Bayesian Analysis (2018) **13**, Number 2, pp. 311–333

# A New Monte Carlo Method for Estimating Marginal Likelihoods

Yu-Bo Wang[*], Ming-Hui Chen[†], Lynn Kuo[‡], and Paul O. Lewis[§]

**Abstract.** Evaluating the marginal likelihood in Bayesian analysis is essential for model selection. Estimators based on a single Markov chain Monte Carlo sample from the posterior distribution include 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 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 optimal weights. Simulation studies were 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.

**Keywords:** Bayesian model selection, cure rate model, harmonic mean estimator, inflated density ratio estimator, ordinal probit regression, power prior.

## 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 (product of likelihood and 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) by 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) and Friel and Pettitt (2008), the constrained GHM estimator with the

[*] Department of Statistics, University of Connecticut, Storrs, CT 06269, USA, yu-bo.wang@uconn.edu

[†] Department of Statistics, University of Connecticut, Storrs, CT 06269, USA, ming-hui.chen@uconn.edu

[‡] Department of Statistics, University of Connecticut, Storrs, CT 06269, USA, lynn.kuo@uconn.edu

[§] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA, paul.lewis@uconn.edu

https://doi.org/10.1214/17-BA1049

highest posterior density (HPD) region of Robert and Wraith (2009) and Marin and Robert (2010), and the steppingstone sampling of Xie et al. (2011) and Fan 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 only existing methods that need only 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 (Lartillot and Philippe, 2006; Xie et al., 2011).

While the IDR estimator has better control over the tails of the kernel than the HM estimator, 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 subset. 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 (minimum variance) PWK estimator. Our simulation studies empirically show that the proposed PWK estimator outperforms both the HM and IDR estimators with respect to root mean square error.

The rest of the article 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 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 simulation studies of a bivariate normal case with the normal-inverse-Wishart prior and a mixture of two bivariate normal distributions to compare the performance and computing time of the HM, IDR and PWK estimators. In Section 6, we compare the results and performance of the PWK estimator to the methods by Chib (1995) and Chen (2005) for an ordinal probit regression model. Moreover, we apply the PWK estimator to the determination of the optimal power prior using two Eastern Cooperative Oncology Group (ECOG) clinical trial data sets. Finally, we conclude with a discussion in Section 7. The proofs of all theorems are given in the Supplementary Web Materials (Wang et al., 2017a).

## 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 $\int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta})d\boldsymbol{\theta}$. The integration is often analytically intractable due to complicated kernel structure.

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

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

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

$$\hat{c}_{HM} = \frac{1}{\frac{1}{T}\sum_{t=1}^{T} \frac{1}{L(\boldsymbol{\theta}_t|D)}}, \tag{2}$$

where the prior $\pi(\boldsymbol{\theta}_t)$ can be viewed as the weight assigned to $1/q(\boldsymbol{\theta}_t)$. Although it has the features of simplicity and 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 (1) is replaced by a lighter-tailed density function $f(\boldsymbol{\theta})$ compared to $q(\boldsymbol{\theta})$:

$$\hat{c}_{GHM} = \frac{1}{\frac{1}{T}\sum_{t=1}^{T} \frac{f(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)}}. \tag{3}$$

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

$$q_r(\boldsymbol{\theta}) = \begin{cases} q(\mathbf{0}) & \text{if } ||\boldsymbol{\theta}|| \le r, \\ q(w(\boldsymbol{\theta})) & \text{if } ||\boldsymbol{\theta}|| > r, \end{cases} \tag{4}$$

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

$$\int_{\boldsymbol{\Theta}} q_r(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{||\boldsymbol{\theta}|| \le r} q_r(\boldsymbol{\theta})d\boldsymbol{\theta} + \int_{||\boldsymbol{\theta}|| > r} q_r(\boldsymbol{\theta})d\boldsymbol{\theta} = q(\mathbf{0})b_r + c, \tag{5}$$

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

$$\frac{q(\mathbf{0})b_r + c}{c} = \int_{\boldsymbol{\Theta}} \frac{q_r(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta})}{c} d\boldsymbol{\theta}, \tag{6}$$

and the IDR estimator is given by

$$\hat{c}_{IDR} = \frac{q(\mathbf{0})b_r}{\frac{1}{T}\sum_{t=1}^{T} \frac{q_r(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)} - 1}. \tag{7}$$

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 estimator, mode finding is essential and standardization of an MCMC sample with respect to the mode and the sample covariance matrix is required.

# 3 A New Monte Carlo Estimator

We first modify (1) and (6) by imposing a working parameter space $\Omega \subset \boldsymbol{\Theta}$, where $\Omega = \{\boldsymbol{\theta} : q(\boldsymbol{\theta})$ 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:

$$\frac{\Delta}{c} = \int_{\Omega} \frac{h(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta})}{c} d\boldsymbol{\theta}. \tag{8}$$

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.

## 3.1 General Monte Carlo Estimator

Suppose $\{A_1, \ldots, A_K\}$ forms a partition of the working parameter space $\Omega$, 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:

$$h(\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k 1\{\boldsymbol{\theta} \in A_k\}. \tag{9}$$

So we can evaluate $\Delta$:

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

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 (9), the PWK estimator for $d \equiv 1/c$ is given by

$$\hat{d} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} \frac{w_k}{q(\boldsymbol{\theta}_t)} 1\{\boldsymbol{\theta}_t \in A_k\}}{\sum_{k=1}^{K} w_k V(A_k)}. \tag{10}$$

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 1.** *Under Assumptions 1 to 2 and certain ergodic (e.g., time-reversible, invariant, and irreducible) conditions, $\hat{d}$ in (10) is a consistent estimator of $d$. In addition, $Var(\hat{d}) < \infty$.*

Note that we consider the estimator for $d$ rather than $c$ because we can obtain an unbiased estimator with finite variance for $d = 1/c$.

**Remark 1.** Another property of $\hat{d}$ in (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, $\boldsymbol{\vartheta}_1 = (\theta_1, \ldots, \theta_q)'$ and $\boldsymbol{\vartheta}_2 = (\theta_{q+1}, \ldots, \theta_p)'$. 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, (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 $\{B_1, \ldots, B_K\}$ 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 $V(B_1), \ldots, V(B_K)$ are the corresponding volumes.

## 3.2 The Optimal Monte Carlo Estimation

Our next step is to find the optimal weight $w_k$ in the class of PWK estimators (10), motivated by Chen and Shao (2002).

Assume $\{\boldsymbol{\theta}_t,\ t = 1, \ldots, T\}$ is an MCMC sample from the posterior distribution $\pi(\boldsymbol{\theta}|D)$. Let $w_k^* = w_k/[\sum_{k=1}^K w_k V(A_k)]$ and $\alpha_k = \mathrm{E}[(1/q^2(\boldsymbol{\theta}))1\{\boldsymbol{\theta} \in A_k\}]$. Write $\hat{d}_t = \sum_{k=1}^K \frac{w_k^*}{q(\boldsymbol{\theta}_t)} 1\{\boldsymbol{\theta}_t \in A_k\}$ such that $\hat{d} = \frac{1}{T}\sum_{t=1}^T \hat{d}_t$. Then, we have $\mathrm{Var}(\hat{d}_t) = \sum_{k=1}^K {w_k^*}^2 \alpha_k - 1/c^2$.

**Theorem 2.** *Letting* $w^*_{k,opt} = V(A_k)/\{\alpha_k[\sum_{k=1}^K V^2(A_k)/\alpha_k]\}$ *for* $k = 1, \ldots, K$, *we have* $Var_{w^*_{k,opt}}(\hat{d}_t) = 1/[\sum_{k=1}^K V^2(A_k)/\alpha_k] - 1/c^2$, *and* $Var_{w^*_{k,opt}}(\hat{d}_t) \leq Var_{w^*_k}(\hat{d}_t)$ *for any weight function* $w_k^*(.)$ *defined on each* $A_k$.

**Remark 2.** In practice, it is quite difficult to estimate the second moment $\alpha_k$. A very large sample size is required in order to obtain an accurate estimate of $\alpha_k$. However, the results shown in Theorem 2 shed light on the choices of $A_1, \ldots, A_K$ and $w_k$. First, it is only required that $w_k$ be 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 (10) reduces to

$$\hat{d} = \frac{\frac{1}{T}\sum_{t=1}^T \sum_{k=1}^K \frac{q(\boldsymbol{\theta}_k^*)}{q(\boldsymbol{\theta}_t)} 1\{\boldsymbol{\theta}_t \in A_k\}}{\sum_{k=1}^K q(\boldsymbol{\theta}_k^*) V(A_k)}. \tag{11}$$

**Remark 3.** Following on Remark 1, when a full conditional density $\pi(\boldsymbol{\vartheta}_1|D, \boldsymbol{\vartheta}_2)$ is available, the estimator $\hat{d}$ in (11) reduces further to

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

**Remark 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

$$\sqrt{\widehat{\mathrm{Var}(\hat{\eta})}} = \left\{\left[\frac{B}{T-B}\right]\frac{\sum_{b=1}^{T-B+1}(\hat{\eta}_b - \bar{\eta})^2}{T-B+1}\right\}^{\frac{1}{2}}, \tag{12}$$

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).

### 3.3 Construction of the Partition with Subsets $A_1, A_2, \ldots, A_K$

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 (11) 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 $A_k = \{\boldsymbol{\psi} : r(k-1)/K \le \|\boldsymbol{\psi}\| < rk/K\}$, with $k = 1, \ldots, K$.

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

**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

$$\hat{d} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\tilde{q}(\boldsymbol{\psi}_k^*)}{\tilde{q}(\boldsymbol{\psi}_t)}1\{\boldsymbol{\psi}_t \in A_k\}}{\sum_{k=1}^{K}\tilde{q}(\boldsymbol{\psi}_k^*)V(A_k)}, \tag{13}$$

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

**Remark 5.** When $K$ is sufficiently large, $\tilde{q}(\boldsymbol{\psi}_t)$ in (13) will be roughly constant over $A_k$ and the PWK estimate will be close to optimal. In addition, each kernel value $\tilde{q}(\boldsymbol{\psi}_t)$ is simply the original kernel value $q(\boldsymbol{\theta}_t)$ multiplied by the absolute value of the Jacobian function.

## 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 extended PWK (ePWK) 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 ePWK 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, the extended form of the general PWK in (10) is given by

$$\hat{d}^* = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K^*}\frac{w_k(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)}1\{\boldsymbol{\theta}_t \in A_k\}}{\sum_{k=1}^{K^*}\int_{A_k} w_k(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{14}$$

**Theorem 3.** *Under Assumption 3 and $q(\boldsymbol{\theta}) > 0$, then the ePWK estimator $\hat{d}^*$ in (14) is a consistent estimator of $d$. In addition, if $\int_{A_k}[w_k(\boldsymbol{\theta})^2/q(\boldsymbol{\theta})]d\boldsymbol{\theta} < \infty$ for $k = 1, \dots, K^*$, then $Var(\hat{d}^*) < \infty$.*

**Remark 6.** It is easy to see that $\hat{d}$ in (10) is a special case of $\hat{d}^*$ in (14). When $K^* = K + 1$ and each fixed weight $w_k$ is assigned to an MCMC sample in each region $A_k$ except $w_{K^*} = 0$, $\hat{d}^*$ reduces to $\hat{d}$.

**Remark 7.** The HM estimator is another special case of $\hat{d}^*$ in (14). 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 8.** In addition, $\hat{d}^*$ in (14) includes the IDR estimator as a special case. Let $K^* = 2$, $A_1 = \{\boldsymbol{\theta} : \ ||\boldsymbol{\theta}|| \le r\}$, $w_1(\boldsymbol{\theta}) = q(\mathbf{0}) - q(\boldsymbol{\theta})$, $A_2 = \{\boldsymbol{\theta} : \ ||\boldsymbol{\theta}|| > r\}$, 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.

**Remark 9.** When the posterior kernel $q(.)$ after the transformation is roughly symmetric, the constant weight $w_k$ assigned to partition set $A_k$ constructed using the rings approach discussed in Section 3.3 often leads to an efficient PWK estimator in (10) as empirically demonstrated in Section 5.1 and Section 6. However, when the posterior kernel $q(.)$ is very skewed or multimodal, the constant weight $w_k$ would result in an inefficient PWK estimator. For such a complex case, we can apply the ePWK estimator in (14). The functional weight $w_k(\boldsymbol{\theta})$ can be constructed as follows. We first divide the $k^{th}$ ring $A_k$ into $m_k$ subsets $A_{k1}, \dots, A_{km_k}$ based on $m_k$ slices such that $A_k = \cup_{\ell=1}^{m_k} A_{k\ell}$ and $A_{k1}, \dots, A_{km_k}$ are disjoint, and then assign $w_k(\boldsymbol{\theta}) = q(\boldsymbol{\theta}^*_{k\ell})$ for $\boldsymbol{\theta} \in A_{k\ell}$, where $\boldsymbol{\theta}^*_{k\ell}$ is a representative point in $A_{k\ell}$, for $\ell = 1, \dots, m_k$. In Section 5.2, we apply this version of the ePWK estimator to an example involving a bimodal distribution to examine its empirical performance.

# 5 Simulation Studies

## 5.1 A Bivariate Normal Example

We apply the PWK estimator for computing the normalizing constant of the posterior of the parameters 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 $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_n)'$ be $n$ observations from a bivariate normal distribution,

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

where $\boldsymbol{\mu} \in R^2$ and $\Sigma$ 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 $\Sigma$ 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 $\boldsymbol{\mu}_0$, $\kappa_0$, $\nu_0$, and $\Lambda_0$. Then, the joint posterior kernel is given by

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

with $\gamma = 2^{\nu_0+1} \pi \Gamma_2(\nu_0/2) |\Lambda_0|^{-\nu_0/2} \kappa_0^{-1}$, where $\Gamma_2(\nu_0/2) = \pi^{1/2} \Gamma(\nu_0/2) \Gamma(\nu_0/2 - 1/2)$. Under this setting, the analytical form of the normalizing constant is available as follows:

$$c = \frac{1}{\pi^n} \frac{\Gamma_2(\nu_n/2)}{\Gamma_2(\nu_0/2)} \frac{|\Lambda_0|^{\nu_0/2}}{|\Lambda_n|^{\nu_n/2}} \left( \frac{\kappa_0}{\kappa_n} \right), \tag{15}$$

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 $\nu_n = \nu_0 + n$. 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 generated 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 $\bar{\boldsymbol{y}}$ was $(-0.029, 0.040)'$, and the sample variance–covariance matrix $S$ was $\begin{pmatrix} 201.987 & 143.330 \\ 143.330 & 192.365 \end{pmatrix}$. Using (15), 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 $\Sigma$ was needed. Here, we used 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 standardized each transformed MCMC sample from its transformed sample mean and standard deviation. In the new parameter space, we constructed the working parameter space and its partition by choosing $r = 1.5$, 2, or 2.5 and $K = 10$, 20, or 100. After selecting a representative point in each spherical shell, we estimated $d = 1/c$ using (13). We compare our method to the HM and IDR methods based on 1,000 independent MCMC samples with $T = 1,000$ or $T = 10,000$ in Table 1. Let $\hat{d}_\ell$ be the estimate of $d$ based on the $\ell^{th}$ MCMC sample for $\ell = 1, 2, \ldots, 1,000$. Then, the simulation estimate (Mean), the MC standard error (MCSE), and the root mean square error (RMSE) of the estimates in log scale are defined as $\widehat{\log c} = \frac{1}{1000}\sum_{\ell=1}^{1000}(-\log \hat{d}_\ell)$, $\{\frac{1}{1000-1}\sum_{\ell=1}^{1000}(-\log \hat{d}_\ell - \widehat{\log c})^2\}^{1/2}$, and $\{\frac{1}{1000}\sum_{\ell=1}^{1000}(-\log \hat{d}_\ell - \log c)^2\}^{1/2}$, respectively.

| | | | | $\log c = -507.2776$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | $T$=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 | -507.262 | 0.053 | 0.055 | -507.264 | 0.016 | 0.021 | 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 | 0.052 | 0.054 | -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 1: Simulation results for the bivariate normal case.

Table 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 1, we see that (i) PWK has the best performance with much smaller MCSE and RMSE than HM and IDR under both $T = 1,000$ and $T = 10,000$; (ii) when $T$ increases, the MCSE and the RMSE of the PWK estimator become smaller under all choices of $r$ and $K$; (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 and the RMSE of the PWK estimator are very similar for all choices of $r$ and $K$ 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.

In this example, we also examine the performance of ePWK by adding a subset $A_{K+1} = \boldsymbol{\Theta} \cap \Omega^c = \{\boldsymbol{\theta} : ||\boldsymbol{\theta}|| > r\}$ such that $K^* = K + 1$ and $\cup_{k=1}^{K^*} A_k = \boldsymbol{\Theta}$. We further specify $w_{K+1}(\boldsymbol{\theta}) = q(\boldsymbol{\theta}^*_{K+1})g(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in A_{K+1}$, where $\boldsymbol{\theta}^*_{K+1}$ is a point on the boundary of $A_{K+1}$ and

$$g(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{2.5}} \exp\left\{-\frac{\boldsymbol{\theta}'\boldsymbol{\theta}}{2}\right\} \Big/ \left[1 - P(\chi^2_{(5)} \leq r^2)\right].$$

Under this specification, we have $\int_{A_{K+1}} w_{K+1}(\boldsymbol{\theta})d\boldsymbol{\theta} = q(\boldsymbol{\theta}^*_{K+1})$. Holding the other subsets $A_1, \ldots, A_K$ and their corresponding weights the same as for PWK, the resulting values of MCSE and RMSE by ePWK are 0.06332 and 0.06579 when $T = 1,000$, $K = 100$, and $r = 1.5$; 0.05167 and 0.05420 when $r = 2.0$; and 0.05500 and 0.05772 when $r = 2.5$. Compared to the results of PWK (0.06375 and 0.06621 when $r = 1.5$; 0.05168 and 0.05420 when $r = 2.0$; and 0.05499 and 0.05772 when $r = 2.5$), ePWK performs very similarly to PWK, which is expected since the posterior kernel has light tails and very low values on $A_{K+1}$.

To evaluate the effect of a vague prior on the precision of the PWK estimator, we extend our simulation study by considering different values of hyperparameters $\kappa_0$ and $\nu_0$. Note that the value of $\log c$ in Table 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 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, we see that the MCSE values under these different values of $(\kappa_0, \nu_0)$ are almost the same while the RMSE values are comparable except the last one with $(\kappa_0, \nu_0) = (1, 10)$, in which the RMSE values are slightly larger.

| | | | $T$=1,000 | | | $T$=10,000 | | |
|---|---|---|---|---|---|---|---|---|
| $\kappa_0$ | $\nu_0$ | $\log c$ | **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 |

Table 2: Simulation results of PWK estimators for different hyperparameters $\kappa_0$ and $\nu_0$.

## 5.2 A Mixture of Two Bivariate Normal Distributions Example

To evaluate the performance of ePWK, we consider the two-dimensional normal mixture in Chen et al. (2006)

$$\pi(\boldsymbol{\mu}) = \sum_{j=1}^{2} \frac{1}{2}\Big[\frac{1}{2\pi}|\Sigma_j|^{-1/2} \exp\Big\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_{0j})'\Sigma_j^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_{0j})\Big\}\Big], \tag{16}$$

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 1(a) is a scatter plot of a random sample with $T = 10,000$ generated from (16). Based on the random sample, we apply ePWK to estimate the normalizing constant in (16), which is known to be 1. Due to

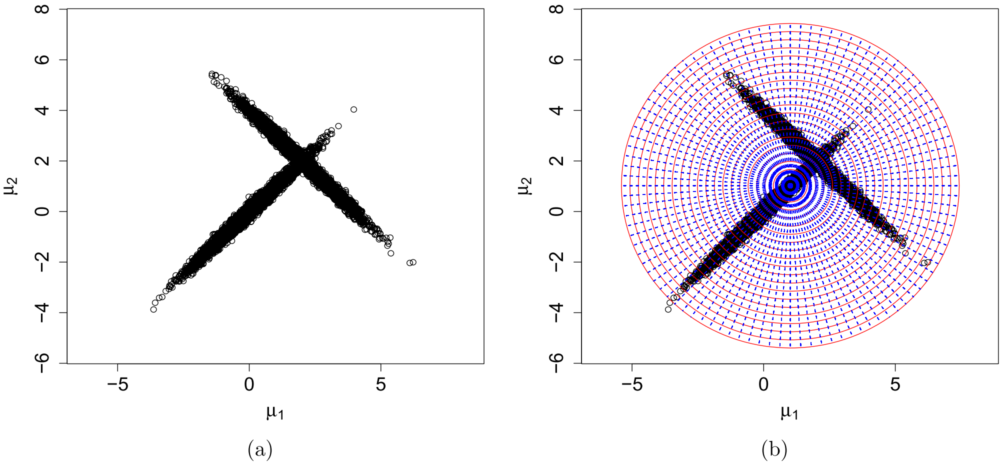


Figure 1: Forming the working parameter space and its partition for a mixture normal distribution with means (0,0) and (2,2).

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, following Remark 9, we additionally slice the existing partition rings by dividing them equally along the angle from 0 to 360 degrees as shown by the dashed lines in Figure 1(b), where the center of the 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. We note that this version of ePWK is the same as PWK except for additional slicing over the partition rings.

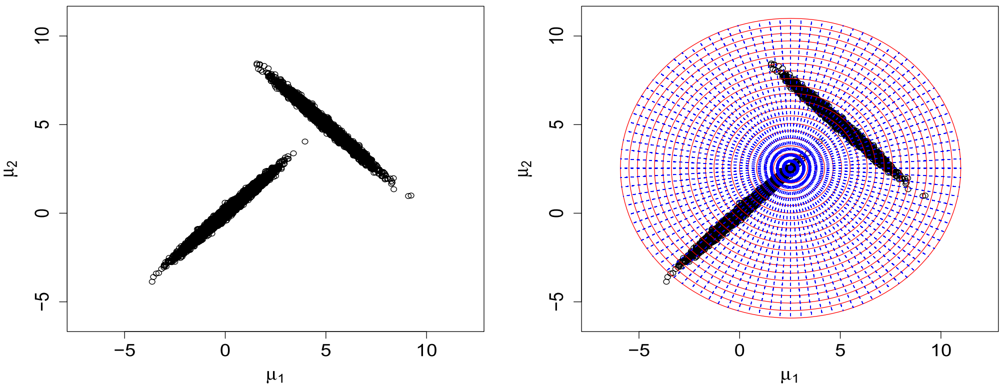
Table 3 shows the results of HM, IDR, and ePWK estimators based on 1,000 independent random samples with $T = 1,000$ or $T = 10,000$ from (16). For ePWK, we consider different values of $K$ (the number of rings) with the same $m_k = m$ (the number of slices) for $k = 1, \ldots, K$ and $r$ (75%, 90%, or 95% $\times \max_{1 \le t \le T} ||\boldsymbol{\mu}_t - \hat{\boldsymbol{\mu}}||$). We use the same values of $r$ for both IDR and ePWK. From Table 3, we see that (i) the RMSE values of the ePWK are considerably smaller than those of HM and IDR; (ii) the performance of ePWK improves when the sample size ($T$) or the number of rings ($K$) increases; and (iii) ePWK takes slightly longer computing time than HM and IDR.

Next, we consider a more challenging case, where $\boldsymbol{\mu}_{02}$ is replaced by $(5, 5)'$ so that the two modes are much further away from each other. Figure 2(a) is a scatter plot of a random sample with $T = 10,000$ and Figure 2(b) shows the partition subsets of the chosen working parameter space.

Table 4 summarizes the simulation results with the same simulation setting as before. We see that ePWK outperforms both HM and IDR under this more challenging case. As expected, the RMSE values in Table 4 are larger than those in Table 3 for all three methods. However, the RMSE values of the ePWK estimator are still quite small when $K$ and $T$ are reasonably large.

| $\log c = 0$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $T$=1,000 | | | $T$=10,000 | | | **Time (sec.)** |
| | $K$ | $m$ | $r$ | **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 |
| ePWK | 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 |

Table 3: Simulation results for the mixture normal with means equal to (0,0) and (2,2).



(a)

(b)

Figure 2: Forming the working parameter space and its partition for a mixture normal distribution with means (0,0) and (5,5).

# 6 Application of the PWK to Real Data Examples

## 6.1 The Ordinal Probit Regression Model

In the first example, we apply the PWK method to computing the marginal likelihood under the ordinal probit regression model. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ denote the vector of observed ordinal responses, each is coded as one value from $0, 1, \ldots, J-1$, $\boldsymbol{X}$ denote the $n \times p$ covariate matrix with the $i^{th}$ row equal to the covariate of the $i^{th}$ subject $x_i'$, and $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)'$ denote the vector of latent random variables. We consider the following hierarchical model as in Albert and Chib (1993) such that

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

| $\log c = 0$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $T$=1,000 | | | $T$=10,000 | | | **Time (sec.)** |
| | $K$ | $m$ | $r$ | **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 |
| ePWK | 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 |

Table 4: Simulation results for the mixture normal with means equal to (0,0) and (5,5).

and

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

where $j = 0, 1, \ldots, J-1$, $\boldsymbol{\beta}$ 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 in Chen (2005)

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

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 $\boldsymbol{\beta}$, $\sigma^2$, 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 (2005). 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 $\boldsymbol{\beta}|\sigma^2 \sim N(0, 10\sigma^2 I_4)$ and $\sigma^2 \sim IG(a_0 = 1, b_0 = 0.1)$. The density function of an inverse gamma distribution $IG(a_0, b_0)$ is 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 (2005) 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 PWK to these two methods under the same MCMC sample sizes $T = 2,500$, or $5,000$ as in Chen (2005), except that Chib's method doubles them.

$$r = 0.75\sqrt{\chi^2_{5,0.95}}$$

| | 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 |

Table 5: The PWK estimates of the marginal likelihood for the prostate cancer data.

For the PWK, we apply a log transformation for $\sigma^2$. 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 the square-root of the $95^{th}$ percentile of the Chi-square distribution with $p = \dim(\boldsymbol{\theta}) = 5$ degrees of freedom, which is derived by computing the norm of $p$ independent standard normal distributions as in Yu et al. (2015). Table 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 (12) with $T/B = 10$. We note that we use the same MCMC sample sizes as in Chen (2005). The results show the PWK estimators are relatively robust to the choice of the radius $r$ and the number $K$ of partition subsets.

From Table 1 of Chen (2005), the estimates of $\log c$ and eMCSE's are $-758.71$ and 0.038 based on Chen's method and $-758.67$ and 0.037 based on Chib's method for $T = 2,500$; and $-758.71$ and 0.024 based on Chen's method and $-758.70$ and 0.023 based on Chib's method for $T = 5,000$. We see from Table 5 that the PWK estimates of $\log c$ are similar to those under both Chen's and Chib's methods but with smaller eMCSE's under the MCMC samples with $T$ = 2,500 and 5,000, respectively. For instance, the PWK estimates of $\log c$ and the corresponding eMCSE's are $-758.70$ and 0.020 for $T = 2,500$ and $-758.70$ and 0.016 for $T = 5,000$ when $r = \sqrt{\chi^2_{5,0.95}}$ and $K = 100$. Thus, the PWK yields a slightly more precise estimate of $\log c$ than the other two methods.

## 6.2 Analysis of ECOG Data

In this subsection, we apply the PWK estimator to the problem of determining the power prior based on 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 that 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 extent to which the historical likelihood is incorporated into the initial prior. Our objective is to find the optimal $a_0$ which maximizes the marginal likelihood for the current data. Ibrahim et al. (2015) point out the difficulty of finding this solution except for normal linear regression models. Therefore, they resort 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 relapse-free survival time for the $i^{th}$ patient, $\nu_i$ denote the censoring status, which is equal to 1 if $y_i$ is a failure time and to 0 if it is 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

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D) = \prod_{i=1}^{n} \Big\{ \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) f(y_i|\boldsymbol{\lambda}) \Big\}^{\nu_i} \exp\{-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) F(y_i|\boldsymbol{\lambda})\}, \tag{17}$$

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 (17), 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 $s_{j-1} \le y < s_j$, $s_0 = 0 < s_1 < s_2 < \ldots < s_5 = \infty$, 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, $\nu_{0i}$ 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 $\text{Exp}(\lambda_0 = 1/100)$ ($\lambda_0$: 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 $\pi_0$ multiplied by the historical likelihood function with power $a_0$ as follows:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0, a_0) \propto \Big[ \prod_{i=1}^{n_0} \Big\{ \exp(\boldsymbol{x}_{0i}'\boldsymbol{\beta}) f(y_{0i}|\boldsymbol{\lambda}) \Big\}^{\nu_{0i}} \exp\{-\exp(\boldsymbol{x}_{0i}'\boldsymbol{\beta}) F(y_{0i}|\boldsymbol{\lambda})\} \Big]^{a_0} \pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda}), \tag{18}$$

where $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0, a_0)$ is called the power prior and $0 \le a_0 \le 1$. 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 $a_0 \ne 0$, the power prior is equal to the right-hand side kernel function in (18) 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 (17) and the power prior in (18), the posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ given $(D, D_0, a_0)$ will be

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D, D_0, a_0) \propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_0, a_0). \tag{19}$$

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 $0 \le a_0 \le 1$. 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 $\lambda_j$ is needed. After the standardization of the transformed MCMC sample, we choose the maximum radius $r = \sqrt{\chi^2_{7,0.95}}$ due to $p = 7$, and the number of spherical shells $K = 100$. By (13) and (12), we can obtain the marginal likelihood estimate and its eMCSE for each chosen $a_0$. We summarize the results in Table 6. Table 6 also includes the PWK estimates under $r = 0.75\sqrt{\chi^2_{5,0.95}}$, $1.25\sqrt{\chi^2_{5,0.95}}$ and $K = 10$, 20 to investigate the robustness of the PWK method.

Note the marginal likelihood function $c$ can be shown to be continuous in $a_0$. Therefore, from Table 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. We also observe that the results are quite robust to the different values of $r$ and $K$, and all point out that the best choice of $a_0$ is between 0.5 and 0.6.

# 7 Discussion

The marginal likelihood is often analytically intractable due to a complicated kernel structure. Nevertheless, an MCMC sample from the posterior distribution is readily available from Bayesian computing software. Additionally, the likelihood values evaluated at the MCMC sample are output in a file. Consequently, we can produce kernel values easily using the output and the prior function. In this paper, we propose a new algorithm, PWK, for estimating the marginal likelihood based on this single MCMC sample and its corresponding kernel values. Unlike some existing algorithms requiring knowledge of the structure of the kernel, we only need to know the kernel values evaluated at the MCMC sample. Therefore, our algorithm can be applied to Bayesian model selection, assessing the sensitivity of conclusions to the prior distribution, and Bayes hypothesis tests. We implement our methodology using the R programming language (R Core Team, 2015). The R codes along with README files are available as Online Supplementary Materials (Wang et al., 2017b).

We extend PWK to handle the parameter space with the full support (ePWK) and we show that HM and IDR are special cases of ePWK. We conduct a simulation study from a bivariate normal distribution with 5 parameters in a Bayesian conjugate

$r = 0.75\sqrt{\chi^2_{7,0.95}}$

| | $K$=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 |

Table 6: PWK estimates for marginal likelihood with different power priors under different choices of $r$ and $K$.

prior inference problem to compare our estimator to HM and IDR; our results show that PWK has the smallest empirical MCSE 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. We conduct another simulation study for a mixture of two bivariate normal distributions to illustrate the ePWK estimator, which is obtained by additionally slicing the partition rings in the partition step of the PWK method. We show that the ePWK method reduces the MCSE and RMSE by a great deal when compared to the HM and IDR methods at the cost of slightly more computation time.

In example analyses of real data, we first consider an ordinal probit regression model, and compare our method to that in Chib (1995) and Chen (2005) with the same MCMC sample size for Chen's method (Chib's method requires twice this sample size). We find the three methods produce comparable estimates for the marginal likelihood and the PWK method produces the smallest eMCSE. 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). We obtain similar results, except that the PWK approach indicates more borrowing of the historical data.

In unimodal problems, we suggest using the square root of the $95^{th}$ percentile in a Chi-square distribution with $p$ degrees of freedom as a guide to choosing a value for the radius $r$ for constructing the working parameter space of the standardized MCMC sample. This is because, after standardizing the MCMC sample, the marginal distribution of each parameter is approximately standard normal. Although the results are quite robust to the choices of $r$ as shown in simulation and case studies, using the Chi-square distribution for guidance ensures that we make use of most of the MCMC sample and avoid the region with posterior density close to 0. For multimodal problems, we suggest using $95\% \times \max_{1\leq t\leq T} ||\boldsymbol{\mu}_t - \hat{\boldsymbol{\mu}}||$ as a guide value for constructing the working parameter space of the transformed MCMC sample. Since this approach may result in many partition subsets with extremely small posterior density in the working parameter space, we can use the spherical rings approach as demonstrated in Section 5.2 to obtain the homogeneity of the MCMC sample in each subset. This new partition approach can also be extended to a $p$-dimensional problem ($p > 2$) by introducing another $p - 2$ angular coordinates as in Lehnen and Wesenberg (2003) and slicing them as in Section 5.2.

## Supplementary Material

Supplementary Web Materials for "A New Monte Carlo Method for Estimating Marginal Likelihoods" (DOI: 10.1214/17-BA1049SUPPA; .pdf).

Online Supplementary Materials for "A New Monte Carlo Method for Estimating Marginal Likelihoods" (DOI: 10.1214/17-BA1049SUPPB; .zip).

# References

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. MR1224394. 323

Chen, M.-H. (2005). "Computing Marginal Likelihoods from a Single MCMC Output." *Statistica Neerlandica*, 59: 16–29. 312, 324, 325, 326, 330

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). "A New Bayesian Model for Survival Data With a Surviving Fraction." *Journal of the American Statistical Association*, 94: 909–919. MR1723307. doi: `https://doi.org/10.2307/2670006`. 327

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. MR2283711. doi: `https://doi.org/10.1214/009053606000000515`. 321

Chen, M.-H. and Shao, Q.-M. (2002). "Partition-Weighted Monte Carlo Estimation." *Annals of the Institute of Statistical Mathematics*, 54: 338–354. MR1910177. doi: `https://doi.org/10.1023/A:1022426103047`. 316

Chib, S. (1995). "Marginal Likelihood from the Gibbs Output." *Journal of the American Statistical Association*, 90: 1313–1321. MR1379473. 311, 312, 325, 330

Chib, S. and Jeliazkov, I. (2001). "Marginal Likelihood from the Metropolis–Hastings Output." *Journal of the American Statistical Association*, 96: 270–281. 311

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. 312

Friel, N. and Pettitt, A. N. (2008). "Marginal Likelihood Estimation via Power Posteriors." *Journal of the Royal Statistical Society, Series B*, 70: 589–607. 311

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. MR1278223. 311, 313

Geweke, J. (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica*, 57: 1317–1339. 311

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. 326, 327, 328, 330

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). "The Power Prior: Theory and Applications." *Statistics in Medicine*. 326

Lartillot, N. and Philippe, H. (2006). "Computing Bayes Factors Using Thermodynamic Integration." *Systematic Biology*, 55: 195–207. 311, 312

Lehnen, A. and Wesenberg, G. E. (2003). “The Sphere Game.” *The AMATYC Review 25*. URL http://faculty.madisoncollege.edu/alehnen/sphere/hypers.htm 330

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. 312, 314

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. MR1700909. doi: https://doi.org/10.1080/00949659608811724. 324

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. 311, 313

Petris, G. and Tardella, L. (2003). “A Geometric Approach to Transdimensional Markov Chain Monte Carlo.” *The Canadian Journal of Statistics*, 31(4): 469–482. 311, 313

Petris, G. and Tardella, L. (2007). “New Perspectives for Estimating Normalizing Constants via Posterior Simulation.” *Technical report, Universitá di Roma “La Sapienza”*. 311, 314

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 328

Robert, C. P. and Wraith, D. (2009). “Computational Methods for Bayesian Model Choice.” In *MaxEnt 2009 Proceedings*. 312, 314

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. 316

Wang, Y.-B., Chen, M.-H., Kuo, L., and Lewis, P. O. (2017a). “Supplementary Web Materials for “A New Monte Carlo Method for Estimating Marginal Likelihoods”.” *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1049SUPPA. 312

Wang, Y.-B., Chen, M.-H., Kuo, L., and Lewis, P. O. (2017b). “Online Supplementary Materials for “A New Monte Carlo Method for Estimating Marginal Likelihoods”.” *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1049SUPPB. 328

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. 312, 313

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. 325

## Acknowledgments

We would like to thank the editor-in-chief, an editor, and an associate editor for their very helpful comments and suggestions, which have led to a much improved version of the paper. This material is based upon work supported by the National Science Foundation under grant number DEB-1354146 to POL, MHC, LK, and Louise Lewis. MHC's research was also partially supported by NIH grant number GM 70335 and P01 CA142538.