Bayesian Methods for Data-Dependent Priors

Dissertation

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

By

William Francis Darnieder, M.S.

Graduate Program in Statistics

The Ohio State University

2011

Dissertation Committee:

Steven N. MacEachern, Advisor

Catherine A. Calder

Mario Peruggia
Abstract

The use of data-dependent priors is strangely ubiquitous in Bayesian statistical modeling given the alarm it raises within some cadres. Indeed, the researcher who employs a data-dependent prior faces the inescapable truth that he or she has used the data twice: the first time in establishing prior belief, and the second in updating the prior with the likelihood to obtain the posterior distribution. In this dissertation, we introduce the Adjusted Data-Dependent Bayesian Paradigm as a principled approach to using data-dependent priors in weak accordance with Bayes’ theorem.

Suppose that a researcher peeks at the data through some summary statistic prior to applying the Bayesian update. This novel method systematically chooses the posterior distribution that captures the information regarding model parameters that is contained in the data above and beyond the information contained in the observed statistic. In special situations the adjusted approach is formally equivalent to other Bayesian methods, but these cases are necessarily rare. The adjusted procedure imposes a null update when the observed statistic is sufficient, choosing the posterior distribution that is equal to the prior. Conversely, observing a non-sufficient statistic will invite a non-trivial update.

Of particular interest is how analyses under the adjusted and naive (unadjusted) procedures compare. Implementation strategies are described for imposing the adjustment in low and high dimensional settings. Posterior simulation is emphasized
using Markov Chain Monte Carlo (MCMC) techniques modified to accommodate the adjusted paradigm. Black box strategies (e.g. preprocessing data) used to fit Dirichlet Process (DP) mixture models are cast as data-dependent and we demonstrate how to apply the adjustment in these settings. Additionally, we compare the predictive power of the naive and adjusted techniques in analyzing the classic galaxies data set popularized in Roeder (1990).
To my mother, Mary,
and in memory of my father,
Donald (1932-2009).
Acknowledgments

Throughout my life I have been fortunate to have had excellent teachers at every stage of my formal education. A few require special mention here because of the impact they have had on my academic and professional arcs. Mr. Edward Mooney planted the first seeds of enthusiasm in mathematical things when I was in the eighth grade at Wilbur Wright Middle School and Mr. Joseph Griesbach continued to nourish my fascination with mathematics at Marquette University High School. It was because of their example and encouragement that I first pursued the mathematics major in college and later began a high school teaching career.

This dissertation is the summit of my academic training. My advisor, Professor MacEachern, has been the best mentor during this time that I could have imagined. I am thankful that he welcomed me as his student, and for the many long discussions regarding this research that would so frequently swirl into the abyss, outside my range of comprehension. He has had a profound influence on my “world view” and, in my professional life, I hope to approximate his model of scientific inquiry as well as I can.

Professor Calder has been very generous in support of my studies while in the Department of Statistics. I have learned a great deal from my research assistantship in the Criminal Justice Research Center under her guidance. I was first introduced to Bayesian statistics by Professor Peruggia who was influential in my desire to embrace the area as my research domain. I thank both Professors Calder and Peruggia for
their membership on my dissertation committee and the helpful discussions we have
had concerning this work.

Graduate school is not easy, and my success would not have been possible without
the love and support of my family. My wife, Melisa, and my young daughters, Emily
and Julia, have brought me so much joy during this time in what is otherwise an
endeavor filled with uncertainty.
Vita

May 12, 1979 ............... Born - Milwaukee, WI

2001 ................................ B.S. Mathematics, *summa cum laude*,
Case Western Reserve University

2003 ................................ M.S. Mathematics,
The Ohio State University

2003 - 2006 ........................ Upper School Mathematics Teacher,
The Wellington School,
Upper Arlington, OH

2008 ................................ M.S. Statistics,
The Ohio State University

2008 - present ....................... Graduate Research Associate,
Criminal Justice Research Center,
The Ohio State University

Fields of Study

Major Field: Statistics
# Table of Contents

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

Dedication . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . iv

Acknowledgments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . v

Vita . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . vii

List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xi

List of Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xii

1. Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

   1.1 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

   1.2 Data-dependent Priors . . . . . . . . . . . . . . . . . . . . . . . . . 3

      1.2.1 Explicitly Data-dependent Priors . . . . . . . . . . . . . . . 4

      1.2.2 Jeffreys’ Prior . . . . . . . . . . . . . . . . . . . . . . . . . . 7

      1.2.3 Zellner’s g-prior . . . . . . . . . . . . . . . . . . . . . . . . . 8

      1.2.4 Wasserman’s Prior . . . . . . . . . . . . . . . . . . . . . . . . 10

      1.2.5 Raftery’s Prior . . . . . . . . . . . . . . . . . . . . . . . . . . 11

   1.3 Motivation for an Alternative Approach to Data-based Prior Selection 12

      1.3.1 Criticism of Data-dependent Techniques . . . . . . . . . . . . 13

      1.3.2 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

2. Background Material . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

   2.1 Empirical Bayes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

   2.2 Importance Sampling . . . . . . . . . . . . . . . . . . . . . . . . . . 18

      2.2.1 Importance Link Functions . . . . . . . . . . . . . . . . . . . 19

   2.3 Metropolis-Hastings Algorithm . . . . . . . . . . . . . . . . . . . . 19
# Table of Contents

2.4 Gibbs Sampling .................................................. 20

3. The Adjusted Data-dependent Bayesian Paradigm .................. 22

3.1 Regularity Conditions ........................................... 22
3.2 Adjusted Data-dependent Bayes ................................... 23
  3.2.1 The Data-Dependent Bayesian Principle ..................... 26
3.3 Properties .......................................................... 27
  3.3.1 Exploiting Conditional Independence in a Hierarchical Model 30
3.4 Equivalent Analyses ............................................... 31
  3.4.1 Equivalence of Naive and Adjusted Data-dependent Procedures 31
  3.4.2 Equivalence of Adjusted and Bayes Procedures ............... 34
  3.4.3 Equivalent Prior Information ................................. 35
3.5 Coherence .......................................................... 36
  3.5.1 Model Coherence ............................................... 36
  3.5.2 Two Researchers, Two Data Sets ............................. 37
  3.5.3 Remark .......................................................... 39

4. Implementation ...................................................... 40

4.1 Analytically Determined Adjustment ............................. 40
  4.1.1 Normal-Normal Compound Decision Problem .................. 40
  4.1.2 Normal Model With Median .................................... 45
4.2 A General MCMC Strategy ........................................ 46
4.3 Low Dimensional Computational Strategy .......................... 47
4.4 High Dimensional Computational Strategy - Importance Sampling 48
4.5 Application .......................................................... 49
  4.5.1 Baseball Batting Averages .................................... 50

5. Adjusted Data-Dependent Bayes in Mixture Modeling ............. 58

5.1 The Null and Alternative Mixture Model .......................... 58
  5.1.1 Breast Cancer Study .......................................... 64
5.2 Dirichlet Process Mixture Models ................................ 68
  5.2.1 Data-Dependent Gaussian Method ............................. 68
  5.2.2 Galaxies Data .................................................. 79
  5.2.3 Use of Order Statistics ....................................... 82

6. Conclusion .......................................................... 96

6.1 Opportunities for Abuse .......................................... 96
6.2 Problems in Data Analysis ....................................... 98
6.3 Black Box Methods ............................................... 99
Bibliography
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Comparison of marginal posterior batting average estimates under the naive and adjusted techniques with rest-of-season batting averages.</td>
<td>57</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>55</td>
</tr>
<tr>
<td>5.1</td>
<td>60</td>
</tr>
<tr>
<td>5.2</td>
<td>65</td>
</tr>
<tr>
<td>5.3</td>
<td>69</td>
</tr>
<tr>
<td>5.4</td>
<td>70</td>
</tr>
<tr>
<td>5.5</td>
<td>71</td>
</tr>
<tr>
<td>5.6</td>
<td>72</td>
</tr>
<tr>
<td>5.7</td>
<td>73</td>
</tr>
<tr>
<td>5.8</td>
<td>74</td>
</tr>
<tr>
<td>5.9</td>
<td>75</td>
</tr>
<tr>
<td>5.10</td>
<td>83</td>
</tr>
</tbody>
</table>
5.11 Naive (red) and adjusted (blue) posterior predictive densities with prior variance inflated by a factor of 4.

5.12 Naive (red) and adjusted (blue) posterior predictive densities with prior variance not inflated.

5.13 Histograms of the difference in predictive log likelihoods (naive minus adjusted) using $X_{(19)}$ and $X_{(56)}$ to generate the data-dependent prior.

5.14 Histograms of the difference in predictive log likelihoods (naive minus adjusted) using $X_{(8)}$ and $X_{(67)}$ to generate the data-dependent prior.

5.15 Histograms of the difference in predictive log likelihoods (naive minus adjusted) using $X_{(1)}$ and $X_{(74)}$ to generate the data-dependent prior.

5.16 Boxplots of the difference of predictive log likelihoods under the naive and adjusted analyses. Each boxplot is labeled by statistic used in forming the data-dependent prior, as well as the size of the hold-out sample.

5.17 Plots of adjusted predictive log likelihood against naive predictive log likelihood using $X_{(19)}$ and $X_{(56)}$ to generate the data-dependent prior.

5.18 Plots of adjusted predictive log likelihood against naive predictive log likelihood using $X_{(8)}$ and $X_{(67)}$ to generate the data-dependent prior.

5.19 Plots of adjusted predictive log likelihood against naive predictive log likelihood using $X_{(1)}$ and $X_{(74)}$ to generate the data-dependent prior.
Chapter 1: Introduction

1.1 Overview

At its root, Bayesian statistics departs from classical statistics in the meaning and operationalization of parameters. To the classicist, most parameters are fixed quantities and inference is aimed at divining some useful statement about them from the data, e.g. point estimate, interval estimate, or hypothesis test. Bayesian statisticians take an alternative viewpoint, that parameters can be described themselves as random variables. If \( \theta \) is the vector of parameters, now viewed as random variables, and \( X \) are the data, then Bayes’ Theorem is the mathematical device that allows for meaningful inference. The posterior density,

\[
p(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)},
\]

fuses prior knowledge about \( \theta \) contained in \( \pi(\theta) \) with the information contained in the data sample. The Bayesian statistician chooses the form of \( \pi(\theta) \) before the data has been collected, but may be influenced by the results of previous experiments, expert opinion, constraints on the parameter space, and other factors in the selection of that form.

Either by desire or necessity, a common, fundamental breach of the Bayesian procedure is to use information from the collected data in informing both the prior and
the posterior densities. Use of the data in the prior modeling leads to a density called a data-dependent prior. The use of data-dependent priors occupies a controversial milieu that draws from both Bayesian and classical paradigms. The central aim of this dissertation is to carve out a space for the use of data-dependent priors in Bayesian modeling that aligns more closely with existing Bayesian procedures. We call this new, more principled approach the Adjusted Data-dependent Bayesian Paradigm. While this new approach does not legitimize the use of data-dependent priors as a purely Bayesian procedure, the adjusted paradigm is decidedly more Bayesian.

The dissertation is organized into six chapters. The remainder of this chapter will formally introduce the reader to data-dependent priors, provide examples of their use, and motivate the need for an improved procedure through pointed criticism of current practice. Chapter 2 reviews statistical literature that underlies many of the existing techniques from which we shall draw. We will repeatedly make use of methods for posterior simulation, so in particular we review the Monte Carlo techniques of importance sampling, the Metropolis-Hastings algorithm, and Gibbs sampling as it pertains to the material contained in this dissertation. Chapter 3 states and justifies the data-dependent Bayesian paradigm, describes several key properties, and sets this new work in the context of existing Bayesian procedures. Chapter 4 describes methods for implementation in low and high dimensional settings. Example analyses illustrate how the adjusted data-dependent procedure might be implemented. Also, we compare the adjusted procedure to results obtained using the current practice, which we label “naive”. Chapter 5 is specifically devoted to application of the adjusted paradigm in the context of mixture modeling. Data-dependent priors are naturally used in mixture modeling, and we highlight the dangers of their current use and propose alternatives
that are more amenable to the adjusted viewpoint. In Chapter 6, we conclude with
a caution to potential users of the adjusted procedure. We also capture a glimpse
of the potentially vast research opportunities this new data-dependent approach may
foster.

1.2 Data-dependent Priors

As a beginning to our study of data-dependent Bayesian methods, it would be wise
to settle on an operating definition of what “data-based prior distributions” are and
how we shall distinguish their various incarnations. For our purposes, data-based
prior distributions will encompass all prior distributions obtained through explicit
consideration of the data generating mechanism or from the realized data themselves.
This definition is broad enough to include all the forms of data-based priors we wish
to study and, when necessary, the primary manifestation of the data-dependence will
be stipulated.

This section aims to introduce the reader to data-based priors through an expo-
sition of several illustrative examples. Although specific examples are described, the
features which generalize will be highlighted. In so doing, we will have explored many
of the most salient contexts in which data-dependent priors arise. As we shall see,
the degree to which these priors depend on the data or the data-generating process
varies hugely across models and methods for their construction.

Wasserman (2000) defines a data-dependent prior as a measurable mapping from
the data space to the set of priors, i.e. $\pi_n: \mathcal{Y}^n \rightarrow \mathcal{P}$. Colloquially, a data-dependent
prior is simply a prior distribution that depends on the data. In what follows there
should be no danger in this second intuitive definition, but we will emphasize any change in this operating definition if needed.

1.2.1 Explicitly Data-dependent Priors

Use of the term *explicitly data-dependent prior* shall be restricted to those prior distributions constructed through overt use of the realized data. In most applications, especially if the prior has a closed functional form, some artifact of the realized data $X$ will be apparent, perhaps only through some summary of the data $T(X)$. An alternative formulation of the resulting prior may, on its face, completely mask the data, but is in fact highly dependent on the data.

Data-analytic Priors

Data-analytic priors are perhaps the most ubiquitous of the data-dependent priors. If little prior information is known regarding the distribution of some modeling parameters, the analyst may be tempted to use the data in forging “prior” beliefs. The choice of using a data-analytic prior may be especially unpalatable to some researchers, but may be necessitated for computational reasons. This may happen when no good objective prior options exist or when vague priors exhibit poor behavior. This is the viewpoint held by Berger (2006) who dismisses vague priors as useful only as approximations to good objective priors. On the other hand, improper objective priors may induce an improper posterior under the probability calculus, a phenomenon we address with respect to mixture modeling.

We shall regard such data-analytic priors as being techniques falling in the domain of the empirical Bayes paradigm. In this vein, we present two examples that would
be considered empirical Bayes and include an additional example that sprawls into Bayesian nonparametrics.

**Compound Decision Problem**

Consider the normal-normal compound decision problem which has been studied extensively in the literature. A few resources on the matter include Morris (1983), Berger (1985), and Carlin and Gelfand (1990). For simplicity, we consider the known variances case. The model is as follows:

\[
X_i | \theta_i \sim \mathcal{N}(\theta_i, \sigma^2), \text{ independently} \tag{1.2}
\]

\[
\theta_i | \eta \sim \mathcal{N}(\eta, \tau^2), \text{ independently} \tag{1.3}
\]

\[
i \in \{1, 2, \ldots, n\} \tag{1.4}
\]

The empirical Bayesian views the hyper-parameter \( \eta \) as a fixed real number. Marginalizing over the \( \theta_i \), we readily see that the \( X_i \) are independent and identically distributed \( \mathcal{N}(\eta, \sigma^2 + \tau^2) \). Thus, we may estimate \( \eta \) using the MLE \( \hat{\eta} = \bar{X} \) and proceed with inference on each of the \( \theta_i \) by using Bayes’ Theorem with \( \eta \) replaced with its estimator \( \hat{\eta} \).

The data-analytic procedure described above suffers from at least one obvious problem: the uncertainty in estimating \( \eta \) has not been taken into account. Many authors have devised methods to account for this uncertainty. Morris (1983) develops a definition of empirical Bayes confidence intervals, and so the reconciliation for having used \( \hat{\eta} \) as a proxy for \( \eta \) is accomplished through a desire to obtain the correct EB coverage. This philosophy is further propounded by Carlin and Gelfand (1990), while bootstrapping methods are developed in Laird and Louis (1987).
In this spirit, we might take the more liberal view that $\eta$ comes from a continuous population and that its prior distribution can be approximated by the data-dependent prior

$$\eta|\bar{X} \sim \mathcal{N}\left(\bar{X}, \frac{\sigma^2 + \tau^2}{n}\right).$$  \hspace{1cm} (1.5)

Written as a density, this is exactly the result of having switched $\bar{X}$ and $\eta$ in the sampling distribution of $\bar{X}|\eta$, a technique which appears in Carlin and Louis (2000). Inferences on the parameters are then made via the approximate posterior distribution

$$p(\theta, \eta|x) \propto \prod_{i=1}^{n} f(x_i|\theta_i) \prod_{i=1}^{n} g(\theta_i|\eta)\pi(\eta|\bar{x}).$$  \hspace{1cm} (1.6)

Another Hierarchical Model

In forecasting mortality projections, Czado et al. (2005) construct a somewhat complex hierarchical Poisson log-bilinear model that estimates each of 7 hyperparameters using maximum likelihood estimation. In this example, the hyperparameters are each three stages away from the actual data stage, one degree further removed than in the compound decision problem of this section. In the researchers’ analysis, no correction for having estimated the parameters is attempted.

Bayesian Nonparametrics

In the Dirchlet process mixture model which we specify as

$$X_i|\theta_i \sim f(x_i|\theta_i)$$  \hspace{1cm} (1.7)

$$\theta_i|G \sim G(\theta_i)$$  \hspace{1cm} (1.8)

$$G|G_0, M \sim DP(M, G_0),$$  \hspace{1cm} (1.9)

the base measure $G_0$ and precision $M$ are frequently chosen in a data analytic way. This is perhaps most common when $G_0$ comes from a parametric family, e.g. $\mathcal{N}(\mu, \sigma^2)$.
and point estimates are obtained for $\mu$ and $\sigma^2$ via maximum likelihood estimation, or through some other point estimation procedure. This implies a data-analytic choice for $G_0$ for this model. MacEachern (1998) analyzes the baseball data of Efron and Morris using a mixture of DP model with a data-analytic choice for $G_0$ from the beta family. McAuliffe et al. (2006) introduce an adaptive empirical Bayes strategy for estimating $G_0$.

1.2.2 Jeffreys’ Prior

We turn now to Jeffreys’ prior, popular in the objective Bayesian domain. As such, it is perhaps surprising that Jeffreys’ prior may be considered “data-based” at all. In the univariate case the Jeffreys approach selects

$$\pi(\theta) = \sqrt{I(\theta)}$$

(1.10)
as the prior density for $\theta$ where $I(\theta)$ is the Fisher information in a single observation.

When $\theta$ is a parameter vector, Jeffreys’ choice becomes

$$\pi(\theta) = \sqrt{|I(\theta)|},$$

(1.11)

where $|I(\theta)|$ is the determinant of the Fisher information matrix (Jeffreys, 1961). The entries of $I(\theta)$ are given by

$$I_{i,j}(\theta) = E \left[ \left( \frac{\partial}{\partial \theta_i} \log f(x|\theta) \right) \left( \frac{\partial}{\partial \theta_j} \log f(x|\theta) \right) \right].$$

(1.12)

What makes this a data-based approach to selecting a prior distribution? It is true that in many important examples, the Jeffreys prior is not explicitly data-dependent as the methods of empirical Bayes are. For example, let $X \sim \mathcal{N}(\mu, \sigma^2)$, then the Jeffreys prior is the improper distribution $\pi(\mu, \sigma^2) \propto \sigma^{-2}$. However, one cannot
discount that Jeffreys deliberately exploits the structure of the data likelihood to arrive at his noninformative prior. In other models, the Jeffreys prior preserves some feature of the data itself as we see next.

Let us consider the autoregressive model of order 1, which we abbreviate as AR(1). Using a formulation similar to Phillips (1991) we have

\[ y_t = \rho y_{t-1} + \epsilon_t \]  
\[ \epsilon_t \sim \mathcal{N}(0, \sigma^2), \text{iid} \]  
\[ t \in \{1, 2, \ldots, T\} \]

The resulting Fisher information matrix \( I(\rho, \sigma) \) is diagonal with entries \( I_{\rho\rho} \) and \( I_{\sigma\sigma} \). The specific expressions are not important, except to remark that \( |I(\rho, \sigma)| = I_{\rho\rho}I_{\sigma\sigma} \) depends explicitly on the initialization \( y_0 \) of the sequence and on the size of the sequence \( T \).

Although this example demonstrates how Jeffreys prior may preserve some feature of the model, the extent to which this happens must actually be quite minimal. By virtue of its construction which effectively involves integrating out the data, the Jeffreys approach lies outside our characterization of those methods which are explicitly data-dependent. However, its formulation is rooted in properties of the data likelihood, and thus its classification as a data-based prior is justified.

### 1.2.3 Zellner’s g-prior

Here we consider the natural conjugate prior in the normal linear multiple regression model written as:

\[ Y = X\beta + \epsilon \]  

\[ \beta \sim \mathcal{N}(0, \text{diag}(\sigma^2_{\text{g}})) \]

\[ \sigma^2_{\text{g}} \sim \text{Gamma}(a, b) \]
where \( Y \) is an \( n \times 1 \) observation vector, \( X \) is a given \( n \times k \) matrix of covariates having rank \( k \), \( \beta \) is a \( k \times 1 \) parameter vector, and \( \epsilon \) is a vector of iid \( \mathcal{N}(0, \sigma^2) \) random errors with \( \sigma^2 \) unknown.

In the notation of Zellner (1986) the natural conjugate prior distribution for \((\beta, \sigma)\) takes the form

\[
\pi(\beta, \sigma) = \pi(\beta|\sigma)\pi(\sigma) 
\]

(1.17)

\[
\pi(\beta|\sigma) = \sigma^{-k} \exp\left\{-(\beta - \bar{\beta})' A (\beta - \bar{\beta})/(2\sigma^2)\right\} 
\]

(1.18)

\[
\pi(\sigma) \propto \sigma^{-(\nu_0+1)} \exp\left\{-\nu_0 s_0^2/(2\sigma^2)\right\} 
\]

(1.19)

In words, \( \beta|\sigma \) has a multivariate normal pdf with prior mean \( \bar{\beta} \) and covariance matrix \( A^{-1}\sigma^2 \), and \( \sigma \) has an inverted gamma prior. Driven by theory involving so-called “rational expectations” due to Muth (1961) leads to this general form of the g-prior, but with the following additional specifications:

\[
A = gX'X 
\]

(1.20)

\[
\nu_0 s_0^2 = (n - k - 2)\sigma_0^2 
\]

(1.21)

\[
\nu_0 = n - k. 
\]

(1.22)

where \( g, \sigma_0^2 > 0 \). An important special instance of the g-prior is to allow \( \sigma \) to have the improper prior distribution \( \pi(\sigma) \propto \sigma^{-1} \). With this formulation, the joint posterior has a particularly nice form as it is just \( p(\beta, \sigma|y) \propto \sigma^{-(n+k+1)} \exp\left\{-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) + g(\beta - \bar{\beta})'X'X(\beta - \bar{\beta})\right\} \). The posterior mean and conditional variance take the elegant forms:

\[
E[\beta|y] = \frac{(X'X)^{-1}X'y + g\beta_0}{1 + g}, \quad (1.23)
\]

\[
V[\beta|\sigma, y] = \frac{(X'X)^{-1}\sigma^2}{1 + g}. \quad (1.24)
\]
Here we easily see how the parameter $g$ influences posterior inference: large $g$ favors the prior mean, while small $g$ favors the least squares estimate.

Like Jeffreys’ prior, the $g$-prior does not explicitly use the realized observations in its form. Yet the degree to which Zellner’s prior is data-based seems a bit stronger than Jeffreys in the sense that the actual values of the covariate matrix $X$ are used. It is possible that the values of the covariates will inform prior understanding of observables moreso than knowledge of the sample size alone.

1.2.4 Wasserman’s Prior

Let us now turn to an important example of a data dependent prior in finite mixture modeling. Let the mixture model be of the form:

$$ f(x|p, \mu, \sigma) = \sum_{j=1}^{k} p_j \mathcal{N}(x|\mu_j, \sigma_j^2). \quad (1.25) $$

where $p_j > 0$ for each $j$ and $\sum_{j=1}^{k} p_j = 1$. Suppose that the mixture probabilities are known and we desire an objective approach to modeling the prior on $(\mu, \sigma)$. An improper prior would be a natural choice, for example Jeffreys’ prior or a flat prior. However, improper priors lead to improper posteriors for the finite mixture model (Wasserman, 2000). Through the introduction of a vector of latent group membership variables $G = (G_1, G_2, \ldots, G_n)$, we re-express the model as:

$$ P(G_i = j) = p_j \quad (1.26) $$

$$ Y_i|G_i, \mu, \sigma \sim \mathcal{N}(\mu_j, \sigma_j^2). \quad (1.27) $$

Let $G^*$ be the set of all group membership vectors placing at least two observations in each cluster, and set $\Delta_n = \sum_{G \in G^*} \mathcal{L}_G$ where $\mathcal{L}_G$ is the likelihood conditional on the group membership vector, Wasserman’s data-dependent prior (Wasserman, 2000).
is then

\[ \pi_n(\mu, \sigma) = \pi(\mu, \sigma) \left(1 - \frac{\Delta_n}{L_n}\right), \]  

(1.28)

where \( L_n \) is the likelihood function and \( \pi(\mu, \sigma) \) is any improper prior.

Wasserman’s prior formally coincides with the the likelihood tampering technique of Diebolt and Robert (1994) who base inference on the posterior obtained by replacing the likelihood \( L_n \) with the pseudo-likelihood \( L_n - \Delta_n \). Implementation within a Gibbs sampler is easy as we can simply discard the samples with prohibited group membership vector. Much more shall be said about Wasserman’s prior in Chapter 5 with respect to the simplest of mixture models, as it serves as an excellent illustrative example. Although we have used a notation that suppresses the influence of the data, Wasserman’s prior is explicitly data-dependent, and so is more akin to data-analytic priors than the examples of Jeffreys and Zellner.

1.2.5 Raftery’s Prior

Also working with the finite mixture of normal densities from above, Raftery (1996) introduces a data-dependent prior based on the range of the data and a few summary statistics. Specifically,

\[ \sigma_j \sim IG(\omega_j/2, \lambda_j/2) \]  

(1.29)

\[ \mu_j | \sigma_j \sim N(\xi_j, \sigma_j/\tau_j). \]  

(1.30)

Additionally, Raftery models the vector \( p \) with a Dirichlet prior. In the model formulation, he sets \( \omega_j = 2.56, \lambda_j = 0.72 \bar{\text{Var}}(y), \xi_j = \bar{y} \) and, \( \tau_j = 2.6 / (y_{\text{max}} - y_{\text{min}})^2 \) for each \( j \). Raftery calls this choice a reference proper prior for Bayes factors. The intent is to have a relatively flat prior for \( \mu_j \) over the range of the data. Simulation studies of Raftery and Wasserman’s approaches demonstrate comparable posterior
distributions for the modeling parameters. Richardson and Green (1997) also use a data-dependent prior based exclusively on the range of the data and which exhibits behavior similar to the approaches of Raftery and Wasserman.

1.3 Motivation for an Alternative Approach to Data-based Prior Selection

Let us consider what is entailed in a typical Bayesian analysis that utilizes a data-based prior distribution. Let \( X = (X_1, X_2, \ldots, X_n) \) be data arising from a distribution with joint probability density \( f(x|\theta) \) with \( \theta = (\theta_1, \theta_2, \ldots, \theta_p) \) representing the vector of parameters. Also, let \( T(X) = (T_1(X), T_2(X), \ldots, T_k(X)) \) denote a vector of statistics. For now we do not assume any special structure to the model and require only that the numbers \( k, n, \) and \( p \) be positive. Lastly, we represent the (data-dependent) joint prior distribution on \( \theta \) as having density \( \pi(\theta|T(X)) \).

If a fully Bayesian approach were to be undertaken, we would have access to a joint prior distribution, denoted by \( \pi(\theta) \), that encapsulates all prior non-experimental information regarding \( \theta \). The Bayesian probability calculus then admits a posterior distribution for \( \theta \) that synthesizes prior knowledge in the form of \( \pi \) with the realized data which we write as \( p(\theta|x) \propto f(x|\theta)\pi(\theta) \). That is, the posterior density for the parameters is proportional to the product of the data likelihood and the prior density.

The description of the analysis performed in the preceding paragraph represents Bayesian thinking and application in its purest form. In real scenarios, very little information may be known about certain parameters, their dependence, and moreover, their very inclusion may be suspect. The choice of data model and prior will only approximate reality, regardless of how finely detailed the settled upon joint probability model actually is. In the event that a proper subjective prior \( \pi(\theta) \) cannot be elicited
from expert opinion or previous experiments, it may be convenient to approximate \( \pi(\theta) \) using \( \pi(\theta|T(x)) \), which is based on a peek at the data through the statistic \( T(x) \). We have already presented several examples of data-based priors together with some motivation for their use, but have deferred criticism until now.

Our focus on formulation of \( \pi(\theta|T(x)) \) as an approximation to a genuine prior density boasts one important feature deserving attention, namely it will actually coincide with some prior density \( \pi(\theta) \). Thus, use of a data-based prior satisfies at least the most basic requirements to constitute a “Bayesian” analysis. To a Bayesian who strictly adheres to the pure methodology, this is likely the point where praise ends and criticism begins.

### 1.3.1 Criticism of Data-dependent Techniques

Curiously, the data-dependent strategy attempts to circumvent the often difficult task of establishing prior belief while still masquerading as a Bayesian technique. In so doing, the data-dependent Bayesian will incorporate aspects of the observed data into his assessment of prior information before experimentation (or at least before analysis), and in the formal posterior updating in accordance with Bayes Theorem. This double use of the data within the Bayesian paradigm is highly objectionable because it can easily be abused to produce virtually any outcome the researcher desires (Gelman et al., 2003). This is true even when the researcher is not performing a deliberately deceptive analysis, but when the researcher may unwittingly be using a data-dependent prior so strong that it completely drives posterior inference. The
concern that prior strength inappropriately drives posterior inference affects a traditional Bayesian analysis as well, and so a sensitivity analysis should be undertaken to assess the extent with which this happens.

Similarly, those critical of data-dependent techniques can originate their arguments with the Likelihood Principle which states that all experimental information is contained in the data likelihood. Generally speaking, empirical Bayes techniques strongly violate this principle. Many objective Bayesian methods also violate the Likelihood Principle, as we have seen in the examples of Jeffreys, Zellner, and Wasserman, but the scope of offense is frequently venial by comparison, e.g. when an objective prior depends only on the sample size. The Likelihood Principle is often violated quite innocently by many analysts who choose a prior after data has been collected: for example, performing exploratory data analysis before establishing prior belief in the form of $\pi(\theta)$ would constitute such an offense. And yet the explicitly data-dependent approach still seems worse because of its brazen disregard of such a principled approach while still invoking the name Bayes.

The above criticisms notwithstanding, the most odious feature of the typical use of data-based priors is how the elegant simplicity of Bayes’ Theorem has been perverted. The statement $p(\theta|x) \propto f(x|\theta)\pi(\theta)$ is true and the usual rules of probability remain in force, while $p(\theta|x) \propto f(x|\theta)\pi(\theta|T(x))$ is patently false. Deely and Lindley (1981) humorously commented that, “It is this use of Bayes formula that presumably accounts for the reference to Bayes in the term, empirical Bayes.” Worse, the data-dependent formulation completely abandons the rules of probability while the notation perfectly begs for greater care.
1.3.2 Conclusion

At this point we have discussed two of the most damning objections to the use of data-dependent priors: they use the data twice and do not admit the calculus of Bayes’ Theorem directly. We will show that these criticisms can largely be overcome through a simply stated adjustment of the data-dependent prior. Violations of the Likelihood Principle and coherence shall remain, but the yawning chasm that exists between purely Bayesian and data-dependent Bayesian procedures will have shrunken considerably.
Chapter 2: Background Material

In this chapter, we orient the reader to view the current use of data-dependent priors in statistical modeling as a natural outgrowth of empirical Bayes methods. We provide a brief overview of the empirical Bayes paradigm and review some of the most salient literature on the subject. In addition, we review Markov Chain Monte Carlo (MCMC) techniques that will be used throughout the rest of the dissertation. Books on the broad subject of MCMC methods include Robert and Casella (1999), and Givens and Hoeting (2005).

2.1 Empirical Bayes

Consider the following Bayesian hierarchical model:

\[ X | \theta \sim f(x | \theta) \]  
\[ \theta | \eta \sim h(\theta | \eta). \]

If \( \eta \) is known, we are in a usual Bayesian setting, where we can compute the posterior as

\[ p(\theta | x, \eta) \propto f(x | \theta)h(\theta | \eta). \]

A fully Bayesian approach specifies an additional distribution \( \pi(\eta) \) that is interpreted as the prior density on the vector of hyperparameters \( \eta \). Again, posterior updating
is straightforward:

\[ p(\theta, \eta | x) \propto f(x|\theta)h(\theta|\eta)\pi(\eta). \]  \tag{2.4}

The empirical Bayes approach blends frequentist and fully Bayesian methods in the following way. Suppose that \( \eta \) is fixed, but unknown. As an alternative to expressing the uncertainty in our knowledge regarding \( \eta \) through a prior distribution, we could estimate \( \eta \) with \( T(x) \). This strategy yields an approximate posterior distribution for \( \theta \):

\[ p(\theta|x) \propto f(x|\theta)h(\theta|T(x)). \]  \tag{2.5}

From notation introduced in Chapter 1, \( h(\theta|T(x)) \) matches what we have described as a data-dependent prior.

The first seminal research in empirical Bayes methods is attributed to Robbins (1956) whose methods are directed primarily towards point estimation in a setting where the \( \theta_i \) are iid from an unspecified distribution \( G \). The scenario we describe in this section has been called parametric empirical Bayes (Morris, 1983) because of the conditioning on the hyperparameter \( \eta \). Carlin and Louis (2000) have written a book devoted to empirical Bayes methods, and Casella (1985) provides a brief introduction to the paradigm.

A concern that naturally arises in (2.5) is that the uncertainty in estimating \( \eta \) has not been taken into account. The remedial measures that have been proposed have typically focused on what are traditionally frequentist properties, with the notable exception of Deely and Lindley (1981) who advocate placing a prior on \( \eta \) and calling the procedure Bayes empirical Bayes. We have referenced much of the work on the nature of these corrections already with respect to the compound decision problem in Chapter 1.
2.2 Importance Sampling

Suppose \( h \) is a function such that \( E[h(X)] = \int h(x)f(x)dx \) exists. By the Law of Large Numbers, if \( X^{(1)}, \ldots, X^{(N)} \sim F, \text{iid} \), then \( \frac{1}{N} \sum_{i=1}^{N} h(X^{(i)}) \to E[h(X)] \text{ a.s.} \) as \( N \to \infty \). Thus, it is possible to obtain estimates of \( E[h(X)] \) by drawing a sample from \( F \) and taking an average. In cases where the relative frequency of the event of primary interest is small, but which would have a relatively great impact on \( E[h(X)] \), this procedure can be modified to instead take draws \( X^{(i)} \) from a density \( g \) with more mass over the range of interesting values than does \( f \). This is possible since we also have \( E[h(X)] = \int h(x)f(x)g(x)dx \). Thus, if \( X^{(1)}, \ldots, X^{(N)} \sim G \), then \( \frac{1}{N} \sum_{i=1}^{N} h(X^{(i)}) \frac{f(X^{(i)})}{g(X^{(i)})} \to E[h(X)] \) as \( N \to \infty \).

In this context, \( g \) is known as the importance sampling function and the sequence of ratios \( \omega_i = \frac{f(X^{(i)})}{g(X^{(i)})} \) for \( i = 1, \ldots, N \) are the importance weights. In the event that \( f \) is known only up to a constant, the weights can be normalized to have either mean 1 or sum 1, by computing either \( \omega_i^* = \frac{N\omega_i}{\sum_{j=1}^{N} \omega_j} \) or \( \omega_i^* = \frac{\omega_i}{\sum_{j=1}^{N} \omega_j} \), respectively.

Of particular importance is the variance of the estimate of \( \hat{\mu} = E[h(X)] \). It is easily seen, as in Robert and Casella (1999), that the variance of \( \hat{\mu} \) is finite only if

\[
\int [h(x)]^2 \frac{[f(x)]^2}{g(x)} dx < \infty. \tag{2.6}
\]

In particular, if \( g \) has thicker tails than \( f \), (2.6) will be satisfied. Rubinstein and Kroese (2008) provide an explicit form for the variance minimizing choice of importance function \( g \). Geweke (1989) provides large sample properties for the importance sample estimates.
A heuristic method which we draw on later to assess the performance of importance sampling is to compute the effective sample size. When the weights are standardized to have mean 1, the effective sample size (Kong et al., 1994) is calculated as

$$E_{SS} = \frac{N}{1 + \hat{var}(\omega^*)}, \quad (2.7)$$

where \(\hat{var}(\omega^*)\) is the sample variance of the \(\omega^*_i\). Intuitively, a large effective sample size is preferred to minimize the Monte Carlo variation.

2.2.1 Importance Link Functions

Related to importance sampling is the use of importance link functions (MacEachern and Peruggia, 2000) to better explore the tails of a target distribution after sampling has already been performed. This can be a troubling issue especially when the target distribution has thicker tails than the importance sampling function. The strategy involves finding a suitable set of transformations \(g_i\) such that \(g_i(x)\) reaches substantial regions of mass (under \(F\)) not reached through the original sample \(x\).

Suppose this objective can be achieved with a single transformation \(g\) that maps the original continuous state space to itself. Then, an estimate of \(E[h(X)]\) is given by:

$$\frac{\sum_{j=1}^{N} |J(x_j)| h(g(x_j)) f(g(x_j))/f(x_j)}{\sum_{j=1}^{N} |J(x_j)| f(g(x_j))/f(x_j)}, \quad (2.8)$$

where \(|J|\) is the absolute value of the determinant of the Jacobian of \(g\).

2.3 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm, Hastings (1970) and Metropolis et al. (1953), provides a very general means for obtaining an approximate sample from a target
distribution $\pi$. Suppose the chain is currently in state $X_t = x_t$. Then the steps needed to obtain $X_{t+1}$ are as follows:

1. From a proposal distribution, $p(\bullet | x_t)$, draw $X^*$.

2. Calculate the Metropolis-Hastings ratio as

$$\alpha(x_t, x^*) = \frac{\pi(x^*) p(x_t | x^*)}{\pi(x_t) p(x^* | x_t)}.$$  \hspace{1cm} (2.9)

3. Set $x_{t+1} = x^*$ with probability $\min(\alpha, 1)$. Otherwise, set $x_{t+1} = x_t$.

If the chain generated by the Metropolis-Hastings algorithm is aperiodic and irreducible, then the chain has unique limiting stationary distribution $\pi$. These criteria can be met if, for example, the proposal density $p$ is independent of the current state $x_t$ and $p(x) > 0$ when $\pi(x) > 0$ (Givens and Hoeting, 2005).

## 2.4 Gibbs Sampling

The Gibbs sampler (Gelfand and Smith, 1990), a special case of the Metropolis-Hastings algorithm, is a very useful technology for producing a Markov chain with target density $\pi$ when it is possible to sample from the set of full conditional densities as follows:

1. Initialize $x_0$. 

20
2. Obtain a draw $x_t$, having dimension $p$, from the full univariate conditional distributions:

$$X^1_{t+1} \sim \pi(\bullet|x_t^2, \ldots, x_t^p) \quad (2.10)$$

$$X^2_{t+1} \sim \pi(\bullet|x_{t+1}^1, x_t^3, \ldots, x_t^p) \quad (2.11)$$

$$\vdots \quad (2.12)$$

$$X^{p-1}_{t+1} \sim \pi(\bullet|x_{t+1}^1, \ldots, x_{t+1}^{p-2}, x_t^p) \quad (2.13)$$

$$X^p_{t+1} \sim \pi(\bullet|x_{t+1}^1, \ldots, x_{t+1}^{p-1}). \quad (2.14)$$

The Gibbs sampler can be used in conjunction with Metropolis Hastings steps to emulate a draw from several of the conditional distributions in step 2. Also, the algorithm can be modified to draw a subcollection of more than one of the variables at a time. A useful extension of the Gibbs sampler and Metropolis-Hastings algorithm was introduced in Neal (2000) to accommodate approximate posterior sampling from a Dirichlet Process mixture model. In particular, we will make use of his Algorithm 5 which introduces a Metropolis-Hastings step to overcome situations where the DP prior is not conjugate.
Chapter 3: The Adjusted Data-dependent Bayesian Paradigm

Having motivated the need for a new paradigm to better accomodate data-dependent priors in Bayesian modeling, we propose a formal method for adjusting the usual (naive) data-dependent analysis. The development requires repeated use of Bayes’ Theorem and leads to a surprisingly simple adjusted procedure. We consolidate the general spirit of the adjusted procedure in what we call the Data-dependent Bayesian Principle. The rest of the chapter carves out the adjusted procedure’s place in the space of existing Bayesian procedures. Fundamental properties of the adjusted data-dependent procedure are also established here.

3.1 Regularity Conditions

Before proceeding, we establish two easily satisfied regularity conditions that validate the probability calculations contained in this dissertation. All subsequent results shall implicitly assume that they hold.

(R1) The probability measures on the random variables $T(X) \mid \theta_1$ and $T(X) \mid \theta_2$ are mutually absolutely continuous for all $\theta_1, \theta_2 \in \Theta$ where $\Theta$ is the parameter space.
(R2) Suppose $g$ is the density of $T(x)$ and that $\pi(\theta|T(x))$ assigns probability 1 to $A \subseteq \Theta$, where
\[ A = \{ \theta \in \Theta | g(T(x)|\theta) > 0 \} . \tag{3.1} \]

Condition (R1) ensures that the set of random variables $T(X)|\theta$ indexed by $\theta$ all share the same null sets with respect to their probability measures. Condition (R2) guarantees that prior mass cannot be assigned to $\theta$ in conflict with $T(X)$. For example, suppose that $X_1, \ldots, X_n|\theta \sim U(0, \theta)$, iid, $T(X) = (X_{(n-1)}, X_{(n)})$, $n > 2$ the largest and second largest order statistics. Let the data-dependent prior be
\[ \pi(\theta|T(x)) = U \left( \frac{x_{(n)} + x_{(n-1)}}{2}, x_{(n)} \right) . \tag{3.2} \]
However, this leads to $g(T(x)|\theta) = 0$ for all $\theta$ in the support of $\pi(\theta|T(x))$.

3.2 Adjusted Data-dependent Bayes

Armed only with the data-dependent prior, $\pi(\theta|T(x))$, as a starting point, we shall develop a more sensible expression for the usual object of interest, the posterior density $p(\theta|x)$. The guiding theme in our derivation will be to create as simple an expression as possible in terms of objects that are usually part of a model formulation. We begin by expressing the joint probability density of $(\theta, x, T(x))$ in two different ways:
\[ p(\theta, x, T(x)) = p(T(x)|\theta, x)p(\theta|x)m(x) \tag{3.3} \]
\[ = f(x|\theta, T(x))\pi(\theta|T(x))m(T(x)) \tag{3.4} \]
A few observations are worth mentioning. First, both the data-dependent prior $\pi(\theta|T(x))$ and the posterior density $p(\theta|x)$ appear in the equations above. Isolating the posterior on the left side and shuffling all of the other objects to the right
hand side, yields
\[ p(\theta|x) = \frac{f(x|\theta, T(x)\pi(\theta|T(x))m(T(x))}{p(T(x)|\theta, x)m(x)}. \] (3.5)

Since \( T(X) \) is completely determined once \( X = x \) has been observed, \( T(X)|x, \theta \) is not random, and we may thus write
\[ p(\theta|x) = \frac{f(x|\theta, T(x))\pi(\theta|T(x))m(T(x))}{m(x)}. \] (3.6)

Finally, we might note that since \( m(T(x))/m(x) \) depends only on the observed data, we can write
\[ p(\theta|x) \propto f(x|\theta, T(x))\pi(\theta|T(x)). \] (3.7)

There is a great deal of beauty in this last line as it gracefully mimics the conventional proportionality statement relating posterior, data likelihood, and prior. The message is clear: if one is going to use a data-dependent prior in place of a usual prior, then one must also replace the unconditional data likelihood \( f(x|\theta) \) with the appropriate conditional likelihood, \( f(x|\theta, T(x)) \).

Formula (3.7) reveals an approximation that can justify the “typical” data-dependent Bayesian analysis. Specifically, the typical data-dependent updating corresponds to (3.7) with the usual data-likelihood substituted in the place of the conditional likelihood. The implication is that, assuming a coherent model structure, the researcher will draw identical inferences by approximating \( \pi(\theta|T(X)) \) (the unadjusted data-dependent approach) or alternatively approximating \( f(x|\theta, T(x)) \) with \( f(x|\theta) \) in (3.7). However, even if the model structure is coherent, the approximations induce incoherence in most non-trivial cases.

Although (3.7) is mathematically appealing, formally updating \( \pi(\theta|T(X)) \) is often not straightforward. Continuing, we aim to reintroduce the data likelihood into our
expressions:

\[ p(\theta|x) \propto f(x|\theta, T(x))\pi(\theta|T(x)) = \frac{p(x, T(x)|\theta)\pi(\theta|T(x))}{g(T(x)|\theta)} \]  

(3.8)

\[ \Rightarrow p(\theta|x) \propto \frac{f(x|\theta)\pi(\theta|T(x))}{g(T(x)|\theta)}. \]  

(3.9)

The last line follows from the fact that \( p(x, T(x)|\theta) = p(T(x)|x, \theta)f(x|\theta) \), where we again make use of \( p(T(x)|\theta, x) \) identically 1 since \( T(X) \) is completely determined once \( X = x \) has been observed.

Expression (3.9) describes \( p(\theta|x) \) in terms of familiar quantities. The numerator is the “naive” posterior arising from the data-dependent prior \( \pi(\theta|T(x)) \) and the unadjusted likelihood. The denominator supplies the requisite conditioning. Note also that (3.9) can be re-expressed as

\[ p(\theta|x) \propto f(x|\theta)\frac{\pi(\theta|T(x))}{g(T(x)|\theta)}, \]  

(3.10)

where the ratio \( \pi(\theta|T(x))/g(T(x)|\theta) \) functions as the prior distribution, ready to be updated with the likelihood \( f(x|\theta) \).

The adjusted procedure arbitrates a compromise between the information contained in the statistic and the information contained in the data regarding parameters modelled using a data-dependent prior. It can be understood with a trivial, but coherent, example that occurs in an analysis of sequential experiments. Suppose a researcher imposes \( \pi(\theta) \) as a prior density on \( \theta \), collects data \( X \) arising from the likelihood \( f(x|\theta) \), and reports the posterior \( \pi(\theta|x) \propto f(x|\theta)\pi(\theta) \). A second experiment might use the posterior density of the first experiment as a prior on \( \theta \) and update using new data \( Y \) in the usual way. This yields the posterior

\[ \pi(\theta|x, y) \propto f(x, y|\theta)\pi(\theta) \]  

that synthesizes the data from both experiments in a completely coherent way. Taking \((X, Y)\) as the data from the experiments, \( f(x|\theta) \)
be viewed as a data-dependent prior with $T(x) = x$. The naive updating would be to report $\pi(\theta|x,y) \propto f(x,y|\theta)\pi(\theta|T(x)) \propto [f(x|\theta)]^2 f(y|\theta)\pi(\theta)$, a particularly egregious example of double dipping. However, the adjusted approach naturally corrects this violation since the sampling distribution of $T(X)$ is quite simply $f(x|\theta)$. The resulting adjusted posterior is identical to that obtained by the orthodox Bayesian, namely $\pi(\theta|x,y) \propto f(y|\theta)\pi(\theta|T(x)) \propto f(x|\theta)f(y|\theta)\pi(\theta)$.

In what is to come, we shall refer to this alternative formulation as adjusted data-dependent Bayes, and the ratio $\pi(\theta|T(x))/g(T(x)|\theta)$ as an adjusted data-dependent prior. Our method differs from a traditional fully Bayesian analysis in that the starting prior information is encoded within $\pi(\theta|T(x))$ instead of $\pi(\theta)$. This change in representation of prior belief may still offend many Bayesians, but the approach derives more principally from Bayes’ Theorem than does the typical Bayesian analysis that neglects the dependence of the prior on the data.

### 3.2.1 The Data-Dependent Bayesian Principle

The preceding development required only the data-dependent prior $\pi(\theta|T(x))$ and data likelihood to obtain expressions (3.7) and (3.9). Importantly, there is no mention of a prior density $\pi(\theta)$ that can be used in conjunction with $f(x|\theta)$ to obtain the posterior density $p(\theta|x)$. Indeed, reasonable priors may not exist for $\theta$, or the priors envisioned are improper and consequently admit an improper posterior. Suppose now that a proper prior $\pi(\theta)$ does exist. Then, joint densities $p(x, \theta)$ and $p(T(x), \theta)$ enjoy the following relationship:

$$1 = \frac{p(\theta|x)m(x)}{f(x|\theta)\pi(\theta)} = \frac{\pi(\theta|T(x))m(T(x))}{g(T(x)|\theta)\pi(\theta)} .$$

(3.11)
Cancellation of $\pi(\theta)$ in both denominators leads to the following proportionality statement which we call the *Data-dependent Bayesian Principle*:

$$
\frac{p(\theta|x)}{f(x|\theta)} \propto \frac{\pi(\theta|T(x))}{g(T(x)|\theta)}. \quad (3.12)
$$

Note that this expression is equivalent to (3.9), but the conclusion we draw is even stronger. An orthodox Bayesian would certainly demand that the Data-dependent Bayesian Principle be satisfied as it follows straightforwardly from Bayes’ Theorem. However, we have shown something a bit stronger: the Data-dependent Bayesian Principle can be satisfied whether $\pi(\theta)$ exists or not. When using $\pi(\theta|T(x))$ one has several choices of approximate posterior $p(\theta|x)$. The adjusted procedure selects the posterior that conforms to an expression implied by Bayes’ Theorem even in the absence of $\pi(\theta)$.

### 3.3 Properties

Here we explore some of the basic theory of the adjusted procedure as a way to lay the foundation for its use. We begin with a pair of results on sufficiency and ancillarity.

**Theorem 3.3.1** *Suppose $T(X)$ is sufficient for $\theta$, then in the adjusted paradigm, the data-dependent prior $\pi(\theta|T(x))$ is also the posterior density.*

**Proof** By the Factorization Theorem, we have

$$
\pi(\theta|x) \propto f(x|\theta) \frac{\pi(\theta|T(x))}{g(T(x)|\theta)} \quad (3.13)
$$

$$
= h(x)g(T(x)|\theta) \frac{\pi(\theta|T(x))}{g(T(x)|\theta)} \quad (3.14)
$$

$$
\propto \pi(\theta|T(x)) \quad (3.15)
$$
Theorem 3.3.2 Suppose $T(X)$ is ancillary for $\theta$, then in the adjusted paradigm, the naive and adjusted posteriors are identical.

Proof Since the adjustment $g(T(x)|\theta) = g(T(x))$ by ancillarity, we have $\pi(\theta|x) \propto f(x|\theta)\pi(\theta|T(x))/g(T(x)|\theta) \propto f(x|\theta)
\pi(\theta|T(x))$. □

The ramifications of these theorems are enormous within the adjusted paradigm. It is hard to imagine that conditioning on an ancillary statistic would be beneficial in any meaningful way, except to note that some objective priors may be dependent on the data only through sample size or through the qualitative form of the likelihood (as with Jeffreys’ prior and with many objective Bayesian methods (Berger and Bernardo, 1992)). This suggests that responsible use of data dependent priors be restricted to priors dependent only on the sample size or conditional on a non-ancillary statistic.

The following result suggests that conditioning on a statistic that induces a coarse partition of the data space is preferred.

Theorem 3.3.3 Let $T_1(x)$ and $T_2(x)$ be statistics with corresponding classes of sets $A_i(x) = \{y|T_i(y) = T_i(x), y \in \mathcal{X}\}$ for $i = 1, 2$ such that $A_1(x) \subseteq A_2(x) \forall x \in \mathcal{X}$. Also, let $p_1(\theta|x)$ and $p_2(\theta|x)$ be the posteriors obtained using data-dependent priors based on $T_1(x)$ and $T_2(x)$, respectively. Suppose $\pi(\theta|T_1(x)) = \pi(\theta|T_2(x))$, then $p_2(\theta|x)/p_1(\theta|x) \propto p(T_1(x)|T_2(x), \theta)$ as a function of $\theta$.

Proof The result follows from the representation given in (3.6). Specifically,

$$p_1(\theta|x) = \frac{f(x|\theta, T_1(x))\pi(\theta|T_1(x))m(T_1(x))}{m(x)} \quad (3.16)$$

and

$$p_2(\theta|x) = \frac{f(x|\theta, T_1(x), T_2(x))p(T_1(x)|T_2(x), \theta)\pi(\theta|T_2(x))m(T_2(x))}{m(x)} \quad (3.17)$$
But, \( f(x|\theta, T_1(x), T_2(x)) = f(x|\theta, T_1(x)) \) since \( A_1(x) \subseteq A_2(x) \). Coupled with 
\( \pi(\theta|T_1(x)) = \pi(\theta|T_2(x)) \) the result follows.

Intuitively, if one were to use the exact same data-dependent prior regardless of whether \( T_1(x) \) or \( T_2(x) \) is used, and \( T_1(x) \) and \( T_2(x) \) are comparable in the sense of Theorem 3.3.3, then \( T_2(x) \) is preferred if there is information in \( T_1(x) \) about \( \theta \) above and beyond the information contained in \( T_2(x) \). As an example, suppose \( T_2(x) \) is the sample median and \( T_1(x) = x \). We already know from Theorem 3.3.1 that conditioning on \( x \) yields a null update from data-dependent prior to posterior. However, if the researcher is satisfied that the sample median provides enough information to formulate a data-dependent prior (and the sample median is not sufficient), then the adjusted posterior will reflect the additional information contained in the full data after accounting for the information provided by the sample median (or other “coarse” statistic). Alternatively, \( T_1(x) \) may be conditionally ancillary given \( T_2(x) \), in which case nothing is lost if \( T_1(x) \) is used instead of \( T_2(x) \).

The propriety of the adjusted posterior is of concern. The following lemma and theorem, whose assumptions are commonly satisfied, are enough to guarantee propriety.

**Lemma 3.3.4** Let \( f \) be continuous on \( \mathbb{R} \) with \( 0 < f < M \) for some \( M \in \mathbb{R} \). Let \( g \geq 0 \) be integrable with \( 0 < \int_{\mathbb{R}} g \, dx < \infty \). Then \( 0 < \int_{\mathbb{R}} fg \, dx < \infty \).

**Proof** Consider a closed bounded interval where \( g \) has mass, then \( f \) attains a minimum (greater than 0) and, consequently, \( \int_{\mathbb{R}} fg \, dx > 0 \). Also, \( \int_{\mathbb{R}} fg \, dx \leq M \int_{\mathbb{R}} g \, dx < \infty \).
Theorem 3.3.5 Let the vector of random variables $X$ and statistic $T(X)$ be such that the ratio of their pdfs, $\frac{f(x|\theta)}{g(T(x)|\theta)}$, is continuous and bounded on $\mathbb{R}$ as a function of $\theta$. Then, if $\pi(\theta|T(x))$ is any proper prior distribution based on $T(x)$, $p(\theta|x)$ is a proper posterior distribution.

Proof First, we recall that $p(\theta|x) \propto \frac{f(x|\theta)}{g(T(x)|\theta)} \pi(\theta|T(x))$. By virtue of the lemma given above $0 < \int_{\mathbb{R}} \frac{f(x|\theta)}{g(T(x)|\theta)} \pi(\theta|T(x)) \, d\theta < \infty$, thereby establishing the propriety of $p(\theta|x)$.

Alternatively, the propriety of $p(\theta|x)$ can be established from (3.7). The representation $p(\theta|x) \propto f(x|\theta, T(x)) \pi(\theta|T(x))$ guarantees that $p(\theta|x)$ is integrable provided that $\pi(\theta|T(x))$ is proper and $f(x|\theta, T(x))$ is a conditional density.

3.3.1 Exploiting Conditional Independence in a Hierarchical Model

Here we address a situation that may arise very naturally, where some subset of parameters is conditionally independent of the statistic upon which the data-dependent prior is based. Suppose the parameter vector $\theta$ can be split into two pieces $\theta = (\psi, \phi)$, such that the data-dependent prior on $\theta$ can be written as $\pi(\theta|T(x)) = \pi(\psi|\phi) \pi(\phi|T(x))$. Via a derivation analogous to the one performed in the previous section, we may write:

$$p(x, \psi, \phi) = f(x|\psi, \phi) \pi(\psi|\phi) \frac{\pi(\phi|T(x))}{g(T(x)|\phi)} m(T(x))$$

and so,

$$p(\psi, \phi|x) \propto f(x|\psi, \phi) \pi(\psi|\phi) \frac{\pi(\phi|T(x))}{g(T(x)|\phi)} .$$

30
The consequence of this result is that if the prior modeling is only performed at one stage of the hierarchy, say only at the most remote stage of hyperprior, then the proposed adjustment is restricted to that level of the hierarchy.

This form also applies to models where one portion of the prior is specified on the basis of scientific knowledge and another portion is data-based. In such cases, \( \pi(\theta|T(x)) \) may take the form \( \pi(\psi)\pi(\phi|T(x)) \), where \( \pi(\psi) \) is the scientific portion of the prior and \( \pi(\phi|T(x)) \) is the data-dependent portion (see e.g. Berliner et al. (2000), Wikle et al. (2001), and Berliner et al. (2008)). We note that the tools of graphical modeling and the introduction of latent variables can lead to useful forms of conditional independence for the analysis.

3.4 Equivalent Analyses

In this section we establish conditions under which the adjusted data-dependent analysis coincides with another form of analysis. We compare the adjusted analysis to: (1) its corresponding naive version, (2) an equivalent fully Bayesian analysis, and (3) another data-dependent Bayesian analysis. In general, we shall see that the analyses will in fact be different, but for special exceptions which we formalize.

3.4.1 Equivalence of Naive and Adjusted Data-dependent Procedures

We consider here when the naive and adjusted data-dependent procedures yield equivalent analyses. As we shall see, the criterion is quite restrictive.

**Theorem 3.4.1** The naive and adjusted analyses will be the same if and only if \( T(X) \) does not depend on \( \theta \).
Proof Using established notation, we begin with inspection of the competing forms of the posterior density.

Naive: $p_n(\theta|x) \propto f(x|\theta)\pi(\theta|T(x))$ (3.20)

Adjusted: $p_a(\theta|x) \propto f(x|\theta)\frac{\pi(\theta|T(x))}{g(T(x)|\theta)}$ (3.21)

Until now, we have only issued a proportionality statement of the naive approach. We now take the simple step of normalizing that expression with respect to $\theta$.

Naive: $p_n(\theta|x) = \frac{f(x|\theta)\pi(\theta|T(x))}{\int f(x|\theta)\pi(\theta|T(x))d\theta}$ (3.22)

Adjusted: $p_a(\theta|x) = \frac{f(x|\theta)\pi(\theta|T(x))m(T(x))}{g(T(x)|\theta)m(x)}$ (3.23)

Generally speaking, a researcher will be led to the same analysis only if $p_n(\theta|x) = p_a(\theta|x)$, which will be true only if

$$\frac{f(x|\theta)\pi(\theta|T(x))}{\int f(x|\theta)\pi(\theta|T(x))d\theta} = \frac{f(x|\theta)\pi(\theta|T(x))m(T(x))}{g(T(x)|\theta)m(x)},$$

which holds only if

$$g(T(x)|\theta) = \frac{m(T(x))}{m(x)} \int f(x|\theta)\pi(\theta|T(x))d\theta.$$ (3.25)

The right hand side of this equation is a function of $x$ alone since $\theta$ has been removed through integration. Thus $g(T(x)|\theta)$ is a function of $x$ alone which is true only if $T(X)$ and $\theta$ are independent, making independence a necessary condition. To show sufficiency, suppose that $T(X)$ and $\theta$ are independent. Then, $\pi(\theta|T(x)) = \pi(\theta)$, which implies

$$\int f(x|\theta)\pi(\theta|T(x))d\theta = \int f(x|\theta)\pi(\theta)d\theta = m(x).$$ (3.26)

We conclude that $T(X)$ and $\theta$ independent implies $g(T(x)|\theta) = m(T(x))$. 

Sample Size Dependent Priors

Does it ever happen in practice that a data-dependent prior $\pi(\theta|T(x))$ features a statistic $T(X)$ that does not depend on $\theta$? The simple answer is no, it does not, for if $T(X)$ does not depend on $\theta$, then $\pi(\theta|T(x)) = \pi(\theta)$ and we would say that a genuine prior on $\theta$ was mistaken for a data-dependent prior. Yet the question is made complicated by the existence of data-dependent priors which exhibit their dependence on the data only through conditioning on the sample size, i.e. $T(X) = T_1(X) = N$.

Let us pause to reconsider the effect of conditioning on the sample size. By Bayes Theorem, we have

$$p(\theta|x, n) = \frac{f(x|\theta, n)\pi(\theta|n)}{m(x|n)}. \quad (3.27)$$

This equation should now alleviate our concerns. The sample size $n$ can really be viewed as intrinsic to the data likelihood $f$ and the marginal likelihood $m$. Thus, it would be just as clear to write $p(\theta|x) = \frac{f_n(x|\theta)n\pi(\theta|n)}{m_n(x)}$, and it would be conventional to simply discard the subscripts since their meaning is implicitly clear. From this form, we see that no adjustment is necessary as long as the sample size truly does not convey any information about $\theta$. Rather, we should say that the adjusted paradigm does not demand that a correction factor be introduced in this scenario. If this makes one uncomfortable, then one solution is to just preserve the conditioning on $n$ by writing $\pi_n(\theta)$ to show that the form of the prior on $\theta$ would change with the sample size. We could interpret this null adjustment as implicitly sanctioning free use of priors dependent only on sample size.

To be clear though, a sequence of unequal priors indexed by sample size $n$ do convey different prior information, and is in the spirit of having a prior that depends fundamentally on a feature of the data.
3.4.2 Equivalence of Adjusted and Bayes Procedures

Suppose that $\pi(\theta | T(x))$ is a proper prior when viewed only as a function of $\theta$. The statistic $T(x)$ being viewed as fixed and known as prior information allows for the computational equivalence of treating $\pi(\theta | T(x))$ as a usual prior and subsequently using Bayes’ Theorem for the posterior updating. That is to say, $\pi(\theta | T(x))$ coincides functionally with a proper prior of the same form, but with no mention of $T(x)$. The same will be true of proper adjusted data-dependent priors, normalized and having the form:

$$
\frac{\pi(\theta | T(x))}{g(T(x)|\theta)} \int \frac{\pi(\theta | T(x))}{g(T(x)|\theta)} d\theta.
$$

(3.28)

Of course, the data-dependent priors, both naive and adjusted, will coincide with different non-data-dependent priors if different statistics $T(x)$ are observed, but the form of $\pi(\theta | T(x))$ does not change when viewed as a function of both $\theta$ and $T(x)$.

It may be possible that the naive data-dependent prior may be transformed into an equivalent non-data-dependent prior through the adjusted procedure. However, the functional form of such priors must be special. Specifically, we demand that

$$
\pi(\theta) \propto \frac{\pi(\theta | T(x))}{g(T(x)|\theta)},
$$

(3.29)

for every vector $x$. Thus, the expression on the right hand side must be a function of $\theta$ alone. This requires that $\pi(\theta | T(x))$ be factorizable as the product $\alpha(\theta)g(T(x)|\theta)$ for a function $\alpha$ of $\theta$ alone. The consequence is that some data-dependent priors will completely lose their dependence on the data through adjustment. As noted, this phenomenon requires a special relationship between the data-dependent prior, the form of the likelihood, and the statistic $T(x)$ being conditioned upon. Understanding
when this is possible is important, although the class of models where this occurs is quite restrictive, and ultimately only of trivial value.

### 3.4.3 Equivalent Prior Information

The data-dependent Bayesian principle allows for the comparison of two data-dependent priors which use different statistics. Suppose there are two priors \( \pi_1(\theta|T_1(x)) \) and \( \pi_2(\theta|T_2(x)) \). If it happens that

\[
\frac{\pi_1(\theta|T_1(x))}{g_1(T_1(x|\theta))} \propto \frac{\pi_2(\theta|T_2(x))}{g_2(T_2(x|\theta))},
\]

(3.30)

for the same data \( x \) and likelihood \( f \), then the resulting posteriors will be identical. When (3.30) holds, we say that the data dependent priors carry equivalent prior information in the adjusted data-dependent paradigm.

Expression (3.30) also provides a method for mapping data-dependent priors based on one statistic to equivalent data-dependent priors based on another. For data-likelihood \( f \), let us denote classes of data-dependent priors \( \Pi_{T_1} \) and \( \Pi_{T_2} \) having priors indexed by \( T_1(x) \) and \( T_2(x) \), respectively for each \( x \in \mathcal{X} \). The classes \( \Pi_{T_1} \) and \( \Pi_{T_2} \) are equivalent in the adjusted paradigm if and only if

\[
\pi_2(\theta|T_2(x)) \propto \frac{g_2(T_2(x|\theta))}{g_1(T_1(x|\theta))} \pi_1(\theta|T_1(x)),
\]

(3.31)

for every \( \pi_1(\theta|T_1(x)) \in \Pi_{T_1} \) and \( \pi_2(\theta|T_2(x)) \in \Pi_{T_2} \).

Equivalence of the classes \( \Pi_{T_1} \) and \( \Pi_{T_2} \) is very strong as it guarantees that updating of priors from either class will result in identical posterior densities for any observable data set in the adjusted paradigm. Note also that equivalence in the adjusted paradigm does not typically imply equivalence in the naive paradigm and vice versa. That is, the priors from \( \Pi_{T_1} \) and \( \Pi_{T_2} \) may be equivalent in the adjusted paradigm, but have produced unequivalent posteriors in the corresponding naive analyses.
3.5 Coherence

Adherence to the data-dependent Bayesian principle allows researchers to use data-dependent priors in a manner that is consistent with the use of proper priors in an orthodox Bayesian analysis. However, imposing the data-dependent adjustment is far from a panacea for the researcher who wishes to use the data in the prior, then pretend the data was not used. (Alas, unringing the bell is not possible.) Thus, it is important to highlight some of the fundamental limitations of the procedure, especially regarding deviations from coherent inference that orthodox Bayesian analyses enjoy.

3.5.1 Model Coherence

Briefly, the structure of a Bayesian model relying on data-dependent priors is incoherent in the sense that usual probability calculations will sometimes produce inconsistent results. Specifying $\pi(\theta|T(x))$ as a data-dependent prior and obtaining $g(T(x)|\theta)$ with the assumption that probability calculus is consistent for all $(\theta, T(x))$, requires that

$$
\pi(\theta|T(x)) = \frac{\pi(\theta)}{m(T(x))} g(T(x)|\theta).
$$

(3.32)

However, if this assumption is met, then the adjusted form of the posterior reverts to the usual posterior

$$
p(\theta|x) = f(x|\theta) \frac{\pi(\theta)}{m(T(x))}.
$$

(3.33)

We draw the conclusion that a “coherent” data-dependent Bayesian model is one which coincides exactly with a non-data-dependent model which uses the prior $\pi(\theta)$. In such instances, any prior information contained in the data has been completely extricated through adjustment. Since it is possible to verify assumption (3.32) before a data set is analyzed, a choice may be possible. If (3.32) is satisfied, peeking at the data
through $T(x)$ gains nothing in the prior, and hence can be avoided altogether. We reiterate that (3.32) is not typically satisfied when $\pi(\theta|T(x))$ is paired with $f(x|\theta)$.

### 3.5.2 Two Researchers, Two Data Sets

As a way of flushing out some of the themes on coherence and equivalence of analyses, we explore the consequences of the adjusted approach in the context of two researchers labeled 1 and 2 who collect data sets $X_1$ and $X_2$, respectively. Independently, the two researchers write down the function $\pi(\theta|T(x))$ and will plug in $T(X_1)$ or $T(X_2)$ for the data set each one has available. Suppose then that the researchers exchange their data so that each researcher can obtain posterior distributions for $\theta$ based on $(X_1, X_2)$. There are now two possibilities for how analysis can proceed, and we explore them both.

**Unequal Priors**

Researcher 1 forms the prior based on the data set he collected, $\pi(\theta|T(x_1))$ and obtains the posterior via the data-dependent Bayesian principle

$$p(\theta|x_1) \propto f(x_1|\theta) \frac{\pi(\theta|T(x_1))}{g(T(x_1)|\theta)}.$$  \hspace{1cm} (3.34)

Researcher 2 acts similarly, but bases his prior on $x_2$ to obtain

$$p(\theta|x_2) \propto f(x_2|\theta) \frac{\pi(\theta|T(x_2))}{g(T(x_2)|\theta)}.$$  \hspace{1cm} (3.35)

At this point the two researchers exchange data sets so they can both obtain $p(\theta|x_1, x_2)$. In general, the posterior densities they obtain from using their initial posterior analyses as priors for the joint analysis will differ. Researcher 1 obtains:

$$p(\theta|x_1, x_2) \propto f(x_1|\theta)f(x_2|\theta) \frac{\pi(\theta|T(x_1))}{g(T(x_1)|\theta)}.$$  \hspace{1cm} (3.36)
and Researcher 2 obtains

$$p(\theta|x_1, x_2) \propto f(x_1|\theta)f(x_2|\theta)\frac{\pi(\theta|T(x_2))}{g(T(x_2)|\theta)}.$$  \hspace{1cm} (3.37)

This is perhaps troubling, but it is not at all surprising: each researcher is fundamentally using a different data-dependent prior. The adjusted priors obtained by appealing to the data-dependent Bayesian principle are also unequal.

**Equal Priors**

The researchers could agree to base their data-dependent priors on $(T(x_1), T(x_2))$. This leads to the posterior based on both data sets

$$p(\theta|x_1, x_2) \propto f(x_1|\theta)f(x_2|\theta)\frac{\pi^*(\theta|T(x_1), T(x_2))}{g(T(x_1), T(x_2)|\theta)}.$$  \hspace{1cm} (3.38)

The researchers have effectively agreed to pool information from both $T(x_1)$ and $T(x_2)$ for data analysis.

There is a hidden peril to another kind of analysis using “equal” priors that the researchers could perform. That is, they could agree on using either $T(x_1)$ or $T(x_2)$ to obtain posterior density either (3.34) or (3.35). The problem with this is that jointly, they have both witnessed the impact of both $T(x_1)$ and $T(x_2)$ and hence should be required to adjust for having peeked at both. This leads to the analysis obtained in (3.38). The effect of *double-dipping* will be felt with respect to the statistic omitted from $(T(x_1), T(x_2))$.

Another option that might be available to the researchers prior to any data analysis being performed is to peek at $T(x_1, x_2)$, only $T(x_1)$ or $T(x_2)$. Using $T(x_1, x_2)$ corresponds to pooling all of the data together first, whereas using $T(x_1)$ or $T(x_2)$ is equivalent to agreeing to analysis (3.34) or (3.35) without knowledge of the unused $T(x_1)$ or $T(x_2)$. 

38
3.5.3 Remark

Adherence to the data-dependent Bayesian principle might cause one to expect that prescribed adjustment effectively washes away the dependence of the prior on the data. Typically this is not the case, as \( \pi(\theta|T(x))/g(T(x)|\theta) \) is still a non-trivial function of both \( \theta \) and \( T(x) \). This will clearly be demonstrated in forthcoming examples. Ultimately, the benefit to following the data-dependent Bayesian principle is that the adjusted data-dependent prior is now in a form that can be acceptably updated with the full likelihood. Intuitively, the adjusted procedure produces a new data-dependent prior modulo \( T(x) \). It is not that the data-dependence is washed away, but rather the double-update with respect to the information contained in the statistic \( T(x) \) has been effectively removed.
Chapter 4: Implementation

4.1 Analytically Determined Adjustment

If a closed form expression for the adjustment \( g(T(x)|\theta) \) is available, then it will often be easiest to implement the adjusted analysis by first fitting the unadjusted model as \( p(\theta|x) \propto f(x|\theta)\pi(\theta|T(x)) \), and then patching the analysis via (3.9). We list this implementation strategy first only because of its convenience. The wrinkle is that deriving the form of \( g(T(x)|\theta) \) demands that the data likelihood \( f(x|\theta) \) and statistic \( T(x) \) be sufficiently easy to work with. Unfortunately, it will be rare that \( g(T(x)|\theta) \) will be in closed form and evaluable, but one important example is in the normal-normal compound decision problem mentioned in Section 1.2.1.

4.1.1 Normal-Normal Compound Decision Problem

We consider here an analysis of the simple normal-normal model specified as in Carlin and Louis (2000):

\[
X_i|\theta_i \sim \mathcal{N}(\theta_i, \sigma^2), \sigma^2 \text{ known, } i \in \{1, 2, \ldots, n\} \tag{4.1}
\]

\[
\theta_i|\eta \sim \mathcal{N}(\eta, \tau^2), \tau^2 \text{ known, } i \in \{1, 2, \ldots, n\} \tag{4.2}
\]

\[
\eta|\bar{X} \sim \mathcal{N}\left(\bar{X}, \frac{\sigma^2 + \tau^2}{\alpha(n)}\right) \tag{4.3}
\]
The choice of $\alpha(n) = n$ corresponds to an empirical Bayes model that attempts to account for the uncertainty in estimating the “fixed” parameter $\eta$ with $\bar{X}$. This is justified by noting that (4.3) with $\alpha(n) = n$ arises naturally by swapping $\eta$ and $\bar{X}$ in the sampling distribution of $\bar{X}$. However, in the interest of remaining as general as possible, we permit $\alpha(n)$ to be any positive function of $n$.

Exploiting the conditional independence in the model, we obtain the form of the adjusted posterior density:

$$p(\theta, \eta|x) \propto f(x|\theta)g(\theta|\eta)\frac{\pi(\eta|\bar{x})}{h(\bar{x}|\eta)} \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta_i)^2 - \frac{1}{2\tau^2} \sum_{i=1}^{n} (\theta_i - \eta)^2 - \frac{\alpha(n) - n}{2(\sigma^2 + \tau^2)} (\eta - \bar{x})^2\right).$$

As this is an adjusted posterior density, it is not immediately clear for what values $\alpha(n)$ this density is proper. We desire to clearly identify this density and formally establish its propriety.

Let $\mu_i = E(\theta_i|x) = E(E(\theta_i|x, \eta)) = \frac{\sigma^2 \bar{x} + \tau^2 x_i}{\sigma^2 + \tau^2}$ and $\mu_\eta = E(\eta|x) = \bar{x}$. We shall handle each of these terms in turn.

First,

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \theta_i)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (\theta_i - \mu_i + \mu_i - x_i)^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} [(\theta_i - \mu_i)^2 + 2(\theta_i - \mu_i)(\mu_i - x_i) + (\mu_i - x_i)^2]$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (\theta_i - \mu_i)^2 - 2 \frac{1}{\sigma^2 + \tau^2} \sum_{i=1}^{n} (\theta_i - \mu_i)(x_i - \bar{x}) + c(x).$$

Next we have

$$\frac{1}{\tau^2} \sum_{i=1}^{n} (\theta_i - \eta)^2 = \frac{1}{\tau^2} \sum_{i=1}^{n} [(\theta_i - \mu_i) + (\mu_i - \bar{x}) + (\bar{x} - \eta)]^2 \quad (4.4)$$

41
\[
\begin{align*}
&= \frac{1}{\tau^2} \sum_{i=1}^{n} (\theta_i - \mu_i)^2 + \frac{1}{\tau^2} \sum_{i=1}^{n} (\eta - \bar{x})^2 + 2 \frac{1}{\sigma^2 + \tau^2} \sum_{i=1}^{n} (\theta_i - \mu_i)(x_i - \bar{x}) \\
&\quad - 2 \frac{1}{\tau^2} \sum_{i=1}^{n} (\theta_i - \mu_i)(\eta - \bar{x}) + c(x),
\end{align*}
\]

where \( c(x) \) indicates that the remaining terms can be expressed as functions of \( x \) alone. Combining the work from above, we may re-express \( p(\theta, \eta|x) \) as proportional to

\[
\exp \left( -\frac{\sigma^2 + \tau^2}{2\sigma^2 \tau^2} \sum_{i=1}^{n} (\theta_i - \mu_i)^2 \right) \times \exp \left( -\frac{n\sigma^2 + \alpha(n)\tau^2}{2\tau^2(\sigma^2 + \tau^2)} \sum_{i=1}^{n} (\eta - \mu_\eta)^2 \right) \\
\times \exp \left( \frac{1}{\tau^2} (\eta - \bar{x}) \sum_{i=1}^{n} (\theta_i - \mu_i) \right).
\]

This is exactly the kernel of a multivariate normal distribution with mean vector \( \mu = (\mu_1, \mu_2, \ldots, \mu_n, \mu_\eta) \) and covariance matrix \( \Sigma \) such that

\[
\Sigma^{-1} = \begin{pmatrix}
\frac{\sigma^2 + \tau^2}{\sigma^2 \tau^2} I_n \\
-\frac{1}{\tau^2} 1_n \\
\end{pmatrix} - \begin{pmatrix}
\frac{1}{\tau^2} 1_n \\
\frac{n\sigma^2 + \alpha(n)\tau^2}{\tau^2(\sigma^2 + \tau^2)} \\
\end{pmatrix} \cdot (4.5)
\]

We now derive the form of \( \Sigma \). To begin we label the partitions as

\[
\Sigma^{-1} = \begin{pmatrix}
B_{11} & B_{12} \\
B_{21} & B_{22}
\end{pmatrix} \quad (4.6)
\]

Then, by formulas contained in Ravishanker and Dey (2002),

\[
\Sigma_{11} = (B_{11} - B_{12}B_{22}^{-1}B_{21})^{-1} \quad (4.7)
\]

\[
\Sigma_{12} = -\Sigma_{11}B_{12}B_{22}^{-1} \quad (4.8)
\]

\[
\Sigma_{21} = -B_{22}^{-1}B_{21}\Sigma_{11} \quad (4.9)
\]

\[
\Sigma_{22} = B_{22}^{-1} + B_{22}^{-1}B_{21}\Sigma_{11}B_{12}B_{22}^{-1}. \quad (4.10)
\]
After some algebra ($J_n$ is an $n$ by $n$ matrix of ones),

\[
\Sigma_{11} = \left( \frac{\sigma^2 + \tau^2}{\sigma^2\tau^2} I_n - \frac{\sigma^2 + \tau^2}{\tau^2(n\sigma^2 + \alpha(n)\tau^2)} J_n \right)^{-1} = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} I_n + \frac{\sigma^4}{\alpha(n)(\sigma^2 + \tau^2)} J_n
\]

\[
\Sigma_{12} = \frac{\sigma^2}{\alpha(n)} 1_n
\]

\[
\Sigma_{21} = \frac{\sigma^2}{\alpha(n)} 1'_n
\]

\[
\Sigma_{22} = \frac{\sigma^2 + \tau^2}{\alpha(n)}.
\]

The positive definiteness of $\Sigma$ is implicit in this derivation, but a formal verification appears below. We now endeavor to show when $\Sigma$ is positive definite. The determinant of a partitioned matrix is given by

\[
|A| = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}|	ag{4.11}
\]

Let these submatrices be expressed as follows:

\[
A_{11} = \left( \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} - \lambda \right) I_n + \frac{\sigma^4}{\alpha(n)(\sigma^2 + \tau^2)} J_n \tag{4.12}
\]

\[
A_{12} = \frac{\sigma^2}{\alpha(n)} 1_n \tag{4.13}
\]

\[
A_{21} = \frac{\sigma^2}{\alpha(n)} 1'_n \tag{4.14}
\]

\[
A_{22} = \frac{\sigma^2 + \tau^2}{\alpha(n)} - \lambda \tag{4.15}
\]

and we proceed with

\[
A_{11} - A_{12}A_{22}^{-1}A_{21} = \left( \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} - \lambda \right) I_n - \frac{\sigma^4\lambda}{(\sigma^2 + \tau^2)(\sigma^2 + \tau^2 - \alpha(n)\lambda)} J_n \tag{4.16}
\]

\[
= \left( \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} - \lambda \right) \left( I_n - \frac{\sigma^4\lambda}{(\sigma^2 + \tau^2 - \alpha(n)\lambda)(\sigma^2\tau^2 - (\sigma^2 + \tau^2)\lambda)} J_n \right) \tag{4.17}
\]

and thus,

\[
|A| = \left( \frac{\sigma^2 + \tau^2}{\alpha(n)} - \lambda \right) \left( \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} - \lambda \right)^n \left( 1 - \frac{n\sigma^4\lambda}{(\sigma^2 + \tau^2 - \alpha(n)\lambda)(\sigma^2\tau^2 - (\sigma^2 + \tau^2)\lambda)} \right) \tag{4.18}
\]

43
\[
\left( \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} - \lambda \right)^{n-1} \left[ \left( \frac{\sigma^2 + \tau^2}{\alpha(n)} - \lambda \right) \left( \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} - \lambda \right) - \frac{n\sigma^4}{\alpha(n)(\sigma^2 + \tau^2)} \right]^{\lambda} \]  

(4.19)

\[
\left( \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} - \lambda \right)^{n-1} \left[ \lambda^2 - \frac{(n+1)\sigma^4 + (2 + \alpha(n))\sigma^2 \tau^2 + \tau^4}{\alpha(n)(\sigma^2 + \tau^2)} \lambda + \frac{\sigma^2 \tau^2}{\alpha(n)} \right], \quad (4.20)
\]

which has only real valued roots since \( \Sigma \) is symmetric. It is readily seen by the sign of the coefficients in the quadratic expression that all the roots are necessarily positive. We conclude that \( \Sigma \) is positive definite for any positive function \( \alpha(n) \) since the eigenvalues of \( \Sigma \) have all been demonstrated to be positive real numbers.

In summary, we have proven that \( (\theta, \eta|x) \) has a multivariate normal distribution with mean vector \( \mu = (\mu_1, \mu_2, \ldots, \mu_n, \mu_\eta) \) where \( \mu_i = \frac{\sigma^2 \bar{x}_i + \tau^2 x_i}{\sigma^2 + \tau^2} \) for each \( i \) and \( \mu_\eta = \bar{x} \), and covariance matrix

\[
\Sigma = \begin{pmatrix}
\frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} I_n + \frac{\sigma^4}{\alpha(n)(\sigma^2 + \tau^2)} \mathbf{1}_n \frac{\sigma^2}{\sigma^2 + \tau^2} \mathbf{1}_n' & \frac{\sigma^2}{\alpha(n)} \mathbf{1}_n \\
\frac{\sigma^2}{\alpha(n)} \mathbf{1}_n' & \frac{\sigma^2}{\alpha(n)} \mathbf{1}_n
\end{pmatrix}. \quad (4.21)
\]

Importantly, we have shown that this posterior is a proper distribution for any positive function \( \alpha(n) \). Also, for \( \alpha(n) = n \), the form of the posterior matches the analysis of Lindley (1971) had we used a flat prior on \( \eta \) instead of the data-dependent prior. That is, the adjusted data-dependent analysis formally coincides with the hyperprior Bayes analysis with \( \pi(\eta) \propto 1 \) for \( \alpha(n) = n \). This is to be expected because when \( \alpha(n) = n \), the sampling distribution of \( \bar{X} \) and the data-dependent prior on \( \eta \) are functionally identical. However, the result is a bit surprising since any \( \alpha(n) > 0 \) leaves room for some truly awful choices of data-dependent priors. Taking \( \alpha(n) < n \) in the adjusted analysis intuitively aligns with having less certainty in the prior information regarding \( \eta \) than the information a flat prior conveys. In fact, the entries of \( \Sigma \) can all be made arbitrarily large by taking \( \alpha(n) \) as small as we want. These would imply strange choices for \( \pi(\eta|\bar{x}) \), but they are all accomodated by the adjusted data-dependent approach.
4.1.2 Normal Model With Median

We consider now a simpler normal model with $\sigma^2$ known, $\mu$ unknown. We consider the ramifications of using the sample median in selecting a data-based prior. The model is specified as follows:

$$Y_i|\mu \sim N(\mu, \sigma^2), \ i = \{1, 2, \ldots, n\}$$

$$\mu|\tilde{Y} \sim \pi(\mu|\tilde{y}).$$

where $\tilde{y}$ is the sample median. For simplicity let us also suppose that $n \geq 3$ is odd. This enables us to easily express the distribution $p(\tilde{y}|\mu)$. The usual formula for the distribution of an order statistic of a random sample from a continuous population applied to the sample median yields (Casella and Berger, 2002):

$$p(\tilde{y}|\mu) = \frac{n!}{(n-1)! \left(\frac{n-1}{2}\right)!} \frac{1}{\sigma} \phi\left(\frac{\tilde{y} - \mu}{\sigma}\right) \left[\Phi\left(\frac{\tilde{y} - \mu}{\sigma}\right)\right]^{\frac{n-1}{2}} \left[1 - \Phi\left(\frac{\tilde{y} - \mu}{\sigma}\right)\right]^{\frac{n-1}{2}}$$

where $\phi$ and $\Phi$ are the standard normal pdf and cdf, respectively.

Let us examine the ratio

$$\frac{\prod_{i=1}^{n} f(y_i|\mu)}{p(\tilde{y}|\mu)} \propto \prod_{i=1}^{n} \frac{\phi\left(\frac{y_i - \mu}{\sigma}\right)}{\phi\left(\frac{\tilde{y} - \mu}{\sigma}\right) \left[\Phi\left(\frac{\tilde{y} - \mu}{\sigma}\right)\right]^{\frac{n-1}{2}} \left[1 - \Phi\left(\frac{\tilde{y} - \mu}{\sigma}\right)\right]^{\frac{n-1}{2}}}$$

$$= \prod_{i=1}^{n} \frac{\phi\left(\frac{y_i - \mu}{\sigma}\right)}{\phi\left(\frac{y_i - \mu}{\sigma}\right) \left[\Phi\left(\frac{\tilde{y} - \mu}{\sigma}\right)\right]^{\frac{n-1}{2}} \left[1 - \Phi\left(\frac{\tilde{y} - \mu}{\sigma}\right)\right]^{\frac{n-1}{2}}}$$

where $y_{(i)}$ denotes the $i^{th}$ order statistic. Each of these terms is continuous, positive and bounded above. The boundedness is a result of each term having a limit equal to zero at $\pm\infty$. Straightforwardly, the product will also be continuous, positive, and bounded above. As a result, any prior distribution based on the sample median will admit a proper posterior distribution for this normal model since the conditions of
Theorem 3.3.5 have been satisfied. The benefits here are two-fold. First, we have performed the simple task of deriving the posterior density using the data dependent Bayesian principle. Second, the posterior obtained is guaranteed to be proper in the adjusted data-dependent paradigm.

We turn now to a host of computational strategies that may be suitable when a complete analytic solution to the adjusted data-dependent problem is unattainable.

### 4.2 A General MCMC Strategy

The normal models of the previous two sections diverge from each other in an important respect. In the compound decision problem, a fully analytic expression for the adjusted posterior density was obtained, while the normal model with data-dependent prior based on the median does not produce a common family of posterior densities, even though the adjustment itself was analytically tractable. In such cases, it is desirable to have a general strategy for simulating from the adjusted posterior density. Indeed, the Metropolis-Hastings algorithm described in Chapter 2 provides the flexibility needed to accommodate oft-encountered non-conjugate models. The Metropolis-Hastings algorithm will also be particularly useful in settings where there is conjugacy in the naive data-dependent models, but where conjugacy is broken when adjustment is imposed. This is because only a proportional form of the posterior density is required in the Metropolis-Hastings acceptance ratio. In the context of the adjusted paradigm, this is expressed succinctly as $f(x|\theta)\pi(\theta|x)/g(T(x)|\theta)$, with no need for determining the normalizing constant.

As an extension of the general Metropolis-Hastings algorithm, Neal’s (2000) Algorithm 5 is particularly useful for implementing the data-dependent adjustment in DP
mixture models. An example analysis accompanies a derivation of a general adjustment for commonly used data-dependent priors in DP mixture modeling in Chapter 5. Having a closed form expression for the adjustment \(g(T(x)|\theta)\) is a very useful component of an adjusted analysis, but the lack of such an expression does not necessarily preclude such an analysis from being performed. Useful approximations to \(g(T(x)|\theta)\) may be available that can be implemented within a computational framework such as importance sampling or MCMC. We explore a host of such possibilities next.

4.3 Low Dimensional Computational Strategy

Except in simple models, it will not be possible to obtain a functional form of the data-dependent adjustment, but it may be that we are able to reasonably approximate the function \(\pi(\theta|T(x))/g(T(x)|\theta)\) over a fine grid of \(\theta\). Let us set the approximation to be \(\pi^*(\theta) = \frac{\pi(\theta|T(x))/g(T(x)|\theta)}{\int \pi(\theta|T(x))/g(T(x)|\theta) d\theta}\). Then the posterior density is approximately \(p(\theta|x) \propto f(x|\theta)\pi^*(\theta)\). Sampling from this distribution might now be performed using the Metropolis-Hastings algorithm which would abandon the need to normalize the densities.

A drawback of this method is that it may be computationally expensive to obtain an approximation to the adjustment that is acceptably precise. This is especially true as the dimensionality of \(\theta\) grows. Thus, functional approximation may be most suitable when the data-dependence of the prior can be shown to exist only in some small subset of the parameters, as may be the case in a Bayesian hierarchical model.
4.4 High Dimensional Computational Strategy - Importance Sampling

Suppose that an analysis has already been performed with the naive data-dependent prior through a sampling technique and that the naive posterior is represented by \( N \) equally weighted samples. With the addition of importance sampling weights, we can represent the adjusted posterior. The importance weighted sample can then be used to make inference under the adjusted model.

Let \( \theta_1, \theta_2, \ldots, \theta_N \) be a random sample from the naive posterior \( p_n(\theta|x) \propto f(x|\theta)\pi(\theta|T(x)) \). We represent the adjusted posterior distribution \( p_a(\theta|x) \) by the weighted sample \( (\theta_i, \omega_i)_{i=1}^N \) where, via (15),

\[
\omega_i \propto \frac{p_a(\theta_i|x)}{p_n(\theta_i|x)} \propto \frac{1}{g(T(x)|\theta_i)}.
\]

(4.27)

The \( \omega_i \) are normalized to have mean 1.

The \( \omega_i \) are used as importance sampling weights, with expectations approximated by weighted averages. Thus, for a \( p_a \)-integrable function \( h(\theta) \),

\[
E_a[h(\theta)] = \int h(\theta)p_a(\theta|x)d\theta
\]

(4.28)

\[
= \int h(\theta)\frac{p_a(\theta|x)}{p_n(\theta|x)}p_n(\theta|x)d\theta
\]

(4.29)

\[
= E_n \left[ h(\theta)\frac{p_a(\theta|x)}{p_n(\theta|x)} \right]
\]

(4.30)

\[
\approx \frac{1}{N} \sum_{i=1}^N h(\theta_i)\frac{p_a(\theta_i|x)}{p_n(\theta_i|x)}
\]

(4.31)

\[
\approx \frac{1}{N} \sum_{i=1}^N h(\theta_i)\omega_i
\]

(4.32)

Geweke (1989) provides large-sample results for estimators of the form given in the right hand side of (4.31). Use of these weights is fully compatible with another layer
of importance sampling, allowing one to sample from a distribution that has greater
dispersion than $p_n(\theta|x)$. Importance link functions (MacEachern and Peruggia, 2000)
provide a convenient means of doing this.

When $g(T(x)|\theta)$ is not available in a closed form, this approach reduces the ad-
justed inference to a density estimation problem. For each value of $\theta$, we draw many
samples $x$ from $f(x|\theta)$ and estimate the density $g(T(x)|\theta)$. This density estimation
is distinct from any MCMC procedure used to fit the model. Two benefits accrue: the
potentially large differences between the limiting distribution of the Markov chain
and the posterior distribution (Hans et al., 2009) that may arise from approximation
of the transition probabilities are avoided, and the weights can be approximated to
any desired accuracy.

Casting the problem as density estimation for $T(x)$ given $\theta$ changes the dimen-
sionality of the problem. The relevant dimension becomes that of $T(x)$ which will
often be small, rather than that of $\theta$ which will often be large, or even infinite, as
with nonparametric Bayesian models.

4.5 Application

In this section, we perform a data analysis using both the naive and adjusted
data-dependent approaches. We highlight implementation issues which are unique to
the adjusted analysis. In this example, we show how the parameter hierarchy and
nature of the data-dependence affect implementation. The results are not surprising,
but serve to illustrate the impact of an adjusted analysis on inference.
4.5.1 Baseball Batting Averages

Our example will use the classic baseball data set first analyzed by Efron and Morris (1975). For eighteen major league baseball players with exactly 45 at bats (AB), the numbers of hits have been recorded. Efron and Morris employ an arcsine transformation on the realized batting averages to stabilize the variance of a binomial distribution. In particular, they take \( X_i = f_{45}(Y_i) \) where \( f_n(y) = \sqrt{n} \arcsin(2y - 1) \), and \( Y_i \) is the sample batting average for each player. They then make use of the approximation \( X_i|\theta_i \sim N(\theta_i, 1) \) where \( \theta_i = f(p_i) \), \( p_i \) the true batting average. James-Stein estimation is performed on the transformed data and the resulting point estimates of transformed rest-of-season batting averages are transformed back to the usual scale.

Model

As in the original analysis, our goal is to predict each player’s batting average for the remainder of the season. Here we construct a simplified hierarchical data-dependent beta-binomial model specified as follows where the data, \( X \) now count the number of hits in 45 at bats:

\[
X_i|\psi_i \sim Binomial(45, \psi_i), i = 1, \ldots, 18 \tag{4.33}
\]

\[
\psi_i|\phi \sim Beta(\phi, c - \phi), c = 215.6 \text{ known} \tag{4.34}
\]

\[
\phi/c|T(X) \sim Beta(\nu_1, \nu_2), \tag{4.35}
\]

where \( T(X) = \sum_{i=1}^{18} X_i, \nu_1/(\nu_1 + \nu_2) = T(X)/n, \nu_1 + \nu_2 = n - 1, \) and \( n = 18 \).
Naive and Adjusted Posterior Densities

The naive analysis is straightforward since a proportional form of the posterior is obtained by simple multiplication of the densities, so that

\[ p_n(\psi, \phi|x) \propto f(x|\psi)g(\psi|\phi)\pi(\phi|T(x)) \]  

(4.36)

We must approach the adjusted analysis more carefully, by writing a form of the full joint probability density:

\[ p(x, \psi, \phi, T(x)) = f(x|\psi, \phi, T(x))g(\psi|\phi, T(x))\pi(\phi|T(x))m(T(x)). \]  

(4.37)

We now make use of the following three alternative forms for \( p(x, \psi, \phi, T(x)) \), \( f(x|\psi, \phi, T(x)) \), and \( g(\psi|\phi, T(x)) \):

\[ p(x, \psi, \phi, T(x)) = p(T(x)|x, \psi, \phi)p(x, \psi, \phi), \]  

(4.38)

\[ f(x|\psi, \phi, T(x)) = \frac{f(x|\psi, \phi, T(x))p(T(x)|\psi, \phi)}{p(T(x)|\psi, \phi)} \]  

(4.39)

\[ = \frac{p(x, T(x)|\psi, \phi)}{p(T(x)|\psi, \phi)} \]  

(4.40)

\[ = \frac{p(T(x)|x, \psi, \phi)f(x|\psi, \phi)}{p(T(x)|\psi, \phi)}, \]  

(4.41)

\[ g(\psi|\phi, T(x)) = \frac{g(\psi|\phi, T(x))h(T(x)|\phi)}{h(T(x)|\phi)} \]  

(4.42)

\[ = \frac{p(\psi, T(x)|\phi)}{h(T(x)|\phi)} \]  

(4.43)

\[ = \frac{p(T(x)|\psi, \phi)g(\psi|\phi)}{h(T(x)|\phi)}. \]  

(4.44)

Substituting each of these expressions into equation (4.37) we have:

\[ p(T(x)|x, \psi, \phi)p(x, \theta, \phi) = \frac{p(T(x)|x, \psi, \phi)f(x|\psi, \phi)p(T(x)|\psi, \phi)g(\psi|\phi)}{p(T(x)|\psi, \phi)h(T(x)|\phi)} \times \pi(\phi|T(x))m(T(x)) \]
and cancelling wherever we can gives:

\[
p(x, \psi, \phi) = f(x|\psi, \phi)g(\psi|\phi)\frac{\pi(\phi|T(x))}{h(T(x)|\phi)}m(T(x)).
\]  

(4.45)

Since \( f(x|\psi, \phi) = f(x|\psi) \) for this model, we obtain the adjusted proportional form of the posterior density:

\[
p_a(\psi, \phi|x) \propto f(x|\psi)g(\psi|\phi)\frac{\pi(\phi|T(x))}{h(T(x)|\phi)}.
\]  

(4.46)

This last expression is especially illuminating because it shows that the adjustment is related only to the parameter that instantiated the data-dependence. We will take the approach of approximating the function \( g(T(x)|\phi) \) as a function of \( \phi \) over a fine grid. Indeed, an exact form for \( g(T(x)|\phi) \) is available, but it is of little use. Let us investigate further by first finding a closed form for \( p(x_i|\phi) \).

**Implementation Challenges**

Since the \( X_i \) are conditionally independent, as are the \( \psi_i \), we have a beta-binomial model, and so

\[
p(x_i|\phi) = \int f(x_i|\psi_i)h(\psi_i|\phi)d\psi_i
\]  

(4.47)

\[
= \int \binom{45}{x_i}\psi_i^{x_i}(1-\psi_i)^{45-x_i}\frac{\Gamma(c)}{\Gamma(\phi)\Gamma(c-\phi)}\psi_i^{\phi-1}(1-\psi_i)^{c-\phi-1}d\psi_i
\]  

(4.48)

\[
= \int \binom{45}{x_i}\frac{\Gamma(c)}{\Gamma(\phi)\Gamma(c-\phi)}\psi_i^{x_i+c-\phi-1}(1-\psi_i)^{45-x_i+c-\phi-1}d\psi_i
\]  

(4.49)

\[
= \binom{45}{x_i}\frac{\Gamma(c)\Gamma(x_i+\phi)\Gamma(45-x_i+c-\phi)}{\Gamma(45+c)\Gamma(\phi)\Gamma(c-\phi)}
\]  

(4.50)

The exact expression for \( g(T(x)|\phi) \) is

\[
g(T(x)|\phi) = \sum_{x: \sum x_i = 215} \left[ \prod_{i=1}^{18} p(x_i|\phi) \right].
\]  

(4.51)

52
The sum on the right hand side has as many terms as there are integral solutions to \( \sum_{i=1}^{18} x_i = 215 \), with \( 0 \leq x_i \leq 45 \) for each \( i \). By a simple application of the inclusion-exclusion formula, the number of solutions is greater than \( 10^{25} \), and so the exact form has far too many terms to make direct computation feasible. We can rewrite the expression above as

\[ g(T(x)|\phi) = \sum_x \mathbb{I} \left[ \sum x_i = 215 \right] p(x|\phi) \]  

(4.52)

and then draw a sample of size \( N \) from \( p(x|\phi) \) to obtain the Monte Carlo estimate

\[ \hat{g}(T(x)|\phi) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I} \left[ \sum_{i=1}^{18} x_i^j = 215 \right]. \]  

(4.53)

In practice, this method will not work either since for many reasonable values of \( \phi \), the probability under \( p(x|\phi) \) of obtaining \( \sum x_i = 215 \) will be negligible. Instead, we implement importance sampling by finding a known distribution that can easily generate \( x \) such that their sum is 215. The Binomial\((45, 215/810)\) performs this job nicely. We proceed by drawing 32 million samples of size 18 from this distribution, keeping the 1,014,962 samples whose sum is 215. We discard the rest since the indicator function will be zero for those samples whose sum is not equal to 215.

Justification for this method is as follows:

\[ g(T(x)|\phi) = \sum_x \mathbb{I} \left[ \sum x_i = 215 \right] p(x|\phi) \]  

(4.54)

\[ = \sum_x \mathbb{I} \left[ \sum x_i = 215 \right] \frac{p(x|\phi)}{f(x)} f(x) \]  

(4.55)

\[ = E_{f} \left[ \mathbb{I} \left[ \sum x_i = 215 \right] \frac{p(x|\phi)}{f(x)} \right] \]  

(4.56)

\[ \approx \frac{1}{N} \sum_{j=1}^{N} \mathbb{I} \left[ \sum x_i^j = 215 \right] \frac{p(x|\phi)}{f(x)} \]  

(4.57)

where \( x^1, x^2, \ldots, x^N \) is a random sample drawn from \( f(x) = \prod_{i=1}^{18} \frac{(45)}{810} x_i \left( \frac{595}{810} \right)^{45-x_i} \).

The same sample of 32 million can be used to approximate \( g(T(x)|\phi) \) at 1000 evenly
spaced points along $[0, c]$, the support of $\phi$. The sampling distribution $g$, now well approximated, allows us to estimate the adjusted data-dependent prior
\[
\pi_a(\phi) = \frac{\pi(\phi|T(x))}{\hat{g}(T(x)|\phi)} \int \frac{\pi(\phi|T(x))}{g(T(x)|\phi)} d\phi
\] (4.58)
by swapping our importance sampling estimate $\hat{g}$ for $g$ into the expression above and estimating the integral by a Riemann sum with $\pi(\phi|T(x))/\hat{g}(T(x)|\phi)$ evaluated at 1000 grid points. Figure 4.1 compares the naive and adjusted data-dependent priors.

Results

We note in particular how the variance of the adjusted prior is noticeably larger than it is for the naive data-dependent prior. We also note that although the spread of each distribution on $\phi$ is markedly different, the centers are essentially unchanged.

Let us now consider the effect of using the adjusted prior on the parameters of primary interest, $\psi$. We will compute the marginal posterior means and variances for each of the $\psi_i$ under both analyses. Using iterated expectations and variances, we have
\[
E[\psi_i|x] = E[E[\psi_i|x, \phi]] = E[E[\psi_i|x, \phi]] = E\left[\frac{x_i + \phi}{45 + c}\right]
\] (4.59)
\[
= \frac{1}{45 + c} [x_i + E[\phi]],
\] (4.60)
and
\[
V[\psi_i|x] = V[E[\psi_i|x, \phi]] + E[V[\psi_i|x, \phi]]
\] (4.61)
\[
= V[E[\psi_i|x, \phi]] + E[V[\psi_i|x, \phi]]
\] (4.62)
\[
= V\left[\frac{x_i + \phi}{45 + c}\right] + E\left[\frac{(x_i + \phi)(45 - x_i + c - \phi)}{(45 + c)^2(46 + c)}\right].
\] (4.63)

Exact results are available for the naive analysis for these two quantities since $\phi$ has a data-dependent beta prior. The adjusted analysis requires approximation,
Figure 4.1: Plot of naive (red) and adjusted (black) data-dependent priors for the baseball example.
appealing again to the grid values on $\phi$ obtained previously to carry out the numerical estimation. Naive and adjusted estimates of batting averages for all 18 baseball players are given in the table on the following page with a comparison to their actual rest-of-season average.

Table 4.1 summarizes the most important findings. The most striking feature of these results is how similar the marginal posterior means under the two analyses are. The adjusted mean is actually a fixed amount (approx. 0.0009) larger than the naive mean. This follows from the form for the difference in expectation values under the two analyses which reduces to the difference of the prior mean on $\phi$ in the competing versions.

The difference in posterior standard deviation under each analysis for each of the $\psi_i$ is not fixed, but the adjusted standard deviations are between 35% and 40% larger than in the naive version. This is to be expected, since the variance of the adjusted prior on $\phi$ is substantially bigger than under the naive prior. The consequence for having peeked at the sum of all hits, and then formally accounting for it is primarily reflected in greater posterior variance of the parameters of interest.
<table>
<thead>
<tr>
<th>Player</th>
<th>H/AB</th>
<th>Season</th>
<th>Naive Mean</th>
<th>Adj. Mean</th>
<th>Naive SD</th>
<th>Adj. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clemente</td>
<td>.400</td>
<td>.346</td>
<td>.289</td>
<td>.290</td>
<td>.031</td>
<td>.042</td>
</tr>
<tr>
<td>F. Robinson</td>
<td>.378</td>
<td>.298</td>
<td>.285</td>
<td>.286</td>
<td>.031</td>
<td>.042</td>
</tr>
<tr>
<td>F. Howard</td>
<td>.356</td>
<td>.276</td>
<td>.281</td>
<td>.282</td>
<td>.031</td>
<td>.042</td>
</tr>
<tr>
<td>Johnstone</td>
<td>.333</td>
<td>.222</td>
<td>.277</td>
<td>.278</td>
<td>.030</td>
<td>.042</td>
</tr>
<tr>
<td>Berry</td>
<td>.311</td>
<td>.273</td>
<td>.273</td>
<td>.274</td>
<td>.030</td>
<td>.042</td>
</tr>
<tr>
<td>Spencer</td>
<td>.311</td>
<td>.270</td>
<td>.273</td>
<td>.274</td>
<td>.030</td>
<td>.042</td>
</tr>
<tr>
<td>Kessinger</td>
<td>.289</td>
<td>.263</td>
<td>.269</td>
<td>.270</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>L. Alvarado</td>
<td>.267</td>
<td>.210</td>
<td>.266</td>
<td>.267</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>Santo</td>
<td>.244</td>
<td>.269</td>
<td>.262</td>
<td>.263</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>Swoboda</td>
<td>.244</td>
<td>.230</td>
<td>.262</td>
<td>.263</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>Unser</td>
<td>.222</td>
<td>.264</td>
<td>.258</td>
<td>.259</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>Williams</td>
<td>.222</td>
<td>.256</td>
<td>.258</td>
<td>.259</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>Scott</td>
<td>.222</td>
<td>.303</td>
<td>.258</td>
<td>.259</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>Petrocelli</td>
<td>.222</td>
<td>.264</td>
<td>.258</td>
<td>.259</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>E. Rodriguez</td>
<td>.222</td>
<td>.226</td>
<td>.258</td>
<td>.259</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>Campaneris</td>
<td>.200</td>
<td>.285</td>
<td>.254</td>
<td>.255</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>Munson</td>
<td>.178</td>
<td>.316</td>
<td>.250</td>
<td>.251</td>
<td>.030</td>
<td>.041</td>
</tr>
<tr>
<td>Alvis</td>
<td>.156</td>
<td>.200</td>
<td>.246</td>
<td>.247</td>
<td>.030</td>
<td>.041</td>
</tr>
</tbody>
</table>

Table 4.1: Comparison of marginal posterior batting average estimates under the naive and adjusted techniques with rest-of-season batting averages.
Chapter 5: Adjusted Data-Dependent Bayes in Mixture Modeling

In this chapter, we apply the adjusted data-dependent procedure in the area of mixture modeling. Existing methods which use data-dependent priors do so with a nod toward objectivity. However, in several important cases, the statistic used to construct the data-dependent prior contains a great deal of information. Consequently, the adjustment will typically be severe for these choices of prior. We explore the ramifications of applying the adjustment in the context of finite mixture models and Dirichlet process mixture models.

5.1 The Null and Alternative Mixture Model

We begin by returning to Wasserman’s prior in the simplest of mixture models, with two equally weighted normal components and a single unknown mean with variances known and equal to 1:

\[
f(x|\mu) = \prod_i f(x_i|\mu) = \prod_i \left\{ \frac{1}{2} \phi(x_i) + \frac{1}{2} \phi(x_i - \mu) \right\}. \tag{5.1}
\]

If we desire an objective Bayes approach, then the most obvious choices for a prior on \( \mu \) are improper, namely Jeffreys’ prior and a flat prior. Let \( \pi(\mu) \) be an improper
prior, then straightforwardly,
\[
\int f(x|\mu)\pi(\mu)d\mu \geq 2^{-n} \prod_i \phi(x_i) \int \pi(\mu)d\mu = \infty. \quad (5.2)
\]

Wasserman follows the strategy of Diebolt and Robert (1994) of simply dropping the piece of the likelihood that causes the posterior to integrate to \(\infty\) giving the pseudo-likelihood:
\[
f^*(x|\mu) = \prod_i \left\{ \frac{1}{2} \phi(x_i) + \frac{1}{2} \phi(x_i - \mu) \right\} - \prod_i \frac{1}{2} \phi(x_i). \quad (5.3)
\]

Inference based on \(f^*(x|\mu)\pi(\mu)\) is identical to inference based on \(f(x|\mu)\pi(\mu|c_x(\mu))\) where \(\pi(\mu|c_x(\mu)) = (1 - \prod_i \{1 + \frac{\phi(x_i - \mu)}{\phi(x_i)}\}^{-1})\pi(\mu)\) is a data-dependent prior that we call Wasserman’s prior. Let us examine the first term of this expression which we have denoted by \(c_x(\mu)\). We provide a picture of the function \(c_x(\mu)\) for a particular set of data with \(n = 10\), \(X_i \sim f(x_i|\mu = 2)\) in Figure 5.1.

In a naive analysis, Wasserman’s prior is very useful, as it places mass near 1 for \(\mu\) over the range of the data while having support on the entire real line. Simulating from the naive posterior is straightforward and resulting posterior inference behaves in a similar manner as the data-dependent procedures of Richardson and Green (1997), and Raftery (1996). However, the formal data-dependent prior Wasserman derives from the technique of Diebolt and Robert is actually highly informative. In fact, Wasserman’s correction function is sufficient for \(\mu\) for the model given in (5.1).

**Theorem 5.1.1**  \(\text{Wasserman’s correction, } c_x(\mu), \text{ is sufficient for } \mu \text{ in the mixture of two normals model.}\)

**Proof**  The proof assumes complete knowledge of \(c_x(\mu)\) so that all derivatives may be evaluated in a neighborhood of zero. In summary, we will demonstrate constructively
Figure 5.1: Plot of Wasserman’s data-dependent prior
how to recover the even sample moments of $x$ along with $\sum_i x_i$. This set of statistics is sufficient for $\mu$ as it is almost surely equivalent to the order statistics of the original data set $x$. The proof relies a bit on some number theory and algebraic geometry, but is not technically difficult.

Consider the one-to-one function of $c_\mu(x)$,

$$f(\mu) = -\log (1 - c_\mu(x)) = \sum_i \log \left( 1 + \exp \left( \mu x_i - \frac{1}{2} \mu^2 \right) \right).$$  \hfill (5.4)

Then,

$$f'(\mu) = \sum_{i=1}^n \frac{\exp \left( -\frac{1}{2} \mu^2 + x_i \mu \right)}{1 + \exp \left( -\frac{1}{2} \mu^2 + x_i \mu \right)} (x_i - \mu).$$  \hfill (5.5)

This may be rewritten as

$$f'(\mu) = \frac{1}{2} \sum_{i=1}^n h(g_{x_i}(\mu))(x_i - \mu),$$  \hfill (5.6)

where

$$h(x) = \frac{2e^x}{1 + e^{2x}} = 1 + \tanh(x)$$  \hfill (5.7)

and

$$g_{x_i}(\mu) = -\frac{1}{4} \mu^2 + \frac{1}{2} x_i \mu.$$  \hfill (5.8)

Each function $g_{x_i}$ is a function of $\mu$ with:

$$g'_{x_i}(\mu) = \frac{1}{2} (x_i - \mu).$$  \hfill (5.9)

Thus, $g_{x_i}(0) = 0$ and $g'_{x_i}(0) = \frac{1}{2} x_i$. We may then write

$$f'(0) = \frac{1}{2} \sum_{i=1}^n h(0) x_i$$  \hfill (5.10)

which implies that

$$2f'(0) = \sum_{i=1}^n x_i.$$  \hfill (5.11)
From facts contained in Hirzebruch (2008) regarding Eulerian polynomials, we
have that \( P_j(-1) = \tanh^{(j)}(0) \) for \( j > 0 \), where \( P_j \) is the \( j \)th Eulerian polynomial.
Consequently, \( h^{(j)}(0) = P_j(-1) \) for each \( j > 0 \). Most important is the additional fact
that \( P_j(-1) = 0 \) for all even \( j > 0 \) and is nonzero for odd \( j > 0 \). Likewise, \( h^{(j)}(0) = 0 \)
for all even \( j > 0 \) and is nonzero for odd \( j > 0 \). The exponential generating function
(Hirzebruch, 2008) for the Eulerian polynomials, \( P_j(t) \) is given by
\[
\sum_{j=0}^{\infty} P_j(t) \frac{x^j}{j!} = (1 - t)e^{(1-t)x}. \tag{5.12}
\]
We proceed by explicitly giving the next two derivatives of \( f \) given as:
\[
f^{(2)}(\mu) = \sum_{i=1}^{n} \left[ \frac{1}{4} h^{(1)}(g_{x_i}(\mu))(x_i - \mu)^2 - \frac{1}{2} h(g_{x_i}(\mu)) \right] \tag{5.13}
\]
and
\[
f^{(3)}(\mu) = \sum_{i=1}^{n} \left[ \frac{1}{8} h^{(2)}(g_{x_i}(\mu))(x_i - \mu)^3 - \frac{3}{4} h^{(1)}(g_{x_i}(\mu))(x_i - \mu) \right]. \tag{5.14}
\]
Now consider functions of the form \( \xi_{x_i}(\mu) = \alpha h^{(j)}(g_{x_i}(\mu))(x_i - \mu)^m \) for \( j, m > 0, \alpha \neq 0 \). We have that
\[
\xi'_{x_i}(\mu) = \frac{\alpha}{2} h^{(j+1)}(g_{x_i}(\mu))(x_i - \mu)^{(m+1)} - \alpha m h^{(j)}(g_{x_i}(\mu))(x_i - \mu)^{(m-1)}. \tag{5.15}
\]
Additionally, if \( m = 0 \) in the expression for \( \xi \), then
\[
\xi'_{x_i}(\mu) = \frac{\alpha}{2} h^{(j+1)}(g_{x_i}(\mu))(x_i - \mu). \tag{5.16}
\]
As a consequence, a function expressible as a sum of terms of the form \( \xi \) where each
power of \((x_i - \mu)\) is even will have derivative with only terms consisting of odd powers
of \((x_i - \mu)\). Similarly, a function expressible as a sum of terms of the form \( \xi \) where
each power of \((x_i - \mu)\) is odd will have a derivative with terms consisting only of even
powers of \((x_i - \mu)\).
It is then apparent that \( f^{(j)} \) will be a summation of expressions consisting only of terms involving even powers of \((x_i - \mu)\) if \( j \) is even, or exclusively terms involving odd powers of \((x_i - \mu)\) if \( j \) is odd. Importantly, the first term will necessarily be
\[
\frac{1}{2j} h^{(j-1)}(g_{x_i}(\mu))(x_i - \mu)^j.
\] (5.17)

Thus, for even \( j \),
\[
f^{(j)}(0) = 2^{-j} P_{j-1}((-1) \sum_{i=1}^{n} x_i^j + R \left( \sum x_i^2, \sum x_i^4, \ldots, \sum x_i^{n-2} \right))
\] (5.18)

where \( R \) is a function of the even sample moments with power less than \( n \).

This allows us to iteratively solve for \( \sum x_i^2, \sum x_i^4, \ldots, \sum x_i^{2n} \). By Bezout’s Theorem, the even sample moments are enough to determine \( \{|x_1|, |x_2|, \ldots, |x_n|\} \) up to a permutation of the indices. This information, combined with the value of \( \sum x_i \), recovers the correct signs almost surely.

This is a serious problem within the adjusted data-dependent paradigm, because we have shown that Wasserman’s correction allows us to reproduce the original data set. Thus, Wasserman’s prior is sufficient for the mixture of normals model, and in the adjusted view, completely specifies the posterior distribution.

Having shown that Wasserman’s correction is sufficient for \( \mu \) in the simple mixture model (5.1), we now focus attention on an alternative data-dependent approach that may show more promise in the adjusted setting. The present application is simpler than those considered by Richardson and Green (1997), and so we modify their approach accordingly. Specifically, we model the prior on \( \mu \) using only information contained in the maximum and minimum of the realized data. Let \( R = Y_{(n)} - Y_{(1)} \) and \( \xi = (Y_{(n)} + Y_{(1)})/2 \), where \( Y_{(i)} \) is the \( i \)th order statistic. Then, the
data-dependent prior for \( \mu \) is \( \pi(\mu|y_1, y_n) = N(\mu|\xi, R^2) \). Although not as elegant to implement as Wasserman’s prior, the modified Richardson and Green prior has the advantage of not conditioning on a sufficient statistic. The sampling distribution \( g(y_1, y_n|\mu) = n(n-1)f(y_1|\mu)f(y_n|\mu)[F(y_n) - F(y_1)]^{n-2} \) lends itself easily to an analytic form of the posterior distribution:

\[
p(\mu|y) \propto \prod_{j=2}^{n-1} f(y_j|\mu) [F(y_n) - F(y_1)]^{n-2} \pi(\mu|y_1, y_n).
\] (5.19)

Figure 5.2 illustrates how the naive and adjusted posteriors compare when the modified Richardson and Green prior is used. For small samples, the adjusted version deviates noticeably from the unadjusted form, while for large samples the adjusted and unadjusted posteriors are qualitatively the same. Under the adjusted paradigm, the choice between Richardson and Green’s prior and Wasserman’s prior is clear. Wasserman’s correction burns all of the information contained in the data in the prior modeling because of its sufficiency. However, the Richardson and Green prior reserves information about the component mean because of their use of an insufficient statistic. Given that the unadjusted posteriors derived under the data-dependent priors of Wasserman, and Richardson and Green exhibit similar behavior, we prefer the Richardson and Green prior since its adjusted form admits a reasonable posterior, whereas Wasserman’s adjusted prior does not.

### 5.1.1 Breast Cancer Study

As an example, we analyze the breast cancer data of Hedenfalk et al. (2001) using naive and adjusted versions of the Richardson and Green data-dependent priors. The gene expression data consist of the p-values, \( P_i \), corresponding to pooled t-statistics of 3170 genes from breast cancer tissue: 7 tumors carrying the BRCA1 and 8 carrying the
Figure 5.2: Comparison of Richardson and Green naive (dash) and adjusted (solid) posterior densities for $\mu$. (left) $N = 5$, (center) $N = 10$, (right) $N = 100$. 

65
BRCA2 mutations compared to a t-distribution with 13 degrees of freedom. Previous analyses can be found in Storey and Tibshirani (2003), Broët et al. (2004), Efron (2004), and McLachlan et al. (2006). The data set we use is that of Storey and Tibshirani which omits 56 cases of genes exhibiting an extremely high degree of differential expression. Under the null hypothesis of no differential expression, $1 - P_i$ follows a standard normal distribution and the model for the transformed data $Z_i = \Phi^{-1}(1 - P_i)$ is:

$$f(z_i) = p\phi(z_i) + (1 - p)\frac{1}{\sigma}\phi\left(\frac{\mu - z_i}{\sigma}\right).$$ (5.20)

The prior modeling is similar to that of Richardson and Green (1997): $\sigma^{-2} \sim G(\alpha, \beta)$, $\mu \sim N(\xi, \kappa^{-1})$, $p \sim U(0, 1)$, $\beta \sim G(a, b)$, $\kappa = R^{-2}$, $\alpha = 2$, $a = 0.2$, $b = 10R^{-2}$, and $\xi$ is the mid-range of the data. Posterior simulation is straightforward via Gibbs sampling for the naive analysis since all of the full conditionals enjoy model conjugacy. An adjusted analysis can be carried out through importance sampling given the realizations of the naive analysis. The weights are proportional to $1/g(z_{(1)}, z_{(n)}|p, \mu, \sigma)$, where $g(z_{(1)}, z_{(n)}|p, \mu, \sigma) = n(n - 1)f(z_{(1)}|p, \mu, \sigma)f(z_{(n)}|p, \mu, \sigma)[F(z_{(n)}) - F(z_{(1)})]^{n-2}$ for each realized $(p, \mu, \sigma)$ obtained through naive posterior simulation. Alternatively, each Gibbs step in the MCMC algorithm used in the naive analysis can be amended to accommodate Metropolis-Hastings steps with acceptance ratios depending on the adjustment $g(z_{(1)}, z_{(n)}|p, \mu, \sigma)$.

The naive posterior simulation was run for 1.1 million iterations, then discarding the first 100,000 as burn-in. Values of parameters requiring arbitrary initialization included $\mu = 1$, $\sigma^2 = 1$, and $p = 0.5$. The adjusted analysis was obtained through an importance weighted sample of the naive simulation. The effective sample size
of 696,284 affirms our belief that the importance weighted sample is a reasonable representation of a sample from the true adjusted posterior distribution.

The naive analysis clearly identifies two posterior modes, a small one near \((0.7, 1.6, 0.3)\) and a much larger one near \((1.5, 1.0, 0.7)\). The adjusted analysis effectively reweights the posterior modes so that the smaller one accumulates more mass compared to the naive analysis. Figures 5.3, 5.4, and 5.5 illustrate the effect of the adjustment on posterior simulation while figures 5.6, 5.7, and 5.8 shows scatterplots of \(\log(w_i)\) against \(\mu_i, \sigma^2_i,\) and \(p_i,\) respectively, where \(w_i\) is the \(i\)th importance weight. The results demonstrate the necessity of performing an adjusted analysis even when the statistics (in this case the maximum and minimum of the data) would \textit{a priori} appear to contain relatively little information about the parameters.

Ultimately, we desire to know the posterior probability of a gene’s p-value landing in the second component of the mixture which corresponds with over-expression. To this end, we compute the following:

\[
f(R = 1|z, \text{data}) = \frac{f(z|R = 1, \text{data}) f(R = 1|\text{data})}{f(z|\text{data})} \approx \sum_{i=1}^{N} w_i \frac{(1 - p_i)\sigma_i^{-1}\phi \left( \frac{z - \mu_i}{\sigma_i} \right)}{p_i\phi(z) + (1 - p_i)\sigma_i^{-1}\phi \left( \frac{z - \mu_i}{\sigma_i} \right)}
\]

In the naive analysis, the \(w_i = 1/N\) for each \(i,\) but in the adjusted they are equal to the corresponding importance sampling weights. Equation 5.22 gives the posterior probability of falling into the second mixture component conditional on the value of \(Z = \Phi^{-1}(1 - P)\) where \(P\) is the p-value from the experiment. Figure 5.9 shows that a noticeable difference between the naive and adjusted analyses appears here as well.

For negative \(z,\) the bend upward is due to larger \(\sigma^2\) values given more weight in the adjusted analysis. However, for the region which the researcher is most interested in,
that is the region where \( Z \) is highest, the probability curves converge and provide virtually identical interpretation.

5.2 Dirichlet Process Mixture Models

While we have seen how data-dependent priors are used in finite mixtures, we now turn to the infinite dimensional setting of Dirichlet process mixture models. Here we derive the appropriate adjustment for the so-called data-dependent Gaussian method and explore a suite of computational strategies for implementing the adjustment. The complexity of the model requires a more careful MCMC strategy than was seen in the finite mixture models of the previous section. We proceed by investigating the use of pre-processed data in conjunction with a “standard” prior, and demonstrate how this corresponds with use of a data-dependent prior.

5.2.1 Data-Dependent Gaussian Method

The data-dependent Gaussian method (McAuliffe, Blei, and Jordan (2006) and Ishwaran and James (2002)) provides a simple recipe for effectively centering the base cdf \( G_0 \) together with the mass parameter \( M \) which informs the prior on \( G \), a procedure briefly introduced in Chapter 1. Simply, \( G_0 \) is specified to be a normal distribution with mean equal to the grand mean of the data, and variance 16 times the sample variance of the data. The method is analogous to pre-processing of the data by standardizing it, i.e. through the transformation of the data \( Y_i = (X_i - \bar{X})/\sqrt{S^2} \) where \( (n - 1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \), then setting \( G_0 \) to be a normal distribution with mean zero and variance 16.
Figure 5.3: Top: histogram of $\mu$ from naive posterior simulation. Bottom: density estimates of $\mu$ from naive (red) and importance sampled adjusted (blue) posterior simulation.
Figure 5.4: Top: histogram of $\sigma^2$ from naive posterior simulation. Bottom: density estimates of $\sigma^2$ from naive (red) and importance sampled adjusted (blue) posterior simulation.
Figure 5.5: Top: histogram of $p$ from naive posterior simulation. Bottom: density estimates of $p$ from naive (red) and importance sampled adjusted (blue) posterior simulation.
Figure 5.6: Scatterplot of log importance weights $\log(w)$ against $\mu$
Figure 5.7: Scatterplot of log importance weights $\log(w)$ against $\sigma^2$
Figure 5.8: Scatterplot of log importance weights $\log(w)$ against $p$
Figure 5.9: Plot of posterior probability of belonging to the second mixture component for the naive (red) and adjusted (blue) analyses.
Pre-processing data onforms to the Gaussian data-dependent method

Here we demonstrate the correspondence between pre-processing the data through the transformation

\[ Y_i = \frac{X_i - \bar{X}}{S} \]  

(5.23)

and the data-dependent Gaussian method which consists of centering and scaling the base measure using the mean and standard deviation of the data.

Suppose we specify the following elements of a DP mixture model:

\[ X_i | \theta_i, \tau^2 \sim \mathcal{N}(\theta_i, \tau^2) \]  

(5.24)

\[ \theta_i | G \sim G \]  

(5.25)

\[ G | M, \bar{X}, S^2 \sim \text{Dir}(MG_0) \]  

(5.26)

\[ G_0(\bullet) = \mathcal{N}(\bullet | \bar{X}, 16S^2). \]  

(5.27)

The quantities \( \bar{X} \) and \( S^2 \) are regarded as “fixed”, extrinsic prior information. Clearly, the prior on \( G \) is data-dependent, but suppose the \( \theta_i \) are transformed such that \( \theta_i^* = (\theta_i - \bar{X})/S \). Since each unique \( \theta_i \) is a random draw from \( G_0 \), each unique \( \theta_i^* \) is a random draw from a \( \mathcal{N}(0, 16) \) distribution. It follows that the transformation of the realized data corresponding to this reparametrized model is indeed \( Y_i = (X_i - \bar{X})/S \).

Thus, pre-processing the data through the transformation \( Y \) conforms to the data-dependent Gaussian method.

The effect of processing the data this way is, in fact, even more substantial since it imposes a reparametrization of the common variance \( \tau^2 \) as well. Now, with \( \tau^* = \tau / S \) together with a prescriptive (“vague”) prior for \( \tau^* \), such as \( \mathcal{IG}(\alpha, \beta) \) where \( \alpha \) and \( \beta \) are small, implicitly corresponds to a data-dependent prior on \( \tau \) in the untransformed space.
These principles generalize to any pair of centering and scaling statistics \( T(X) = (T_1(X), T_2(X)) \) again viewed as extrinsic prior information. If \( T_2(X) > 0 \), then the model

\[
X_i|\theta_i, \tau^2 \sim \mathcal{N}(\theta_i, \tau^2) \tag{5.28}
\]

\[
\theta_i|G \sim G \tag{5.29}
\]

\[
G|M, \bar{X}, S^2 \sim \text{Dir}(MG_0) \tag{5.30}
\]

\[
G_0(\bullet) = \mathcal{N}(\bullet|T_1(X), T_2(X)) \tag{5.31}
\]

can alternatively be obtained through the transformation

\[
Y_i = \frac{X_i - T_1(X)}{\sqrt{T_2(X)}} \tag{5.32}
\]

corresponding to the model

\[
Y_i|\theta_i^*, \tau^{*2} \sim \mathcal{N}(\theta_i^*, \tau^{*2}) \tag{5.33}
\]

\[
\theta_i^*|G \sim G \tag{5.34}
\]

\[
G|M, T_1(X), T_2(X) \sim \text{Dir}(MG_0) \tag{5.35}
\]

\[
G_0(\bullet) = \mathcal{N}(\bullet|0, 1) \tag{5.36}
\]

where \( \theta_i^* = (\theta_i - T_1(X))/\sqrt{T_2(X)} \) and \( \tau^* = \tau/\sqrt{T_2(X)} \).

**The adjustment in the Gaussian data-dependent method**

Consider the following normal model with common, unknown variance:

\[
X_i|\theta_i, \tau^2 \sim \mathcal{N}(\theta_i, \tau^2), \ i = 1, \ldots, n. \tag{5.37}
\]

We consider now, in isolation and without special regard to the higher stages of the model hierarchy, what the appropriate adjustment is if the data-dependent prior for
\((\theta, \tau^2)\) is based on \((\bar{X}, S^2)\). The required adjustment is the density of \((\bar{X}, S^2)\), given the parameters of the model, \(g(\bar{X}, S^2|\theta_1, \theta_2, \ldots, \theta_n, \tau^2)\), regarded as a function of the parameters. Let \(\theta = (\theta_1, \theta_2, \ldots, \theta_n)\). If \(X \sim \mathcal{N}(\theta, \tau^2 I)\), then \(X/\tau \sim \mathcal{N}(\theta/\tau, I)\), then \(\bar{X}\) and \(S^2\) are independent (Christensen, 2002), as are one-to-one transformations of \(\bar{X}\) and \(S^2\). Consequently, we may rewrite

\[
g(\bar{X}, S^2|\theta, \tau^2) = g_1(\bar{X}|\theta, \tau^2)g_2(S^2|\theta, \tau^2). \tag{5.38}
\]

It will be somewhat easier to work with the distributions of \(n\bar{X}\) and \((n-1)S^2\). Expressing the adjustment in terms of these simple transformations of \(\bar{X}\) and \(S^2\) is acceptable since the Jacobians of the transformations depend only on functions of \(X\) (which is fixed), and not on any of the parameters \((\theta, \tau^2)\). Thus, the adjustments \(g\) computed at each desired \((\theta, \tau^2)\) with respect to either \(\bar{X}\) and \(S^2\), or a one-to-one transformation, will be proportional to each other.

Now, \(n\bar{X} \sim \mathcal{N}(\sum_{i=1}^{n} \theta_i, n\tau^2)\). If the data have been standardized as described above, then

\[
g_1(n\bar{Y}|\theta, \tau^2) = \mathcal{N} \left( 0 \left| \sum_{i=1}^{n} \theta_i, n\tau^2 \right. \right). \tag{5.39}
\]

The distribution of \((n-1)S^2\) is not much more difficult to obtain. Let \(A = (I - J/n)\) where \(I\) is the \(n\) by \(n\) identity matrix and \(J\) is the \(n\) by \(n\) matrix of ones. Then, letting \(Z = X/\tau\), we have \(Z \sim \mathcal{N}(\theta/\tau, I)\), \((n-1)S^2/\tau^2 = Z'AZ\), and consequently

\[
Z'AZ \sim \chi^2(n-1, \theta'A\theta/2\tau^2). \tag{5.40}
\]

That is, \(Z'AZ\) has a non-central chi-squared distribution with \(n-1\) degrees of freedom and non-centrality parameter \(\theta'A\theta/2\tau^2\) (Christensen, 2002). Finally,

\[
p(\tau^2Z'AZ|\theta, \tau^2) = \chi^2(Z'AZ|n-1, \theta'A\theta/2\tau^2)/\tau^2 \tag{5.41}
\]
and
\[
g_2((n - 1)S^2|\theta, \tau^2) = \frac{\chi^2\left(\frac{(n-1)S^2}{\tau^2} \bigg| n - 1, \frac{\theta'\Lambda\theta}{2\tau^2}\right)}{\tau^2}. \tag{5.42}
\]

In words, \((n - 1)S^2\) has a scaled non-central chi-squared distribution.

In summary, the model in (5.37) yields the analytic adjustment:
\[
g(\bar{X}, S^2|\theta, \tau^2) \propto \frac{1}{\tau^2} N\left(n\bar{X} \bigg| \sum_{i=1}^{n} \theta_i, n\tau^2\right) \chi^2\left(\frac{(n-1)S^2}{\tau^2} \bigg| n - 1, \frac{\theta'\Lambda\theta}{2\tau^2}\right). \tag{5.43}
\]

The result is in fact quite general as it applies to any model where the data-likelihood is of the form (5.37) and where the data-dependent prior is of the form \(\pi(\theta, \tau^2|\bar{X}, S^2)\). The formula (5.43) is useable when the data \(X\) are in pre-processed form, \(i.e.\) the result of a transformation of the form \(X_i = (X_i^* - \bar{X}^*)/S_X\), and where the \(\theta_i\) and \(\tau^2\) have a commensurate meaning. This transformation of the data, with the statistics utilized for the data-dependent prior viewed as fixed, extrinsic quantities, imposes a similar transformation on the space of statistics. The Jacobian of the transformation is a function of \(S_X\) alone, thereby retaining the proportionality statement of (5.37).

Since the data-dependent Gaussian method conforms to this structure, the adjustment applies there as well. The ramifications of its use in Dirichlet process mixture models may now be explored in the context of a classic data set.

### 5.2.2 Galaxies Data

For the galaxies data of Postman, Huchra, and Geller (1986), later popularized in the mixture modeling literature by Roeder (1990), we perform naive and adjusted analyses and consider the unique computational issues that arise. The data set consists of the velocities of 82 galaxies relative to the Milky Way observed in six distinct
conic sections of space. Our goals are two-fold: we aim to demonstrate how the adjustment can be implemented while also outlining failed attempts to do so, and to compare the adjusted and naive analyses using the data-dependent Gaussian method, primarily through an examination of the resulting posterior predictive densities for each method of analysis. We assume the following Dirichlet process mixture (DP mixture) model, reserving specification of the prior on $\tau^2$ after a comment:

\[
X_i|\theta_i, \tau^2 \sim \mathcal{N}(\theta_i, \tau^2), \ i = 1, \ldots, n \tag{5.44}
\]

\[
\theta_i|G \sim G, \ i = 1, \ldots, N \tag{5.45}
\]

\[
G|M, \bar{X}, S^2 \sim \text{Dir}(MG_0), \text{ where } G_0(Y) = \mathcal{N}(Y|\bar{X}, 16S^2) \tag{5.46}
\]

\[
M = 1.5 \tag{5.47}
\]

We begin the analysis by preprocessing the data according to (5.23). The prior distribution on $\tau^2$ is intended to be “vague” and conjugate, and we defer specification of such a prior to one on the corresponding parameter in the transformed parameter space. Pre-processing implies the following transformed model:

\[
Y_i|\theta^*_i, \tau^{*2} \sim \mathcal{N}(\theta^*_i, \tau^{*2}), \ i = 1, \ldots, n \tag{5.48}
\]

\[
\theta^*_i|G \sim G, \ i = 1, \ldots, N \tag{5.49}
\]

\[
G|M \sim \text{Dir}(MG_0), \text{ where } G_0(\bullet) = \mathcal{N}(\bullet|0, 16) \tag{5.50}
\]

\[
\tau^{*2} \sim IG(.01, .01) \tag{5.51}
\]

\[
M = 1.5 \tag{5.52}
\]

where $\theta^*_i = (\theta_i - \bar{X})/S^2_X$ and $\tau^* = \tau/S_X$. Treated naively, posterior simulation is straightforward due to conjugacy and we implement the basic algorithm described in MacEachern (1998). The algorithm is also ideal, since we presently do not care to
draw an individual $G^{(i)}$ during iteration $i$, preferring to marginalize over the random density $G$.

Initial attempts at an adjusted analysis all began with the goal of applying the adjustment via importance sampling to posterior draws obtained through the naive analysis. When useable, this method is highly desirable for two main reasons. First, naive posterior simulation is very fast. Second, the importance weights are proportional to the adjustment and are also relatively easy to compute since a closed form expression is available.

Unfortunately, and upon repeated attempts, the importance weights exhibited symptoms of having an infinite variance with only a handful of posterior draws for $\theta$ and $\tau^2$ receiving substantial mass out of a sample of 100,000. The root cause is simple in that the adjusted posterior is suspected to have much heavier tails than its corresponding naive cousin. Because of the substantial savings in computation that the importance sampling method confers, further attempts were made to rein in the hypothesized tails of the adjusted posterior through importance link functions. These attempts focused on increasing dispersion around the realized cluster means by factors that depended on the cluster size. Ultimately, these techniques failed to bear fruit as the importance weights obtained continued to exhibit an unacceptably large variance. In turn, this led to the development of a Metropolis-Hastings simulation strategy to obtain an approximate sample from the adjusted posterior density directly as described in Chapter 4.

Using the adjustment derived in (5.42) together with the Metropolis-Hastings strategy of Chapter 4, we obtain an approximate sample from the adjusted posterior
distribution. The posterior predictive density for $Y_{n+1}|Y_1, \ldots, Y_n$ is approximated by

$$p(y_{n+1}|y) = \frac{1}{N} \left[ \sum_{j=1}^{N} \frac{M}{M+n^2} \mathcal{N}(y_{n+1}|0, 16 + \tau^2_{(j)}) + \frac{1}{M+n} \sum_{i=1}^{n} \mathcal{N}(y_{n+1}|\theta_{i(j)}, \tau^2_{(j)}) \right],$$

(5.53)

where $(\theta_{(j)}, \tau^2_{(j)})$ is the $j$th realized posterior iterate. We illustrate the difference between the naive and adjusted posterior predictive densities in figures 5.10, 5.11, and 5.12 for ten MCMC runs each. Figure 5.10 displays results from the traditional Gaussian data-dependent method as described here which inflates the variance of the density $G_0$ by 16 times the sample variance of the data. The plot clearly shows tails that are exaggerated in their thickness providing mass well outside a believable range. We also include results from an analysis which uses the sample variance without inflation and with reduced inflation as the variance of $G_0$. Here, the conclusion is similar, but less severe in its comical departure from the naive analysis. The adjusted posterior predictive density deviates sharply from the naive analysis, yet allocates most of its mass over a plausible interval. The posterior predictive densities have been transformed back to the original scale of the data.

5.2.3 Use of Order Statistics

From the galaxies data analysis of the previous section, we learn that the use of the sample mean and sample variance to construct a data-dependent prior requires a severe adjustment. The posterior predictive densities under the naive and adjusted analyses simply convey very different information in forecasting future observations. This behavior may be contrasted with the results of the Hedenfalk breast cancer analysis, which show only a minor departure between the naive and adjusted versions. The key difference seems to be in the amount of information provided by the summary
Figure 5.10: Naive (red) and adjusted (blue) posterior predictive densities with prior variance inflated by a factor of 16.
Figure 5.11: Naive (red) and adjusted (blue) posterior predictive densities with prior variance inflated by a factor of 4.
Figure 5.12: Naive (red) and adjusted (blue) posterior predictive densities with prior variance not inflated.
statistic balanced against the form of the prior used. Standardizing the data as in the Gaussian data-dependent method provides an easily implemented, generic approach, but perhaps is inefficient in maximally using the information contained in the sample mean and variance. We might expect a milder adjustment using a similarly generic technique, but with summary statistics that are significantly less informative.

We shall explore this hypothesis further by conducting an experiment that uses the order statistics to center and scale the galaxies data, and compares the posterior predictive likelihood of hold-out samples of various sizes under the naive and adjusted procedures. In each of 50 trials, 8 of the 82 galaxies were randomly selected to form the hold-out sample. The mean and variance of the data were estimated via maximum likelihood estimation from symmetrically chosen order statistics. Three sets of order statistics were separately chosen: $X_{(1)}$ and $X_{(74)}$, $X_{(8)}$ and $X_{(67)}$ (closest to the 1st and 9th deciles), $X_{(19)}$ and $X_{(56)}$ (closest to the 1st and 3rd quartiles). The procedure for selecting the data-dependent prior assumes that the data come from a normal population with mean $\mu$ and variance $\sigma^2$. Then, $\hat{\mu}$ and $\hat{\sigma}^2$ are found by maximizing the joint density of the order statistics $X_{(i)}$ and $X_{(j)}$ under an assumption of normality:

$$f(x_{(i)}, x_{(j)} | \mu, \sigma^2) = \frac{n!}{(i - 1)!(j - 1 - i)!(n - j)!} \frac{1}{\sigma^2} \phi \left( \frac{x_{(i)} - \mu}{\sigma} \right) \frac{1}{\sigma^2} \phi \left( \frac{x_{(j)} - \mu}{\sigma} \right) \left[ \Phi \left( \frac{x_{(i)} - \mu}{\sigma} \right) \right]^{i-1} \left[ \Phi \left( \frac{x_{(j)} - \mu}{\sigma} \right) - \Phi \left( \frac{x_{(i)} - \mu}{\sigma} \right) \right]^{j-1-i} \left[ 1 - \Phi \left( \frac{x_{(j)} - \mu}{\sigma} \right) \right]^{n-j}.$$

Calculation of these estimates was performed using the optim function in the statistical programming language R. The data-dependent Gaussian method was then employed with these estimates for $\mu$ and $\sigma^2$. It should be noted that the 50 hold-out samples were independently selected for each of the three sets of order statistics. That
is, the 50 hold-out sets when using $X_{(1)}$ and $X_{(74)}$ are different from the 50 used with $X_{(8)}$ and $X_{(67)}$, and likewise with $X_{(19)}$ and $X_{(56)}$. In its entirety, the model may be summarized as:

\[
Y_i \mid \theta^*_i, \tau^* \sim N(\theta^*_i, \tau^*), \ i = 1, \ldots, n
\]  
(5.54)

\[
\theta^*_i \mid G \sim G, \ i = 1, \ldots, N
\]  
(5.55)

\[
G \mid M \sim Dir(MG_0), \text{ where } G_0(\bullet) = N(\bullet|0, 1)
\]  
(5.56)

\[
\tau^* \sim IG(.01, .01)
\]  
(5.57)

\[
M = 1.5,
\]  
(5.58)

where $Y_i = (X_i - \hat{\mu})/\hat{\sigma}$.

As in the analysis which used the sample mean and variance of the data, the Metropolis-Hastings algorithm was used to directly obtain an approximate sample from the adjusted posterior distribution. For each hold-out set, the chain was run for 110,000 iterations, then dropping the first 10,000 cases as burn-in. Importance sampling was used to generate a representation of the naive posterior density from the adjusted version. We note that it is convenient to draw $G$ conditional on the current values of $\theta_i$ during each iteration of the MCMC, both in obtaining the appropriate acceptance ratio corresponding to the adjusted analysis, and also for evaluating the predictive log-likelihood for the out-of-sample cases. The joint posterior likelihood for subsets of size 1, 2, 4, and 8 from each hold-out sample was estimated for both the adjusted and naive versions. Since the observations are assumed to be conditionally iid, the joint predictive density of any subcollection is simply a product of univariate densities. Thus, the log-likelihood for each individual observation of the hold-out sample can be evaluated separately, then combined mutually as needed to obtain
joint likelihoods, weighted appropriately in the case of the naive analysis version. For
each hold-out sample of size 8, there are 8 subcollections of size 1, 28 of size 2, 70 of
size 4, and 1 of size 8.

The histograms shown in figures 5.13, 5.14, and 5.15 display the distribution of
the difference of log likelihoods for each subcollection size (naive log likelihood minus
adjusted log-likelihood). The boxplots in 5.16 summarize the same information. Gen-
erally speaking, there is no appreciable predictive advantage in using one of either the
naive or adjusted analyses over the other. Although, use of the approximate quartiles
appears to slightly favor the adjusted analysis, whereas use of the approximate deciles
slightly favors the naive analysis. The final three figures (5.17, 5.18, and 5.19) seek
to find systematic departures in the predictive quality of the competing analyses. No
general pattern seems to emerge, except to say that rare events are predicted with
slightly greater likelihood in the naive analysis in all three examples.
Figure 5.13: Histograms of the difference in predictive log likelihoods (naive minus adjusted) using $X_{(19)}$ and $X_{(50)}$ to generate the data-dependent prior.
Figure 5.14: Histograms of the difference in predictive log likelihoods (naive minus adjusted) using $X_{(8)}$ and $X_{(67)}$ to generate the data-dependent prior.
Figure 5.15: Histograms of the difference in predictive log likelihoods (naive minus adjusted) using $X_{(1)}$ and $X_{(74)}$ to generate the data-dependent prior.
Figure 5.16: Boxplots of the difference of predictive log likelihoods under the naive and adjusted analyses. Each boxplot is labeled by statistic used in forming the data-dependent prior, as well as the size of the hold-out sample.
Figure 5.17: Plots of adjusted predictive log likelihood against naive predictive log likelihood using $X_{(19)}$ and $X_{(56)}$ to generate the data-dependent prior.
Figure 5.18: Plots of adjusted predictive log likelihood against naive predictive log likelihood using $X_{(8)}$ and $X_{(67)}$ to generate the data-dependent prior.
Figure 5.19: Plots of adjusted predictive log likelihood against naive predictive log likelihood using $X_{(1)}$ and $X_{(74)}$ to generate the data-dependent prior.
Chapter 6: Conclusion

This dissertation has introduced the adjusted data-dependent Bayesian paradigm as a way to more faithfully use Bayes’ Theorem in statistical modeling. We have demonstrated several properties of the adjusted technique and have shown how to implement the procedure. Adjusted analyses have also been compared to their corresponding naive (unadjusted) analyses, with extensive treatment in the case of mixture modeling.

Here we offer several concluding remarks on the subject to potential users. As with other Bayesian methods, there is the prospect for serious abuse in implementing the paradigm, an example upon which we shed some light. We also discuss some of the new directions that this research may take in the future.

6.1 Opportunities for Abuse

Suppose that upon learning of the adjusted approach one wishes to compare an already completed naive analysis with its adjusted version. If the researcher used an explicitly data-dependent prior (as opposed to a sample size dependent prior), the resulting posterior inferences may be strikingly different. The extent to which the analyses differ can be so severe that the results based on the adjusted version effectively invalidate the original conclusions of the study. In light of the virtues of
performing an adjusted analysis, the researcher may be tempted to alter his or her
original analysis \textit{a posteriori}. In other words, the researcher may try to demonstrate
that the original analysis is equivalent to \textit{some} adjusted data-dependent technique.

In principle, this is not difficult to do, perhaps not difficult to hide, but is ex-
tremely dishonest! The problem reduces to comparing the data-dependent prior with
its adjusted version obtained by dividing the original naive prior by the sampling
distribution.

\begin{equation}
\text{Original Naive: } \pi(\theta|T(x))
\end{equation}

\begin{equation}
\text{Original Adjusted: } \frac{\pi(\theta|T(x))}{g(T(x)|\theta)}
\end{equation}

The researcher might now argue that he or she should have used a different data-
dependent prior from the outset having the specific form \( \pi(\theta|T(x))g(T(x)|\theta) \). It may
also be possible to hide the functional form by claiming the analytic form is unavail-
able and resorting to a graph of the function as a testament to its reasonableness.

Since \( \pi(\theta|T(x))g(T(x)|\theta) \) depends on the data only through \( T(x) \) the correct adjust-
ment is to divide through by \( g(T(x)|\theta) \), yielding the original naive data-dependent prior.

To summarize, the recipe to cheat is as follows:

1. Choose the data-dependent prior you really want: \( \pi(\theta|T(x)) \).
2. Do not reveal \( \pi(\theta|T(x)) \), but claim \( \pi(\theta|T(x))g(T(x)|\theta) \) is a reasonable data-
dependent prior.
3. The adjusted data-dependent prior \( \frac{\pi(\theta|T(x))g(T(x)|\theta)}{g(T(x)|\theta)} = \pi(\theta|T(x)) \) is the data-
dependent prior you were angling for.

The highest hurdle in executing the plan outlined above is accomplishing step 2, but in some special cases it can be deceptively easy. For example, \( \pi(\theta|T(x)) \)

and \( g(T(x)|\theta) \) have identical functional forms and step 2 reduces to the claim that 
\( \pi(\theta|T(x))^2 \) is a reasonable data-dependent prior. Thus, the dishonest researcher is
able to engineer a data-dependent prior that produces an adjusted analysis equivalent
to any naive analysis he or she envisions or has already completed.

### 6.2 Problems in Data Analysis

We have seen that the adjusted approach applies a correction to the data-dependent
prior that is determined by the sampling distribution of the statistic that is being con-
ditioned upon. In this paradigm, sufficient statistics are bad choices and sample size
dependent priors generally do not require adjustment, but is there a clearly defined
middle ground between these extremes that makes for sound data analysis? There
seems to be no clear answer to this question.

Certainly the adjustment can be enforced (or at least approximately enforced) in
any analysis that uses a data-based approach in principle. But this may lead to even
more objectionable analytic outcomes than in the naive scheme. Critics will argue
fairly that the adjustment purports to remove the prior data-dependence, but in prac-
tice it does not. So, many of the criticisms leveled against data-dependent Bayesian
methods before will apply here because in the common data-analytic scenario the
adjusted method is still data-dependent.

Fundamentally, the disharmony that we have seen adjusted analyses create is
caused by a failure to adequately impose reasonable prior distributions on param-
ters that are conditional on the statistics used. Viewed as fixed, extrinsic information,
\( \pi(\theta|T(x)) \) might genuinely be a reasonable, informative prior on \( \theta \), but when viewed
as a conditional density produces incredible results that deviate sharply from reality.
Thus, if the adjusted paradigm is to be embraced by the statistical community, then the meaning of $\pi(\theta|T(x))$ must change from “proxy” for $\pi(\theta)$, to “conditional density”. Such a definition is already implicit in the notation of $\pi(\theta|T(x))$, but may face resistance from a community that may be disinclined to alter the meaning of symbols as basic as these.

### 6.3 Black Box Methods

The process of standardizing data by centering and scaling to have mean 0 and variance 1 is common, and we have shown how this practice is equivalent to use of a data-dependent prior. We have demonstrated that using the sample mean and variance in this way can betray the true magnitude of the information these statistics contain about model parameters. Yet the goal of using these statistics is modest: the researcher wishes primarily to place prior mass near some measure of the center of the data and endow the prior with tails that are nearly guaranteed to cover the sample range. Additionally, the sample mean and variance are very easy to compute and interpret.

While it is not in our interest to promote the use of automatic procedures for constructing data-dependent priors, there is room to explore such possibilities within the adjusted paradigm. In particular, the use of order statistics in Chapter 5 to center and scale the data were shown empirically to accomplish the same general task as using the sample mean and variance, but the imposed adjustment was much milder. A rule of thumb that has seemed to emerge from this research is that if only vague information is known about certain parameters of the model and a data-dependent prior will be employed, then as little information as possible from the data should
be used. In the case of the data-dependent Gaussian method modified to use order statistics, this was exactly the case: only two carefully chosen points informed the prior.

Lastly, the routine practice of inflating the variance of the data-dependent prior appears to be significantly at odds with the adjusted paradigm. Variance inflation is used to inject vagueness in an attempt to mimic a non-informative prior. Through adjustment, the inflation is exposed as completely artificial. We conclude that in the adjusted paradigm a more careful approach must be taken to consider the bearing that the statistic $T(x)$ really has on our understanding of $\theta$. Ultimately, this is what will make the adjusted methods useable or not.

Identifying good statistics and distributions to use as data-dependent priors in a variety of other modeling settings is a sprawling task that must be left to future work. However, it is our hope that the content of this dissertation encourages the Bayesian community to adopt the adjusted data-dependent paradigm and that the area will advance beyond this initial work.
Bibliography


