Scalable Bayesian regression utilising marginal information



Tristan Daniel Gray-Davies

Supervisor: Prof. Chris Holmes

Department of Statistics

University of Oxford

This dissertation is submitted for the degree of

Doctor of Philosophy

Christ Church July 2017

I should like to dedicate this thesis to my family.

Declaration

Some work in this thesis is the product of a collaborative effort. Specifically:

- Chapter 2 is joint work with Stephen Walker and my supervisor Chris Holmes
- Chapter 3 is joint work with Francois Caron and Chris Holmes, which has been published in the Electronic Journal of Statistics (Gray-Davies *et al.*, 2016)
- Chapter 4 is joint work with James Watson, Luis Nieto-Barajas and Chris Holmes

Tristan Daniel Gray-Davies

July 2017

Acknowledgements

I thoroughly enjoyed my time spent in Oxford at the Department of Statistics, and should like to thank the people who supported me throughout my time here.

First of all, I sincerely thank my supervisor, Chris Holmes. His enthusiasm, creativity and originality opened up many interesting questions and ideas. He taught me a great deal over the course of my DPhil, and provided thoughtful guidance throughout. I particularly appreciate his openness and friendliness, which made the experience all the more enjoyable.

I am also very grateful to my collaborators, Francois Caron, Stephen Walker and Luis Nieto-Barajas for their help, and for being a pleasure to work with. I would especially like to thank Stephen Walker for his warm hospitality during my visit to the University of Texas at Austin.

In the department, I benefited greatly from the support and company of the Holmes group. I'd particularly like to thank my friends who helped me throughout: James Watson, Aimee Taylor, George Nicholson, Louis Aslett, Pierre Jacob and Jeremy Heng.

At Christ Church, I've been lucky enough to have met some great friends who have been a huge part of my student experience: Chris Lander, Alexandra Schultz, Tristan Smart, Alex Drong, Luke Lewis and Kate Roll. Thanks also to the Landers for giving me a place to stay in the final stages.

Above all, I should like to thank my parents for all their care, love and support over the years.

Abstract

This thesis explores approaches to regression that utilise the treatment of covariates as random variables. The distribution of covariates, along with the conditional regression model $Y \mid X$, define the joint model over (Y,X), and in particular, the marginal distribution of the response Y. This marginal distribution provides a vehicle for the incorporation of prior information, as well as external, marginal data. The marginal distribution of the response provides a means of parameterisation that can yield scalable inference, simple prior elicitation, and, in the case of survival analysis, the complete treatment of truncated data. In many cases, this information can be utilised without need to specify a model for X.

Chapter 2 considers the application of Bayesian linear regression where large marginal datasets are available, but the collection of response and covariate data together is limited to a small dataset. These marginal datasets can be used to estimate the marginal means and variances of Y and X, which impose two constraints on the parameters of the linear regression model. We define a joint prior over covariate effects and the conditional variance σ^2 via a parameter transformation, which allows us to guarantee these marginal constraints are met. This provides a computationally efficient means of incorporating marginal information, useful when incorporation via the imputation of missing values may be implausible. The resulting prior and posterior have rich dependence structures that have a natural "analysis of variance" interpretation, due to the constraint on the total marginal variance of Y. The concept of "marginal coherence" is introduced, whereby competing models place the same prior on the marginal mean and variance of the response. Our marginally constrained prior

can be extended by placing priors on the marginal variances, in order to perform variable selection in a marginally coherent fashion.

Chapter 3 constructs a Bayesian nonparametric regression model parameterised in terms of F_Y, the marginal distribution of the response. This naturally allows the incorporation of marginal data, and provides a natural means of specifying a prior distribution for a regression model. The construction is such that the distribution of the ordering of the response, given covariates, takes the form of the Plackett-Luce model for ranks. This facilitates a natural composite likelihood approximation that decomposes the likelihood into a term for the marginal response data, and a term for the probability of the observed ranking. This can be viewed as a extension to the partial likelihood for proportional hazards models. This convenient form leads to simple approximate posterior inference, which circumvents the need to perform MCMC, allowing scalability to large datasets. We apply the model to a US Census dataset with over $1,300,000$ data points and more than 100 covariates, where the nonparametric prior is able to capture the highly non-standard distribution of incomes.

Chapter 4 explores the analysis of randomised clinical trial (RCT) data for subgroup analysis, where interest lies in the optimal allocation of treatment $D(X)$, based on covariates. Standard analyses build a conditional model $Y \mid X, T$ for the response, given treatment and covariates, which can be used to deduce the optimal treatment rule. We show that the treatment of covariates as random facilitates direct testing of a treatment rule, without the need to specify a conditional model. This provides a robust, efficient, and easy-to-use methodology for testing treatment rules. This nonparametric testing approach is used as a splitting criteria in a random-forest methodology for the exploratory analysis of subgroups. The model introduced in Chapter 3 is applied in the context of subgroup analysis, providing a Bayesian nonparametric analogue to this approach: where inference is based only on the order of the data, circumventing the requirement to specify a full data-generating model. Both approaches to subgroup analysis are applied to data from an AIDS Clinical Trial.

Table of contents

1 Introduction **1**

1.1 Statistical fundamentals 1

1.2 Nonparametric statistics 4

1.2.1 Bayesian Nonparametrics 5

1.3 Regression 9

1.3.1 Survival analysis 11

1.4 Motivation 14

2 Constrained marginal variance priors for linear regression **17**

2.1 Introduction 17

2.2 Bayesian linear model 18

2.2.1 As a missing data problem 22

2.3 Fixed marginal variance prior 24

2.3.1 In conjunction with sparsity priors 31

2.4 Marginal coherence in variable selection 36

2.5 Illustrations 44

2.5.1 Simulation example 44

2.5.2 US Census application 49

2.6 Discussion 55

3 A Scalable Bayesian Nonparametric regression model **59**

3.1 Introduction . . . 59

3.2 Bayesian nonparametric regression . . . 60

3.2.1 Semiparametric regression . . . 61

3.2.2 Nonparametric regression . . . 63

3.3 The statistical model . . . 67

3.4 Approximations for posterior inference and prediction . . . 70

3.4.1 Asymptotics for the marginal composite posteriors . . . 72

3.4.2 Posterior predictive . . . 73

3.4.3 Connection to proportional hazards . . . 74

3.5 Illustrations . . . 77

3.5.1 Simulation example . . . 78

3.5.2 US Census application . . . 82

3.6 Discussion . . . 88

4 Model-free testing for subgroup analysis **91**

4.1 Introduction . . . 91

4.2 Subgroup Analysis . . . 93

4.2.1 Testing for treatment heterogeneity . . . 95

4.2.2 Exploratory methods . . . 97

4.3 Testing treatment rules . . . 101

4.3.1 Multiple testing issues . . . 106

4.4 Random forests of treatment trees . . . 110

4.4.1 Comparison study with OWL . . . 114

4.4.2 AIDS Clinical Trials Group Study . . . 117

4.5 Bayesian approach to subgroup analysis . . . 124

4.5.1 AIDS Clinical Trial application . . . 125

4.6 Discussion 131

5 Discussion **133**

References **137**

Appendix A Posterior predication for Chapter 3 **145**

Chapter 1

Introduction

1.1 Statistical fundamentals

Central to the study of statistics is the notion of a probability model. Data y is treated as an observation of a random variable Y, which is assumed to be randomly generated according to an unknown probability distribution F_Y. Within this model-based approach, the data y are used to make inferences about the unknown probability distribution F_Y. Learning about the data-generating mechanism allows the statistician to answer fundamental questions regarding the processes underlying data generation, and the prediction of data yet to be seen. A common setup is the following: a dataset $\{y_i\}_{i=1}^n$ consists of n observations of random variables $\{Y_i\}_{i=1}^n$, where each Y_i is independently distributed according to the probability distribution F_Y. A first step of parametric inference is to make assumptions regarding the form of this distribution, representing our statistical model. For example, if $\{y_i\}_{i=1}^n$ represent independent arrival times, it may be appropriate to assume that the distribution of each Y_i is exponential, that is $Y_i \sim \text{Exp}(\theta)$, where θ represents the rate parameter of the exponential distribution. More generally, θ can represent a vector of parameters associated with a parametric family. In this example we "model" the real-world data-generating procedure by assuming that each datum is independently exponentially distributed with the same rate

parameter θ, which is unknown. In theory, the model should be chosen to best represent the structural and distributional properties the statistician deems most appropriate, given the nature of the data. In practice, additional factors such the associated computational task and the interpretability of parameters become a large part of the model choice.

A model constitutes an incomplete specification of the data generating mechanism, with the final piece in this case being the value of the parameter θ. The remaining task is to make inferences about this parameter θ, given the available data $\{y_i\}_{i=1}^n$. In our example, given the task to identify F_Y, the statistical model limits the search from the space of all possible distributions to the subspace associated with exponential distributions, which can be indexed by the rate parameter θ. The likelihood plays a key role in this inferential procedure, in our example:

$$L(\theta : \{y_i\}_{i=1}^n) = \prod_{i=1}^{n} \theta \exp(-\theta y_i).$$

Frequentist statistical analysis centres around the maximum likelihood estimate $\hat{\theta}(y)$, the value of θ that maximises this likelihood function given the data, informally, the parameter value under which the observed data is most probable. This point estimate of the unknown parameter value θ represents a "best guess", which comes with attractive asymptotic guarantees. In contrast, the aim of the Bayesian statistician is to provide a full description of uncertainty surrounding the value of θ. This is achieved via the representation of the subjective uncertainty surrounding the value of θ in terms of a "posterior distribution". Mathematically, the parameter is treated as a random variable with a distribution that reflects subjective uncertainty of the parameter value. A "prior" distribution $\pi(\cdot)$ represents the uncertainty surrounding the value of θ before the data has been seen. In light of the data, our beliefs are updated to reflect any information gained from the data. This process of updating beliefs, given data, is achieved by the application of Bayes' rule:

$$\pi(\theta \mid y) = \frac{\pi(\theta)L(\theta;y)}{\int_\theta L(\theta;y)\pi(\theta)}$$

whereby the information in the data elegantly transforms our prior distribution into a posterior distribution that represents our updated subjective beliefs about θ in light of the data. In practice, this calculation is not always tractable, in which case, rather than the calculation of a posterior distribution, a sample $\{\theta_j\}_{j=1}^{M}$ from the posterior distribution can be generated using Markov Chain Monte Carlo (MCMC), a numerical technique for generating random samples from a distribution. The posterior probability measure can then be approximated by the Monte Carlo measure: $\sum_{j=1}^{M} \delta_{\theta_j}$. The posterior distribution is a rich, complete description of uncertainty surrounding the parameter value, which provides a means for propagating this uncertainty for the purposes of prediction and decision-making.

A common objection to Bayesian statistics is the need to specify a prior distribution, which can be viewed as a burden in cases where little is known about the distribution of the data. In cases where there is a lot of information available regarding the distribution in question, the prior provides a convenient means of incorporating all known information to bear on the problem. In cases where there is less known about the distribution of the data, the prior should reflect this level of subjective uncertainty. A maximum-likelihood analysis does not require the specification of a prior distribution for parameters, but is not free of assumptions. The specification of a model is built on the premise of ruling out all distributions other than a small parametric family of distributions. Indeed, in our example, this represents the strong assumption that the data is exponentially distributed. The prior in a Bayesian analysis simply reflects beliefs of the relative plausibility of the parameter value. It is useful to think of a prior on a parameter as simply a convenient means of placing a prior over the space of possible distributions from which the data is generated. Constructing an exponential model, with a prior on the rate parameter, is a convenient means of placing all prior mass onto the subspace of all models corresponding to exponential distributions, with mass distributed among this subset of distributions in accordance with the prior on the rate parameter. I would argue that a claim of complete ignorance of a parameter value *a priori* in

a parametric model is inconsistent with the ability to assume that the distribution belongs to a parametric family. A lot is asked of the Bayesian statistician, but the posterior distribution represents a rich reward.

1.2 Nonparametric statistics

Nonparametric statistical methods are motivated by the desire to move away from strict distributional assumptions associated with parametric models. Frequentist nonparametric statistics covers a broad range of methods for applications such as testing, estimation, confidence intervals and regression. Wasserman (2006) serves as a useful resource for many such methods. An elegant, historic category of these methods are nonparametric hypothesis tests, which utilise symmetries and often rankings of the data, circumventing the need to specify a model. An example of such a model-free method is the Mann-Whitney-U test (Mann & Whitney, 1947) for the stochastic ordering of the distributions of two samples. Given samples $\{y_i\}_{i=1}^n$ and $\{z_j\}_{j=1}^m$ of independent observations of random variables $\{Y_i\}_{i=1}^n$ and $\{Z_j\}_{j=1}^m$, under the null hypothesis

$$H_0 : \mathbb{P}(Y > Z) = \mathbb{P}(Z > Y)$$

(or under the stronger null hypothesis that the distributions are equal) a datum from sample 1 is as likely to be greater than a datum from sample 2 as it is to be the smaller of the two. The alternative hypothesis H_1 is chosen as appropriate, under which these probabilities are not equal. Let R denote the sum of the ranks of $\{y_i\}_{i=1}^n$ in the pooled sample $\{\{y_i\}_{i=1}^n, \{z_i\}_{j=1}^m\}$, where a rank of 1 denotes the smallest value. Then, under H_0, the distribution of:

$$U = R - \frac{n(n+1)}{2}$$

is known, so a p-value can be calculated to test H_0 against H_1. The point of note is that this procedure makes no assumptions regarding the distribution of the data, rather utilising the form of the null hypothesis to construct a suitable test statistic. As a result, very few assumptions are required, and such methods are safeguarded against model-misspecification. In addition, rank-based methods provide further robustness against outliers, as ranks reduce the effect of extreme values. However, these tests have very limited scope, providing a tool for specific questions in certain applications.

1.2.1 Bayesian Nonparametrics

Classical nonparametric statistics limit assumptions made about the distribution of the data, in some cases leaving the distribution completely unspecified. In contrast, Bayesian statistics is fundamentally model-based. The model provides a means for capturing information in the data through the likelihood, which allows the updating of beliefs surrounding the model via Bayes' rule. As such, Bayesian analysis cannot be model-free. In a Bayesian setting, "nonparametric" describes the use of priors over large, often infinite-dimensional parameter spaces (see Hjort *et al.*, 2010, for a thorough introduction). This provides a large support, which "casts a wide net" and aims to capture as many distributions as possible. This large prior support is the Bayesian analogue of limiting distributional assumptions. The appeal of such priors is the flexibility offered by such a large support over the space of distributions, allowing for a wide range of possible distributions to be captured by such priors. The large support safeguards against model misspecification, and provides robustness to outliers, as the high number of degrees of freedom reduces the effect of outliers.

Rather than expressing uncertainty regarding the data-generating mechanism via a prior on a finite-dimensional parameter, nonparametric priors directly define a distribution over distributions. As such, nonparametric priors are often conveniently defined in terms of the construction of a random distribution via a stochastic process. The Pólya tree provides

an intuitive example of nonparametric priors for distributions F_Y with support on $\mathbb{R}$, via the construction of a random probability measure (Ferguson, 1974; Lavine, 1992, 1994; Mauldin *et al.*, 1992). $\mathbb{R}$ can be recursively bisected *ad infinitum* to define a sequence of nested partitions of $\mathbb{R}$. It is convenient to use the language of "trees" to describe this nested sequence of partitions, where each level of the tree defines a partition of $\mathbb{R}$. A distribution can be characterised in terms of the allocation of probability mass to elements of the partition of $\mathbb{R}$ defined by every level of the tree. This provides a natural means of constructing a random probability measure, via a random allocation of probability mass at each level of the tree, which is consistent across levels. Formally, a Pólya tree prior is defined as follows: Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k) \in E^k = \{0,1\}^k$, and define a sequence of embedded partitions of $\mathbb{R}$ by $\Gamma_k = \{B_{\boldsymbol{\varepsilon}} : \{\varepsilon \in E^k\}$, where the $B_{\boldsymbol{\varepsilon}}$ are defined recursively, such that $B_{\boldsymbol{\varepsilon}0} \cup B_{\boldsymbol{\varepsilon}1} = B_{\boldsymbol{\varepsilon}}$. Now let $E^* = \cup_{k\geq 1} E^k$, the set of all countable sequences of zeros and ones, and let $\mathscr{A} = \{\alpha_{\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} \in E^*\}$ be a set of nonnegative real numbers. Then, a random probability measure P is a Pólya tree process with respect to $\Gamma = \{\Gamma_k : k \geq 1\}$ and $\mathscr{A}$ (which we denote by $P \sim \mathrm{PT}(\Gamma, \mathscr{A})$) if $P(B_{\boldsymbol{\varepsilon}0} \mid B_{\boldsymbol{\varepsilon}}) \sim \mathrm{Beta}(\alpha_{\boldsymbol{\varepsilon}0}, \alpha_{\boldsymbol{\varepsilon}1})$, independently for all $\boldsymbol{\varepsilon} \in E^*$. Typically, the recursive partition is generated with respect to a base probability measure P_0 by splitting the space by the quantiles $P_0^{-1}(\frac{k}{2^m})$ at level m. An example of $\mathscr{A}$ is the sequence defined by $\alpha_{\varepsilon} = cm^r$ at level m, where $r = 2$ provides almost-sure absolute continuity of the random measure P. Posterior updating, given an observation $y \sim P$, can simply be viewed as a sequence of Bernoulli observations as we observe the path down the tree corresponding to the observation y. Since the relative probability mass at each split has a beta distributed prior, the update of these parameters is conjugate, leading to the posterior distribution: $P \mid y \sim \mathrm{PT}(\Gamma, \mathscr{A}^*)$, where:

$$\alpha_{\varepsilon}^* = \begin{cases} \alpha_{\varepsilon}, & \text{if } y \notin B_{\varepsilon} \\ \alpha_{\varepsilon} + 1, & \text{if } y \in B_{\varepsilon} \end{cases}$$

Given a dataset of size n, this is simply extended by induction. This conjugacy property aids posterior computation. For example, posterior predictive sampling is achieved using a simple algorithm that cascades down the tree. At stage m we set $\varepsilon_m = 0$ with probability $P(B_{\varepsilon_{1:(m-1)}0} \mid B_{\varepsilon_{1:(m-1)}})$, a random beta-distributed probability $\text{Beta}(\alpha_{\varepsilon_{1:(m-1)}0}, \alpha_{\varepsilon_{1:(m-1)}1})$, otherwise setting $\varepsilon_m = 1$. This is repeated until either a pre-specified truncation level is reached, or until the subset B_ε contains no data. In either case, we can draw from the posterior predictive by sampling from the base measure P_0, restricted to B_ε. The Pólya tree provides a computationally tractable prior with convenient specification through the base measure P_0, with infinite dimensionality providing flexibility to capture a wide range of possible distributions. A drawback of the prior is the dependence on the somewhat arbitrary partition structure, which effects the prior as well as posterior updating. Mixtures of Pólya trees - constructed by mixing over the base measure - mitigate this issue somewhat, reducing the dependence on the partition structure.

The most widely-used and influential prior in Bayesian nonparametrics is the Dirichlet Process prior (Ferguson, 1974). The Dirichlet process is a probability distribution on probability measures that induces finite-dimensional Dirichlet distributions on finite partitions of the sample space of the random measure. Formally, given a finite (base) measure α on a Polish space Ψ, a random measure P on Ψ is a Dirichlet process if, for every finite measurable partition $\{B_1, \ldots, B_k\}$ of Ψ, $(P(B_1), \ldots, P(B_k))$ is Dirichlet distributed with parameters $(\alpha(B_1), \ldots, \alpha(B_k))$. P is distributed according to a Dirichlet process with base measure α, denoted by $P \sim \mathscr{D}_\alpha$.

From this we can deduce some elementary properties of the Dirichlet process. First, the mass assigned to any set A is $\text{Beta}(\alpha(A), \alpha(A^c))$ distributed, so $\mathbb{E}[P(A)] = \frac{\alpha(A)}{\alpha(A)+\alpha(A^c)} = G(A)$, where G is the normalised measure given by $G(A) = \frac{\alpha(A)}{\alpha(\Phi)}$. Letting $M := \alpha(\Psi)$, it is common to parameterise the Dirichlet process in terms of G and M, as $Y \sim P$ and $P \sim \text{DP}(M, G)$ (a natural alternative notation), implies that $X \sim G$, marginally. Like the

Pólya tree, the Dirichlet process is a conjugate prior given i.i.d data from an unknown distribution. Formally, if $Y_1, \ldots, Y_n \sim P$ with $P \sim \mathscr{D}_\alpha$, then $P \mid Y_1, \ldots, Y_n \sim \mathscr{D}_{\alpha + \sum_{i=1}^n \delta_{Y_i}}$. This makes the DP a highly convenient prior, computationally.

Draws from a Dirichlet process prior are almost surely discrete probability measures. In fact, a draw from the DP can be constructed via a "stick-breaking" process (Sethuraman, 1994) as follows:

$$P = \sum_{i=1}^{\infty} V_i \delta_{\theta_i}, \qquad \theta_i \sim G, \qquad V_i = \left[\prod_{j=1}^{i-1} (1 - S_j) \right] S_i, \qquad S_i \sim \text{Beta}(1, M).$$

The stick-breaking interpretation is that starting with a stick of length 1, representing probability mass, the stick is broken randomly at the point $S_1 \sim \text{Beta}(1, M)$ and this probability mass is assigned to a random draw $\theta_1 \sim G$ from the base measure. This process is repeated, splitting the remaining stick of probability mass of length $1 - S_1$ by the proportion S_2 and assigning this probability to another draw from the base measure, θ_2. This stick-breaking repeated *ad infinitum*, distributing the probability mass among points drawn from the base measure.

The discreteness of draws from the Dirichlet process results in ties in data generated from the random distribution. Specifically, given data $Y_1, \ldots, Y_n \sim P$, where $P \sim \text{DP}(M, G)$, the predictive distribution of $Y_{n+1} \mid Y_1, \ldots, Y_n$ is given by the so-called Pólya urn scheme (Blackwell & MacQueen, 1973):

$$Y_{n+1} \mid Y_1, \ldots, Y_n \sim \frac{M}{M+n} G + \sum_{i=1}^{n} \frac{1}{M+n} \delta_{Y_i}$$

whereby new observations duplicate previous observations with probability proportional to the frequency of the observation, otherwise taking a value drawn from the base probability measure G. As with the Pólya tree, this prediction is a trivial computation as a result of

conjugacy of the prior. These properties provide insight into properties of the Dirichlet process, and the associated posterior behaviour.

The discreteness of draws from the Dirichlet process will not be appropriate in applications where the distribution of interest is thought to be continuous. In such applications, it is possible to use the Dirichlet process prior as a mixing measure for the parameters of a parametric distribution, resulting in an infinite mixture model. A popular class of model is the Dirichlet Process Mixture (Lo, 1984), whereby a Dirichlet process prior is placed on the distribution of the parameters of a parametric family. The result is an "infinite mixture model". Precisely:

$$f_Y(y) = \int K(y|\theta) dP(\theta)$$
$$P \sim \mathrm{DP}(M, G)$$

where K is the density of the chosen parametric family. Since draws from a Dirichlet Process are almost surely atomic measures, there is positive probability of observations sharing a parameter value given the random measure P, with groupings distributed in accordance with the Pólya urn scheme. The result is an effect of clustering within a sample, with a random, limitless number of clusters. This has proved to be an extremely popular model as it models heterogeneity within a sample well, and is highly flexible, with large support. Efficient MCMC schemes (Escobar & West, 1995; MacEachern & Müller, 1998; Neal, 2000) have lead to the widespread use of the Dirichlet Process Mixture (DPM) in density estimation.

1.3 Regression

Regression analysis is the investigation of the relationship between a variable of interest Y and other "covariates", predictive variables $X_1, \ldots, X_p$. Interest lies in the nature of the relationship, the identification and analysis of important predictive variables, and the

prediction of new Y^* given new covariate values $x_1^*, \ldots, x_p^*$. As such, it is natural to consider the conditional distribution of $Y \mid X = x$, as this describes the nature of the dependence of Y on X and allows predictions to be made of the response, given specific covariate values. Mathematically, a regression model is a set of conditional distributions $f_x(y;\theta_1)$, parameterised by θ_1, the object of statistical interest. A prior $\pi_1(\cdot)$ on θ_1 reflects uncertainty of the parameter value before the observation of the data. Then, given data $\{y_i, x_i\}_{i=1}^n$, the posterior distribution reflects the updated beliefs about θ_1. The posterior can then be used to calculate probabilities relating to covariate effects, inform decision-making and to calculate the posterior predictive distribution of the response Y^* given covariate value x^*:

$$p(y^* \mid x^*, \{y_i, x_i\}_{i=1}^n) = \int_{\theta_1} f_{x^*}(y^* \mid \theta_1)\pi_1(\theta_1 \mid \{y_i, x_i\}_{i=1}^n)d\theta_1.$$

Typically, rather than specifying a joint model for $(Y, X_1, \ldots, X_p)$, the covariates are treated as fixed, despite actually being random in many applications. This is justified using the following likelihood argument. Given a dataset $\{y_i, x_i\}_{i=1}^n$, treating covariates as fixed gives us the likelihood:

$$L(\theta_1; \{y_i, x_i\}_{i=1}^n) = \prod_{i=1}^n f_{x_i}(y_i; \theta_1).$$

If we were to treat X as random variables, with density $g(x;\theta_2)$, our likelihood becomes:

$$L(\theta_1, \theta_2; \{y_i, x_i\}_{i=1}^n) = \prod_{i=1}^n f_{x_i}(y_i; \theta_1)g(x_i; \theta_2) = L(\theta_1; \{y_i, x_i\}_{i=1}^n) \prod_{i=1}^n g(x_i; \theta_2).$$

Therefore, if independent priors π_1 and π_2 are placed on θ_1 and θ_2, the joint posterior will factorise into an independent posterior for θ_1 and θ_2 as follows:

$$\begin{aligned}
\pi(\theta_1, \theta_2 \mid \{y_i, x_i\}_{i=1}^n) &\propto \pi_1(\theta_1)\pi_2(\theta_2)L(\theta_1, \theta_2; \{y_i, x_i\}_{i=1}^n) \\
&= \pi_1(\theta_1)\pi_2(\theta_2)\prod_{i=1}^n f_{x_i}(y_i; \theta_1)g(x_i; \theta_2) \\
&= [\pi_1(\theta_1)L(\theta_1; \{y_i, x_i\}_{i=1}^n)][\pi_2(\theta_2)L(\theta_2; \{y_i, x_i\}_{i=1}^n)] \\
&\propto \pi_1(\theta_1 \mid \{y_i, x_i\}_{i=1}^n)\pi_2(\theta_2 \mid \{x_i\}_{i=1}^n)
\end{aligned}$$

leading to the same posterior for θ_1 as when X is treated as fixed. Inference for θ_1, and therefore prediction via the posterior predictive $p(y^* \mid x^*, \{y_i, x_i\}_{i=1}^n) = \int_{\theta_1} f_{x^*}(y^* \mid \theta_1)\pi_1(\theta_1 \mid \{y_i, x_i\}_{i=1}^n)d\theta_1$, is not dependent on the treatment of X as random or fixed. In this sense, a full regression analysis can be done without consideration of the distribution of the covariates. However, by treating the covariates as fixed, our regression model is simply a set of conditional distributions for Y given different values of X, rather than a full joint distribution. As a result, the marginal distribution of Y, upon integrating out X, is not defined. By treating X as random variables, our conditional model $f_x(y; \theta_1)$ for Y implies a marginal distribution for Y, given by: $f(y; \theta_1, \theta_2) = \int_x f_x(y; \theta_1)g(x; \theta_2)dx$. This thesis explores approaches to regression that utilise the marginal distribution of the response, which is not possible when covariates are treated as fixed.

1.3.1 Survival analysis

A fascinating subset of regression models can be found in the rich literature of survival analysis. The response $Y \geq 0$ denotes a positive waiting time until an event. Commonly, applications are in a medical setting, where this represents time until death, but Y can represent time until a disease onset, time until remission, or more generally any waiting-time, such as time until the failure of a mechanical object. Missing data, in the form of

censoring and truncation of the response, is common in such applications. As a consequence of the positivity of the response, survival models are commonly parameterised differently to standard regression models. The typical approach to defining a regression model - via a conditional mean and an error term - will not guarantee positive values for Y, so is inappropriate. As a result, rather than parameterisation in terms of probability density and distribution functions, survival models are typically defined in terms of hazard and survivor functions. The survivor function $S(y) = \mathbb{P}(Y > y)$ returns the probability that the event occurs after the time y. The hazard function (or hazard rate) is defined as an "age-specific failure rate":

$$h(y) = \lim_{\delta \to 0} \frac{\mathbb{P}(y \leq Y < y + \delta \mid y \leq Y)}{\delta}$$

The hazard function is key to the interpretation of many survival models. For example, if Y is exponentially distributed, its hazard function is constant, so the probability of failure within a small interval, given survival until that point, is fixed. The cumulative hazard is defined as the integral of the hazard function: $H(y) = \int_0^y h(u)du$. From these definitions we can deduce further relationships between the hazard function and density function. By the rules of conditional probability, we have:

$$h(y) = \frac{f(y)}{S(y)}$$

so

$$h(y) = -\frac{S^{'}(y)}{S(y)} = -\frac{d(log(S(y)))}{dy}$$

since $S(0) = 1$, gives us

$$S(y) = \exp\{-H(y)\}$$

and, therefore

$$f(y) = h(y)\exp\{-H(y)\}.$$

This provides a simple expression for the likelihood in terms of hazard functions. Censoring and truncation are easily incorporated to the likelihood, with right censored data (observations corresponding to $\{Y \geq y\}$) contributing $\exp\{-H(y)\}$ and left truncated data (observations corresponding to $\{Y = y \mid Y \geq y_t\}$) contributing $h(y)\exp\{-(H(y)-H(y_t))\}$ to the likelihood.

Regression models in survival analysis are constructed via hazard functions $h_x(y)$ indexed by covariates, rather than density functions $f_x(y)$. The accelerated failure time model is built on the premise that the effect of a covariate is to accelerate (or decelerate) the lifetime by a constant. This is achieved by setting:

$$h_x(y) = \lambda(x)h_0(\lambda(x)y)$$

which implies

$$S_x(y) = S_0(\lambda(x)y).$$

where $S_0(\cdot)$ and $h_0(\cdot)$ are called the baseline survivor function and baseline hazard function, the survivor and hazard functions corresponding to an individual with $x = 0$. $\lambda(x) \geq 0$ is simply a positive function of x, commonly taken to be $\lambda(x) = \exp(\beta x)$, which determines the degree of acceleration, as a function of x. The most popular model in survival analysis is Cox's proportional hazards model (Cox, 1972), which has over 41,000 citations at the time of writing. Simply, the model specifies:

$$h_x(y) = \lambda(x)h_0(t).$$

That is, at any time, the instantaneous failure rate for an individual is proportional to the rate corresponding to an individual with $x = 0$, where the constant is a function of the covariate x. A consequence of this definition is that the probability of an individual having an event before another individual is given by $\frac{\lambda(x_1)}{\lambda(x_1)+\lambda(x_2)}$, and importantly does not depend on h_0. Therefore, if primary interest lies in covariate effect, rather than the distribution of the response, the

probability of the observed ranking provides a "partial likelihood", that only depends on λ. By essentially transforming the data into ranks from event times, this allows circumvention of the specification of the form of h_0, in what is referred to as a "semiparametric" approach. The partial likelihood is robust in the sense that the use of ranks safeguards against outliers, and also in the sense that the partial likelihood holds for any baseline hazard h_0. The simple form makes maximum (partial) likelihood inference straightforward and computationally efficient. The proportional hazard model serves as an excellent example of a model popular on the basis of computation and robustness, rather than purely for modeling considerations.

1.4 Motivation

We have seen that there are a number of approaches to any statistical task. I believe the Bayesian paradigm provides a complete, descriptive analysis, via the reporting of the posterior distribution. In any model-based approach, the statistical model represents the majority of the assumptions associated with the statistical process; as such, I believe that a prior distribution is a reasonable requirement for a statistical analysis. However, there are specific statistical questions that may be addresses without need to build a full model for the data. Where applicable, I feel that nonparametric hypothesis tests provide a valid alternative to a model-based approach. This thesis explores the use of Bayesian parametric, Bayesian nonparametric and frequentist nonparametric approaches. I believe that each of these provides a powerful tool for the analysis of data, the merits of which will depend on the application.

The central theme of this thesis is the treatment of covariates X as random variables within a regression context. We show how this facilitates the definition of the marginal distribution of Y upon integrating out covariates. In many applications, this marginal distribution is highly interpretable and provides a vehicle for the incorporation of prior information. Chapter 2 explores the incorporation of known marginal means and variances of Y and X into a prior for linear regression. The known marginal means and variances induce a condition that contains

all parameters, leading to a rich joint structure in the prior and posterior. In the absence of large marginal datasets, this framework can still be used with a prior on unknown marginal moments, in order to incorporate the natural dependence structure. We show how treatment of X as random variables raises the notion of "marginal coherence" in the context of variable selection, which is satisfied when each model assigns the same distribution to the sample mean and variance of the response.

In Chapter 3, treatment of covariates as random allows the construction of a Bayesian nonparametric regression model which is parameterised in terms of the marginal distribution of Y. This facilitates a composite likelihood approximation that can be viewed as an extension of the partial likelihood for the proportional hazards model, which provides scalable computation that aids the application of Bayesian nonparametric priors in a regression context on a large scale. In addition, we show how this approach aids inference for covariates effects, given truncated data.

Finally, Chapter 4 explores the area of subgroup analysis: the task of using data from a randomised clinical trial to identifying an optimal treatment rule $D : X \to T$, which allocates treatment (T) on the basis of covariates. Treatment of the covariates as random allows direct consideration of the distribution of the response under a treatment rule $Y \mid T = D(X)$, which we show can be tested against treatment at random. This approach allows complete circumvention of the specification of a statistical model. This provides a robust and simple alternative to the standard two-step approach that first constructs a conditional model for the response $Y \mid X = x, T = t$, and then uses the fit of this model to determine a course of treatment D.

Chapter 2

Constrained marginal variance priors for linear regression

2.1 Introduction

This chapter examines linear regression where the marginal means and variances of both Y and X are known. Potential applications might include regression within a large population, where separate records for each of Y and X are common, but their collection together is limited to a small dataset. This situation is increasingly common in the age of "big data", with the availability of large population-wide data repositories, such as the UK Biobank[1] and the US census[2]. However, this information about marginal distributions is seldom incorporated into analysis, despite Bayesian statistics providing the opportunity to incorporate this prior information. This is because the treatment of covariates as non-random leaves these marginal distributions undefined. We propose a joint prior over regression coefficients and the noise variance that incorporates constraints imposed by the marginal means and variances. This is done via a simple parameter transformation that results in a prior defined over the parameter

[1] http://www.ukbiobank.ac.uk/
[2] http://www.census.gov/programs-surveys/acs/data/pums.html

subspace that satisfies the parameter constraint. This leads to a richly structured prior across the parameters and the noise variance which is easy to specify in terms of relative magnitudes of variance components. The notion of "marginal coherence" is introduced in the context of hypothesis testing, whereby each competing model is constrained to place the same prior on the marginal mean and variance of the response. This is achieved by placing priors on the unknown marginal response mean and variance and applying the marginally constrained prior conditional on these values. We apply the marginally constrained priors to a simulation example and a small subset of the US census data, where the remainder of the data is used to estimate marginal means and variances.

2.2 Bayesian linear model

In theory, Bayesian analysis of the linear model is incredibly simple. A conditional model for the response $Y \in \mathbb{R}$, given covariates $X \in \mathbb{R}^p$ is specified in terms of parameters $\alpha, \beta_1, \ldots, \beta_p$ and σ^2 as follows:

$$Y \mid X, \alpha, \beta_1, \ldots, \beta_p, \sigma^2 \sim \mathrm{N}(\alpha + \sum_{j=1}^{p} \beta_j X_j, \sigma^2)$$

commonly expressed in the form:

$$Y = \alpha + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. A dataset of size n consists of a vector $\mathbf{y}$ of size n and a matrix of covariates $\mathbf{x}$ of size $n \times p$, where row j contains the covariates corresponding to y_j. It is typical to treat $\mathbf{x}$ as fixed, giving the likelihood:

$$L(\alpha, \beta_{1:p}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \alpha - \sum_{j=1}^{p} x_{i,j}\beta_j)^2}{2\sigma^2})$$

If **x** is treated as random, it would be natural to parameterise its distribution with parameters independent of those used for the conditional distribution of $Y \mid X$. In this case, treating covariates as random results in the likelihood being multiplied by a term that does not contain $(\alpha, \beta_{1:p}, \sigma^2)$; therefore, inference for these parameters is identical to that when **x** is treated as fixed, justifying this common practice.

All that remains is for the investigator to express prior beliefs about the model parameters in the form of a prior distribution $p(\alpha, \beta_{1:p}, \sigma^2)$. The prior offers the opportunity to incorporate all available information about parameters in a coherent way; however, in practice, a small number of priors are overwhelmingly popular. Textbook Bayesian analyses of the linear model (for example: Box & Tiao, 2011; Broemeling, 1984; Gelman *et al.*, 2014) are dominated by conjugate priors. The posterior distributions corresponding to conjugate priors for α, β, σ^2, given regression data, take the same form as the prior. As such, posteriors can be reported analytically, without need for significant computation. An additional benefit to conjugate priors is that the effect of the prior on the posterior is clear, providing a good introduction to the Bayesian approach. For the remainder of this review, β is extended to include α as the first element, with the matrix $\boldsymbol{x}$ adjusted to include a column of 1s as the first column accordingly.

A common "noninformative" prior is given by:

$$p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}.$$

This is the product of a "flat" prior on $\boldsymbol{\beta}$ and a Jeffreys prior on σ^2. Clearly this is improper, as its integral is infinite; however, the resulting posterior is proper for $n > p$. The motivation behind this prior is to provide as little information as possible in the prior, for use in the complete absence of prior information. No value of $\boldsymbol{\beta}$ is any more likely than any other in our prior. The posterior marginal for σ^2 takes the form of a scaled inverse-χ^2, and conditional on σ^2, $\boldsymbol{\beta}$ is normally distributed in the posterior.

The conjugate normal-inverse-gamma prior with hyperparameters $(\boldsymbol{\mu}, \Sigma, a, b)$ takes the form:

$$p(\boldsymbol{\beta}, \sigma^2) = |\Sigma|^{-\frac{1}{2}}(2\pi)^{-\frac{p}{2}}\frac{b^a}{\Gamma(a)}(\frac{1}{\sigma^2})^{\frac{p}{2}+a+1}\exp(-\frac{2b+(\boldsymbol{\beta}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})}{2\sigma^2})$$

Which leads to the convenient factorisation:

$$\sigma^2 \mid a, b \sim \Gamma^{-1}(a, b)$$
$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\mu}, \Sigma \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\Sigma^{-1})$$

A common approach is to let $\Sigma = \boldsymbol{v}I_p$, so each $\beta_j \mid \sigma^2 \sim \mathrm{N}(\mu_j, \sigma^2 v_j)$ independently. From here, specification is completed by fixing $\mu, \{v\}_{j=1}^{p+1}, a$ and b. In the interests of scientific reporting, $\mu = 0$ is a common default choice. $\boldsymbol{v}$ can be specified to reflect any information regarding the relative magnitudes of covariate effects.

Contained in this family of conjugate priors is Zellner's g-prior (Zellner, 1986), corresponding to the case where $\Sigma = g^{-1}(X^TX)$, which offers tractability in calculation of Bayes factors. The prior covariance matrix is equal to a scalar g multiplied by the covariance matrix of the maximum likelihood estimator $\hat{\beta}$, which depends on the data only through the covariates, given σ^2.

A greater diversity of priors can be found in problems concerning sparsity. In such applications, usually with large p, there is prior belief that many of the covariate effects are zero. To reflect this, Park & Casella (2008) use a conditional Laplace prior for regression coefficients as follows:

$$p(\sigma^2) \propto \sigma^{-2}$$
$$p(\boldsymbol{\beta} \mid \sigma^2) = \prod_{j=1}^{p+1}\frac{\lambda}{2\sigma}\exp(\frac{-\lambda|\beta_j|}{\sigma}).$$

A popular alternative approach to model sparsity is the family of scale mixture of normals (see Polson & Scott, 2010, for a comprehensive review). Particular examples include the horseshoe prior (Carvalho *et al.*, 2010), given by:

$$\begin{aligned} p(\sigma^2) &\propto \sigma^{-2} \\ \tau \mid \sigma &\sim \mathrm{C}^+(0,\sigma) \\ \lambda_j \mid \tau &\sim \mathrm{C}^+(0,\tau) \quad j = 1,\ldots,p+1 \\ \beta_j \mid \lambda_j &\sim \mathrm{N}(0,\lambda_j^2) \quad j = 1,\ldots,p+1 \end{aligned}$$

where $\mathrm{C}^+(0,a)$ denotes a standard half-Cauchy distribution with scale parameter a. In this setup, each β_j is mixed over its own λ_j. This allows a greater diversity of β_j values than afforded by a standard prior. Similarly, the normal-gamma prior (Griffin & Brown, 2010) is given by:

$$\begin{aligned} p(\sigma^2) &\propto \sigma^{-2} \\ \lambda_j &\sim \Gamma(a,b) \quad j = 1,\ldots,p+1 \\ \beta_j \mid \lambda_j &\sim \mathrm{N}(0,\lambda_j^2) \quad j = 1,\ldots,p+1. \end{aligned}$$

Griffin & Brown (2010) provide interpretation of hyperparameters, showing that large values of a result in more even coefficients, with small a encouraging a diversity of covariate effects, resulting in the majority of the regression variation being attributed to a small number of coefficients. This provides an intuitive means for incorporating prior information about the expected spread of covariate effects.

Conjugacy simplifies posterior analysis, but it is easy to lose sight of the fact that the prior is supposed to reflect subjective uncertainty in the parameter values. Robert (2007) notes "attacks on g-priors hardly ever consider its major drawback, namely, that its choice is

not based on prior information". Gelman (2008) also expresses skepticism that these priors represent actual prior beliefs “What a coincidence that, of all the infinite variety of priors that could be chosen, it always seems to be the normal, gamma, beta, etc., that turn out to be the right choices?”. Jordan (2011), states "Elicitation remains a major source of open problems". The problem is that the nature of prior information has to be assessed on a case-by-case basis. These priors provide a way to specify the mean and variance of parameters in the prior, and it could be argued that practitioners are unlikely to have any more information than this. However, the treatment of covariates as deterministic has led to the oversight of the existence of the marginal distribution of Y, when covariates are in fact random. We consider marginal data as a common source of prior information in modern statistics, with the increased availability of population-level data such as the UK Biobank. A simple example of this is the analysis of heritability of height, where there is plenty of population-wide height data available, but the collection of parent and child height together is limited to smaller datasets. This chapter explores how information regarding marginal means and variances can be incorporated into a linear regression analysis in a coherent and principled way. When these marginal moments are known, they provide a simple parameter constraint. A transformation of parameters allows us to specify a prior such that this constraint is satisfied, leading to a highly structured prior with interpretable behaviour. Before exploring such priors, it is important to understand the standard analysis of marginal data.

2.2.1 As a missing data problem

In a general regression setting, let $f(y \mid x, \theta_1)$ denote the conditional density for $Y_i \mid X_i = x$, parameterised by θ_1. X_i are random variables with density $g(x \mid \theta_2)$ parameterised by θ_2. Each paired observation (y_i, x_i) is independent and identically distributed. This implies a

marginal distribution for Y_i, by integrating out X_i:

$$f(y \mid \theta_1, \theta_2) = \int_x f(y \mid x, \theta_1) g(x \mid \theta_2) dx$$

which is parameterised by both θ_1 and θ_2. A prior $p(\theta_1, \theta_2)$ is specified, and in addition to our usual paired data $\{y_i, x_i\}_{i=1}^n$, unpaired, marginal datasets $\{y_i^*\}_{i=1}^M$ and $\{x_i^\dagger\}_{i=1}^N$ are available. The complete data available takes the following form:

$$\begin{pmatrix} y_1^* & \emptyset \\ \vdots & \vdots \\ y_M^* & \emptyset \\ \emptyset & x_1^\dagger \\ \vdots & \vdots \\ \emptyset & x_N^\dagger \\ y_1 & x_1 \\ \vdots & \vdots \\ y_n & x_n \end{pmatrix}$$

The likelihood for (θ_1, θ_2) has the form:

$$L(\theta_1, \theta_2; \{y_i, x_i\}_{i=1}^n, \{y_i^*\}_{i=1}^M, \{x_i^\dagger\}_{i=1}^N) = \\ [\prod_{i=1}^M \int_x f(y_i^* \mid x, \theta_1) g(x \mid \theta_2) dx][\prod_{i=1}^n f(y_i \mid x_i; \theta_1) g(x_i \mid \theta_2)][\prod_{i=1}^N g(x_i^\dagger \mid \theta_2)]$$

Direct posterior inference will be difficult as a result of the integral in the likelihood. A common solution is to treat the problem as one of missing data. $\{Y_i^\dagger\}_{i=1}^M$ and $\{X_i^*\}_{i=1}^N$ are missing, so we now perform inference on $\theta_1, \theta_2, \{Y_i^\dagger\}_{i=1}^M$ and $\{X_i^*\}_{i=1}^N$. Practically, this is simple within a Gibbs sampler, where missing values are "imputed", given current parameter values, followed by parameter values being updated given the full imputed dataset.

We are concerned with applications where marginal datasets are very large, making imputation computationally demanding. In addition, the requirement to specify a full model for possibly high-dimensional covariates can prove a difficult task.

Recently, Chatterjee *et al.* (2016) outlined a constrained maximum likelihood approach to incorporating summary level information from external data sources. This approach shares the motivations of the approach we consider, but within a frequentist maximum likelihood framework.

2.3 Fixed marginal variance prior

Consider linear regression when the following marginal quantities are well-known: $\mathbb{E}[Y_i]$, $\text{Var}[Y_i]$, $\mathbb{E}[X_{i,j}]$, $\text{Var}[X_{i,j}]$, and $\text{Cov}[X_{i,j}, X_{i,k}]$. This information would typically come from massive population-wide marginal datasets of Y and X. This marginal information can be leveraged into a prior in the linear regression analysis of a smaller paired dataset $\{y_i, x_i\}_{i=1}^n$.

As a pre-processing step, covariates can be normalised so that $\mathbb{E}[X_{i,j}] = 0$ and $\text{Var}[X_{i,j}] = 1$ for all j, in order to simplify notation and isolate the mean of Y. Treatment of categorical variables is covered in detail in Section 2.5.2. Let's introduce the notation for the known quantities:

$$\begin{aligned} \mathbb{E}[Y_i] &= \mu_0 \\ \text{Var}[Y_i] &= \sigma_0^2 \\ \text{Cov}[X_{i,j}, X_{i,k}] &= \rho_{j,k} \end{aligned} \tag{2.1}$$

The task is to incorporate this known information into a a simple linear model:

$$Y_i = \alpha + \sum_{j=1}^{p} X_{i,j}\beta_j + \varepsilon_i \tag{2.2}$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$.

The linear model, along with the known covariate covariance structure implies the following marginal moments for Y:

$$\begin{aligned} \mathbb{E}[Y_i \mid \alpha, \beta, \sigma] &= \alpha \\ \mathrm{Var}[Y_i \mid \alpha, \beta, \sigma] &= \sum_{j=1}^{p} \beta_j^2 + 2 \sum_{1 \leq l} \sum_{l < m \leq p} \beta_l \beta_m \mathrm{Cov}(X_{i,l}, X_{i,m}) + \sigma^2 \end{aligned} \tag{2.3}$$

Matching these theoretical moments with the known moments from marginal population data gives us the constraints:

$$\mu_0 = \alpha \tag{2.4}$$

$$\sigma_0^2 = \sum_{j=1}^{p} \beta_j^2 + 2 \sum_{1 \leq l} \sum_{l < m \leq p} \beta_l \beta_m \rho_{l,m} + \sigma^2 \tag{2.5}$$

The first constraint fixes α. With this in mind, it is convenient to centre Y as a pre-processing step, giving $\alpha = 0$; as such, α is omitted henceforth. This highlights a benefit to centering covariates X: we are now simply left with one constraint.

The second constraint is more complex, as it involves all remaining parameters $\beta_1, \ldots, \beta_p, \sigma^2$. The aim is to specify a prior that will guarantee this condition is met. The key to achieving this is to note that multiplying $\beta_1, \ldots, \beta_p$ and σ by a constant c scales the implied marginal variance by a factor of c^2, as each term in the sum is either a squared parameter, or a product of two parameters. This makes it easy to see that we can meet the condition simply by scaling the parameters together appropriately. To do this we introduce unscaled parameters $(\theta_1, \ldots, \theta_p, \phi^2)$, then transform them into $(\beta_1, \ldots, \beta_p, \sigma^2)$ in a way that the condition is met. This approach to parameter construction is similar to that in Gelman *et al.* (1996), where a transformation of parameters ensures a parameter sum constraint in a toxicology application where total blood flow is related to blood flow within different compartments of the body.

Formally define our parameters $(\beta_1, \ldots, \beta_p, \sigma^2)$ as a transformation of $(\theta_1, \ldots, \theta_p, \phi^2)$ by:

$$\begin{aligned}\beta_k &= \sigma_0 \frac{\theta_k}{\sqrt{\sum_{j=1}^{p} \theta_j^2 + 2\sum_{1\leq l}\sum_{l<m\leq p} \theta_l \theta_m \rho_{l,m} + \phi^2}} \\ \sigma^2 &= \sigma_0^2 \frac{\phi^2}{\sum_{j=1}^{p} \theta_j^2 + 2\sum_{1\leq l}\sum_{l<m\leq p} \theta_l \theta_m \rho_{l,m} + \phi^2}\end{aligned} \tag{2.6}$$

With this construction, we have:

$$\begin{aligned}&\sum_{j=1}^{p} \beta_j^2 + 2\sum_{1\leq l}\sum_{l<m\leq p} \beta_l \beta_m \rho_{l,m} + \sigma^2 \\ &= \sigma_0^2 \frac{\sum_{j=1}^{p} \theta_j^2 + 2\sum_{1\leq l}\sum_{l<m\leq p} \theta_l \theta_m \rho_{l,m} + \phi^2}{\sum_{j=1}^{p} \theta_j^2 + 2\sum_{1\leq l}\sum_{l<m\leq p} \theta_l \theta_m \rho_{l,m} + \phi^2} \\ &= \sigma_0^2\end{aligned}$$

Therefore, placing a prior on $p(\theta_1, \ldots, \theta_p, \phi^2)$ implies a prior on $\beta_1, \ldots, \beta_p, \sigma^2$ which satisfies the marginal variance constraint. The induced prior is given by:

$$p(\theta_1(\beta_1, \ldots, \beta_p, \sigma^2), \ldots, \theta_p(\beta_1, \ldots, \beta_p, \sigma^2), \phi^2(\beta_1, \ldots, \beta_p, \sigma^2)) \left| \frac{\delta(\beta_1, \ldots, \beta_p, \sigma^2)}{\delta(\theta_1, \ldots, \theta_p, \phi^2)} \right|$$

Rather than deal with calculation of the implied prior, it is simpler to proceed with inference via MCMC on $\theta_1, \ldots, \theta_p, \phi^2$. The output constitutes of a posterior sample of $\theta_1, \ldots, \theta_p, \phi^2$ which can simply be transformed into the parameters of interest $\beta_1, \ldots, \beta_p, \sigma^2$ to provide a posterior sample. The log likelihood for $\theta_1, \ldots, \theta_p, \phi^2$ is simply calculated by substituting for $\beta_1, \ldots, \beta_p, \sigma^2$ into our Gaussian likelihood:

$$-\frac{n}{2}\log\sigma_0^2\phi^2+\frac{n}{2}\log[\sum_{j=1}^{p}\theta_j^2+2\sum_{1\leq l}\sum_{l<m\leq p}\theta_l\theta_m\rho_{l,m}+\phi^2]$$
$$-\frac{\sum_{i=1}^{n}(y_i-\sum_{k=1}^{p}\frac{x_{i,k}\sigma_0\theta_k}{\sqrt{\sum_{j=1}^{p}\theta_j^2+2\sum_{1\leq l}\sum_{l<m\leq p}\theta_l\theta_m\rho_{l,m}+\phi^2}})^2}{\frac{2\sigma_0^2\phi^2}{\sum_{j=1}^{p}\theta_j^2+2\sum_{1\leq l}\sum_{l<m\leq p}\theta_l\theta_m\rho_{l,m}+\phi^2}}$$

or, alternatively:

$$-\frac{n}{2}\log\sigma_0^2\phi^2+\frac{n}{2}\log[\sum_{j=1}^{p}\theta_j^2+2\sum_{1\leq l}\sum_{l<m\leq p}\theta_l\theta_m\rho_{l,m}+\phi^2]$$
$$-\frac{\sum_{i=1}^{n}(y_i\sqrt{\sum_{j=1}^{p}\theta_j^2+2\sum_{1\leq l}\sum_{l<m\leq p}\theta_l\theta_m\rho_{l,m}+\phi^2}-\sum_{k=1}^{p}x_{i,k}\sigma_0\theta_k)^2}{2\sigma_0^2\phi^2}$$

For example, define a prior over $\theta_1,\ldots,\theta_p,\phi$:

$$\begin{aligned}\theta_k &\sim \mathrm{N}(\mu_k,s_k^2)\ k=1,\ldots,p\\ \phi^2 &\sim \Gamma^{-1}(a,b).\end{aligned}\tag{2.7}$$

Consider the case where $\boldsymbol{\mu}=0$, which is commonplace on the basis of neutrality in scientific reporting. Then,

$$\begin{aligned}-\beta_k &= \sigma_0\frac{-\theta_k}{\sqrt{\sum_{j=1}^{p}\theta_j^2+2\sum_{1\leq l}\sum_{l<m\leq p}\theta_l\theta_m\rho_{l,m}+\phi^2}}\\ &\stackrel{d}{=} \sigma_0\frac{-(-\theta_k)}{\sqrt{\sum_{j=1}^{p}(-\theta_j)^2+2\sum_{1\leq l}\sum_{l<m\leq p}(-\theta_l)(-\theta_m)\rho_{l,m}+\phi^2}}\\ &= \beta_k\end{aligned}$$

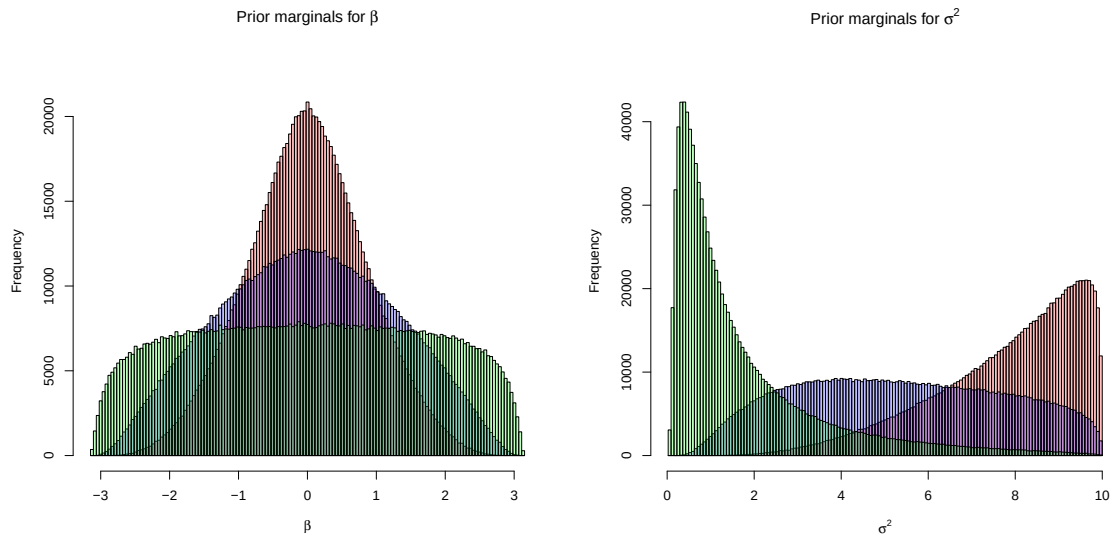


Fig. 2.1 Marginal priors induced by Equation 2.7 with independent covariates, $\sigma_0^2 = 10, \mu = 0, a = 2, b = 1, p = 3$ for $s^2 = 0.25, 0.5$ and 1.5 in red, blue and green respectively

So symmetry of θ around 0 implies symmetry of β around 0, giving $\mathbb{E}[\beta] = 0$. Similarly, it is clear that multiplying each of $\theta_1, \ldots, \theta_p, \phi$ by a constant c will have no effect on $\beta_1, \ldots, \beta_p, \sigma^2$ as the constant c will cancel in the fraction. Multiplying these parameters by a constant c is equivalent to giving them the prior:

$$\theta_k \sim \mathrm{N}(0, s_k^2 c^2)\ k = 1, \ldots, p \tag{2.8}$$

$$\phi^2 \sim \Gamma^{-1}(a, bc^2). \tag{2.9}$$

Therefore, this prior induces the same prior for $\beta_1, \ldots, \beta_p, \sigma^2$ for any constant c. This suggests that the important thing to consider when specifying a prior becomes the relative magnitudes of the parameters, rather than their absolute magnitudes. Figure 2.1 shows the prior marginals for β_1 and σ^2 for different variances of θ, with the prior for ϕ^2 fixed. Clearly when the prior variance of θ is small relative to ϕ^2, β becomes more concentrated around 0 and the prior for σ^2 shifts to make up the unexplained variance. Similarly, when the prior variance for θ becomes large, more of the variation in Y becomes attributable to

variance in the covariates, resulting in the prior putting more mass on smaller σ^2 values. In fact, a good heuristic for specifying the prior for $\theta_1, \ldots, \theta_p, \phi$ is to treat them as if $Y \mid X, \boldsymbol{\theta}, \phi^2 \sim \mathrm{N}(X\boldsymbol{\theta}, \phi^2)$; that is to ignore the transformation, and then let the transformation incorporate the desired joint distribution. In this case the marginal response variance would be given by $\sum_{j=1}^{p} \theta_j^2 + 2\sum_{1 \leq l} \sum_{l < m \leq p} \theta_l \theta_m \rho_{l,m} + \phi^2$, which under our prior has expected value $\sum_{j=1}^{p} s_j^2 + \frac{b}{a-1}$. So, under the unconstrained prior, the expectation of the marginal response variance decomposes into a sum of terms corresponding to each parameter. A prior belief in the relative degrees of variance explained by noise, and each covariate, can be translated into hyperparameters with this expression in mind; this mirrors the parameter interpretation of normal-gamma priors explained in Griffin & Brown (2010). This decomposition does not hold for the transformed parameters, but it serves as a guideline for assigning relative magnitudes in the prior.

Though in the prior each θ_j and ϕ^2 are independent, the transformation will restrict $\beta_1, \ldots, \beta_p, \sigma^2$ to the subspace where the marginal variance constraint holds, leading to a strong dependence structure, as seen in Figure 2.2. The dependence between β_1 and σ^2 when using the constrained priors is clear, in contrast to an independent prior.

The incorporation of the marginal constraints introduces an analysis of variance (ANOVA) interpretation. In this context, the interpretation of ε changes from that of a "noise" variable to one that explains the variance in Y unexplained by X. The conjugate prior in Figure 2.2 seems particularly inappropriate in this context, as larger values of σ^2 are associated in larger absolute values of β. In the simplest case, where covariates are uncorrelated, the marginal variance becomes: $\mathrm{Var}(Y \mid \beta, \sigma^2) = \sum_{j=1}^{p} \beta^2 + \sigma^2$. In this case we have a very simple and interpretable decomposition of the variation in Y into: variation attributable to each covariate, and additional variation. In the more general case, this interpretation becomes clouded by the covariance terms: variance can no longer be attributed to a covariate, as the covariance terms cannot be attributed to a single covariate. If there are blocks of covariates that are

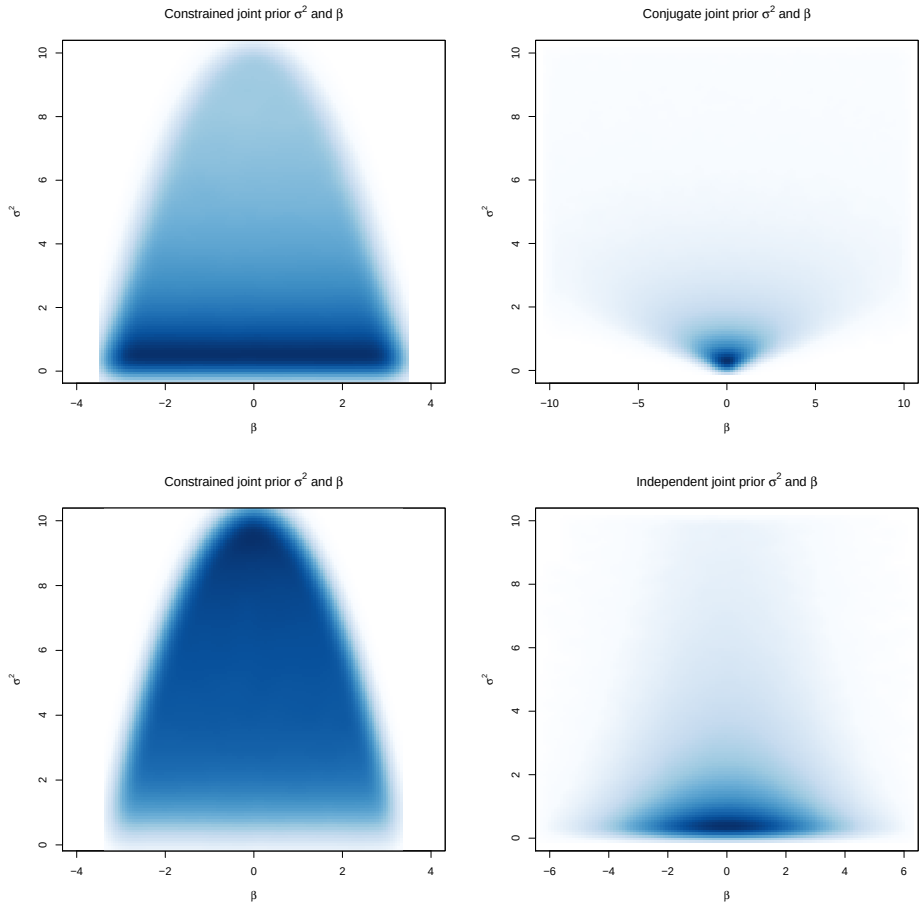


Fig. 2.2 Joint prior for (β_1, σ^2) under the marginal constraint $\sigma_0^2 = 10$ as in Figure 2.1 with $s^2 = 1.5$ (top left),compared to the corresponding untransformed independent prior (bottom right), the conjugate prior (top right) and the constrained prior with $s^2 = 0.5$ (bottom left)

only correlated within blocks, we can consider variance as attributable to each block. For example, in heritability, where covariates are genetic and environmental, it is common to create experiments in such a way that genetic covariates are independent of environmental covariates, which are each independent of the noise. Then the variance sum decomposes into terms involving genetics, environment, and noise. Hertability is then defined as the proportion of the total variance attributable to genetics. In such applications it is natural to think of the total variance as fixed, with interest being in the nature of the decomposition. It is useful to bear in mind this analysis of variance interpretation when analyzing properties of the prior and posterior.

2.3.1 In conjunction with sparsity priors

So far, we have only considered normal priors for θ within the context of the constrained marginal prior. In certain applications, it is known *a priori* that only a small subset of β are non-zero. This has led to the popularity of sparsity priors such as the Bayesian Lasso (Park & Casella, 2008), the horseshoe (Carvalho *et al.*, 2010) and normal-gamma priors (Griffin & Brown, 2010); Polson & Scott (2010) contains a thorough review of such approaches. These priors allow for greater diversity of covariate effects than standard priors, accounting for both large and near-zero effects. We consider the application of such priors to $\boldsymbol{\theta}$ within a marginally constrained prior. θ_j being close to zero should result in a corresponding β_j close to zero, so we might expect to inherit some properties from the sparsity prior on θ. Following Tibshirani (1996), the ability of sparsity priors to induce "shrinkage" - where smaller $\hat{\beta}$ are pulled to zero more than for larger values of $\hat{\beta}$ - has been of particular interest. Since the marginally constrained prior is joint over (β, σ^2), standard results cannot be applied in this context. As such, examination of the "shrinkage factor" is beyond the scope of this thesis. Rather, we show qualitatively that a sparsity prior on θ leads to many intuitive properties of

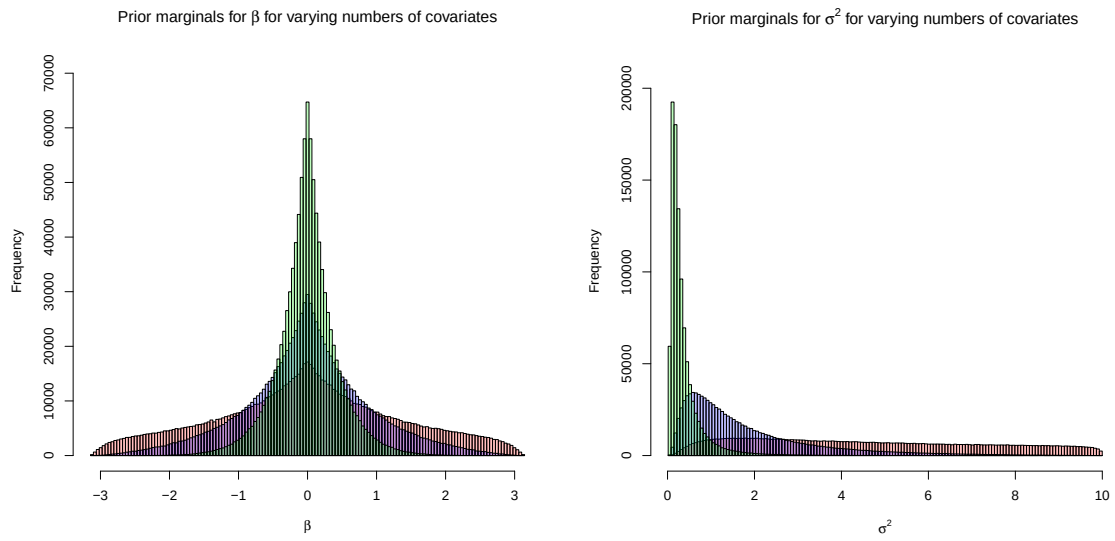


Fig. 2.3 Priors for β_1 (left) and σ^2 (right) using a Laplace prior on θ with 3, 10 and 50 covariates in red, blue and green respectively.

the implied prior on β. For simplicity, in the examples in this section, covariates are assumed to be independent and the marginal variance is known to be 10.

First, consider a Laplace prior for θ in the cases $p = 3, 10$ and 50:

$$\begin{aligned} \theta_k &\sim \text{Laplace}(0,2) \quad k = 1,\ldots,p \\ \phi^2 &\sim \Gamma^{-1}(2,1). \end{aligned} \tag{2.10}$$

In the context of the ANOVA interpretation, it becomes clear how the addition of more covariates should effect the constrained prior. As more potential sources of variation are introduced, the magnitude of each must fall in order to satisfy the variance constraint. In practice, in the presence of large numbers of covariates, it is common to reduce the prior variance of the parameters, for example in genetics. This is a natural consequence of the marginal constraint, as Figure 2.3 displays. Intuitively, as the number of covariates increases, each covariate effect, and the noise variance, must fall to satisfy the constraint. The shape of the β prior can be seen to mimic the shape of the Laplace, when compared to the use of a Gaussian prior on θ in Figure 2.1.

Next, we examine the use of the normal-gamma prior, as its parameterization leads to a natural prior specification within the context of a marginally constrained prior. Standard implementation of the prior is as follows:

$$\begin{aligned} p(\sigma^2) &\propto \sigma^{-2} \\ \lambda_j &\sim \Gamma(a,b) \quad j = 1,\ldots,p+1 \\ \beta_j \mid \lambda_j &\sim \mathrm{N}(0,\lambda_j^2) \quad j = 1,\ldots,p+1. \end{aligned}$$

Griffin & Brown (2010) examine the marginal variance of the response, given hyperparameters $\lambda_1,\ldots,\lambda_p$:

$$\mathrm{Var}[Y \mid \boldsymbol{\lambda},\sigma^2] = \sum_{j=1}^{p} \lambda_j + \sigma^2$$

where, as usual, we assume Y has been centred to give $\alpha = 0$, and covariates have been centred and scaled to have 0 expectation and variance 1. Note that this marginal variance is different in nature to the marginal variance $\mathrm{Var}[Y \mid \boldsymbol{\beta},\sigma^2]$ used in the marginal constraint. $\mathrm{Var}[Y \mid \boldsymbol{\beta},\sigma^2]$ conditions on $\boldsymbol{\beta}$, which gives an expression for the marginal variance in a sample of the response. Here, we instead condition on hyperparameters, effectively taking the expectation of the marginal variance under the prior on $\beta \mid \lambda$ (since Y has expectation 0). Since the priors for each β_j are independent, $\beta_j\beta_k\mathrm{Cov}(X_j,X_k)$ has expectation 0. This marginal variance can be used to inform hyperparameter specification. In particular, Griffin & Brown (2010) note that $\frac{\lambda_k}{\sum_{j=1}^{p}\lambda_j}$ - which can be interpreted as the proportion of regression variability attributable to x_j - follows a $\mathrm{Dir}(a,\ldots,a)$ distribution. As such, the distribution of the proportions of variation attributable to each covariate is governed by a. Large a will result in similar proportions of variation attributable to each covariate, whereas small a results in the majority of variation being attributable to a small number of covariates. This provides an interpretable way to specify prior information of this form. Here we apply this prior to parameters θ and inspect the behaviour induced on the transformed parameters β.

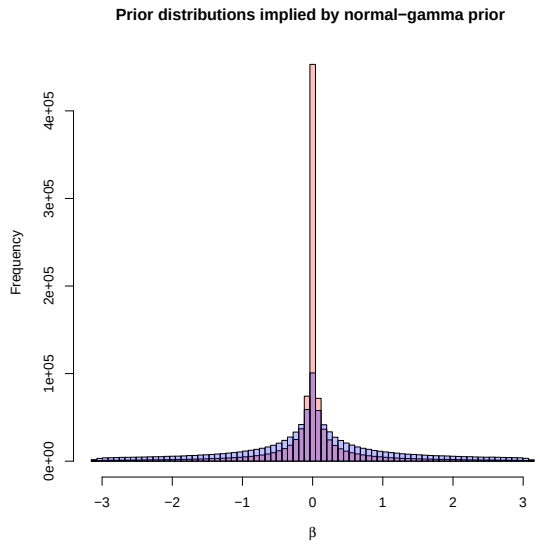


Fig. 2.4 Marginal priors for β_1 given by the normal-gamma prior in equation 2.11 for $a = 1$ (blue) and $a = 0.3$ (red).

Concretely, we use the prior:

$$\begin{aligned} \phi^2 &\sim \Gamma^{-1}(2,5) \\ \lambda_j &\sim \Gamma(a,1) \quad j = 1,\ldots,p \\ \theta_j \mid \lambda_j &\sim \mathrm{N}(0,\lambda_j^2) \quad j = 1,\ldots,p. \end{aligned} \tag{2.11}$$

with (β, σ^2) defined by equation 2.6.

Figure 2.4 shows the implied marginal prior for β_1 for $a = 1$ and $a = 0.3$. This prior places large mass near 0, but also allows for larger values of β. Clearly $a = 0.3$ leads to increased support close to zero, further supported by Figure 2.5, which shows the joint prior distribution of (β_1, β_2) for $a = 1$ and $a = 0.3$. This joint prior has a similar form to the use of sparsity priors in a regular setting. One key feature of the normal-gamma prior is the tendency to generate β that are either large, or close to zero, in contrast to a normal prior with fixed variance. Figure 2.6 shows the distribution of $\frac{\beta_1^2}{\sum_{j=1}^p \beta_j^2}$ - the proportion of non-noise variance in Y attributable to the first covariate - using constrained normal-gamma priors with $a = 1$

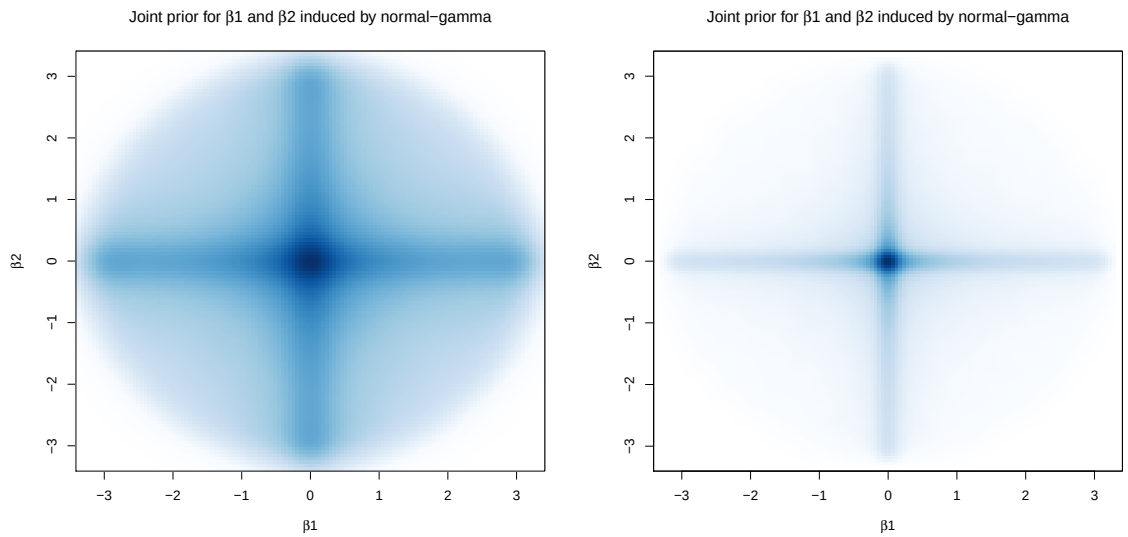


Fig. 2.5 Joint priors for (β_1, β_2) given by the normal-gamma prior in equation 2.11 for $a = 1$ (left) and $a = 0.3$ (right).

and $a = 0.3$ in contrast to the constrained prior with $\theta \sim \mathrm{N}(0, 1)$. This is done in the case of few covariates ($p = 3$) and for many covariates ($p = 20$). Normal-gamma priors for θ result in a more skewed distribution of variance explained by each parameter. In the more extreme case of $a = 0.3$, this proportion is highly likely to be very low or represent a significant proportion of the variation attributable to covariates. When $p = 3$, this is rather extreme, with a relatively high probability that nearly all variation is explained by one covariate. As p increases to 20, the proportion naturally decreases, as there are more sources of covariate variation. Nonetheless, in each case the normal-gamma prior leads to more extreme values of this proportion than a standard normal prior. The effect of a is similar to that when used in an unconstrained normal-gamma prior, with smaller a leading to fewer sources of covariate variation. This should be tuned in accordance with prior beliefs about the number of predictive covariates. Given this hyperparameter, b and the distribution of ϕ^2 should be set to reflect any beliefs in the proportion of variation explained by parameters relative to noise.

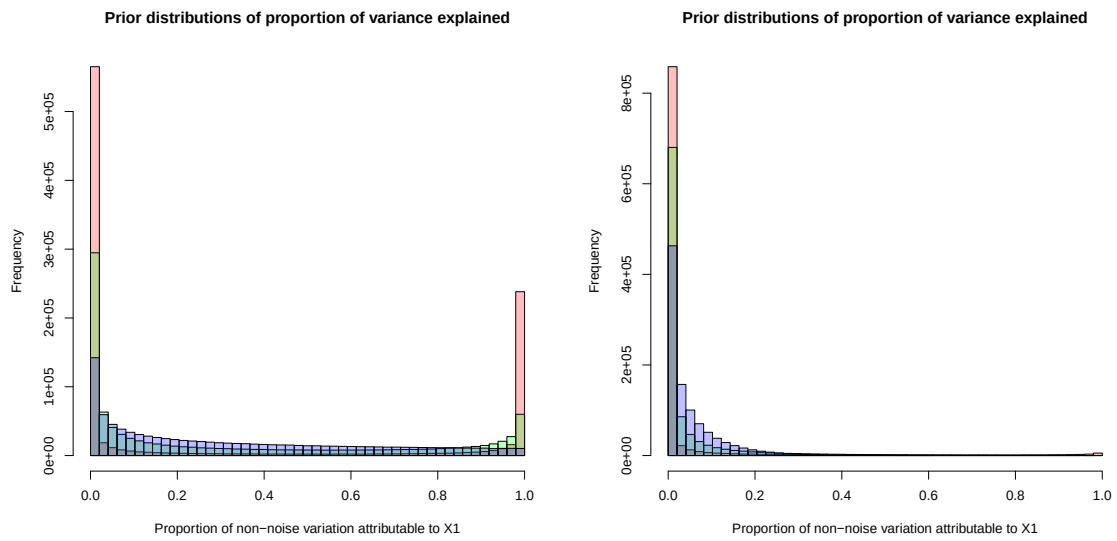


Fig. 2.6 Distribution of the variation due to the first covariate, as a proportion of the total variation due to covariates, for $p = 3$ (left) and $p = 20$ (right) using the standard normal prior (blue) compared to the normal-gamma prior with $a = 1$ (green) and $a = 0.3$ (red).

2.4 Marginal coherence in variable selection

A common aim of a linear regression analysis is the identification of the subset of covariates that have an effect on the response. As mentioned in section 2.3.1, sparsity priors are often employed to shrink small estimated covariate effects to near 0 in order to determine the covariates with significant effect. Another approach is the use of "spike and slab" priors (Mitchell & Beauchamp, 1988), where each regression parameter is a mixture of a standard prior and a point mass at 0:

$$p(\beta_j) = z_j \mathrm{N}(\beta_j \mid 0, \tau_j) + (1 - z_j)\delta_{\beta_j = 0}$$

where $p(z_j = 1) = w$ specifies the probability that $\beta_j = 0$. The resulting posterior provides the probability of inclusion of each parameter in the model. The problem can alternatively be framed in terms of multiple hypotheses, where each hypotheses corresponds to a different model. In the case of linear regression, each hypotheses is defined by the subset of covariates

contained in the model. See Chipman *et al.* (2001) for a comprehensive review. This section is concerned with coherence in the specification multiple hypotheses in the context of linear regression. In particular, we examine a property "marginal coherence", which is defined as the equality in distribution of the marginal mean and variance of Y across each hypothesis. This is achieved via an application of marginally constrained priors and leads to an interpretable change in prior as more covariates are included in the model.

Start by considering a very simple model selection application: Y represents a clinical response in a drug trial, for example red blood cell count. X is a binary variable, taking the value 1 if the patient is assigned the drug, and $X = 0$ if the patient is assigned to a placebo. Interest lies in the effectiveness of the drug, formally, in the choice of hypothesis:

$$H_0 : Y \sim \mathrm{N}(\alpha^{(0)}, {\sigma^{(0)}}^2)$$
$$H_1 : Y \sim \mathrm{N}(\alpha^{(1)} + X\beta, {\sigma^{(1)}}^2)$$

When forming such hypotheses, the superscript is often omitted, but care must be taken to consider the interpretation of each parameter within the context of the corresponding hypothesis. In a Bayesian framework, priors are placed on each parameter to reflect prior beliefs under each hypothesis. It is important to consider if these prior beliefs are correctly represented within each hypothesis. The hypotheses result in the following conditional means and variances:

$$\mathbb{E}[Y \mid X = x, \alpha^{(0)}, H_0] = \alpha^{(0)}$$
$$\mathbb{E}[Y \mid X = x, \alpha^{(1)}, \beta^{(1)}, H_1] = \alpha^{(1)} + \beta^{(1)} x$$
$$\mathrm{Var}[Y \mid X = x, \alpha^{(0)}, H_0] = {\sigma^{(0)}}^2$$
$$\mathrm{Var}[Y \mid X = x, \alpha^{(1)}, \beta^{(1)}, H_1] = {\sigma^{(1)}}^2.$$

Placing priors on these parameters induces a prior distribution on each of these conditional moments. Within the context of the drug trial, it seems natural to have fixed beliefs about the conditional distribution of the red blood cell count of an individual given the placebo, regardless of hypothesis. The presence or absence of drug effect should have no bearing on the distribution of red blood cell count among patients not given the drug. Since the conditional distributions are Gaussian, it is enough to match means and variances to reflect this belief; using identical priors for $(\alpha^{(0)}, {\sigma^{(0)}}^2)$ and $(\alpha^{(1)}, {\sigma^{(1)}}^2)$ achieves this, resulting in the equality of the conditional expectations and variances of the response given $X = 0$ across each hypothesis. In some sense, the hypotheses are now "conditionally coherent" (for $X = 0$). Of course, the conditional expectations given $X = 1$ differ in this case. The hypotheses can be made "conditionally coherent" for any one choice of $X = x_0$ for a given x_0; the requirement of the priors is $\alpha^{(0)} \overset{d}{=} \alpha^{(1)} + \beta^{(1)} x_0$, where x_0 is the value of interest. The simplest way to achieve this is to shift the covariate so that $X = 0$ corresponds to the value of interest; then, conditional upon $X = 0$, the response distribution is invariant to change of hypothesis. The extension to multiple hypotheses and covariates is natural, where $X_j = 0$ for $j = 1, \ldots, p$.

This highlights the care required to express prior beliefs across hypotheses in a coherent fashion: if $X = 1$ for the treated, but now $X = -1$ (rather than $X = 0$) for the placebo, using the same priors for α and σ^2 across hypotheses will result in inconsistency in the prior distributions for the mean blood cell count for patients not taking the drug between the two hypotheses. This drug trial example clearly calls for equality of this distribution between hypotheses of drug effect, so it is appropriate to use $X = 0$ to correspond to the placebo, or to specify priors such that $\alpha^{(0)} \overset{d}{=} \alpha^{(1)} + \beta^{(1)} x_0$, where x_0 is the value of X corresponding to the placebo. However, this does not represent all applications: take an example of the same regression model where now Y corresponds to height and X corresponds to hours of sleep per night, where the dataset corresponds to adult males sampled from the UK population. In this case, rather than equality of conditional means and variances given a

particular value of X, it seems natural to instead fix the marginal expectation and variance across hypotheses. Intuitively, one might believe that the mean and variance of the height of UK males is invariant to hypotheses regarding the effect of sleep, and wish to reflect this in the prior. Rather than a sleep effect introducing variation in heights, any sleep effect is thought to explain variation observed in heights. This raises the question of a potential notion of "marginal coherence", which we define as when the marginal expectation and variance of Y (integrating out X) is given the same prior distribution across hypotheses. (Note that the marginal distribution of Y depends on the distribution of X, so is no longer Gaussian in general). Therefore, equating means and variances does not result in the same distribution., but nonetheless, equality of mean and variance in the prior is a natural requirement of a cohere. The hypotheses imply the following properties of the marginal response:

$$\begin{aligned}
&\mathbb{E}[Y \mid \alpha^{(0)}, H_0] = \alpha^{(0)} \\
&\mathbb{E}[Y \mid, \alpha^{(1)}, \beta^{(1)}, H_1] = \alpha^{(1)} + \beta^{(1)}\mathbb{E}[X] \\
&\mathrm{Var}[Y \mid \alpha^{(0)}, H_0] = {\sigma^{(0)}}^2 \\
&\mathrm{Var}[Y \mid X = 0, \alpha^{(1)}, \beta^{(1)}, H_1] = {\beta^{(1)}}^2 \mathrm{Var}[X] + {\sigma^{(1)}}^2
\end{aligned}$$

The simplest way to equate the marginal mean is to centre covariates to give $\mathbb{E}[X] = 0$. Then simply placing the same prior on $\alpha^{(0)}$ and $\alpha^{(1)}$ achieves this. In order to equate the marginal variances, clearly ${\sigma^{(0)}}^2$ and ${\sigma^{(1)}}^2$ must be treated differently to one another, as $\mathrm{Var}[X] \neq 0$. The task is to specify priors for ${\sigma^{(0)}}^2$ and $(\beta, {\sigma^{(1)}}^2)$ that equate the marginal variances in distribution. To achieve this, the marginally constrained prior of the previous section can be applied with a prior placed on the now unknown marginal response variance, σ_0^2. This is presented in the general variable selection setup.

Given response Y and p covariates in a vector X, interest lies in the choice of hypothesis, where each hypothesis corresponds to the assertion that only a subset of these covariates have

non-zero effects, denoted by a vector of 1s and 0s, γ. For example, if $p = 5$, $\gamma = (0,1,0,0,1)$ corresponds to the hypothesis that the only non-zero β are β_2 and β_5. Let X_γ denote the vector of covariates restricted to just the covariates where $\gamma_j = 1$. The maximum number of hypotheses in this setup is 2^p, with one hypothesis for each possible value of γ, where each hypothesis takes the form:

$$H_\gamma : Y \sim \mathrm{N}(\alpha^{(\gamma)} + X_\gamma \beta^{(\gamma)}, {\sigma^{(\gamma)}}^2)$$

where $\beta^{(\gamma)}$ is a column vector of parameters of length $\sum_{j=1}^{p} \gamma_j$. It is important to note that these parameters are different between hypotheses, despite possibly misleading notation. For example, $\beta^{(1,0,0,0,0)}$ is not the same parameter as the first value of the vector $\beta^{(1,1,1,1,1)}$. Intuitively, the coefficient of a covariate is interpreted as the change in response given a unit change, with all other covariates fixed. Clearly, this definition depends on the nature of the other covariates contained in the model, so should be treated as such. It shall become apparent that different treatment of these parameters is a requirement of marginal coherence. First, priors are set for parameters common to each hypothesis. For each γ, the prior for H_γ is given by:

$$\begin{aligned}
\alpha^{(\gamma)} &\sim p_\alpha \\
{\sigma_0^{(\gamma)}}^2 &\sim p_{\sigma_0^2} \\
\theta^{(\gamma)} &\sim p_{\theta^{(\gamma)}} \\
{\phi^{(\gamma)}}^2 &\sim p_{{\phi^{(\gamma)}}^2} \\
\beta^{(\gamma)} &= \frac{\sigma_0^{(\gamma)} \theta^{(\gamma)}}{\sqrt{{\theta^{(\gamma)}}^T \Sigma_x \theta^{(\gamma)} + {\phi^{(\gamma)}}^2}} \\
{\sigma^{(\gamma)}}^2 &= \frac{{\sigma_0^{(\gamma)}}^2 {\phi^{(\gamma)}}^2}{{\theta^{(\gamma)}}^T \Sigma_x \theta^{(\gamma)} + {\phi^{(\gamma)}}^2}.
\end{aligned}$$

From the previous section:

$$\mathbb{E}[Y \mid H_\gamma, \alpha^{(\gamma)}, \beta^{(\gamma)}, {\sigma^{(\gamma)}}^2] \sim p_\alpha$$
$$\mathrm{Var}[Y \mid H_\gamma, \alpha^{(\gamma)}, \beta^{(\gamma)}, {\sigma^{(\gamma)}}^2] \sim p_{\sigma_0^2}.$$

Using this construction, the prior distributions for these values is the same under all hypotheses, satisfying our definition of "marginal coherence" in expectation and variance. This holds for any choice of priors, so it is convenient to simplify the prior distributions, setting $p_{\theta^{(\gamma)}}(\theta^{(\gamma)}) = \prod_{j=1}^{|\gamma|} p_\theta(\theta_j^{(\gamma)})$ and $p_{\phi^{(\gamma)^2}} = p_{\phi^2}$. To complete the prior, $\mathbb{P}(H_\gamma)$ must be specified for each γ.

To ease the burden of notation, consider a simple nested hypothesis example:

$$\begin{aligned}
&H_0 : Y \sim \mathrm{N}(\alpha^{(0)}, {\sigma^{(0)}}^2) \\
&H_1 : Y \sim \mathrm{N}(\alpha^{(1)} + \beta_1^{(1)} X_1, {\sigma^{(1)}}^2) \\
&\vdots \\
&H_p : Y \sim \mathrm{N}(\alpha^{(p)} + \sum_{j=1}^{p} \beta_j^{(p)} X_j, {\sigma^{(p)}}^2)
\end{aligned}$$

where covariates are centered about their mean, scaled to have unit variance and are known to be independent, for simplicity. A typical analysis might drop all superscripts and place priors on each parameter:

$$\begin{aligned}
\alpha &\sim \mathrm{N}(0,1) \\
\beta_j &\sim \mathrm{N}(0,1) \;\; j = 1, \ldots, p \\
\sigma^2 &\sim \chi^2(1)
\end{aligned}$$

In this toy example the non-standard χ^2 prior for variance is used for analytical tractability.

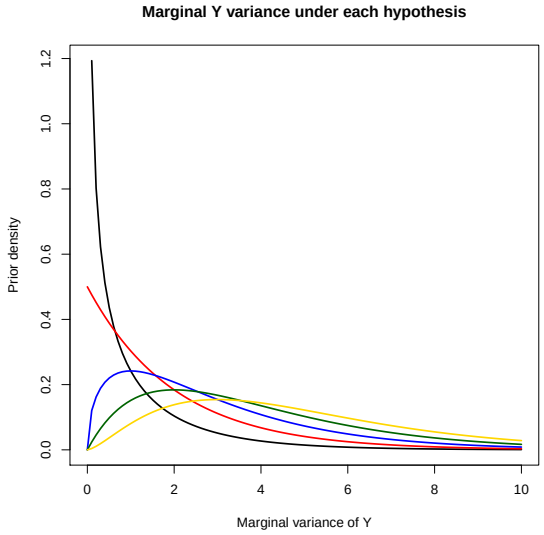


Fig. 2.7 Distribution of the marginal variance of Y induced by standard priors for H_k for $k = 0,1,2,3,4$ in black, red, blue, green and yellow respectively. Under the marginally coherent prior all hypotheses induce the blue prior.

A marginally coherent approach specifies priors by placing a $\chi^2(3)$ prior on the marginal response variance:

$$
\begin{aligned}
\alpha^{(k)} &\sim \mathrm{N}(0,1) \\
{\sigma_0^{(k)}}^2 &\sim \chi^2(3) \\
\theta_j^{(k)} &\sim \mathrm{N}(0,1) \;\; j = 1,\ldots,p \\
{\phi^{(k)}}^2 &\sim \chi^2(1) \\
\beta^{(k)} &= \frac{\sigma_0^{(k)}\theta^{(k)}}{\sqrt{{\theta^{(k)}}^T \Sigma_X \theta^{(k)} + {\phi^{(k)}}^2}} \\
{\sigma^{(k)}}^2 &= \frac{{\sigma_0^{(k)}}^2 {\phi^{(k)}}^2}{{\theta^{(k)}}^T \Sigma_x \theta^{(k)} + {\phi^{(k)}}^2}.
\end{aligned}
$$

Both approaches result in a $\mathrm{N}(0,1)$ prior on the marginal variance of Y (because covariates are centred at their means), but the traditional approach results in a $\chi^2(k+1)$ prior on the marginal variance of Y given H_k. Figure 2.7 shows this inconsistency between hypotheses

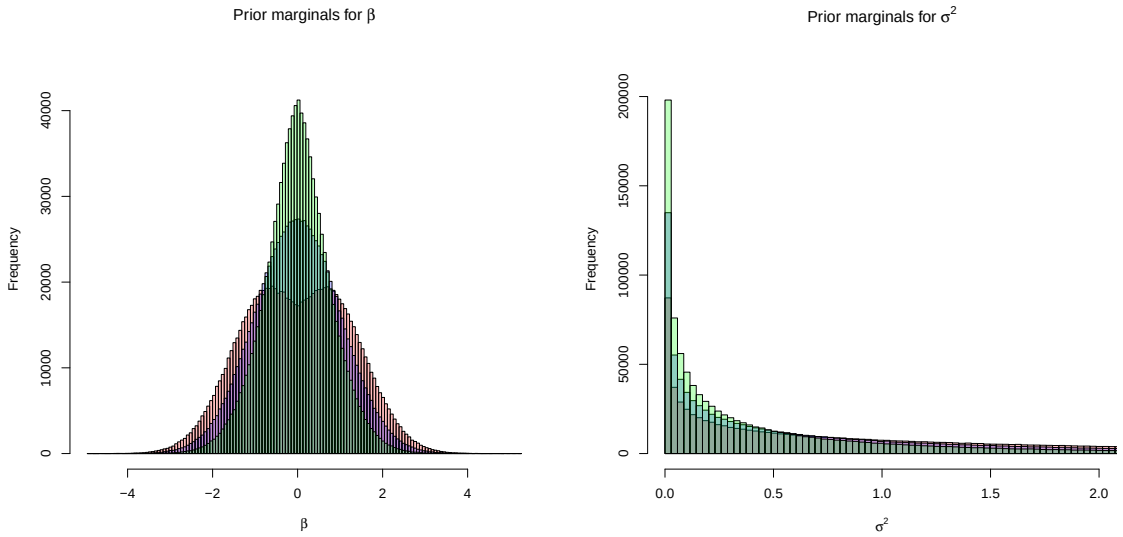


Fig. 2.8 Prior marginals for β_1 (left) and σ^2 (right) using the marginally coherent prior given H_1 (red), H_4 (green) and the corresponding prior for all hypotheses in the standard setting (blue).

using the traditional approach. In contrast, the marginally coherent prior allows us to specify the common distribution, in this case set at a compromise between the distributions seen in the traditional approach. A result of this is a change in the prior distributions of β and σ^2 parameters between hypotheses, compared to the fixed priors traditionally used. Figure 2.8 shows this property, which has the same "decomposition of variance" interpretation as the constrained marginal variance prior in section 2.3. The β_1 prior under H_1 has higher variance than under the higher dimensional H_4, as there are fewer potential sources of variation. The prior even becomes multi-modal with modes either side of zero, in order to explain the variance expected in the population. Even if Y's marginal distribution is not of particular interest, if prior beliefs are such that hypotheses should not change this marginal distribution, a standard prior is misspecified. Practically, this could mean parameter variance being inflated in higher-dimensional models and deflated variance for lower-dimensional models, compared to what a marginally coherent prior would reflect, as seen in Figure 2.8.

2.5 Illustrations

2.5.1 Simulation example

In this section the marginally constrained prior is applied to a simulated dataset of size $n = 200$. There are three covariates, which are mean-zero, multivariate Gaussian random variables with covariance matrix given by:

$$\begin{pmatrix} 1 & 0.3 & -0.2 \\ 0.3 & 1 & 0.5 \\ -0.2 & 0.5 & 1 \end{pmatrix}$$

.

The response is given by:

$$Y_i = \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon_i$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$; $\beta_1 = 1; \beta_2 = 2; \beta_3 = -0.4$; and $\sigma^2 = 2$. This implies a marginal response variance of 7.7. Assume this variance is known *a priori*, along with the covariance structure of X.

The prior on $\theta_1, \ldots, \theta_p, \phi^2$ is given by:

$$\theta_j \sim \mathrm{N}(0, 0.5) \;\; j = 1, 2, 3$$
$$\phi^2 \sim \Gamma^{-1}(2, 1).$$

The choice of hyperparameters reflects a belief that variation due to “noise" is about twice that of from any one covariate, with each covariate treated equally, as $\mathbb{E}[\phi^2] = 2\mathbb{E}[\theta_1^2]$. Of course, in practice, this specification would reflect prior belief in the relative magnitudes of effects, possibly simulating from the prior to produce a plot similar to Figure 2.1, aiding this

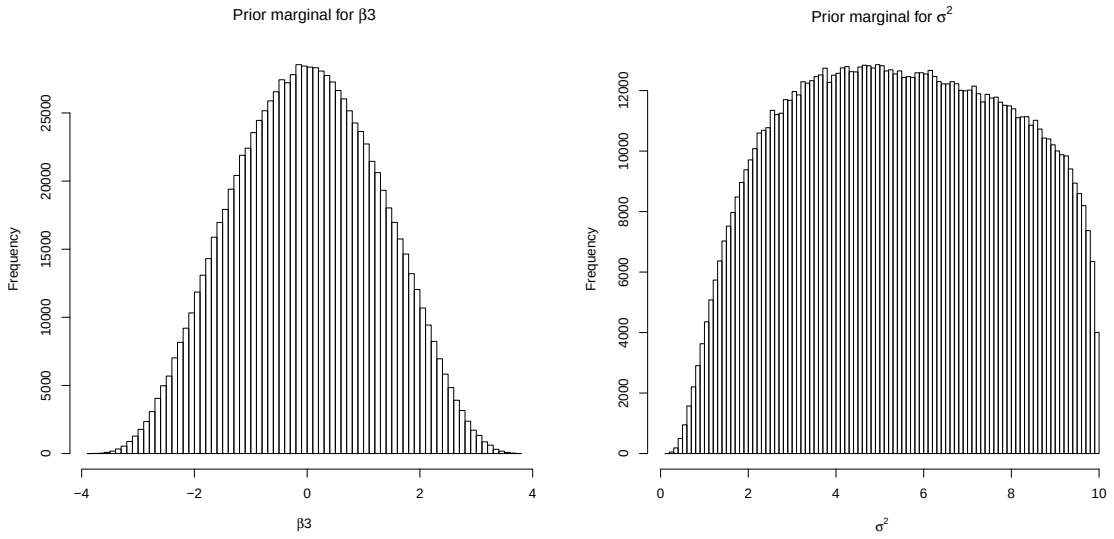


Fig. 2.9 Marginal priors for β_1 (left) and σ^2 (right) under marginal constraint $\sigma_0^2 = 7.7$

choice. Using the known marginal variance σ_0^2 and covariate covariance matrix, these are transformed into the parameters of interest $\beta_1, \ldots, \beta_p, \sigma^2$, as shown in Figure 2.9, which is similar - but not identical - to the blue prior in Figure 2.1, due to the difference in covariate covariance. Clearly the marginal prior for each β is non-Gaussian after the transformation and σ^2 is bounded by the known marginal variance.

The covariance structure of the covariates has an interesting effect on the corresponding β: as $X_{i,1}$ and $X_{i,2}$ are positively correlated, as β_1 increases and is positive, the probability that β_2 is negative increases; effectively in order to ensure the $2\rho_{1,2}\beta_1\beta_2$ term is negative, which dampens the increase in β_1^2. Alternatively, the negative $\rho_{1,3}$ results in increased probability that β_1 and β_3 share the same sign. Figure 2.10 displays this structure, which mimics the nature of the dependence when using the g-prior. However, two major differences are highlighted between these priors: firstly, the g-prior has heavy Gaussian tails not present in the constrained prior, which are essentially truncated by the constraint; and secondly, the shaded plots provide a visualisation of the three-dimensional dependence between β_1, β_2 and σ^2. As expected, in the constrained prior, large absolute values of β_1 and β_3 result in small

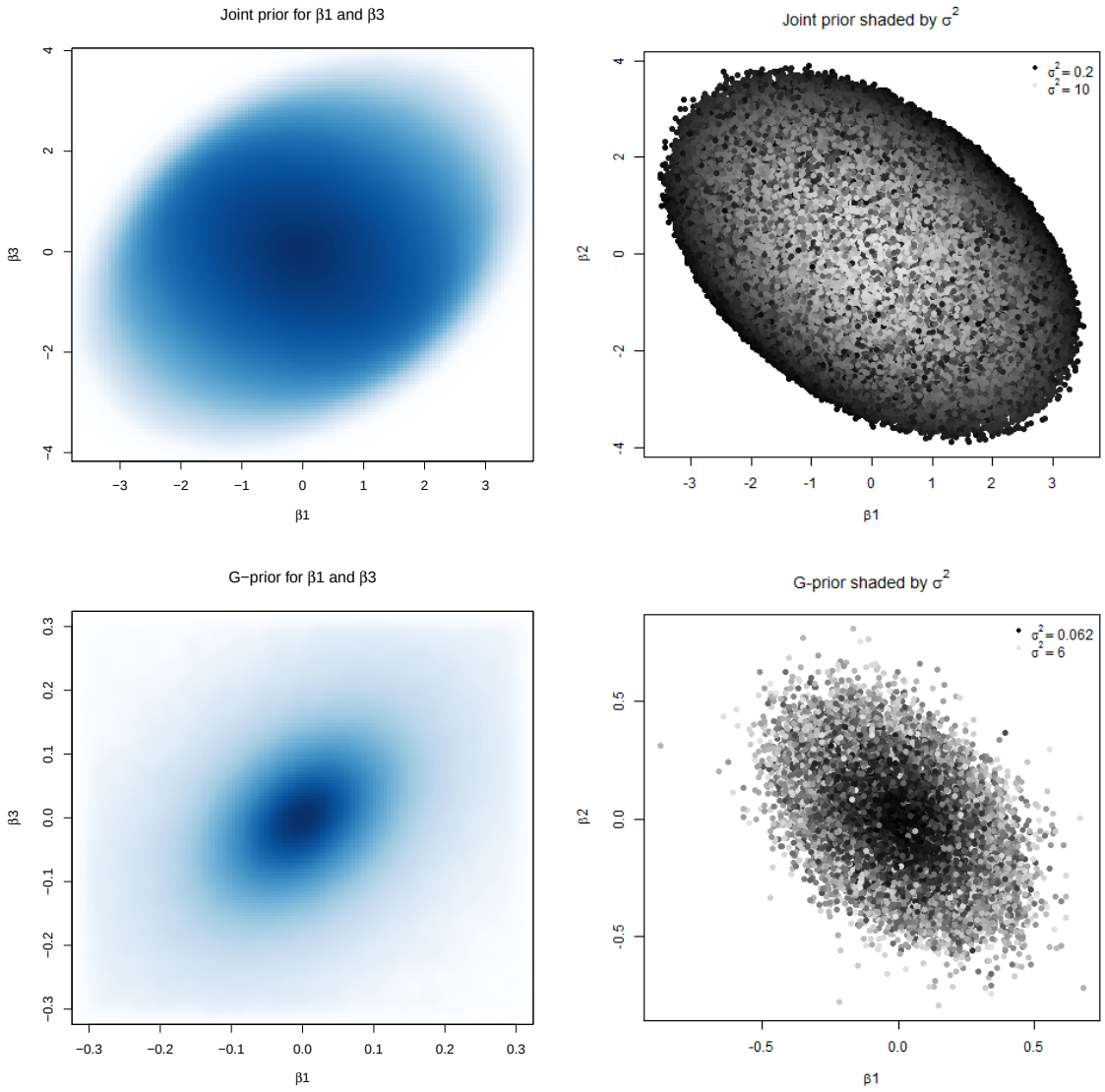


Fig. 2.10 Joint prior of (β_1, β_3) (top left) and (β_1, β_2) shaded by σ^2 (top right) with the marginal constraint (top left). Joint g-prior for (β_1, β_3) (bottom left) and (β_1, β_2) shaded by σ^2 (bottom right).

values of σ^2, with larger values found when their absolute values are small. The opposite relationship found in the g-prior is a direct consequence of the definition.

Posterior Simulation

Posterior sampling is simple using random walk metropolis. The simulated data is analyzed using the marginal constraint prior, and using the corresponding prior without the parameter transformation, with adjusted hyperparameters to reflect similar prior marginals using both approaches. With $n = 200$, the likelihood is dominant enough to result in similar marginal posteriors for each parameter $\beta_1, \beta_2, \beta_3, \sigma^2$ using both priors. However, Figure 2.11 displays how the dependence structure from the marginally constrained prior transfer into the posterior, resulting in a structure not seen in a traditional analysis. The joint posterior of β_1, β_2 shaded by σ^2 demonstrates a posterior dependence structure similar to that in the prior, shown in Figure 2.10. This utilisation of the extra marginal information should aid prediction.

The hyperparameters used in the independent prior were manually tuned to best match the constrained prior to aid comparison. This is required because the marginally constrained prior automatically scales the prior to have support that is in accordance with the known marginal variance. This raises an interesting point regarding the scale of the prior: we've established that priors 2.8 are equivalent for any choice of c, where a standard independent prior is clearly not robust to this form of misspecification, which in extreme cases can have a hugely harmful effect on the posterior. The conjugate prior is also robust to this misspecification in the sense that if the true parameter values are all significantly larger than in the prior, information is shared between the parameters through the built in dependence between σ^2 and β.

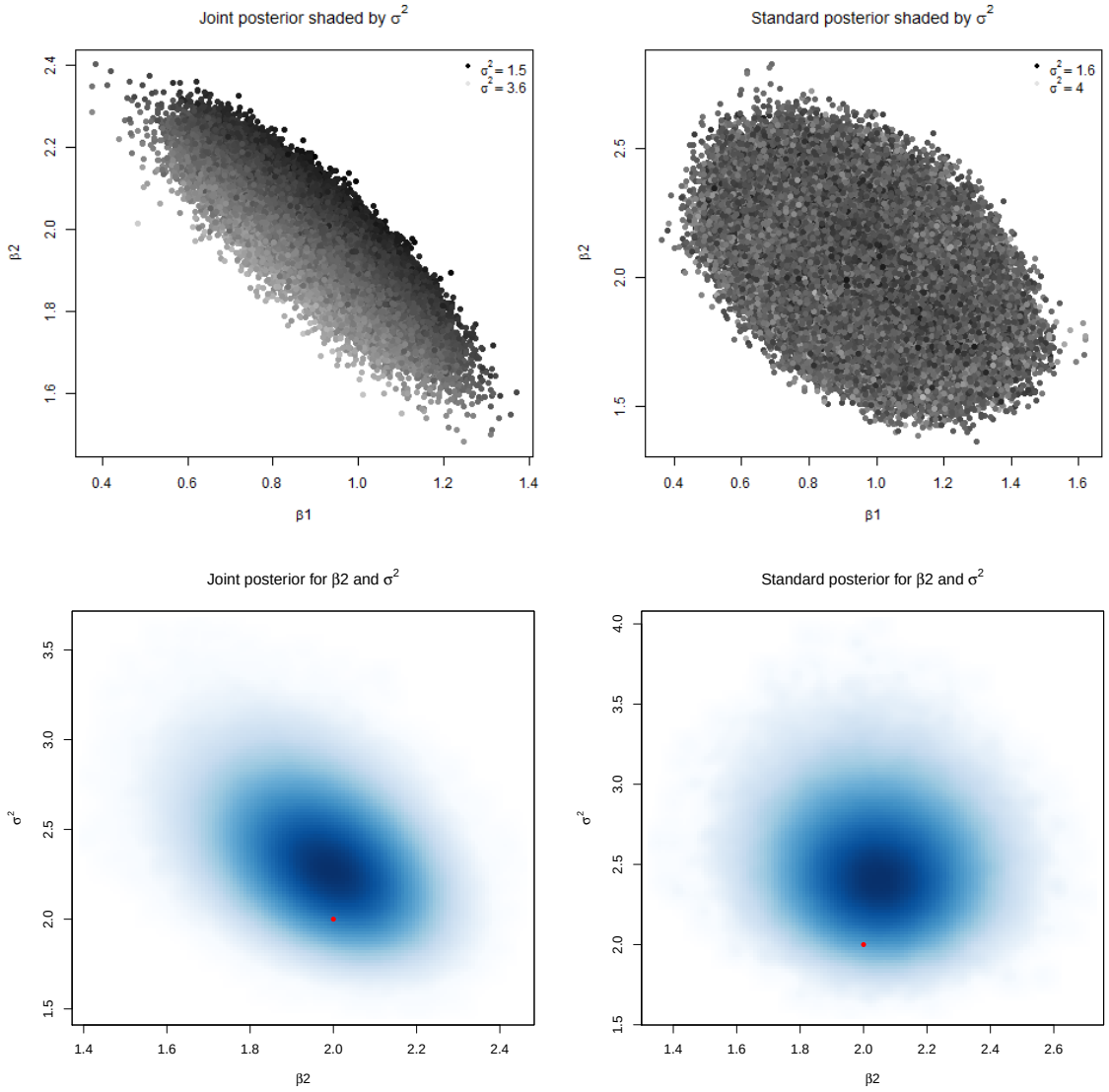


Fig. 2.11 Joint posterior of (β_1, β_2) shaded by σ^2 using the constrained prior (top left) and an independent prior with $\mu = 0, s^2 = 1, a = 5, b = 15$ (top right). Joint posterior of (β_2, σ^2) using the constrained prior (bottom left) and the independent prior (bottom right) with the red dot corresponding to the correct parameter value.

2.5.2 US Census application

In this section the constrained marginal variance prior is applied to a subset of the US census data[3] for personal annual income. The response is given by annual income in US dollars, which we predict using covariates: marital status, quarter of birth, age, gender, average hours worked per week and weeks worked per year. We split the data into: a marginal Y dataset $\{y_i^\dagger\}_{i=1}^{500,000}$, a marginal X dataset$\{x_i^*\}_{i=1}^{500,000}$, and joint data $\{y_i, x_i\}_{i=1}^n$ for both $n = 50$ and $n = 200$. Marginal means and variances are estimated using these marginal datasets, which are then used to define a marginally constrained prior for the linear regression analysis of the small joint dataset.

The categorical variable quarter of birth takes 4 values, each corresponding to a quarter of the calendar year. The categorical variable for marital status takes 5 values, corresponding to: married, widowed, divorced, separated, and never married. Rather than a standard dummy coding scheme, the requirement of centred covariates motivates the use of an alternative coding scheme. A variant of effects coding provides centered dummy variables with natural interpretation (see, e.g. Cohen *et al.*, 2013, Chapter 8 for an introduction to coding schemes for categorical variables). One value of the categorical variable is used as a reference level, then each other value of the categorical variable is given a corresponding dummy variable, which takes one of three values: one corresponding to the value itself, another corresponding to the reference level, and the value 0, taken otherwise. We wish to choose these values based on the estimated distribution of the categorical variable from the marginal X data in order for each dummy variable to have mean 0. For example, marital status takes 5 values, with probabilities p_1, p_2, p_3, p_4 and p_5 respectively, which are estimated from the marginal data. Never married ($x = 5$) is chosen as a baseline, so there are 4 dummy variables, which we denote by e_1, e_2, e_3 and e_4. Then, for $j = 1, 2, 3, 4$:

[3] http://www.census.gov/acs/www/data_documentation/pums_data/

$$e_j = \begin{cases} a_j, & \text{if } x = j \\ b_j, & \text{if } x = 5 \\ 0, & \text{otherwise} \end{cases} .$$

In order to ensure $\mathbb{E}[e_j] = 0$, we must have $b_j = -a_j \frac{p_j}{p_5}$; the choice of a_j determines the variance of e_j. Throughout, we have scaled covariates to have variance 1 in order to ease notation; however, in the case of categorical variables, this can hinder parameter interpretation. The parameter β_j corresponds to the change in mean response for a unit change in e_j. If all e_j have unit variance, the a_j will differ, resulting in differing interpretations for β_js. By fixing the value of a_j for all values of j, all effects for each (non-reference) value of the categorical variable will be on the same scale. Concretely, we use the coding in Table 2.1, so that each variable has mean 0, each parameter is on the same scale, and the variance of the first dummy variable is 1.

Table 2.1 Coding for categorical variables

Variable value	e_1	e_2	e_3	e_4
1	$\frac{p_5}{p_1}\sqrt{\frac{p_1}{p_5p_1+p_5^2}}$	0	0	0
2	0	$\frac{p_5}{p_1}\sqrt{\frac{p_1}{p_5p_1+p_5^2}}$	0	0
3	0	0	$\frac{p_5}{p_1}\sqrt{\frac{p_1}{p_5p_1+p_5^2}}$	0
4	0	0	0	$\frac{p_5}{p_1}\sqrt{\frac{p_1}{p_5p_1+p_5^2}}$
5	$-\sqrt{\frac{p_1}{p_5p_1+p_5^2}}$	$-\frac{p_2}{p_1}\sqrt{\frac{p_1}{p_5p_1+p_5^2}}$	$-\frac{p_3}{p_1}\sqrt{\frac{p_1}{p_5p_1+p_5^2}}$	$-\frac{p_4}{p_1}\sqrt{\frac{p_1}{p_5p_1+p_5^2}}$

All other covariates are centered and scaled in the usual way. Denote by Σ_x the covariance matrix of X estimated using $\{x_i^*\}_{i=1}^{500,000}$, after covariates are coded and scaled appropriately. Using the marginal dataset $\{y_i^\dagger\}_{i=1}^{500,000}$, the mean and variance of Y are estimated to be $\$49,000$ and 3.7×10^9. This is used to center Y to have mean 0, and for the sake of numerical

stability, is scaled to have variance 1. The standard linear model is given by:

$$Y_i \sim \mathrm{N}(\alpha + \sum_{j=1}^{11} X_{i,j}\beta_j, \sigma^2).$$

We compare analysis using an independent prior:

$$\begin{aligned}
\alpha &\sim \mathrm{N}(0, 0.1^2) \\
\beta_j &\sim \mathrm{N}(0, 0.1^2) \;\; j = 1, \ldots, 11 \\
\sigma^2 &\sim \Gamma^{-1}(2, 1).
\end{aligned}$$

and the marginally constrained prior, where $\sigma_0^2 = 1$:

$$\begin{aligned}
\alpha &= 0 \\
\theta_j &\sim \mathrm{N}(0, 0.1^2) \;\; j = 1, \ldots, 11 \\
\phi^2 &\sim \Gamma^{-1}(2, 1) \\
\boldsymbol{\beta} &= \frac{\boldsymbol{\theta}}{\sqrt{\boldsymbol{\theta}^T \Sigma_X \boldsymbol{\theta} + \phi^2}} \\
\sigma^2 &= \frac{\phi^2}{\boldsymbol{\theta}^T \Sigma_x \boldsymbol{\theta} + \phi^2}.
\end{aligned}$$

Priors for non-variance parameters are mean 0 for neutrality. The variance of 0.1^2 was chosen to reflect a belief that the change in salary for unit change in any covariate is in the interval $(-\$12,000, \$12,000) = (-0.2\sqrt{3.7 \times 10^9}, 0.2\sqrt{3.7 \times 10^9})$ with probability 0.95. This seems reasonable given the standardisation of covariates. The same prior was used for α to reflect a 0.95 probability that the average salary is in the interval $(\$37,000, \$61,000)$. The prior for the variance is chosen to have mean 0.8, corresponding to 80% of the variation in Y being unexplained by covariates, in expectation. For the marginally constrained prior, we use the heuristic of placing the standard independent prior on $(\boldsymbol{\theta}, \phi^2)$. Figure 2.12 displays the

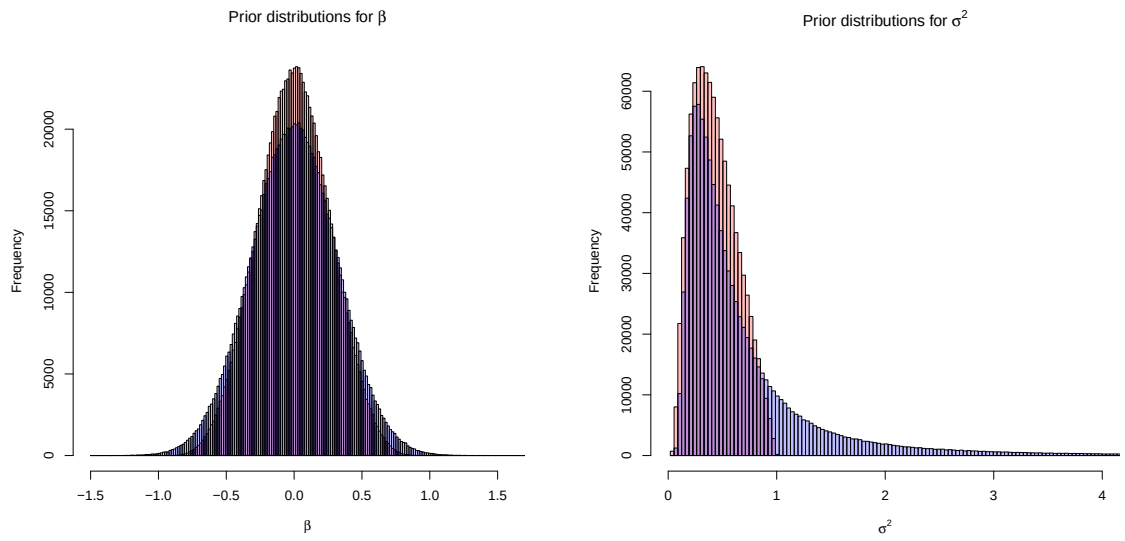


Fig. 2.12 Priors for β_1 (left) and σ^2 (right) under marginally constrained prior (red) and an independent prior (blue).

priors for β_1 and σ^2 under the independent and marginally constrained priors. The marginal priors for β_1 are similar, with the constrained prior more tight around 0. However, there is a stark contrast between the priors for σ^2. The marginally constrained prior is similar to the independent prior constrained to be on the interval $[0,1]$. The "known" marginal variance of 1 constitutes an upper bound for σ^2, as highlighted by the constrained prior. In contrast, the independent prior is far more flat, with a tail supporting values over 3.

Figure 2.13 shows the joint prior distribution of the covariate effects corresponding to weeks worked a year and hours worked a week, shaded by σ^2, using both the constrained and the independent priors. Since these covariates are positively correlated ($\rho_{10,11} = 0.34$), the corresponding covariate effects are negatively correlated in the constrained prior. In addition, there is a clear dependence on σ^2, with large absolute values of covariate effects corresponding to lower σ^2. The independent prior naturally does not exhibit any structure of this kind. Figure 2.14 shows how this translates into the posterior (for $n = 200$). Both posteriors now have negative correlation between β_{10} and β_{11}, but while there is no dependence on σ^2 when using the independent prior, there is a very clear structure in the constrained posterior. As we

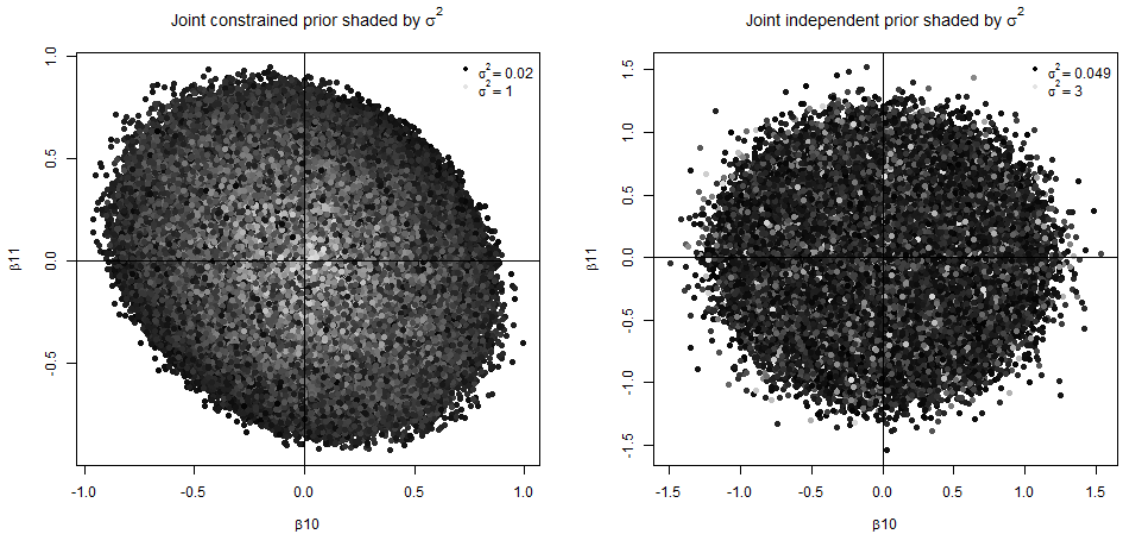


Fig. 2.13 Joint priors for covariate effects corresponding to hours worked a week (β_{10}) and weeks worked a year (β_{11}) under the constrained prior (left) and the independent prior (right), shaded by the prior for σ^2.

have seen, larger covariate effects result in smaller σ^2 as there is less residual variance to explain.

Now we examine the posteriors using both priors for both $n = 50$ and $n = 200$. Figure 2.15 shows how the posterior for σ^2 under the constraint supports higher values than under the independent prior, despite the independent prior supporting higher values. It would seem that the small ($n = 50$) dataset has lower variation than that seen in the marginal dataset $\{y_i^\dagger\}_{i=1}^{500,000}$. There is simply more information incorporated into the constrained posterior, which is apparent from this posterior, where the standard posterior seems to indicate a smaller variance than one would expect, given the marginal variance. For the larger dataset, the posteriors start to converge, and both support higher values of σ^2 than in the smaller dataset.

There is a clear trend in the marginal posteriors for covariate effects. Given the smaller dataset, the constrained posteriors have a higher variance, and modes further from 0 than the corresponding independent posteriors. It seems that the large marginal variation is being spread around the covariate effects. There is not much evidence in the small dataset for

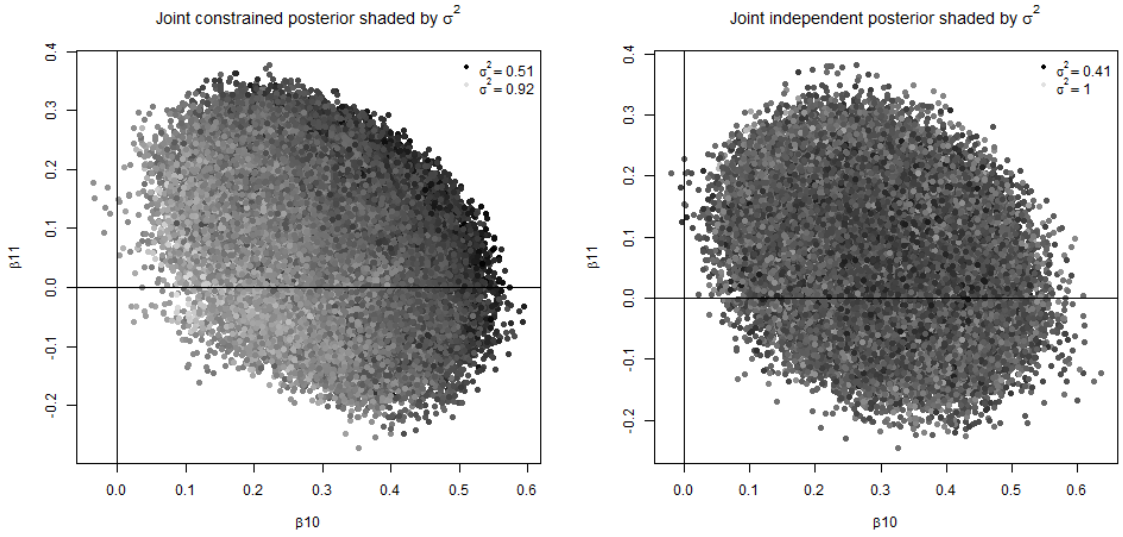


Fig. 2.14 Joint posteriors, given data of size $n = 200$, for covariate effects corresponding to hours worked a week (β_{10}) and weeks worked a year (β_{11}) under the constrained prior (left) and the independent prior (right), shaded by the posterior for σ^2.

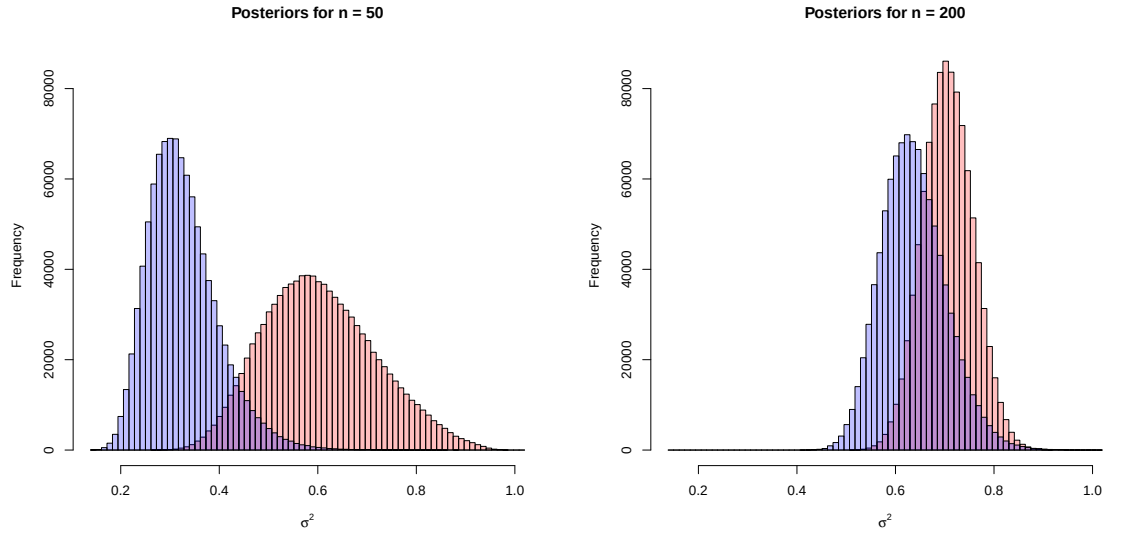


Fig. 2.15 Posteriors for σ^2 given data of size $n = 50$ (left) and $n = 200$ (right), under the marginally constrained prior (red) and the independent prior (blue).

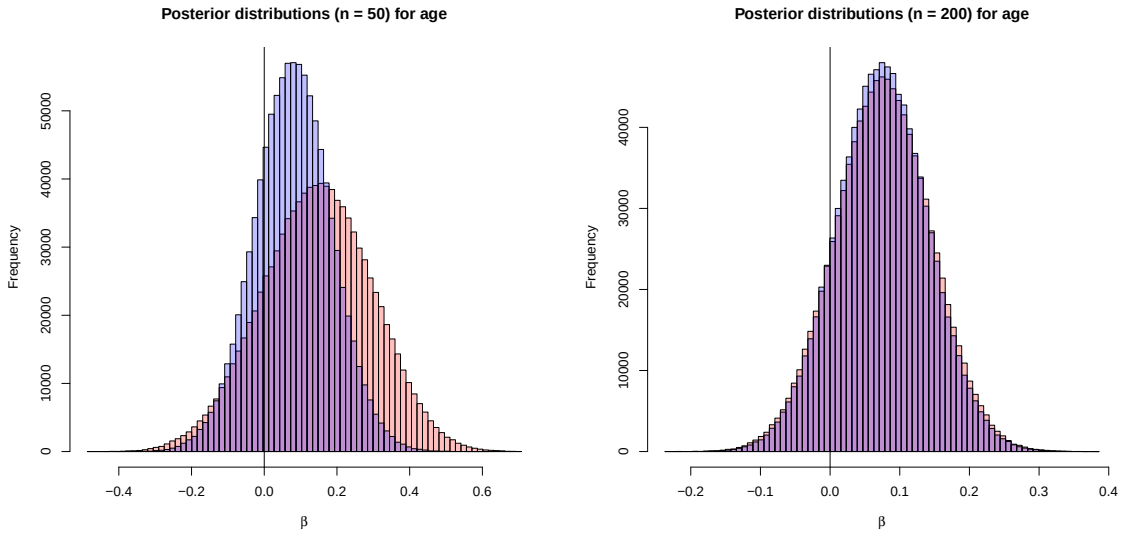


Fig. 2.16 Posteriors for the effect of age.

significant deviation of these effects from 0, but the constrained prior must account for the known variation, leading to larger supported absolute values of covariate effects. Using the larger dataset, the likelihood starts to dominate and the posteriors become very similar. In addition, this larger dataset seems to provide sufficient evidence of covariate effects to satisfy the constrained model. There is one exception to this trend: the covariate effect corresponding to "widowed", a rare value of the "marital status" categorical variable, which has no direct observations in the $n = 50$ dataset. The posterior under both priors is diffuse, due to lack of information, but in this case the constrained prior is more tight around zero, as in the prior.

2.6 Discussion

We have seen the impact of incorporating marginal information through a constraint on the implied marginal mean and variance in the linear model. One particularly compelling aspect is the strong, intuitive dependence structure gained through such a constraint, which should aid prediction accuracy. In the absence of marginal response data, the prior provides a

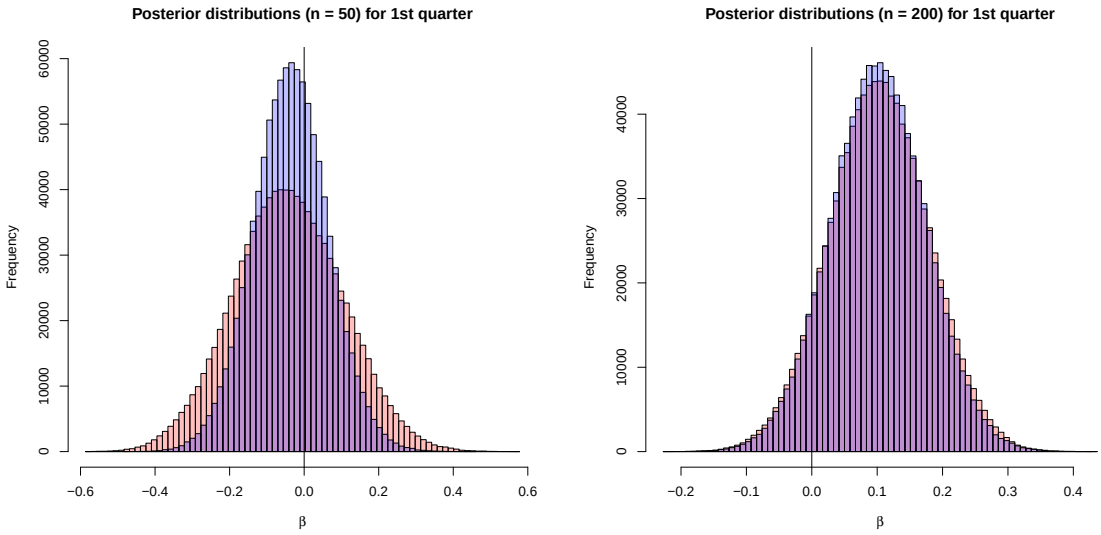


Fig. 2.17 Posteriors for the effect of age.

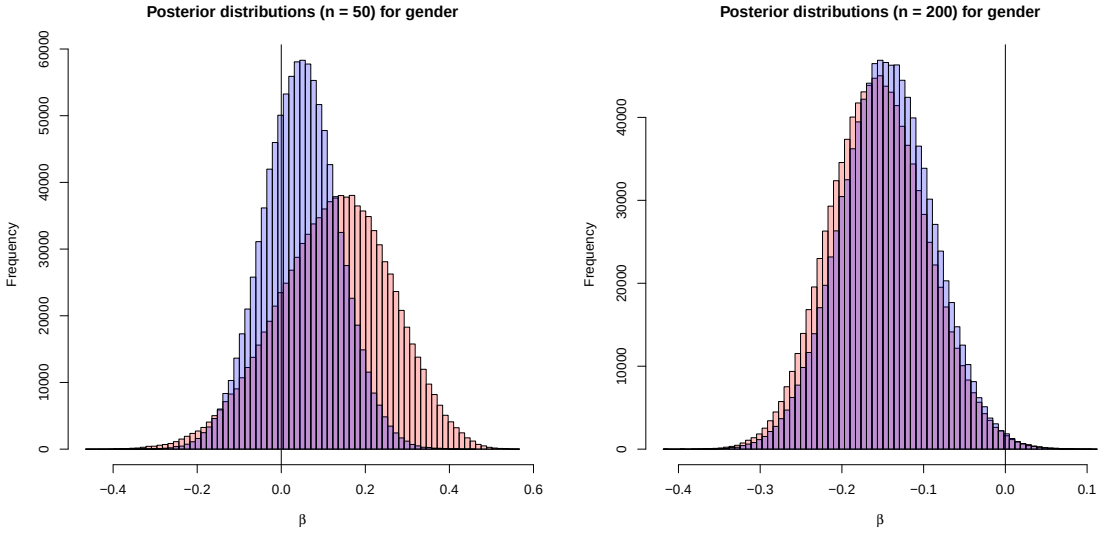


Fig. 2.18 Posteriors for the effect of quarter of birth.

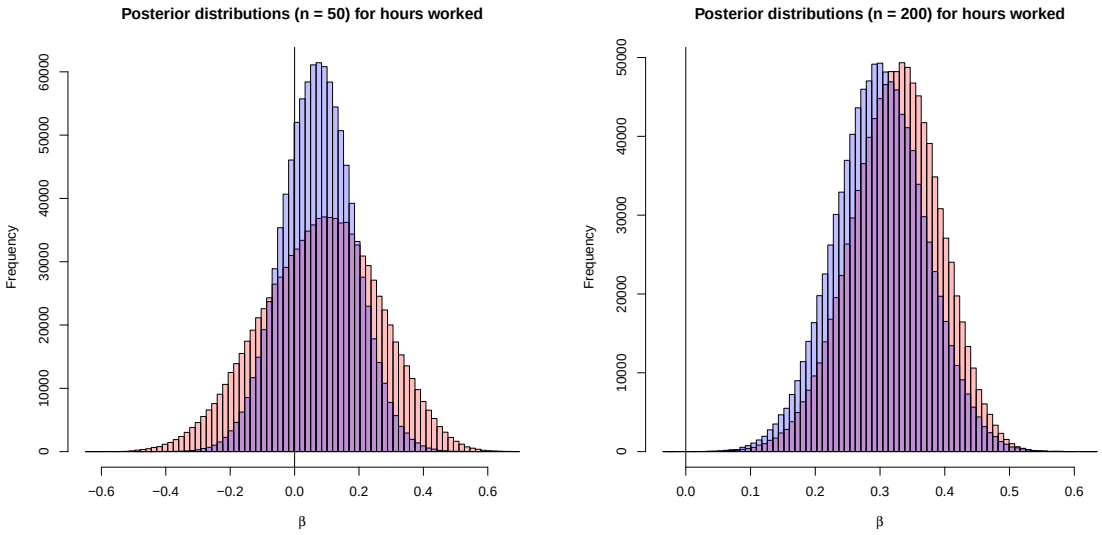


Fig. 2.19 Posteriors for the effect of hours worked per week.

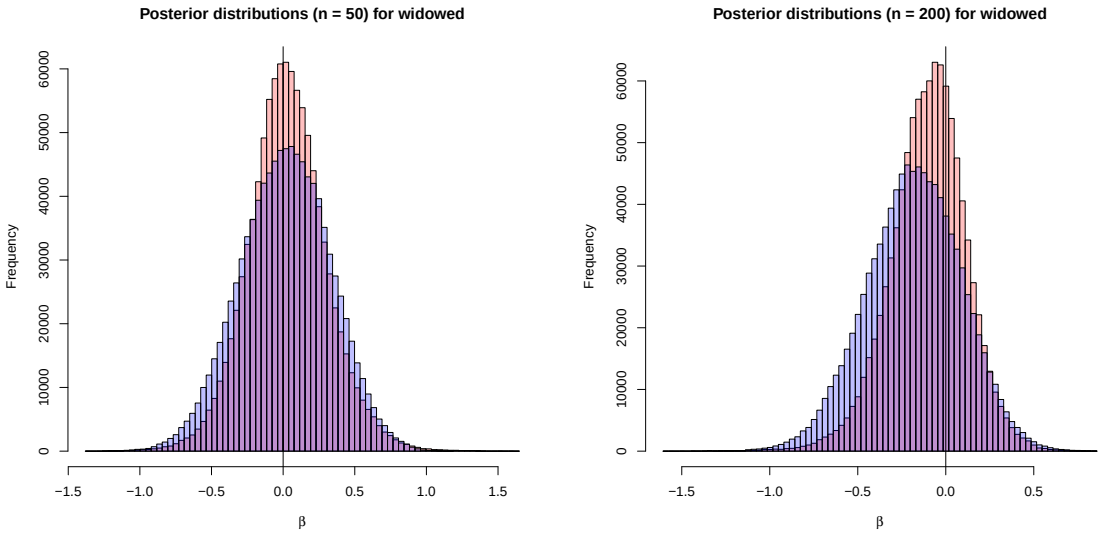


Fig. 2.20 Posterior for the categorical variable indicating that the individual is widowed

means of specifying priors for variable selection that are marginally coherent. Fundamentally, when suitable, large marginal datasets are available, the constraint incorporates a great deal of information, reducing the dimensionality of the model space by 2; however, care must be taken to ensure the marginal data corresponds to the same population as the joint data of interest. In the case of the US census data, a normal model is essentially misspecified, due to lack of negative values and heavy tales of the true distribution. In the face of such misspecification, the use of marginal constraints only further restricts the model space, accentuating any major misspecification. This lack of normality might be the cause of inflated covariate effects in this example, as there is more variance in the sample than can be explained by the linear model. In such cases, it is more appropriate to turn to larger classes of models to avoid misspecification; Bayesian nonparametrics provides such a class of flexible models. In the next chapter we explore a Bayesian nonparametric regression model that is constructed using the full marginal distribution of Y, enabling simple incorporation of marginal data into the flexible model.

Chapter 3

A Scalable Bayesian Nonparametric regression model

3.1 Introduction

Bayesian nonparametric regression offers a flexible and robust way of modeling the dependence between covariates $x \in \mathscr{X}$ and a response variable $Y \in \mathbb{R}$ by using models with larger support than their parametric counterparts. Nonparametric statistical models are motivated by robustness and their ability to capture effects such as outliers, strong nonlinearities or multimodalities, whilst providing probabilistic measures of predictive uncertainty. These models allow complexity to grow with the data size. However, as a result, Bayesian nonparametric regression models are often computationally demanding, posing a problem for application to massive datasets. This chapter aims to address this issue in part, by developing a regression model that lends itself to a useful composite likelihood approximation that enables application on a massive scale. Building on Chapter 2, treating covariates as random variables allows specification of the conditional model in terms of the marginal distribution of the response F_Y and covariate effects β. The model is constructed in such a way that the ranking of responses, given covariates x, has closed form. This allows approximation of

the full likelihood by a convenient factorisation in terms of the marginal distribution of Y and the distribution of the ordering of Y, given covariates. Inclusion of marginal Y and X data becomes trivial given the form of this composite likelihood. The model can be viewed as a re-parameterisation of the proportional hazards model that is marginally coherent in a stronger sense than introduced in Chapter 2, as the marginal distribution of the response is independent of covariate parameters. This lends to a natural extension of the partial likelihood that includes the marginal distribution of the response, allowing efficient computation of the whole model. The form of the model allows easy incorporation of extra information related to covariate bias, given truncated data. The chapter is organised as follows: Section 3.2 provides background on approaches to nonparametric regression, Section 3.3 describes the proposed model, the marginal composite likelihood approach for scalable inference is presented in Section 3.4, and Section 3.5 presents some results of this approach on simulated data and on the US Census dataset.

3.2 Bayesian nonparametric regression

There are many existing approaches to nonparametric regression. The vast majority of approaches directly specify the conditional distribution $Y \mid X = x$ of the response, given covariates. These methods can be broadly categorised as follows: methods that allow for a flexible relationship between the covariates and the response; methods that allow for a flexible conditional distribution for the response, given covariates; and approaches that allow for flexibility of both the conditional distribution and the covariate-response relationship. The first group of models can be thought of as “mean-function methods". This class of model assumes that the model can be expressed in the form:

$y = f(x) + \varepsilon$

where ε has a standard (often Gaussian) distribution. The function f is the object of interest, which in the classical literature is estimated using several methods such as Kernel

smoothing (Gasser & Müller, 1984), spline smoothing (Craven & Wahba, 1978) and Lowess (Cleveland, 1979). Bayesian approaches place a prior on the function f. Such approaches include Gaussian processes (see e.g. Rasmussen, 2006), basis function representations such as splines or kernels (Denison *et al.*, 2002; Müller & Quintana, 2004), wavelets (see Vidakovic, 2009, for a review) and Bayesian regression trees (Chipman *et al.*, 2010).

3.2.1 Semiparametric regression

While many mean-function methods can be considered semiparametric, this section reviews semiparametric models where the conditional distribution of the response, given covariates, has a nonparametric form. Many of the models used for survival analysis fall into this category; the most popular of which is the Cox proportional hazards model (Cox, 1972). Simply, the model specifies:

$$h_x(y) = \lambda(x)h_0(y)$$

where $h_0(.)$ is called the “baseline hazard", the hazard function corresponding to an individual with $x = 0$. $\lambda(x)$ is simply a function of x, commonly taken to be $\lambda(x) = \exp(\beta x)$. Classical analysis of the proportional hazards model uses a “partial likelihood", rather than the full likelihood; this is motivated by the desire to make inferences for β when the baseline hazard function h_0 is completely unknown. Within the context of survival analysis, it is convenient to use the language of event times to describe the response. Let $t_1 < t_2 < \ldots < t_n$ denote the ordered event times, and let $l_j = \sum_{i=1}^{n} i\mathbb{I}[y_i = t_j]$ be the label of the individual that has an event at time t_j. As a principle for deriving the partial likelihood, Cox notes that “In the absence of knowledge of $h_0(.)$, the t_i can provide little or no information about β, for their distribution will depend heavily on h_0."; as such, the partial likelihood uses $\{l_j\}_{j=1}^{n}$ as the data for inference on β. Given that there is an event at time t_1, the probability of $l_1 = j$ -

that this event corresponds to individual j - is exactly a ratio of hazard functions evaluated at t_1:

$$\frac{h_j(t_1)}{\sum_{k=1}^{n} h_k(t_1)} = \frac{\lambda_j}{\sum_{k=1}^{n} \lambda_k}.$$

For each subsequent observation, all that changes is that the sum in the denominator only contains individuals that have not yet had an event; this set of individuals is denoted $R(j)$ at time t_j. Given an ordering l, the partial likelihood becomes:

$$\prod_{i=1}^{n} \frac{\lambda_{l_i}}{\sum_{j \geq i} \lambda_{l_j}} = \prod_{i=1}^{n} \frac{\exp\{x_{l_i}\beta\}}{\sum_{j \in R(i)} \exp\{x_{l_j}\beta\}}$$

This likelihood has a closed form, independent of h_0, allowing inference for β without consideration of h_0, as desired. This has become the standard analysis of survival data in practice, due to the simplicity and specific interest in covariate effects.

This partial likelihood has been applied in a Bayesian context, giving a partial posterior for β where h_0 is ignored (Bissiri *et al.*, 2016; Carlin *et al.*, 1995; Raftery *et al.*, 1995; Ritter & Tanner, 1992); however, this is the exception to the rule, with Bayesian analysis of proportional hazards models rarely utilising the form of the partial likelihood. Instead, it is typical to deal with the full likelihood in order to model h_0 (or alternatively H_0) and β. This enables a full posterior analysis, including posterior prediction. Specifically, a parametric prior is placed on the coefficients β and a nonparametric prior on the baseline cumulative hazard function (Aslanidou *et al.*, 1998; Hjort, 1990; Ibrahim *et al.*, 2001, 1999; Kalbfleisch, 1978). In particular, Kalbfleisch (1978) introduced the use of the Gamma process prior on the baseline hazard and later Hjort (1990) introduced the Beta process prior. These nonparametric priors for H_0 or h_0 can capture very flexible conditional distributions for the response, but there is some restriction imposed on the covariate-response relationship through the proportional hazards assumption. A potential issue with this parameterisation is that h_0

corresponds to the hazards for an individual with $x = 0$, which might prove a difficult object upon which to specify prior beliefs.

Alternatively, Pólya trees have been used as a prior for a “baseline" distribution function in such models. Initially, Walker & Mallick (1999) used Pólya trees for a prior on the error distribution in an accelerated failure time model in log form, as in Ying *et al.* (1995), which results in a linear regression form. Hanson & Johnson (2002) then used a mixture of Pólya trees for a prior on the error term in a linear regression model. Later Hanson (2006) outlined inference methods for using mixtures of Pólya trees in multiple standard regression models, such as generalised linear mixed models, the proportional hazards model, and the proportional odds model.

3.2.2 Nonparametric regression

MacEachern (1999) introduced the Dependent Dirichlet Process (DDP): a probability model over the set of distributions $F_{\mathscr{X}} := \{F_x : x \in \mathscr{X}\}$, which is indexed by covariates x in the covariate space $\mathscr{X}$. This is done in such a way that, marginally, each F_x has a Dirichlet Process prior, and these random distributions are dependent across the covariate space. This is achieved by constructing the random distributions by the Sethuraman stick-breaking process (Sethuraman, 1994) in parallel, letting the weights and locations depend on x. MacEachern (1999) emphasised the “single p" case, where only the locations are allowed to depend on x, and the weights are fixed across different values of x:

$$F_x = \sum_{h=1}^{\infty} p_h \delta(\theta_{h,x}), \text{ for each } x \in \mathscr{X}$$

where $\theta_{h,\mathscr{X}} = \left\{\theta_{h,x} : x \in \mathscr{X}\right\}$ represents a stochastic process for each h. These stochastic processes are independent and identically distributed. The result is that, marginally, each F_x has a DP prior with base measure F_{0x}, where F_{0x} represents the distribution of $\theta_{h,x}$ induced

by the stochastic process $\theta_{h,\mathscr{X}}$. The dependence between these random measures is also dictated by the stochastic process $\theta_{h,\mathscr{X}}$. This general probability model for constructing a collection of dependent random measures has given rise to many special cases that result in interesting models.

Griffin & Steel (2006) proposed another method of manipulating the Sethuraman stick-breaking process in order to incorporate covariates into the Dirichlet Process. The stick-breaking process is governed by the relationship $p_i = V_i \prod_{j=1}^{i-1}(1-V_j)$, where $V_i \sim \text{Beta}(1,\alpha)$. The random vectors $\mathbf{V}$ and θ together define the random distribution F. Changing the order π of these vectors will result in a different distribution F^*, since this will change the p_i defined by the process, changing the weights associated with each atom θ_i, the positions of which are unchanged. Allowing the ordering $\pi(x)$ to depend on x allows us to define a probability model over a family of dependent random distributions $F_{\mathscr{X}}$, as before. Since atoms that appear earlier in the order of the stick-breaking representation are typically assigned higher mass, similarity in the orderings $\pi(x_1)$ and $\pi(x_2)$ will imply similarity between the generated distributions F_{x_1} and F_{x_2}. The form of the ordering process $\{\pi(x)\}_{x\in\mathscr{X}}$ defines this dependence.

Dunson *et al.* (2007) defines the Weighted Mixture of Dirichlet Processes prior to introduce covariates. Given data $\{y_{1:n}, x_{1:n}\}$, a prior distribution over $F_{\mathscr{X}}$ is defined by:

$$F_x = \sum_{j=1}^{n} b_j(x) F^*_{x_j}, \qquad F^*_{x_j} \sim DP(\alpha F_0), \qquad \text{for } j = 1,\ldots,n. \quad \forall x \in \mathscr{X}$$

Where $\mathbf{b}(\mathbf{x}) = (\mathbf{b_1}(\mathbf{x}),\ldots,\mathbf{b_n}(\mathbf{x}))$ is a weight function $\mathbf{b} : \mathscr{X} \to \mathscr{P}^{\mathbf{n}}$, where $\mathscr{P}^n$ denotes the n-dimensional probability simplex. Each b_i depends on the respective x_i. The model assigns a Dirichlet process to each observation, and subsequent observations are drawn from a mixture of these Dirichlet Processes, with weights dependent on similarities between the covariate of the subsequent observation and the observed covariates. A form of the weight function is suggested to be of the form $b_j(x) = \exp(\|x - x_j\|^2)$ re-weighted and normalised.

Thus, observations will cluster together according to this metric on the covariate space. This results in a generalisation of the Blackwell & MacQueen (1973) Pólya Urn scheme, where the conditional distribution on clusters depends on the covariate similarity as well as the usual dependence on the population of each cluster.

Dunson & Park (2008) defines the general class of Kernel Stick-Breaking processes as a probability model for $F_{\mathscr{X}}$. The process is defined as follows:

$$F_x = \sum_{h=1}^{\infty} p_{h,x} F_h^* \qquad p_{h,x} = V_h K(x, Z_h) \prod_{l=1}^{h-1} (1 - V_l K(x, Z_l))$$

$$V_h \sim \text{Be}(a_h, b_h) \qquad F_h^* \sim \mathscr{Q} \qquad Z_h \sim H$$

where H is a probability measure on the covariate space $\mathscr{X}$, $\mathscr{Q}$ is a probability measure on the space of distribution functions on the domain of y, and $K : \mathscr{X} \times \mathscr{X} \to [0,1]$ is a kernel function. If x_1 is similar to x_2, then the weights $K(x_1, Z_h)$, for each h, will be similar to the weights $K(x_2, Z_h)$, meaning that F_{x_1} and F_{x_2} will allocate similar weights to the F_h^*. This results in dependence between the random distributions. In the special case where $F_h^* = \delta_{\theta_h}$ where $\theta_h \sim F_0$, and when $K(x,z) = 1 \; \forall (x,z)$, the process reduces to a standard Dirichlet Process with base measure F_0. Relaxing the assumption on K gives a version of the Dependent Dirichlet Process. Assuming that $K(x,z) = 1 \; \forall (x,z)$ and that Q is a Dirichlet Process measure, results in a Dirichlet Process mixture of Dirichlet processes. Dunson & Park (2008) provide a generalised Pólya Urn scheme corresponding to the Kernel Stick-Breaking process.

In a similar fashion to the construction of Dependent Dirichlet Processes, Trippa *et al.* (2011) define a class of random probability distributions using Pólya trees. They first define what they call a Multivariate Beta process for dependent Beta random variables; these dependent Beta random variables are used to couple Pólya trees by using these random

variables in the construction of the Pólya tree. The result is a class of dependent random distributions that can be indexed by covariates for use in a regression setting. This model, and an extension using mixtures of Pólya trees are applied to lung cancer data.

These regression models are all constructed via the conditional distribution $Y \mid X = x$. An alternative approach to regression is to model the covariates and the response (Y, X) jointly, as outlined in Müller *et al.* (1996); this approach can capture both flexible conditional distributions and a flexible covariate-response relationship, so falls into the fully nonparametric category of models. Issues with high dimensional covariates are raised and addressed by Wade *et al.* (2014).

A major drawback of current Bayesian methods for semiparametric or nonparametric regression is that many methods do not scale well with the number of samples and/or with the dimensionality of the covariates. This chapter proposes a joint Bayesian nonparametric regression model $F_{X,Y}$ that affords an approximation which can scale easily to large data applications. The model is parameterised in terms of the marginal distributions of the response F_Y and covariates F_X, and then a conditional regression model that utilises the two marginal distributions, governed by a function λ_β parameterised by β,

$$F_X \sim \mathbb{P}_X \tag{3.1}$$

$$F_Y \sim \mathbb{P}_Y \tag{3.2}$$

$$\beta \sim \pi_\beta \tag{3.3}$$

where $\mathbb{P}_X$ and $\mathbb{P}_Y$ are nonparametric priors over probability distributions and $\lambda_\beta : \mathscr{X} \to \mathbb{R}_+$ is some parametric regression function of the covariates that characterises the dependence between covariates and response. The parameterisation is similar to that found in copula models (see e.g. Nelsen, 2013); however, rather than constructing dependence between X and Y by a joint distribution $C(F_Y(Y), F_X(X))$, our construction forms a conditional distribution for $F_Y(Y) \mid X$. This formulation is simpler and more intuitive in a regression setting. Using

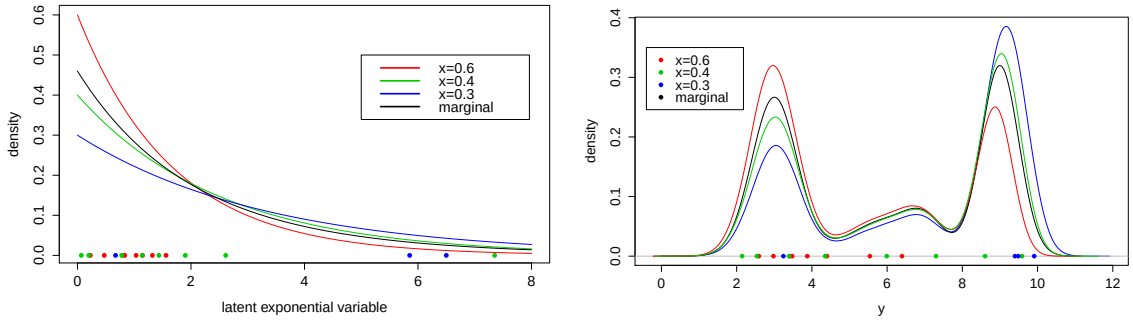
latent exponential variables, we are able to specify a Plackett-Luce model (Luce, 1959; Plackett, 1975) for the order of $F_Y(Y)$ (and therefore Y), given covariates X. The positive function λ_β tunes the stochastic ordering of the responses, given the covariates, such that the ratio $\lambda_\beta(X_i)/\left(\lambda_\beta(X_i)+\lambda_\beta(X_j)\right)$ represents the conditional probability $Pr(Y_i < Y_j|X_i,X_j)$. This property is similar to that found in probabilistic index models (Thas *et al.*, 2012). There is therefore a natural interpretation of the parameters: λ_β tunes the relative ordering of the responses at different covariate values, and F_Y sets the marginal distribution of the responses. This strong interpretability is an important feature as it provides a good vehicle for specifying prior beliefs.

For inference, we propose to use a marginal composite likelihood approach, splitting the likelihood into a component for the ranking probability, and a component for the marginal density for the response; this approach is somewhat similar to the extended rank likelihood for copula estimation in multivariate analysis (D. Hoff, 2007; Murray *et al.*, 2013). However, since our construction leads to a closed-form Plackett-Luce rank distribution, this allows the model to scale to large data applications and allows for the use of standard existing software for Bayesian nonparametric density estimation and Plackett-Luce ranking estimation. As an illustration, we show an application of our approach to a US Census dataset, with over 1,300,000 data points and more than 100 covariates. Code is provided to allow replication of these results [1].

3.3 The statistical model

Let (X_i,Y_i), $i=1,\ldots,n$ be the covariates and responses, and let $\lambda_\beta : \mathscr{X} \to \mathbb{R}_+$ denote the regression function. To build the dependence we introduce a latent random variable Z_i that is

[1] http://www.http://www.stats.ox.ac.uk/~tgdavies



(a) (b)

Fig. 3.1 Illustration of the latent variable used to capture the regression dependence. In (a) we show the distribution of the conditional latent variable, Z, at various points in X assuming a log-linear dependence. In (b) we see the corresponding predictive distributions using a Gaussian mixture model for the marginal, F_Y, shown as the black line. The points in Z shown in (a) are mapped to the points in Y shown in (b) where the ordering is preserved

used to capture the underlying relative level of the response via,

$$Z_i \mid X_i = x_i \sim \text{Exp}(\lambda_{\beta(x_i)}) \tag{3.4}$$

where $\text{Exp}(a)$ denotes the standard exponential distribution of rate a.

The model can be summarised as follows, for $i = 1, \ldots, n$

$$X_i \overset{\text{iid}}{\sim} F_X \tag{3.5}$$

$$Z_i \mid X_i, \beta \overset{\text{ind}}{\sim} \text{Exp}(\lambda_\beta(X_i)) \tag{3.6}$$

$$Y_i = F_Y^{-1}(F_Z(Z_i)) \tag{3.7}$$

where

$$\begin{aligned} F_Z(z) &= \int_{\mathscr{X}} F_{Z|X=x}(z) dF_X(x) \\ &= \int_{\mathscr{X}} \left(1 - e^{-\lambda_\beta(x)z}\right) dF_X(x). \end{aligned}$$

By construction, $F_Z(Z_i)$ is marginally uniformly distributed on $[0,1]$; as a result, $Y_i = F_Y^{-1}(F_Z(Z_i))$ is marginally distributed according to F_Y. The construction via the inverse probability integral transform of a probability integral transform, is similar to that of a Gaussian copula model (see e.g. Song, 2000). A copula C specifies the joint distribution function of $(F_Y(Y), F_X(X))$ in order to define a dependence structure between Y and X. However, here the regression function λ_β - which controls the rate of the exponential random variable Z - defines this relationship. Figure 3.1 shows the correspondence between the conditional exponential random variables, $Z \mid X$, shown in Figure 3.1(a) for differing covariate values, and the resulting conditional response distributions in 3.1(b), where the marginal F_Y is a Gaussian mixture model shown as the black line. We can see visually that the distributions in 3.1(b) are stochastically ordered under the model. The colored points shown in (a) are mapped to the points shown in (b), where again, ordering is preserved.

As F_Y and F_Z are cumulative density functions, $F_Y^{-1} \circ F_Z$ is a monotonically increasing function and

$$\mathbb{P}(Y_i \leq Y_j) = \mathbb{P}(Z_i \leq Z_j) = \frac{\lambda_\beta(x_i)}{\lambda_\beta(x_i) + \lambda_\beta(x_j)}.$$

This clarifies the role of the regression function. More generally, given an ordering $\nu = (\nu_1, \ldots, \nu_n)$ (a permutation of $\{1, 2, \ldots, n\}$), we have:

$$\mathbb{P}(Y_{\nu_1} \leq Y_{\nu_2}, \ldots, \leq Y_{\nu_n}) = \mathbb{P}(Z_{\nu_1} \leq Z_{\nu_2}, \ldots, \leq Z_{\nu_n}) = \prod_{i=1}^{n} \frac{\lambda_\beta(x_{\nu_i})}{\sum_{j \geq i} \lambda_\beta(x_{\nu_j})}.$$

The above model is the Plackett-Luce model (Luce, 1959; Plackett, 1975) - popular in the ranking literature - which also corresponds to the partial likelihood used for Cox proportional hazards models (Cox, 1972).

The conditional distribution function can be expressed as:

$$F_{Y|X=x}(y) = 1 - \exp[-\lambda_\beta(x) F_Z^{-1}(F_Y(y))].$$

Given λ_β, the random variables $Y \mid X = x$ are stochastically ordered, a property inherited from the latent exponential variables Z. For x_1, x_2 such that $\lambda_\beta(x_1) \geq \lambda_\beta(x_2)$

$$F_{Y|X=x_1}(y) \leq F_{Y|X=x_2}(y) \;\; \forall y \in \mathbb{R}.$$

If F_Y has a density with respect to Lebesgue measure, f_Y, then we can use a change of variables to calculate the conditional density as follows:

$$\begin{aligned} f_{Y|X=x}(y) &= f_Y(y)\frac{f_{Z|X=x}(z(y))}{f_Z(z(y))} \\ &= f_Y(y)\frac{f_{Z|X=x}(F_Z^{-1}(F_Y(y)))}{f_Z(F_Z^{-1}(F_Y(y)))} \\ &= f_Y(y)\frac{\lambda_\beta(x)\exp[-\lambda_\beta(x)F_Z^{-1}(F_Y(y))]}{\int_{\mathscr{X}}\lambda_\beta(x')\exp[-\lambda_\beta(x')F_Z^{-1}(F_Y(y))]dF_X(x')}. \end{aligned}$$

We end the construction of the model by assuming a prior over the finite-dimensional parameter β and a Bayesian nonparametric prior over the marginal distributions F_X and F_Y:

$$\beta \sim \pi_\beta \tag{3.8}$$

$$F_Y \sim \mathbb{P}_Y \tag{3.9}$$

$$F_X \sim \mathbb{P}_X \tag{3.10}$$

where π_β is some parametric prior, and $\mathbb{P}_X$ and $\mathbb{P}_Y$ may be a Dirichlet process mixture or a Pólya tree prior, as described in the introduction 1.2.1.

3.4 Approximations for posterior inference and prediction

Assume that both F_X and F_Y admit a density with respect to Lebesgue measure, noted f_X and f_Y. The unknown quantities for our regression model are therefore (f_Y, β, f_X). Given data

$(x_{1:n}, y_{1:n})$, where $x_{1:n} = (x_1, \ldots, x_n)$ and $y_{1:n} = (y_1, \ldots, y_n)$, we have the following likelihood:

$$L(f_Y, \beta, f_X; (x_{1:n}, y_{1:n})) = \prod_{i=1}^{n} f_Y(y_i) \frac{\lambda_\beta(x_i) \exp[-\lambda_\beta(x_i) F_Z^{-1}(F_Y(y_i))]}{\int_{\mathscr{X}} \lambda_\beta(x') \exp[-\lambda_\beta(x') F_Z^{-1}(F_Y(y_i))] dF_X(x')} f_X(x_i). \quad (3.11)$$

Inference could proceed using numerical methods such as MCMC but for large datasets this is cumbersome. Hence we consider here a Bayesian composite marginal likelihood approach (Cox & Reid, 2004; Lindsay, 1988; Pauli *et al.*, 2011; Ribatet *et al.*, 2012; Varin *et al.*, 2011) that we show offers computational tractability and the use of standard Bayesian methods. Define $y^*_{1:n}$ to be $y_{1:n}$ ordered from lowest to highest, and let $\nu_{1:n} = (\nu_1, \ldots, \nu_n)$ be a vector representing the order of $y_{1:n}$, so that $y^*_i = y_{\nu_i}$. Then we can re-write our data $\{y_{1:n}, x_{1:n}\}$ equivalently as $\{y^*_{1:n}, \nu_{1:n}, x_{1:n}\}$. Now, let L_C denote the composite marginal likelihood based on $\{y^*_{1:n}\}$ and $\{\nu_{1:n}, x_{1:n}\}$; that is, the product of the likelihood terms associated with each of these observations:

$$\begin{aligned} L_C(f_Y, \beta, f_X; \{y_{1:n}, x_{1:n}\}) &= L(f_Y, \beta, f_X; \{y^*_{1:n}\}) \times L(f_Y, \beta, f_X; \{\nu_{1:n}, x_{1:n}\}) \\ &= n! \left[\prod_{i=1}^{n} f_Y(y_i)\right] \times \left[\prod_{i=1}^{n} \frac{\lambda_\beta(x_{\nu_i})}{\sum_{j \geq i} \lambda_\beta(x_{\nu_j})}\right] \times \left[\prod_{i=1}^{n} f_X(x_i)\right]. \quad (3.12) \end{aligned}$$

We can see that this composite likelihood approach factors the likelihood into separate terms involving f_Y, β and f_X, leading to the following pseudo posterior distribution:

$$\pi_C(f_Y, \beta, f_X \mid \{y_{1:n}, x_{1:n}\}) = \pi_C(f_Y \mid y^\star_{1:n}) \pi_C(f_X \mid x_{1:n}) \pi_C(\beta \mid \nu_{1:n}, x_{1:n}) \quad (3.13)$$

Inference over the parameters f_Y, β, f_X can thus be carried out independently under the composite likelihood approach. Standard software for Bayesian nonparametric univariate density estimation can be used for f_Y and f_X, and software for fitting Plackett-Luce/Cox proportional hazard can be used for fitting β to facilitate a Laplace approximation. Overall,

the advantages of the approximate composite likelihood approach include computational tractability and scalable inference using standard software, hence good numerical reproducibility, and high interpretability as the components in the composite likelihood have explicit form and meaning. This latter point aids in prior elicitation as it allows the analyst to separate out and represent their beliefs on the marginal distributions, which are simpler to specify than the full conditionals, and then consider the dependence given the marginals.

The Bayesian composite likelihood approach has attracted some attention over recent years (Pauli *et al.*, 2011; Ribatet *et al.*, 2012; Varin *et al.*, 2011). In particular, Ribatet *et al.* (2012) considered two adjustments to the marginal likelihood approach in order to retain some of the desirable properties of the usual likelihood. However, their adjustments apply to a specific form of composite likelihood, where it factorises as a product of composite likelihoods for each observation: $L_c^{total}(y \mid \theta) = \prod_{i=1}^{n} L_c(y_i \mid \theta)$, where $L_c(y_i \mid \theta)$ is the composite likelihood for observation i. Our composite likelihood approach does not fit in this framework, as we do not have this product form over the observations, and we cannot therefore apply the adjustments suggested by Ribatet *et al.* (2012).

3.4.1 Asymptotics for the marginal composite posteriors

Consider first the pseudo-posterior for f_Y:

$$\begin{aligned}\pi_C(f_Y \mid \{y_{1:n}, x_{1:n}\}) &\propto \pi(f_Y) L_C(f_Y; \{y_{1:n}, x_{1:n}\}) \\ &\propto \pi(f_Y) \prod_{i=1}^{n} f_Y(y_i).\end{aligned}$$

So, our pseudo-posterior is exactly the posterior based on the i.i.d sample $\{y_{1:n}\}$, where $y_{1:n} \sim F_Y$. This is the standard setting for posterior inference, so we can apply consistency results from Bayesian nonparametric inference for F_Y, see for example Ghosal & Van der Vaart (2013). The same is true for f_X. Now consider the log-linear form for λ: $\lambda(x) =$

$\exp(-\beta x)$. Then, we have the pseudo-posterior:

$$\begin{aligned}\pi_C(\beta \mid \{y_{1:n}, x_{1:n}\}) &\propto \pi(\beta) L_C(\beta; \{y_{1:n}, x_{1:n}\}) \\ &\propto \pi(\beta) \prod_{i=1}^{n} \frac{e^{\beta x_{v_i}}}{\sum_{j \geq i} e^{\beta x_{v_j}}}.\end{aligned}$$

This is exactly the posterior considered by Kim (2006) in a different setting where a Bernstein-Von Mises theorem is proven, which can be applied here.

3.4.2 Posterior predictive

We can use simulation methods such as MCMC to easily generate samples $\{F_Y^{(j)}, \beta^{(j)}\}_{j=1}^m$ from the pseudo-posterior (3.13); the predictive distribution can then be approximated by:

$$p(y' \mid x', \{y_{1:n}, x_{1:n}\}) \simeq \frac{1}{m} \sum_{j=1}^{m} p(y' \mid x', \beta^{(j)}, F_Y^{(j)}).$$

To simulate from this distribution, we can use the forward generating process of our model, given $X = x'$:

$$Z' \mid X' = x' \sim \text{Exp}\left(\lambda_\beta(x')\right) \tag{3.14}$$

$$Y' = F_Y^{-1}(F_Z(Z')). \tag{3.15}$$

In many applications, modeling F_X might be cumbersome, and not the primary object of interest. In this case we propose to use an empirical Bayes approach by setting $F_X = \hat{F}_X$ at the empirical CDF. So, to generate a posterior predictive sample, given a posterior sample $\{F_Y^{(j)}, \beta^{(j)}\}_{j=1}^m$, Equation 3.15 becomes:

$$Y'^{(j)} = F_Y^{-1(j)} \left(1 - \frac{1}{n} \sum_{i=1}^{n} e^{-Z'^{(j)} \lambda_{\beta^{(j)}}(x_i)}\right)$$

where we note that $Z'^{(j)}$ is conditional on $X' = x'$, and the CDF inversion is tractable, depending on the form of F_Y. Some particular examples are discussed in the appendix. The Pólya tree serves as a convenient prior for F_Y, since this CDF inversion can be simulated directly, without sampling the entirety of $F_Y^{(j)}$.

3.4.3 Connection to proportional hazards

This section shows how our model can be viewed as a re-parameterisation of the proportional hazards model that is marginally coherent, and lends to a natural extension of the partial likelihood to include all model parameters. The survivor function implied for our model is given by:

$$\begin{aligned}\mathbb{P}(Y > y \mid X = x) &= \mathbb{P}(F_Y^{-1}(F_Z(Z_i)) > y \mid X = x) \\ &= \mathbb{P}(Z_i > F_Z^{-1}(F_Y(y)) \mid X = x) \\ &= \exp\{-\lambda_\beta(x)\, F_Z^{-1}(F_Y(y))\}\end{aligned}$$

Therefore, the cumulative hazard is given by:

$$H_x(y) = \lambda_\beta(x) F_Z^{-1}(F_Y(y))$$

implying proportional hazards, and a correspondence between the baseline hazard and F_Y. We argue that parameterisation in terms of F_Y provides a natural vehicle for prior information, in contrast to the baseline cumulative hazard function. Furthermore, our construction facilitates the incorporation of marginal Y data, which would be difficult within a standard proportional hazards model. In addition, let's explore the nature of the use of the partial likelihood (Cox, 1972), when viewed as an approximate model assumption. The use of the probability of the ordering of Y as the likelihood for β equates to the conditional independence assumption: $Y \perp\!\!\!\perp X \mid \nu$, where β parameterises the relationship between X and Y. This is exactly the

conditional independence used in the composite likelihood (Equation 3.12); in this sense, our composite likelihood uses the same approximation as the traditional partial likelihood. By facilitating a natural factorisation of the likelihood into the partial likelihood and a term for the marginal response, parameterisation in terms of F_Y gives a simple approximate likelihood for all parameters in the model, rather than just β. When parameterised in terms of H_0 or h_0, there is no such approximate likelihood for all parameters, so inference must use the full likelihood, which can be cumbersome for large datasets. Another effect of constructing the model via the marginal distribution for the response, is that any variable selection priors will be marginally coherent, in the sense that the marginal distribution of Y is the same for each subset of variables; a stronger coherence than defined for means and variances in Chapter 2. This is because the construction guarantees that F_Y is independent of β.

In the next section we discuss the incorporation of truncated and censored data. We shall see that treatment of the covariates as random provides an additional term in the likelihood that depends on β compared to an analysis in which covariates are treated as fixed. The fact that F_Y is independent of β in our model allows simple calculation of this term, enabling the incorporation of this additional information. The downside to our approach is the requirement to model X; however, as we have seen, this only need enter at the stage of prediction.

Censored and truncated data

Censored and truncated data are common in survival analysis. Truncated data is observed subject to a condition, such as delayed entry into a study. If an individual enters the study at time s, and has an event at time y, the contribution to the likelihood is given by $\frac{f_Y(y)}{1-F_Y(s)}$, the density of Y conditional that $Y \geq s$. Censored data constitute inexact observations, for example, that the time is within an interval, rather than a specific observed event time. Commonly, this takes the form of right censoring, the observation that $Y > c$, where c is the time that the individual left the study. The contribution to the likelihood in this example is

$1 - F_Y(c)$, the probability that Y exceeds this censoring value. In a regression setting, since covariates x are typically treated as fixed, the likelihood corresponding to a left truncated datum is given by:

$$\begin{aligned} L(\theta; y \mid x) &= f_{Y|X}(y \mid \{Y > s\}, x; \theta) \\ &= \frac{f_{Y|X}(y \mid x; \theta)}{1 - F_{Y|X}(s \mid x; \theta)} \end{aligned} \tag{3.16}$$

where θ parameterises the conditional density $f_{Y|X}(y \mid x)$. The likelihood for other forms of truncation can be calculated in a similar way. However, intuitively, in the delayed-entry example, there may be information in the distribution of the covariates. Covariates of left truncated data are not sampled from the population covariate distribution, rather, they are subject to a survival bias, which depends on θ. Concretely,

$$\begin{aligned} L(\theta; y, x) &= f_{Y,X}(y, x \mid \{Y > s\}; \theta) \\ &= \frac{f_{Y|X}(y, x; \theta)}{1 - F_Y(s; \theta)} \\ &= \frac{f_{Y|X}(y \mid x; \theta) f_X(x)}{1 - F_Y(s; \theta)}. \end{aligned} \tag{3.17}$$

The ratio of these two likelihoods is simply $f_X(x \mid Y > s)$, by Bayes' rule. This matches our intuition: treating covariates as fixed ignores the survival bias in the covariate sample. The likelihood 3.17 has been previously explored by Bergeron *et al.* (2012), who employ a joint modeling approach to calculate the maximum likelihood estimator with covariates treated as random. However, this complete likelihood has seen limited use as a result of the requirement to model X. However, using our model, the $F_Y(s; \theta)$ term depends only on F_Y, not on β. As a result, the likelihood for a sample with left truncation times $\{s_i\}_{i=1}^n$ is simply given by:

$$L(f_Y, \beta, f_X; \{x_i, y_i, s_i\}_{i=1}^n) = \frac{L(f_Y, \beta, f_X; \{x_i, y_i\}_{i=1}^n)}{\prod_{i=1}^n 1 - F_Y(s_i)}$$

Therefore, we can again apply the composite likelihood approximation, resulting in the approximate likelihood:

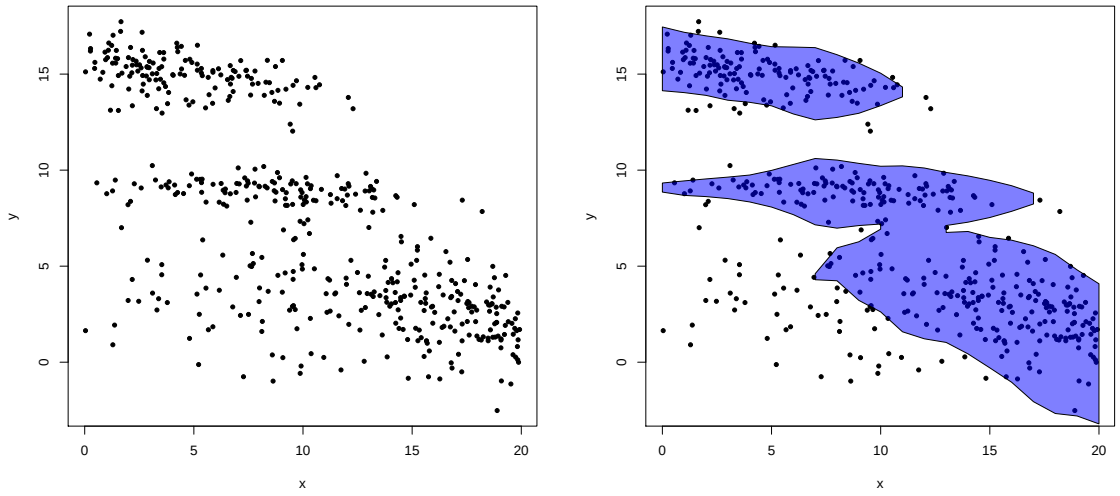
$$L_C(f_Y,\beta,f_X;\{y_{1:n},x_{1:n},s_{1:n}\}) = \frac{L_C(f_Y,\beta,f_X;\{y_{1:n},x_{1:n},s_{1:n}\})}{\prod_{i=1}^{n} 1-F_Y(s_i)}$$
$$= n!\left[\prod_{i=1}^{n}\frac{f_Y(y_i)}{1-F_Y(s_i)}\right]\times\left[\prod_{i=1}^{n}\frac{\lambda_\beta(x_{\nu_i})}{\sum_{j\geq i}\lambda_\beta(x_{\nu_j})}\right]\times\left[\prod_{i=1}^{n}f_X(x_i)\right].$$

That is, F_Y is updated on the basis of a truncated sample, but β is updated based on the non-truncated observations. Therefore, if interest lies in β, inference can proceed for β without specification of a prior for F_Y or F_X using this full partial likelihood. This is in contrast to the standard treatment of truncated data (for example, Cox & Oakes, 1984; Klein & Moeschberger, 2005), where the partial likelihood is adjusted for truncation. Quite simply, this covariate bias term cancels out the standard truncation adjustment, leaving us with the standard non-truncated likelihood. The use of our model allows more information to be incorporated into the likelihood, through the term for the covariate bias, without need to specify F_X or F_Y.

Censoring is incorporated in a way in which is consistent with the standard treatment of censored data (e.g. Klein & Moeschberger, 2005). Censored observations contribute to a partial, rather than full ordering ν, and F_Y is updated on the basis of censored observations $\{Y > c\}$, contributing $1-F_Y(c)$ to the likelihood. The composite likelihood for β in this case matches the partial likelihood in the proportional hazard model with censoring.

3.5 Illustrations

In this section we apply our method to two examples. The first is a simulation example where we generate from a multi-modal conditional and explore the ability of our method to fit the



(a) (b)

Fig. 3.2 (a) Data simulated from the model with mixture of three Gaussians marginal distribution for Y. (b) 80% highest posterior density intervals of the predictive distribution at each value of x.

data. The second is a large real-world application in the regression analysis of US Census data.

3.5.1 Simulation example

In this section we apply the model to a dataset simulated from our model to consider how well we can recover known dependence. The marginal distribution of Y is set to a mixture of three Gaussian distributions, with means 3, 9 and 15, standard deviations of 2, 0.5 and 1 with mixture weights of 0.5, 0.2 and 0.3 respectively. β is set to 0.25, with $\lambda_\beta(x) = \exp(\beta x)$. $X \sim \text{Unif}(0,20)$ and $n = 500$. The data is shown in Figure 3.2(a).

Clearly any type of linear or non-linear regression with a parametric noise distribution will be inappropriate here. The conditional distribution of Y given x is multi-modal, rendering many popular regression models inappropriate.

We compared our approach to a linear dependent Dirichlet process mixture of normals (LDDPM) (De Iorio *et al.*, 2004), using the R package DPpackage (Jara, 2007; Jara *et al.*, 2011). This model specifies that:

$$Y_i \mid x_i \sim \int \mathrm{N}(y_i; x_i\beta, \sigma^2) G(d\beta, d\sigma^2)$$
$$G \mid \alpha, \mu_b, s_b \sim \mathrm{DP}(\alpha, G_0)$$

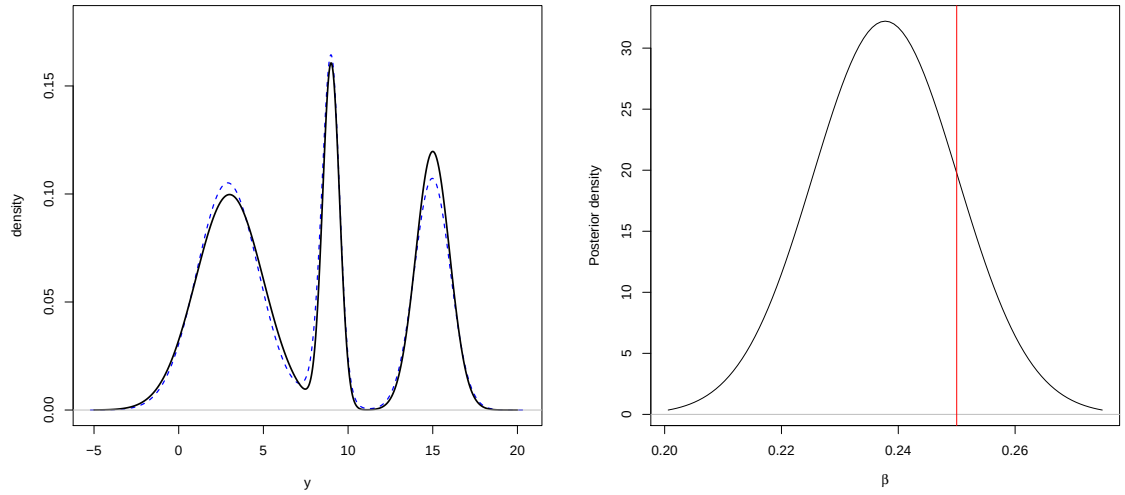
where $G_0 = \mathrm{N}(\mu_b, s_b)\Gamma(\tau_1/2, \tau_2/2)$ and

$$s_b \mid \nu, \psi \sim IW(\nu, \psi)$$

with $\alpha = 1, \mu_b = (9,0)^T, \nu = 4, \tau_1 = 1, \tau_2 = 2, \psi = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and $s_b = \left(\begin{smallmatrix} 36 & 0 \\ 0 & 36 \end{smallmatrix}\right)$.

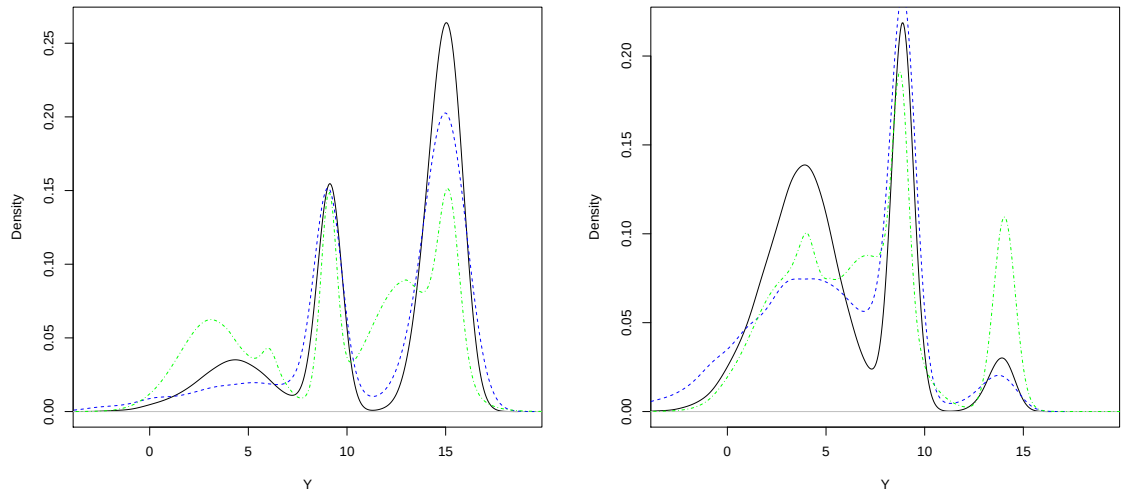
We apply our model, modeling the marginal as a Dirichlet Process mixture of Gaussian distributions using $\alpha = 1$ and a normal-inverted-Wishart distribution for the base measure. That is, our base measure $G_0(\mu, \sigma^2) = \mathrm{N}(\mu \mid \mu_1, \frac{\sigma^2}{\kappa_1}) IW(\sigma^2 \mid \nu_1, \psi_1)$, where $\mu_1 = 9, \kappa_1 = 0.5, \nu_1 = 4$ and $\psi_1 = 1$. A Gaussian prior centered at 0 with unit variance is used for β.

In Figure 3.2(b) the simulated data is shown, with the 80% highest posterior density (HPD) intervals of the predictive distribution at each value of x. Qualitatively, we see that the model can capture the nonlinearities in the data and demonstrates the flexibility to model the multi-modal conditional response. In Figure 3.3 we show the predictive marginal, $\hat{F}_Y$ and the posterior distribution for β. Clearly the marginal distribution for Y is very well recovered from the data. This parameterisation of the model in terms of the marginal distribution for the response allows this to be estimated from the complete dataset, without reliance on other aspects of the model. The strength of information available is apparent in the quality of the fit to the sampling distribution. The posterior for the parameter β shows reasonable support around the true value, being slightly pulled towards 0 by the prior.



(a) (b)

Fig. 3.3 (a) The posterior predictive marginal for y under our model in blue, compared to the actual sampling distribution in black. (b) The posterior distribution for β, compared to the true value of 0.25 marked in red.



(a) (b)

Fig. 3.4 Predictive densities for (a) $x = 5$ and (b) $x = 12$. The true predictive is shown in black, the predictive distributions under our model in blue, and the predictive under the linear DDP mixture in green.

We can further inspect how these come together in the posterior predictive conditional distribution for Y given x. Consider this distribution for $x = 5$ and $x = 12$, for both our model and the linear DDP mixture model, as shown in Figure 3.4. Again, our model provides a reasonable fit. The predictive distribution is not as accurate as the marginal distribution for Y, but this is to be expected, since the conditional distribution is a product of the whole model, compounding uncertainties from both β and the marginal distribution for y. Nonetheless, the fit is good and noticeably better than the flexible linear DDP mixture, as you would expect, given that the sampling distribution is within the support of our model. Concretely, the L1-distance between the estimated conditional and the true conditional distribution can be calculated in each case. When $x = 5$ the distance to our prediction is 0.00869, whereas the distance to the linear DDP is 0.0214, and when $x = 12$ the distance to our prediction is 0.0127 and the distance to the linear DDP is 0.0146.

A point of note is that these posterior predictive plots are smoothed kernel density estimates of MCMC samples. Therefore, Gaussian shapes are slightly exaggerated. Whilst not entirely clear from the plot, both our predictive and the sampling distribution comprise of slightly skewed Gaussian distributions, since the conditional distribution is the marginal distribution for Y weighted across the quantiles.

To illustrate that the model is capable of modeling a range of distributions, we consider data sampled from a Gaussian linear model. The covariates are simulated uniformly on $[0, 10]$, with $Y \sim N(3 + 2x, 2)$ and $n = 300$. We use our model, modeling the marginal for Y with a Pólya tree prior whose partition is set on a Gaussian distribution with mean 12.5 and standard deviation 6, and $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = m^2$. A Gaussian prior centered at 0 with variance 8 is used for for β. The posterior predictive 80% HPD intervals display a reasonable fit of the linear data, shown in Figure 3.5. The variance seems slightly inflated, but this is a consequence of the large support of the model.

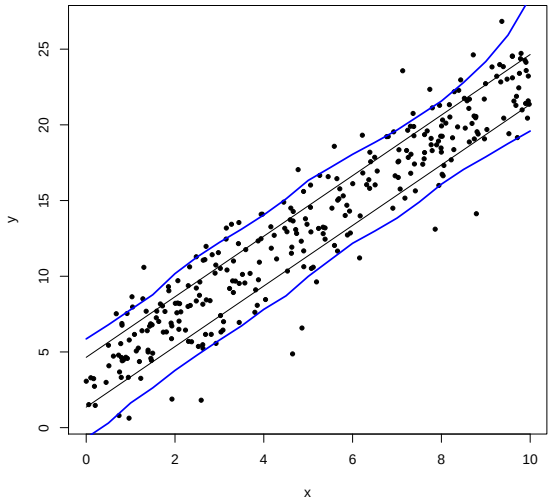


Fig. 3.5 Simulated data from a linear model with Gaussian residuals (black dots). 80% HDP intervals of the predictive distribution of our model at each value of x are represented in blue, and true HDP interval in black

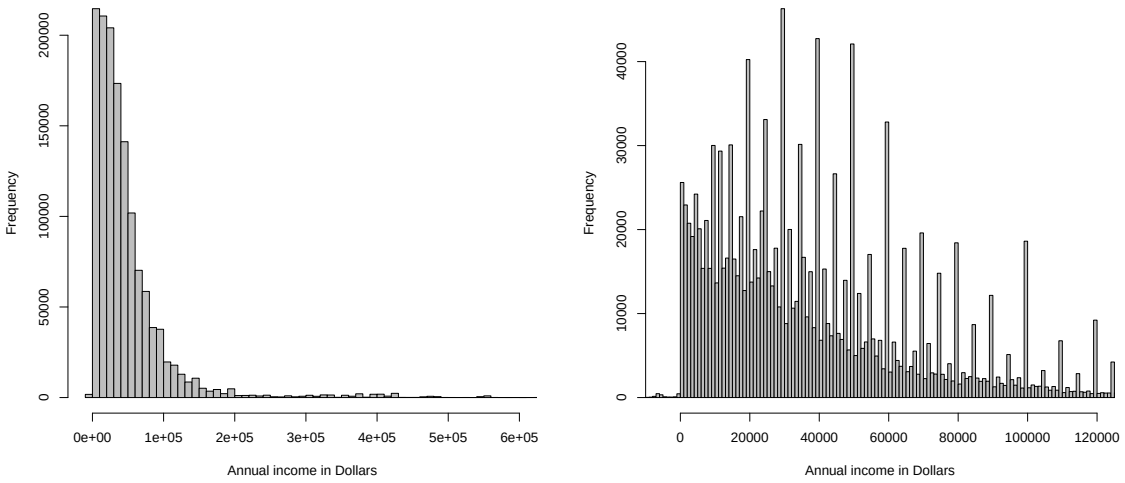
3.5.2 US Census application

We apply the methodology to a regression task using US census data[2] for personal annual income.

We use the American Community Survey data from 2013, which comprises of responses to questions on the survey given to a 1% sample of the US population. Since we are interested in income, the subset of $1,371,401$ employed civilians over the age of 16 is used. We have used a relevant, linearly independent subset of the data as covariates, excluding highly informative questions such as occupation, which would almost completely explain the response. This leaves 15 explanatory variables, 10 of which are categorical variables, some of which have many levels. The result is a $1,371,401 \times 114$ design matrix.

The covariates are: US state (Texas as a baseline), age, class of worker (employee of private for-profit company as a baseline), travel time to work, means of transportation to work (works from home as a baseline), language other than English spoken at home (none as

[2]http://www.census.gov/acs/www/data_documentation/pums_data/



(a) (b)

Fig. 3.6 Histograms of annual income on different scales. Right hand plot is zoomed in on incomes up to 120,000.

a baseline), marital status (married as a baseline), educational attainment (regular high school diploma as a baseline), gender (male as a baseline), hours worked a week, weeks worked last year, disability status (without a disability as a baseline), quarter of birth (first quarter as a baseline), and world area of birth (United States of America as a baseline).

The levels of annual income shown in Figure 3.6 can be seen to be heavy tailed, which requires a flexible model to capture. Another noticeable feature of the data is that the income levels are discontinuous, with large spikes in frequency at particular income levels. This could in part be due to standardised salary structures resulting in certain salary levels becoming common. This motivates the use of a nonparametric approach as it is difficult to imagine how a parametric density could conditionally capture the features shown in Figure (3.6). However, standard Bayesian nonparametric models simply cannot be applied to a problem of this scale. Attempting to apply existing methods in this literature, such as the linear dependent Dirichlet process mixture, failed to run due to the dimensionality and scale of the data.

For the analysis, we consider both the empirical distribution function and a Pólya tree prior for the marginal distribution of y. The partition of the Pólya tree is set on the quantiles of a Gaussian distribution with mean 35,000 and standard deviation 20,000, and $\alpha_{\varepsilon_1 \ldots \varepsilon_m} = m^2$. We use a log-linear regression function $\lambda(x) = \exp(\beta x)$ and place independent Gaussian priors with mean 0 and unit variance on the coefficients in β.

Predictive performance

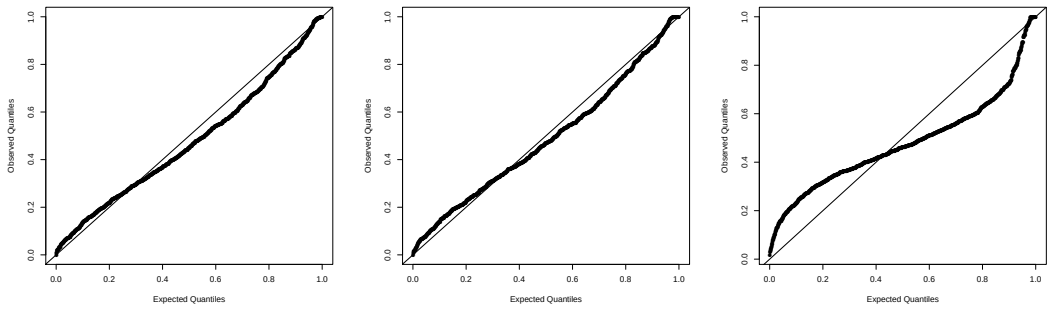
We compare the out-of-sample predictive performance of our model with four competing non-Bayesian approaches namely, a standard linear regression model, a median regression model ,a LASSO and random forests[3]. For our model we investigated three distinct priors for the marginal distribution of the response: a Pólya tree centred on a Gaussian, a Pólya tree centred on a Laplace, and an empirical Bayes approach using the empirical CDF. To compare methods we use repeated random subsets of 1000 test samples and train each model on the remaining data, with 10 repeats. Predictive accuracy is judged by mean squared-error (MSE), mean absolute error (MAE) and qualitatively via a qq-plot. To create the qq-plots we compute the predictive distribution function $F(y \mid x)$ evaluated at the observed value for each of these test samples. Under the assumption that we have a perfect predictive distribution, these values should be independent uniform random variables. A deviation from this distribution implies a mis-match of the posterior predictive and the actual distribution. We are unable to apply this approach to the median regression model, as it does not provide a predictive distribution and would require fitting the model for a large number of quantiles. In the case of the linear model we used maximum likelihood estimates for prediction, rather than a fully Bayesian approach. With such a large dataset the strength of any reasonable default prior would be significantly diminished, so this should mimic a Bayesian approach well.

[3]These models were fitted in R using the functions lm, rq (from the quantreg package), lars (from the lars package) and randomForest (from the randomForest package). For the random forest only 50 trees were used to ease the computational burden. For LASSO the regularisation parameter was chosen using cv.lars. Default settings were used for each.

Table 3.1 Mean out-of-sample prediction errors with standard deviation of this error after 10 repetitions

	Mean square error (10^9)	Mean absolute error(10^4)
Empirical model*	2.79 ± 0.51	2.44 ± 0.15
Pólya tree (Gaussian)*	2.81 ± 0.64	2.41 ± 0.16
Pólya tree (Laplace)*	2.71 ± 0.52	2.44 ± 0.14
Linear model	2.66 ± 0.59	2.67 ± 0.16
LASSO	2.99 ± 0.65	2.81 ± 0.16
Median regression	2.99 ± 0.68	2.48 ± 0.17
Random Forests	2.44 ± 0.52	2.36 ± 0.15

Summary statistics of predictive fit are shown in Table 3.1. Perhaps unsurprisingly on such a large data set the linear model targeting the conditional mean does well of the on MSE but this is at the expense of the median under MAE. In addition, studying the predictive qq-plot in Figure 3.7b shows the inadequacy of the linear model to provide calibrated predictions. The LASSO performs relatively poorly, suggesting that most covariates are influential for prediction. The median regression, as expected, provides relative accuracy on the MAE, but does so at the expense of MSE; in addition, as mentioned above, suffers from the lack of a fully predictive model. The Bayesian nonparametric methods perform relatively well on both summary measures, with perhaps that based on the Laplace marginal showing greatest accuracy. Random Forests is the best performer for both measures of estimating the mean. This is to be expected, as the method provides very flexible model for the mean. However, the random forest does not provide a flexible conditional distribution for the response, which is Gaussian. In Figure 3.7a we show the predictive qq-plot from this model, demonstrating that the full predictive distribution is captured well. These diagnostics suggest that even non-linear regression models with parametric noise would not provide a satisfactory fit for the data, since the unusual conditional distribution of the response cannot be captured by such models. Even the flexible random forest methodology assumes Gaussian noise, so also falls short in this respect. This highlights the benefit of our nonparametric approach, in



(a) (b) (c)

Fig. 3.7 qq-plots (a) under our model using the empirical distribution for Y, (b) our model with a Laplace-centered Pólya tree prior for F_Y, and (c) using a standard linear model.

the ability to capture wide classes of distributions. This ability to capture the conditional distribution should lead to more accurate prediction.

We next consider inference for covariate effects. In order to gain a measure of the relevance of each covariate we quantified the concentration of the posterior probability measure away from the prior "null" centring of $\beta_j = 0$. To do this we estimated the Bayesian sign-probability from the posterior marginal for each covariate as,

$$PrSign_j = \max\left[\int_{\beta<0} \pi(\beta_j \mid \cdot)d\beta, \int_{\beta>0} \pi(\beta_j \mid \cdot)d\beta\right] \tag{3.18}$$

where $\pi(\beta_j \mid \cdot)$ is the posterior marginal for β_j. This measures the relative tail area in the posterior marginal laying to the left or right of 0. A large value of *PrSign* suggests there is strong evidence against $\beta_j = 0$. In certain respects this is akin to a Bayesian marginal version of a p-value, and is trivially calculated from MCMC output, or from normal approximations to the posterior distribution. Table 3.2 shows the most relevant covariates as ranked by this measure.

Unsurprisingly, hours worked a week and weeks worked last year show high certainty of a positive effect on income. After these, educational achievement measured via degrees unsurprisingly imply higher earnings compared to the regular high school diploma. Since

Table 3.2 Top covariate parameters ranked by (3.18): the log posterior probability of the parameter being a different sign to the posterior mean. A negative parameter value has a positive effect on income.

	Log probability of different sign	Posterior mean
Hours worked a week	-1.1×10^5	-0.044
Weeks worked last year	-1.0×10^5	-0.045
Bachelor's degree	-4.2×10^4	-0.80
Master's degree	-4.16×10^4	-1.0
Professional degree	-2.7×10^4	-1.4
Age	-2.6×10^4	-0.018
Female	-1.8×10^4	0.35
Doctorate degree	-1.5×10^4	-1.2
Never Married	-1.3×10^4	0.37
Associate's degree	-6.5×10^3	-0.39
Travel time to work	-3.2×10^3	-0.0035
1 or more years of college credit	-2.0×10^3	-0.18
Self employed (incorporated)	-2.0×10^3	-0.30
Grade 11 in school	-1.8×10^3	0.38
Walks to work	-1.5×10^3	0.36
Disabled	-1.3×10^3	0.19

these are part of the same variable, it is simple to compare the effects due to these degrees. Despite Bachelor's degree providing the most certainty of a positive effect, a further professional degree beyond a Bachelor's degree has the highest posterior mean effect. The ranking in Table 3.2 reflects the greater evidence in the data for a non-zero Bachelor effect, due to a much higher number of observations of those with Bachelor's degrees, and hence lower variance in the effect size compared with those with a higher degree. There is also strong evidence for female workers earning less than their male counterparts, as well as increasing income with age and even travel time to work.

Finally, we show it is simple to provide the full posterior predictive distribution of annual income of somebody in the test sample, using the Pólya tree model. We choose, as a hypothetical person, a 57 year old female from North Carolina, who is self employed, married, who works from home, speaks English at home, went to college but for less than

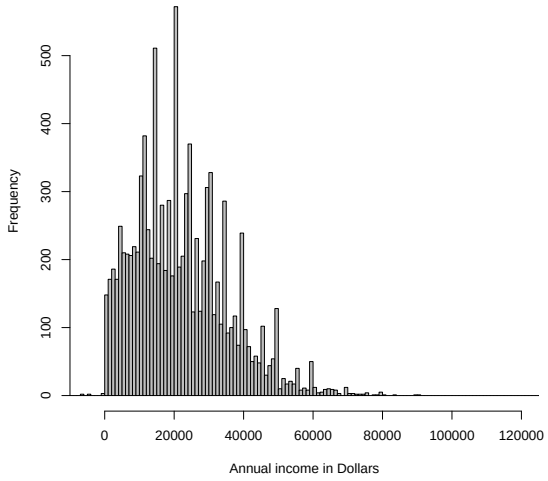


Fig. 3.8 Posterior predictive distribution using a Pólya tree for the marginal distribution, zoomed in on incomes up to 120,000

a year, who usually works 30 hours a week for 43.5 weeks last year, and was born in the first quarter of the year in the USA. The structure and shape of the posterior predictive, represented in Figure 3.8, match that of the marginal distribution for Y in the data, just on a narrower range.

3.6 Discussion

We introduced a Bayesian semiparametric regression model that is designed to scale to large data applications. In doing so we make use of an interpretable model for ranks, via a Plackett-Luce copula method, and nonparametric density models for the marginals. We used a composite marginal likelihood approximation that leads to a number of advantages. It affords computationally tractability, aids in the interpretation of the model, and makes prior specification explicit on known objects.

The key to the scalability of the method is the use of the composite likelihood approximation, which splits the inference into two simpler tasks. The use of the Laplace approximation

for the covariate effect and the Pólya tree for the marginal response allow for fast posterior inference, without requiring any MCMC sampling methods. In fact, sampling methods are only used for prediction, which is by far the slowest part of the inference procedure.

Going forward, it would be interesting to see if theoretical bounds on the approximation error as a function of sample size could be derived. It may also be possible to apply results such as those found in Kim (2006) to provide further guarantees of asymptotic behaviour such as properties of the predictive distribution. In addition, it would be interesting to explore non-linear models for the regression function $\lambda(x)$, such as those based on a random forests methodology. In fact, random forests applied to the US Census dataset (with restricted node size to enable application to this scale) gives a highly competitive MSE to our tested models. This is likely because random forests is able to capture interaction terms between covariates, which seem highly plausible *a priori* in this particular dataset. It should be interesting to incorporate such flexibility into a Bayesian nonparametric approach using Plackett-Luce regression functions.

Chapter 4

Model-free testing for subgroup analysis

4.1 Introduction

This chapter considers the analysis of two-arm randomised clinical trial data, with the aim of identifying the optimal allocation of treatment based on covariates X. In the medical literature, this is typically posed as a regression problem, where heterogeneity of treatment effect is identified via testing interaction terms in a regression model (for example, Assmann *et al.*, 2000; Brookes *et al.*, 2004). In order to safeguard against false-positive results associated with multiple testing, it has been recommended that every proposed subgroup should correspond to a pre-specified, clinically motivated hypothesis. However, in the statistical literature, many exploratory methods for the *post hoc* analysis of RCT data have been proposed, motivated by potentially large gains in efficiency of treatment. For example, in a study where genetic data is available, there are unlikely to be a small number of clinically-derived potential subgroups. In such an application, careful *post hoc* analysis could serve as an important hypothesis generation tool. Typically, such exploratory approaches build a conditional model for the response $Y \mid X = x, T = t$, given covariates and treatment status, and use this to determine optimal treatment allocation (Foster *et al.*, 2011; Qian & Murphy, 2011). More recently, Zhao *et al.* (2012) and later Zhou *et al.* (2015), proposed a direct approach of

optimizing the expected response under a treatment rule D, a function that maps covariates to treatment allocation. By directly considering $\mathbb{E}[Y \mid T = D(X)]$, it is possible to estimate optimal treatment allocation without specifying a conditional model $Y \mid X = x, T = t$. The appeal of this is to safeguard against model misspecification by making minimal model assumptions. This robustness is especially important in the context of subgroup analysis, as model misspecification can lead to spurious results, which are a major concern in an exploratory analysis. There are additionally a number of practical benefits to a model-free approach, such as ease of implementation and reproducibility. However, the "Outcome weighted learning" (OWL) approach has calibration issues, namely that a uniform shift of the response will change the eventual allocation of treatment of the algorithm. This produces a bias towards an allocation of treatment matching the observed treatment allocation when the response is positive. This issue is addressed by "residual weighted learning" (RWL) (Zhou *et al.*, 2015), where OWL is applied to the residuals of a linear regression model for $Y \mid X$. As a result, the approach is no longer model-free. Following OWL, and the work in the previous chapters, we believe there are substantial benefits to be made by framing the problem in terms of $Y \mid T = D(X)$, the marginal distribution of the response under a treatment rule D, rather than the full conditional distribution of the response $Y \mid X = x, T = t$. Framed directly in terms of the treatment rule D, subgroup analysis is naturally stated as a testing problem. Given a proposed treatment rule D, RCT data can be viewed in terms of two samples: one sample of individuals treated in accordance with D and one sample of individuals treated in accordance with the "mirror rule" D^c, that assigns opposite treatment. By symmetry, the size of these two samples are equal, in distribution, if the RCT is balanced. Therefore, it is possible to test for a difference between the distribution of the response under D and under D^c using a simple two-sample test. We show that this is equivalent to testing D against treatment at random, a natural null hypothesis. This allows us to fully circumvent the modeling stage and directly test the efficacy of treatment rules using nonparametric tests such

as the Mann-Whitney-U test for ranks. In addition to testing proposed treatment rules, this nonparametric approach can be extended to exploratory subgroup analysis. The test statistic of the nonparametric tests for treatment rules can be used as as a splitting criteria in a random forest of "treatment trees", which provides a flexible means for the exploratory proposal and evaluation of treatment rules. We introduce the notion of "selective logging" - the forest-level analogue of pruning - to combat issues associated with combining decision rules with covariate subsampling. Permutation methods can be used to correcting for associated multiple testing, resulting in a complete, robust, easy-to-use, model-free methodology for exploratory subgroup analysis. We propose that any subgroups found this way should provide motivation for future trials, rather than used as evidence to guide future treatment. Section 4.5 explores the application of an extension of the Bayesian nonparametric model introduced in Chapter 3 to subgroup analysis, providing a Bayesian analogue of the frequentist approach. Both approaches are applied to AIDS clinical trial data, leading to similar findings regarding the efficacy of treatment in subgroups.

As in the previous chapters, treatment of covariates as random variables allows us to define marginal distributions where covariates are integrated out. In previous chapters, this provides a means to define the marginal distribution of the response, which can be used to elicit prior information into a regression model. In the subgroup analysis setting, the treatment of X as random allows direct consideration of $Y \mid T = D(X)$, the distribution of the response under a treatment rule. We argue that this is the quantity of interest, and that conditional models simply provide a proxy tool for the analysis of this distribution.

4.2 Subgroup Analysis

A two-arm randomised controlled clinical trial (RCT) is an experiment designed to investigate the efficacy of a clinical intervention. In our canonical two-arm example, patients are randomly split into two groups, where one group is given experimental treatment and the

other "control" group is given standard care. Randomisation addresses bias issues associated with the potential dependence between treatment and covariates in observational studies; this allows causal relationships to be inferred. As such, RCTs are now standard, and their use is widespread in clinical and pharmaceutical research. (See Matthews, 2006, for a thorough introduction to RCTs). Increasingly, interest lies in the identification of a "subgroup" of the population that will benefit from treatment. Of phase III oncology RCTs conducted from 2011 to 2013, 85% were reported with a subgroup analysis (Zhang *et al.*, 2015). The ultimate aim of such an analysis is to provide personalised treatment to individuals based on covariates. This is already the case in the treatment of some cancers, where therapy type is determined by gene expression (Van't Veer & Bernards, 2008). However, subgroup analysis poses many problems, including a lack of a unified framework in applications, and issues with multiple testing. We hope to address these issues in part by posing the problem as one of the hypothesis testing of treatment rules. This allows us to use model-free methods that can be widely-used.

We consider RCT data of the form $\{Y_i, X_i, T_i\}_{i=1}^n$. $Y_i \in \mathbb{R}$ denotes the response of individual i, the primary outcome of interest. $T_i \in \{0,1\}$ is an indicator of treatment, taking the value 1 if individual i is in the experimental arm, and the value 0 if individual i is in the control group. $X_i \in \mathbb{R}^p$ is a vector of covariates, which may be predictive of Y. A standard analysis of RCT data is concerned with the effect of treatment T on the response Y, possibly incorporating X to increase power. A subgroup analysis is ultimately concerned with the identification of a treatment rule D that assigns a treatment status, given a covariate value $D : \mathbb{R}^p \to \{0,1\}$. Additionally, in the interest of scientific reporting, we believe that it is important to be able to associate a level of evidence in favour of any proposed treatment rule, such as a p-value. However, if the same data are used to determine the optimal treatment rule as those used to calculate each p-value or Bayes factor, this introduces a clear bias, rendering these uninterpretable. Therefore, multiple testing issues must be given careful consideration.

4.2.1 Testing for treatment heterogeneity

In the medical literature, the problem is framed in terms of the detection of heterogeneity of treatment effect. This is often done via interaction tests in regression models (Assmann *et al.*, 2000; Brookes *et al.*, 2004; Rothwell, 2005; Sun *et al.*, 2012; Zhang *et al.*, 2015). Consider a "subgroup" A, defined as a subset of the covariate space $\mathbb{R}^p$. To test for a difference in treatment effect for $X \in A$ and $X \in A^c$, one specifies a conditional model for the response $Y \mid X, T$ given covariates and treatment, for example:

$$Y = \alpha + X\beta + T\gamma + T\mathbb{I}\{X \in A\}\theta + \varepsilon \tag{4.1}$$

where $\mathbb{I}$ is the indicator function. Testing for a difference of effect for $X \in A$ and $X \in A^c$ amounts to testing the hypothesis $H_0 : \theta = 0$. Extension to logistic regression, generalized linear regression and Cox proportional hazards is simple. The p-value associated with this test indicates a level of evidence in favour of this heterogeneity. Identification of treatment heterogeneity, together with the estimated global treatment effect can be used to inform an optimal allocation of treatment. It is important to note that treatment heterogeneity is a necessary, but not sufficient condition for heterogeneity in optimal treatment allocation. This important point seems often ignored - p-values associated with interaction terms measure certainty of treatment heterogeneity, rather than relative efficacy of treatment. It is important to distinguish between "crossover" interactions (Gail & Simon, 1985) - where in this context, interaction results in a difference in optimal treatment - and "non-crossover" interactions, where despite treatment heterogeneity, one treatment is universally optimal. An appropriate test for optimality of treatment in the subgroup A should take the form $H_0 : \theta + \gamma = 0$, which tests the net benefit of treatment for $X \in A$. Figure 4.1 illustrates this with point with toy examples. In the case that the first case, the interaction term results in a change in optimal

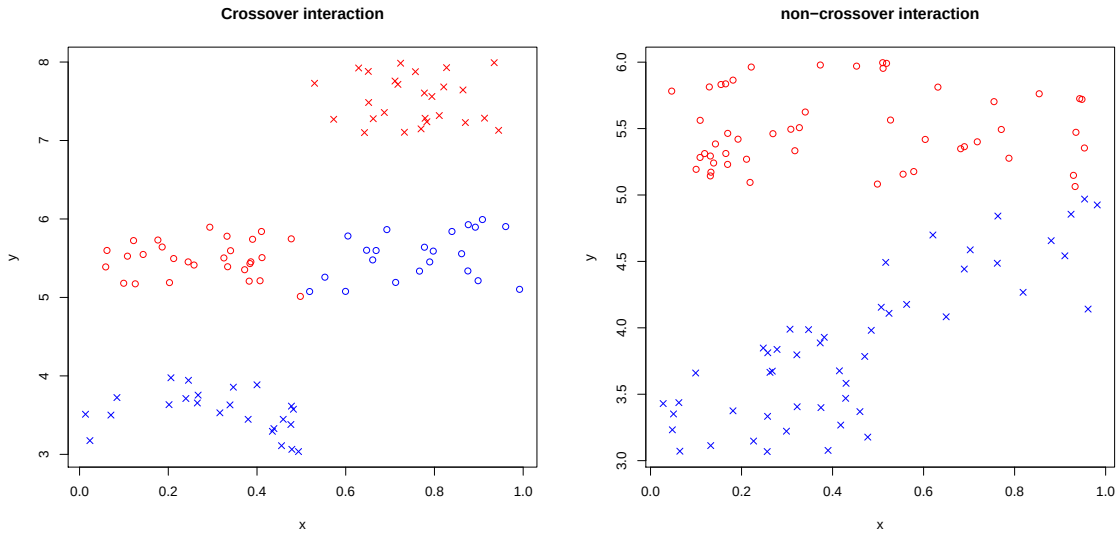


Fig. 4.1 Plots of response y given covariate x for experimental treatment (crosses) and standard treatment (circles) for two toy examples. Left, the interaction between treatment and x results in treating only those with $x > 0$ as the optimal course of treatment (red). Right, there is an interaction between treatment and x, but placebo for all patients is nonetheless the optimal treatment rule (red).

treatment, whereas in the second case, the negative global treatment effect is strong enough to negate the interaction effect.

Subgroup analysis has become a divisive topic on the basis of issues of multiple-testing. Rothwell (2005) recommends skepticism when interpreting the results of *post hoc* analysis of clinical trial data, instead suggesting a conservative approach. He proposed guidelines for subgroup analysis, interpretation and reporting. Importantly, Rothwell proposes that a small number of subgroups should be proposed in advance of the trial, on the basis of clinical input. Then, significance values calculated using interaction tests for each proposed subgroup should be reported, in order to eliminate publication bias. Following these proposals, Sun *et al.* (2012) conducted a systematic review of the analysis subgroups in RCTs in 2007, finding that very few adhered to the proposed criteria of credibility. In particular, "About two thirds of subgroup analyses associated with claims were not clearly prespecified". Similarly, in a study of the reporting of subgroup analysis in oncology between 2011 and 2013, Zhang *et al.*

(2015) found that analyses were "neither uniform nor complete". Following this, Altman (2015) recommended editors require authors to adhere to relevant reporting guidelines, in order to reduce the risk of publishing misleading results.

Yet, despite these proposals, there is clearly demand for a rigorous *post hoc* analysis of RCT data due to potentially large gains in efficiency of treatment. Accounting for multiple testing via cross-validation, or permutation methods, represents an important aspect of any such analysis, which are outlined in Section 4.3.1. We stress that any potential subgroups found this way should provide motivation for future trials, rather than used as evidence to guide future treatment.

4.2.2 Exploratory methods

Many methods for exploratory subgroup analysis have been proposed in the statistical literature. The vast majority of approaches are model-based, where a conditional model is specified for the response $Y \mid X = x, T = t$, given treatment and covariates. Once such a model is fitted, the output can be used to determine which covariate values will benefit from treatment. For example, to determine optimal treatment for an individual with $X = x_0$, one would compare the conditional means: $\mathbb{E}[Y \mid X = x_0, T = 0]$ and $\mathbb{E}[Y \mid X = x_0, T = 1]$. Any such model must include some potential interaction between X and T in order to model heterogeneity of treatment effect. Qian & Murphy (2011) use L_1-penalised least squares to estimate these conditional expectations, and from this derive an estimate for the optimal treatment rule $D_0 : X \to \{0, 1\}$ that assigns optimal treatment for each covariate value. Bounds are derived for the difference between the mean response under the estimated and optimal treatment rules.

Cai *et al.* (2011) use a 2-stage procedure, where first they fit a generalized linear model to estimate the conditional mean of the response separately for the control group and for the treated. The difference between the estimated means provides a parametric estimate

for the benefit of treatment at each covariate value. The covariate space is then split into regions with similar parametrically estimated benefit, and the benefit of treatment is estimated nonparametrically as a function of this parametric estimate. Matsouaka *et al.* (2014) take a similar approach in order to provide robustness to misspecification of the parametric model.

Tree-based methods such as classification and regression trees (Breiman *et al.*, 1984), and random forests (Breiman, 2001) have proven particularly popular methods as they allow for interaction and are able to capture potential subgroups via a recursive partitioning of the covariate space. Foster *et al.* (2011) apply such an approach to binary response data, forming a predictive model for $\mathbb{P}(Y = 1 \mid X, T)$ using a random forest taking as input $(X, T, X\mathbb{I}\{T = 0\}, X\mathbb{I}\{T = 1\})$. This fits a standard predictive random forest, which is used to calculate an estimate of the distribution of the response for each datum under treatment and no treatment, formally $\mathbb{P}(Y = 1 \mid X = x_i, T = 0)$ and $\mathbb{P}(Y = 1 \mid X = x_i, T = 1)$. The difference in this probability, defined as $Z_i := \mathbb{P}(Y = 1 \mid X = x_i, T = 1) - \mathbb{P}(Y = 1 \mid X = x_i, T = 0)$, is regarded as the treatment effect. A second regression tree is then fitted with Z as the response and X as covariates, in order to determine subsets of the covariate space corresponding to benefit of treatment above a pre-specified level.

Su *et al.* (2009) use an "interaction tree" approach, where the splitting criterion in the regression tree is the p-value of a test for interaction between treatment and the indicator variable $\mathbb{I}\{X_j < c\}$, where the proposed split is at the value c on the jth covariate. A fitted tree then divides the covariate space recursively on the basis of the presence of interaction effects. In the context of a drug trial, the fitted tree identifies subsets of the covariate space that have constant treatment effects. This output could then be used together with an estimation of the treatment effect in each subset in order to determine a subgroup that benefits from treatment. However, this second stage highlights the fact that interactions are not the same as a difference in optimal treatment. In the context of subgroup analysis, we are only interested in crossover interactions, those that lead to a difference in optimal treatment. The analysis

of wages via interaction trees (Su *et al.*, 2009) serves as an example of when the majority of detected iterations are non-crossover, which would be of little interest within a subgroup analysis.

Ruberg *et al.* (2010) use random forests regression on the control group, where each tree splits the covariate space into subgroups that differentiate the response in the control group. These are then used as proposed subgroups, where treatment effect is tested against placebo. However, by generating potential subgroups via analysis of only the control group, it is possible to miss treatment heterogeneity within a subgroup of the covariate space that has no primary effect. For example, a genetic marker may not be predictive of the response in the control group, and therefore none of the subgroups generated by the random forest would split on the genetic marker. If treatment results in a benefit for one genotype at this marker, but a detriment to the others, treatment could be deemed ineffective as it is never tested separately for different genotypes.

All of these approaches essentially split the task of finding optimal treatment into two stages: the data-generating procedure is modeled, and this model fit is used to determine an optimal treatment rule. Zhao *et al.* (2012) forgo the modeling stage and attempt to directly optimize the treatment rule in a technique called "Outcome weighted learning" (OWL). Rather than trying to estimate the conditional expectation $\mathbb{E}[Y \mid X, T]$, this approach considers the expectation of the response under a treatment rule $\mathbb{E}_D[Y] = \mathbb{E}[Y \mid T = D(X)]$, where $D : \mathbb{R}^p \to \{0,1\}$ maps covariates to treatment. Following the derivation of Qian & Murphy (2011), this expectation is expressed in the form:

$$\mathbb{E}[Y \mid T = D(X)] = \mathbb{E}[2Y\mathbb{I}\{T = D(X)\}]$$

when treatment is given to half of the patients in the RCT. Interest lies in the function D that maximises this expectation. Zhao *et al.* (2012) approximate this expectation using the

empirical distribution, and attempts to maximise:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{I}\{T_i = \text{sign}(f(X_i))\} \tag{4.2}$$

where the treatment rule $D(x)$ is represented as $\text{sign}(f(X))$ for some function f. To combat nonconvexity of the identity function this is then approximated by a hinge loss, the task becomes to minimise:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \max\left(1 - T_i \,\text{sign}(f(X_i)), 0\right) + \lambda_n ||f||^2.$$

where $||f||$ is some norm for the function f and λ_n is a parameter that penalises complexity of f. This optimisation is achieved using support vector machine techniques. However, there is a problem with approximating the expectation using the empirical distribution. It is easy to see that, when Y are all positive, Equation 4.2 is maximised by assigning the same treatment as seen in the data. Similarly, if Y is negative, this is maximised by giving the opposite treatment to that seen in the data. The complexity penalty prevents a trivial matching of observed and assigned treatments, but a bias is still present. This calibration issue is tackled by first regressing the covariates on the response and proceeding with the residuals, which are centered. This is called "Residual weighted learning" (Zhou *et al.*, 2015). In this sense, even taking this direct approach, the requirement of calibration necessitates the construction of a regression model.

We aim to build on this approach of direct consideration of the distribution of the response under a treatment rule, in effect marginalising out covariates. This can allow us to circumvent a conditional statistical model, and safeguard from the associated potential for misspecification.

4.3 Testing treatment rules

Following the direct approach of Zhao *et al.* (2012), we frame the question of subgroup analysis in terms of treatment rules $D : \mathbb{R}^p \to \{0,1\}$ that map the covariate space to treatment assignment. Cost of treatment is easily incorporated within this framework by using a cost-adjusted response: $Y_i := Y_i^{\text{clinical}} - \kappa T_i$. This is a simple pre-processing step, where any monetary cost is converted onto the response scale κ. First, we present a simple nonparametric test for testing a treatment rule D against its mirror rule D^c, the rule that assigns the opposite treatment in every case. This is shown to be equivalent to testing against treatment at random, a natural null rule when there is no established optimal course of treatment. The test is simple to adapt to the case where there is a current preferred course of treatment. This approach provides a simple and intuitive test that can be entirely model-free, resulting in robustness to misspecification. As with residual weighted learning, it is possible to first regress Y on X and use the following tests on the residuals rather than the response. This comes at the cost of making modeling assumptions, but when X are predictive of the response, this will result in increased power. We argue that the testing of treatment rules answers a more appropriate question than interaction tests, as they directly test the efficacy of a proposed treatment rule, rather than testing model parameters which are then combined to determine a treatment rule. This point is highlighted by the possibility of non-uniqueness of optimal treatment rules. Figure 4.2 shows a simple example of this, where both "treat all" and "treat if and only if $x > 0.5$", are optimal. This displays the lack of importance of the exact interaction region, and highlights the fact that finding interactions answers a different question to that posed when conducting subgroup analysis.

First note that there is a one-to-one correspondence between treatment rules and subgroups via the simple definition $D_A(X) := \mathbb{I}\{X \in A\}$. D_A is the treatment rule that assigns treatment if and only if covariates are contained in the subgroup A. Thus, any proposed subgroup can instead be viewed as the corresponding treatment rule. Given RCT data of the

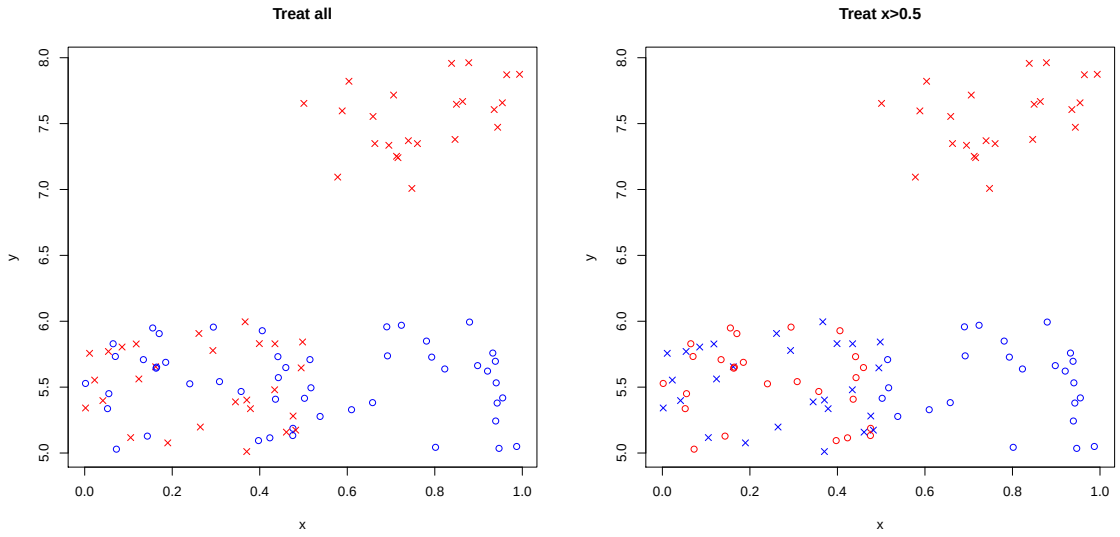


Fig. 4.2 Crosses represent treatment and circles represent no treatment. In this example it is optimal to treat those with $x > 0.5$. Any treatment rule that satisfies this criterion is optimal, such as: the rule to treat all (red points in the left panel), and the rule to only treat those with $x > 0.5$ (red points in the right panel).

form data of the form $\{X_i, Y_i, T_i\}_{i=1}^n$, interest lies in testing the efficacy of a treatment rule D. A natural measure of this is the probability of treatment benefit under D, defined by:

$$\begin{aligned}\alpha_D &:= \mathbb{P}(Y^D > Y) + \frac{1}{2}\mathbb{P}(Y^D = Y) \\ &= \int_X \mathbb{P}(Y^D > Y \mid X) + \frac{1}{2}\mathbb{P}(Y^D = Y \mid X)\pi(dX)\end{aligned}$$

where π denotes the covariate distribution, Y denotes the response under randomisation and Y^D denotes the response under the treatment rule D. Values of α_D greater than 0.5 indicate that the treatment rule is better than a randomised treatment. Randomised treatment provides a good reference point where there is no established treatment rule.

Since treatment at random treats in accordance with any rule D (that is, $T_i = D(X_i)$) with probability $\frac{1}{2}$ and in accordance with D^c ($T_i = D^c(X_i)$) with probability $\frac{1}{2}$, D is better than random if and only of it is better than D^c. This argument is key to our approach, and allows

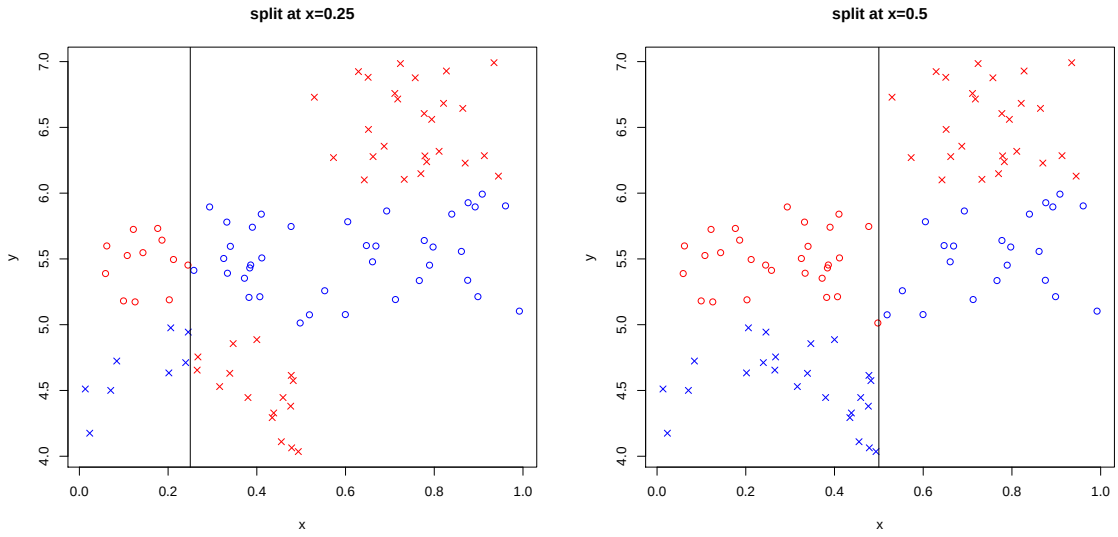


Fig. 4.3 Crosses represent treatment and circles represent no treatment. For subgroups $A_1 := \{X \leq 0.25\}$ (left) and $A_2 := \{X \leq 0.25\}$ (right), data can be split into two samples: those treated in accordance with D_{A_i} (blue) and $D^c_{A_i}$ (red).

the definition of hypotheses regarding α_D, the probability of benefit under D, in terms of hypothesis comparing D and its mirror D^c.

For each individual in a sample, for a given treatment rule D, either $T_i = D(X_i)$, or $T_i = D^c(X_i)$. Therefore, we have two samples: one sample of responses under the treatment rule D, and one sample under the treatment rule D^c. This is illustrated by a toy example in Figure 4.3 for the subgroups $\{X \leq 0.25\}$ and $\{X \leq 0.25\}$. Since treatment is typically administered to half of the patients at random, the sizes of these samples have the same distribution, and so, on average, are of equal size. Given these two samples $\{y^D\} \mid T = D(X)$ and $\{y^{D^c}\} \mid T = D^c(X)$ of size n_1 and n_2, we can perform a simple test to test for a difference in the distribution of the two samples. Motivated by robustness, we use a Mann-Whitney-U test, which has test statistics given by:

$$U = \sum_{i=1}^{n} R_i \mathbb{I}\{T_i = D(X_i)\}$$

where R_i represents the rank (1 corresponding to the lowest value) of Y_i within the sample $\{Y_j\}_{j=1}^n$. The distribution of this statistic is known under the null hypothesis:

$$H_0 : \mathbb{P}(Y^D > Y^{D^c}) + \frac{1}{2}\mathbb{P}(Y^D = Y^{D^c}) = 0.5$$

where Y^D denotes the response under treatment in accordance with the treatment rule D. The one-sided alternative hypothesis is given by:

$$H_1 : \mathbb{P}(Y^D > Y^{D^c}) + \frac{1}{2}\mathbb{P}(Y^D = Y^{D^c}) > 0.5$$

We can relate these hypotheses to the probability of treatment benefit under D, α_D:

$$\begin{aligned}
\alpha_D =& \mathbb{P}(Y^D > Y) + \frac{1}{2}\mathbb{P}(Y^D = Y) \\
=& \mathbb{P}[T = D(X)] \left(\mathbb{P}[Y^D > Y \mid T = D(X)] + \frac{1}{2}\mathbb{P}[Y^D = Y \mid T = D(X)] \right) + \\
& \mathbb{P}[T \neq D(X)] \left(\mathbb{P}[Y^D > Y \mid T \neq D(X)] + \frac{1}{2}\mathbb{P}[Y^D = Y \mid T \neq D(X)] \right) \\
=& \frac{1}{2}(0 + \frac{1}{2}) + \frac{1}{2} \left(\mathbb{P}[Y^D > Y^{D^c}] + \frac{1}{2}\mathbb{P}[Y^D = Y^{D^c}] \right) \\
=& \frac{1}{4} + \frac{1}{2} \left(\mathbb{P}[Y^D > Y^{D^c}] + \frac{1}{2}\mathbb{P}[Y^D = Y^{D^c}] \right)
\end{aligned}$$

That is, we can equivalently write our hypotheses as:

$$H_0 : \alpha_D = 0.5$$

$$H_1 : \alpha_D > 0.5$$

Therefore, testing for treatment benefit of a treatment rule D versus randomisation is achieved via a simple Mann-Whitney-U test of those treated in accordance with D versus those treated

in accordance with the mirror rule D^c. A similar argument can be applied for the use of alternative two-sample tests. For example, the null hypothesis can be formulated using expectations as follows:

$$H_0 : \mathbb{E}[Y^D - Y] = 0.$$

Similarly, we have:

$$\begin{aligned}\mathbb{E}[Y^D - Y] =& \mathbb{E}[Y^D - Y \mid T = D(X)]\mathbb{P}[T = D(X)] + \\ & \mathbb{E}[Y^D - Y \mid T \neq D(X)]\mathbb{P}[T \neq D(X)] \\ =& \frac{1}{2}\mathbb{E}[Y^D - Y^{D^c}].\end{aligned}$$

So, H_0 is equivalent to $H_0 : \mathbb{E}[Y^D - Y^{D^c}] = 0$, which can be tested for example using a t-test. In general, the Mann-Whitney-U test is recommended due to being a model-free nonparametric test, but a suitable test can be chosen by context. For example, when the response constitutes survival data with truncation and censoring, the log-rank test may provide a suitable nonparametric test.

The above relates the test of samples under a treatment rule D against a sample under D^c, to a test of D against randomised treatment. In some contexts, there will be a current preferred treatment rule (for example, when the treatment is expensive, a natural preferred course of treatment would be to not treat at all), and comparison to treatment at random is inappropriate. In this case, let D_* denote the preferred treatment rule, the standard against which other treatment rules are to be tested. To compare a treatment rule D to D_*, it is important to note where the treatments differ; $A_{D \neq D_*} := \{X \in \mathbb{R}^p : D(X) \neq D_*(X)\}$ defines the subset of the covariate space where the treatment rules differ. D is better (or worse) than D_* if and only if it is better (or worse) on the subspace $A_{D \neq D_*}$, since the treatment rules are identical elsewhere. Therefore, it is natural to pose a hypothesis test in terms of a difference

of effect on the subspace $A_{D\neq D_*}$. Our hypotheses become:

$$H_0 : \alpha_{D-D_*} = 0.5$$
$$H_1 : \alpha_{D-D_*} > 0.5$$

where:

$$\alpha_{D-D_*} := \mathbb{P}(Y^D > Y^{D_*} \mid X \in A_{D\neq D_*}) + \frac{1}{2}\mathbb{P}(Y^{D_*} = Y^D \mid X \in A_{D\neq D_*}).$$

Alternatively, for a test of difference of means, we now have:

$$H_0 : \mathbb{E}(Y^D - Y^{D_*} \mid X \in A_{D\neq D_*}) = 0$$
$$H_1 : \mathbb{E}(Y^D - Y^{D_*} \mid X \in A_{D\neq D_*}) > 0.$$

Again, this is simply a two-sample test of D against its mirror, after restricting the data to where D differs from D_*, that is, where $X_i \in A_{D\neq D_*}$. A natural consequence of this is that there is reduced confidence in benefit of a treatment rule when the two treatment rules are similar, differing only on a small subsample of the data.

4.3.1 Multiple testing issues

The p-values that come from these tests of treatment rules only have a valid interpretation if the treatment rules are proposed in advance, in small number, as described by Rothwell (2005). However, the exploratory methods outlined in section REF provide a means for generating subgroups (and therefore, treatment rules) from RCT data. Since the data have been used to identify these treatment rules, the p-value from a test using the same RCT data will not be uniformly distributed, so loses interpretation. One simple solution is to split the RCT data into "training" data and "test" data. Exploratory methods can be used on the

training data to identify a potential treatment rule, the validity of which can be tested using our simple test on the test data. Since different data are used to generate the treatment rule and test it, the p-value is valid. However, RCTs are often expensive, and as a result, small. Small datasets will result in test datasets too small to generate enough power to test the treatment rule effectively. An alternative approach is to use permutation methods to calibrate the p-value (see e.g. Davison & Hinkley, 1997, chapter 4.3). The idea is to generate datasets from the null hypothesis via permutation of some aspects of the data. Then the exploratory procedure is used to generate a treatment rule based for each dataset. A p-value is then calculated for each treatment rule on its respective dataset. This gives us an approximate distribution for p-values under multiple-testing, which can be used to adjust our p-value onto the usual scale. So far, the null hypothesis has not fully specified, so generation of data under the null in some cases requires additional information. We shall explore what assumptions are necessary in each case.

We wish to generate data under the null, but the null hypothesis is different for the test of each treatment rule; therefore, we generate data from a "global null", that is, a null that satisfies the null hypothesis corresponding to every possible treatment rule D. It is easy to see that the "sharp null" corresponding to the assumption of no treatment effect will satisfy each H_0, since all treatment rules become equivalent. Permuting the values of T_i generates datasets under this null, where any relationship between Y and T is lost, but any relationship between Y and X is kept intact. Lee & Rubin (2015) use such a procedure in the context of subgroup analysis to generate counterfactual data under this null to correct for multiple testing. This null hypothesis of no treatment effect is appropriate when there is no clear default treatment rule, as deviation from this null results in disparity between the performance of treatment rules, implying that some rules are better than others. The procedure is simply:

- Apply the exploratory method on the data $\{Y,T,X\}$, to get an estimate D^0 of the optimal treatment rule.

- Calculate a p-value p_0 for the test of $H_0 : \alpha_{D^0} = 0.5$ against $H_1 : \alpha_{D^0} > 0.5$, using the data $\{Y, T, X\}$.
- For $j = 1, \ldots, M$:
  - Generate v_j, a random permutation of $(1, \ldots, n)$
  - Apply the exploratory method on the data $\{Y, T_{v_j}, X\}$, to get an estimate D^j of the optimal treatment rule.
  - Calculate a p-value p_j for the test of $H_0 : \alpha_{D^j} = 0.5$ against $H_1 : \alpha_{D^j} > 0.5$, using the data $\{Y, T_{v_j}, X\}$.
- $\frac{1}{M}\sum_{j=1}^{M} \mathbb{I}\{p_0 > p_j\}$ gives a corrected p-value for $H_0 : \alpha_{D^0} = 0.5$ against $H_1 : \alpha_{D^0} > 0.5$.

Where T_v denotes the vector of treatment $(T_1, \ldots, T_n)$ re-ordered according to the permutation v. The procedure essentially estimates the distribution of p-values attained using this method of exploration followed by testing, under the assumption of no treatment effect. This is used to calculate a corrected p-value that is uniformly distributed under this assumption.

When there is a current preferred treatment rule, it is inappropriate to generate under the assumption of no treatment effect, as this preference implies a perceived difference between the treated and control groups. In order to generate data under the assumption that the optimal treatment rule is D_*, one would have to assume a conditional model for Y that is consistent with D_* being optimal. Datasets generated from this model could be used as before to calculate the distribution of p-values under multiple testing. The construction of such a model will depend on the beliefs surrounding the currently preferred treatment rule, and such an approach is highly sensitive to model choice. However, in a specific case, it is possible to augment the permutation method to generate data under a suitable null, rather than specifying a full model. This is the case where treatment is expensive, so the *a priori* preferred treatment rule is to not treat anyone, that is $D_*(X) = 0 \; \forall X$. Recall that we have been using a cost-adjusted response $Y_i := Y_i^{\text{clinical}} - \kappa T_i$, where κ is the cost of treatment on

the scale of the clinical response. A natural null might be that treatment has no effect on Y^{clinical}, so the cost κ makes not treating an optimal treatment rule. It is easy to generate under such a null, by using the clinical response, permuting treatments and then adjusting to the cost-adjusted response based on the permuted treatment. Under this sampling scheme, any Y dependence on X is preserved, the treatment cost is incorporated and clinical response is independent of treatment. A treatment rule is estimated via exploration on each dataset, which is then tested against D_*. To clarify, the procedure becomes:

- Apply the exploratory method on the data $\{Y,T,X\}$, to get an estimate D^0 of the optimal treatment rule.
- Calculate a p-value p_0 for the test of $H_0 : \alpha_{D^0-D_*} = 0.5$ against $H_1 : \alpha_{D^0-D_*} > 0.5$, using the data $\{Y,T,X\}$.
- For $j = 1,\ldots,M$:
  - Generate v_j, a random permutation of $(1,\ldots,n)$
  - Apply the exploratory method on the data $\{Y + \kappa(T_{v_j} - T), T_{v_j}, X\}$, to get an estimate D^j of the optimal treatment rule.
  - Calculate a p-value p_j for the test of $H_0 : \alpha_{D^j-D_*} = 0.5$ against $H_1 : \alpha_{D^j-D_*} > 0.5$, using the data $\{Y + \kappa(T_{v_j} - T), T_{v_j}, X\}$.
- $\frac{1}{M}\sum_{j=1}^{M} \mathbb{I}\{p_0 > p_j\}$ gives a corrected p-value for $H_0 : \alpha_{D^0-D_*} = 0.5$ against $H_1 : \alpha_{D^0-D_*} > 0.5$.

These permutation methods allow calculation of valid p-values for the validity of treatment rules proposed via an exploratory analysis of the data. In each case, the permutation method makes minimal assumptions about the data-generating procedure.

4.4 Random forests of treatment trees

In this section, we propose an exploratory method for finding an optimal treatment rule, based on the use of nonparametric tests for treatment rules, as outlined in section 4.3. The idea is to test large number of treatment rules using the RCT data, choosing the treatment rule that gives the most significant p-value for a test of benefit of treatment. The space of treatment rules is explored via recursive partitioning of the covariate space, as in classification and regression trees. Throughout, we shall use the Mann-Whitney-U test, but as before, any two-sample test (log rank test, t-test) can be used. As before, if there is thought to be a strong primary effect of X on the response, regressing X on Y first and proceeding with the residuals will increase the power to detect treatment heterogeneity. Of course, in this case the approach is no longer model-free.

Decision trees provide a means of recursively partitioning a covariate space. Classification and regression trees (see e.g. Breiman *et al.*, 1984) are a popular machine-learning tool, where a prediction model is fitted within each partition of a decision tree. Each partition is chosen on the basis of a splitting criterion, for example in regression a reduction in the sum of squared errors. The recursive partitioning allows a complex structure between covariate and response. In other settings, Su *et al.* (2009) (interaction trees) and Ishwaran *et al.* (2008) (survival forests) have used decision trees in a non-standard fashion, using an interaction test statistic and a log-rank test statistic respectively as a splitting criteria. Decision trees provide a natural way to construct treatment rules, where a treatment is assigned to each partition. The tests for treatment rules in section 4.3 provide a natural splitting criterion, constructing "treatment trees" in a similar way to the trees in Ishwaran *et al.* (2008) and Su *et al.* (2009). However, the use of treatment rule tests avoids splits corresponding to "non crossover" interactions, that would be present in interaction trees, which are not of interest when determining an optimal treatment rule.

Random forests (Breiman, 2001) are an ensemble method that uses a "forest" of decision trees; each tree is trained using a bootstrapped sample of the data and a subset of covariates. In a regression setting, the collection of regression trees is utilized by averaging over the predictions of each tree. Similarly, in a classification setting, the classification of the forest is simply the most popular classification among the constituent trees. The central idea is that each tree will have low bias but high variance, with trees in the forest having low pairwise correlation. The aggregation of these into a forest should therefore have low bias and low variance (Friedman *et al.*, 2001, chapter 15). In the same way, we aggregate treatment trees constructed using nonparametric tests in order to construct a consensus treatment rule with low bias and low variance. In addition, "selective logging" is introduced as a forest-level analogue of "pruning", which combats potential issues with covariate sub-sampling. First, we address the construction of trees for treatment rules, and then describe how these are assembled into a forest.

Given data $\{Y,X,T\}_{i=1}^{n}$, a treatment tree is constructed as follows. For every covariate, a split is proposed at each quintile of the covariate sample. Each split represents two treatment rules: treat to the right, or treat to the left. Concretely, denote the split k on covariate j by $q_{j,k}$, and the corresponding two treatment rules: $D(X) = \mathbb{I}\{X_j < q_{j,k}\}$ and $D^c(X) = \mathbb{I}\{X_j \geq q_{j,k}\}$. Visually, for each covariate, we generate a sequence of rules, as illustrated by Figure 4.3. This gives us a total of 10 potential treatment rules for each covariate. Add to these the constant treatment rules: $D(X) = 1$ and $D(X) = 0$. Then, for each treatment rule D, calculate the test statistic:

$$U = \sum_{i=1}^{n} R_i \mathbb{I}\{T_i = D(X_i)\}$$

where R_i represents the rank (1 corresponding to the lowest value) of Y_i within the sample $\{Y_j\}_{j=1}^{n}$. That is, at split k on covariate j, we calculate:

$$U = \sum_{i=1}^{n} R_i \mathbb{I}\{T_i = \mathbb{I}\{X_{i,j} < q_{j,k}\}\}.$$

Then calculate the corresponding p-value associated with the one-sided alternative hypothesis. The split corresponding to the lowest p-value is then chosen, with children nodes being passed the data to the left and the right of the split, respectively. If the lowest p-value is above some pre-specified value p_{cutoff}, or if the lowest p-value corresponds to either $D(X) = 1$ or $D(X) = 0$, no split is made. In addition, each tested rule must result in a two-sample test above a pre-specified size $\min_{\text{test}}$, and must split the data into child nodes above a pre-specified size $\min_{\text{child}}$. This procedure is iterated on each child node until no split is made, or until a pre-specified maximum depth D is reached. Once constructed, the tree represents a treatment rule where each leaf node is assigned treatment according to that chosen by the parent splitting rule. If the tree is just a root, treatment is simply $D(X) = 1$ or $D(X) = 0$, determined by the lowest p-value.

As with the standard extension to random forests, a forest comprises of N trees trained on a bootstrapped subsample of the data with a random subsample of Q covariates, where N and Q are pre-specified. Since each tree is trained on a bootstrapped subsample of the data, the remaining data can be used to test the treatment rule corresponding to the tree. This is a form of cross-validation, since different data is used to train and test the rule. As such, each tree is associated with a p-value. Given a "selective logging" parameter $J \in \{1,\ldots,N\}$ (a pre-specified constant), the final forest is given by the J trees with the lowest p-values. Finally, this forest is used to define a treatment rule by the consensus of the constituent trees. That is, given a forest $F = \{DT_j\}_{j=1}^{J}$, where DT_j is the jth decision tree, the forest rule is given by: $F(X) = \mathbb{I}\{\sum_{j=1}^{J} DT_j(X) > 0.5\}$.

The reasoning behind this final selective logging step is illustrated by the following simple example. Consider a case where there are 50 covariates, where only one binary covariate X_1 produces treatment heterogeneity, where $X = 0$ results in a slightly harmful treatment effect and $X = 1$ results in a hugely beneficial treatment effect. Then, for all trees that contain X_1 (after covariate sub-sampling) the treatment rule should be $D(X) = \mathbb{I}\{X_1 = 1\}$. For all

trees that do not contain X_1, treatment will be universally better than no treatment, since the benefit for $X = 1$ is larger than the detriment for $X = 0$ (a case where $X = 1$ is more frequent than $X = 0$ represents a similar example of this effect). As such, these trees should give the treatment rule $D(X) = 1$. If the number of covariates to be sub-sampled, Q is small enough, this would result in the rule $D(X) = \mathbb{I}\{X_1 = 1\}$ being "washed out" by the more frequent $D(X) = 1$. As a result, the forest would assign treatment to those with $X_1 = 0$ despite every tree containing X_1 assigning the opposite. Selective logging combats this problem, since the forest is now comprised of a subset of trees associated with better treatment, in this case the trees containing X_1. As such, J and Q should be specified with this in mind. If Q is small relative to the total number of covariates p, J should accordingly be small relative to the total number of trees N. As a rule-of-thumb, note that approximately $(\frac{Q}{p})^2$ of the trees will contain a particular pair of covariates. So, in order to not wash out any covariate interaction, one might wish to specify $J = N(\frac{Q}{p})^2$. However, the use of data out of the "bag" corresponding to each tree to determine inclusion of the tree will create higher pairwise correlation between trees than found in a standard random forest. As such, there is something of a balancing act, where selective logging is most important in high-dimensional applications, where the effect of covariate sub-sampling is more pronounced.

As with standard random forest algorithms, it can be difficult to interpret or visualise the output in high dimensions, in contrast to a single tree. However, there are some simple techniques that can aid interpretation. The relative importance of covariates is difficult to measure concretely, but the number of trees in the forest that contain a split on a given covariate provides a reasonable indication of the relevance of the covariate. Alternatively, the average p-value associated with trees that contain a split on a given covariate can be used to order covariates by importance. Once important covariates have been identified, heat-map plots of $\sum_{j=1}^{J} DT_j(X)$ can be made by varying X in two dimensions with the other dimensions

fixed. Heat-maps over important covariates can provide an indication of the treatment rule boundary; an example of this visualisation is given in the next section.

4.4.1 Comparison study with OWL

In this section we compare performance of our random forest algorithm with that of OWL (Zhao *et al.*, 2012) using examples found in the original paper. The authors kindly provided software to implement OWL using a linear kernel function. The examples each have a highly complex optimal treatment rule, which serves to illustrate the limitations of our algorithm. In practice, it is highly implausible that optimal rules should take such a form, but it is interesting to observe to what extent tree structures and a linear kernel can approximate such rules.

The covariate X space is 50 dimensional, with each element being independently uniformly distributed on the interval $[-1,1]$. In contrast to our standard notation, the treatment T takes values in $\{-1,1\}$ uniformly. Given treatment T_i and covariate X_i, the conditional distribution of the response is given by:

$$Y_i \mid X_i, T_i \sim \mathrm{N}(1 + 2X_{i,1} + X_{i,2} + 0.5X_{i,3} + f(X_i, T_i), 1)$$

where for each example, the treatment effect is given by:

1. $f(X,T) = 0.442(1 - X_1 - X_2)T$
2. $f(X,T) = (X_2 - 0.25X_1^2 - 1)T$
3. $f(X,T) = (0.5 - X_1^2 - X_2^2)(X_1^2 + X_2^2 - 0.3)T$

Clearly, the response only depends on the first three covariates - X_1, X_2 and X_2 - with the treatment effect only depending on X_1 and X_2. Therefore, the optimal treatment rule is simply a function of the covariates X_1 and X_2. A dataset of size 250 is sampled, and both OWL and

our algorithm are applied. The random forest is parameterized as follows: $D = 4$ (maximum tree depth), $Q = 25$ (number of covariates used by each tree), $N = 1000$ (initial number of trees), $J = 500$ (number of trees after selective logging), $p_{\text{cutoff}} = 0.1$ (threshold p-value required to split).

The results are visualised by plotting the fitted treatments of points on a 50×50 grid of points on $[-1,1]^2$ (the values of X_1 and X_2), with a random vector for the remaining 48 covariates. The random forest fit is visualised by plotting $\sum_{j=1}^{J} DT_j(X)$ for each X value. Since treatment takes the value 1 or -1, values greater than 0 will be given treatment 1 and values below 0 will be given treatment -1. The sign of the fitted linear kernel function indicates the treatment given by OWL. Figure 4.4 displays the fits alongside the true optimal treatment rule, where red corresponds to treatment 1 and blue corresponds to treatment -1. Clearly OWL fits the first example very well, with our algorithm picking up the optimal treatment rule, but struggling to fit it using a mixture of tree structures. This is as expected, as OWL optimizes over such linear treatment rules. In the second example, both OWL and the random forest identify that -1 is the best treatment for the majority, with some high X_2 value individuals being assigned treatment 1; our methods seems to perform better here, picking up a larger portion of this region. Clearly, the limiting factor in both cases is the inability to fit a parabola with a linear function or using a "grid" structure. In the final example, both OWL and our algorithm assign treatment -1 to each individual in the plot. Neither the linear kernel or the average of trees is able to capture the ring. This highlights the limitations of both approaches, but we would argue that these examples represent extreme scenarios and are not representative of biological applications, that would likely take much simpler forms.

It is possible to calculate a p-value for the test of each treatment rule against its mirror. These use the same data as was used to train the treatment rules, so do not constitute calibrated p-values. Rather they provide a relative source of evidence in favour of a treatment rule versus treatment at random. As such, these can be used to compare the fits of our algorithm

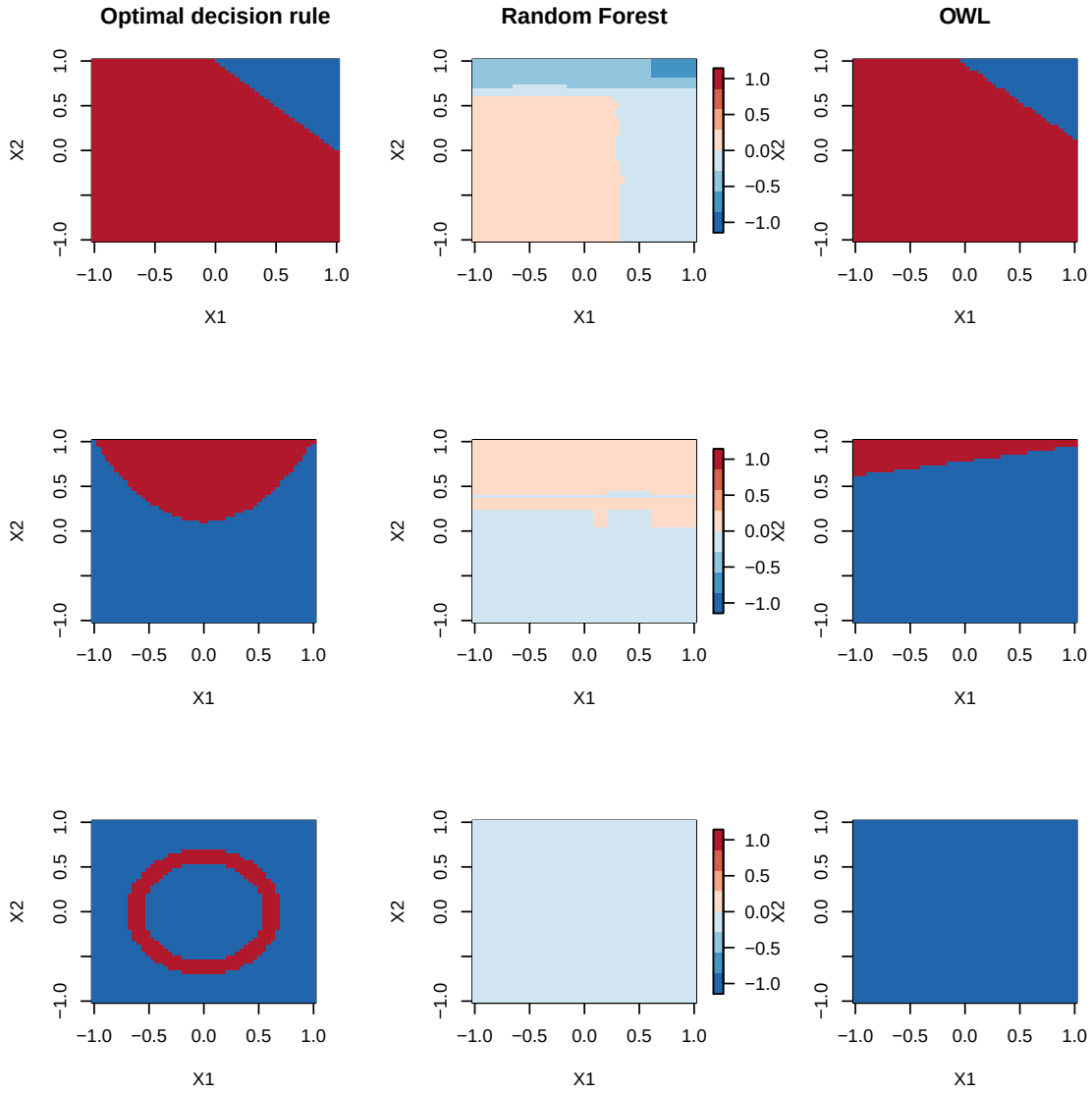


Fig. 4.4 Each row corresponds to a different data-generating model, with the first column indicating the optimal treatment rule, the second column the average treatment allocation among trees in the random forest and the third column the treatment rule output of OWL. Red corresponds to treatment 1 and blue treatment -1. The consensus forest treatment rule takes the value 1 when the tree average treatment is positive and -1 when the average is negative.

with OWL in each example. In the first example, the OWL rule gives a p-value of 8.5×10^{-16} compared with 1.1×10^{-4} for the random forest, indicating a far better OWL fit, supporting the qualitative observations from Figure 4.4. In the second example, the OWL treatment rule gives a p-value of 10^{-11}, compared to 1.7×10^{-13} for the random forest, indicating a slightly better fit by our algorithm, in accordance with our observations. For the third example, the plots would seem to indicate equality of the treatment rules; however, these are the treatment rules for a random value of $X_3, \ldots, X_{50}$. Over the entire covariate space, the two rules both assign a mixture of treatment 1 and -1 (though, dominated by -1). The resulting p-values of these treatment rules are 5.8×10^{-12} for OWL and 1.8×10^{-14} for the random forest. Again, the random forest provides a slightly better fit. All things considered, performance of the two algorithms is comparable on these particular examples, with OWL performing particularly well in the linear case.

4.4.2 AIDS Clinical Trials Group Study

In this section we apply the developed methodology to AIDS Clinical Trials Group Study (ACTG 320) data which is publicly available[1]. The data is from a double-blind, placebo-controlled trial stratified by initial CD4 count, where each of the 1151 patients were randomly assigned one of two daily treatment regimes: a two-drug regimen of lamivudine (3TC) with idovudine (ZDV) or stavudine (d5T); or the three-drug regime also containing indinavir (IDV). We group these into those who received IDV treatment and those who did not receive IDV. The trial was designed to test for the effect of IDV on the time to development of AIDS or death (disease-free survival time). The response Y is censored survival data, with 96 events and 1055 censored observations. We consider the exploratory analysis of optimal treatment rules with our random forest algorithm using the binary, ordinal, continuous covariates:

[1] https://www.umass.edu/statdata/statdata/data/

gender, IV drug use history, Hemophilia, Karnofsky Performance Scale, baseline CD4 count, prior ZDV use and age at enrollment, as specified in Table 4.1.

Table 4.1 ACTF 320 Covariates

Variable name	**Variable type**	**Variable coding/Units**
Gender	Binary	1 = Male
IV drug use history	Ordinal	0 = Never 1 = Previously 2 = Currently
Hemophilia	Binary	1 = Hemophiliac
Karnofsky Performance Scale	Ordinal	0 = Normal activity not possible 1 = Normal activity with effort 2 = Normal activity possible 3 = Normal
Baseline CD4 count	Continuous	Cells/milliliter
Prior ZDV use	Continuous	Months
Age at Enrollment	Continuous	Years

Following such a trial, the natural first-step is to test for treatment effect. Within our framework, this is simply the test of the treatment rule $D(X) = 1$ versus the treatment rule $D(X) = 0$. Since the response consists of censored survival data, we use the Log-Rank test. The resulting p-value is 0.00117, indicating highly significant evidence in favour of treatment resulting in an increase in disease-free survival time. With high levels of censoring, the Kaplan–Meier plot of estimated survivor functions provides a means of visualising the distribution of the response. Figure 4.5 shows these estimates for treated and non-treated individuals, supporting the hypothesis of a treatment effect. The ACTG study was in fact stopped early as a result of the strong evidence of efficacy of IDV. However, there is still the potential for a superior treatment rule. With the evidence of treatment effect in mind, it would be desirable to test the treatment rule output of our random forest algorithm against the "treat all" treatment rule $D(X) = 1$. Accounting for multiple testing for such a test does not fall within standard permutation methods (see Section 4.3.1), and would require model

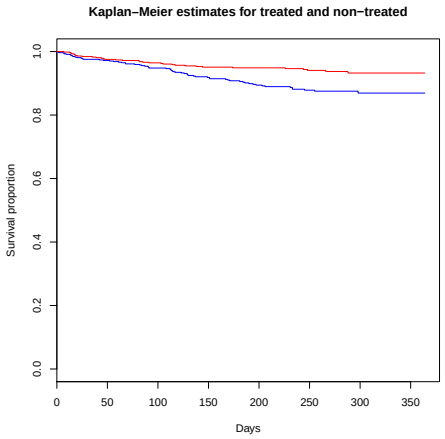


Fig. 4.5 Kaplan–Meier estimates of survivor functions under treatment (red) and no treatment (blue).

assumptions to be made in order to generate data under such a null hypothesis. With this in mind, we turn to cross-validation, fitting our random forest on a random "training" subset of the data of size 700, using the remaining 451 datum to test the generated treatment rule.

We apply our random forest with $D = 2$ (maximum tree depth), $Q = 4$ (number of covariates used by each tree), $N = 300$ (initial number of trees), $J = 100$ (number of trees after selective logging), $p_{\text{cutoff}} = 0.025$ (threshold p-value required to split). The tree depth is set with respect to the small number of covariates, and a relatively low cutoff is chosen to avoid over-fitting of the data. Let $RF(X)$ denote the treatment rule corresponding to the consensus of our fitted random forest. When applied to the individuals in the study, RF assigns treatment to 996 of the 1151 individuals in the study; this would be expected, following evidence of a strong global treatment effect. We can examine the fit of RF by looking at the variables that appear in the most trees in the forest. The two most important variables appear to be prior ZDV use and age, appearing in 35% and 23% of trees respectively. The next most important variable was gender, in only 16% of trees, with all other variables in at most 5% of trees. We can examine the fit visually via a sequence of pairs plots of the treatment assignment of RF for each individual in the study, where in each plot the axes are a pair of the most important variables. Figure 4.6 displays the assignment of treatment under

RF in this way using the three most important variables. The majority are given treatment, but a small subgroup benefit from no treatment. This subgroup seems to correspond to individuals with fewer than 20 months prior ZDV use and ages in the range $0-30$, along with some with low prior ZDV use at higher ages. The Log-Rank test tests the efficacy of the *RF* treatment rule versus random on the test data gives a p-value of 0.036, indicating that the rule is significantly better than treatment at random. We would expect this in the context of this particular trial as treatment seems to have a global effect, so any rule that treats the vast majority of individuals should be effective relative to randomisation. Therefore, the more appropriate question is: is *RF* better than the "treat all" rule? Again, we use a Log-Rank test on the test data, resulting in a p-value of 0.87, indicating there is not enough evidence to reject the "treat all" rule in favour of *RF*. Our analysis would conclude with a recommendation of treatment to all, in accordance with the original findings.

In many cases there is a cost associated with treatment; in some cases, this is monetary, but in others the cost is in terms of quality of life, due to, for example, side-effects of treatment. This is often incorporated by using quality-adjusted life-years (QALYs), (Zeckhauser & Shepard, 1976), rather than survival time. Here we consider a case where treatment results in a reduction of quality of life by a factor of 0.5; that is, one year of treatment-free life is equivalent to two years of life with treatment. (Note that this is purely for illustration, and is not based on the effects of treatment using IDV). We define the response $Y_i^{QALY} = Y \times 0.5^{T_i}$, with censored data $\{Y_i > y_i\}$ treated in the same way: $\{Y_i^{QALY} > y_i \times 0.5^{T_i}\}$, since $\{Y_i^{QALY} > y_i \times 0.5^{T_i}\} \iff \{Y_i > y_i\}$. Again, it is natural to first test the efficacy of treatment using a Log-Rank test. This time, the p-value is 0.43; insignificant evidence to indicate a benefit of blanket treatment. In contrast to using the unadjusted response, this time we have no *a priori* preferred treatment rule following what is, essentially, a failed trial; this represents a typical application of exploratory subgroup analysis. The QALY factor of 0.5 was chosen to reflect such a scenario. In this case, we are interested in any treatment rule that

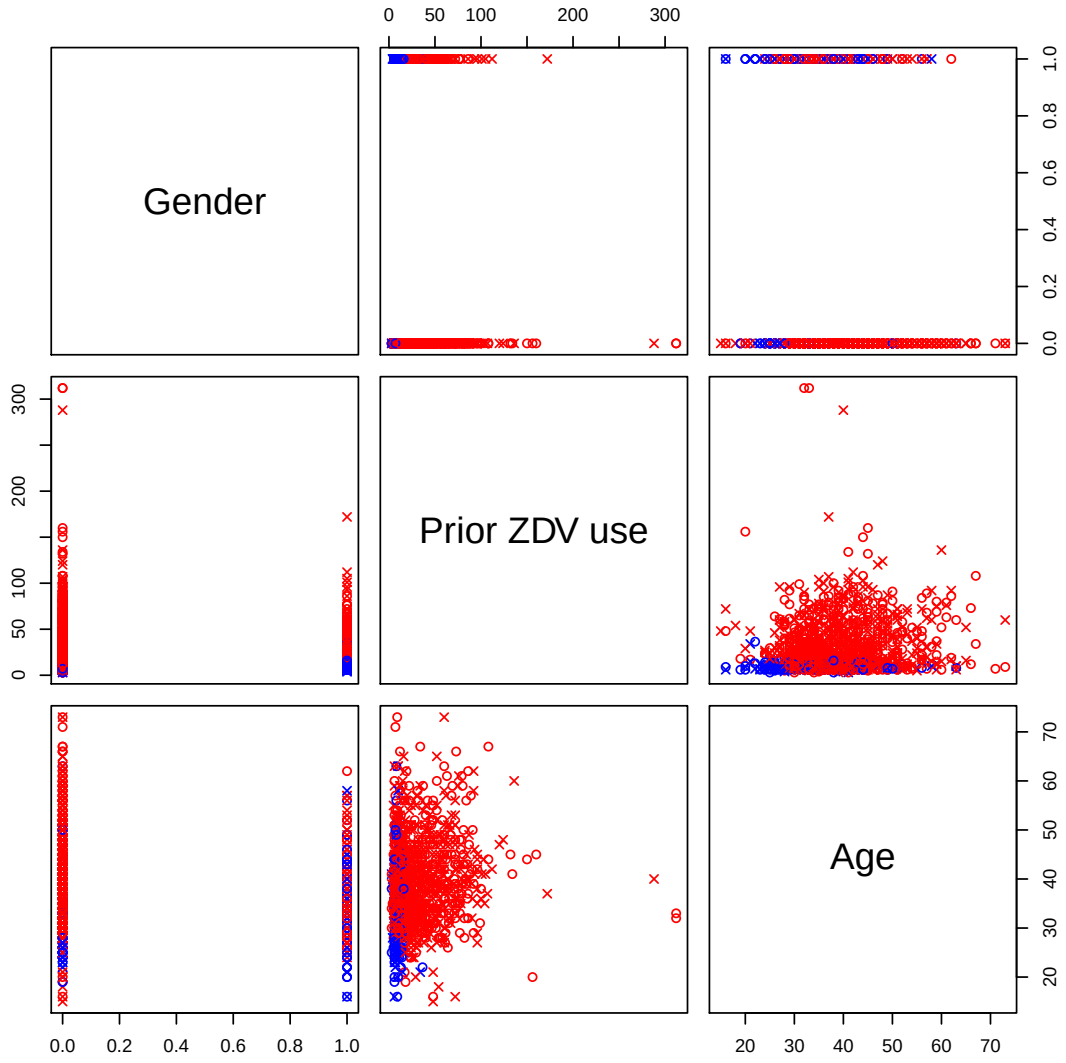


Fig. 4.6 Random forest assignment of treatment to individuals in the study, where those recommended IDV treatment are in red and those recommended no treatment in blue. Individuals treated in the study are represented with crosses with those not given IDV represented with circles.

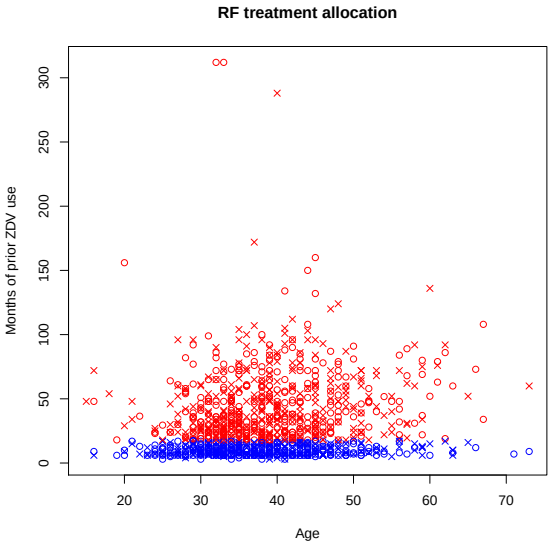


Fig. 4.7 Random forest assignment of treatment to individuals in the QALY study, where those recommended IDV treatment are in red and those recommended no treatment in blue. Individuals treated in the study are represented with crosses with those not given IDV represented with circles. The clear trend is the recommendation of treatment to those with more than approximately 17 months prior ZDV use.

is significantly better than treatment at random. Multiple testing is accounted for in this case using permutation methods, so this time we use the entire dataset to train our random forest algorithm. Again, we apply our random forest algorithm using the same parameterization. We asses variable importance simply by counting the number of trees in the forest containing a split on a given variable. 95% of the trees split on prior ZDV use, 1% of trees split on Karnofsky Performance Scale, and no trees split on any of the other variables. This indicates a clear-cut treatment rule that is essentially a decision tree with a single split, on Prior ZDV use. The average value of this split is 17.7, indicating that the corresponding treatment rule is approximately the rule that treats only the individuals with more than 17.7 months of prior ZDV use. Figure 4.7 shows the allocation of treatment recommended by RF to individuals in the study, sorted by age and prior ZDV use, confirming the nature of the treatment rule.

We wish to test RF versus its mirror rule RF^c, equivalent to testing the null hypothesis that the distribution of Y^{QALY} - the quality-of-life adjusted response - under RF is the same as

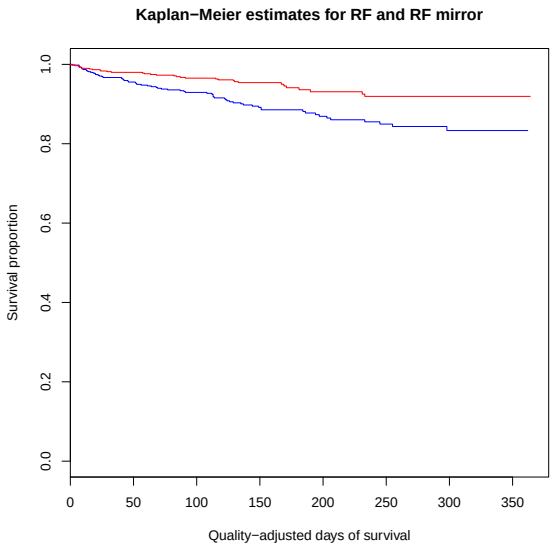


Fig. 4.8 Kaplan-Meier estimates of survivor functions of those treated in accordance with *RF* (red) and those not treated in accordance with *RF* (blue).

that under treatment at random. The p-value for this test is 2.4×10^{-4}; however, this clearly needs calibration due to multiple testing. To do this, we simply permute the treatment vector, as this generates data under the assumption that treatment has no influence on Y^{QALY}. This is a natural null hypothesis under the assumption that any gain in longevity due to treatment is counterbalanced by the effect on quality of life; this certainly seems reasonable with the lack of any *a priori* preferred treatment rule. The random forest is applied to each of 100 permuted datasets, each giving a corresponding p-value. 2 of these p-values are smaller than our p-value of 2.4×10^{-4}, resulting in a corrected p-value of 0.02, and a rejection of the null hypothesis that the distribution of quality-adjusted lifetime is equal under treatment according to *RF* and under randomisation.

Figure 4.8 shows the estimated survival functions for individuals treated in accordance with *RF*, and those treated in accordance with RF^c, indicating a clear benefit of treatment under *RF*. This example highlights the potential advantage of exploratory subgroup analysis when applied with careful consideration to multiple-testing issues. In practice, if such a quality-of-life adjustment was appropriate, this analysis would motivate the construction of a

new trial to confirm the efficacy of the *RF* treatment rule to treat only those with more than 17.7 months prior ZDV use.

4.5 Bayesian approach to subgroup analysis

Throughout, we have considered a frequentist nonparametric testing approach to subgroup analysis, motivated by the desire to circumvent model specification and associated assumptions. Here, we present a Bayesian alternative, based on the application of the nonparametric model in Section 3. This model is parameterised in terms of: the marginal distribution of the response, and a parameter β_j for each covariate x_j. We conduct a subgroup analysis via the specification of a model $\mathscr{M}_k$ corresponding to each potential subgroup A_k. Each subgroup is incorporated by extending the covariate matrix to include an indicator for treatment in accordance with the associated treatment rule. For each "subgroup" A of the covariate space, there are two associated treatment rules: $D_A(X) := \mathbb{I}\{X \in A\}$, and the mirror rule $D_A^c(X) := \mathbb{I}\{X \notin A\} = D_{A^c}(X)$. Define the variable $S_{i,A} := \mathbb{I}\{\{X_i \in A \cap T_i = 1\} \cup \{X_i \notin A \cap T_i = 0\}\}$, which takes the value 1 if individual i is treated in accordance with D_A, and 0 if treated in accordance with D_{A^c}. For model $\mathscr{M}_k$, we include the vector S_{A_k} in the covariate matrix, where positive values of β indicate $D_{A_k^c}$ is preferable, and negative values indicate D_{A_k} is preferable. As such, we only include one subgroup for each $\{A, A^c\}$ pair, as consideration of one covers the analysis of both.

We consider a subset of the subgroups defined by trees of depth 2, namely: subgroups of the form: $\{X_{i,j} \leq c_{j,k}\}$ and all combinations of two such splits on different covariates. This represents the subset of trees of depth 2 that have the same split in both branches of the second level of the tree. This greatly reduces the search space, which aids computation, whilst still enabling the capture of interactions. The "level 0" subgroup is simply $\mathscr{X}$, inducing the treatment rules $D(X) = 1$ (treat everyone) and $D(X) = 0$ (treat no one). The "level 1" subgroups are of the form $\{X_{i,j} \leq c_{j,k}\}$, where $c_{j,k}$ is the kth quintile of covariate j. The

hyperplanes $X_{i,j} = b$ and $X_{i,l} = c$ split the covariate space into 4 subspaces; each of which defines a subgroup, as does the "diagonal", defined by $\{[X_{i,j} \leq b \cap X_{i,j} \leq c] \cup [X_{i,j} > b \cap X_{i,j} > c]\}$. These constitute the subgroups at level 2. All other subsets generated by $X_{i,j} = b$ and $X_{i,l} = c$ are either complements of these, or are subgroups contained in level 0 or level 1.

For each model, the same priors are used for F_X and F_Y, which we do not specify. Rather attention is focused on β and γ, the covariate and treatment rule effects. For model $\mathscr{M}_k$, we define:

$$\lambda(X,S) = \exp\{X\beta + S_{A_k}\gamma\}$$

where S_{A_k} is the vector containing $S_{i,A_k} := \mathbb{I}\{\{X_i \in A_k \cap T_i = 1\} \cup \{X_i \notin A_k \cap T_i = 0\}\}$ for $i = 1,\ldots,n$. β and γ are given mean 0 Gaussian priors, and the model specification is completed via the allocation of prior probability to each model $\mathscr{M}_k$.

4.5.1 AIDS Clinical Trial application

We apply this Bayesian nonparametric approach to the AIDS Clinical Trial Study Group data analysed in Section 4.4.2 using the model-free approach. As described in the previous section, a model is defined for each subgroup A_k, a subset of trees of depth 2. We set the prior probability of each model via:

$$P_0[\text{Level 0 tree}] = 0.9$$
$$P_0[\text{Level 1 tree}] = 0.05$$
$$P_0[\text{Level 2 tree}] = 0.05$$

where each tree within a level is given equal mass. As before, we consider splits on the quintiles of each continuous covariate, and splits on each value of ordinal covariates. The result is 1 level 0 tree, 19 level 1 trees and 745 level 2 trees. Therefore, the prior probability

for the model corresponding to a particular level 2 tree is $0.05 \times \frac{1}{745} = \frac{1}{14900}$. Clearly the Bayesian approach provides a natural means for the correction of "multiple testing" through the prior. Within each model, we use the prior:

$$\beta_j \sim \mathrm{N}(0,4) \quad j = 1,\ldots,7$$
$$\gamma \sim \mathrm{N}(0,0.5)$$

As we would expect the impact of treatment to be of smaller magnitude than primary covariate effects. Since all models have the same prior for F_Y and F_X, the independent form of the composite likelihood allows us to proceed with inference based solely on (β,γ), using the partial likelihood term corresponding to the probability of the observed (partial) ranking. The posterior distributions of covariate effects within each model can be approximated using a Laplace approximation, which is simply computed using the output of a suitable ridge regression fit[2] - which mimics our Gaussian prior - to calculate the MAP and the matrix of second derivatives of the posterior at this point. For each k, this gives us an approximation $q^{(k)}_{\mathrm{Laplace}}(\beta,\gamma \mid \mathscr{M}_k)$ of the posterior of (β,γ) under $\mathscr{M}_k$:

$$q^{(k)}_{\mathrm{Laplace}}(\beta,\gamma) \approx p(\beta,\gamma \mid \{y_i,x_i,t_i,c_i\}_{i=1}^n, \mathscr{M}_k)$$

[2]using the "coxph" function in the R "survival" package

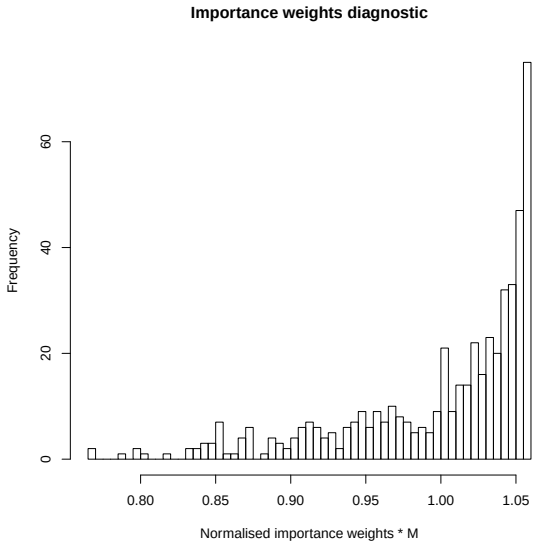


Fig. 4.9 Histogram of the sample of normalised importance weights ($\times M$) for the approximation of the marginal likelihood. Efficient importance sampling is associated with these weights being close to 1, as is the case here.

This approximation can be used to calculate the marginal (composite) likelihood under each model $\mathscr{M}_k$ via importance sampling (see e.g. Ripley, 2009) as follows:

$$\begin{aligned}
L_C(\mathscr{M}_k;\{y_i,x_i,t_i,c_i\}_{i=1}^n) &= \int_{\beta,\gamma} L_C(\beta,\gamma,\mathscr{M}_k;\{y_i,x_i,t_i,c_i\}_{i=1}^n)p(\beta,\gamma\mid\mathscr{M}_k)d\beta d\gamma \\
&= \int_{\beta,\gamma} q^{(k)}_{\text{Laplace}}(\beta,\gamma\mid\mathscr{M}_k)\frac{L_C(\beta,\gamma,\mathscr{M}_k;\{y_i,x_i,t_i,c_i\}_{i=1}^n)p(\beta,\gamma\mid\mathscr{M}_k)}{q^{(k)}_{\text{Laplace}}(\beta,\gamma\mid\mathscr{M}_k)}d\beta d\gamma \\
&= \mathbb{E}_{q^{(k)}_{\text{Laplace}}}\left[\frac{L_C(\beta,\gamma,\mathscr{M}_k;\{y_i,x_i,t_i,c_i\}_{i=1}^n)p(\beta,\gamma\mid\mathscr{M}_k)}{q^{(k)}_{\text{Laplace}}(\beta,\gamma\mid\mathscr{M}_k)}\right] \\
&\approx \frac{1}{M}\sum_{r=1}^{M}\frac{L_C(\beta_r,\gamma_r,\mathscr{M}_k;\{y_i,x_i,t_i,c_i\}_{i=1}^n)p(\beta_r,\gamma_r\mid\mathscr{M}_k)}{q^{(k)}_{\text{Laplace}}(\beta_r,\gamma_r\mid\mathscr{M}_k)}
\end{aligned}$$

where $\{\beta_r,\gamma_r\}_{r=1}^M$ is a sample from $q^{(k)}_{\text{Laplace}}$.

The accuracy of this approximation can be examined via examination of the "importance weights" $w_r = \frac{L_C(\beta_r,\gamma_r,\mathscr{M}_k;\{y_i,x_i,t_i,c_i\}_{i=1}^n)p(\beta_r,\gamma_r|\mathscr{M}_k)}{q^{(k)}_{\text{Laplace}}(\beta_r,\gamma_r|\mathscr{M}_k)}$. A popular diagnostic is to plot the normalized weights $w_r^{\text{Norm}} := \frac{w_r}{\sum_{t=1}^M w_t}$, multiplied by M, which should concentrate around 1,

as illustrated by Figure 4.9 in our importance sampling for the level 0 model $\mathscr{M}_1$. Another diagnostic is to calculate the "effective sample size", given by $\left(\sum_{r=1}^{M}(w_r^{\text{Norm}})^2\right)^{-1}$ (Kong *et al.*, 1994). In the approximation for $\mathscr{M}_1$, an importance sample of size $M = 500$ results in an effective sample size of 498, indicating the importance sampler is extremely efficient, and that the Laplace approximation provides a extremely good approximation of the posterior. We use this technique to calculate the marginal likelihood for each model, the results of which are displayed in Figure 4.10. The marginal likelihood under the level 0 model ranks fairly highly, with the split on the second quintile of the 6th covariate (months of prior ZDV) having the largest marginal likelihood of level 1 models, and the highest overall likelihood given by a level 2 model. Our results can be summarised by the posterior distribution over models. The posterior places 0.9 probability on the level 0 model and 0.02 probability on the model with subgroup $\{X_6 \leq 16\}$, that is, those with fewer than 16 months prior ZDV treatment. All other models have posterior probabilities less than 0.01. The posterior for γ within the level 0 model puts the majority of mass on negative values of γ, indicating a benefit to treatment of all. The posterior for γ within the $\{X_6 \leq 16\}$ model places the majority of mass on positive values of γ, indicating that treatment of this subgroup is associated with lower response values. As such, the corresponding treatment rule is to treat if $\{X_6 > 16\}$. This posterior represent a wealth of support for the "treat all" rule supporting the findings of the treatment trees approach.

As before, we consider a hypothetical scenario where treatment halves the quality of life. We conduct our analysis using the response $Y_i^{QALY} := Y \times 0.5^{T_i}$ in the same way. We use the same subgroup models, as well as the same prior for (β, γ). Figure 4.11 displays the marginal likelihoods under each model. It is immediately apparent that the level 0 model ranks lower in terms of marginal likelihood than in the non-adjusted analysis. Of the level 1 models, the two models with splits on the 2nd and 3rd quintiles of covariate 6 (months of prior ZDV) have much higher likelihoods than other models, ranking highly out of level

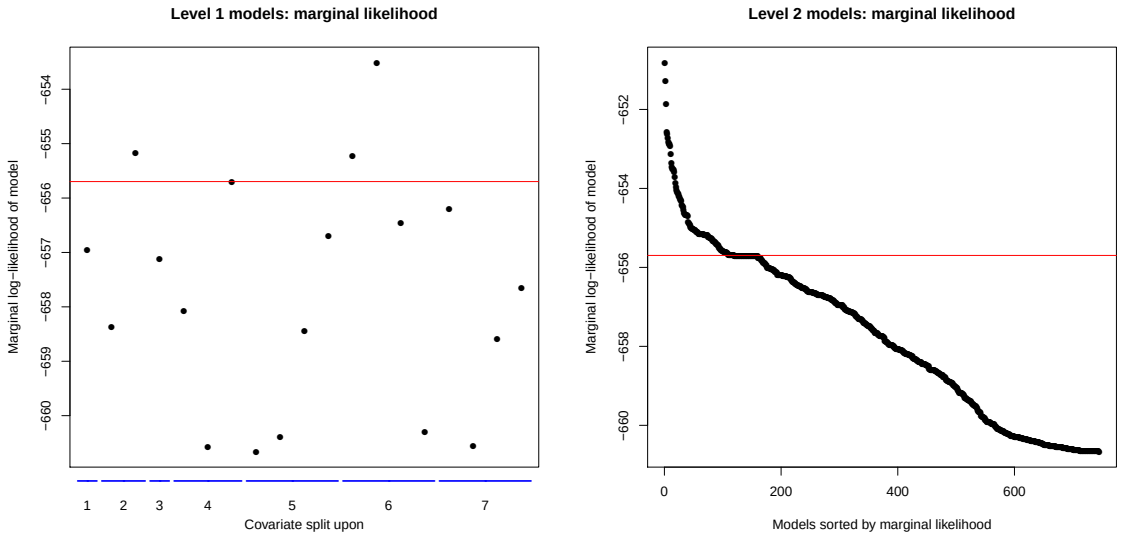


Fig. 4.10 Left, marginal log-likelihoods of level 1 models, separated by splitting covariate and ordered within covariate by split point. Right, the ordered marginal log-likelihoods of level 2 models. The red line denotes the marginal log-likelihood of the level 0 model.

2 models. The posterior distribution over models this time spreads mass among a number of potential models. The models with most posterior mass correspond to the following subgroups:

- $\mathscr{X}$ with probability 0.32
- $\{X_6 \leq 27\}$ with probability 0.21
- $\{X_6 \leq 16\}$ with probability 0.17
- $\{X_5 > 15, X_6 \leq 27\}$ with probability 0.06
- $\{X_5 \leq 98, X_6 > 16\}$ with probability 0.01.

All other models have posterior probability less than 0.01. The distribution of the sign of γ in each case dictates the treatment rule associated with each subgroup. The resulting treatment rules are given by:

- $D(X) = 1$

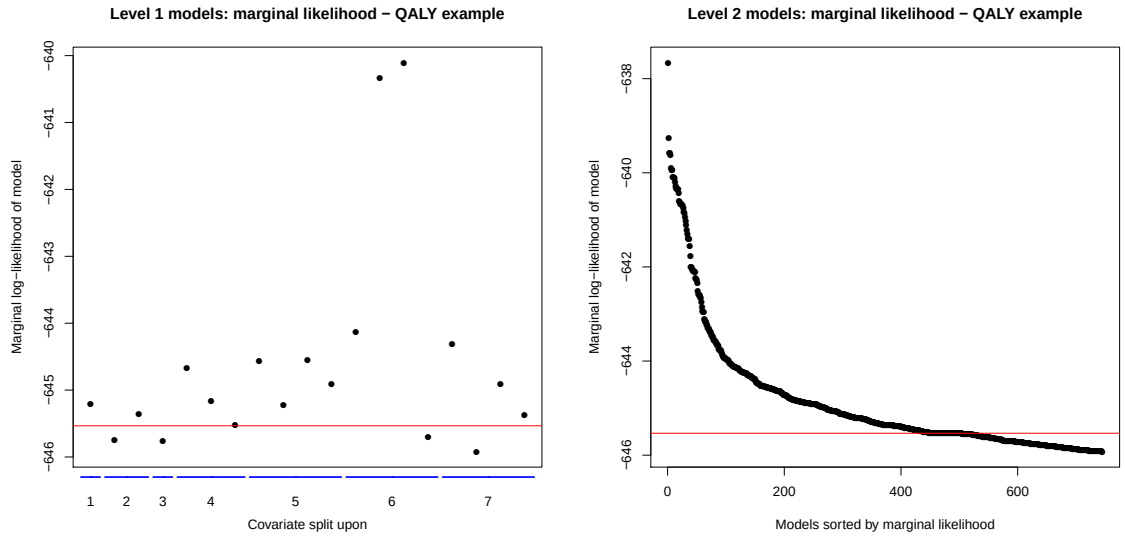


Fig. 4.11 Marginal log-likelihoods for the QALY example. Left, marginal log-likelihoods of level 1 models, separated by splitting covariate and ordered within covariate by split point. Right, the ordered marginal log-likelihoods of level 2 models. The red line denotes the marginal log-likelihood of the level 0 model.

- $D(X) = \mathbb{I}\{X_6 > 27\}$
- $D(X) = \mathbb{I}\{X_6 > 16\}$
- $D(X) = \mathbb{I}\{X_5 \leq 15, X_6 > 27\}$
- $D(X) = \mathbb{I}\{X_5 \leq 98, X_6 > 16\}$

We see that the "treat all" rule is the rule with highest posterior mass, but multiple models advocate similar treatment rules, mainly of just the subgroup $\{X_6 > 16\}$. This analysis of the QALY response results in similar findings to the treatment trees analysis, though reported in a different format. Both methods conclude to treat the majority of patients, despite the QALY adjustment, and identify the subgroup corresponding to individuals with longer periods of previous ZDV treatment as certainly benefiting from treatment.

4.6 Discussion

This chapter has developed two methodologies for subgroup analysis, one based on nonparametric tests and random forests, the other using the Bayesian nonparametric model of Chapter 3. The key defining feature of these approaches is the focus on the treatment rule $D(X)$, rather than the effect of treatment. The treatment trees methodology provides a scalable, model-free approach, which should be easy to replicate on account of the ease of use. The Bayesian approach circumvents modeling of the distributions of Y and X, rather, effectively modeling the distribution of the ranks of Y within the sample. The posterior distribution over models represents a highly informative report of beliefs, which automatically accounts for multiple testing through the prior. The two methodologies provide alternative approaches, the suitability of which will depend on the amount of prior information available.

Chapter 5

Discussion

This thesis has investigated approaches to various topics in regression that utilise the treatment of covariates as random variables. In particular, we have used the marginal distribution of the response as a way to incorporate prior information, ease prior specification and to improve computation. I believe the convenient standard treatment of covariates as fixed can lead to oversight of potential sources of data, prior information and model interpretation.

Chapter 2 introduced a joint prior for the parameters in the Bayesian linear model that constrains the marginal means and variance of the response. This provides a means for incorporating very large marginal datasets that would otherwise be computationally infeasible to include via imputation. The ability to incorporate marginal data is not immediately apparent within a framework where the covariates are treated as fixed, leading to the oversight of this form of analysis. The joint prior has a rich, interpretable structure through the fixed marginal variance constraint. In addition, this constraint automatically calibrates the prior. We quantitatively explored the behaviour of sparsity priors using our parameter transformation, where this calibration and structure might prove particularly useful. The constraint on the marginal variance of the response could provide a useful tool for the controlling of covariate effects in high-dimensional applications. The development of theoretical guarantees of sparsity induction in the transformed prior represents an interesting line of future work.

There are many potential extensions to other classes of models, such as generalised linear models and non-linear regression. The extent to which marginal data provides information is related to the form of the response. For example, in logistic regression, the maximum possible information available would be the population-wide incidence probability, which would enforce just one constraint within a similar methodology. A multinomial or ordinal response could potentially provide large amounts of marginal information. Another possible extension of the work in Chapter 2 is the investigation of asymptotic theory of variable and model selection using a marginally coherent approach when compared to a marginally incoherent approach. Since marginally incoherent priors specify different priors for the marginal variance, it is intuitive that one of these priors will be closer to reality than the other, introducing a bias. Therefore, we may expect gains by using a marginally coherent alternative, where this bias is eliminated. It would also be interesting to explore Lindley's paradox within our framework of marginally coherent priors, where "flat" priors would be placed on the untransformed parameters. The marginal constraint may provide a way to specify a flat prior over parameters that circumvents the associated bias towards the smaller model, since the transformation would bound this flat prior over the space of parameters of interest.

Chapter 3 introduced a Bayesian nonparametric regression model via a construction parameterised in terms of F_Y. In practice, this serves as a useful vehicle for prior information and external marginal data. The parameterisation leads to a convenient composite likelihood approximation that allows scalable computation, providing a method to utlise Bayesian nonparametric priors for density estimation within a regression framework. One limitation of the model in current form is the use of the regression function $\lambda(x) = \exp(X\beta)$. This limits flexibility of the covariate-response relationship, restricting any covariate effect to be monotone. We saw the detrimental effect this can have in the US Census application, where the flexibility of the covariate-response relationship in the random forests methodology

provides a superior fit of the conditional mean. This can be achieved within our model by using basis functions or tree structures in the definition of λ, similar to that used in the application of the model to subgroup analysis.

We explored how the model can be used to analyse truncated and censored survival data, using the composite likelihood; however, full posterior inference in this case is only possible if the posterior for F_Y can be calculated, given truncated and censored data. One natural approach to combat this is to parameterise the model in terms of the marginal hazard h_Y or marginal cumulative hazard H_Y, in conjunction with a Beta or Gamma process prior. In this setting, sampling schemes are available for posterior sampling, given truncated and censored data. The use of the composite likelihood would allow scalability of such a model to large survival datasets.

Chapter 4 utilised the treatment of covariates as random variables to test treatment rules $Y \mid D(X) = T$, where covariates are marginalised. This facilitated the use of simple two-sample tests for a treatment rule D versus its mirror rule D^c, providing a robust, scalable and model-free approach to subgroup analysis. We explored a Bayesian approach by using our regression model from Chapter 3. In this Bayesian approach, we define a model for a number of potential subgroups, defined using a subset of trees of depth 2. A true analogue of our testing trees approach would define each model using a forest, rather than a tree structure. To do this, one could look to adapt the Bayesian additive regression trees approach to subgroup analysis. It would also be interesting to conduct a full decision-theoretic analysis of subgroups using this approach, where the space of actions represents the space of treatment rules, and the loss function is related to the distribution of the response under such a treatment rule. A treatment rule is chosen to minimise the expected posterior loss, using our Bayesian model. This approach provides a full incorporation of the posterior into the choice of treatment rule, but represents a significant computational task. It would be interesting to explore how this could be implemented in a computationally efficient, scalable manner.

References

Altman, Douglas G. 2015. Clinical trials: Subgroup analyses in randomized trials - more rigour needed. *Nature Reviews Clinical Oncology*, **12**(9), 506–507.

Aslanidou, Helen, Dey, Dipak K., & Sinha, Debajyoti. 1998. Bayesian analysis of multivariate survival data using Monte Carlo methods. *Canadian Journal of Statistics*, **26**(1), 33–48.

Assmann, Susan F, Pocock, Stuart J, Enos, Laura E, & Kasten, Linda E. 2000. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, **355**(9209), 1064–1069.

Bergeron, Pierre-Jerome, Asgharian, Masoud, & Wolfson, David B. 2012. Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association*.

Bissiri, PG, Holmes, CC, & Walker, Stephen G. 2016. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Blackwell, David, & MacQueen, James B. 1973. Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, **1**(2), pp. 353–355.

Box, George EP, & Tiao, George C. 2011. *Bayesian inference in statistical analysis*. Vol. 40. John Wiley & Sons.

Breiman, Leo. 2001. Random forests. *Machine learning*, **45**(1), 5–32.

Breiman, Leo, Friedman, Jerome, Stone, Charles J, & Olshen, Richard A. 1984. *Classification and regression trees*. CRC press.

Broemeling, Lyle. 1984. *Bayesian Analysis of Linear Models*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis.

Brookes, Sara T, Whitely, Elise, Egger, Matthias, Smith, George Davey, Mulheran, Paul A, & Peters, Tim J. 2004. Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test. *Journal of clinical epidemiology*, **57**(3), 229–236.

Cai, Tianxi, Tian, Lu, Wong, Peggy H, & Wei, LJ. 2011. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, **12**(2), 270–282.

Carlin, Bradley P., Chaloner, Kathryn M., Louis, Thomas A., & Rhame, Frank S. 1995. *Elicitation, Monitoring, and Analysis for an AIDS Clinical Trial*. Springer New York. Pages 48–89.

Carvalho, Carlos M, Polson, Nicholas G, & Scott, James G. 2010. The horseshoe estimator for sparse signals. *Biometrika*, asq017.

Chatterjee, Nilanjan, Chen, Yi-Hau, Maas, Paige, & Carroll, Raymond J. 2016. Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information From External Big Data Sources. *Journal of the American Statistical Association*, **111**(513), 107–117.

Chipman, H.A., George, E.I., & McCulloch, R.E. 2010. BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, **4**(1), 266–298.

Chipman, Hugh, George, Edward I, McCulloch, Robert E, Clyde, Merlise, Foster, Dean P, & Stine, Robert A. 2001. The Practical Implementation of Bayesian Model Selection. *Lecture Notes-Monograph Series*, **38**, 65–134.

Cleveland, William S. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, **74**(368), 829–836.

Cohen, Jacob, Cohen, Patricia, West, Stephen G, & Aiken, Leona S. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Cox, D. R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(2), pp. 187–220.

Cox, D. R., & Reid, N. 2004. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, **91**(3), 729–737.

Cox, D.R., & Oakes, D. 1984. *Analysis of Survival Data*. Monographs on Statistics and Applied Probability. Taylor & Francis.

Craven, Peter, & Wahba, Grace. 1978. Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**(4), 377–403.

D. Hoff, Peter. 2007. Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.*, **1**(1), 265–283.

Davison, Anthony Christopher, & Hinkley, David Victor. 1997. *Bootstrap methods and their application*. Vol. 1. Cambridge university press.

De Iorio, Maria, Müller, Peter, Rosner, Gary L, & MacEachern, Steven N. 2004. An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, **99**(465), 205–215.

Denison, D.G.T., Holmes, C.C., Mallick, B.K., & Smith, A.F.M. 2002. *Bayesian methods for nonlinear classification and regression*. John Wiley & Sons.

Dunson, D.B., & Park, J.-H. 2008. Kernel stick-breaking processes. *Biometrika*, **95**(2), 307–323.

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

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

Ferguson, T.S. 1974. Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics*, **2**(4), pp. 615–629.

Foster, Jared C, Taylor, Jeremy MG, & Ruberg, Stephen J. 2011. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, **30**(24).

Friedman, Jerome, Hastie, Trevor, & Tibshirani, Robert. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin.

Gail, M., & Simon, R. 1985. Testing for Qualitative Interactions between Treatment Effects and Patient Subsets. *Biometrics*, **41**(2), 361–372.

Gasser, Theo, & Müller, Hans-Georg. 1984. Estimating Regression Functions and Their Derivatives by the Kernel Method. *Scandinavian Journal of Statistics*, **11**(3), pp. 171–185.

Gelman, Andrew. 2008. Objections to Bayesian statistics. *Bayesian Anal.*, **3**(3), 445–449.

Gelman, Andrew, Bois, Frederic, & Jiang, Jiming. 1996. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, **91**(436), 1400–1412.

Gelman, Andrew, Carlin, John B, Stern, Hal S, & Rubin, Donald B. 2014. *Bayesian data analysis*. Vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Ghosal, S., & Van der Vaart, A.W. 2013. *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, New York.

Gray-Davies, Tristan, Holmes, Chris C., & Caron, François. 2016. Scalable Bayesian nonparametric regression via a Plackett-Luce model for conditional ranks. *Electron. J. Statist.*, **10**(2), 1807–1828.

Griffin, J.E., & Steel, M.F.J. 2006. Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association*, **101**(473), 179–194.

Griffin, Jim E., & Brown, Philip J. 2010. Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.*, **5**(1), 171–188.

Hanson, T., & Johnson, W.O. 2002. Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, **97**(460).

Hanson, Timothy E. 2006. Inference for Mixtures of Finite Polya Tree Models. *Journal of the American Statistical Association*, **101**(476), 1548–1565.

Hjort, N.L. 1990. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 1259–1294.

Hjort, N.L., Holmes, C.C., Müller, P., & Walker, S.G. 2010. *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Ibrahim, J.G., Chen, M.H., & Sinha, D. 2001. *Bayesian Survival Analysis*. Springer Series in Statistics. Springer.

Ibrahim, Joseph G., Chen, Ming-Hui, & MacEachern, Steven N. 1999. Bayesian variable selection for proportional hazards models. *Canadian Journal of Statistics*, **27**(4), 701–717.

Ishwaran, Hemant, Kogalur, Udaya B, Blackstone, Eugene H, & Lauer, Michael S. 2008. Random survival forests. *The annals of applied statistics*, 841–860.

Jara, A. 2007. Applied Bayesian Non- and Semi-parametric Inference Using DPpackage. *R News*, **7**(3), 17–26.

Jara, A., Hanson, T., Quintana, F., Müller, P., & Rosner, G. 2011. DPpackage: Bayesian Semi- and Nonparametric Modeling in R. *Journal of Statistical Software*, **40**(5), 1–30.

Jordan, M. 2011. What are the open problems in Bayesian statistics. *The ISBA Bulletin*, **18**(1), 568.

Kalbfleisch, John D. 1978. Non-Parametric Bayesian Analysis of Survival Time Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **40**(2), pp. 214–221.

Kim, Y. 2006. The Bernstein–von Mises theorem for the proportional hazard model. *The Annals of Statistics*, **34**(4), 1678–1700.

Klein, John P, & Moeschberger, Melvin L. 2005. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kong, Augustine, Liu, Jun S, & Wong, Wing Hung. 1994. Sequential imputations and Bayesian missing data problems. *Journal of the American statistical association*, **89**(425), 278–288.

Lavine, M. 1992. Some Aspects of Polya Tree Distributions for Statistical Modelling. *The Annals of Statistics*, **20**(3), pp. 1222–1235.

Lavine, M. 1994. More Aspects of Polya Tree Distributions for Statistical Modelling. *The Annals of Statistics*, **22**(3), pp. 1161–1176.

Lee, Joseph J, & Rubin, Donald B. 2015. Valid randomization-based p-values for partially post hoc subgroup analyses. *Statistics in medicine*, **34**(24), 3214–3222.

Lindsay, B. G. 1988. Composite likelihood methods. *Contemporary Mathematics*, **80**(1), 221–39.

Lo, A.Y. 1984. On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, **12**(1), pp. 351–357.

Luce, R..D. 1959. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley and sons.

MacEachern, S.N. 1999. Dependent Nonparametric Processes. *In: Proceedings of the Bayesian Statistical Science Section*. American Statistical Association.

MacEachern, S.N., & Müller, P. 1998. Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, **7**(2), 223–238.

Mann, H. B., & Whitney, D. R. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Statist.*, **18**(1), 50–60.

Matsouaka, Roland A, Li, Junlong, & Cai, Tianxi. 2014. Evaluating marker-guided treatment selection strategies. *Biometrics*, **70**(3), 489–499.

Matthews, John NS. 2006. *Introduction to randomized controlled clinical trials*. CRC Press.

Mauldin, R. Daniel, Sudderth, William D., & Williams, S. C. 1992. Polya Trees and Random Distributions. *The Annals of Statistics*, **20**(3), pp. 1203–1221.

Mitchell, Toby J, & Beauchamp, John J. 1988. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, **83**(404), 1023–1032.

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

Müller, P., Erkanli, A., & West, M. 1996. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**(1), 67–79.

Murray, Jared S, Dunson, David B, Carin, Lawrence, & Lucas, Joseph E. 2013. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, **108**(502), 656–665.

Neal, R.M. 2000. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **9**(2), 249–265.

Nelsen, Roger B. 2013. *An introduction to copulas*. Vol. 139. Springer Science & Business Media.

Park, Trevor, & Casella, George. 2008. The Bayesian Lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.

Pauli, F., Racugno, W., & Ventura, L. 2011. Bayesian composite marginal likelihoods. *Statistica Sinica*, **21**(1), 149.

Plackett, R. L. 1975. The Analysis of Permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **24**(2), pp. 193–202.

Polson, Nicholas G, & Scott, James G. 2010. *Shrink globally, act locally: sparse Bayesian regularization and prediction*. Oxford University Press.

Qian, Min, & Murphy, Susan A. 2011. Performance guarantees for individualized treatment rules. *Ann. Statist.*, **39**(2), 1180–1210.

Raftery, Adrian, Madigan, David, & Volinsky, Chris T. 1995. Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance. *Pages 323–349 of: Bayesian Statistics 5*.

Rasmussen, C.E. 2006. *Gaussian processes for machine learning*. MIT Press.

Ribatet, M., Cooley, D., & Davison, A. C. 2012. Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, **22**, 813–845.

Ripley, Brian D. 2009. *Stochastic simulation*. Vol. 316. John Wiley & Sons.

Ritter, Christian, & Tanner, Martin A. 1992. Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, **87**(419), 861–868.

Robert, Christian. 2007. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Rothwell, Peter M. 2005. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, **365**(9454), 176–186.

Ruberg, Stephen J, Chen, Lei, & Wang, Yanping. 2010. The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. *Clinical trials*.

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

Song, Peter Xue-Kun. 2000. Multivariate Dispersion Models Generated from Gaussian Copula. *Scandinavian Journal of Statistics*, **27**(2), 305–320.

Su, Xiaogang, Tsai, Chih-Ling, Wang, Hansheng, Nickerson, David M, & Li, Bogong. 2009. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, **10**(Feb), 141–158.

Sun, Xin, Briel, Matthias, Busse, Jason W, You, John J, Akl, Elie A, Mejza, Filip, Bala, Malgorzata M, Bassler, Dirk, Mertz, Dominik, Diaz-Granados, Natalia, *et al.* 2012. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *Bmj*, **344**, e1553.

Thas, Olivier, Neve, Jan De, Clement, Lieven, & Ottoy, Jean-Pierre. 2012. Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(4), 623–671.

Tibshirani, Robert. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

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

Van't Veer, Laura J, & Bernards, René. 2008. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, **452**(7187), 564–570.

Varin, C., Reid, N., & Firth, D. 2011. An overview of composite likelihood methods. *Statistica Sinica*, **21**(1), 5–42.

Vidakovic, Brani. 2009. *Statistical modeling by wavelets*. Vol. 503. John Wiley & Sons.

Wade, S., Dunson, D.B., Petrone, S., & Trippa, L. 2014. Improving Prediction from Dirichlet Process Mixtures via Enrichment. *Journal of Machine Learning Research*, **15**, 1041–1071.

Walker, S., & Mallick, B.K. 1999. A Bayesian Semiparametric Accelerated Failure Time Model. *Biometrics*, **55**(2), 477–483.

Wasserman, L. 2006. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York.

Ying, Z., Jung, S. H., & Wei, L. J. 1995. Survival Analysis with Median Regression Models. *Journal of the American Statistical Association*, **90**(429), pp. 178–184.

Zeckhauser, R, & Shepard, D. 1976. Where now for saving lives? *Law and contemporary problems*, **40**(4), 5–45.

Zellner, A. 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Pages 233–243 of:* Goel, P., & Zellner, A. (eds), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Amsterdam: North-Holland Publishing Co.

Zhang, Sheng, Liang, Fei, Li, Wenfeng, & Hu, Xichun. 2015. Subgroup Analyses in Reporting of Phase III Clinical Trials in Solid Tumors. *Journal of Clinical Oncology*, **33**(15), 1697–1702.

Zhao, Yingqi, Zeng, Donglin, Rush, A John, & Kosorok, Michael R. 2012. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, **107**(499), 1106–1118.

Zhou, Xin, Mayer-Hamblett, Nicole, Khan, Umer, & Kosorok, Michael R. 2015. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 00–00.

Appendix A

Posterior predication for Chapter 3

Here we provide additional information on the simulation from the predictive distribution for the Bayesian nonparametric model in Chapter 3, which depends on the choice of the prior for F_Y.

It might be the case that n is so large that simply using the empirical CDF is a reasonable approximation for an empirical Bayes approach. In this case, the inversion of the cdf becomes trivial, and MCMC is only required for the β posterior sample.

- Simulate $\beta^{(j)}$ from the pseudo posterior
- $Z' \sim \text{Exp}(\lambda_{\beta^{(j)}}(x'))$
- Calculate $u^{(j)} := 1 - \frac{1}{n}\sum_{i=1}^{n} e^{-Z'\lambda_{\beta^{(j)}}(x_i)}$
- Set $Y'^{(j)} = y_{(\lceil nu^{(j)} \rceil)}$

An alternative approach might be to use a Bayesian Bootstrap on F_Y. This works out similarly to using the empirical CDF, but we must simulate the Dirichlet weights for each of the atoms. The sampling scheme becomes:

- Simulate $\beta^{(j)}$ from the pseudo posterior

- Sample $Z' \sim \text{Exp}(\lambda_{\beta^{(j)}}(x'))$
- Calculate $u^{(j)} := 1 - \frac{1}{n}\sum_{i=1}^{n} e^{-Z'\lambda_{\beta^{(j)}}(x_i)}$
- Simulate $(W_1, W_2, \ldots, W_n) \sim \text{Dirichlet}(1, 1, \ldots, 1)$
- Set $Y'^{(j)} = y_{(\min\{k:\sum_{i=1}^{k} W_i \geq u^{(j)}\})}$

For Pólya tree priors, under our composite likelihood scheme, the posterior for F_Y is also a Pólya tree, due to conjugacy of the Pólya tree prior. Simulation then proceeds as follows:

- Simulate $\beta^{(j)}$ from the pseudo posterior
- $Z' \sim \text{Exp}(\lambda_{\beta^{(j)}}(x'))$
- Calculate $u^{(j)} := 1 - \frac{1}{n}\sum_{i=1}^{n} e^{-Z'\lambda_{\beta^{(j)}}(x_i)}$

Then all we need to calculate is $F_Y^{-1(j)}(u^{(j)})$. Pólya trees make this easy. A sample from a Pólya tree distribution is a random probability measure, which is constructed by assigning random masses to each branch in a partition tree of the space. So, given the first partition point in the tree, we can simply generate the random mass either side of this point, and trivially deduce which branch $F_Y^{-1(j)}(u^{(j)})$ lies in. We repeat this process down the tree until we reach the truncation point often used when using Pólya trees.

Formally, given a Pólya tree truncated at level M, let $a_0 = 0, B_0 = 1, \varepsilon_0 = \emptyset$ and for k from 1 to M:

- $\theta_{\varepsilon_k 0} \sim \text{Beta}(\alpha_{\varepsilon_k 0}, \alpha_{\varepsilon_k 1})$
- if $u \in [a_k, \theta_k(b_{k-1} - a_{k-1}) + a_{k-1}]$
  - $\varepsilon_k = \varepsilon_{k-1} 0$
  - $a_k = a_{k-1}$

  - $b_k = a_{k-1} + \theta_k(b_{k-1} - a_{k-1})$

- Otherwise

  - $\varepsilon_k = \varepsilon_{k-1} 1$
  - $a_k = a_{k-1} + \theta_k(b_{k-1} - a_{k-1})$
  - $b_k = a_{k-1}$

When applying Dirichlet Process Mixture models the difficulty here becomes the inversion of F_Y, since this has no closed form. A simple Monte Carlo approximation can be used to approximate this inversion for each posterior sample.

- Simulate $\beta^{(j)}$ from the pseudo posterior
- Sample $Z' \sim \text{Exp}(\lambda_{\beta^{(j)}}(x'))$
- Simulate $F_Y^{(j)}$ from the pseudo posterior
- Simulate $Y_k'^{(j)} \sim F_Y^{(j)}$ for $k = 1, \ldots, N$
- Calculate $u^{(j)} := 1 - \frac{1}{n} \sum_{i=1}^{n} e^{-Z' \lambda_{\beta^{(j)}}(x_i)}$
- Set $Y'^{(j)} = Y'^{(j)}_{(\lceil N u^{(j)} \rceil)}$