Robust Statistical Modeling through Nonparametric Bayesian Methods

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Ju Hee Lee, M.S.

Graduate Program in Statistics

The Ohio State University

2010

Dissertation Committee:

Dr. Steven N. MaEachern, Advisor

Dr. Mario Peruggia

Dr. Christopher M. Hans
© Copyright by

Ju Hee Lee

2010
Nonparametric Bayesian models are commonly used to obtain robust statistical inference, and the most popular nonparametric Bayesian model is, arguably, the mixture of Dirichlet processes (MDP) model. In this study, we examine the question of how to obtain more robustness than under a conventional MDP model. In answer to this question, we develop two models from a nonparametric Bayesian viewpoint, and we investigate their properties: (i) the limiting Dirichlet process ($\limdir$) model, and (ii) the local-mass preserving mixture of Dirichlet process ($LMDP$) model. The $\limdir$ model addresses the question of how to perform a “noninformative” nonparametric Bayesian analysis. Rather than being noninformative, the model requires a slight amount of input, and so provides us with a minimally informative prior distribution with which to conduct a nonparametric Bayesian analysis. The $\limdir$ prior distribution can be viewed as the limit of a sequence of mixture of Dirichlet process models. This model requires only modest input, and yet provides posterior behavior which has a number of important qualitative features, including robustness. Second, the $LMDP$ prior distribution focuses on local mass (defined in the paper). To specify such a prior distribution, we carefully consider the behavior of parameters of interest in some small region, and we then select a prior distribution which preserves mass in the region. Local mass preservation ties the mass of the base measure to its dispersion, resulting in robust inference. These two strategies for constructing a prior distribution can be
applied to any model based on the Dirichlet process. Calibration of the prior distribution is considered. We use the limdir for the compound decision problem and the one-way analysis of variance problem, and compare its performance to that of mixture of Dirichlet processes models and to parametric Bayesian models on actual data sets. We apply the LMDP model for the one-way analysis of variance problem, and compare its performance to that of a mixture of Dirichlet processes model with a conventional prior structure. In addition to developing the robust nonparametric Bayesian models, the latter part of the study describes a general form of consistency which does not necessarily rely on correct specification of the likelihood. We carefully investigate issues of consistency and inconsistency for a variety of functions of interest, such as equality of subsets of treatment means, without the assumption that the model is correct. We prove that Bayes estimators achieve (asymptotic) consistency under some suitable regularity conditions on the assumed likelihood. More importantly, we find a need to distinguish between the notions of two parameters being “equal to one another” and “close to one another”, and we illustrate differences in asymptotic inference for these two statements. This distinction carries with it implications for Bayesian tests of a point null hypothesis.
This is dedicated to my family.
ACKNOWLEDGMENTS

I am immeasurably indebted to my advisor, Professor MacEachern for providing countless help. I really appreciate his warm guidance and cheerful encouragement. I also thank Professor Peruggia for his advice and support throughout my graduate studies. I am grateful to Professor Hans for his helpful advice. My heartfelt gratitude goes to all professors in the department. I truly appreciate the great support that they provided.

Finally, I wish to thank my family and my friend, Hakwoo Kim for their love and support.
VITA

1999 ..................................................... B.S. Ewha Womans University

2006 ..................................................... M.S. The Ohio State University

2008-present ............................... Graduate Research Associate,
The Nationwide Center for Advanced
Customer Insights, The Ohio State
University.

PUBLICATIONS

Research Publications


FIELDS OF STUDY

Major Field: Statistics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>1</td>
</tr>
<tr>
<td>Dedication</td>
<td>3</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>4</td>
</tr>
<tr>
<td>Vita</td>
<td>5</td>
</tr>
<tr>
<td>List of Tables</td>
<td>9</td>
</tr>
<tr>
<td>List of Figures</td>
<td>12</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Literature Review</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Nonparametric Bayesian models</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Computations related to nonparametric Bayesian models based on the Dirichlet process prior distribution</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Consistency of Bayes estimators</td>
<td>17</td>
</tr>
<tr>
<td>3. Minimally Informative Prior Distributions for Nonparametric Bayesian Analysis</td>
<td>23</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>23</td>
</tr>
<tr>
<td>3.2 The limdir model</td>
<td>24</td>
</tr>
<tr>
<td>3.2.1 Limits of Dirichlet processes</td>
<td>28</td>
</tr>
<tr>
<td>3.2.2 Local mass</td>
<td>32</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The Bell exponential number, $B_n$ for some values of $n$.</td>
</tr>
<tr>
<td>3.1</td>
<td>Predictive Performance for the Remaining Season ($X'_j$) Based on the Early Season Data ($X_j$). The mass parameter, $M$, was recalibrated for each month.</td>
</tr>
<tr>
<td>3.2</td>
<td>Predictive Performance of the Models when a Player is Held Out. Evaluation is specific to the National League (NL), the American League (AL), or Major League Baseball (MLB). The mass parameter, $M$, was recalibrated for each of the three evaluations. Performance is measured by $SSPE$, $TSE$, and sum of log marginal likelihoods.</td>
</tr>
<tr>
<td>3.3</td>
<td>Predictive Performance for Sales Rates in 2007 ($X'_j$) Based on Sales Rates in 2006 ($X_j$). Performance is measured by $SSPE$ and sum of log marginal likelihoods. The mass parameter, $M$, was recalibrated for each quarter and each month. For the quarterly data and the monthly data, the mean across quarter (month) is presented.</td>
</tr>
<tr>
<td>3.4</td>
<td>Predictive Performance of the Models when a DMA is Held Out. Evaluation is specific to the full year, each quarter or each month of 2006. Performance is measured by $SSPE$ and sum of log marginal likelihoods. For the quarterly data and the monthly data, the mean across quarter (month) is presented.</td>
</tr>
<tr>
<td>3.5</td>
<td>Example 3. Sucrose Solution Data from Myers and Well. MSE from a one-way ANOVA is 1.706.</td>
</tr>
</tbody>
</table>
3.6 Pairwise Posterior Probabilities of Equality of Any Two Treatment Means, \( P(\theta_i = \theta_j|x) \) for the Sucrose solution data. The posterior probabilities are estimated under the limit of Dirichlet process model and the mixture of Dirichlet processes model. To study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by 1×RMSE, 3×RMSE, 5×RMSE or 10×RMSE.

3.7 Posterior Estimates of the Treatment Means, \( \theta_i \) for \( i = 1, \ldots, 4 \) and the Common Variance for the Observations, \( \sigma^2 \) for the Sucrose solution data. The parameters are estimated under the limit of Dirichlet process model and the mixture of Dirichlet processes model. To study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by 1×RMSE, 3×RMSE, 5×RMSE or 10×RMSE.

4.1 Pairwise Posterior Probabilities of Equality of Any Two Treatment Means, \( P(\theta_i = \theta_j|x) \) for the Sucrose solution data. The posterior probabilities are estimated under the local-mass preserving mixture of Dirichlet process model and the conventional mixture of Dirichlet processes model. To study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by 1×RMSE, 3×RMSE, 5×RMSE or 10×RMSE.

4.2 Posterior Estimates of the Treatment Means, \( \theta_i \), for \( i = 1, \ldots, 4 \), and the Common Variance for the Observations, \( \sigma^2 \), for the Sucrose solution data. The posterior probabilities are estimated under the local-mass preserving mixture of Dirichlet process model and the conventional mixture of Dirichlet processes model. To study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by 1×RMSE, 3×RMSE, 5×RMSE or 10×RMSE.

4.3 Pairwise Posterior Probabilities of Equality of Any Two Treatment Means, \( P(\theta_i = \theta_j|x) \) for the Sucrose solution data. The posterior probabilities are estimated under the local-mass preserving mixture of Dirichlet process model and the conventional mixture of Dirichlet processes model. We increase the scale parameter of the inverse-gamma prior distribution for \( \tau^2 \) by 25 times to inflate the dispersion of the base measure of the Dirichlet process. In addition, to study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by 1×RMSE, 3×RMSE, 5×RMSE or 10×RMSE.
4.4 Posterior Estimates of the Treatment Means, $\theta_i$, for $i = 1, \ldots, 4$, and the Common Variance for the Observations, $\sigma^2$, for the Sucrose solution data. The parameters are estimated under the local-mass preserving mixture of Dirichlet process model and the conventional mixture of Dirichlet processes model. We increase the scale parameter of the inverse-gamma prior distribution for $\tau^2$ by 25 times to inflate the dispersion of the base measure of the Dirichlet process. In addition, to study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by $1 \times$ RMSE, $3 \times$ RMSE, $5 \times$ RMSE or $10 \times$ RMSE.
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Intuition behind the Definition of the Middle and the Tails. Panel $a$ shows a set of normalized likelihoods for individuals with mean $\theta_i$ and $k = 3$. Panel $b$ shows three normal measures in a sequence that we envision tending to a multiple of Lebesgue measure. Panel $c$ shows three measures whose distribution functions match those in panel $b$, but the mass parameters for the three are identical.</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>Example 3. Sucrose Solution Data from Myers and Well. The blue rectangles are the sample mean for each treatment and the red dotted line is the overall mean. MSE from a one-way ANOVA is 1.706.</td>
<td>52</td>
</tr>
<tr>
<td>4.1</td>
<td>Intuition behind Local Mass Preservation. Panel $a$ shows three normal measures with identical total mass parameter $M$. As the dispersion increases, the measure assigns larger mass in the tail region, but it assigns smaller mass in the middle region. Panel $b$ shows three normal measures by matching mass in the middle region of the sharpest measure. The measures assign approximately constant mass in the middle, but they assign more mass in the tail as the dispersion increases.</td>
<td>66</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

In statistical modeling, it is assumed that data are generated from a probability model and that unknown parameters characterize the probability model. The conventional, frequentist approach assumes that the unknown parameters are fixed, and treats the probability model as a function of these fixed, but unknown quantities. To obtain the estimates of the parameters, the frequentist commonly uses the Maximum Likelihood Method, the Method of Moments, the Generalized Method of Moments, Penalized Maximum Likelihood Method, etc.

In contrast to the frequentist approach, the Bayesian approach treats the unknown parameters as random variables. It requires the determination of a prior distribution, a probability distribution of the unknown parameters before collecting the data, based on prior information. Then Bayes theorem combines prior information and the observed data into the posterior distribution, the probability distribution of the parameters after observing the data. Through this procedure, the posterior distribution contains all the information about the parameters from prior information and the data, so one can learn about the parameters solely based on the posterior distribution.
From the above brief description, the prior distribution plays an important role in Bayesian analysis in the sense that it contains all currently available information about the parameters. It helps Bayesian analysis to effectively utilize prior information in modeling. According to assumptions on the prior distribution, Bayesian analysis methods can be categorized into two different approaches, parametric Bayes and nonparametric Bayes. In the parametric Bayes, one assumes that the prior distribution is in some family of parametric distributions with hyperparameters. Since the hyperparameters are usually unknown, one may use observed data to estimate them (Empirical Bayes Analysis) or use hierarchical prior distributions for them (Hierarchical Bayes Analysis). Once a model is specifically stated in either way, the analysis can proceed following the Bayes theorem.

However, oftentimes elicitation of a “correct” form of the prior distribution is impossible due to incomplete prior information. The possibility that a form of the prior distribution is misspecified given rise to some inevitable issues in parametric Bayesian modeling such as sensitivity of posterior distributions to specification of prior distributions. Many researchers have investigated the robustness of Bayesian analysis to possible misspecification of the prior distribution, and developed a variety of robust techniques within the parametric Bayesian framework. However, the more natural way of dealing this problem is nonparametric Bayes (Berger 1993). In the nonparametric Bayes, one assumes only that the parameters independently and identically follow a common but unknown distribution, and puts a second stage prior on the class of all possible prior distributions. Because of no restriction on the form of the prior distribution, nonparametric Bayes models are known to produce more robust inference to misspecification of the form of the prior distribution.
Due to their inherent robustness, the nonparametric Bayesian models have become increasingly used with the recent development of the Markov chain Monte Carlo method. Their theoretical properties, such as asymptotics of the resulting estimators, have been rigorously investigated as well. However, there are many outstanding issues such as the lack of development of noninformative nonparametric Bayesian prior distributions and the need for more carefully structured forms of nonparametric Bayesian models, designed to produce robust inference. In this dissertation, we discuss research questions on how to properly use the nonparametric Bayesian models based on the Dirichlet process. We develop two nonparametric Bayesian models involving the Dirichlet process prior. In additional work, we investigate the consistency issue of Bayes estimators without the assumption that a model is correctly specified.

The outline of this dissertation is as follows: In Chapter 2, we discuss the existing literature on the following topics, nonparametric Bayes models, computations related to nonparametric Bayes models based on the Dirichlet process prior, and consistency of Bayes estimators. First, we review the nonparametric Bayes models with focus on the Dirichlet process. We discuss the properties of nonparametric Bayes models involving the Dirichlet process and their applications to a variety of statistical problems. Understanding the properties is essential to develop the models in the following chapters. Then, we review computational issues that arise under various nonparametric Bayesian models. Finally, we review consistency of Bayes estimators under different settings, such as estimators under parametric Bayesian models and nonparametric Bayesian models. More interestingly we discuss the literature on consistency of Bayes estimators when the assumed model is not correct, which will be developed further in this dissertation.
In Chapter 3, we address the problem of how to conduct a nonparametric Bayesian analysis with minimally informative prior distributions. Noninformative prior distributions are often used in Bayesian analysis to avoid subjective prior specification. They often successfully produce reasonable inference results in parametric Bayesian analysis. In this chapter, we discuss the inadequacies of noninformative prior distributions in nonparametric Bayesian analysis that people use in practice, and develop the limiting Dirichlet process model providing reasonable posterior distributions with only modest input. We illustrate the performance of the limiting Dirichlet process model with actual data sets and compare it to that of other models.

In Chapter 4, we further develop the concept of local mass introduced in Chapter 3, and apply it to specification of hierarchical prior distributions in nonparametric Bayesian models. We develop a strategy of calibrating a class of prior distributions that focus on local mass, resulting in more robust inference. We apply the strategy to mixture of Dirichlet processes models, and develop the local-mass preserving structure of prior distributions for mixture of Dirichlet processes models. We compare mixture of Dirichlet processes models having the local-mass preserving structure of prior distributions to those having conventional structure of prior distributions with an actual data set.

In Chapter 5, we carefully investigate issues of consistency and inconsistency for a variety of functions of interest without the assumption that a model is correctly specified. Assuming that the densities lie in a $K$ dimensional minimal standard exponential family, we examine the asymptotic behavior of Bayes estimators of a parameter and Bayesian hypothesis test. We prove that the Bayes estimator achieves asymptotic consistency under suitable conditions. More importantly, while examining
the consistency of Bayesian hypothesis test, we find the need of distinction between
the hypotheses formulations, “equal to each other” and “close to each other” even in
an asymptotic manner. We discuss the implications of our findings.

Finally, in Chapter 6, we summarize the results presented in Chapters 3 through
5. We also give pointers to some future work that could be carried out in areas related
to those accomplished in this dissertation.
CHAPTER 2

LITERATURE REVIEW

2.1 Nonparametric Bayesian models

Nonparametric Bayesian models have recently received attention from many statisticians as an attractive alternative to parametric Bayesian models. The nonparametric Bayesian models do not rely on parametric assumptions on the form of distribution functions. They treat the unknown distribution function itself as a random quantity. To do this, it requires a random probability measure on a space of probability distribution functions. The most well-known random probability measure is the Dirichlet process, and a good starting point for a discussion of the literature on nonparametric Bayesian models is the paper of Ferguson (1973).

Ferguson (1973) introduced the Dirichlet process as a random probability measure. To set notation, let \( \mathcal{X} \) be a sample space and let \( \mathcal{A} \) be a \( \sigma \)-field of subsets of the sample space. Also, let \( \alpha \) be a finite, positive measure defined on \( (\mathcal{X}, \mathcal{A}) \). Ferguson defined the Dirichlet process in the following way: a random probability distribution, \( F \), is generated from a Dirichlet process with parameter \( \alpha \) if for every \( k = 1, 2, \ldots \) and any measurable partition \( (A_1, \ldots, A_k) \) of \( \mathcal{X} \), the random vector \( (F(A_1), \ldots, F(A_k)) \)
follows a Dirichlet distribution with parameters \((\alpha(A_1), \ldots, \alpha(A_k))\). In his main theorem, he showed that if random variables, \(X_1, \ldots, X_n\) are generated from such \(F\), then the posterior distribution of \(F\) is also a Dirichlet process, but with updated parameter \(\alpha + \sum_{i=1}^{n} \delta_{x_i}\) where \(\delta_x\) denotes a point mass at a point \(x\).

Ferguson applied the Dirichlet process to various statistical problems including the problem of estimating \(F\). While discussing the problem, he provided an interpretation of \(\alpha\) in the Dirichlet process. Let \((\mathcal{X}, \mathcal{F}) = (\mathbb{R}, \mathcal{B})\) where \(\mathbb{R}\) is the real line and \(\mathcal{B}\) the \(\sigma\)-field of Borel sets. Then, \(F_0(t) = \alpha((-\infty, t])/\alpha(\mathbb{R})\) represents the prior guess at the shape of the unknown \(F(t)\), and its total mass, \(M = \alpha(\mathbb{R})\), represents a measure of faith in the prior guess at \(F\), measured in units of numbers of observations. Following this interpretation of \(M\), he viewed the Dirichlet process prior acquired by passing \(M\) to zero as the “noninformative” Dirichlet prior. He supported his interpretation by showing that the resulting Bayes estimates under integrated squared error loss converge to the empirical distribution function.

Ferguson also pointed out a disadvantage of using a Dirichlet process: the distribution \(F\), chosen by a Dirichlet process, generates a new observation exactly equal to one of the previously observed observations with positive probability, and so any \(F\) sampled from a Dirichlet process is almost surely (a.s.) discrete. Antoniak (1974) argued that the a.s. discreteness is not appropriate in many situations, and introduced the mixture of Dirichlet processes (MDP) to remove the constraint of the a.s. discreteness of \(F\). He employed a Dirichlet process for the mixing distribution of a parameter, \(G\), rather than a sampling distribution, \(F\), generating observations, and he called the resulting posterior distribution a MDP (see his Corollary 3.1); suppose that \(G\) is generated from a Dirichlet process on a standard Borel space \((\Theta, \mathcal{A}_0)\) where
\( \Theta \) is a set and \( \mathcal{A}' \) is a \( \sigma \)-field of subsets of \( \Theta \). Let \( \theta \) be a sample from \( G \). If the conditional distribution of \( X \) given \( G \) and \( \theta \) is \( F(\cdot|\theta) \), then the posterior distribution of \( G \) given \( X \) is a MDP. In addition to the derivation, he provided expressions for the posterior distribution, and illustrated applications of the MDP to problems such as estimation of \( G \) and \( \theta \).

The Dirichlet process has a wide range of applicability since it allows one to make flexible inference. Early works on its application include the following: Susarla and Van Ryzin (1976) estimated a survival function from incomplete or censored observations with a Dirichlet process, and compared the resulting nonparametric Bayesian estimator to a frequentist estimator, the product limit estimator. Berry and Christensen (1979) applied a MDP to a binomial problem, and compared its performance to that of other empirical Bayes estimators. Lo (1984) used a MDP to produce Bayesian nonparametric density estimates. Kuo (1988) applied a Dirichlet process to estimate the potency curve in bioassay. Bush and MacEachern (1996) used a Dirichlet process to model a distribution of random effects in a mixed-model ANOVA setting. MacEachern and Müller (2000) briefly reviewed the literature in application of the Dirichlet process to various statistical inference problems.

Blackwell and MacQueen (1973) presented an alternative definition of the Dirichlet process by connecting it to a generalized Pólya urn scheme. When we integrate out \( G \), the prior distribution of \( \theta_i \) given \( \theta_1, \ldots, \theta_{i-1} \) can be represented as

\[
\theta_i|\theta_1, \ldots, \theta_{i-1} \sim \frac{1}{i - 1 + M} \sum_{j=1}^{i-1} \delta_{\theta_j} + \frac{M}{i - 1 + M} G_0.
\]

(2.1)

From this representation (Equation 2.1), we see that \( G \) is a discrete distribution with a countably infinite number of \( \theta \).
Another useful representation of the Dirichlet process was presented by Sethuraman (1994), following work by Sethuraman and Tiwari (1982). Following his constructive definition of the Dirichlet process, the Dirichlet process can also be represented as follows: let $U_h \sim \text{Beta}(1, M)$. Let $\theta_i$ be independent from $U_1, U_2, \ldots$ and $\theta_i \sim G_0$. Define $p_1 = U_1$ and $p_h = U_h \prod_{l=1}^{h-1} (1 - U_l)$. For any $G$ generated from the Dirichlet process with parameter $\alpha$ where $\alpha = MG_0$, $G(A)$ can be defined as $\sum_{h=1}^{\infty} p_h \delta_{\theta_h}(A)$.

With these alternative representations of the Dirichlet process, the computations needed for inference under MDP models can be accelerated. Issues arising from the computations are discussed in Section 2.2.

Sethuraman and Tiwari (1982) showed that as $M \to 0$, $G$ converges to a degenerate probability measure at a particular point in $\Theta$ randomly chosen from $G_0$. Consequently, sending $M$ to zero can be viewed as conveying strong information about $G$ rather than conveying no information. Following this interpretation, as $M \to 0$ in MDP models, under a mild condition on the likelihood, the posterior distribution becomes more and more concentrated on the partition where all the $\theta_i$s are identical. The resulting estimates are the same as the estimates under a parametric model where $\theta \sim G_0$ and $X_i \overset{i.i.d.}{\sim} f(X|\theta), i = 1, \ldots, n$. On the other hand, as $M \to \infty$, a Dirichlet process puts little weight on joining existing clusters and the posterior estimate of $G$ converges to $G_0$. Thus the posterior distribution becomes more and more concentrated on the partition that all the $\theta_i$s are distinct. That is, a typical inference for the vector, $\theta$ tends to the inference under the model where $\theta_i \overset{i.i.d.}{\sim} G_0$ and $X_i \overset{\text{indep.}}{\sim} f(X|\theta_i)$ (Ferguson 1983).

As briefly discussed above, the great advantage of using a Dirichlet process prior is the gain in robustness obtained by removing a parametric assumption on the form of
the prior. The support of a Dirichlet process prior includes all possible distributions for $G$, so it is flexible. On the other hand, parametric models restrict $G$ to be in a parametric family. Inherently, MDP models are more robust to misspecification of $G$ due to consistency results which show that the posterior distribution concentrates near the true distribution generating the $\theta_i$s, say $G_T$, under suitable conditions (see Section 2.3 for more details). Therefore, as MacEachern and Müller (2000) stated, posterior inference under the MDP model converges to posterior inference under $G_T$, as the sample size tends to infinity. However, inference under a parametric model does not converge to $G_T$ if $G_T$ is not in the assumed parametric family. They also raised another interesting issue about a lack of robustness under extended MDP models: a badly outlying single observation or a badly outlying small pocket of observations can have a large effect on overall inference and on inference for each individual $\theta_i$. To be specific, consider the following model discussed in MacEachern and Müller (2000):

$$X_i|\theta_i, \nu \sim \text{N}(\theta_i, \sigma^2) \quad i = 1, \ldots, n,$$

$$\theta_i|\nu \sim G(\theta|\nu)$$

$$G \sim \text{DP}(MG_0) \text{ where } G_0 \sim N(\mu, \tau^2),$$

$$\mu \sim N(\mu_0, \tau_0^2),$$

where $\sigma^2$, $\mu_0$, and $\tau_0^2$ are known, and $\tau^2$ is bounded above. Here, $\nu$ represents parameters, other than $\theta$, added into this hierarchy. This extends the simple MDP model by placing a conjugate normal distribution on the mean $\mu$ of $G_0$. MacEachern and Müller (2000) assumed that an aberrant observation $X_i$ goes to infinity, and studied the behavior of the resulting posterior distribution. As an observation $X_i$ goes to infinity, the center of the posterior distribution of $\mu$ is also pulled toward infinity.
Since the variance is finite, the base measure of the Dirichlet process assigns negligible mass to $\theta$ near the rest of the data, and the posterior distribution puts all the other $\theta_j, j \neq i$, into one cluster. Moreover, the common value of $\theta_j, j \neq i$ is also pulled toward infinity. Despite the inherent robustness of MDP models, adding a hierarchy with a conjugate prior distribution certainly destroys the entire inference, as well as the inference for $\theta_j$, due to the presence of aberrant observations. MacEachern and Müller (2000) suggested two solutions; the first is to plug in a robust estimate for $\mu$ rather than putting a prior distribution on $\mu$. The second is to use a more robust form for the base measure, such as an improper base measure or a more carefully constructed proper base measure.

One may replace a Dirichlet process prior with another random probability measure. Alternatives include product partition models (Hartigan 1990), Pólya trees (Lavine 1992) and Bernstein polynomials (Petrone 1999a and Petrone 1999b). Pólya trees and product partition models include the Dirichlet process prior as a special case. Müller and Quintana (2004) reviewed nonparametric Bayesian methods with application to various statistical problems including density estimation, regression and survival analysis.

2.2 Computations related to nonparametric Bayesian models based on the Dirichlet process prior distribution

MDP models place a prior distribution on countably infinite mixtures, resulting in a posterior distribution of $\theta$ given $X$ involving a mixture over all possible partitions of the $\theta_i$s. Analytical evaluation of the posterior distribution is feasible only for relatively small $n$, and in general, exact evaluation is not feasible. Because of the
computational difficulties involved with MDP models, a great deal of research has been conducted on the development of computational techniques for MDP models.

Despite the large variety of applications discussed in Section 2.1, the heart of an MDP model can be described as a simple Bayes model:

\[
X_i | \theta_i, \nu \sim \text{ind } f(X|\theta_i, \nu) \quad i = 1, \ldots, n, \\
\theta_i | \nu \sim \text{i.i.d. } G(\theta|\nu).
\]

To model uncertainties about the functional form of \(G\), \(G\) is assumed to be randomly generated from a Dirichlet process with parameter \(\alpha = MG_0\). Here, \(\nu\) represents parameters, other than \(\theta\), added into this hierarchy, as before. Under a Dirichlet process prior, \(G\) gives a positive probability to the event that some of the \(\theta_i\)s are equal, and the \(n\) \(\theta_i\)s reduce to some \(p \leq n\) distinct values. We let \(\{\gamma_1, \ldots, \gamma_p\}\) represent the set of the \(p\) distinct values among \(n\) elements of \(\theta\). We introduce an \(n\)-dimensional vector, \(s\), that indicates to which cluster each of the \(\theta_i\) belongs. Let \(s\) be defined by the relation \(s_i = j\) if and only if \(\theta_i\) is in the \(j^{th}\) cluster. We also denote the size of the \(j^{th}\) cluster in a given partition by \(c_j\).

To compute the posterior estimate of \(\theta_i\) under a MDP model, one computes the posterior estimate conditional on a particular partition, and sums those estimates over all possible partitions with the posterior probabilities of the partitions as weights. However, the number of possible partitions among \(\theta_1, \ldots, \theta_n\) increases extremely quickly in \(n\). The number of partitions of \(\theta\) is known to be equal to the Bell exponential number, \(B_n\), which can be defined in a recursive form as \(B_n = \sum_{i=0}^{n-1} \binom{n-1}{i} B_i\), for any \(n \geq 1\), where \(B_0 = 1\) (Berry and Christensen 1979). The values of \(B_n\) for some values of \(n\) are shown in Table 2.1. This implies that analytical evaluation of the posterior estimators under a MDP model is not feasible, except for small values
of \( n \). For estimation, one needs either simplification (as in Susarla and Van Ryzin 1976 for survival analysis) or approximation.

Berry and Christensen (1979) provided a way of parametrically approximating \( G \) to avoid extensive computations, and proceeded to perform the posterior calculation. Kuo (1988) and Lo (1984) developed similar Monte Carlo integration algorithms to evaluate the posterior distribution under the MDP model. However, the computation of Bayes estimators based on an MDP model can be effectively and efficiently conducted through a Gibbs sampling scheme. We construct a Markov chain which has the posterior distribution as its equilibrium distribution, and use a sample path of the Markov chain to approximate inference under the posterior distribution.

Escobar (1994) used a Dirichlet process prior to estimate a vector of normal means, and he presented a Gibbs sampling algorithm to sample the \( \theta_i \)s conditional on \( X \). He integrated out \( G \), and used the Pólya urn scheme. Then, because of exchangeability of the \( \theta_i \)s, the \( \theta_i \) can be sequentially generated as follows:

1. Sample \((\theta_i, s_i)\) for \( i = 1, \ldots, n \), given all other parameters:

   Remove \( \theta_i \) from \( \theta \). Let \( c_j^- \) and \( p^- \) be the cluster size and the number of clusters without \( \theta_i \). If \( \theta_i \) is clustered alone, removing \( \theta_i \) leads to fewer clusters, i.e., \( p^- = p - 1 \), and so relabeling is needed as follows: let \( \gamma_{s_i} = \gamma_p \) and \( s_l = s_i \) for any \( l \) with \( s_l = p \). Then, remove \( \gamma_p \) from \( \gamma \). Simulate each \( \theta_i \) from the

<table>
<thead>
<tr>
<th>( n )</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>10</th>
<th>30</th>
<th>50</th>
<th>70</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_n )</td>
<td>1</td>
<td>5</td>
<td>52</td>
<td>877</td>
<td>1.16E+05</td>
<td>8.47E+23</td>
<td>1.86E+47</td>
<td>1.81E+73</td>
<td>4.76E+115</td>
</tr>
</tbody>
</table>

Table 2.1: The Bell exponential number, \( B_n \) for some values of \( n \).
conditional described below:

\[
P(\theta_i = \gamma_j | \theta_{-i}, X) \propto c_j f(X_i | \gamma_j, \nu) \text{ for } j = 1, \ldots, p^- \\
P(\theta_i \neq \gamma_j | \theta_{-i}, X) \propto M \int f(X_i | \theta, \nu) dG_0 \text{ for any } j = 1, \ldots, p^-, 
\]

where \( \theta_{-i} \) is the vector \( \theta \) without \( \theta_{-i} \). If \( \theta_i \neq \gamma_j \) for all \( j = 1, \ldots, p^- \), then \( \theta_i \) is generated from \( h(\theta | X_i) \) where \( h(\theta | X_i) = g_0(\theta)f(X_i | \theta, \nu)/\int f(X_i | \theta, \nu) dG_0 \)

where \( g_0 \) is the density function corresponding to \( G_0 \).

The vector \((\theta, s)\) is updated after each \((\theta_i, s_i)\) is generated and before \((\theta_{i+1}, s_{i+1})\) is generated.

2. Sample other parameters.

Through sampling the \( \theta_i \)'s, \( s \) is implicitly updated. Escobar and West (1995) and MacEachern (1994) independently extended this computational method to a more general case. These algorithms produce an ergodic Markov chain, but convergence to the posterior distribution may be slow when groups of observations are associated with the same \( \theta \) value with high probability. The algorithm changes the values of \( \theta \) for a single observation at a time, so that the values of \( \theta \) for a large group is only rarely moved (MacEachern 1994). Similarly, the transitions between various other states can occur only rarely (Jain and Neal 2004).

MacEachern (1994) was concerned with the slow convergence rate in the case described above and introduced important algorithms that improve the rate of convergence. He marginalized over \( \theta_i \) and developed an algorithm using \( s \) as the state space of the Gibbs sampler. The algorithm updates \( s_i | s_{-i}, X \) instead of \( \theta_i | \theta_{-i}, X \), that is, \( s_i | s_{-i}, X \) is viewed as a multinomial random variate taking on the value of
one of \((p^- + 1)\) clusters \((p^-\) for existing clusters among the \(\theta_{-i}\) and a \((p^- + 1)^{th}\) cluster to start a new cluster) as follows:

1. Sample \(s_i\) given all other parameters:

\[
P(s_i = j|\theta_{-i}, X) \propto c_j \int f(X_i|\theta, \nu)dH_{-i,j} \text{ for } j = 1, \ldots, p^-
\]

\[
P(s_i = p^- + 1|\theta_{-i}, X) \propto M \int f(X_i|\theta, \nu)dG_0,
\]

where \(H_{-i,j}\) is the posterior distribution of \(\theta\) based on a prior \(G_0\) and observations \(x_l\) for \(s_l = j, l = 1, \ldots, n,\) and \(l \neq i\). If \(s_i = (p^- + 1)\), then generate a new value of \(\theta\) for \(\gamma_{p^-+1}\) from \(h(\theta|X_i)\).

2. Sample other parameters.

MacEachern (1994) illustrated an improvement by his algorithm with examples in which the posterior distribution concentrates on partitions having a small number of clusters. When less conjugacy is present, the algorithm developed by Bush and MacEachern (1996) moves the locations of the clusters each time through the algorithm to achieve faster mixing over posterior. That is, sampling \(\gamma\) conditional on \(s\) and other parameters as follows:

Sample \(\gamma_j\) from the posterior \(p(\gamma_j|s, X, \nu)\), that is,

\[
\gamma_j|s, X \sim p(\gamma_j|s, X, \nu) \propto g_0(\gamma_j|\nu) \prod_{i=1\mid s_i=j}^{n} f(x_i|\gamma_j, \nu).
\]

When \(G_0\) and \(f(X|\theta, \nu)\) are a conjugate pair, the evaluation of \(\int f(X_i|\theta, \nu)dG_0\) in the above algorithms can be easily done. Especially, the algorithm developed in MacEachern (1994) integrates out \(G\) and \(\gamma\) by fully exploiting the conjugacy, so that it updates \(s\) only and achieves better mixing. However, in the non-conjugate case,
evaluation is not trivial. This may result in computational difficulties. To develop a conceptual framework for computational strategies in general situations, MacEachern and Müller (1998) defined a model augmentation based on the “no-gap” algorithm, and presented a MCMC implementation which allows for the use of a non-conjugate pair of $G_0$ and $f(x|\theta, \nu)$. They augmented $\gamma$ to

$$\left\{ \gamma_1, \ldots, \gamma_p, \gamma_{p+1}, \ldots, \gamma_n \right\},$$

by introducing the $(n - p)$ empty clusters. In their algorithm, to avoid the need for integration, they instead generate the locations of empty clusters, and evaluate $f(X|\theta_i, \nu)$. They also discussed an augmentation based on the “complete” algorithm in their technical report. Neal (2000) updated either $s$ or $\theta$ through the Metropolis-Hastings algorithm for non-conjugate cases. He generated the proposals from the conditional prior distributions, $p(s_i|s_{-i})$ or $p(\theta_i|\theta_{-i})$.

The algorithms discussed so far update $s$ or $\theta$ through an incremental sampler. As discussed in Jain and Neal (2004), if any two or more $\gamma_i$ are close to each other, the Markov chain might be trapped in a local mode where $\theta$s associated with those $\gamma$s are merged. Non-incremental Markov chain samplers for MDP models, which simultaneously move a group of $\theta$s to a new cluster in each update, can be alternative for better mixing over the posterior distribution and often leads to faster convergence in such cases. Green and Richardson (2001) and Jain and Neal (2004) have proposed such Markov chain samplers based on split/merge moves. Green and Richardson (2001) introduced a split/merge sampler for MDP models with non-conjugate prior distributions. Their split/merge sampler is implemented within the reversible jump framework since a split/merge move increases/decreases the model dimension, $p$. Jain
and Neal (2004) developed a non-incremental Markov chain sampler based on the Metropolis-Hastings algorithm with split or merge proposals for MDP models with conjugate prior distributions. In their split/merge algorithm, any two $i$ and $i'$ are randomly selected. If they are in the same cluster, say cluster $j$, then propose a partition by splitting $\theta_l$ for $s_l = j$, $l = 1, \ldots, n$. If they are in different clusters, say clusters $j$ and $j'$, then propose a partition by merging $\theta_s$ for $s_l = j$ or $j'$, and evaluate the proposals by the Metropolis-Hastings acceptance probability. In addition, Jain and Neal (2004) elaborated their algorithm by using restricted Gibbs sampling proposals. Jain and Neal (2007) modified the restricted Gibbs sampling scan and extended their split/merge technique to accommodate models with any conditionally conjugate prior distributions.

2.3 Consistency of Bayes estimators

One area of research that attracts many statisticians is asymptotics of Bayes estimators. Because it assumes that there exists a “true value” for parameters, asymptotic consistency is often viewed as a frequentiest criterion of an estimator. However, the property of consistency is also useful to Bayesians since few would want to use a prior distribution producing undesirable or unexpected posterior results (Kleijn and van der Vaart 2006). To set a problem, let observations, $X_1, \ldots, X_n$, be generated from a probability distribution, $f$, with an unknown parameter, $\theta$. The parameter $\theta$ is to be estimated from the data. Bayesian statisticians formulate a probability distribution, $G$ over the unknown $\theta$. It is of interest that as more and more data come in, the posterior distribution concentrates near the true parameter, $\theta_T$. 

17
There is an extensive literature on consistency and inconsistency of Bayesian procedures. Early work focused on establishing conditions under which Bayesian methods produced consistent estimators, often driven by an assumption of a locally constant prior density and an asymptotically normal likelihood. The general tenor of the results is that Bayesian estimators behave similarly to maximum likelihood estimators. Thus, the consistency results and Bayesian central limit theorems of Bernstein and von Mises, with subsequent work in sources such as Savage (1954), Edwards and Savage (1963), Walker (1969), Johnson (1970), Le Cam (1986), and Christensen (2009). In this stream of work, a model is laid out, and a sequence of (often independent and identically distributed) data arises from a sampling model determined by some point in the parameter space. The definition of consistency varies with technical conditions, but is either convergence to the degenerate distribution at the true parameter value or is almost sure convergence to this value. A second stream of research, arguably more Bayesian in nature, was initiated by Doob (1949), who examined behavior of the posterior distribution under the assumption that the prior and likelihood accurately describe the joint marginal distribution of the data. He showed that, under very weak conditions, the Bayes estimators are consistent for all $\theta \in \Theta$ having non zero prior mass.

Another interesting area of research in asymptotics of Bayes estimators is consistency of Bayes estimators in infinite dimensional cases. Nonparametric Bayesian models involve infinite dimensional parameters, and an infinite dimensional parameter space is essential to the creation of nonparametric Bayesian models. In infinite dimensional problems, construction of the prior distribution is more difficult, as is the asymptotic analysis. It is of great interest to Bayesian statisticians to understand how
the prior distribution in the nonparametric Bayesian case influences the asymptotic behavior of the posterior distribution.

Schwartz (1965) proved that if a prior distribution puts positive mass on Kullback-Leibler neighborhoods of a true density function, \( f_T \), then the posterior distribution is weakly consistent at \( f_T \), implying that the posterior distribution concentrates in weak neighborhoods of \( f_T \). Later Ghosal et al. (1999a) showed that the condition of Schwartz is not necessary for consistency, with an example where weak consistency holds although the prior distribution does not put positive mass on the Kullback-Leibler neighborhood of \( f_T \).

On the other hand, Freedman (1963, 1965), and later Diaconis and Freedman (1986a, b) and the associated discussions, highlighted the differences between finite dimensional parameter spaces and infinite dimensional parameter spaces. They showed the difficulty of obtaining consistency for infinite dimensional parameter spaces with several interesting examples.

Freedman (1963) proved that when \( X_i \) are i.i.d discrete samples taking on only finitely many values, the Bayes estimator is consistent for all \( f \) to which a prior distribution assigns positive mass. He proved asymptotic normality of the posterior distribution in the finite case. In addition, he showed that if \( X_i \) is a random sample from an infinite set, even weak consistency can fail to hold for some \( f \) in the support of the prior distribution. Diaconis and Freedman (1986a, b) considered the location problem, \( X_i = \theta + \epsilon_i \), where the \( \epsilon_i \) are independent with unknown distribution \( F \). They assumed that \( F \) follows a Dirichlet process with parameter \( \alpha \), and showed that there exist \( F \)'s with density symmetric about zero, such that, under certain choices of base measures for the Dirichlet process, the posterior distribution of \( \theta \) oscillates...
between two false values. This result again provides an example of a prior distribution which puts positive mass on all weak neighborhoods of \( f_T \), but for which the resulting posterior distribution is not weakly consistent.

Since Freedman (1963, 1965) and Diaconis and Freedman (1986a, b), a host of papers have investigated consistency (and rates of convergence for consistent estimators) in a variety of infinite dimensional settings, or in finite dimensional settings where the dimension of the parameter space grows as the sample size grows. Ghosal et al. (1999b) considered a symmetrized Pólya tree prior for \( G \) instead of a Dirichlet process prior in the example of Diaconis and Freedman (1986a, b). They proved that the posterior distribution of \( \theta \) is consistent for a large class of true symmetric distribution, including the cases of Diaconis and Freedman (1986a, b), by appealing to Schwartz’ theorem.

Weak neighborhoods may be too large, in the sense that they contain many distributions unlike \( f_T \). Many researchers began to examine the posterior consistency for stronger neighborhoods of \( f_T \). Barron et al. (1999) considered a prior distribution putting positive mass in all Kullback-Leibler neighborhoods of \( f_T \), and they provided additional conditions to ensure that the posterior distribution concentrates in every Hellinger neighborhood of \( f_T \). Ghosal et al. (1999c) presented alternative sufficient conditions for consistency in Hellinger neighborhoods. Later, Walker (2003) argued that the conditions in Barron et al. (1999) are not necessary, and he presented new sufficient conditions for strong consistency.

Many studies on rates of convergence have also been conducted in Bayesian nonparametric models. It is known that convergence rates in Bayesian infinite dimensional models are often slower than those from the maximum likelihood estimators
in classical models. Investigations on how to improve the rate are currently active. Among many other references on the convergence rates, Ghosal et al. (2000) and Shen and Wasserman (2001) studied the rate under several metrics. They pointed out two main quantities that determine the rate: the “entropy rate” (a measure of the model complexity), and the amount of prior mass given to suitable neighborhoods around $f_T$. Lijoi et al. (2007) also discussed the convergence rate, but presented alternative elements characterizing the rate.

A separate thread of the literature departs from these traditions. Rather than assuming that the model is correct (either from a frequentist or Bayesian point of view), consistency is studied under the assumption that the likelihood is incorrectly specified. Berk (1966) examined the limiting behavior of a posterior distribution without the assumption that the model used for analysis is correctly specified. He used the term “carrier” to denote the smallest relatively closed set having probability one, and showed that, under certain conditions on the model, the carrier of the prior distribution, and the actual distribution of data, the posterior distribution asymptotically converges to a set (he called the set the asymptotic carrier), which may contain more than one point. He pointed out that even when the model is incorrect, the asymptotic carrier may contain one point, and the posterior distribution becomes degenerate at the value asymptotically, resulting in the posterior consistency. He contrasted his result to that when the model is correct, and provided an example where the asymptotic behavior of the posterior distribution is inconsistent.

Similar to the work of Berk (1966), Bunke and Milhaud (1998) examined convergence properties of the posterior distribution and created a Bayesian central limit theorem under a possibly incorrect model. They defined the pseudotrue parameter(s)
as the point(s) in Θ minimizing the Kullback-Leibler divergence from the true distribution to the assumed likelihood. They provided a different set of assumptions than those of Berk (1966), to establish weak convergence to the pseudotrue parameters. In addition, they proved that, under regularity conditions, the posterior distribution a.s. concentrates on the pseudotrue parameters in a stronger sense. They also presented sufficient conditions under which Bayes estimators can achieve asymptotic normality in the case of a possibly incorrect model under the assumption that the pseudotrue value is unique.

Kleijn and van der Vaart (2006) treated the infinite dimensional case where the true distribution generating data does not belong to the support of the prior distribution. Their main theorem demonstrates sufficient conditions under which there exist points in the support of the prior distribution minimizing the Kullback-Leibler divergence with respect to the true distribution, and the posterior distribution concentrates near the points. They also characterized the rate of concentration near these points with an entropy condition and a prior-mass condition.
CHAPTER 3

MINIMALLY INFORMATIVE PRIOR DISTRIBUTIONS
FOR NONPARAMETERIC BAYESIAN ANALYSIS

3.1 Introduction

An outstanding problem in the area of nonparametric Bayesian analysis is how to perform a noninformative or minimally informative analysis. Technical details of the models render the usual approach of simply selecting a diffuse prior distribution useless: Applying the standard approaches to developing noninformative analyses in this setting leads to inference that, in important ways, does not depend on the data. We develop an alternative formulation which requires only modest input and yet provides reasonable behavior, a posteriori. The analysis that we perform also provides insight into how to use models based on the Dirichlet process when one wishes to use an informative prior distribution.

A new class of models lies at the heart of our analysis. The models, which we term limit of Dirichlet process (limdir) models, are derived from the limit of a sequence of mixture of Dirichlet process models. We describe how to take the limit to ensure that the portion of the posterior distribution with which we are concerned has a proper limiting distribution. We present conditions that ensure this propriety.
We apply the new class of limdir models to one-way analysis of variance and to the compound decision problem. We provide a prior elicitation strategy that allows us to calibrate our minimally informative Bayesian analysis with either subjective input or classical procedures. We note that our approach to nonparametric Bayesian analysis of variance differs from those Tomlinson and Escobar (1999), Müller et al. (2004) and De Iorio et al. (2004).

Section 3.2 describes the new class of models and presents theoretical results for them. Section 3.3 discusses elicitation of the prior distribution. Sections 3.4, 3.5 and 3.6 present a data analysis and comparisons with other methods. The final section contains conclusions.

### 3.2 The limdir model

The development of the limdir model is driven by the goal of providing a sound posterior analysis. The limdir process is the limit of a sequence of Dirichlet processes which, in our setting, tends to an “improper process.” It will be used in a hierarchical model and is thus an extension of the mixture of Dirichlet processes model that has become a staple of the nonparametric Bayesian diet (Dey et al. (1998), Müller and Quintana (2004) and Walker et al. (1999)).

The model for the data, conditional on the treatment means and variances, is the standard one-way analysis of variance model. There are $k$ treatments, with $n_i$ observations on the $i^{th}$ treatment. The observations are mutually independent. Thus, with $\theta_i$ and $\sigma_i^2$ representing the mean and variance of treatment $i$,

$$X_{i1}, \ldots, X_{in_i} | (\theta_i, \sigma_i^2) \overset{iid}{\sim} N(\theta_i, \sigma_i^2) \text{ for } i = 1, \ldots, k.$$  \(3.1\)
For the compound decision problem, we take \( n_i = 1 \). The limdir process is formally defined as the limit, as \( t \to \infty \), of a sequence of Dirichlet processes with base measures \( \{ \alpha_t \}_{t=1}^{\infty} \):

\[
G \sim \text{Dir}(\alpha_t); \quad \theta_1, \ldots, \theta_k \mid G \overset{\text{i.i.d.}}{\sim} G.
\]

(3.2)

The Dirichlet process was described in detail by Ferguson (1973). The parameter of the Dirichlet process, \( \alpha \), is a measure which may be split into two parts. The first is the marginal distribution for \( \theta_i \), say \( G_0 \), while the second is the total mass of the measure, \( M \). Thus \( \alpha \) is often written as \( MG_0 \). The prior (and posterior) for \( \theta = (\theta_1, \ldots, \theta_k) \) is a mixture of several components. Each component corresponds to a different partition of \( \theta \) into \( p \leq k \) subsets, or clusters, where all \( \theta_i \) in a cluster are equal while those in different clusters may differ. The \( k \)-dimensional vector \( s \) indicates to which cluster each of the \( \theta_i \) belongs. It is defined by the relation \( s_i = j \) if and only if \( \theta_i \) is in the \( j \)th cluster. When comparing more than one partition, the first subscript on \( s \) indicates the partition, while a second subscript indicates the position in the vector. For each partition, we represent the sizes of the \( p \) clusters by \( c = (c_1, \ldots, c_p) \) and the locations of the clusters by \( \gamma = (\gamma_1, \ldots, \gamma_p) \).

The Dirichlet process determines the distribution of \((p, \gamma)\). Each partition with \( p \) clusters and cluster sizes \( c \) has a prior probability of \( \pi(s) = M^p \prod_{i=1}^{p} \Gamma(c_i) / \prod_{i=1}^{k} (M + i - 1) \) (Antoniak 1974). Given \( p \), the \( \gamma_i \) form a random sample from \( G_0 \). The description of \( s \) may be governed by many conventions. We adopt the simple convention that, for a partition with \( p \) clusters, the first occurrences of the integers from 1 to \( p \) must occur in order, as \( s \) is read from left to right. This is the “no gaps” scheme for describing \( s \) that is used for computational purposes in MacEachern (1998).
The division of \( \theta \) into clusters is closely tied to the description of \( \theta \) arising from a Pólya urn scheme. Sethuraman (1994) uses this feature to provide a constructive definition of the Dirichlet process. Gopalan and Berry (1998) exploit the clustering to develop multiple comparisons procedures based on a single mixture of Dirichlet processes model: treatment means are judged equivalent if they belong to the same cluster and different if they do not. The probability distribution over partitions of \( \theta \) thus directly provides multiple comparison inferences. Product partition models (Crowley (1997) and Quintana and Iglesias (2003)) also have this feature, as do certain models of partial exchangeability (Mallick and Walker (1997)).

The posterior distribution on \( \theta \), for a fixed \( t \) under the model given by (3.1)-(3.2) with \( G_0 \) a \( N(0, \tau^2_t) \) measure, is a mixture of normal distributions. A component of the mixture results from the partition of \( \theta \) into \( p \) clusters. We use an * to represent observations tied to treatments in a particular cluster. Hence \( n_i^* = \sum_{l=1}^{k} n_l \), and \( \sum_{j=1}^{n_i^*} X_{ij}^* = \sum_{l=1}^{k} \sum_{s_l=i}^{n_l} X_{lj} \) is the sum of all observations tied to cluster \( i \). We define \( SSE_i^* = \sum_{j=1}^{n_i^*} (X_{ij}^* - \bar{X}_i^*)^2 \). With likelihoods (3.1) and known variances, the location vector \( \gamma \) can be marginalized (e.g., MacEachern (1994)). The posterior distribution over the partitions is given by the expression

\[
P_t(s|X) \propto M^p \prod_{i=1}^{p} \left[ \frac{\Gamma(c_i)}{(1 + n_i^* \tau^2_t \sigma^{-2})^{1/2}} \exp \left\{ \frac{\sigma^{-4}(\sum_{j=1}^{n_i^*} X_{ij}^*)^2}{2(\tau^{-2}_t + n_i^* \sigma^{-2})} - \frac{\sum_{j=1}^{n_i^*} (X_{ij}^*)^2}{2\sigma^2} \right\} \right],
\]

when \( \sigma_i^2 = \sigma^2 \) for \( i = 1, \ldots, k \). The conditional (on \( s \) and \( X \)) distribution of the \( p \) cluster locations is a product of independent normal distributions, with means \( \sigma^{-2} n_i^* \bar{X}_i^*/(\tau^{-2}_t + n_i^* \sigma^{-2}) \) and variances \( (\tau^{-2}_t + n_i^* \sigma^{-2})^{-1} \), respectively. The distribution on \((s, \gamma)\) induces a distribution on \( \theta_1, \ldots, \theta_k \). The distribution across partitions leads to the behavior exploited by Berry and Christensen (1979) and Escobar (1994) for the compound decision problem, and described in a general setting by George (1986)
as multiple shrinkage. We refer to the limdir version of this known variance model as Model 0. Formula (3) also applies to the case of known, but differing $\sigma_i^2$. To use it, adjust the $n_i$ to account for the differing $\sigma_i^2$.

The typical analysis of variance problem involves unknown $\sigma_i^2$, so, to be of general use, we must create models that allow for this behavior. We introduce four models to capture the main sorts of normal likelihood multiple comparison problems. In the next sub-section, we formally define the limdir distribution, or limit of Dirichlet processes distribution.

The first model deals with the most common multiple comparisons problem which arises from the one-way analysis of variance: Conditional on the mean vector, $\theta$, and a common variance, $\sigma^2$, the observations follow (3.1). This produces

Model 1. $\theta_1, \ldots, \theta_k \sim \text{limdir}(\{\alpha_t\}); \quad \sigma^2 \text{ has density } \pi(\sigma^2) \propto \sigma^{-2}$.

The second through fourth models provide approaches when the variances associated with the treatments may vary. The second model ensures that all of the treatment variances will differ, and provides a Bayesian solution to a $k$ treatment version of the Behrens-Fisher problem.

Model 2. $\theta_1, \ldots, \theta_k \sim \text{limdir}(\{\alpha_t\}); \quad \sigma_i^2 \text{ has density } \pi(\sigma_i^2) \propto \sigma_i^{-2} \text{ for } i = 1, \ldots, k$.

The third model provides for data driven decisions about whether variances are equal or not. Rather than writing a mixture of Models 1 and 2, we place the limdir distribution on $(\sigma_1^2, \ldots, \sigma_k^2)$ to allow equality of subsets of the variances. In the special case of $k = 2$, Model 3 is exactly a mixture of Models 1 and 2.

Model 3. $\theta_1, \ldots, \theta_k \sim \text{limdir}(\{\alpha_t\}); \quad \sigma_1^2, \ldots, \sigma_k^2 \sim \text{limdir}(\{\beta_t\})$. 

27
The fourth model is targeted at an inferential decision of identical or different treatments. Treatments in the same cluster will have identical mean and variance. Hence a high posterior probability that two treatments are in the same cluster implies a high posterior probability that they share a common mean and variance. Coupling a common two moments with the assumption of normality implies identical distributions. We note that Models 3 and 4 are quite different models.

Model 4. \((\theta_1, \sigma^2_1), \ldots, (\theta_k, \sigma^2_k) \sim \text{limdir}(\{\alpha_t \beta_t\}).\)

For Models 1 through 3, the sequence \(\{\alpha_t\}\) is as previously described. For Model 3, the sequence \(\{\beta_t\}\) consists of inverse gamma distributions with parameters \(a_t\) and \(b_t\). In Model 4, the sequence \(\{\alpha_t \beta_t\}\) consists of the product measure derived from \(\alpha_t\) and \(\beta_t\).

3.2.1 Limits of Dirichlet processes

The limits that we take are motivated by the posterior, conditional view (e.g., Berger (1993) Section 4.4). We seek a limit for which the limiting posteriors have acceptable behavior. We first discuss two standard limits, demonstrating their ineffectiveness for the multiple comparisons and compound decision problems, and then present our choice of limit.

The standard limit when noninformative analyses are discussed for the Dirichlet process is to let \(M \to 0\) (Ferguson (1973)). When \(M \to 0\), and \(\theta\) is observed directly, the posterior estimate of \(G\) under an integrated squared error loss function is just \(\hat{G}\), the empirical cdf. However, when \(\theta\) is observed indirectly, with information about its components filtered through the likelihood, the result is quite different. We
marginalize $\gamma$ and express the posterior as a sum over the partitions of $\theta$.

$$
\pi(s_i|X) = \frac{\pi(s_i)f(X|s_i)}{\sum_{s_j \in S} \pi(s_j)f(X|s_j)}.
$$

(3.4)

The likelihood $f(X|s_i)$ remains constant as $M \to 0$ while $\lim_{M \to 0} \pi(s_i)/\pi(s_j)$ is $\infty$ if $p_{s_i} < p_{s_j}$, and is $\pi(s_i)/\pi(s_j)$ if $p_{s_i} = p_{s_j}$. For our model, the likelihoods are all non-zero, and so the limiting posterior concentrates on the partition with $p = 1$ implying that all $\theta_i$ are equal. In the ANOVA problem, all treatment means are equal with probability 1. In the compound decision problem, all $\theta_i$ are equal with probability 1, and so under typical loss functions all estimates of the $\theta_i$ are the same. In addition, under Models 3 and 4, the posterior also concentrates on the event that all treatment variances are equal. The data have no effect on these limits!

In practice, many have taken $M$ small but positive to represent a minimally informative analysis. However, since the limiting behavior is degenerate, such a choice is highly informative rather than minimally informative. For discussion of this behavior in a variety of settings, see Sethuraman and Tiwari (1982), Newton et al. (1996), or MacEachern (1998).

The standard limit in parametric Bayes models for normal means would take $\tau_i^2 \to \infty$ resulting in an improper, uniform prior on the line for the marginal distribution of a treatment mean. The usual simultaneous limit for the distribution on the variances would result in a limiting prior density proportional to $\sigma^{-2}$ (see, for example, Bernardo and Smith (2000), p.329). Discussion of this limit is complicated by the impropriety of the limiting prior. We assume that $\sum_{i=1}^{k} n_i \geq k + 1$, so that the limiting posterior on $(\theta, \sigma^2)$ is proper. Then, for each pair of partitions $s_i$ and $s_j$,

$$
\lim_{t \to \infty} \frac{\pi(s_i|X)}{\pi(s_j|X)} = \lim_{t \to \infty} \frac{\pi(s_i)f_t(X|s_i)}{\pi(s_j)f_t(X|s_j)}.
$$

(3.5)
Letting $s_j$ denote the partition with $p = 1$ and $s_i$ denote another arbitrary partition, we have, under Model 1,

$$f_t(X|s_j) = \int \int \frac{1}{\tau_l} \phi\left(\frac{\gamma_l}{\tau_l}\right) \sigma^{-2} \prod_{j=1}^{n_j^*} \left\{ \frac{1}{\sigma} \phi\left(\frac{X_{1j}^* - \gamma_l}{\sigma}\right) \right\} d\gamma_1 d\sigma^2 \quad (3.6)$$

and

$$f_t(X|s_i) = \int \ldots \int \sigma^{-2} \prod_{l=1}^{p \mathbf{s}_i} \left[ \frac{1}{\tau_l} \phi\left(\frac{\gamma_l}{\tau_l}\right) \prod_{j=1}^{n_{lj}^*} \left\{ \frac{1}{\sigma} \phi\left(\frac{X_{lj}^* - \gamma_l}{\sigma}\right) \right\} \right] d\gamma_1 \ldots d\gamma_p \mathbf{s}_i d\sigma^2, \quad (3.7)$$

where $\phi$ represents the standard normal probability density function. The ratio $f_t(X|s_i) / f_t(X|s_j)$ tends to 0, implying that the posterior probability of $s_i$ also tends to 0. Considering the result for each partition with $p > 1$ leads to the conclusion that $\pi(s_j|X) = 1$ in the limit. Thus the limiting posterior distribution concentrates on the event $p = 1$, resulting in poor inference. Again, the data have no effect on this limit!

We repair the inadequacies of the two standard limits by focusing on the goal of obtaining a limiting posterior distribution which does not concentrate on the event that all treatment means are equal. Instead, we seek a posterior that assigns positive probability to each partition of $\theta$. We also wish to have a marginal prior distribution for $\theta_i$ that corresponds to the usual parametric reference prior, as in the second limit above. Our solution is to take a limit under which the local mass of the Dirichlet measure tends to a non-zero constant for each compact, non-null measurable set. The particular limit that we take involves a sequence of Dirichlet processes, each with measure $\alpha_t$ in Models 0, 1 and 2 or with measures $\alpha_t$ and $\beta_t$ in Models 3 and 4. As the next subsection shows, the particulars of the sequence that takes one to a given local mass are of little importance, and so we have the freedom to choose
a convenient sequence. The conjugate form for likelihood (3.1) is normal, and this simplifies calculation.

We return briefly to the case of known variance for ease of exposition. Focusing purely on the sequence of measures, \( \{\alpha_t\}_{t=1}^\infty \), we take \( G_0 \) to be normal with mean 0 and variance \( \tau_t^2 \). Our limit will send \( \tau_t^2 \rightarrow \infty \). In order to stabilize the local mass, we define \( M_t = M \sqrt{2\pi \tau_t} \) for some constant \( M > 0 \). The end result is a sequence of measures \( \{\alpha_t\}_{t=1}^\infty \). The measure \( \alpha_t \) assigns to the compact interval \([a,b]\) a mass of 
\[
M_t (\Phi(b/\tau_t) - \Phi(a/\tau_t)) = M \sqrt{2\pi \tau_t} (\Phi(b/\tau_t) - \Phi(a/\tau_t)),
\]
with \( \Phi \) representing the standard normal cdf. The limit of these masses is \( M(b - a) \). Importantly, this limiting value exists, is positive when \( b > a \), and is finite.

Applying this limit to 3.3,
\[
P_t(s|X) \propto (M \sqrt{2\pi \tau_t})^p \prod_{i=1}^p \left[ \frac{\Gamma(c_i)}{(1 + n_i^* \tau_t^2 \sigma^{-2})^{1/2}} \exp \left\{ - \frac{\sum_{j=1}^{n_i^*} X_{ij}^2 - \sigma^{-2}(\sum_{j=1}^{n_i^*} X_{ij}^2)^2}{2\sigma^2} \right\} \right]
\rightarrow M^p (2\pi \sigma^2)^{\frac{p}{2}} \prod_{i=1}^p \left\{ \frac{\Gamma(c_i)(n_i^*)^{-1/2}}{(\sigma^2 + SSE_i^*)} \exp \left( - \frac{\sigma^{-2}SSE_i^*}{2\sigma^2} \right) \right\}.
\]
The conditional limiting posterior distributions on the \( \gamma_i \) are independent normal distributions with means \( \bar{X}_i^* \) and variances \( \sigma^2/n_i^* \). We note that, conditional upon a partition, this simple limdir model captures the standard reference analysis. Differently from the standard reference prior in a parametric Bayes model, the limdir model additionally provides the distribution over the partitions.

The limits that we take for the unknown variance models parallel the known variance case. For the treatment mean Dirichlet process, we select measures \( \alpha_t \) that are proportional to normal distributions with mean 0 and variance \( \tau_t^2 \). The mass of \( \alpha_t \) will be \( M_t = M_t \sqrt{2\pi \tau_t} \). For the variance Dirichlet processes, we select measures \( \beta_t \) that are proportional to inverse gamma distributions with parameters \( a_t \) and \( b_t = \)
1/\alpha_t$. The mass of $\beta_t$ will be $M_t = M_2 \Gamma(\alpha_t) b_t^{a_t}$. In each instance, the limit sends $\tau_t^2 \to \infty$ and $a_t \to 0$. For Model 4, we take $M_1 M_2$ to equal some specified mass. The normalizing constants are included in the expressions to simplify algebraic calculation. The limiting posterior distribution for Models 1 through 4 is described in Appendix A.

There are many ways to generate a noninformative prior distribution for a model. With a single treatment, the model we have written is the standard prior in a location-scale setting (Berger (1993), p. 88). It preserves an a priori independence between the treatment mean and variance. It also gives the standard model, conditional on each partition of $\theta$. We note that limdir versions of other noninformative analyses can be created.

### 3.2.2 Local mass

The results in this subsection justify the concept of local mass. To do so, we consider a limdir model based on a likelihood (to replace (3.1)), satisfying only minimal conditions. The results show that a wide variety of sequences $\{\alpha_t\}_{t=1}^\infty$ lead to the same limiting behavior. The formal condition on $\alpha_t$ in the middle (a compact set that contains the region of substantial likelihood) and the tails (a co-compact set where the likelihood is negligible) is that the tail measure not contribute too much to integrals relative to the middle measure.

Throughout this subsection, we assume that $\theta_i$ lies in $\mathbb{R}$ for $i = 1, \ldots, k$ and that the $P(X_i|\theta_i)$ are mutually absolutely continuous in $\theta_i$. These restriction can be relaxed, though at a cost of considerable complexity in the mathematics. The mutual absolute continuity condition avoids division by 0 in the proof of Theorem 3.2.2.
Definition 3.2.1. Given a configuration $s$ and data $X$, for any fixed $0 < \epsilon < 1$, a set $K = [-k_1, k_1] \times \ldots \times [-k_p, k_p]$ in $\mathbb{R}^p$ is said to be the middle if $(1 - \epsilon) \int \ldots \int P(X|s, \gamma) \alpha(d\gamma_1) \ldots \alpha(d\gamma_p) \leq \int \ldots \int_K P(X|s, \gamma) \alpha(d\gamma_1) \ldots \alpha(d\gamma_p)$ where $\alpha$ is a positive, sigma-finite (possibly infinite), measure. The set $K^c$ is said to be the tail.

Figure 3.1 provides intuition behind this definition of the middle and tail of the measure. Panel $a$ shows a set of normal likelihoods for individuals with means $\theta_i$ and $k = 3$, ignoring potential clustering. Taking $k_1 = 10\sigma + \max_{i=1,\ldots,k}|x_i|$, and setting all other $k_i = k_1$, we note that $\theta$ in the tail have far, far smaller likelihoods than do values of $\theta$ supported by the data. The middle contains all values of $\theta$ that are even remotely supported by the data. Panel $b$ shows three normal measures in a sequence that we envision tending to a multiple of Lebesgue measure. The sharpest, and first in the sequence, assigns little mass in the neighborhood of the likelihoods, and so the posterior distribution assigns great probability to a small number of large clusters. The second measure, with a modest spread, assigns more mass around the likelihoods. With the off-center likelihoods, the posterior distribution is tipped toward a smaller number of larger clusters than is the prior distribution. The third measure is, in the region of the plot, close to a multiple of Lebesgue measure. Posterior inference on $\theta$ under this model is well along the way to the limiting posterior inference obtained under the limdir model. Panel $c$ shows three measures whose distribution functions match those in panel $b$. However, the mass parameters for the three measures are identical. This sequence illustrates the breakdown of the second limit in Section 2.1. The lack of local mass in the region supported by the likelihood implies that the posterior assigns a high probability to a single cluster.
Figure 3.1: Intuition behind the Definition of the Middle and the Tails. Panel a shows a set of normalized likelihoods for individuals with mean $\theta_i$ and $k = 3$. Panel b shows three normal measures in a sequence that we envision tending to a multiple of Lebesgue measure. Panel c shows three measures whose distribution functions match those in panel b, but the mass parameters for the three are identical.

Convergence of the posterior distribution of $s$ in a well-defined sense relies on two conditions.

1. Condition on the Middle

For each $\epsilon > 0$, there exists $T_1$ such that for all $t > T_1$,

$$\left| \int \cdots \int_K P(X|s, \gamma)\alpha_t(d\gamma_1)\cdots\alpha_t(d\gamma_p) - \int \cdots \int_K P(X|s, \gamma)\alpha(d\gamma_1)\cdots\alpha(d\gamma_p) \right| < \epsilon \int \cdots \int_K P(X|s, \gamma)\alpha(d\gamma_1)\cdots\alpha(d\gamma_p).$$

2. Condition on the Tail

For each $\epsilon > 0$, there exists $T_2$ such that for all $t > T_2$,

$$\int \cdots \int_{K^c} P(X|s, \gamma)\alpha_t(d\gamma_1)\cdots\alpha_t(d\gamma_p) < \frac{\epsilon}{1 - \epsilon} \int \cdots \int_K P(X|s, \gamma)\alpha(d\gamma_1)\cdots\alpha(d\gamma_p).$$
The condition on the middle is very mild. The main case of interest is when $\alpha_t$ and $\alpha$ admit densities with respect to Lebesgue measure. In this setting, if a relative difference in densities tends to 0, the condition will hold. Denoting the implied product densities over $K$ by $f_t$ and $f$, the condition will be satisfied whenever a version of the densities exists such that $\sup_{\gamma \in K} |f_t(\gamma) - f(\gamma)|/f(\gamma) \rightarrow 0$. Examples of such convergence include a sequence of scaled normal measures or a sequence of bounded, uniform measures converging to a multiple of Lebesgue measure. Alternatively, if the likelihood is nicely behaved, the condition is satisfied if $\alpha_t - \alpha$ converges to the null measure on $K$. Such nice behavior occurs if, conditional on each $s$, we have $\sup_{\gamma, \gamma' \in K} P(X|s, \gamma)/P(X|s, \gamma') \leq B$ for some $B < \infty$. This would hold, for example, for any likelihood in a full exponential family.

The condition on the tail is best understood by considering an example which violates it. Suppose that the likelihood is standard normal with $n_i = 1$ and $k = 2$. Define the measure $\alpha_t$ to be Lebesgue measure on $[-t, t]$ if $t$ is odd, and Lebesgue measure on $[-t, t]$ with a point mass of measure $\exp\{(t+1)^4\}$ at $t+1$. The condition on the tail (and the condition on the middle) holds for the odd subsequence of measures, and the limiting measure, $\alpha$, is Lebesgue measure on the real line. The limiting posterior distribution assigns positive mass to each of the two possible values of $s$.

For the even subsequence, the posterior distributions are dominated by the spikes of mass. For each $\gamma_i$, these increase as $\exp(t^4)$ while the likelihood decreases at a mere rate of $\exp(-t^2/2)$. The resulting probabilities in the sequence of posterior distributions increase at a rate of $\exp(t^4)$ for $s_1 = s_2$ and at a rate of $\exp(2t^4)$ for $s_1 \neq s_2$. Thus, the even subsequence assigns posterior probability tending to 1 to $s_1 \neq s_2$, and so has different limiting behavior than does the odd subsequence.
Together there is no well-defined limit. Imposition of the condition on the tail rules out sequences of measures that have enormous, growing spikes of mass which move progressively further out in the tails.

**Theorem 3.2.2.** Suppose that there is a sequence of measures, \( \{\alpha_t\}_{t=1}^{\infty} \), and a measure, \( \alpha \). If the sequence \( \{\alpha_t\}_{t=1}^{\infty} \) and \( \alpha \) satisfy the aforementioned conditions on the middle and the tail, then the posterior distribution of \( s \) under the mixture of Dirichlet processes models with base measures \( \alpha_t \) converges to a well-defined limit. We describe this limit as the posterior distribution on \( s \) under the limdir model with base measure \( \alpha \).

**Proof.** Choose a pair of partitions \( u \) and \( v \). Fix \( \epsilon > 0 \) and less than 1, and find a middle and tail for \( u \) that applies to the \( \alpha_t \) and \( \alpha \). Do the same for \( v \). Then, with the second subscript on \( T \) indicating the partition, \( t > \max(T_1u, T_1v, T_2u, T_2v) \),

\[
\frac{P_t(s = u | X)}{P_t(s = v | X)} \leq \frac{(1 - \epsilon^2 + \epsilon) \int \cdots \int_{K_u} P(X | u, \gamma) \prod_{i=1}^{p_U} \Gamma(c_{u_i}) \alpha(d_{\gamma_1}) \cdots \alpha(d_{\gamma_{pU}})}{(1 - \epsilon) \int \cdots \int_{K_v} P(X | v, \gamma) \prod_{i=1}^{p_V} \Gamma(c_{v_i}) \alpha(d_{\gamma_1}) \cdots \alpha(d_{\gamma_{pV}})} \leq \frac{(1 - \epsilon^2 + \epsilon) \int \cdots \int_{K_u} P(X | u, \gamma) \prod_{i=1}^{p_U} \Gamma(c_{u_i}) \alpha(d_{\gamma_1}) \cdots \alpha(d_{\gamma_{pU}})}{(1 - \epsilon)^3 P(u | X)}.
\]

\[
\frac{P_t(s = u | X)}{P_t(s = v | X)} \geq \frac{(1 - \epsilon) \int \cdots \int_{K_u} P(X | u, \gamma) \prod_{i=1}^{p_U} \Gamma(c_{u_i}) \alpha(d_{\gamma_1}) \cdots \alpha(d_{\gamma_{pU}})}{(1 - \epsilon^2 + \epsilon) \int \cdots \int_{K_v} P(X | v, \gamma) \prod_{i=1}^{p_V} \Gamma(c_{v_i}) \alpha(d_{\gamma_1}) \cdots \alpha(d_{\gamma_{pV}})} \geq \frac{(1 - \epsilon) P(u | X)}{(1 - \epsilon^2 + \epsilon) P(v | X)}.
\]

Thus,

\[
\lim_{t \to \infty} \frac{P_t(u | X)}{P_t(v | X)} = \frac{P(u | X)}{P(v | X)}.
\]

36
Applying this argument to all pairs \((u, v)\) yields the result.

To show convergence of the posterior distribution of \(s\), it is enough to show that \(\{\alpha_t\}_{t=0}^\infty\) and \(\alpha\) satisfy the conditions of Theorem 3.2.2.

**Theorem 3.2.3.** Let \(\{\alpha_t\}_{t=1}^\infty\) be a sequence of measures of the Dirichlet process where 
\[ M_t = M \sqrt{2\pi} \tau_t, \quad G_0 \text{ is normal with mean 0 and variance } \tau_t^2, \quad \lim_{t \to \infty} \tau_t^2 = \infty. \]
Define the measure of the limdir process, \(\alpha\), to be \(M\) times Lebesgue measure on \(\mathbb{R}\).
Assume that there exists some finite \(B\) for which \(\int \ldots \int P(X|s, \gamma) \alpha(d\gamma_1) \ldots \alpha(d\gamma_p) < B\) for all \(s\). Then the posterior distribution of \(s\) under the Dirichlet process model with base measure \(\alpha_t\) converges to the posterior distribution of \(s\) under the limdir model with base measure \(\alpha\).

**Proof.** Fix \(s\), \(X\) and \(\epsilon\). By assumption, \(\int \ldots \int P(X|s, \gamma) \alpha(d\gamma_1) \ldots \alpha(d\gamma_p) < B\), and so there exists some \(k_1\) for which the set \(K = [-k_1, k_1]^p\) is a middle of \(\alpha\) as in Definition 3.2.1. Let \(f(\cdot)\) denote the \(N(\cdot; 0, \tau_t^2)\) density. Note that \(M_t f(k_1) < \alpha_t(d\gamma)\) inside the set \(K\) while the inequality is reversed for \(K^c\). This ensures that \(K\) is a middle for \(\alpha_t\) for all \(t\).

1. **Condition on the Middle**

\[
| \int \ldots \int_K P(X|s, \gamma) \alpha_t(d\gamma_1) \ldots \alpha_t(d\gamma_p) - \int \ldots \int_K P(X|s, \gamma) \alpha(d\gamma_1) \ldots \alpha(d\gamma_p) | \\
\leq |M - M_t f(k_1)|^p \int \ldots \int_K P(X|s, \gamma) \alpha(d\gamma_1) \ldots \alpha(d\gamma_p). \]

The constant in front of the integral is \(M^p[1 - \exp(-k_1^2/(2\tau_t^2))]^p\). A large enough \(t\) ensures that the condition holds.
2. Condition on the Tails

We use the fact that \( \alpha_t(d\gamma) / \alpha_t(d\gamma) = \exp\{-\gamma^2/(2\tau_t^2)\} < 1 \). Thus,

\[
\int \ldots \int_{K^c} P(X|s,\gamma) \alpha_t(d\gamma_1) \ldots \alpha_t(d\gamma_p) < \int \ldots \int_{K^c} P(X|s,\gamma) \alpha(d\gamma_1) \ldots \alpha(d\gamma_p)
< \frac{\epsilon}{1-\epsilon} \int \ldots \int_{K} P(X|s,\gamma) \alpha(d\gamma_1) \ldots \alpha(d\gamma_p).
\]

\[\Box\]

3.2.3 Propriety and non-degeneracy of the limiting posteriors

In order to provide effective inferential procedures, we require posterior distributions that are proper for the main quantities of interest (i.e., a proper joint distribution of the treatment means and variance). We also require that the posterior based on a finite sample be non-degenerate. These two properties are satisfied for Models 0 and 1 under mild conditions on the sampling distributions of the \( X_i|\theta_i,\sigma^2_i \). The next result follows from formal calculation.

**Theorem 3.2.4.** The limiting posterior distribution of \( \theta \) calculated under Model 0 with the sequence of measures described in Theorem 3.2.3 is proper. Furthermore, all partitions of \( \theta \) receive positive posterior probability.

When \( \sigma^2 \) is unknown, a continuity condition and an additional sample size condition are needed to ensure the posterior distribution on \( (\theta, \sigma^2) \) is proper. Formal calculation produces the following result for Model 1. Appendix A contains the calculations.

**Theorem 3.2.5.** Assume that \( \sum_{i=1}^k SSE_i > 0 \). Also assume that \( n_i > 1 \) for some \( i = 1, \ldots, k \). Then the limiting posterior distribution of \( (\theta, \sigma^2) \) calculated under Model
with the sequence of measures described in Theorem 3.2.3 is proper. Furthermore, all partitions of $\theta$ receive positive posterior probability.

The next theorem describes a similar result for Models 2, 3 and 4. The additional sample sizes are required because the $\sigma_i$ are allowed to differ. To confirm the non-degeneracy of the limiting distribution on configurations, comparison of ratios of posterior probabilities of configurations is made. The proof consists of straightforward evaluation of the limiting posterior, and so has been omitted.

**Theorem 3.2.6.** Assume that $P(X_{ij} = X_{i'j'}) = 0$ for all pairs of observations $X_{ij}$ and $X_{i'j'}$. Under Models 2, 3 and 4, the limiting posterior distribution of $(\theta, \sigma^2)$ will be proper provided $n_i > 1$ for all $i = 1, \ldots, k$. This condition also ensures that the distribution over partitions of $\theta$ will be non-degenerate in all three models; it ensures that the distribution over partitions of $\sigma^2$ will be non-degenerate in Models 3 and 4.

To conclude the section, we note that the entire limiting posterior distribution is not proper. Under both models, the limiting posterior distribution for a new treatment, say $\theta_{k+1}$, is uniform on the real line. The limiting posterior probability that $\theta_{k+1} = \theta_i$, $i \leq k$, is 0. Under Models 2, 3 and 4, the limiting posterior distribution on $\sigma^2_{k+1}$ is improper, proportional to $\sigma^{-2}_{k+1}$. However, should data be collected on this new treatment, the updated limiting posterior on the expanded $(\theta, \sigma^2)$ would be proper under Models 0 and 1 if $n_{k+1} \geq 1$ and proper under Models 2 through 4 if $n_{k+1} > 1$. Inference on $\theta_{k+1}$ (or $\theta_{k+1}$ and $\sigma^2_{k+1}$) can be made, conditional on its joining an existing cluster. To do so, select one of the $\theta_i$ (or $\theta_i$ and $\sigma^2_i$), $i = 1, \ldots, k$, at random, and set $\theta_{k+1}$ (or $\theta_{k+1}$ and $\sigma^2_{k+1}$) equal to it. We study the performance of this conditional method in the second predictive exercise of Sections 3.4 and 3.5.
3.3 Elicitation of the prior distribution

Our recommended elicitation of the prior distribution consists of two stages. In the first, a limiting model is selected. In the second, a value is chosen for the limiting local mass of the Dirichlet process’ base measure. The choice of limiting local mass may be made to match classical concerns or it may be made to match subjective assessment. With classical concerns in the analysis of variance setting, if the main concern is the pairwise error rate, we recommend a choice of $M$ that yields the targeted pairwise type I error rate. If the main concern is the experimentwise error rate, we recommend a choice of $M$ (or $M_1$ and $M_2$) that yields the targeted experimentwise type I error rate. For both types of calibration, in order to determine whether an error occurs, we need to set a cutoff threshold, $\lambda$.

We calibrate the local mass through a simulation so that the chosen mass parameter achieves a certain pairwise type I error rate. We assume that there are only two treatments, and compute the pairwise error rate with simulation data as follows: Defining $a$ to be the pairwise error rate, we set $a = \frac{1}{N} \sum_{i=1}^{N} I(P^{(i)}(\theta_1 = \theta_2 | \sigma_1^2, \sigma_2^2, X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}) < \lambda)$, where $i = 1, \ldots, N$ indexes the replicate in the simulation. The data, $X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}$, are generated under the assumption that $\theta_1 = \theta_2$. Under Model 1, we can write $P^{(i)}(\theta_1 = \theta_2 | \sigma^2, X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}) = A/(A + MB)$, where the variables $A$ and $B$ are easily simulated. Working with Formula A.1 from Appendix A, we have $A = (n_1 + n_2)^{-\frac{1}{2}} \Gamma(\frac{n_1 + n_2 - 1}{2}) \frac{1}{2} (Y_1 + Y_2)^{\frac{n_1 + n_2 - 1}{2}}$ and $B = (2\pi)^{\frac{1}{2}} n_1^{-\frac{1}{2}} n_2^{-\frac{1}{2}} \Gamma(\frac{n_1 + n_2 - 2}{2}) (\frac{1}{2} Y_1)^{\frac{1}{2}} (\frac{1}{2} Y_2)^{\frac{1}{2}}$, where $Y_1$ is $\sigma^2 \chi^2(n_1 + n_2 - 2)$ and $Y_2$ is independently $\sigma^2 \chi^2(1)$. If the experiment was designed with a particular value of $\sigma^2$ in mind, we use that value of $\sigma^2$. Otherwise, we take $\sigma^2$ to be the mean square error (MSE) from a one-way analysis of variance on the actual data. Working
from these expressions, a simulation is performed to find a value of $M$ for which $a$ is approximately the targeted type I error probability. Experimentwise calibration proceeds in a similar fashion. We generate the data from the null hypothesis that $\theta_1 = \ldots = \theta_k$ and average the experimentwise errors. The simplicity of the simulation enables us to perform the computations to an arbitrary degree of accuracy.

Subjective assessment of the local mass parameter, as with subjective elicitation of any prior distribution, can be done in many ways (e.g., Hartigan (1983), Chapter 1). In our context, one approach is to focus on $E[P(\theta_1 = \theta_2) \mid \sigma^2, X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}]$ under the assumption that $\theta_1 = \theta_2$. The expectation can be evaluated with the help of formulas in Appendix A and either a simulation or numerical integration.

3.4 Example 1: The baseball data

In this section, we illustrate our techniques with a baseball batting average data set. The data set consists of the results of at-bats for players in Major League Baseball in the 2005 season from April through September obtained at http://sports.espn.go.com/mlb/statistics, and broken down by month. For each month, we used the batting averages from at-bats from April through that month to predict performance for the remainder of the season. From the beginning of the season until the $j$th month, the $i$th player had $H_{ji}$ hits in $AB_{ji}$ at-bats. We assume that $H_{ji} \sim Bin(AB_{ji}, p_i)$. The parameter $p_i$ describes the $i$th player’s true but unknown ability. Following Brown (2008), we use the variance stabilizing transformation, $X_{ji} = arcsin(\sqrt{(H_{ji} + 1)/4}/(AB_{ji} + 1/2))$, so that $X_{ji}$ is approximately normal with
mean $\theta_i = \arcsin(\sqrt{p_i})$ and variance $\sigma^2_{ji} = 1/(4AB_{ji})$. To make the normal approximation accurate, our analysis included only batters having $H_{ji} > 10$ and having more than 10 at-bats after the $j$th month.

We illustrate the limdir model and compare it to a mixture of Dirichlet processes model and a parametric hierarchical Bayes model through two predictive exercises. The first is prediction of the remainder of the season based upon the first portion of the season. The second is prediction of a player’s performance based on other players’ performances. To study the impact of including pitchers in the analysis, we group the players by their league (American League (AL) or National League (NL)) in addition to pooling all of the players (Major League Baseball (MLB)).

For each predictive exercise, we examined a variety of calibrations of the models. The distribution of the $X_{ji}$ appears to be unimodal, with noticeable tail decay, and with left skewness. In our judgement, the normal distribution with mean $\mu$ and variance $\tau^2$ is a reasonable choice for a base measure of the mixture of Dirichlet processes model. For the parametric Bayes model, an improper prior was chosen for $\mu$ and $\tau^2$ which was $\propto 1/\tau^2$.

The limdir model was implemented with a base measure proportional to Lebesgue measure. The mass parameter for the mixture of Dirichlet processes model was chosen on the basis of the pairwise error rate as described in Section 3.3. The pairwise type I error rate was taken to be 0.05 and the cutoff threshold was set at 0.5. The variance used in the calibration was the mean of the $\sigma^2_i$. The resulting mass parameters ranged from 0.714 to 0.977. The local mass for the limdir model was calibrated in a similar fashion. This led to local mass parameters ranging from 0.472 to 1.133. Note that the
local mass and the mass for the mixture of Dirichlet processes model are not directly comparable, as constants have been swept into the local mass to simplify calculations.

The estimators are compared on the basis of sum of squared prediction error, approximate total squared error, and the sum of log marginal likelihood. When predicting the performance of the remaining months, sum of squared prediction error is given by \( SSPE = \sum_{i=1}^{n_j} (X'_{ji} - \hat{\theta}_i)^2 \), where \( X'_{ji} \) is a performance of the \( i \)th player after the \( j \)th month, and \( \hat{\theta}_i \) is the posterior mean of \( \theta_i \). The approximate total squared error adjusts \( SSPE \) for sampling variation in the remaining season performance. Thus \( TSE = SSPE - \sum_{i=1}^{n_j} \sigma_{ji}^2 \). The sum of log marginal likelihoods is \( \sum_{i=1}^{n_j} \log(m(X'_{ji}|X_j)) \). For a single player, this is the log marginal likelihood. Summing across players provides an indication of the fit of competing models on a log-likelihood scale. Computationally, it is far more stable than is the log joint marginal likelihood of all players.

We evaluate \( \hat{\theta}_i \) with a Gibbs sampling run for the limdir model and the mixture of Dirichlet processes models. 200,000 iterates were used for estimation, after a burn-in period of 10,000 iterates. For the parametric Bayes model, we used numerical integration over a fine grid for \( \tau^2 \) to evaluate \( \hat{\theta}_i \). Table 3.1 presents the results of the first predictive exercise. From the preliminary analysis, we note that the distribution of the \( X_{ji} \) for the AL is approximately symmetric, but the distribution of the \( X_{ji} \) for the NL and MLB is skewed to the left primarily due to pitchers’ poor batting performance. The amount of player-specific information in the early season data increases as the cut-off month moves later, providing more information about the underlying distribution of players’ abilities and providing a sharper likelihood for each player. The additional information changes the relative performance of the
Month | April | May | June | July | August
--- | --- | --- | --- | --- | ---
National League
Limiting Dirichlet Process Model
\( M \) | 0.472 | 0.555 | 0.622 | 0.697 | 0.873
\( SSPE \) | 0.655 | 1.241 | 1.366 | 1.856 | 2.216
\( \widetilde{TSE} \) | 0.387 | 0.653 | 0.564 | 0.762 | 0.951
\( \log(m(X'_j)|X_j) \) | 241.02 | 255.89 | 281.06 | 256.21 | 167.56
Mixture of Dirichlet Processes Model
\( M \) | 0.950 | 0.804 | 0.772 | 0.745 | 0.714
\( SSPE \) | 0.755 | 1.267 | 1.486 | 1.863 | 2.104
\( \widetilde{TSE} \) | 0.487 | 0.679 | 0.684 | 0.769 | 0.839
\( \log(m(X'_j)|X_j) \) | 219.19 | 258.91 | 274.64 | 255.36 | 169.63
Parametric Bayes Model
\( SSPE \) | 0.748 | 1.373 | 1.640 | 2.086 | 2.138
\( \widetilde{TSE} \) | 0.480 | 0.785 | 0.838 | 0.993 | 0.874
\( \log(m(X'_j)|X_j) \) | 225.59 | 245.03 | 260.49 | 244.37 | 169.46
American League
Limiting Dirichlet Process Model
\( M \) | 0.560 | 0.751 | 0.869 | 0.963 | 1.133
\( SSPE \) | 0.241 | 0.240 | 0.341 | 0.423 | 0.724
\( \widetilde{TSE} \) | 0.073 | 0.004 | -0.009 | 0.001 | 0.000
\( \log(m(X'_j)|X_j) \) | 256.70 | 260.95 | 267.86 | 238.13 | 191.83
Mixture of Dirichlet Processes Model
\( M \) | 0.945 | 0.886 | 0.900 | 0.977 | 0.945
\( SSPE \) | 0.242 | 0.242 | 0.348 | 0.425 | 0.724
\( \widetilde{TSE} \) | 0.073 | 0.006 | -0.002 | 0.003 | 0.000
\( \log(m(X'_j)|X_j) \) | 257.34 | 262.36 | 268.84 | 238.09 | 191.93
Parametric Bayes Model
\( SSPE \) | 0.239 | 0.241 | 0.344 | 0.420 | 0.717
\( \widetilde{TSE} \) | 0.071 | 0.005 | -0.005 | -0.002 | -0.007
\( \log(m(X'_j)|X_j) \) | 258.85 | 262.92 | 269.42 | 239.46 | 192.87
Major League Baseball
Limiting Dirichlet Process Model
\( M \) | 0.508 | 0.618 | 0.699 | 0.779 | 0.967
\( SSPE \) | 0.948 | 1.502 | 1.811 | 2.423 | 2.954
\( \widetilde{TSE} \) | 0.511 | 0.677 | 0.659 | 0.907 | 0.966
\( \log(m(X'_j)|X_j) \) | 492.45 | 514.33 | 544.63 | 492.55 | 359.48
Mixture of Dirichlet Processes Model
\( M \) | 0.949 | 0.852 | 0.836 | 0.850 | 0.817
\( SSPE \) | 0.995 | 1.704 | 2.174 | 2.731 | 2.915
\( \widetilde{TSE} \) | 0.559 | 0.879 | 1.022 | 1.216 | 0.927
\( \log(m(X'_j)|X_j) \) | 475.75 | 502.92 | 524.02 | 480.88 | 361.88
Parametric Bayes Model
\( SSPE \) | 0.989 | 1.734 | 2.202 | 2.781 | 2.911
\( \widetilde{TSE} \) | 0.553 | 0.910 | 1.050 | 1.265 | 0.923
\( \log(m(X'_j)|X_j) \) | 481.07 | 497.47 | 520.18 | 478.78 | 363.29

Table 3.1: Predictive Performance for the Remaining Season \((X'_j)\) Based on the Early Season Data \((X_j)\). The mass parameter, \( M \), was recalibrated for each month.
models. For the AL, the parametric Bayes model performs slightly better than the ylimdir model. For the NL and MLB the ylimdir model performs the best overall. The difference is greatest for the earlier splits where pooling information across players is most important. A reversal occurs in the August split, where many pitchers drop out of the data set. Overall, the mixture of Dirichlet processes model’s performance is between that of the ylimdir model and the parametric Bayes model.

The second predictive exercise examines prediction for a new individual. For this exercise, we included only batters satisfying the screening rule for the first exercise as of June. We hold out one player at a time and predict the performance of the player based on the rest of the players. Appendix B contains the calculations. The results are shown in Table 3.2. Overall, the ylimdir model performs the best among the three models although the results from the mixture of Dirichlet processes models are close to those from the ylimdir model. The ylimdir model outperforms the parametric Bayes model for the NL and MLB since it predicts well for pitchers who have fewer at-bats and poor batting performance. Recall that the prediction under the ylimdir model is conditional, requiring the new player to join an existing cluster.

3.5 Example 2: The new-sales insurance data

In this section, we illustrate our techniques with a new-sales data set from an insurance company. The data consist of monthly counts of new policies and number of households by designated marketing area (DMA) in 2006 and 2007. A DMA is a geographic region which is covered by the same broadcast television and radio stations and the same local newspapers. A DMA typically consists of a large city and the surrounding area, and individuals within the DMA are exposed to the same
<table>
<thead>
<tr>
<th>Limiting Dirichlet Process Model</th>
<th>League</th>
<th>NL</th>
<th>AL</th>
<th>MLB</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M )</td>
<td>0.935</td>
<td>1.318</td>
<td>1.054</td>
<td></td>
</tr>
<tr>
<td>( SSPE )</td>
<td>1.482</td>
<td>0.227</td>
<td>1.834</td>
<td></td>
</tr>
<tr>
<td>( \hat{TSE} )</td>
<td>2.017</td>
<td>0.079</td>
<td>1.296</td>
<td></td>
</tr>
<tr>
<td>( log(m(X_j'</td>
<td>X_j)) )</td>
<td>275.67</td>
<td>297.29</td>
<td>564.85</td>
</tr>
<tr>
<td>Mixture of Dirichlet Processes Model</td>
<td>League</td>
<td>NL</td>
<td>AL</td>
<td>MLB</td>
</tr>
<tr>
<td>( M )</td>
<td>1.684</td>
<td>2.500</td>
<td>1.942</td>
<td></td>
</tr>
<tr>
<td>( SSPE )</td>
<td>1.486</td>
<td>0.228</td>
<td>1.838</td>
<td></td>
</tr>
<tr>
<td>( \hat{TSE} )</td>
<td>1.095</td>
<td>0.080</td>
<td>1.300</td>
<td></td>
</tr>
<tr>
<td>( log(m(X_j'</td>
<td>X_j)) )</td>
<td>273.25</td>
<td>294.97</td>
<td>562.78</td>
</tr>
<tr>
<td>Parametric Bayes Model</td>
<td>League</td>
<td>NL</td>
<td>AL</td>
<td>MLB</td>
</tr>
<tr>
<td>( SSPE )</td>
<td>1.542</td>
<td>0.227</td>
<td>1.918</td>
<td></td>
</tr>
<tr>
<td>( \hat{TSE} )</td>
<td>1.151</td>
<td>0.079</td>
<td>1.380</td>
<td></td>
</tr>
<tr>
<td>( log(m(X_j'</td>
<td>X_j)) )</td>
<td>237.64</td>
<td>298.88</td>
<td>522.50</td>
</tr>
</tbody>
</table>

Table 3.2: Predictive Performance of the Models when a Player is Held Out. Evaluation is specific to the National League (NL), the American League (AL), or Major League Baseball (MLB). The mass parameter, \( M \), was recalibrated for each of the three evaluations. Performance is measured by \( SSPE \), \( \hat{TSE} \), and sum of log marginal likelihoods.

advertising campaigns. Companies tailor marketing and sales efforts to the individual DMA.

Index the DMA by \( i \). The data consist of the number of households in DMA \( i \), \( HH_i \), and the count of new sales in month \( j \), \( NS_{ji} \), \( i = 1, \ldots, 200, \ j = 1, \ldots, 24 \). We assume that \( NS_{ji} \sim Poisson(p_i HH_i) \). The parameter \( p_i \) describes the \( i \)th DMA’s true but unknown sales rate. Following McCullagh and Nelder (1999), we use the variance stabilization transformation, \( X_{ji} = \sqrt{NS_{ji}/HH_i} \), so that \( X_{ji} \) is approximately normal with mean \( \theta_i = \sqrt{p_i} \) and variance \( \sigma_{ji}^2 = 1/(4HH_i) \). To make the normal approximation accurate, our analysis included only DMAs having \( HH_i > 50000 \). For the aggregated analyses, the data is collapsed to quarters or years before transformation.
Insurance sales show seasonal variation and are affected by many factors, among them pricing, marketing activities, and economic conditions. Preliminary analysis of the data confirmed these effects. To account for them, we limit comparisons to the same time period in successive years, we adjusted new sales in 2007 by multiplying the ratio of the means of the two years, and we adjusted for overdispersion. We use the prime symbol (′) to denote the adjusted variables for 2007, so that $X'_{ji}$ represents the transformed, adjusted sales rate for 2007. To account for overdispersion, we estimated an overdispersion factor for each month of $\hat{\rho}_j = \frac{\sum_{i=1}^{n_j} (X_{ji} - X'_{ji})^2}{\sum_{i=1}^{n_j} (\sigma_{ji}^2 + \sigma'_{ji}^2)}$, and inflated $\sigma_{ji}^2$ and $\sigma'_{ji}^2$ by $\hat{\rho}_j$. We denote the inflated variances by $\sigma_{ji}^2$ and $\sigma'_{ji}^2$ for the two years, respectively.

We illustrate the limdir model and compare it to a mixture of Dirichlet processes model and a parametric hierarchical Bayes model through two predictive exercises. The first is prediction of each month of 2007 based upon the corresponding month of 2006. The second is prediction of a DMA’s new sales for each month of 2006 based on other DMAs’ new sales. To study how the amount of DMA-specific information impacts prediction, we merge the monthly data sets into quarterly and full year data sets, and conduct the same predictive exercises.

The distribution of the $X_{ji}$ appears to be unimodal, with noticeable tail decay, and with right skewness. In our judgement, the normal distribution with mean $\mu$ and variance $\tau^2$ is a reasonable choice for a base measure of the mixture of Dirichlet processes model. For the parametric Bayes model, we placed a normal prior, $N(\mu, \tau^2)$, on $\theta$ and an improper prior on $\mu$ and $\tau^2$ which was $\propto 1/\tau^2$.

The limdir model was implemented with a base measure proportional to Lebesgue measure. The mass parameter for the mixture of Dirichlet processes model was chosen.
on the basis of the pairwise error rate as described in Section 3.3. The pairwise type I error rate was taken to be 0.05 and the cutoff threshold was set at 0.5. The variance used in the calibration was the mean of the $\sigma^2_{ji}$. The resulting mass parameters ranged from 0.743 to 1.156. The local mass for the limdir model was calibrated in a similar fashion. This led to local mass parameters ranging from 12.095 to 24.677. Note that the local mass and the mass for the mixture of Dirichlet processes model are not directly comparable, as constants have been swept into the local mass to simplify calculations.

The estimators are compared on the basis of sum of squared prediction error and sum of log marginal likelihood. When predicting the sales rate of 2007, sum of squared prediction error is given by $SSPE = \sum_{i=1}^{n_j} (X'_{ji} - \hat{\theta}_i)^2$, where $X'_{ji}$ is the transformed, adjusted sales rate of the $i$th DMA in the $j$th month of 2007, and $\hat{\theta}_i$ is the posterior mean of $\theta_i$. The sum of log marginal likelihoods is $\sum_{i=1}^{n_j} \log(m(X'_{ji} | X_j))$. For a single DMA, this is the log marginal likelihood. Summing across DMAs provides an indication of the predictive ability of competing models on a log-likelihood scale. Computationally, it is far more stable than is the log joint marginal likelihood of all DMAs.

We evaluate $\hat{\theta}_i$ with a Gibbs sampling run for the limdir model and the mixture of Dirichlet processes models. 200000 iterates were used for estimation, after a burn-in period of 10000 iterates. For the parametric Bayes model, we used numerical integration over a fine grid for $\tau^2$ to evaluate $\hat{\theta}_i$.

Table 3.3 presents the results of the first predictive exercise, forecasting 2007 performance from 2006 results. The three models are essentially equivalent in terms of $SSPE$, with minor differences in trailing decimal places. In terms of log marginal
<table>
<thead>
<tr>
<th>Model</th>
<th>( M )</th>
<th>Yearly</th>
<th>Quarterly</th>
<th>Monthly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limiting Dirichlet Process Model</td>
<td></td>
<td>12.095</td>
<td>17.631</td>
<td>24.675</td>
</tr>
<tr>
<td>SSPE</td>
<td>0.0047</td>
<td>0.0022</td>
<td>0.0011</td>
<td></td>
</tr>
<tr>
<td>( \log(m(X'_j</td>
<td>X_j)) )</td>
<td>740.826</td>
<td>821.450</td>
<td>912.296</td>
</tr>
<tr>
<td>Mixture of Dirichlet Processes Model</td>
<td>1.156</td>
<td>0.875</td>
<td>0.743</td>
<td></td>
</tr>
<tr>
<td>SSPE</td>
<td>0.0047</td>
<td>0.0022</td>
<td>0.0011</td>
<td></td>
</tr>
<tr>
<td>( \log(m(X'_j</td>
<td>X_j)) )</td>
<td>718.728</td>
<td>821.283</td>
<td>908.773</td>
</tr>
<tr>
<td>Parametric Bayes Model</td>
<td>0.0047</td>
<td>0.0022</td>
<td>0.0011</td>
<td></td>
</tr>
<tr>
<td>( \log(m(X'_j</td>
<td>X_j)) )</td>
<td>744.724</td>
<td>819.290</td>
<td>909.154</td>
</tr>
</tbody>
</table>

Table 3.3: Predictive Performance for Sales Rates in 2007 (\( X'_j \)) Based on Sales Rates in 2006 (\( X_j \)). Performance is measured by SSPE and sum of log marginal likelihoods. The mass parameter, \( M \), was recalibrated for each quarter and each month. For the quarterly data and the monthly data, the mean across quarter (month) is presented.

likelihood, the limdir model is the best overall performer. Performance varies with the amount of data used to create the forecasts. When a full year’s data is used to forecast the next year’s performance, \( \sigma^2_{ji} \) is quite small, the likelihood for the DMA is quite strong, and so there is little benefit to pooling information across DMAs. When the likelihood is weaker, as is the case for the quarterly and monthly data, the benefit of pooling information is greater. This, coupled with the skewed distribution of the \( X_{ij} \) leads to superior performance for the nonparametric methods. We examined which DMAs contribute to the difference in evaluation. There is little difference between the three models for the DMAs with middling \( X_{ji} \). The main difference is in the treatment of DMAs with extreme values of \( X_{ji} \). In the mixture of Dirichlet processes model, these DMAs tend to fall in larger clusters than in the limdir model due to the much smaller local mass in the tails. This excessive tail clustering leads to poorer predictive performance.
The second predictive exercise examines prediction for a new DMA. We hold out one DMA at a time and predict the sales rate of the DMA based on the rest of the DMAs. The prediction under the limdir model is conditional, requiring the new DMA to join an existing cluster. Appendix B contains the calculations. The results are shown in Table 3.4. As in the first predictive exercise, there is very little difference in $SSPE$. This follows from the fact that we are not using any data to distinguish between DMAs. The point predictions and hence $SSPE$ are essentially the mean for the remaining DMAs under all three models. There are, however, dramatic differences in log marginal likelihood. The two nonparametric models have the flexibility to capture the non-normality of the distribution of the $X_{ji}$, and so have far greater log marginal likelihood than the parametric model. The mixture of Dirichlet processes model is the best performer for the yearly data, while the limdir model is the best performer when the likelihoods are weaker in the quarterly and monthly data.

<table>
<thead>
<tr>
<th></th>
<th>Yearly</th>
<th>Quarterly</th>
<th>Monthly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limiting Dirichlet</td>
<td>$M$</td>
<td>12.095</td>
<td>17.632</td>
</tr>
<tr>
<td>Process Model</td>
<td>$SSPE$</td>
<td>0.144</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>$\log(m(X'_j</td>
<td>X_{j}))$</td>
<td>2901.124</td>
</tr>
<tr>
<td>Mixture of Dirichlet</td>
<td>$M$</td>
<td>1.155</td>
<td>0.876</td>
</tr>
<tr>
<td>Processes Model</td>
<td>$SSPE$</td>
<td>0.144</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>$\log(m(X'_j</td>
<td>X_{j}))$</td>
<td>2965.649</td>
</tr>
<tr>
<td>Parametric Bayes</td>
<td>$SSPE$</td>
<td>0.144</td>
<td>0.037</td>
</tr>
<tr>
<td>Model</td>
<td>$\log(m(X'_j</td>
<td>X_{j}))$</td>
<td>2046.349</td>
</tr>
</tbody>
</table>

Table 3.4: Predictive Performance of the Models when a DMA is Held Out. Evaluation is specific to the full year, each quarter or each month of 2006. Performance is measured by $SSPE$ and sum of log marginal likelihoods. For the quarterly data and the monthly data, the mean across quarter (month) is presented.
3.6 Example 3: Sucrose Solution Data

In this section, we analyze the data presented in Myers and Well (1995, p. 89). Rats were given one of four concentrations of sucrose solution. The speed with which the rats traversed a runway was measured. Eight rats were assigned to each treatment in a completely randomized design. The data appear in Table 3.5 and Figure 3.2.

<table>
<thead>
<tr>
<th>Treatments</th>
<th>Response</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1</td>
<td>1.4 2.0 3.2 1.4 2.3 4.0 5.0 4.7</td>
<td>3.000</td>
<td>1.441</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>3.2 6.8 5.0 2.5 6.1 4.8 4.6 4.2</td>
<td>4.650</td>
<td>1.404</td>
</tr>
<tr>
<td>Treatment 3</td>
<td>6.2 3.1 3.2 4.0 4.5 6.4 4.4 4.1</td>
<td>4.488</td>
<td>1.229</td>
</tr>
<tr>
<td>Treatment 4</td>
<td>5.8 6.6 6.5 5.9 5.9 3.0 5.9 5.6</td>
<td>5.650</td>
<td>1.125</td>
</tr>
</tbody>
</table>

Table 3.5: Example 3. Sucrose Solution Data from Myers and Well. MSE from a one-way ANOVA is 1.706.

We index a treatment, and a rat within treatment, by $i$ and $j$, respectively. Let $X_{ij}$ be the measured speed of the $j^{th}$ rat for the $i^{th}$ treatment. We assume that the $X_{ij}$ independently follow a normal distribution with unknown mean $\theta_i$ and variance $\sigma^2$. By placing an improper prior distribution on $\sigma^2$, we apply Model 1 to this data. We compare the results to those from the mixture of Dirichlet processes (MDP) model. We choose a normal distribution with mean $\mu$ and variance $\tau^2$ for the base measure of the MDP model. The parameters $\mu$ and $\tau^2$ are estimated based on the data. We use the overall sample mean, $\bar{X}$, and the mean square treatment, $\sum_i \sum_j (X_{ij} - \bar{X}_i)^2/k$, for the estimates of $\mu$ and $\tau^2$, respectively, where $k$ represents the number of treatments.

The mass parameter for the limdir model was chosen on the basis of the pairwise error rate, as described in Section 3. The pairwise Type I error rate was taken to
Figure 3.2: Example 3. Sucrose Solution Data from Myers and Well. The blue rectangles are the sample mean for each treatment and the red dotted line is the overall mean. MSE from a one-way ANOVA is 1.706.

be 0.05 and the cutoff threshold was set at 0.5. The variance used in the calibration was the MSE. The resulting mass parameter was chosen to be 0.698. For the MDP model, we place a gamma prior distribution on $M$ under the MDP model. The two parameters, $a$ and $b$, for the gamma prior distribution were calibrated through the R function $DPelicit$, in $DPpackage$, with the prior expected number of clusters equal to 2 and the variance of the number of clusters equal to 1. Since the data set has only four treatments, computations can be done analytically under Model 1. We use a Gibbs sampling run for the MDP model to compute the posterior probabilities of interest. 200,000 iterates were used for estimation, after a burn-in period of 10,000 iterates. Convergence diagnostics suggest that a run of this length produces good estimates of posterior quantities.
To study the impact of an outlying treatment on the overall inference, we define the root mean square error (RMSE) to be the square root of mean square error, and move the eight observations on treatment 1 down to slower speeds by subtracting $1 \times \text{RMSE}$, $3 \times \text{RMSE}$, $5 \times \text{RMSE}$ or $10 \times \text{RMSE}$ from the actual values. As discussed in Section 2.1, MacEachern and Müller (2000) suggest using estimates of hyperparameters robust to aberrant observations to avoid a lack of robustness under MDP models in the case where data contain some outlying observations. To illustrate how differently the MDP model behaves with the methods of estimating $\mu$ and $\tau^2$, we estimate them based on the original data (MDP model 2) or each perturbed set of data (MDP model 1).

The estimates under the three models are compared on the basis of posterior probabilities of equality of $\theta_i$ and $\theta_{i'}$, and posterior estimates of the treatment means and the variance of an observation. Table 3.6 gives the pairwise posterior probabilities of equality, $P(\theta_i = \theta_j|x)$, for all pairs. Table 3.7 gives the estimates of the treatment means, $\theta_i$, for $i = 1, \ldots, 4$, and the estimate of $\sigma^2$. Under the limdir model, the posterior probabilities of the equalities, $\theta_1 = \theta_i$, for $i = 2, 3, 4$ tend to zero as the $X_{1j}$ are moved down gradually, and yet the posterior probabilities of the other equalities remain approximately constant. This implies that the limdir model successfully recognizes the difference between the treatment 1 and the other treatments, due to local mass preservation. The successful isolation of $X_1$ yields robust estimates of $\theta$ and $\sigma^2$.

Under the MDP model 1, the estimate of $\mu$ also gets pulled toward $X_1$ as $X_1$ moves. This implies that the base measure under the MDP model 1 tends to assign little mass around the other $\theta_i$, $i \neq 1$, and so the model puts them into the same
<table>
<thead>
<tr>
<th>Model</th>
<th>Equality</th>
<th>Orig.</th>
<th>1RMSE</th>
<th>3RMSE</th>
<th>5RMSE</th>
<th>10RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limit of Dirichlet Process Model</td>
<td>$\theta_1 = \theta_2$</td>
<td>0.038</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_3$</td>
<td>0.053</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_4$</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_3$</td>
<td>0.420</td>
<td>0.433</td>
<td>0.434</td>
<td>0.434</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_4$</td>
<td>0.206</td>
<td>0.212</td>
<td>0.212</td>
<td>0.212</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td>$\theta_3 = \theta_4$</td>
<td>0.170</td>
<td>0.178</td>
<td>0.178</td>
<td>0.178</td>
<td>0.178</td>
</tr>
<tr>
<td>Mixture of Dirichlet Processes Model 1</td>
<td>$\theta_1 = \theta_2$</td>
<td>0.070</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_3$</td>
<td>0.091</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_4$</td>
<td>0.014</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_3$</td>
<td>0.410</td>
<td>0.520</td>
<td>0.654</td>
<td>0.733</td>
<td>0.830</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_4$</td>
<td>0.271</td>
<td>0.343</td>
<td>0.463</td>
<td>0.554</td>
<td>0.687</td>
</tr>
<tr>
<td></td>
<td>$\theta_3 = \theta_4$</td>
<td>0.229</td>
<td>0.305</td>
<td>0.428</td>
<td>0.523</td>
<td>0.663</td>
</tr>
<tr>
<td>Mixture of Dirichlet Processes Model 2</td>
<td>$\theta_1 = \theta_2$</td>
<td>0.070</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.165</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_3$</td>
<td>0.091</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
<td>0.171</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_4$</td>
<td>0.014</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_3$</td>
<td>0.410</td>
<td>0.422</td>
<td>0.417</td>
<td>0.407</td>
<td>0.324</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_4$</td>
<td>0.271</td>
<td>0.280</td>
<td>0.293</td>
<td>0.319</td>
<td>0.326</td>
</tr>
<tr>
<td></td>
<td>$\theta_3 = \theta_4$</td>
<td>0.229</td>
<td>0.243</td>
<td>0.258</td>
<td>0.291</td>
<td>0.323</td>
</tr>
</tbody>
</table>

Table 3.6: Pairwise Posterior Probabilities of Equality of Any Two Treatment Means, $P(\theta_i = \theta_j|x)$ for the Sucrose solution data. The posterior probabilities are estimated under the limit of Dirichlet process model and the mixture of Dirichlet processes model. To study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by $1\times$RMSE, $3\times$RMSE, $5\times$RMSE or $10\times$RMSE.
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Orig.</th>
<th>1RMSE</th>
<th>3RMSE</th>
<th>5RMSE</th>
<th>10RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limit of Dirichlet</td>
<td>$\theta_1$</td>
<td>3.063</td>
<td>1.695</td>
<td>-0.918</td>
<td>-3.531</td>
<td>-10.061</td>
</tr>
<tr>
<td>Process Model</td>
<td>$\theta_2$</td>
<td>4.679</td>
<td>4.705</td>
<td>4.706</td>
<td>4.706</td>
<td>4.706</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>4.563</td>
<td>4.602</td>
<td>4.603</td>
<td>4.603</td>
<td>4.603</td>
</tr>
<tr>
<td></td>
<td>$\theta_4$</td>
<td>5.482</td>
<td>5.479</td>
<td>5.479</td>
<td>5.479</td>
<td>5.479</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>1.868</td>
<td>1.845</td>
<td>1.844</td>
<td>1.844</td>
<td>1.844</td>
</tr>
<tr>
<td>Mixture of Dirichlet</td>
<td>$\theta_1$</td>
<td>3.326</td>
<td>1.877</td>
<td>-0.806</td>
<td>-3.450</td>
<td>-10.013</td>
</tr>
<tr>
<td>Processes Model 1</td>
<td>$\theta_2$</td>
<td>4.658</td>
<td>4.719</td>
<td>4.760</td>
<td>4.791</td>
<td>4.833</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>4.551</td>
<td>4.632</td>
<td>4.693</td>
<td>4.738</td>
<td>4.799</td>
</tr>
<tr>
<td></td>
<td>$\theta_4$</td>
<td>5.282</td>
<td>5.304</td>
<td>5.264</td>
<td>5.215</td>
<td>5.134</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>1.892</td>
<td>1.861</td>
<td>1.872</td>
<td>1.882</td>
<td>1.896</td>
</tr>
<tr>
<td>Mixture of Dirichlet</td>
<td>$\theta_1$</td>
<td>3.326</td>
<td>2.149</td>
<td>0.041</td>
<td>-1.767</td>
<td>1.696</td>
</tr>
<tr>
<td>Processes Model 2</td>
<td>$\theta_2$</td>
<td>4.658</td>
<td>4.702</td>
<td>4.706</td>
<td>4.706</td>
<td>4.207</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>4.551</td>
<td>4.610</td>
<td>4.617</td>
<td>4.624</td>
<td>4.174</td>
</tr>
<tr>
<td></td>
<td>$\theta_4$</td>
<td>5.282</td>
<td>5.282</td>
<td>5.262</td>
<td>5.208</td>
<td>4.414</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>1.892</td>
<td>1.897</td>
<td>2.118</td>
<td>2.816</td>
<td>40.510</td>
</tr>
</tbody>
</table>

Table 3.7: Posterior Estimates of the Treatment Means, $\theta_i$ for $i = 1, \ldots, 4$ and the Common Variance for the Observations, $\sigma^2$ for the Sucrose solution data. The parameters are estimated under the limit of Dirichlet process model and the mixture of Dirichlet processes model. To study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by $1 \times$ RMSE, $3 \times$ RMSE, $5 \times$ RMSE or $10 \times$ RMSE.
clusters with larger probabilities. Tables 3.6 and 3.7 show that the posterior prob-
abilities of any equality, $\theta_i = \theta_{i'}, i \neq 1, i' \neq 1, i \neq i'$, are increasing, and this leads
to poor estimates of $\theta$ and $\sigma^2$. On the other hand, the estimates of $\mu$ of the MDP
model 2 are fixed at the estimate based on the original data, so that the likelihood
of $X_1$ tends to move to the region where the base measure under the MDP model 2
assigns less mass, as $X_1$ moves down. Tables 3.6 and 3.7 show that with moderately
outlying $X_1$, the posterior estimates remains relatively robust although the estimate
of $\theta_1$ is heavily affected by $\mu$. However, it is apparent that when $X_1$ is moved down
by $10 \times$ RMSE, the MDP model 2 also provides unreasonable inference.

3.7 Conclusions

The limdir model developed in this chapter extends the popular mixture of Dirich-
let processes model. It allows us to use an improper base measure for the Dirichlet
process, thus creating a sort of noninformative, nonparametric Bayesian analysis.
The noninformative analysis provides insight into how to use the proper mixture of
Dirichlet processes model, and it suggests several interesting extensions.

First, the development of the limdir models presented in this paper has been
tied to normal locations. The strategy is to first identify a noninformative analysis
for a single component problem, to then create a sequence of base measures that
preserve local mass and that lead to this analysis for the single component, and to
then explicitly describe the limiting local mass. This strategy applies quite generally
as indicated by the results in Section 3.2.2. This strategy allows us to create Dirichlet
based Jeffreys’ analyses, to adapt Bernardo (1979) reference priors, etc.
A traditional advantage of noninformative analyses is that they simplify calculation of the posterior distribution. This advantage is less important in the modern environment of simulation based fits, but it is still relevant when it comes to calibration of a prior distribution and also for problems involving a large number of treatments. The advantages carry over from the parametric setting to the limdir setting, enabling us to implement very efficient computational algorithms to fit the models. The simplicity of the calculations should also prove useful in developing extensions of fractional Bayes factors (O’Hagan 1995) and intrinsic Bayes factors (Berger and Pericchi 1996) to a nonparametric Bayesian setting. Such extension would allow routine use of the models for automated Bayesian analysis. The limdir models provide a springboard from which these techniques can be developed.

The development of the limdir models carries a strong message for mixture of Dirichlet processes modelling. The main message is the importance of local mass. In selecting the mass parameter of the Dirichlet process, the relevant quantity is local mass — not the entire mass of the base measure. We recommend replacing the standard independent prior distributions on \( G_0 \) and \( M \) (Escobar and West 1995), with a distribution on \( G_0 \) and a distribution on \( M|G_0 \). The latter distribution would be chosen to keep local mass toward the center of the distribution approximately constant.

In practice, many users of mixture of Dirichlet process models deliberately choose an overdispersed base measure in the mistaken belief that this reduces the impact of the prior distribution on the analysis. An overdispersed analysis which ignores the concept of local mass results in an artificially small number of clusters.
For analyses with an overdispersed base measure, calibration via local mass is essential. However, the posterior distribution of $\theta_{k+1}$ remains unrealistic. The entire mass of an overdispersed base measure will be large, leading to too much weight given to $\theta_{k+1}$ initiating a new cluster. For limdir models with an improper marginal distribution, we have shown that use of the conditional distribution of $\theta_{k+1}$, given that it joins an already established cluster, performs well. This fix can be used with proper, overdispersed distributions as well. In addition, a second fix is available: one can average this conditional distribution with a downweighted and less dispersed substitute for the base measure.

We believe that the limdir models will find a useful place in the arsenal of nonparametric Bayesian techniques. We also believe that consideration of the models will improve the practice of nonparametric Bayesian analysis, and that the models open a valuable collection of research problems.
CHAPTER 4

LOCAL-MASS PRESERVING PRIOR DISTRIBUTIONS
FOR NONPARAMETRIC BAYESIAN MODELS

4.1 Introduction

In this chapter, we address the problem of prior specification for models involving
the Dirichlet process. The flatness of the prior distribution’s tail relative to the tail of
the likelihood is known to have an impact on the robustness of the resulting inference.
In the case where the likelihoods are concentrated in the tails of a sharp-tailed prior
distribution, the prior distribution heavily influences the posterior inference due to
negligible mass in the tails. However, if any prior distribution preserves reasonable
amount of mass in the tails, the prior distribution becomes less influential on the
likelihood which are off from the center of the prior distribution and so the resulting
analysis is robust with respect to misspecification of a prior distribution. Due to this
argument, flat-tailed prior distributions are often considered to be robust (Berger
1993).

In parametric Bayesian modeling, intentional inflation of the dispersion of a prior
distribution is commonly used to represent vague prior information, and it often
produces reasonably robust inference. Similarly, in many applications of Bayesian
nonparametric models involving a Dirichlet process, an overdispersed prior is deliberately chosen in an attempt to achieve robustness. However, as discussed in Chapter 3, an overdispersed prior distribution in a mixture of Dirichlet processes model does not represent vague prior beliefs, nor does it result in robust inference. To avoid this misstep, we introduce the concept of local mass, defined in Chapter 3, to this prior specification problem, and we propose a new structure for the prior distribution that focuses on local mass. This alternative strategy of constructing prior distributions preserves local mass in any region of the parameter space by tying the total mass of the base measure to its dispersion.

Section 4.2 describes the new class of local-mass preserving prior distributions and presents technical details with an application to a Dirichlet process model. Section 4.3 discusses calibration of the prior distribution. Section 4.4 presents a data analysis with an one-way analysis of variance example, and makes comparisons to similar models with conventional prior distributions. The final section contains conclusions.

4.2 Models

We begin the development of the local-mass preserving mixture of Dirichlet process (LMDP) model with a presentation of the conventional structure of prior distributions in a mixture of Dirichlet processes (MDP) model. The MDP model applies the Dirichlet process prior in Bayesian hierarchical modeling by placing a Dirichlet process prior on a latent mixing distribution. To be precise, we present a hierarchical MDP model as follows:
\[
X_i | \theta_i \overset{\text{ind}}{\sim} F(\cdot | \theta_i) \quad i = 1, \ldots, n,
\]
\[
\theta_i | G \overset{\text{i.i.d.}}{\sim} G(\cdot | \nu),
\]
\[
G | M, \nu \sim DP(\alpha) \text{ where } \alpha = MG_0(\cdot | \nu),
\]
\[
\nu \sim H_\nu(\cdot),
\]
\[
M \sim H_M(\cdot),
\]

where \( \alpha \) is the parameter of the Dirichlet process. \( \alpha \) is a measure which can be decomposed into two parts, the total mass parameter, \( M \) and the base distribution, \( G_0 \). \( G_0 \) determines the features of the prior distribution such as shape, location and dispersion, and \( G_0 \) may be characterized with some other hyperparameters, \( \nu \). \( M \) is the total mass of the measure. The model described above extends the basic MDP model by extending the hierarchy to place a prior distribution on \( M \) and \( \nu \). The above model can also be viewed as an analogue of a Bayesian, parametric hierarchical model (e.g., MacEachern 1998).

The Dirichlet process prior probabilistically partitions \( \theta \) into \( p \leq n \) clusters. Following the definitions and notation in Chapter 3, we let the \( n \)-dimensional vector, \( s \), denote a partition. We also let the \( p \)-dimensional vectors, \( c \) and \( \gamma \), denote the sizes of the \( p \) clusters and their locations, respectively. The Dirichlet process assigns probability to a partition as follows:

\[
\pi(s) = M^p \prod_{i=1}^{p} \Gamma(c_i) / \prod_{i=1}^{n} (M + i - 1).
\]

Given \( s \), the \( p \) \( \gamma_i \)'s are a random sample from \( G_0 \). We describe \( s \) using the “no gaps” scheme (MacEachern 1998).
4.2.1 Independence structure of prior distributions

One of the concerns in hierarchical modeling is the sensitivity of inferences produced by assumed forms of component distributions (Escobar and West 1998). Since the choice of prior distribution can have an influence on the inferences, the prior distribution should ideally be elicited carefully. Ironically, although one may place hierarchical prior distributions on $M$ and $\nu$ due to the great appropriate values for uncertainty about them, specification of the prior distribution requires an extraordinary amount of prior information about them. One common way to reflect limitations of prior information is to use a flat-tailed prior distribution. This is because flat-tailed prior distributions often produce reasonably robust posterior inference with respect to aberrant observations. One example illustrating this idea is the use of a Cauchy distribution as a prior distribution for a normal mean, instead of a normal distribution. The resulting model does yield robust inference, but at the cost of losing the natural conjugacy, and sacrificing easy computation. An alternative way to produce relatively flat tails in the prior distribution while maintaining conjugacy is to inflate the dispersion of the normal prior distribution by purposely increasing its variance. Following this idea, one may consider inflating the dispersion of $G_0$ through $\nu$ to reflect vague prior information.

With regard to placing a prior distribution on $M$, the traditional way is to place a prior distribution independent of $\nu$ (Escobar and West 1995). This independence prior structure leads to conditional posterior independence of $M$ and $\nu$. Escobar (1994) applied the hierarchical MDP model to the normal mean estimation problem and used the relationship between the mass parameter and the expected number of clusters to select a prior distribution for $M$. He used a discrete distribution for $M$. 
As discussed in Section 2.2, Escobar (1994) considered a normal distribution for $G_0$ with unknown mean and variance because of conjugacy to the normal likelihood, and he estimated the unknown hyperparameters from data.

Escobar and West (1995) developed an extension of the model developed in Escobar (1994) for a more general setting. Similar to Escobar (1994), they used the expected number of clusters to elicit a prior distribution on $M$ and put a gamma prior on $M$ to simplify computation. They introduced a latent variable, $\eta$, following a beta distribution, and sampled $M$ conditional on the most current number of clusters and the latent variable in a Gibbs sampler. Because of unknown $\sigma_i^2$, they considered the normal/inverse-gamma distribution for $G_0$ which is a choice of convenience, to facilitate computation.

As a specific illustration, we consider the normal mean estimation problem with known variance, similar to the problem in Escobar (1994). We assume that $G_0$ is a normal distribution with unknown mean $\mu$ and variance $\tau^2$, because of conjugacy to the normal likelihoods. Again, we place a normal distribution with known mean $\mu_0$ and variance $\tau_0^2$ on $\mu$, and we place an inverse-gamma distribution with shape parameter $a_0$ and scale parameter $b_0$ (hence, mean $b_0/(a_0 - 1)$) on $\tau^2$. Following Escobar and West (1995), we place a gamma prior distribution with shape parameter $a$ and scale parameter $b$ (hence, mean $ab$) on $M$. The complete prior specification for
the model can be written as

\[ X_i | \theta_i, \sigma_i^2 \overset{ind}{\sim} N(\cdot | \theta_i, \sigma_i^2) \quad i = 1, \ldots, n, \]

\[ G | M, \mu, \tau^2 \sim DP(MG_0) \text{ where } G_0 \text{ is } N(\mu, \tau^2), \]

\[ \mu \sim N(\mu_0, \tau_0^2), \text{ and } \tau^2 \sim IG(a_0, b_0), \]

\[ M \sim Gamma(a, b). \]

In this conventional hierarchical MDP model, \( M \) is independent of the dispersion parameter of \( G_0, \tau^2 \). Due to this independence structure in the prior distribution, intentional inflation of the base distribution’s dispersion may result in unreasonable inference: for any fixed \( M \), \( \alpha \) assigns smaller mass to the central portion of the parameter space under a more diffuse prior distribution. The resulting measure does have more mass in the tails, and so the resulting posterior may not be heavily affected by aberrant observations. However, due to the fact that a Dirichlet process prior puts any two components into a cluster with larger probability when the two components are in a region to which \( \alpha \) assigns small mass, we are left with poor, unintended inference for parameters in the middle region. To avoid this problem, we should consider the concept of local mass when deciding on the structure of the prior distribution.

### 4.2.2 Local-mass preserving structure of prior distributions

Our goal is to develop a new structure for prior distribution for the hierarchical MDP model, such that the resulting model shows more robustness than does the conventional MDP model. Local mass is defined in Chapter 4 as mass assigned by a measure \( \alpha \) to a small measurable set in any region of the parameter space. Our strategy is to construct local-mass preserving prior distributions through tying the mass of the base measure to its dispersion. As a result, a class of our prior distributions
is more appropriate for modeling the tail behavior without losing reasonable inference in the non-tail regions.

To preserve local mass, we consider the following prior specification for $M$:
\[
M | \tau \sim \text{Gamma}(a', b' \tau).
\]

$M$ and $\tau$ are positively associated in the prior structure. Mass in the middle region under this independence structure decreases as the dispersion increases for a fixed value of the total mass. However, due to dependence of $M$ on $\tau$, the total mass increases, and consequently, local mass remains approximately constant as the dispersion varies.

Figure 4.1 provides intuition behind preservation of local mass. Panel $a$ shows a sequence of normal measures with the same mean 0, but increasing variances for a fixed value of $M$. The sequence demonstrates that the sharpest measure assigns negligible mass in the tails, whereas it assigns a great amount of mass in the middle region. As the dispersion increases, the measure assigns more mass in the tails, but less mass in the middle. On the other hand, panel $b$ illustrates a sequence of normal measures which assigns approximately constant mass in the middle region of the sharpest measure. The middle region where the likelihoods are concentrated gets the same amount of mass under this sequence despite the increasing overdispersion, and the resulting inference on the components in this region is not affected much by the overdispersed prior distributions under this sequence. At the same time, along the sequence, the measure assigns more mass in the tail region, so that the overdispersion can represent vague prior beliefs under this structure.
Figure 4.1: Intuition behind Local Mass Preservation. Panel (a) shows three normal measures with identical total mass parameter $M$. As the dispersion increases, the measure assigns larger mass in the tail region, but it assigns smaller mass in the middle region. Panel (b) shows three normal measures by matching mass in the middle region of the sharpest measure. The measures assign approximately constant mass in the middle, but they assign more mass in the tail as the dispersion increases.
Mathematically, for any interval \([l_1, l_2]\), the base measure under the local-mass preserving structure assigns the following mass conditional on \(\mu\) and \(\tau^2\),

\[
\alpha([l_1, l_2]|\mu, \tau^2) = E(M|\tau) \int_{l_1}^{l_2} g_0(\theta|\mu, \tau^2) d\theta = a'b' \int_{l_1}^{l_2} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2\tau^2} (\theta - \mu)^2\} d\theta.
\]

On the other hand, under the conventional prior structure, the base measure assigns mass in a different way, as follows:

\[
\alpha([l_1, l_2]|\mu, \tau^2) = E(M) \int_{l_1}^{l_2} g_0(\theta|\mu, \tau^2) d\theta = ab \{\Phi(l_2 - \mu/\tau) - \Phi(l_1 - \mu/\tau)\},
\]

where \(\Phi\) is a cdf of the standard normal distribution.

We note that under the conventional prior structure, the mass assigned to \([l_1, l_2]\) converges to zero as \(\tau^2\) tends to infinity, and yet under the LMDP model, it converges to \(a'b'(l_2 - l_1)/\sqrt{2\pi}\). Under this new model, the dispersion of the base measure is independent of local mass in the limit.

### 4.3 Calibration of the prior distribution

The hyperparameters of the LMDP model can be assigned in many ways, ranging from direct specification to indirect specification. One possibility is to use a calibration based on a type I error rate as in Chapter 3. Alternatively, we might focus on other aspects of the prior distribution. Our calibration of the hyperparameters in the LMDP model can be conducted by first specifying the expected number of clusters and the variance of the number of clusters given \(n\) observations. From the Pólya urn representation of the Dirichlet process, we can induce the expected number of clusters and the variance of the number of clusters. Given \(n\) observations and the
mass parameter $M$, the expected number of clusters is

$$E(p_n|M) = E(p_{n-1}|M) + \frac{M}{M + n - 1}$$

$$= \sum_{i=1}^{n} \frac{M}{M + i - 1}. \quad (4.1)$$

Similarly, the variance of the number of clusters is

$$Var(p_n|M) = E\{Var(p_n|p_{n-1}, M)\} + Var\{E(p_n|p_{n-1}, M)\}$$

$$= \frac{M}{M + n - 1} \frac{n - 1}{M + n - 1} + Var(p_{n-1} + \frac{M}{M + n - 1}|M)$$

$$= \frac{M(n - 1)}{(M + n - 1)^2} + Var(p_{n-1}|M)$$

$$= \sum_{i=1}^{n} \frac{M(i - 1)}{(M + i - 1)^2}. \quad (4.2)$$

The calibration procedure searches for values of $a$ and $b$ such that they generate the specified values for the expected number of clusters and the variance of the number of clusters. The values of $a$ and $b$ can be found through the Newton-Raphson method. Due to a difficulty in differentiation of Equations 4.1 and 4.2, the Jacobian in the Newton-Raphson method is estimated via the forward difference approximation.

Under the LMDP model, $M$ is not independent of $\tau$, so Equations 4.1 and 4.2 should be marginalized with respect to $\tau$ as well. The marginalization requires conditional evaluations as follows,

$$E(p_n|M) = E\{E(p_n|M, \tau)\}, \quad (4.3)$$

$$Var(p_n|M) = E\{Var(p_n|M, \tau)\} + E\{Var(p_n|M, \tau)\}. \quad (4.4)$$

Since $p_n$ does not depend on $\tau$, Equations 4.3 and 4.4 produce similar calibration as do Equations 4.1 and 4.2. The difference lies in the form of the marginal prior distribution for $M$. 68
4.4 Example: Sucrose Solution Data

In this section, we revisit the data presented in Myers and Well (1995, p. 89) to illustrate the performance of the local-mass preserving mixture of Dirichlet process (LMDP) model and compare it to that of the conventional mixture of Dirichlet process (MDP) model. As we described in Section 3.6, in an experiment with a completely randomized design, a researcher treated eight rats with one of four concentrations of sucrose solution, and he measured the speeds with which the rats traverse a runaway. The data are shown in Table 3.5 and Figure 3.2.

We follow the notation and definitions in Section 3.6. We let $X_{ij}$ denote the measured speed of the $j^{th}$ rat for the $i^{th}$ treatment for $i = 1, \ldots, 4$ and $j = 1, \ldots, 8$. We assume that the $X_{ij}$ independently follow a normal distribution with unknown mean $\theta_i$ and variance $\sigma^2$. We place an improper prior distribution on $\sigma^2$, i.e., $\pi(\sigma^2) \propto 1/\sigma^2$, and we apply MDP models to model the $\theta_i$s. We choose a normal distribution with mean $\mu$ and variance $\tau^2$ for the base measure of the MDP models. We estimate $\mu$ with the overall sample mean, $\hat{\mu} = \bar{X}$, and we place an inverse gamma prior distribution, $IG(a_0, b_0)$ on $\tau^2$. We set $a_0$ at 6 to ensure that it has a reasonable shape. We set $b_0 = \hat{\tau}^2(a_0 - 1)$ where $\hat{\tau}^2$ is an estimate of $\tau^2$ under a parametric Bayesian model with a normal prior distribution, so that $E(\tau^2) = \hat{\tau}^2$.

We also consider a prior distribution for the total mass parameter of the Dirichlet process prior, $M$. The LMDP ties the mass of the base measure to its dispersion to preserve local mass. We put a gamma prior distribution, $Gamma(a', b')$ on $M$ for the LMDP. Here, the two parameters, $a'$ and $b'$, are determined through the method described in Section 4.3, with the prior expected number of clusters equal to 2 and the variance of the number of clusters equal to 1. Similarly, we consider a gamma prior
distribution, $Gamma(a,b)$, for the conventional MDP model. The two parameters, $a$ and $b$, for the conventional MDP model were calibrated through the R function $DPelicit$, in $DPpackage$, with the same prior expected number of clusters and the same variance of the number of clusters.

Similar to the study in Section 3.6, we subtract $1\times RMSE$, $3\times RMSE$, $5\times RMSE$ or $10\times RMSE$ from the actual values on the treatment 1, and we move them down to slower speeds. The $\hat{\mu}$ and $\hat{\tau}^2$ are fixed on the estimates based on the original data to examine how the two models deal with an unexpected outlying treatment. In addition to this, we increase the parameter, $b_0$, of the prior distributions for $\tau^2$ by 25 times to make the normal base measure overdispersed.

Tables 4.1 and 4.2 show the results under the two models with well-calibrated parameters. Both of the models yield reasonable estimates of the probabilities of the equalities except for the case where $X_1$ is pulled down by $10\times RMSE$. Similarly, the estimates of $\theta_i$ and $\sigma^2$ are reasonable under both of the two models although the estimates of $\theta_1$ and $\theta_4$ under the LMDP are more robust. In the case where the $X_1$ is pulled down by $10\times RMSE$, the two models break down in the sense that none of them distinguishes $\theta_1$ from the other $\theta_i$s successfully. However, the LMDP produces more reasonable estimates, and from the pattern along the columns, the conventional MDP model breaks down more quickly than the LMDP model.

Tables 4.3 and 4.4 illustrate the results under the two models with an artificially inflated dispersion parameter. By comparing the first column of Tables 4.3 and 4.4 to that of Tables 4.1 and 4.2, the artificially overdispersed base measure moderately changes the inference under the LMDP model. However, the conventional MDP model puts any two $\theta_i$s into a cluster with larger probabilities, and so the inference under
the conventional MDP model is considerably changed with the overdispersed base measure. Even along the perturbation of the data, the LMDP model produces more robust estimates especially for $\theta_1$ and $\theta_4$ than the conventional MDP model does.

The conventional MDP model does not seem to be broken even in the case where $X_1$ is moved down by $10\times$RMSE, but the probabilities of the equalities, $\theta_i = \theta_{i'}$ for $i, i' = 2, 3, 4$, $i \neq i'$ are clearly overestimated due to the artificial overdispersion of the base measure.

![Table 4.1](image)

Table 4.1: Pairwise Posterior Probabilities of Equality of Any Two Treatment Means, $P(\theta_i = \theta_j|x)$ for the Sucrose solution data. The posterior probabilities are estimated under the local-mass preserving mixture of Dirichlet process model and the conventional mixture of Dirichlet processes model. To study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by $1\times$RMSE, $3\times$RMSE, $5\times$RMSE or $10\times$RMSE.
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Orig.</th>
<th>1RMSE</th>
<th>3RMSE</th>
<th>5RMSE</th>
<th>10RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local-Mass Preserving</td>
<td>$\theta_1$</td>
<td>3.356</td>
<td>2.111</td>
<td>0.085</td>
<td>-1.754</td>
<td>-1.365</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>4.670</td>
<td>4.728</td>
<td>4.732</td>
<td>4.720</td>
<td>4.347</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>4.571</td>
<td>4.649</td>
<td>4.657</td>
<td>4.650</td>
<td>4.300</td>
</tr>
<tr>
<td>Mixture of Dirichlet</td>
<td>$\theta_4$</td>
<td>5.228</td>
<td>5.246</td>
<td>5.202</td>
<td>5.166</td>
<td>4.617</td>
</tr>
<tr>
<td>Process Model</td>
<td>$\sigma^2$</td>
<td>1.926</td>
<td>1.909</td>
<td>2.219</td>
<td>3.210</td>
<td>28.450</td>
</tr>
<tr>
<td>Conventional</td>
<td>$\theta_1$</td>
<td>3.408</td>
<td>2.237</td>
<td>0.272</td>
<td>-1.086</td>
<td>1.820</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>4.664</td>
<td>4.719</td>
<td>4.727</td>
<td>4.717</td>
<td>4.095</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>4.566</td>
<td>4.644</td>
<td>4.657</td>
<td>4.654</td>
<td>4.059</td>
</tr>
<tr>
<td></td>
<td>$\theta_4$</td>
<td>5.200</td>
<td>5.207</td>
<td>5.170</td>
<td>5.085</td>
<td>4.266</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>1.932</td>
<td>1.942</td>
<td>2.320</td>
<td>3.878</td>
<td>41.900</td>
</tr>
</tbody>
</table>

Table 4.2: Posterior Estimates of the Treatment Means, $\theta_i$, for $i = 1, \ldots, 4$, and the Common Variance for the Observations, $\sigma^2$, for the Sucrose solution data. The posterior probabilities are estimated under the local-mass preserving mixture of Dirichlet process model and the conventional mixture of Dirichlet processes model. To study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by 1×RMSE, 3×RMSE, 5×RMSE or 10×RMSE.

4.5 Conclusions

The LMDP model developed in this chapter robustifies the conventional hierarchical MDP models. Through this modeling strategy, we can preserve local mass in regions of interest even under the overdispersed base measure, which leads to robust inference. This local-mass preserving modeling strategy also provides insights into how to robustfy other nonparametric Bayesian models, and it indicates interesting extensions.

In this chapter, we presented the development of the LMDP model for the normal mean estimation problem. The essence in the development can be easily extended to other models. The strategy is to first identify parameters related to the dispersion of the base measure, and then to relate the identified parameters to construction of
<table>
<thead>
<tr>
<th>Model</th>
<th>Equality</th>
<th>Orig.</th>
<th>1RMSE</th>
<th>3RMSE</th>
<th>5RMSE</th>
<th>10RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local-Mass Preserving</td>
<td>$\theta_1 = \theta_2$</td>
<td>0.098</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_3$</td>
<td>0.121</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_4$</td>
<td>0.029</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_3$</td>
<td>0.621</td>
<td>0.638</td>
<td>0.640</td>
<td>0.640</td>
<td>0.638</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_4$</td>
<td>0.415</td>
<td>0.428</td>
<td>0.431</td>
<td>0.432</td>
<td>0.430</td>
</tr>
<tr>
<td></td>
<td>$\theta_3 = \theta_4$</td>
<td>0.377</td>
<td>0.394</td>
<td>0.395</td>
<td>0.398</td>
<td>0.397</td>
</tr>
<tr>
<td>Mixture of Dirichlet Process Model</td>
<td>$\theta_1 = \theta_2$</td>
<td>0.194</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_3$</td>
<td>0.218</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 = \theta_4$</td>
<td>0.097</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_3$</td>
<td>0.777</td>
<td>0.811</td>
<td>0.811</td>
<td>0.811</td>
<td>0.811</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 = \theta_4$</td>
<td>0.624</td>
<td>0.656</td>
<td>0.656</td>
<td>0.657</td>
<td>0.655</td>
</tr>
<tr>
<td></td>
<td>$\theta_3 = \theta_4$</td>
<td>0.586</td>
<td>0.630</td>
<td>0.631</td>
<td>0.631</td>
<td>0.630</td>
</tr>
</tbody>
</table>

Table 4.3: Pairwise Posterior Probabilities of Equality of Any Two Treatment Means, $P(\theta_i = \theta_j|x)$ for the Sucrose solution data. The posterior probabilities are estimated under the local-mass preserving mixture of Dirichlet process model and the conventional mixture of Dirichlet processes model. We increase the scale parameter of the inverse-gamma prior distribution for $\tau^2$ by 25 times to inflate the dispersion of the base measure of the Dirichlet process. In addition, to study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by 1×RMSE, 3×RMSE, 5×RMSE or 10×RMSE.
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Orig.</th>
<th>1RMSE</th>
<th>3RMSE</th>
<th>5RMSE</th>
<th>10RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local-Mass Preserving</td>
<td>θ₁</td>
<td>3.157</td>
<td>1.718</td>
<td>-0.875</td>
<td>-3.464</td>
<td>-9.939</td>
</tr>
<tr>
<td></td>
<td>θ₂</td>
<td>4.706</td>
<td>4.764</td>
<td>4.766</td>
<td>4.765</td>
<td>4.765</td>
</tr>
<tr>
<td></td>
<td>θ₄</td>
<td>5.312</td>
<td>5.317</td>
<td>5.316</td>
<td>5.316</td>
<td>5.315</td>
</tr>
<tr>
<td></td>
<td>σ²</td>
<td>1.925</td>
<td>1.866</td>
<td>1.864</td>
<td>1.868</td>
<td>1.872</td>
</tr>
<tr>
<td>Mixture of Dirichlet</td>
<td>θ₁</td>
<td>3.290</td>
<td>1.722</td>
<td>-0.866</td>
<td>-3.452</td>
<td>-9.916</td>
</tr>
<tr>
<td></td>
<td>θ₄</td>
<td>5.133</td>
<td>5.159</td>
<td>5.159</td>
<td>5.158</td>
<td>5.159</td>
</tr>
<tr>
<td></td>
<td>σ²</td>
<td>2.008</td>
<td>1.891</td>
<td>1.891</td>
<td>1.892</td>
<td>1.897</td>
</tr>
</tbody>
</table>

Table 4.4: Posterior Estimates of the Treatment Means, θᵢ, for i = 1, ..., 4, and the Common Variance for the Observations, σ², for the Sucrose solution data. The parameters are estimated under the local-mass preserving mixture of Dirichlet process model and the conventional mixture of Dirichlet processes model. We increase the scale parameter of the inverse-gamma prior distribution for τ² by 25 times to inflate the dispersion of the base measure of the Dirichlet process. In addition, to study the impact of having an outlying treatment, we move eight observations on treatment 1 down by subtracting by 1×RMSE, 3×RMSE, 5×RMSE or 10×RMSE.
a prior distribution for the total mass parameter such that local mass remains approximately constant. As a specific example, ANOVA based on a binomial likelihood for which a conjugate base measure is a beta distribution, \( \text{Beta}(\alpha, \beta) \), can be treated in a similar fashion. To preserve local mass, we relate \( \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta) \) to the specification of a prior distribution for \( M \) leading to mass \( \int_{l_1}^{l_2} \theta^{(\alpha-1)}(1 - \theta)^{(\beta-1)} \, d\theta \) for any interval \([l_1, l_2], l_2 > l_1\).

Furthermore, due to the complexity of high dimensional problems, the concept of local mass carries a strong message to inferences in the application to those problems. The problem of specifying a reasonable joint prior distribution over many parameters is often very difficult, and the requisite computations are often very complicated. One may find a prior structure resulting in relatively easy computations, yet yielding reasonably robust inferences. The strategy developed here allows us to keep the conjugacy structure, leading to simplification of the calculation of the posterior distribution. Moreover, by focusing on local behavior of parameters of interest, the resulting posterior inference is less sensitive to prior misspecification. Together, these two features make our strategy an attractive tool in nonparametric Bayes.
CHAPTER 5

CONSISTENCY OF BAYES ESTIMATORS WITHOUT THE ASSUMPTION THAT THE MODEL IS CORRECT

5.1 Introduction

In this chapter, we investigate issues of consistency and inconsistency of Bayes estimators. The question of consistency arises when one examines the asymptotic behavior of an estimator. One envisions an experiment that grows ever larger, or a sequence of experiments of increasing size, and asks whether the sequence of estimators converge to the true value of a parameter. Convergence may be in one of several forms.

The asymptotic behavior of the maximum likelihood estimator (MLE), which is the most commonly used frequentist estimator, has been intensively investigated. It is known that, under suitable regularity conditions on the likelihood, the MLE is consistent, and that it is also asymptotically efficient (it is asymptotically unbiased, with minimum asymptotic variance). On the other hand, the posterior distribution in a Bayesian analysis is determined by the prior distribution and likelihood together. Thus, choosing an appropriate probability model for both prior and likelihood is essential to produce reasonable asymptotic behavior of the posterior distribution.
The importance of both prior and likelihood makes consideration of consistency more difficult.

There are many interesting topics in asymptotics of Bayes estimators, as reviewed in Section 2.3. One significant result in Bayesian asymptotics is that, under relatively mild regularity conditions, the impact of the prior distribution on the posterior distribution washes out as the sample size increases. More importantly, the Bayes estimator becomes asymptotically concentrated near the true parameter value for any point in the parameter space if the prior distribution assigns positive mass to its neighborhood. This implies that, as the sample size tends to infinity, the data determine the asymptotic behavior of the posterior distribution, regardless of the form of the prior distribution used for analysis, as long as the prior distribution puts some positive mass around the true parameter value. The hedge to this broad statement of consistency lies in the regularity conditions and in a careful definition of the topology that defines a neighborhood.

The stream of research that developed out of Berk (1966) is particularly relevant to our work on the topic of Bayesian asymptotics. This research studies the asymptotic behavior of Bayes estimators and posterior distributions under a possibly incorrect model. Berk (1966) studied the asymptotic behavior of posterior distributions without the assumption that the model is correct. He showed that under certain conditions on the model, the carrier of the prior distribution, and the actual distribution of the data, the posterior distribution asymptotically concentrates on a set which may contain more than one point. McCulloch (1988) examines implications of this view. Bunke and Milhaud (1998) provide results on concentration of the posterior distribution and a Bayesian central limit theorem, while Kleijn and van der Vaart
(2006) treat the infinite dimensional case. In this chapter, we investigate issues of consistency and inconsistency for a variety of functions of interest, such as equality of components of a $K$-dimensional parameter vector, $\theta$, without the assumption that the model is correct.

First, we study the asymptotic behavior of the posterior distribution. We study analyses under the assumption that the densities lie in a $K$-dimensional minimal standard exponential family. After defining meaningful versions of the parameters, we show that, if the family is full, the asymptotic carrier contains a single point. Therefore, under a wide variety of loss functions, the Bayes estimator is consistent. This would seem to close the issue. However, we find a need to distinguish between the notions of $\theta_i = \theta_j$ and $||\theta_i - \theta_j|| < \epsilon$, even for vanishingly small $\epsilon$, when the model is not correct. We show that “equal to each other” and “close to each other” behave differently, even in an asymptotic manner. This distinction carries with it implications for a Bayesian test of a point null hypothesis.

Section 5.2 briefly describes the definition and the properties of an exponential family. Sections 5.3 and 5.4 examine consistency of parameter estimation and hypotheses testing, respectively. Section 5.5 briefly discusses other Bayesian inference problems. The final section contains conclusions.

5.2 Definition of the exponential family

We present the definition of a $K$-dimensional standard exponential family (Brown 1986).

**Definition 5.2.1.** Let $\nu$ be a $\sigma$-finite measure on the Borel subsets of $\mathbb{R}^K$. Let

$$
N = N_\nu = \left\{ \theta : \int \exp\{\theta' X\} \nu(d\mathbf{x}) < \infty \right\}.
$$
Let
\[ \lambda(\theta) = \int \exp\{\theta' X\} \nu(dx). \quad (5.1) \]
(Define \( \lambda(\theta) = \infty \) if the integral in Equation 5.1 is infinite.) Let
\[ \psi(\theta) = \log \lambda(\theta), \]
and define
\[ f(x|\theta) = \exp\{\theta' x - \psi(\theta)\}, \quad \theta \in \mathcal{N}. \]

Let \( \Theta \subset \mathcal{N} \). The family of probability densities \( \{f(\cdot|\theta) : \theta \in \Theta\} \) is called a \( K \)-dimensional standard exponential family (of probability densities).

Since any \( m \) dimensional exponential family which is not minimal can be reduced to a \( K \)-dimensional minimal standard family through sufficiency, reparameterization, and proper choice of \( \nu \) for some \( K \leq m \) (Brown (1986) 1.9 Theorem p13), we consider minimal standard exponential families only.

Brown (1986) listed the following fundamental properties of exponential families:

1. \( \mathcal{N} \) is a convex set and \( \psi \) is convex on \( \mathcal{N} \).

2. \( \psi \) is lower semi-continuous on \( \mathbb{R}^K \) and is continuous on \( \mathcal{N}^o \) where \( \mathcal{N}^o \) denotes the interior of \( \mathcal{N} \).

3. If the family is minimal, then \( \psi \) is strictly convex on \( \mathcal{N} \) and \( f(\cdot|\theta) \neq f(\cdot|\theta') \) for any \( \theta \neq \theta' \in \mathcal{N} \).

4. If \( \mathcal{N} \) is open, then the family is regular.
All of these fundamental properties will be used in the later proofs.

The family is full if $\Theta = \mathcal{N}$. If $\Theta$ is a $K$-dimensional differentiable manifold in $\mathcal{N}$, then such a standard exponential family is called a differentiable subfamily. One example of a differentiable subfamily is a curved exponential family whose parameter space $\Theta$ is a differentiable curve.

### 5.3 Consistency for the Bayesian parameter estimation problem

The typical discussion of consistency in a Bayesian setting requires that the conditional distribution of the data given the parameters fall in the family of models indexed by the parameters and that the “true value” of parameters fall in the support of the prior distribution. The usual argument for consistency follows from the prior distribution assigning positive mass to all Kullback-Liebler neighborhoods of the “true” parameter value. The posterior then concentrates in the weak neighborhoods of the true parameter (Schwartz 1965). We step outside this usual restrictive framework and consider the more general form of consistency that does not necessarily rely on correct specification of the likelihood.

The usual argument for consistency in Bayes models follows from the concentration of the posterior near an appropriate point in the parameter space. With the assumed likelihood in a minimal standard exponential family, $\mathbf{X}$ is the sufficient statistic for $\theta$. To ensure the needed concentration of the posterior, we will require that these sufficient statistics converge.

McCulloch (1988) discussed the likelihood function and the maximum likelihood estimator (MLE) for models in an exponential family when the true distribution generating the data does not lie in the exponential family. Under regularity conditions
on the true distribution, he showed that there exists $\theta^*$ in $\Theta$ minimizing the Kullback-Liebler divergence between the true distribution and the assumed likelihood. He also showed that the MLE of $\theta$ converges almost surely to $\theta^*$ as the size of a random sample tends to infinity.

The next theorem extends this type of consistency to the Bayesian setting. It relies on the assumption that the model used for analysis is this: the sequence of variates is i.i.d. from a $K$-dimensional full standard exponential family. The results demonstrate that the posterior distribution for a parameter is asymptotically confined to $A_0$ without the assumption that the model is correct. It is also shown that the asymptotic carrier contains a single point. Therefore, the Bayesian estimator converges to a unique point when the prior distribution has large enough support.

Bunke and Milhaud (1998) and Kleijn and van der Vaart (2006) contain similar results, though the regularity conditions differ slightly (see, for example, Remark 3 in Bunke and Milhaud (1998)). We follow the notation and definitions in Berk (1966). Throughout this paper, if there is no other specification, an expectation is the expectation with respect to the true distribution generating the data.

**Definition 5.3.1.**

1. $\eta(\theta) = E \log f(x|\theta)$, and $\eta^* = \sup_{\theta \in \Theta} \eta(\theta)$.

2. The asymptotic carrier, $A_0 = \{ \theta \in \Theta : \eta(\theta) = \eta^* \}$.

**Assumption 5.3.2.**

(1) $f(X|\theta)$ is in a $K$-dimensional full minimal standard exponential family where $\theta \in \Theta$.

(2) For all $\theta \in \Theta$, $F_0\{x; f(x|\theta) > 0\} = 1$.

(3) The prior distribution assigns positive probability to every (open) neighborhood of every point in $\Theta$. 

81
(4) \( E X_{jk} \) exists and \( E|X_{jk}| \) is finite for every \( k = 1, \ldots, K \) where \( X_{jk} \) is the \( k^{\text{th}} \) element of a \( K \)-dimensional random vector \( X_j \).

(5) \( f(X|\theta) \) is regular.

**Theorem 5.3.3.** Suppose that a model used for analysis specifies that \( K \)-dimensional random vectors \( X_j \) are i.i.d. with one of the densities \( f(\cdot|\theta) \), where \( \theta \in \Theta \). Suppose that \( \{X_j\}_{j=1}^{\infty} \) are actually independently distributed with density \( f_0(\cdot|\theta) \) (with distribution function \( F_0 \)), which need not be any of the densities in the model. If Assumption 5.3.2 (1)-(4) hold, \( A_0 \) contains a unique point \( \theta^* \), and the posterior distribution almost surely \([F_0]\) becomes asymptotically degenerate at \( \theta^* \) in the sense that if \( U \) is an open set containing \( A_0 \), \( \lim_{n \to \infty} P(\theta \in U|X_1, \ldots, X_n) = 1[F_0] \).

**Proof.** Since \( f(X|\theta) = \exp\{\theta^t X - \psi(\theta)\} \), \( \eta(\theta) = E \log f(X|\theta) = E[\theta^t X - \psi(\theta)] = EX^t \theta - \psi(\theta) \). \( \eta(\theta) \) is strictly convex on \( \Theta \). Thus, there exists a unique point \( \theta^* \) satisfying \( \eta^* = \eta(\theta^*) \) and then \( A_0 \) contains one point, \( \theta^* \).

Let \( U \) be a fixed open set containing \( A_0 \), and define \( \delta' = \inf_{\theta \in U^c}\{\eta^* - \eta(\theta)\} > 0 \).

Also, let \( \bar{X}_n = \sum_{j=1}^n X_j/n \) and define a set \( B_n = \{\theta \in \Theta : \bar{X}_n^t \theta - \psi(\theta) \geq \eta^* - \delta\} \) for some \( 0 < \delta < \delta' \).

We observe that \( \bar{X}_n \overset{a.s.}{\to} E X \) by the strong law of large numbers (SLLN), so that \( B_n \subset U \) and \( B_n \) is not empty with probability tending to 1.

Now, we consider the posterior odds;

\[
\frac{P(\theta \in U|X_n)}{P(\theta \in U^c|X_n)} > \frac{P(\theta \in B_n|X_n)}{P(\theta \in U^c|X_n)} = \frac{\int_{B_n} \exp\{n\bar{X}_n^t \theta - n\psi(\theta)\} \pi(\theta) d\theta}{\int_{U^c} \exp\{n\bar{X}_n^t \theta - n\psi(\theta)\} \pi(\theta) d\theta} \geq \left( \frac{\min_{\theta \in B_n} \exp\{\bar{X}_n^t \theta - \psi(\theta)\}}{\sup_{\theta \in U^c} \exp\{\bar{X}_n^t \theta - \psi(\theta)\}} \right)^n \frac{\int_{B_n} \pi(\theta) d\theta}{\int_{U^c} \pi(\theta) d\theta}.
\]
We note the following almost sure limits:

\[
\lim_{n \to \infty} \min_{\theta \in B_n} \exp\{\mathbf{X}_n^t \theta - \psi(\theta)\} = \exp(\eta^* - \delta), \quad \text{and} \quad (5.2)
\]

\[
\lim_{n \to \infty} \sup_{\theta \in U^c} \exp\{\mathbf{X}_n^t \theta - \psi(\theta)\} = \exp(\eta^* - \delta'). \quad (5.3)
\]

Together, Equations 5.2 and 5.3 imply that

\[
\lim_{n \to \infty} \min_{\theta \in B_n} \exp\{\mathbf{X}_n^t \theta - \psi(\theta)\} \sup_{\theta \in U^c} \exp\{\mathbf{X}_n^t \theta - \psi(\theta)\} = \exp(\delta' - \delta).
\]

Therefore, we can conclude that

\[
\left[ \min_{\theta \in B_n} \exp\{\mathbf{X}_n^t \theta - \psi(\theta)\} \right] \left[ \sup_{\theta \in U^c} \exp\{\mathbf{X}_n^t \theta - \psi(\theta)\} \right] \frac{n \int_{B_n} \pi(\theta) d\theta}{\int_{U^c} \pi(\theta) d\theta} \to \infty,
\]

which implies that

\[ P(\theta \in U | \mathbf{X}_n) \to 1 \text{ in probability}. \]

**Remark 5.3.4.** If \( \theta^* \) is an interior point of \( \Theta \), then \( \theta^* = (\nabla \psi)^{-1}(E\mathbf{X}) \).

Theorem 5.3.3 shows that even when \( \{\mathbf{X}_j, j \geq 1\} \) are not actually generated from a distribution in the full minimal standard exponential family, the asymptotic carrier contains one point only, so that the posterior asymptotically concentrates at that point in the parameter space. Therefore, consistency for \( \theta \) is asymptotically obtained. Remark 5.3.4 shows that if \( \theta^* \) is in the interior of \( \Theta \), then \( A_0 \) contains the true value. We note that this result does not depend on the form of the prior distribution when the prior distribution has large enough support.

A more straightforward argument relies on the limiting behavior of the Bayes estimators under the presumed model and the insight that the posterior distribution is determined by the sufficient statistics, regardless of their provenance. For
example, with an analysis under a full minimal standard exponential family, when \( \hat{X} \to \nabla \psi(\theta^*) \), the posterior distribution of \( \theta \) concentrates near \( \theta^* \). It matters not one whit that the data may have come from any other distribution, that the \( X_j \) are dependent or that \( \text{Var}(X_j) \) is infinite. The analysis is based on the vector of sufficient statistics, \( \hat{X} \), and so the large sample behavior of \( \hat{X} \) determines the large sample behavior of \( \pi(\theta|X) \).

However, the result of Theorem 5.3.3 may not hold if the assumed model is not full. The following two examples show that for a curved exponential family, consistency of the estimators under a Bayesian model is not always achieved. In the examples, \( Y \) denotes a point in the sample space. Bunke and Milhaud (1998) presented similar examples based on normal distributions showing inconsistency of Bayes estimators. The following examples differ from theirs in terms of the prior distribution and the form of the true distribution generating the data.

**Example 5.3.5.** For the normal family, assume that \( \xi_0^2 = \sigma_0^2 \) where \( EY = \xi_0 \neq 0 \) and \( \text{Var}(Y) = \sigma_0^2 \), so that

\[
g(Y|\xi) = \frac{1}{\sqrt{2\pi} |\xi|} \exp\left\{ \frac{1}{\xi} Y - \frac{1}{2\xi^2} Y^2 - \frac{1}{2} \right\}, \quad \xi \neq 0.
\]

(5.4)

Although this is formally a two-parameter exponential family with natural parameter \( \theta = (1/\xi, -1/(2\xi^2))^t \) and statistic \( X = (Y, Y^2)^t \), the two-dimensional parameter \( \theta \) lies on a curve in \( \Theta = \mathbb{R}^2 - \{0\} \), making the family of distribution in Equation 5.4 a curved exponential family. Although \( g(Y|\xi) \) is not full, \( A_0 \) always contains one value only and the posterior distribution is almost surely \([F_0] \) asymptotically carried on \( A_0 \).

In addition, \( A_0 \) contains \( \xi_0 \), only when \( \xi_0^2 = \sigma_0^2 \).
Proof. Let

\[ f(Y|\xi) = \frac{1}{|\xi|} \exp\left\{ \frac{1}{\xi} Y - \frac{1}{2} \frac{Y^2}{\xi^2} \right\}, \quad \xi \neq 0. \]

Then,

\[
\eta(\xi) = E[- \log |\xi| + \frac{Y}{\xi} - \frac{Y^2}{2\xi^2}] - \frac{1}{2} = - \log |\xi| + \frac{\xi_0}{\xi} - \frac{\xi_0^2 + \sigma_0^2}{2\xi^2} - \frac{1}{2}.
\]

Also, we have

\[
\eta'(\xi) = - \frac{1}{\xi} - \frac{\xi_0}{\xi^2} + \frac{\xi_0^2 + \sigma_0^2}{\xi^3} = - \frac{1}{\xi^3} \{ \xi^2 + \xi_0 \xi - (\xi_0^2 + \sigma_0^2) \},
\]

\[
\eta''(\xi) = \frac{2\xi_0}{\xi^3} - 3 \frac{\xi_0^2 + \sigma_0^2}{\xi^4} = \frac{1}{\xi^3} \{ \xi^2 + 2\xi_0 \xi - 3(\xi_0^2 + \sigma_0^2) \}.
\]

Let \( \xi_1 = (-\xi_0 - \sqrt{5\xi_0^2 + 4\sigma_0^2})/2 \) and \( \xi_2 = (-\xi_0 + \sqrt{5\xi_0^2 + 4\sigma_0^2})/2. \) We note that \( \xi_1 < 0 < \xi_2 \) and that

\[
\eta'(\xi_1) = \eta'(\xi_2) = 0,
\]

\[
\eta''(\xi_1) < 0, \text{ and } \eta''(\xi_2) < 0.
\]

Thus, \( \eta(\xi) \) has local maxima at \( \xi_1 \) and \( \xi_2. \)

Without loss of generality, assume that \( \xi_0 > 0. \) We observe that \( |\xi_1| > 0 \) and \( \eta(\xi_2) > \eta(|\xi_1|). \) Also,

\[
\eta(|\xi_1|) - \eta(\xi_1) = \frac{\xi_0}{|\xi_1|} - \frac{\xi_0}{\xi_1} > 0, \text{ that is, } \eta(|\xi_1|) > \eta(\xi_1).
\]

Therefore, \( A_0 \) contains either one of \( \xi_1 \) or \( \xi_2 \) only depending on the sign of \( \xi_0. \) Also, if the assumption that \( \xi_0^2 = \sigma_0^2 \) is satisfied, then \( A_0 \) contains \( \xi_0. \)

In Example 5.3.5, \( A_0 \) contains a single value, and the value could lead to asymptotic means and variances matching actual means and variances, depending on whether
the assumption $\xi_0^2 = \sigma_0^2$ is satisfied. The following situation illustrates different behavior of a curved exponential family. Under this model, $A_0$ may contain more than one value, which is in contrast to the conclusion of Theorem 5.3.3 for full exponential families.

**Example 5.3.6.** For the normal family, assume that

$$
\sigma_0^2 = \begin{cases} 
\xi_0^2, & \text{if } \xi > 0, \\
\frac{c\xi_0^2}{\xi^2}, & \text{if } \xi < 0,
\end{cases}
$$

(5.5)

for some $c > 0$, where $EY = \xi_0 (\neq 0)$ and $\text{Var}(Y) = \sigma_0^2$, so that

$$
g(Y|\xi) = \begin{cases} 
\frac{1}{\sqrt{2\pi|\xi|}} \exp\left\{\frac{1}{\xi} Y - \frac{1}{2c} Y^2 - \frac{1}{2}\right\}, & \text{if } \xi > 0, \\
\frac{1}{\sqrt{2\pi|\xi|c}} \exp\left\{\frac{1}{c\xi} Y - \frac{1}{2c} Y^2 - \frac{1}{2}\right\}, & \text{if } \xi < 0,
\end{cases}
$$

This is an example of a curved exponential family with natural parameter $\theta = (1/\xi, -1/(2\xi^2))^t$ and a statistic $X = (Y, Y^2)^t$. The two-dimensional parameter $\theta$ lies on a curve in $\Theta = \mathbb{R}^2 - \{0\}$. Although Assumption 5.3.2 (1)-(4) hold, $A_0$ contains one or two values depending on the values of $\xi_0, \sigma_0^2$ and $c$. In addition, when only when Equation 5.5 is satisfied, $A_0$ contains $\xi_0$ only.

**Proof.** If $\xi_0 > 0$, then $\eta(\xi), \eta'(\xi)$ and $\eta''(\xi)$ are the same as in Example 5.3.5. On the other hand, if $\xi_0 < 0$, then

$$
\eta(\xi) = E[-\log|\xi| + \frac{Y}{c\xi} - \frac{Y^2}{2c\xi^2} - \frac{1}{2c} - \frac{1}{2} \log(c)] \\
= -\log|\xi| + \frac{\xi_0}{c\xi} - \frac{\xi_0^2 + \sigma_0^2}{2c\xi^2} - \frac{1}{2c} - \frac{1}{2} \log(c).
$$

Also, we have

$$
\eta'(\xi) = -\frac{1}{\xi} - \frac{\xi_0}{c\xi^2} + \frac{\xi_0^2 + \sigma_0^2}{c\xi^3} = -\frac{1}{c\xi^3}\{c\xi^2 + \xi_0 \xi - (\xi_0^2 + \sigma_0^2)\},
$$

$$
\eta''(\xi) = \frac{1}{\xi^2} + \frac{2\xi_0}{c\xi^3} - \frac{3(\xi_0^2 + \sigma_0^2)}{c\xi^4} = \frac{1}{c\xi^4}\{c\xi^2 + 2\xi_0 \xi - 3(\xi_0^2 + \sigma_0^2)\}.
$$
Let \( \xi_1 = (-\xi_0 - \sqrt{\xi_0^2 + 4c(\xi_0^3 + \sigma_0^2)})/(2c) \) and \( \xi_2 = (-\xi_0 + \sqrt{5\xi_0^2 + 4\sigma_0^2})/2 \). The other two roots of the quadratics determined by \( \eta'(\xi) = 0 \) lie outside the ranges of \( \xi_0 \) where the quadratics hold (See Equation 5.5). We note that \( \xi_1 < 0 < \xi_2 \) and that
\[
\eta'(\xi_1) = \eta'(\xi_2) = 0,
\]
\[
\eta''(\xi_1) < 0, \text{ and } \eta''(\xi_2) < 0.
\]
Thus, \( \eta(\xi) \) has local maxima at \( \xi_1 \) and \( \xi_2 \).

Now we consider \( \eta(\xi_2) - \eta(\xi_1) \).
\[
\eta(\xi_2) - \eta(\xi_1) = \log \left| \frac{\xi_0}{\xi_0} \right| - \frac{1}{2} \frac{1}{2c} - \frac{1}{2} \log(c)
\]
\[
+ \left\{ \begin{array}{l}
\frac{2\xi_0}{\xi_0 + \sqrt{5\xi_0^2 + 4\sigma_0^2}} \{\xi_0 + \sqrt{5\xi_0^2 + 4c(\xi_0^3 + \sigma_0^2)}\} \\
2(5\xi_0^2 + \sigma_0^2) \end{array} \right\}
\]
\[
- \left\{ \begin{array}{l}
\frac{2\xi_0}{\xi_0 + \sqrt{5\xi_0^2 + 4\sigma_0^2}} \{\xi_0 + \sqrt{5\xi_0^2 + 4c(\xi_0^3 + \sigma_0^2)}\} \\
2(5\xi_0^2 + \sigma_0^2) \end{array} \right\}
\]
\[
\left\{ \begin{array}{l}
\frac{2\xi_0}{\xi_0 + \sqrt{5\xi_0^2 + 4\sigma_0^2}} \{\xi_0 + \sqrt{5\xi_0^2 + 4c(\xi_0^3 + \sigma_0^2)}\} \\
2(5\xi_0^2 + \sigma_0^2) \end{array} \right\}^2.
\]

Without loss of generality, assume that \( \xi_0 > 0 \), and then we have that
\[
\lim_{\sigma_0^2 \to 0} \eta(\xi_2) - \eta(\xi_1) = \left\{ \begin{array}{l}
-\log(1 + \sqrt{5}) + \frac{2}{1+\sqrt{5}} - \frac{2}{(-1 + \sqrt{5})^2} - \frac{1}{2} \\
\frac{2c}{(1+\sqrt{1+4c})^2} - \frac{1}{2c}\end{array} \right\},
\]
\[
\lim_{\sigma_0^2 \to \infty} \eta(\xi_2) - \eta(\xi_1) = \frac{1}{2} + \frac{1}{2c}.
\]
For any value \( c > 0 \), \( \lim_{\sigma_0^2 \to 0} \eta(\xi_2) - \eta(\xi_1) > 0 \), but \( \lim_{\sigma_0^2 \to \infty} \eta(\xi_2) - \eta(\xi_1) < 0 \) for \( 0 < c < 1 \) and \( \lim_{\sigma_0^2 \to \infty} \eta(\xi_2) - \eta(\xi_1) > 0 \) for \( c > 1 \). We observe that \( \eta(\xi_2) - \eta(\xi_1) \) is continuous in \( \sigma_0^2 \) and \( c \). Therefore, for any fixed \( \xi_0 \), there exist some values of \( \sigma_0^2 \) and \( c \) such that \( \eta(\xi_2) - \eta(\xi_1) = 0 \) so that \( A_0 \) contains both \( \xi_1 \) and \( \xi_2 \). Also, if the assumed condition Equation 5.5 is satisfied, then \( A_0 \) contains either \( \xi_1 \) or \( \xi_2 \) only and the value is equal to \( \xi_0 \).
5.4 Consistency for the Bayesian hypothesis testing problem

Theorem 5.3.3 seems to cover consistency for the standard inferences, allowing us to decide whether (and which) $\theta$ are nearly equal or are different. However, there is a subtle difference between parameters being equal and merely being nearly equal that is highlighted by the consistency issue. We believe that this should have an impact on formulation of hypothesis testing as a Bayesian decision problem.

In the following, we consider consistency issues in Bayesian hypothesis testing problems, $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$, where $\theta_i$ denotes the “true mean” of group $i$ for $i = 1, 2$, and illustrate the concern in the following theorem and examples.

Assumption 5.4.1. 
(1) $E X_i^2$ is finite for every $k = 1, \ldots, K$ and $i = 1, 2$.

(2) The prior assigns positive probability to the event $\theta_1 = \theta_2$. If $\theta_1 = \theta_2$, the common value of $\theta$ comes from a distribution, which has density with respect to Lebesgue measure $\pi(\theta)$. If $\theta_1 \neq \theta_2$, then $\theta_i$ are i.i.d. from $\pi(\theta)$.

(3) $\pi(\theta)$ satisfies the following:

(a) $\pi(\theta)$ is continuous in a (open) neighborhood of $\theta_i$ for $i = 1, 2$.

(b) $\pi(\theta_i) > 0$ for $i = 1, 2$.

Theorem 5.4.2. Suppose that the model used for analysis specifies that $\{X_{ij}\}_{j=1}^{\infty}$, $i = 1, 2$ are mutually independent sequences of $K$-dimensional random vectors, where $X_{ij}$ has density $f(\cdot|\theta_i)$, with $\theta_i \in \Theta$ for $i = 1, 2$. Let $P(\cdot|X_n)$ denote a probability with respect to the posterior distribution of $\theta$ given $X_{11}, \ldots, X_{1n}, X_{21}, \ldots, X_{2n}$. Suppose that $\{X_{ij}\}_{j=1}^{\infty}$ are actually independently distributed with density $f_0(\cdot|\theta_i)$, which need not to be any of the densities in the model. If Assumptions 5.3.2 (1), (2) and (5),
and 5.4.1 (1)-(3) hold, then for any fixed $M > 0$

$$\lim_{n \to \infty} P \left( \frac{P(\theta_1 = \theta_2 | X_n)}{P(\theta_1 \neq \theta_2 | X_n)} > M \right) = 1,$$

when $\theta_1 = \theta_2$.

**Proof.** The proof of this theorem is presented in Appendix C. □

**Remark 5.4.3.** The result of Theorem 5.4.2 holds even when the data have $\text{cor}(X_{ij}, X_{ij'}) \neq 0$ or $\text{cor}(X_{ij}, X_{i'j}) \neq 0$ for $i \neq i'$ and $j \neq j'$. It is sufficient to have $(\bar{X}_1^t, \bar{X}_2^t)^t$ tend to its mean at an $O_p(n^{-1/2})$ rate. In this case, the proof of Theorem 5.4.2 remains essentially unchanged.

Theorem 5.4.2 and Remark 5.4.3 tell us that consistency for Bayesian hypotheses testing is asymptotically obtained with a full minimal standard exponential family model. It does not matter that the data may have come from any other distribution or that the $X_{ij}$ show relatively strong dependence. In the following example, it is shown that we can obtain a stronger conclusion with assumed normal likelihoods because $\nabla^2 \psi(\theta) = 1$.

**Example 5.4.4.** Suppose that Assumptions 5.4.1 (1) - (3) hold, and let $f(X_{ij} | \theta_i) = \exp\{-X_{ij} \theta_i - \theta_i^2/2\}$, $i = 1, 2$, where $\Theta = \mathbb{R}$. The model specifies that the $X_{ij}$ are independently drawn from $N(\theta_i, 1)$ distributions. Then, when $\theta_1 = \theta_2$, $P(\theta_1 = \theta_2 | X_n)$ converges to 1 a.s. as $n$ goes to infinity.

**Proof.** Since $\psi(\theta) = \theta^2/2$, $\hat{\theta} = \bar{X}$, $\hat{\theta}_i = \bar{X}_i$ and $\psi''(\theta) = 1$.

From Equation C.1 in the proof of Theorem 5.4.2, the Bayes factor can be bounded below as

$$\frac{P(\theta_1 = \theta_2 | X_n)}{P(\theta_1 \neq \theta_2 | X_n)} \geq \frac{\int_{B_n} \exp\{2n\bar{X} \theta - 2n\psi(\theta)\} \pi(\theta) d\theta}{\int_{\mathbb{R}} \exp\{nX_1 \theta - n\psi(\theta)\} \pi(\theta) d\theta \int_{\mathbb{R}} \exp\{nX_2 \theta - n\psi(\theta)\} \pi(\theta) d\theta} \quad (5.6)$$
From Equation C.3,
\[
\int_{B_n} \exp\{2n \bar{X} \theta - 2n \psi(\theta)\} \pi(\theta) d\theta \geq L_n \sqrt{\frac{\pi}{n}} \exp(n \bar{X}^2) \int_{B_n} \psi(\theta | \bar{X}, \frac{1}{2n}) d\theta. \tag{5.7}
\]

From Equation C.7, for each \(i\)
\[
\int_{\mathbb{R}} \exp\{n \bar{X}_i \theta - n \psi(\theta)\} \pi(\theta) d\theta \leq \frac{U_{in}}{P(\theta \in B_{in})} n \sqrt{\frac{2\pi}{n}} \exp(\frac{n \bar{X}_i^2}{2}). \tag{5.8}
\]

By combining Equations 5.7 and 5.8, let
\[
R_n = \frac{\sqrt{n} L_n P(\theta \in B_{2n}) P(\theta \in B_{2n})}{U_{in} U_{2n} 2 \sqrt{\pi}} \exp\{-\frac{n}{4} (\bar{X}_1 - \bar{X}_2)^2\},
\]

and then \(R_n\) provides a lower bound for the RHS of Inequality 5.6 a.s.

The law of the iterated logarithm (Breiman (1992)) states that
\[
\liminf_{n \to \infty} \exp\{-\frac{n}{4} (\bar{X}_1 - \bar{X}_2)^2\} = (\log n)^{-\sigma_0^2/2} \text{ a.s., where } \sigma_0^2 = Var(X_1 - X_2).
\]

Therefore,
\[
\lim_{n \to \infty} R_n = \infty,
\]

which implies that
\[
P(\theta_1 = \theta_2 | X_n) = 1 \text{ a.s. when } \theta_1 = \theta_2.
\]

This example illustrates consistency of the Bayesian hypothesis test when both \(X_1\) and \(X_2\) have finite variance. In contrast, the following example shows that if either one of the variances is not finite, one cannot achieve consistency in testing with the assumed normal model.
Example 5.4.5. Suppose that Assumptions 5.4.1 (2)-(3) hold, and let
\[ f(X_{ij} | \theta) = \exp\{-X_{ij}\theta - \theta_i^2/2\}, \quad i = 1, 2 \text{ where } \Theta = \mathbb{R}, \quad i = 1, 2. \]
Assume that \( E\chi_{ij} \) exists but \( E|X_{ij}|^{1+c} \) does not exist for some \( 0 < c < 1 \), for either \( i = 1 \) or \( i = 2 \). Then, when \( \theta_1 = \theta_2 \), \( P(\theta_1 = \theta_2 | X_n) \) does not tend to 1.

Proof. Similar to Equation C.1 in the proof of Theorem 5.4.2, the Bayes factor can be bounded above as
\[
P(\theta_1 = \theta_2 | X_n) \leq \frac{\int_{\mathbb{R}} \exp\{2n\bar{X}\theta - 2n\psi(\theta)\} \pi(\theta) d\theta}{\prod_{i=1}^{2} \int_{B_n} \exp\{nX_i\theta - n\psi(\theta)\} \pi(\theta) d\theta}.
\] (5.9)

Similar to Equations C.7,
\[
\int_{\mathbb{R}} \exp\{2n\bar{X}\theta - 2n\psi(\theta)\} \pi(\theta) d\theta \leq \frac{U_n}{P(\theta \in B_n)} \sqrt{n \exp\left(\frac{n\bar{X}^2}{2}\right)}.
\] (5.10)

Similar to Equations C.3,
\[
\int_{\mathbb{R}} \exp\{n\bar{X}_i\theta - n\psi(\theta)\} d\theta \geq L_n \sqrt{\frac{2\pi}{n} \exp\left(\frac{n\bar{X}^2_i}{2}\right)}.
\] (5.11)

By combining Equations 5.10 and 5.11, let
\[
R'_n = \frac{U_n \sqrt{n}}{2\sqrt{\pi}L_1L_1nP(\theta \in B_n)} \exp\{- \frac{n}{4}(\bar{X}_1 - \bar{X}_2)^2\},
\] (5.12)

and then \( R'_n \) is an upper bound for the RHS of Inequality 5.9 a.s.

Let \( Y_j = X_{1j} - X_{2j} \) and \( S_n = \sum_{j=1}^{n} Y_j \). Since \( E|Y|(1+\delta) = \infty \) for some \( 0 < \delta < 1 \) and \( EY = 0 \), the event that \( |S_n| > n^{1/(1+\delta)} \) takes place for infinitely many \( n \) has probability 1 (Feller (1946), Theorem 1). Also, we note that if \( |S_n| > n^{1/(1+\delta)} \),
\[
\exp\{- \frac{n(n\bar{X}_1 - \bar{X}_2)^2}{4}\} = \exp\{- \frac{(\sqrt{n}\bar{Y})^2}{4}\} = \exp\{- \frac{1}{4}(\frac{S_n}{\sqrt{n}})^2\} < \exp\{- \frac{1}{4}n^{\frac{1}{1+\delta}}\}.
\]

Then,
\[
\liminf_{n \to \infty} R'_n = 0 \text{ with probability 1},
\]

91
which implies that
\[
\liminf_{n \to \infty} P(\theta_1 = \theta_2|X_n) = 0 \text{ in probability, even when } \theta_1 = \theta_2.
\]

The preceding two examples illustrate that, in the broad context of consistency issues when the sampling model lies outside the support of the prior distribution, we must distinguish between null hypotheses of the form \( \theta_1 = \theta_2 \) and \( ||\theta_1 - \theta_2|| < \epsilon \) for some small \( \epsilon \). Consistency for the latter is obtained more easily than is consistency for the former. As a consequence of this result, when we believe that the data indicate a fat-tailed sampling distribution, or where there are a number of potentially outlying observations, we recommend rephrasing the declaration of no difference from \( \theta_i = \theta_j \) to \( ||\theta_i - \theta_j|| < \epsilon \) for a suitably small \( \epsilon \).

Finally, we note that the distinction drawn above applies quite generally to Bayesian testing problems. We generally believe \( ||\theta_i - \theta_j|| < \epsilon \) to be a preferable formulation of the concept of “no difference”. This is in accord with much of the discussion of whether any point null hypothesis should receive positive probability. However, much of this discussion has been at a philosophical level. With this example, we provide a mathematical demonstration of a difference between point nulls and nulls that concentrate near a point. This difference even holds up asymptotically.

### 5.5 Consistency in other problems

There is a wide range of goals in Bayesian inference problems. These goals include parameter estimation and hypothesis testing, as discussed in Sections 5.3 and 5.4. Another goal is to obtain the predictive distribution. With strong conditions on the
likelihood, consistency results for these more general objectives can be easily derived. However, with weak conditions on the likelihoods, statements about consistency often degenerate into a morass of special cases and hedged statements.

Suppose that the model used for analysis is one of normal likelihoods with a common variance, and consider the probability that an observation exceeds some threshold, \( t \), i.e., \( p_i = P(X_{i,n_i+1} \in A) \) where \( X_{ij} \) are observed for group \( i = 1, \ldots, K \) and \( j = 1, \ldots, n_i \), and \( A = (t, \infty) \). We may wish to determine equality and inequality between \( p_i \), etc. Under the assumed likelihoods, statements about the \( p_i \) correspond directly to statements about the means, \( \theta_i \). Furthermore, consistency for the entire predictive distribution can be obtained only with the correct model. The strongest conclusions on consistency for the \( p_i \) rely on this type of strong condition on the likelihood.

For some statements about the \( p_i \), the condition of having the correct likelihood can be replaced by a weaker condition. For example, our goal may be to select the group with the largest \( p_i \). If the groups are all members of some location family with full support, finite mean and common, finite variance, then \( p_i \) is a monotonic, strictly increasing function of \( \theta_i \). This implies that, exclusive of ties among the \( p_i \), we may well be able to asymptotically determine the ordering of the \( \theta_i \) and hence the ordering of the \( p_i \). This is enough to guarantee an asymptotically correct selection of the best group.

For other goals, the conditions cannot be relaxed. Under the general conditions of consistency that we consider, where the groups may have different variances and the likelihoods may be non-normal, there is no direct connection between the \( p_i \) and
the $\theta_i$. Incorrect specification of the form of the likelihood leads to inconsistency. For some sets $A$, the predictive probabilities converge to the wrong number.

Hedges to statements of inconsistency arise because, for some sets $A$, the true predictive probabilities will match the limit with the assumed likelihood. As an example, consider the situation where the true family of sampling distributions is the gamma family, the assumed is the normal family, and estimation of the $p_i$ is our goal. An easy indirect proof of the contrapositive establishes the result that, for normal and gamma distributions with matching means and variances, there is some threshold $t$ for which the correspondingly defined $p_i$ are the same: Suppose the conclusion does not hold. Then the gamma cdf is always less than (greater than) the normal cdf (i.e., $F_N(x) < F_G(x)$ for all $x$). But this implies that the gamma variate is stochastically greater than (less than) the normal variate, and hence that it has a larger (smaller) mean. This is a contradiction, and so we have proven our statement. This proof in no way relies on the gamma family, but applies to any continuous random variable for which the variance exists.

Further examination of the result for the normal and gamma shows that it is exceedingly rare (in almost any sense) for the two $p_i$s to match. We refer to this sort of consistency as accidental consistency. As a general rule, one can handle accidental consistency by a slight reformulation of the problem. To do so, we would place a topology on an intermediate object of inference or on the eventual object of inference and demand consistency for every parameter value in an open neighborhood about the point. If consistency does not obtain throughout the neighborhood, we would term the result accidentally consistent. Such a development enables us to make much clearer and stronger statements about inconsistency. We avoid this formulation of the
problem here, as substantial space and technical notation is required to fully describe this notion.

The development of the above argument can be paralleled for other estimands. The general predictive ranking and selection problem has, as its goal, choice of a group which maximizes the value of a future observation on the group. The value of an observation is determined by $V_i = \int f_i(x)w(x)dx$, where $w(x)$ satisfies conditions that guarantee finite values for all groups. Consistency of the predictive distributions along with suitable conditions on $w$ ensure consistent estimation of the $V_i$, and hence selection of a best group. Inconsistent estimation of the predictive distribution will typically result in inconsistent estimation of the $V_i$.

5.6 Conclusions

The consistency (and inconsistency) results discussed in this chapter apply to a much wider class of models than the standard minimal exponential family models considered here. The differentiation between $\theta_i = \theta_j$ and $||\theta_i - \theta_j|| < \epsilon$ for a suitably small $\epsilon$ is relevant to much modern applied Bayesian work. Current work on model selection and model averaging is but one example of the importance of this distinction.

Since the seminal work of George and McCulloch (1993) and Raftery et al. (1997) on Bayesian model selection/model averaging via MCMC simulation, many authors have joined the fray, writing meta-regression-models which consist of a collection of models (the models are defined by the set of non-zero regression coefficients) and distributions over parameters conditional on each of the models. Much work in the area selects a single model or a subset of models for averaging based on evaluation of the probability that coefficients in the models are non-zero. Our discussion of consistency
highlights the difference between “equal to zero” and “near zero”, demonstrating even asymptotic differences in behavior. In the case of ANOVA, where models correspond to exact equality of groups of the treatment means, the implication is that the wrong model may be selected. In a more general regression setting, the implication is also that the wrong model may be selected, even asymptotically.
CHAPTER 6

CONTRIBUTIONS AND FUTURE WORK

In this dissertation, we have discussed robust statistical modeling through non-parametric Bayesian methods. For this purpose, we developed two nonparametric Bayesian models: the limiting Dirichlet process model and the local-mass preserving mixture of Dirichlet processes model. We also examined the asymptotic behavior of Bayes estimators in the case where a model is incorrect. In this chapter, we summarize the material presented in Chapters 3 to 5. We also describe directions for future work in areas related to the topics discussed in this dissertation.

In Chapter 3, we examined the question of how to create a “noninformative” prior distribution in the context of nonparametric Bayesian analysis. We described qualitative features of acceptable posterior behavior. We showed that the standard approaches to creating noninformative prior distributions for parametric Bayesian models produce unacceptable posterior behavior in our context. In contrast, the limiting Dirichlet process model leads to a reasonable posterior distribution with only modest input. The limiting Dirichlet process model is derived as the limit of a sequence of mixture of Dirichlet processes models. The key to obtaining a proper posterior distribution, developed in this work, is the concept of local mass. The limiting Dirichlet process model was compared to the mixture of Dirichlet processes
model on several data sets. We found that the limiting Dirichlet process model performs better than the mixture of Dirichlet processes model, especially when there is a great deal of information on the individual case.

In Chapter 4, we applied the concept of local mass to specification of hierarchical prior distributions in nonparametric Bayesian models. We developed a class of prior distributions which preserve local mass. To do so, we carefully considered behavior of parameters of interest in some small region, and we developed a class of prior distributions that preserve mass in the region. Our strategy of local mass preservation is to tie the mass of the base measure to its dispersion, so that for any non-null measurable set in the region of interest, local mass remains approximately constant as the dispersion varies. We showed that local mass preservation results in robust inference. We applied the strategy of constructing such a prior distribution to Dirichlet process models and discussed calibration of the prior distribution. We demonstrated performance of the new model with actual data, and we compared it to mixture of Dirichlet processes models having traditional prior structure.

The development of the models — the limiting Dirichlet process model and the local-mass preserving mixture of Dirichlet process model — provides insight into how to use nonparametric Bayesian models: The concept of local mass is important, as it enables us to properly capture tail behavior and eventually to obtain robust inference. The importance of the concept extends to other nonparametric Bayesian models. A future research topic is how to extend the concept of local mass to the wide variety of nonparametric Bayesian models which are currently in use, such as completely random measures, product partition models, and Pólya tree models.
In addition, we illustrated the use of the two models with simple applications to the compound decision problem and to the one-way ANOVA problem. The models are compatible with hierarchical Bayesian modelling strategies, and so their use can be extended to address a broad range of statistical problems such as regression, survival analysis, clustering, etc. As one example, we describe application to the nonparametric Bayesian density estimation problem. The traditional density estimation method in this area uses a mixture of Dirichlet processes model. Since the two models developed in this dissertation preserve local mass in the tails, they capture the tail behavior in a different fashion. The difference between the traditional model and our models can be viewed as analogous to the differences between the two main approaches to the classical kernel density estimation. One approach uses a kernel with constant bandwidth; the other uses kernels of differing widths, enabling one to better capture the tail of the density. Our methods are similar to the latter approach while the traditional method is similar to the former approach.

The last portion of this dissertation is concerned with consistency for a variety of functions of interest in the Bayesian setting, without the assumption that the model is correctly specified. First, we studied asymptotic concentration of the posterior distribution around the true value of a parameter. After defining meaningful versions of the parameters, we provided a set of regularity conditions under which the asymptotic carrier contains a single point. From there, it is a short step to show that the Bayes estimator is consistent. This would seem to close the issue. However, we found a need to distinguish between the notions of \( \theta_i = \theta_j \) and \( ||\theta_i - \theta_j|| < \epsilon \), even for vanishingly small \( \epsilon \), when the model is not correct. We showed that “equal to one another” and “close to one another” behave differently, even in an asymptotic
manner. This distinction carries with it implications for a Bayesian test of a point null hypothesis. This chapter also sketched a general approach to notions of consistency, distinguishing between traditional statements of consistency and the notion of accidental consistency.

The results in Chapter 5 can be linked to nonparametric Bayesian models involving a Dirichlet process prior or other prior distribution with a mixture structure. A Dirichlet process prior assigns positive probability to the event that each specific pair of components have identical parameter values. The result in Chapter 5 provides a caution for inference about the equality of any two such components under those models. Furthermore, the Dirichlet process prior assigns probability to each partition of components into clusters, with members of a cluster sharing a common parameter value. Partitioning components into clusters is one goal of a cluster analysis, and so the Dirichlet process prior is often used to drive nonparametric Bayesian clustering. The results of Chapter 5 suggest that there may be a benefit in redefining a cluster as a collection of components with similar parameter values rather than as a collection of components with a common parameter value.
This section contains the derivation of the posterior probability (up to the normalizing constant) of a particular configuration under Model 1. For this particular configuration, we assume that there are \( p \) clusters, of size \( c_1, \ldots, c_p \). The star notation denotes observations within a particular cluster and \( n \) denotes \( \sum_{i=1}^{k} n_i \). For clarity, we write \( v \) in place of \( \sigma^2 \).

To compute the posterior probabilities under Model 1, we extend the previous calculation under Model 0 to account for an unknown variance. Carrying that portion of the posterior dealing with the variance throughout the Model 0 calculation, we write, with \( P_a \) denoting the probability under some fixed \( a > 0 \)

\[
P_a(s|X) \propto M^p \int (2\pi v)^{-n+p} \prod_{i=1}^{p} \left\{ \Gamma(c_i)(n_i^*)^{-\frac{1}{2}} \exp\left(-\frac{SSE_i^*}{2v}\right) \right\} \frac{b^{-a}}{\Gamma(a)} v^{-(a+1)} \exp\left(-\frac{1}{v b'}\right) dv
\]

\[
\rightarrow M^p (2\pi)^{-\frac{n-p}{2}} \prod_{i=1}^{p} \left\{ \Gamma(c_i)(n_i^*)^{-\frac{1}{2}} \right\} \int v^{-\frac{n-p}{2}} \exp\left(-\frac{1}{v} \sum_{i=1}^{p} \frac{SSE_i^*}{2}\right) dv
\]

\[
\propto M^p (2\pi)^{\frac{p}{2}} \prod_{i=1}^{p} \left\{ \Gamma(c_i)(n_i^*)^{-\frac{1}{2}} \right\} \Gamma\left(\frac{n-p}{2}\right) \left(\sum_{i=1}^{p} \frac{SSE_i^*}{2}\right)^{-\frac{n-p}{2}}. \tag{A.1}
\]

Conditional on a particular cluster vector, \( s \), the posterior distribution of \( v \) is inverse gamma with parameters \( a' = (n-p)/2 \) and \( b' = (\sum_{i=1}^{p} SSE_i^*/2)^{-1} \). In order for
this inverse gamma distribution to exist, we must have $a' > 0$ (hence $n > p$) and we must also have $b' > 0$. This leads to the conditions for the existence of a proper prior distribution given in Section 3.2.3. Since the final expression is positive and finite for all partitions, and since the same constant of proportionality applies to all partitions, each partition receives positive posterior probability. Given $s$ and $v$, the posterior distributions of the $\gamma_i$ are independent normal distributions with means $\bar{X}_i^\ast$ and variances $v/n_i^\ast$, respectively.

The simplicity of the formulas for Models 0 and 1 allow us to directly evaluate the posterior distribution on $s$ for experiments with a modest number of treatments. For larger experiments, following MacEachern (1994) refinement of Escobar (1994), we can run a Gibbs sampler on $s$.

Formulas for Models 2, 3 and 4 follow from similar calculation. The final formula for Model 2 is

$$P(s, v|X) \propto M^p (2\pi)^{\frac{p^2}{2}} \prod_{l=1}^{k} \left( v_l^{-\frac{n_l}{2} + 1} \exp\left( -\frac{1}{2} \sum_{j=1}^{n_l} X_{ij}^2 \right) \right) \prod_{i=1}^{p} \left\{ \Gamma(c_i) \left( \sum_{l=1|s_l=i}^{k} n_l v_l^{-1} \right)^{-\frac{1}{2}} \exp\left( \frac{1}{2} \sum_{l=1|s_l=i}^{k} \frac{n_l \bar{X}_l v_l^{-1}}{n_l v_l^{-1}} \right) \right\}.$$  

Conditional on $(s, v, X)$, the $\gamma_i$ follow independent normal distributions with means $\frac{\sum_{l=1|s_l=i}^{k} n_l v_l^{-1} \bar{X}_l}{\sum_{l=1|s_l=i}^{k} n_l v_l^{-1}}$ and variances $(\sum_{l=1|s_l=i}^{k} n_l v_l^{-1})^{-1}$, respectively. Since the $v_l$ cannot be marginalized in any easy fashion, computations can be performed via Markov chain Monte Carlo techniques on the parameter $(s, v)$. In this model, all calculations involving the limiting Dirichlet process are conditionally conjugate, and so a technique such as that in Bush and MacEachern (1996) can be used for simulation.
The formulas for Model 3 can be simplified in a variety of fashions, leading to more complex simulation strategies for fitting the model. Since there are two limiting Dirichlet processes in this model, additional subscripts are needed on $s$, $M$, $p$, $c$ and $n^*$ to indicate which limiting Dirichlet process is being considered. A basic formula is the following, for a simulation based fit on the parameter $(s_1, s_2, v, \gamma)$:

$$P(s_1, s_2, v, \gamma|X) \propto M_1^{p_1} M_2^{p_2} \prod_{i=1}^{p_1} \Gamma(c_{1i}) \prod_{i=1}^{p_2} \left\{ \Gamma(c_{2i}) v_i^{-\frac{n_i^*}{2}} \right\} \prod_{i=1}^k \exp \left\{ -\frac{v_i^{s_{2i}}}{2} (SSE_l + n_l (\bar{X}_l - \gamma_{s1l})^2) \right\}.$$  

This formula is sufficient to develop a simulation based fit, following the algorithm in Bush and MacEachern (1996).

Model 4 admits more simplification than do Models 2 and 3. Under this model, there is a single limiting Dirichlet process, and this leads to the final formula

$$P(s|X) \propto M_p (2\pi)^{\frac{p}{2}} \prod_{i=1}^p \left\{ \Gamma(c_i) \Gamma(n_i^* - \frac{1}{2}) n_i^*^{-\frac{1}{2}} \left( \frac{SSE_i^*}{2} \right)^{-\frac{n_i^* - 1}{2}} \right\}.$$  

Conditional on $(s, X)$, the $v_i$ have independent inverse gamma distributions with parameters $a' = \frac{n_i^* - 1}{2}$ and $b' = \frac{2}{SSE_i^*}$, respectively; also conditioning on these parameters, the $\gamma_i$ have independent normal distributions with means $\bar{X}_i^*$ and variances $\frac{v_i}{n_i^*}$, respectively. The formula for this model is sufficient to allow one to directly evaluate the posterior distribution on $s$ for experiments with a moderate number of treatments. For experiments with more treatments, a simulation based fit can be used, working only with the vector $s$.  

103
APPENDIX B

DERIVATION OF THE PREDICTIVE POSTERIOR DISTRIBUTIONS UNDER THE LIMDIR MODEL

This section contains the derivation of the predictive posterior probability (up to the normalizing constant) for a new observation under the limdir model. Again, the prediction under the limdir model is conditional requiring that a new observation joins one of the existing cluster. We assume that a variance associated with an observation is known, but differs, that is, $\sigma_i^2$.

From the derivation presented in Section 3.2, conditional on $s$ and $X = (X_1, \ldots, X_n)$, the $p$ cluster locations are independently from a normal distribution as follows,

$$
\gamma_l|X, s \sim N(\mu_l, \tau_l^2) \text{ for } l = 1, \ldots, p,
$$

where $\mu_l = \left( \sum_{j=1}^{n_l} \frac{1}{\sigma_{lj}^2} \right)^{-1} \left( \sum_{j=1}^{n_l} \frac{X_{lj}^*}{\sigma_{lj}^2} \right)$, and $\tau_l^2 = \left( \sum_{j=1}^{n_l} \frac{1}{\sigma_{lj}^2} \right)^{-1}$,

where $*$ represents observations tied to a particular cluster.

To set the notation, let $X'$ a new observation. Assume that $X'$ follows a normal distribution with unknown mean $\theta'$ and known variance $\sigma'^2$. 

104
Given a partition $s$ and the observed data $X$, the probability that $X'$ joins the $l$th cluster is:

$$P(\theta' = \gamma_l | s, X, X') \propto \frac{c_l}{n} \frac{1}{\sqrt{\sigma'^2 + \tau_l^2}} \phi\left( \frac{X' - \mu_l}{\sqrt{\sigma'^2 + \tau_l^2}} \right),$$

where $c_l$ is the size of the $l$th cluster and $\phi$ represents the standard normal probability density function.

Conditional that $X'$ is in the $l$th cluster, the posterior distribution of $\gamma_l$ follows

$$N\left( \frac{\mu_l}{\tau_l^2} + \frac{X'}{\sigma'^2}, \frac{1}{\tau_l^2} + \frac{1}{\sigma'^2} \right).$$

Then by letting $\eta_l = P(\theta' = \gamma_l | s, X, X')$, the predictive estimate for $\theta'$ conditional on $s, X$ and $X'$ is $\mu_l/\tau_l^2 + X'/\sigma'^2$.

Finally, by summing over all the possible partition, we compute the predictive estimate of $\theta'$ given $X$ as follows:

$$\hat{\theta}' | X, X' = \sum_s P(s | X) \sum_{l=1}^{p_S} \eta_l \left( \frac{\mu_l}{\tau_l^2} + \frac{X'}{\sigma'^2} \right).$$
APPENDIX C

PROOF OF THEOREM 5.4.2

This section begins with Lemma C.0.1.

Lemma C.0.1. Suppose that the model used for analysis specifies that \( \{X_{ij}\}_{j=1}^{\infty} \), \( i = 1, 2 \) are mutually independent sequences of \( K \)-dimensional random vectors, where \( X_{ij} \) has density \( f(\cdot|\theta_i) \), with \( \theta_i \in \Theta \) for \( i = 1, 2 \). Assume that Assumptions 5.3.2 (1), (2) and 5.4.1 (1)-(3) hold.

(1) There exist some \( \delta > 0 \) and \( B = \{\theta \in \Theta : E X^t \theta - \psi(\theta) \geq \eta^* - \delta\} \) such that \( \pi(\theta) \) is continuous on \( B \) and \( \min_{\theta \in B} \pi(\theta) > 0 \).

(2) Define \( \bar{X}_i = \sum_{j=1}^{n} X_{ij}/n \) and \( \bar{X} = \sum_{i=1}^{2} \sum_{j=1}^{n} X_{ij}/(2n) \). When \( \theta_1 = \theta_2 \),

(a) \( B_{in} \overset{a.s.}{\rightarrow} B \) for \( i = 1, 2 \) where \( B_{in} = \{\theta \in \Theta : \bar{X}_i^t \theta - \psi(\theta) \geq \eta^* - \delta\} \).

(b) \( B_n \overset{a.s.}{\rightarrow} B \) where \( B_n = \{\theta \in \Theta : \bar{X}^t \theta - \psi(\theta) \geq \eta^* - \delta\} \).

where the distance between \( B_n \) and \( B \) or between \( B_{in} \) and \( B \) is the Lebesgue measure of their symmetric difference.

Proof. Define \( \theta_0 = \theta_1 = \theta_2 \) and let \( E X = \nabla \psi(\theta_0) \). Since \( \pi(\theta) \) is continuous in a neighborhood of \( \theta_0 \) and \( \pi(\theta_0) > 0 \), there exists \( \epsilon \) such that \( \pi(\theta) \) is continuous on
$B' = (\theta_0 - \epsilon, \theta_0 + \epsilon)$ and $\inf_{\theta \in B'} \pi(\theta) > 0$. Let $\delta' = \inf_{\theta \in B'} \{\eta^* - \eta(\theta)\}$. Because of the strict concavity of $\eta(\theta) = E X^t \theta - \psi(\theta)$ on $\Theta$, $\delta' > 0$. Then, for any $0 < \delta < \delta'$, let $B = \{\theta \in \Theta : E X^t \theta - \psi(\theta) \geq \eta^* - \delta\}$. $B \subset B'$. Therefore, statement (1) holds.

Since $X_a.s. \to E X$ and $X_a.s. \to E X$ by the SLLN, $B_{in} a.s. \to B$ and $B_n a.s. \to B$. Therefore, statement (2) holds.

The proof of Theorem 5.4.2 is presented as follows:

Proof. We use the same definitions as in Lemma C.0.1 throughout this proof.

Consider the Bayes factor,

$$\frac{\mathbb{P}(\theta_1 = \theta_2 \mid X_n)}{\mathbb{P}(\theta_1 \neq \theta_2 \mid X_n)} = \frac{\int_{\Theta} \exp\{2n \bar{X}^t \theta - 2n \psi(\theta)\} \pi(\theta) d\theta}{\prod_{i=1}^2 \int_{\Theta} \exp\{n \bar{X}_i^t \theta - n \psi(\theta)\} \pi(\theta) d\theta} \geq \frac{\int_{B_n} \exp\{2n \bar{X}^t \theta - 2n \psi(\theta)\} \pi(\theta) d\theta}{\prod_{i=1}^2 \int_{\Theta} \exp\{n \bar{X}_i^t \theta - n \psi(\theta)\} \pi(\theta) d\theta}. \quad (C.1)$$

First, consider the numerator of Equation C.1. Let $\bar{X} = \nabla \psi(\hat{\theta})$. Taking a Taylor series expansion of $\psi(\theta)$ around $\hat{\theta}$, for any $\theta \in B_n$,

$$2n \bar{X}^t \theta - 2n \psi(\theta) = -2n \psi(\hat{\theta}) + 2n \hat{\theta}^t \bar{X} - n(\theta - \hat{\theta})^t \nabla^2 \psi(\hat{\theta}^*)(\theta - \hat{\theta}), \quad (C.2)$$

where $\hat{\theta}^*$ is on a line segment connecting $\theta$ and $\hat{\theta}$.

Define $\lambda_1^*(B_n) = \max_{\theta \in B_n} \lambda_1(\nabla^2 \psi(\hat{\theta}^*))$ where $\lambda = (\lambda_1, \ldots, \lambda_K)$ denotes the eigenvalues of a $K \times K$ positive definite matrix $\nabla^2 \psi$ with $\lambda_1 \geq \ldots \geq \lambda_K$. Note that $\lambda_1^*(B_n) > 0$. 

107
Therefore,
\[
\int_{B_n} \exp\{2n\bar{X}^t\theta - 2n\psi(\theta)\} \pi(\theta) d\theta \geq L_n \int_{B_n} \exp\{2n\bar{X}^t\theta - 2n\psi(\theta)\} d\theta \\
= L_n \exp\{-2n\psi(\hat{\theta}) + 2n\hat{\theta}^t \bar{X}\} \\
\geq L_n \exp\{-2n\psi(\hat{\theta}) + 2n\hat{\theta}^t \bar{X}\} \\
\geq L_n \exp\{\frac{\pi}{2n\lambda^*_1(B_n)} \frac{1}{2n\lambda^*_1(B_n)} (\text{d}\theta)\}
\]
where \(L_n = \min_{\theta \in B_n} \pi(\theta)\) and \(\phi(\cdot | \mu, \tau^2)\) is a pdf of a normal distribution with mean \(\mu\) and variance \(\tau^2\).

We note that by Lemma C.0.1 the following is true:

(a) \(L_n \stackrel{a.s.}{\to} L\) where \(L = \min_{\theta \in B} \pi(\theta) > 0\).

(b) \(\lambda^*_1(B_n) \stackrel{a.s.}{\to} \lambda_1(B) = \max_{\theta \in B} \lambda_1(\nabla^2 \psi(\hat{\theta}^{**})) > 0\) where \(\hat{\theta}^{**}\) is on a line segment connecting \(\theta\) and \(\theta_0\) for \(\theta \in B\).

(c) \[
\int_{B_n} \prod_{k=1}^{K} \phi(\theta_k | \hat{\theta}_k, \frac{1}{2n\lambda^*_1(B_n)}) d\theta \stackrel{a.s.}{\to} 1.
\]

Now, consider the denominator of Equation C.1,
\[
\int_{\Theta} \exp\{n\bar{X}^t\theta - n\psi(\theta)\} \pi(\theta) d\theta = \frac{\int_{\Theta} \exp\{n\bar{X}^t\theta - n\psi(\theta)\} \pi(\theta) d\theta}{\int_{B_n} \exp\{n\bar{X}^t\theta - n\psi(\theta)\} \pi(\theta) d\theta} \\
\times \int_{B_n} \exp\{n\bar{X}^t\theta - n\psi(\theta)\} \pi(\theta) d\theta \tag{C.4}
\]

Let \(\bar{X}_i = \nabla \psi(\hat{\theta}_i)\). Taking a Taylor series expansion of \(\psi(\theta)\) around \(\hat{\theta}_i\), for any \(\theta \in B_{in}\),
\[
n\bar{X}^t \theta - n\psi(\theta) = -n\psi(\hat{\theta}_i) + n\hat{\theta}_i^t \bar{X}_i - \frac{n}{2} (\theta - \hat{\theta}_i)^t \nabla^2 \psi(\hat{\theta}_i)(\theta - \hat{\theta}_i),
\]
where \( \hat{\theta}_i^* \) is on the line segment connecting \( \theta \) and \( \hat{\theta}_i \).

Define \( \lambda_K^*(B_{in}) = \min_{\theta \in B_{in}} \lambda_K(\nabla^2 \psi(\hat{\theta}_i^*)) \). \( \lambda_K^*(B_{in}) \) is a continuous function of \( \theta \), and \( \nabla^2 \psi \) is positive definite for all \( \theta \in \Theta \). Since \( B_{in} \) is compact, \( \lambda_K^*(B_{in}) > 0 \).

Therefore, similar to Equation C.3,

\[
\int_{B_{in}} \exp\{n\hat{X}_t^i\theta - n\psi(\theta)\} \pi(\theta) d\theta \leq U_{in} \exp\{n\hat{X}_t^i\hat{\theta}_i - n\psi(\hat{\theta}_i)\} \int_{B_{in}} \exp\{-\frac{n}{2} (\theta - \hat{\theta}_i)^t \nabla^2 \psi(\hat{\theta}_i)(\theta - \hat{\theta}_i)\} d\theta
\]

\[
\leq U_{in} \exp\{n\hat{X}_t^i\hat{\theta}_i - n\psi(\hat{\theta}_i)\} \left( \frac{2\pi}{n\lambda_K^*(B_{in})} \right)^{K/2} \int_{B_{in}} \prod_{k=1}^K \phi(\theta_k | \hat{\theta}_{ik}, \frac{1}{n\lambda_K^*(B_{in})}) d\theta
\]

\[
\leq U_{in} \exp\{n\hat{X}_t^i\hat{\theta}_i - n\psi(\hat{\theta}_i)\} \left( \frac{2\pi}{n\lambda_K^*(B_{in})} \right)^{K/2}, \quad \text{(C.5)}
\]

where \( U_{in} = \max_{\theta \in B_{in}} \pi(\theta) \). The second inequality follows from the non-negative definiteness of \( \nabla^2 \psi(\theta) - \lambda_K^*(B_{in})I \) for all \( \theta \in B_{in} \).

Also,

\[
\frac{\int_{\Theta} \exp\{n\hat{X}_t^i\theta - n\psi(\theta)\} \pi(\theta) d\theta}{\int_{B_{in}} \exp\{n\hat{X}_t^i\theta - n\psi(\theta)\} \pi(\theta) d\theta} = 1 + \frac{\int_{B_{in}} \exp\{n\hat{X}_t^i\theta - n\psi(\theta)\} \pi(\theta) d\theta}{\int_{B_{in}} \exp\{n\hat{X}_t^i\theta - n\psi(\theta)\} \pi(\theta) d\theta}
\]

\[
\leq 1 + \frac{\exp[n\sup_{\theta \in B_{in}} \{\hat{X}_t^i\theta - \psi(\theta)\}] P(\theta \in B_{in}^c)}{\exp[n\inf_{\theta \in B_{in}} \{\hat{X}_t^i\theta - \psi(\theta)\}] P(\theta \in B_{in})}
\]

\[
= \frac{1}{P(\theta \in B_{in})}, \quad \text{(C.6)}
\]

By taking Equations C.5 and C.6, Equation C.4 is bounded above by

\[
\frac{U_{in}}{P(\theta \in B_{in})} \left( \frac{2\pi}{n\lambda_K^*(B_{in})} \right)^{K/2} \exp\{n\hat{X}_t^i\hat{\theta}_i - n\psi(\hat{\theta}_i)\}. \quad \text{(C.7)}
\]

We note that the following is true:

(a) \( U_{in} \xrightarrow{a.s.} U \) where \( U = \max_{\theta \in B} \pi(\theta) \).
(b) \( \lambda_K(B_m) \xrightarrow{a.s.} \lambda_K^*(B) = \min_{\theta \in B} \lambda_K(\nabla^2 \psi(\hat{\theta}^*_i)) > 0 \) where \( \hat{\theta}^*_i \) is on a line segment connecting \( \theta \) and \( \theta_0 \) for \( \theta \in B \).

(c) \( 1/P(\theta \in B_m) \xrightarrow{a.s.} 1/P(\theta \in B) \).

Let
\[
R_n = \frac{L_n P(\theta \in B_1 n) P(\theta \in B_2 n)}{U_1 n U_2 n} \left\{ \frac{n \lambda_K(B_1 n) \lambda_K(B_2 n)}{4 \pi \lambda(1 n)} \right\}^{K/2} \exp\left\{ n \left\{ 2 \tilde{X}^t \hat{\theta} - \tilde{X}_1^t \hat{\theta}_1 - \tilde{X}_2^t \hat{\theta}_2 + \psi(\hat{\theta}_1) + \psi(\hat{\theta}_2) - 2 \psi(\hat{\theta}) \right\} \right\},
\]
and then \( R_n \) is the lower bound of Equation C.1 almost surely.

Let \( \zeta = (\nabla \psi)^{-1} \) and \( \tilde{X} = (\tilde{X}_1, \tilde{X}_2)^t \). Define \( H_n(\tilde{X}) \) as follows;
\[
H_n(\tilde{X}) = 2 \tilde{X}^t \zeta(\tilde{X}) - \tilde{X}_1^t \zeta(\tilde{X}_1) - \tilde{X}_2^t \zeta(\tilde{X}_2) + \psi(\tilde{X}_1) + \psi(\tilde{X}_2) - 2 \psi(\tilde{X}) = 2 \tilde{X}^t \zeta(\tilde{X}) - \tilde{X}_1^t \zeta(\tilde{X}_1) - \tilde{X}_2^t \zeta(\tilde{X}_2) + \psi(\tilde{X}_1) + \psi(\tilde{X}_2) - 2 \psi(\tilde{X}).
\]

Then,
\[
R_n = \frac{L_n P(\theta \in B_1 n) P(\theta \in B_2 n)}{U_1 n U_2 n} \left\{ \frac{n \lambda_K(B_1 n) \lambda_K(B_2 n)}{4 \pi \lambda(1 n)} \right\}^{K/2} \exp\left\{ n H_n(\tilde{X}) \right\}.
\]
Taking a Taylor’s expansion of \( H_n(\tilde{X}) \) around \( \omega_0^* = (\omega_0^1, \omega_0^2)^t = (\nabla \psi(\theta_0)^t, \nabla \psi(\theta_0)^t)^t \),
\[
n H_n(\tilde{X}) = \frac{n}{2} (\tilde{X} - \omega_0^*)^t \nabla^2 H_n(\omega^*)(\tilde{X} - \omega_0^*),
\]
where \( \omega^* = (\omega_1^*, \omega_2^*)^t \) is on a line segment connecting \( \tilde{X} \) and \( \omega_0^* \).

We note that
\[
\nabla^2 H_n(\omega^*) = \begin{pmatrix} W_{11} - 2V_1 & W_{12} \\ W_{21} & W_{22} - 2V_2 \end{pmatrix},
\]
where
\[
[W_{iv}]_{jj'} = \frac{1}{2} \frac{\partial \zeta_j(\omega^*)}{\partial X_{i'j'}}, [V_i]_{jj'} = \frac{1}{2} \frac{\partial \psi(\omega^*)}{\partial X_{i'j'}}, \text{ and }
\]
\[
\omega^* = (\omega_1^* + \omega_2^*)/2.
\]
We observe that by the SLLN,

\[ W_{ii'} \text{ and } V_i \xrightarrow{a.s.} \Lambda \text{ where } [\Lambda]_{ij'} = \frac{1}{2} \frac{\partial \zeta_j(\omega_0)}{\partial X_{ij'}}, \]

and by the CLT,

\[ \sqrt{n}(\bar{X}_i - \omega_0) \xrightarrow{D} Z_i \text{ where } Z_i \sim N_K(0, \text{Var}(X_i)) \text{ for } i = 1, 2. \]

Finally, by Slutsky’s theorem, \( nH_n(\bar{X}) \) converges in distribution to the distribution of

\[ -\frac{1}{2}(Z_1 - Z_2)'\Lambda(Z_1 - Z_2). \]

In turn, \( n^{-K/2}R_n \) converges in distribution to the distribution of

\[ \frac{LP(\theta \in B)^2}{U^2} \left\{ \frac{\lambda_K(B)^2}{4\pi\lambda_1(B)} \right\}^{K/2} \exp\left\{ -\frac{1}{2}(Z_1 - Z_2)'\Lambda(Z_1 - Z_2) \right\}, \]

which is a real-valued random variable.

This implies that for any fixed \( M > 0 \),

\[ \lim_{n \to \infty} P(R_n > M) = 1. \]

Since \( R_n \) is the lower bound of Equation C.1,

\[ \lim_{n \to \infty} P\left( \frac{P(\theta_1 = \theta_2 | X_n)}{P(\theta_1 \neq \theta_2 | X_n)} > M \right) = 1. \]
Bibliography


