6-15-2016

Adaptive Partition Weighted Monte Carlo Estimation

Yu-Bo Wang
yu-bo.wang@uconn.edu

Adaptive Partition Weighted Monte Carlo Estimation

Yu-Bo Wang, Ph.D.

University of Connecticut, 2016

ABSTRACT

This dissertation mainly focuses on the development of new Monte Carlo estimators for marginal likelihood and marginal posterior density with minimal assumption of a known nonnormalized posterior density and a single MCMC sample from the posterior distribution. We use the ideas of partitioning the parameter space and assigning an adaptive weight to the points of MCMC sample within different partition subsets. The estimators are shown to be consistent with the targets and their optimal performances in terms of minimizing the variance of estimators can be achieved by increasing the number of partition subsets. The proposing methods provide efficient ways to the problems including but not limited to Bayesian model or variable selection, the choices of power prior by empirical Bayes method, and phylogenetic model selection for a variable topology. Moreover, when multiple MCMC samples are available from the posterior density and conditional posterior densities, we provide a hybrid method, which is benefited from the dimension reduction.

Adaptive Partition Weighted Monte Carlo Estimation

Yu-Bo Wang

B.S., Statistics, National Chengchi University, Taiwan, 2006

M.S., Statistics, National Chengchi University, Taiwan, 2009

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Adaptive Partition Weighted Monte Carlo Estimation

Presented by

Yu-Bo Wang, B.S. Statistics, M.S. Statistics

Major Advisor _____
Lynn Kuo

Major Advisor _____
Ming-Hui Chen

Associate Advisor _____
Paul O. Lewis

University of Connecticut

2016

ACKNOWLEDGEMENTS

First, I want to exalt my God through Jesus Christ.

I would like to express my utmost and sincere gratitude to my major advisers Professor Lynn Kuo and Professor Ming-Hui Chen. With their wisdom, commitment, enthusiasm, and patience, they inspire and guide me in life and research, and establish great models of being a great scholar to me. Because of them, I can be what I am now.

I am grateful to my associate adviser Professor Paul O. Lewis for his kind help and valuable advice. He enriches my knowledge in phylogeny and programming, which are so important to me in the whole process of my graduate studies and to my future research.

I would also like to thank Professor James Grady. Because of him, I can have a valuable 4-year consulting experience in UConn Health Center. With his guidance and training, I have accessed to many different projects and become confident.

I would also like to thank Professor Natiee Ting. He enriches my experience in applied statistics. Through him, I have a deeper understanding about clinical trials, see the different applications of statistics, and cultivate my another research interest.

Many thanks are to all faculty members and my fellow graduate students. Special thanks to Tracy Burke and Megan Petsa for their helpful assistance, and to my former colleagues at UConn Health Center and Boehringer Ingelheim for offering me the precious opportunities to gain various experiences and continuous support. Lastly, I want to dedicate this dissertation to my wife, Yi-Chun, and my parents. Their everlasting love and support always encourage me to achieve my goals.

Contents

Chapter 1

Introduction

In this dissertation, marginal likelihood and marginal posterior density estimation is the main focus with special discussions on computation of the Bayes factor, the choice of power prior by empirical Bayes method, dimension reduction when multiple MCMC samples are available, and the application in phylogenetic models.

1.1 Marginal Likelihood Estimation

Evaluating the marginal likelihood in Bayesian analysis is essential for model selection. There are existing estimators based on a single Markov chain Monte Carlo sample from the posterior distribution, including the harmonic mean estimator and the inflated density ratio estimator. We propose a new class of Monte Carlo estimators based on this single Markov chain Monte Carlo sample. This class can be thought of as a generalization of the harmonic mean and inflated density ratio estimators using a partition weighted kernel (likelihood times prior). We also show that our estimator is

consistent and has better theoretical properties than the harmonic mean and inflated density ratio estimators. In addition, we provide guidelines on choosing the optimal weights. Simulation studies are conducted to examine the empirical performance of the proposed estimator. We further demonstrate the desirable features of the proposed estimator with two real data sets: one is from a prostate cancer study using an ordinal probit regression model with latent variables; the other is for the power prior construction from two Eastern Cooperative Oncology Group phase III clinical trials using the cure rate survival model with similar objectives.

1.2 Marginal Posterior Density Estimation

The computation of marginal posterior density in Bayesian analysis is essential in that it can provide complete information about parameters of interest. Furthermore, the marginal posterior density can be used for computing Bayes factors, posterior model probabilities, and diagnostic measures. The conditional marginal density estimator (CMDE) is theoretically the best for marginal density estimation but requires the closed-form expression of the conditional posterior density, which is often not available in many applications. We develop an Adaptive Partition weighTed (APT) method to realize the CMDE. This unbiased estimator requires only a single MCMC output from the joint posterior distribution and the known unnormalized posterior density. The theoretical properties and various applications of the APT estimator are examined in detail. The APT method is also extended to the estimation of conditional posterior densities. We further demonstrate the desirable features of the proposed method with two real data sets: one is from a study of dissociative identity disorder patients

using an analysis of variance model with constrained inequalities; the other is from a prostate cancer study, where model selection is investigated using ordinal probit regression models with latent variables.

1.3 Overview of the Dissertation

This rest of the dissertation is organized as follows:

Chapter 2 reviews some established methods for marginal likelihood. A detailed development of the proposed Monte Carlo marginal likelihood estimator is presented and its various properties are examined. We also provide a guideline on how to implement this method in different types of parameter space. A simulation study of comparing its performance with other existing methods is illustrated in a bivariate normal distribution with both unknown mean and covariance matrix. Then a real data from a prostate cancer study and ECOG data are analyzed to demonstrate the usefulness of this new method. A sensitivity analysis of this new method is also included.

Chapter 3 introduces the new Monte Carlo estimator for marginal posterior density with inspiration from the new method in Chapter 2. Its development and properties are detailed examined and discussed. In first real data, we empirically show the precision of this method in an inequality-constrained analysis of variance model. In second example, the application of this method to computing the Bayes factor is demonstrated. Besides, the benefit of dimension reduction is empirically shown when multiple MCMC samples are available.

Chapter 4 proposes the new marginal likelihood estimator for the phylogenetic

models with variable topology. It is inspired by the new method in Chapter 2. By using the concept of working parameter space, the estimator can avoid those topologies with few or no visiting of MCMC sample, which is a common phylogeny problem when a variable topology is considered.

Chapter 5 makes conclusions from the established theorems and the results from the simulation study and real data analysis. Discussions and directions for future research are also provided.

Chapter 2

Marginal Likelihood Estimation

2.1 Introduction

The Bayes factor quantifying evidence of one model over a competing model is commonly used for model comparison or variable selection in Bayesian inference. The Bayes factor is a ratio of two marginal likelihoods, where the marginal likelihood is essentially the average fit of the model to the data. However, the integration for the marginal likelihood is often analytically intractable due to the complex kernel (the likelihood times the prior) structure. To deal with this computational problem, several Monte Carlo methods have been developed. They include the importance sampling (IS) of Geweke (1989), the harmonic mean (HM) of Newton and Raftery (1994) and its generalization (GHM) of Gelfand and Dey (1994), the serial approaches of Chib (1995) and Chib and Jeliazkov (2001), the inflated density ratio method (IDR) of Petris and Tardella (2003) and Petris and Tardella (2007), the thermodynamic integration (TI) of Lartillot and Philippe (2006), the constrained GHM estimator with

the highest posterior density (HPD) region of Robert and Wraith (2009) and Marin and Robert (2010), and the steppingstone sampling of Xie et al. (2011). Under some mild conditions, they are all shown to be asymptotically convergent to the marginal likelihood by the ergodic theorem. They vary in using Monte Carlo samples or kernels in the Monte Carlo integration.

We assume only a single Markov chain Monte Carlo (MCMC) sample from the posterior distribution, which may be readily available from standard Bayesian software, and the known kernel function for computing the marginal likelihood. The HM and IDR estimators are the two existing ones, which only need these two minimal assumptions. The main difference between the HM and the IDR estimators lies in the different weights assigned to the inverse of the kernel function. The former uses the prior function as a weight, while the latter uses the difference between a perturbed density and its kernel function. Although the HM estimator has been used in practice because of its simplicity, it can be unstable when the prior has heavier tails than the likelihood function and it is known to overestimate the marginal likelihood (Xie et al., 2011).

While the IDR estimator has better control over the tails of the kernel than the HM estimator in controlling the tails of the likelihood function, it requires reparameterization, posterior mode calculation, and a careful selection of radius. Under the aforementioned two minimal assumptions, we extend the HM and IDR methods to develop a new Monte Carlo method, namely, the partition weighted kernel (PWK) estimator. The PWK estimator is constructed by first partitioning the working parameter space (where the kernel is bounded away from zero), and then estimating the marginal likelihood by a weighted average of the kernel values evaluated at the MCMC sample, where weights are assigned locally using a representative kernel value

in each partition. We show the PWK estimator is consistent and has finite variance. When the partition is refined enough to make the kernel values in the same region similar, we can construct the best PWK estimator with the minimum variance. Our simulation study empirically shows that the proposed PWK estimator outperforms both the HM and IDR estimators.

The rest of Chapter 2 is organized as follows. Section 2 is a review of the HM, GHM and IDR methods that motivate the PWK estimator. In Section 3, we develop the PWK estimator and its theoretical properties. Additionally, in the class of the general PWK estimator, we find the best (minimum variance) PWK estimator and provide a spherical shell approach to realize it. In Section 4, an extended general PWK estimator which is defined on the full support of the kernel function is investigated. Besides the theoretical properties, we show that the HM and IDR estimators are special cases in this family. In Section 5, we conduct a simulation study of a bivariate normal case with the normal-inverse-Wishart prior to compare the performance and computing time of the HM, IDR and PWK estimators. In Section 6, we compare the performance of the PWK estimator to the methods by Chib (1995) and Chen (2005b) for a ordinal probit regression model. Moreover, we apply the PWK estimator to the determination of the optimal power prior using two ECOG clinical trial data sets. Finally, we conclude with a discussion in Section 7.

2.2 Preliminary

We review several Monte Carlo methods that only require a known kernel function and an MCMC sample from the posterior distribution to compute the marginal likelihood.

Suppose $\boldsymbol{\theta}$ is a p-dimensional vector of parameters and D denotes the data. Then, the kernel function for the joint posterior density $\pi(\boldsymbol{\theta}|D)$ is $q(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})$, where $L(\boldsymbol{\theta}|D)$ is the likelihood function and $\pi(\boldsymbol{\theta})$ is a proper prior density. Assume $\boldsymbol{\Theta} \subset R^p$ is the support of $q(\boldsymbol{\theta})$. The unknown marginal likelihood c is defined to be $\int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta})d\boldsymbol{\theta}$. Due to the complicated kernel structure, the integration is often analytically intractable.

To estimate the normalizing constant c, Newton and Raftery (1994) suggest the following equation to motivate the HM method,

$$\frac{1}{c} = \int_{\boldsymbol{\Theta}} \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta})}{c} d\boldsymbol{\theta}. \tag{2.1}$$

Let $\{\boldsymbol{\theta}_t,\ t = 1, \ldots, T\}$ be an MCMC sample from the posterior distribution $\pi(\boldsymbol{\theta}|D) = q(\boldsymbol{\theta})/c$. The HM estimator is then given by

$$\hat{c}_{HM} = \frac{1}{\frac{1}{T}\sum_{t=1}^{T} \frac{1}{L(\boldsymbol{\theta}_t|D)}}, \tag{2.2}$$

where the prior $\pi(\boldsymbol{\theta}_t)$ can be viewed as the weight assigned to $1/q(\boldsymbol{\theta}_t)$. Although it has the feature of simplicity and has asymptotic convergence to the marginal likelihood, the finite variance is not guaranteed. Xie et al. (2011) also point out that the HM estimator tends to overestimate the marginal likelihood.

Gelfand and Dey (1994) suggest the GHM estimator where $\pi(\boldsymbol{\theta})$ in equation (2.1) is replaced by a lighter-tailed density function $f(\boldsymbol{\theta})$ compared to $q(\boldsymbol{\theta})$:

$$\hat{c}_{GHM} = \frac{1}{\frac{1}{T}\sum_{t=1}^{T} \frac{f(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)}}. \tag{2.3}$$

By proposing a light-tailed density, the ratio $f(\boldsymbol{\theta}_t)/q(\boldsymbol{\theta}_t)$ can be controlled. Consequently, the estimator has finite variance. However, in high dimensional problems, finding a suitable density $f(\boldsymbol{\theta})$ may be a challenge.

Petris and Tardella (2003) propose the IDR estimator. They use the difference between a perturbed distribution $q_r(\boldsymbol{\theta})$, which is inflated in the center of the kernel, and the posterior kernel $q(\boldsymbol{\theta})$ as the weight. The perturbed density $q_r(\boldsymbol{\theta})$ is defined as

$$q_r(\boldsymbol{\theta}) = \begin{cases} q(\mathbf{0}) & \text{if } ||\boldsymbol{\theta}|| \leq r, \\ q(w(\boldsymbol{\theta})) & \text{if } ||\boldsymbol{\theta}|| > r, \end{cases} \tag{2.4}$$

where r is the chosen radius and $w(\boldsymbol{\theta}) = \boldsymbol{\theta}\left(1 - r^p/||\boldsymbol{\theta}||^p\right)^{1/p}$. It follows,

$$\int_{\boldsymbol{\Theta}} q_r(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{||\boldsymbol{\theta}||\leq r} q_r(\boldsymbol{\theta}) d\boldsymbol{\theta} + \int_{||\boldsymbol{\theta}||>r} q_r(\boldsymbol{\theta}) d\boldsymbol{\theta} = q(\mathbf{0}) b_r + c, \tag{2.5}$$

where b_r = Volume of the ball $\{\boldsymbol{\theta} : \ ||\boldsymbol{\theta}|| \leq r\} = \pi^{p/2} r^p / \Gamma(p/2+1)$. Then, we can have the following equation,

$$\frac{q(\mathbf{0}) b_r + c}{c} = \int_{\boldsymbol{\Theta}} \frac{q_r(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta})}{c} d\boldsymbol{\theta}, \tag{2.6}$$

and the IDR estimator is given by

$$\hat{c}_{IDR} = \frac{q(\mathbf{0}) b_r}{\frac{1}{T}\sum_{t=1}^{T} \frac{q_r(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)} - 1}. \tag{2.7}$$

Under some mild conditions, the estimator is shown to have finite variance by Petris and Tardella (2007). However, the method requires a careful selection of radius and unbounded support of $q(\boldsymbol{\theta})$. Any bounded parameter must be reparameterized to the

full real line. Also, in order to have a more efficient estimate, mode finding is essential and standardization of an MCMC sample with respect to the mode and the sample covariance matrix is required.

2.3 A New Monte Carlo Estimator

We first modify (2.1) and (2.6) by imposing a working parameter space $\Omega \subset \boldsymbol{\Theta}$, where $\Omega = \{\boldsymbol{\theta} : q(\boldsymbol{\theta}) \text{ is bounded away from zero}\}$ to avoid regions with extremely low kernel values. Then we assume there is a function $h(\boldsymbol{\theta})$ such that $\int_\Omega h(\boldsymbol{\theta}) d\boldsymbol{\theta} = \Delta$ can be evaluated. Consequently, we have the identity:

$$\frac{\Delta}{c} = \int_\Omega \frac{h(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta})}{c} d\boldsymbol{\theta}. \tag{2.8}$$

We next partition the working parameter space into K subsets, where the ratio of $h(\boldsymbol{\theta})$ over $q(\boldsymbol{\theta})$ has similar values within each subset, to reduce the variance of the Monte Carlo estimator. The general form of the PWK estimator with unspecified local weights is essentially a weighted average for the harmonic mean estimator for $q(\boldsymbol{\theta})$ with the same weights assigned locally to an MCMC sample in a subset.

The working parameter space is essentially the constrained support considered by Robert and Wraith (2009) and Marin and Robert (2010). However, we do not require $h(\boldsymbol{\theta})$ to be a density function as in GHM or constrained GHM. Consequently, we allow a larger class of estimators to be considered.

2.3.1 General Monte Carlo Estimator

Suppose $\{A_1, \ldots, A_K\}$ forms a partition of the working parameter space Ω, where for an integer $K > 0$, $w_1, \ldots, w_K$ are the weights assigned to these K regions, respectively.

Let the weight function be the step function:

$$h(\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k 1\{\boldsymbol{\theta} \in A_k\}. \tag{2.9}$$

Evaluate Δ:

$$\Delta = \int_{\Omega} h(\boldsymbol{\theta}) d\boldsymbol{\theta} = \sum_{k=1}^{K} w_k V(A_k),$$

where $V(A_k)$ is the volume of the k^{th} subset in the partition, that is, $V(A_k) = \int_{\Omega} 1\{\boldsymbol{\theta} \in A_k\} d\boldsymbol{\theta}$.

Using the step function $h(.)$ in (2.9), the PWK estimator for $d \equiv 1/c$ is given by

$$\hat{d} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{w_k}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t \in A_k\}}{\sum_{k=1}^{K} w_k V(A_k)}. \tag{2.10}$$

In order to establish consistency and finite variance of the PWK estimator, we introduce two assumptions.

Assumption 1: The volume of each region $V(A_k) < \infty$ for $k = 1, 2, \ldots, K$.

Assumption 2: $q(\boldsymbol{\theta})$ is positive and continuous on $\overline{A}_k$, where $\overline{A}_k$ is the closure of A_k for $k = 1, \ldots, K$.

Theorem 2.3.1. *Under Assumptions 1 to 2 and certain ergodic conditions, $\hat{d}$ in (2.10) is a consistent estimator of $1/c$. In addition, Var$(\hat{d}) < \infty$.*

Proof:

$$\begin{aligned}
&\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{w_k}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t\in A_k\}\\
=&\sum_{k=1}^{K}w_k\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t\in A_k\}\\
=&\sum_{k=1}^{K}w_k\int_{\{\boldsymbol{\theta}\in A_k\}}\frac{1}{q(\boldsymbol{\theta})}\frac{q(\boldsymbol{\theta})}{c}d\boldsymbol{\theta}\\
=&d\sum_{k=1}^{K}w_kV(A_k),
\end{aligned}$$

which implies that $\hat{d}\xrightarrow{a.s.}1/c$. Let $q_{k,min}=\min_{\{\boldsymbol{\theta}_t\in A_k\}}q(\boldsymbol{\theta}_t)$. Under Assumption 2, we have $q_{k,min}>0$. Write $g(\boldsymbol{\theta}_t)=\sum_{k=1}^{K}w_k/q(\boldsymbol{\theta}_t)1\{\boldsymbol{\theta}_t\in A_k\}$. Under Assumptions 1 and 2, we have

$$\begin{aligned}
\mathrm{E}[g(\boldsymbol{\theta}_t)]^2=&\sum_{k=1}^{K}\mathrm{E}\Big(\Big[\frac{w_k}{q(\boldsymbol{\theta}_t)}\Big]1\{\boldsymbol{\theta}_t\in A_k\}\Big)^2\\
\le&\sum_{k=1}^{K}\frac{w_k^2}{q_{k,min}}\mathrm{E}\Big(\Big[\frac{1}{q(\boldsymbol{\theta}_t)}\Big]1\{\boldsymbol{\theta}_t\in A_k\}\Big)\\
\le&\sum_{k=1}^{K}\frac{w_k^2V(A_k)}{q_{k,min}c}<\infty, \qquad (2.11)
\end{aligned}$$

which implies that $\mathrm{Var}[g(\boldsymbol{\theta}_i)] < \infty$. Using Cauchy–Schwarz Inequality, we obtain

$$\begin{aligned}
\mathrm{Var}\Big[\frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{\theta}_t)\Big] =& \frac{1}{T^2}\mathrm{Var}\Big[\sum_{t=1}^{T} g(\boldsymbol{\theta}_t)\Big] \\
=& \frac{1}{T^2}\Big\{\sum_{t=1}^{T}\mathrm{Var}[g(\boldsymbol{\theta}_t)] + 2\sum_{t'<t''}\sum \mathrm{Cov}[g(\boldsymbol{\theta}_{t'}), g(\boldsymbol{\theta}_{t''})]\Big\} \\
\le& \frac{1}{T^2}\Big\{\sum_{t=1}^{T}\mathrm{Var}[g(\boldsymbol{\theta}_t)] + 2\sum_{t'<t''}\sum \sqrt{\mathrm{Var}[g(\boldsymbol{\theta}_{t'})]\mathrm{Var}[g(\boldsymbol{\theta}_{t''})]}\Big\}. \quad (2.12)
\end{aligned}$$

Thus, $\mathrm{Var}(\hat{d}) < \infty$ directly follows from (2.12). □

REMARK 3.1: Another property of $\hat{d}$ in (2.10) is that when a certain full conditional density is available, the computation can be lessened. This is often the case in the generalized linear model with latent variables or random effects, and in any Gibbs sampler or its hybrid. To be specific, let $(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)$ be 2 blocks of parameters, $\boldsymbol{\vartheta}_1 = (\theta_1, ..., \theta_q)'$ and $\boldsymbol{\vartheta}_2 = (\theta_{q+1}, ..., \theta_p)'$. Assume that a full conditional density, $\pi(\boldsymbol{\vartheta}_1|D, \boldsymbol{\vartheta}_2)$, is available. Then, the p-dimensional estimation problem can be reduced to $p - q$ dimensions:

$$\begin{aligned}
1 &= \int_{R^p} \frac{q(\boldsymbol{\theta})}{c} d\boldsymbol{\theta} \\
&= \int_{R^{p-q}}\int_{R^q} \frac{q(\boldsymbol{\vartheta}_2)\pi(\boldsymbol{\vartheta}_1|D, \boldsymbol{\vartheta}_2)}{c} d\boldsymbol{\vartheta}_1 d\boldsymbol{\vartheta}_2 \\
&= \int_{R^{p-q}} \frac{q(\boldsymbol{\vartheta}_2)}{c}\int_{R^q} \pi(\boldsymbol{\vartheta}_1|D, \boldsymbol{\vartheta}_2) d\boldsymbol{\vartheta}_1 d\boldsymbol{\vartheta}_2 \\
&= \int_{R^{p-q}} \frac{q(\boldsymbol{\vartheta}_2)}{c} d\boldsymbol{\vartheta}_2,
\end{aligned}$$

where $q(\boldsymbol{\vartheta}_2) = \int_{R^q} q(\boldsymbol{\theta}) d\boldsymbol{\vartheta}_1$, which has a closed form expression. Therefore, instead of investigating the kernel $q(\boldsymbol{\theta})$, we can work on the kernel $q(\boldsymbol{\vartheta}_2)$. In this case, (2.10)

becomes

$$\hat{d} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{w_k}{q(\boldsymbol{\vartheta}_{2_t})}1\{\boldsymbol{\vartheta}_{2_t} \in B_k\}}{\sum_{k=1}^{K} w_k V(B_k)},$$

where $\{B_1, ..., B_K\}$ is a partition of the working parameter space $\Omega_2, \Omega_2 \subset \boldsymbol{\Theta}_2$, which is the support of $q(\boldsymbol{\vartheta}_2)$, and $V(B_1), ..., V(B_K)$ are the corresponding volumes, respectively.

2.3.2 The Optimal Monte Carlo Estimation

Our next step is to find the optimal weight w_k in the class of PWK estimators (2.10), motivated by Chen and Shao (2002).

Theorem 2.3.2. *Assume* $\{\boldsymbol{\theta}_t,\ t = 1, \ldots, T\}$ *is an MCMC sample from the posterior distribution* $\pi(\boldsymbol{\theta}|D)$*, and let* $w_k^* = w_k / \Big[\sum_{k=1}^{K} w_k V(A_k)\Big]$ *and* $\alpha_k = E[(1/q^2(\boldsymbol{\theta}))1\{\boldsymbol{\theta} \in A_k\}]$*. Then,* $Var(\hat{d}) = \Big(\sum_{k=1}^{K} w_k^{*2}\alpha_k - 1/c^2\Big)/T$*. Moreover, the optimal variance=* $\Big\{1/\Big[\sum_{k=1}^{K} V^2(A_k)/\alpha_k\Big] - 1/c^2\Big\}/T$*, which is obtained by*

$$w_{k,opt}^* = V(A_k)/\Big\{\alpha_k\Big[\sum_{k=1}^{K} V^2(A_k)/\alpha_k\Big]\Big\}$$

for $k = 1, ..., K$*.*

Proof: First,

$$\begin{aligned}
\mathrm{Var}\left(\hat{d}\right) =& \mathrm{Var}\Big[\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{w_k^*}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t \in A_k\}\Big] \\
=& \frac{1}{T}\mathrm{Var}\Big[\sum_{k=1}^{K}\frac{w_k^*}{q(\boldsymbol{\theta})}1\{\boldsymbol{\theta} \in A_k\}\Big] \\
=& \frac{1}{T}\Big\{\mathrm{E}\Big[\sum_{k=1}^{K}\frac{w_k^*}{q(\boldsymbol{\theta})}1\{\boldsymbol{\theta} \in A_k\}\Big]^2 - \Big(\mathrm{E}\Big[\sum_{k=1}^{K}\frac{w_k^*}{q(\boldsymbol{\theta})}1\{\boldsymbol{\theta} \in A_k\}\Big]\Big)^2\Big\} \\
=& \frac{1}{T}\Big\{\mathrm{E}\Big[\sum_{k=1}^{K}\frac{w_k^{*2}}{q^2(\boldsymbol{\theta})}1\{\boldsymbol{\theta} \in A_k\}\Big] - \Big(\sum_{k=1}^{K}w_k^*\mathrm{E}\Big[\frac{1}{q(\boldsymbol{\theta})}1\{\boldsymbol{\theta} \in A_k\}\Big]\Big)^2\Big\} \\
=& \frac{1}{T}\Big\{\sum_{k=1}^{K}w_k^{*2}\mathrm{E}\Big[\frac{1}{q^2(\boldsymbol{\theta})}1\{\boldsymbol{\theta} \in A_k\}\Big] - \Big[\sum_{k=1}^{K}\frac{w_k^*V(A_k)}{c}\Big]^2\Big\} \\
=& \frac{1}{T}\Big\{\sum_{k=1}^{K}w_k^{*2}\alpha_k - \frac{1}{c^2}\Big\}.
\end{aligned}$$

Secondly, with the constraint $\sum_{k=1}^{K} w_k^* V(A_k) = 1$, the optimal weights directly follow from the Lagrange multiplier method,

$$\begin{aligned}
&\frac{\partial}{\partial w_k}\Big[\frac{1}{T}\Big(\sum_{k=1}^{K}w_k^{*2}\alpha_k - \frac{1}{c^2}\Big) - \lambda\Big(\sum_{k=1}^{K}w_k^*V(A_k) - 1\Big)\Big] = 0 \\
\Rightarrow& \frac{1}{T}(2w_k^*\alpha_k) - \lambda V(A_k) = 0 \\
\Rightarrow& w_k^* = \frac{T\lambda V(A_k)}{2\alpha_k}, \quad \text{for } k = 1, ..., K.
\end{aligned}$$

Replacing w_k^* by $T\lambda V(A_k)/(2\alpha_k)$ in the constraint, we can obtain $\lambda = 1/\Big[\sum_{k=1}^{K} TV^2(A_k)/(2\alpha_k)\Big]$ and $w_{k,opt}^* = TV(A_k)/\Big[2\alpha_k \sum_{k=1}^{K} TV^2(A_k)/(2\alpha_k)\Big] = V(A_k)/\Big[\alpha_k \sum_{k=1}^{K} V^2(A_k)/\alpha_k\Big]$, for $k = 1, ..., K$. So, $w_{k,opt}$ is proportion to $V(A_k)/\alpha_k$

for $k = 1, ..., K$. Under this setting, the variance can be simplified to

$$\text{Var}\left(\hat{d}\right) = \frac{1}{T}\Big\{\frac{1}{\sum_{k=1}^{K} V^2(A_k)/\alpha_k} - \frac{1}{c^2}\Big\}. \tag{2.13}$$

□

REMARK 3.2: In practice, it is quite difficult to estimate the second moment α_k. A very large sample size is required in order to obtain an accurate estimate of α_k. However, the results shown in Theorem 2.3.2 sheds light on the choices of $A_1, \ldots, A_K$ and w_k. First, it is only required that w_k is proportional to $V(A_k)/\alpha_k$. Second, if $q(\boldsymbol{\theta})$ is roughly constant over A_k, then $\alpha_k \approx V(A_k)/[q(\boldsymbol{\theta}_k^*)c]$, where $\boldsymbol{\theta}_k^* \in A_k$. Thus, in this case, we can simply choose $w_k = q(\boldsymbol{\theta}_k^*)$ and $\hat{d}$ in (2.10) reduces to

$$\hat{d} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{q(\boldsymbol{\theta}_k^*)}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t \in A_k\}}{\sum_{k=1}^{K} q(\boldsymbol{\theta}_k^*)V(A_k)}. \tag{2.14}$$

REMARK 3.3: Followed by the Remark 3.1, when a full conditional density $\pi(\boldsymbol{\vartheta}_1|D, \boldsymbol{\vartheta}_2)$ is available, the estimator $\hat{d}$ in (2.10) reduces further to

$$\hat{d} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{q(\boldsymbol{\vartheta}_2^*)}{q(\boldsymbol{\vartheta}_{2_t})}1\{\boldsymbol{\vartheta}_{2_t} \in B_k\}}{\sum_{k=1}^{K} q(\boldsymbol{\vartheta}_2^*)V(B_k)}.$$

REMARK 3.4: In practice, the marginal likelihood is often reported in log scale. Considering the dependence within the MCMC sample, we use the Overlapping Batch Statistics (OBS) of Schmeiser et al. (1990) to estimate the Monte Carlo (MC) standard error of $-\log(\hat{d})$. Let $\hat{\eta}_b$ denote an estimate of the reciprocal of the marginal likelihood in log scale using the b^{th} batch, $\{\boldsymbol{\theta}_t, t = b, b+1, \ldots, b+B-1\}$, of the MCMC sample

for $b = 1, 2, \ldots, T - B + 1$, where $B < T$ is the batch size. Then, the OBS estimated MC standard error of $\hat{\eta} = -\log(\hat{d})$ is given by

$$\sqrt{\widehat{\mathrm{Var}(\hat{\eta})}} = \left\{ \left[\frac{B}{T-B} \right] \frac{\sum_{b=1}^{T-B+1} (\hat{\eta}_b - \bar{\eta})^2}{T-B+1} \right\}^{\frac{1}{2}}, \tag{2.15}$$

where $\bar{\eta} = \sum_{b=1}^{T-B+1} \hat{\eta}_b/(T-B+1)$ and a batch size B is suggested to be $10 \leq T/B \leq 20$ in Schmeiser et al. (1990).

2.3.3 Construction of the Partition with Subsets $A_1, \ldots, A_K$

In order to make $q(\boldsymbol{\theta})$ roughly constant over A_k, for each k, which is a sufficient condition for the PWK estimator in (2.14) to be optimal, we provide the following rings approach for achieving it:

Step 1: Assume $\boldsymbol{\Theta}$ is R^p; if not, then a transformation $\boldsymbol{\phi} = G_1(\boldsymbol{\theta})$ is needed so that the parameter space of $\boldsymbol{\phi}$ is R^p.

Step 2: Use the MCMC sample to compute the mean $\overline{\boldsymbol{\phi}}$ and the covariance matrix $\widehat{\Sigma}$ of $\boldsymbol{\phi}$ and then standardize $\boldsymbol{\phi}$ by $\boldsymbol{\psi} = G_2(\boldsymbol{\phi}) = \widehat{\Sigma}^{-1/2}(\boldsymbol{\phi} - \overline{\boldsymbol{\phi}})$.

Step 3: Construct a working parameter space for $\boldsymbol{\psi}$ by choosing a reasonable radius r such that $\|\boldsymbol{\psi}\| < r$ for most of the standardized MCMC sample.

Step 4: Partition the working parameter space into a sequence of K spherical shells such that $A_k = \{\boldsymbol{\psi} : r(k-1)/K \leq \|\boldsymbol{\psi}\| < rk/K\}$, with $k = 1, \ldots, K$.

Step 5: Select a $\boldsymbol{\psi}_k^*$ in A_k as a representative point, for example, $\|\boldsymbol{\psi}_k^*\| = r[k/K - 1/(2K)]$.

Sept 6: Compute the new kernel value $\tilde{q}(\boldsymbol{\psi}_k^*) = q(G_1^{-1}G_2^{-1}(\boldsymbol{\psi}_k^*))|J|_{\boldsymbol{\psi}=\boldsymbol{\psi}_k^*}$, where $J = |\partial\boldsymbol{\theta}/\partial\boldsymbol{\phi}||\partial\boldsymbol{\phi}/\partial\boldsymbol{\psi}|$. Also compute the new kernel value $\tilde{q}(\boldsymbol{\psi}_t), t = 1, \ldots, T$, for the standardized MCMC sample.

Step 7: Estimate $d = 1/c$ by

$$\hat{d} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\tilde{q}(\boldsymbol{\psi}_k^*)}{\tilde{q}(\boldsymbol{\psi}_t)}1\{\boldsymbol{\psi}_t \in A_k\}}{\sum_{k=1}^{K}\tilde{q}(\boldsymbol{\psi}_k^*)V(A_k)}, \tag{2.16}$$

where $V(A_k) = \{(rk/K)^p - [r(k-1)/K]^p\}\pi^{p/2}/\Gamma(p/2+1)$.

REMARK 3.5: When K is big enough, $\tilde{q}(\boldsymbol{\psi}_t)$ in (2.16) will be roughly constant over A_k so that the best PWK estimate will be obtained. Besides, each kernel value $\tilde{q}(\boldsymbol{\psi}_t)$ is simply the original kernel value $q(\boldsymbol{\theta}_t)$ multiplied by the absolute value of Jacobian function.

2.4 Extension of the General PWK Estimator

In this section, we generalize the PWK estimator from the working parameter space to the full support space and from the locally constant weight function to a general weight function of $\boldsymbol{\theta}$. We call this class to be variable PWK (vPWK) estimators.

Suppose $\{A_1, \ldots, A_{K^*}\}$ is a partition of $\boldsymbol{\Theta}$, and $w_k(\boldsymbol{\theta})$ is a weight function defined on A_k. We need the following assumption to define this vPWK class:

Assumption 3 : The weight function w_k is integrable, that is, $\int |w_k(\boldsymbol{\theta})|d\boldsymbol{\theta} < \infty$ for $k = 1, \ldots, K^*$.

Under Assumption 3, we can extend the general PWK in (2.10) to a variable weighted

Monte Carlo estimator of $1/c$, which is given by

$$\hat{d}^* = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K^*}\frac{w_k(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t \in A_k\}}{\sum_{k=1}^{K^*}\int_{A_k} w_k(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{2.17}$$

Theorem 2.4.1. *Under Assumption 3 and $q(\boldsymbol{\theta}) > 0$, then the vPWK estimator $\hat{d}^*$ in (2.17) is a consistent estimator of $1/c$. In addition, if $\int_{A_k}[w_k(\boldsymbol{\theta})^2/q(\boldsymbol{\theta})]d\boldsymbol{\theta} < \infty$ for $k = 1, \ldots, K^*$, then $Var(\hat{d}^*) < \infty$.*

Proof: Under certain ergodic conditions and Assumption 3, the consistency property can be proven similarly as that in the general PWK estimator in (2.10). Specifically, we have

$$\begin{aligned}
&\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K^*}\frac{w_k(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t \in A_k\}\\
=&\sum_{k=1}^{K^*}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\frac{w_k(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t \in A_k\}\\
=&\sum_{k=1}^{K^*}\int_{\{\boldsymbol{\theta}\in A_k\}}\frac{w_k(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\frac{q(\boldsymbol{\theta})}{c}d\boldsymbol{\theta}\\
=&d\sum_{k=1}^{K^*}\int_{\{\boldsymbol{\theta}\in A_k\}} w_k(\boldsymbol{\theta})d\boldsymbol{\theta},
\end{aligned}$$

which implies that $\hat{d}^* \xrightarrow{a.s.} 1/c$. □

REMARK 4.1: It is easy to see that $\hat{d}$ in (2.10) is a special case of $\hat{d}^*$ in (2.17). When $K^* = K + 1$ and assigning an MCMC sample in each region with an equal weight, w_k, among which $w_{K^*} = 0$, $\hat{d}^*$ reduces to $\hat{d}$.

REMARK 4.2: The HM estimator is another special case of $\hat{d}^*$ in (2.17). When using

the prior $\pi(\boldsymbol{\theta}_i)$ as weights, the inverse of $\hat{d}^*$ is the HM estimator.

$$\begin{aligned}\hat{d}^*|_{w_k(\boldsymbol{\theta})=\pi(\boldsymbol{\theta})} &= \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K^*}\frac{\pi(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t \in A_k\}}{\sum_{k=1}^{K^*}\int_{A_k}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &= \frac{\frac{1}{T}\sum_{t=1}^{T}\frac{\pi(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)}\sum_{k=1}^{K^*}1\{\boldsymbol{\theta}_t \in A_k\}}{\int_{\boldsymbol{\Theta}}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{L(\boldsymbol{\theta}_t|D)}.\end{aligned}$$

REMARK 4.3: In addition, $\hat{d}^*$ in (2.17) includes the IDR estimator as a special case. Let $K^* = 2$, $A_1 = \{\boldsymbol{\theta} : \ ||\boldsymbol{\theta}|| \leq r\}$, $w_1(\boldsymbol{\theta}) = q(\mathbf{0}) - q(\boldsymbol{\theta})$, $A_2 = \{\boldsymbol{\theta} : \ ||\boldsymbol{\theta}|| > r\}$, and $w_2(\boldsymbol{\theta}) = q_r(\boldsymbol{\theta}) - q(\boldsymbol{\theta})$. We can show that $\int_{A_1} w_1(\boldsymbol{\theta})d\boldsymbol{\theta} = q(\mathbf{0})b_r - \int_{A_1} q(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $\int_{A_2} w_2(\boldsymbol{\theta})d\boldsymbol{\theta} = c - \int_{A_2} q(\boldsymbol{\theta})d\boldsymbol{\theta}$, implying $\sum_{k=1}^{2}\int_{A_k} w_k(\boldsymbol{\theta})d\boldsymbol{\theta} = q(\mathbf{0})b_r$. Thus, the inverse of $\hat{d}^*$ reduces to the IDR estimator. Note $w_1(\boldsymbol{\theta}_t)$ and $w_2(\boldsymbol{\theta}_t)$ in IDR are allowed to be negative.

2.5 Simulation Studies

2.5.1 A Bivariate Normal Example

We apply the PWK estimator for computing the normalizing constant of a bivariate normal distribution with the normal-inverse-Wishart prior. We consider both location and scale parameters to be unknown. Including the scale parameters makes computation challenging. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n)'$ be n observations from a bivariate normal distribution,

$$\boldsymbol{y}_i|\boldsymbol{\mu}, \Sigma \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}, \Sigma), i = 1, \ldots, n,$$

where $\boldsymbol{\mu} \in R^2$ and Σ are unknown parameters. The likelihood function is

$$L(\boldsymbol{\mu}, \Sigma|\boldsymbol{y}) = (2\pi)^{-n}|\Sigma|^{-n/2} \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\Big\}.$$

The prior for $\boldsymbol{\mu}$ and Σ is specified as follows:

$$\boldsymbol{\mu}|\Sigma \sim N(\boldsymbol{\mu}_0, \Sigma/\kappa_0) \text{ and } \Sigma \sim IW_{\nu_0}(\Lambda_0^{-1})$$

with hyperparameters $\boldsymbol{\mu}_0$, κ_0, ν_0, and Λ_0. Then, the joint posterior kernel is given by

$$\begin{aligned} q(\boldsymbol{\mu}, \Sigma) &= L(\boldsymbol{\mu}, \Sigma|\boldsymbol{y})\pi(\boldsymbol{\mu}|\Sigma)\pi(\Sigma) \\ &= (2\pi)^{-n}|\Sigma|^{-(n+\nu_0+2)/2-1}\frac{1}{\gamma} \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\Big\} \\ &\times \exp\Big\{ -\frac{\kappa_0}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\Big\} \exp\Big\{ -\frac{1}{2}trace(\Lambda_0\Sigma^{-1})\Big\} \end{aligned}$$

with $\gamma = 2^{\nu_0+1}\pi\Gamma_2(\nu_0/2)|\Lambda_0|^{-\nu_0/2}\kappa^{-1}$, where $\Gamma_2(\nu_0/2) = \pi^{1/2}\Gamma(\nu_0/2)\Gamma(\nu_0/2 - 1/2)$ is the bivariate gamma function. Under this setting, the analytical form of the normalizing constant is available as follows:

$$c = \frac{1}{\pi^n}\frac{\Gamma_2(\nu_n/2)}{\Gamma_2(\nu_0/2)}\frac{|\Lambda_0|^{\nu_0/2}}{|\Lambda_n|^{\nu_n/2}}(\frac{\kappa_0}{\kappa_n}), \tag{2.18}$$

where $\Lambda_n = \Lambda_0 + \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})' + \frac{\kappa_0 n}{\kappa_0+n}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{y}})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{y}})'$, $\kappa_n = \kappa_0 + n$, and $\nu_n = \nu_0 + n$. In the scenario, we set the hyperparameters $\boldsymbol{\mu}_0 = (0, 0)'$, $k_0 = 0.01$, $\nu_0 = 3$, and $\Lambda_0 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$. We generate a random sample $\boldsymbol{y}$ with $n = 200$ from a

bivariate normal distribution with $\boldsymbol{\mu} = (0, 0)$ and $\Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$. The corresponding sample mean $\bar{\boldsymbol{y}}$ is $(-0.029, 0.040)'$, and the sample variance-covariance matrix S is $\begin{pmatrix} 201.987 & 143.330 \\ 143.330 & 192.365 \end{pmatrix}$. Using equation (2.18), the marginal likelihood in log scale is -507.278. In this example, in order to apply the spherical shell approach in Section 3.3, a transformation of Σ is needed. Here, we use the log transformation for each variance parameter and the Fisher z-transformation for the correlation coefficient parameter to have unbounded support for each of them. Then, we standardize each transformed MCMC sample from its transformed sample mean and standard deviation. In the new parameter space, we construct the working parameter space and its partition by choosing $r = 1.5$, 2, or 2.5 and $K = 10$, 20, or 100. After picking up a representative point in each spherical shell, we estimate $d = 1/c$ using equation (2.16). We compare our method to the HM and IDR methods based on $M = 1,000$ independent MCMC samples with $T = 1,000$ or $T = 10,000$ in Table 2.1. Let $\hat{d}_m$ be the estimate of d based on the m^{th} MCMC sample for $m = 1, 2, \ldots, M$. Then, the simulation estimate (Mean), the MC standard error (MCSE), and the root mean squared error (RMSE) of the estimates in log scale are defined as $\widehat{\log c} = \frac{1}{M}\sum_{m=1}^{M}(-\log \hat{d}_m)$, $\{\frac{1}{M-1}\sum_{m=1}^{M}(-\log \hat{d}_m - \widehat{\log c})^2\}^{1/2}$, and $\{\frac{1}{M}\sum_{m=1}^{M}(-\log \hat{d}_m - \log c)^2\}^{1/2}$, respectively.

Table 2.1: Simulation results for the bivariate normal case

$\log c = -507.2776$									
			T=1,000			T=10,000			**Time (sec.)**
	K	r	**Mean**	**MCSE**	**RMSE**	**Mean**	**MCSE**	**RMSE**	
HM			-494.671	0.908	12.639	-495.142	0.762	12.159	0.644
IDR		1.5	-509.064	0.302	1.811	-509.123	0.145	1.851	1.638
		2.0	-509.095	0.537	1.895	-509.284	0.387	2.043	1.634
		2.5	-508.926	0.710	1.795	-509.216	0.629	2.038	1.621
PWK	10	1.5	-507.260	0.064	0.067	-507.264	0.020	0.025	0.329
		2.0	-507.260	0.057	0.059	-507.264	0.018	0.022	0.596
		2.5	-507.259	0.057	0.060	-507.264	0.019	0.023	0.784
	20	1.5	-507.260	0.064	0.066	-507.264	0.020	0.024	0.327
		2.0	-507.262	0.053	0.055	-507.264	0.016	0.021	0.596
		2.5	-507.259	0.055	0.058	-507.264	0.018	0.023	0.792
	100	1.5	-507.260	0.064	0.066	-507.264	0.020	0.024	0.426
		2.0	-507.261	0.052	0.054	-507.264	0.016	0.021	0.660
		2.5	-507.260	0.055	0.058	-507.264	0.018	0.022	0.877

Table 2.1 shows the results, where the average computing time (in seconds) per MCMC sample on an Intel i7 processor machine with 12 GB of RAM memory using a Windows 8.1 operating system is given in the last column. From Table 2.1, we see that (i) PWK has the best performance with much smaller MCSE's and RMSE's than HM and IDR under both $T = 1,000$ and $T = 10,000$; (ii) when T increases, the MCSE's and the RMSE's of the PWK estimator becomes smaller under all choices of r's and K's; (iii) the performance of the HM estimator slightly improves but the IDR estimator does not when T increases; and (iv) the computing time of the PWK estimator is comparable to that of the HM estimator while the IDR estimator requires the most computing time. It is interesting to mention that the MCSE's and the RMSE's of the PWK estimator are very similar for all choices of r's and K's

under each T, implying the robustness of the PWK estimator with respect to the specification of the working parameter space and the number of partition subsets.

To evaluate the effect of the vague prior on the precision of the PWK estimator, we extend our simulation study by considering different values of hyperparameters κ_0 and ν_0. Note that the value of $\log c$ in Table 2.1 is computed under $\kappa_0 = 0.01$ and $\nu_0 = 3$, which corresponds to a relatively vague prior for $(\boldsymbol{\mu}, \Sigma)$. Table 2.2 shows the simulation results of the PWK estimators with $r = 2$ and $K = 100$ for $(\kappa_0, \nu_0) =$ (0.0001, 3), (1, 3), and (1, 10) in addition to (0.01, 3). From Table 2.2, we see that the MCSE's under these different values of (κ_0, ν_0) are almost the same while these RMSE's are comparable except the last one with $(\kappa_0, \nu_0) = (1, 10)$, in which the RMSE's are slightly larger.

Table 2.2: Simulation results of PWK estimators for different hyperparameters κ_0 and ν_0

			T=1,000			T=10,000		
κ_0	ν_0	$\log c$	**Mean**	**MCSE**	**RMSE**	**Mean**	**MCSE**	**RMSE**
0.0001	3	-511.883	-511.866	0.052	0.054	-511.869	0.016	0.021
0.01	3	-507.278	-507.261	0.052	0.054	-507.264	0.016	0.021
1	3	-502.682	-502.665	0.052	0.054	-502.669	0.016	0.021
1	10	-512.773	-512.721	0.053	0.074	-512.725	0.016	0.050

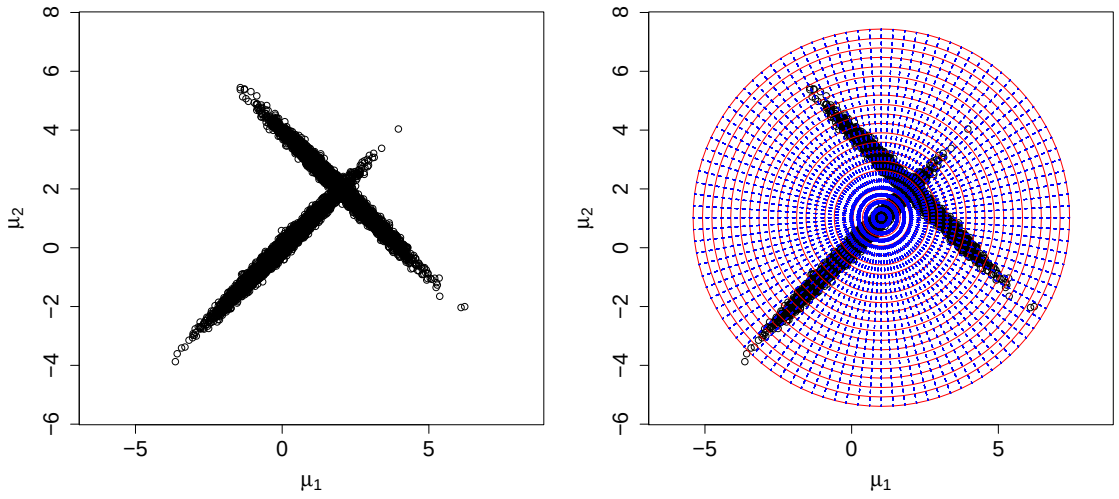
2.5.2 A Mixture Normal Example

To further evaluate the performance of the PWK, we consider a two-dimensional normal mixture in Chen et al. (2006) as follows

$$\pi(\boldsymbol{\mu}) = \sum_{j=1}^{2} \frac{1}{2}\Big[\frac{1}{2\pi}|\Sigma_j|^{-1/2} \exp\Big\{ -\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_{0j})'\Sigma_j^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_{0j})\Big\}\Big], \tag{2.19}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, $\boldsymbol{\mu}_{01} = (0, 0)'$, $\boldsymbol{\mu}_{02} = (2, 2)'$ and $\Sigma_j = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_j \\ \sigma_1\sigma_2\rho_j & \sigma_2^2 \end{pmatrix}$ with $\sigma_1 = \sigma_2 = 1$, $\rho_1 = 0.99$, and $\rho_2 = -0.99$. Figure 2.1(a) is a scatter plot of a random sample with $T = 10,000$ generated from (2.19). Based on the random sample, we apply the PWK to estimate the normalizing constant in (2.19), which is known to be 1. Due to the high but opposite correlations (i.e., $\rho_1 = 0.99$ and $\rho_2 = -0.99$), $\pi(\boldsymbol{\mu})$ cannot be homogeneous over a partition ring formed by the spherical shell approach in Section 3.3. To circumvent this difficulty, we additionally slice (dash lines) the existing partition rings by dividing equally along the angle from 0 to 360 degrees as shown in Figure 2.1(b), where the center of circle is the sample posterior mean (denoted as $\hat{\boldsymbol{\mu}}$). Now, the heterogeneity of $\pi(\boldsymbol{\mu})$ over each partition subset is effectively eliminated by this additional slicing step.

Table 2.3 shows the results of HM, IDR, and PWK estimators based on $M = 1,000$ independent random samples with $T = 1,000$ or $T = 10,000$ from (2.19). For PWK, we consider different K's (number of rings $\times$ number of slices) and r's (75%, 90%, or 95% $\times$ $\max_{1 \le t \le T} ||\boldsymbol{\mu}_t - \hat{\boldsymbol{\mu}}||$). We use the same values of r's for both IDR and PWK. From Table 2.3, we see that (i) the RMSE's of PWK are considerably smaller than those of HM and IDR; (ii) the performance of PWK improves when the sample size (T) or the number of partition subsets (K) increases; and (iii) PWK takes slightly longer computing time than HM and IDR.



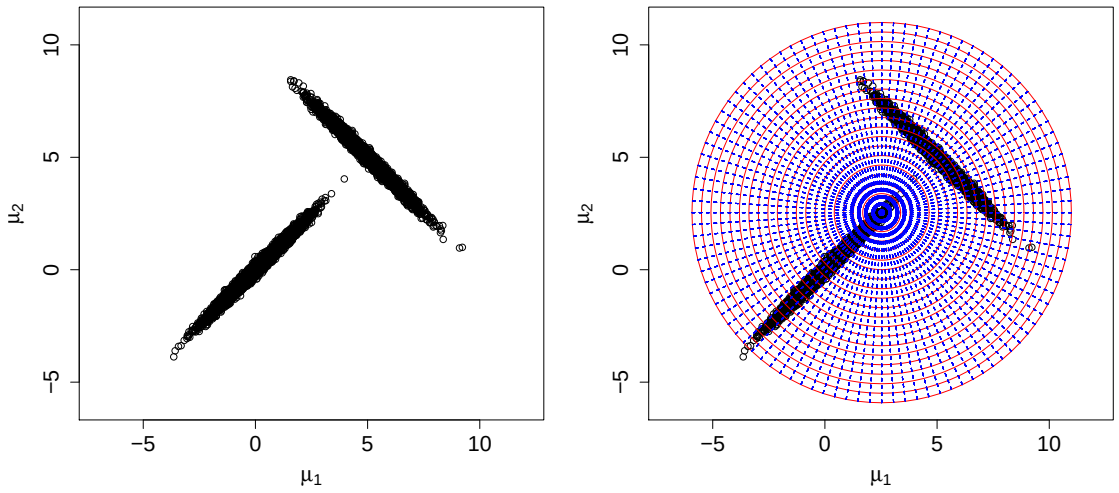
(a) (b)

Figure 2.1: Forming the working parameter space and its partition for a mixture normal distribution with means (0,0) and (2,2).

Table 2.3: Simulation results for the mixture normal with means equal to (0,0) and (2,2)

$\log c = 0$									
			T=1,000			T=10,000			**Time (sec.)**
	K	r	**Mean**	**MCSE**	**RMSE**	**Mean**	**MCSE**	**RMSE**	
HM			-2.868	0.685	2.948	-3.069	0.519	3.113	1.647
IDR		5.065	1.879	0.639	1.985	1.706	0.448	1.764	1.680
		6.078	2.149	0.650	2.245	1.935	0.485	1.995	1.839
		6.415	2.243	0.659	2.337	2.015	0.485	2.073	1.717
PWK	20×100	5.065	0.001	0.020	0.020	0.000	0.006	0.006	2.167
		6.078	0.000	0.025	0.025	0.000	0.008	0.008	2.375
		6.415	0.000	0.025	0.025	-0.001	0.008	0.008	2.187
	100×100	5.065	0.000	0.011	0.011	0.000	0.003	0.003	2.933
		6.078	0.000	0.011	0.011	0.000	0.004	0.004	3.037
		6.415	0.000	0.011	0.011	0.000	0.004	0.004	2.929

Next, we consider a more challenging case, where $\boldsymbol{\mu}_{02}$ is replaced by $(5,5)'$ so that the two modes are much far away from each other. Figure 2.2 (a) is a scatter plot of a random sample with $T = 10,000$ and Figure 2.2 (b) shows the partition subsets of the chosen working parameter space.



(a) (b)

Figure 2.2: Forming the working parameter space and its partition for a mixture normal distribution with means (0,0) and (5,5).

Table 2.4 summarizes the simulation results with the same simulation setting as before. We see that PWK outperforms both HM and IDR under this more challenging case. As expected, the RMSE's in Table 2.4 are larger than those in Table 2.3 for all three methods. However, the RMSE's of the PWK estimator are still quite small when K and T are reasonably large.

Table 2.4: Simulation results for the mixture normal with means equal to (0,0) and (5,5)

$\log c = 0$									
			T=1,000			T=10,000			**Time (sec.)**
	K	r	**Mean**	**MCSE**	**RMSE**	**Mean**	**MCSE**	**RMSE**	
HM			-2.915	0.681	2.993	-3.107	0.500	3.147	1.728
IDR		6.675	2.340	1.586	2.825	2.791	1.695	3.263	1.763
		8.011	1.658	1.780	2.429	2.409	1.350	2.760	1.730
		8.456	1.568	1.985	2.524	2.216	1.430	2.636	1.822
PWK	20×100	6.675	-0.001	0.035	0.035	0.000	0.011	0.011	2.277
		8.011	0.003	0.060	0.060	0.000	0.019	0.019	2.253
		8.456	0.000	0.060	0.060	0.000	0.018	0.018	2.374
	100×100	6.675	0.000	0.018	0.018	0.000	0.006	0.006	3.022
		8.011	0.000	0.018	0.018	0.000	0.006	0.006	2.933
		8.456	0.000	0.019	0.019	0.000	0.006	0.006	3.114

2.6 Application of the PWK to Real Data Examples

2.6.1 The Ordinal Probit Regression Model

In the first example, we apply the PWK method to compute the marginal likelihood under the ordinal probit regression model. Let $\boldsymbol{y} = (y_1, y_2, ..., y_n)'$ denote the vector of observed ordinal responses, each is coded as one value from $0, 1, ..., J-1$, $\boldsymbol{X}$ denote the $n \times p$ covariate matrix with the i^{th} row equal to the covariate of the i^{th} subject $\boldsymbol{x}_i'$, and $\boldsymbol{u} = (u_1, u_2, ..., u_n)'$ denote the vector of latent random variables. We consider

the following hierarchical model as in Albert and Chib (1993) such that

$$y_i = j, \text{ if } \gamma_j \leq u_i < \gamma_{j+1}$$

and

$$u_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

where $j = 0, 1, ..., J-1$, $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients, and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Based on the reparameterization of Nandram and Chen (1996), the cutpoints for dividing the latent variable u_i can be specified as $-\infty = \gamma_0 < \gamma_1 = 0 \leq \gamma_2 \leq \cdots \leq \gamma_{J-1} = 1 < \gamma_J = \infty$. Under this setting, the likelihood function is given by in Chen (2005b)

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} \big[\Phi\big(\frac{\gamma_{y_{i+1}} - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\big) - \Phi\big(\frac{\gamma_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\big)\big],$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma, \gamma_2, \ldots, \gamma_{J-2})'$ if $J \geq 4$, otherwise, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma)'$, and $\Phi(.)$ is the cumulative standard normal distribution function. Then, we specify normal, inverse gamma, and uniform priors for the parameters $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\gamma}$, respectively.

To examine the performance of the PWK estimator under this model, we consider the prostate cancer data of $n = 713$ patients as in Chen (2005b). In this data set, Pathological Extracapsular Extension (PECE, y) is a clinical ordinal response variable, and Prostate Specific Antigen (PSA, x_1), Clinical Gleason Score (GLEAS, x_2), and Clinical Stage (CSTAGE, x_3) are three covariates. PECE takes values of 0, 1, or 2, where 0 means that there is no cancer cell present in or near the capsule, 1 denotes that the cancer cells extend into but not through the capsule, and 2 indicates that cancer cells extend through the capsule. PSA and GLEAS are continuous variables

while CSTAGE is a binary outcome, which was assigned to 1 if the 1992 American Joint Commission on cancer clinical stage T-category was 1, and assigned to 2 if the T-category was 2 or higher.

In this application, $J = 3$ so that all four cutpoints can be assigned to fixed values: $-\infty = \gamma_0 < \gamma_1 = 0 < \gamma_2 = 1 < \gamma_3 = \infty$. Then, the prior distribution is specified as

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2),$$

where $\boldsymbol{\beta}|\sigma^2 \sim N(\mathbf{0}, 10\sigma^2 I_4)$ and $\sigma^2 \sim IG(a_0 = 1, b_0 = 0.1)$, an inverse gamma distribution with density proportional to $(\sigma^2)^{-(a_0+1)} \exp(-b_0/\sigma^2)$.

The marginal likelihood is not analytically available. Nevertheless, the estimates of this are obtained in Table 1 of Chen (2005b) using the method proposed by Chen (called Chen's method) and the method proposed by Chib (1995) (called Chib's method). Chen's method needs only a single MCMC sample from the joint posterior distribution $\pi(\boldsymbol{\beta}, \sigma^2|D)$. However, Chib's method with two blocks requires an additional MCMC sample from the conditional posterior distribution $\pi(\sigma^2|\boldsymbol{\beta}^*, D)$, where $\boldsymbol{\beta}^*$ is the posterior mean of $\boldsymbol{\beta}$. We compare the PWK method with $r = 3.327$ and $K = 100$ to these two methods under the same MCMC sample sizes $T = 2,500$, or $5,000$, except Chib's method doubles them. We have obtained the PWK estimates $\log \hat{c} = -758.70$, or -758.70 respectively with corresponding OBS estimated MCSE to be 0.020, or 0.016. So we observe that the PWK estimates are comparable to that of the other two methods and the PWK method has the smallest OBS estimated MCSE among the three methods.

For the PWK, the log transformation of σ^2 is needed. Then, after the standardization of the transformed MCMC sample, we consider $K = 10$, 20, and 100 and

$r = 0.75\sqrt{\chi^2_{5,0.95}}$, $\sqrt{\chi^2_{5,0.95}}$, and $1.25\sqrt{\chi^2_{5,0.95}}$ to investigate robustness of the PWK estimates with respect to these choices. We note that $\sqrt{\chi^2_{5,0.95}}$ is chosen based on Yu et al. (2015). Table 2.5 shows the PWK estimates and the corresponding estimated MCSE (eMCSE) under the MCMC samples with T = 2,500 and 5,000, where eMCSE is computed using (2.15) with $T/B = 10$. We note that we use the same MCMC sample sizes as in Chen (2005b). The results show the PWK estimators are relatively robust to the choices of the radius r and number of partitioned subsets K.

Table 2.5: The PWK estimates of the marginal likelihood for the prostate cancer data

$r = 0.75\sqrt{\chi^2_{5,0.95}}$						
	PWK (K=10)		PWK (K=20)		PWK (K=100)	
T	$-\log \hat{d}$	eMCSE	$-\log \hat{d}$	eMCSE	$-\log \hat{d}$	eMCSE
$2,500$	-758.73	0.026	-758.73	0.025	-758.73	0.025
$5,000$	-758.70	0.021	-758.70	0.020	-758.70	0.020

$r = \sqrt{\chi^2_{5,0.95}} = 3.327$						
	PWK (K=10)		PWK (K=20)		PWK (K=100)	
T	$-\log \hat{d}$	eMCSE	$-\log \hat{d}$	eMCSE	$-\log \hat{d}$	eMCSE
$2,500$	-758.70	0.020	-758.70	0.019	-758.70	0.020
$5,000$	-758.70	0.016	-758.70	0.016	-758.70	0.016

$r = 1.25\sqrt{\chi^2_{5,0.95}}$						
	PWK (K=10)		PWK (K=20)		PWK (K=100)	
T	$-\log \hat{d}$	eMCSE	$-\log \hat{d}$	eMCSE	$-\log \hat{d}$	eMCSE
$2,500$	-758.69	0.020	-758.69	0.019	-758.69	0.017
$5,000$	-758.70	0.018	-758.70	0.015	-758.69	0.014

2.6.2 Analysis of ECOG Data

In this subsection, we apply the PWK estimator to the problem of determining the power prior based on the historical data for the current analysis. Assume we have

conducted two clinical trials for the same objective. A natural way to combine these two trials is to consider the power prior setting, which allows us to borrow information from the historical data to construct the prior for the current analysis. Assume we have an initial prior for the unknown parameters which is determined before observing the historical data. To quantify the heterogeneity between the current data and the historical data, the power prior weights the historical likelihood function by the power a_0, where $0 \leq a_0 \leq 1$, to indicate the degree of incorporating the historical likelihood to the initial prior. Our objective is to find the optimal a_0 which maximizes the marginal likelihood for the current data with respect to the power prior. Ibrahim et al. (2015) point out the difficulty of finding this solution except for normal linear regression models. Therefore, they resolve to using the deviance information criterion (DIC) and the logarithm of pseudo marginal likelihood (LPML) criterion for constructing the parameter a_0 of the power prior in Ibrahim et al. (2012, 2015). To evaluate DIC, we need to plug the MCMC sample into the sum of the log likelihood over all data points; to evaluate LPML, we need to take the sum of the log transformation of each CPO, where the i^{th} CPO is the harmonic mean of the i^{th} likelihood evaluated at the MCMC sample from the posterior distribution based on the full sample. Both methods yield much less computational burden than the marginal likelihood method. We will show how the PWK estimator can circumvent the computational burden in evaluating the marginal likelihood.

The effectiveness of Interferon Alpha-2b (IFN) in immunotherapy for melanoma patients has been evaluated by two observation-controlled clinical trials: Eastern Cooperative Oncology Group (ECOG) phase III, E1684, followed by E1690. The first trial E1684 was conducted with 286 patients randomly assigned to either IFN or Observation. The IFN arm demonstrated a significantly better survival curve, but

with substantial side effects due to high dose regimen. To confirm the results of the E1684 and the benefit of IFN at a lower dosage, a later trial E1690 was conducted with three arms: high dose IFN, low dose IFN, and Observation. We use the data in E1684 as the historical data and a subset (high dose arm and Observation) of the E1690 trial as our current data. There are 427 patients in this subset.

For $n = 427$ patients in the current trial (E1690), we follow the model in Chen et al. (1999). Let y_i denote the survival time for the i^{th} patient, ν_i denote the censoring status, which is equal to 1 if y_i is a failure time and to 0 if it is the right censored, $\boldsymbol{x}_i = (1, \text{trt}_i)'$ denote the vector of covariates, where $\text{trt}_i = 1$ if the i^{th} patient received IFN and $\text{trt}_i = 0$ if the i^{th} patient was assigned to Observation. Then, the likelihood function is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D) = \prod_{i=1}^{n} \Big\{ \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) f(y_i|\boldsymbol{\lambda}) \Big\}^{\nu_i} \exp\{-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) F(y_i|\boldsymbol{\lambda}))\}, \tag{2.20}$$

where $D = (n, \boldsymbol{y}, \boldsymbol{\nu}, X)$ is the observed current data, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, and $F(y|\boldsymbol{\lambda})$ is the cumulative distribution function and $f(y|\boldsymbol{\lambda})$ is the corresponding density function. In (2.20), we use the same piecewise exponential model for $F(y|\boldsymbol{\lambda})$ as Ibrahim et al. (2012), which is given by

$$F(y|\boldsymbol{\lambda}) = 1 - \exp\Big\{ -\lambda_j(y - s_{j-1}) - \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1}) \Big\},$$

where $s_{j-1} \leq y < s_j$, $s_0 = 0 < s_1 < s_2 < \ldots < s_5 = \infty$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_5)'$.

For $n_0 = 286$ patients in the historical trial (E1684), we attempt to extract some of its information to set up the prior distribution for the current analysis. Similarly, we let y_{0i} denote the survival time for the i^{th} patient, ν_{0i} denote the censoring status,

and $\boldsymbol{x}_{0i} = (1, \text{trt}_{0i})'$ denote the vector of covariates. So $D_0 = (n_0, \boldsymbol{y}_0, \boldsymbol{\nu}_0, X_0)$ is the observed historical data. Assume $\pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is an initial prior. Here, we specify an initial proper prior $N(0, 100I_2)$ for $\boldsymbol{\beta}$ and $Exp(\lambda_0 = 1/100)$ (λ_0: rate parameter) for each $\lambda_j, j = 1, \ldots, 5$, to come close to the flat prior in Ibrahim et al. (2012). To update the initial prior with the historical data, the power prior is intuitively set as the initial prior π_0 multiplied by the historical likelihood function with power a_0 as follows:

$$\begin{aligned}
&\pi(\boldsymbol{\beta}, \boldsymbol{\lambda} | D_0, a_0) \\
&\propto \Big[\prod_{i=1}^{n_0} \Big\{ \exp(\boldsymbol{x}_{0i}'\boldsymbol{\beta}) f(y_{0i}|\boldsymbol{\lambda}) \Big\}^{\nu_{0i}} \exp\{-\exp(\boldsymbol{x}_{0i}'\boldsymbol{\beta}) F(y_{0i}|\boldsymbol{\lambda}))\} \Big]^{a_0} \pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda}), \qquad (2.21)
\end{aligned}$$

where $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0, a_0)$ is called the power prior and $0 \leq a_0 \leq 1$. In this setting, we can see when $a_0 = 0$, the power prior is exactly equal to the initial prior, which integrates to be 1, and when $a_0 \neq 0$, the power prior is equal to the right-hand side kernel function in (2.21) divided by $c_0 = \int L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0)^{a_0} \pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda}) d\boldsymbol{\beta} d\boldsymbol{\lambda}$. Combining the likelihood function in (2.20) and the power prior in (2.21), the posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ given (D, D_0, a_0) will be

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D, D_0, a_0) \propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D) \pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0, a_0). \qquad (2.22)$$

In this framework, we compare the marginal likelihoods of $L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0, a_0)$ for $0 \leq a_0 \leq 1$. The one with the highest marginal likelihood is our final model, and its corresponding a_0 determines the power prior.

However, as we point out earlier, except for $a_0 = 0$, $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0, a_0)$ is known up to a normalizing constant c_0. Hence, a two-step evaluation is needed to obtain the

marginal likelihood:

$$\begin{aligned} c &= \int L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0, a_0)d\boldsymbol{\beta}d\boldsymbol{\lambda} \\ &= \frac{\int L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0)^{a_0}\pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda})d\boldsymbol{\beta}d\boldsymbol{\lambda}}{\int L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0)^{a_0}\pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda})d\boldsymbol{\beta}d\boldsymbol{\lambda}} \\ &= \frac{c_1}{c_0} = \frac{d_0}{d_1}. \end{aligned}$$

We apply the PWK to estimate the numerator, $L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0)^{a_0}\pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda})$, and the denominator, $L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0)^{a_0}\pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda})$, respectively.

For each choice of a_0 with an increment of 0.1 from 0 to 1, an MCMC sample size is fixed at 10,000. The log transformation of each λ_j is needed. After the standardization of transformed MCMC sample, we choose the maximum radius $r = 3.751$ and the number of spherical shells $K = 100$. By equations (2.16) and (2.15), we can obtain the marginal likelihood estimate and its eMCSE in each chosen a_0. We summarize the results in Table 2.6. Table 2.6 also includes the PWK estimates under different choices of r's and K's.

Table 2.6: PWK estimates for marginal likelihood with different power priors under different choices of r's and K's

$r = 0.75\sqrt{\chi^2_{7,0.95}}$						
	K=10		K=20		K=100	
a_0	$\ln(\hat{d_0}/\hat{d_1})$	eMCSE	$\ln(\hat{d_0}/\hat{d_1})$	eMCSE	$\ln(\hat{d_0}/\hat{d_1})$	eMCSE
0.0	-552.717	0.028	-552.713	0.026	-552.709	0.028
0.1	-523.619	0.055	-523.614	0.051	-523.621	0.053
0.2	-522.091	0.044	-522.078	0.044	-522.073	0.044
0.3	-521.408	0.043	-521.420	0.043	-521.419	0.043
0.4	-521.336	0.046	-521.332	0.047	-521.338	0.045
0.5	-521.201	0.057	-521.229	0.060	-521.229	0.060
0.6	-521.189	0.037	-521.202	0.034	-521.187	0.033
0.7	-521.356	0.050	-521.363	0.044	-521.353	0.044
0.8	-521.553	0.054	-521.558	0.056	-521.576	0.058
0.9	-521.592	0.061	-521.618	0.051	-521.612	0.050
1.0	-521.702	0.052	-521.724	0.055	-521.732	0.050

$r = \sqrt{\chi^2_{7,0.95}} = 3.751$						
	K=10		K=20		K=100	
a_0	$\ln(\hat{d_0}/\hat{d_1})$	eMCSE	$\ln(\hat{d_0}/\hat{d_1})$	eMCSE	$\ln(\hat{d_0}/\hat{d_1})$	eMCSE
0.0	-552.732	0.022	-552.707	0.025	-552.708	0.027
0.1	-523.633	0.059	-523.646	0.049	-523.624	0.054
0.2	-522.098	0.052	-522.093	0.050	-522.077	0.045
0.3	-521.433	0.039	-521.432	0.040	-521.417	0.043
0.4	-521.309	0.046	-521.321	0.048	-521.339	0.043
0.5	-521.179	0.062	-521.187	0.059	-521.230	0.059
0.6	-521.186	0.039	-521.174	0.037	-521.187	0.033
0.7	-521.365	0.034	-521.361	0.042	-521.349	0.044
0.8	-521.535	0.055	-521.568	0.056	-521.573	0.056
0.9	-521.627	0.047	-521.613	0.055	-521.613	0.050
1.0	-521.746	0.059	-521.739	0.049	-521.732	0.050

$r = 1.25\sqrt{\chi^2_{7,0.95}}$						
	K=10		K=20		K=100	
a_0	$\ln(\hat{d_0}/\hat{d_1})$	eMCSE	$\ln(\hat{d_0}/\hat{d_1})$	eMCSE	$\ln(\hat{d_0}/\hat{d_1})$	eMCSE
0.0	-552.740	0.039	-552.719	0.033	-552.708	0.027
0.1	-523.551	0.057	-523.622	0.052	-523.622	0.053
0.2	-522.105	0.045	-522.077	0.044	-522.071	0.045
0.3	-521.427	0.048	-521.422	0.045	-521.421	0.042
0.4	-521.311	0.048	-521.317	0.046	-521.335	0.044
0.5	-521.239	0.052	-521.232	0.057	-521.227	0.059
0.6	-521.186	0.037	-521.171	0.033	-521.184	0.032
0.7	-521.381	0.047	-521.376	0.045	-521.350	0.043
0.8	-521.569	0.067	-521.578	0.063	-521.578	0.057
0.9	-521.597	0.052	-521.621	0.054	-521.609	0.049
1.0	-521.705	0.060	-521.740	0.046	-521.730	0.051

Note the marginal likelihood function c can be shown to be continuous in a_0. Therefore, from the results in Table 2.6, we see that the best choice of a_0 is between 0.5 and 0.6 under the marginal likelihood criterion. This result is quite comparable to the result of $a_0 = 0.4$ in Ibrahim et al. (2012) obtained by DIC and LPML criteria, where a suitable marginal likelihood computation was not accessible to them at the time. We also observe that the results are quite robust to the different r's and K's, and all point out the best choice of a_0 is between 0.5 and 0.6.

2.7 Discussion

The marginal likelihood is often analytically intractable due to the complicated kernel structure. Nevertheless, an MCMC sample from the posterior distribution is readily available from Bayesian computing software, for example, MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), Beast (Drummond et al., 2012), and Phycas (Lewis et al., 2015) in Bayesian phylogenetics. Additionally, the likelihood values evaluated at the MCMC sample are outputted in a file. Consequently, we can produce kernel values easily using the output and the prior function. In this chapter, we propose an easily implemented algorithm PWK for estimating the marginal likelihood based on this single chain of the MCMC sample and its corresponding kernel values. Unlike some existing algorithms requiring knowing the structure of the kernel, which is rare in Bayesain phylogenetics, we only need to know the kernel values evaluated at the MCMC sample. Therefore, our algorithm can be applied to model selection in Bayesian phylogenetics with a fixed topology. It may have potential for the variable topology problems.

We extend our PWK to the variable PWK that can handle the parameter space with full support and we show the HM and IDR are special cases of this vPWK. We conduct a simulation study from a bivariate normal model with 5 parameters in a Bayesian conjugate prior inference problem to compare our estimator to HM and IDR; our results show the PWK has the smallest empirical SE and RMSE. The computation time for our method is only slightly longer than that for the HM which indicates our spherical shell partition approach is very efficient.

In real data analysis, we first consider an ordinal probit regression model with a latent variable structure, and compare our method to that in Chib (1995) and Chen (2005b). We find the three methods are comparable and our method has the smallest MCSE. In the second example, we consider a cure rate survival model with the piecewise constant baseline hazard function and a power prior construction based on two clinical trial data sets. We obtain the optimal power prior using the marginal likelihood criterion as opposed to the DIC and LPML methods considered by Ibrahim et al. (2012). Although we obtain similar results, except we prefer borrowing more of the historical data, it would be interesting to investigate the effects of the criterion and the initial prior on the choice of a_0. We implemented our methodology using the R programming language (Team (2014)).

In an unimodal problem, we suggest using the square root of the 95^{th} percentile in Chi-square distribution with p degree of freedom as the guide value (r_{GV}) of radius r for constructing the working parameter space of the standardized MCMC sample. This is because after standardizing the MCMC sample, each parameter can be marginally viewed as a normal distribution. Although the results are quite robust to the choices of r's as shown in simulation and case studies, this way can insure that we can make use of most of the MCMC sample and avoid the region with posterior density close

to 0. For a multimodal problem, we suggest using $95\% \times \max_{1 \leq t \leq T} ||\boldsymbol{\mu}_t - \hat{\boldsymbol{\mu}}||$ as r_{GV} for constructing the working parameter space of the transformed MCMC sample. Since this approach may include many place with extremely small posterior density in the working parameter space, we propose an advanced spherical rings approach with additional slices on the partition rings to assure the homogeneity of the MCMC sample in each subset. This new partition approach can also be extended to n-dimensional problem by introducing another $n - 2$ angular coordinates.

Chapter 3

Marginal Posterior Density Estimation

3.1 Introduction

Posterior density estimation is one of the most important topics in Bayesian inference because it provides the complete information about parameters of interest in the model. Chen (2005a) provides an overview of the usefulness of the posterior density estimation in computing Bayes factors, marginal likelihoods, and posterior model probabilities. Posterior density estimation has also been used in various applications, including selection of the best predictors for the development of AIDS or death using historical data (Chen et al., 1999) in the AIDS study, the development of computational algorithms in molecular population genetics (Stephens and Donnelly, 2000), estimation of the functional Bregman divergence for Bayesian model diagnostics (Goh and Dey, 2014), and the intensity bias correction in endorectal multi-parametric MRI

(Lui, 2014; Lui et al., 2015).

When all parameters in the model are of concern, the joint posterior density is investigated using the unnormalized posterior density, which is the product of the likelihood function and a joint prior distribution. The normalizing constant, which is the marginal likelihood when the prior is proper, is often analytically intractable due to the complicated model structure. In this case, the computational problem reduces to estimation of the marginal likelihood. For estimating the marginal likelihood, many Monte Carlo methods have been developed. An efficient method with minimal assumptions is the *partition weighted kernel* (PWK) estimation in Chapter 2. It requires only a single Markov chain Monte Carlo (MCMC) sample from the posterior distribution and the known unnormalized posterior density, both available from standard Bayesian software. By assigning a weight to each point of an MCMC sample adaptively within a working parameter space where the unnormalized posterior density is bounded away from zero, the PWK estimator is consistent for the reciprocal of the normalizing constant and of finite variance, and its minimum variance estimator can be achieved.

In many practical problems, however, investigators may be interested only in specific parameters rather than all parameters. As a result, the calculation and display of the marginal posterior density of the focal parameters are of most interest. Similar to the joint posterior density, the marginal posterior density is not analytically available for most cases. Hence, several methods in the literature have been developed for estimating marginal posterior densities using an MCMC sample from the joint posterior distribution. One common approach is the *kernel density estimator* (KDE) proposed by Rosenblatt et al. (1956) and Parzen (1962). KDE is easily implemented, but leaves room for more efficient estimators that can make use of more of the available

information. Another approach is the *conditional marginal density estimator* (CMDE) method proposed by Gelfand et al. (1992). This method simply takes the average of the known conditional posterior density of the interested parameters given the other parameters from the MCMC sample. The CMDE is most efficient since it is a Rao-Blackwell estimator. Unfortunately, CMDE requires the closed form of the conditional posterior density, which is often known only up to an unknown normalizing constant for many Bayesian problems, especially when the parameter space is constrained. To overcome this difficulty, Chen (1994) proposed the *importance weighted marginal density estimation* (IWMDE) method, which can be viewed as a generalization of the CMDE and only requires a careful selection of the conditional weight density. Some other approaches including two block structures of the IWMDE by Oh (1999) and the *Gibbs stopper* (GS) estimator by Yu and Tanner (1999) are described in Chen (2005a). However, when the conditional posterior density is analytically intractable, the realization of the CMDE still remains an open research problem.

We use ideas motivated by the PWK estimator to compute the marginal posterior density. We propose a new Monte Carlo method, the *Adaptive Partition weighTed* (APT) marginal density estimator, which assumes only the availability of the unnormalized posterior density and an MCMC sample from the joint posterior distribution. The APT method is constructed by first partitioning a subset of the support of the conditional posterior distribution, and then estimating the marginal posterior density at a fixed point of the focused parameters. An adaptive weighted average is assigned to the ratios of the unnormalized posterior density evaluated at the MCMC sample, except the focused parameters in the numerator are set at this fixed point, where weights are assigned locally using a representative value of the unnormalized posterior density in each partitioned subset. Both the partition and the weights change adap-

tively at each MCMC sample point. We show that the APT estimator is unbiased, and most of all, its optimal result is realizable and approximates the CMDE, which is known to be the best solution but whose widespread use is limited by unavailability of the conditional posterior density analytically. In addition, we extend the APT method to estimate the conditional posterior density when an MCMC sample from the conditional posterior density is available. In our first real data example, where closed form conditional posterior densities are available, we show that our estimator produces results similar to the gold standard (CMDE). In our second real data example, we demonstrate an excellent performance of the APT method in estimating Bayes factors. The proposed APT method has the potential to become a powerful tool for computing posterior densities, Bayes factors, and marginal likelihoods for complex Bayesian models with a large number of parameters.

The rest of the article is organized as follows: Section 2 is a review of existing methods and a summary of preliminaries needed for our method. In Section 3, we develop the APT method for estimating marginal and conditional posterior densities, and examine its theoretical properties and various applications. In Section 4, we compare the performances of KDE and APT with the CMDE in the inequality-constrained analysis of variance model. We use the data set from an amnesia study of dissociative identity disorder patients to demonstrate the empirical performance of the APT. In Section 5, we present a novel application of the APT method for computing the marginal posterior densities in Bayesian model selection. A complete Bayesian analysis is carried out under an ordinal probit regression model with latent variables using the real data from a prostate cancer study. To avoid the phenomenon of the Bartlett's or Lindley's paradox (Jeffreys, 1998; Lindley, 1957), we first apply the PWK estimator to find the best prior distribution in the full model by the empirical

Bayes method and we then apply APT to compute the Bayes factors under the best full model using the same MCMC sample used by the PWK estimator for model selection. Finally, we conclude the paper with a brief discussion in Section 6.

3.2 Preliminaries

We first introduce some notation. Suppose $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\xi})$ is a v-dimensional vector of parameters, where $\boldsymbol{\theta}$ is a vector of parameters of interest having length p. Let D denote the data, $L(\boldsymbol{\zeta}|D)$ denote the likelihood function, and $\pi(\boldsymbol{\zeta})$ denote the prior distribution for $\boldsymbol{\zeta}$. Then, the joint posterior density is

$$\pi(\boldsymbol{\zeta}|D) = \frac{L(\boldsymbol{\zeta}|D)\pi(\boldsymbol{\zeta})}{c} = \frac{q(\boldsymbol{\zeta})}{c}, \tag{3.1}$$

where c is the normalizing constant, and $q(\boldsymbol{\zeta})$ is the unnormalized posterior density. The equation (3.1) shows that the joint posterior density is known up to a normalizing constant. Hence, when $p = v$, the computation problem is the estimation of c. For this value, the PWK estimator in Chapter 2 can be used. Assume $\boldsymbol{\Omega}$ is the support of $\pi(\boldsymbol{\zeta}|D)$ with $\boldsymbol{\Omega}' \subset \boldsymbol{\Omega}$ being the working parameter space, and $\{A_1, A_2, ..., A_K\}$ forms a partition of $\boldsymbol{\Omega}'$. If we have an MCMC sample $\{\boldsymbol{\zeta}_t = (\boldsymbol{\theta}_t, \boldsymbol{\xi}_t), t = 1, 2, \ldots, T\}$ from the joint posterior distribution $\pi(\boldsymbol{\zeta}|D)$, the PWK estimator for $1/c$ is

$$\widehat{c^{-1}} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{q(\boldsymbol{\zeta}_k^*)}{q(\boldsymbol{\zeta}_t)}1\{\boldsymbol{\zeta}_t \in A_k\}}{\sum_{k=1}^{K} q(\boldsymbol{\zeta}_k^*)V(A_k)}, \tag{3.2}$$

where $\boldsymbol{\zeta}_k^*$ is a representative point in region A_k, $1\{\boldsymbol{\zeta}_t \in A_k\}$ is the indicator function, and $V(A_k) = \int_{\boldsymbol{\Omega}'} 1\{\boldsymbol{\zeta} \in A_k\} d\boldsymbol{\zeta}$. As discussed in Chapter 2, $q(\boldsymbol{\zeta}_t)$ in (3.2) can be made

close to $q(\boldsymbol{\zeta}_k^*)$ by increasing the number of partition subsets to improve estimation precision.

When $p < v$, the marginal posterior density of interest is defined by

$$\pi(\boldsymbol{\theta}_0|D) = \int_{\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}} \pi(\boldsymbol{\zeta}|D)_{|\boldsymbol{\theta}=\boldsymbol{\theta}_0} d\boldsymbol{\xi} = \int_{\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}} \pi(\boldsymbol{\theta}_0, \boldsymbol{\xi}|D) d\boldsymbol{\xi}, \tag{3.3}$$

where $\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} = \{\boldsymbol{\xi} : (\boldsymbol{\theta}_0, \boldsymbol{\xi}) \in \boldsymbol{\Omega}\}$. To estimate the marginal posterior density, a nonparametric kernel density estimator (KDE) can be used. It is similar to the KDE of the frequentist literature except the sample is replaced by the MCMC sample of interested parameters. Although it is easily implemented and requires no further assumptions, it may be less efficient because it does not use the information from the MCMC sample of non-focal parameters and the known structure of the posterior distribution.

Another common approach is the *conditional marginal density estimator* (CMDE) proposed by Gelfand et al. (1992). Assume the analytical form of the conditional posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}, D)$ is available. Then, (3.3) can be re-written as

$$\pi(\boldsymbol{\theta}_0|D) = \int_{\boldsymbol{\Omega}} \pi(\boldsymbol{\theta}_0, \boldsymbol{\xi}|D)\pi(\boldsymbol{\theta}|\boldsymbol{\xi}, D) d\boldsymbol{\zeta} = \int_{\boldsymbol{\Omega}} \pi(\boldsymbol{\theta}_0|\boldsymbol{\xi}, D)\pi(\boldsymbol{\zeta}|D) d\boldsymbol{\zeta}. \tag{3.4}$$

Then

$$\hat{\pi}_{\text{CMDE}}(\boldsymbol{\theta}_0|D) = \frac{1}{T}\sum_{t=1}^{T} \pi(\boldsymbol{\theta}_0|\boldsymbol{\xi}_t, D). \tag{3.5}$$

It can be shown that under some mild regularity conditions, $\hat{\pi}_{\text{CMDE}}(\boldsymbol{\theta}_0|D)$ is an unbiased and consistent estimator of the marginal posterior density, that is,

$$E(\hat{\pi}_{\text{CMDE}}(\boldsymbol{\theta}_0|D)) = \pi(\boldsymbol{\theta}_0|D),$$

and

$$\lim_{T\to\infty} \hat{\pi}_{\text{CMDE}}(\boldsymbol{\theta}_0|D) = \pi(\boldsymbol{\theta}_0|D) \;\; a.s.$$

In addition, the use of the conditional structure of the posterior density makes the CMDE a Rao-Blackwell estimator so that it is optimal for this estimation problem. However, the closed form of $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}, D)$ is often not available in many Bayesian problems. To overcome this difficulty, Chen (1994) proposed the *importance weighted marginal density estimation* (IWMDE) method, which can be considered as a generalization of CMDE. Consider the following identity:

$$\pi(\boldsymbol{\theta}_0|D) = \int \frac{w(\boldsymbol{\theta}|\boldsymbol{\xi})\pi(\boldsymbol{\theta}_0, \boldsymbol{\xi}|D)}{\pi(\boldsymbol{\zeta}|D)}\pi(\boldsymbol{\zeta}|D)d\boldsymbol{\zeta}, \tag{3.6}$$

where $w(\boldsymbol{\theta}|\boldsymbol{\xi})$ is a proposed conditional density whose support is contained in the support of $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}, D)$. Using the identity in (3.6), the IWMDE of $\pi(\boldsymbol{\theta}_0|D)$ is given by

$$\hat{\pi}_{\text{IWMDE}}(\boldsymbol{\theta}_0|D) = \frac{1}{T}\sum_{t=1}^{T} w(\boldsymbol{\theta}_t|\boldsymbol{\xi}_t)\frac{\pi(\boldsymbol{\theta}_0, \boldsymbol{\xi}_t|D)}{\pi(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t|D)}. \tag{3.7}$$

The IWMDE method is attractive since it does not require the conditional posterior density to be known, the only requirement is that one needs to choose a good weight function $w(\boldsymbol{\theta}|\boldsymbol{\xi})$. Under mild regularity conditions, the IWMDE also has the properties of unbiasedness and consistency to the marginal posterior density.

3.3 The Proposed Method for Estimating Posterior Densities

3.3.1 Estimating Marginal Posterior Density

Chen (1994) showed that the optimal weight function minimizing the variance of (3.7) is $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}, D)$, and in this case the IWMDE in (3.7) reduces to CMDE in (3.5). However, this optimal weight function is unavailable in most cases, and proposing a similar one is nontrivial. To circumvent the difficulties, we exploit the main idea behind the PWK estimator to obtain an approximate distribution of the conditional posterior density so that the CMDE can be realized for marginal posterior density estimation.

Let $\Theta_{\boldsymbol{\xi}} = \{\boldsymbol{\theta} : \ q(\boldsymbol{\theta}, \xi) > 0\}$ denote the support of the conditional posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}, D)$ and $\tilde{\Theta}_{\boldsymbol{\xi}}$ be any subset of $\Theta_{\boldsymbol{\xi}}$ such that $\int_{\tilde{\Theta}_{\boldsymbol{\xi}}} q(\boldsymbol{\theta}, \boldsymbol{\xi}) d\boldsymbol{\theta} > 0$. We call $\tilde{\Theta}_{\boldsymbol{\xi}}$ the conditional working parameter space. We also let $\{A_k(\boldsymbol{\xi}), \ k = 1, 2, \ldots, K\}$ be the partition of $\tilde{\Theta}_{\boldsymbol{\xi}}$. We consider a weight function $w(\boldsymbol{\theta}|\boldsymbol{\xi})$ which satisfies the following two conditions: (i) $w(\boldsymbol{\theta}|\boldsymbol{\xi}) \geq 0$ and (ii) $\int_{\tilde{\Theta}_{\boldsymbol{\xi}}} w(\boldsymbol{\theta}|\boldsymbol{\xi}) d\boldsymbol{\theta} = 1$. Then, we propose a new estimator of $\pi(\boldsymbol{\theta}_0|D)$ using the idea behind PWK:

$$\hat{\pi}(\boldsymbol{\theta}_0|D) = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} w(\boldsymbol{\theta}_t|\boldsymbol{\xi}_t) \frac{q(\boldsymbol{\theta}_0, \boldsymbol{\xi}_t)}{q(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)} 1\{\boldsymbol{\theta}_t \in A_k(\boldsymbol{\xi}_t)\}. \tag{3.8}$$

Write

$$\hat{\pi}_t(\boldsymbol{\theta}_0|D) = \sum_{k=1}^{K} w(\boldsymbol{\theta}_t|\boldsymbol{\xi}_t) \frac{q(\boldsymbol{\theta}_0, \boldsymbol{\xi}_t)}{q(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)} 1\{\boldsymbol{\theta}_t \in A_k(\boldsymbol{\xi}_t)\}$$

such that $\hat{\pi}(\boldsymbol{\theta}_0|D) = \frac{1}{T}\sum_{t=1}^{T}\hat{\pi}_t(\boldsymbol{\theta}_0|D)$. Then, we have

$$\begin{aligned} E[\hat{\pi}_t(\boldsymbol{\theta}_0|D)] &= \int\int \sum_{k=1}^{K} w(\boldsymbol{\theta}|\boldsymbol{\xi})\frac{q(\boldsymbol{\theta}_0,\boldsymbol{\xi})}{q(\boldsymbol{\theta},\boldsymbol{\xi})}1\{\boldsymbol{\theta}\in A_k(\boldsymbol{\xi})\}\frac{q(\boldsymbol{\theta},\boldsymbol{\xi})}{c}d\boldsymbol{\theta}d\boldsymbol{\xi} \\ &= \int \frac{q(\boldsymbol{\theta}_0,\boldsymbol{\xi})}{c}\int w(\boldsymbol{\theta}|\boldsymbol{\xi})\sum_{k=1}^{K}1\{\boldsymbol{\theta}\in A_k(\boldsymbol{\xi})\}d\boldsymbol{\theta}d\boldsymbol{\xi} \\ &= \int \frac{q(\boldsymbol{\theta}_0,\boldsymbol{\xi})}{c}\int_{\tilde{\Theta}_{\boldsymbol{\xi}}} w(\boldsymbol{\theta}|\boldsymbol{\xi})d\boldsymbol{\theta}d\boldsymbol{\xi} \\ &= \pi(\boldsymbol{\theta}_0|D), \end{aligned} \tag{3.9}$$

which ensures that $\hat{\pi}(\boldsymbol{\theta}_0|D)$ in (3.8) is an unbiased estimator of $\pi(\boldsymbol{\theta}_0|D)$. After some algebra, we obtain

$$\mathrm{Var}_w\{\hat{\pi}_t(\boldsymbol{\theta}_0|D)\} = \int \Big[q(\boldsymbol{\theta}_0,\boldsymbol{\xi})\pi(\boldsymbol{\theta}_0,\boldsymbol{\xi}|D)\int_{\tilde{\Theta}_{\boldsymbol{\xi}}}\frac{w^2(\boldsymbol{\theta}|\boldsymbol{\xi})}{q(\boldsymbol{\theta},\boldsymbol{\xi})}d\boldsymbol{\theta}\Big]d\boldsymbol{\xi} - \pi^2(\boldsymbol{\theta}_0|D). \tag{3.10}$$

Now, we establish the following useful result.

Theorem 3.3.1. *Let*

$$w_{opt}(\boldsymbol{\theta}|\boldsymbol{\xi}) = \frac{q(\boldsymbol{\theta},\boldsymbol{\xi})}{\int_{\tilde{\Theta}_{\boldsymbol{\xi}}} q(\ddot{\boldsymbol{\theta}},\boldsymbol{\xi})d\ddot{\boldsymbol{\theta}}}, \tag{3.11}$$

which is the conditional posterior density defined on $\tilde{\Theta}_{\boldsymbol{\xi}}$*. Then, we have*

$$Var_{w_{opt}}\{\hat{\pi}_t(\boldsymbol{\theta}_0|D)\} = \int\left[\frac{q(\boldsymbol{\theta}_0,\boldsymbol{\xi})\pi(\boldsymbol{\theta}_0,\boldsymbol{\xi}|D)}{\int_{\tilde{\Theta}_{\boldsymbol{\xi}}} q(\boldsymbol{\theta},\boldsymbol{\xi})d\boldsymbol{\theta}}\right]d\boldsymbol{\xi} - \pi^2(\boldsymbol{\theta}_0|D) \tag{3.12}$$

and

$$Var_{w_{opt}}\{\hat{\pi}_t(\boldsymbol{\theta}_0|D)\} \leq Var_w\{\hat{\pi}_t(\boldsymbol{\theta}_0|D)\} \tag{3.13}$$

for any conditional density $w(.)$ defined on $\tilde{\Theta}_{\boldsymbol{\xi}}$.

The result established in Theorem 3.3.1 is an extension of Theorem 2.1 in Chen (1994). The proof of this theorem is also similar.

Proof of Theorem 3.3.1. Using the Cauchy-Schwarz inequality, we have

$$1 = \left(\int_{\tilde{\Theta}_{\boldsymbol{\xi}}} \frac{w(\boldsymbol{\theta}|\boldsymbol{\xi})}{\sqrt{q(\boldsymbol{\theta},\boldsymbol{\xi})}} \sqrt{q(\boldsymbol{\theta},\boldsymbol{\xi})} d\boldsymbol{\theta} \right)^2 \leq \int_{\tilde{\Theta}_{\boldsymbol{\xi}}} \frac{w^2(\boldsymbol{\theta}|\boldsymbol{\xi})}{q(\boldsymbol{\theta},\boldsymbol{\xi})} d\boldsymbol{\theta} \int_{\tilde{\Theta}_{\boldsymbol{\xi}}} q(\boldsymbol{\theta},\boldsymbol{\xi}) d\boldsymbol{\theta}.$$

Subsequently, we obtain

$$\begin{aligned}
&\frac{1}{\int_{\tilde{\Theta}_{\boldsymbol{\xi}}} q(\boldsymbol{\theta},\boldsymbol{\xi}) d\boldsymbol{\theta}} \leq \int_{\tilde{\Theta}_{\boldsymbol{\xi}}} \frac{w^2(\boldsymbol{\theta}|\boldsymbol{\xi})}{q(\boldsymbol{\theta},\boldsymbol{\xi})} d\boldsymbol{\theta} \\
\Rightarrow &\int \left[q(\boldsymbol{\theta}_0,\boldsymbol{\xi}) \pi(\boldsymbol{\theta}_0,\boldsymbol{\xi}|D) \frac{1}{\int_{\tilde{\Theta}_{\boldsymbol{\xi}}} q(\boldsymbol{\theta},\boldsymbol{\xi}) d\boldsymbol{\theta}} \right] d\boldsymbol{\xi} \\
&\qquad \leq \int \left[q(\boldsymbol{\theta}_0,\boldsymbol{\xi}) \pi(\boldsymbol{\theta}_0,\boldsymbol{\xi}|D) \int_{\tilde{\Theta}_{\boldsymbol{\xi}}} \frac{w^2(\boldsymbol{\theta}|\boldsymbol{\xi})}{q(\boldsymbol{\theta},\boldsymbol{\xi})} d\boldsymbol{\theta} \right] d\boldsymbol{\xi} \\
\Rightarrow &\mathrm{Var}_{w_{\mathrm{opt}}}\{\hat{\pi}_t(\boldsymbol{\theta}_0|D)\} \leq \mathrm{Var}_{w}\{\hat{\pi}_t(\boldsymbol{\theta}_0|D)\},
\end{aligned}$$

which completes the proof. □

REMARK 3.1: Note that $w_{\mathrm{opt}}(\boldsymbol{\theta}|\boldsymbol{\xi})$ in (3.11) is not the conditional posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{\xi}, D)$ unless $\tilde{\Theta}_{\boldsymbol{\xi}} = \Theta_{\boldsymbol{\xi}}$. Consequently, $\hat{\pi}_{w_{\mathrm{opt}}}(\boldsymbol{\theta}_0|D)$, that is equation (3.8) with plug-in weights of (3.11), is not the CMDE. From Theorem 3.3.1, we see that $\mathrm{Var}_{w_{\mathrm{opt}}}\{\hat{\pi}_t(\boldsymbol{\theta}_0|D)\}$ decreases when the conditional working parameter space $\tilde{\Theta}_{\boldsymbol{\xi}}$ gets larger. Once $\tilde{\Theta}_{\boldsymbol{\xi}} = \Theta_{\boldsymbol{\xi}}$, $\mathrm{Var}\{\hat{\pi}_{w_{\mathrm{opt}}}(\boldsymbol{\theta}_0|D)\}$ is equal to $[\int \pi(\boldsymbol{\theta}_0|\boldsymbol{\xi}, D)\pi(\boldsymbol{\theta}_0,\boldsymbol{\xi}|D)d\boldsymbol{\xi} - \pi^2(\boldsymbol{\theta}_0|D)]/T$ and exactly the variance of the CMDE. Thus, the CMDE is the best among all of the estimators given in (3.8). As discussed in Chen (1994), the CMDE

is analytically intractable and thus it is almost impossible to compute the CMDE in practice for most applications. However, the result established in Theorem 3.3.1 sheds light on how to obtain an estimator of the marginal posterior density, which is approximately as good as the CMDE.

Although $w_{\text{opt}}(\boldsymbol{\theta}|\boldsymbol{\xi})$ is analytically intractable, we can borrow the idea behind the PWK estimator to obtain an approximate optimal estimator $\hat{\pi}_w(\boldsymbol{\theta}_0|D)$ using discretization. First, we take $\tilde{\Theta}_{\boldsymbol{\xi}}$ such that $\int_{\tilde{\Theta}_{\boldsymbol{\xi}}} d\boldsymbol{\theta} < \infty$. Second, we specify $w(.)$ as

$$w_{\text{APT}}(\boldsymbol{\theta}|\boldsymbol{\xi}) = \frac{q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi})}{\sum_{k=1}^{K} q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi}) V(A_k(\boldsymbol{\xi}))}, \quad \boldsymbol{\theta} \in A_k(\boldsymbol{\xi}), \tag{3.14}$$

where $\boldsymbol{\theta}_k^*$ is a fixed point in $A_k(\boldsymbol{\xi})$ and $V(A_k(\boldsymbol{\xi}))$ is the volume of $A_k(\boldsymbol{\xi})$. Then, plugging w_{APT} in (3.8) gives

$$\hat{\pi}_{\text{APT}}(\boldsymbol{\theta}_0|D) = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi}_t) 1\{\boldsymbol{\theta}_t \in A_k(\boldsymbol{\xi}_t)\}}{\sum_{k=1}^{K} q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi}_t) V(A_k(\boldsymbol{\xi}_t))} \frac{q(\boldsymbol{\theta}_0, \boldsymbol{\xi}_t)}{q(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)}. \tag{3.15}$$

We see from (3.15) that the partition $\{A_k(\boldsymbol{\xi}_t),\ k = 1, 2, \ldots, K\}$ changes at each t and, consequently, the weight $\frac{q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi}_t) 1\{\boldsymbol{\theta}_t \in A_k(\boldsymbol{\xi}_t)\}}{\sum_{k=1}^{K} q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi}_t) V(A_k(\boldsymbol{\xi}_t))}$ adaptively changes with t. Therefore, we call this the Adaptive Partition weighTed (APT) marginal density estimator.

REMARK 3.2: Under very mild conditions, we can show that $\hat{\pi}_{\text{APT}}(\boldsymbol{\theta}_0|D)$ is an unbiased and consistent estimator of $\pi(\boldsymbol{\theta}_0|D)$. In addition, when $K \to \infty$ and $V(A_k(\boldsymbol{\xi})) \to 0$, $w_{\text{APT}}(\boldsymbol{\theta}|\boldsymbol{\xi}) \to w_{\text{opt}}(\boldsymbol{\theta}|\boldsymbol{\xi})$. In this case, $\hat{\pi}_{\text{APT}}(\boldsymbol{\theta}_0|D)$ is the best estimator for the given subset $\tilde{\Theta}_{\boldsymbol{\xi}}$. Furthermore, when $\tilde{\Theta}_{\boldsymbol{\xi}} \to \Theta_{\boldsymbol{\xi}}$, $K \to \infty$, and $V(A_k(\boldsymbol{\xi})) \to 0$, $\hat{\pi}_{\text{APT}}(\boldsymbol{\theta}_0|D) \to \hat{\pi}_{\text{CMDE}}(\boldsymbol{\theta}_0|D)$. Thus, the proposed APT has a potential to be as good as the best solution (CMDE).

3.3.2 Estimating Conditional Posterior Density

Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Let $\pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_{01}, D)$ denote the conditional posterior density of $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{01}$. Using the same notation as in Section 3.1, the APT estimator of $\pi(\boldsymbol{\theta}_{02}|\boldsymbol{\theta}_{01}, D)$ is given by

$$\hat{\pi}_{\text{APT}}(\boldsymbol{\theta}_{02}|\boldsymbol{\theta}_{01}, D) = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} \frac{q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}^*_{2k}, \boldsymbol{\xi}_t)1\{\boldsymbol{\theta}_{2t} \in A_k(\boldsymbol{\theta}_{01}, \boldsymbol{\xi}_t)\}}{\sum_{k=1}^{K} q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}^*_{2k}, \boldsymbol{\xi}_t)V(A_k(\boldsymbol{\theta}_{01}, \boldsymbol{\xi}_t))} \frac{q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02}, \boldsymbol{\xi}_t)}{q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{2t}, \boldsymbol{\xi}_t)}, \tag{3.16}$$

where $\{(\boldsymbol{\theta}_{2t}, \boldsymbol{\xi}_t), t = 1, 2, ..., T\}$ is an MCMC sample from $\pi(\boldsymbol{\theta}_2, \boldsymbol{\xi}|\boldsymbol{\theta}_{01}, D)$, $\boldsymbol{\theta}^*_{2k}$ is a representative point in $A_k(\boldsymbol{\theta}_{01}, \boldsymbol{\xi})$, and $V(A_k(\boldsymbol{\theta}_{01}, \boldsymbol{\xi}))$ is the volume of $A_k(\boldsymbol{\theta}_{01}, \boldsymbol{\xi})$.

Under the mild regularity conditions, we can show that $\hat{\pi}_{\text{APT}}(\boldsymbol{\theta}_{02}|\boldsymbol{\theta}_{01}, D)$ is an unbiased estimator of $\pi(\boldsymbol{\theta}_{02}|\boldsymbol{\theta}_{01}, D)$. This result can be established as follows. Write $c(\boldsymbol{\theta}_{01}) = \int q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_2, \boldsymbol{\xi})d\boldsymbol{\theta}_2 d\boldsymbol{\xi}$. Then, we have $\pi(\boldsymbol{\theta}_2, \boldsymbol{\xi}|\boldsymbol{\theta}_{01}, D) = q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_2, \boldsymbol{\xi})/c(\boldsymbol{\theta}_{01})$, $\pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_{01}, D) = \int \pi(\boldsymbol{\theta}_2, \boldsymbol{\xi}|\boldsymbol{\theta}_{01}, D)d\boldsymbol{\xi}$, and

$$\begin{aligned}
&E[\hat{\pi}_{\text{APT}}(\boldsymbol{\theta}_{02}|\boldsymbol{\theta}_{01}, D)] \\
&= \int\int \sum_{k=1}^{K} \frac{q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}^*_{2k}, \boldsymbol{\xi})1\{\boldsymbol{\theta}_2 \in A_k(\boldsymbol{\theta}_{01}, \boldsymbol{\xi})\}}{\sum_{k=1}^{K} q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}^*_{2k}, \boldsymbol{\xi})V(A_k(\boldsymbol{\theta}_{01}, \boldsymbol{\xi}))} \frac{q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02}, \boldsymbol{\xi})}{q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_2, \boldsymbol{\xi})} \pi(\boldsymbol{\theta}_2, \boldsymbol{\xi}|\boldsymbol{\theta}_{01}, D)d\boldsymbol{\theta}_2 d\boldsymbol{\xi} \\
&= \int\int \sum_{k=1}^{K} \frac{q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}^*_{2k}, \boldsymbol{\xi})1\{\boldsymbol{\theta}_2 \in A_k(\boldsymbol{\theta}_{01}, \boldsymbol{\xi})\}}{\sum_{k=1}^{K} q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}^*_{2k}, \boldsymbol{\xi})V(A_k(\boldsymbol{\theta}_{01}, \boldsymbol{\xi}))} \frac{q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02}, \boldsymbol{\xi})}{c(\boldsymbol{\theta}_{01})} d\boldsymbol{\theta}_2 d\boldsymbol{\xi} \\
&= \int \frac{q(\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02}, \boldsymbol{\xi})}{c(\boldsymbol{\theta}_{01})} d\boldsymbol{\xi} = \pi(\boldsymbol{\theta}_{02}|\boldsymbol{\theta}_{01}, D).
\end{aligned} \tag{3.17}$$

There are two major applications of (3.16). First, when the dimension of $\boldsymbol{\theta}_0$ is high, $\hat{\pi}_{\text{APT}}(\boldsymbol{\theta}_0|D)$ in (3.15) may not be efficient. The dimension reduction can be achieved via the identity $\pi(\boldsymbol{\theta}_0|D) = \pi(\boldsymbol{\theta}_{01}|D)\pi(\boldsymbol{\theta}_{02}|\boldsymbol{\theta}_{01}, D)$. Now, instead of estimating $\pi(\boldsymbol{\theta}_0|D)$ directly, we use (3.15) and (3.16) to estimate $\pi(\boldsymbol{\theta}_{01}|D)$ and $\pi(\boldsymbol{\theta}_{02}|\boldsymbol{\theta}_{01}, D)$,

respectively. These two estimators require one MCMC sample from the full posterior distribution $\pi(\boldsymbol{\zeta}|D)$ and another MCMC sample from the conditional posterior distribution $\pi(\boldsymbol{\theta}_2, \boldsymbol{\xi}|\boldsymbol{\theta}_{01}, D)$. We note that once the sampling code for the full posterior distribution is readily available, the very same code with minimal changes can be used to generate an MCMC sample from the conditional posterior distribution.

Second, (3.16) can also be applied to the estimation of the marginal likelihood through Chib's identity (Chib, 1995). Suppose we group $\boldsymbol{\zeta}$ into G blocks as $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1', \boldsymbol{\zeta}_2', ..., \boldsymbol{\zeta}_G')'$. Following Chib (1995), we have

$$c = \frac{q(\boldsymbol{\zeta}^*)}{\pi(\boldsymbol{\zeta}_1^*, \boldsymbol{\zeta}_2^*, ..., \boldsymbol{\zeta}_G^*|D)} = \frac{q(\boldsymbol{\zeta}^*)}{\pi(\boldsymbol{\zeta}_1^*|D)\pi(\boldsymbol{\zeta}_2^*|\boldsymbol{\zeta}_1^*, D)\ldots\pi(\boldsymbol{\zeta}_G^*|\boldsymbol{\zeta}_1^*, \boldsymbol{\zeta}_2^*, ..., \boldsymbol{\zeta}_{G-1}^*, D)}, \quad (3.18)$$

where $\boldsymbol{\zeta}^* = ((\boldsymbol{\zeta}_1^*)', (\boldsymbol{\zeta}_2^*)', ..., (\boldsymbol{\zeta}_G^*)')'$ is a high posterior density point such as the posterior mean or mode of $\boldsymbol{\zeta}$. In (3.18), $\pi(\boldsymbol{\zeta}_1^*|D)$ can be estimated by (3.15) using an MCMC sample from the full posterior density while an estimate of each conditional density $\pi(\boldsymbol{\zeta}_g^*|\boldsymbol{\zeta}_1^*, \boldsymbol{\zeta}_2^*, ..., \boldsymbol{\zeta}_{g-1}^*, D)$ can be obtained by (3.16) using an MCMC sample from the conditional posterior distribution $\pi(\boldsymbol{\zeta}_g, \boldsymbol{\zeta}_{g+1}, \ldots, \boldsymbol{\zeta}_G|\boldsymbol{\zeta}_1^*, \boldsymbol{\zeta}_2^*, ..., \boldsymbol{\zeta}_{g-1}^*, D)$ for $g = 2, 3, \ldots, G$. This approach does not require knowing the full conditional posterior density analytically (i.e., CMDE) or the closed-form expression of the full conditional density after introducing additional latent variables (namely, the augmented CMDE) as discussed in Chib (1995).

3.4 Inequality-Constrained Analysis of Variance

In this section, we apply APT to the inequality-constrained analysis of variance model to find the marginal posterior density. We use the data in Hoijtink et al. (2008)

collected to study amnesia in patients with dissociative identity disorder (DID). The experiment is designed to compare memory performance scores among four groups, including DID, mimic normal, symptom simulated, and true amnesic subjects. Chen and Kim (2008) were interested in testing several hypotheses for the mean scores. Based on their best model, we further investigate the marginal posterior density of each mean. In each case, the closed form of the conditional posterior density is available so that the CMDE can be used for comparison. Furthermore, we obtain two joint marginal posterior densities: the means of the DID and true amnesia groups, and the means of DID and symptom simulated groups. Note that the conditional posterior density of the former is analytically available while the latter requires numerical integration in the process.

The experiment investigates whether a DID patient really suffers from amnesia when switching from one identity to another. To objectively evaluate the subjective experience of amnesia in DID patients and avoid the iatrogenic problem (i.e., patients' behaviors are induced by therapists) or some suggestive influence of media, DID patients ($n_1 = 19$) were implicitly measured for amnesia in the following procedure: after being told a brief story and shown some figures, they were asked to change their identity and then answer recognition questions about the story and figure details. To avoid the issue of symptom simulation in DID patients, three control groups composed of normal healthy people were recruited for comparison. The second (first control) group ($n_2 = 25$) was a normal control group. They were told the story and shown the figures, and were then asked to answer the questions. Without any intervention, their performance would be the performance of normal people. The same procedure was used for the third (second control) group ($n_3 = 25$), but they were extensively informed about the behaviors of DID in advance and asked to deliberately simulate

inter-identity amnesia. They can be viewed as the symptom simulation group. The fourth (third control) group ($n_4 = 25$) were asked to directly answer the questions without experiencing the story and graphs. Due to their complete lack of knowledge, this group represented a true amnesic group.

We model memory performance score y_{ij} for the i^{th} subject in group j as an independent observation from a normal distribution with mean μ_j and variance σ^2 for $i = 1, 2, ..., n_j$ and $j = 1, ..., J$. Let Θ denote the corresponding constrained parameter space for $\boldsymbol{\theta}$ under the hypothesis, for example $\{\mu_1 < \mu_2 < ....\}$, and let $D = \{n, \boldsymbol{y}\} = \{\sum_{j=1}^{J} n_j, y_{11}, y_{21}, \ldots, y_{n_1 1}, \ldots, y_{1J}, y_{2J}, \ldots, y_{n_J J}\}$ denote the observed data. Then, the likelihood function is given by

$$L(\boldsymbol{\theta}|D) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \mu_j)^2\right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2) = (\mu_1, \ldots, \mu_J, \sigma^2)$. For this model, a prior distribution in Chen and Ibrahim (2003) is considered: a power prior based on the prior predictive values of the performance score is assigned for μ given σ^2, and an inverse gamma prior is assumed for σ^2. Specifically, the joint prior is

$$\pi(\boldsymbol{\mu}, \sigma^2|\boldsymbol{y}_0, a_0) \propto \left[(\sigma^2)^{-\frac{n_0}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_{0j}}(y_{0ij} - \mu_j)^2\right\}\right]^{a_0} (\sigma^2)^{-b_{01}-1} \times \exp\left(-\frac{b_{02}}{\sigma^2}\right) 1\{(\boldsymbol{\mu}, \sigma^2) \in \Theta\},$$

where $\boldsymbol{y}_0 = (\boldsymbol{y}_{01}, \boldsymbol{y}_{02}, ..., \boldsymbol{y}_{0J})$ with $\boldsymbol{y}'_{0j} = (y_{01j}, y_{02j}, \ldots, y_{0n_{0j}j})$ denotes the prior predictive values of the response variables, a_0 is a scalar parameter determining the degree of involvement of the prior predictive values in forming the prior distribution,

$n_0 = \sum_{j=1}^{J} n_{0j}$, and $b_{01} > 0$ and $b_{02} > 0$ are prespecified hyperparameters. Under this constrained structure, the marginal posterior density of μ_j is analytically intractable, but can be evaluated by using an MCMC approach.

Assume $\sum_{i=1}^{n_j} y_{0ij} = 0$ for $j = 1, 2, ..., 4$ and $b_{01} = b_{02} = 0.0001$. For comparing our proposed method with KDE and CMDE, we choose one of the best hypotheses: $\Theta = \{\mu_2 > (\mu_1, \mu_4) > \mu_3, \sigma^2 > 0\}$ under $a_0 = 0.01$ based on Bayesian analyses by Chen and Kim (2008), where Bayesian model selection of different hypotheses on mean scores is carried out with various criteria including L measure, deviance information criterion (DIC), the conditional predictive ordinate (CPO) statistic, and the marginal likelihood or Bayes factor. Under this hypothesis, the CMDE is applicable in estimating the marginal posterior density of each mean since the closed form of

the conditional posterior density is available as follows:

$$\pi(\mu_1|\boldsymbol{\mu}_{(-1)},\sigma^2,D,\boldsymbol{y}_0,a_0)=\frac{\sqrt{\frac{n_1+a_0n_1}{2\pi\sigma^2}}\exp\left(-\frac{(n_1+a_0n_1)[\mu_1-\frac{\sum y_{i1}}{n_1+a_0n_1}]^2}{2\sigma^2}\right)}{\left[\Phi\left(\frac{\mu_2-\frac{\sum y_{i1}}{n_1+a_0n_1}}{\sqrt{\frac{\sigma^2}{n_1+a_0n_1}}}\right)-\Phi\left(\frac{\mu_3-\frac{\sum y_{i1}}{n_1+a_0n_1}}{\sqrt{\frac{\sigma^2}{n_1+a_0n_1}}}\right)\right]},$$

$$\pi(\mu_2|\boldsymbol{\mu}_{(-2)},\sigma^2,D,\boldsymbol{y}_0,a_0)=\frac{\sqrt{\frac{n_2+a_0n_2}{2\pi\sigma^2}}\exp\left(-\frac{(n_2+a_0n_2)[\mu_2-\frac{\sum y_{i2}}{n_2+a_0n_2}]^2}{2\sigma^2}\right)}{\left[1-\Phi\left(\frac{\max(\mu_1,\mu_4)-\frac{\sum y_{i2}}{n_2+a_0n_2}}{\sqrt{\frac{\sigma^2}{n_2+a_0n_2}}}\right)\right]},$$

$$\pi(\mu_3|\boldsymbol{\mu}_{(-3)},\sigma^2,D,\boldsymbol{y}_0,a_0)=\frac{\sqrt{\frac{n_3+a_0n_3}{2\pi\sigma^2}}\exp\left(-\frac{(n_3+a_0n_3)[\mu_3-\frac{\sum y_{i3}}{n_3+a_0n_3}]^2}{2\sigma^2}\right)}{\Phi\left(\frac{\min(\mu_1,\mu_4)-\frac{\sum y_{i3}}{n_3+a_0n_3}}{\sqrt{\frac{\sigma^2}{n_3+a_0n_3}}}\right)},$$

$$\pi(\mu_4|\boldsymbol{\mu}_{(-4)},\sigma^2,D,\boldsymbol{y}_0,a_0)=\frac{\sqrt{\frac{n_4+a_0n_4}{2\pi\sigma^2}}\exp\left(-\frac{(n_4+a_0n_4)[\mu_4-\frac{\sum y_{i4}}{n_4+a_0n_4}]^2}{2\sigma^2}\right)}{\left[\Phi\left(\frac{\mu_2-\frac{\sum y_{i4}}{n_4+a_0n_4}}{\sqrt{\frac{\sigma^2}{n_4+a_0n_4}}}\right)-\Phi\left(\frac{\mu_3-\frac{\sum y_{i4}}{n_4+a_0n_4}}{\sqrt{\frac{\sigma^2}{n_4+a_0n_4}}}\right)\right]},$$

where $\Phi(.)$ denotes the cumulative standard normal distribution function. To implement APT, we first choose the conditional working parameter space $\tilde{\Theta}_{\boldsymbol{\mu}_{(-jt)},\sigma_t^2}$ of each $\pi(\mu_j|\boldsymbol{\mu}_{(-j)},\sigma^2,D,\boldsymbol{y}_0,a_0)$ to be $\tilde{\Theta}_{\boldsymbol{\mu}_{(-1t)},\sigma_t^2}=\{\mu_{3t}<\mu_{1t}<\mu_{2t}\}$, $\tilde{\Theta}_{\boldsymbol{\mu}_{(-2t)},\sigma_t^2}=\{\max(\mu_{1t},\mu_{4t})<\mu_{2t}<15\}$, $\tilde{\Theta}_{\boldsymbol{\mu}_{(-3t)},\sigma_t^2}=\{0<\mu_{3t}<\min(\mu_{1t},\mu_{4t})\}$, and $\tilde{\Theta}_{\boldsymbol{\mu}_{(-4t)},\sigma_t^2}=\{\mu_{3t}<\mu_{4t}<\mu_{2t}\}$, where μ_{jt} is the t^{th} MCMC sample point for the j^{th} parameter. Notice that the chosen conditional working parameter space for $\tilde{\Theta}_{\boldsymbol{\mu}_{(-jt)},\sigma_t^2},j=1,4$ is exactly equal to the whole support of the conditional posterior distribution $\pi(\mu_j|\boldsymbol{\mu}_{(-j)},\sigma^2,D,\boldsymbol{y}_0,a_0),j=1,4$ since μ_1 and μ_4 are bounded on both sides. For μ_2 and μ_3, which are each bounded on one side, we set 15 as the upper bound and 0

as the lower bound for each respective parameter such that the working space can cover most of the MCMC sample. Then, we decide on the number of subsets K in the conditional working parameter space.

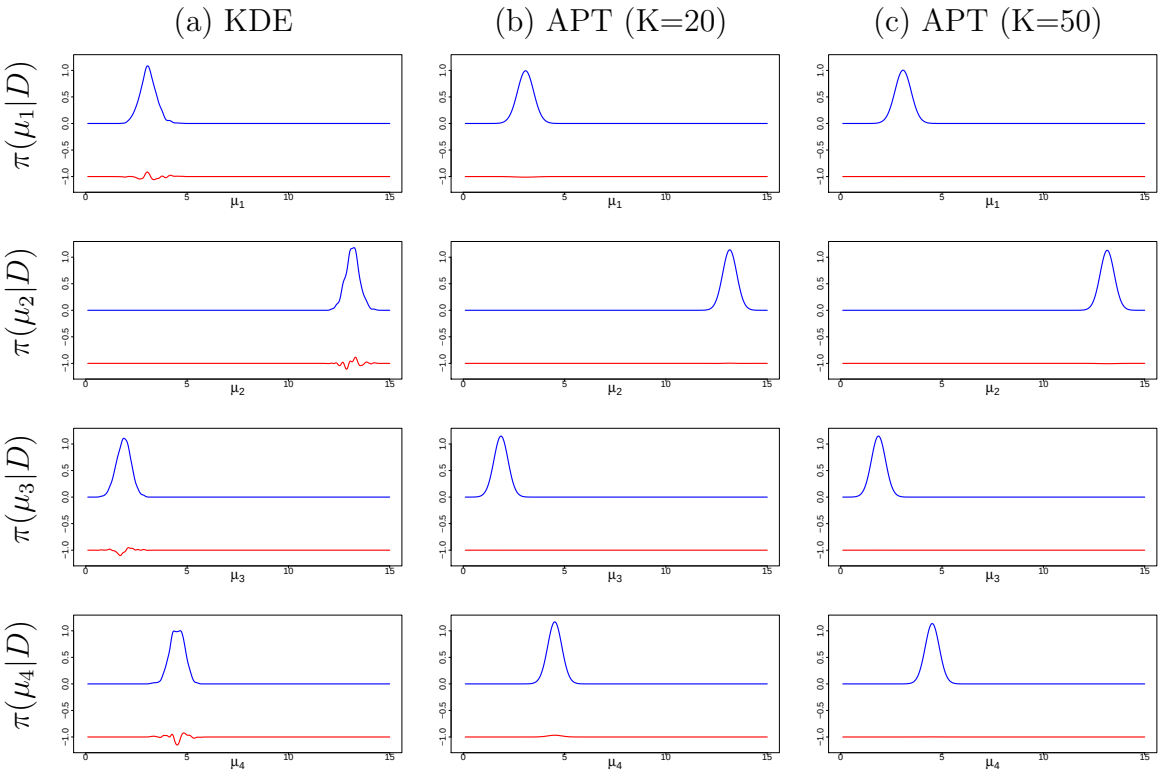


Figure 3.1: The estimated marginal posterior density curve for each μ_j with the fixed MCMC sample size $T = 1,000$ by (a) KDE, (b) APT (K=20), and (c) APT (K=50). The blue curve is the estimated curve by the corresponding approach, and the red one is its difference from the estimated curve by the CMDE with a further subtraction of 1 to separate the two curves.

Figure 3.1 includes the four estimated curves of KDE in the first column (using the function, density(.), with the default setting in R software), and APT curves with $K = 20$ and $K = 50$ in the second and third columns, respectively. In each graph, a blue curve denotes the estimated curve, while a red curve (offset by 1 vertically)

denotes the difference of the estimated curve from the CMDE estimated curve. It shows that KDE deviates from the CMDE for each parameter in the region around its posterior mode. Whereas, APT can produce results very similar to CMDE even when the number of partition subsets is small. When K is increased to $K = 50$, APT completely overlaps with the CMDE, empirically confirming the results established in Remarks 3.2 and 3.3.

Next, we examine the performance of APT in the joint marginal posterior density of μ_1 and μ_4. As in the one-dimensional problem, the CMDE is applicable because the closed form for the conditional posterior density is available:

$$\pi(\mu_1, \mu_4 | \boldsymbol{\mu}_{(-1,-4)}, \sigma^2, D, \boldsymbol{y}_0, a_0) = \prod_{j=1,4} \frac{\sqrt{\frac{n_j + a_0 n_j}{2\pi\sigma^2}} \exp\left(- \frac{(n_j + a_0 n_j)[\mu_j - \frac{\sum_i y_{ij}}{n_j + a_0 n_j}]^2}{2\sigma^2} \right)}{\left[\Phi\left(\frac{\mu_2 - \frac{\sum_i y_{ij}}{n_j + a_0 n_j}}{\sqrt{\frac{\sigma^2}{n_j + a_0 n_j}}} \right) - \Phi\left(\frac{\mu_3 - \frac{\sum_i y_{ij}}{n_j + a_0 n_j}}{\sqrt{\frac{\sigma^2}{n_j + a_0 n_j}}} \right) \right]}.$$

For APT, the conditional working parameter space $\tilde{\Theta}_{\boldsymbol{\theta}_{(-1t,-4t)}}$ is set to be $\{\mu_{3t} < (\mu_{1t}, \mu_{4t}) < \mu_{2t}\}$, then the working parameter space in each dimension is equally divided into 20 or 50 pieces so that $K = 400$ or $K = 2,500$. Figure 3.2 shows the estimated joint marginal posterior densities by KDE (using the function, kde2d(.), with the default setting in R software), APT (K=400), and APT (K=2,500) based on an MCMC sample of size $T = 1,000$. The difference of each from the CMDE estimate is shown below the estimated density surface (using an offset of -1). Figure 3.2 shows that KDE is quite different from the CMDE in this two-dimensional problem, while only minor differences are detectable in the center using APT even with a small number of partition subsets. These slight differences disappear when K is increased.

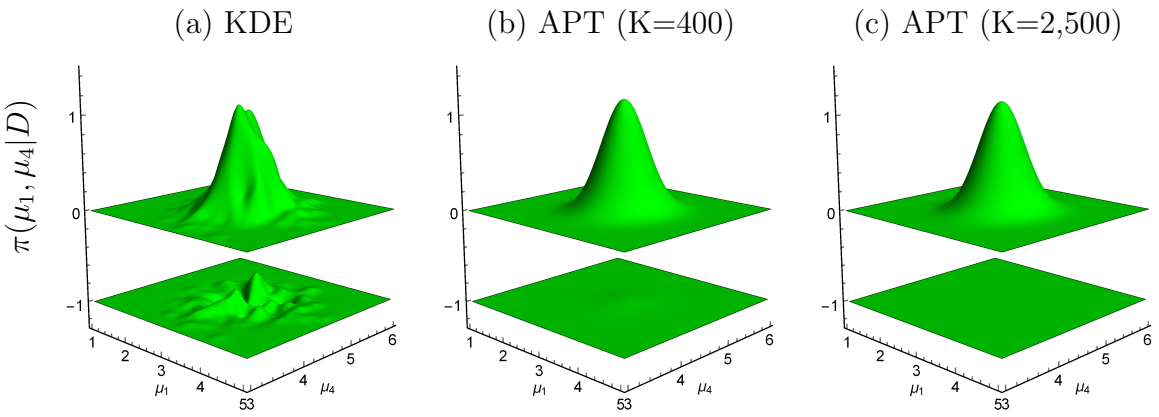


Figure 3.2: The estimated curves of joint marginal posterior density of μ_1 and μ_4 (top half) and their differences from the CMDE (lower half).

Finally, we estimate the joint marginal posterior density of μ_1 and μ_3, where the two parameters of interest are constrained with respect to each other ($\mu_1 > \mu_3$) and the corresponding univariate marginal posterior densities are not far away (Figure 3.1). Under this special structure, the evaluation of the conditional posterior density, $\pi(\mu_1, \mu_3|\boldsymbol{\mu}_{(-1,-3)}, \sigma^2, D, \boldsymbol{y}_0, a_0)$, requires numerical integration. The conditional parameter space $\Theta_{\boldsymbol{\mu}_{(-1t,-3t)},\sigma^2}$ for APT is the trapezoid area bordered by lines $\mu_1 = \mu_2$, $\mu_3 = \mu_4$ and $\mu_1 = \mu_3$ and the conditional working parameter space $\tilde{\Theta}_{\boldsymbol{\mu}_{(-1t,-3t)},\sigma^2}$ is the union of small rectangles inside of $\Theta_{\boldsymbol{\mu}_{(-1t,-3t)},\sigma^2}$ as shown in Figure 3.3.

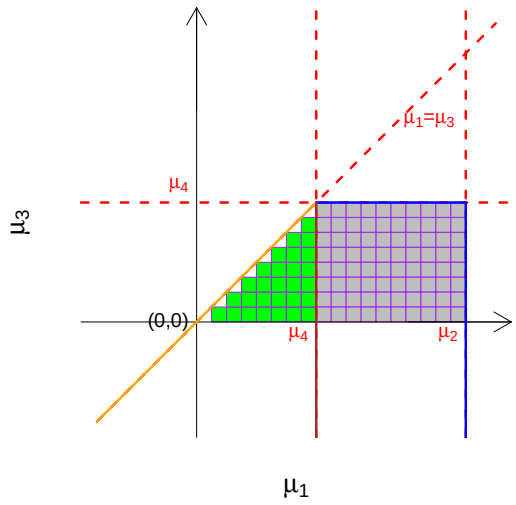


Figure 3.3: The chosen conditional working parameter space of μ_1 and μ_3 and its partition (small rectangles) for APT.

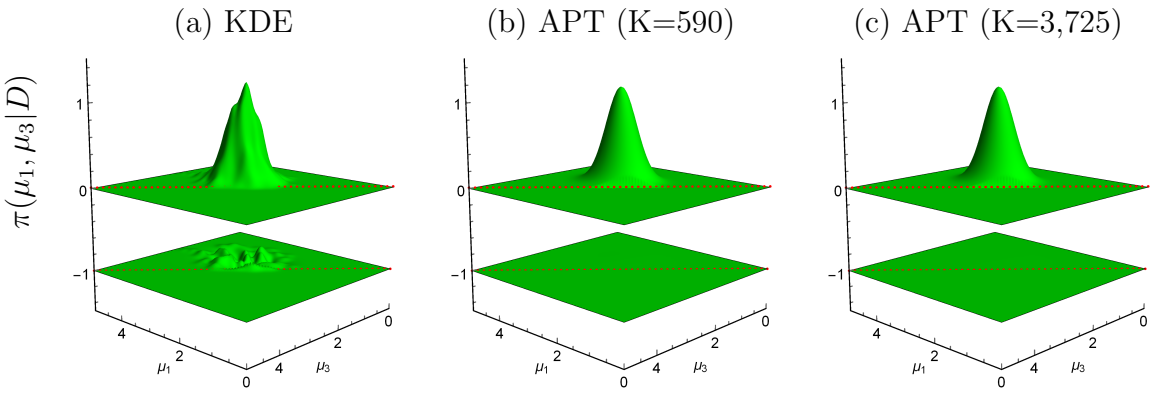


Figure 3.4: The estimated curves of joint marginal posterior density of μ_1 and μ_3 (top half) and their differences from the CMDE (lower half). The red dashed line is the boundary $\mu_1 = \mu_3$ of the constraint $\mu_1 > \mu_3$.

Figure 3.4 shows the performance of KDE (using the function, kde2d(.), with the default setting in the R software) and APT based on an MCMC sample of size $T = 1,000$. APT performs well even for the smaller number of partition subsets, $K = 590$ (190 for green area and 400 for gray area). In contrast, KDE again deviates

from the CMDE. Because it does not use the structure of the known unnormalized posterior density, the KDE has positive values in the area beyond the red dashed line, where the joint marginal posterior density values are supposed to be 0.

3.5 APT for Bayesian Variable Selection

3.5.1 The Basic Formulation

Suppose there are p regression parameters, $\boldsymbol{\beta}$, in the full model M, and $\boldsymbol{\beta} \in R^p$. In variable selection, we let $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\beta}^{(-m)}$ denote the regression parameters included in and excluded from the reduced model m, respectively. Then, $\boldsymbol{\beta}^{(M)} = \boldsymbol{\beta} = ((\boldsymbol{\beta}^{(m)})', (\boldsymbol{\beta}^{(-m)})')'$ holds for all m, and $\boldsymbol{\beta}^{(-M)} = \emptyset$. We further let $\boldsymbol{\alpha}$ denote the nuisance parameters in the model. Under this setting, two conditions required for the Savage-Dickey density ratio in Dickey (1971) are given in Chen et al. (1999) as follows:

Condition I. $L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}|D, m) = L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(-m)} = \mathbf{0}, \boldsymbol{\alpha}|D, M)$, where

$L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}|D, m)$ is the likelihood function under the reduced model m and $L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}^{(-m)} = \mathbf{0}, \boldsymbol{\alpha}|D, M)$ is the likelihood function under the full model evaluated at $\boldsymbol{\beta}^{(m)}$, $\boldsymbol{\beta}^{(-m)} = \mathbf{0}$ and $\boldsymbol{\alpha}$;

Condition II. $\pi(\boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}|m) = \pi(\boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}|\boldsymbol{\beta}^{(-m)} = \mathbf{0}, M)$, where $\pi(\boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}|m)$ is the prior distribution specified under the reduced model m and $\pi(\boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}|\boldsymbol{\beta}^{(-m)} = \mathbf{0}, M)$ is the conditional prior distribution of $\boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}$ given $\boldsymbol{\beta}^{(-m)} = \mathbf{0}$ under the full model.

If both conditions hold, the Bayes factor of the reduced model m over the full model M can be simplified to

$$BF = \frac{c(D|m)}{c(D|M)} = \frac{\pi(\boldsymbol{\beta}^{(-m)} = \mathbf{0}|D, M)}{\pi(\boldsymbol{\beta}^{(-m)} = \mathbf{0}|M)}, \tag{3.19}$$

where $c(D|m)$ and $c(D|M)$ are the marginal likelihoods under the reduced model and the full model, respectively, and $\pi(\boldsymbol{\beta}^{(-m)} = \mathbf{0}|D, M)$ and $\pi(\boldsymbol{\beta}^{(-m)} = \mathbf{0}|M)$ denote the marginal posterior and prior densities evaluated at $\boldsymbol{\beta}^{(-m)} = \mathbf{0}$ from the full model, respectively. It shows that the Bayes factor can be simplified as the function of the marginal posterior density evaluated at zero values of the parameters excluded from the reduced model m so that the model comparison can be done by using a single MCMC sample from the posterior density of the full model. To estimate the marginal posterior density in (3.19), we can use the APT in (3.15) by assigning $\boldsymbol{\theta} = \boldsymbol{\beta}^{(-m)}$, $\boldsymbol{\theta}_0 = \mathbf{0}$, and $\boldsymbol{\xi} = ((\boldsymbol{\beta}^{(m)})', \boldsymbol{\alpha}')'$. Since the parameters in the marginal posterior density are all regression parameters with no constraints, the partition of the conditional working parameter space can be easily constructed via elliptical rings. Assuming the dimension of $\boldsymbol{\theta}$ is q, the elliptical rings are defined as

$$A_k(\boldsymbol{\xi}) = A_k = \{\boldsymbol{\theta} : \ r(k-1)/K \leq \|(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})'\overline{\Sigma}^{-\frac{1}{2}}\| \leq rk/K\}, \ k = 1, 2, \ldots, K,$$

where $\bar{\boldsymbol{\theta}}$ and $\overline{\Sigma}$ are the posterior mean and the posterior covariance matrix of $\boldsymbol{\theta}$, respectively, and r is the radius chosen for the conditional working parameter space. The volume of $A_k(\boldsymbol{\xi})$ can be calculated as follows

$$V(A_k) = V_q|\overline{\Sigma}|^{1/2}\Big[\Big(\frac{rk}{K}\Big)^q - \Big(\frac{r(k-1)}{K}\Big)^q\Big], \tag{3.20}$$

where $|.|$ is the determinant and $V_q = \pi^{q/2}/\Gamma(q/2+1)$ is the volume of a q-dimensional unit hypersphere for $k = 1, 2, \ldots, K$.

Suppose $\{(\boldsymbol{\beta}_t^{(m)}, \boldsymbol{\beta}_t^{(-m)}, \boldsymbol{\alpha}_t), t = 1, ..., T\}$ is an MCMC sample from the posterior distribution under the full model. Let $\boldsymbol{\theta}_t = \boldsymbol{\beta}_t^{(-m)}$ and $\boldsymbol{\xi}_t = ((\boldsymbol{\beta}_t^{(m)})', \boldsymbol{\alpha}_t')'$ for $t = 1, 2, \ldots, T$. Then, we have

$$\hat{\pi}_t(\boldsymbol{\theta} = \mathbf{0}|D, M) = \sum_{k=1}^{K} \frac{q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi}_t) 1\{\boldsymbol{\theta}_t \in A_k\}}{\sum_{k=1}^{K} q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi}_t) V(A_k)} \frac{q(\boldsymbol{\theta} = \mathbf{0}, \boldsymbol{\xi}_t)}{q(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)},$$

where $\boldsymbol{\theta}_k^* \in A_k$ is a representative point in the k^{th} elliptical ring for $k = 1, 2, \ldots, K$. Assume that the marginal prior density $\pi(\boldsymbol{\theta} = \mathbf{0}|M)$ is analytically available. Then, the Bayes factor in log scale is estimated as

$$\log \widehat{BF}_{\text{APT}} = \log \left\{ \frac{1}{T} \sum_{t=1}^{T} \hat{\pi}_t(\boldsymbol{\theta} = \mathbf{0}|D, M) \right\} - \log \pi(\boldsymbol{\theta} = \mathbf{0}|M). \tag{3.21}$$

We use the Overlapping Batch Statistics (OBS) of Schmeiser et al. (1990) to estimate the Monte Carlo (MC) standard error of $\log \widehat{BF}_{\text{APT}}$. Write $\hat{\eta} = \log \widehat{BF}_{\text{APT}}$. Let $\hat{\eta}_b$ denote an estimate of the Bayes factor in log scale via (3.21) using the b^{th} batch, $\{(\boldsymbol{\beta}_t^{(m)}, \boldsymbol{\beta}_t^{(-m)}, \boldsymbol{\alpha}_t), t = b, ..., b + B - 1\}$, of the MCMC sample for $b = 1, ..., T - B + 1$, where $B < T$ is the batch size. Then, the OBS estimated MC standard error of $\hat{\eta}$ is given by

$$\sqrt{\widehat{\text{Var}(\hat{\eta})}} = \left\{ \left[\frac{B}{T - B} \right] \frac{\sum_{b=1}^{T-B+1} (\hat{\eta}_b - \bar{\eta}_{\text{OBS}})^2}{T - B + 1} \right\}^{1/2}, \tag{3.22}$$

where $\bar{\eta}_{\text{OBS}} = \sum_{b=1}^{T-B+1} \hat{\eta}_b/(T - B + 1)$. According to Schmeiser et al. (1990), a reasonable choice of batch size B is $10 \leq T/B \leq 20$.

REMARK 5.1: A reasonable choice of the representative point in each elliptical ring

A_k can be based on the following equation

$$\boldsymbol{\theta}_k^* = r_k^* \overline{\Sigma}^{\frac{1}{2}} \boldsymbol{d} + \bar{\boldsymbol{\theta}}, \tag{3.23}$$

where $r_k^* = r[k/K - 1/(2K)]$, and $\boldsymbol{d}$ is a normalized vector, that is $\|\boldsymbol{d}\| = 1$, which decides the direction of $\boldsymbol{\theta}_k^*$. Since the largest eigenvalue of $\overline{\Sigma}$ can explain the most variation of the MCMC sample, we suggest using its corresponding eigenvector for $\boldsymbol{d}$.

3.5.2 The Ordinal Probit Regression Model

We consider the model selection problem under the ordinal probit regression model, in which the prior distribution does not satisfy Condition II. Thus, the Savage-Dickey density ratio does not hold. Under this situation, we show how marginal posterior density estimation can still be used for estimating ratios of marginal likelihoods between each reduced model and the full model.

Let $\boldsymbol{y} = (y_1, y_2, ..., y_n)'$ denote the vector of observed ordinal responses, each is coded as one value from $0, 1, ..., J-1$, and $\boldsymbol{X}$ denote the $n \times p$ covariate matrix with the i^{th} row equal to the covariates of the i^{th} subject $\boldsymbol{x}_i'$. Let $D = (\boldsymbol{y}, \boldsymbol{X}, n)$. Following Nandram and Chen (1996), the likelihood function is then given by

$$L(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}, \sigma | D) = \prod_{i=1}^{n} \left[\Phi\left(\frac{\breve{\gamma}_{y_{i+1}} - \boldsymbol{x}_i' \breve{\boldsymbol{\beta}}}{\sigma} \right) - \Phi\left(\frac{\breve{\gamma}_{y_i} - \boldsymbol{x}_i' \breve{\boldsymbol{\beta}}}{\sigma} \right) \right], \tag{3.24}$$

where $\breve{\boldsymbol{\gamma}} = (\breve{\gamma}_2, \ldots, \breve{\gamma}_{J-2})'$, $-\infty = \breve{\gamma}_0 < \breve{\gamma}_1 = 0 \leq \breve{\gamma}_2 \leq \cdots \leq \breve{\gamma}_{J-1} = 1 < \breve{\gamma}_J = \infty$ are the cutoff points, $\breve{\boldsymbol{\beta}}$ is a p-dimensional vector of the regression coefficients, and $\sigma > 0$ is a scale parameter. The likelihood function in (3.24) is derived based on the reparameterization of the ordinal regression model with latent variables proposed by

Albert and Chib (1993) to accelerate convergence of the Gibbs sampling algorithm. Another attractive feature of (3.24) is that when $J = 3$, there are no unknown cutoff points. The priors of $(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}, \sigma^2)$ considered in Chen (2005b) are given by

$$\pi(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}, \sigma^2) = \pi(\breve{\boldsymbol{\beta}}|\sigma^2)\pi(\sigma^2)\pi(\breve{\boldsymbol{\gamma}}), \tag{3.25}$$

where $\breve{\boldsymbol{\beta}}|\sigma^2 \sim N(0, \tau_0^{-1}\sigma^2 I_p)$, σ^2 follows an inverse gamma distribution $IG(b_{01}, b_{02})$, $\breve{\boldsymbol{\gamma}}$ is assigned a uniform prior on the constrained space of $\breve{\boldsymbol{\gamma}}$, I_p is the $p \times p$ identity matrix, and $\tau_0 > 0$, $b_{01} > 0$, and $b_{02} > 0$ are prespecified hyperparameters. Then, the resulting posterior distribution is given by

$$\pi(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}, \sigma^2|D) \propto L(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}, \sigma|D)\pi(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}, \sigma^2). \tag{3.26}$$

To carry out Bayesian variable selection under the ordinal probit regression model, we let $(\boldsymbol{x}_i^{(m)}, \breve{\boldsymbol{\beta}}^{(m)})$ and $(\boldsymbol{x}_i^{(-m)}, \breve{\boldsymbol{\beta}}^{(-m)})$ denote the p_m and $p - p_m$ covariates and the corresponding regression coefficients included in and excluded from the m^{th} reduced model, respectively, such that $0 < p_m \leq p$, $\boldsymbol{x}_i = ((\boldsymbol{x}_i^{(m)})', (\boldsymbol{x}_i^{(-m)})')'$, and $\breve{\boldsymbol{\beta}} = ((\breve{\boldsymbol{\beta}}^{(m)})', (\breve{\boldsymbol{\beta}}^{(-m)})')'$. To ensure that the prior of $(\breve{\boldsymbol{\beta}}^{(m)}, \sigma^2, \boldsymbol{\gamma})$ under the reduced model m has the same structure as the one given in (3.25) under the full model, we take

$$\pi(\breve{\boldsymbol{\beta}}^{(m)}, \breve{\boldsymbol{\gamma}}, \sigma^2|m) = \pi(\breve{\boldsymbol{\beta}}^{(m)}|\sigma^2)\pi(\sigma^2)\pi(\breve{\boldsymbol{\gamma}}), \tag{3.27}$$

where $\breve{\boldsymbol{\beta}}^{(m)}|\sigma^2 \sim N(0, \tau_0^{-1}\sigma^2 I_{p_m})$, and $\pi(\sigma^2)$ and $\pi(\breve{\boldsymbol{\gamma}})$ are defined in (3.25). The likelihood function $L(\breve{\boldsymbol{\beta}}^{(m)}, \breve{\boldsymbol{\gamma}}, \sigma|D, m)$ under the reduced model is given by (3.24) with $\boldsymbol{x}_i'\breve{\boldsymbol{\beta}}$ replaced by $(\boldsymbol{x}_i^{(m)})'\breve{\boldsymbol{\beta}}^{(m)}$. For this application, $\boldsymbol{\alpha} = (\breve{\boldsymbol{\gamma}}', \sigma^2)'$. Since $\boldsymbol{x}_i'\breve{\boldsymbol{\beta}}\Big|_{\breve{\boldsymbol{\beta}}^{(-m)}=\mathbf{0}} =$

$(\boldsymbol{x}_i^{(m)})'\breve{\boldsymbol{\beta}}^{(m)}$, Condition I holds. From (3.25), we have

$$\begin{aligned}&\pi(\breve{\boldsymbol{\beta}}^{(m)}, \breve{\boldsymbol{\gamma}}, \sigma^2|\boldsymbol{\beta}^{(-m)} = \mathbf{0})\\ &\propto(\sigma^2)^{-p/2}\exp\Big\{-\frac{\tau_0}{2\sigma^2}(\breve{\boldsymbol{\beta}}^{(m)})'(\breve{\boldsymbol{\beta}}^{(m)})\Big\}(\sigma^2)^{-(b_{01}+1)}\exp(-b_{02}/\sigma^2),\end{aligned}$$

which is clearly not equal to $\pi(\breve{\boldsymbol{\beta}}^{(m)}, \breve{\boldsymbol{\gamma}}, \sigma^2|m)$ given in (3.27) when $p_m < p$. Thus, Condition II is not satisfied. Under this setting, the Bayes factor of the reduced model m over the full model cannot be calculated using (3.19). To circumvent this problem, we take the following one-to-one transformations:

$$\boldsymbol{\beta} = \frac{\breve{\boldsymbol{\beta}}}{\sqrt{\sigma^2}},\ \ \gamma_j = \frac{\breve{\gamma}_j}{\sqrt{\sigma^2}},\ j = 2, \ldots, J-2,\ \text{ and }\ \gamma_{J-1} = \frac{1}{\sqrt{\sigma^2}}. \tag{3.28}$$

The absolute value of the determinant of the Jacobian of the transformations from $(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}, \sigma^2)$ to $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is $\gamma_{J-1}^{-(p+J)}$. Write $\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_{J-1})'$. After the transformations, the likelihood function of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}|D) = \prod_{i=1}^{n}\left[\Phi\left(\gamma_{y_{i+1}} - \boldsymbol{x}_i'\boldsymbol{\beta}\right) - \Phi\left(\gamma_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right], \tag{3.29}$$

and the prior of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is given by

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \exp\Big(-\frac{\tau_0}{2}\boldsymbol{\beta}'\boldsymbol{\beta}\Big)\gamma_{J-1}^{-J+2b_{01}+2}\exp\Big(-b_{02}\gamma_{J-1}^2\Big), \tag{3.30}$$

where $\boldsymbol{\beta} \in R^p$ and $-\infty = \gamma_0 < \gamma_1 = 0 \le \gamma_2 \le \cdots \le \gamma_{J-1} < \gamma_J = \infty$. Now, applying the Bayesian variable selection procedure to $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with the likelihood function and the prior given by (3.29) and (3.30), we can show that both Conditions 1 and 2 are satisfied. Therefore, we can use the APT via (3.19) to compute the Bayes factor.

Although the use of (3.29) and (3.30) leads to easy computation of the Bayes factor, sampling $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ from the posterior distribution induced by (3.29) and (3.30) is not as efficient as sampling $(\breve{\boldsymbol{\beta}}, \breve{\gamma}, \sigma^2)$ from (3.26) (e.g. Nandram and Chen, 1996). Since the transformations in (3.28) are one-to-one, MCMC samples of $(\breve{\boldsymbol{\beta}}, \breve{\gamma}, \sigma^2)$ from (3.26), which can be generated by Nandram-Chen algorithm, can be directly used to obtain MCMC samples of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ via (3.28). Therefore, by combining these two settings of the ordinal probit regression model, we can achieve both convenient implementation of MCMC sampling and efficient computation of Bayes factors.

3.5.3 Analysis of the Prostate Cancer Data

We apply the ordinal probit regression model to the prostate cancer data ($n = 713$) as in Chen (2005b). We examine the relationships between a clinical ordinal response variable, Pathological Extracapsular Extension (PECE, y), and three covariates: Prostate Specific Antigen (PSA, x_1), Clinical Gleason Score (GLEAS, x_2, x_3), and Clinical Stage (CSTAGE, x_4). Here, PECE takes values of 0, 1, or 2, where 0 means that there is no cancer cell present in or near the capsule, 1 denotes that the cancer cells extend into but not through the capsule, and 2 indicates that cancer cells extend through the capsule. PSA is a continuous variable while CSTAGE is a binary outcome, which is assigned to 1 if the 1992 American Joint Commission on Cancer (AJCC) clinical stage T-category is 1, and assigned to 2 if the T-category is 2 or higher. As for GLEAS, considering similar biologic behaviors of tumors (well-differentiated, moderately-differentiated, or poorly-differentiated), it is trichotomized by two dummy variables: $x_2 = 1$ if GLEAS$= 7$, otherwise $x_2 = 0$; $x_3 = 1$ if GLEAS> 7, otherwise $x_3 = 0$. We note that in Chen (2005b), GLEAS was

treated as a continuous variable. Since x_2 and x_3 are defined from the same variable GLEAS, we either include both x_2 and x_3 in the model or exclude them together from the model.

In this example, since $J = 3$, there are no unknown cutpoints in (3.24). In (3.25), we take $b_{01} = 1$ and $b_{02} = 0.1$. Due to lack of historical information about τ_0 and for avoiding the phenomenon of the Bartlett's or Lindley's paradox (see Jeffreys, 1998; Lindley, 1957), we first use an empirical Bayes method to find the best prior setting of τ_0 for the full model based on the marginal likelihood criterion. The PWK estimator was used to calculate the marginal likelihoods under different τ_0. Table 3.1 summarizes the marginal likelihood estimate and its Monte Carlo standard error (MCSE) for each τ_0 using an MCMC sample of size $T = 10,000$. The MCSE is calculated by (3.22) with the batch size $B = 1,000$.

Table 3.1: Marginal likelihood estimate under different precision τ_0

τ_0	$\widehat{c(D\|M)}$	MCSE
0.1	-766.398	0.009
1	-760.791	0.007
10	-756.690	0.011
15	-756.516	0.011
20	-756.641	0.009
30	-757.240	0.005
40	-758.065	0.009
50	-758.928	0.011

From the results in Table 3.1, we choose $\tau_0 = 15$ as the best prior under the full

model. On this basis, we use the Bayes factor for variable selection. Following (3.28), we first transform the regression parameters to satisfy the conditions for Savage-Dickey density ratio. Then, we apply APT to estimate the marginal posterior density at zero values in (3.19) by constructing an elliptical-ring partition of the working parameter space with $r = 4$ and $K = 1, 5,$ or 10. Table 3.2 shows the 95% HPD interval for each β, Bayes factor estimates of the reduced models over the full model in log scale, their estimated MCSE's using (3.22) with the batch size $B = 1,000$, and their relative MCSEs denoted by rMCSEs, each of which is defined as MCSE divided by its $|\log \widehat{BF}|$. The results are obtained based on an MCMC sample of size $10,000$. From Table 3.2, we see that the full model is preferred because all the values of $\log BF$ are negative. This result is also supported by the 95% HPD intervals, which do not contain 0. In addition, we observe that each MCSE dramatically drops when a reasonable number of subsets ($K = 5$ or 10) is used. The fact that each MCSE is relatively small compared to the magnitude of $\log BF$ empirically demonstrates that our proposed method is very accurate. Table 3.2 also shows the results for $\tau_0 = 0.1$, where the inconsistency between the Bayes factors and the 95% HPD intervals is evident. Even though the 95% HPD intervals suggest that all covariates are important for the case of $\tau_0 = 0.1$, the Bayes factor tends to favor the reduced model, which is known as the Bartlett's or Lindley's paradox.

Table 3.2: HPD interval and Bayes Factor in log scale when $\tau_0 = 15$ and $\tau_0 = 0.1$

	$\tau_0 = 15$			$\tau_0 = 0.1$		
Variable	Mean	SD	95% HPD	Mean	SD	95% HPD
x_0	0.573	0.071	(0.434, 0.716)	0.581	0.070	(0.446, 0.718)
x_1	0.494	0.103	(0.303, 0.701)	0.512	0.103	(0.322, 0.723)
x_2	0.129	0.063	(0.013, 0.257)	0.128	0.062	(0.008, 0.253)
x_3	0.241	0.070	(0.108, 0.378)	0.239	0.069	(0.103, 0.373)
x_4	0.146	0.064	(0.021, 0.270)	0.145	0.063	(0.027, 0.273)
	APT(K=1)			APT(K=1)		
Model	$\log BF$	MCSE	rMCSE	$\log BF$	MCSE	rMCSE
(x_1, x_2, x_3)	-1.035	0.059	5.72%	1.437	0.094	6.52%
(x_2, x_3, x_4)	-17.184	0.091	0.53%	-15.126	0.191	1.27%
(x_1, x_4)	-4.740	0.092	1.93%	0.424	0.065	15.30%
(x_2, x_3)	-20.023	0.067	0.34%	-15.722	0.167	1.06%
(x_1)	-6.404	0.085	1.33%	1.125	0.080	7.10%
(x_4)	-28.045	0.126	0.45%	-21.985	0.081	0.37%
	APT(K=5)			APT(K=5)		
(x_1, x_2, x_3)	-1.033	0.006	0.59%	1.471	0.010	0.70%
(x_2, x_3, x_4)	-17.086	0.039	0.23%	-15.425	0.060	0.39%
(x_1, x_4)	-4.645	0.024	0.52%	0.330	0.034	10.23%
(x_2, x_3)	-19.848	0.061	0.31%	-15.791	0.069	0.44%
(x_1)	-6.263	0.035	0.56%	1.213	0.053	4.33%
(x_4)	-27.993	0.047	0.17%	-21.902	0.060	0.28%
	APT(K=10)			APT(K=10)		
(x_1, x_2, x_3)	-1.034	0.006	0.60%	1.471	0.010	0.69%
(x_2, x_3, x_4)	-17.088	0.039	0.23%	-15.424	0.057	0.37%
(x_1, x_4)	-4.649	0.023	0.49%	0.327	0.031	9.51%
(x_2, x_3)	-19.844	0.059	0.30%	-15.783	0.065	0.41%
(x_1)	-6.268	0.032	0.51%	1.200	0.048	3.96%
(x_4)	-27.989	0.046	0.16%	-21.893	0.058	0.26%

We also use the KDE with several choices of the kernel function to compute $\pi(\boldsymbol{\beta}^{(-m)} = \mathbf{0}|D, M)$ in (3.19). The KDE estimates and corresponding MCSEs of $\log BF$ for the model (x_1, x_2, x_3) when $\tau_0 = 0.1$ are 1.466 and 0.060 under the Gaussian kernel; 1.475 and 0.059 under the Epanechnikov kernel; 1.446 and 0.069 under the rectangular kernel; 1.472 and 0.059 under the triangular kernel; 1.473 and 0.058 under the biweight kernel; 1.472 and 0.058 under the cosine kernel; and 1.475 and 0.058 under the optcosine kernel. These MCSEs are larger than 0.010 shown in Table

3.2 using APT with $K = 5$ or $K = 10$.

To examine empirical performance of the estimates (3.15) and (3.16) for $\pi(\boldsymbol{\theta}_{01} = 0|D)$ and $\pi(\boldsymbol{\theta}_{02} = 0|\boldsymbol{\theta}_{01} = 0, D)$ using two MCMC samples, we consider models (x_1, x_4) and (x_2, x_3) when $\tau_0 = 0.1$. In order to make a fair comparison, we generate 5,000 MCMC sample points from both the full posterior distribution and the conditional posterior distribution. The resulting values of $\log BF$, MCSE, and rMCSE are 0.331, 0.021, and 6.42%, respectively, under model (x_1, x_4), and -15.849, 0.053, and 0.33%, respectively, under model (x_2, x_3). In these computations, we fixed $r = 4$ and $K = 10$. In both cases, the estimates of $\log BF$ using (3.15) and (3.16) for $\pi(\boldsymbol{\theta}_{01} = 0|D)$ and $\pi(\boldsymbol{\theta}_{02} = 0|\boldsymbol{\theta}_{01} = 0, D)$ have smaller MCSEs and rMCSE. These results empirically demonstrate that the approach based on the dimension reduction is quite promising in obtaining a more accurate estimate of $\log BF$.

3.6 Discussion

Marginal posterior density estimation provides complete information for the parameters of interest in Bayesian inference. While it is known that the conditional marginal density estimator is the best (minimizing the variance) in the class of importance weighted marginal density estimators (Chen, 1994), the method for realizing this estimator has not been accessible when the closed form of the conditional posterior density is not available. To circumvent it, we propose an adaptive partition weighted (APT) marginal density estimator that requires only the unnormalized posterior density and an MCMC sample from the joint posterior distribution. Our method is constructed by first partitioning the subset of the support of the conditional posterior

distribution, and then estimating the marginal posterior density at a fixed point of the focal parameter vector by assigning a weighted average to the ratios of the unnormalized posterior density evaluated at the MCMC sample. The focal parameters in the numerator are set to this fixed point, and weights are assigned locally using a representative value in each partitioned subset. We show that our estimator is unbiased, and approaches the conditional marginal density estimator when the number of partition subsets is large. We illustrate our method APT with two examples. The first one is a constrained parameter space problem, where the conditional marginal density estimator for the parameters of interest can be evaluated. We show our estimator performs as well as the gold standard (the conditional marginal density estimator) in this case. The other example involves variable selection under the ordinal probit regression model. In this example, we demonstrate the usefulness of the marginal posterior density estimation for computing Bayes factors between the reduced model with a subset of variables and the full model with all variables using a single MCMC sample from the posterior distribution under the full model.

As discussed in Section 3 and shown in Section 5, APT is useful not only in computing marginal posterior densities but also in estimating marginal likelihoods through Chib's identity and Bayes factors via the Savage-Dickey density ratio. As further demonstrated in Section 5, the APT method in conjunction with the conditional posterior density estimate leads to a more efficient estimate, which will be potentially useful for Bayesian computation problems with high-dimensional parameters. Currently, KDE is a standard method for computing and displaying marginal posterior densities using MCMC samples in existing Bayesian software. As empirically shown in Sections 4 and 5, APT can be much closer to CMDE and produce a more accurate estimate of Bayes factor than KDE. Since the proposed method requires only

the unnormalized posterior density and an MCMC sample from the joint posterior distribution, APT has a potential to become a "black-box" algorithm, which can be implemented in existing software including SAS, OpenBugs (Thomas et al., 2006), as well as more specialized Bayesian software such as that used in phylogenetics: MrBayes 3.2 (Ronquist et al., 2012) and BEAST2 (Bouckaert et al., 2014).

Chapter 4

Marginal Likelihoods of Phylogenetic Models Using a Posterior Sample

4.1 Introduction

In Bayesian phylogenetics model selection, many Monte Carlo (MC) methods have been developed in recent decades to estimate the marginal likelihood (normalizing constant) for a fixed tree topology under a substitution model. This constant value measures a average fit of the model to the data over the prior information for the specific tree, and hence can be used as a criterion to select the best model. The available methods include the harmonic mean (HM) method by Newton and Raftery (1994), the inflated density ratio (IDR) method by Petris and Tardella (2003, 2007), the thermodynamic integration (TI) method by Lartillot and Philippe (2006), the stepping-stone (SS) method by Xie et al. (2011), and the generalized stepping-stone

(GSS) method by Fan et al. (2011). Under certain ergodic conditions, they are all shown to produce consistent marginal likelihood estimators.

Wu et al. (2014) and Holder et al. (2014) further extend the HM, IDR, SS, and GSS methods to estimating the marginal likelihood for variable tree topology. This value is preferred by systematists, since the marginal likelihood is evaluated by summing over all tree topologies, and hence, models can be compared based on the overall marginal likelihood rather than being restricted to a certain topology. In the HM method, the joint prior distribution of the tree and parameters in the substitution model is used as a weight to the inverse of the joint posterior kernel evaluated at the points of an Markov chain Monte Carlo (MCMC) sample. For the IDR method, the prior of topology times a perturbed density, which is based on the conditional posterior kernel given a topology T, is assigned as the weight. For SS and GSS, the computation is decomposed into a series of telescoping ratios of two marginal likelihoods using power posterior kernels, where each ratio requires an MCMC sample from a power posterior distribution and is estimated by the importance sampling approach. Then, the overall marginal likelihood estimate is obtained simply by the multiplication of all these estimated ratios.

The partition weighted kernel estimator (PWK) developed in Chapter 2 can also be applied to the marginal likelihood calculation for a fixed topology problem. The PWK estimator is constructed by first partitioning the working parameter space (where the posterior kernel is bounded away from zero), and in each partition subset, assigning a local weight to the inverse of the posterior kernel evaluated at an MCMC sample. This method is essentially a generalization of the HM and IDR methods, but produces an estimator with smaller variance as shown in Chapter 2. In addition, compared to SS and GSS, the PWK is more attractive since it only needs a single

MCMC sample from the joint posterior distribution. In this chapter, we extend the working parameter space to the discrete tree topology space and propose the variable-topology partition weighted kernel (VPWK) estimator for the marginal likelihood. Furthermore, the use of estimated posterior density of selected tree as a re-weighted function of this new method also improves the efficiency.

In the rest of this chapter, we first formulize the problem and introduce the existing approaches in Section 2. In Section 3, we update the PWK estimator to variable tree topology and examine its theoretical properties. Based on this method, the VPWK estimator and its related properties are developed. In Section 4, we use the *rcbl* data set in Lewis and Trainor (2012) for the real data analysis and compare the results of VPWK to the VSS and VGSS estimates in the general time reversible plus Gamma model. Finally, we conclude with a discussion in Section 5.

4.2 Preliminary

Suppose $\boldsymbol{\theta}_\tau$ is a p_τ-dimensional vector of parameters affiliated with a binary labeled unrooted tree topology τ on a set of S taxa for $\tau \in \Gamma = \{\tau_1, \tau_2, ..., \tau_T\}$, where Γ is the discrete tree topology space, and T is the total number of distinct tree topologies and equal to $(2S-5)!! = (2S-5)\ (2S-7)\dots(3)(1)$. Let $\boldsymbol{y}$ denote the data. Also let $f(\boldsymbol{y}|\boldsymbol{\theta}_\tau, \tau)$ be the likelihood function and $\pi(\boldsymbol{\theta}_\tau, \tau) = \pi(\boldsymbol{\theta}_\tau|\tau)\pi(\tau)$ be the prior distribution for $(\boldsymbol{\theta}_\tau, \tau)$. Then, the posterior kernel $q(\boldsymbol{\theta}_\tau, \tau)$ is $f(\boldsymbol{y}|\boldsymbol{\theta}_\tau, \tau)\pi(\boldsymbol{\theta}_\tau, \tau)$ with support $\boldsymbol{\Omega} = \Gamma \times \boldsymbol{\Theta}_\tau$, where $\boldsymbol{\Theta}_\tau = \{\boldsymbol{\theta}_\tau : (\boldsymbol{\theta}_\tau, \tau) \in \boldsymbol{\Omega}\}$ is the support of conditional posterior function $\boldsymbol{\theta}_\tau$ given τ, and the unnormalized conditional posterior function

$\boldsymbol{\theta}_\tau$ given τ is defined as

$$q(\boldsymbol{\theta}_\tau|\tau) = f(\boldsymbol{y}|\boldsymbol{\theta}_\tau, \tau)\pi(\boldsymbol{\theta}_\tau|\tau).$$

Hence, the conditional marginal likelihood given τ, that is the marginal likelihood for a fixed topology, is

$$c(\tau) = \int_{\boldsymbol{\Theta}_\tau} q(\boldsymbol{\theta}_\tau|\tau) d\boldsymbol{\theta}_\tau, \tag{4.1}$$

and the overall marginal likelihood, that is the marginal likelihood for a variable topology, is given by

$$c = \sum_{\tau \in \Gamma} c(\tau)\pi(\tau). \tag{4.2}$$

Although c is a function of $c(\tau)$ for $\tau \in \Gamma$, it is impractical to evaluate this summation by brute-force approach since the size of Γ is often large. To estimate (4.2), Wu et al. (2014) and Holder et al. (2014) update the HM, IDR, SS, and GSS estimators originally designed for the marginal likelihood estimation under a fixed tree topology to a variable tree topology. We examine each approach in the remaining of this section.

Suppose $\{(\boldsymbol{\theta}_\tau^{(n)}, \tau^{(n)}), n = 1, 2, \ldots, N\}$ is an MCMC sample from the posterior distribution $\pi(\boldsymbol{\theta}_\tau, \tau|\boldsymbol{y}) = q(\boldsymbol{\theta}_\tau, \tau)/c$, the HM estimator is given by

$$\hat{c}_{\text{VHM}} = \left[\frac{1}{N} \sum_{n=1}^{N} \frac{1}{f(\boldsymbol{y}|\boldsymbol{\theta}_{\tau^{(n)}}^{(n)}, \tau^{(n)})} \right]^{-1}. \tag{4.3}$$

Under certain ergodic conditions, it is shown that

$$\hat{c}_{\text{VHM}} \overset{a.s.}{\to} c.$$

For the IDR, based on the conditional posterior kernel $q(\boldsymbol{\theta}_\tau|\tau)$, an perturbed density $q_r(\boldsymbol{\theta}_\tau|\tau)$ is proposed as

$$q_r(\boldsymbol{\theta}_\tau|\tau) = \begin{cases} q(\mathbf{0}|\tau) & \text{if } ||\boldsymbol{\theta}_\tau|| \leq r_\tau, \\ q(w(\boldsymbol{\theta}_\tau)|\tau) & \text{if } ||\boldsymbol{\theta}_\tau|| > r_\tau, \end{cases}$$

where r_τ is the chosen radius for each tree topology and $w(\boldsymbol{\theta}_\tau) = \boldsymbol{\theta}_\tau \times (1 - r_\tau^{p_\tau}/||\boldsymbol{\theta}_\tau||^{p_\tau})^{1/p_\tau}$. It follows,

$$\int_{\boldsymbol{\Theta}_\tau} q_r(\boldsymbol{\theta}_\tau) d\boldsymbol{\theta}_\tau = \int_{||\boldsymbol{\theta}_\tau|| \leq r_\tau} q_r(\boldsymbol{\theta}_\tau|\tau) d\boldsymbol{\theta}_\tau + \int_{||\boldsymbol{\theta}_\tau|| > r_\tau} q_r(\boldsymbol{\theta}_\tau|\tau) d\boldsymbol{\theta}_\tau = q(\mathbf{0}|\tau) b_{r_\tau} + c,$$

where b_{r_τ} = Volume of the ball $\{\boldsymbol{\theta}_\tau : \ ||\boldsymbol{\theta}_\tau|| \leq r_\tau\} = \pi^{p_\tau/2} r^{p_\tau}/\Gamma(p_\tau/2+1)$. Then, the IDR estimator is given by

$$\hat{c}_{\text{VIDR}} = \frac{1}{\frac{1}{N}\sum_{n=1}^{N} \frac{1}{q(\mathbf{0}|\tau^{(n)}) b_{\tau^{(n)}}} \left[\frac{q_r(\boldsymbol{\theta}_\tau^{(n)}|\tau^{(n)})}{q(\boldsymbol{\theta}_\tau^{(n)}|\tau^{(n)})} - 1 \right]}. \tag{4.4}$$

It is also shown that

$$\hat{c}_{\text{VIDR}} \overset{a.s.}{\to} c.$$

In the GSS, the power posterior distribution is proposed as

$$p_\beta(\boldsymbol{\theta}_\tau, \tau) = \frac{q_\beta(\boldsymbol{\theta}_\tau|\tau)\pi_\beta(\tau)}{c_\beta}, \tag{4.5}$$

where $0.0 \leq \beta \leq 1.0$,

$$\pi_\beta(\tau) = [\pi(\tau)]^\beta [\pi^*(\tau)]^{1-\beta},$$

$$q_\beta(\boldsymbol{\theta}_\tau|\tau) = [q_\beta(\boldsymbol{\theta}_\tau|\tau)]^\beta [\pi^*_\beta(\boldsymbol{\theta}_\tau|\tau)]^{1-\beta},$$

$\pi^*_\beta(\boldsymbol{\theta}_\tau|\tau)$ is a conditional reference prior distribution given τ, and $\pi^*(\tau)$ is a reference prior distribution for the tree topology. When $\beta = 0$, $c_0 = 1$ due to a proper reference distribution proposed; when $\beta = 1$, $c_1 = c$ is the overall marginal likelihood. With this setting, the computation of c is decomposed into

$$c = \prod_{h=1}^{H} \frac{c_{\beta_h}}{c_{\beta_{h-1}}}, \tag{4.6}$$

where β_h is often chosen as h/H, $h = 1, 2, \ldots, H$. Given an MCMC sample is available from $p_{\beta_{h-1}}(\boldsymbol{\theta}_\tau, \tau)$, the estimator for $u_h = c_{\beta_h}/c_{\beta_{h-1}}$ is given by

$$\hat{u}_{\text{VSS},h} = \frac{1}{N} \sum_{n=1}^{N} \left[\frac{q(\boldsymbol{\theta}^{(n)}_{\tau^{(n)}}|\tau^{(n)})}{\pi^*(\boldsymbol{\theta}^{(n)}_{\tau^{(n)}}|\tau^{(n)})} \right]^{\beta_h - \beta_{h-1}}. \tag{4.7}$$

Then,the overall marginal likelihood is calculated by

$$\hat{c}_{\text{VSS}} = \prod_{h=1}^{H} \hat{u}_{\text{VSS},h}. \tag{4.8}$$

Under certain ergodic conditions,

$$\hat{u}_{\text{VSS},h} \xrightarrow{a.s.} \frac{c_{\beta_h}}{c_{\beta_{h-1}}},$$

so that

$$\hat{c}_{\text{VSS}} \xrightarrow{a.s.} c.$$

For the GSS, it improves the SS method by proposing a reference distribution close to the posterior distribution. Then, a more efficient estimator is developed since a smaller value of H is required to fill in the gap of dissimilarity between the posterior and reference prior densities, or more stable estimates of u_h can be obtained given the same H is used.

In spite of consistency to the overall marginal likelihood for all four estimators, both HM and IDR estimators are shown being less efficient than the PWK in Chapter 2, and both SS and GSS require MCMC samples from a series of power posterior distribution so that more computation time is expected. The motivation is initialized by by these two facts and the PWK. We propose a new estimator, a variable-topology partition weighted kernel (VPWK) estimator, in next section. This estimator only requires a known posterior kernel and an MCMC sample from the posterior distribution, and allows us to only focus on few frequently sampled trees rather than all sampled trees.

4.3 PWK Estimator in Variable Topology

To estimate the marginal likelihood in (4.2), we first derive the PWK estimator in Chapter 2 for a variable tree topology. We then examine its theoretical properties, and further develop an improved estimator, variable-topology partition weighted kernel (VPWK) estimator. The proposed method only requires the known posterior kernel and a single MCMC sample from the posterior distribution. With minimal assumptions, the VPWK is consistent to the reciprocal of overall marginal likelihood and has a finite variance. It can be optimalized by increasing the number of partition

subsets in each chosen tree. Most of all, this new estimator is free from the curse of unsampled or seldom-sampled topologies in an MCMC sample, which are common in a variable-topology problem.

4.3.1 General Monte Carlo Estimator

Suppose $\tilde{\boldsymbol{\Theta}}_\tau = \{\boldsymbol{\theta}_\tau : q(\boldsymbol{\theta}_\tau|\tau) > 0\}$ is the working parameter space of $\boldsymbol{\Theta}_\tau$, and $\{A_1(\tau), A_2(\tau), \ldots, A_{K_\tau}(\tau)\}$ forms a partition of $\tilde{\boldsymbol{\Theta}}_\tau$. Given an MCMC sample $\{(\boldsymbol{\theta}^{(n)}_{\tau^{(n)}}, \tau^{(n)}), n = 1, 2, \ldots, N\}$ is available from the posterior distribution $\pi(\boldsymbol{\theta}_\tau, \tau|\boldsymbol{y})$, we update the PWK estimator for the reciprocal of the marginal likelihood in a variable-topology problem as follows

$$\frac{\hat{1}}{c_{\text{PWK}}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\Gamma} 1\{\tau^{(n)} = t\} \sum_{k_t=1}^{K_t} \frac{1}{\sum_{k_t=1}^{K_t} w_{k_t} V(A_{k_t}(t))} \left[\frac{w_{k_t}}{q(\boldsymbol{\theta}_t^{(n)}|t)} 1\{\boldsymbol{\theta}_t^{(n)} \in A_{k_t}(t)\}\right]. \tag{4.9}$$

It can be shown that under some mild regularity conditions, $\widehat{1/c}_{\text{PWK}}$ is consistent to $1/c$.

Proof: Under certain ergodic conditions, we first have

$$\begin{aligned}
&\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\Gamma}1\{\tau^{(n)}=t\}\sum_{k_t=1}^{K_t}\frac{1}{\sum_{k_t=1}^{K_t}w_{k_t}V(A_{k_t}(t))}\left[\frac{w_{k_t}}{q(\boldsymbol{\theta}_t^{(n)}|t)}1\{\boldsymbol{\theta}_t^{(n)}\in A_{k_t}(t)\}\right]\\
=&\sum_{\tau\in\Gamma}\int_{A_{k_\tau}(\tau)}\sum_{t\in\Gamma}1\{\tau=t\}\sum_{k_t=1}^{K_t}\frac{1}{\sum_{k_t=1}^{K_t}w_{k_t}V(A_{k_t})}\frac{w_{k_t}}{q(\boldsymbol{\theta}_t|t)}\frac{q(\boldsymbol{\theta}_\tau,\tau)}{c}d\boldsymbol{\theta}_\tau\\
=&\sum_{t\in\Gamma}\sum_{k_t=1}^{K_t}\frac{1}{\sum_{k_t=1}^{K_t}w_{k_t}V(A_{k_t}(t))}\int_{A_{k_t}(t)}\frac{w_{k_t}}{q(\boldsymbol{\theta}_t|t)}\frac{q(\boldsymbol{\theta}_t,t)}{c}d\boldsymbol{\theta}_t\\
=&\sum_{t\in\Gamma}\sum_{k_t=1}^{K_t}\frac{1}{\sum_{k_t=1}^{K_t}w_{k_t}V(A_{k_t}(t))}\int_{A_{k_t}(t)}\frac{w_{k_t}\pi(t)}{c}d\boldsymbol{\theta}_t\\
=&\sum_{t\in\Gamma}\frac{\pi(t)}{c}=\frac{1}{c}
\end{aligned}$$

so that

$$\frac{\hat{1}}{c}_{\text{PWK}}\xrightarrow{a.s.}\frac{1}{c}.$$

□

According to Chapter 2, the variance of $\sum_{k_t=1}^{K_t}\frac{1}{\sum_{k_t=1}^{K_t}w_{k_t}V(A_{k_t}(t))}\left[\frac{w_{k_t}}{q(\boldsymbol{\theta}_t^{(n)}|t)}\,1\{\boldsymbol{\theta}_t^{(n)}\in A_{k_t}(t)\}\right]$ can be minimized by partitioning $\tilde{\boldsymbol{\Theta}}_t$ into an abundant number of subsets so that the sufficient condition for the optimality, that is homogeneity of $q(\boldsymbol{\theta}_t|t)$ in each subset, is satisfied. For each subset, we assign $q(\boldsymbol{\theta}_{k_t}^*|t)$ as the local weight w_{k_t}, where $\boldsymbol{\theta}_{k_t}^*$ is a representative point for the subset $A_{k_t}(t)$, and obtain an approximately optimal PWK estimator. Based on the same idea, suppose for each $\tau\in\Gamma$, K_τ is large enough to insure the homogeneity of $q(\boldsymbol{\theta}_\tau|\tau)$ in each subset, the optimal estimator of

(4.9) with minimum variance is given by

$$\frac{\hat{1}}{c_{\text{PWK}}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\Gamma} 1\{\tau^{(n)}=t\}\sum_{k_t=1}^{K_t}\frac{1}{\sum_{k_t=1}^{K_t} q(\boldsymbol{\theta}^*_{k_t}|t)V(A_{k_t}(t))}\left[\frac{q(\boldsymbol{\theta}^*_{k_t}|t)}{q(\boldsymbol{\theta}^{(n)}_t|t)}1\{\boldsymbol{\theta}^{(n)}_t\in A_{k_t}(t)\}\right] \tag{4.10}$$

We see that (4.10) is a consistent estimator to $1/c$ by the ergodic theorem and approximates the optimal estimator in the class of estimators in (4.9) when $K_\tau \to \infty$ for $\tau \in \Gamma$. Nevertheless, when the number of taxa S increases, the size of Γ increases dramatically resulting in a huge computation burden from partitioning the working parameter space for each sampled topology. In addition, the fact that most of the MCMC sample comprises few dominant topologies motivates us to improve (4.10) by excluding topologies of low frequency in the MCMC sample. In the next section, we extend the idea of the working parameter space from continuous parameters to the discrete tree topology space, and propose the variable-topology partition weighted kernel (VPWK) estimator to estimate the marginal likelihood in a variable-topology problem.

4.3.2 New Monte Carlo Estimator

The number of possible tree topology increases dramatically with number of taxa, which results in the impossibility of constructing the working parameter space and its partition subset for each tree. Moreover, knowing the fact that few tree topologies account for most of the MCMC sample, we improve (4.10) by introducing the idea

of a subset Ξ of $\{\tau, \tau = 1, 2, \ldots, T\}$. Essentially, it is an extended concept of the working parameter space in Chapter 2 to a discrete parameter. We show that the VPWK estimator is also consistent to $1/c$. Under certain ergodic conditions and Assumptions 1 and 2, it is shown to have finite variance.

Suppose Ξ is a subset of Γ, where Ξ excludes any topology with a small value of $\hat{\pi}(\tau|\boldsymbol{y})$. We propose the new estimator as

$$\begin{aligned}
&\hat{\frac{1}{c}}_{VPWK} \\
=&\frac{\frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\Xi}1\{\tau^{(n)}=t\}\sum_{k_t=1}^{K_t}\frac{1}{\sum_{k_t=1}^{K_t}q(\boldsymbol{\theta}^*_{k_t}|t)V(A_{k_t}(t))}\left[\frac{q(\boldsymbol{\theta}^*_{k_t},t)}{q(\boldsymbol{\theta}^{(n)}_t,t)}1\{\boldsymbol{\theta}^{(n)}_t\in A_{k_t}(t)\}\right]}{\sum_{t\in\Xi}\pi(t)}.
\end{aligned} \tag{4.11}$$

In order to establish consistency and finite variance of the new estimator, we introduce two assumptions.

Assumption 1: The volume of each region $V(A_{k_\tau}(\tau)) < \infty$ for $k = 1, 2, \ldots, K_\tau$, and $\tau \in \Xi$.

Assumption 2: $q(\boldsymbol{\theta}_{k_\tau}, \tau)$ is positive and continuous on $\overline{A}_{k_\tau}(\tau)$, where $\overline{A}_{k_\tau}(\tau)$ is the closure of $A_{k_\tau}(\tau)$ for $k = 1, \ldots, K_\tau$, and $\tau \in \Xi$.

Theorem 4.3.1. *Under Assumptions 1 to 2 and certain ergodic (e.g., time-reversible, invariant, and irreducible) conditions,* $\widehat{1/c}_{VPWK}$ *in (4.11) is a consistent estimator of* $1/c$. *In addition,* $Var(\widehat{1/c}_{VPWK}) < \infty$.

Proof: Under certain ergodic conditions, we have

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\Xi}1\{\tau^{(n)}=t\}\sum_{k_t=1}^{K_t}\frac{1}{\sum_{k_t=1}^{K_t}q(\boldsymbol{\theta}^*_{k_t}|t)V(A_{k_t}(t))}\left[\frac{q(\boldsymbol{\theta}^*_{k_t},t)}{q(\boldsymbol{\theta}^{(n)}_t,t)}1\{\boldsymbol{\theta}^{(n)}_t\in A_{k_t}(t)\}\right]$$
$$\xrightarrow{a.s.}\frac{\sum_{t\in\Xi}\pi(t)}{c}.$$

□

Although (4.11) has most of the desirable properties, we note that the weight assigned to $q(\boldsymbol{\theta}^*_{k_t},t)1\{\boldsymbol{\theta}^{(n)}_t\in A_{k_t}(t)\}/q(\boldsymbol{\theta}^{(n)}_t,t)$ is $\pi(t)/[\sum_{k_t=1}^{K_t}q(\boldsymbol{\theta}^*_{k_t},t)V(A_{k_t}(t))]$, which may not be efficient especially when a vague prior distribution of tree topology is used. To improve (4.11), we multiply this weight by $\hat{\pi}(t|\boldsymbol{y})/\pi(t)$ so that the VPWK estimator is given by

$$\begin{aligned}&\frac{\hat{1}}{c_{VPWK}}\\ =&\frac{\frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\Xi}1\{\tau^{(n)}=t\}\sum_{k_t=1}^{K_t}\frac{\hat{\pi}(t|\boldsymbol{y})}{\sum_{k_t=1}^{K_t}q(\boldsymbol{\theta}^*_{k_t},t)V(A_{k_t}(t))}\left[\frac{q(\boldsymbol{\theta}^*_{k_t},t)}{q(\boldsymbol{\theta}^{(n)}_t,t)}1\{\boldsymbol{\theta}^{(n)}_t\in A_{k_t}(t)\}\right]}{\sum_{t\in\Xi}\hat{\pi}(t|\boldsymbol{y})},\end{aligned}\tag{4.12}$$

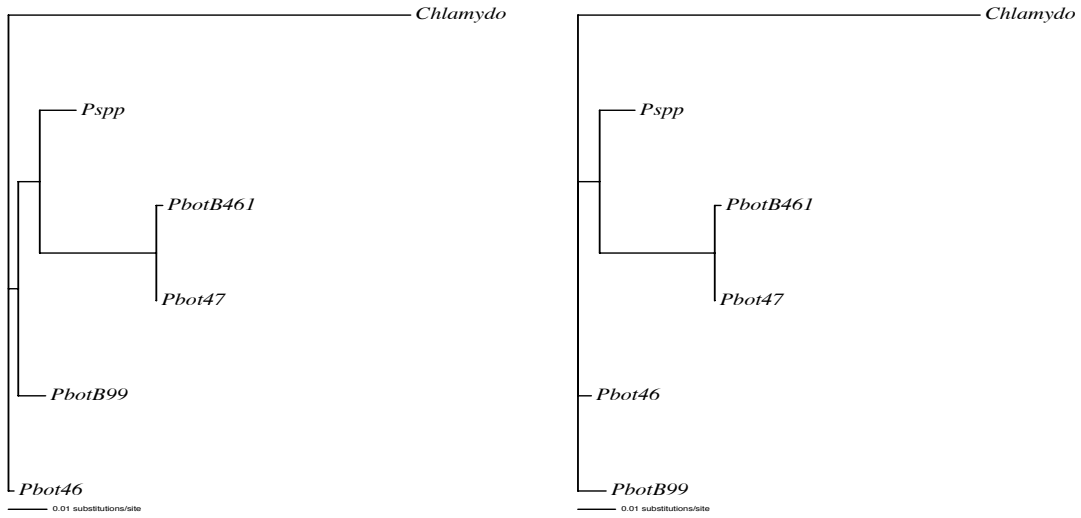
where $\hat{\pi}(t|\boldsymbol{y})=\sum_{n=1}^{N}1\{\tau^{(n)}=t\}/N$. This re-weighting approach allows an MCMC sample drawn from the dominant trees have more weights so that increases the efficiency. Additionally, we note that although both numerator and denominator in (4.12) need to be estimated, it does not require an extra MCMC sample. Instead, using the same MCMC sample in both parts is more efficient as proven in Chen et al. (2008). Also, (4.12) has all desirable properties of that in (4.11).

4.4 6-taxon *rcb*L Data Set

In this section, we use the same *rcb*L data set as in Lewis and Trainor (2012), where the survival under desiccation of green algae was investigated in soil collected from Storrs, Connecticut, in October 1958 and dried for 43 years. Following a series of long-term studies, in 2001 and 2008, Lewis and Trainor (2012) repeated the growth experiment on the same soil, and found the presence of the green alga *Protosiphon botryoides* from liquid preparations of soil dried for 43 years and its absence from cultures prepared with soils dried for 50 years.

The *rcb*L data set includes 6 taxa: *Chlamydopodiumvacuolatum* (Chlamydo), *Protosiphon* FRT2000 (Pspp), and the *Protosiphon* cultures from UTEX (the University of Texas Culture Collection of Algae): UTEX B99 (PbotB99), UTEX 46 (Pbot46), UTEX 47 (Pbot47), and UTEX B461 (PbotB461), each with the sequence length equal to 1376. The number of unrooted tree topologies for 6 taxa is 105. Figure 4.1 shows the majority rule (which equals the most probable tree topology) and second most probable trees. Around 80% of the MCMC sample are from these two dominant trees. For this data set, we consider the general time reversible plus Gamma (GTR+G) model, which involves 18 unknown parameters (9 edge lengths, 3 nucleotide relative frequencies, 5 exchangeabilities, and 1 gamma shape) in $\boldsymbol{\theta}_\tau$. To examine performance of the VPWK estimator in this variable-topology problem, we will compare the VPWK marginal likelihood estimate with those estimated using the variable-topology stepping-stone approach in MrBayes (using the method of Xie et al. (2011)) and Phycas (using the generalized stepping stone method of Fan et al. (2011)). Additionally, the sensitivity analysis of the VPWK based on different chosen Ξ's will be included. To apply the VPWK to this model, we first transform the MCMC sam-

ple of continuous parameters: the log transformation of branch length parameters, the log-ratio transformation of nucleotide frequency and GTR exchangeability parameters, and the log transformation of the gamma shape parameter, so that the new conditional parameter space is R^p. Then, we standardize the transformed MCMC sample and choose the radius as 5 to form the working parameter space. Note that the chosen radius is a round-up value of $\sqrt{\chi^2_{p=18,0.95}}$, which is a suggested value to cover the most of an MCMC sample in Yu et al. (2015). On this basis, we apply the spherical shell approach in Chapter 2 to construct the partitioned subsets of working parameter space for each chosen tree topology.



(a) (b)

Figure 4.1: (a) is the majority rule tree (around 65% MCMC sample points), and (b) is the second most probable tree (around 15% MCMC sample points). Though impossible to see due to very short edge lengths, Pbot46 is sister to (Pspp, (PbotB461, Pbot47)) in tree (b).

The VSS and VPWK estimates are summarized in Table 4.1 based on an MCMC sample with size $N = 60,000$, where 15 out of 105 trees were visited by the MCMC

sample. Table 4.1 provides two estimates based on the two most probable trees and the ten most probable trees. The former only uses 81.26% of MCMC sample while the latter uses 99.62% of MCMC sample. We see the results are robust to the size of Ξ. In both cases, we only need to construct 2 or 10 working parameter spaces and their partition subsets rather than creating a working parameter space for all 105 possible tree topologies, or all 15 tree topologies included in the MCMC sample. We also observe that, based on a single MCMC sample from the joint posterior density, the VPWK can produce comparable results to VSS by MrBayes (Ronquist et al., 2012) and VGSS by Phycas (Lewis et al., 2015), which requires MCMC samples from a series of power posterior distributions. Note that H is chosen as 29 and the size of each MCMC sample from the power posterior distribution is $3,584$ in VSS; while H is chosen as 30 and N from the power posterior distribution is $1,000$ in VGSS.

Table 4.1: Marginal likelihood estimates of VPWK and VSS in log scale

Method	No. of Tree Considered	% of the MCMC sample Used	Estimate
VPWK	2	81.26%	-2682.79
	10	99.62%	-2682.78
VSS by MrBayes	all trees in the MCMC sample	100.00%	-2682.79
VGSS by Phycas	all trees in the MCMC sample	100.00%	-2682.77

4.5 Results and Discussion

In this chapter, we develop a new estimator for a variable tree topology based on the PWK estimator in Chapter 2. In this particular application, we introduce the concept of working parameter space to a tree topology space. The computation time is much lessened by only focusing on frequently sampled trees rather than all trees in the

MCMC sample. Additionally, this new estimator is desirable due to only requiring a single MCMC sample from the posterior density. In addition to theoretical properties such as consistency to the reciprocal of overall marginal likelihood, finite variance, and optimization of the VPWK estimator, it also makes use of the estimated posterior distribution of tree topology as a re-weighted function to increase the efficiency.

In real data analysis, we use the *rcb*L data set in Lewis and Trainor (2012) and fit the data with the general time reversible plus Gamma model, where there are 18 parameters involved in. We show that the VPWK estimates are comparable to the results of VSS and VGSS, both of which require a series of MCMC samples from the power posterior distributions. We also show that the VPWK estimates are robust to the choices of the size of Ξ as long as the selected trees contain reasonable size of an MCMC sample.

At this stage, we are still developing software of the VPWK method for public use. Besides the GTR+G model, more complicated models such as the GTR+G+I model will be included.

Chapter 5

Concluding Remarks and Future Work

5.1 Concluding Remarks

In this research work, a series of Monte Carlo estimators are developed for calculating the marginal likelihood, marginal posterior density, and marginal likelihood for a variable tree topology. All methods are inspired by the idea of partitioning the working parameter space and only require the known posterior kernel function and a single MCMC sample from the posterior distribution.

The partition weighted kernel estimator is essentially an extension of the harmonic mean (Newton and Raftery, 1994) and inflated density ratio (Petris and Tardella, 2003, 2007) approaches, but has more desirable properties including but not limited to consistency, finite variance, optimization. In two simulation studies, we show the PWK has smaller MCSE and RMSE, and approaches proposed for forming the

partition subsets of the working parameter space can insure the homogeneity of the posterior kernel in each subset so that provide stable estimates. In the real data example, we compare the PWK method with Chen's (Chen, 2005b) and Chib's (Chib, 1995) methods. We also show that the PWK produce a smaller estimated Monte Carlo standard error by Overlapping batch statistics (Schmeiser et al., 1990). In the second example, we show the usage of the PWK estimator in constructing the power prior by the empirical Bayes approach using the marginal likelihood as a criterion.

In Chapter 3, we show the adaptive partition weighted estimator can approximate the gold standard approach, the conditional marginal density estimator, when the number of subsets increases. In the first real data example, we empirically show this new method works well in the inequality-constrained analysis of variance. Then, in the second example, we show the usefulness of this method in Bayesian variable selection when the conditions for Savage-Dickey density ratio are satisfied.

In Chapter 4, we develop the variable-topology partition weighted kernel estimator based on the PWK for the overall marginal likelihood calculation. This new method contains all desirable properties as the PWK. In addition, we use the re-weighted function to improve the efficiency of the VPWK. In real data example, the general time reversible plus Gamma model is considered, where there are 18 unknown parameters, for this 6-taxon data. We show this new estimator can produce the comparable results as the VSS and GSS but needs much less computation.

5.2 Future Work

Since our new methods only need minimum assumptions, we are developing the open packages, which can construct the working parameter and its partition subsets by the default setting based on the input of the posterior kernel function and the MCMC sample from the investigators, and produce the marginal likelihood and marginal posterior density estimates.

For Bayesian phylogenetics, we want to make use of the idea of Savage-Dickey density ratio and develop a method to do the phylogenetics model selection by a single MCMC sample from the posterior distribution based on the most complicated model considered. This would be a challenge problem especially for a variable topology.

Additionally, although we have done many sensitivity analyses about the robustness of the methods to the chosen working parameter space and the number of subsets, we are still interested in developing some measures about the convergent speeds of the number of partition subsets.

We leave these three interests as our future works.

Bibliography

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American of Statistical Assocation*, 88: 669–679.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). "BEAST 2: a software platform for Bayesian evolutionary analysis." *PLoS Computational Biology*, 10(4): e1003537.

Chen, M.-H. (1994). "Importance-weighted marginal Bayesian posterior density estimation." *Journal of the American Statistical Association*, 89(427): 818–824.

— (2005a). "Bayesian computation: from posterior densities to Bayes factors, marginal likelihoods, and posterior model probabilities." In Dey, D. and Rao, C. (eds.), *Handbook of Statistics 25, Bayesian Thinking: Modeling and Computation*, 437–457. Elsevier.

— (2005b). "Computing marginal likelihoods from a single MCMC output." *Statistica Neerlandica*, 59: 16–29.

Chen, M.-H., Huang, L., Ibrahim, J. G., and Kim, S. (2008). "Bayesian variable selection and computation for generalized linear models with conjugate priors." *Bayesian Analysis*, 3(3): 585.

Chen, M.-H. and Ibrahim, J. G. (2003). "Conjugate priors for generalized linear models." *Statistica Sinica*, 13(2): 461–476.

Chen, M.-H., Ibrahim, J. G., and Yiannoutsos, C. (1999). "Prior elicitation, variable selection and Bayesian computation for logistic regression models." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 223–242.

Chen, M.-H. and Kim, S. (2008). "The Bayes factor versus other model selection criteria for the selection of constrained models." In Hoijtink, H., Klugkist, I., and Boelen, P. (eds.), *Bayesian Evaluation of Informative Hypotheses*, 155–180. Springer.

Chen, M.-H., Kim, S., et al. (2006). "Discussion of Equi-energy sampler by Kou, Zhou and Wong." *The Annals of Statistics*, 34(4): 1629–1635.

Chen, M.-H. and Shao, Q.-M. (2002). "Partition-Weighted Monte Carlo Estimation." *Annals of the Institute of Statistical Mathematics*, 54: 338–354.

Chib, S. (1995). "Marginal likelihood from the Gibbs output." *Journal of the American Statistical Association*, 90(432): 1313–1321.

Chib, S. and Jeliazkov, I. (2001). "Marginal likelihood from the Metropolis-Hastings output." *Journal of the American Statistical Association*, 96: 270–281.

Dickey, J. M. (1971). "The weighted likelihood ratio, linear hypotheses on normal location parameters." *The Annals of Mathematical Statistics*, 204–223.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). "Bayesian Phylogenetics with BEAUti and the BEAST 1.7." *Molecular Biology And Evolution*, 29: 1969–1973.

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2011). "Choosing among partition models in Bayesian phylogenetics." *Molecular Biology and Evolution*, 28(1): 523–532.

Gelfand, A. E. and Dey, D. K. (1994). "Bayesian Model Choice: Asymptotics and Exact Calculations." *Journal of the Royal Statistical Society, Series B*, 56: 501–514.

Gelfand, A. E., Smith, A. F., and Lee, T.-M. (1992). "Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling." *Journal of the American Statistical Association*, 87(418): 523–532.

Geweke, J. (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica*, 57: 1317–1339.

Goh, G. and Dey, D. K. (2014). "Bayesian model diagnostics using functional Bregman divergence." *Journal of Multivariate Analysis*, 124: 371–383.

Hoijtink, H., Klugkist, I., and Boelen, P. (2008). *Bayesian Evaluation of Informative Hypotheses*. Springer Science & Business Media.

Holder, M. T., Lewis, P. O., Swofford, D. L., and Bryant, D. (2014). "Variable tree topology stepping-stone marginal likelihood estimation." In Chen, M.-H., Kuo, L., and Lewis, P. O. (eds.), *Bayesian Phylogenetics: Methods, Algorithms, and Applications*, 95–112. Chapman & Hall/CRC Mathematical and Computational Biology.

Huelsenbeck, J. P. and Ronquist, F. (2001). "MrBayes: Bayesian inference of phylogeny." *Bioinformatics*, 17: 754–755.

Ibrahim, J. G., Chen, M.-H., and Chu, H. (2012). "Bayesian Methods in Clinical Trials: a Bayesian Analysis of ECOG Trials E1684 and E1690." *BMC Medical Research Methodology*, 12: 170–183.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). "The Power Prior: Theory and Applications." *Statistics in Medicine*.

Jeffreys, H. (1998). *The Theory of Probability*. OUP Oxford.

Lartillot, N. and Philippe, H. (2006). "Computing Bayes Factors Using Thermodynamic Integration." *Systematic Biology*, 55: 195–207.

Lewis, L. A. and Trainor, F. R. (2012). "Survival of Protosiphon botryoides (Chlorophyceae, Chlorophyta) from a Connecticut soil dried for 43 years." *Phycologia*, 51(6): 662–665.

Lewis, P. O., Holder, M. T., and Swofford, D. L. (2015). "Phycas: software for Bayesian phylogenetic analysis." *Systematic Biology*, 525–531.

Lindley, D. V. (1957). "A statistical paradox." *Biometrika*, 187–192.

Lui, D. (2014). *Monte Carlo framework for prostate cancer correction and reconstruction in endorectal multi-parametric MRI*. M.S. Thesis, Department of Systems Design Engineering, University of Waterloo, Canada.

Lui, D., Modhafar, A., Haider, M. A., and Wong, A. (2015). "Monte Carlo-based noise compensation in coil intensity corrected endorectal MRI." *BMC Medical Imaging*, 15(1): 1.

Marin, J. M. and Robert, C. P. (2010). "Importance Sampling Methods for Bayesian Discrimination between Embedded Models." In Chen, M.-H., Dey, D. K., Muller, P., Sun, D., and Ye, K. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, 513–527. New York: Springer.

Nandram, B. and Chen, M.-H. (1996). "Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence." *Journal of Statistical Computation and Simulation*, 54: 129–144.

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian Inference by the Weighted Likelihood Bootstrap." *Journal of the Royal Statistical Society, Series B*, 56: 3–48.

Oh, M.-S. (1999). "Estimation of posterior density functions from a posterior sample." *Computational Statistics & Data Analysis*, 29(4): 411–427.

Parzen, E. (1962). "On estimation of a probability density function and mode." *The Annals of Mathematical Statistics*, 1065–1076.

Petris, G. and Tardella, L. (2003). "A Geometric Approach to Transdimensional Markov Chain Monte Carlo." *The Canadian Journal of Statistics*, 31(4): 469–482.

— (2007). "New Perspectives for Estimating Normalizing Constants via Posterior Simulation." *Technical report, Universita di Roma "La Sapienza."*.

Robert, C. P. and Wraith, D. (2009). "Computational Methods for Bayesian Model Choice." *MaxEnt 2009 Proceedings*.

Ronquist, F. and Huelsenbeck, J. P. (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." *Bioinformatics*, 19: 1572–1574.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). "MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space." *Systematic Biology*, 61(3): 539–542.

Rosenblatt, M. et al. (1956). "Remarks on some nonparametric estimates of a density function." *The Annals of Mathematical Statistics*, 27(3): 832–837.

Schmeiser, B. W., Avramidis, T. N., and Hashem, S. (1990). "Overlapping batch statistics." In *Proceedings of the 22nd Conference on Winter Simulation*, 395–398. IEEE Press.

Stephens, M. and Donnelly, P. (2000). "Inference in molecular population genetics." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 605–635.

Team, R. C. (2014). "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013."

Thomas, A., O´Hara, B., Ligges, U., and Sturtz, S. (2006). "Making BUGS open." *R news*, 6(1): 12–17.

Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2014). "Consistency of marginal likelihood estimation when topology varies." In Chen, M.-H., Kuo, L., and Lewis, P. O. (eds.), *Bayesian Phylogenetics: Methods, Algorithms, and Applications*, 113–128. Chapman & Hall/CRC Mathematical and Computational Biology.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). "Improving Marginal

Likelihood Estimation for Bayesian Phylogenetic Model Selection." *Systematic Biology*, 60(2): 150–160.

Yu, F., Chen, M.-H., Kuo, L., Talbott, H., and Davis, J. S. (2015). "Confident difference criterion: a new Bayesian differentially expressed gene selection algorithm with applications." *BMC Bioinformatics*, 16(1): 245.

Yu, J. Z. and Tanner, M. A. (1999). "An analytical study of several Markov chain Monte Carlo estimators of the marginal likelihood." *Journal of Computational and Graphical Statistics*, 8(4): 839–853.