + n\big(j, k(j, y_{n+1}), y\big)}{2cj^2 + n\big(j-1, k(j-1, y_{n+1}), y\big)}$$

where $k(j, y_{n+1})$ is the partition at level j into which y_{n+1} falls. Lavine [1994] shows that the updated predictive density g for $J \to \infty$ can be bounded above and that by truncating the Polya tree at a finite level J, $g(y_{n+1}|y)$ can be estimated within a factor $\delta \leq \exp(\frac{n}{2} \sum_{j=J}^{\infty} j^{-2})$. Hanson and Johnson [2002] show that Condition 4 in Definition 1.3 leads to predictive distributions that are exact if J is chosen to be sufficiently large, in the sense that in any partition into which no elements of y fall, the predictive density from a finite PT is exactly the same as from an infinite PT.

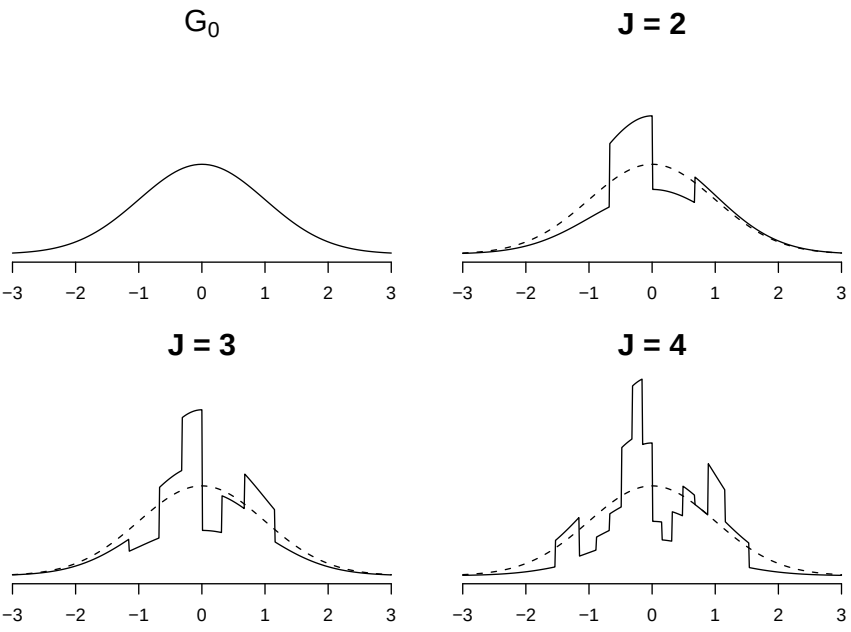


Figure 1.2: Possible posterior density estimates from a finite PT distribution for various levels J

They suggest the rule of thumb of choosing $J \doteq \log_2 n$, which allows for a more detailed estimation of G when more data are available. This choice of J derives from the prior expectation that at least one observation should fall into each set at level J.

Figure 1.2 visualizes what the densities of posterior iterates from a finite PT might look like for levels $J = 2, 3$ and 4 when $G_0 = N(0, 1)$. With increasing J, the shape of the density becomes more flexible and is able to capture any arbitrary distribution found in data. Condition 4 in Definition 1.3 ensures that in the case that all branch probabilities $Y_{e_j(k)}$ are equal to 0.5, the centering distribution G_0 is obtained.

The densities in Figure 1.2 also exemplify a problem that naturally arises from finite Polya trees: iterates from the posterior Polya tree distribution are necessarily discontinuous at the partition points $G_0^{-1}(k/2^j)$. Paddock et al. [2003] address this problem by proposing a randomized Polya tree, which adds random jitter to the partition points. An second issue with simple Polya tree arises with a choice of G_0 that puts a lot of prior mass on an interval of Ω in which little or no data occur. This

results in all the posterior mass in the tails of G_0 where sets $B(j,k)$ are larger and densities are thus fit with less precision and convergence of the posterior in sampling algorithms will be very slow [Barron et al., 1999].

Employing a mixture of Polya trees rather than a simple Polya tree avoids issues arising from having to choose a single centering distribution G_0. Here, we replace G_0 with G_θ and place a prior distribution on θ. In posterior calculations, the sample distribution now informs the choice of the centering distribution, avoiding the problem of a bad choice of G_0. Additionally, with varying θ, the partitions in Π change at each step of the Gibbs sampler, and mixing over the different partitions results in a smoother (differentiable) predictive density [Hanson, 2006].

The general mixture of Polya trees model is

$$G|\theta \sim PT(\Pi_\theta, \mathcal{A})$$
$$\theta \sim p_\theta(\theta)$$

where now $\Pi_\theta^J = \{B_\theta(j,k) = (G_\theta^{-1}((k-1)2^{-j}), G_\theta^{-1}(k2^{-j})) : j = 1,\ldots,J, k = 1,\ldots,2^j\}$.

We can define a mixture of finite Polya trees analogously by truncating the tree at a fixed level $J < \infty$:

$$G|\theta \sim PT(\Pi_\theta^J, \mathcal{A}^J)$$
$$\theta \sim p_\theta(\theta)$$

Empirical studies have shown that the particular choice of J affects results only slightly or not at all. As a result, (mixtures of) finite Polya trees are the model most used in PT applications, and they will be used throughout this dissertation.

Polya trees have been used in a variety of data analysis problems. Applications include nonparametric error distributions in regression models [Hanson and Johnson,

2002, Hanson, 2006], and distributions of mixed effects in hierarchical generalized linear models as presented in Walker and Mallick [1997]. Polya trees have also been employed in analysis of survival data [Walker and Mallick, 1997, 1999, Hanson, 2006, Zhao et al., 2009], nonparametric meta-analysis [Branscum and Hanson, 2008], time series [Denison and Mallick, 2006], and modeling ROC curves [Hanson et al., 2008]. Applications that involved testing a parametric model versus a nonparametric alternative have been presented in Berger and Guglielmi [2001] and Hanson [2006]. Multivariate versions of Polya trees have been developed in Hanson [2006], Yang et al. [2008], Trippa et al. [2011] and Hanson et al. [2011], of which the latter proposes an efficient approximate sampling algorithm for the complex model.

1.1.3 Fitting Polya tree models

To outline computational aspects of fitting Polya tree models, we first introduce some additional notation. Let $n_\theta(j,k,y)$ be the number of elements in the data vector y that fall into set $B_\theta(j,k)$, and let $k_\theta(j,y_i) \in \{1,\ldots,2^j\}$ identify the set at level j into which observation y_i falls.

The partition cut points at level j are $\{G_\theta^{-1}(k/2^j)\}_{k=1}^{2^j-1}$, from which we obtain the following computational formulas [Hanson, 2006]:

$$n_\theta(j,k,y) = \sum_{i=1}^{n} I\{\lfloor 2^j G_\theta(y_i)\rfloor = k-1\}$$

$$k_\theta(j,y_i) = \lfloor 2^j G_\theta(y_i)\rfloor + 1$$

We take c and $\rho(.)$ to be fixed (usually at $c=1$ or smaller for moderate sample sizes, and $\rho(j)=j^2$). G is completely defined by $\mathcal{Y}$ and θ and $G[B_\theta(j,k)|\mathcal{Y},\theta] = \prod_{i=1}^{j} Y_{e_i(\lceil k2^{i-j}\rceil)}$. We will define $p_{\mathcal{Y}}(k)$ as the probability of the k-th partition on the lowest level (J) of a finite tree:

$$p_{\mathcal{Y}}(k) = G[B_\theta(J,k)|\mathcal{Y},\theta] = \prod_{j=1}^{J} Y_{e_j(\lceil k2^{j-J}\rceil)}$$

The cumulative distribution function $G(y|\mathcal{Y},\theta)$ is given by

$$G(y|\mathcal{Y},\theta) = \sum_{k=1}^{k_\theta(J,y)-1} p_{\mathcal{Y}}(k) + p_{\mathcal{Y}}(k_\theta(J,y))[2^J G_\theta(y) - k_\theta(J,y) + 1]. \tag{1.1}$$

The corresponding density function is

$$g(y|\mathcal{Y},\theta) = 2^J p_{\mathcal{Y}}(k_\theta(J,y)) g_\theta(y). \tag{1.2}$$

Sampling from the posterior Polya tree distribution in a Gibbs sampler for a general model is straightforward. The likelihood function is calculated using a form of (1.2) and the current iterates of the branch probabilities. After drawing samples from the full conditional distributions of each of the other model parameters, a new set of branch probabilities $\mathcal{Y}^{(i)}$ is generated as a random draw from the updated beta distribution of each branch probability.

At the same time, explicit estimation of both (1.1) and (1.2), as well as functionals of G, is possible. For example, the q^{th} quantile of G can be estimated as

$$G^{-1}(q|\mathcal{Y},\theta) = G_\theta^{-1}\Big\{\frac{q - \sum_{k=1}^{K} p_{\mathcal{Y}}(k) + K p_{\mathcal{Y}}(K)}{2^J p_{\mathcal{Y}}(K)}\Big\},$$

where K is such that $\sum_{k=1}^{K-1} p_{\mathcal{Y}}(k) < q \leq \sum_{k=1}^{K} p_{\mathcal{Y}}(k)$.

Figure 1.3 shows samples from the posterior distribution $G|y$ for a sample y from a bimodal distribution. The observations were generated by selecting the $(i-0.5)/100$ quantiles for $i = 1,\ldots,100$ from the mixture $(0.5N(5,1) + 0.5N(13,1))$. To these

data, we fit the mixture of finite PTs model

$$y_i|G \overset{iid}{\sim} G$$
$$G|(\mu,\sigma) \sim FPT(c, j^2, N(\mu,\sigma))$$
$$(\mu,\sigma) \sim N(0,100) \times \Gamma(2,2)$$

truncating the tree at $J = 4$. To show the effect of the scale parameter, c was fixed at 0.1, 1, 5 and 10.

We plotted 50 samples from the posterior PT distribution, randomly chosen from 10,000 iterations after a burn-in period of 1,000 iterations. It becomes clear that for smaller values of c the posterior distribution more closely follows the empirical distribution of the data and the samples have greater variability. For larger c, samples from the posterior are more concentrated around the normal centering distribution, and at the same time the functions are smoother. Figure 1.4 graphs estimated distribution functions $G|y$ from the mixture of Polya trees and as expected, for smaller values of c, $G|y$ is able to capture the bimodality of the distribution more closely.

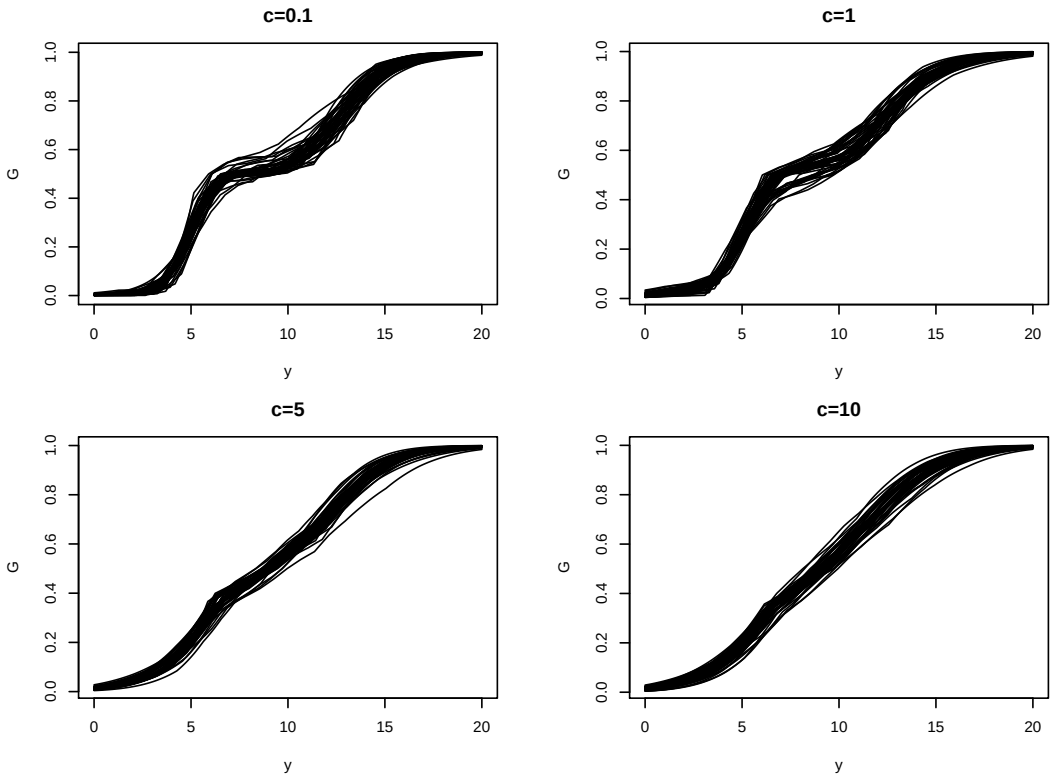


Figure 1.3: Posterior samples from a mixture of Polya trees for bimodal data ($n = 100$) with a normal centering distribution for $c = 0.1, 1, 5$ and 10, $J = 4$

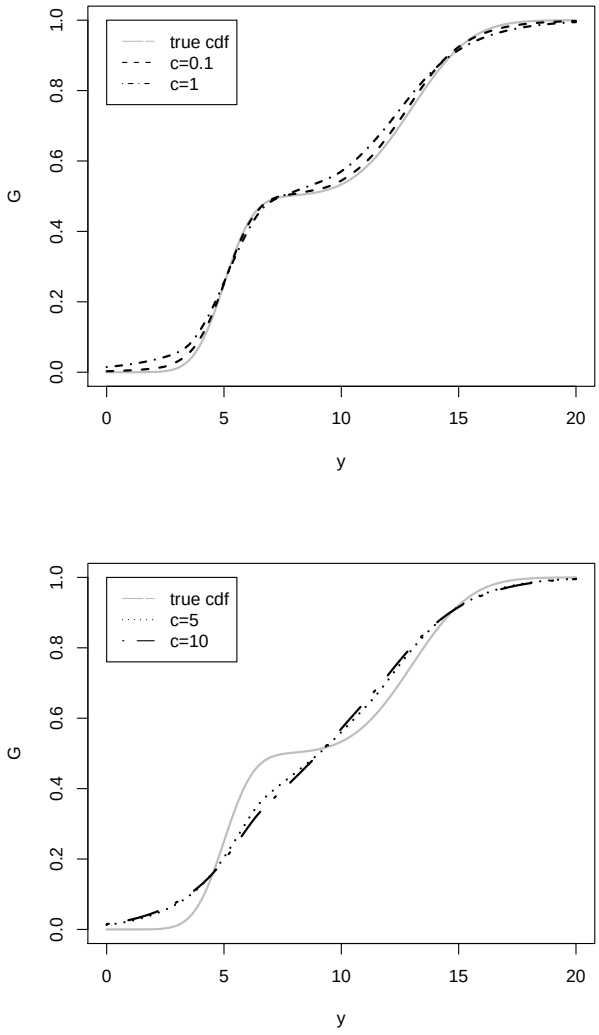


Figure 1.4: Estimated distributions from a mixture of Polya trees for bimodal data ($n = 100$) with a normal centering distribution for $c = 0.1, 1, 5$ and 10, $J = 4$

1.2 Bayes factors

Bayes factors (BFs) are a method of comparing competing models or hypotheses in the Bayesian framework. In general, the Bayes factor comparing two hypotheses H_0 and H_1 is [Kass and Raftery, 1995]

$$BF = \frac{Pr(y|H_1)}{Pr(y|H_0)} = \frac{Pr(H_1|y)p(H_0)}{Pr(H_0|y)p(H_1)}. \tag{1.3}$$

In the non- or semiparametric context, Bayes factors have been employed to test parametric goodness of fit, generally by nesting the parametric model within a more general nonparametric alternative. In a sense the Bayes factor measures how strongly the data support or contradict the parametric model. Testing for goodness of fit has been proposed for various nonparametric prior families, such as Dirichlet process mixtures [Carota and Parmigiani, 1996, Basu and Chib, 2003], Polya trees [Ghosal et al., 1998, Berger and Guglielmi, 2001, Hanson, 2006], and Gaussian process priors [Verdinelli and Wasserman, 1998].

Gelfand and Dey [1994] provide a discussion of asymptotic behavior and calculations for Bayes factors in the case that the two hypotheses are parametric models. For comparing parametric priors to a Polya tree alternative, conditions for the consistency of Bayes factors have been presented by Ghosal et al. [1999], Dass and Lee [2004] and McVinish et al. [2009]. Ghosal et al. [2008] give general sufficient conditions for consistency of the BF for nonparametric hierarchical priors. For similar model comparisons for continuous data using Dirichlet processes, problems of inconsistency of the BF for some models are discussed by Berger and Guglielmi [2001] and Carota [2006].

Table 1.1 lists the cutoff values as suggested by Jeffreys [1961] for rejecting H_0 based on values of the Bayes factor. The table also assigns numbers to each of the categories, which will be used in evaluating test performance in a later chapter. Note

Table 1.1: Cutoff values for Bayes factors according to Jeffreys [1961, Appendix B] with categories 0-5 assigned for later reference

$\log_{10}(BF)$	BF	Evidence against H_0	Category
< 0	< 1	no evidence	0
0 - 0.5	1 - 3.2	barely worth mentioning	1
0.5 - 1	3.2 - 10	substantial	2
1 - 1.5	10 - 32	strong	3
1.5 - 2	32 - 100	very strong	4
> 2	> 100	decisive	5

that in the Bayesian framework, the two hypotheses are interchangeable, and the ratio in (1.3) can be reversed and the Bayes factor may be interpreted as evidence against H_1 or evidence *for* H_0.

Alternative methods for Bayesian model choice include, for example, the deviance information criterion (DIC) [Spiegelhalter et al., 2002], posterior predictive p-values [Gelman et al., 1996] or distance measures [Goutis and Robert, 1997]. For example, in a nonparametric setting, the Kullback-Leibler distance has been employed to measure the distance between the prior and the posterior distribution [Carota et al., 1996, Carota and Parmigiani, 1998], or the distance between a parametric family and the distribution that generated the data [Viele, 2007].

In settings in which prediction of individual observations is of interest, one measure for model selection among models M_k is based on the conditional predictive ordinate $\text{CPO}_i = f_i(y_i|y_{(-i)}, M_k)$ proposed by Geisser and Eddy [1979] and Geisser [1980], where $y_{(-i)}$ are the data with the i^{th} observation omitted. Gelfand and Dey [1994] utilize the CPO to calculate a pseudo Bayes factor based on a simple sampling approach. They propose estimating the CPO based on MC Gibbs-sampler iterates $\psi^{(m)}$ from the posterior distribution of the parameter vector ψ

$$\hat{f}(y_i|y_{(-i)}, M_k) = E^{-1}_{M_k|y}\Big\{\frac{1}{f(y_i|M_k)}\Big\} = \text{MC}\sum_{m=1}^{\text{MC}}\Big\{\frac{1}{f(y_i|\psi^{(m)}, M_k)}\Big\}^{-1}.$$

The log-pseudo marginal likelihood (LPML) for a given model is then defined as

$$\text{LPML}_{M_k} = \log \prod_{i=1}^{n} \text{CPO}_i.$$

The pseudo Bayes factor $PSBF$ for comparing two models is

$$\text{PSBF} = \exp(\text{LPML}_{M_2} - \text{LPML}_{M_1}),$$

which can then be interpreted analogously to Jeffrey's categories for Bayes factors.

1.3 Dissertation outline

The remainder of this dissertation presents two Polya tree models in regression settings. In Chapter 2, a Bayesian semiparametric model for risk regression with continuous response data is proposed. The method includes an Empirical Bayes test procedure for evaluating goodness of fit of a parametric residual distribution. Both theoretical and computational results about the performance of the test and risk estimation procedure are presented.

Chapter 3 presents a novel approach for defining dependent priors on function spaces. In a regression setting, this method models nonparametric error distributions across ordinal covariates flexibly while allowing dependencies between errors at different covariate values.

Appendix A.1 contains a summary of notation. Symbols defined for Polya trees are summarized in Table A.1. Additionally, notation for parametric distributions used in this dissertation is outlined in Table A.2.

Chapter 2 Bayesian Semiparametric Risk Regression with Measurement Data

Logistic regression models are a popular tool for risk estimation in medical and biological data analysis. With continuous response (i.e., measurement) data, it is common to create a dichotomous outcome by specifying a threshold for positivity. Fitting a linear regression via least squares to the original, non-dichotomized response assuming a logistic error distribution has previously been shown to yield more efficient estimators of odds ratios than ordinary logistic regression of the dichotomized endpoint. This chapter develops a novel test for assessing goodness of fit of logistic regression based on a Bayesian semiparametric Polya tree model.

Bayes factors are calculated using the Savage-Dickey ratio for testing the null hypothesis of logistic regression versus a semiparametric generalization. The proposed empirical Bayes approach is computationally efficient since it does not require MCMC sampling, and we show that results from it are equivalent to results from a fully Bayesian implementation for large sample sizes. A method for semiparametric estimation of risks, risk ratios, and odds ratios is developed, which can be employed when the hypothesis of a logistic error distribution is rejected.

2.1 Introduction

In the context of medical or public health research, interest often lies in quantifying the risk of adverse outcomes and identifying at-risk subpopulations. Although outcomes may be communicated as binary, they are often defined based on an underlying continuous variable. The actual endpoint of interest may not be directly observable because procedures to precisely determine a patient's status are invasive or even destructive. In such a case, biomarkers or other variables may serve as surrogate measures. As an example, the gold-standard for determining lower than normal bone turnover to identify renal osteodystrophy in patients with chronic kidney disease

is an invasive and time consuming bone biopsy, the classification of which should be made only by highly trained experts. As an alternative, clinicians may use levels of the parathyroid hormone (PTH), of which levels below 150 in a particular PTH assay indicate low bone turnover [Malluche and Monier-Faugere, 2006].

The metabolic syndrome *diabetes mellitus* is marked by elevated blood sugar levels and glucose intolerance due to insulin deficiency or impaired effectiveness of insulin action [Zimmet et al., 2004]. For the purpose of individual diagnosis of diabetes, multiple testing and other criteria would be considered by the diagnosing clinician, however for epidemiologic purposes testing is rarely repeated and fasting plasma glucose measures are most commonly used to identify a subpopulation with diabetes. A person with fasting plasma glucose level at or above 126mg/dL is considered diabetic [WHO06].

In other scenarios, the endpoint of interest is directly defined based on a threshold for a continuous variable. For example, the classification of an overweight or obese individual is generally based on the body mass index, which is a continuous variable calculated from a person's height and weight. A person with a BMI of $30kg/m^2$ or above is considered obese, while the cutoff for considering a person overweight is $25kg/m^2$. This classification is of epidemiologic interest, as overweight and obesity are risk factors for other diseases such as diabetes and heart disease.

In a multitude of scenarios, well-established thresholds are used by practitioners to make clinical diagnoses and treatment decisions, or by epidemiologic studies to quantify the health of subpopulations. Traditionally, risk assessment models for binary outcomes are built using logistic regression. If the outcome is based on an underlying continuous variable, much of the information contained in the original variable is lost by reducing it to a 0/1 outcome. From a statistical perspective, it would be preferable to retain all the information of the continuous response and create a model that predicts mean response. Not only will the loss of information result

in more uncertainty about the model parameters, but the adherence to rigid cutoffs classifies individuals into groups, with no measure of how different they are in terms of the original variable. For example, an individual with a BMI of 29.8 and another with a BMI of 30.2 are most likely very similar in terms of body fat percentage and other physical measures, but they are classified as not-obese and obese, respectively, the same way that two individuals with BMIs of 25 and 40 would be classified, who would undoubtedly have greater physiological dissimilarities.

However, if clinical diagnoses or epidemiologic characterizations are based on established, hard cutoff values, models of the mean continuous response may not directly address the clinical questions at hand [Ragland, 1992]. Additionally, risks and related measures are easier to interpret and communicate to clinicians, patients, policy makers and the general public. It is therefore desirable to retain all the information of a continuous response throughout the model-building process and then translate the model into risk inference for a binary outcome at the end of the analysis process.

Moser and Coombs [2004] show that by fitting a linear model via least squares with the original, non-dichotomized response variable assuming a logistic error distribution, risk and odds ratio parameters are equivalent to those under the ordinary logistic regression model for the dichotomized data. This connection has been employed and empirically confirmed, for example by Bakhshi et al. [2008]. Moreover, Moser and Coombs [2004] illustrate that large gains in efficiency are achieved by modeling the original continuous response data. Specifically, much smaller sample sizes are needed for the same power seen in ordinary logistic regression.

The connection between parameters of interest for continuous and binary logistic regression depends on the condition that data follow a logistic distribution. Sections 2.2 and 2.3 will demonstrate this equivalence and explore biases in estimates when the data distribution deviates from logistic. In Section 2.4, a Polya tree-based goodness of fit test for the parametric distributional assumption is proposed, and two com-

putational approaches are compared. An Empirical Bayes approach that provides a one-step estimation of the Savage-Dickey ratio is compared to a fully Bayesian MCMC sampling approach and theoretical results on consistency of the Bayes factor under the Empirical Bayes approach are presented. Methods for semiparametric estimation of risks, risk ratios, and odds ratios are presented in Section 2.5. Performance of the proposed method on simulated data and on survey data sets is evaluated in Sections 2.6 and 2.7, respectively.

2.2 Background

Dichotomizing a continuous outcome according to a cutoff d may arguably have some interpretative advantages. For statistical modeling and inference, however, reducing the information contained in a continuous variable to a binary outcome results in loss of efficiency, as explored theoretically by Selvin [1987] and demonstrated empirically by Moser and Coombs [2004] and Ragland [1992]. Also, building risk prediction models using measurement data does not preclude subsequent thresholding to aid in decision making. For instance, we can determine the predictive density of BMI for a certain type of person and base decisions on whether that density largely supports BMI values above 30.

Moser and Coombs [2004] investigated differences in statistical efficiency for logistic linear regression of measurement data and ordinary logistic regression of dichotomized data. They start with a standard linear model for continuous responses $y_i = x_i'\beta + \varepsilon_i, i = 1, \ldots, n$, where $x' = (1, x_1, \ldots, x_{p-1})$ and $\beta = (\beta_0, \ldots, \beta_{p-1})'$. If the residuals are independent, identically distributed and follow a logistic distribution with mean 0 and standard deviation σ, a natural connection between β, σ, odds ratios, and coefficients from logistic regression of dichotomized data arises.

The cumulative distribution function for a random variable Y that follows a lo-

gistic distribution with mean $x'\beta$ and standard deviation σ is

$$P(Y \leq d|x, \beta, \sigma) = \frac{1}{1 + \exp[-\lambda(d - x'\beta)/\sigma]} \tag{2.1}$$

where $\lambda = \pi/\sqrt{3}$. Letting $x_{(-1,j)} = (1, x_1, \ldots, x_j - 1, \ldots, x_{p-1})$, the odds ratio for the event $Y > d$ comparing individuals that differ by one unit on x_j but are otherwise the same, can be expressed as

$$\begin{aligned} OR_j &= \frac{P(Y > d|x, \beta, \sigma)/[1 - P(Y > d|x, \beta, \sigma)]}{P(Y > d|x_{(-1,j)}, \beta, \sigma)/[1 - P(Y > d|x_{(-1,j)}, \beta, \sigma)]} \\ &= \frac{\exp[\lambda(d - x'_{(-1,j)}\beta)/\sigma]}{\exp[\lambda(d - x'\beta)/\sigma]} \\ &= \exp(\lambda\beta_j/\sigma). \end{aligned}$$

Now consider the common alternative in which a dichotomized variable $Y_i^* = I(Y_i > d)$ is modeled by ordinary logistic regression with

$$P(Y > d|x, \phi) = P(Y^* = 1|x, \phi) = \frac{\exp(x'\phi)}{[1 + \exp(x'\phi)]}$$

where $\phi = (\phi_0, \ldots, \phi_{p-1})'$.

The odds ratio for the same effect under this model is

$$OR_j^* = \frac{P(Y^* = 1|x, \phi)/[1 - P(Y^* = 1|x, \phi)]}{P(Y^* = 1|x_{(-1,j)}, \phi)/[1 - P(Y^* = 1|x_{(-1,j)}, \phi)]}$$

$$= \frac{\exp(x'\phi)}{\exp(x'_{(-1,j)}\phi)}$$

$$= \exp(\phi_j).$$

Since the odds ratios OR_j and OR_j^* are defined equivalently, it follows that

$$OR_j = OR_j^* \Rightarrow \exp(\lambda\beta_j/\sigma) = \exp(\phi_j),$$

which leads to the following connection between regression coefficients from the two modeling approaches:

$$\lambda\beta_j/\sigma = \phi_j.$$

Therefore, we can derive statistical tests and estimates for the usual odds ratio corresponding to the effect of x_j for any cutoff d based on the least squares estimates for the regression model. Statistical inference for risks, risk ratios, and other related parameters are also available from this model.

Citing the similarity between the two distributions, Moser and Coombs [2004] substituted a normal for the logistic distribution on the residuals and applied standard linear model theory to determine approximate confidence interval formulas for odds ratios. Instead of applying normal theory to non-normal data, an alternative approach would generate asymptotic confidence intervals from theory for a logistic accelerated failure time (AFT) model [Hosmer and Lemeshow, 1999], which is commonly used in

the field of survival analysis. Numerical procedures for fitting AFT models are built into most statistical software packages, including R and SAS.

2.3 Impact of model misspecification

The direct connection between regression coefficients from the models for the dichotomized and the continuous response data, as well as the proportionality of the odds independent of the cutoff d, hinges upon the assumption of a logistic error distribution. Real data, however, very often do not meet parametric assumptions. The following demonstrations visualize the effect of deviations from the logistic error distribution on bias of risk and odds ratio estimates.

For the linear model with a single continuous covariate x_i and $y_i = \beta x_i + \varepsilon_i$, with $\varepsilon_i \sim \text{logistic}(0, \sigma)$, the log odds ratio is equal to $\lambda\beta/\sigma$ for any value of x and any cutoff d. For other error distributions, however, the odds ratio is no longer constant with x. To demonstrate this effect, 10,000 residuals for this simple linear model were simulated from three different distributions: the normal, skew-normal [Azzalini, 1985] and student-$t(3)$ distribution. All distributions were normalized to have mean 0 and standard deviation $\beta\lambda$, which in the case of a logistic distribution would result in a log odds ratio of 1. Values of x were generated as a sequence of evenly distributed values between -1 and 10. The log odds ratio for the events $Y > 4$ and $Y > 5$ were estimated using ordinary logistic regression and the AFT model with logistic error for $\beta = 0.5, 1,$ and 2.

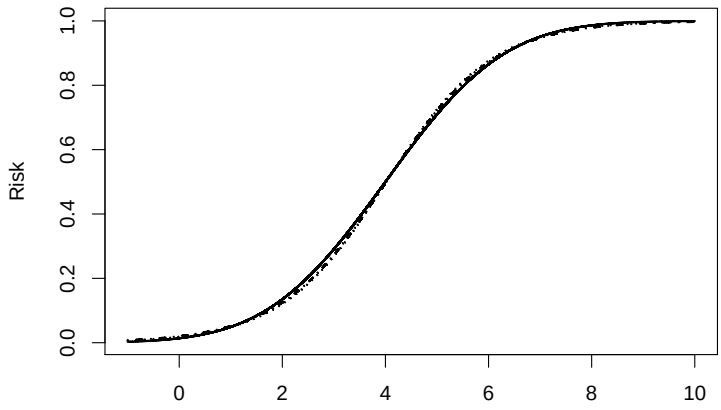
Figure 2.1 compares the true risks of $Y > 4$ to the estimated risks from the two parametric models for $\beta = 1$. In the case of the two symmetric distributions (Figures 2.1(a) and 2.1(c)) the risk estimates from the two parametric models are very similar, but do not model the shape of the risk function appropriately. The deviation in shape is less dramatic in the case of the normal distribution than for $t(3)$, because the normal is similar to the logistic distribution. For the skew-normal distribution (Figure 2.1(b)) the two models result in different estimates, and both models are not

able to capture the general shape of the asymmetric risk function.

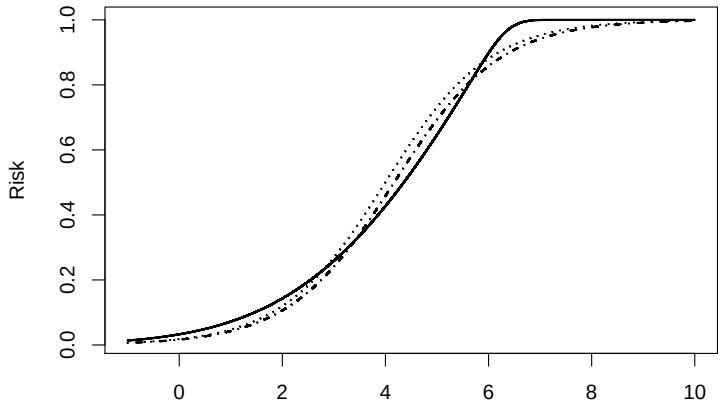
The inadequacy of the parametric models becomes even more evident when comparing true and estimated log odds ratios (Figure 2.2). The shapes of the true odds ratio functions cannot be captured by the estimates from either logistic regression or the linear model of the continuous response, as they estimate a constant close to the value of 1 (estimates of log odds ratio functions not presented). Note that the odds ratios are not properly modeled although the mean structure in this simple scenario is correctly specified, i.e., the shape of the log odds ratio function is solely due to the shape in the error distribution. The location of the true (log) odds ratio curves here is not independent of the cutoff. Changing the cutoff from $d = 4$ to $d = 5$ retains the shape of the curves, but shifts them along the x-axis. In addition, the shape of the functions changes with the error standard deviation, even though the ratio β/σ is held constant at λ.

Although we focus on logistic regression without cutoffs, alternative families of parametric distributions may be fit to the data. In the case of a normal error distribution, a standard normal model could be fit, and a direct link between its parameters and those from a probit regression of dichotomized data can be made. The relationship between the probability of the event $Y > d$ and linear predictors x in probit regression is modeled as $P(Y > d|x, \phi) = P(Y^* = 1|x, \phi) = \Phi(x'\phi)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal. Fitting the model $y_i = x_i'\beta + \varepsilon_i$ to continuous data with a normal error distribution results in the relationship $\phi_j = \beta_j/\sigma$.

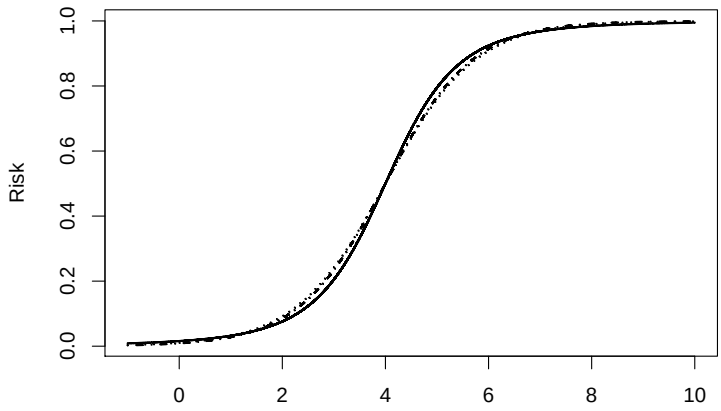
However, such models for continuous response data are not necessarily readily available for all parametric distributions, and this strategy fails if the correct family of parametric error distributions is not identified. For dichotomous response models, nonparametric generalizations of link functions have been proposed, for example, using mixtures of beta distributions [Mallick and Gelfand, 1996] or Polya trees [Hanson,



(a) Normal error

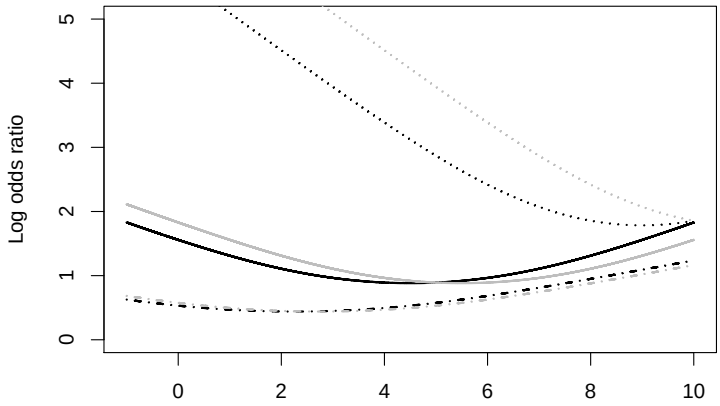


(b) Skew-normal error

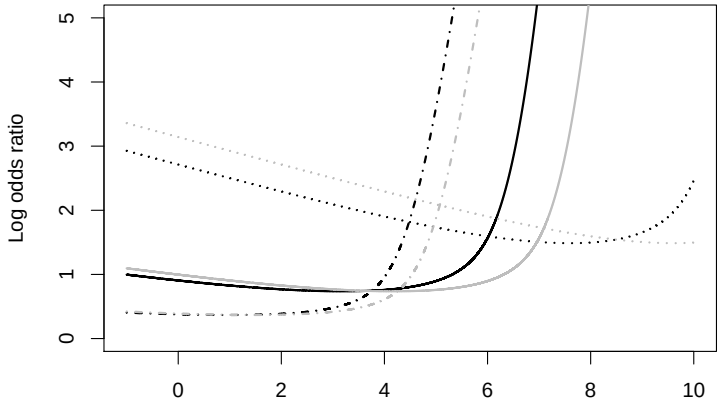


(c) Student-t(3) error

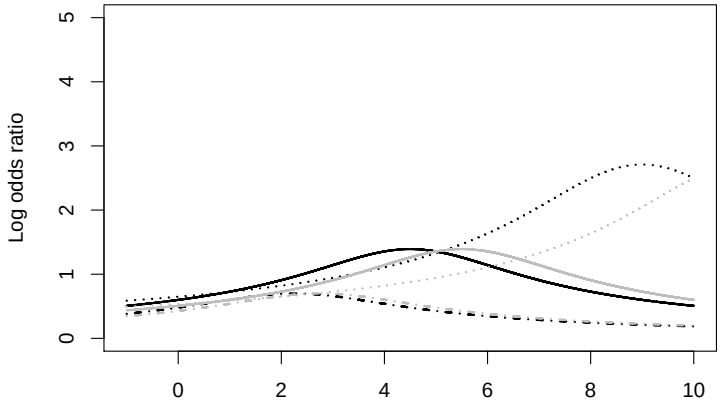
Figure 2.1: True risk (solid line) and estimated risk from ordinary logistic regression (dotted line) and AFT model of the continuous response (dot-dashed line) for three non-logistic error distributions.



(a) Normal error



(b) Skew-normal error



(c) Student-t(3) error

Figure 2.2: True log odds ratios for the events $Y > 4$ (black) and $Y > 5$ (grey), for $\beta = 0.5$ (dotted lines), $\beta = 1$ (solid lines) and $\beta = 2$ (dot-dashed line) and three non-logistic error distributions.

2006]. To retain the efficiency gains from modeling continuous response data and at the same time model risk and odds functions of arbitrary shape more appropriately, a new model based on nonparametric error distributions will be developed in the following section.

2.4 Testing for goodness of fit in logistic regression

We use Bayes factors to test whether the assumption of a logistic error distribution is violated in a linear model for measurement data. To this end, we embed a logistic regression in a semiparametric Polya tree model. Specifically, for $i = 1, \ldots, n$,

$$y_i = x_i'\beta + \sigma\varepsilon_i; \quad \varepsilon_i | G \overset{iid}{\sim} G; \quad G \sim PT(c, \rho(\cdot), G_0).$$

We refer to G as the residual distribution, although it is the distribution of the usual errors scaled by σ.

The Polya tree prior expectation of G is $G_0(y) = [1 + e^{-y\lambda}]^{-1}$, the logistic distribution with mean 0 and variance 1. To ensure identifiability of the intercept β_0, the median of G is fixed at 0 by setting the probabilities at the first level of the tree, namely Y_0 and Y_1, equal to 0.5. The underlying logistic distribution is obtained when $H_0 : \mathcal{Y} = \mathcal{Y}_0 \equiv 0.5$ is true, i.e., when all PT probabilities are equal to 0.5, so that $(Y_{e_j(2k-1)}, Y_{e_j(2k)}) = (0.5, 0.5), j = 1, \ldots, J; k = 1, \ldots, 2^{j-1}$.

To test the null hypothesis that the ε_i's follow a logistic distribution against a nonparametric alternative, we will employ the Savage-Dickey ratio [Verdinelli and Wasserman, 1995, Hanson, 2006]. The Savage-Dickey ratio gives the general form of a Bayes factor for testing nested hypotheses, and is used in this study for the particular case of a logistic distribution nested within a flexible alternative that is a generalization of logistic regression.

In the case of the particular hypotheses considered here, $\mathcal{Y}$ and $\theta = (\beta, \sigma)$ are assumed to be a priori independent, and we can assume that $p(\theta|H_0) = \int p(\theta, \mathcal{Y}|H_1)\, d\mathcal{Y}$,

which implies that $p(\theta|H_0) = p(\theta|H_1)$. Following some general results in Kass and Raftery [1995] and Verdinelli and Wasserman [1995], we can derive the Savage-Dickey ratio for this particular model.

Proposition 2.1. *The two conditions stated above are sufficient to simplify the Bayes factor $BF = Pr(y|H_1)/Pr(y|H_0)$ to the Savage-Dickey ratio [Kass and Raftery, 1995, Verdinelli and Wasserman, 1995]*

$$BF = \frac{p(\mathcal{Y}_0)}{p(\mathcal{Y}_0|y)}, \tag{2.2}$$

where $y = (y_1, \ldots, y_n)$.

Proof. [Verdinelli and Wasserman, 1995]

$$\begin{aligned}
BF = \frac{Pr(y|H_1)}{Pr(y|H_0)} &= \frac{\iint p(y|\mathcal{Y}, \theta) p(\mathcal{Y}, \theta)\, d\mathcal{Y}\, d\theta}{\int p(y|\mathcal{Y}_0, \theta) p_0(\theta)\, d\theta} \\
&= \frac{p(y)}{p(\mathcal{Y}_0|y)} \int \frac{p(\mathcal{Y}_0|y)}{p(y|\mathcal{Y}_0, \theta) p_0(\theta)}\, d\theta \\
&= \frac{p(y)}{p(\mathcal{Y}_0|y)} \int \frac{p(\mathcal{Y}_0|y) p(\theta|\mathcal{Y}_0, y)}{p(y|\mathcal{Y}_0, \theta) p_0(\theta) p(\theta|\mathcal{Y}_0, y)}\, d\theta \\
&= \frac{p(y)}{p(\mathcal{Y}_0|y)} \int \frac{p(\mathcal{Y}_0, \theta, y)/p(y)}{p(y|\mathcal{Y}_0, \theta) p_0(\theta) p(\theta|\mathcal{Y}_0, y)}\, d\theta \\
&= \frac{1}{p(\mathcal{Y}_0|y)} \int \frac{p(\mathcal{Y}_0, \theta)}{p_0(\theta) p(\theta|\mathcal{Y}_0, y)}\, d\theta \\
&= \frac{1}{p(\mathcal{Y}_0|y)} \int \frac{p(\mathcal{Y}_0) p(\theta)}{p(\theta) p(\theta|\mathcal{Y}_0, y)}\, d\theta
\end{aligned}$$

$$= \frac{p(\mathcal{Y}_0)}{p(\mathcal{Y}_0|y)} \int \frac{1}{p(\theta|\mathcal{Y}_0, y)}\, d\theta$$

$$= \frac{p(\mathcal{Y}_0)}{p(\mathcal{Y}_0|y)}$$

□

For fixed c, β and σ, the Savage-Dickey ratio is

$$BF = \frac{p(\mathcal{Y}_0)}{p(\mathcal{Y}_0|\bar{\varepsilon})} \tag{2.3}$$

where $\bar{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$,

$$p(\mathcal{Y}_0) = \prod_{j=1}^{J} \prod_{k=1}^{2^{j-1}} beta(0.5|c\rho(j), c\rho(j)) \tag{2.4}$$

is the joint prior density of all branching probabilities evaluated at 0.5, and

$$p(\mathcal{Y}_0|\bar{\varepsilon}) = \prod_{j=1}^{J} \prod_{k=1}^{2^{j-1}} beta(0.5|c\rho(j) + n(j, 2k-1, \bar{\varepsilon}), c\rho(j) + n(j, 2k, \bar{\varepsilon})) \tag{2.5}$$

is the joint posterior density of all branching probabilities evaluated at 0.5, given the residuals.

2.4.1 Empirical Bayes test

As a single-step approximation of the Savage-Dickey ratio in (2.3), least-squares estimates $\hat{\beta}$ and $\hat{\sigma}$ may be calculated as consistent and unbiased estimators of β and σ, which gives residuals $\hat{\varepsilon}_i = (y_i - x_i'\hat{\beta})/\hat{\sigma}$ that are substituted into (2.4) and (2.5). The

least-squares estimators are defined as

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$\hat{\sigma} = \sqrt{y'(I_n - X(X'X)^{-1}X')y/(n-p)},$$

where X is the $n \times p$ matrix $[x_1, \cdots, x_n]'$.

This gives an approximation to the Bayes factor that is much less computationally expensive than a traditional MCMC sampling approach for the fully Bayesian analysis described in Section 2.6.1, which involves sampling branching probabilities in $\mathcal{Y}$ and calculating Polya tree density estimates at each step of the Gibbs sampler. On the other hand, this single-step calculation has negligible computational requirements and, as shown in later sections, performs similar to a full MCMC approach.

2.4.2 Theoretical results

In the context of Bayes factors, consistency is defined as follows [McVinish et al., 2009], [Diaconis and Freedman, 1986]: The Bayes factor BF_n for testing $H_0 : f = f_0$ versus $H_1 : f \neq f_0$ is said to be consistent if for $f = f_0$, $\lim_{n\to\infty} BF_n = 0$, in probability, and for any $f \neq f_0$, $\lim_{n\to\infty} BF_n = \infty$, in probability. A stricter definition demands that convergence under both hypotheses be almost surely [Dass and Lee, 2004].

Consistency under H_0. The numerator of the Bayes factor defined in (2.3) is constant for fixed c and J, and only its denominator depends on the data y. Under the null hypothesis it is therefore sufficient to show that $1/p(\mathcal{Y}_0|\bar{\varepsilon})$ as defined in (2.5) converges to 0 with increasing sample size.

The consistency of $\hat{\beta}$ and $\hat{\sigma}$ under relatively weak assumptions has been established [Lai et al., 1978], even if the error distribution is misspecified [Gould and Lawless, 1988]. Considering only the left branch probabilities $Y_{e_j(2k-1)}$ of the Polya tree, it

follows from the conjugacy result that

$$Y_{e_j(2k-1)}|\bar{\varepsilon} \sim \text{beta}(cj^2 + n(j, 2k-1, \bar{\varepsilon}), cj^2 + n(j, 2k, \bar{\varepsilon})),$$

and therefore

$$E(Y_{e_j(2k-1)}|\bar{\varepsilon}) = \frac{cj^2 + n(j, 2k-1, \bar{\varepsilon})}{2cj^2 + n(j, 2k-1, \bar{\varepsilon}) + n(j, 2k, \bar{\varepsilon})}$$

as well as

$$\begin{aligned} Var(Y_{e_j(2k-1)}|\bar{\varepsilon}) = &\frac{[cj^2 + n(j, 2k-1, \bar{\varepsilon})][cj^2 + n(j, 2k, \bar{\varepsilon})]}{[2cj^2 + n(j, 2k-1, \bar{\varepsilon}) + n(j, 2k, \bar{\varepsilon})]^2} \\ &\times \frac{1}{[2cj^2 + n(j, 2k-1, \bar{\varepsilon}) + n(j, 2k, \bar{\varepsilon}) + 1]}. \end{aligned}$$

By the Strong Law of Large Numbers and the consistency of $\hat{\beta}$ and $\hat{\sigma}$, $n(j, k, \bar{\varepsilon})/n \xrightarrow{a.s.} P\big(\epsilon_i \in B(j, k)\big)$ as $n \to \infty$, for all j, k. Under H_0, $P\big(\epsilon_i \in B(j, k)\big) = 2^{-j}$ and therefore

$$\begin{aligned} E(Y_{e_j(2k-1)}|\bar{\varepsilon}) &= \frac{cj^2/n + n(j, 2k-1, \bar{\varepsilon})/n}{2cj^2/n + n(j, 2k-1, \bar{\varepsilon})/n + n(j, 2k, \bar{\varepsilon})/n} \\ &\xrightarrow{a.s.} \frac{0 + 2^{-j}}{0 + 2^{-j} + 2^{-j}} = 0.5 \end{aligned}$$

as $n \to \infty$. Additionally,

$$\begin{aligned} Var(Y_{e_j(2k-1)}|\bar{\varepsilon}) = &\frac{[cj^2/n + n(j, 2k-1, \bar{\varepsilon})/n][cj^2/n + n(j, 2k, \bar{\varepsilon})/n]}{[2cj^2/n + n(j, 2k-1, \bar{\varepsilon})/n + n(j, 2k, \bar{\varepsilon})/n]^2} \\ &\times \frac{1/n}{[2cj^2/n + n(j, 2k-1, \bar{\varepsilon})/n + n(j, 2k, \bar{\varepsilon})/n + 1/n]} \\ &\xrightarrow{a.s.} 0 \end{aligned}$$

as $n \to \infty$.

As a result, the posterior distribution of each $Y_{e_j(2k-1)}$ is consistent, i.e., deteriorates to point mass at 0.5 with $n \to \infty$. Therefore, expression (2.5) does indeed diverge to $+\infty$ and in turn $BF_n \overset{n\to\infty}{\longrightarrow} 0$ under H_0.

Consistency under H_1. The following proposition will be used to demonstrate consistency under the alternative hypothesis:

Proposition 2.2. *Let F and f be the true distribution and density functions generating the data and $B(J, k(J, x))$ be the set at level J of the Polya tree partition induced by some centering distribution G, into which observation x falls. Then, $J \log 2 \geq -\int f(x) \log F\big(B(J, k(J, x))\big)dx$, with equality only if $f(x) = g_0(x)$.*

Proof.

$$\begin{aligned}
\int f(x) \log F\big(B(J, k(J, x))\big)dx &= \int_{B(J,1)} f(x) \log F\big(B(J, 1)\big)dx + \dots \\
&\quad + \int_{B(J,2^J)} f(x) \log F\big(B(J, 2^J))\big)dx \\
&= \log F\big(B(J, 1)\big) \int_{B(J,1)} f(x)dx + \dots \\
&\quad + \log F\big(B(J, 2^J)\big) \int_{B(J,2^J)} f(x)dx \\
&= \sum_{k=1}^{2^J} F\big(B(J, k)\big) \log F\big(B(J, k)\big).
\end{aligned}$$

As entropy $-\sum_{i=1}^{n} p_i \log p_i$ is maximized if $p_i = 1/n, \forall i = 1, \dots, n$, 2.6 is minimized if $F\big(B(J, k)\big) = 2^{-J}$, i.e., under the null hypothesis, and generally

$$\sum_{k=1}^{2^J} F\big(B(J, k)\big) \log F\big(B(J, k)\big) \geq \sum_{k=1}^{2^J} 2^{-J} \log 2^{-J} = -J \log 2. \tag{2.6}$$

□

To show consistency in the case of the alternative hypothesis when the true error

distribution is not logistic, let the *Kullback-Leibler neighborhood* of the density f be defined as

$$K_\epsilon(f) = \Big\{ g \in \mathcal{G} : \int f(x) \log \frac{f(x)}{g(x)} \mu(dx) < \epsilon \Big\} \text{ for } \epsilon > 0,$$

where $\mathcal{G}$ is the class of all densities with respect to Lebesgue measure on $\mathbb{R}$. Let P denote the probability under the nonparametric Polya tree prior. Then f is said to be in the *Kullback-Leibler support* of P if $P\big(K_\epsilon\big(f)\big) > 0$ for all $\epsilon > 0$.

Dass and Lee [2004, Theorem 3] show that if the true density f, where $f \neq f_0$, is in the Kullback-Leibler support of the prior distribution, then the Bayes factor is consistent under H_1. For the specific case of an infinite Polya tree prior, Ghosal et al. [1998] showed that f will be in the Kullback-Leibler support if $\sum_{j=1}^{\infty} \rho(j)^{-1/2} < \infty$. This property is not satisfied for our choice of $\rho(j) = j^2$, however, we do not employ an infinite Polya tree and can therefore invoke the following approximation.

In the case of a finite Polya tree truncated at level J, the true density f is not guaranteed to lie in the support of the finite Polya tree prior. We can, however, find a $\delta > 0$ where

$$\delta = \inf \Big\{ \int f(x) \log \frac{f(x)}{g(x)} dx : g \text{ is in the support of } FPT(c, \rho(j), G_0) \Big\} \tag{2.7}$$

and δ is no larger than $\delta_0 = \int f(x) \log[f(x)/g_0(x)]dx$, which is the Kullback-Leibler divergence $D(g_0, f)$ for the logistic centering distribution g_0. To establish that $\delta < \delta_0$,

note that

$$\begin{aligned} D(g,f) &= \int f(x)\log\frac{f(x)}{g(x)}dx \\ &= \int f(x)\log\frac{f(x)}{g_0(x)2^J p_{\mathcal{Y}}(k(J,x))}dx \\ &= \int f(x)\Big\{\log\frac{f(x)}{g_0(x)} - \log\big[2^J p_{\mathcal{Y}}\big(k(J,x)\big)\big]\Big\}dx \\ &= \int f(x)\log\frac{f(x)}{g_0(x)}dx - \int f(x)J\log 2dx - \int f(x)\log p_{\mathcal{Y}}\big(k(J,x)\big)dx. \end{aligned}$$

By the definition of a finite Polya tree and the fact that the interval $[0,1]$ is the support for each branch probability, we can guarantee that there is a set of branch probabilities $\mathcal{Y}$ for which $p_{\mathcal{Y}}\big(k(J,x)\big) = F\big[B\big(J,k(J,x)\big)\big]$ for all x, where F is the distribution function corresponding to the true density f. Therefore,

$$\int f(x)\log\frac{f(x)}{g(x)}dx = \delta_0 - J\log 2 - \int f(x)\log F\big(B(J,k(J,x))\big)dx,$$

where by proposition 2.2, $J\log 2 \geq -\int f(x)\log F\big(B(J,k(J,x))\big)dx$, with equality only if $f(x) = g_0(x)$. Therefore, under the alternative hypothesis there exists a $\delta < \delta_0$, i.e., there is a density in the support of the FPT that is closer to f in the Kullback-Leibler sense than the logistic density g_0.

Walker et al. [2004] rewrite the Bayes factor as $BF = I_1/I_0$, where $I_j = \int\prod_{i=1}^n g_j(x_i)/f(x_i)P(dg_j)$ and g_j is the (conditional) sampling model for the data under model M_j, $j = 0,1$. They show that if a $\delta > 0$ as in (2.7) exists and $\liminf_n D(g_n,f) \geq \delta$, a.s., then $n^{-1}\log I_1 \to -\delta$. Additionally, since $I_0 = \prod_{i=1}^n \frac{g_0(x_i)}{f(x_i)}$, we have

$$\begin{aligned}\frac{1}{n}\log I_0 &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i)}{g_0(x_i)} \\ &\to -E_f\Big[\log\frac{f(x)}{g_0(x)}\Big] \\ &= -\int f(x)\log\frac{f(x)}{g_0(x)}dx = -\delta_0.\end{aligned}$$

As a result,

$$\begin{aligned}n^{-1}\log BF_n &= n^{-1}\log I_{1n}/I_{0n} \\ &\to \delta_0 - \delta, \text{a.s.}\end{aligned} \tag{2.8}$$

Therefore, when H_1 holds the Bayes factor tends to infinity as n goes to infinity, so the goodness of fit test is consistent under H_1.

2.5 Estimation

Should the goodness of fit test indicate that the residuals do not follow a logistic distribution, our methodology has a built-in semiparametric approach to estimating risks, relative risks, odds ratios and related measures. We use the nonparametric residual distribution modeled by the Polya tree to estimate the risk of an observation falling above any cutoff d, which is equivalent to the risk of a residual falling above $(d - x'\beta)/\sigma$. This risk can be estimated for any covariate vector of interest, as visualized in Figure 2.3. Note that risks can be estimated for multiple cutoffs simultaneously within the same model, and estimation for multinomial outcomes can be performed analogously to the case of dichotomous classification.

Hanson [2006] presents several computational aspects of finite Polya tree models. In particular, we are interested in probabilities at the lowest level of a Polya tree. If c and $\rho(\cdot)$ are fixed, G is completely defined by $\mathcal{Y}$, with $G[B(j,k)|\mathcal{Y}] = \prod_{i=1}^{j} Y_{e_i(k\lceil 2^{i-j}\rceil)}$.

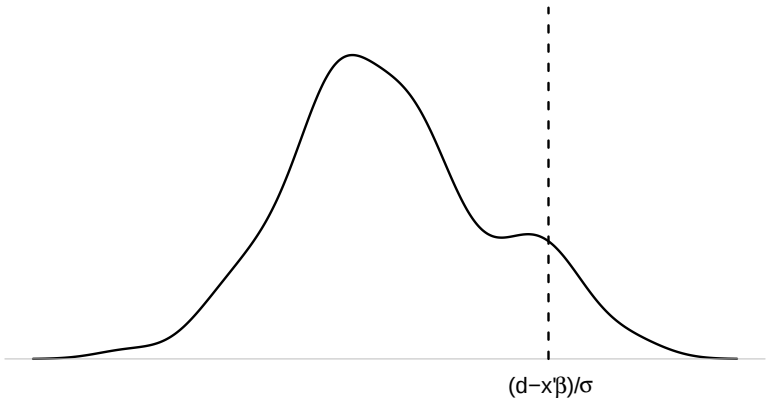


Figure 2.3: Risk estimation for the event $Y > d$ can be performed for any covariate vector x and any arbitrary shape of the residual distribution.

Let $k(J, \varepsilon) \in \{1, \ldots, 2^J\}$ index the set at level J of the tree into which scalar ε falls, and let $p_{\mathcal{Y}}(k)$ be the probability of the k-th partition on the lowest level (J) of a finite tree. Where $d^* = (d - x'\beta)/\sigma$ and M is the number of generated samples from the posterior distribution, the risk $P(Y > d|x)$ of a response greater than cutoff d for an individual with covariate vector x is

$$\begin{aligned} P((Y - x'\beta)/\sigma > (d - x'\beta)/\sigma) &= P(\varepsilon > d^*) \\ &= \sum_{m=k(J,d^*)+1}^{2^J} p_{\mathcal{Y}}(m) + p_{\mathcal{Y}}(k(J, d^*))[k(J, d^*) - 2^J G_0(d^*)] \\ \doteq \frac{1}{M} \sum_{i=1}^{M} \Big\{ \sum_{m=k(J,d^{*(i)})+1}^{2^J} & p_{\mathcal{Y}^{(i)}}(m) + p_{\mathcal{Y}^{(i)}}(k(J, d^{*(i)}))[k(J, d^{*(i)}) - 2^J G_0(d^{*(i)})] \Big\}. \quad (2.9) \end{aligned}$$

For the Empirical Bayes approach, where β and σ are set equal to their least-squares estimates, we generate posterior realizations of Polya trees by sampling from the posterior distributions of the PT probabilities. Risks and functions thereof (e.g., odds ratios) are calculated according to (2.9), and credible intervals for any of these parameters are derived based on appropriate posterior percentiles. For this calculation, the two PT probabilities at level 1 of the tree are not fixed at 0.5, as we are no

longer estimating a location parameter, but merely the shape of the error distribution.

In the case of a fully Bayesian implementation, a risk value would be generated from its posterior distribution at each iteration of the Gibbs sampler based on the current values $\beta^{(i)}$, $\sigma^{(i)}$ and $\mathcal{Y}^{(i)}$. Estimates for other quantities of interest such as risk ratios, risk differences, odds and odds ratios, can be obtained by both methods based on the risks estimated at each iteration of the Gibbs sampler.

2.6 Simulation study

Simulations on data from known distributions compare the performance of the Empirical Bayes goodness of fit test to that of a full Bayesian MCMC implementation. For risk estimation, we compare results from the Empirical Bayes estimation procedure to estimates from traditional logistic regression and a linear model fit to the continuous response.

2.6.1 Full Bayesian approach

The Bayes factor under a fully Bayesian analysis for testing goodness of fit of a logistic distribution is $BF = p(\mathcal{Y}_0)/p(\mathcal{Y}_0|y)$, where $p(\mathcal{Y}_0|y) = \int p(\mathcal{Y}_0|\theta, y)p(\theta|y)d\theta \doteq \frac{1}{M}\sum_{i=1}^{M} p(\mathcal{Y}_0|\theta^{(i)}, y)$. Here, $(\theta^{(1)}, \ldots, \theta^{(M)})$ is an MCMC sample from the posterior distribution of θ, and $p(\mathcal{Y}_0|\theta^{(i)}, y)$ is the full conditional of $\mathcal{Y}$ evaluated at 0.5 at iteration i, namely $p(\mathcal{Y}_0|\theta^{(i)}, y) = p(\mathcal{Y}_0|\bar{\varepsilon}^{(i)}) = \prod_{j=1}^{J}\prod_{k=1}^{2^{j-1}} beta(0.5|c\rho(j) + n(e_j(2k-1), \bar{\varepsilon}^{(i)}), c\rho(j) + n(e_j(2k), \bar{\varepsilon}^{(i)}))$.

We use independent priors of the following form on β and σ:

$$\log\sigma \sim N(\mu_\sigma, s_\sigma^2); \quad \beta \sim N_p(\mu_\beta, \Sigma_\beta).$$

These priors may be diffuse or informative, the latter being constructed using methods similar to those detailed in Bedrick et al. [1996]. The posterior distribution of (β, σ) is approximated numerically by using output from a Gibbs sampler that contains a Metropolis-Hastings step for sampling $(\beta, \log\sigma)$, specifically using a random

walk chain with a multivariate normal proposal distribution. In our applications, the covariance matrix of the proposal distribution was determined by running the chain for an initial 5000 iterations and computing the sample covariance of the simulated $(\beta^{(i)}, \sigma^{(i)})$ iterates. This matrix was scaled to achieve reasonable mixing and acceptance rates. The full conditionals used in the Gibbs sampler to generate samples from the posterior distribution are listed below.

Sampling branch probabilities. Based on the conjugacy result for Polya trees, the distribution of a branch probability, given all other parameters and the data, is an updated beta distribution, namely

$$Y_{e_j(2k-1)}|(\beta, \sigma, c, y) \sim \text{beta}\big(c\rho(j) + n(j, 2k-1, \bar{\varepsilon}), c\rho(j) + n(j, 2k, \bar{\varepsilon})\big)$$

for k in $\{1, \ldots, 2^{j-1}\}$ and $Y_{e_j(2k)} = 1 - Y_{e_j(2k-1)}$.

Sampling (β, σ). The full conditional density for the distribution of $\beta, \sigma|\mathcal{Y}, c, y$ is

$$\begin{aligned} p(\beta, \sigma|\mathcal{Y}, c, y) &\propto p(y|\mathcal{Y}, \beta, \sigma)p(\beta)p(\sigma)p(\mathcal{Y}|c) \\ &\propto \left\{\prod_{j=1}^{n} g(y_j|\mathcal{Y}, \beta, \sigma)\right\} p(\beta)p(\sigma)p(\mathcal{Y}|c), \end{aligned}$$

where $g(y_j|\mathcal{Y}, \beta, \sigma)$ is the Polya tree density as defined in (1.2).

After drawing $\theta^{cand} = (\beta^{cand}, \sigma^{cand})$ from the proposal distribution, the candidate iterate is accepted with probability

$$\min\left\{1, \frac{p(\theta^{cand}|\mathcal{Y}, c, y)}{p(\theta^{curr}|\mathcal{Y}, c, y)}\right\} = \min\left\{1, \frac{\prod_{j=1}^{n} g(y_j|\mathcal{Y}, \beta^{cand}, \sigma^{cand})p(\beta^{cand})p(\sigma^{cand})}{\prod_{j=1}^{n} g(y_j|\mathcal{Y}, \beta^{curr}, \sigma^{curr})p(\beta^{curr})p(\sigma^{curr})}\right\}. \tag{2.10}$$

Each iteration of the Gibbs sampler will result in an iterate of the Bayes factor, calculated from the current iterates $\mathcal{Y}^{(i)}, \beta^{(i)}, \sigma^{(i)}$ according to (2.3). The final Bayes factor is determined as the mean across all post burn-in MCMC iterates.

2.6.2 Simulation data

The algorithms were implemented and all simulations were run in R version 2.7.0 or 2.7.1 [R Development Core Team, 2009]. Data were generated for scenarios with and without covariates. The model that generated the data for the case without covariates was $y_i = 25 + \varepsilon_i$, with ε_i being generated from the following distributions: logistic$(0, 2)$, $N(0, 2)$, $t(3)$, a mixture of two normals, the $N(-4, 2)$ with probability 0.4 and $N(4, 2)$ with probability 0.6, and exp(1). For the scenarios with covariates, the generating model for the simulated observations had $y_i = 15 + x_{1i} + 0.3x_{2i} + \varepsilon_i$, where the distributions generating ε_i were the same as in the no-covariate case. For each observation, x_{1i} was generated from a Bernoulli(0.4)-distribution, while x_{2i} was generated from the $N(40, 8)$. For each scenario, 100 simulated data sets of size $n = 50, 100, 200,$ and 400 were generated.

To investigate the performance of the Empirical Bayes approach to testing and estimation, results were generated based on 10,000 posterior samples from the updated Polya tree distributions. For the full MCMC implementation, the prior distribution for β was chosen to be $N(0, I_p \cdot 100)$ and for $\log \sigma$ it was $N(2, 2)$. The MC chains were run for 100,000 iterations with the first 20,000 iterations discarded as a burn-in period.

2.6.3 Test performance

One-sample data

As a first look at the performance of our proposed goodness of fit test, we considered a scenario without covariates. Tables 2.1 and 2.2 present the number of simulated data sets for which the Bayes factor fell into each category of Jeffreys' classification

(see Table 1.1) for each of the error distributions investigated. These tables present results for both the full Bayesian model and the Empirical Bayes (EB) approach. To consolidate the tables, and because there is little practical difference between "strong evidence" and "very strong evidence," categories 3 and 4 have been collapsed. For the full Bayesian approach, results are listed for $J = 4$, while for the EB approach we compare results for $J = 4$ to those for $J = 8$ to investigate the influence of the size of the finite Polya tree. The two methods give comparable results across all scenarios considered, especially with relatively large sample sizes ($n \geq 100$).

The size of the parameter c affects how far the nonparametric posterior distribution of the residuals can deviate from the logistic centering distribution, with larger values of c putting more prior weight on logistic regression. Results in Tables 2.1 and 2.2 are presented for $c = 0.1, 0.5, 1, 5, 10$. In the case of a truly logistic error distribution, for c as large as 5 or 10, Bayes factors are somewhat less likely to accept the null hypothesis. In the case of alternative error distributions, Bayes factors under larger values of c are less likely to pick up deviations from the null distribution for smaller sample sizes. A value for c as small as 0.5 is therefore recommended, in particular for small sample sizes, to allow data-driven deviations from the centering distribution. With increasing sample size, the data will overwhelm the effect even of larger c in the posterior branch probabilities, particularly in higher levels of the tree.

Comparing the full Bayesian MCMC sampling approach to the EB method, we find that in the case of a true logistic distribution, we are slightly more likely to find evidence against H_0 with the MCMC approach, although the proportion of cases for which we find no substantial or no evidence against H_0 is still at 0.94 or higher. In the case of a logistic error distribution, the EB test finds no evidence against H_0 in almost all cases (97%-100% for $n = 50$, 100% for greater n), and never results in a Bayes factor higher than category 1 ("barely worth mentioning"). The full MCMC implementation indicates at least "substantial" evidence against H_0 in about 1% -

Table 2.1: From 100 simulated data sets, the number of Bayes factors that fall into the categories defined by Jeffreys (see Table 1.1) for a scenario without covariates for the fully Bayesian and empirical Bayesian goodness of fit test of logistic distribution.

			Fully Bayesian approach					Empirical Bayes approach									
			J=4					J=4					J=8				
			Jeffreys' categories														
Distribution	c	n	0	1	2	3&4	5	0	1	2	3&4	5	0	1	2	3&4	5
logistic	0.1	50	99	1	0	0	0	100	0	0	0	0	100	0	0	0	0
		100	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0
		200	98	0	0	1	1	100	0	0	0	0	100	0	0	0	0
		400	98	0	0	1	1	100	0	0	0	0	100	0	0	0	0
	0.5	50	95	5	0	0	0	99	1	0	0	0	99	1	0	0	0
		100	97	1	1	1	0	100	0	0	0	0	100	0	0	0	0
		200	95	2	2	0	1	100	0	0	0	0	100	0	0	0	0
		400	99	0	0	1	0	100	0	0	0	0	100	0	0	0	0
	1	50	96	4	0	0	0	97	3	0	0	0	97	3	0	0	0
		100	95	3	1	1	0	100	0	0	0	0	100	0	0	0	0
		200	94	4	1	1	0	99	1	0	0	0	99	1	0	0	0
		400	98	2	0	0	0	100	0	0	0	0	100	0	0	0	0
	5	50	95	5	0	0	0	96	3	1	0	0	96	3	1	0	0
		100	95	4	0	1	0	98	2	0	0	0	97	3	0	0	0
		200	91	7	1	1	0	97	3	0	0	0	97	2	1	0	0
		400	96	3	0	0	1	100	0	0	0	0	100	0	0	0	0
	10	50	91	9	0	0	0	96	3	1	0	0	96	3	1	0	0
		100	95	4	1	0	0	96	4	0	0	0	96	4	0	0	0
		200	92	7	0	1	0	95	5	0	0	0	95	5	0	0	0
		400	99	0	0	1	0	100	0	0	0	0	99	1	0	0	0
normal	0.1	50	95	1	2	1	1	100	0	0	0	0	100	0	0	0	0
		100	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0
		200	97	2	0	0	1	100	0	0	0	0	100	0	0	0	0
		400	93	0	1	3	3	100	0	0	0	0	100	0	0	0	0
	0.5	50	96	0	1	2	1	100	0	0	0	0	100	0	0	0	0
		100	88	7	4	1	0	100	0	0	0	0	100	0	0	0	0
		200	91	2	2	2	3	100	0	0	0	0	99	1	0	0	0
		400	87	2	4	4	3	100	0	0	0	0	100	0	0	0	0
	1	50	96	0	2	2	0	100	0	0	0	0	100	0	0	0	0
		100	86	8	3	3	0	99	1	0	0	0	98	2	0	0	0
		200	88	5	3	3	1	100	0	0	0	0	99	0	1	0	0
		400	83	3	5	4	5	98	2	0	0	0	98	2	0	0	0
	5	50	96	4	0	0	0	97	3	0	0	0	97	3	0	0	0
		100	87	12	1	0	0	88	11	1	0	0	89	10	1	0	0
		200	87	7	6	0	0	93	3	4	0	0	93	3	4	0	0
		400	75	14	3	6	2	91	5	2	2	0	88	9	0	3	0
	10	50	96	4	0	0	0	96	4	0	0	0	96	4	0	0	0
		100	86	14	0	0	0	86	13	1	0	0	87	12	1	0	0
		200	86	9	5	0	0	91	5	4	0	0	91	5	4	0	0
		400	74	15	6	4	1	83	13	1	3	0	82	14	1	3	0

Table 2.2: From 100 simulated data sets, the number of Bayes factors that fall into the categories defined by Jeffreys (see Table 1.1) for a scenario without covariates for the fully Bayesian and empirical Bayesian goodness of fit test of logistic distribution.

			Fully Bayesian approach					Empirical Bayes approach									
			J=4					J=4					J=8				
								Jeffreys' categories									
Distribution	c	n	0	1	2	3&4	5	0	1	2	3&4	5	0	1	2	3&4	5
t(3)	0.1	50	22	7	2	7	62	30	11	5	12	42	36	8	4	15	37
		100	1	2	4	3	90	9	1	1	4	85	12	2	1	5	80
		200	1	0	0	0	99	0	0	0	0	100	2	0	0	1	97
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	0.5	50	11	6	8	15	60	10	7	3	16	64	12	5	6	15	62
		100	1	0	0	3	96	1	1	0	2	96	2	0	0	3	95
		200	0	0	0	1	99	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	1	50	10	8	6	24	52	6	5	7	19	63	6	7	6	17	64
		100	1	0	1	2	96	1	0	1	1	97	1	1	0	1	97
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	5	50	7	23	30	32	8	3	11	15	32	39	3	11	15	32	39
		100	1	1	2	17	79	0	1	1	2	96	0	1	1	2	96
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	10	50	6	39	38	15	2	2	20	24	31	23	2	20	24	31	23
		100	1	1	7	27	64	0	1	2	13	84	0	1	2	13	84
		200	0	0	0	1	99	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
bimodal	0.1	50	4	7	2	13	74	66	9	7	7	11	75	5	4	7	9
		100	0	0	0	0	100	31	6	8	14	41	46	3	8	14	29
		200	0	0	0	0	100	1	1	1	0	97	5	1	1	4	89
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	0.5	50	3	6	8	22	61	24	15	15	26	20	25	15	15	25	20
		100	0	0	1	1	98	2	1	4	15	78	2	0	2	14	82
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	1	50	3	7	13	28	49	12	19	23	28	18	12	20	22	28	18
		100	0	0	1	2	97	0	1	2	16	81	0	1	1	14	84
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	5	50	3	40	34	20	3	5	40	37	18	0	6	39	36	19	0
		100	0	0	3	29	68	0	0	4	29	67	0	0	4	27	69
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	10	50	3	70	22	5	0	5	61	29	5	0	6	60	29	5	0
		100	0	2	17	57	24	0	2	10	59	29	0	1	10	60	29
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
exp(1)	0.1	50	24	8	13	11	44	52	10	8	11	19	54	6	6	16	18
		100	0	0	0	0	100	5	0	0	4	91	5	1	5	6	83
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	0.5	50	8	14	16	27	35	24	4	19	22	31	23	5	13	27	32
		100	0	0	0	0	100	0	0	2	2	96	0	0	0	3	97
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	1	50	8	20	25	26	21	16	12	22	27	23	15	12	17	31	25
		100	0	0	0	1	99	0	0	1	3	96	0	0	0	3	97
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	5	50	20	43	27	9	1	10	35	24	26	5	9	33	26	27	5
		100	0	0	6	41	53	0	0	3	12	85	0	0	3	11	86
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	10	50	21	56	18	5	0	7	47	26	18	2	7	46	26	19	2
		100	0	6	25	55	14	0	0	6	42	52	0	0	6	41	53
		200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
		400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100

2% of cases.

One would not expect great power to detect the subtle difference between a Normal and a logistic distribution, as they are very similar in shape. For larger samples ($n = 400$), the MCMC approach is able to pick up the slight deviation from a logistic distribution in only a handful of cases and finds at least substantial evidence against H_0 in 7% - 14% of simulated data sets. As with the logistic distribution, the EB approach is less likely to find evidence against the null hypothesis and in our simulations does not detect any deviation from the null hypothesis when c is small.

For scenarios with a $t(3)$, bimodal or exponential(1) error distribution, each of which deviates from the null hypothesis more prominently than the Normal distribution does, we find that with sample sizes of 200 and 400, both methods for calculating the Bayes factor find decisive evidence against the null in almost 100% of the cases.

With data from an exponential distribution, differences between the two approaches are overall small; with a sample size of 100 or greater, the deviation from the null hypothesis was picked up in all cases by both methods. Similarly, for regression error from a $t(3)$ distribution, with the recommended value of c the performance of the two approaches is practically identical. At the smaller sample sizes considered here, the MCMC test proves more powerful than the EB method. These differences show particularly in the data sets generated from a bimodal distribution, where, for $c = 0.5$ or 1, the EB approach finds little or no evidence against H_0 in $30\% - 40\%$ of the data sets of size 50, while for the full Bayesian approach these percentages drop to about 10%.

For small sample sizes the full Bayesian implementation appears to have some advantages over the EB approach. In the simulations presented here, a sample size of 100 is however large enough to diminish those advantages. This is particularly encouraging as computational savings achieved by the EB method are especially relevant for larger sample sizes for which running a full MCMC implementation would

be costly or not feasible at all.

For all practical purposes, results from the EB approach were the same for $J = 4$ and $J = 8$. Using a Polya tree with $2^8 = 256$ partitions versus a tree with $2^4 = 16$ partitions at the lowest level does not appear to affect the conclusion of this goodness of fit test. Given that computational effort increases exponentially with the number of levels in a Polya tree, this is an important finding that supports the use of a relatively small J when interest lies in testing for a logistic residual distribution. However, estimation is affected by J, so that for estimation purposes a larger value of J should be preferred.

For a different look at the simulation results, Table 2.3 presents the median Bayes factor (on the $\log_{10}$ scale) for each method for $J = 4$. As discussed above, in most scenarios both methods seem to agree on when to reject H_0 or not. In the case of a logistic or a Normal distribution, although both methods tend to not find evidence against H_0, the Bayes factor derived by the EB method tends to be smaller than that based on the full Bayesian approach, i.e., the EB methods tends to find more evidence for H_0. In the case of the other three distributions investigated, Bayes factors again tend to be higher when calculated based on MCMC iterates, which leads to the higher rates of rejecting H_0 discussed above. The exponentially distributed data is the only case in which there is no consistent relation between the Bayes factors calculated by the two approaches; for small c the full Bayesian approach tends to find more evidence against H_0, while for larger values of c there tends to be less evidence.

Regression data

In a regression setting, in addition to testing for a parametric logistic error distribution, parameters β and σ need to be estimated either using least-squares (EB) or MCMC sampling. As results indicate in the previous section, values of c larger than 1 tend to give posterior estimates very close to the centering distribution and thus tests do not indicate evidence against a logistic error distribution except for large sample

Table 2.3: Median of $\log_{10}(BF)$ across 100 simulations without covariates, $J = 4$

		Fully Bayesian approach					Empirical Bayes approach				
		Distribution									
c	n	logistic	Normal	t(3)	bimodal	exp(1)	logistic	Normal	t(3)	bimodal	exp(1)
0.1	50	-3.21	-2.74	2.99	3.62	1.35	-4.91	-4.88	1.25	-1.10	-0.08
	100	-4.40	-3.80	11.44	11.57	7.06	-6.56	-6.34	8.47	1.35	5.17
	200	-6.26	-5.35	27.36	27.55	18.50	-8.53	-8.13	22.18	8.25	15.87
	400	-8.71	-6.81	60.01	55.25	42.47	-10.85	-10.01	52.80	20.67	38.23
0.5	50	-1.39	-1.30	2.52	2.43	1.37	-1.97	-1.91	2.82	0.73	1.18
	100	-1.88	-1.53	9.53	7.38	6.57	-2.76	-2.63	10.85	3.61	6.10
	200	-2.72	-2.16	26.54	19.97	18.13	-3.90	-3.49	24.68	10.69	16.37
	400	-4.07	-2.88	57.50	48.93	41.58	-5.59	-4.78	55.00	23.55	38.16
1	50	-0.97	-0.92	2.05	1.96	0.94	-1.34	-1.30	2.78	0.89	1.03
	100	-1.24	-1.03	7.55	5.82	5.27	-1.84	-1.72	10.32	3.56	5.35
	200	-1.84	-1.46	23.47	17.26	15.91	-2.62	-2.28	24.09	10.38	14.86
	400	-2.74	-1.79	52.39	38.31	38.75	-3.99	-3.19	53.83	23.50	35.71
5	50	-0.37	-0.34	0.85	0.60	0.30	-0.50	-0.50	1.57	0.53	0.58
	100	-0.53	-0.41	3.64	2.42	2.04	-0.67	-0.58	6.33	2.39	2.85
	00	-0.76	-0.55	12.83	9.91	8.27	-1.02	-0.76	17.45	7.45	9.03
	400	-1.14	-0.65	33.75	27.04	24.39	-1.65	-1.02	44.28	18.83	24.03
10	50	-0.23	-0.22	0.55	0.32	0.17	-0.32	-0.31	1.09	0.36	0.46
	100	-0.33	-0.27	2.39	1.54	1.24	-0.43	-0.37	4.68	1.69	2.07
	00	-0.51	-0.36	8.69	6.18	5.21	-0.67	-0.45	13.27	5.62	6.57
	400	-0.81	-0.43	25.23	19.27	17.30	-1.12	-0.62	36.47	15.28	18.28

sizes. Results presented in this section are therefore limited to values $c = 0.1, 0.5, 1$.

For true logistic error (Table 2.4), the results show only small differences compared to simulations without covariates. The full Bayesian approach is now even more likely to find evidence against H_0. While without covariates 94%-100% of tests found no evidence against the null hypothesis, now only 84%-99% correctly fall into this category. On the other hand, the percentage of Bayes factors that falsely find at least "substantial evidence again H_0" is as great as 10% in one scenario ($c = 0.5, n = 100$).

The EB method is once again less powerful in detecting the difference between a logistic and a normal distribution. The full Bayesian implementation is in fact more powerful in this scenario with covariates than in the previous setting. This gain, however, comes with the increase of falsely discovered deviations from the logistic distribution discussed above.

In the three cases where the error distribution deviates more strongly from the null hypothesis (Table 2.5), the need to estimate additional parameters results in a less powerful test for small sample sizes, indicated by overall smaller median Bayes

Table 2.4: From 100 simulated data sets, the number of Bayes factors that fall into the categories defined by Jeffreys (see Table 1.1) for a scenario with two covariates for the fully Bayesian and empirical Bayesian goodness of fit test of logistic distribution.

			Fully Bayesian approach					Empirical Bayes approach									
			J=4					J=4					J=8				
			Jeffreys' categories														
Distribution	c	n	0	1	2	3&4	5	0	1	2	3&4	5	0	1	2	3&4	5
logistic	0.1	50	92	2	1	0	1	100	0	0	0	0	100	0	0	0	0
	0.1	100	96	2	1	0	1	100	0	0	0	0	100	0	0	0	0
	0.1	200	99	0	0	1	0	100	0	0	0	0	100	0	0	0	0
	0.1	400	99	0	0	1	0	100	0	0	0	0	100	0	0	0	0
	0.5	50	91	5	3	1	0	100	0	0	0	0	100	0	0	0	0
	0.5	100	84	6	5	4	1	99	1	0	0	0	100	0	0	0	0
	0.5	200	89	6	4	1	0	100	0	0	0	0	100	0	0	0	0
	0.5	400	96	2	1	1	0	100	0	0	0	0	100	0	0	0	0
	1	50	93	6	1	0	0	99	1	0	0	0	99	1	0	0	0
	1	100	88	6	2	3	1	99	0	1	0	0	99	1	0	0	0
	1	200	88	4	6	1	1	99	1	0	0	0	99	1	0	0	0
	1	400	94	3	2	0	1	100	0	0	0	0	100	0	0	0	0
normal	0.1	50	84	5	5	4	2	100	0	0	0	0	100	0	0	0	0
	0.1	100	91	1	2	2	4	100	0	0	0	0	100	0	0	0	0
	0.1	200	95	3	0	1	1	100	0	0	0	0	100	0	0	0	0
	0.1	400	96	0	0	0	4	100	0	0	0	0	100	0	0	0	0
	0.5	50	86	5	5	1	3	100	0	0	0	0	100	0	0	0	0
	0.5	100	83	6	6	3	2	98	1	1	0	0	98	1	1	0	0
	0.5	200	84	5	3	4	4	100	0	0	0	0	100	0	0	0	0
	0.5	400	82	5	6	2	5	99	0	0	1	0	100	0	0	0	0
	1	50	91	3	4	2	0	100	0	0	0	0	99	1	0	0	0
	1	100	84	5	6	5	0	96	3	0	1	0	96	3	0	1	0
	1	200	73	14	3	8	2	99	1	0	0	0	99	1	0	0	0
	1	400	72	5	8	6	9	99	0	0	1	0	99	0	0	1	0

factors (see Table 2.6). This loss of power affects both the Empirical Bayes and the MCMC implementations, therefore observations about comparisons between MCMC and Empirical Bayes implementations remain the same as made above for the case of no covariates. Again, the full Bayesian approach has more power to detect deviations in the error distribution. However, at a sample size of 100, differences in conclusions according to Jeffreys' categories are diminishing. Moreover, the differences disappear at a sample size of 200, even though differences in the median value of the Bayes factor are maintained.

Table 2.5: From 100 simulated data sets, the number of Bayes factors that fall into the categories defined by Jeffreys (see Table 1.1) for a scenario with two covariates for the fully Bayesian and Empirical Bayesian goodness of fit test of logistic distribution.

			Fully Bayesian approach					Empirical Bayes approach									
			J=4					J=4					J=8				
			Jeffreys' categories														
Distribution	c	n	0	1	2	3&4	5	0	1	2	3&4	5	0	1	2	3&4	5
t(1)	0.1	50	27	5	4	9	55	71	0	4	9	16	71	2	7	4	16
	0.1	100	8	3	0	1	88	16	3	8	6	67	31	4	2	4	59
	0.1	200	0	0	0	1	99	0	0	0	2	98	4	0	0	2	94
	0.1	400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	0.5	50	23	7	14	18	38	32	6	11	17	34	33	6	8	21	32
	0.5	100	2	2	1	5	90	3	1	2	7	87	3	3	0	4	90
	0.5	200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	0.5	400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	1	50	23	13	15	18	31	23	11	8	23	35	22	12	7	26	33
	1	100	2	2	1	7	88	1	2	3	2	92	1	2	3	1	93
	1	200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	1	400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
bimodal	0.1	50	2	6	7	12	73	83	7	2	4	4	88	1	3	4	4
	0.1	100	0	0	0	2	98	38	7	6	16	33	57	7	4	13	19
	0.1	200	0	0	0	0	100	3	1	0	3	93	8	1	2	3	86
	0.1	400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	0.5	50	13	12	14	16	45	51	16	12	11	10	51	15	12	13	9
	0.5	100	0	1	0	1	98	6	3	5	11	75	6	3	7	10	74
	0.5	200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	0.5	400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	1	50	16	20	16	21	27	27	34	14	16	9	29	33	12	18	8
	1	100	1	0	0	5	94	1	3	6	12	78	1	4	4	12	79
	1	200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	1	400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
exp(1)	0.1	50	28	9	7	15	41	74	4	3	12	7	76	3	5	10	6
	0.1	100	1	0	2	2	95	20	5	4	14	57	28	6	9	15	42
	0.1	200	0	0	0	0	100	0	0	0	2	98	1	1	2	1	95
	0.1	400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	0.5	50	16	13	21	27	23	46	12	11	15	16	46	10	14	14	16
	0.5	100	0	0	0	1	99	2	1	4	10	83	1	2	1	11	85
	0.5	200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	0.5	400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	1	50	18	31	20	24	7	41	14	9	24	12	39	15	12	21	13
	1	100	0	0	1	1	98	1	1	3	11	84	1	1	2	9	87
	1	200	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	1	400	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100

Table 2.6: Median of $\log_{10}(BF)$ across 100 simulations with two covariates, $J = 4$.

		Fully Bayesian approach					Empirical Bayes approach				
		Distribution									
c	n	logistic	Normal	t(3)	bimodal	exp(1)	logistic	Normal	t(3)	bimodal	exp(1)
0.1	50	-2.22	-1.49	2.28	3.46	1.19	-4.87	-4.99	-1.36	-2.21	-1.62
	100	-3.19	-2.75	9.96	9.71	6.5	-6.75	-6.47	4.94	0.92	2.51
	200	-4.98	-4.3	25.57	20.3	16.82	-8.72	-8.34	17.3	7.24	11.48
	400	-7.26	-6.03	54.12	36.94	43.25	-10.8	-10.13	46.07	20.92	35.73
0.5	50	-0.9	-0.87	1.52	1.72	0.99	-1.93	-1.94	1.07	-0.03	0.22
	100	-1.1	-0.95	8.33	8.55	5.8	-2.9	-2.61	7.14	3.18	4.32
	200	-1.78	-1.24	22.88	18.64	16.08	-4.08	-3.71	19.93	9.8	12.78
	400	-3.34	-2.15	51.55	37.24	41.15	-5.58	-4.96	48.61	23.86	36.17
1	50	-0.71	-0.69	0.93	0.98	0.55	-1.28	-1.27	1.19	0.3	0.38
	100	-0.91	-0.76	6.87	6.74	4.69	-1.93	-1.7	7.06	3.16	4.06
	200	-1.32	-0.83	19.29	16.59	14.54	-2.8	-2.49	19.65	9.74	11.86
	400	-2.11	-1.17	46.18	36.06	38.51	-3.94	-3.38	47.64	23.52	33.87

2.6.4 Estimation

Performance of risk estimation under the proposed EB approach is compared to estimation under frequentist logistic regression and accelerated failure time modeling using 1000 simulated data sets that were generated under two scenarios. In the first case, the underlying error distribution is logistic. In the second scenario, residuals are simulated from the exponential distribution with scale parameter 1. Risk estimates and their credible intervals under the EB approach are obtained as outlined in Section 2.5. The Polya tree priors were chosen with $J = 8$ levels and $c = 0.5$. Confidence intervals for the two parametric methods can be easily generated from standard output by statistical software.

For data with a logistic error distribution, maximum likelihood and posterior mean risk estimates with root mean squared error (MSE) in parentheses, are presented in Table 2.7. In this scenario, odds ratios as described in Section 2.2 are constant across covariates for all values of the cutoff d. Therefore, assumptions for both logistic regression and the accelerated failure time model are fulfilled. All three methods correctly estimated the risk for all covariate combinations even for small sample sizes. Notably, the proposed EB approach does not overfit these data sets and provides accurate and precise risk inference. Comparing mean squared error, logistic regression results in greater error than the other two methods, due to the fact that the other methods directly model the linear relationship between covariates and response, while logistic regression models a dichotomized response. Empirical Bayes estimation and the accelerated failure time model give comparable MSE, with a slightly higher MSE for the EB approach. This stems from the fact that using a nonparametric model for the error distribution introduces additional uncertainty and thus variability into the estimation procedure.

Both parametric logistic regression and the AFT model assume constant odds ratios, and are therefore by design unable to account for the fact that for non-logistic

Table 2.7: Results for risk estimation under logistic regression (LR), accelerated failure time (AFT) model with a logistic baseline distribution, and empirical Bayes (EB) with $J = 8, c = 0.5$. The true error distribution is logistic$(0, 1)$. Results across 1000 simulated data sets are maximum likelihood and posterior mean estimates of risk, with root mean squared error in parentheses.

x_1	x_2	true risk	n	LR		AFT		EB	
40	0	0.062	50	0.057	(0.047)	0.058	(0.026)	0.061	(0.029)
			100	0.062	(0.034)	0.062	(0.018)	0.063	(0.021)
			200	0.061	(0.023)	0.061	(0.013)	0.062	(0.015)
			400	0.061	(0.017)	0.061	(0.009)	0.061	(0.011)
	1	0.140	50	0.130	(0.094)	0.136	(0.056)	0.140	(0.059)
			100	0.139	(0.064)	0.141	(0.039)	0.144	(0.043)
			200	0.139	(0.048)	0.139	(0.027)	0.141	(0.029)
			400	0.139	(0.033)	0.139	(0.019)	0.139	(0.021)
45	0	0.204	50	0.194	(0.111)	0.198	(0.064)	0.202	(0.067)
			100	0.206	(0.070)	0.206	(0.045)	0.208	(0.049)
			200	0.204	(0.051)	0.205	(0.031)	0.206	(0.034)
			400	0.203	(0.036)	0.204	(0.021)	0.204	(0.023)
	1	0.389	50	0.382	(0.178)	0.383	(0.102)	0.385	(0.106)
			100	0.391	(0.116)	0.390	(0.071)	0.390	(0.075)
			200	0.390	(0.084)	0.388	(0.052)	0.388	(0.055)
			400	0.388	(0.057)	0.388	(0.035)	0.389	(0.038)

error distributions, odds ratios are not generally independent of the covariate vector x_i or the cutoff d. The results presented in Table 2.8 for exponentially distributed error show the resulting bias in risk estimates under these parametric models. For logistic regression and the AFT model, the bias does not decrease with increasing sample size, whereas as n increases, the EB estimates approach the true risks.

For the AFT model in this setting, bias tends to be smaller than for logistic regression when the true risk is close to 0.5 (i.e., 0.223 or 0.607). For risks closer to the edges of the parameter space, in our simulations with risks of 0.05 and 0.14, the AFT estimates are farther from the true risks than estimates from logistic regression. Among the three methods, only the EB approach allows for sufficient flexibility in modeling risks so that estimates tend towards the true values with increasing sample size.

2.7 Examples

To demonstrate differences in estimated odds ratios provided by the semiparametric Empirical Bayes model when compared to the two parametric alternatives, estimation

Table 2.8: Results for risk estimation under logistic regression (LR), accelerated failure time (AFT) model with a logistic baseline distribution, and empirical Bayes (EB) with $J = 8, c = 0.5$. The true error distribution is exponential(1). Results across 1000 simulated data sets are maximum likelihood and posterior mean estimates of risk, with root mean squared error in parentheses.

x_1	x_2	true risk	n	LR		AFT		EB	
40	0	0.050	50	0.035	(0.045)	0.015	(0.037)	0.029	(0.031)
			100	0.034	(0.036)	0.015	(0.036)	0.031	(0.025)
			200	0.032	(0.028)	0.014	(0.036)	0.035	(0.020)
			400	0.031	(0.024)	0.014	(0.036)	0.039	(0.015)
	1	0.135	50	0.145	(0.131)	0.096	(0.069)	0.123	(0.060)
			100	0.141	(0.091)	0.094	(0.056)	0.128	(0.040)
			200	0.145	(0.062)	0.092	(0.051)	0.130	(0.028)
			400	0.146	(0.045)	0.092	(0.047)	0.133	(0.019)
45	0	0.223	50	0.281	(0.191)	0.213	(0.082)	0.224	(0.070)
			100	0.279	(0.136)	0.213	(0.058)	0.224	(0.047)
			200	0.279	(0.097)	0.214	(0.043)	0.224	(0.034)
			400	0.280	(0.079)	0.214	(0.031)	0.224	(0.024)
	1	0.607	50	0.697	(0.222)	0.672	(0.106)	0.642	(0.127)
			100	0.683	(0.148)	0.670	(0.086)	0.631	(0.094)
			200	0.687	(0.119)	0.666	(0.073)	0.619	(0.070)
			400	0.685	(0.098)	0.668	(0.068)	0.616	(0.050)

will be performed on data sets from two nationwide surveys. Risk factors for obesity will be modeled on a subset drawn from the Health and Retirement Study, and risk factors for diabetes, defined as plasma glucose levels above a clinical threshold, are modeled on a data set from the National Health and Nutrition Examination Survey. The intention of these investigations is not to discover new relationships between risk factors and outcomes, but instead to demonstrate that distributional assumptions may be violated in commonly analyzed data sets and how inferences may be affected.

2.7.1 Risk factors for obesity in the Health and Retirement Study

Moser and Coombs [2004] considered a subset of data collected by the Health and Retirement Study, which is sponsored by the National Institute of Aging and conducted by the University of Michigan [Health and Retirement Study, 1992]. The survey has been conducted yearly since 1992, and the collected data have been the source of a large number of publications. A search of the publication list on the study's website (http://hrsonline.isr.umich.edu/ index.php?p=biblio) retrieved over 1,800 related publications between 1992 and 2011. Moser and Coombs employed a

Table 2.9: Risk factors included in the analysis of the Health and Retirement data.

Variable	Range
Exercise	1 = vigorous exercise at least once a week, 0 = otherwise
Smoking	1 = regular smoker, 0 = otherwise
Alcohol	1 = at least occasional alcohol consumption, 0 = otherwise

subsample of the data set in 1992 to demonstrate that their suggested method of estimating odds ratios without dichotomizing yielded similar point estimates of odds ratios, but smaller confidence intervals than ordinary logistic regression. The data analyzed here were selected from the survey in 1992, and included a total of 4673 Caucasian women who were between 40 and 70 years old. The risk factors included in the analysis were exercise, smoking and alcohol. An attempt at a perfect reproduction of the data set used by Moser and Coombs was not successful due to the limited information provided on how exactly their data were selected. The definitions of the binary covariates are outlined in Table 2.9.

Residuals from a parametric AFT model fit to the data using the risk factors exercise, smoking and alcohol, as well as the covariates age and education (ranging from 0 to 17 years of completed education) indicate that the distribution of residuals is skewed to the right (Figure 2.4), implying that the assumption of a logistic distribution is not appropriate here. The Bayes factor for testing a logistic error distribution versus a nonparametric PT alternative ($J = 8$, $c = 0.5$) is $10^{97.5}$, providing decisive evidence against the null hypothesis.

Figure 2.5 presents risk curve estimates as a function of age for the different levels of exercise, smoking and alcohol consumption for logistic regression, the AFT model, and the semiparametric Empirical Bayes estimation procedure. In each plot, black lines represent the estimated risk of obesity for a woman who does not exercise, smoke, or consume alcohol and has completed 12 years of education. The grey lines represent the risk estimates for a woman who indicated 1 on either exercise, smoking or alcohol,

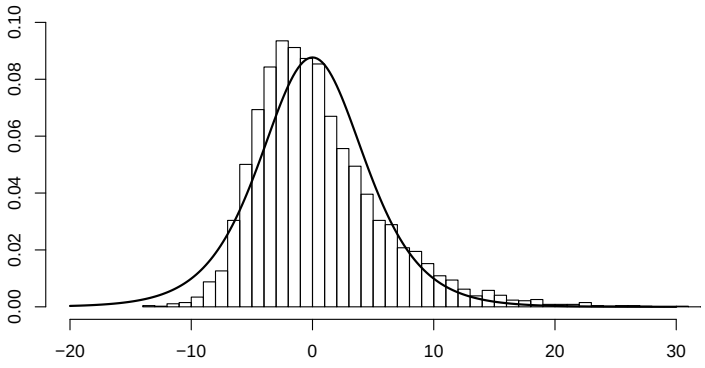
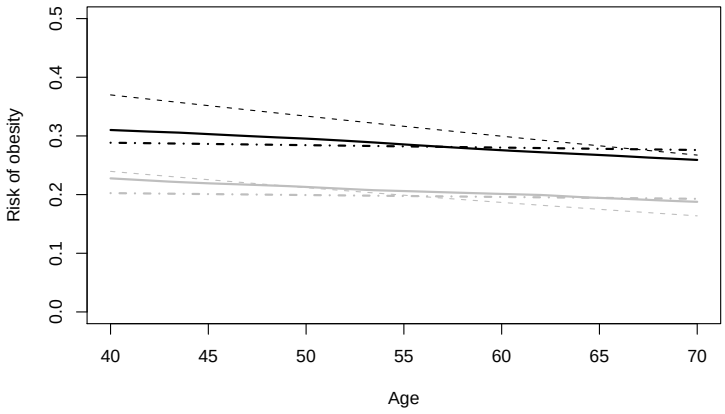


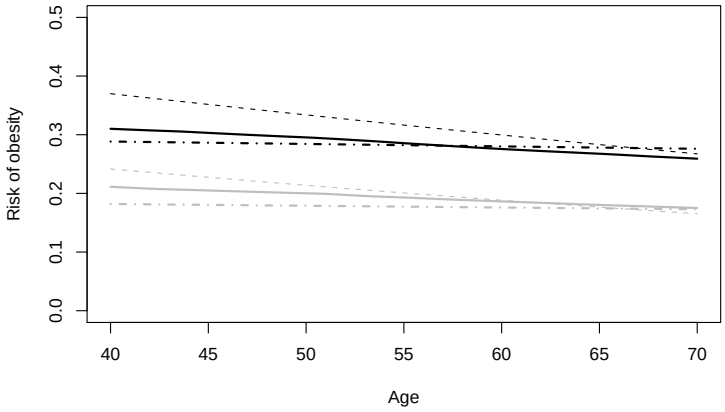
Figure 2.4: Histogram of residuals from the AFT model fit to the Health and Retirement data, compared to a logistic distribution with empirically estimated standard deviation.

but has the same values on all other predictors. All three methods give slightly different results, which agrees with simulation results that showed that disagreement even between the two parametric methods is greatest when the error distribution is skewed. Logistic regression suggests a higher risk of obesity for a woman who smokes or consumes alcohol than the other two methods. The AFT model, on the other hand, tends to suggest a lower risk than the two alternatives. Overall, EB estimation tends to suggest less of a risk difference than the two parametric methods.

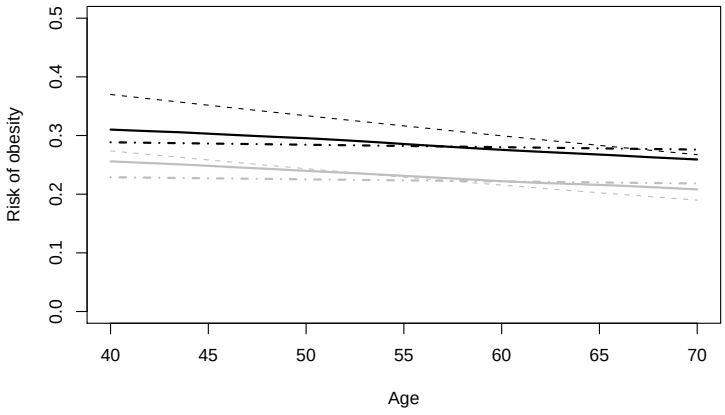
Odds ratios corresponding to the three risk factors as a function of age (in black) and their point-wise 95% confidence/credible intervals (in grey) are graphed in Figure 2.6. For the variables exercise and alcohol the odds ratio estimates from the EB and the AFT model are similar, while the odds ratios estimated by logistic regression are smaller than the EB estimates for all three factors. In these two cases, the credible intervals from the EB estimates just barely overlap with the logistic regression confidence intervals. Overall, EB odds ratio estimates are closer to 1, suggesting a smaller "protective" effect of exercise, smoking and alcohol than the two parametric models. Based on simulation results, we can speculate that the estimates from logistic regression and the AFT model are biased, and the EB estimates are more representative of



(a) Exercise



(b) Smoking



(c) Alcohol

Figure 2.5: Estimated risk of obesity from the Empirical Bayes method (solid lines), AFT (dot-dashed lines) and logistic regression (dotted lines) for variables exercise, smoking and alcohol equal to 0 (black lines) compared to an individual with only exercise = 1 (a), smoking = 1 (b) or alcohol = 1 (c).

the true relationship of the odds of obesity.

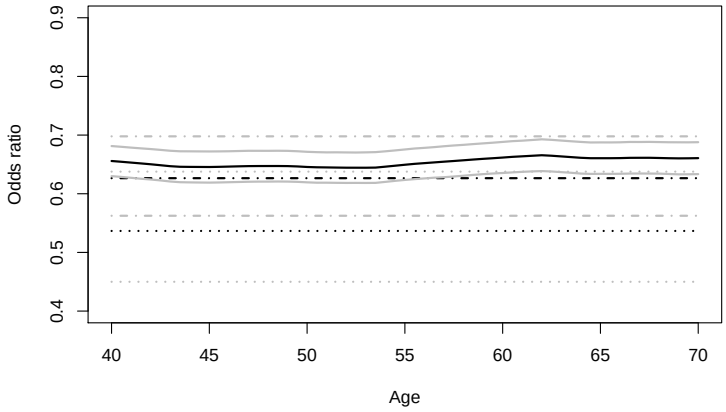
2.7.2 Risk of diabetes in NHANES

Diagnosis of diabetes is generally driven by repeated observation of several clinical factors. One common indicator of diabetes is fasting plasma glucose (FPG). A patient with (repeated) FPG levels at or above 126 mg/dL is considered diabetic. With diabetes being an increasing public health concern, monitoring diabetes levels across populations and identifying risk factors for diabetes has been of great concern in the public health community. Fasting plasma glucose is a common measure of identifying at-risk groups in epidemiologic studies, even if collecting multiple observations on individuals is not feasible [WHO06].

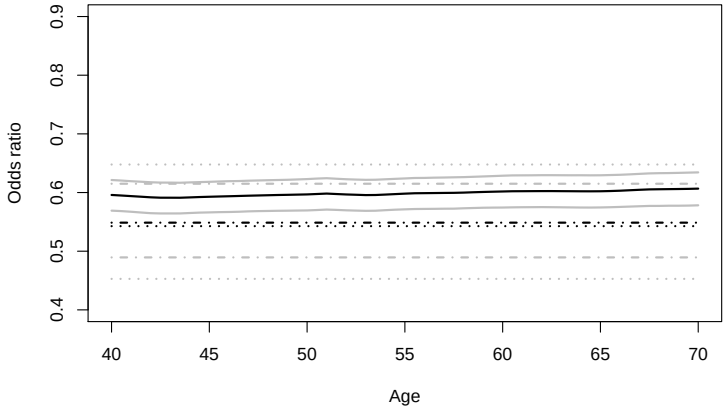
The National Health and Nutrition Examination Study (NHANES) is a U.S. survey conducted by the National Center for Health Statistics, which is part of the Centers for Disease Control and Prevention [NCHS11]. Survey data are released in two-year intervals. Among hundreds of variables, fasting plasma glucose levels, measured after 9 hours of fasting, are measured on thousands of participants. The data set analyzed here was selected from the 2007-08 database [CDC09]. The data set included the variables gender, age (18-80 years) and BMI, collected on 2699 individuals.

As a first step toward demonstrating the differences between the three estimation techniques that have been considered in this chapter, an AFT model was fit with quadratic terms in age and BMI as well as all possible interactions. This model was then reduced using backward step-wise selection. The final model using either AIC or p-values below 0.05 as elimination criterion was

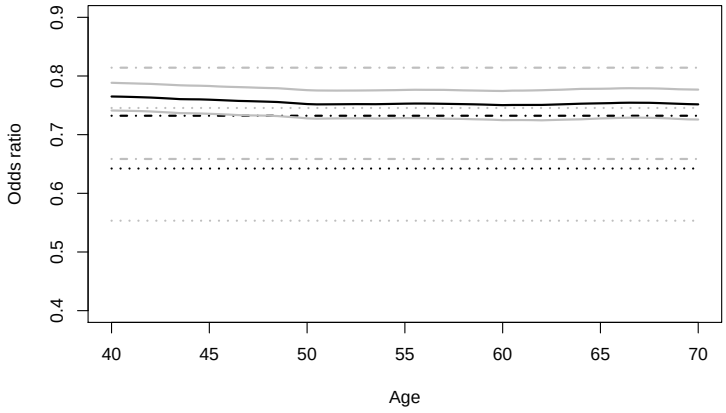
$$FPG_i = \beta_0 + \beta_1 * I(female)_i + \beta_2 * age_i + \beta_3 * BMI_i + \beta_4 * age_i * BMI_i + \varepsilon_i \tag{2.11}$$



(a) Exercise



(b) Smoking



(c) Alcohol

Figure 2.6: Estimated odds ratios of obesity for risk factors exercise, smoking and alcohol with 95% confidence/credible intervals in grey from the Empirical Bayes method (solid lines), AFT (dot-dashed lines) and logistic regression (dotted lines).

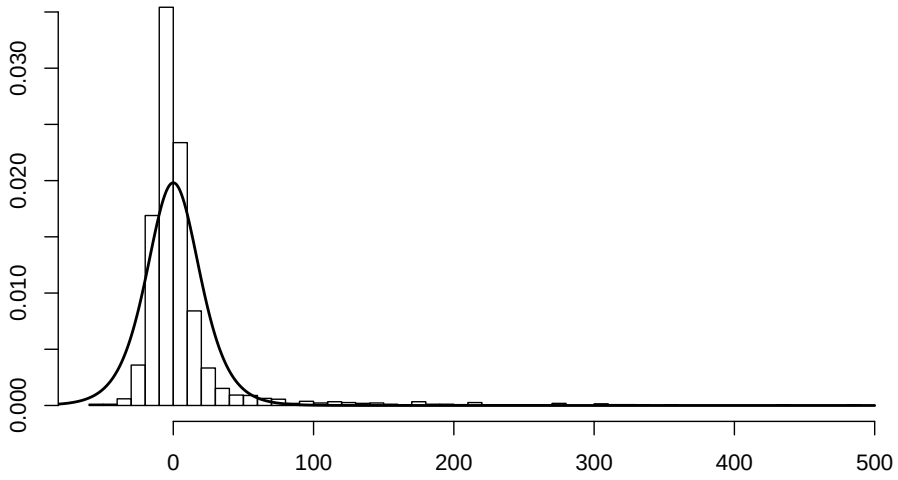


Figure 2.7: Histogram of residuals from the AFT model fit to the NHANES data set, compared to a logistic distribution with empirically estimated standard deviation.

The same variables were retained when model selection was performed on a logistic regression model.

Figure 2.7 shows the histogram of residuals for the AFT model. As with the previous data set, the distribution of the residuals is greatly skewed to the right. Reasons for the observed skewness might lie in a combination of a natural skewness of the fasting plasma glucose measure, measurement and recording errors, as well as the fact that potentially not all measures were in fact taken after at least 9 hours of fasting. Fitting the model in (2.11) using the proposed EB method, the Bayes factor derived by the goodness of fit test is $10^{568.9}$, indicating decisive evidence against a logistic error distribution.

Figure 2.8 plots the odds ratio for diabetes comparing female versus male across a range of BMI values for individuals that are 20, 40 and 60 years old. The logistic regression and AFT models did not indicate an interaction between gender and BMI or age, therefore odds ratios are estimated to be constant by these two methods. However, odds ratio estimates from the EB method are decreasing with BMI values, indicating that differences in odds between genders increase with BMI. Specifically, odds of diabetes are greater for men than for women, and this gender effect increases

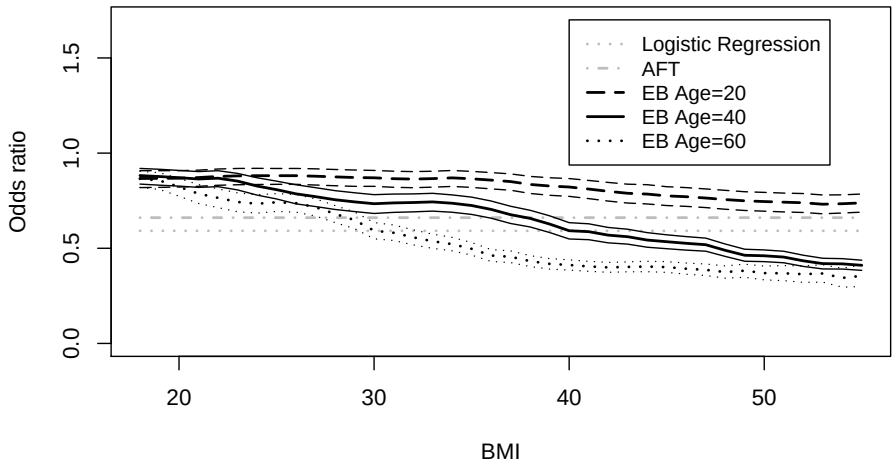
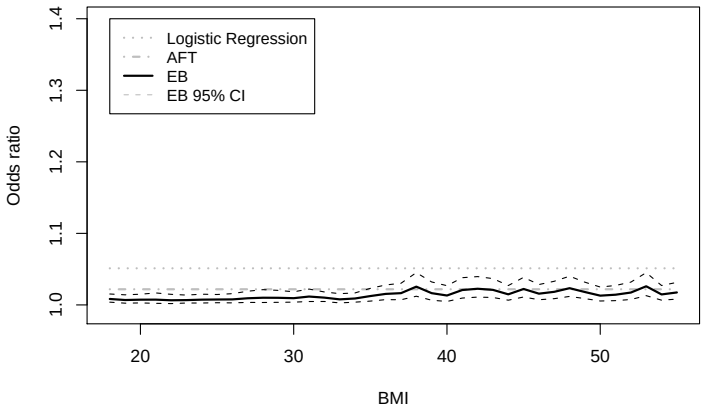


Figure 2.8: Odds ratios for gender estimated by the EB method with 95% credible intervals for ages 20, 40 and 60, compared to estimates from logistic regression and an accelerated failure time model (AFT).

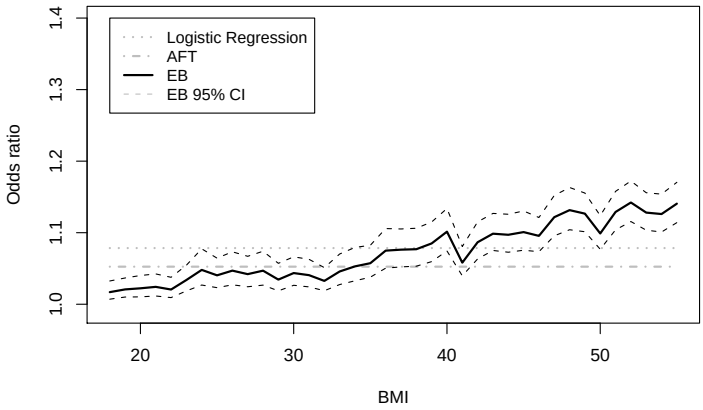
with BMI. Also, odds ratio estimates differ greatly between ages. With increasing age, the effect of gender appears to be greater, particularly for large BMI values.

Figure 2.9 presents graphs of odds ratio estimates comparing two male individuals who differ by one BMI unit for ages 20, 40 and 60 years. Due to the age-by-BMI interaction that all models include, odds ratio estimates differ with age even for the logistic regression and AFT models. However, in the EB approach an additional interaction between age and BMI becomes apparent. Overall, the adverse effect of increased BMI becomes more dramatic in higher ranges of BMI, and additionally this effect is exacerbated with increasing age.

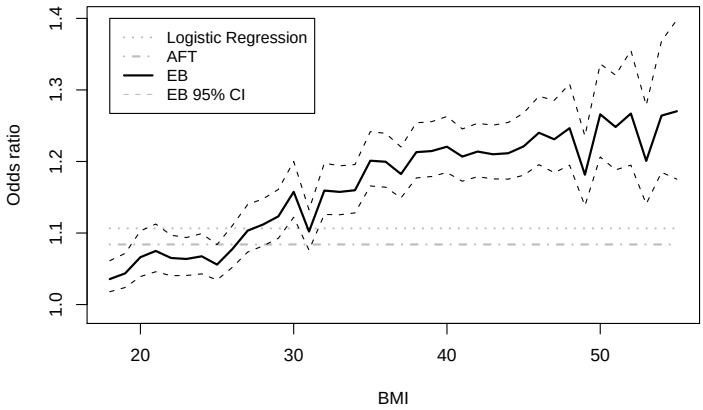
More strongly than in the example of risk estimation on data from the Health and Retirement Study, differences in inference based on the three models become apparent in this analysis of an NHANES data set. In the linear models, interactions between gender and age or gender and BMI were not significant, although the EB model suggests that these variables do not affect odds independently. An interaction between age and BMI is modeled even by the two parametric models, however, this interaction



(a) 20 years



(b) 40 years



(c) 60 years

Figure 2.9: Estimated odds ratios of diabetes for males differing by one unit in BMI for different ages based on EB methods with 95% credible intervals (CI).

does not seem to completely capture the interplay between the two variables.

2.8 Summary

Building upon the observation that risk estimation is more efficient when based on continuous measurement data rather than dichotomized data, this chapter presented a method for flexible risk estimation for any arbitrary residual distribution. A one-step test procedure can verify the goodness of fit of a logistic error distribution and a semiparametric estimation framework models risk-related measures without parametric assumptions about the shape of the error distribution. Theoretical results have demonstrated the consistency of our goodness of fit EB test. Simulations have demonstrated that if the true underlying distribution is not logistic, the increased flexibility to model the distribution found in the data results in reduced bias in risk estimates compared to accelerated failure time and logistic regression models. The applications of our novel semiparametric model to subsets of two large-scale surveys show that deviations from parametric assumptions can be found in data sets that have been the basis of many investigations, and that modeling the error distributions nonparametrically can lead to inferences different from those based on parametric models.

The Polya tree model that has been proposed here can be implemented in a traditional Bayesian fashion using a Gibbs sampler, but as a fast alternative an estimation procedure that samples only from the posterior distribution of the residual distribution has been proposed. This Empirical Bayes procedure results in dramatic computational savings but equivalent results compared to the full Bayesian implementation. The intended advantage of such a computational approach is that application of the model is facilitated for practitioners, as both programming effort and computational time are reduced. Additionally, with capacities to collect and store data ever increasing, full MCMC implementations for large scale analyses are often no longer feasible and approximate solutions are necessary. Simulations presented in this chap-

ter show that with large sample sizes, results from the Empirical Bayes method are equivalent to those from the full MCMC implementation, so that our method does not compromise statistical inference.

Chapter 3 A Dependent Polya Tree Model for Regression Analysis

This chapter introduces a semiparametric model for linear regression analysis using dependent Polya trees. The methods are particularly suited to longitudinal repeated measures data. By modeling error distributions at consecutive time points using separate, but dependent Polya tree priors, distributional information can be pooled across time points while allowing for enough flexibility to accommodate changes in error distributions.

3.1 Introduction

Standard linear models assume a homoscedastic, parametric error distribution, which generally does not represent patterns in real data. Due to the central limit theorem and its variations, the effect of deviations from these assumptions may not dramatically affect estimation of mean response, however, proper models are of special importance in prediction of individual observations [MacEachern, 1999].

In many scenarios, not only are error distributions not parametric, but they also are not homogenous across covariates. In the case of a longitudinal study following different treatments groups, for example, individuals within a certain group may respond differently to the treatment, resulting in a skewed or multimodal distribution. There may also be a time-varying treatment effect, resulting in changes in the response distribution over time.

The method that will be presented here is a new way of describing correlated error distributions, irrespective of the median structure. The method is highly flexible in that it can be incorporated into a wide range of models, such as linear and nonlinear fixed effects, random effects, and mixed effects models. Standard posterior inferences about model parameters are still possible in the sense of median regression, but more flexibility and predictive accuracy are gained by modeling the error distribution nonparametrically.

Section 3.2 gives an overview of previous approaches to introducing dependencies among prior distributions on function spaces. A new dependent Polya tree model and its theoretical properties are outlined in Section 3.3. The performance of the proposed model on simulated data and on real data is evaluated in Sections 3.4 and 3.5.

3.2 Background

The model that will be developed in this chapter introduces a dependence structure among residual distributions, which can be incorporated into a model with arbitrary mean structure. In the most general setting, we consider a nonlinear mixed model

$$y|u \sim f_{y|u}(g(X, Z; \beta, u), \Sigma) \tag{3.1}$$

where β is a vector of fixed effects, u a vector of random effects, X and Z are the fixed-effects and random-effects covariate matrices, and Σ a collection of scale parameters.

A special case of this model is ordinary linear regression. The ordinary linear mixed model

$$y|u \sim N(X'\beta + Z'u, \Sigma), \qquad u \sim N(0, \Sigma_u)$$

is also a special case, along with the commonly used random intercept, random slope, and random intercept and slope models.

In equation 3.1, the density f is traditionally modeled parametrically, although examples listed in Chapter 1 show that the density can be extended to be nonparametric within the Bayesian framework. In the simplest models, the scale and shape of f are homogeneous across covariate values, and observations are independently distributed, which will not accommodate many scenarios encountered with real data.

Several models for introducing dependencies among random probability distribu-

tions have been suggested and employed in a variety of applications. One of the earliest attempts to define dependencies among Dirichlet processes was presented in Cifarelli and Regazzini [1978], who introduced a regression for the baseline measure G_{0x} of marginally DP-distributed random measures in a product of mixtures of Dirichlet processes [De Iorio et al., 2004]. Such models were used by Muliere and Petrone [1993], who defined a regression of the base measure, $G_{0x} = N(x\beta, \sigma^2)$, and defined the dependent processes as $F_x \sim DP(\alpha G_{0x})$, and Carota and Parmigiani [2002], who employed regression for the baseline measure for count data.

MacEachern [1999] defined a dependent Dirichlet process (DDP) based on Sethuraman's representation of a Dirichlet process (see Section 1.1.1), by defining the point masses θ_i as realizations of stochastic processes $\theta_{ix}, x \in \mathcal{X}$. Gelfand et al. [2005] applied this idea to a model for point-referenced spatial data using realizations of Gaussian processes. Additional applications of this idea include linear regression $\theta_{ix} = x_i'\beta$ in the context of an ANOVA model [De Iorio et al., 2004] and nonproportional hazards survival modeling [De Iorio et al., 2009]. Additionally, the shape of the base measure G_{0x} may vary with $x \in \mathcal{X}$, and variates v_i from the Sethuraman construction of the DP may also be replaced by stochastic processes, $v_{i\mathcal{X}}$, resulting in marginal distributions $G_x \sim DP(\alpha_x G_{0x})$. Griffin and Steel [2006] propose a related dependent Dirichlet process, in which dependence is introduced by modeling the parameters of the beta distributions of the weights v_i as a function of covariates. MacEachern [1999] points out that a similar approach, replacing countable sets of variates by stochastic processes, can be applied to other nonparametric methods, such as Polya trees.

Dunson et al. [2007] propose a weighted mixture of Dirichlet process priors, by mixing over independent samples G_{x_j} from a Dirichlet process with common base measure G_0 and α. The weight function is dependent on the relative distances between the covariate values. For the estimate of G_x, greater weight is assigned to those G_{x_j} for which x and x_j are close. The approaches for dependent Dirichlet processes

listed here introduce a relationship between covariates and elements of the Dirichlet process prior, and then mix a smooth kernel $f(y_i|x_i) = \int_\Psi f(y_i|x_i, \psi)G_{x_i}(d\psi)$ over the nonparametric distribution G_{x_i}.

Until recently, little work on dependent Polya trees had been done. Jara and Hanson [in press] propose a class of dependent processes centered around the idea of Polya trees, or more generally, around tailfree processes, that regress density shape on predictors. Rather than modeling branch probabilities Y_{jk} in a tree as arising from beta distributions, they are defined as $Y_{jk} = h\{\eta_{jk}(x, \omega)\}$, where h is a strictly increasing continuous function with range [0,1]. Specifically, Jara and Hanson use the logistic link $Y_{jk}(x, \omega) = \exp\{\eta_{jk}(x, \omega)\}/[1 + \exp\{\eta_{jk}(x, \omega)\}]$. Although this model is not technically a Polya tree process, marginal asymptotic equivalence can be shown for the logistic link if the η_{jk}'s are realizations from independent zero-mean Gaussian processes [Jara and Hanson, in press, Prop. 3]. While this model presents a way to introduce dependencies across continuous covariate values, in the case of categorical covariates, such as discrete points of observation in longitudinal or repeated measures data, the resulting distributions are independent.

3.3 Dependent Polya tree model

In the case of continuous response data, either from repeated measures studies or cross-sectional studies with ordinal covariates indexed by t, error distributions G_t may evolve both in terms of scale and shape across levels of the covariate $t = 1, \ldots, T$. For concreteness, our presentation focuses on longitudinal data, but other data structures can be modeled. We consider the following model:

$$
\begin{aligned}
y_{it} &= g(x_{it}, z_{it}; \beta, u) + \varepsilon_{it} \\
\varepsilon_{it} &\sim G_t \\
G_t &\sim FPT(c, \rho(j), G_0).
\end{aligned}
$$

Additionally, the Polya tree prior for each G_t may be based on a different centering distribution $G_{0(t)}$, which can, for example, differ by a scale parameter σ_t. However, due to temporal or spatial dependencies, the error distributions might not be completely independent for different t. In fact, if we suspect a gradual evolution of the error process over time, information can be gained by letting estimates of the error distribution at time point t be informed by error distributions at previous time points.

Dependencies between error distributions might be modeled by introducing a Markov-type relationship between the branch probabilities of Polya trees at consecutive time points. We model G_1 using a standard Polya tree prior: $G_1 \sim PT(cj^2, G_{0(1)})$. In keeping with standard choices for mixed models, and to potentially test for deviations from parametric normal models, the centering distribution $G_{0(1)}$ might be chosen to be normal, i.e., define the prior distribution $G_1|\sigma_1 \sim PT\big(cj^2, N(0, \sigma_1)\big)$.

Starting at $t = 2$, the dependent Polya tree (DPT) prior distribution on G_t is defined such that it is dependent on the Polya tree at point $t-1$. For $t > 1$, $G_t|G_{t-1} \sim DPT(cj^2, G_{0(t)})$, where G_{t-1} and G_t are related through their branch probabilities. Left branch probabilities $Y_{e_j(k),t}, k = 1, 3, ..., 2^j - 1$ are modeled conditional on the corresponding branch probability at time $t - 1$:

$$Y_{e_j(k),t}|Y_{e_j(k),t-1} \sim \text{beta}(cj^2 Y_{e_j(k),t-1}, cj^2(1 - Y_{e_j(k+1),t-1})). \tag{3.2}$$

A sequence of Polya tree priors defined in this way and truncated at a finite level J will be referred to as *dependent finite Polya trees* (DFPT). The posterior distributions of the conditional branch probabilities are updated analogously to updating in a simple Polya tree, i.e., for $k = 1, 3, \ldots, 2^j - 1$

$$Y_{e_j(k),t}|Y_{e_j(k),t-1}, y \sim \text{beta}\big(cj^2 Y_{e_j(k),t-1} + n(j,k,y), cj^2(1 - Y_{e_j(k),t-1}) + n(j,k+1,y)\big) \tag{3.3}$$

3.3.1 Theoretical structure

The prior expected value of each branch probability at time t for $t > 1$ is equal to the corresponding branch probability at time $t-1$:

$$E[Y_{e_j(k),t}|Y_{e_j(k),t-1}] = \frac{cj^2 Y_{e_j(k),t-1}}{cj^2 Y_{e_j(k),t-1} + cj^2(1 - Y_{e_j(k),t-1})} = Y_{e_j(k),t-1}.$$

The prior variance of each branch probability at time $t = 1$ is

$$Var(Y_{e_j(k),1}) = \frac{c^2 j^4}{(2cj^2)^2(2cj^2+1)} = \frac{1}{4(2cj^2+1)},$$

while for branch probabilities at all following time points it is

$$\begin{aligned} Var[Y_{e_j(k),t+1}|Y_{e_j(k),t}] &= \frac{c^2 j^4 Y_{e_j(k),t}(1 - Y_{e_j(k),t})}{[cj^2 Y_{e_j(k),t} + cj^2(1 - Y_{e_j(k),t})]^2(cj^2 Y_{e_j(k),t} + cj^2(1 - Y_{e_j(k),t}) + 1)} \\ &= \frac{c^2 j^4 Y_{e_j(k),t}(1 - Y_{e_j(k),t})}{c^2 j^4 (cj^2+1)} \\ &= \frac{Y_{e_j(k),t}(1 - Y_{e_j(k),t})}{(cj^2+1)}. \end{aligned}$$

The covariance between each pair of corresponding branch probabilities $Y_{e_j(k),t}$ and $Y_{e_j(k),t+1}$ given $Y_{e_j(k),t-1}$ is

$$Cov[Y_{e_j(k),t}, Y_{e_j(k),t+1}|Y_{e_j(k),t-1}] = \frac{Y_{e_j(k),t-1}(1 - Y_{e_j(k),t-1})}{(cj^2+1)},$$

and the correlation is

$$Cor[Y_{e_j(k),t}, Y_{e_j(k),t+1}|Y_{e_j(k),t-1}] = Cor[Y_{e_j(k),t}, Y_{e_j(k),t+1}] = \sqrt{\frac{cj^2+1}{2cj^2+1}}. \tag{3.4}$$

Proof. Let $u = Y_{e_j(k),t-1}$,$v = Y_{e_j(k),t}$, $w = Y_{e_j(k),t+1}$ and $B(v,w) = \frac{\Gamma(v)\Gamma(w)}{\Gamma(v+w)}$ be the beta

function.

$$\begin{aligned}
E[VW|U] &= \int\int vwp(w|v)p(v|u)dwdv \\
&= \int\int v\frac{w^{cj^2v}(1-w)^{cj^2(1-v)-1}}{B\big(cj^2v, cj^2(1-v)\big)}p(v|u)dwdv \\
&= \int vp(v|u)\frac{\Gamma\big(cj^2\big)\Gamma\big(cj^2v+1\big)\Gamma\big(cj^2(1-v)\big)}{\Gamma\big(cj^2v\big)\Gamma\big(cj^2(1-v)\big)\Gamma\big(cj^2+1\big)}dv \\
&= \int v^2p(v|u)dv \\
&= \int \frac{v^{cj^2u+1}(1-v)^{cj^2(1-u)-1}}{B\big(cj^2u, cj^2(1-u)\big)}dv \\
&= \frac{u(cj^2u+1)}{cj^2+1}.
\end{aligned}$$

Hence

$$Cov[V,W|U] = E[VW|U] - E[V|U]E[W|U] = \frac{u(cj^2u+1)}{cj^2+1} - u^2 = \frac{u(1-u)}{cj^2+1}.$$

By a similar derivation, $Var[W|U] = \frac{(2cj^2+1)u(1-u)}{(cj^2+1)^2}$ and therefore

$$Cor[V,W|U] = \frac{u(1-u)}{cj^2+1}\sqrt{\frac{(cj^2+1)(cj^2+1)^2}{[u(1-u)]^2(2cj^2+1)}} = \sqrt{\frac{(cj^2+1)}{(2cj^2+1)}}.$$

□

The posterior predictive density of a finite Polya tree with J levels is [Hanson and Johnson, 2002]

$$p(w|y) = g_0(w)2^J\prod_{j=2}^{J}\frac{cj^2+n(j,k,y)(w)}{2cj^2+n(j-1,k,y)(w)}, \tag{3.5}$$

where $n(j,k,y)(w)$ is the number of elements in the data vector y that fall into the same partition as w at level j of the Polya tree.

For a finite Polya tree, with dependencies modeled as in (3.2), the predictive

density is defined as in (3.5) only for the first time point. For all following time points, the posterior predictive density is

$$\begin{aligned} p(w|y, G_{t-1}) &= g_0(w) \prod_{j=1}^{J} 2 \cdot P\Big[w \in B_0\big(j, k(j,w)\big)\Big] \\ &= g_0(w) \prod_{j=1}^{J} 2 \cdot Y_{e_j(k),t} \\ &\approx g_0(w) 2^J \prod_{j=1}^{J} \frac{cj^2 Y^*_{e_j(k),t-1} + n(j,k,y)(w)}{cj^2 Y^*_{e_j(k),t-1} + cj^2(1 - Y^*_{e_j(k),t-1}) + n(j-1,k,y)(w)} \\ &= g_0(w) 2^J \prod_{j=1}^{J} \frac{cj^2 Y^*_{e_j(k),t-1} + n(j,k,y)(w)}{cj^2 + n(j-1,k,y)(w)} \end{aligned}$$

where $Y^*_{e_j(k),1} = \frac{2cj^2+2n(j,k,y)(w)}{2cj^2+n(j-1,k,y)(w)}$ and $Y^*_{e_j(k),t} = \frac{2cj^2 Y^*_{e_j(k),t-1}+2n(j,k,y)(w)}{cj^2+n(j-1,k,y)(w)}$. The approximation in the third step of this derivation follows from Lavine [1992, Theorem 2].

3.4 Simulation study

Performance of the proposed dependent Polya tree model is investigated for a longitudinal data model, with three scenarios of evolving distributions: a normal distribution changing into a skewed distribution or a bimodal distribution over time, and a distribution changing its shape from left to right skewed. Additionally, we consider scenarios in which the shape of the error distribution remains the same over time, but the scale may vary. The predictive performance of the proposed model is compared to two alternative Polya tree models, namely a model that defines independent Polya tree priors for the error distribution at each time point, and a single common Polya tree prior for errors across all time points.

3.4.1 Error models

The general model fit to each simulated data set is

$$y_{it} = \mu_t + \varepsilon_{it}$$
$$(\mu_1, \ldots, \mu_T) \sim N(0, I_T \cdot 100).$$

The distribution of the errors ε_{it} is modeled according to the proposed dependent Polya tree prior, as well as by two alternative finite Polya tree models. The three models are defined as follows.

Dependent finite Polya tree model (DFPT)

We model the distribution of ε_{it} as G_t, relating the distributions on the individual G_t as:

$$\begin{aligned}
\varepsilon_{it}|G_t &\overset{iid}{\sim} G_t \\
G_1|\sigma_1 &\sim FPT(cj^2, N(0, \sigma_1)) \\
G_t|\sigma_t &\sim DFPT(cj^2, N(0, \sigma_t), G_{t-1}) \text{ for } t \geq 2 \\
\sigma_t^2 &\sim p(\sigma_t^2) \text{ for } t = 1, \ldots, 4.
\end{aligned}$$

Independent finite Polya tree model (IFPT)

If great changes in error distributions are expected, and preceding time points are not expected to inform the shape of the distribution at subsequent time points, one might consider imposing an independent Polya tree prior at each time point, resulting

in the model

$$\begin{aligned}
\varepsilon_{it}|G_t &\overset{iid}{\sim} G_t \\
G_t|\sigma_t &\sim FPT(cj^2, N(0,\sigma_t)) \text{ for } t = 1,\ldots,4 \\
\sigma_t &\sim p(\sigma_t^2) \text{ for } t = 1,\ldots,4.
\end{aligned}$$

Single finite Polya tree model

The two previous models are compared to a model with errors arising from a single error distribution G_1 across all time points.

$$\begin{aligned}
\varepsilon_{it}|G_1 &\overset{iid}{\sim} G_1 \text{ for } t = 1,\ldots,4 \\
G_1|\sigma_1 &\sim FPT(cj^2, N(0,\sigma_1)) \\
\sigma_1 &\sim p(\sigma_1^2).
\end{aligned}$$

3.4.2 Simulation settings

To simulate a scenario in which the shape of the error distribution changes from a normal distribution into a right-skewed distribution, data y_{it} were generated according to $y_{i1} \sim N(0,1)$, and for $t = 2,3,4$, $y_{it} = u_{it} + v_{it}$, where $u_{it} \sim N(0,1)$, and $v_{it} \sim \Gamma(1, t-1)$. A sequence of distributions that change from a normal distribution into a bimodal distribution over time was generated with $y_{i1} \sim N(0,1)$, and for $t = 2,3,4$, $y_{it} \sim N(a \cdot m_t, 0.5)$ where $m_2 = .4$, $m_3 = .8$, $m_4 = 1.2$ and $a = -1$ with probability 0.5 and $a = 1$ otherwise. Finally, a distribution that changes its direction of skew was generated from a skew normal distribution, which was first introduced as a distributional family by O'Hagan and Leonard [1976]. Using notation from Azzalini [1985], the density function of a skew-normal random variable Z is defined as

$$f(z|\lambda) = 2\phi(z)\Phi(\lambda z) \qquad\qquad (-\infty < z < \infty)$$

where ϕ and Φ are the density and distribution function of a standard normal distribution, respectively. For the simulation scenario described here, the shape parameter was chosen to be $\lambda = -6, -3, 3,$ and 6 for $t = 1, 2, 3,$ and 4, respectively, resulting in a distribution that changes from left to right skewed. For the two cases in which the shape of the error distribution did not change over time, errors were modeled as coming from a skew-normal distribution with $\lambda = 3$ that was scaled to ensure standard deviation $\sigma_t = t/4$ or $\sigma_t = 1$ for $t = 1, \ldots, 4$, respectively.

Data were generated for sample sizes $n = 50, 100$ and 200 by selecting $\left[(100i - 50)/n\right]^{th}$ percentiles for $i = 1, \ldots, n$ of the respective distributions. This ensures samples in which the empirical distribution function at each observation corresponds to the true generating distribution function. Finite Polya trees were truncated at depth $J = 4$, and the parameter c was fixed at 1. Prior distributions for σ_t were chosen to be $\Gamma(2, 2)$. Samples from the posterior distributions were generated from a Gibbs sampler using the Metropolis-Hastings algorithm. Chains stabilized very quickly and were run for $50,000$ iterations after a burn-in period of $5,000$ iterations.

Posterior iterates for μ_t were generated using a random-walk Metropolis-Hastings algorithm, while iterates of σ_t were generated using an independence chain in Metropolis-Hastings. The likelihood function for the DFPT model is

$$\begin{aligned} L(\mu, \sigma, \mathcal{Y}_1, \ldots, \mathcal{Y}_4) &= p(y|\beta, \sigma, \mathcal{Y}_1, \ldots, \mathcal{Y}_4) \\ &= \prod_{t=1}^{4} \prod_{i=1}^{n} g(\varepsilon_{it}|\mathcal{Y}_t, \sigma_t) \\ &= \prod_{t=1}^{4} \prod_{i=1}^{n} 2^J p_{\mathcal{Y}}(k_{\sigma_t}(J, \varepsilon_{it})) g_{\sigma_t}(\varepsilon_{it}) \end{aligned}$$

where $\mu = (\mu_1, \ldots, \mu_4)'$ and $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)'$.

The full conditional distribution for μ is

$$p(\mu|y,\sigma,\mathcal{Y}_1,\ldots,\mathcal{Y}_4) \propto L(\mu,\sigma,\mathcal{Y}_1,\ldots,\mathcal{Y}_4)p(\mu_1,\ldots,\mu_4). \tag{3.6}$$

The resulting acceptance probability for μ^* generated from the proposal distribution $N_4(\mu, 0.2\cdot I_4)$ given the previous iterate μ in a random-walk Metropolis-Hastings step is

$$\min\Big\{1, \frac{\prod_{t=1}^4\prod_{i=1}^n g(\hat{\varepsilon}^*_{it}|\mathcal{Y}_t,\sigma_t)}{\prod_{t=1}^4\prod_{i=1}^n g(\hat{\varepsilon}_{it}|\mathcal{Y}_t,\sigma_t)}\frac{\phi_4(\mu^*)}{\phi_4(\mu)}\Big\} \tag{3.7}$$

where ϕ_4 is the density function of $N_4(0, I_4\cdot 100)$. Iterates of $\sigma_1,\ldots,\sigma_4$ were generated separately using an independence chain algorithm. The full conditional for each σ_t is

$$p(\sigma_t|y,\mu,\mathcal{Y}_1,\ldots,\mathcal{Y}_4) \propto L(\mu,\sigma,\mathcal{Y}_1,\ldots,\mathcal{Y}_4)p(\sigma_t) \tag{3.8}$$

and the acceptance probability for σ_t^* generated from the proposal distribution $\Gamma(2,1)$ with density function $\gamma_{2,1}$ is

$$\begin{aligned}
&\min\Big\{1, \frac{\prod_{i=1}^n p_{\mathcal{Y}_t}(k_{\sigma_t^*}(J,\varepsilon_{it}))g_{\sigma_t^*}(\varepsilon_{it})}{\prod_{i=1}^n p_{\mathcal{Y}_t}(k_{\sigma_t}(J,\varepsilon_{it}))g_{\sigma_t}(\varepsilon_{it})}\frac{\gamma_{2,2}(\sigma_t^*)}{\gamma_{2,2}(\sigma_t)}\frac{\gamma_{2,1}(\sigma_t)}{\gamma_{2,1}(\sigma_t^*)}\Big\}\\
=&\min\Big\{1, \frac{\prod_{i=1}^n p_{\mathcal{Y}_t}(k_{\sigma_t^*}(J,\varepsilon_{it}))g_{\sigma_t^*}(\varepsilon_{it})}{\prod_{i=1}^n p_{\mathcal{Y}_t}(k_{\sigma_t}(J,\varepsilon_{it}))g_{\sigma_t}(\varepsilon_{it})}\frac{\Gamma(2)2^2\sigma_t^*\exp(-\sigma_t^*/2)}{\Gamma(2)2^2\sigma_t\exp(-\sigma_t/2)}\frac{\Gamma(2)1^2\sigma_t\exp(-\sigma_t/1)}{\Gamma(2)1^2\sigma_t^*\exp(-\sigma_t^*/1)}\Big\}\\
=&\min\Big\{1, \frac{\prod_{i=1}^n p_{\mathcal{Y}_t}(k_{\sigma_t^*}(J,\varepsilon_{it}))g_{\sigma_t^*}(\varepsilon_{it})}{\prod_{i=1}^n p_{\mathcal{Y}_t}(k_{\sigma_t}(J,\varepsilon_{it}))g_{\sigma_t}(\varepsilon_{it})}\exp\frac{\sigma_t^*-\sigma_t}{2}\Big\}.
\end{aligned}$$

3.4.3 Results

Model fit was evaluated using log-pseudo marginal likelihood (see section 1.2). LPML values for the three models and five distributional scenarios are listed in Table 3.1 and predictive error densities generated by the three models are presented in Figures 3.2, 3.3 and 3.4.

For data with homoscedastic error across all four time points, the single finite Polya tree model, which indeed models one common error distribution, performs best in terms of LPML. Out of the two alternative models, the DFPT model has slightly higher LPML, indicating that borrowing information across time points does improve the model fit compared to a model with four independent trees. These results demonstrate that the proposed model presents a compromise between independent error distributions and a single error distribution for all time points. For identically distributed error, the single FPT model appropriately pools information from residuals across all time points to estimate their distribution. The DFPT takes advantage of only part of this information, resulting in lower LPML values. The IFPT model, on the other hand, considers residuals at each time point separately, and thus performs the worst among the models considered here.

For all four scenarios with changing error distributions, the model with a dependent finite Polya tree prior has the highest LPML values across all simulated scenarios. For skewed error with increasing variance, the DFPT model results in LPML values that are 0.9, 2.6 and 4.8 units higher than the IFPT model, for sample sizes 50, 100 and 200, respectively. This results in Bayes factors on the $\log_{10}$-scale of 0.4, 1.1 and 2.1, indicating strong evidence for the predictive superiority for sample sizes 100 and 200. Visually, the differences in posterior density estimates are subtle (Figure 3.1).

For the case in which error distributions were constructed to evolve from normal to skewed or from normal to bimodal, the differences in LPML values between the dependent Polya tree and the independent Polya tree models are substantial, of a magnitude between 4.2 and 7.6, which corresponds to Bayes factors on the $\log_{10}$-scale between 1.8 and 3.3 and indicates superior predictive power of the newly proposed model. The differences in predictive densities (Figures 3.2 and 3.3) are especially apparent at later time points.

For the scenario of error distributions that change shape from left skewed to

right skewed, the differences in the distributions generating the data were far more subtle than for the previous two scenarios. But even with such a slight evolution in distributions over time, the DFPT model continues to outperform both alternative models in terms of LPML. The differences in LPML values are of smaller magnitude – in the range of 0.7 to 3.2 compared to the IFPT model, and 2.3 to 10.1 compared to the SFPT models. Visually, the differences between predictive densities generated by the DFPT and IFPT models (Figure 3.4) are also much smaller than in the previous two scenarios.

By construction of the simulation scenarios, the model with only a single Polya tree prior for errors across all time points is the least appropriate model of the three considered here, and does indeed result in the lowest values of LPML for all sample sizes and all four scenarios in which the error distribution evolves. Represented by the dotted lines in Figures 3.1 through 3.4, it is clear that this simple model is not able to capture the shape of the true error distribution in any of these scenarios.

Simulation results show that in scenarios in which error distributions change over time, predictive power may be gained by pooling information across time points. Even in scenarios in which the shape of the error distribution changed dramatically (e.g., the distribution changed the direction of the skew, or developed a bimodal shape), introducing dependencies outperformed the model that assumed completely independent error distributions.

Table 3.1: LPML values for model comparison for three sample sizes n.

		n		
Distribution	Model	50	100	200
right skewed	Dependent Finite Polya Trees	-284.7	-562.6	-1116.7
homoscedastic	Independent Finite Polya Trees	-285.4	-565.2	-1122.1
	Single Finite Polya Tree	-281.7	-559.5	-1114.0
right skewed	Dependent Finite Polya Trees	-166.4	-326.2	-643.7
heteroscedastic	Independent Finite Polya Trees	-167.3	-328.8	-648.5
	Single Finite Polya Tree	-198.8	-391.3	-774.9
normal	Dependent Finite Polya Trees	-390.2	-774.2	-1524.2
to right skew	Independent Finite Polya Trees	-394.4	-781.0	-1531.5
	Single Finite Polya Tree	-411.0	-814.0	-1600.7
normal	Dependent Finite Polya Trees	-376.6	-745.6	-1485.3
to bimodal	Independent Finite Polya Trees	-381.0	-751.9	-1492.9
	Single Finite Polya Tree	-438.9	-880.5	-1692.3
left to	Dependent Finite Polya Trees	-192.2	-376.5	-737.3
right skew	Independent Finite Polya Trees	-192.9	-378.7	-740.5
	Single Finite Polya Tree	-194.5	-386.4	-767.4

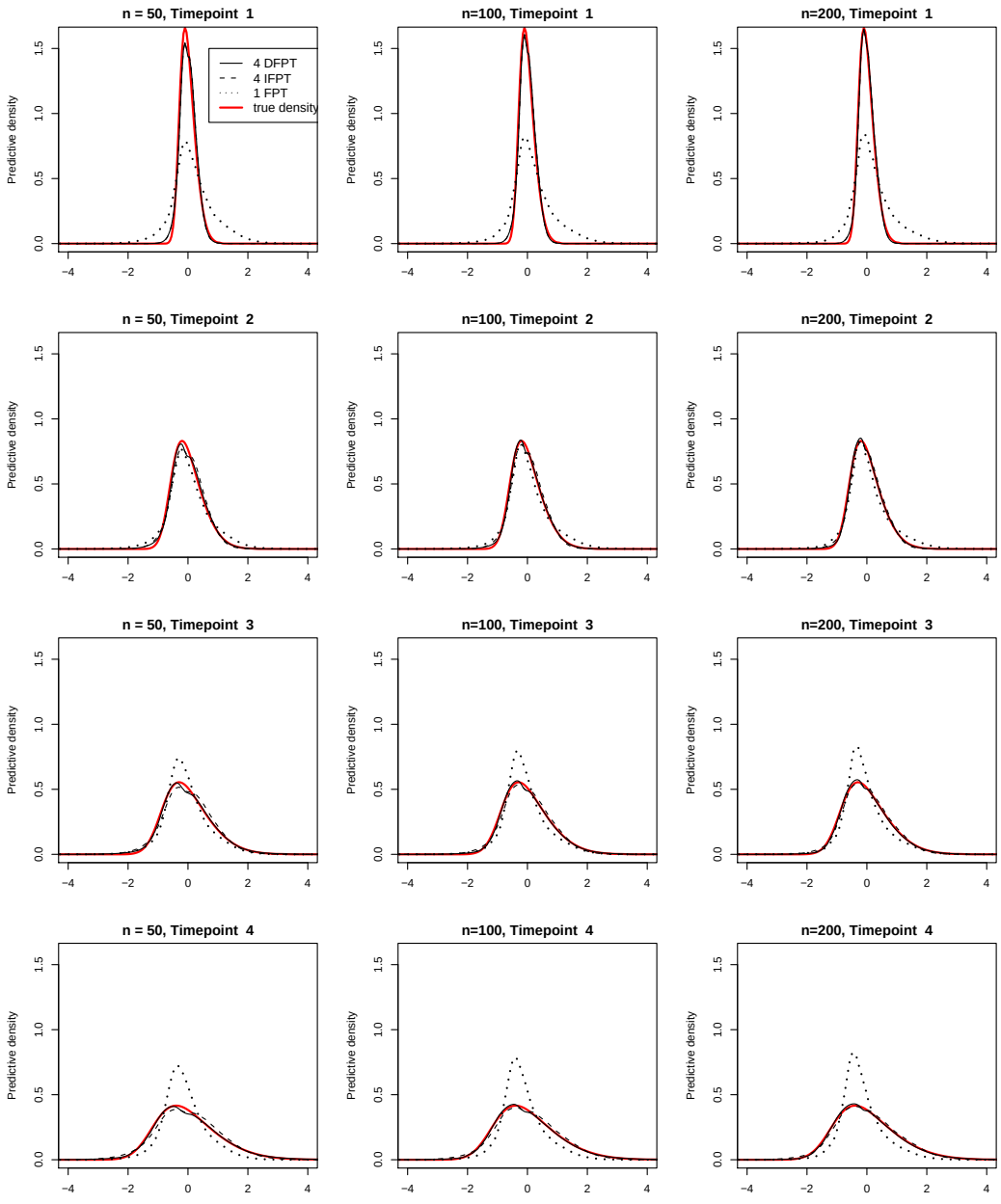


Figure 3.1: Predictive error densities from a dependent finite Polya tree model (DFPT), an independent finite Polya tree model (IFPT), and a single finite Polya tree model (FPT) compared to the true error density, which is a skew-normal distribution with skew equal to 0.66 and variance increasing over time.

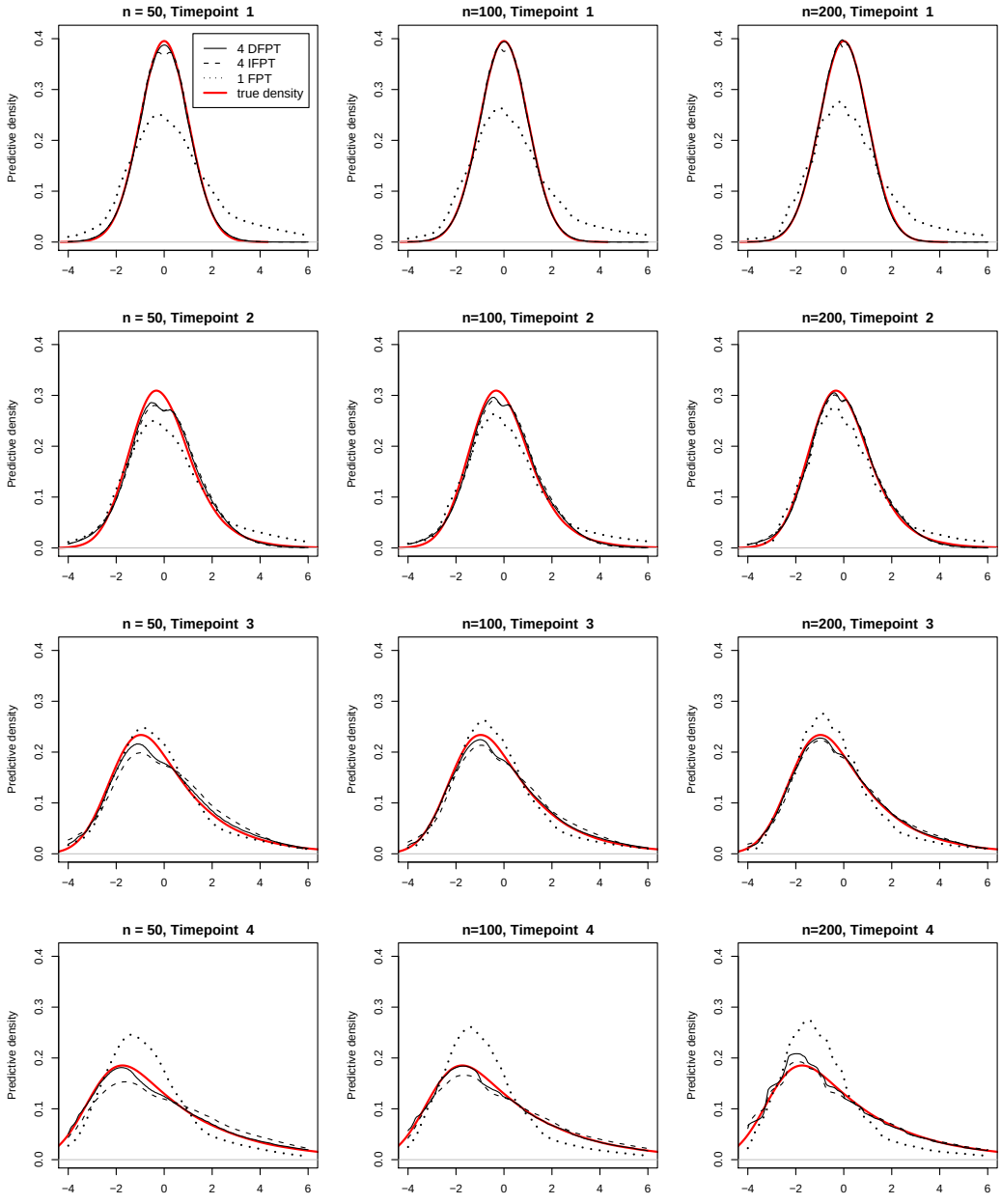


Figure 3.2: Predictive error densities from a dependent finite Polya tree model (DFPT), an independent finite Polya tree model (IFPT), and a single finite Polya tree model (FPT) compared to the true error density, which changes from a normal to a right skewed distribution over time.

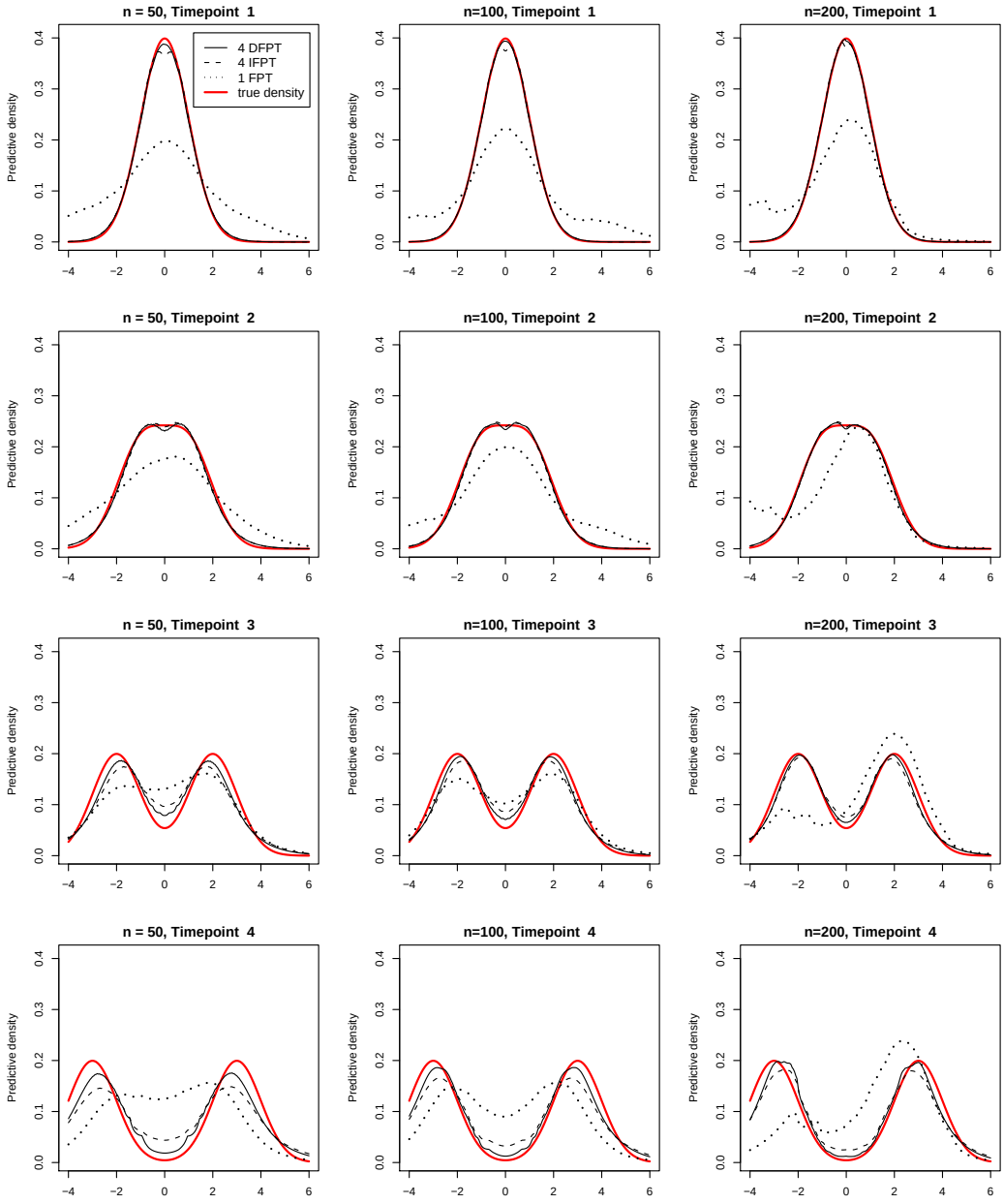


Figure 3.3: Predictive error densities from a dependent finite Polya tree model (DFPT), an independent finite Polya tree model (IFPT), and a single finite Polya tree model (FPT) compared to the true error density, which changes from a normal to a bimodal error distribution over time.

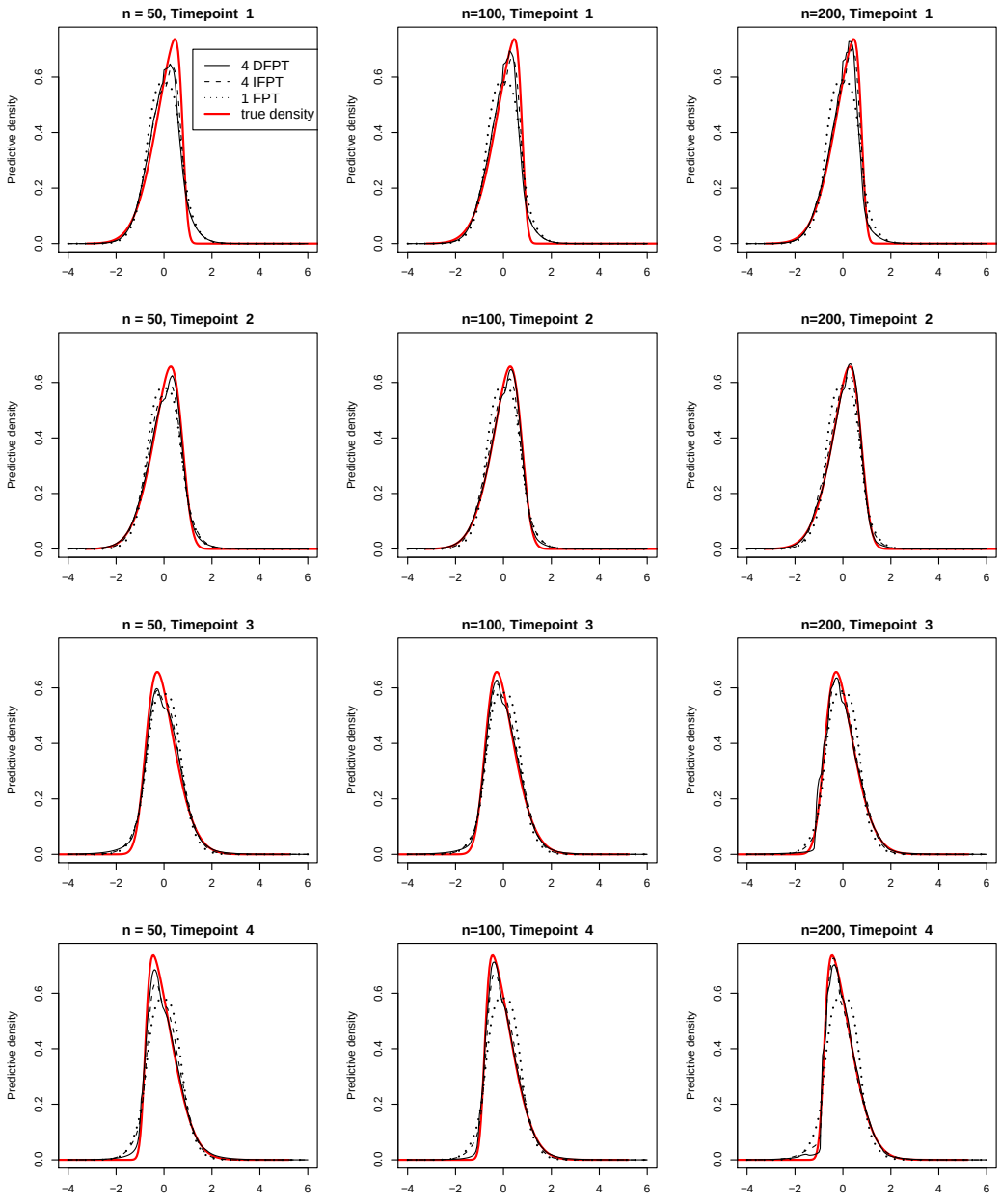


Figure 3.4: Predictive error densities from a dependent finite Polya tree model (DFPT), an independent finite Polya tree model (IFPT), and a single finite Polya tree model (FPT) compared to the true error density, which changes from a left skewed to a right skewed distribution over time.

3.5 Example

3.5.1 Growth data

Jara and Hanson [in press] demonstrate the performance of their proposed model of dependent tailfree processes on a data set previously explored by Royston and Wright [1998] and Kapitula and Bedrick [2005]. The data set consists of measurements of the serum concentration of immunoglobulin G (IgG) for 298 children between the ages of 6 and 72 months [Isaacs et al., 1983]. All 298 observations are independent, i.e., there are no true repeated measures on the same individual in the data set. Therefore, unlike in the simulation study of longitudinal data, this analysis illustrates how our method can be applied to modeling ordinary regression error with a dependent Polya tree.

Kapitula and Bedrick [2005] fit an exponential normal growth model, which is a parametric approach to estimating percentile curves, to the log-transformed IgG values using the covariate age (x). Under their model, the density for the log-transformed response is

$$f(z_i) = \frac{1}{\sqrt{2\pi}} \exp\Big[-\frac{\{\exp(\gamma z_i) - 1\}^2}{2\gamma^2} + \gamma z_i\Big] \frac{1}{\Phi(1/|\gamma|)} \tag{3.9}$$

where $z_i = (\log y_i - \mu_i)/\sigma_i$, $\sigma_i = \theta_0 + \theta_1 x_i^{-2}$, $\Phi(\cdot)$ is the standard normal distribution function, and

$$\mu_i = \beta_0 + \beta_1 x_i^2 + \beta_2 x_i^{-2} + \varepsilon_i. \tag{3.10}$$

Jara and Hanson [in press] fit the median regression model

$$\log(y_i) = \beta_0 + \beta_1 \sqrt{x_i} + \beta_2 x_i^{-2} + \varepsilon_i, \tag{3.11}$$

Table 3.2: Age groups defined for the IgG data set.

Age group $a(x)$	Age in months (x)	Age group $a(x)$	Age in months (x)
1	6-11	4	36-47
2	12-23	5	48-59
3	24-35	6	60-72

and define $\varepsilon_i | G_{x_i} \overset{\text{indep}}{\sim} G_{x_i}$ where $\{G_x : x \in \mathcal{X}\}$ has a dependent tailfree process prior. It is not clear that the difference in the median function between the two models was intentional, and in what follows the two alternatives will be compared.

We allow for changes in error distributions over time by modeling a different nonparametric error distribution for the six age groups defined in Table 3.2. The distribution of ages in the data set (see Figure 3.5) shows spikes at 6, 12, 24, and 36 months of age, suggesting that for a large number of children, age was possibly not recorded precisely but rounded to full years. The resulting models are thus

$$\begin{aligned} \log(y_i) &= \beta_0 + \beta_1 x_i^2 + \beta_2 x_i^{-2} + \varepsilon_i \\ \varepsilon_i &\sim G_t, \quad t = a(x_i) \end{aligned} \tag{3.12}$$

where $a(x_i)$ is the age group for subject i, and

$$\begin{aligned} \log(y_i) &= \beta_0 + \beta_1 \sqrt{x_i} + \beta_2 x_i^{-2} + \varepsilon_i \\ \varepsilon_i &\sim G_t, \quad t = a(x_i). \end{aligned} \tag{3.13}$$

Posterior estimates were generated using a Gibbs sampler with Metropolis-Hastings sampling. The prior distribution for $\beta = (\beta_0, \beta_1, \beta_2)$ was $N(0, I_3 \cdot 100)$ and a $\Gamma(2, 2)$ prior was used for the standard deviations σ_t of the centering distributions of the

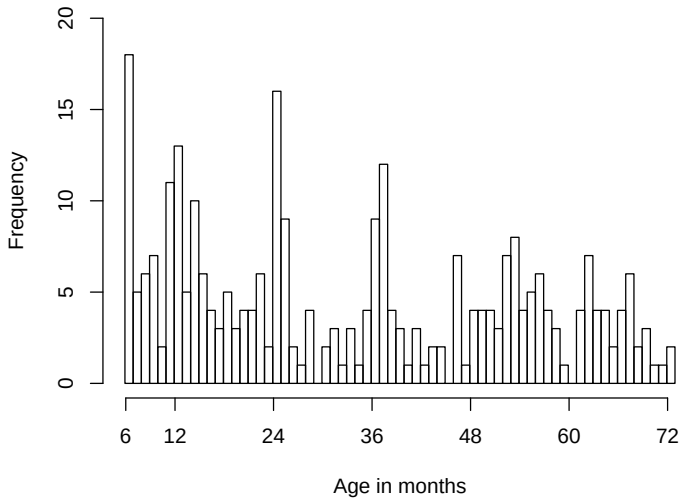


Figure 3.5: Distribution of ages of 298 children

Polya trees. The likelihood function for the dependent Polya tree model is

$$\begin{aligned} L(\beta, \sigma, \mathcal{Y}_1, \ldots, \mathcal{Y}_6) &= p(y|\beta, \sigma, \mathcal{Y}_1, \ldots, \mathcal{Y}_6) \\ &= \prod_{t=1}^{6} \prod_{i \in A_t} g(\varepsilon_i | \mathcal{Y}_t, \sigma_t) \\ &= \prod_{t=1}^{6} \prod_{i \in A_t} 2^J p_{\mathcal{Y}}(k_\sigma(J, \varepsilon_i)) g_\sigma(\varepsilon_i) \end{aligned}$$

where $A_t = \{i : a(x_i) = t\}$ is the set of indices for the observations that fall into age category t. The full conditional distribution for $\beta = (\beta_0, \beta_1, \beta_2)'$ is

$$p(\beta|\sigma, \mathcal{Y}_1, \ldots, \mathcal{Y}_6) \propto L(\beta, \sigma, \mathcal{Y}_1, \ldots, \mathcal{Y}_6) p(\beta),$$

resulting in an acceptance of the proposal β^* generated from $N(\beta, \Sigma_\beta)$ in a random-

walk Metropolis-Hastings step with probability

$$\min\Big\{1, \frac{\prod_{t=1}^{6}\prod_{i\in A_t} p_{\mathcal{Y}_t}(k_{\sigma_t}(J,\varepsilon_i^*))g_{\sigma_t}(\varepsilon_i^*)}{\prod_{t=1}^{6}\prod_{i\in A_t} p_{\mathcal{Y}_t}(k_{\sigma_t}(J,\varepsilon_i))g_{\sigma_t}(\varepsilon_i)} \frac{\phi(\beta^*/100)}{\phi(\beta/100)}\Big\}. \tag{3.14}$$

Each σ_t was sampled in an independent step from the full conditional distribution

$$p(\sigma_t|\beta,\mathcal{Y}_t) \propto L(\beta,\sigma,\mathcal{Y}_1,\ldots,\mathcal{Y}_6)p(\sigma_t),$$

and consequently the acceptance probability of the proposal σ_t^* generated from the proposal distribution $\Gamma(2,1)$ in the independence chain algorithm is

$$\begin{aligned}&\min\Big\{1, \frac{\prod_{i\in A_t} p_{\mathcal{Y}_t}(k_{\sigma_t^*}(J,\varepsilon_i))g_{\sigma_t^*}(\varepsilon_i)}{\prod_{i\in A_t} p_{\mathcal{Y}_t}(k_{\sigma_t}(J,\varepsilon_i))g_{\sigma_t}(\varepsilon_i)} \frac{\gamma_{2,2}(\sigma_t^*)}{\gamma_{2,2}(\sigma_t)} \frac{\gamma_{2,1}(\sigma_t)}{\gamma_{2,1}(\sigma_t^*)}\Big\} \\ &\min\Big\{1, \frac{\prod_{i\in A_t} p_{\mathcal{Y}_t}(k_{\sigma_t^*}(J,\varepsilon_i))g_{\sigma_t^*}(\varepsilon_i)}{\prod_{i\in A_t} p_{\mathcal{Y}_t}(k_{\sigma_t}(J,\varepsilon_i))g_{\sigma_t}(\varepsilon_i)} \exp\Big\{\frac{\sigma_t^*-\sigma_t}{2}\Big\}\Big\}.\end{aligned} \tag{3.15}$$

The full conditionals for the branch probabilities in sets $\mathcal{Y}_1,\ldots,\mathcal{Y}_6$ are derived as outlined in previous sections. The Gibbs sampler was run for 150,000 iterations after discarding iterates from a burn-in period of 50,000 iterations. Polya tree parameters were fixed at $J = 4$ and $c = 1$ or $c = 0.5$. The covariance matrix Σ_β for the proposal distribution of β was updated to a scaled version of the empirical covariance matrix based on the first 25,000 iterations of the burn-in period to improve mixing of the chains. The scale factor was chosen to be 0.2, which allowed the acceptance probabilities for each parameter to lie within the range of approximately 0.2 to 0.35.

Table 3.3 presents the LPML values obtained by the two median models considered here. For both models, $c = 0.5$ results in a clear improvement over $c = 1$. Simulations with $c = 0.1$ were attempted, but failed to mix appropriately. Comparing the two different median structures, the model proposed by Kapitula and Bedrick [2005] (see equation 3.12) performs better, with a difference in LPML of about 3 for either value of c, which results in a Bayes factor of $10^{1.3}$, indicating substantial evidence in favor

Table 3.3: LPML values obtained for the IgG data set from the DFPT model.

Model	c	LPML
$\mu(x_i) = \beta_0 + \beta_1 x_i^2 + \beta_2 x_i^{-2}$	1	-118.8
	0.5	-114.2
$\mu(x_i) = \beta_0 + \beta_1 \sqrt{x_i} + \beta_2 x_i^{-2}$	1	-121.8
	0.5	- 117.4

Table 3.4: Parameter estimates for two models from the DFPT model, $J = 4$, $c = 0.5$.

Model	Covariate	Estimate	95% Credible interval
$\mu(x_i) = \beta_0 + \beta_1 x_i^2 + \beta_2 x_i^{-2}$	β_0	1.508	[1.419, 1.663]
	β_1	0.018	[0.008, 0.022]
	β_2	-0.141	[-0.223,-0.056]
$\mu(x_i) = \beta_0 + \beta_1 \sqrt{x_i} + \beta_2 x_i^{-2}$	β_0	1.089	[0.879, 1.380]
	β_1	0.349	[0.201, 0.484]
	β_2	-0.087	[-0.167, -0.004]

Kapitula and Bedrick's model.

Parameter estimates from the two DFPT models for $c = 0.5$ are presented in Table 3.4. The DFPT estimates obtained for Kapitula and Bedrick's median model are comparable to the point estimates reported from their analysis ($\hat{\beta}_0 = 1.569$, $\hat{\beta}_1 = 0.013$ and $\hat{\beta}_2 = -0.167$).

Predictive densities calculated by the two models for five arbitrary time points as well as the estimated median functions and their 95% credible intervals are presented in Figure 3.6. Differences in the two models become particularly apparent with increasing age. Starting at about 50 months, the two median functions deviate, and differences in predictive densities become more prominent.

Jara and Hanson [in press] compared the performance of their proposed model in terms of LPML to several alternative models Jara and Hanson [in press], including a model with normal error on the log-scale, the exponential normal model suggested by Kapitula and Bedrick [2005], a Dirichlet Process mixture model, and a linear

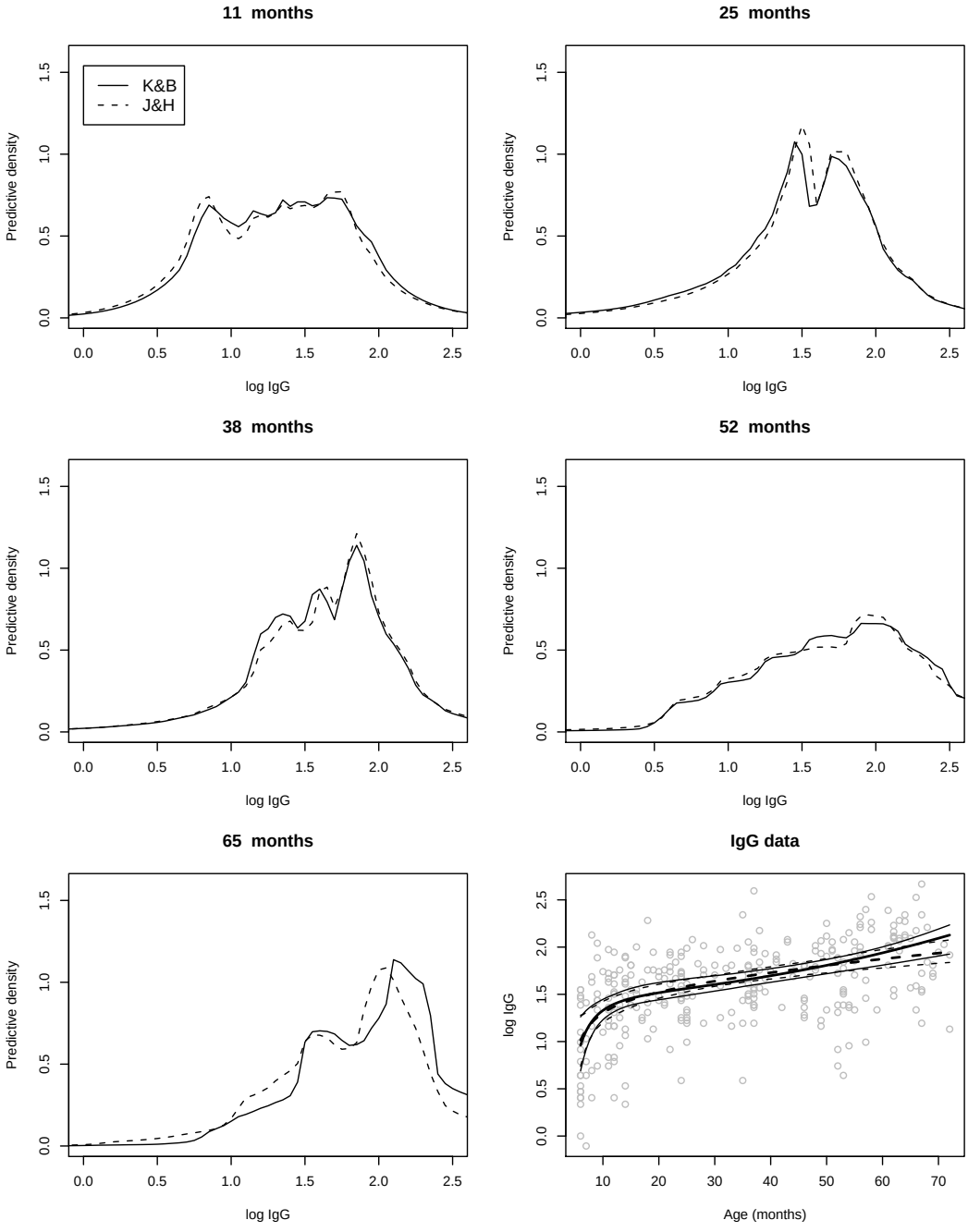


Figure 3.6: Predictive densities for five selected time points and estimated median function with 95% credible interval from the DFPT model ($J = 4, c = 0.5$) for median functions 3.12 (K&B, solid lines) and 3.13 (J&H, dashed lines).

dependent Dirichlet Process model [De Iorio et al., 2004, 2009]. Table 3.5 lists the LPML values that Jara and Hanson reported for these models.

Clearly, the Normal error model on the log-scale and the Dirichlet process mixture model, both of which model a uniform error distribution, are not appropriate for this data set, having the smallest LPML values of the models considered here. Overall, the models employing Polya trees seem to outperform all other models in this comparison. There is a significant advantage over modeling the error distribution using a linear dependent Dirichlet process. However, an important finding is that the proposed DFPT model, which models data at discrete time intervals, has equivalent predictive power as the dependent tailfree process proposed by Jara and Hanson, which models a continuously evolving error distribution.

Table 3.5: LPML values from various models for the IgG data set presented by Jara and Hanson [in press]) for the median function $\mu(x_i) = \beta_0 + \beta_1\sqrt{x_i} + \beta_2 x_i^{-2}$.

Model	LPML
Dependent Tailfree Process	-121
Normal Error on log-scale	-143
Exponential Normal Model	-136
Dirichlet Process Mixture	-143
Linear Dependent Dirichlet Process	-139

3.6 Summary

A novel approach to defining dependent priors on error distributions in regression models has been presented in this chapter. For data scenarios in which error distributions slowly evolve over time or over an ordinal covariate, the proposed model allows for information to be pooled across ordered points. The model specifically assumes an ordered relationship between points, i.e., the shape of the error distribution at time point t may be informed by the shape at time point $t-1$.

In particular in the context of describing predictive densities for individual observations, flexible modeling of the error distribution becomes important. This model

poses a solution between the two extremes of homogeneous error and independent error distributions for each category, and simulations have shown that for gradual changes in distributions, this method results in increased predictive accuracy.

Chapter 4 Summary and Outlook

Two semiparametric generalizations of popular linear models using Polya trees have been presented in this dissertation. We have demonstrated that in the case of deviations from parametric assumptions, a one-step approximation of the mean structure still leads to a consistent goodness of fit test and reduced bias in estimation. The second proposed method is a novel approach to introducing dependencies between nonparametric error models.

Chapter 2 presented a generalized model for risk estimation that extends the parametric idea of logistic regression for a dichotomized response that is based on measurement data. The gains in efficiency resulting from modeling the continuous response rather than the dichotomized outcome leads to smaller samples sizes required to detect an effect of a given size. To take advantage of such savings in terms of sample size, a method for sample size calculation for this semiparametric approach needs to be developed.

This model can easily be extended to more general ordinal responses. For example, risk factors for an individual falling into either of the categories *overweight* or *obese* can be modeled simultaneously. In each sampling step, the nonparametric risk would simply be calculated for both cutoffs. The Empirical Bayes version of the Savage-Dickey ratio provides a powerful test to detect deviations from a logistic distribution, but equally applies to other parametric distributions. For example, by centering the Polya tree prior on G at a normal rather than a logistic distribution, a test for the appropriateness of probit regression can be implemented, while the framework for nonparametric risk estimation would remain the same as described above.

In the context of epidemiologic studies, it would be of great interest to extend the risk estimation procedure to various sampling schemes, such as case-control studies. The additional challenge of incorporating priors beliefs about outcome prevalence would need to be addressed carefully. For clinical diagnoses that are based on mul-

tiple outcomes, a multivariate extension of the approach could be developed. For example, the diagnosis of diabetes is generally based on measurements of several different factors. We could formulate a general rule that a patient is diagnosed with a disease if at least a certain number of factors indicate the disease. Then, the risk of an individual falling above the critical cutoff for at least that number of indicators could be calculated. Recent developments in the area of multivariate Polya trees [Trippa et al., 2011, Hanson et al., 2011] might be extended to such a model.

In its current form, the semiparametric model for risk estimation assumes a fixed error distribution. Future work will develop a more flexible model that allows for changes in error distribution with covariate values. The basic idea for nonparametric risk estimation would remain the same, however, the finite Polya tree prior for the error distribution would be extended to accommodate such changes. Methods developed in Chapter 3 are an obvious starting point for such a model.

Chapter 3 presented a novel approach to introducing dependencies between Polya tree distributions associated with ordered covariates. The approach might be extended to allow for covariate-dependent effects on the shape of the distribution. Pursuing an extension of the proposed approach with some aspect of the dependent tailfree processes developed by Jara and Hanson [in press] would be of great interest.

Another potential extension of the model is a generalization of the dependency structure to define relationships in multiple dimensions to model, for example, spatio-temporal processes. In addition to the theoretical development of such an approach, computational challenges specific to Polya trees would need to be addressed, as an increase in dimensions greatly increases the computational cost of any MCMC implementation of the model.

Chapter A Appendix

A.1 Notation

Table A.1: Polya tree notation

$e_j(k)$	the j-fold binary representation of the number $k-1$
$B_\theta(j,k)$	$= (G_\theta^{-1}((k-1)2^{-j}), G_\theta^{-1}(k2^{-j})]$ for $k = 1, ..., 2^j$
	the k-th set in the partition of Ω at level j
$Y_{e_j(k)}$	random branch probability: conditional probability of set $B_\theta(j,k)$
$\mathcal{Y}$	set of random branch probabilities $Y_{e_j(k)}$, $j = 1, 2, ...$
Π_θ^J	$= \{B_\theta(j,k) : j = 1, ..., J, k = 1, ...2^j\}$
	collection of partitions induced by G_θ up to level J
$n_\theta(j,k,y)$	the number of observations in y that fall into set $B_\theta(j,k)$
$k_\theta(j,y)$	$\in 1, ..., 2^j$, index of set on level j into which observation y falls
$p_{\mathcal{Y}}(k)$	$= G\{B_\theta(J,k)\|\mathcal{Y},\theta\} = \prod_{j=1}^{J} Y_{e_j(\lceil k2^{j-J}\rceil)}$
$g(y\|\mathcal{Y},\theta)$	$= 2^J p_{\mathcal{Y}}(k_\theta(J,y)) g_\theta(y)$, Polya tree density

Table A.2: Distributions

$N(\mu,\sigma)$	Normal distribution with mean μ and standard deviation σ
logistic(μ,σ)	Logistic distribution with mean μ and standard deviation σ
$\Gamma(a,b)$	Gamma distribution with shape parameter a and scale parameter b

Bibliography

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

A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985.

E. Bakhshi, M.R. Eshraghian, K. Mohammad, and B. Seifi. A comparison of two methods for estimating odds ratios: Results from the national health survey. *BMC Medical Research Methodology*, 8(78):published online, 2008.

A. Barron, M.J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

S. Basu and S. Chib. Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association*, 98(461):224–235, 2003.

E.J. Bedrick, R. Christensen, and W. Johnson. A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, 91(436): 1450–1460, 1996.

J.O. Berger and A. Guglielmi. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, 96(453):174–184, 2001.

A.J. Branscum and T.E. Hanson. Bayesian nonparametric meta-analysis using Polya tree mixture models. *Biometrics*, 64(3):825–833, 2008.

C.A. Bush and S.N. MacEachern. A semiparametric Bayesian model for randomized block designs. *Biometrika*, 83(2):275–285, 1996.

C. Carota. Some faults of the Bayes factor in nonparametric model selection. *Statistical Methods and Applications*, 15(1):37–42, 2006.

C. Carota and G. Parmigiani. *Bayesian statistics*, volume 5, chapter On Bayes factors for nonparametric alternatives, pages 507–511. Oxford University Press, London, 1996.

C. Carota and G. Parmigiani. A Dirichlet process elaboration diagnostic for binomial goodness of fit. *Sociedad de Estadística e Investigación Operativa*, 7(1):133–145, 1998.

C. Carota and G. Parmigiani. Semiparametric regression for count data. *Biometrika*, 89(2):265–281, 2002.

C. Carota, G. Parmigiani, and N.G. Polson. Diagnostic measures for model criticism. *Journal of the American Statistical Association*, 91(434):753–762, 1996.

G. Casella and E.I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

CDC Centers for Disease Control and National Center for Health Statistics (NCHS) Prevention. National Health and Nutrition Examination Survey data. U.S. Department of Health and Human Services, Hyattsville, MD, 2009. URL `http://www.cdc.gov/nchs/nhanes.htm`.

D.M. Cifarelli and E. Regazzini. Problemi statistici non parametrici in condizioni di scambiabilitá parziale: impiego di medie associative. Technical report Ser. III, 1–36, Istituto di Matematica Finanziaria dell'Universitá di Torino, 1978.

S.C. Dass and J. Lee. A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives. *Journal of Statistical Planning and Inference*, 119(1):143–152, 2004.

M. De Iorio, P.M. Müller, G.L. Rosner, and S.N. MacEachern. An ANOVA model for dependent random measures. *Biometrics*, 99(465):205–215, 2004.

M. De Iorio, W.O. Johnson, P. Müller, and G.L. Rosner. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771, 2009.

D. Denison and B.K. Mallick. *Bayesian Statistics and its Applications*, chapter Analyzing financial data using Polya trees, pages 122–132. Anamaya Publishers, New Delhi, 2006.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.

D.B. Dunson, N. Pillai, and J.-H. Park. Bayesian density regression. *Journal of the Royal Statistical Society, Series B*, 69(2):163–183, 2007.

R.L. Dykstra and P. Laud. A Bayesian nonparametric approach to reliability. *The Annals of Statistics*, 9(2):356–367, 1981.

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

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

J. Fabius. Asymptotics behavior of Bayes' estimates. *The Annals of Mathematical Statistics*, 35(2):846–856, 1964.

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

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

D.A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403, 1963.

S. Geisser. Discussion on sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, 143(4):416–417, 1980.

S. Geisser and W.F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

A.E. Gelfand. *Asymptotics, nonparametrics, and time series*, chapter Approaches for Semiparametric Bayesian Regression, pages 615–638. Marcel Dekker, New York, 1^{st} edition, 1999.

A.E. Gelfand and D.K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B*, 56(3):501–514, 1994.

A.E. Gelfand and A. Kottas. A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11(2):289–305, 2002.

A.E. Gelfand, A. Kottas, and S.N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

S. Ghosal, J.K. Ghosh, and R.V. Ramamoorthi. Consistent semiparametric estimation about a location parameter. *Journal of Statistical Planning and Inference*, 77 (2):181–193, 1998.

S. Ghosal, J.K. Ghosh, and R.V. Ramamoorthi. *Asymptotic Nonparametrics and Time Series: A Tribute to Madan Lal Puri*, chapter Consistency issues in Bayesian nonparametrics, pages 639–667. Marcel Dekker, New York, 1999.

S. Ghosal, J. Lember, and A.W. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008.

A. Gould and J.F. Lawless. Consistency and efficiency of regression coefficient estimates in location-scale models. *Biometrika*, 75(3):535–540, 1988.

C. Goutis and C.P. Robert. Choice among hypotheses using estimation criteria. *Annals of Economics and Statistics*, 46, 1997.

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

T.E. Hanson. Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101(476):1548–1565, 2006.

T.E. Hanson and W.O. Johnson. Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Associations*, 97(460):1020–1033, 2002.

T.E. Hanson, J. Sethuraman, and L. Xu. On choosing the centering distribution in Dirichlet process mixture models. *Statistics & Probability Letters*, 72(2):153–162, 2005.

T.E. Hanson, A.J. Branscum, and I.A. Gardner. Multivariate mixtures of Polya trees for modeling ROC data. *Statistical Modelling*, 8(1):81–96, 2008.

T.E. Hanson, J.V.D. Monteiro, and A. Jara. The Polya tree sampler: toward efficient and automatic independent Metropolis-Hastings proposals. *Journal of Computational and Graphical Statistics*, 20(1):41–62, 2011.

Health and Retirement Study. Wave I/Year 1992, Public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG09740). Ann Arbor, MI, 1992. URL `http://hrsonline.isr.umich.edu`.

N.L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *the Annals of Statistics*, 18(3):1259–1294, 1990.

D.W. Hosmer and S. Lemeshow. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley & Sons, Inc., New York, U.S., 1999.

D. Isaacs, D.G. Altman, C.E. Tidmarsh, H.B. Valman, and A.D.B. Webster. Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: reference ranges for igg, iga, igm. *Journal of Clinical Pathology*, 36:1193 – 1196, 1983.

A. Jara and T. Hanson. A class of mixtures of dependent tailfree processes. *Biometrika*, in press.

H. Jeffreys. *Theory of Probability*. Oxford, The Clarendon Press, Oxford, UK, 3^{rd} edition, 1961.

J.D. Kalbfleisch. Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B*, 40(2):214–221, 1978.

L.R. Kapitula and E.J. Bedrick. Diagnostics for the exponential normal growth curve model. *Statistics in Medicine*, 24(1):95–108, 2005.

R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

K.P. Kleinman and J.G. Ibrahim. A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine*, 17(22):2579–2596, 1998.

A. Kottas and A.E. Gelfand. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96(456):1458–1468, 2011.

C.H. Kraft. A class of distribution function processes which have derivatives. *Journal of Applied Probability*, 1(2):385–388, 1964.

C.H. Kraft and C. van Eeden. Bayesian bio-assay. *The Annals of Mathematical Statistics*, 35(2):886–890, 1964.

L. Kuo and B. Mallick. Bayesian semiparametric inference for the accelerated failure-time model. *The Canadian Journal of Statistics*, 25(4):457–472, 1997.

T.L. Lai, H. Robbins, and C.Z. Wei. Strong consistency of least squares estimates in multiple regression. *Proceedings of the National Academy of Science*, 75(7): 3034–3036, 1978.

M. Lavine. Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 20(3):1222–1235, 1992.

M. Lavine. More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 22(3):1161–1176, 1994.

S.N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association*, pages 50–55, Alexandria, VA, 1999. American Statistical Association.

S.N. MacEachern and P. Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

B.K. Mallick and A.E. Gelfand. Semiparametric errors-in-variables models - a Bayesian approach. *Journal of Statistical Planning and Inference*, 52:307–321, 1996.

H.H. Malluche and M.C. Monier-Faugere. Renal osteodystrophy: what's in a name? Presentation of a clinically useful new model to interpret bone histologic findings. *Clinical Nephrology*, 65(4):235–242, 2006.

R.D. Mauldin, W.D. Sudderth, and S.C. Williams. Polya trees and random distributions. *The Annals of Statistics*, 20(3):1203–1221, 1992.

R. McVinish, J. Rousseau, and K. Mengersen. Bayesian goodness of fit testing with mixtures of triangular distributions. *Scandinavian Journal of Statistics*, 36:337–354, 2009.

B.K. Moser and L.P. Coombs. Odds ratios for a continuous outcome variable without dichotomizing. *Statistics in Medicine*, 23(12):1843–1860, 2004.

P. Muliere and S. Petrone. A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models. *Journal of the Italian Statistical Society*, 3:349–364, 1993.

P. Müller and F.A. Quintana. Nonparametric Bayesian data analysis. *Statistical Science*, 19(1):95–110, 2004.

NCHS National Center for Health Statistics; Centers for Disease Control and Prevention. About the National Health and Nutrition Examination Survey data. U.S. Department of Health and Human Services, Hyattsville, MD, 2011. URL `http://www.cdc.gov/nchs/nhanes/about_nhanes.htm`.

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

A. O'Hagan and T. Leonard. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63(1):201–203, 1976.

S.M. Paddock, F. Ruggeri, M. Lavine, and M. West. Randomized Polya tree models for nonparametric Bayesian inference. *Statistica Sinica*, 13:443–460, 2003.

M.L. Pennell and D.B. Dunson. Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics*, 62:1044–1052, 2006.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`.

D.R. Ragland. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutoff point. *Epidemiology*, 3(5): 434–440, 1992.

P. Royston and E.M. Wright. A method for estimating age-specifc reference intervals ('normal ranges') based on fractional polynomials and exponential transformation. *Journal of the Royal Statistical Society, Series A*, 161:79 –101, 1998.

S. Selvin. Two issues concerning the analysis of grouped data. *European Journal of Epidemiology*, 3(3):284–287, 1987.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2), 1994.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B*, 64 (4):583–639, 2002.

L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

L. Trippa, P. Müller, and W. Johnson. The multivariate beta process and an extension of the Polya tree model. *Biometrika*, 98(1):17–34, 2011.

I.V. Verdinelli and L. Wasserman. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618, 1995.

I.V. Verdinelli and L. Wasserman. Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *The Annals of Statistics*, 26(4):1215 – 1241, 1998.

K. Viele. Nonparametric estimation of Kullback-Leibler information illustrated by evaluating goodness of fit. *Bayesian Analysis*, 2(2):239–280, 2007.

S. Walker, P. Damien, and P. Lenk. On priors with a Kullback-Leibler property. *Journal of the American Statistical Association*, 99(466):404–408, 2004.

S.G. Walker and B.K. Mallick. Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society, Series B*, 59(4):849–860, 1997.

S.G. Walker and B.K. Mallick. A Bayesian semiparametric accelerated failure time model. *Biometrics*, 55(2):477–483, 1999.

S.G. Walker, P. Damien, P.W. Laud, and A.F.M. Smith. Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society, Series B*, 61(3):485–527, 1999.

WHO World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hypoglycemia. World Health Organization, Geneva, Switzerland. International Diabetes Federation. Report of a WHO/IDF consultation, 2006. URL `http://www.idf.org/webdata/docs/WHO_IDF_definition_diagnosis_of_diabetes.pdf`.

M. Yang, T. Hanson, and R. Christensen. Nonparametric Bayesian estimation of a bivariate density with interval censored data. *Computational Statistics and Data Analysis*, 52(12):5202–5214, 2008.

L. Zhao, T.E. Hanson, and B.P. Carlin. Mixtures of Polya trees for flexible spatial frailty survival modelling. *Biometrika*, 96(2):263–276, 2009.

P. Zimmet, C. Cowie, J.-M. Ekoe, and J.E. Shaw. *International Textbook of Diabetes Mellitus*, chapter Classification of diabetes mellitus and other categories of glucose intolerance, pages 3–14. John Wiley and Sons Ltd, 3^{rd} edition, 2004.

Vita

Date of Birth: March 10, 1983

Place of Birth: Linz, Austria

Education

M.S. degree in Statistics, University of Kentucky, USA, May 2008
M.S. degree in Computer Science - Intelligent Systems, Vienna University of Technology, Austria, June 2006
B.S. degree in Computer Science - Data Engineering and Statistics, Vienna University of Technology, Austria, September 2004

Employment

Peer Training Leader, Department of Statistics, University of Kentucky
August 2010 - present

Primary Instructor, Department of Statistics, University of Kentucky
August 2009 - May 2010

Statistical consultant, College of Agriculture, University of Kentucky
August 2007 -May 2009

Teaching Assistant, Department of Statistics, University of Kentucky
August 2006 - June 2007

Instructor in Adult Education, *BBRZ Vienna* ("vocational education and rehabilitation center"), Vienna, Austria
September 2005 - March 2006

Internship at *Arithmetica - Actuarial and Financial Consulting Ltd.*, Vienna, Austria
February 2004

Publications

Refereed Publications

[7] Huggins, P., Johnson, C.K., **Schörgendorfer, A.**, Putta, S., Stromberg, A.J., Voss, S.R. "Identification of Differentially Expressed Thyroid Hormone Responsive Genes from the Brain of the Mexican Axolotl *(Ambystoma mexicanum)*", 5th Aquatic Animal Models of Human Disease Conference, Symposium Volume, Comparative Biochemistry and Physiology, accepted.

[6] **Schörgendorfer, A.**, Madden, L.V. and Bathke, A.C. "Choosing Appropriate Covariance Matrices for a Nonparametric Analysis of Factorials in Block Designs." Journal of Applied Statistics, 38(4): 833-850.

[5] Pike, A.C., Mueller, T.G., **Schörgendorfer, A.**, Luck, J.D., Shearer, S.A., and Karathanasis, A.D. "Locating Eroded Waterways with United States Geological Survey Elevation Data." Agronomy Journal 102(4), 2010: 1269–1273.

[4] Royse, J., Arthur, M., **Schörgendorfer, A.**, and Loftis, D.L. "Establishment and growth of oak (Quercus alb, Q. prinus) seedlings in burned and fire-excluded upland forests on the Cumberland Plateau." Forest Ecology and Management 260, 2010: 502 –510.

[3] Pike, A.C., Mueller, T.G., **Schörgendorfer, A.**, Shearer, S.A., and Karathanasis, A.D. "Erosion indices derived from terrain attributes using Logistic Regression and Neural Networks." Agronomy Journal 101(5), 2009: 1068 –1079.

[2] McEwan, R.W., Birchfield, M.K., **Schörgendorfer, A.**, and Arthur, M. "Leaf phenology and freeze tolerance of the invasive shrub Amur honeysuckle and potential native competitors." The Journal of the Torrey Botanical Society, 136(2), 2009: 212–220.

[1] Mercer, D.R., **Schörgendorfer, A.**, and Vandyke, R. "Sexual Differences in Larval Molting Rates in a Protandrous Mosquito (Diptera: Culicidae) Species, Aedes sierrensis." Journal of Medical Entomology 45(5), 2008: 861–866.

Proceedings

[4] Lee, B.D., Wilson, C.L., **Schörgendorfer, A.**, Haight-Maybriar,L., and Webb, J. "Subwatershed clustering based on geomorphic and human induced landscape modifications: the Licking river basin." Contributed talk at Kentucky Water Resources Annual Symposium, Lexington, Kentucky, March 22, 2010.

[3] Lee, B.D., **Schörgendorfer, A.**, and Linebach, C. "Watershed Clustering Based on Geomorphic and Human Induced Landscape Modifications: A Central Kentucky

Example." Contributed talk at Kentucky Water Resources Annual Symposium, Lexington, Kentucky, March 2, 2009.

[2] **Schörgendorfer, A.** and Elmenreich, W. "Extended Confidence-Weighted Averaging in Sensor Fusion." Proceedings of the Junior Scientist Conference 2006. Vienna, Austria, April 2006.

[1] Elmenreich, W. and **Schörgendorfer, A.** "Fusion of Continuous-Valued Sensor Measurements using Statistical Analysis." Proceedings of the International Symposium on Mathematical Methods in Engineering. Ankara, Turkey, April 2006.

Posters and Presentations

[7] **Schörgendorfer, A.**, Branscum, A., and Hanson, T. "A Bayesian nonparametric test for logistic distribution and odds ratio estimation without dichotomizing." Poster, Joint Statistical Meetings. Vancouver, Canada. August 2010.

[6] **Schörgendorfer, A.**, Branscum, A., and Hanson, T. "A Bayesian nonparametric test for logistic distribution and odds ratio estimation without dichotomizing." Poster, International Meetings on Bayesian Statistics. Valencia, Spain. June 2010.

[5] **Schörgendorfer, A.**, Branscum, A., and Hanson, T. "A Semiparametric Generalization of Logistic Regression with Continuous Response Data." Poster, Conference on Nonparametric Statistics and Statistical Learning. Columbus, Ohio. May 2010.

[4] **Schörgendorfer, A.**, Branscum, A., and Hanson, T. "A Bayesian Nonparametric Goodness of Fit Test for Logistic Regression with Continuous Response Data". Poster, ENAR 2010 Spring Meeting., New Orleans, Louisiana. March 2010.

[3] **Schörgendorfer, A.**, Madden, L.V., and Bathke A. "Heterogeneous Compound Symmetry in a Nonparametric Analysis of Block Designs." Contributed talk, Joint Statistical Meetings. Washington, D.C. August 2009.

[2] **Schörgendorfer, A.**, Madden, L.V., and Bathke A. "Heterogeneous Compound Symmetry in a Nonparametric Analysis of Block Designs." Contributed, Graduate Student Interdisciplinary Conference. University of Kentucky, Lexington, Kentucky. April 2009.

[1] **Schörgendorfer, A.**, and Elmenreich, W. "Extended Confidence-Weighted Averaging in Sensor Fusion." Contributed talk, Junior Scientist Conference 2006. Vienna, Austria. April 2006.

Honors

R.L. Anderson Teaching Award, Department of Statistics, University of Kentucky, April 2010

R.L. Anderson Research Award, Department of Statistics, University of Kentucky, April 2009

(Angela Schoergendorfer)

(Date)