NOTE TO USERS

The original manuscript received by UMI contains broken, slanted and or light print. All efforts were made to acquire the highest quality manuscript from the author or school. Pages were microfilmed as received.

This reproduction is the best copy available

UMI
ANALYSIS OF DENSITIES

by

George Andrew Tomlinson

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Public Health Sciences
University of Toronto

© Copyright by George Andrew Tomlinson, 1998
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.
Abstract

Analysis of Densities. Ph.D. 1998. George Andrew Tomlinson, Department of Public Health Sciences, University of Toronto

A fully Bayesian method is developed for modelling the distribution of probability density functions across a population. Probability densities for individual subjects are modelled as samples from a “parent” distribution. The probability density functions for these subjects are modelled as mixtures of Dirichlet processes. A common baseline prior distribution for the mixing distribution is assumed. The similarity between density estimates for different subjects is controlled by the precision parameters of the subject-specific Dirichlet processes. The common baseline prior is also estimated using a mixture-of-Dirichlet-processes model. For this second Dirichlet process, the precision parameter controls how much detail from the individual density estimates is captured in the estimate of the parent distribution. Most of the effort is computational involving non-conjugate priors and a large Gibbs-sampling scheme to obtain samples of predictive densities from their posterior distribution. To measure how far any density deviates from the average, the Hellinger distance is used. The presentation of results is graphical: distances based on predictive densities sampled from their posterior distribution are used to highlight key observations and other structure in the data being modeled.
To R. L. Cohen.
Contents

Abstract .......................... ii

Contents .......................... iv

List of Tables .................... viii

List of Figures .................... ix

0 Notation ......................... 1

1 Introduction ..................... 2
  1.1 Motivation ........................ 2
  1.2 Outline of Thesis ................ 3

2 A First Approach to the Problem 5
  2.1 The Data ......................... 5
  2.2 The Analysis ...................... 7
  2.3 Fitting the Density Estimates .... 10
  2.4 Results and Further Analysis ... 10
  2.5 Discussion ....................... 19

3 Background ...................... 23
  3.1 The Dirichlet Process Prior .... 23
  3.2 The Dirichlet Process Posterior 30
3.3 The Dirichlet Process in Density Estimation .......................... 30
3.4 Gibbs Sampling With The Dirichlet Prior ............................ 33
3.5 Learning About The Dirichlet
   Precision Parameter ........................................ 35
   3.5.1 Parametric Prior for $\alpha_0$ ............................ 36
   3.5.2 Discrete Prior For $\alpha_0$ ................................ 37
3.6 Summary ...................................................... 37

4 Model for Multiple Samples ............................................. 39
  4.1 Structure of the Data ........................................... 39
  4.2 The Need for a Flexible $G_0$ ................................ 41
  4.3 Posterior Analysis of $G_0$ ................................... 44
  4.4 Preview of Model Fitting ..................................... 47
  4.5 Summary ...................................................... 48

5 Sampling the Distribution of $\pi_{ij}|Y_{ij}$ ......................... 49
  5.1 Computing the Marginal Density $f_0(Y)$ .......................... 51
  5.2 Sampling From the Baseline
   Posterior Distribution $G_0(\pi_{ij}|Y_{ij})$ ...................... 53
  5.3 Which Part of $G_0(\pi_{ij}|Y_{ij})$ to Sample .................... 56
  5.4 Summary ...................................................... 57

6 Updating the Distribution $G_0$ ....................................... 58
  6.1 Hierarchical Model for $\pi^*$ ................................ 58
  6.2 Sampling Distribution of $\tau_r|\pi^*_r$ ........................... 61
  6.3 Computing the Marginal Density $f_{00}(\pi^*)$ .................... 62
  6.4 Sampling from the Baseline
   Posterior Distribution $G_{00}(\tau_r|\pi^*_r)$ .................... 62
   6.4.1 Sampling $w$ ............................................. 63
  6.5 Remi\xing $T$ .................................................. 65
  6.6 Summary ...................................................... 66
7 Additional Learning About Hyperparameters

7.1 Learning About $\alpha_0$ and $\alpha_{00}$ ............................................. 68
7.2 Learning About $\Delta_2$ ................................................................. 69
7.3 Learning About $G_{00}$: $\Delta_1$ and $m_0$ ........................................... 70
  7.3.1 Learning About $\Delta_1$ ............................................................ 70
  7.3.2 Learning About $m_0$ ............................................................... 70
  7.3.3 Changes to Candidate $\tau$ ....................................................... 71

8 Inference With the Model

8.1 Predictive Distributions ............................................................... 72
8.2 Measuring Dissimilarity ............................................................... 75
8.3 Simulating Data From the Posterior .............................................. 77

9 Miscellaneous Topics

9.1 Overview of Algorithm .................................................................... 80
  9.1.1 Hierarchical Model ................................................................. 80
  9.1.2 Sampling $\pi_{ij}|Y_{ij}$ ............................................................... 81
  9.1.3 Sampling $\tau_r|\pi_r^*$ ............................................................. 81
  9.1.4 Entire Algorithm ....................................................................... 82
9.2 Choice of Prior Distributions ......................................................... 83
  9.2.1 Dirichlet Precision Parameters .................................................. 83
  9.2.2 Parametric Priors ..................................................................... 84
9.3 Computational Issues ....................................................................... 88

10 Simulated Application: Summarizing a Set of Samples

10.1 Data Generation ............................................................................. 97
10.2 Application of Model ..................................................................... 99
10.3 Results .......................................................................................... 102

11 A Real Application

11.1 Description of Data ..................................................................... 121
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.2</td>
<td>Application of Model</td>
<td>122</td>
</tr>
<tr>
<td>11.3</td>
<td>Results</td>
<td>123</td>
</tr>
<tr>
<td>11.4</td>
<td>Special Cases</td>
<td>130</td>
</tr>
<tr>
<td>12</td>
<td>Conclusions</td>
<td>137</td>
</tr>
<tr>
<td>12.1</td>
<td>Summary of Findings</td>
<td>137</td>
</tr>
<tr>
<td>12.2</td>
<td>Future Work</td>
<td>138</td>
</tr>
<tr>
<td>A</td>
<td>Gaussian Quadrature</td>
<td>141</td>
</tr>
<tr>
<td>B</td>
<td>Sampling From a Log-Concave Density</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>B.1 Details of the Algorithm for Log-Concave Distributions</td>
<td>143</td>
</tr>
<tr>
<td>C</td>
<td>Tabled Values of $P_n(k</td>
<td>\alpha)$</td>
</tr>
<tr>
<td>D</td>
<td>Computing $f_0(Y)$ and $f_{00}(\pi^*)$</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>D.1 Computing $f_0(Y)$</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>D.2 Computing $f_{00}(\pi^*)$</td>
<td>152</td>
</tr>
<tr>
<td>E</td>
<td>Sampling from $G_0(\pi</td>
<td>Y)$ and $G_{00}(\tau</td>
</tr>
<tr>
<td></td>
<td>E.1 Sampling from $G_0(\pi</td>
<td>Y)$</td>
</tr>
<tr>
<td></td>
<td>E.2 Sampling from $G_{00}(\tau</td>
<td>\pi^*)$</td>
</tr>
<tr>
<td>F</td>
<td>Sampling From Remixing Density of $\tau^*$</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>F.1 Gibbs Sub-Sampler</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>F.2 The Generalized Inverse Gaussian Distribution</td>
<td>160</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>162</td>
</tr>
</tbody>
</table>
List of Tables

3.1 General formulation of density estimation and Escobar and West implementation ........................................ 31
4.1 Analogy between estimation of density of $Y_{ij}$ and density of $\pi_r^*$ ...................................................... 46
10.1 Sets of priors used in analyses of simulated data ..................................................................................... 100
C.1 Probability distribution of number of distinct components for $n = 25$
and various $\alpha$ ............................................................................................................................................. 147
C.2 Probability distribution of number of distinct components for $n = 50$
and various $\alpha$ ............................................................................................................................................. 148
List of Figures

2.1 Histograms of Reaction Time Data .......................... 6
2.2 Plots of All Kernel Density Estimates and Extreme Functions .......................... 11
2.3 Ordered Array of Plots of Kernel Density Estimates .......................... 12
2.4 Hellinger Distances Based On Kernel Density Estimates .......................... 14
2.5 Null Hypothesis Bootstrap Distribution of Ordered Hellinger Distances .......................... 16
2.6 Boxplots of Bootstrapped Within-Subject Hellinger Distances .......................... 17
2.7 Two-Step Bootstrap Distribution of Hellinger Distance .......................... 20

3.1 $P(K = k|\alpha_0, n = 25)$ .......................... 27
3.2 $P(K = k|\alpha_0, n = 50)$ .......................... 28

4.1 Small $\alpha_0$ in $G_i \sim \mathcal{D}(G_0, \alpha_0)$ .......................... 42
4.2 Large $\alpha_0$ in $G_i \sim \mathcal{D}(G_0, \alpha_0)$ .......................... 43

6.1 The Joint Density $f(v, w)$ .......................... 60

9.1 Understanding the Hyperparameters: $\Delta_2 > 1$ .......................... 86
9.2 Understanding the Hyperparameters: $\Delta_2 < 1$ .......................... 87
9.3 Normal-Gamma Mixture over Samples From $F: \alpha = 2, \beta = 2, \Delta_2 = 0.5$ .......................... 89
9.4 Normal-Gamma Mixture over Samples From $F: \alpha = 2, \beta = 2, \Delta_2 = 1$ .......................... 90
9.5 Normal-Gamma Mixture over Samples From $F: \alpha = 4, \beta = 4, \Delta_2 = 1$ .......................... 91

10.1 Sample PDFs From the Model Used to Generate Data .......................... 98
10.2 Set 1: Plots of Sampled $f_{obs}^B$ and $f_i^B$ .......................... 104
10.3 Set 1: Posterior Samples of Selected Hyperparameters .......................... 105
10.4 Set 2: Plots of Sampled $f^B_{\text{obs}}$ and $f^B_i$ ................................. 107
10.5 Set 1: Posterior ECDF of $\alpha_0$ for Common (solid) and Independent (dashed) Priors .................................................. 108
10.6 Set 4: Plots of Sampled $f^B_{\text{obs}}$ and $f^B_i$ .................................. 109
10.7 Posterior Samples of $\Delta_1$ and $\Delta_2$ from Sets 2 and 3 ................. 111
10.8 Posterior Samples of Number of Distinct $\tau$ .................................. 112
10.9 Set 3: Plots of Sampled $f^B_{\text{obs}}$ and $f^B_i$ .................................. 113
10.10 Sets 2 and 3: Selected Generated PDFs ...................................... 114
10.11 Sets 2 and 3: QQ-plot of $n^*$ ...................................................... 116
10.12 Sampled Cluster Centres and $\pi^*$ ............................................. 117
10.13 Sampled Cluster Centres and $\pi^*$ ............................................. 118
10.14 Sampled Cluster Centres and $\pi^*$ ............................................. 119
10.15 Sampled Cluster Centres and $\pi^*$ ............................................. 120
11.1 Trace of Number of Distinct Elements in $\mathbf{T}$ ............................... 124
11.2 Trace of Number of Distinct Elements in $\mathbf{U}^*$ ............................. 125
11.3 PDFs for Reaction Time Data ..................................................... 126
11.4 Empirical Cumulative Distribution Function for the Hellinger Distances $D(f^B_i)$ ........................................................... 127
11.5 Boxplots of Hellinger Distances by Group for Reaction Time Data ....... 129
11.6 Mean of $\alpha^I_0$ by Group for Reaction Time Data ............................ 131
11.7 Reaction Time Data: Sampled Cluster centres (a) ............................ 132
11.8 Reaction Time Data: Sampled Cluster centres (b) ............................ 133
11.9 Tiny $\alpha_0$ in Model for Reaction Time Data ................................... 135
11.10 Huge $\alpha_0$ in Model for Reaction Time Data ............................... 136
Chapter 0

Notation

\[ \Gamma(a) \] refers to the gamma function evaluated at \( a \)

\[ \text{Gamma}(x|a, b) \] refers to the gamma density function of the random variable \( x \).

\[ \text{Gamma}(x|a, b) = \frac{b^a e^{-bx} x^{a-1}}{\Gamma(a)} \]

\[ N(x|\mu, 1/v) \] refers to the normal density of the random variable \( x \).

\[ N(x|\mu, 1/v) = \sqrt{\frac{v}{2\pi}} \exp\left[-\frac{v}{2}(x - \mu)^2\right] \]

Sometimes, this will be written as \( N(x|\pi) \), where \( \pi \) is the vector \((\mu, v)\).

\[ M(\pi|\tau) \] refers to the normal-gamma density function of the random variable \( \pi = (\mu, v) \), conditional on \( \tau = (m, w) \) with known hyperparameters \( \alpha, \beta \) and \( \Delta \).

\[ M(\pi|\tau) = N(\mu|m, 1/(\Delta w)) \Gamma(\nu|\alpha, \beta/w) \]

\[ D(F, \alpha) \] refers to a Dirichlet process with expectation \( F \) and precision \( \alpha \).

\[ \delta_{x_0}(x) \] refers to a degenerate distribution (point mass) at \( x = x_0 \)

\[ \text{Beta}(x|r, s) \] refers to the beta density function of the random variable \( x \).

\[ \text{Beta}(x|r, s) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} x^{r-1}(1 - x)^{s-1} \]

\[ F(x) \] refers to the cumulative distribution function of \( x \)

\[ f(x) \] refers to the probability density function of \( x \).
Chapter 1

Introduction

1.1 Motivation

When the data collected on an individual is an unordered and independent set of measurements of a single type (e.g., sizes of a sample of white blood cells, reaction times to various stimuli), summarizing the data with a single number or pair of numbers (e.g., mean and standard deviation) may obscure some important features of the data, unless its distribution is close to some standard distribution. In the case of white blood cell sizes, for example, the distribution of sizes is expected to be trimodal and the relative proportions of the data that are centred at the three modes, as well as the locations of the modes are both important indicators of health. For a single set of measurements on one individual, either a histogram or a smooth density estimate is a good method of presenting the important features of the data. This thesis deals with the situation where a number of samples of the above type are collected on different individuals. The data on an individual can be thought of as a histogram or density estimate. In general, not all individuals will share the same true underlying density: the densities will vary from person to person.

For this type of data, there are no standard methods of analysis or presentation. This thesis has two main goals:
1. To develop a method of estimating a typical probability density function - a population average - and some indication of how individuals' density functions vary around this average. This will be purely descriptive: an analogy is the calculation of the mean and standard deviation of a simple univariate dataset. Except here, the results will be presented graphically.

2. To use the knowledge acquired in (1) to classify individuals as to how typical or atypical they are. This is a form of outlier detection — the univariate analog is the z-score. The goal here is inferential and the result will be a measurement of how far any particular subject is from the average of all subjects.

Rather than assuming any particular distributional form, the approach taken is quite general. As a result, most of the effort in this thesis is focused not directly on the above goals, but on the computational methods needed for calculating the quantities needed in (1) and (2). In fact, while this was not initially a primary goal, the development of the algorithm is the main contribution of this thesis.

1.2 Outline of Thesis

Chapter 2 presents a straightforward analysis of a dataset where 30 response times are measured on each of 17 individuals and interest focuses on the distribution of times within an individual. The weaknesses of this analysis motivate the development of the methods presented in the remainder of the thesis. Familiarity with the background necessary to understanding these methods is confined to a segment of the Bayesian community, so Chapter 3 gives a brief summary of these methods, including the Dirichlet process prior and its use in density estimation. Chapter 4 describes the basic hierarchical model used in the present multiple sample setting, and interprets the meanings of the components. Chapters 5 and 6 contain descriptions of the algorithms used to obtain posterior distributions and the methods used for key computations required by the algorithms. Chapter 7 extends the basic algorithm to
obtain information about key parameters of the prior distributions. Chapter 8 discusses inference from the model, including methods for sampling from the posterior distributions, for measuring dissimilarity between samples, and for presenting results. Chapter 9 gives an overview of the model and some issues related to computation. Chapter 10 presents results from using this algorithm to model data simulated from a mixture of normals where the mixture components vary from subject to subject. Comparisons are made between features of the true model from which the data was simulated and features of data simulated from the posterior distributions as computed by the algorithm. Chapter 11 presents results from using the algorithm on the same data analyzed in Chapter 2. Chapter 12 discusses some open questions related to the methods presented in the thesis and makes suggestions for further work.
Chapter 2

A First Approach to the Problem

This chapter will examine a dataset where the observation on each individual can be thought of as a set of independent measurements on a single feature. A straightforward approach to an analysis will be presented that addresses the two goals laid out in the introduction — (1) summarization and (2) ranking according to typicality. The strengths and weaknesses of this approach will be discussed and this discussion will highlight the characteristics of any good approach to analyzing this sort of data.

2.1 The Data

This is the data first analyzed by Belin and Rubin [3]. Reaction times were measured 30 times on each of 17 subjects — 6 schizophrenics and 11 non-schizophrenics. The data are highly skewed and analysis of the data is made easier if the natural logarithms of the data are used. More information about the data set is found in chapter 11. Histograms of the logarithms of the data are shown in figure 2.1.

The subjects will be indexed by \( i = 1, \ldots, n \) and the observations within subject \( i \) will be indexed by \( j = 1, \ldots, n_i \). The data for one subject is then \( y_i = \{y_{i1}, \ldots, y_{in_i}\} \). Here, \( n = 17 \) and \( n_i = 30 \) for all \( i \).
Figure 2.1: Histograms of Reaction Time Data
2.2 The Analysis

It seems natural to assume that the data for a particular subject come from a probability density function unique to that subject. This function is sampled from some parent distribution and might be thought of as the population average function plus some "noise". The "noise" is the effect of genetic and environmental factors unique to that subject. Refer to the probability density function giving rise to the data for subject \( i \) as \( f_i \). Refer to the parent or average probability density function as \( f_0 \). To continue in this line of thought, one approach to analyzing this data is:

1. Estimate the probability density function for each subject. \( \hat{f}_i(y) \).
2. Estimate the average probability density function \( \hat{f}_0(y) \).
3. Calculate the deviation \( D(\hat{f}_0, \hat{f}_i) \) of each individual from the average.
4. Assuming that these deviations capture important variation from subject to subject, use them in devising summaries of the data and for making inferences.

For steps (1) and (2) a method of computing an estimate of a density function is needed. After the histogram, the simplest density estimate to compute is the kernel density estimate [24]. With data \( \{y_1, \ldots, y_n\} \) and a given window width \( h \), this is computed as

\[
\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{y - y_i}{h} \right).
\] (2.1)

The standard normal density function is commonly used as the function \( K(\cdot) \). For step (1), the estimate for subject \( i \) using window width \( h \) is:

\[
\hat{f}_i(y) = \frac{1}{n_i h} \sum_{j=1}^{n_i} K \left( \frac{y - y_{ij}}{h} \right).
\] (2.2)

There are at least two potential estimators for \( f_0 \). One estimator for \( f_0 \) is the average of the estimated densities \( \hat{f}_i \). Call this \( \hat{f}_0^1 \).

\[
\hat{f}_0^1(y) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_i(y)
\] (2.3)
A different estimator is formed by estimating the density of the combined data \( \{y_{11}, \ldots, y_{m_1}, y_{21}, \ldots, y_{m_2}, \ldots, y_{n_1}, \ldots, y_{m_n}\} \). Call this \( \hat{f}_0^2 \):

\[
\hat{f}_0^2(y) = \frac{1}{\sum_{i=1}^{m_n} n_i h_0} \sum_{i=1}^{m_n} \sum_{j=1}^{n_i} K\left(\frac{y - y_{ij}}{h_0}\right).
\]  

(2.4)

The notation \( h_0 \) was used for the window width to emphasize that a different (likely smaller) window width may be used for \( \hat{f}_0^2 \). If \( h_0 = h \), the estimator \( \hat{f}_0^2 \) is equivalent to the weighted average of the individual densities:

\[
\hat{f}_0^2 = \frac{1}{\sum_{i=1}^{m_n} n_i} \sum_{i=1}^{m_n} n_i \hat{f}_i.
\]

If \( h_0 = h \) and all the sample sizes \( n_i \) are equal, then the two estimators (2.3) and (2.4) give the same value. In the remainder of this chapter, equation (2.4) will be used to estimate \( f_0 \).

For step (3), many measures of the distance between two density functions exist, depending on what type of differences it is important to measure. Since the actual distance measure used is not important for understanding the ideas involved, and the Hellinger distance is used later in this thesis, it will also be used here (see section 8.2, for the motivation for this distance measure). The definition of the Hellinger distance between two density functions \( p \) and \( q \), as used in this thesis, is

\[
D(p, q) = \left[ \int \left( \sqrt{p(y)} - \sqrt{q(y)} \right)^2 p(y) \, dy \right]^{\frac{1}{2}}. 
\]  

(2.5)

Notice that \( D(p, q) \) might not be equal to \( D(q, p) \). Refer to the distance between \( \hat{f}_i \) and \( \hat{f}_0 \) as \( \tilde{D}_i \).

\[
\tilde{D}_i = D(\hat{f}_0, \hat{f}_i)
\]

Step (4) will be expanded upon here. If the individual \( \hat{f}_i \) are similar to \( \hat{f}_0 \), then \( \hat{f}_0 \) may be a reasonable summary of the entire dataset. If there is substantial variation from subject to subject, the average \( \hat{f}_0 \) may not resemble any particular \( \hat{f}_i \). Thus, \( \hat{f}_0 \) itself may not be a particularly good summary of the set of \( \hat{f}_i \). For example, consider \( y_{ij} | \mu_i \sim N(y_{ij} | \mu_i, 1), j = 1, \ldots, n_i \) and \( \mu_i \sim N(\mu_i | 0, 1/w), i = 1, \ldots, n \) where \( w \) (a
precision) is small, so that the $\mu_i$ are very spread out. The true average density $f_0$ will be a function with approximately $n$ standard normal density “bumps” spread along the real line. This would in the least be a misleading summary of the data.

On a computer, one effective method of summarizing density functions on a number of subjects is to draw a single plot, then add and remove the estimated densities one by one. The resulting animation gives a clear picture of the variation over the population. The closest to this approach on a single printed plot is to superimpose the estimated densities; however, for anything more than a small number of subjects, the result often is a thick band of lines surrounding the average with no fine detail visible.

One approach to presenting variability in functions has been to use principal components analysis of the functions [16]. With this method, a set of functions evaluated at common data points is treated as a set of multivariate observations and subjected to a principal components analysis. Functions are ordered by their scores on the first principal component – the first mode of variation. The functions corresponding to the minimum, median and maximum score on this principal component are plotted as a way of illustrating this mode of variation. Similar plots are made for the second principal component. The important idea, as far as this thesis is concerned, is that the functions are ordered with respect to some metric.

If the distance measure $D$ captures the important variation between subjects, then ranking the $\hat{f}_i$ by $\bar{D}_i$ makes sense. The distance measure $D$ provides a mapping from the infinite-dimensional space of probability density functions to the one-dimensional space $\mathbb{R}^+$. Furthermore, each different distance measure $D$ imposes a certain order on the space of probability density functions. With the right choice of $D$, examination of the estimated $\hat{f}_i$ in order of $\bar{D}_i$ should allow interpretation in terms of the underlying scientific problem.
2.3 Fitting the Density Estimates

Although it is known that there are two groups, this analysis will ignore group membership and see if it picks up the heterogeneity in the data. The only remaining decision is the choice of window widths $h$ to use in equation (2.1) for estimating $f_0$ and $f_i$. The choice of window width can have a large effect on the smoothness of the kernel density estimate. Here, values of $h$ corresponding to slightly undersmoothed estimates were used, so that some detail in the data was preserved. These values were chosen by eye. A window width of $h = 0.05$ was used for estimating $f_0$. For estimating $f_i$, a larger window width $h = 0.1$ was used, reflecting the smaller sample size.

2.4 Results and Further Analysis

Figure 2.2(a) contains a plot of $f_0$ as well as the individual kernel density estimates $f_i$. The schizophrenic and non-schizophrenic subjects are plotted in different colours. From this plot, it is quite apparent that the $f_i$ for some of the schizophrenic subjects are both shifted and more dispersed. Figure 2.2(b) confirms this. The subjects with the four largest $\bar{D}_i$ are schizophrenic. The estimate $f_0$ can also be seen to be a poor summary of the data for even the non-schizophrenic population.

Figure 2.3 plots all 17 $f_i$ in order of increasing $\bar{D}_i$. Interestingly, subject 14, who on the evidence of the histograms is distinctly unlike the non-schizophrenic subjects, is not among the four subjects furthest from $f_0$. This is due to the choice of the Hellinger distance for $D$. Subject 14 has a mode at the same location as $f_0$ and a long right tail. Much of the mass under $f_{14}$ lies in the region where $f_0$ is small. Examination of the definition 2.5 shows that differences between $p$ and $q$ where $p$ is small will be given small weight. A distance measure other than the Hellinger distance is more appropriate for detecting differences between $f_i$ and $f_0$ that lie in the tail of $f_0$.

Figure 2.4 shows the empirical cumulative distribution function for the $\bar{D}_i$ with
Figure 2.2: Plots of All Kernel Density Estimates and Extreme Functions: Plot (a) shows \( \hat{f}_0 \) (black line), and the estimated \( \hat{f}_i \). Plot (b) shows \( \hat{f}_0 \) and the four \( \hat{f}_i \) that correspond to the four largest distances \( \hat{D}_i \). In the legend, "Max" indicates the \( \hat{f}_i \) with the largest \( \hat{D}_i \), "Max-1" indicates the \( \hat{f}_i \) with the largest \( \hat{D}_i \), and so on.

(a) Densities of Combined Sample and Individual Subjects

(b) Selected Estimated PDFs
Figure 2.3: Ordered Array of Plots of Kernel Density Estimates: Starting at the top left and moving across the rows, the $f_i$ are plotted in order of increasing distances $D_i$. 
subject numbers for cross-referencing this graph with the histograms and figure 2.3. There is a gap between the $\bar{D}_i$ for subjects 15, 17, 16 and 13 and those for the remaining subjects. The distribution of the distances is needed in order to determine how much meaning to attach to the observed differences between subjects.

One method of obtaining this distribution is to use the bootstrap [9]. There are several ways in which the bootstrap could be applied to this problem. One way is to put the problem in a simple hypothesis-testing framework. The null hypothesis is

\[ H_0: \text{data for each of the subjects are sampled from } f_0. \]

The implication is that any observed differences between $\hat{f}_i$ on different subjects are a result of randomly sampling only 30 observations from $f_0$ for each subject. Label the ordered values of the observed distances as $\hat{D}[1], \ldots, \hat{D}[17]$. The distribution of these 17 ordered distances under the above null hypothesis is simulated as follows. Repeat for $b = 1, \ldots, B$:

1. Take 17 bootstrap samples $y_{bi}^*, i = 1, \ldots, 17$ each of size 30, from the combined set $y_{ij}, i = 1, \ldots, n; j = 1, \ldots, n_i$.

2. Compute $\hat{f}_{bi}^*$ for each sample $y_{bi}^*$.

3. Compute $\hat{f}_{b0}^*$ for the combined data $\{y_{b1}^*, \ldots, y_{b17}^*\}$

4. Compute the distances $D_{bi}^* = D(\hat{f}_{bi}^*, \hat{f}_{b0}^*)$.

5. Label the ordered values of the distances in bootstrap sample $b$ as $D_{bi}^*[1], \ldots, D_{bi}^*[17]$.

The values $D_{i1}^*[1], \ldots, D_{i17}^*[1]$, the smallest distances within each bootstrapped sample, can be used to assign a p-value to $\hat{D}[1]$, the smallest observed distance among the $\hat{D}_i$. Similarly, p-values can be assigned to any $\hat{D}[i]$ by comparisons to the bootstrap distribution of $D_{bi}^*[i]$.

Figure 2.5 shows the results of carrying out this procedure. The graph labelled D[i] is a histogram of 2000 values of $D_{b_i}^*[i]$. For the first 10 ordered distances, the observed distance is shown on the histogram by a vertical line. The other 7 observed distances were all so large that it was not feasible to represent them with vertical lines on the same graphs as the bootstrap samples. Instead, the size of each observed
Figure 2.4: Empirical Cumulative Distribution Function of the Hellinger Distances Based On Kernel Density Estimates. The plotting character indicates group membership and the number printed next to each point corresponds to the subject number.
\( \hat{D}[i] \) is printed on each graph. Every observed ordered distance \( \hat{D}[i] \) was larger than the maximum of the corresponding \( D_0^*[i] \). The null distribution of \( D_0^*[i] \) concentrates on much smaller values than the distribution of \( \hat{D}[i] \). The explanation is this: With random sampling from the combined sample of 510 \( y_{ij} \), there is low probability that a bootstrap sample of 30 will contain a majority of its points in a region corresponding to one of the extreme \( f_i \).

This result of this approach is quite unsatisfactory — all that can be concluded is that data for these subjects do not appear to be sampled i.i.d from \( f_0 \). This is not surprising: the beginning of this chapter postulates a different \( f_i \) for each subject, with the \( f_i \) sampled from some distribution. Notice that there is no explicit representation of \( f_i \) in this bootstrap sampling scheme.

For a more sophisticated use of the bootstrap, the different sources of variability in the observed data \( y \) need to be determined. There is between-subject variability in the underlying densities \( f_i \) that give rise to the data on a particular subject \( i \). And there is the within-person variability that comes from sampling \( y_i \) from \( f_i \). This is something like a random effects analysis of variance: for each \( i \), the subject-specific mean \( f_i \) is randomly sampled from some distribution \( G_0 \) (perhaps with population mean \( f_0 \)): the data on subject \( i \) are sampled from \( f_i \). In this kind of model it is of more interest to estimate distributions than to test hypotheses.

The following approach was taken to examine the within-subject variability in \( \hat{D}_i \). For \( b = 1, \ldots, 200 \):

1. For \( i = 1, \ldots, 17 \), draw a bootstrap sample \( y_{bi}^* \) of size 30 from \( y_i \).
2. Compute \( \hat{f}_{bi}^* \) on \( y_{bi}^* \).
3. Compute \( \hat{f}_{b0}^* \) on the combined data \( \{ y_{b1}^*, \ldots, y_{b17}^* \} \).
4. Compute the 17 distances \( D_{bi}^* = D(\hat{f}_{b0}^*, \hat{f}_{bi}^*) \).

Figure 2.6 shows boxplots of these 200 within-subject bootstrapped distances for each subject.
Figure 2.5: Null Hypothesis Bootstrap Distribution of Ordered Hellinger Distances: The plot labelled $D_{[i]}$ shows a histogram of the 2000 values of $D^*_b[i]$. $D^*_b[i]$ is the $i$th ordered Hellinger distance among the 17 values of $D^*_b$ in bootstrap sample $b$. The null hypothesis and bootstrap sampling scheme are described on page 13.
Figure 2.6: Distribution of 200 bootstrap values of $D(\hat{f}_{i0}, \hat{f}_{i1})$ on each subject. Each $\hat{f}_{i0}$ was computed on a bootstrap sample of 30 observations drawn from $y_i$. Each bootstrapped distance was computed between the individual bootstrap density estimates and $f_{i0}$ for the combined data in bootstrap samples $\{y_{i1}, \ldots, y_{i17}\}$.

**Within-Subject Bootstrap Samples of Hellinger Distance**
Figure 2.4 is the empirical distribution of $\tilde{D}_i$, but it is quite sparse as a result of the small sample size. A smoother distribution can be simulated with the bootstrap. The random effects structure above suggests a two-level bootstrap, to take into account both between- and within-subject variation. To formalize this, suppose the data has the following structure:

\begin{align*}
y_{ij} & \sim f_i, j = 1, \ldots, n_i \\
f_i & \sim G_0, i = 1, \ldots, n
\end{align*}

Under the assumption that the densities $f_i$ are sampled from some parent distribution $G_0$, the empirical distribution $\hat{G}_0$ is represented by the 17 estimated $\hat{f}_i$. Bootstrap samples from $\hat{G}_0$ are properly sampled from among the $\hat{f}_i$ with replacement. For B bootstrap replications, this would give indices $i_b^*, b = 1, \ldots, B$: $i_b^* \in \{1, \ldots, 17\}$. The next step might be to simulate a set of 30 observations $y_{i_b^*}$ from the sampled $f_{i_b^*}$. The bootstrap density estimate $\hat{f}_{i_b^*}$ would be computed on $y_{i_b^*}$.

The following approximation to the above sampling scheme was used: the major difference is that the indices $i_b^*$ were used to sample among the data $y_i$ instead of among the $\hat{f}_i$. To obtain the bootstrap distribution of $D$, repeat for $b = 1, \ldots, B$:

1. Sample a subject $i_b^*$ from among subjects 1 to 17.
2. Sample 30 observations on subject $i_b^*$ with replacement to get $y_{i_b^*}$.
3. Compute the density estimate $\hat{f}_{i_b^*}$ on $y_{i_b^*}$.
4. Compute the distance $D_{i_b^*} = D(\hat{f}_0, \hat{f}_{i_b^*})$

Notice that the subjects $i$ do not necessarily occur an equal number of times among the $B$ bootstrap samples with this bootstrap scheme. Also, this scheme does not presuppose a sample size of 17, but aims to construct something close to the population distribution of $D$. This is why the observed $\hat{f}_0$ is used as the mean density: for large $B$, computing $\hat{f}_0$ on the combined sample of $30 \times B$ $y$-values gives an estimate very close to the observed $f_0$. Figure 2.7 shows the bootstrap distribution of $B = 2000$.
bootstrap replications of $D_0^*$ from the two-step bootstrap scheme. This follows the same general shape as the empirical distribution, but is "filled in". Only one of the observed $\hat{D}_i$ is larger than the 95th percentile, which is just about right for a sample size of 17. Of course, an outlying subject will make the tails of this bootstrap distribution appear longer, since each subject contributes to the estimation of the distribution of the distance. Perhaps a better approach would be to compare the distance of subject $i$ from the mean to the distribution of $D$ formed without subject $i$. That is, the distance between $\hat{f}_0^{-i}$ and $\hat{f}_i$ could be compared to the bootstrap distribution of $D_0^*$ based on the 16 observations.

Similarly the bootstrap distribution in Figure 2.7 could be used to decide whether data on a new subject comes from the same distribution as data for these 17 subjects. The distance $D_{\text{test}} = D(\hat{f}_0, \hat{f}_{\text{new}})$ can be referred to the bootstrap distribution of $D_0^*$. Large values of $\{D_0^* < D_{\text{test}}\}$ would indicate that the new subject is unlike the others.

2.5 Discussion

The nonparametric methods in this chapter estimate densities $\hat{f}_i$ for each subject. Summaries and inference are by necessity based on these estimated densities, as there is no underlying model. Some of the summaries and comparisons in this chapter are useful and simple to compute. The two-step bootstrap scheme appears to best represent the different modes of variation. Importantly, the sample average $\hat{f}_0$ plays no role in this bootstrap distribution of $f_0^*$. However, the meanings of some of the key parts of this nonparametric approach are uncertain. For example, it is not clear how to interpret the probability density $f_0$, or even how to decide which estimate of $f_0$ should be used; there is also no well-defined relationship between $f_0$ and the $f_i$, or between $f_i$ and $G_0$. Although intuition says that the density functions themselves vary from subject-to-subject, there is no explicit representation of this variation in this approach.
Figure 2.7: Empirical CDF of the distribution of 2000 bootstrap values of $D(\hat{f}_0, \hat{f}_b^*)$. The observed values are also plotted. The bootstrap samples are obtained in a two-step process. First a subject is sampled from $1, \ldots, 17$. Then a bootstrap sample $y_b^*$ is taken from the data for subject $i$. The estimate $\hat{f}_b^*$ is computed on $y_b^*$. 
Furthermore, there may be important information that it is not possible to extract with this model-free analysis. For example, Gelman and Rubin’s analysis of the response time data postulate three groups of response times [15]. Non-schizophrenics constitute the first group. Schizophrenics as a group exhibit a delayed response and some proportion of the responses on each schizophrenic subject is further delayed. There is scientific interest in the mean reaction times and the probability of a delay. The authors fit a parametric hierarchical Bayesian model to this data with parameters representing means, delays and the probability of delay. Posterior distributions are obtained for the various parameters and inference is carried out in the parameter space. This type of analysis is not possible with a purely nonparametric approach.

Gelman and Rubin’s approach requires two things: (1) that there be an underlying scientific model for the data; and (2) that the data come from some finite mixture of parametric distributions. The purpose of this thesis is to develop a general approach to summarizing distributions on a set of subjects and in general neither of these conditions can be assumed. What is needed is an approach that combines the flexibility of the nonparametric approach in this chapter with an environment in which quantities of interest can be explicitly modelled. And as the two-step bootstrap procedure shows, characterizing the nature of the subject-to-subject variability appears essential to a successful treatment of this sort of data.

In the two-step bootstrap sampling scheme, it was assumed that

\[ f_i \sim G_0. \]

The distribution \( G_0 \) was not estimated in the usual sense, although the samples \( \hat{f}_i \) (ordered by \( \hat{D}_i \)) might be thought of as constituting the empirical distribution \( \hat{G}_0 \). To estimate \( G_0 \) in a more conventional manner, as a function for example, it is necessary to specify how \( f_i \) is sampled from \( G_0 \). Then, given a collection of \( f_i \), \( G_0 \) could presumably be estimated by maximum likelihood. Or, in a Bayesian setting, given a prior distribution on \( G_0 \) and samples from \( G_0 \), one could obtain the posterior distribution of \( G_0 \).

One approach that allows a density function to be sampled from a distribution is
the Bayesian density estimation model of Escobar and West [12]. The next chapter goes into some detail in explaining this model. Suffice it to say here that this model is as flexible as the kernel density estimator in that it assumes no particular distributional form. But it has the advantage of being parametric, so it can be extended to answer the types of questions that the kernel density estimate cannot. In fact, the density estimation model in Escobar and West has as special cases something close to the kernel density estimate and its parametric competitor the finite mixture model
Chapter 3

Background

This chapter briefly reviews some of the statistical background needed to understand the new work in the thesis. It begins with a general description of the Dirichlet process, a distribution on the space of probability distribution functions. The use of the Dirichlet process as a prior and calculation of the corresponding posterior are described. Finally, there is a review of some methods relating to the use of the Dirichlet process in density estimation.

3.1 The Dirichlet Process Prior

The methods developed later in this thesis are based on Bayesian techniques to estimate densities which are developed in Escobar [11, 10] and Escobar and West [12]. These in turn rely on the Dirichlet process prior, a useful tool for so-called non-parametric Bayesian analysis. A Dirichlet process is a probability distribution on a space of probability distributions; since it was introduced by Ferguson [13], it is sometimes referred to as a Ferguson distribution. This chapter will first describe the Dirichlet process, then the use of a Dirichlet process in a Bayesian setting, and finally give a summary of the work of Escobar and West.

In a Bayesian setting, placing a prior distribution on a parameter expresses both the prior belief in the average value of the parameter and the strength of that prior
belief — the precision with respect to the average. To express prior belief about the value of a probability density function, a natural prior distribution is a distribution over the space of probability density functions. One such distribution is the Dirichlet process prior.

The notation \( G \sim \mathcal{D}(F, \alpha_0) \) is used when a distribution function \( G \) has a Dirichlet process prior. The distribution \( F \) is the expected value of \( G \) and this expected value will sometimes be referred to as the baseline prior distribution of the Dirichlet process. The parameter \( \alpha_0 \) is the precision parameter, a positive scalar that measures the strength of the prior belief that sampled distributions \( G \) will be close to \( F \). The larger \( \alpha_0 \), the larger the belief that samples of \( G \) will be close to \( F \).

Ferguson defined the Dirichlet process in measure theoretic terms. Without resorting to measure theory, for a distribution of a single random variable, this definition can be roughly summarized as follows. Let \( G \) be a random distribution whose distribution is a Dirichlet process \( \mathcal{D}(F, \alpha_0) \). Let \( \chi \) be the domain of \( G \) and consider any partition of \( \chi \) into sets \( B_1, \ldots, B_N \). Now consider the vector of random probabilities

\[
P = (P_1, \ldots, P_N) = (P(B_1), \ldots, P(B_N)) = (\int_{B_1} dG(x), \ldots, \int_{B_N} dG(x)).
\]

where of course \( \sum_{i=1}^N P_i = 1 \), since the sets \( \{B_i\} \) are a partition of \( \chi \). The vector \( P \) has the following joint density function

\[
\frac{\Gamma(\alpha_0)}{\prod \Gamma(\alpha_0 F_i)} P_1^{\alpha_0 F_1-1} \cdots P_N^{\alpha_0 F_N-1} (1-P_1-P_2-\cdots-P_{N-1})^{\alpha_0 F_N-1}.
\]

where \( F_i = \int_{B_i} dF(x) \). This is the density of the Dirichlet distribution with parameter vector \((\alpha_0 F_1, \ldots, \alpha_0 F_N)\). If \( X \) is sampled from the random distribution \( G \), the expected probability of \( X \) falling into a region \( B_i \) is \( \int_{B_i} dF(x) \). For small \( \alpha_0 \), the variance of the probability is large. As \( \alpha_0 \) gets larger, the variance of this probability becomes smaller, and distributions sampled from \( \mathcal{D}(F, \alpha_0) \) become more like \( F \).

The practical implications of this are best seen in the structure of a sample from \( G \) when \( G \sim \mathcal{D}(F, \alpha_0) \) \([5]\). Suppose \( X \sim G \) and \( F \) is a continuous distribution; to sample random variables \( X_1, \ldots, X_n \) from \( G \), the following sequence of distributions
is sampled:

\[
X_1 \sim F \\
X_2 \sim \frac{\alpha_0}{\alpha_0 + 1} F + \frac{1}{\alpha_0 + 1} \delta_x \\
X_3 \sim \frac{\alpha_0}{\alpha_0 + 2} F + \frac{1}{\alpha_0 + 2} (\delta_{x_1} + \delta_{x_2}) \\
\vdots \\
X_n \sim \frac{\alpha_0}{\alpha_0 + n - 1} F + \frac{1}{\alpha_0 + n - 1} \sum_{i=1}^{n-1} \delta_{x_i}. \tag{3.1}
\]

where \( \delta_x = 1 \) is a point mass at \( x \).

The first observation sampled from \( G \) is sampled from the prior expectation \( F \). The second observation is sampled from a mixture: it is either sampled from \( F \) (with probability \( \frac{\alpha_0}{\alpha_0 + 1} \)) or is equal to the first observation \( X_1 \) (with probability \( \frac{1}{\alpha_0 + 1} \)). For \( X_n \), the probability of obtaining a new value is \( \frac{\alpha_0}{\alpha_0 + n - 1} \). With equal probabilities of \( \frac{1}{\alpha_0 + n - 1} \) on each, \( X_n \) is equal to one of the previously sampled \( X_i \). As \( n \to \infty \), the probability of drawing \( X_n \) from \( F \) approaches zero, and \( X_n \) is equal to one of the existing \( X_1, \ldots, X_{n-1} \) with probability close to 1. For a sample of any size greater than 1, it is clear that there is a positive probability that at least 2 of the \( X_i \) will have a common value, in which case the number of distinct values will be less than \( n \). The unique values among the \( X_1, \ldots, X_n \) are an i.i.d. sample from the baseline prior distribution \( F \), the only source of new values [5].

The role of \( \alpha_0 \) can now easily be seen. If \( \alpha_0 \) is very large, then the probability that \( X_i \) is drawn from \( F \) will be large (\( \frac{\alpha_0}{\alpha_0 + i - 1} \approx 1 \)). There will be large number of distinct values in a sample, all drawn from \( F \), and the distribution of the sample will approximate the distribution \( F \). If \( \alpha_0 \) is very small, then the probability that \( X_i \) is drawn from \( F \) will be small; it will have a high probability of being equal to an existing \( X_j, j < i \). The sample will cluster on a small number of distinct values. The distribution of the sample will be quite different from \( F \).

As the previous paragraph suggests, placing a Dirichlet process prior on \( G \) with precision parameter \( \alpha_0 \) establishes a prior distribution on the number of distinct values in a sample of size \( n \) from \( G \). The distribution of the number of distinct values \( K \) in a
sample of size $n$ can be computed from a recursive formula that follows immediately from the distribution of the sample above (equation 3.1). Let $P_n(K = k|\alpha_0)$ be the probability of obtaining $k$ distinct values in a sample of size $n$ when the precision parameter is $\alpha_0$.

$$P_1(1|\alpha_0) = 1$$

$$P_n(1|\alpha_0) = \frac{\alpha_0^n}{\alpha_0} \frac{\alpha_0}{\alpha_0 + 1} \cdots \frac{\alpha_0}{\alpha_0 + n - 1}$$

$$= \frac{\alpha_0^n \Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)}$$

$$P_n(K = k|\alpha_0) = \frac{\alpha_0 P_{n-1}(K = k - 1|\alpha_0) + (n - 1) P_{n-1}(K = k|\alpha_0)}{\alpha_0 + n - 1}$$

$n \geq k > 1$

Results in Antoniak [1] show that the probabilities calculated from this recursive relationship have to be evaluated only for $\alpha_0 = 1$, as the following relationship holds:

$$P_n(K = k|\alpha_0) = P_n(K = k|\alpha_0 = 1)n! \alpha_0^k \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)}. \quad (3.2)$$

Figures 3.1 and 3.2 show $P_n(K = k|\alpha_0)$ for a range of values of $\alpha_0$. for $n = 25$ and $n = 50$. Although these distributions are discrete, they are plotted with lines for the sake of clarity. The actual values can be found in Appendix C.

For a given sample size, choice of the value for $\alpha_0$ has a profound effect on the distribution of the number of distinct values. Similarly, the value of $\alpha_0$ as a precision parameter can be interpreted only for a particular sample size. Antoniak [1] derives an approximation for the expected number of distinct values. For the $i^{th}$ draw of an observation from $G$, the probability of a new value is $\frac{\alpha_0}{\alpha_0 + i - 1}$. Let $W_i = 1$ if observation $i$ is a new value and $0$ if it is equal to an existing value. so $E(W_i)$ is equal to the probability of a new value on the $i^{th}$ draw. Let the number of distinct values in a sample of size $n$ be $K_n = \sum_{i=1}^{n} W_i$. Then $E(K_n) = \sum_{i=1}^{n} \frac{\alpha_0}{\alpha_0 + i - 1} \approx \alpha_0 \int_{0}^{\infty} (\alpha_0 + x)^{-1}dx = \alpha_0 \log\left(\frac{\alpha_0 + n}{\alpha_0}\right)$. Since the sum $\sum_{i=1}^{\infty} \frac{\alpha_0}{\alpha_0 + i - 1}$ is divergent (by the integral test [21]), as $n \to \infty$, $E(K_n) \to \infty$, even though the probability of a new value gets rarer and rarer with increasing $n$. Blackwell and MacQueen [5] show that as $n \to \infty$, the distribution of the sequence $X_1, \ldots, X_n$ converges with probability 1 to a discrete distribution.
Figure 3.1: \( P(K = k | \alpha_0, n = 25) \)
Figure 3.2: $P(K = k | \alpha_0, n = 50)$
Not only does the probability of a new value become rarer, but for any particular sample from a distribution with a Dirichlet prior, the empirical distribution changes less and less with increasing sample size. Any functional of the sample (e.g., sample mean, sample variance) will stabilize in the same way. The difference between the probability measure of a finite sample and the probability measure of a hypothetical infinite sample can be made arbitrarily small by taking large enough samples from $G$. Since the finite sample of size $n$ will cluster on $k < n$ unique values, only $k$ values need be generated from the baseline prior. The probability weights for the $k$ atoms can be calculated separately (see section 8.3). This ability of finite samples from $G \sim \mathcal{D}(F, \alpha_0)$ to capture important features of $G$ is detailed in a recent report by Muliere and Tardella [20].

Some results from that report are used in Section 8.3 in an algorithm that generates distribution functions from a Dirichlet process posterior. To give the flavour of the algorithm, a simple example, borrowed from Muliere and Tardella, is given here. Suppose that $G \sim \mathcal{D}(F, \alpha_0)$ with $F$ continuous and interest focuses on the median of $G$. Each function $G$ sampled from $\mathcal{D}(F, \alpha_0)$ will have a median, so that placing a prior on $G$ also places a prior on the median of $G$. Specifying how close the empirical distribution of a sample from $G$ should be to the true distribution $G$ determines the number of atoms $k$ to sample from the baseline prior $F$. Now, repeat the following steps a large number $N$ times:

1. Simulate $k$ atoms independently from $F$.

2. Calculate $k$ probability weights for these $k$ atoms.

3. Compute the median of this weighted sample.

The distribution of $N$ of these simulated sample medians is close to the true distribution of the median of $G$. The authors show that the distribution of any continuous functional of $G$ can be approximated this way.
3.2 The Dirichlet Process Posterior

An appealing feature of the Dirichlet process prior for a distribution function \( G \) is that the posterior distribution of \( G \), given a sample from \( G \) is also a Dirichlet process [13]. Under the model

\[
X|G \sim G \\
G|F, \alpha_0 \sim \mathcal{D}(F, \alpha_0).
\]

the posterior distribution of \( G \), given an observation \( x \) is

\[
G|x, \alpha_0, F \sim \mathcal{D}\left(\frac{\alpha_0 F + \delta_x}{\alpha_0 + 1}, \alpha_0 + 1\right).
\]

where \( \delta_x \) represents a unit mass at the point \( x \) and \( \frac{\alpha_0 F + \delta_x}{\alpha_0 + 1} \) represents the distribution which is a mixture of \( F \) and \( \delta_x \), with weights of \( \frac{\alpha_0}{\alpha_0 + 1} \) and \( \frac{1}{\alpha_0 + 1} \) respectively. By induction, the posterior distribution given observations \( x = \{x_1, \ldots, x_n\} \) is

\[
G|x, \alpha_0 \sim \mathcal{D}\left(\frac{\alpha_0 F + \sum_{i=1}^{n} \delta_{x_i}}{\alpha_0 + n}, \alpha_0 + n\right). \tag{3.3}
\]

The posterior distribution of \( G|x \) is a Dirichlet process with expectation equal to a weighted average of the prior mean \( F \) and unit masses at each of the observations \( \{x_1, \ldots, x_n\} \). This allows an interpretation of \( \alpha_0 \) distinct from its role in the expression for the number of atoms in a finite sample from \( G \). The value of \( \alpha_0 \) expresses the strength of the prior belief that \( G = F \) in terms of numbers of observations. In this simple Dirichlet process model, the posterior places most weight on the baseline prior \( F \) until \( n \geq \alpha_0 \)

3.3 The Dirichlet Process in Density Estimation

Escobar and West [12] develop an algorithm for density estimation using mixtures of normal densities where the distribution function for the means and variances of the components of the mixtures has a Dirichlet process prior. This algorithm will be the basis of the work in this thesis, so will be examined in some detail below. As much
Table 3.1: General formulation of density estimation and Escobar and West implementation

<table>
<thead>
<tr>
<th>General Formulation</th>
<th>Escobar and West</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_i</td>
<td>\pi_i \sim f(Y_i</td>
</tr>
<tr>
<td>$\pi_i</td>
<td>G \sim G$</td>
</tr>
<tr>
<td>$G</td>
<td>G_0 \sim \mathcal{D}(G_0, \alpha_0)$</td>
</tr>
<tr>
<td>$G_0$ is a distribution on $\pi$</td>
<td>$G_0 \equiv N(\mu</td>
</tr>
</tbody>
</table>

as possible, the algorithm will be presented in general terms and references to the implementation in the Escobar and West paper will be used as concrete examples. Since the paper will be repeatedly referenced, it will be referred to in shorthand as EW.

The essence of EW is the hierarchical model shown in the left column of Table 3.1. The observations $Y_n = \{Y_1, \ldots, Y_n\}$ each have a distribution $f(Y_i|\pi_i)$ parametrized by $\pi_i$, and are conditionally independent, given the values $\Pi_n = \{\pi_1, \ldots, \pi_n\}$. The prior distribution of the $\pi_i$ is the distribution $G$, which is a randomly sampled function. It has a Dirichlet process prior distribution with mean $G_0$ and precision $\alpha_0$. In the EW paper (right column of Table 3.1), the distribution $f(Y_i|\pi_i)$ is a normal distribution, where the parameter vector $\pi_i$ represents the mean and precision of the normal distribution for $Y_i$. The mean of the Dirichlet process prior, $G_0$, is a normal-gamma distribution. To sample $\pi = (\mu_i, \nu_i)$ from this $G_0$, first a precision $\nu_i$ is sampled from a gamma distribution with shape $a$ and scale $b$. Then a mean $\mu_i$ is sampled from a normal distribution with mean $m$ and precision $v$ times a known scalar $\Delta$.

The Bayesian density estimate that results from this model is the predictive distribution of a new observation, given the observed data, $f(Y_{n+1}|Y_n)$. For example, in EW a new observation $Y_{n+1}|\pi_{n+1}$ is normally distributed with parameter $\pi_{n+1}$. Conditional on the values $\Pi_n$, equation (3.1) is used to write the distribution of $\pi_{n+1}$ as

$$F(\pi_{n+1}|\Pi_n) = \frac{\alpha_0 G_0(\pi_{n+1}) + \sum_{i=1}^{n} \delta_{\pi_i}(\pi_{n+1})}{\alpha_0 + n}.$$
There will likely be \( k < n \) distinct values in \( \Pi_n \), so by collecting \( \pi_i \) that share common values, this can be rewritten as

\[
F(\pi_{n+1}|\Pi_n) = \frac{\alpha_0 G_0(\pi_{n+1}) + \sum_{j=1}^{k} n_j^* \delta_{\pi_j^*}(\pi_{n+1})}{\alpha_0 + n}
\]

The \( \pi_j^* \) are the \( k \) distinct values in \( \Pi_n \) and \( n_j^* \) is the number of members of \( \Pi_n \) that are equal to \( \pi_j^* \). The notion of a configuration [27] is introduced here, as it makes some of the subsequent discussion a little more concise. The set of indices of the members of \( \Pi_n \) that are equal to \( \pi_j^* \) is called \( S_j \). For example, if \( \pi_3, \pi_7 \) and \( \pi_8 \) have a common value that is indexed as \( \pi_2^* \), then \( S_2 = \{3, 7, 8\} \). In this case, \( n_2^* = 3 \). The set \( \{S_1, \ldots, S_k\} \) is called a configuration. The configuration and the ordered set of values \( \{\pi_1^*, \ldots, \pi_k^*\} \) are a description of the relationship between the data values \( Y_i \) and the parameters \( \Pi_n \).

The predictive density \( f(Y_{n+1}|\Pi_n) \) is given by integrating over the distribution of \( \pi_{n+1} \):

\[
f(Y_{n+1}|\Pi_n) = \int f(Y_{n+1}|\pi_{n+1}) dF(\pi_{n+1}|\Pi_n) = \frac{\alpha_0 f_0(Y_{n+1}) + \sum_{j=1}^{k} n_j^* f(Y_{n+1}|\pi_j^*)}{\alpha_0 + n}.
\]

The density \( f_0(Y_{n+1}) \) is the marginal distribution of \( Y \) under the prior \( G_0(\pi) \) and the likelihood \( f(Y|\pi) \). For the model in EW, \( f_0(Y_{n+1}) = f f(Y_{n+1}|\pi_{n+1}) dG_0(\pi_{n+1}) \) turns out to be a t-distribution, with mean, precision and degrees of freedom determined by the parameters of \( G_0 \). The predictive distribution of \( Y_{n+1} \), given \( \Pi_n \), is a mixture of this t-distribution and \( k \) normal distributions.

So far, the observed data \( Y_n \) are not included in the model. To use the observed data, write out the conditional predictive distribution and notice that \( f(Y_{n+1}|\Pi_n, Y_n) \) as evaluated above, is free of \( Y_n \).

\[
f(Y_{n+1}|Y_n) = \int f(Y_{n+1}|\Pi_n, Y_n) dF(\Pi_n|Y_n) = \int f(Y_{n+1}|\Pi_n) dF(\Pi_n|Y_n).
\]
This last integral can be approximated by the sum
\[
f(Y_{n+1} | Y_n) \approx \frac{1}{B} \sum_{b=1}^{B} f(Y_{n+1} | \Pi_n^b).\tag{3.6}
\]
where each vector \( \Pi_n^b \) is sampled from \( F(\Pi_n | Y_n) \) and \( B \) is suitably large. Each of the summands in equation (3.6) is formed by plugging the sampled values \( \Pi_n^b \) into equation (3.4). The result is a sample from the posterior distribution of \( f(Y_{n+1} | \Pi_n) \).
Their average value gives the posterior predictive distribution.

In E.W. there is no explicit expression for \( F(\Pi_n | Y_n) \). This will generally be true when the distribution of \( G(\pi_i) \) is given a Dirichlet prior. Fortunately, it is possible [10] to obtain approximate samples from \( F(\Pi_n | Y_n) \) through Gibbs sampling. These sampled values of \( \Pi_n^b \) can then be used in the Monte Carlo integration in equation (3.6). The Gibbs sampling method is described in the next section.

### 3.4 Gibbs Sampling With The Dirichlet Prior

Write \( \Pi_n \) without \( \pi_i \) as \( \Pi_n^{-i} \). In this model, Gibbs sampling works on the presumption that the full conditional distributions \( f(\pi_i | \Pi_n^{-i}, Y_n) \) are available and can be sampled from. Each of the \( \pi_i \) is updated in turn for \( i = 1, \ldots, n \), with the most recently sampled values being entered into \( \Pi_n \). One complete \( (n \text{ step}) \) cycle of updating \( f(\pi_i | \Pi_n^{-i}, Y_n) \) constitutes one Gibbs sampler update of \( f(\Pi_n | Y_n) \). After a large number (say \( l \)) of such cycles, the distribution of the sampled values of \( \Pi_n \) is considered to have converged to the required joint conditional distribution, \( F(\Pi_n | Y_n) \). Values sampled from this joint conditional distribution after cycle \( l \) can be used for Monte Carlo integration.

The conditional independence of the members of \( Y_n \), given \( \Pi_n \), means that
\[
f(Y_n | \Pi_n) = f(Y_1 | \pi_1) f(Y_2 | \pi_2) \ldots f(Y_n | \pi_n).
\]
This allows the simplification between the second and fourth lines of the expressions below for the conditional distributions required by the Gibbs sampler.

\[
f(\pi_i | \Pi_n^{-i}, Y_n) \propto f(Y_n | \pi_i, \Pi_n^{-i}) f(\pi_i | \Pi_n^{-i})
\]
There is one minor adjustment for this to be correct. If \( i \in S_j \), then \( n_j^* \) is replaced by \( n_j^* - 1 \). Finally, the conditional distribution is written in a way that makes it useful:

\[
f(\pi_i|\Pi_n^{-1}, Y_n) \propto \alpha_0 f_0(Y_i) G_0(\pi_i|Y_i) + \sum_{j=1}^{k} f(Y_i|\pi_j^*) n_j^* \delta_{\pi_j^*}(\pi_i)
\]

(3.7)

where \( G_0(\pi_i|Y_i) \) is the baseline posterior, given data \( Y_i \) and \( f_0(Y) \) is the marginal distribution of \( Y \) under the baseline prior \( G_0 \). Inspection of equation (3.7) shows that the distribution of \( (\pi_i|\Pi_n^{-1}, Y_n) \) is a mixture that can be sampled as follows.

1. \( \pi_i \) is sampled from the posterior \( G_0(\pi_i|Y_i) \) with probability proportional to \( \alpha_0 \) times \( f_0(Y_i) \), the marginal distribution of \( Y \) evaluated at \( Y_i \).

2. \( \pi_i \) is equal to one of the \( \pi_j^* \) with probability proportional to \( n_j^* \) times the likelihood \( f(Y_i|\pi_j^*) \).

The constant of proportionality is the same in (1) and (2) above. Notice that each new value of \( \pi_i|Y_i \) is always sampled from the baseline posterior distribution \( G_0(\pi_i|Y_i) \). For the specified baseline prior \( G_0(\pi_i) \) and conditional distribution \( f(Y_i|\pi_i) \), this method requires two things:

1. The evaluation of the marginal distribution of \( Y \) at \( Y_i \). In the EW model, this is a t-distribution.

2. That it be possible to sample from the posterior distribution \( G_0(\pi_i|Y_i) \). In EW, this is a normal-gamma distribution.

Notice that this sampling method is independent of the actual distributions involved and results solely from the structure of the model. The development above will be used with only slight modifications in Chapters 4, 5, and 6.
Once it is judged that the distribution of the sampled values of $\mathbf{\Pi}_n$ has converged to the required distribution, the sampled values themselves can be used for the Monte Carlo integration. At each step $b$ of the Gibbs sampler, the current value of the vector $\mathbf{\Pi}_n$ is used as $\mathbf{\Pi}_n^b$ in equation (3.6). This gives a sample of the predictive distribution for the parameter $\pi_{n+1}$, conditional on the observed data. The individual samples $f(Y_{n+1}|\mathbf{\Pi}_n^b)$ can be used to assess the variability around the posterior mean $f(Y_{n+1}|\mathbf{Y}_n)$. The mean value of the predictive distribution can be written in full as

$$f(Y_{n+1}|\mathbf{Y}_n) \approx \frac{1}{B}(\alpha_0 + n)^{-1} \sum_{b=1}^{B} \left[ \alpha_0 f_0(Y_{n+1}) + \sum_{j=1}^{k^b} n_j^{*b} f(Y_{n+1}|\pi_j^{*b}) \right].$$  

(3.8)

where $k^b$, $\pi_j^{*b}$, and $n_j^{*b}$ have their usual meanings, and the superscript $b$ indicates that the values are particular to step $b$ of the Gibbs sampler. Notice that each sampled predictive distribution is a mixture of (1) a baseline predictive distribution that is determined solely by the baseline prior $G_0$; and (2) distributions parametrized by the the sampled values $\mathbf{\Pi}_n^b$. In EW, the second part of this distribution is a mixture of normals, but the number of unique values in $\mathbf{\Pi}_n^b$, thus the number of mixture components and their respective weights, varies over cycles of the Gibbs sampler.

This approach to density estimation has been successfully applied to several scientific problems. Escobar and West [12] show an application to the distribution of the velocity of galaxies where scientific evidence focuses on the number of clusters $k$. West [28] contains a detailed example of an application to neurological transmission data. This approach has also been used for bivariate density estimation [29].

### 3.5 Learning About The Dirichlet

#### Precision Parameter

The distances between the means $\mu_i$, the sizes of the precisions $\nu_i$, and the size of $\alpha_0$ all combine to determine the amount of smoothing in the sampled predictive distributions. The posterior distribution of the number of distinct elements among $\mathbf{\Pi}_n$ in the density estimation model is a function not only of $\alpha_0$, but also of these
other parameters. The interplay between these variables is quite complex, so it is important that as much as possible is learned from the data about their values. EW shows one way that learning about $\alpha_0$ can be added to the Gibbs sampling analysis.

### 3.5.1 Parametric Prior for $\alpha_0$

From equation (3.2), $f_n(k|\alpha_0) \propto \alpha_0^k \Gamma(\alpha_0) \Gamma^{-1}(\alpha_0 + n)$. Conditional independence arguments regarding $\alpha_0$, $k$, $Y_n$, and $\Pi_n$ lead to the equality

$$f(\alpha_0|k, \Pi_n, Y_n) = f(\alpha_0|k, n) \propto f(\alpha_0)f(k|\alpha_0, n).$$

Placing a gamma prior $\text{Gamma}(\alpha_0|x, X)$ on $\alpha_0$ and using the identity

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} = \frac{(\alpha_0 + n)B(\alpha_0 + 1, n)}{\alpha_0\Gamma(n)}$$

(where $B(\ldots)$ is the beta function) gives the expression

$$f(\alpha_0|k, n) \propto \alpha_0^{k+\varepsilon-2}e^{-X\alpha_0(\alpha_0 + n)}B(\alpha_0 + 1, n) \times \alpha_0^{k+\varepsilon-2}e^{-X\alpha_0(\alpha_0 + n)} \int_0^1 z^{\alpha_0}(1 - z)^{n-1} dz.$$

Following EW, this means that $f(\alpha_0|k, n)$ is the marginal distribution from the bivariate distribution of $\alpha_0$ and a variable $\eta$:

$$f(\alpha_0, \eta|k, n) \propto (\alpha_0 + n)\alpha_0^{k+\varepsilon-2}e^{-X\alpha_0\eta(1 - \eta)^{n-1}}. \quad 0 < \eta < 1.$$

The full conditional distribution for $\alpha_0$ is now the mixture of two gamma distributions:

$$f(\alpha_0|\eta, k, n) \propto (\alpha_0 + n)\alpha_0^{k+\varepsilon-2}e^{-X\alpha_0\eta(1 - \eta)^{n-1}}$$

$$\times \alpha_0^{k+\varepsilon-1}e^{-\alpha_0(X - \log(\eta)) + n\alpha_0^{k+\varepsilon-2}e^{-\alpha_0(X - \log(\eta))}}$$

$$\times \frac{\Gamma(x + k)}{(X - \log(\eta))^{x+k}}\text{Gamma}(\alpha_0|k + x, X - \log(\eta))$$

$$+ \frac{\Gamma(x + k + 1)}{n(X - \log(\eta))^{k+x-1}}\text{Gamma}(\alpha_0|k + x - 1, X - \log(\eta)). \quad (3.9)$$

The conditional distribution of $\eta$ is $\text{Beta}(\eta|\alpha_0 + 1, n)$:

$$f(\eta|\alpha_0, n) \propto \eta^{\alpha_0}(1 - \eta)^{n-1}$$
Updating \( \alpha_0 \) is now straightforward. Given the current value of \( \alpha_0 \) and the sample size \( n \), draw \( \eta \) from \( f(\eta|\alpha_0, n) \). Given this value of \( \eta \) and the current value of \( k \), draw a new \( \alpha_0 \) from the mixture of gamma distributions (3.9). The posterior distribution \( f(\alpha_0|Y_n) \) can be estimated by the average of the \( B \) sampled conditional distributions

\[
f(\alpha_0|Y_n) = \frac{1}{B} \sum_{b=1}^{B} f(\alpha_0|\eta_b, k^b)
\]

### 3.5.2 Discrete Prior For \( \alpha_0 \)

There is at least one disadvantage to using a parametric prior distribution for \( \alpha_0 \): it may not be possible to choose hyperparameters such that the distribution adequately represents one's prior belief. One solution is to choose a prior that is a mixture of parametric priors. Another solution is to use a discrete prior: on a grid of \( \alpha_0 \) values \( \{a_1, \ldots, a_R\} \), choose prior probabilities \( \{p_1, \ldots, p_R\} \). For \( \alpha_0 \), the posterior distribution under this prior is easily calculated as:

\[
f(\alpha_0 = a_i|k, n) = \frac{a_i^k \Gamma(a_i)p_i/\Gamma(a_i + n)}{\sum_{i=1}^{R} a_i^k \Gamma(a_i)p_i/\Gamma(a_i + n)}
\]

Escobar [11] suggests a grid based on the sample size raised to exponents evenly spaced between -1 and 2. The value \( \alpha_0 = n^{-1} \) corresponds approximately to an expectation of 1 distinct value and the value \( \alpha_0 = n^2 \) corresponds approximately to \( n \) distinct values. For example, with \( R = 301 \), this would give a grid of \( \alpha_0 \) values \( \{n^{-1}, n^{-0.1}, \ldots, n^{2}\} \).

### 3.6 Summary

This chapter has described the Dirichlet process prior and its use in density application as developed by Escobar and West. Since it will be used in subsequent chapters, particular attention was given to the Gibbs sampling scheme of Escobar that makes it possible to obtain samples from the analytically intractable joint posterior distribution \( F(\Pi_n|Y_n) \). The presentation of this algorithm was general enough that
its application to slightly different models should be transparent. See section 5 of Escobar [11] for an explicit statement of the general extension to different models.
Chapter 4

Model for Multiple Samples

4.1  Structure of the Data

The structure of the data that this thesis deals with is typified by the dataset analyzed in Chapter 2. A number of independent measurements of some feature are taken on each subject, and there are several different subjects. Of interest is the way in which the distribution of the measurements varies from subject to subject. The general structure of the data will be described more precisely below.

The distribution of $Y$: Data are observed on each of $n$ subjects, with subjects being indexed by $i$. On the $i^{th}$ subject, $n_i$ independent data points are observed, each indexed by $j$, so that $Y_{ij}$ refers to the $j^{th}$ data point on the $i^{th}$ subject. For each subject $i$, the data are assumed to come from the following hierarchical model:

$$Y_{ij}|\pi_{ij} \sim N(Y_{ij}|\pi_{ij})$$

$$\pi_{ij}|G_i \sim G_i$$

$$G_i \sim \mathcal{D}(G_0,\alpha_0). \quad (4.1)$$

As before, conditional on the parameters $\Pi_i = \{\pi_{ij} : j = 1, \ldots , n_i\}$, the data $Y_i = \{Y_{ij} : j = 1, \ldots , n_i\}$ are independent. Each $Y_{ij}$ is distributed normally with mean $\mu_{ij}$ and precision $\nu_{ij}$. 
The distribution of $\pi_{ij}$: The parameters $\mathbf{I}_i$ for subject $i$ come from the distribution $G_i$, so that the distribution of $Y_i$ varies between subjects when not all $G_i$ are equal.

The Distribution of $G_i$: The distributions $G_i$ have independent Dirichlet process priors with a common baseline prior $G_0$ and a common precision $\alpha_0$. The similarity between any $G_i$ and $G_0$ is controlled by the precision parameter for the Dirichlet process. Hence the similarity between $G_i$ and $G_{i'}$ for $i \neq i'$ is also controlled by this precision parameter. Consider a sample $\mathbf{I}_1$ from $G_1$ and $\mathbf{I}_2$ from $G_2$ each of size $n$. The conditional distribution for a new value of $\pi$ for sample 1 is

$$
\pi_{1,n+1} | \mathbf{I}_{1n} \sim \frac{\alpha_0}{n + \alpha_0} G_0(\pi_{1,n+1}) + \frac{1}{n + \alpha_0} \sum_{k=1}^{n} \delta_{\pi_{1k}}(\pi_{1,n+1}).
$$

Likewise, the conditional distribution for a new value for sample 2 is

$$
\pi_{2,n+1} | \mathbf{I}_{2n} \sim \frac{\alpha_0}{n + \alpha_0} G_0(\pi_{2,n+1}) + \frac{1}{n + \alpha_0} \sum_{k=1}^{n} \delta_{\pi_{2k}}(\pi_{2,n+1}).
$$

If $\alpha_0$ is big, the two conditional distributions above are both close to $G_0$, hence close to each other; on the other hand, if $\alpha_0$ is small, samples from $G_i$ may be quite different from each other.

This interpretation of the meaning of large and small $\alpha_0$ in this model is seen by comparing figures 4.1 and 4.2. (A fuller description of these figures is given in section 4.4.) On each of the two figures, the plot labelled $G_0$ is the baseline prior for the Dirichlet processes in the plots below; $G_0$ is the same in each figure. The top row of three plots, labelled $G_1$, $G_2$ and $G_3$, shows the distribution of a new value of $\pi$, given an existing sample of size 25:

$$
\pi_{i,26} \sim \frac{\alpha_0 G_0 + \sum_{j=1}^{25} \delta_{\pi_{ij}}}{\alpha_0 + 25}.
$$

To aid the comparison of the distributions corresponding to different values of $\alpha_0$, the distributions $f_i(Y)$ corresponding to each $G_i$ have also been plotted:

$$
f_i(Y) = \frac{\alpha_0 \int N(Y|\pi) dG_0(\pi) + \sum_{j=1}^{25} N(Y|\pi_{ij})}{\alpha_0 + 25}.
$$
The dotted vertical reference lines are located at the same abscissas on each graph. In Figure 4.1, \( \alpha_0 \) has the relatively small value 3; some similarities in the three densities \( f(Y) \) reflect the influence of the baseline prior \( G_0 \) – all three have modes at similar locations. However, the order by location of the heights of the modes is different in each of the three plots of \( f(Y) \). In Figure 4.2, \( \alpha_0 \) has the relatively large value 25, and all three sampled densities \( f(Y) \) are quite similar – both in locations of the modes, as well as the order of the their heights.

### 4.2 The Need for a Flexible \( G_0 \)

If \( G_0 \) is a fixed baseline distribution, as in EW, then the model so far amounts to \( n \) separate density estimation problems, all shrunk toward the common \( G_0 \), the prior expectation of the \( G_i \). The amount of shrinkage is determined by \( \alpha_0 \). If \( \alpha_0 \) is large, then the predictive densities for all subjects will tend to look alike. If \( \alpha_0 \) is small, then the densities are estimated essentially independently and the analysis is similar to that in Chapter 2, where a separate kernel density estimate was computed for each subject. The amount and type of variation between subjects is not modelled. There are several reasons why this model is insufficient for the two purposes set out in the motivation section.

In EW, interest lies only in the distribution \( G \) and a single predictive distribution \( f(Y) \); the baseline prior \( G_0 \) is employed only as a source of new \( \pi \). So long as \( G_0 \) gives support to a wide range of candidate values for \( \pi \), the functional form of \( G_0 \) does not matter and it can be chosen for computational convenience. Although \( G_0 \) is the prior expected value of \( G \), there is very little weight given to the belief that the distribution of \( Y \) is actually the marginal based on this baseline prior – a t-distribution. This is evidenced by the small values of \( \alpha \) in EW; they use values near 1 for a sample size of 82.

In the present model for multiple samples, it is assumed that the distributions \( G_i \) have a common parent distribution, which is itself of interest. This is a natural role
Figure 4.1: Here, $\alpha_0 = 3$ in $G_i \sim \mathcal{D}(G_0, \alpha_0)$: For the sake of simplicity, all parameters have been represented in one-dimension. Graphs of $G_1$, $G_2$ and $G_3$ show the distribution of $f(\pi_{n+1} | \Pi_n) \propto \alpha_0 G_0 + \sum_{j=1}^{n} \delta_{\pi_j}$. The bottom row of graphs shows that the distributions $f_1(Y_{n+1} | \Pi_n)$ are quite dissimilar.
Figure 4.2: Here, $\alpha_0 = 25$ in $G_i \sim D(G_0, \alpha_0)$: For the sake of simplicity, all parameters have been represented in one-dimension. Graphs of $G_1$, $G_2$ and $G_3$ show the distribution of $f(\pi_{n+1}|\Pi_n) \propto \alpha_0 G_0 + \sum_{j=1}^{n} \delta_{\pi_j}$. The bottom row of graphs shows that the distributions $f_i(Y_{n+1}|\Pi_n)$ are quite alike.
for $G_0$ to play. Giving weight to large values of $\alpha_0$ expresses the belief that the $G_i$ are all close to this parent distribution. Unless the observed data are of a very unusual form, this way of thinking about $G_0$ is not consistent with using a diffuse distribution chosen for its computational convenience. The distribution $G_0$ should be close to the $G_i$. Of course it is not known in advance what the $G_i$ will be; this suggests that the distribution $G_0$ should be adapted to the "observed" $G_i$.

EW does this to a certain extent, by placing prior distributions on the parameters of the normal-gamma baseline prior $G_0$, then updating these priors conditional on the data. But the prior remains a normal-gamma with an associated $t$-distribution marginal for $Y$. What is needed here is a way of modifying $G_0$ conditional on the $G_i$ so that it can adaptively change not only its location and scale, but also its functional form. Then, in some sense, $G_0$ gives the population average distribution for $\pi_{ij}$, and the mixture distribution

$$F(\pi_{i,n+1}|\Pi_i) \propto \alpha_0 G_0(\pi_{i,n+1}) + \sum_{j=1}^{n} \delta_{\pi_{ij}}(\pi_{i,n+1})$$

allows new values $\pi_{i,n+1}$ for subject $i$ to come from a mixture of $G_0$ and atoms particular to subject $i$.

With the additional data available in the multiple samples, it is possible and desirable to treat the prior expectation $G_0$ itself as a random function. A vague prior distribution is placed on $G_0$ and, given the information in the set of $G_i$, samples are obtained from the posterior distribution of $G_0$. This approach is explained in the next section.

4.3 Posterior Analysis of $G_0$

The prior distribution of $G_i$ is a Dirichlet process with baseline distribution $G_0$. The posterior distribution of $G_i$ given data $Y_i$ is also a Dirichlet process. The baseline distribution for this process (call it $G^*$) is not known explicitly, but since it is a Dirichlet process, is known that the unique elements among $\Pi_i$ are an i.i.d. sample
Let $\mathbf{I}^*$ be the set of the unique elements among the $n$ sets $\mathbf{I}_i, i = 1, \ldots, n$ and label each of the $n_{\pi^*}$ unique value as $\pi_r^*, r = 1, \ldots, n_{\pi^*}$. The values $\mathbf{I}^*$ can be used to estimate the baseline distribution $G^*$. Whatever method is used to estimate this distribution, it is important that it be easy to sample from the resulting estimate.

One approach that keeps the model in the same Bayesian framework is to place a Dirichlet process prior on $G_0$, say $D(G_{00}, \alpha_{00})$, and obtain samples from its posterior distribution. Under this simple Dirichlet process model, the posterior distribution of $G_0$ given $\mathbf{I}^*$ is

$$G_0|\mathbf{I}^* \sim D\left(\frac{\alpha_{00} G_{00} + \sum_{r=1}^{n_{\pi^*}} \delta_{\pi_r^*}}{\alpha_{00} + n_{\pi^*}}, \alpha_{00} + n_{\pi^*}\right).$$

and a new value of $\pi$ sampled from $G_0$ has positive probability of being equal to an existing $\pi_r^*$. However, for $G_0$ to be a useful baseline prior in the model laid out in equation (4.1), it should be a continuous distribution: each time a value of $\pi$ is sampled from $G_0(\pi|Y)$, it should be a unique value. The results relating to the number of distinct elements for a given $\alpha$ and sample size $n$ all rely on the assumption that the baseline prior is a continuous distribution. Furthermore, the estimate of $G_0(\pi|Y)$ under this simple Dirichlet process model is little more than the empirical distribution based on the $n_{\pi^*}$ values in $\mathbf{I}^*$. In the two dimensions spanned by $\pi$, this may not be a useful estimate, unless $n_{\pi^*}$ is quite large.

To improve on this, using an approach similar to that in EW, a continuous $G_0$ is formed by putting "noise" around a distribution $F$. The distribution $F$ is given a Dirichlet process prior. To highlight the structural similarity to $G_i \sim D(G_{00}, \alpha_{00})$, the mean of the Dirichlet process prior for $F$ is called $G_{00}$ and the precision is called $\alpha_{00}$.

$$F \sim D(G_{00}, \alpha_{00})$$

Values sampled from $F$ will be referred to as $\tau$. The distribution $G_0$ is formed by adding normal-gamma "noise" around $\tau$-values drawn from $F$. The notation $M(\pi_r^*|\tau_r)$ will be used to denote that $\pi_r^*$ has a normal-gamma distribution with parameters determined by $\tau_r$. Chapters 5 and 6 contain details of the relationship between $\pi_r^*$. 
and \( \tau \). Here, it will suffice to think of the parameter \( \tau \) acting as the "centre" of the distribution of \( \pi^* \). In sections of the thesis dealing with estimation of \( G_0 \), the notation \( G_0(\pi^*) \) will be used to emphasize that \( G_0 \) is the distribution of the unique values \( \pi^* \). Given \( F \), \( G_0 \) is formed as

\[
G_0(\pi^*) = \int M(\pi^*|\tau) dF(\tau).
\]

(4.2)

The distribution \( F \) provides the parameters and weights for the mixture distribution \( G_0 \). This model ensures that each sample from \( G_0 \) yields a unique value, while allowing the use of the machinery for sampling \( F \) from a Dirichlet process posterior. When \( F \) has a Dirichlet process prior and samples from \( F \) parametrize the distribution of \( x \), then the posterior distribution of \( F|x \) is called a mixture of Dirichlet processes, a phrase introduced by Antoniak [1]. Here, values of \( \tau \) sampled from \( F \) parametrize the distribution of \( \Pi^* \), so the distribution of \( F|\Pi^* \) is a mixture of Dirichlet processes.

Each unique value \( \pi^*_r \) has \( \tau \) as the parameter of its normal-gamma distribution, so corresponding to the set \( \Pi^* \) is the set comprising \( n_{\pi^*_r} \) (probably non-unique) atoms \( \mathbf{T} = \{\tau_1, \tau_2, \ldots, \tau_{n_{\pi^*_r}}\} \). Integration over the distribution \( F \) gives an integral over the baseline prior \( G_{00} \) and a sum over the point masses at the atoms \( \tau \). Equation 4.3 below for the distribution of a new value \( \pi_{n_{\pi^*_r}+1}^* \) given the set \( \mathbf{T} \) is similar to equation (3.4) for the distribution of \( Y_{n+1} \) given the set \( \Pi_n \).

\[
G_0(\pi_{n_{\pi^*_r}+1}^*|\mathbf{T}) \propto \sum_{r=1}^{n_{\pi^*_r}} M(\pi_{n_{\pi^*_r}+1}^*|\tau_r) dG_{00}(\tau_r)
\]

(4.3)

The similarity between the models for the density of the \( \Pi^* \) and the density of a single sample of \( Y_i \) is shown in Table 4.1. Structurally, there are no differences between the two models. The collection of \( \Pi^* \) takes the place of \( Y_i \) as data. The normal-gamma

<table>
<thead>
<tr>
<th>\text{Estimation of } f(Y_{ij})</th>
<th>\text{Estimation of } G_0(\pi^*_r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_{ij}</td>
<td>\pi_{ij} \sim \mathcal{N}(Y_{ij}</td>
</tr>
<tr>
<td>( \pi_{ij}</td>
<td>G_i \sim G_i )</td>
</tr>
<tr>
<td>( G_i</td>
<td>G_0 \sim \mathcal{D}(G_0, \alpha) )</td>
</tr>
</tbody>
</table>

Table 4.1: Analogy between estimation of density of \( Y_{ij} \) and density of \( \pi^*_r \)
takes the place of the normal as the distribution of the data given a parameter. The set \( T \) takes the place of the set \( \Pi_i \) as the parameters of the distribution. Both \( T \) and \( \Pi_i \) come from a distribution that has a Dirichlet process prior. Just as the left column of the table does not explicitly refer to the density \( f(Y) \), the right column does not mention the density \( G_0(\pi^*) \). The only substantial difference is that one model has univariate data \( Y \) and the other bivariate data \( \pi^* \), but this makes no difference as far as the algorithm for model fitting is concerned. Given the collection of \( \Pi^* \), a set of \( T \) can be sampled from the posterior distribution of \( F \). Monte Carlo integration over \( F \) can make use of these values: integration over \( G_0 \) uses the smoothed values from equation (4.3).

### 4.4 Preview of Model Fitting

Fitting a model to the entire collection of subjects entails alternating between the following two steps, each of which uses the Gibbs sampling algorithm described in section 3.4.

1. Using the current estimate of \( G_0 \) (see equation 4.3) as the baseline prior for the Dirichlet processes \( G_i \sim D(G_0, \alpha_0) \), obtain samples from the posterior distributions of \( \Pi_i|Y_i, T \) for \( i = 1, \ldots, n \).

2. Using the (fixed) distribution \( G_{00} \) as the baseline prior for the single Dirichlet process \( F \sim D(G_{00}, \alpha_{00}) \), obtain a sample from the posterior distribution of \( T|\Pi^* \).

After cycling through the two steps above a large number of times, parameters are considered to be sampled from their joint posterior distributions. In particular, values of \( T|\Pi^* \) are considered to be a realization from a distribution sampled from the posterior distribution of \( F \). Likewise, the values of \( \Pi_i|Y_i \) are a considered to be realization from a distribution sampled from the posterior distribution of \( G_i \). These posterior quantities can be used in a number of ways to summarize and interpret the distribution underlying the observed data.
Figures 4.2 and 4.1, as well as depicting the effect of the size of $\alpha_0$, also illustrate the structure of the hierarchical model for multiple samples. For the sake of simplicity, the parameters $\tau$ and $\pi$ have been represented in one-dimension (keeping the precision components constant). The top graph is the baseline prior $G_{00}$, a normal distribution. Below this is the distribution of a new $\tau$ to be sampled from $F$, given some existing values of $\tau$. This is represented by a mixture of point masses $1/(\alpha_{00} + n\pi^*)$ at several values of $\tau$ and a weight on the the baseline prior $G_{00}$ equal to $\alpha_{00}/(\alpha_{00} + n\pi^*)$. Below this, the graph labelled $G_0$ shows the result of integrating over $F$ as in equation (4.3). For the purposes of the graphs, only Gaussian noise is added to the distribution $F$. The top row containing three graphs represents the distribution of a new value of $\pi$, with point masses $1/(\alpha_0 + n_i)$ at the $\pi^*$ and a weight on the the baseline distribution $G_0$ equal to $\alpha_0/(\alpha_0 + n_i)$. The bottom row of three graphs shows the distribution of $Y$, formed by adding Gaussian noise around the atoms $\pi^*$ and the baseline distribution $G_0$.

4.5 Summary

This chapter has outlined the hierarchical model for multiple samples and given the general framework for inference. By using Dirichlet process priors at both levels of sampling, the whole model is kept in a Bayesian framework. Individual density estimates are fitted to each subject’s data, conditional on the observed data and the current estimate of the parent distribution. Then, conditional on these individual estimates, the parent distribution is updated. The posterior distribution of an individual density estimate is mostly a reflection of the data for that subject, but also incorporates information from other subjects, through their influence on the parent distribution.
Chapter 5

Sampling the Distribution of $\pi_{ij}|Y_{ij}$

The data for each subject, $Y_i$, is visited separately and treated essentially the same as in the one-sample model described in section 3.3. Subjects are indexed by $i$ and observations within a subject are indexed by $j$. There are $n_i$ observations for subject $i$. Refer to the set of $\pi_{ij}, j = 1, \ldots, n_i$ for subject $i$ as $\Pi_i$ and this same set without observation $j$ as $\Pi_i^{-j}$. Each distribution $G_i$ has a Dirichlet process prior with the same baseline prior $G_0$ and precision $\alpha_0$.

$$
Y_{ij}|\pi_{ij} \sim f(Y_{ij}|\pi_{ij})
$$

$$
\pi_{ij}|G_i \sim G_i
$$

$$
G_i \sim \mathcal{D}(G_0, \alpha_0).
$$

The distribution $f(Y|\pi)$ is the normal density with mean $\mu$ and precision $\nu$, where $\pi = (\mu, \nu)$.

The difference between this model and the single sample model in EW lies in the specification of the baseline prior distribution $G_0$. In the EW single sample model, $G_0$ was a normal-gamma distribution, the parameters of which were given by hyperpriors. Section 4.3 introduced the idea of $G_0$ being a random function, formed by adding normal-gamma noise around a random function $F$. Another way to think about this is that the parameters sampled from $F$ parametrize the mixture distribution that acts
as the baseline prior $G_0$, as in Antoniak [1]:

$$G_0(\pi^*) = \int M(\pi^* | \tau) \ dF(\tau). \quad (5.2)$$

The distribution $F$ is sampled from a Dirichlet process $\mathcal{D}(G_{00}, \alpha_{00})$ in a separate step of the Gibbs sampler. Once sampled, the function $F$ is essentially considered fixed with respect to visiting each subject and sampling $\pi_{ij} | Y_{ij}$.

Recall that random variables sampled from $F$ are labelled $\tau$. Each $\pi^*_\tau$ has as its parameter vector $\tau$, so there are $n_{\pi^*}$, probably non-unique, elements $\{\tau_1, \tau_2, \ldots, \tau_{n_{\pi^*}}\}$: the entire sample of $\tau$ from $F$ is referred to as $T$. Given $T$, the distribution of a new $\tau_{n_{\pi^*}+1}$ is

$$\tau_{n_{\pi^*}+1} | T \sim \frac{1}{\alpha_0 + n_{\pi^*}} \left[ \alpha_0 G_{00} (\tau_{n_{\pi^*}+1}) + \sum_{r=1}^{n_{\pi^*}} \delta_{\tau_r} (\tau_{n_{\pi^*}+1}) \right]. \quad (5.3)$$

The conditional distribution of a new value of $\pi$ is:

$$\pi_{n_{\pi^*}+1} | \tau_{n_{\pi^*}+1} \sim M (\pi_{n_{\pi^*}+1} | \tau_{n_{\pi^*}+1}). \quad (5.4)$$

Putting together 5.4 and 5.3, and integrating over the distribution of $\tau_{n_{\pi^*}+1}$ gives the marginal distribution of $\pi_{n_{\pi^*}+1}$.

$$f(\pi_{n_{\pi^*}+1} | T) = \frac{1}{\alpha_0 + n_{\pi^*}} \left[ \alpha_0 \int M (\pi_{n_{\pi^*}+1} | \tau) \ dG_{00}(\tau) + \sum_{r=1}^{n_{\pi^*}} M (\pi_{n_{\pi^*}+1} | \tau_r) \right]. \quad (5.5)$$

This distribution will be used as the baseline prior $G_0$ for the Dirichlet process $G_t \sim \mathcal{D}(G_0, \alpha_0)$. Here, writing $f(\pi_{n_{\pi^*}+1} | T)$ emphasizes that this is the density of a new value of $\pi^*$ and that the set $T$ can change. From here on, the conditioning will be omitted and the notation $G_0(\pi)$ will be used to refer to the expression in equation (5.5).

Section 3.3 outlines the algorithm of Escobar (1994) for sampling the conditional distributions needed for updating in a Gibbs sampling routine for the parameters $\Pi_i$. Rewriting equation (3.7) for multiple samples, the only change is in the subscripting. The conditional distribution of $\pi_{ij}$, given the datum $Y_{ij}$ and $\Pi_i^{-j}$ is given by

$$\pi_{ij} | \Pi_i^{-j}, Y_{ij} \sim \frac{\alpha_0 f_0 (Y_{ij}) G_0 (\pi_{ij} | Y_{ij}) + \sum_{k \neq j} f (Y_{ij} | \pi_{ik}) \delta_{\pi_{ik}} (\pi_{ij})}{\alpha_0 f_0 (Y_{ij}) + \sum_{k \neq j} f (Y_{ij} | \pi_{ik})}. \quad (5.6)$$
where \( f_0(Y_{ij}) \) is the marginal density of \( Y \) under the baseline prior \( G_0 \), evaluated at \( Y_{ij} \).

\[
f_0(Y_{ij}) = \int f(Y_{ij} | \pi) dG_0(\pi) \tag{5.7}
\]

The other term in the denominator, \( \sum_{k \neq j} f(Y_{ij} | \pi_{ik}) \), is the integral of the sum in the numerator, so that the density above is normalized.

To sample \( \pi_{ij} \) from this distribution, as in section 3.3, \( \pi_{ij} \) is either equal to one of the \( \Pi^{-1}_i \) or it is sampled from the baseline posterior distribution. Specifically, the distribution of \( \pi_{ij} \) is sampled according to

\[
\pi_{ij} | \Pi^{-1}_i \cdot Y_{ij} \begin{cases} 
\pi_{ik} & \text{w. prob } \frac{f(Y_{ij} | \pi_{ik})}{\alpha_0 f_0(Y_{ij}) + \sum_{r \neq j} f(Y_{ij} | \pi_r)} \\
\sim G_0(\pi_{ij} | Y_{ij}) & \text{w. prob } \frac{\alpha_0 f_0(Y_{ij})}{\alpha_0 f_0(Y_{ij}) + \sum_{r \neq j} f(Y_{ij} | \pi_r)}
\end{cases}
\]

The baseline marginal distribution \( f_0(Y) \) and the baseline posterior distribution \( G_0(\pi_{ij} | Y_{ij}) \) are both different from those in EW and require some computation. The next two sections describe how the computations were carried out.

### 5.1 Computing the Marginal Density \( f_0(Y) \)

The marginal density of \( Y \) under the baseline prior \( G_0 \) looks quite daunting at first, but can be broken down into components so that only one part is not handled by standard conjugate prior methods. Combining equations (5.5) and (5.7) gives the following integral for the marginal density \( f_0(Y) \).

\[
f_0(Y) \propto \int f(Y | \pi) \left[ \alpha_{00} \int M(\pi | \tau) dG_{00}(\tau) + \sum_{r=1}^{n_{\tau^*}} M(\pi | \tau_r) \right] d\pi
\]

\[
\propto \int \alpha_{00} f(Y | \pi) \left[ \int M(\pi | \tau) dG_{00}(\tau) \right] d\pi
\]

\[
+ \int f(Y | \pi) \sum_{r=1}^{n_{\tau^*}} M(\pi | \tau_r) d\pi \tag{5.8}
\]

\[
\propto \int f(Y | \pi) \sum_{r=1}^{n_{\tau^*}} M(\pi | \tau_r) d\pi \tag{5.9}
\]
The proportionality constant here is equal to \((a_{00} + n_{\pi^*})^{-1}\), as the marginal is a mixture of \(n_{\pi^*}\) densities with a weight of 1 and a density weighted by \(a_{00}\).

The integral on line (5.9) is a sum of standard conjugate forms. When the conditional distribution \(M(\pi|\tau_r) = M(\mu, v|m_r, w_r)\) has the following definition. Given \(w_r\), the distribution of \(v\) is \(\text{Gamma}(\alpha, \beta/w_r)\). Then, \(\mu|v, m_r \sim N(\mu|m_r, 1/(\Delta_2 v))\), where \(\Delta_2\) is a known scalar. With this definition of \(M(\pi|\tau_r)\), the hierarchical model underlying line (5.9) is

\[
Y|\mu, v \sim N(Y|\mu, 1/v) \\
\mu|m_r, v_r \sim N(m_r, 1/(\Delta_2 v)) \\
v_r|w_r \sim \text{Gamma}(v_r|\alpha, \beta/w_r).
\]

The marginal distribution of \(Y\) is then a t-distribution with \(2\alpha\) degrees of freedom. mode \(m_r\) and precision \(X = \frac{\alpha w_r \Delta_2}{\beta (1 + \Delta_2)}\).

\[
Y \sim T_{2\alpha} \left( m_r, \frac{\alpha w_r \Delta_2}{\beta (1 + \Delta_2)} \right)
\]

The marginal density corresponding to atom \(\tau_r\) is

\[
f_r(Y) = \int f(Y|\pi)M(\pi|\tau_r)d\pi
\]

\[
= \frac{X^{\frac{3}{2}} \Gamma[\alpha + 1/2]}{(2\alpha \pi)^{\frac{3}{2}} \Gamma(\alpha)} \left[1 + \frac{X}{2\alpha (Y - m_r)^2}\right]^{-\alpha - 1/2}.
\]

Details of this derivation are given in Appendix D.

Refer to the contribution of \(G_{00}\) to the marginal density of \(Y\) (line 5.8) as \(f_{00}(Y)\).

\[
f_{00}(Y) = \int f(Y|\pi) \left( \int M(\pi|\tau)dG_{00}(\tau) \right) d\pi.
\]

The density \(f_{00}(Y)\) is evaluated partly analytically and partly using a Gaussian quadrature routine (this also is shown in Appendix D). The marginal density of \(Y\) under the prior \(G_0(\pi)\) is the weighted sum

\[
f_0(Y) = \frac{a_{00} f_{00}(Y) + \sum_{r=1}^{n_{\pi^*}} f_r(Y)}{a_{00} + n_{\pi^*}}.
\]
The value of the marginal evaluated at \( Y_{ij} \), \( f_0(Y_{ij}) \) is used to compute the probabilities needed for sampling \( \pi_{ij}|Y_{ij} \). A computational shortcut is available here. Since the number of distinct values in \( T \) will be smaller than \( n_\pi \), the sum is calculated as the weighted sum of t-distributions at the unique values. Further details concerning the number of unique values in \( T \) can be found in Chapter 6.

### 5.2 Sampling From the Baseline

**Posterior Distribution** \( G_0(\pi_{ij}|Y_{ij}) \)

By a standard application of Bayes' rule,

\[
G_0(\pi_{ij}|Y_{ij}) = \frac{f(Y_{ij}|\pi_{ij})G_0(\pi_{ij})}{f_0(Y_{ij})},
\]

where in this model

\[
f(Y_{ij}|\pi_{ij}) = N(Y_{ij}|\mu_i, 1/v_i)
\]

and

\[
G_0(\pi_{ij}) \propto \alpha_0 \int M(\pi_{ij}|\tau) dG_{00}(\tau) + \sum_{r=1}^{n_\pi} M(\pi_{ij}|\tau_r)
\]

The distribution \( G_0 \) is a mixture of normal-gamma distributions at the atoms \( \tau_r \) and a normal-gamma distribution convolved with the baseline prior \( G_{00} \). Therefore, the posterior \( G_0(\pi_{ij}|Y_{ij}) \) is a mixture of posteriors centred at the atoms \( \tau_r \) and the posterior for the baseline prior. Expanding each of these components in terms of normal and gamma densities gives the following.

\[
f(Y_{ij}|\pi_{ij}) \sum_{s=1}^{n_\pi} M(\pi_{ij}|\tau_r)
\]

\[
= N(Y_{ij}|\mu_i, 1/v_i) \sum_{r=1}^{n_\pi} N(\mu_i|m_r, 1/(\Delta_2v_i))\text{Gamma}(v_i|\alpha, \frac{\beta}{w_r}) \quad (5.11)
\]

\[
f(Y_{ij}|\pi_{ij}) \int M(\pi_{ij}|\tau)dG_{00}(\tau)
\]

\[
= N(Y_{ij}|\mu_i, 1/v_i) \int N(\mu_i|m, 1/(\Delta_2v_i))\text{Gamma}(v_i|\alpha, \frac{\beta}{w})
\]

\[
\times N(m_0|1/(\Delta_1w))\text{Gamma}(w|a, b) \, dm \, dw \quad (5.12)
\]
Since \( m_r \) and \( w_r \) are considered fixed, equation (5.11) is a sum of \( n_{*r} \) standard conjugate forms. To recap, the relevant portion of the hierarchical model is

\[
Y_{ij}|\mu_{ij}, v_{ij} \sim N(Y_{ij}|\mu_{ij}, 1/v_{ij})
\]

\[
\mu_{ij}|v_{ij} \sim N(\mu_{ij}|m_r, 1/(\Delta_2 v_{ij}))
\]

\[
v_{ij} \sim \text{Gamma}(v_{ij}|\alpha, \beta/w_r).
\]

This gives the following normal-gamma posterior for \( \pi_{ij}|Y_{ij}, \tau_r \) [7].

\[
\mu_{ij}|v_{ij}, Y_{ij}, m_r \sim N\left(\mu_{ij}\frac{\Delta_2 m_r + Y_{ij}}{\Delta_2 + 1}, 1/((\Delta_2 + 1)v_{ij})\right)
\]

\[
v_{ij}|Y_{ij}, w_r, m_r \sim \text{Gamma}\left(v_{ij}|\alpha + 1/2, \frac{1}{w_r} + \frac{\Delta_2(Y_{ij} - m_r)^2}{2(\Delta_2 + 1)}\right).
\] (5.13)

So, equation (5.11) becomes a sum of \( n_{*r} \) normal-gamma densities, corresponding to the \( n_{*r} \) atoms \( T \).

Equation (5.12) above can be somewhat simplified by integration over \( m \). The details can be found in Appendix E. However, the resulting integral in \( w \) is not analytically tractable. Two approximate approaches were tried for obtaining samples from this integral.

\[
f(Y_{ij}|\pi_{ij}) \int M(\pi_{ij}|\tau)dG_{00}(\tau)
\]

\[
= f(Y_{ij}|\pi_{ij}) \int N(\mu_{ij}|m, 1/((\Delta_2 v_{ij}))\text{Gamma}(v_{ij}|\alpha, \beta/w)\]

\[
\times N(m|m_0, 1/(\Delta_1 w))\text{Gamma}(w|a, b)dmdw
\]

\[
= \frac{\psi_{ij}^a \psi_{ij}^a \sqrt{\Delta_1 \Delta_2}}{\Gamma(a)\Gamma(\alpha)(2\pi)} \exp\left(-\frac{v_{ij}}{2}(Y_{ij} - \mu_{ij})^2\right)
\]

\[
\times \int \frac{1}{\sqrt{\Delta_2 v_{ij} + \Delta_1 w}} \exp\left[-\frac{1}{2(\Delta_1 w + \Delta_2 v_{ij})}(\mu_{ij} - m_0)^2 - \frac{\beta v_{ij}}{w} - bw\right] w^{a-\alpha-\frac{1}{2}} dw
\]

First, a Gibbs sampling routine was applied to the trivariate density \( f(\mu, v, w|Y) \) that comes from dropping the integration over \( w \). After making a transformation to \( (\mu, r, s) \), it was possible to obtain conditional distributions \( f(\mu|r, s, Y) \) (a normal distribution) and \( f(r|\mu, s, Y) \) (a gamma distribution) that could be easily sampled.

The distribution \( f(s|\mu, r, Y) \) resembled a log-concave density (see appendix B) and
was treated as one. Using these three full conditional distributions, a short-run Gibbs sampler yielded an approximate sample from $f(\mu, r, s|Y)$ which was back-transformed to $(\mu, v, w)$. Integration over $w$ was approximated by using only $\mu$ and $v$ from this triplet.

There were two problems with this method of approximation. First, there was the practical problem of the number of cycles to run the Gibbs sampler for. This Gibbs sampler routine was called many thousands of times by the main routine, so to keep computational time down, it was not feasible to run a large number of cycles. For the same reason, it was not feasible to assess convergence of this sub-chain. Secondly, the assumption of log-concavity, for $f(s|\mu, r, Y)$, which depended on changing values of $\mu$ and $r$ was often violated. When this happened, sampled values of $s$ were not from the required distribution, casting doubt on the validity of this entire Gibbs sub-sampling routine.

A second approach was much more straightforward and relied on approximating the continuous distribution $G_{00}$ in equation (5.12) with a discrete distribution at a large number ($N$) of points sampled from the baseline prior $G_{00}$. This gives the following approximation, where the $N$ values $\overline{\tau}_k, k = 1, \ldots, N$ are a sample from the baseline prior $G_{00}(\tau)$.

$$f(Y_{ij}|\pi_{ij}) \int_{\tau} M(\pi_{ij}|\tau) dG_{00}(\tau) \approx f(Y_{ij}|\pi_{ij}) \sum_{k=1}^{N} M(\pi_{ij}|\overline{\tau}_k)$$

This approach bears some similarity to the algorithms described by MacEachern and Müller [19, 18] for computations in mixture of Dirichlet process models with non-conjugate baseline priors.

Notice that the approximation gives the same form for this part of the posterior as shown in line (5.11) for the part of the posterior centred at the atoms. The same rules that gave a sample from the posteriors at the atoms can now be used for obtaining a sample from the approximate baseline posterior. Equation (5.13) gives the sampling rule, except the atoms $\tau_r$ are replaced by the randomly sampled $\overline{\tau}_k$. When $\pi_{ij}|Y_{ij}$ is to be sampled from the part of the baseline prior in line (5.12), there are two steps.
First, the probability that $\pi_{ij}$ is sampled from the posterior at $\bar{\pi}_k$ is calculated for each $\bar{\pi}_k$. These calculations use the t-distribution formula (5.10) to calculate weights and normalize by the sum of the weights. Secondly, once a particular $\bar{\pi}_k$ has been chosen (using multinomial sampling), $\pi|Y, \bar{\pi}_k$ is sampled from the normal-gamma posterior (5.13), as if the parameters for the prior were set by $\bar{\pi}_k$. Once a particular $\bar{\pi}_k$ is chosen, it is replaced by a new sample from $G_{00}$. Otherwise, to reduce the number of computations, the set of $\bar{\pi}_k$ is reused. Section 7.3 discusses posterior analysis for the hyperparameters of $G_{00}$ and how the values of $\bar{\pi}$ are modified when the values of the hyperparameters change.

### 5.3 Which Part of $G_0(\pi_{ij}|Y_{ij})$ to Sample

The components of $f_0(Y_{ij})$ evaluated in section 5.1 are used to calculate the probabilities of sampling from the baseline posterior component (5.11) or one of the posterior components in the sum (5.12). This is seen by rewriting the posterior $G_0(\pi_{ij}|Y_{ij})$ (and taking some liberties with notation). The conditional $f_{00}(\pi_{ij}|Y_{ij})$ is the posterior distribution of $\pi_{ij}$ if its prior is formed by adding "noise" to $G_{00}$. The distribution $M(\pi_{ij}|\tau_r, Y_{ij})$ is the normal-gamma posterior corresponding to the normal-gamma prior on $\pi_{ij}$ at the atom $\tau_r$.

\[
G_0(\pi_{ij}|Y_{ij}) \propto \alpha_{00} f(Y_{ij}|\pi_{ij}) \int M(\pi_{ij}|\tau)dG_{00}(\tau) + \sum_{\tau} f(Y_{ij}|\pi_{ij}) M(\pi_{ij}|\tau_r) \\
= \alpha_{00} f(Y_{ij}|\pi_{ij}) f_{00}(\pi_{ij}) + \sum_{\tau} f(Y_{ij}|\pi_{ij}) M(\pi_{ij}|\tau_r)
\]

As in section 5.1, let $f_0(Y_{ij}) \propto \alpha_{00} f_{00}(Y_{ij}) + \sum f_r(Y_{ij})$. Then, the rule for deciding which part of $G_0$ to sample from becomes:

\[
\pi_{ij}|Y_{ij} \begin{cases} \\
\sim M(\pi_{ij}|\tau_r, Y_{ij}) \text{ with probability } f_r(Y_{ij})/f_0(Y_{ij}) \\
\sim f_{00}(\pi_{ij}|Y_{ij}) \text{ with probability } \alpha_{00} f_{00}(Y_{ij})/f_0(Y_{ij})
\end{cases}
\]

The probability that a new value of $\pi_{ij}$ is drawn from a particular posterior distribution is proportional to the marginal of $Y_{ij}$ under the corresponding prior.
Sampling from \( f_{00}(\pi_{ij}|Y_{ij}) \) is accomplished through a similar rule. Recall that \( G_{00} \) is approximated by \( N \) independent samples from \( G_{00}: \{\pi_k, k = 1, \ldots, N\} \). Let \( f_{\pi_k}(Y) \) be the marginal distribution of \( Y \) under the candidate value \( \pi_k \). Given that \( \pi_{ij} \) is to be sampled from \( f_{00}(\pi_{ij}|Y_{ij}) \).

\[
\pi_{ij}|Y_{ij} \sim M(\pi_{ij}|\pi_k, Y_{ij}) \quad \text{with probability} \quad \frac{f_{\pi_k}(Y_{ij})}{\sum_{k=1}^{N} f_{\pi_k}(Y_{ij})}.
\]

To summarize, sampling \( \pi_{ij}|Y_{ij} \) from the baseline prior distribution \( G_{00}(\pi_{ij}|Y_{ij}) \) proceeds as follows.

1. Using the rule just derived, decide whether to sample from a normal-gamma posterior at one of the atoms \( \tau_r \) or from the posterior of the \( G_{00} \) component.

2. If an atom \( \tau_r \) is selected, sample from \( M(\pi|\tau_r, Y_{ij}) \).

3. If the posterior of \( G_{00} \) is selected, select one of the \( \pi_k \) with probability proportional to \( f_{\pi_k}(Y_{ij}) \) and sample from the corresponding \( M(\pi|\pi_k, Y_{ij}) \).

### 5.4 Summary

This chapter has presented the methods needed for one update of the parameters \( \Pi_i|Y_i \) and results in each subject being represented by a new configuration. Section 9.1 contains pseudo-code for the algorithm. The next step is to estimate \( G_0 \) based on the new values \( \Pi^* \).
Chapter 6

Updating the Distribution $G_0$

6.1 Hierarchical Model for $\pi^*$

The distribution $G_0$ is the baseline prior for the $n$ Dirichlet processes $G_i \sim \mathcal{D}(\alpha_0, \alpha_0)$. $i = 1, \ldots, n$. To form $G_0$, normal-gamma "noise" is added around the distribution $F$, which is itself sampled from a Dirichlet process. The Dirichlet process prior on $F$ places a prior distribution on $G_0$. This chapter shows how $G_0$ can be updated, given the data in the $G_i$.

Assume that samples of $\Pi_i | Y_i$, $i = 1, \ldots, n$ are available through use of the algorithm of Chapter 5. Let $n_{\pi^*}$ be the number of distinct values among the $\Pi_i$ for all subjects and let $\Pi^* = \{\pi_i^*, r = 1, \ldots, n_{\pi^*}\}$ be those distinct values. The distinct set of values $\Pi^*$ are used to estimate the shared baseline distribution for the Dirichlet posteriors $G_i | Y_i$.

The hierarchical model, treating $\Pi^*$ as data in a density estimation problem, is as follows. As before, the shorthand $\pi$ will be used for $(\mu, v)$ and $\tau$ will be used for the mean precision pair $(m, w)$.

\[
\begin{align*}
\pi_i^* | \tau_r & \sim M(\pi_i^* | \tau_r) \\
\tau | F & \sim F(\tau) \\
F | G_{00} & \sim \mathcal{D}(G_{00}, \alpha_{00})
\end{align*}
\]
Each unique value $\pi^*_r$ has a normal-gamma prior distribution with parameters given by $\tau_r$. Given $F$, the values of $\tau_r$ have the distribution $F$. This distribution $F$ is random with the prior distribution a Dirichlet process. Specifically, the distribution $G_{00}$ is a normal-gamma distribution which generates values of $\tau$ as follows:

$$w \sim \text{Gamma}(w|a,b)$$
$$m|w,\Delta_1, m_0 \sim N(m|m_0, 1/(\Delta_1 w)). \quad \Delta_1 > 0. \quad (6.1)$$

The conditional density $M(\pi|\tau)$ has the following meaning:

$$v|w \sim \text{Gamma}(v|\alpha, \beta/w)$$
$$\mu|m, v, \Delta_2 \sim N(\mu|m, 1/(\Delta_2 v)). \quad \Delta_2 > 0. \quad (6.2)$$

The expected value of $\mu$ given $m$ is $m$. The parameters $\alpha$ and $\beta$ determine the relationship between $v$ and $w$—both the expected value and variance of $v$ given $w$. Given $w$, the expected value and variance of $v$ are given by

$$E(v|w) = \frac{\alpha}{\beta w}$$
$$\text{Var}(v|w) = \frac{\alpha}{\beta^2 w^2}.$$

The joint distribution of $v$ and $w$ when $w$ has a Gamma(5,5) distribution is shown in figure 6.1 for various values of $\alpha$ and $\beta$. When $\alpha = \beta = 1$, $v|w$ has an exponential distribution which, although it has a mean of $w$, places most of its mass on values near 0. The values $\alpha = \beta = 2$ represent a slightly stronger association between $v$ and $w$, but $v$ is still quite dispersed for a given $w$. The values $\alpha = \beta = 10$ represent a strong relationship between $v$ and $w$, with most mass in their joint density lying along the line $v = w$. Finally, $\alpha = 5$ and $\beta = 10$ represents a relationship where $v|w$ is about half as big as $w$ and not very dispersed. For values of $v$ to be near values of $w$, $\alpha = \beta$ and $\alpha \gg 1$.

Further discussion of the interpretation of priors is deferred to section 9.2.
Figure 6.1: The Joint Density $f(v, w)$

- $\alpha=1; \beta=1$
- $\alpha=2; \beta=2$
- $\alpha=10; \beta=10$
- $\alpha=5; \beta=10$
6.2 Sampling Distribution of $\tau_r | \pi_r^*$

As before, the set of the $\tau_r, r = 1, \ldots, n_{\pi^*}$ is $T$: the same set without $\tau_r$ is $T^{-r}$. The general formula for updating in the Gibbs sampling scheme (3.7) is easily rewritten for this model:

$$f(\tau_r | T^{-r}, \pi_r^*) = \frac{\alpha_{00} f_{00}(\pi_r^*) G_{00}(\tau_r | \pi_r^*) + \sum_{i \neq r} M(\pi_r^* | \tau_i^*) \delta_{\tau_r}(\tau_r)}{\alpha_{00} f_{00}(\pi_r^*) + \sum_{i \neq r} M(\pi_r^* | \tau_i)}. \quad (6.3)$$

The denominator is just the normalizing constant. formed by integrating the numerator with respect to $\tau_r$. The term $f_{00}(\pi_r^*)$ is the baseline marginal density of $\pi^*$ evaluated at $\pi_r^*$. It is computed as the integral with respect to $G_{00}$ of the likelihood of $\pi^*$, namely

$$f_{00}(\pi_r^*) = \int M(\pi_r^* | \tau) dG_{00}(\tau)$$

The other term in the denominator. $\sum_{i \neq r} M(\pi_r^* | \tau_i^*)$. is the integral of the sum in the numerator with respect to $\tau$. Writing this in the form that reflects that there are $n_{\pi^*} < n_{\pi}$ distinct values among the $T$ gives

$$f(\tau_r | T^{-r}, \pi_r^*) \propto \alpha_{00} f_{00}(\pi^*) G_{00}(\tau_r | \pi_r^*) + \sum_{s=1}^{n_{\pi^*}} M(\pi_r | \tau_s^*) n_s^* \delta_{\tau_r^*}(\tau_r). \quad (6.4)$$

Using the same notation as before. $\tau_s^*, s = 1, \ldots, n_{\pi^*}$ are the $n_{\pi^*}$ distinct values in $T$ and $n_s^*$ is the number of $\tau_r$ sharing the common value $\tau_s^*$. Again. there is one slight modification to this formula that is stated here to reduce any further complication in the notation: when $r \in S_s$, $n_s^*$ is replaced by $n_s^* - 1$.

Sampling from the distribution in line 6.4 proceeds according to the rule:

$$\begin{align*}
\tau_r | T^{-r}, \pi_r^* = \tau_s^* & \quad \text{w. prob } \frac{n_s^* M(\pi_r^* | \tau_s^*)}{\alpha_{00} f_{00}(\pi_r^*) + \sum_{s=1}^{n_{\pi^*}} n_s^* M(\pi_r^* | \tau_s^*)} \\
\sim G_{00}(\tau_r | \pi_r^*) & \quad \text{w. prob } \frac{\alpha_{00} f_{00}(\pi_r^*)}{\alpha_{00} f_{00}(\pi_r^*) + \sum_{s=1}^{n_{\pi^*}} n_s^* M(\pi_r^* | \tau_s^*)}
\end{align*}$$

With probability proportional to $n_s^*$ times the likelihood $M(\pi_r^* | \tau_s^*)$, $\tau_r$ is sampled to be equal to $\tau_s^*$. With probability proportional to $\alpha_{00}$ times the baseline marginal distribution of $\pi^*$ evaluated at $\pi_r^*$, $\tau_r$ is sampled from the baseline posterior $G_{00}(\tau_r | \pi_r^*)$. 

To be able to use the Escobar algorithm, it must be possible to calculate $f_{00}(\pi^*_r)$ and sample from the posterior density $G_{00}(\tau|\pi^*_r)$. The next two sections show how to do this.

### 6.3 Computing the Marginal Density $f_{00}(\pi^*)$

This section briefly describes the evaluation of $f(\pi^*)$, which is defined as

$$f(\pi^*) = \int M(\pi^*|\tau) dG_{00}(\tau).$$

Using the specification of $G_{00}(\tau)$ and $M(\pi^*|\tau)$ described in the previous section, this marginal density is evaluated as

$$f(\pi^*) = \int_0^\infty \int_{-\infty}^{\infty} N(\mu^*|m, \Delta_2 v^*) \text{Gamma}(v^*|\alpha, \beta/w) N(m|m_0, 1/(\Delta_1 w)) \text{Gamma}(w|a, b) \, dm \, dw$$

The two normal densities combine to give a normal density in $m$ which can be integrated over $(-\infty, \infty)$. The remaining integral in $w$ is approximated by Gaussian quadrature using Laguerre weights (see Appendix A).

Appendix D shows the details of the integration. The values $\pi^*_r$ at which $f(\pi^*)$ is evaluated may change after each cycle of sampling $\pi_{ij}|Y_{ij}$. Furthermore, Chapter 7 introduces posterior analysis for the parameters $\Delta_1$, $\Delta_2$, and $m_0$. Even for values in $\Pi^*$ that do not change, changing values of the hyperparameters mean that the marginal distribution $f(\pi^*)$ must be evaluated at the $n_{\pi^*}$ values in $\Pi^*$ before the values $\tau_r|\pi^*_r$ can be resampled.

### 6.4 Sampling from the Baseline

**Posterior Distribution** $G_{00}(\tau_r|\pi_r^*)$

This section shows how the baseline prior $G_{00}(\tau)$ (6.1) and the likelihood $M(\pi^*|\tau)$ (6.2) combine to give the baseline posterior $G_{00}(|\pi^*)$. All subscripts on $\pi^*, \tau, \mu^*, v^*, m$, and $w$ are the same, ranging over the values $1, \ldots, n_{\pi^*}$, so subscripts are omitted to simplify
notation. Details of the computations that lead to the conditional distributions below are shown in Appendix E.

\[ G_00(\tau|\pi^*) = \frac{M(\pi^*|\tau)G_00(\tau)}{f(\pi^*)} \times N(\mu^*|m_1/(\Delta_2 v^*))\text{Gamma}(v^*|\alpha, \beta/w)N(m|m_0, 1/(\Delta_1 w))\text{Gamma}(w|\alpha, \beta) \times f(m|w, \pi^*)q(w|\pi^*) \]  

(6.5)

The joint density of \( m \) and \( w \) given \( \pi^* \) is proportional to the product of two densities. Sampling \( \tau|\pi^* \) can proceed in two steps. First, \( w \) is sampled from its density, proportional to

\[ q(w|\pi^*) = \exp \left[ -\frac{1}{2} \frac{\Delta_1 \Delta_2 v^* w (\mu^* - m_0)^2 - bw - \beta v^*}{w} \right] w^{a - \frac{1}{2} - \alpha}(A)^{-\frac{1}{2}}. \]  

(6.6)

where \( A = \Delta_2 v^* + \Delta_1 w \). Then, given \( w \), \( m \) has the density \( N(m|\frac{\Delta_2 v^* A}{\Delta_2 v^* + \Delta_1 w m_0}, A) \):

\[ f(m|w, \pi^*) \propto \sqrt{A} \exp \left[ -\frac{A}{2} \left( m - \frac{\Delta_2 v^* \mu^* + \Delta_1 w m_0}{A} \right)^2 \right]. \]

6.4.1 Sampling \( w \)

The density \( q(w|\pi^*) \) is non-standard - no ready-made algorithms exist for sampling \( w \). The density is unimodal and looks somewhat like a gamma density. In fact, as \( w \to \infty \), if \( a - \alpha > 0 \), it behaves like a \( \text{Gamma}(w|a - \alpha, b) \). If \( a - \alpha < 0 \), the limiting density has a right tail that is dominated by the exponential \( \exp\{-|bw|/w\} \).

\[ \lim_{w \to \infty} q(w|\mu^*, v^*) = \exp \left[ -\frac{1}{2} \Delta_2 v^*(\mu^* - m_0)^2 - bw \right] w^{a - \frac{1}{2} - \alpha}(\Delta_1 w)^{-\frac{1}{2}} \times \exp\{-bw\} w^{a - \alpha - 1} \]

The limit as \( w \to 0 \) is also a known density if \( \alpha - a > \frac{1}{2} \), the inverse gamma distribution, \( IG(\alpha - a - \frac{1}{2}, \beta v^*) \)

\[ \lim_{w \to 0} q(w|\mu^*, v^*) = \exp\{-\beta v^*/w\} w^{a - \alpha - \frac{1}{2}}. \]

It is encouraging that \( q(w) \) behaves well in some limiting cases, but sampling from \( q(w) \) poses some difficulties. In Devroye's comprehensive 1986 monograph [8], there
is an algorithm for simulating random variables from a log-concave density. Details of the algorithm are given in Appendix B. Let \( h(w) = \log |q(w)| \). Then log-concavity of \( q(w) \) over a region is equivalent to \( h''(w) < 0 \) over the whole region. Unfortunately, there is no guarantee that \( q(w) \) is log-concave over \((0, \infty)\). The second derivative \( h''(w) \) has a numerator that is a fourth-degree polynomial in \( w \) and a denominator that is always positive; the sign of the numerator determines whether \( h''(w) \) is negative.

\[
 h(w) = -\frac{1}{2} \frac{\Delta_1 \Delta_2 v^* w}{\Delta_2 v^* + \Delta_1 w} (\mu^* - m_0)^2 - bw \\
 -\beta \frac{v^*}{w} + (a - \frac{1}{2} - \alpha) \log(w) - \frac{1}{2} \log(\Delta_2 v^* + \Delta_1 w) \\
 h'(w) = -\frac{\Delta_1 \Delta_2 v^*^2 (\mu^* - m_0)^2}{2(\Delta_2 v^* + \Delta_1 w)^2} - b + \beta \frac{v^*}{w^2} + \frac{a - \frac{1}{2} - \alpha}{w} - \frac{\Delta_1}{2(\Delta_2 v^* + \Delta_1 w)} \\
 h''(w) = \frac{N(w)}{D(w)} \text{ where } D(w) > 0 \text{ and} \\
 N(w) = \left(1 - 2(a - \frac{1}{2} - \alpha)\right) c_1^3 w^4 \\
 + \left(2c_2 v^*(\mu^* - m_0)^2 + \Delta_2 - 4\Delta_1 \beta - 6\Delta_2 (a - \frac{1}{2} - \alpha)\right) c_1^2 v^* w^3 \\
 - \left(12\Delta_1 \beta + 6\Delta_2 (a - \frac{1}{2} - \alpha)\right) \Delta_1 \Delta_2 v^*^2 w^2 \\
 - \left(12\Delta_1 \beta + 2\Delta_2 (a - \frac{1}{2} - \alpha)\right) c_2^2 v^*^3 w \\
 - 4c_2^3 \beta v^*^4
\]

The coefficients of the polynomial are determined both by parameters of the priors (some of which are resampled as part of the Gibbs sampling process) and by sampled values of \( \mu^* \) and \( v^* \). If \( a - \alpha - \frac{1}{2} > \frac{1}{2} \), then only the coefficient of \( w^3 \) is not guaranteed to be negative. The relative sizes of the coefficients for the other powers of \( w \) determine whether the value of the entire polynomial is negative. For certain values of \( a \) and \( \alpha \), the log-concave algorithm can be used for \( w \).

The value of \( a - \alpha - \frac{1}{2} \) will typically be negative for the model in this thesis. The values of \( a \) and \( b \) will be small to give weight over a large region of the prior space of \( w \). The values of \( \alpha \) and \( \beta \) will typically be large, so that values of \( v|w \) are near \( w \). The fact that the exponent of \( w \) in equation (6.6) will be negative suggests that the
density of \( z = 1/w \) might be more amenable to simulation.

Let \( r(z) \) be the density of \( z = 1/w \). Then, with \( h(z) = \log(r(z|\pi^*)) \), the derivative information is as follows

\[
\begin{align*}
    h(z) &= -\frac{1}{2} \frac{\Delta_1 \Delta_2 v^*}{\Delta_2 v^* z + \Delta_1} (\mu^* - m_0)^2 - b\frac{1}{z} - v^* z + (\alpha - a - 1) \log(z) - \frac{1}{2} \log(\Delta_2 v^* z + \Delta_1) \\
    h'(z) &= \frac{1}{2} \frac{\Delta_1 \Delta_2 v^*^2 (\mu^* - m_0)^2}{(\Delta_2 v^* z + \Delta_1)^2} + b\frac{1}{z^2} - v^* z + \frac{\alpha - a - 1}{z} - \frac{1}{2} \frac{\Delta_2 v^*}{\Delta_2 v^* z + \Delta_1} \\
    h''(z) &= \frac{N(z)}{D(z)} \text{ where } D(z) > 0 \text{ and } \\
    N(z) &= -\left\{ 2(\alpha - a - 1) - 1 \right\} \Delta_2^3 v^* z^4 \\
    &\quad - \left\{ 2\Delta_1 \Delta_2 v^* (\mu^* - m_0)^2 + 4b\Delta_2 v^* + \Delta_1 (6(\alpha - a - 1) - 1) \right\} \Delta_2^2 v^* z^3 \\
    &\quad - \left\{ 12b\Delta_2 v^* + 6(\alpha - a - 1)\Delta_1 \right\} \Delta_1 \Delta_2 v^* z^2 \\
    &\quad - \left\{ 12b\Delta_2 v^* + 2\Delta_1 (\alpha - a - 1) \right\} \Delta_1^2 z \\
    &\quad - 4b\Delta_1^3 
\end{align*}
\]

For this second derivative to be negative for all values of \( z \), it is sufficient that \( \alpha - a > \frac{7}{6} \). This is only a minor restriction on these hyperparameters, since \( \alpha > a \) is consistent with the structure of the model. The algorithm for log-concave densities can be used to generate \( z \) from \( r(z) \) and then \( w = 1/z \). The only information needed for the log-concave algorithm that is not shown above is the mode of \( r(z) \). Finding the mode of \( r(z) \) by setting \( h'(z) = 0 \) involves solving a fourth degree polynomial in \( z \). Instead, a numerical search algorithm, using MNBRAK and GOLDEN [22], was used to locate the mode.

### 6.5 Remixing \( T \)

In this model, \( \alpha_{\infty} \) tends to be small (see section 9.2), so there will typically be few distinct values in \( T \), each with substantial weight. There will be low probability that sampling from \( \tau_r|\pi_r^* \) will yield a new value. If \( T \) gets stuck on the same values
from iteration to iteration, then convergence to the posterior distribution is slowed down. One way to alleviate this problem is to resample $T$ conditional on the configurations [6]. This has been called “remixing”. Let $S_s$ be the set of indices of the members of $\Pi^* \tau$ with the common parameter value $\tau^*_s$. Then $\Pi^*_s$ will refer to the set $\{\tau^*_s : r \in S_s\}$. Conditional on the elements of $\Pi^*_s$, $\tau^*_s$ is resampled from the baseline posterior $G_{\Pi^*_s}(\tau^*_s|\Pi^*_s)$.

It is possible to factor the joint posterior distribution of $(m^*_r, w^*_r|\Pi^*_s)$ into the product of a normal distribution in $m^*_r|w^*_r$ and a distribution in $w^*_r$ similar to $q(w)$, equation (6.6). However, sampling from the distribution in $w^*_r$ is subject to the same problem as sampling from $q(w)$, namely the possibility of non-log-concavity. In fact, the parameters of this distribution vary much more widely than those of $q(w)$ and the log-concavity criterion is often violated.

While the joint conditional distribution is difficult to sample, the individual conditionals of $(m^*_r|\Pi^*_s)$ and $(w^*_r|\Pi^*_s)$ are readily available. A short-run Gibbs subsampler iterated over these two conditionals will yield an approximate draw from the joint conditional. Since the number of distinct values $n_{\tau^*}$ is small, and the remixing is done only once per update of $G_{0}$, the additional computational time is not an issue. The joint conditional and individual conditional densities are shown in Appendix F.

### 6.6 Summary

This chapter has shown how to obtain a sample of $T$ conditional on the current values in $\Pi^*$. The prior distributions are not of standard conjugate form, so computations of the required marginals and posteriors are not straightforward. Gaussian quadrature (with Laguerre weights) is used for computing the marginal distribution $f_{00}(\pi^*)$. The posterior distribution of the $w$ component of the parameter $\tau$ does not have a log-concave density, but its reciprocal $z = 1/w$ does, so can be simulated through an efficient rejection algorithm. To alleviate concerns that the parameters $T$ might get stuck on a the same values, the unique values are resampled based on the
configurations.
Chapter 7

Additional Learning About Hyperparameters

7.1 Learning About $\alpha_0$ and $\alpha_{00}$

The parameters $\alpha_0$ and $\alpha_{00}$ are crucial to the multiple sample model. The treatment of the parameter $\alpha_{00}$ is exactly the same as in EW. A gamma distribution prior is placed on the parameter $\alpha_{00}$; posterior analysis is based on the number of distinct elements in $T$ after updating $T|\Pi^*$ and involves data augmentation (see section 3.5.1).

It is not so clear how $\alpha_0$ should be treated, as different subjects may have different numbers of distinct elements among their $\Pi_i$. The simplest approach is to obtain a new sample of $\alpha_0$ after each updating of $\Pi_i|Y_i$. A modification of this standard approach is investigated. This places identical but independent gamma prior distributions on $\alpha_0$ for each subject $i$, so that the Dirichlet process priors are now

$$G_i \sim \mathcal{D}(G_{00}, \alpha_0^i).$$

Since the meaning of $\alpha_0$ as a precision parameter depends on the sample size, this allows $\alpha_0$ to adapt to the particular sample size of each subject. Separate $\alpha_0$ parameters for each subject also allow varying closeness of each subject’s $G_i$ to the baseline prior $G_0$. Significant differences between the posterior for $\alpha_0^i$ and the posteriors for
the other subjects may indicate that subject \( i \) is different from the others.

### 7.2 Learning About \( \Delta_2 \)

The parameter \( \Delta_2 \) is the ratio of the precision of \( \mu|m \) to the precision of \( Y|\mu \). As such, it relates to the smoothness of the predictive density. In the complicated model in this thesis, it is difficult to specify the amount of smoothing purely through \( \Delta_2 \). The number of distinct components in each sampled \( G_i \), as well as the spread of the elements in \( G_0 \) also enter into the smoothness of the predictive densities. Fortunately, there is much more data here to learn about \( \Delta_2 \) from posterior analysis than there usually is in a density estimation problem.

Conditional on \( \Delta_2 \) and the values of \( T \), the members of \( \Pi^* \) are conditionally independent (and independent of any other parameters in the model), each with a normal-gamma density. Writing only the part of this distribution relevant to \( \Delta_2 \).

\[
\begin{align*}
    f(\Delta_2|T, \Pi^*) &\propto f(\Pi^*|T, \Delta_2)f(\Delta_2) \\
    &\propto \left\{ \prod_{s=1}^{n^*_s} \prod_{j=1}^{n^*_s} \left( \Delta_2 \nu^*_j \right)^{\frac{1}{2}} \exp \left[ -\frac{\nu^*_j \Delta_2}{2} (\mu^*_j - m_s)^2 \right] \right\} f(\Delta_2).
\end{align*}
\]

Now let \( f(\Delta_2) \) be the gamma density Gamma\( (\Delta_2|A_{\Delta_2}, B_{\Delta_2}) \). Notice that \( \sum_{s=1}^{n^*_s} n^*_s = n^*_\pi \), the “sample size” of the \( \Pi^* \). Then the posterior distribution of \( \Delta_2 \), given \( \Pi^* \) and \( T \) is also a gamma distribution.

\[
\begin{align*}
    f(\Delta_2|T, \Pi^*) &\propto \sum_{s=1}^{n^*_s} \Delta_2^{n^*_s - \frac{1}{2}} \exp \left[ \sum_{j=1}^{n^*_s} -\frac{\nu^*_j \Delta_2}{2} (\mu^*_j - m_s)^2 \right] \Delta_2^{A_{\Delta_2} - 1} \exp(-B_{\Delta_2} \Delta_2) \\
    &\propto \text{Gamma}(A_{\Delta_2} + \frac{n^*_\pi}{2}, B_{\Delta_2} + \sum_{s=1}^{n^*_s} \sum_{j=1}^{n^*_s} \frac{\nu^*_j}{2} (\mu^*_j - m_s)^2)
\end{align*}
\]

Sampling from the posterior distribution of \( \Delta_2 \) is easily incorporated into the Gibbs sampling scheme: \( \Delta_2 \) is sampled from the gamma posterior determined by the values of \( \Pi^* \) and \( T \) at the present stage of the Gibbs sampler. Neither the actual values of \( \Delta_2 \) nor its posterior distribution are of primary interest here. Sampling \( \Delta_2 \) from
its posterior distribution ensures that $\Delta_2$ is the right size. and reduces the bias that might result from improper specification of a constant $\Delta_2$.

### 7.3 Learning About $G_{00}$: $\Delta_1$ and $m_0$

#### 7.3.1 Learning About $\Delta_1$

A similar argument can be made about sampling $\Delta_1$. Given $m_0$ and $\Delta_1$, $T^*$, the distinct elements in $T$, are independent samples from the baseline normal-gamma distribution, so:

$$f(\Delta_1|m_0, T^*) \propto f(T^*|m_0, \Delta_1)f(\Delta_1) \propto \left\{ \prod_{r=1}^{n_{**}} (\Delta_1 w_r)^{\frac{1}{2}} \exp \left[ -\frac{\Delta_1 w_r (m_r - m_0)^2}{2} \right] \right\} f(\Delta_1).$$

The gamma prior $\text{Gamma}(\Delta_1|A_{\Delta_1}, B_{\Delta_1})$ for $\Delta_1$ gives a gamma posterior:

$$f(\Delta_1|T, m_0) \propto (\Delta_1)^{\frac{n_{**}}{2}} \exp \left[ -\sum_{r=1}^{n_{**}} \frac{\Delta_1 w_r (m_r - m_0)^2}{2} \right] \Delta_1^{A_{\Delta_1} - 1} \exp(-B_{\Delta_1} \Delta_1) \propto \text{Gamma}(\Delta_1|A_{\Delta_1} + \frac{n_{**}}{2}, B_{\Delta_1} + \sum_{r=1}^{n_{**}} \frac{w_r}{2} (m_r - m_0)^2).$$

As with $\Delta_2$, $\Delta_1$ is sampled from the above conditional distribution determined by the present values of $T$ and $m_0$.

#### 7.3.2 Learning About $m_0$

The same conditional independence arguments that allowed posterior analysis for $\Delta_1$ are easily adapted for posterior analysis of $m_0$. A normal prior $N(m_0|d, 1/D)$ on $m_0$ leads to a normal posterior.

$$f(m_0|T, m_0, \Delta) \propto \exp \left[ -\frac{1}{2} \sum_{r=1}^{n_{**}} w_r \Delta_1 (m_r - m_0)^2 - \frac{D}{2} (m_0 - d)^2 \right]$$

$$\propto \exp \left[ -\frac{\Delta_1 \sum_{r=1}^{n_{**}} w_r + D}{2} \left( m_0 - \frac{\sum_{r=1}^{n_{**}} w_r m_r + D}{\Delta_1 \sum_{r=1}^{n_{**}} w_r + D} \right)^2 \right]$$
7.3.3 Changes to Candidate $\tau$

Section 5.2 describes the method for obtaining samples from the baseline posterior $G_0(\pi_{ij} | Y_{ij})$. Recall that $\pi_{ij} | Y_{ij}$ is sampled either from the distribution $M(\pi_{ij} | \tau_r)$ centred at the atom $\tau_r$ or from $\int M(\pi_{ij} | \tau) dG_0(\tau)$. In the latter case, an approximate method replaces the integral over the prior $G_0$ with a sum over $N$ independent samples $\tau_k, k = 1, \ldots, N$ from the prior. Changes to the values of $m_0$ and $\Delta_1$ change the baseline prior $G_0$ and should be reflected in the values $\tau_k$. Notice that new values of $\Delta_1$ and $m_0$ affect only the distribution of $\bar{m}_k$. It is not necessary to resample $N$ new values of $\bar{m}_k$. If the old and new values of $\Delta_1$ are called $\Delta^{old}_1$ and $\Delta^{new}_1$, and the corresponding values of $m_0$ are called $m^{old}_0$ and $m^{new}_0$, the mean component $\bar{m}_k$ of each candidate value $\tau_k$ is rescaled according to:

$$\bar{m}_k^{new} = m_k^{new} + (\bar{m}_k^{old} - m_k^{old}) \sqrt{\frac{\Delta_1^{old}}{\Delta_1^{new}}}.$$

The precision component does not change.
Chapter 8

Inference With the Model

8.1 Predictive Distributions

The two goals laid out in the introduction were: (1) to present the population average density and the variability from subject to subject; and (2) to assign a measure of typicality to any particular subject. This chapter shows how predictive densities sampled from the model described in the previous chapters can be used to meet both of these goals.

Notice that there are two types of predictive densities. First, there is the predictive density for a new observation on a given subject \( i \). A sample of the density is obtained by conditioning on the values \( \Pi_i^b \) at the current step \( b \) in the Gibbs sampler. The parameters \( \Pi^b = \{ m_0^b, \alpha_0^b, \alpha_0^b, \Delta_1^b, \Delta_2^b \} \) are also held at their current values. The density is evaluated as

\[
    f_i^b(Y_{n_i+1}) \propto \alpha_0^b \rho_0^b(Y_{n_i+1}) + \sum_j N(Y_{n_i+1} | \pi_{ij}^b), \quad (8.1)
\]

This is a mixture of normal distributions with parameters \( \pi_{ij}^b \) and the distribution \( \rho_0^b(Y_{n_i+1}) \), the marginal distribution of \( Y \) under the current value of the baseline prior \( G_0^b(\pi) \), as evaluated in section 5.1. Ignoring the contribution of the marginal distribution, equation (8.1) has similarities to both the kernel density estimate of Chapter 2 and a normal mixture density estimate. If there are \( n_i \) distinct elements

72
in $\Pi_{b}^{i}$, then the sum $\sum_{j}^{n} N(Y_{n+1} | \pi_{ij}^{b})$ acts like a kernel density estimate, except that the window width is not fixed, being given by $\sqrt{1/\theta_{ij}^{b}}$. If there are few distinct values in $\Pi_{b}^{i}$, then the sum is better thought of as a normal mixture density estimate.

As the Gibbs sampler progresses, variability in the sampled predictive densities comes from two sources. Variability in the estimate of the parent distribution is reflected in changing values of the marginal distribution $p_{0}^{b}(Y_{n+1})$: variability in the estimate for subject $i$ is reflected in the changing values of $\pi_{ij}^{b}$. Of course, these are related, as the distinct values in $\Pi_{i}^{b-1}$ play a role in determining $G_{0}^{b}$ (and therefore $p_{0}^{b}$) and the distinct values in $\Pi_{b}^{i}$ are sampled from $G_{0}^{b}$.

The average predictive distribution for subject $i$ is the average over all $B$ Gibbs samples:

$$f_{i}^{B}(y) = B^{-1} \sum_{b=1}^{B} f_{i}^{b}(Y_{n+1}).$$

The sampled predictive densities $f_{i}^{b}(Y_{n+1})$ are analogous to the kernel density estimates based on the bootstrap samples $(f_{b}^{*})^{i}$ in Chapter 2, as they indicate the uncertainty in the estimation of the density for subject $i$. The average predictive distribution for subject $i$ is the Bayesian density estimate, so it is analogous to the kernel density estimate based on the observed data for subject $i$.

Averaging these individual mean predictive distributions across the $n$ subjects gives the average observed predictive distribution.

$$\overline{f_{obs}^{B}} = n^{-1} \sum_{i=1}^{n} f_{i}^{B}(y).$$

As with the average kernel density estimate, this average density might not have a useful interpretation as the likely density of any subject.

A second type of predictive density is the density of an observation on a new subject. Conditioning on the current value of the parent distribution $G_{0}^{b}(\pi)$ and the current value of the parameters $H_{b}$, a distribution $G_{\text{new}}$ for a new subject is drawn from the Dirichlet process:

$$G_{\text{new}} \sim \mathcal{D}(\alpha_{0}^{b}, G_{0}^{b}).$$

(8.2)
Section 8.3 presents an approximate method of obtaining a sample of $G_{new}$. The sampled predictive distribution $f_{new}^b(Y)$ is formed through integration over $G_{new}$.

$$f_{new}^b(Y) = \int f(Y|\pi)dG_{new}(\pi) \quad (8.3)$$

Densities generated this way for a given $G_0^b$ will be alike to the extent that $\alpha_0^b$ is large. All differences between them are governed by the Dirichlet sampling process in (8.2). In generating densities for new subjects at step $b$ of the Gibbs sampler, this Dirichlet process represents all the variation of the individual densities around the current value of the average density.

This is in contrast to densities generated at different steps of the Gibbs sampler. As the Gibbs sampler progresses, and $G_0^b$ varies, densities generated from equations (8.2) and (8.3) for different $b$ will exhibit two sources of variation. The first is the variation due to the estimation of $G_0$. The other is the Dirichlet process variation described in the previous paragraph. In the implementation of the algorithm in this thesis, a single $f_{new}^b(Y)$ was generated for a given $b$. This means that no two generated $f_{new}^b(Y)$ share the same sample of the parent distribution and ensures that any subset of the sampled $f_{new}^b(Y)$ incorporates both sources of variation in the posterior distribution.

The average predictive density for a new subject is the average over the $B$ Gibbs samples: $\overline{f_{new}^b} = B^{-1} \sum_{b=1}^{B} f_{new}^b(Y)$. The calculated functions $\overline{f_{new}^b}$ and $\overline{f_{obs}}$ are very close and converge to the same value. To reduce correlation between generated predictive distributions, the Gibbs sampler can be run $k - 1$ steps between generating new densities. Then the sums above are over $b' = k, 2k, \ldots, B/k$. This also saves some computational cost, as generating $G_{new}$ and evaluating integral (8.3) are computationally intensive.

Note that it is possible to sample densities for new subjects in the Dirichlet process model because the distribution of densities has an explicit and parametric representation. In the bootstrap analysis of Chapter 2, the distribution of densities was considered to be

$$f_i \sim G_0.$$
However, since $G_0$ was estimated (nonparametrically) by its empirical distribution — for the reaction time data, the set of 17 density estimates — there were only 17 distinct values for a density to act as $f^\text{new}$. This is somewhat limiting if interest lies in the densities that fall in the extreme of the distribution. In fact, Efron and Tibshirani [9] cite estimation of the distribution of a maximum as a situation where the nonparametric bootstrap fails.

### 8.2 Measuring Dissimilarity

The approach to summarizing the results will be the same as that introduced in section 2.2 and is based on ranking densities in order of typicality. For the purposes of outlier detection, a useful measure of the typicality of a sampled density function is one which measures how far the function is from the mean of all the sampled functions. Many metrics exist which measure the distance between density functions and in any particular application, a distance measure would be chosen for application-specific reasons. In this thesis, since there is no particular scientific question to be answered, the Hellinger metric is used, largely on the grounds that it has been used in proofs of convergence for models using Dirichlet processes [2].

To measure the distance between two densities $p(x)$ and $q(x)$ with respect to some density $h(x)$, calculate

$$D(p, q) = \left[ \int \left( \sqrt{p(x)} - \sqrt{q(x)} \right)^2 h(x) \, dx \right]^{\frac{1}{2}}.$$  

The function $h(x)$ acts as a weighting function for the difference between the two densities. Consider the situation where $p(x)$ is a known density and the question is: how close is $q(x)$ to $p(x)$? One natural choice for $h$ is $h(x) = p(x)$, indicating that differences in regions of relatively high density are the more important. Hence $D(p, q)$ becomes:

$$D(p, q) = \left[ \int \left( \sqrt{p(x)} - \sqrt{q(x)} \right)^2 p(x) \, dx \right]^{\frac{1}{2}}.$$
Other weight functions $h(x)$ might be used that identify differences only in key portions of the space of $x$. In fact, there is no reason that another distance measure altogether could not be used if it was more suitable for the analysis at hand.

There are several pairs of densities $p$ and $q$ for which the distance $D(p, q)$ is of interest. First, consider the distance between each sampled predictive density on a subject ($q = f_i^b$) and the average for that subject ($p = f_i^B$):

$$D(f_i^B, f_i^b)$$

The distribution over $b$ of this distance reflects the uncertainty in estimating the predictive density for subject $i$.

Next, consider the distance between the $i^{th}$ average predictive distribution ($q = f_i^B$) and the mean across all $n$ individuals ($p = \overline{f_i^B}$):

$$D(\overline{f_i^B}, f_i^B)$$

The distribution over $i$ of this distance is a measure of the variability between subjects. Calculation of these $n$ distances allows individuals to be ranked according to their closeness to the average.

The final distance that will be used in summaries of the data is the distance between each of the simulated predictive densities ($q = f_{new}^b$) and the average new predictive density ($p = \overline{f_{new}^B}$):

$$D(\overline{f_{new}^B}, f_{new}^b)$$

These $B$ distances constitute a sample from the posterior distribution of the distance. The sampled functions $f_{new}^b$ can be ordered by their distance from the average. Plotting sampled functions corresponding to selected quantiles from the posterior distribution of the distance $D$ is analogous to calculating the highest density region of a posterior distribution of a finite-valued parameter.

None of the distributions $f_i^b$ or $f_{new}^b$ will be known analytically, so a numerical approximation to the distance measures above will be necessary. If $p$, $q$ and $h$ are evaluated over the same equally spaced grid of points $y_t$, $t = 1, \ldots, n_y$, with $d = y_2 - y_1$
and
\[ D_t = \left( \sqrt{p(y_t)} - \sqrt{q(y_t)} \right)^2 h(y_t). \]
the distance measure is approximated by the trapezoid rule:
\[ D(p, q) \approx \left[ \frac{dD_1}{2} + \sum_{t=2}^{n_2-1} dD_t + \frac{dD_{n_2}}{2} \right]^{\frac{1}{2}}. \]

8.3 Simulating Data From the Posterior

In section 8.1, it was assumed that it was possible to obtain a sample from the Dirichlet process \( G_{\text{new}} \sim \mathcal{D}(\alpha_0^b, G_0^b) \). Throughout this thesis, samples of functions from Dirichlet processes have been avoided: the functions \( G_i \) and \( G_0^b \) have always been integrated out of sampling distributions, so that samples have been taken *from* \( G_i \) and \( G_0^b \), but not of \( G_0^b \) and \( G_i \). This section outlines the algorithm of Muliere and Tardella for approximating distributions of functionals of samples from Dirichlet processes [20].

To sample a distribution \( g(.) \) from the prior \( G \sim \mathcal{D}(\alpha_0, G_0) \) would require an infinite number of points to be sampled from \( G_0 \). The exact expression for a sampled function \( g(X) \) is
\[ g(X) = \sum_{i=1}^{\infty} p_i \delta_{X_i}, \]
where the \( X_i \) are sampled independently from \( G_0 \) and the \( p_i \) are the probability weights at the atoms \( X_i \). Following Sethuraman [23], the \( p_i \) can be constructed as follows:
\[
\begin{align*}
B_i & \sim \text{Beta}_B(1, \alpha_0) \quad i = 1, 2, 3, \ldots \\
p_1 &= B_1 \\
p_i &= B_i(1 - B_{i-1})(1 - B_{i-2}) \cdots (1 - B_1) \quad \forall \ i \geq 2. 
\end{align*}
\]
\[ (8.4) \]
The authors construct what they call an \( \varepsilon \)-Dirichlet process which is within a specified amount \( \varepsilon \) in total variation distance of the true Dirichlet process, \( \varepsilon \in (0, 1) \). A sample
from this process is given by truncating the series above at a finite number \( n_c \) of atoms. All the remaining probability \( r_c = 1 - \sum_{i=1}^{n_c} p_i \) is assigned to an atom \( X_0 \) sampled independently from \( G_0 \). If \( G^* \) has this \( \varepsilon \)-Dirichlet distribution, the expression for a sampled density is:

\[
g^*(X) = \sum_{i=1}^{n_c} p_i \delta_{X_i} + r_c \delta_{X_0}. \tag{8.5}
\]

The distribution of \( n_c \) is shown to be Poisson with mean \( \lambda = -\alpha_0 \log(\varepsilon) \). To sample a distribution \( g \) from an \( \varepsilon \)-Dirichlet process with precision parameter \( \alpha_0 \):

1. Choose a value of \( \varepsilon \).
2. Sample \( n_c \) from a Poisson(\( \lambda \)) distribution, where \( \lambda = -\alpha \log(\varepsilon) \).
3. Calculate \( p_i, i = 1, \ldots, n_c \) according to (8.4) and let \( r_c = 1 - \sum_{i=1}^{n_c} p_i \).
4. Sample \( X_i, i = 0, \ldots, n_c \) from \( G_0 \).
5. Use equation (8.5) to calculate \( g^* \).

This algorithm depends on samples of \( X \) from the baseline prior \( G_0 \) only in step (4). To sample \( \pi^* \) from \( G_0^b \), recall that \( G_0^b \) is formed as a mixture of normal-gamma distributions. Rewriting equation (5.5) to include values of hyperparameters and \( T \) at step (4) of the Gibbs sampler gives:

\[
G_0^b(\pi^* | T^b, H^b) = \frac{1}{\alpha_{00}^b + n_{\pi^*}^b} \left[ \alpha_{00}^b f_{00}(\pi^* | H^b) + \sum_{s=1}^{n_{\pi^*}^b} (n_s^b)^b M (\pi^* | (\tau_s^b)^b, H^b) \right].
\]

The \( n_c + 1 \) distinct values of \( \pi^* \) required by step (4) are sampled from the baseline marginal \( f_{00} \) with probability \( \alpha_{00}^b / (\alpha_{00}^b + n_{\pi^*}^b) \); and from \( M (\pi^* | (\tau_s^b)^b, H^b) \) with probability \( (n_s^b)^b / (\alpha_{00}^b + n_{\pi^*}^b) \). The distribution \( G_{new} \) sampled from the \( \varepsilon \)-Dirichlet process is formed as

\[
G_{new}(\pi) = \sum_{i=1}^{n_c} p_i \delta_{\pi_i^*} + r_c \delta_{\pi_0^*}.
\]
The new predictive distribution $f_{\text{new}}(y)$ is formed by integrating the conditional normal density $f(Y|\pi)$ with respect to the distribution $G_{\text{new}}(\pi)$.

$$f_{\text{new}}(Y) = \int f(Y|\pi) dG_{\text{new}}(\pi)$$

Since $G_{\text{new}}(\pi)$ is a discrete distribution, the new predictive distribution is evaluated as the mixture of $n_{\pi} + 1$ normal densities

$$f_{\text{new}}(Y) = \sum_{i=1}^{n_{\pi}} p_i N(Y|\pi_i^*) + r \cdot N(Y|\pi_0^*)$$

The rest of this section is a sketch of a proof that the distribution of the new densities generated from the $\varepsilon$-Dirichlet process converges to the distribution of the densities based on the full Dirichlet process. Muliere and Tardella show the following result.

- $T(G) = \int h(\pi) dG(\pi)$ is defined as a linear functional of $G$ if $h$ is a measurable function from the support of $G$ to $\mathbb{R}^d$.

- If $T(G) < \infty$ and $T(G)$ is a linear functional, then $T(G)$ converges in probability to $T(G)$ as $\varepsilon \downarrow 0$.

In calculating the new predictive density,

$$h(\pi) = f(Y|\pi) \propto \sqrt{v} \exp\left(-\frac{v}{2}(Y - \mu)^2\right).$$

and

$$T(G) = f_{\text{new}}(Y) = \int h(\pi) dG_{\text{new}}(\pi) = \int h(\mu, v) dG_{\text{new}}(\mu, v)$$

Thus, $h(\pi) = f(Y|\pi)$ is a measurable function and $T(G) = f_{\text{new}}(Y) < \infty$. Therefore for small enough $\varepsilon$, at each point $Y$, the distribution of the new densities based on $G_{\text{new}}$ sampled from the $\varepsilon$-Dirichlet process converges in probability to the distribution of densities based on $G_{\text{new}}$ sampled from the full Dirichlet process.
Chapter 9

Miscellaneous Topics

9.1 Overview of Algorithm

Admittedly, there are many details in the rather big model in this thesis that may obscure what is essentially a straightforward approach when seen from the right level. This section presents the core hierarchical model model and then the algorithm for fitting the model in the form of pseudo-code.

9.1.1 Hierarchical Model

The core hierarchical model (absent the priors on several hyperparameters) can be represented as follows:

\[
\begin{align*}
Y_{ij} | \pi_{ij} & \sim N(Y | \pi_{ij}), \ j = 1, \ldots, n_i \\
\pi_{ij} | G_i & \sim G_i, \ j = 1, \ldots, n_i \\
G_i & \sim \mathcal{D}(G_0, \alpha_0), \ i = 1, \ldots, n \\
\pi_r^* | \tau_r & \sim M(\pi_r^* | \tau_r), \ r = 1, \ldots, n_{\pi^*} \\
\tau_r | F & \sim F \\
F & \sim \mathcal{D}(G_{00}, \alpha_{00}) \\
G_{00} & \equiv \text{Normal-Gamma}
\end{align*}
\]
9.1.2 Sampling $\pi_{ij}|Y_{ij}$

This section of the algorithm is described in Chapter 5. Considering the distribution $G_0$ and all other parameters as fixed, a single Gibbs sampler cycle of updating all values of $\pi_{ij}|Y_{ij}$ proceeds as follows.

DO OVER SUBJECTS $i = 1$ TO $n$
    DO OVER OBSERVATIONS $j = 1$ TO $n_i$
        CALCULATE $f_0(Y_{ij})$ and $f(Y_{ij}|\pi_{ik}), k \neq j$
        SELECT $G_0(\pi_{ij}|Y_{ij})$ OR $\pi_{ik}$
        IF (SELECTED $G_0(\pi_{ij}|Y_{ij})$)
            CALCULATE $f_r(Y_{ij}), f_00(Y_{ij})$
            SELECT ATOM $\tau_r$ OR BASELINE PRIOR
            IF (SELECTED ATOM $\tau_r$)
                SAMPLE $\pi_{ij}|Y_{ij}$ NORMAL-GAMMA AROUND $\tau_r$
            ELSE IF (SELECTED BASELINE PRIOR)
                CALCULATE $f_{\pi_k}(Y_{ij})$
                SELECT CANDIDATE $\pi_k$
                SAMPLE $\pi_{ij}|Y_{ij}$ NORMAL-GAMMA AROUND $\pi_k$
        END IF
        ELSE
            SET $\pi_{ij}$ EQUAL TO $\pi_{ik}$
        END IF
    END DO
END DO

9.1.3 Sampling $\tau_r|\pi_r^*$

This section of the algorithm is described in Chapter 6. Considering the distribution $G_{00}$ and all other parameters as fixed, a single Gibbs sampler cycle of updating $T|\Pi^*$
proceeds as follows.

DO $r = 1$ TO $n_r$

     CALCULATE $f_{00}(\pi^*_r)$ and $f(\pi^*_r|\tau_k), k \neq r$

     SELECT $G_{00}(\tau_r|\pi^*_r)$ OR $\tau_k, k \neq r$

     IF (SELECTED $G_{00}(\tau_r|\pi^*_r)$)

     SAMPLE $w_r = 1/z$ where $z \sim r(z|\pi^*_r)$

     SAMPLE $m_r|w_r$ FROM NORMAL DISTRIBUTION

     ELSE

     SET $\tau_r$ EQUAL TO $\tau_k$

     END IF

END DO

DO $s = 1$ TO $n_r$ (REMIK)

SAMPLE $G_{00}(\tau^*_s|\Pi^*_s)$ THROUGH SHORT-RUN GIBBS SAMPLER

DO $i = 1$ TO $K$

     SAMPLE $w^*_s(i)|\Pi^*_s, m^*_s(i - 1)$

     SAMPLE $m^*_s(i)|\Pi^*_s, w^*_s(i)$

END DO

SET $\tau^*_s = (w^*_s(K), m^*_s(K))$

END DO

9.1.4 Entire Algorithm

The entire algorithm includes the above two big steps, as well as updating of hyperparameters $H = \{m_0, \alpha_0, \alpha_0, \Delta_1, \Delta_2\}$, generation of densities from the posterior and outputting of sampled posterior quantities of interest. These may include, for example, samples from the posteriors of the hyperparameters, sampled $G^b_0$ and sampled distances.
DO UNTIL "CONVERGENCE" THEN B MORE TIMES
UPDATE T PARAMETERS GIVEN $\Pi^*$
REMIX T GIVEN CONFIGURATIONS
UPDATE HYPERPARAMETERS $\Delta_1, \Delta_2, m_0$
RESCALE CANDIDATE VALUES FROM $G_{00}$: $\pi_k$
UPDATE $\alpha_{00}$
DO $i=1$ TO $n$
  UPDATE $\Pi_i$ PARAMETERS OF SAMPLE $i$ GIVEN $Y_i$
  UPDATE $\alpha_0$
  CALCULATE $f_i^b(y)$
END DO
IF "CONVERGED" = TRUE THEN
  SIMULATE $f_{new}^b(y)$
  OUTPUT CURRENT PARAMETER VALUES OF INTEREST
END IF
END DO

9.2 Choice of Prior Distributions

The appropriate choice of priors is essential for useful analysis with this model. Some of the priors will be left uninformative and some will be very informative. Below is a list of the hyperparameters, their interpretation and where possible, a guide to choices of values.

9.2.1 Dirichlet Precision Parameters

$\alpha_0$: A very informative prior will be used for $\alpha_0$. The whole premise underlying this model is that the observed data arise from the same parent distribution. The prior for $\alpha_0$ will be one that favours large values, so that the $G_i$ can be forced to be close to the distribution $G_0$. Large $\alpha_0$ means that there will be a large number of $\pi^*_k$. In
turn, when $G_0$ itself is updated, it will be similar to the $G_t$. In a sample of size $n$ from a distribution with a Dirichlet process prior, recall that the expected number of distinct elements for a given $\alpha_0$ is $\alpha_0 \log(1 + n/\alpha_0)$. A prior for $\alpha_0$ that favours values between $n$ and $2n$ gives a prior expectation of approximately $0.7n$ to $0.8n$ distinct elements. A discrete prior with similar weighting but a tiny mass on a value of $\alpha_0$ near 1 is an option that allows for small posterior values of $\alpha_0$ only when the likelihood is overwhelmingly large at these small values.

$\alpha_{00}$: A very informative prior will be used here, but one favouring small values. The distribution $F$ (or its smoothed out equivalent $G_0$) is the parent distribution and is not believed to have much fine structure. It will concentrate its mass on a few atoms. The size of the sample used to estimate $G_0$ is $n_{\tau}$, and this depends on the size of $\alpha_0$, the size of the sample for each individual, and the number of individuals $n$. The expected sample size is $\sum_{i=1}^{n} n_i \alpha_0 \log(1 + n_i/\alpha_0)$, which will be on the order of $nn_1$ if all samples sizes are equal and $\alpha_0$ is chosen as above. The prior expected number of atoms in $T$ is then $\alpha_{00} \log(1 + n n_1/\alpha_{00}) \approx \alpha_{00} \log(n) + \log(n_1)$ for small $\alpha_{00}$. This suggests a prior that puts almost no weight on values of $\alpha_{00}$ above 1.

### 9.2.2 Parametric Priors

Ignoring the Dirichlet process sampling, the hierarchical model is

\[
Y_{ij} | \mu_{*,*}^*, v_{*,*}^* \sim N(Y_{ij} | \mu_{*,*}^*, 1/v_{*,*}^*) \quad \forall \{ij\} \in S_{\tau}^{**}
\]

\[
v_{*,*}^* | w_s^* \sim \text{Gamma}(v_{*,*}^* | \alpha, \beta/w_s^*)
\]

\[
\mu_{*,*} | m_s^*, v_{*,*}^* \sim N(\mu_{*,*} | m_s^*, 1/(\Delta_2 v_{*,*}^*)) \quad \forall r \in S_{\tau}^{**}
\]

\[
w_s^* \sim \text{Gamma}(w_s^* | a, b)
\]

\[
m_s^* | w_s^* \sim N(m_s^* | m_0, 1/(\Delta_1 w_s^*))
\]

\[
m_0 \sim N(m_0 | d, 1/D)
\]

The set $S_{\tau}^{**}$ contains the indices of the $Y_{ij}$ that have $\pi_{*,*}^*$ as parameter and $S_{\tau}^{**}$ contains the indices of the $\pi_{*,*}^*$ that have $\tau_{*,*}^*$ as parameter.
\textbf{m}_0: \text{ The normal density prior for } m_0 \text{ should be very dispersed, which means using small } D, \text{ so that samples of the cluster means } m^*_j \text{ are taken from a distribution that has support over the range of the data.}

\textbf{v|w}: \text{ Recall that } v|w \sim \text{Gamma}(v|\alpha, \beta/w). \text{ Setting } \alpha = \beta \text{ means that } E(v|w) = w \text{ and } \text{Var}(v|w) = w^2 \frac{\alpha}{\beta^2} = \frac{w^2}{\beta} \text{ so that the parameter } \beta \text{ is a scale parameter for the variance of } v|w: \text{ the larger } \beta, \text{ the closer } v \text{ is to } w. \text{ In the models examined here, } \alpha \text{ will always be equal to } \beta, \text{ and both will be fairly large, so that } v \approx w.

\textbf{w}: \text{ Using } \alpha = \beta \text{ and } \alpha \gg 1 \text{ means that } v \approx w, \text{ so that the prior for } w \text{ is essentially the prior for the precision of } Y_{ij} \text{ around its mean } \mu_{ij}, \text{ and should be chosen accordingly.}

\Delta_2: \text{ The parameter } v^*_j \text{ is the precision of each of the } Y_{ij} \text{ that share the value of the parameter } \pi^*_i. \text{ The scaled value } \Delta_2 v^*_j \text{ is the precision of } \mu^*_j \text{ around its mean } m^*_j. \text{ This means that the scale factor } \Delta_2 \text{ is the ratio of these two precisions. Simulated data from models with different sizes of } \Delta_2 \text{ are illustrated in figures 9.1 and 9.2. The precision } v \text{ has been fixed at a value of } 1/2 \text{ for the purposes of this illustration. The first graph in each figure shows two fixed values of } m_j: m_1 \text{ and } m_2. \text{ The second graph shows six values of } \mu^*_j: \text{ the leftmost three values were sampled from the density } N(m_1, 1/(\Delta_2 v)) \text{ and the other three values were sampled from } N(m_2, 1/(\Delta_2 v)). \text{ The bottom graph shows 12 values of } Y_i; \text{ two values were sampled from each of the six normal densities } N(Y_i|\mu^*_j, 1/v). \text{ For a given value of } v^*_j, \text{ small } \Delta_2 \text{ (figure 9.2) means that the the } \mu^*_j \text{ will be spread out around their mean } m_j, \text{ with their respective } Y_i \text{'s relatively close to them. If } \Delta_2 \text{ is large (figure 9.1), then the converse is true, and the } \mu^*_j \text{ are closer to their respective } m_j \text{ than the } Y_i \text{'s are to their respective } \mu^*_j. \text{ With large } \alpha_0, \text{ it is common that only only 1 or 2 of the } \pi_{ij} \text{ map to a particular } \pi^*. \text{ The } Y_{ij} \text{ will be close to their } \mu_{ij}, \text{ so their precision should be higher than that of the } \mu_{ij} \text{ around } m_j \text{ and the prior for } \Delta_2 \text{ should favour values smaller than 1.}

\Delta_1: \text{ The scale factor } \Delta_1 \text{ acts in a similar way for the precisions of } m^*_j \text{ and } \mu^*. \text{ Assuming that } v \approx w, \text{ then the ratio } \Delta_1/\Delta_2 \text{ indicates how much less variable the } m^*_j \text{'s are than the } \mu^*.\]
Figure 9.1: Understanding the Hyperparameters: $\Delta_2 > 1$ means that the parameters $\mu_i^*$ are closer to their means $m_s$ than the data $Y_{ij}$ are to their means $\mu_{ij}$.
Figure 9.2: Understanding the Hyperparameters: $\Delta_2 < 1$ means that the parameters $\mu_2^*$ are more dispersed around their means than the data $Y_{ij}$ are around their means $\mu_{ij}$. This is the more realistic situation for the model in this thesis.
The choice of priors for $\Delta_1$ and $\Delta_2$ should keep all this in mind. A given value of $w_\ast^s$ is essentially the precision of all the $Y_{ij}$ mapping up to $r_\ast^s$. The value $\Delta_1$ indicates the precision of the cluster centres $m_\ast^s$ around $m_0$ in units of the precision of $Y_{ij}$. Likewise, the value $\Delta_2$ indicates the precision of the cluster centres $\mu_\ast^s$ around $m_\ast^s$ in the same units. Priors for both $\Delta_1$ and $\Delta_2$ will usually give weight to values much less than 1, with their relative sizes being determined by the particular data analysis.

Figures 9.3, 9.4, and 9.5 show how $\alpha$, $\beta$ and $\Delta_2$ affect the estimate of $G_0(\pi^\ast)$ formed by mixing the normal-gamma priors centered at the atoms $\pi^\ast$. In these plots, the contours represent equiprobable levels of $G_0(\pi^\ast) = \sum_{s=1}^3 p_s M(\pi^\ast | \tau_\ast^s)$, where the black dots represent $\tau_\ast^s$ with weights of $p_1 = 0.3$, $p_2 = 0.3$ and $p_3 = 0.4$, reading from left to right. The $m_\ast^s$ and $w_\ast^s$ values themselves can be read from the $\mu$ and $\nu$ axes respectively, as they are on the same scale as the $\mu$ and $\nu$. The parameters $\alpha$ and $\beta$ act to scale the prior in the $\nu$-direction and $\Delta_2$ scales it in the $\mu$-direction.

### 9.3 Computational Issues

Each step in the algorithm for Gibbs sampling the Dirichlet posterior relies on the conditional distribution of a parameter given the datum that the parameter "belongs" to, i.e.: 

$$f(\pi_{ij}|Y_{ij}, \Pi_i^{-1}) \text{ and } f(\tau_\ast|\pi_\ast^\ast, T^{-\tau}).$$

This means that at each step in the Gibbs sampler, the parameter for each datum must be clearly identified. Each $Y_{ij}$ must have a $\pi_{ij}$ and each $\pi_\ast^s$ must have a $\tau_\ast^s$. The small example below shows the logical relationships that must be maintained between elements of the $Y$, $\Pi_i$, $\Pi^\ast$, and $T$. In this example, there are two subjects, each with 3 observed data points $Y_{ij}$. Initially, at the stage of the Gibbs sampler shown in the first diagram below, there are 3 elements in $\Pi^\ast$ and 2 unique elements in $T$, the first
Figure 9.3: Normal-gamma mixture over values of $\tau^*$ from $F$ with hyperparameters $\alpha = 2; \beta = 2; \Delta_2 = 0.5$
Figure 9.4: Normal-gamma mixture over values of $\tau^*$ from $F$ with hyperparameters $\alpha = 2; \beta = 2; \Delta_2 = 1$
Figure 9.5: Normal-gamma mixture over values of $\tau^*$ from $F$ with hyperparameters $
alpha = 4; \beta = 4. \Delta_2 = 1$
with 2 \( \pi^* \) values mapping to it, the second with 1.

\[
\begin{align*}
Y_{11} & \rightarrow \pi_{11} \\
Y_{12} & \rightarrow \pi_{12} \\
Y_{13} & \rightarrow \pi_{13} \\
Y_{21} & \rightarrow \pi_{21} \\
Y_{22} & \rightarrow \pi_{22} \\
Y_{23} & \rightarrow \pi_{23}
\end{align*}
\]

\( = \pi_1^* \rightarrow \tau_1 \)

\( = \tau_1^* \)

\( = \pi_2^* \rightarrow \tau_2 \)

\( = \tau_2^* \)

\( = \pi_3^* \rightarrow \tau_3 \)

\( = \tau_3^* \)

Now consider updating the parameter \( \pi_{23} | Y_{23}, \Pi_2^{-3} \). There are three immediate possibilities which it is important to distinguish, as well as one which follows from the third:

Case 1 Sample \( \pi_{23} = \pi_{21} \). In this case, the initial diagram changes to

\[
\begin{align*}
Y_{11} & \rightarrow \pi_{11} \\
Y_{12} & \rightarrow \pi_{12} \\
Y_{13} & \rightarrow \pi_{13} \\
Y_{21} & \rightarrow \pi_{21} \\
Y_{23} & \rightarrow \pi_{23} \\
Y_{22} & \rightarrow \pi_{22}
\end{align*}
\]

\( = \pi_1^* \rightarrow \tau_1 \)

\( = \tau_1^* \)

\( = \pi_2^* \rightarrow \tau_2 \)

\( = \tau_2^* \)

\( = \pi_3^* \rightarrow \tau_3 \)

\( = \tau_3^* \)

or sample \( \pi_{23} = \pi_{22} \). in which case, the initial diagram is unchanged. Neither of these choices involves any changes to the values in \( \Pi^* \). or to the values of the weights associated with the members of \( T \).

Case 2 Sample \( \pi_{23} \) equal to a new value \( \pi_4^* \) from \( G_0(\pi | Y_{23}) \); one possibility is to sample the value from around one of the distinct atoms in \( T \). For example, if \( \tau_1^* \) is
chosen, the diagram changes to

\[
\begin{align*}
Y_{11} & \rightarrow \pi_{11} \\
Y_{12} & \rightarrow \pi_{12} \\
Y_{13} & \rightarrow \pi_{13} \\
Y_{21} & \rightarrow \pi_{21} \\
Y_{23} & \rightarrow \pi_{23} \\
Y_{22} & \rightarrow \pi_{22}
\end{align*}
\]

\[
= \pi_1^* \rightarrow \tau_1
\]

\[
= \pi_2^* \rightarrow \tau_2
\]

\[
= \pi_4^* \rightarrow \tau_4
\]

\[
= \pi_3^* \rightarrow \tau_3
\]

Here, the index 4 on \( \pi^* \) emphasizes that \( \pi_{23} \) is equal to a new value. The total sample size of \( \Pi^* \) increases to 4; however, there are still only 2 distinct elements in \( \mathbf{T} \). Notice that the number of \( \pi^* \) having \( \tau_2^* \) as parameter remains at 1, so \( n_{\tau_2}^* \) is still 1; but the new value of \( \pi_4^* \) has \( \tau_1^* \) as its parameter, so that \( n_{\tau_1}^* \) increases to 3.

Case 3 As in Case 2, sample \( \pi_{23} \) equal to a new value \( \pi_4^* \) from \( G_0(\pi|Y_{23}) \). The distinction is that unlike Case 2, \( \pi^* \) is sampled around a new atom \( \tau_3^* \) drawn from \( G_{00}(\tau|Y_{23}) \). The new diagram is

\[
\begin{align*}
Y_{11} & \rightarrow \pi_{11} \\
Y_{12} & \rightarrow \pi_{12} \\
Y_{13} & \rightarrow \pi_{13} \\
Y_{21} & \rightarrow \pi_{21} \\
Y_{22} & \rightarrow \pi_{22} \\
Y_{23} & \rightarrow \pi_{23}
\end{align*}
\]

\[
= \pi_1^* \rightarrow \tau_1
\]

\[
= \pi_2^* \rightarrow \tau_2
\]

\[
= \pi_3^* \rightarrow \tau_3
\]

\[
= \pi_4^* \rightarrow \tau_4
\]

In this case, the sample size of \( \Pi^* \) increases to 4 and a new value is added to the set \( \mathbf{T} \).

Case 4 Next consider updating the parameter \( \pi_{22}|Y_{22}, \Pi_{22}^{-2} \) starting from this last configuration. Whatever the sampled value for this parameter, it cannot be equal
to $\pi_3^i$: consequently, $\pi_3^i$ will have no data mapping to it. In future Gibbs sampler steps, it will have zero probability of being sampled, so can be omitted from further calculations. Now that $\pi_3^i$ is no longer part of the sample from the posterior distribution $G_0(\pi^*|Y)$, it should not be used as data in updating $G_0$: the size of the configuration mapping to atom $\pi_3^i$ is zero. The value of the atom $\pi_3^i$ can also be dropped from further calculations.

None of this is problematic from the point of view of programming the algorithm. With care, the changing relationships implied by the Gibbs sampling can be tracked as needed.

However, the astute reader might have spotted something unusual with regard to the sampling procedure in cases 2, 3 and 4. Throughout Chapter 5, the baseline posterior was said to be treated as fixed when sampling $\pi_{ij}|Y_{ij}$. But in these examples, sampling from $G_0(\pi_{ij}|Y_{ij})$ appears to change $G_0$, either through adding new atoms $\pi^*$, changing the weights on existing atoms, or dropping atoms. The apparent anomaly results from two sampling steps being conflated into one. These are the correct two steps to take in obtaining the new value $\pi_{n_{x_1}+1}^*$ and the new value $\pi_{n_{x_1}+1}$:

1. With fixed $G_0$: Obtain the new value of $\pi^*$. Sample $\pi_{n_{x_1}+1}^*|Y_{ij}$ from the existing $G_0$. Here, there are two possible sources of "centres" for $\pi_{n_{x_1}+1}^*$. It is sampled either around $\tau_3$ or around $\pi_3^i$.

2. Update $G_0$: Obtain a new value of $\tau^*$ by sampling $\tau_{n_{x_2}+1}|\pi_{n_{x_1}+1}^*$. The distribution of $\tau_{n_{x_2}+1}|\pi_{n_{x_1}+1}^*$ is given by modifying equation (6.4), namely:

$$F(\tau_{n_{x_2}+1}|T,\pi_{n_{x_1}+1}^*) \propto \alpha_{00} f_{00}(\pi^*) G_{00}(\tau_{n_{x_2}+1}|\pi_{n_{x_1}+1}^*)$$

$$+ \sum_{s=1}^{n_{x_*}} M(\pi_{n_{x_2}+1}|\pi_{n_{x_1}+1}^*) n_s \delta_{\pi^*_s}(\tau_{n_{x_2}+1}). \quad (9.1)$$

This is the exact way to sample the value $\tau_{n_{x_2}+1}|\pi_{n_{x_1}+1}^*$, but its implementation in the computer program was slightly different. Step 1 was carried out as described, but step 2 was modified. The baseline prior $G_0(\pi|Y)$ is sampled often in this large-$\alpha_0$
model, so new values of $\pi^*$ occur often. The distribution (9.1) would have to be sampled equally often, with the attendant calculation of $f_{00}(\pi^*)$, and the possible evaluation of a normal-gamma density at all $N$ candidate values $\tau_k$. Furthermore, with small values of $\alpha_0$, $\tau_{n,s+1}$ will most likely be sampled equal to an existing $\tau^*_s$, so the calculation of $f_{00}(\pi^*)$ is of little value.

To reduce computations, in step 2 above, $\tau_{n,s+1}^*$ is sampled from this approximation to distribution (9.1):

$$F(\tau_{n,s+1}^*|\mathbf{T}, \pi_s^*, \pi_{n,s+1}^*) \propto \sum_{s=1}^{n_{\tau^*}} M(\pi_{n,s+1}^*|\tau^*_s) n_{\tau^*}^* \delta_{\tau^*_s}(\tau_{n,s+1}^*).$$

(9.2)

That is, $\tau_{n,s+1}^*$ is set equal to an existing $\tau^*_s$ with probability equal to the likelihood that $\pi_{n,s+1}^*$ has $\tau^*_s$ as its “centre”. With his approximation, case (3) in the above diagrams cannot happen as shown, as a new value of $\tau^*$ is never introduced when sampling a new value of $\pi^*$. Otherwise, the weights $n_{\tau^*}$ are modified as if $\tau_{n,s+1}^*$ had been sampled through the exact procedure. In particular, if the updating of $\pi_{ij}|Y_{ij}$ means that there is no data $\pi^*$ for a particular value of $\tau^*$, as in case 4 above, then the corresponding $n_{\tau^*}$ is set to zero before the approximate distribution 9.2 above is sampled. Since new values of $\tau|\pi_{n,s+1}^*$ will be sampled from $G_{00}(\tau|\pi_{n,s+1}^*)$ as needed when the distribution $G_0$ is updated and when $\mathbf{T}$ is remixed, this approximation should not prevent convergence to the posterior: “bad” values sampled here from among the existing $\mathbf{T}$ will be replaced by “good” values later.
Chapter 10

Simulated Application:
Summarizing a Set of Samples

The motivation for modelling the sort of data in this thesis arose from work with a haematologist and the distributions of blood cell volumes. When the thesis work began, there was an understanding that it would soon be possible to extract data on blood cell volumes directly from the machine that analyzed the blood samples. The plan was that this data would provide a good first application for the thesis work. At the time, the machine produced printed summaries and graphs that appeared to be based on fitting a mixture of normals with the number of components determined by physiological theory. As it turned out, the manufacturer of the blood analysis machine was loathe to make it possible to have access to the raw data. This unwillingness may have been related to work that the manufacturer was doing on improving the machine’s internal analysis routines. In any case, as the thesis work neared completion it became apparent that there would be no blood cell volume data to analyze. Instead, some data was simulated to demonstrate key features of the algorithm; this chapter deals with this simulated application. A real application was found in the reaction time data of Gelman et al. [4] introduced in Chapter 2 and this data is also examined in the next chapter in light of what is already known about it.

This chapter has two purposes. First, by analysing data from a known model, it
will be shown that the methods described in this thesis do indeed provide summaries of the data that capture its important features. Secondly, the sensitivity of the model to the specification of some of the prior distributions will be investigated. It is hoped that this provides some insights into the types of smoothing that occur in the model.

10.1 Data Generation

Data was generated from the following model for \( n = 20 \) subjects \( (i = 1 \ldots 20) \), each with equal sample size \( n_i = 25 \) \( (j = 1 \ldots 25) \).

\[
Y_{ij} \sim \sum_{k=1}^{3} p_k N(Y_{ij} | \mu_{ik}, 1/v_{ik})
\]

\[
\mu_{ik} \sim N(\mu_{ik} | m_k, 1/w_k)
\]

The values of the parameters \( v_{ik}, m_k \) and \( w_k \) were set as below (remembering that \( r \) and \( w \) are precisions):

\[
p_1 = 0.5 \quad p_2 = 0.2 \quad p_3 = 0.3
\]
\[
v_{i1} = 1 \quad v_{i2} = \frac{1}{1.5^2} \quad v_{i3} = \frac{1}{1.5^2} \quad \forall i
\]
\[
m_1 = -5 \quad m_2 = 0 \quad m_3 = 5
\]
\[
w_1 = \frac{1}{2^2} \quad w_2 = 1 \quad w_3 = \frac{1}{0.25^2}
\]

Figure 10.1 shows 9 probability distributions generated from this model. The true probability distributions all have three components and have a high probability of having at least two modes. The location of the leftmost component varies considerably, but it has the largest precision conditional on its location. The location of the rightmost component varies relatively little, but it is less precise around this location. The middle component can combine with either the first or third components so that \( f(Y) \) has only two modes.
Figure 10.1: Sample PDFs From the Model Used to Generate Data

Sample 1

Sample 2

Sample 5

Sample 6

Sample 8

Sample 12

Sample 15

Sample 16

Sample 18
10.2 Application of Model

The data were analyzed with the algorithm summarized in Section 9.1. Details specific to the application to this simulated data are presented below. **Starting values:** The initial values for \( \mu_{ij} \) were set equal to the observed \( Y_{ij} \). All \( v_{ij} \) for subject \( i \) were set equal to \( n_i \) times the sample precision of the data on subject \( i: v_{ij}^{-1} = n_i^{-2} \sum_{j=1}^{25}(Y_{ij} - \overline{Y}_i)^2 \). Conditional on these \( N = \sum_{i=1}^{n} n_i \) values of \( \pi_i^* \). \( N \) initial values for \( \pi_i^* \) were sampled from \( G_{00}(\tau_r|\pi_i^*) \). This ensured that the initial potential cluster centres \( m_i^* \) would be sampled from \( m_r \) spread across the space of \( Y \). The starting values of other parameters were set to their prior means.

**Number of Iterations:** An informal method was used to decide the number of iterations before using samples from the sequence of Gibbs samples as representative of the posterior. Inspection of a graph of the number of atoms in \( T \) was used as a guide: since this number changes slowly, it was assumed that the entire sequence of Gibbs samples was coming from the joint posterior distribution when this number stabilized. Typically, this happened after no more than 500 iterations. The Gibbs sampler was run for a further \( B = 5000 \) iterations of the entire updating algorithm, with samples from posterior distributions being saved to files every 25 iterations. These posterior samples were used for summarizing results. Although for some parameters, posterior distributions could have been calculated by averaging the analytical conditional distributions over the \( B \) Gibbs steps. For the sample predictive distributions \( f_B(Y_{n+1}^t) \), data from every iteration are included in the average \( f_B(Y_{n+1}^t) \).

**Choice of Priors** For the first goal of this chapter — to demonstrate that the model can capture the important features of the dataset — priors were chosen to be consistent with the model that generated the data. For the second goal — investigation of the sensitivity to priors — a variety of priors were used. The sets of priors used are shown in table 10.1. In every case, the prior for \( \alpha_{00} \) was Gamma(60, 100). This corresponds to a prior expectation of around 4 distinct components in \( T \).

Two types of priors were used for \( \alpha_0 \). In set 1, an \( \alpha_0 \) common to all subjects was given a Gamma(50, 1) prior. In the other 3 sets, precision parameters \( \alpha_0^t \) which
Table 10.1: Sets of priors used in analyses of simulated data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Set 1</th>
<th>Set 2</th>
<th>Set 3</th>
<th>Set 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Shape</td>
<td>Scale</td>
<td>Shape</td>
<td>Scale</td>
</tr>
<tr>
<td>$w$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$v</td>
<td>w$</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>$\alpha_0^i$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>50</td>
<td>1</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>$\Delta_1$</td>
<td>0.4</td>
<td>10</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$\Delta_2$</td>
<td>2</td>
<td>10</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Some general comments on the sets of priors:

- **Set 1:** The Gamma(50, 1) prior for $\alpha_0$ corresponds to a prior expectation of around 20 distinct $\pi^*$ for each subject. Keep in mind that the posterior distribution of the number of $\pi^*$ for each subject is modified by the other parameters of the model. The precision parameter $w$ was given a Gamma(2, 2) prior, which is fairly dispersed, containing 95% of its mass in a highest density region (HDR) extending from 0.021 to 2.38. To guarantee that $v$ is near $w$, the prior for $v|w$ was Gamma(10, 10/w), which has a 95% HDR of 0.43 to 1.63. The expected values of the precisions of $Y, v$ and $m$ were set approximately in the ratio $25:5:1$. by using priors $\Delta_1 \sim \text{Gamma}(0.4, 10) \ (97.5^{th} \text{ percentile} = 0.22)$ and $\Delta_2 \sim \text{Gamma}(2, 10) \ (\text{HDR: 0.004 to 0.476})$. These priors represent a great deal of prior knowledge about the relative sizes of the precisions of the various means.

- **Set 2:** The prior for each $\alpha_0^i$ was Gamma(25, 1) (HDR: 15.6 - 34.96), corresponding to a prior expectation of around 17 distinct elements $\pi^*$ for each
subject. Another way of thinking of an $\alpha_0$ of this size is that in simple Dirichlet Gibbs sampling, it gives equal importance to the data for subject $i$ and the baseline prior. The prior weight given to the baseline prior $G_0$ in set 2 is about half of that in set 1. As a result, the estimates for predictive densities on different subjects are allowed to vary more, giving more weight to the data for each particular subject. The precision parameter $w$ was given the same Gamma$(2, 2)$ prior but the prior for $v|w$ was changed to Gamma$(5.5/w)$ (95% HDR: 0.24 - 1.88). This places less of a constraint on the size of the individual precisions $v$ in relation to the precision components $w$ of the few cluster centres $T$. By allowing more heterogeneity among the $\pi^*$. the result should be predictive densities that accommodate differences between individuals. No strong prior assumptions were made about the precision scale factors $\Delta_1$ and $\Delta_2$; the distribution Gamma$(0.01, 0.01)$ was used for both. For data on the scale measured here, the likelihood dominates any information contained in this prior. This set of priors represents substantially less prior knowledge than the priors in the first scenario. It will be worth examining the relative sizes of posterior values of sampled $\Delta_1$ and $\Delta_2$; it is hoped that the posterior values for these parameters are in the order $1 > \Delta_2 > \Delta_1$.

- **Set 3**: The prior on $w$ was Gamma$(4.4)$ (HDR: 0.18 - 1.98) and the prior on $v|w$ was Gamma$(10.10/w)$. This reduces the prior variability on $v_{ij}$ the precision of $Y_{ij}|\mu_{ij}$ and in fact, this set of priors may restrict $v_{ij}$ to too small a range. If the density estimate is thought of as a kernel density estimate with varying window width, then placing too restrictive a prior on $v_{ij}$ means that the window width is not allowed to vary enough. The same vague priors used in set 2 were used for $\Delta_1$ and $\Delta_2$. Through comparison of the results from this set of priors to the results from set 2, the effect of an overly precise specification of the values of $v_{ij}$ will be examined.

- **Set 4**: This is the same as set 1 except for the prior on the Dirichlet precision parameter $\alpha_0$. For the simulated data in this section, no particular subject
should deviate by an especially large amount from the average. This is meant in the general sense that there is only one subject population and all differences are a result of random sampling from the data generation model. There should be little difference between the results allowing independent precision parameters or just one. Set 1 and set 4 differ only on this distinction on the prior for \( \alpha_0 \), so can be used to assess the effect of independent priors on the posterior distribution of \( \alpha_0 \) and the consequences for the rest of the model. If there are no real differences in the results of analyses based on prior sets 1 and 4, then the use of the individual priors is not introducing an artefact. It may then be worth examining whether the posterior distribution of \( \alpha_0 \) can be used as a measure of similarity between a particular subject and the parent distribution when there are differences between subjects. If the posterior distribution of \( \alpha_0^I \) for subject 1 and the posterior distributions of \( \alpha_0^I \) of the other subjects are sufficiently different, this might be evidence that subject 1 is less like \( C_n \) than the other subjects are.

10.3 Results

Set 1: Figure 10.2(a) plots the 20 mean predictive distributions \( f^B_i \) on the same graph. With only 20 subjects, it is still possible to see that the modes vary in height and location but the overall impression is merely one of a set of distributions that are similar, but not quite alike. With more subjects, it might not be possible to see even this.

Hellinger distances were computed between each mean predictive distribution \( f^B_i \) and their overall mean \( \overline{f^B_{obs}} \). The predictive distributions were ranked according to this distance and the \( f^B_i \) corresponding to the four largest distances were plotted alongside \( \overline{f^B_{obs}} \) in figure 10.2(b). The use of colour helps to distinguish the five plotted distributions and the following can be seen: (1) the mode near \(-5\) varies substantially in location but is always largest mode; (2) the mode near \(+5\) does not vary in height
or location; and (3) the mode near 0 is sometimes subsumed into one of the two larger modes. Actually, the predictive densities near \(-5\) sometimes appear to have two local modes. This will be discussed further below.

The prior favouring large values for \(\alpha_0\) does not stem from a belief that nearly 20 distinct normal components are needed to fit a mixture for each subject. Rather, the large number of elements \(\pi^*\) guarantees that the parent distribution \(G_0\) will be estimated from data that is representative of the distributions of \(Y\). If \(\alpha_0\) were given an uninformative prior, its posterior would favour values corresponding to a small number of \(\pi^*\) within each subject. 3 or so for the simulated data here. Then the distribution \(G_0\) would be based on a sample of only about \(3n\) values of \(\pi^*\), \(n\) near the centre of each mode. The result would be a \(G_0\) that gave equal weight to each of the three modes.

Of course there is a penalty in using too large a value of \(\alpha_0\). As \(\alpha_0 \to \infty\), all the individual \(G_i\) converge on \(G_0\). Differences between individual predictive densities disappear. Something of this can be seen in figure 10.2(b), where \(\alpha_0\) has a prior mean of 50. The predictive densities with the second and fourth largest Hellinger distance, those near the mode at \(-5\) have a minor mode at \(-5\) that appears to correspond to the prior \(G_0\). Figure 10.3(b) is a histogram of 4000 values of \(\alpha_0\) sampled from the posterior. The posterior is shifted slightly downward from the prior; 20 distinct components per subject are not needed to fit this mixture of three normals. But the mean is still around 45. This corresponds to the prior \(G_0\) having almost twice as much weight in the estimate of \(G_i\) as the individual elements \(\pi^*\). The key to successful use of this model is finding a balance between an \(\alpha_0\) that is so big that it obscures differences between subjects and one that is so small that the parent distribution is in no way representative of the individuals. Figure 10.3(a) is a histogram of 200 values of \(\alpha_{00}\) sampled from its posterior. The posterior is little changed from the highly informative prior. Figure 10.3(c) is a histogram of Hellinger distances of 200 generated \(f_{\text{new}}^B\) from their mean value \(\overline{f_{\text{new}}^B}\). Figure 10.3(c) is a histogram of the distances of the 20 subject-means \(f_i^B\) from their average \(\overline{f_{\text{obs}}^B}\). The generated distances
Figure 10.2: Set i: Plot (a) shows $\bar{f}_{\text{obs}}$ (black line), and $f_i^B$. Plot (b) shows $f_i^B$ corresponding to the four largest Hellinger distances from $\bar{f}_{\text{obs}}$.

(a) Mean Predictive Distribution and Samples

(b) Selected Sample PDFs
Figure 10.3: Set 1: Posterior Samples of Selected Hyperparameters

(a) Sampled alpha00 With Prior

(b) Sampled alpha10 With Prior

(c) Generated Distances

(d) Observed Distances
are larger since they incorporate two types of uncertainty: (1) the variability of a particular individual's mean function around the average of all individuals; and (2) the uncertainty in estimating an individual's mean based on the observed data arising from the mean function.

**Set 1 compared to set 2:** These two sets of priors differ in more than one aspect. So differences in posterior distributions cannot be traced to a single factor. But it is worth comparing figures 10.2(b) and 10.4(b) in light of set 2 having smaller a prior expectation for $\alpha_0$. For the results under prior set 2, the influence of $\omega_0$ is seen only through a "shoulder" in the PDFs near -5. The PDFs are not shrunk so much towards their common mean, so they reach values of $f(y)$ close to those in the model used to generate the data. Furthermore, these PDFs capture all of the important features of $f(y)$ noted earlier.

**Set 1 compared to set 4:** The effect of allowing a separate Dirichlet precision parameter for each subject is minimal for this data set. Figure 10.5 shows plots of the empirical cumulative distribution functions (ECDF) based on posterior samples of $\alpha_0$ from the common and independent prior models. The plots are for 9 randomly chosen individuals: each value of $\alpha_0$ for individual $i$ was sampled conditionally on the current number of distinct components in $\Pi^p$ at 200 widely spaced steps of the Gibbs sampler. Figure 10.2 and the corresponding figure for set 4 (figure 10.6) are almost identical, with the functions from set 4 being just slightly less smoothed.

**Set 2 compared to set 3** The priors for $w$ and $v|w$ in set 3 lead to a narrower prior distribution for the different precisions of $Y|\mu$. In both prior sets the prior mean of $w$ is 1. However, the mean posterior value for $w$ in set 2 is 1.78, whereas in set 3 it is 1.35. Thinking of this in terms of the likelihood and the prior, the likelihood favours values larger than the prior. Placing a narrower prior on $w$ in set 3 shifts the posterior distribution of $w$ towards its prior distribution, on values the likelihood indicates as being too small. These small values of $w$, combined with the tight prior on $v|w$ mean that $v$ is also too small. This manifests itself in several ways. (1) The posterior means for $\Delta_1$ and $\Delta_2$ are each about 50% larger for set 3 than for set 2.
Figure 10.4: Set 2: Plots of Sampled $f_{\text{obs}}^B$ and $f_i^B$

(a) Mean Predictive Distribution and Samples

(b) Selected Sample PDFs
Figure 10.5: Set 1: Posterior ECDF of $\alpha_0$ for Common (solid) and Independent (dashed) Priors
Figure 10.6: Set 4: Plots of Sampled $f_{obs}^B$ and $f_i^B$

(a) Mean Predictive Distribution and Samples

(b) Selected Sample PDFs
(Figure 10.7), as the model “tries” to increase the precisions of \( m \) and \( \mu \). (2) Smaller \( v^*_i \) in set 3 means that more \( Y_{ij} \) can map to a single \( \pi^*_i \), so there are fewer elements in the posterior samples of \( \Pi^* \) for set 3. (3) The tight prior on \( w|v \) also means that the posterior \( w|v \) is tight around \( v \) — there are more distinct \( \tau^* \) in samples from \( F \) for set 3, even though the precisions of \( m|w \) are smaller (Figure 10.8) Surprisingly, given these differences, the posterior predictive distributions of \( y \) for sets 2 and 3 are very close. This can be seen by comparing Figures 10.4 and 10.9. There are no noteworthy differences in either the mean function or the subjects ranked furthest from the mean under the Hellinger metric. The vague priors on \( \Delta_1 \) and \( \Delta_2 \) mean that the overly precise priors on \( w \) and \( v|w \) had little effect on these estimates.

Figures 10.10(a) and 10.10(b) each show the mean of 200 generated new predictive distributions \( \bar{f}_{\text{new}}^B \) and the new predictive distributions corresponding to large percentiles of the Hellinger distance. It should be kept in mind that the new predictive distributions \( f_{\text{new}}^B \) are naturally more variable than the individual mean predictive distributions \( f_i^B \), which are averages over 5000 sampled functions \( f_i^B \).

The priors in set 3 appear to cause some problems with the generated new predictive distributions. The extreme sampled \( f_{\text{new}} \) from set 2 exhibit some of the characteristics of the original data: the leftmost mode varies around the value \(-5\); the value of \( f(y) \) for these modes is near 0.2; and three of the four extreme functions have their lower mode exactly at \( +5 \). The corresponding plot for set 3 contains distributions that look distinctly unlike either the model used to generate this data or the density estimates for the 20 subjects. These generated \( f_{\text{new}} \) are far from the mean. it is true, but in a way that is not suggested by the observed data. The problem may lie in the smaller number of distinct elements in \( \Pi^* \) in set 3. Figure 10.11 is a qq-plot of the number of elements in \( \Pi^* \) for prior sets 2 and 3, at the 200 steps in the Gibbs sampler where data was saved to a file. There are on the average about 30 fewer elements in a sample of \( \Pi^* \) in set 3 than in set 2. As discussed above, some of the important features of the distribution of \( Y \) may be filtered out if there are too few \( \pi^* \) used in the estimation of \( C_0 \). However, a difference of 30 elements works out
Figure 10.7: Posterior Samples of $\Delta_1$ and $\Delta_2$ from Sets 2 and 3
Figure 10.8: Posterior Samples of Number of Distinct $\tau$

Set 2

Set 3
Figure 10.9: Set 3: Plots of Sampled $f^B_{obs}$ and $f^B_i$

(a) Mean Predictive Distribution and Samples

(b) Selected Sample PDFs
Figure 10.10: Sets 2 and 3: Selected Generated PDFs: The darkest line is the mean function $f_{new}^B$. The others are functions corresponding to large percentiles of the distribution of the Hellinger distance. For example, the function $f_{new}^{new}$ labelled 96%ile has the $0.96 \times 200^{th}$ largest distance from $f_{new}^B$ out of the 200 $f_{new}^{new}$ generated.

(a) Set 2: Selected Generated PDFs

(b) Set 3: Selected Generated PDFs
to fewer than 2 per subject, so this is not the only explanation for the poor quality of $f^{new}$ for prior set 3.

In generating $G_{new}^b \sim D(\alpha_0^b, C_0^b)$, if $\alpha_0^b$ is too small, then $G_{new}$ can be quite unlike $C_0^b$. This is the other likely source of the aberrant $f^{new}$ in set 3. Fewer distinct values of $\pi^*$ for each subject under prior set 3 (as discussed above) lead to smaller posterior values of $\alpha_0$ than for prior set 2. There is a small difference of 1.5 between the posterior means of $\alpha_0$ for sets 2 and 3, and this makes little difference in the mean predictive distributions. But in the lower tail of the posterior distribution of $\alpha_0$, this small shift translates into a large difference in the proportion of low values. For set 3, 19.6% of the posterior values of $\alpha_0$ lie below 15, whereas for set 2, only 11.4% are smaller than 15. The effect of a larger proportion of small posterior values of $\alpha_0$ under prior set 3 is to produce a larger proportion of $f^{new}$ with the potential to exhibit unusual behaviour.

Finally, figures 10.12 to 10.15 are presented to give an idea of the sort of information that is available in posterior samples from the model, and also as an aid to understanding the role of $G_0$. These plots use data from a run using prior set 2. Each plot shows the values in $T$ at a given stage of the Gibbs sampler with a large coloured plotting character and the $\pi^*$ values that map to each $\tau^*$ with a smaller plotting character of the same colour. The data were saved just after updating $T|\Pi^*$. The contour lines show the shape of $G_0$ that results from mixing over the parameters in $T$. The patterns observed in the data can be seen in abstract in these graphs: for example, the fact that the precision $\nu$ has higher values at the mode near $-5$. 
Figure 10.11: QQ-plot of Posterior Values of Number of Elements in \( \Pi^* \) in Prior Sets 2 and 3

Set 2: Number of Distinct Values
Dotted line is \( y=x \) reference
Figure 10.12: Sampled Cluster centres and $\pi^*$: The large characters are the elements of $\mathbf{T}$ and the small characters are the elements of $\mathbf{II}^*$. Figures 10.12 to 10.15 are sampled at different steps in one long run using prior set 2.
Figure 10.13: Sampled Cluster centres and $\pi^*$: The large characters are the elements of $T$ and the small characters are the elements of $\Pi^*$. 
Figure 10.14: Sampled Cluster centres and $\pi^*$: The large characters are the elements of $T$ and the small characters are the elements of $\Pi^*$. 
Figure 10.15: Sampled Cluster centres and \( \pi^* \): The large characters are the elements of \( T \) and the small characters are the elements of \( \Pi^* \).
Chapter 11

A Real Application

11.1 Description of Data

This is the data introduced in Chapter 2 and first analyzed by Belin and Rubin [3]; it was analyzed again by Belin and Rubin [4] and again in chapter 16 of Gelman et al. [14]. To recap, reaction times were measured 30 times on each of 17 subjects — 6 schizophrenics and 11 non-schizophrenics. The response times of the schizophrenic subjects are on the average higher and appear to have a higher variance. Belin and Rubin cite psychological theory which states that schizophrenics suffer two types of delays: a general motor reflex delay on all responses and an additional delay on some that is related to attention deficit. Their analysis fits a model that is designed specifically to answer questions about the sizes and probabilities of the two types of delays. The data are highly skewed and the normal mixture model underlying the work in this thesis is not entirely appropriate. As in Belin and Rubin, the natural logarithms of the data will be used. Histograms of the logarithms of the data were shown in figure 2.1.
11.2 Application of Model

It is not possible with the present incarnation of the algorithm in this thesis to explicitly model differences between subjects with and without a diagnosis of schizophrenia. To examine these differences without extending the algorithm, one approach would be to fit separate models to each of the two groups and make comparisons of the predictive distributions, distances, $G_0$ estimates and other quantities of interest. This will not be done here and instead, the analysis will be exploratory and descriptive: the data will be treated without knowledge of subjects' diagnoses. No strong prior assumptions will be made about the number or sizes of delays.

Choice of Priors Choosing priors on a logarithmic scale is a little more difficult than choosing priors on the original scale of the data. Belin and Rubin use an improper uniform for their small set of parameters, which sidesteps the issue of how to elicit prior information on a logarithmic scale. An improper prior is not a possibility here, as exact calculation of marginal distribution and normalizing constants plays a large part in the sampling algorithm. Largely because of a lack of alternatives, fairly flat but proper prior distributions were used for $\Delta_1$ and $\Delta_2$:

$$\Delta_1 \sim \text{Gamma}(0.2, 0.1)$$
$$\Delta_2 \sim \text{Gamma}(0.4, 0.1)$$

The difference in these priors reflects the belief that the variation of $\mu$ around the cluster centres $m$ is smaller than the variation of $m$ around $m_0$. The prior on $w$ was Gamma(4.0,1) (HDR: 7.1 - 79.5), again quite flat over the range of a priori plausible values. A Gamma(10.10/w) prior was used for $v|w$. In the previous chapter, it appeared that the best results came from using a prior on $\alpha_0$ with expectation equal to the sample size. The Dirichlet precision parameters $\alpha_0^j$ were given independent priors Gamma(30,1). This gives a prior expectation of about 21 distinct elements in $\Pi_i$. A Gamma(5.5) prior was placed on $\alpha_{00}$. With 17 samples and 21 $\tau^*$ in each, this corresponds to a prior expectation of about 6 distinct elements in $T$. This expectation is larger than the psychological theory suggests, but the distribution on the distinct
number of elements gives support to a wide range of values, from 1 to 10.

**Starting Values and Number of Iterations:** These were evaluated the same way as in the previous chapter.

### 11.3 Results

Figure 11.1 is a plot of the number of elements in $T$ plotted against iteration number and figure 11.2 similar figure for the number of elements in $\Pi^*$. These were used as a guide to convergence. It appears that the numbers of elements reach a fairly stationary distribution well before 500 iterations. Values after the 500th iteration were used in posterior analysis.

Figure 11.3(a) shows the overall mean predictive distribution and the predictive distributions for all 17 subjects. Figure 11.3(b) shows the mean distribution and the distributions for the subjects with the four largest Hellinger distances $D(f_{\text{obs}}^B, f_i^B)$. These figures show that the 4 largest Hellinger distances correspond to schizophrenic subjects. These four subjects appear to deviate from the population mean function in two different ways. Two subjects have a mode near the population mode, then a second, more dispersed component near a log-time of 6.5; the other two subjects appear to have only a single mode, slightly higher than the population mode, at a value of 6.0. Figure 11.4 plots the empirical cumulative distribution function of the Hellinger distances based on the 17 mean predictive densities. This plot also indicates which subject corresponds to each distance. If the histograms are numbered from the top left, across the rows, the largest distances corresponded to histograms 13, 16, 17, and 15. Histogram number 14 appears to the eye to be quite unlike the histograms for the non-schizophrenic subjects, but has one of the smallest Hellinger distances. The mean function for subject 14 has the longest tail of all the subjects, but also has a mode in the same region as the overall mean function. Since the mean function gives little weight to values larger than 6.5, differences in the region are not reflected in the Hellinger distance. This suggests it is necessary to carefully choose the distance...
Figure 11.1: Trace of Number of Distinct Elements in T
Figure 11.2: Trace of Number of Distinct Elements in $\Pi^*$
Figure 11.3: PDFs for Reaction Time Data. These are the estimates based on the priors in section 11.2. Figure (a) shows the mean predictive density for each subject and the overall mean across subjects. Figure (b) shows the four predictive densities corresponding to the largest Hellinger distances.

(a) Mean Predictive Distribution and Samples

(b) Selected Sample PDFs
Figure 11.4: Empirical Cumulative Distribution Function for the Hellinger Distances $D(f_i^B)$
measure to reflect the types of differences that are of interest. For this particular problem, the Hellinger distance may not be ideal.

Comparison of Figure 11.4 and Figure 2.4 shows that the ordering of the subjects by Hellinger distance is remarkably similar whether kernel density estimates or the model in this thesis are used. Only subject 14 need be moved for the two rankings to be identical. The shrinking of all densities toward a common parent density is reflected by the fact that the Hellinger distances in Figure 11.4 are about half the size of their counterparts under kernel density estimation.

The Hellinger distances between the sampled functions $f_i^h$ and the overall mean $f_{obs}^{hh}$ are easily computed. For each subject, these distances are a sample from the posterior distribution of the distance. Figure 11.5 shows boxplots of 200 sampled distances for each subject. The vertical line separates the schizophrenics from the non-schizophrenics. The differences between the 4 schizophrenics with the largest mean distances and the other subjects, with the exception of subject 13, are seen to be rather small in relation to the amount of variation in the posterior distributions. Again, the interpretation of this relates to the choice of a distance function.

Figure 11.6 plots the posterior mean of $\alpha_0^i$ against its rank within the set of 17 subjects. The two smallest posterior means belong to schizophrenic subjects. The other two schizophrenic subjects with data that appear different from the rest do not have remarkable values of $\alpha_0$. On the whole, $\alpha_0$ does not appear as useful as the distance measure for assessing closeness of individual distributions to the average.

Figures 11.7 and 11.8 show the location of the sampled cluster centres $r^*$ and their associated $\pi^*$ from a single step in the Gibbs sampler. The visual impression gained from the histograms is confirmed — the $r^*$ values with the smaller mean component have larger precision components than the $r^*$ values with larger means. There is definitely one large cluster that appears to coincide with the non-schizophrenic subjects' measurements. The number of clusters needed to represent the delayed responses varies between 2 and 5 over all the Gibbs samples. This variability in the number of delayed response clusters makes it difficult to assess the size of the two delays that
Figure 11.5: Boxplots of Hellinger Distances by Group for Reaction Time Data
theory suggests: but this model does demonstrate that it is not clear that there are
three and only three components to the distribution of response times.

11.4 Special Cases

The role of \( \alpha_0 \) as a smoothing parameter is discussed in sections 9.2 and 10.2. The
analysis in the first part of this chapter uses a Gamma(30, 1) prior. This section will
show the results from two additional analyses, one using a "tiny" \( \alpha_0 \) and one using a
"huge" \( \alpha_0 \).

Using a tiny \( \alpha_0 \) estimates each density function essentially independently of the
others. Recall the equation for a sampled predictive density function for a new ob-
servation on subject \( i \) (8.1):

\[
    f_i^b(Y_{n+1}) \propto \alpha_0 p_0^b(Y_{n+1}) + \sum_j N(Y_{n+1} | \pi_{ij}^b).
\]

The prior \( G_0 \) is estimated from \( \Pi^* \) on all the subjects. So the influence of the other
subjects is expressed through the value of \( G_0 \) in calculating the baseline marginal
distribution \( p_0^b \). Small \( \alpha_0 \) means three things: (1) There will be few distinct values
among the \( \pi^* \). (2) The individual \( G_i \) will be unlike. (3) The parent distribution
\( G_0 \) will not capture the features of the individual \( G_i \). Considered individually, the
sampled density estimates will be similar to finite mixture densities, with a possibly
different number of components in the mixture at each step \( b \) of the Gibbs sampler.

Using a huge \( \alpha_0 \) forces the individual \( G_i \) to resemble \( G_0 \), since values of \( \pi_{ij} | Y_{ij} \)
are almost always sampled from \( G_0 \). Furthermore, in the equation for the predictive
density, the baseline marginal \( f_0(Y) \) is heavily weighted, and atoms in \( \Pi_i \) receive little
weight. The predictive densities \( f_i^b \) are alike for all \( i \) and are approximately equal
to \( f_0(Y) \). In effect, all the points \( Y_{ij} \) are used to estimate \( G_0 \), through the almost
one-to-one mapping of each \( Y_{ij} \) to a \( \pi_j^* \). The estimate of \( \widehat{f_{0|Y}}(Y) \) based on huge \( \alpha_0 \) is
similar to the kernel density estimate based on the data pooled across all subjects.

Figures 11.9 and 11.10 illustrate the effects of these two extreme sizes of \( \alpha_0 \) for
the reaction time data set. While small \( \alpha_0 \) allows for good estimation of individual
Figure 11.6: Mean of $\alpha_0$ by Group for Reaction Time Data

Graph showing the comparison of mean $\alpha_0$ between non-schizophrenic and schizophrenic groups.
Figure 11.7: Reaction Time Data: One Set of Sampled Cluster Centres $\tau^*$ and associated $\pi^*$. For a given colour, the large characters is a $\tau^*$ and the small characters are the $\pi^*$ that map to it.
Figure 11.8: Reaction Time Data: One Set of Sampled Cluster Centres $\tau^*$ and associated $\pi^*$, with colour-coding described in 11.7.
densities, there is no sense of the densities being sampled from a common parent. At the other extreme, extremely large values of $\alpha$ force all density estimates to be equal to each other and a common parent. Individual differences are completely smoothed out. Intermediate values of $\alpha_0$ as used in the rest of the thesis provide a middle ground, both modelling the parent distribution and allowing deviations from this parent.
Figure 11.9: Plot (a) shows the mean predictive distributions for each of the 17 subjects under a Gamma(3, 1) prior for $\alpha_0$. Plot (b) shows the 4 predictive distributions with the largest Hellinger distances. The individual predictive densities are averages of sampled densities that are mostly mixtures of 1 to 4 components.

(a) Mean Predictive Distribution and Samples

(b) Selected Sample PDFs
Figure 11.10: Plot (a) shows the mean predictive distributions for each of the 17 subjects under a Gamma(500,1) prior for $\alpha_0$. Plot (b) shows the 4 predictive distributions with the largest Hellinger distances. The individual predictive densities are indistinguishable from the mean. Note the resemblance of the predictive densities to the kernel density estimate $f_0$ in Figure 2.2.

(a) Mean Predictive Distribution and Samples

(b) Selected Sample PDFs
Chapter 12

Conclusions

12.1 Summary of Findings

A fully Bayesian hierarchical model was designed to estimate both the probability distribution functions for \( n \) sets of data and the underlying probability distribution that gave rise to these \( n \) distributions. Although the full description of both the details of the model and the computations involved are quite lengthy, at bottom, the model has an intuitive design. Individual densities are estimated and a template for these densities is created from the information in the individual estimates. The distance (the exact form of which can be specified for the particular task at hand) between each individual density and the overall average is used to order the data for the observed subjects. New distribution functions, simulated from the posterior distribution of \( f(y) \), can be ranked by the same distance measure to give an idea what new data might look like and also for model checking.

Non-conjugate priors were used in a couple of places in this work. While this necessitated some tedious algebra, it did not prevent the numerical calculation of "exact" marginal or posterior distributions. The flexibility allowed by non-conjugate priors is worth the additional programming and computational time that they require. By introducing adaptive rejection sampling and Gaussian quadrature to the Gibbs sampling algorithm, a model that involves several types of non-conjugacy can be used.
without having to resort to the use of approximate distributions.

By focusing attention on the distributions $G_i$, conceptually this work uses samples from a Dirichlet process as distributions. This falls into a tradition that began with Ferguson’s first paper [13], where he discusses estimation of a distribution. Further work in this vein includes Susarla and Van Ryzin’s work on estimation of a survival function [25, 26] and Kuo’s work on estimating the tolerance distribution in bioassays [17]. In this thesis the large number of distinct values sampled from $G_i$ means that any functional of these distinct values acts like a functional of $G_i$ itself. This use of the Dirichlet process contrasts with its use as a convenient way to detect departures from the specified baseline prior where the sample from the Dirichlet process itself has not been of interest. For example, the work of Escobar and West uses the Dirichlet process as a way of assigning a prior to the few mean-variance parameters of a mixture model.

Extensive use is made of the Dirichlet precision parameter, mostly as a smoothing parameter but also as a potential measure of difference between a given $G_i$ and $G_0$. Highly informative priors for $\alpha_0$ and $\alpha_{00}$ are at the heart of the model, determining the amount of detail in the data that is modelled at each level in the hierarchy.

### 12.2 Future Work

Most of the work in this thesis involves the development of the algorithm for modeling a group of distributions. The use of the distance from the mean for ranking observed subjects was introduced to meet the goals set out when this work began. It is only one possible use of the wealth of information available in the posterior distributions. For any particular application, different aspects of the posterior distributions might be of interest. For example, with the reaction time data, the number of clusters and the locations of the mean components of the cluster centres can be associated with physiological processes, delays in processing. A judicious choice of priors could concentrate most of the posterior distribution on three clusters and the distances
between their means could be calculated at each stage of the Gibbs sampler, giving posterior distributions of the delays. Since the group memberships are known with this data, it would seem sensible to make use of this and introduce some parameters representing group mean differences. There appears to be a lot of potential for shaping this model to a number of specific applications.

Although the use of non-conjugate priors did not prevent computation of the necessary quantities, it did increase the computational time needed. To speed up computations, it might be worth searching for conjugate priors that provide a similar level of flexibility in the model specification.

An issue not discussed here is convergence. The results in the simulated applications appear consistent with the model used to generate the data, but it would be reassuring to have theoretical proofs that the posterior distributions obtained from the sampling schemes do converge to the true ones. Escobar has shown that the basic Dirichlet process Markov chain Monte Carlo (DP-MCMC) algorithm at the heart of this work converges [11], so it may be straightforward to show that the paired DP-MCMC here also converges. Another interesting issue related to convergence is the effect of the approximations that were made (numerical integration, truncation of distributions, Gibbs sub-sampler) on the limiting distribution produced by the Gibbs sampler. This is not a straightforward question, as these approximations enter into the sampling scheme variously as distributions to be sampled and sampling probabilities.

Other measures of dissimilarity could be used to rank the observed subjects, again depending on the particular application. The sampled predictive densities for each subject are saved to a file and can be examined with any measure of dissimilarity that makes sense. Use of several different measures might highlight different aspects of the variability between subjects. The analyses in the last two chapters that hinged on the use of the Hellinger distance should be seen merely as examples of the type of analysis that could be done with posterior samples from this model. Before any serious work is done in the area of detecting outliers in the space of distributions,
the distance measure being used should be thoroughly investigated. Its properties – what it calls close and what it does not – should be appropriate to the task at hand.

The individual sample sizes here were moderate. For larger sample sizes, it might be necessary to consider binning the data and using weighted analysis — the present technique requires too much computational time to be useful with large datasets that need to be analyzed quickly. Fortunately, with increasingly affordable increasingly fast computers, this may not be an issue for very long.
Appendix A

Gaussian Quadrature

If an integral can be written in the form

\[ I = \int_a^b f(x)w(x) \, dx \]

where \( f(x) \) is reasonably smooth and \( w(x) \) is one of a small group of weight functions, then it is possible to obtain an accurate approximation to the integral by a method known in general as Gaussian quadrature. On a finite interval with \( w(x) = 1 \), it is known as Legendre quadrature. On the interval \([0, \infty)\) with \( w(x) = e^{-x}x^B \) and \( B > -1 \), it is known as generalized Laguerre quadrature [22].

For each weight function \( w(x) \), there is an algorithm that produces a set of weights \( W_i \) and values \( t_i \) at which to evaluate the function \( f(.) \). These are used in the following equation to obtain the approximate value of the integral:

\[ I \approx \sum_{i=1}^{n} W_i f(t_i). \]

The integer \( n \) is chosen beforehand; as \( n \) increases, the approximation becomes more accurate at the expense of computation of additional weights \( W_i \).

For generalized Laguerre quadrature, computer code is freely available to compute both the \( t_i \) and the \( W_i \), given \( n \) and \( B \) for a number of weight functions \( w(.) \). The algorithm used for this thesis was obtained from the online software repository at www.netlib.org. It is available at URL http://www.netlib.org/go/gaussq.f
Appendix B

Sampling From a Log-Concave Density

The algorithm for sampling from a log-concave density is a special case of the modified composition method [8]. The domain of a function $f$ is divided into regions and a dominating curve $g_i$ is found for each region.

If $g_i$ is the section of $g$ that dominates $f$ in the region $R_i$, and $q_i$ is the proportion of the area under $g$ in the region $R_i$, a random variable can be generated from $f$ according to the following algorithm:

**REPEAT**

Generate a r.v $Z$ on $\{1, \ldots, K\}$ with probability vector $q_1, \ldots, q_K$
Generate a r.v. $W$ with density $g_Z/q_Z$
Generate $U \sim U[0,1]$
**UNTIL** $Ug_Z(W) \leq f(W)$
RETURN $W$

Below is the algorithm for generating random variables from a log-concave density proportional to $q(w)$. The domain of $w$ is broken into 3 regions. In the central region, the dominating function is uniform with a value equal to the maximum of $q(w)$. Outside of this region, the dominating functions are exponential. To simplify
notation, let \( h(w) = \log(q(w)) \). (from Devroye. 1986. VII.2.6)

1. Evaluate the mode \( m \) of \( q(w) \)

For the density \( r(z|\pi^*) \) in section 6.4.1, a numerical search algorithm is needed to locate the density. For the GIG density, the mode has a closed form solution.

2. Locate \( a \) and \( b \) at which \( q(m + a) = q(m - b) = e^{-1}q(m) \). Finding exact values \( a \) and \( b \) would require numerical solution; instead a crude step search was used. Starting at a value equal to a fraction of \( m \), successively larger candidate values \( b \) were considered until \( q(m - b) < e^{-1}q(m) \). Similarly, for \( a \), successively larger candidate values \( a \) were considered until \( q(m + a) < e^{-1}q(m) \).

3. Simulate candidate random variables \( W \) from a density \( g(w) \) with a uniform central region and exponential tails above and below this central region. In each region, the function \( g(w) \) dominates \( q(w) \).

\[
g(w) = \min \{ q(m - b) \exp \{ [w - (m - b)]h'(m - b) \}, q(m), q(m + a) \exp \{ [w - (m + a)]h'(m + a) \} \} \quad (B.1)
\]

4. Accept or reject candidate values \( W \) according to standard rejection rule inequalities. With \( U \sim \text{Unif}(0, 1) \), accept if \( U \leq \frac{f(W)}{g(W)} \)

### B.1 Details of the Algorithm for Log-Concave Densities

The following initializations are needed. The mixture probabilities \( p_l, p_m, p_r \) are proportional to the areas under \( g(w) \) in the three regions \((-\infty, m - b^*], (m - b^*, m + a^*)\).
and \(|m + a^*; \infty\). Also, \(U\) and \(V\) are sampled from a uniform distribution on \((0, 1)\).

\[
q_m = q(m) \\
\lambda_r = \frac{-1}{h'(m+a)} \\
\lambda_I = \frac{1}{h'(m-b)} \\
a^* = a + \lambda_r \log\left(\frac{q(m+a)}{f_m}\right) \\
b^* = b + \lambda_l \log\left(\frac{q(m-b)}{f_m}\right) \\
s = \lambda_l + \lambda_r + a^* + b^* \\
p_l = \frac{\lambda_l}{s} \\
p_r = \frac{\lambda_r}{s} \\
p_m = \frac{a^* + b^*}{s}
\]

Random variate generation proceeds according to the following algorithm. First, \(U\) is compared to \(p_l, p_m,\) and \(p_r\) to determine whether a candidate value will be generated from one of the exponential tail dominating curves or the central uniform region. A rejection rule then determines whether this value will be used. The value of \(W\) when \(\text{ACCEPT} = 1\) is a sample from \(q(w)\).

REPEAT

\begin{align*}
&\text{GENERATE iid } U, V \\
&\text{IF } U \leq p_m \text{ THEN} \\
&\quad \text{GENERATE } Y \sim U[0, 1] \\
&\quad W = m - b^* + Y(a^* + b^*) \\
&\quad \text{IF } V f_m \leq f(W) \\
&\quad \quad \text{ACCEPT } = 1 \\
&\text{ELSE IF } p_m < U \leq p_m + p_r \text{ THEN} \\
&\quad \text{GENERATE AN EXPONENTIAL R.V. } E = -\log\left(\frac{U-p_m}{p_r}\right) \\
&\quad W = m + a^* + \lambda_r E \\
&\quad \text{IF } V f_m e^{-E} \leq f(W) \\
&\quad \quad \text{ACCEPT } = 1 \\
&\text{ELSE GENERATE AN EXPONENTIAL R.V. } E = -\log\left(\frac{U-p_m-p_r}{1-p_m-p_r}\right) \\
&\quad W = m - b^* - \lambda_l E \\
&\quad \text{IF } V f_m e^{-E} \leq f(W) \\
&\quad \quad \text{ACCEPT } = 1 \\
&\text{UNTIL } \text{ACCEPT} = 1
\end{align*}
For the algorithm described above, with exponential tails and a uniform central region, the rejection rule is slightly obsfucated. A little algebra shows the development of the rejection rule for candidate values from the upper exponential tail. Recall:

\[ g_3(w) = f(m + u)e^{-[w-(m+a)]/\lambda_r} \]
\[ W = m + u - \lambda_r + \lambda_r E. \]

where \( E \sim \exp(1) \). The first line below shows the usual statement of the acceptance condition. The following statements are equivalent.

\[ \int_{g_3(W)} \leq f(W) \]
\[ \int f(m + u) \exp \left\{ -(W - (m + u))/\lambda_r \right\} \leq f(W) \]
\[ \int f(m + u) \exp \left\{ -E + 1 \right\} \leq f(W) \]
\[ \int f(m) \exp \left\{ -E \right\} \leq f(W) \]

Notice that the normalizing constant for \( f(.) \) appears in both sides of the comparison, so it does not have to be calculated.

The area from \( m + u - \lambda_r \) to \( \infty \) under the right exponential tail is given by

\[ \int_{m+a-\lambda_r}^{\infty} g_r(w) \, dw = \int_{m+a-\lambda_r}^{\infty} f(m + u)e^{-[w-(m+a)]/\lambda_r} \, dw \]
\[ = \left[ -f(m + u)\lambda_r e^{-[w-(m+a)]/\lambda_r} \right]_{m+a-\lambda_r}^{\infty} \]
\[ = f(m + u)\lambda_r e^1 \]
\[ = f(m)\lambda_r. \]

Similarly, the area under the left exponential tail between 0 and \( m - b + \lambda_l \) is \( f(m)\lambda_l \).

The area under the uniform density between \( m - b + \lambda_l \) and \( m + u - \lambda_r \) is \( (a + b - \lambda_r - \lambda_l)f(m) \). The total area under \( g(w) \) is \( (a + b)f(m) \). This gives the \( q_i \) for the modified composition method.

\[ q_1 = \lambda_l/(a + b) \quad q_3 = \lambda_r/(a + b) \quad q_2 = 1 - q_1 - q_3 \]
Appendix C

Tabled Values of $P_n(k|\alpha)$.

The tables on the next two pages show the distribution of the number of distinct components ($k$) in samples of sizes 25 and 50 from distributions that have a Dirichlet process prior. The probabilities were computed from the recursive formulas

\[
P_1(1|\alpha) = 1
\]

\[
P_n(k|\alpha) = (\alpha P_{n-1}(k - 1) + (n - 1) P_{n-1}(k)) (\alpha_0 + n)^{-1}, n \geq k > 1
\]

The distributions are shown for $\alpha$ equal to 1, 5, 10, 25 and 50. The tables run up to the largest $k$ such that the probability in the $\alpha = 50$ column is larger than 0.001. Figures 3.1 and 3.2 plot these distributions.
Table C.1: Probability distribution of number of distinct components for $n = 25$ and various $\alpha$
<table>
<thead>
<tr>
<th>k</th>
<th>1.</th>
<th>5.</th>
<th>10.</th>
<th>25.</th>
<th>50.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.200</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.096</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.184</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.234</td>
<td>0.005</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>0.209</td>
<td>0.021</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>6</td>
<td>0.140</td>
<td>0.069</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>7</td>
<td>0.074</td>
<td>0.183</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>8</td>
<td>0.031</td>
<td>0.390</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>9</td>
<td>0.012</td>
<td>0.691</td>
<td>0.009</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>10</td>
<td>0.003</td>
<td>1.032</td>
<td>0.027</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>11</td>
<td>0.000</td>
<td>1.319</td>
<td>0.068</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>12</td>
<td>0.000</td>
<td>1.450</td>
<td>0.150</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>13</td>
<td>0.000</td>
<td>1.409</td>
<td>0.291</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>14</td>
<td>0.000</td>
<td>1.198</td>
<td>0.494</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>15</td>
<td>0.000</td>
<td>0.902</td>
<td>0.744</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>0.604</td>
<td>0.997</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>17</td>
<td>0.000</td>
<td>0.362</td>
<td>1.195</td>
<td>0.005</td>
<td>0.000</td>
</tr>
<tr>
<td>18</td>
<td>0.000</td>
<td>0.195</td>
<td>1.287</td>
<td>0.013</td>
<td>0.000</td>
</tr>
<tr>
<td>19</td>
<td>0.000</td>
<td>0.095</td>
<td>1.248</td>
<td>0.033</td>
<td>0.000</td>
</tr>
<tr>
<td>20</td>
<td>0.000</td>
<td>0.041</td>
<td>1.094</td>
<td>0.071</td>
<td>0.000</td>
</tr>
<tr>
<td>21</td>
<td>0.000</td>
<td>0.016</td>
<td>0.868</td>
<td>0.142</td>
<td>0.000</td>
</tr>
<tr>
<td>22</td>
<td>0.000</td>
<td>0.000</td>
<td>0.626</td>
<td>0.255</td>
<td>0.000</td>
</tr>
<tr>
<td>23</td>
<td>0.000</td>
<td>0.000</td>
<td>0.410</td>
<td>0.417</td>
<td>0.001</td>
</tr>
<tr>
<td>24</td>
<td>0.000</td>
<td>0.000</td>
<td>0.244</td>
<td>0.622</td>
<td>0.004</td>
</tr>
<tr>
<td>25</td>
<td>0.000</td>
<td>0.000</td>
<td>0.133</td>
<td>0.844</td>
<td>0.010</td>
</tr>
<tr>
<td>26</td>
<td>0.000</td>
<td>0.000</td>
<td>0.066</td>
<td>1.044</td>
<td>0.024</td>
</tr>
<tr>
<td>27</td>
<td>0.000</td>
<td>0.000</td>
<td>0.030</td>
<td>1.178</td>
<td>0.055</td>
</tr>
<tr>
<td>28</td>
<td>0.000</td>
<td>0.000</td>
<td>0.012</td>
<td>1.213</td>
<td>0.113</td>
</tr>
<tr>
<td>29</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.139</td>
<td>0.213</td>
</tr>
<tr>
<td>30</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.975</td>
<td>0.364</td>
</tr>
<tr>
<td>31</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.761</td>
<td>0.568</td>
</tr>
<tr>
<td>32</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.540</td>
<td>0.806</td>
</tr>
<tr>
<td>33</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.348</td>
<td>1.039</td>
</tr>
<tr>
<td>34</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.204</td>
<td>1.215</td>
</tr>
<tr>
<td>35</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.108</td>
<td>1.286</td>
</tr>
<tr>
<td>36</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.051</td>
<td>1.228</td>
</tr>
<tr>
<td>37</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.022</td>
<td>1.055</td>
</tr>
<tr>
<td>38</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.012</td>
</tr>
<tr>
<td>39</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.556</td>
</tr>
<tr>
<td>40</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.338</td>
</tr>
<tr>
<td>41</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.180</td>
</tr>
<tr>
<td>42</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.084</td>
</tr>
<tr>
<td>43</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.034</td>
</tr>
<tr>
<td>44</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Table C.2: Probability distribution of number of distinct components for n = 50 and various α.
Appendix D

Computing $f_0(Y)$ and $f_{00}(\pi^*)$

D.1 Computing $f_0(Y)$

The marginal distribution of $Y$ under the baseline prior $G_0(\pi)$ is needed in order to use Escobar’s algorithm for Gibbs sampling the posterior distribution $\Pi_n|Y_n$. In the multiple sample model presented here, the baseline prior $G_0$ places “noise” around $F$, a mixture of $G_{00}(\tau)$ and atoms $\tau_r$ sampled from $G_{00}$. For each of the fixed atoms of $G_0$, the following hierarchical model determines the distribution of $Y$.

\[
Y|\mu, v \sim N(Y|\mu, 1/v) \quad f(Y|\mu, v) \propto v^{1/2} \exp \left(-\frac{v}{2}(Y - \mu)^2\right) \\
\mu|v, m \sim N(\mu|m, 1/(\Delta_2v)) \quad f(\mu|v, m) \propto v^{1/2} \exp \left(-\frac{\Delta_2}{2}(\mu - m)^2\right) \\
v|w \sim \text{Gamma}(v|\alpha, \beta/w) \quad f(v|w) \propto w^{-1} \exp(-\beta v/w)
\]

The marginal distribution of $Y$ for a fixed atom $\tau_r$ under this hierarchical model is calculated by integrating over $\mu$ and $v$ the full joint density of $(Y, \mu, v)$. For a given atom $\tau_r$, call this $f_r(Y)$.

\[
f_r(Y) \propto \int v^\alpha \exp \left[-\frac{1}{2} \left( v(Y - \mu)^2 + \Delta_2v(\mu - m_r)^2 + \frac{2\beta v}{w_r} \right) \right] dv \, d\mu \\
\propto \int v^\alpha \exp \left[-\frac{1}{2} (v + \Delta_2v) \left( \mu - \frac{vY + \Delta_2vm_r}{v + \Delta_2v} \right)^2 + \frac{\Delta_2v^2}{v + \Delta_2v} (Y - m_r)^2 + \frac{2\beta v}{w_r} \right] dv \, d\mu \\
\propto \int v^\alpha (v + \Delta_2v)^{-1/2} \exp \left[-\frac{1}{2} \left\{ \frac{\Delta_2v}{1 + \Delta_2} (Y - m_r)^2 + \frac{2\beta v}{w_r} \right\} \right] dv
\]

149
\[ \propto \int v^{\alpha - 1/2} \exp \left[ -\frac{v}{2} \left( \frac{\Delta_2}{1 + \Delta_2} (Y - m_r)^2 + \frac{2\beta}{w_r} \right) \right] dv \]

The function inside the integral is the kernel of the gamma density with shape parameter \( \alpha + 1/2 \) and scale \( \frac{1}{2} (\frac{\Delta_2}{1 + \Delta_2} (Y - m_r)^2 + \frac{2\beta}{w_r}) \), so that the integration is straightforward.

\[ f_r(Y) \propto \left[ \frac{1}{2} \left( \frac{\Delta_2}{1 + \Delta_2} (Y - m_r)^2 + \frac{2\beta}{w_r} \right) \right]^{-\alpha - 1/2} \Gamma(\alpha + \frac{1}{2}) \]

Writing the proportionality constant back in gives

\[ f_r(Y) = \left( \frac{\beta}{w_r} \right)^\alpha \Gamma \left( \frac{\alpha}{2} + \frac{1}{2} \right) \Gamma \left( \alpha \sqrt{2\pi} \right) \frac{\Delta_2}{1 + \Delta_2} \left[ \frac{1}{2} \frac{\Delta_2}{21 + \Delta_2} (Y - m_r)^2 + \frac{\beta}{w_r} \right]^{-\alpha + \frac{5}{2}} \]

\[ \times \frac{\Gamma \left( \frac{\alpha}{2} + \frac{1}{2} \right)}{\sqrt{2\pi}} \Gamma \left( \frac{\alpha}{2} \right) \frac{\Delta_2}{1 + \Delta_2} \left[ \frac{w_r \Delta_2}{\beta (1 + \Delta_2)} (Y - m_r)^2 + 1 \right]^{-\alpha + \frac{5}{2}} \]

This is the \( t \)-distribution [7] with \( 2\alpha \) degrees of freedom, precision parameter \( \frac{\alpha w_r \Delta_2}{\beta (1 + \Delta_2)} \) and location parameter \( m_r \).

Recall that the baseline prior distribution \( C_{00}(\tau) \) in equation (5.9) is a normal-gamma density. With \( \tau = (m, w) \).

\[ m|w, m_0, \Delta_1 \sim N(m|m_0, 1/(\Delta_1 w)) \]

\[ w \sim \text{Gamma}(w|a, b) \]

The contribution of the baseline prior on \( \tau \) to the marginal distribution of \( Y \) is evaluated by integrating over \( \mu, v, m \) and \( w \) the full joint density \( f(Y, \mu, v, m, w) \). Call this \( P_{00}(Y) \).

\[ f_{00}(Y) = \int N(Y|\mu, 1/v)N(\mu|m, 1/(\Delta_2 v)) \text{Gamma}(v|\alpha, \beta/w)N(m|m_0, 1/(\Delta_1 w)) \text{Gamma}(w|a, b) \]

\[ \propto \int \left( \frac{v}{2\pi} \right)^{\frac{1}{2}} \exp \left[ -\frac{1}{2} (Y - \mu)^2 \right] \left( \frac{\Delta_2 v}{2\pi} \right)^{\frac{1}{2}} \exp \left[ -\frac{\Delta_2 v}{2} (\mu - m)^2 \right] \left( \frac{\beta}{w} \right)^{\alpha} \exp \left( -\frac{\beta v}{w} \right) \Gamma(\alpha) \]

\[ \times \left( \frac{\Delta_1 w}{2\pi} \right)^{\frac{1}{2}} \exp \left[ -\frac{\Delta_1 w}{2} (m - m_0)^2 \right] \left( \frac{b^a \exp \left( -bw \right) w^{a-1}}{\Gamma(a)} \right) dw dm d\mu dv \]
Integration over \( m \) and \( \mu \) through “completing the square” gives a double integral over \( w \) and \( v \):

\[
\mathcal{f}_{\alpha \alpha}(Y) \propto \int_{w,v} \frac{v^{a-\frac{1}{2}}}{\sqrt{\Delta_1 \Delta_2 w + \Delta_1 w + \Delta_2 v}} \times \exp \left[ \frac{-\Delta_1 \Delta_2 vw}{2(\Delta_1 \Delta_2 w + \Delta_1 w + \Delta_2 v)} (m_0 - Y)^2 - \frac{\phi v}{w} - bw \right] w^{a-\frac{1}{2}-\alpha} dw dv.
\]

Making the change of variables \( r = \phi v/w \) and \( s = bw \), integration over \( s \) can be accomplished by noticing that \( (s|r,Y) \) has a gamma distribution with the shape parameter \( a + \frac{1}{2} \) and the scale parameter \( \frac{1}{2b} \cdot \frac{\Delta_1 \Delta_2 (m_0 - Y)^2}{\beta((\Delta_1 \Delta_2 + \Delta_1) + \Delta_2 r)} + 1 \)

\[
\mathcal{P}_{\alpha \alpha}(Y) \propto \int_{r,s} \frac{s^{a-\frac{1}{2}} r^{a-\frac{1}{2}}}{\sqrt{\Delta_1 \Delta_2 + \Delta_1 + \Delta_2 r}} \exp \left[ -s \left( \frac{\Delta_1 \Delta_2 r (m_0 - Y)^2}{(2b)((\beta(\Delta_1 \Delta_2 + \Delta_1) + \Delta_2 r)} + 1 \right) \right] dr ds
\]

Thus, integration over \( s \) is possible and the result is rewritten below. Since the integral can be written as \( \int f(r)r^\beta e^{-r} dr \) it can be approximated using Gaussian quadrature with Laguerre weights (Appendix A).

\[
f_{\alpha \alpha}(Y) = K \int_{\mathcal{r}} (\beta \Delta_1(\Delta_2 + 1) + \Delta_2 r)^{-\frac{a}{2}} \left( \frac{\Delta_1 \Delta_2 r (m_0 - Y)^2}{2(\beta(\Delta_1 \Delta_2 + 1) + \Delta_2 r)} + b \right) \frac{-a-\frac{1}{2}}{r^{a-\frac{1}{2}} e^{-r}} dr
\]

where

\[
K = \frac{\Gamma(a + \frac{1}{2})b^a \sqrt{\Delta_1 \Delta_2}}{\Gamma(a)\Gamma(\alpha)\sqrt{2\pi}}.
\]

For a set of \( \mathbf{T} \) at any given stage in the Gibbs sampler, the marginal \( f(Y) \) can now be calculated as a weighted sum of the contribution from the baseline marginal \( P_{\alpha \alpha} \) and the contributions from \( P_r \), the marginals centred at the atoms \( \tau_r \).

\[
f(Y) = \frac{\alpha_{\alpha \alpha} f_{\alpha \alpha}(Y) + \sum_{r=1}^{n_r} f_r(Y)}{\alpha_{\alpha \alpha} + n_r}
\]

Of course, the number of distinct elements among the elements of \( \mathbf{T} \) will be less than \( n_r \), so this calculation typically requires fewer than \( n_r \) different evaluations of the t-distribution.
D.2 Computing $f_{00}(\pi^*)$

The baseline marginal distribution $f(\pi^*)$ is computed from the following hierarchical model which was laid out in Chapter 6. The superscript $*$ is omitted from $\mu$ and $\nu$ for the sake of notational clarity:

\[
\begin{align*}
\mu|v, m &\sim N(\mu|m, 1/(\Delta v v)) : f(\mu|v, m) \propto v^{1/2} \exp \left( -\frac{\Delta v^2}{2}(\mu - m)^2 \right) \\
v|w &\sim \text{Gamma}(v|\alpha, \beta/w) : f(v|w) \propto v^{(\alpha - 1)} \exp(-\beta v/w) \\
m|w &\sim N(m|m_0, 1/(\Delta w w)) : f(m|w) \propto w^{1/2} \exp \left( -\frac{\Delta w^2}{2}(m - m_0)^2 \right) \\
w &\sim \text{Gamma}(w|a, b) : f(w) \propto w^{a-1} \exp(-bw) 
\end{align*}
\]

Combining these gives:

\[
\begin{align*}
f(\pi^*) &= \int_0^\infty \int_{-\infty}^\infty f(\mu|v, m) f(v|w) f(m|w) f(w) \, dm \, dw \\
&= \int_0^\infty \int_{-\infty}^\infty \sqrt{\frac{\Delta v v}{2\pi}} \exp \left[ -\frac{\Delta v^2}{2}(\mu - m)^2 \right] \frac{(\beta/w)^{\alpha - 1}}{\Gamma(\alpha)} \\
&\quad \times \sqrt{\frac{\Delta w w}{2\pi}} \exp \left[ -\frac{\Delta w^2}{2}(m - m_0)^2 \right] \frac{b^a e^{-bw} w^{a-1}}{\Gamma(a)} \, dm \, dw.
\end{align*}
\]

The two normal densities act as a prior and a likelihood with a sample size of 1, and combine into a normal posterior for $m$. After collecting all terms not involving $m$ and $w$ into a normalizing constant, this becomes:

\[
\begin{align*}
f(\mu, v) &\propto \int_0^\infty \int_{-\infty}^\infty \exp \left[ -\frac{\Delta v^2 + \Delta_1 w}{2} \left\{ m - \frac{\mu \Delta_2 v + m_0 \Delta_1 w}{\Delta_2 v + \Delta_1 w} \right\}^2 \right] \, dm \\
&\quad \times \exp \left[ -\frac{1}{2\Delta_2 v + \Delta_1 w} (\mu - m_0)^2 \right] w^{a-1/2} \exp \left[ -\frac{v^2}{w} - bw \right] \, dw.
\end{align*}
\]

The function of $m$ is proportional to the normal density $N(m|\mu_{\Delta_2 v + m_0 \Delta_1 w}/(\Delta_2 v + \Delta_1 w), 1/(\Delta_2 v + \Delta_1 w))$ so integration over $m$ gives $\sqrt{\frac{2\pi}{\Delta_2 v + \Delta_1 w}}$. The remaining integral in $w$ is proportional to:

\[
\int_0^\infty (\Delta_2 v + \Delta_1 w)^{-\frac{1}{2}} \exp \left[ -\frac{1}{2\Delta_2 v + \Delta_1 w} (\mu - m_0)^2 \right] e^{-bw} w^{a-1/2} \, dw. \tag{D.1}
\]

where the normalizing constant is

\[
\frac{\sqrt{\Delta_2 \Delta_1} v^{\alpha-1/2} \Gamma(a)}{\sqrt{2\pi} \Gamma(a) \Gamma(\alpha)}
\]
The approach from here is contingent on the value of $a - \alpha - 1/2$. If $a - \alpha - 1/2 > -1$, the integrand is proportional to the product of a function $f(w)$ and the gamma density $\text{Gamma}(w|a - \alpha + 1/2, b)$. The transformation $t = bw$ puts the integral in the right form for use of Gaussian quadrature with weights calculated using the Laguerre formulas:

$$f(\pi^*) = \int_0^\infty g(t)t^B e^{-t} dt,$$

where

$$g(t) = \left(b \Delta_2 v + \Delta_1 t\right)^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{\Delta_1 \Delta_2 vt}{b \Delta_2 v + \Delta_1 t} (\mu - m_0)^2 - \frac{i \beta v}{t} \right].$$

and $B = a - \alpha - \frac{1}{2}$ (see Appendix A). The integral is approximated by

$$f(\pi^*) = K \sum_{i=1}^{n} g(t_i)W_i$$

where the normalizing constant is

$$K = \frac{\sqrt{\Delta_1 \Delta_2 v^{a-\frac{1}{2}} (b \beta)^{\alpha}}}{\sqrt{2\pi \Gamma(a) \Gamma(b)}}.$$

The points $t_i$ at which to evaluate the function and the weights $W_i$ are both generated by the quadrature algorithm for the exponent $B$ and the chosen value of $n$.

If $a - \alpha - 1/2 < -1$, the transformation $t = \beta v/w$ is needed to put the integral in (D.1) in the form $\int_0^\infty g(t)t^B e^{-t} dt$. with $B > -1$. Here, $B = \alpha - a - 1$ and

$$g(t) = \left(\Delta_2 t + \Delta_1 \beta\right)^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \frac{\Delta_1 \Delta_2 vt \beta}{b \Delta_2 t + \Delta_1 \beta} (\mu - m_0)^2 - \frac{i \beta v}{t} \right].$$

The integral is again approximated by

$$f(\pi^*) = K \sum_{i=1}^{n} g(t_i)W_i$$

where the normalizing constant is

$$K = \frac{\sqrt{\Delta_1 \Delta_2 v^{a-\frac{1}{2}} \beta^{\alpha+\frac{1}{2}} b^{\alpha}}}{\sqrt{2\pi \Gamma(a) \Gamma(b)}}.$$

In simulations testing the accuracy of the quadrature algorithm on this function, $n = 10$ gave adequate approximations. Two criteria were used for checking the adequacy of the approximation, for a single set of hyperparameters $a, b, \Delta_1, \Delta_2$ and $m_0$:
increasing \( n \) changed \( f(\pi^*) \) little for each trial value of \( \pi^* \): and secondly. Simpson's rule integration of the evaluated \( f(\pi^*) \) over a grid of \( \pi^* \) values gave a value very close to 1. Partly through analytical solution and partly through numerical integration, accurate and fairly rapid computation of \( f(\pi^*) \) is possible.
Appendix E

Sampling from $G_0(\pi|Y)$ and $G_{00}(\tau|\pi^*)$

E.1 Sampling from $G_0(\pi|Y)$

This appendix includes details of sampling from the part of $G_0(\pi|Y)$ shown in equation (5.12). The terms in $m$ form the kernel of a normal density with precision $\Delta_2v + \Delta_1w$. so integration over $m$ gives the value $(\Delta_2v + \Delta_1w)^{-\frac{1}{2}}$.

$$G_0(\pi|Y) = f(Y|\pi) \int \frac{M(\pi|\tau)dG_{00}(\tau)}{\int f(Y|\pi) \int_{m,w} \frac{N(\mu|m, 1/(\Delta_2v))\Gamma\mu(a, \Delta_1w)G(m|a, b) \Gamma\mu(a)\mu(b)}{\int_{m,w} \frac{\Delta_1w}{2\pi} \exp \left[ -\frac{\Delta_1w}{2} (m - m_0)^2 \right] \Gamma\mu(a)\mu(b)}}$$

$$\times e^{-\frac{v}{w} (Y - \mu)^2} \int w \exp \left[ -\frac{\Delta_2v}{2} (\mu - m)^2 \right] \Gamma\mu(a)\mu(b)}$$

$$\times v^\alpha \exp \left[ -\frac{v}{2} (Y - \mu)^2 \right] \int w \exp \left[ -\frac{\Delta_2v + \Delta_1w}{2} (m - \frac{\Delta_2v\mu + \Delta_1wm_0}{\Delta_2v + \Delta_1w}) \right]$$

$$\times \exp \left[ -\frac{1}{2} \frac{\Delta_1\Delta_2vw}{\Delta_1w + \Delta_2v} (m - m_0)^2 \right] \frac{1}{\sqrt{\Delta_2v + \Delta_1w}}$$

155
The proportionality constant is
\[ \frac{\beta^\alpha a^b \sqrt{\Delta_1 \Delta_2}}{\Gamma(a) \Gamma(\alpha)(2\pi)} \cdot \]

The final integral in w could not be solved analytically, so it was not possible to obtain samples of the function \( G_0(\pi|Y) \) directly.

Since integration over w looked formidable, a different approach was tried. The required density is the marginal from a trivariate density in \((\mu, v, w)\). To make it easier to sample these random variables, make the transformation \( s = \frac{v}{w} \) and \( r = w \). If each of the three full conditional densities could easily be sampled then the three densities would constitute a Gibbs sub-sampler. After a suitable number of iterations, the variables \( s, \mu \) and \( r \) would be a sample from their joint density (conditional on other parameters). The variables \( s \) and \( r \) would be transformed back to \( v \) and \( w \) and then \( w \) would be dropped. The sampled \( \mu \) and \( v \) would be a sample from the baseline posterior \( G_0(\pi|Y) \). Unfortunately, it was not possible to efficiently obtain samples from the third conditional density in \( s \). The partial results are presented below as evidence that conjugate priors are still important building blocks in a large hierarchical model.

The Jacobian of the transformation is \( r \), so the full transformed density becomes:

\[
\begin{align*}
  f(\mu, r, s) &= \frac{\beta^\alpha a^b \sqrt{\Delta_1 \Delta_2}}{\Gamma(a) \Gamma(\alpha)(2\pi)} \exp \left( -\frac{rs}{2} (Y - \mu)^2 \right) \frac{1}{\sqrt{\Delta_2 rs + \Delta_1 r}} \\
  &\times \exp \left[ -\frac{1}{2} \frac{\Delta_1 \Delta_2 r^2 s}{\Delta_1 r + \Delta_2 rs} (\mu - m_0)^2 - br - b s \right] r^{a-\alpha+\frac{1}{2}} \\
  &= \frac{\beta^\alpha a^b \sqrt{\Delta_1 \Delta_2}}{\Gamma(a) \Gamma(\alpha)(2\pi)} e^{-\beta s a^b e^{-br}} \\
  &\times \exp \left[ -\frac{rs}{2} (Y - \mu)^2 - \frac{1}{2} \left( \frac{\Delta_1 \Delta_2 rs}{\Delta_1 + \Delta_2 s} \right) (\mu - m_0)^2 \right] \sqrt{\frac{1}{\Delta_2 s + \Delta_1}}.
\end{align*}
\]

Although this is not a standard density, 2 of the 3 conditional densities are easy to sample.
Note that the full conditional density of $\mu$ can be written as:

$$f(\mu | r, s) \propto \exp \left[ -\frac{1}{2} rs \frac{\Delta_1 + \Delta_2 s + \Delta_1 \Delta_2}{\Delta_1 + \Delta_2 s} \left( \mu - \frac{(\Delta_1 + \Delta_2 s)Y + \Delta_1 \Delta_2 m_0}{\Delta_1 + \Delta_2 s + \Delta_1 \Delta_2} \right)^2 \right]$$

so that:

$$\mu | r, s \sim N \left( \frac{(\Delta_1 + \Delta_2 s)Y + \Delta_1 \Delta_2 m_0}{\Delta_1 + \Delta_2 s + \Delta_1 \Delta_2}, rs \frac{\Delta_1 + \Delta_2 s + \Delta_1 \Delta_2}{\Delta_1 + \Delta_2 s} \right)$$

And the full conditional density of $r$ can be written as:

$$f(r | \mu, s) \propto r^a \exp \left[ -r \left( b + \frac{s}{2}(Y - \mu)^2 + \frac{1}{2} \frac{\Delta_1 \Delta_2 s}{\Delta_1 + \Delta_2 s} (\mu - m_0)^2 \right) \right]$$

so that:

$$r | \mu, s \sim \text{Gamma} \left( a + 1, b + \frac{s}{2}(Y - \mu)^2 + \frac{1}{2} \frac{\Delta_1 \Delta_2 s}{\Delta_1 + \Delta_2 s} (\mu - m_0)^2 \right)$$

The conditional density of $s$ can be written as:

$$f(s | \mu, r) \propto \frac{s^a}{\sqrt{\Delta_1 + \Delta_2 s}} \exp \left[ -\frac{s}{2} \left( \frac{2\alpha + r(Y - \mu)^2 + \frac{\Delta_1 \Delta_2 r}{\Delta_1 + \Delta_2 s} (\mu - m_0)^2}{2} \right) \right]$$

This third density can only be tackled with an approach that requires a fair degree of computation. There are two drawbacks to following up the Gibbs sub-sampler approach here. First, the computations involved in obtaining one sample are multiplied by the length of the sub-sampler. And secondly, the convergence of the subsampler cannot be assessed without further computation, so it is possible that the drawn $(\mu, v)$ might be from the incorrect posterior.

### E.2 Sampling from $G_{00}(\tau | \pi^*)$

The baseline posterior distribution for $\tau | \pi^*$ is proportional to the product of the prior $G_{00}(\tau)$ and the likelihood $M(\pi^* | \tau)$.

$$G_{00}(\tau | \pi^*) = \frac{M(\pi^* | \tau)G_{00}(\tau)}{f(\pi^*)} \propto N(\mu^* | m, \Delta_2 v^*) \text{Gamma}(v^* | \alpha, \beta | w) N(m | m_0, \Delta_1 w) \text{Gamma}(w | a, b)$$

$$\propto \exp \left[ -\frac{\Delta_2 v^*}{2} (\mu^* - m)^2 \right] \left( \frac{\beta}{w} \right)^\alpha \exp \left[ -\frac{\beta v^*}{w} \right]$$

$$\times \sqrt{w} \exp \left[ -\frac{\Delta_1 w}{2} (m - m_0)^2 \right] \exp [-bw] w^{a-1}$$
Collecting terms in $m$ and $w$ gives

$$G_{00}(\pi|\pi^*) \propto \exp \left[ -\frac{1}{2} \left( \Delta_2 v^*(\mu^* - m)^2 + \Delta_1 w(m - m_0)^2 \right) - bw - \frac{\beta v^*}{w} \right] w^{a-\frac{1}{2}-\alpha}$$

$$\propto \exp \left[ -\frac{1}{2} \left( \Delta_2 v^* + \Delta_1 w \right) \left( m - \frac{\Delta_2 v^* \mu^* + \Delta_1 w m_0}{2 \Delta_2 v^* + \Delta_1 w} \right)^2 + \frac{\Delta_1 \Delta_2 v^* w}{\Delta_2 v^* + \Delta_1 w} (\mu^* - m_0)^2 \right]$$

$$\times \exp \left[ -bw - \frac{\beta v^*}{w} \right] w^{a-\frac{1}{2}-\alpha}$$

$$\propto \sqrt{\Delta_2 v^* + \Delta_1 w} \exp \left[ -\frac{1}{2} \left( \Delta_2 v^* + \Delta_1 w \right) \left( m - \frac{\Delta_2 v^* \mu^* + \Delta_1 w m_0}{\Delta_2 v^* + \Delta_1 w} \right)^2 \right]$$

$$\times \exp \left[ -\frac{1}{2} \frac{\Delta_1 \Delta_2 v^* w}{\Delta_2 v^* + \Delta_1 w} (\mu^* - m_0)^2 - bw - \frac{\beta v^*}{w} \right] w^{a-\frac{1}{2}-\alpha} (\Delta_2 v^* + \Delta_1 w)^{-\frac{1}{2}}$$

In the last two lines, the density is written as the product of a normal density in $m|w$ and a density in $w$. Sampling proceeds by first sampling $w|\pi^*$ and then sampling $m|w, \pi^*$ from its conditional normal density. This is described in section 6.4.1
Appendix F

Sampling From Remixed Density of $\tau^*$

F.1 Gibbs Sub-Sampler

The likelihood for the configuration $s$ is given by

$$f(\Pi_s^*|\tau^*_s) \propto \prod_{i=1}^{n^*_s} N(\mu_i^*|m_s^*, 1/(\Delta_2 w_s^*)) \Gamma(m^*_s|\alpha, \beta/w_s^*).$$

and the prior for $m_s^*$ and $w_s^*$ is given by

$$f(m_s^*, w_s^*) = N(m_s^*|m_0, 1/(\Delta_1 w_s^*)) \Gamma(w_s^*|a, b).$$

So the joint posterior distribution of $(m_s^*, w_s^*|\Pi_s^*)$ is given by

$$f(m_s^*, w_s^*|\Pi_s^*) \propto \prod_{i=1}^{n_s^*} N(\mu_i^*|m_s^*, 1/(\Delta_2 w_s^*)) \Gamma(m^*_s|\alpha, \beta/w_s^*)$$

$$\times N(m_s^*|m_0, 1/(\Delta_1 w_s^*)) \Gamma(w_s^*|a, b)$$

$$\times \prod_{i=1}^{n_s^*} \exp \left[ -\frac{\Delta_2 v_i^*}{2} (\mu_i^* - m_s^*)^2 \right] w_s^*^{-(\alpha)} \exp \left[ -\beta \frac{v_i^*}{w_s^*} \right]$$

$$\times \sqrt{w_s^*} \exp \left[ -\frac{\Delta_1 w_s^*(m_s^* - m_0)^2}{2} \right] w_s^{(\alpha-1)} \exp \left[ -bw_s^* \right]$$

159
The full conditional for $m^*_s$ is given by

$$f(m^*_s|w^*_s, \Pi^*_s) \propto \exp \left[ -\frac{\Delta_2}{2} \sum_{i} \Delta_i^2 \left( \mu^*_i - m^*_s \right)^2 - \frac{\Delta_1}{2} w^*_s \left( m^*_s - m_0 \right)^2 \right]$$

$$\propto N(m^*_s|\mu^*_i, \Sigma_{SS}^{-1}) N(m_0|w^*_s, 1/(\sum_{i} \Delta_i^2 + \Delta_1 w^*_s)) \quad (F.1)$$

The full conditional for $w^*_s$ is given by

$$f(w^*_s|m^*_s, \Pi^*_s) \propto \exp \left[ -\frac{1}{2} \left( \left( \Delta_1 (m^*_s - m_0)^2 + 2b \right) w^*_s + \frac{2\beta \sum_{i} \Delta_i^2 v^*_i}{w^*_s} \right) \right]$$

$$\times w^*_s^{(-n^*_s \alpha + a - \frac{1}{2})} \quad (F.2)$$

Iterating between these two conditional densities will eventually yield an approximate draw from the joint posterior for $\tau^*_s$. Sampling from $f(m^*_s|w^*_s, \Pi^*_s)$ is straightforward. The next section shows how to sample from $f(w^*_s|m^*_s, \Pi^*_s)$.

### F.2 The Generalized Inverse Gaussian Distribution

The conditional density in $w^*_s$ in equation (F.2) is a generalized inverse Gaussian with parameters $\lambda = -n^*_s \alpha + a + \frac{1}{2}, \Psi = \Delta_1 (m^*_s - m_0)^2 + 2b$ and $\chi = 2\beta \sum_{i} \Delta_i^2 v^*_i$ [8].

Let $x = \text{GIG}(\lambda, \Psi, \chi)$ be a random variable with the generalized inverse Gaussian distribution function with parameters $\lambda \in \mathbb{R}, \chi > 0, \Psi > 0$. The density function of $x$ is

$$f(x) = \frac{\left( \frac{\Psi}{x} \right)^{\frac{\lambda}{2}}}{2K_{\lambda}(\sqrt{\Psi x})} x^{\lambda-1} \exp \left[ -\frac{1}{2} \left( \chi + \Psi x \right) \right] (x > 0)$$

where $K_{\lambda}(.)$ is the modified Bessel function of the third kind, namely

$$K_{\lambda}(u) = \frac{1}{2} \int_{-\infty}^{\infty} \cosh(\lambda u) \exp(-u \cosh(u)) du$$

It is easily shown that

$$\text{GIG}(\lambda, \Psi, \chi) = \sqrt{\frac{\chi}{\Psi}} \text{GIG}(\sqrt{\chi}, \sqrt{\chi})$$

$$\text{GIG}(\lambda, \Psi, \chi) = \frac{1}{\text{GIG}(-\lambda, \chi, \Psi)} \quad (F.3)$$

$$\text{GIG}(\lambda, \Psi, \chi) = \frac{1}{\text{GIG}(-\lambda, \chi, \Psi)} \quad (F.4)$$
so it is necessary only to consider the situation where \( \lambda > 0 \) and \( \Psi = \chi \), as these transformations take care of the other situations. In particular, if \( \lambda < 0 \), as will usually be the case in this model.

\[
\text{GIG}(\lambda, \Psi, \chi) = \sqrt{\frac{\Psi}{\chi}} \frac{1}{\text{GIG}(-\lambda, \sqrt{\chi \Psi}, \sqrt{\chi \Psi})}.
\]

If \( \lambda > 0 \),

\[
\text{GIG}(\lambda, \Psi, \chi) = \sqrt{\frac{\Psi}{\chi}} \text{GIG}(\lambda, \sqrt{\chi \Psi}, \sqrt{\chi \Psi})
\]

For \( \lambda > 1 \), \( f(x) \) is log-concave and can be handled by the algorithm described in Appendix B. If \( 0 < \lambda < 1 \), \( f(x) \) is log-concave for \( x \leq \Psi/(1-\lambda) \); for \( x > \Psi i/(1-\lambda) \), the algorithm is modified by replacing the dominating upper exponential tail with a gamma tail [8]. The mode of \( f(x) \) is incorrect in [8]. The correct value is (for the general case where \( \Psi \) may not be equal to \( \chi \))

\[
m = \sqrt{\left(\frac{\lambda - 1}{\Psi}\right)^2 + \left(\frac{\chi}{\Psi}\right)^2} + \frac{\lambda - 1}{\Psi}.
\]
Bibliography


IMAGE EVALUATION
TEST TARGET (QA-3)

1.0  1.1  1.25

1.4  1.6

150mm

6"

APPLIED IMAGE Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved