Statistics and Computing manuscript No.
(will be inserted by the editor)

Properties of the bridge sampler with a focus on splitting the MCMC sample

Jackie S. T. Wong [a,*] · **Jonathan J. Forster** [b] · **Peter W. F. Smith** [c]

Received: date / Accepted: date

E-mail addresses:
jw19203@essex.ac.uk (J.S.T. Wong)
J.J.Forster@warwick.ac.uk (J.J. Forster)
P.W.Smith@soton.ac.uk (P.W.F. Smith)

Abstract Computation of normalizing constants is a fundamental mathematical problem in various disciplines, particularly in Bayesian model selection problems. A sampling based technique known as bridge sampling (Meng and Wong 1996) has been found to produce accurate estimates of normalizing constants and is shown to possess good asymptotic properties. For small to moderate sample sizes (as in situations with limited computational resources), we demonstrate that the (optimal) bridge sampler produces biased estimates. Specifically, when one density (we denote as p_2) is constructed to be close to the target density (we denote as p_1) using method of moments, our simulation based results indicate that the correlation induced bias through the moments-matching procedure is non-negligible. More crucially, the bias amplifies as the dimensionality of the problem increases. Thus, a series of theoretical as well as empirical investigations is carried out to identify the nature and origin of the bias. We then examine the effect of sample size allocation on the accuracy of bridge sampling estimates and discovered that one possibility of reducing both the bias and standard error with little increase in computational effort is by drawing extra samples from the moments-matched density p_2 (which we assume easy to sample from), provided that the evaluation of p_1 is not too expensive. We proceed to show how the simple adaptive approach we termed "splitting" manages to alleviate the correlation induced bias at the expense of a higher standard error, irrespective of the dimensionality involved. We also slightly modified the strategy suggested by Wang et al. (2019) to address the issue of the increase of standard error due to splitting, which is later generalized to further improve the efficiency. We conclude the paper by offering our insights of the application of a combination of these adaptive methods to improve the accuracy of bridge sampling estimates in Bayesian applications (where posterior samples are typically expensive to generate) based on the preceding investigations, with an application to a practical example.

Keywords Normalizing constants · Bridge sampling · Method of moments · Correlation induced bias · Bayesian applications.

1 Introduction

Estimating normalizing constants is a well-known problem, solutions of which often revolve around developing new or modifying current numerical computational algorithms to circumvent this issue that hinders subsequent statistical/scientific inferences. To give a few examples: likelihood inference in the presence of missing data where computation of the observed-data likelihood is essentially the problem of estimating the normalizing constant of the complete-data likelihood, a rather common application in genetic linkage analysis (see Irwin et al. 1994, Augustine Kong et al. 1994, Jensen and Kong 1999 etc.); computation of free energy differences (e.g. Bennett 1976, Frenkel 1986, Neal 1993 etc.

* Corresponding author
[a] Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK
[b] Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[c] Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, Highfield, Southampton, SO17 1BJ, UK

); estimation of marginal likelihoods and Bayes factors within the Bayesian framework (e.g. Kass and Raftery 1995, Carlin and Louis 2000, Sinharay and Stern 2005 etc.).

In Bayesian computations, evaluation of the normalizing constant, known as the marginal likelihood, can initially be avoided during the parameter estimation stage using Markov Chain Monte Carlo (MCMC) sampling methods since it is not the parameter of interest. However, this very quantity plays a central role in Bayesian model comparison and model averaging. To be exact, denoting θ as the parameter and M as the model parameter, Bayes theorem stipulates that the posterior distribution given data, X, is

$$f_M(\theta|X) = \frac{f_M(X|\theta)f_M(\theta)}{f_M(X)}, \tag{1}$$

where $f_M(X|\theta)$ is the model likelihood and $f_M(\theta)$ is the prior distribution of θ. It is clear that the numerator of Equation (1), $f_M(X) = \int f_M(X|\theta)f_M(\theta)\mathrm{d}\theta$ (the marginal likelihood), does not depend on θ and is regarded as the normalizing constant of the posterior distribution. The Bayes factor, defined as the ratio of marginal likelihoods of two competing models, is the key quantity in Bayesian model selection because it encodes the evidence of model preference given by the data (Kass and Raftery 1995). Computing Bayes factors is extremely challenging, and is the primary reason why Bayesian inference was not popular (since exact posterior computations are prohibited) until the discovery of MCMC methods. Therefore, it is crucial to be able to estimate the aforementioned quantities to carry out a fully Bayesian computational approach. Typically, posterior samples are very expensive to generate (mostly using MCMC methods) given the actual computational constraints. Occasionally, the evaluation of likelihoods can also be rather costly (e.g. in the presence of latent data/parameters). Hence, the aim in this context is often to maximize the statistical efficiency of the estimates produced, given a fixed number of posterior samples.

A range of possible computational techniques are available for computing marginal likelihoods/Bayes factors; see Carlin and Louis (2000) for a comprehensive review. Simulation based (Monte Carlo) approximation is commonly used by most statisticians due to its general applicability and their knowledge of sampling based inference. Some examples include the importance sampling method (e.g. Geweke 1989), Chib's method (Carlin and Chib 1995), harmonic mean estimator (Newton and Raftery 1994), generalized harmonic mean estimator (Gelfand and Dey 1994), reversible jump MCMC method (Green 1995), path sampling (Gelman and Meng 1998) etc. In this paper, we focus on the bridge sampling method, which is a technique originally developed by Bennett (1976) in the specific context of free energy estimation. This technique was later refined and formulated by Meng and Wong (1996) into a more general setting involving estimation of ratio of two normalizing constants, which can also be constructed to estimate a single normalizing constant when one of the densities is normalized. The bridge sampling technique has been widely applied in various research areas including: missing data analysis (Jensen and Kong 1999, Lee et al. 2003), factor analysis (Meng and Schilling 1996, Lopes and West 2004), statistical regressions (Mira and Nicholls 2004, Bartolucci et al. 2006, Overstall and Forster 2010 etc.), Markov mixture models (Frühwirth-Schnatter 2004) etc. More recent applications include Guy et al. (2013), Tan (2013), Wong et al. (2018), Gronau et al. (2017a). The package "bridgesampling" in R (Gronau et al. 2017b) can now be used to implement the bridge sampling estimation conveniently. We also provide our version of R code (see Appendix) which focuses on estimating marginal likelihoods using the bridge sampling technique, with various algorithms to increase efficiency (to be introduced in the paper), given a set of posterior samples.

Bridge sampling estimates are empirically found to be rather accurate (e.g. Sinharay and Stern 2005, Frühwirth-Schnatter 2004), leading to its popularity. While known to be asymptotically unbiased, bridge sampling technique produces biased estimates in practical usage for small to moderate sample sizes. Meng and Schilling (1996) carried out an empirical analysis of the optimal bridge sampling estimator and illustrated that the estimator yields positive bias that worsens with increasing distance between the two distributions. The second type of bias arises when the approximation density is determined from the posterior samples using the method of moments, resulting in a systematic underestimation of the normalizing constant due to the correlation induced through the moments-matching procedure, as demonstrated by Overstall and Forster (2010). Wong (2017) also showed how the issue of underestimation worsens in high-dimensional problems. Additionally, Wang et al. (2019) pointed out a similar issue of using the sample moments of the U-warped distribution to construct a mixture Gaussian approximation resulting in biased bridge sampling estimates. They proposed a similar approach as Overstall and Forster (2010) to eliminate the bias, and also suggested a modification to avoid an increase in the estimates' standard errors. In this paper, we perform a bias analysis on the bridge sampling estimator by breaking it down into smaller

components, providing some theoretical insights of the origin of the two types of biases. We then focus on the correlation induced bias. The effect of sample size allocation on bridge sampling estimates is examined, which lead to reduced bias and standard error when applied appropriately in certain scenarios. Several alternatives capable of improving bridge sampling estimates (either by mitigating the bias or reducing the standard error) are then presented and explored in detail. A series of simulation studies is conducted to ascertain our conjecture, putting emphasis on not just the relative mean square error as an overall measure of efficiency, but also on a detailed analysis of the empirical bias and standard error separately.

The rest of the paper is structured as follows. First, we introduce the bridge sampling estimator and describe some examples to showcase the empirical bias of bridge sampling estimates. We proceed to identify the source of the bias by breaking down the bridge sampling estimator into smaller components for ease of explanation (Section 2). Secondly, we examine the effect of different allocation of sample sizes on the behaviour of bridge sampling estimates (Section 3). We then describe and extend the idea of splitting, which alleviates the correlation induced bias, but at the same time result in an increased standard error (Section 4). Our investigation also reveals the optimal way of applying the partitioning based on various situations. The approach by Wang et al. (2019) to avoid an increase in standard error due to splitting is modified and extended (Sections 5 and 6). Finally, some matters of consideration during the practical implementation of the bridge sampling method in Bayesian computations are presented on the basis of the preceding investigations, and concluded with an illustrative example (Section 7).

2 The Bridge Sampling Estimator

Suppose that $p_i(\theta)$ $(i = 1, 2)$ are two densities with parameter spaces $\Theta_i \subset \mathbb{R}^d$ respectively, where d is the dimension of θ, and are known up to a normalizing constant, i.e. $p_i(\theta) = \frac{q_i(\theta)}{c_i}$, with c_i as the corresponding normalizing constants of the unnormalized densities, $q_i(\theta)$. The fundamental usage of bridge sampling is based on the following key identity,

$$r \equiv \frac{c_1}{c_2} = \frac{\mathbb{E}_2[q_1(\theta)\omega(\theta)]}{\mathbb{E}_1[q_2(\theta)\omega(\theta)]}, \tag{2}$$

where $\omega(\theta)$ is the so called bridge function (defined on the common support $\Theta_1 \cap \Theta_2$) satisfying $0 < \left|\int_{\Theta_1\cap\Theta_2} p_1(\theta)p_2(\theta)\omega(\theta)\mathrm{d}\theta\right| < \infty$, so that the ratio in Equation (2) is well defined (Meng and Wong 1996). According to Meng and Wong (1996), the existence of $\omega()$ for Equation (2) (and hence the bridge sampler to be valid) is ensured as long as the two densities have non-trivial common support (which is almost always satisfied in practice). Given that the above condition is satisfied, the Monte Carlo estimate of r is simply

$$\hat{r} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} q_1(\theta_2^i)\omega(\theta_2^i)}{\frac{1}{N_1}\sum_{j=1}^{N_1} q_2(\theta_1^j)\omega(\theta_1^j)}, \tag{3}$$

where $\{\theta_1^{1:N_1}\} \equiv \{\theta_1^1, \ldots, \theta_1^{N_1}\}$ and $\{\theta_2^{1:N_2}\} \equiv \{\theta_2^1, \ldots, \theta_2^{N_2}\}$ are sets of random (possibly dependent) realizations from $p_1(\theta)$ and $p_2(\theta)$ respectively. Under certain regularity conditions, $\hat{r}$ converges asymptotically to the true value, r (i.e. the sample averages in Equation (3) converge to their respective population averages). It is also worth noting here that the bridge sampling estimate is essentially a maximum likelihood estimate under the interesting semi-parametric formulation by Kong et al. (2003).

The choice of the bridge function, $\omega()$, is arbitrary, but defines the resulting estimator formed. For instance, choosing $\omega_I(\theta) = 1/q_2(\theta)$ leads to the well known importance sampling estimator,

$$\hat{r}_I = \frac{1}{N_2}\sum_{i=1}^{N_2} \frac{q_1(\theta_2^i)}{q_2(\theta_2^i)}. \tag{4}$$

While choosing $\omega_{RI}(\theta) = 1/q_1(\theta)$ leads to the so-called reciprocal importance sampling estimator (Gelfand and Dey 1994),

$$\hat{r}_{RI} = \left[\frac{1}{N_1}\sum_{i=1}^{N_1} \frac{q_2(\theta_1^i)}{q_1(\theta_1^i)}\right]^{-1}. \tag{5}$$

Additionally, the estimation method developed by Chib (1995) is also a special case of the bridge sampling estimator (see Gelman and Meng 1998 for more examples). Thus, the bridge sampling estimator is a generalization of several algorithms that encompass a wide range of sampling-based normalizing constants estimation methods.

Meng and Wong (1996) proposed that an optimal choice of $\omega()$, in the sense of minimizing the asymptotic Relative Mean Square Error (RMSE), is given by the reciprocal of a mixture between the two densities,

$$\omega_O(\theta) \propto \frac{1}{N_1 q_1(\theta) + r N_2 q_2(\theta)}, \tag{6}$$

provided draws from both distributions are independent. Since $\omega_O()$ still involves the unknown r, Meng and Wong (1996) suggested the following iterative computational procedure:

$$\hat{r}_O^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\left[\frac{l(\theta_2^i)}{N_1 l(\theta_2^i) + N_2 \hat{r}_O^{(t)}}\right]}{\frac{1}{N_1}\sum_{j=1}^{N_1}\left[\frac{1}{N_1 l(\theta_1^j) + N_2 \hat{r}_O^{(t)}}\right]}, \tag{7}$$

where $\hat{r}_O^{(t)}$ is the t^{th} iteration of the estimator and $l(\theta) = \frac{q_1(\theta)}{q_2(\theta)}$. Starting with an initial guess, $\hat{r}_O^{(0)}$, the optimal bridge estimate, $\hat{r}_O$, can be obtained by iterating (7) until convergence.

2.1 The Empirical Bias of Optimal Bridge Estimates

Meng and Wong (1996) only considered the asymptotic behaviour of $\hat{r}_O$ in terms of the RMSE. However, the practical behaviour of $\hat{r}_O$ computed using finite number of samples (due to limited computational resources) is often of significant interest too. Moreover, it is also insightful to investigate the bias and standard error of $\hat{r}_O$ separately rather than using the RMSE as a measure of overall efficiency, which considers the bias and standard error altogether, where

$$\text{RMSE} = \frac{\mathbb{E}[(\hat{r}-r)^2]}{r^2} = \frac{[\mathbb{E}(\hat{r})-r]^2}{r^2} + \frac{\mathbb{E}[(\hat{r}-\mathbb{E}(\hat{r}))^2]}{r^2}$$
$$= (\text{Relative Bias})^2 + (\text{Relative standard error})^2.$$

The existence of the non-negligible bias of $\hat{r}_O$ can be illustrated using a toy example as described below. Suppose that we are interested in evaluating the integral $\int q_1(\theta)\mathrm{d}\theta$, given that we are able to generate a sample of size N from $q_1()$, $\{\theta_1^{1:N}\}$. In order to use the optimal bridge sampling estimator, we would choose a normalized density, $q_2(\theta) = p_2(\theta)$, so that the answer to the above integral is intended to be c_1 (unknown). According to Meng and Wong (1996), the efficiency of the bridge sampling estimator will be minimized when the area of "overlapping" (the "harmonic" divergence in their definition) is large. An immediate choice of q_2 for this purpose is then a normal distribution with moments chosen to match the sample moments of $\{\theta_1^{1:N}\}$, or more generally, denoting $\{\theta_1\}$ as $\{\theta_1^{1:N}\}$, we write $q_2 = q_2^{\{\theta_1\}}$ and $p_2 = p_2^{\{\theta_1\}}$ as densities that depend on the sample from p_1. Throughout, we also use the notation $p_2 \leftarrow \{\theta_1\}$ to denote the case when p_2 is dependent on the samples from p_1, while $p_2 \nleftarrow \{\theta_1\}$ indicates that p_1 and p_2 are independently chosen. Making use of the information contained within the samples from p_1 to derive p_2 guarantees that the "overlapping" between p_1 and p_2 is large. However, as we demonstrate in the simulation study below, this also introduces bias to the corresponding estimate, $\hat{r}_O$, due to the correlation induced between the samples from p_1 and p_2.

As an illustration, let p_1 be the density of a univariate standard normal distribution, $N(0,1)$ with $\{\theta_1^{1:N}\}$ as the corresponding sample. Then let p_2 be the density of $N(\bar{\theta}_1, \hat{\sigma}_1^2)$, where $\bar{\theta}_1$ and $\hat{\sigma}_1^2$ are the sample mean and variance derived from $\{\theta_1^{1:N}\}$, i.e. $\bar{\theta}_1 = \frac{\sum_{i=1}^N \theta_1^i}{N}$ and $\hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (\theta_1^i - \bar{\theta})^2}{N-1}$. A sample of size N, $\{\theta_2^{1:N}\}$, is then generated from p_2. For simplicity, we assume that $q_1 = p_1$ and $q_2 = p_2$ so that the true value of r is known a priori to be one. We then evaluate Equation (7) at the entirety of samples from p_1 and p_2 (whence $N_1 = N_2 = N$), and consider the behaviour of $\hat{r}_O$ with varying sample sizes from the set $N \in \{100, 200, \ldots, 10000\}$. Each computation is also replicated $R = 10\,000$ times to learn about the underlying distribution of $\hat{r}_O$ for each N.

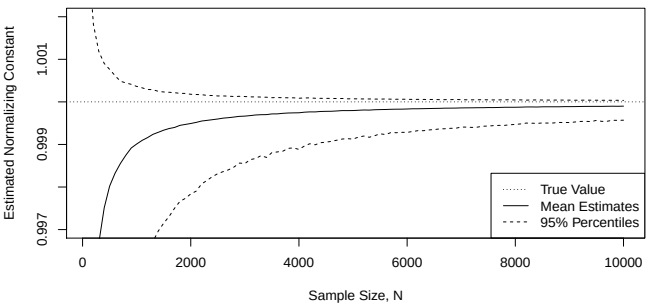

Fig. 1: Plot of the mean estimates of the ratio of normalizing constants, r, against sample size using $\hat{r}_O$, accompanied by the associated 95% intervals.

Figure 1 depicts the mean and 95% intervals (constructed from the sample percentiles) of $\hat{r}_O$, plotted against N. Evidently, there is a systematic underestimation of the value of $r = 1$, where the bias slowly diminishes as N increases, confirming the assertion by Meng and Wong (1996) that the bias term is asymptotically negligible. However, it is clear that for small to moderate N, the bias is non-negligible. Even though the magnitude of the bias appears to be non-significant in this uni-dimensional case, the negative bias will be further amplified as the dimension of the parameter increases. To put this into perspective, performing the bridge sampling on a 100-dimensional standard normal distribution with a sample size of 10,000 yields an estimate of $\hat{r}_O \approx 0.77$, which is considerably lower than the actual value. Therefore, it is imperative to understand the behaviour of the bias so that we could identify the optimal bridge estimate produced in a specific practical application with certain level of confidence.

2.2 Investigating the Origin of the Bias of $\hat{r}_O$

It is challenging to derive theoretical properties of the iteratively produced $\hat{r}_O$. Thus, the bias analysis is achieved by breaking $\hat{r}_O$ down into smaller components, $\hat{r}_I$ and $\hat{r}_{RI}$, the biases of which are analysed separately. $\hat{r}_I$ is shown to produce unbiased estimates, and hence, the bias of $\hat{r}_O$ can be traced from $\hat{r}_{RI}$. There

are two types of biases for $\hat{r}_{RI}$, which we attempt to describe in two steps: first, Taylor expansion is used to show that the (positive) bias of $\hat{r}_O$ depends on the distance between p_1 and p_2 when $p_2 \nleftarrow \{\theta_1\}$; second, we demonstrate that even when p_2 is constructed to resemble p_1 using samples from p_1 (using method of moments), i.e. $p_2 \leftarrow \{\theta_1\}$, the correlation induced bias through the moments-matching procedure (as observed in Section 2.1) is of negative magnitude and is amplified in high-dimensional problems.

Recall from Equation (6) that ω_O is essentially the reciprocal of a mixture between the two densities, q_1 and q_2, which can be alternatively expressed as

$$\omega_O(\theta) \propto \left(\frac{N_1}{\omega_{RI}(\theta)} + \frac{rN_2}{\omega_I(\theta)} \right)^{-1}. \tag{8}$$

In other words, $\hat{r}_O$ is essentially formed from a combination (in some way) between $\hat{r}_I$ and $\hat{r}_{RI}$. Hence, we expect $\hat{r}_O$ to inherit some properties from both $\hat{r}_I$ and $\hat{r}_{RI}$, even though $\hat{r}_O$ is regarded as an improved version in the sense of having a smaller RMSE than both, as proven by Meng and Wong (1996). On a side note, this is the reason why the bridge sampling technique is robust with respect to the tail behaviour of q_2 as compared to $\hat{r}_I$ and $\hat{r}_{RI}$, because the requirements of heavier-tailed and lighter-tailed important sampling densities respectively (as explained by Frühwirth-Schnatter 2004) counteract each other upon "averaging".

As a crude indication of the above relationship between $\hat{r}_O$, $\hat{r}_I$ and $\hat{r}_{RI}$, Figure 2 is created, where the estimates of $\hat{r}_I$ and $\hat{r}_{RI}$ (computed in similar set up as in Section 2.1) are included as a comparison. Clearly, the estimates of $\hat{r}_O$ lie within those of $\hat{r}_I$ and $\hat{r}_{RI}$ (including the percentiles). Thus, it is a plausible strategy to break down the problem of investigating the bias of $\hat{r}_O$ into investigating the bias of $\hat{r}_I$ and $\hat{r}_{RI}$ separately, which is much easier.

Importance sampling estimates are known to be unbiased (Chen et al. 2000, p. 127). Even in the case where $p_2 \leftarrow \{\theta_1\}$, it can be shown that the resulting $\hat{r}_I$ is unbiased (visibly evident in Figure 2). To see this, we note that conditional on $\{\theta_1\}$, $p_2^{\{\theta_1\}}$ is just an ordinary density function, then $\mathbb{E}[\hat{r}_I] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\theta_1}\left[\mathbb{E}_{\theta_2^i|\theta_1}\left[\frac{q_1(\theta_2^i)}{q_2^{\{\theta_1\}}(\theta_2^i)}\right]\right] = \frac{c_1}{c_2}$ (see Appendix A for proof).

On the contrary, $\hat{r}_{RI}$, is notorious for producing biased estimates (e.g. Neal 1994). $\hat{r}_{RI}$ belongs to the ratio estimator, which is known to overestimate the normalizing constant when $p_2 \nleftarrow \{\theta_1\}$. In particular, using

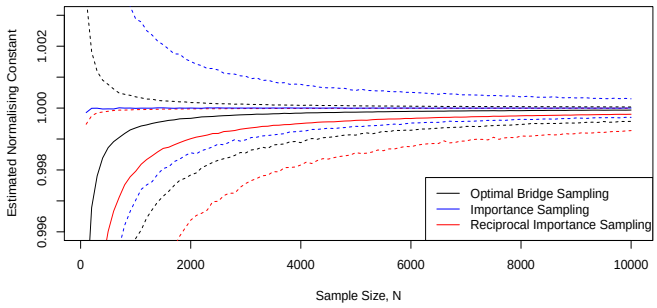

Fig. 2: Plot of the mean estimates (solid lines) of the ratio of normalizing constants against sample size using $\hat{r}_O$, $\hat{r}_I$, and $\hat{r}_{RI}$, accompanied by the associated 95% intervals (dotted lines).

Jensen's inequality, it can be shown that

$$\mathbb{E}[\hat{r}_{RI}] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\frac{q_2(\theta_1^i)}{q_1(\theta_1^i)}\right)^{-1}\right] \geq \left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\frac{q_2(\theta_1^i)}{q_1(\theta_1^i)}\right]\right)^{-1} = \frac{c_1}{c_2}.$$

More specifically, we can derive the approximate magnitude of the overestimation by using a Taylor expansion. If $\bar{\eta} = \frac{1}{N}\sum_{i=1}^{N}\frac{q_2(\theta_1^i)}{q_1(\theta_1^i)}$, then we have $\eta \equiv \mathbb{E}[\bar{\eta}] = \frac{c_2}{c_1}$ when $p_2 \nleftarrow \{\theta_1\}$. Applying a Taylor series expansion about $\eta = \mathbb{E}[\bar{\eta}]$ gives $\mathbb{E}[\hat{r}_{RI}] \approx \frac{c_1}{c_2} + \left(\frac{c_1}{c_2}\right)^3 \times \mathrm{Var}[\bar{\eta}] - \left(\frac{c_1}{c_2}\right)^4 \times \mathbb{E}[(\bar{\eta}-\eta)^3]$ to the third order approximation (see Appendix B). When $p_2 \nleftarrow \{\theta_1\}$, and $\{\theta_1^1, \ldots, \theta_1^N\}$ are random independent realizations from p_1, then $\mathrm{Var}[\bar{\eta}] = \frac{1}{N}\mathrm{Var}\left[\frac{q_2(\theta_1^i)}{q_1(\theta_1^i)}\right] = O\left(\frac{1}{N}\right)$ and $\mathbb{E}[(\bar{\eta}-\eta)^3] = \frac{1}{N^2}\mathbb{E}\left[\left(\frac{q_2(\theta_1^i)}{q_1(\theta_1^i)} - \eta\right)^3\right] = O\left(\frac{1}{N^2}\right)$. The term $\mathbb{E}[(\bar{\eta}-\eta)^3]$ typically possesses negligible value and can be ignored. Therefore, $\hat{r}_{RI}$ carries a positive bias of magnitude $\left(\frac{c_1}{c_2}\right)^3 \times \mathrm{Var}[\bar{\eta}]$ (order $1/N$) approximately, which vanishes as $N \to \infty$. Note that $\mathrm{Var}[\bar{\eta}]$ can be expressed as

$$\mathrm{Var}[\bar{\eta}] = \frac{1}{N}\left(\frac{c_2}{c_1}\right)^2\left(\int \frac{p_2^2(\theta)}{p_1(\theta)}\mathrm{d}\theta - 1\right) = \frac{1}{N}\left(\frac{c_2}{c_1}\right)^2\left[\mathbb{E}_{p_2}\left[\frac{p_2}{p_1}\right] - 1\right],$$

where $\mathbb{E}_{p_2}\left[\frac{p_2}{p_1}\right]$ resembles the Kullback-Leibler divergence (Kullback and Leibler 1951), $\mathbb{E}_{p_2}\left[\log\left(\frac{p_2}{p_1}\right)\right]$, to a certain degree. This implies that $\mathrm{Var}[\bar{\eta}]$ measures the divergence of p_2 from p_1, which is an indication of how much they overlap. The smaller the overlap between p_1 and p_2, the larger the value of $\mathrm{Var}[\bar{\eta}]$, and hence, the

larger the bias (see Appendix C). Of course, $\mathbb{E}_{p_2}\left[\frac{p_2}{p_1}\right]$ is not always finite as $\frac{p_2}{p_1}$ is not always square integrable with respect to p_1 as pointed out by Meng and Wong (1996), but the multiplicative factor of $\frac{1}{N}$ ensures that the practical bias vanishes as N increases. This phenomenon can also be observed from Meng and Schilling (1996), where the positive bias of $\hat{r}_O$ in their simulation increases as the divergence between p_1 and p_2 (measured in Hellinger distance) increases.

The above derivation does not explain the underestimation in Figure 2 when $p_2 \leftarrow \{\theta_1\}$. Again, we focus on situation where $p_2^{\{\theta_1\}}$ is constructed to be close to p_1 (using method of moments). Intuitively, this is due to the correlation between the samples from p_1 and p_2 through the sample moments, which then manifests itself in the form of a systematic bias. More specifically, the bias switches sign because the term $\mathrm{Var}[\bar{\eta}]$ becomes smaller when p_1 and p_2 are close. Hence, the supposedly positive bias of $\hat{r}_{RI}$ is dominated by another source of bias, which originates from $\frac{1}{\eta}$, where $\eta = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\theta_1^{-i}}\left[\mathbb{E}_{\theta_1^i|\theta_1^{-i}}\left[\frac{q_2^{\{\theta_1\}}(\theta_1^i)}{q_1(\theta_1^i)}\right]\right] \neq \frac{c_2}{c_1}$ with $\{\theta_1^{-i}\} = \{\theta_1^1, \ldots, \theta_1^{i-1}, \theta_1^{i+1}, \ldots, \theta_1^N\}$. It is difficult to derive a simpler mathematical expression for η here, even in a very simple case involving normal distributions (a sketch proof when p_1 and p_2 are both exponential is provided in Appendix D). But empirically, it has been found that $\eta > \frac{c_2}{c_1}$, resulting in an underestimation of the true value of $r = \frac{c_1}{c_2}$, i.e. $\mathbb{E}[\hat{r}_{RI}] < \frac{c_1}{c_2}$ (as observed in Figure 2).

Using the delta method, the variance of $\hat{r}_{RI}$ (to the second order) can be expressed as

$$\mathrm{Var}\left[\frac{1}{\bar{\eta}}\right] \approx \frac{c_2^2}{N^2\eta^4 c_1^2}\mathrm{Var}\left[\sum_{i=1}^{N}\frac{p_2^{\{\theta_1\}}(\theta_1^i)}{p_1(\theta_1^i)}\right]. \tag{9}$$

3 The Effect of Sample Size Allocation on the Accuracy of $\hat{r}_O$

N_1 and N_2 appear in $\omega_O()$ as the mixture proportions of q_1 and q_2 respectively (see Equation (6)). Given that $\omega_O()$ plays the role to provide an optimal linkage between the two densities, it is logical that the allocation of samples sizes directly influences the efficiency of the resulting bridge sampling estimate. To the best of our knowledge, the effect of relative sample sizes on the efficiency of $\hat{r}_O$ has yet to be investigated. Although Chen et al. (2000, p. 129) vaguely stated that the optimal choice of $\omega()$ is more vital than the optimal allocation of sample sizes, a more thorough study on the effect of relative sample sizes could potentially lead to ways of improving the efficiency of $\hat{r}_O$.

By inspecting Equation (8), we note that the relative sizes of N_1 and N_2 determine the resulting behaviour of $\hat{r}_O$ based on the weights given on the mixture components. Allocating a larger N_1 relative to N_2 corresponds to prioritizing the ω_{RI} component, implying that the resulting $\hat{r}_O$ behaves more similarly to $\hat{r}_{RI}$. By contrast, allocating a smaller N_1 relative to N_2 corresponds to prioritizing the ω_I component, meaning the behaviour of $\hat{r}_O$ is more inclined towards $\hat{r}_I$. Using $N_1 = N_2$ corresponds to the original bridge sampling estimate recommended by Meng and Wong (1996). In the extreme case where $N_1 = 0$ (where none of the samples from p_1 is used to evaluate the estimator), then the bridge sampling procedure produces $\hat{r}_I$ exactly.

Since it was discovered from Section 2.1 that $\hat{r}_I$ is unbiased, while $\hat{r}_{RI}$ produces biased estimates, the relative values of N_1 and N_2 indirectly govern the bias of $\hat{r}_O$. Here, we focus on a scenario where it is computationally expensive to simulate from p_1, while it is relatively cheaper to simulate samples from p_2 and to evaluate these samples at p_1. Thus, we investigate the possibility of using $N_2 > N_1 = N$ to improve the efficiency of $\hat{r}_O$ with little increase in the computational effort, under the computational constraint that N could not be freely increased. Generally speaking, using a larger N_2 corresponds to allocating more weight to ω_I (and hence the unbiased $\hat{r}_I$), which then reduces the associated bias for $\hat{r}_O$ since less weight is given to the biased $\hat{r}_{RI}$ when N_1 is relatively small. Moreover, using a larger N_2 reduces the standard error of $\hat{r}_O$ since the estimator is evaluated at a greater number of samples. Therefore, in theory, we expect that using a larger N_2 not only diminishes the bias, but also decreases the standard error of $\hat{r}_O$.

Returning to the simulation study in Section 2.1, rather than only setting $N_2 = N$, three different sample size allocations are examined:

i. Naive approach, $N_2 = N$.
ii. A constant multiple of N, $N_2 = 10N$.
iii. Some relatively large number, $N_2 = 50\,000$.

$\hat{r}_O$ is then evaluated at the entire samples from both p_1 and p_2, so that $N_1 = N$ and N_2 is from one of the above.

As shown in Figure 3, a larger N_2 generally leads to better estimates by reducing both the bias and standard error as hypothesized. For $N_2 = 50\,000$ (blue), the bias remarkably shrinks to almost zero for all N. For $N_2 = 10N$ (red), the performance of $\hat{r}_O$ with respect to N is consistently better relative to using $N_2 = N$ (black), overtaking that of using $N_2 = 50\,000$ at $N = 5000$ (when blue and red lines cross each other), where the red outperforms the blue by having a larger

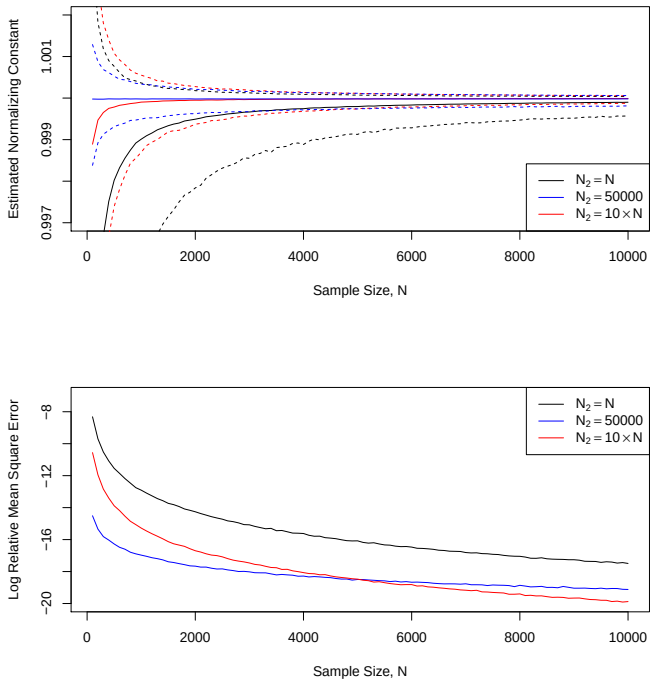

Fig. 3: Top panel shows the mean estimates of $\hat{r}_O$ plotted against N for various N_2, accompanied by the associated 95% intervals (dotted lines). The bottom panel shows the corresponding log RMSE.

N_2. The RMSE for $N_2 = 10N$ also appears to decrease indefinitely as N increases, while the improvement for $N_2 = 50\,000$ slowly decelerates with increasing N (mainly because its standard error does not reduce considerably towards the end). This indicates that the efficiency of $\hat{r}_O$ could be further improved by using an even larger N_2. Therefore, it is clearly possible to improve the efficiency of $\hat{r}_O$ (by reducing both the bias and standard error) by using a relatively larger N_2 in this particular example.

The previous result is to be expected since using a relatively large N_2 implies that $\hat{r}_O$ behaves more closely to $\hat{r}_I$. Using a normal distribution as an important sampling distribution to compute the normalizing constant of another normal distribution is certainly going to behave well as they possess similar tail behaviour. It is perhaps more interesting to consider a heavier tailed p_1 (where importance sampling procedures are known to be less efficient) and assess if the improvement due to a larger N_2 is as apparent as it was previously. Suppose now that p_1 and q_1 are densities of Student's t-distribution with three degrees of freedom (t_3), using a similar set up as before, the behaviour of $\hat{r}_O$ in response to N is examined (refer to Figure 4). Remarkably, similar patterns are observed even though the improvement is less substantial when compared to the previous case. The bias and standard error of $\hat{r}_O$ are now larger due to the difference in nature between p_1 and p_2 (mostly due to different tail behaviours), resulting in a larger overall RMSE than the previous case. In conclusion, it can be deduced that it is generally beneficial to use a larger N_2 given a fixed samples from p_1 during the evaluation of $\hat{r}_O$ if p_2 is the sample moments-matched normal density, since this reduces both the bias and standard error of $\hat{r}_O$ with little increase in computational effort (assuming that p_1 is quick to evaluate). More specifically, using a relatively larger N_2 corresponds to altering the priority between the importance sampling and reciprocal importance sampling method with which the evaluation of $\hat{r}_O$ is based upon, favouring the unbiased importance sampling method more while partially retaining the good behaviour of the reciprocal importance sampling method.

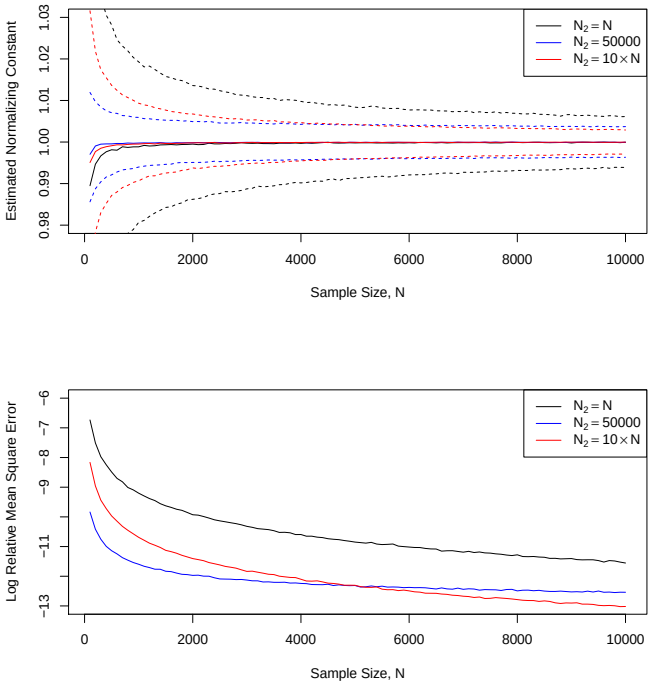

Fig. 4: Top panel shows the mean estimates of $\hat{r}_O$ plotted against N for various N_2, accompanied by the associated 95% intervals (dotted lines), when p_1 is the density of t_3. The bottom panel shows the corresponding log RMSE.

4 The Splitting Approach

The splitting approach is first introduced and applied to $\hat{r}_{RI}$, where the results are examined as motivation

for $\hat{r}_O$ in the next subsection. Since the negative bias of $\hat{r}_{RI}$ originates mainly from using $p_2^{\{\theta_1\}}$ through the moments-matching procedure, the easiest method to mitigate this issue is to compute the sample moments using only a portion of the sample from p_1, and then evaluate the estimator at the remaining sample. More formally, suppose k is the proportion of the samples from p_1 where the moments are computed, where we then write $p_2^s = p_2^{\{\theta_1^{1:kN}\}}$ as the resulting density. In what follows, for ease of exposition, we assume that kN is an integer. Where kN is not an integer, it should be replaced by the largest integer not exceeding kN. For example, in our simulation study before (see Section 2.2), p_2^s is the density of $N(\bar{\theta}_1^s, (\hat{\sigma}_1^s)^2)$, where $\bar{\theta}_1^s = \frac{\sum_{i=1}^{kN}\theta_1^i}{kN}$ and $(\hat{\sigma}_1^s)^2 = \frac{\sum_{i=1}^{kN}(\theta_1^i - \bar{\theta}_1^s)^2}{kN-1}$. Equation (5) is then appropriately evaluated at the remaining samples from p_1 to give $\hat{r}_{RI}^s$, i.e.

$$\hat{r}_{RI}^s = \left[\frac{1}{(1-k)N}\sum_{i=kN+1}^{N}\frac{q_2^s(\theta_1^i)}{q_1(\theta_1^i)}\right]^{-1}, \tag{10}$$

where $N_1 = (1-k)N$ here and q_2^s is defined analogously as p_2^s. As an illustration, the splitting approach can be represented by the diagram in Figure 5.

$$\underbrace{\theta_1^1, \ldots\ldots\ldots\ldots\ldots, \theta_1^{kN}}_{\text{Compute } \bar{\theta}_1^s \text{ and } (\hat{\sigma}_1^s)^2}, \underbrace{\theta_1^{kN+1}, \ldots\ldots\ldots\ldots\ldots, \theta_1^N}_{\text{Evaluate Equation (10)}} \longrightarrow \hat{r}_{RI}^s$$

Fig. 5: A summary of the splitting approach, where the first subset of the samples from p_1 is used to derive moments for constructing p_2, while the second subset is used to evaluate the estimator.

This technique has been applied by Overstall and Forster (2010), Wong (2017), and Wang et al. (2019), but the underlying mathematical principles were not discussed in detail. To see how the splitting approach manages to alleviate the bias, we let $\bar{\eta}^s = \frac{1}{(1-k)N}\sum_{i=kN+1}^{N}\frac{q_2^s(\theta_1^i)}{q_1(\theta_1^i)}$. Then, it can be shown that $\eta^s = \mathbb{E}[\bar{\eta}^s] = \frac{c_2}{c_1}$, implying that $\hat{r}_{RI}^s$ is unbiased to the second order (see Appendix E for technical details).

Unfortunately, the elimination of bias occurs at the expense of yielding a larger standard error for the resulting estimate. The primary intuition behind this is the fact that $\hat{r}_{RI}^s$ is only evaluated at a shorter portion of the original sample. More specifically, the variance of $\hat{r}_{RI}^s$ can be approximated (see Appendix E) as

$$\mathrm{Var}[\hat{r}_{RI}^s] \approx \frac{c_1^2}{(1-k)Nc_2^2} \times \mathbb{E}_{\theta_1^{1:kN}}\left[\mathrm{Var}_{\theta_1^N|\theta_1^{1:kN}}\left(\frac{p_2^{\{\theta_1^{1:kN}\}}(\theta_1^N)}{p_1(\theta_1^N)}\right)\right]. \tag{11}$$

It is then of interest to compare $\mathrm{Var}[\hat{r}_{RI}]$ (in Equation (9)) with $\mathrm{Var}[\hat{r}_{RI}^s]$ (in Equation (11)). The following crude calculation is presented to illustrate the approximate increase in the variance due to the splitting approach (exact calculation involves intractable expressions). Firstly, we momentarily assume that $\eta \approx \frac{c_2}{c_1}$ (even though we know $\eta > \frac{c_2}{c_1}$) since the misestimation is relatively small here. Secondly, for the purpose of illustration, we assume that

$$\mathrm{Var}\left[\sum_{i=1}^{N}\frac{p_2^{\{\theta_1^{1:N}\}}(\theta_1^i)}{p_1(\theta_1^i)}\right] \approx \sum_{i=1}^{N}\mathrm{Var}\left[\frac{p_2^{\{\theta_1^{1:N}\}}(\theta_1^i)}{p_1(\theta_1^i)}\right] = N\cdot\mathrm{Var}\left[\frac{p_2^{\{\theta_1^{1:N}\}}(\theta_1^N)}{p_1(\theta_1^N)}\right],$$

i.e. the individual components within the summation are independent (which is generally not true in practice). Expression in (9) then becomes $\mathrm{Var}\left[\frac{1}{\bar{\eta}}\right] \approx \frac{c_1^2}{Nc_2^2} \times \mathrm{Var}\left[\frac{p_2^{\{\theta_1^{1:N}\}}(\theta_1^N)}{p_1(\theta_1^N)}\right]$. Finally, we note that the terms $\mathrm{Var}\left[\frac{p_2^{\{\theta_1^{1:N}\}}(\theta_1^N)}{p_1(\theta_1^N)}\right]$ and $\mathbb{E}_{\theta_1^{1:kN}}\left[\mathrm{Var}_{\theta_1^N|\theta_1^{1:kN}}\left(\frac{p_2^{\{\theta_1^{1:kN}\}}(\theta_1^N)}{p_1(\theta_1^N)}\right)\right]$ carry similar interpretation of being the average of the conditional variance of $\frac{p_2}{p_1}$ over the samples involved in constructing p_2, and hence, we have $\frac{\mathrm{Var}[\hat{r}_{RI}^s]}{\mathrm{Var}[\hat{r}_{RI}]} \approx \frac{1}{1-k}$, which means that the splitting approach increases the variance of the resulting estimate by a factor of $\frac{1}{1-k}$ approximately. Crudely speaking, the ratio of the variances is approximately the ratio of the proportion of samples used to evaluate the estimators (which is what we observed empirically), e.g. when $k = 1/2$, then $\mathrm{Var}[\hat{r}_{RI}^s] \approx 2\mathrm{Var}[\hat{r}_{RI}]$. The crude calculation above does not hold in general due to the simplifying assumptions used, but nevertheless, it provides an intuition of how the splitting approach leads to increased standard error. Regardless, it is evident that the splitting approach manages to correct the bias by avoiding the use of the same samples for moments-matching and evaluation of the estimator, but at the same time introduces more variations to the resulting estimates (by having a smaller sample size to work with). Therefore, the efficacy of the splitting approach in improving the estimator is dictated by the trade-off between the bias and variance.

Now, consider simulation study similar to that conducted in Section 2.1, except now we are interested in investigating the empirical behaviour of $\hat{r}_{RI}$ and $\hat{r}^s_{RI}$. Suppose p_1 and q_1 are densities of $N(0,1)$, with a sample of size N, $\{\theta_1^{1:N}\}$, being made available. We assess the following two approaches of constructing p_2.

1. Approach 1 (naive): p_2 is the density of $N(\bar{\theta}_1, \hat{\sigma}_1^2)$, where $\bar{\theta}_1 = \frac{\sum_{i=1}^N \theta_1^i}{N}$ and $\hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (\theta_1^i - \bar{\theta}_1)^2}{N-1}$. Hence, the reciprocal importance sampling estimator is evaluated at the entire samples, $\{\theta_1^{1:N}\}$, where $N_1 = N$, producing $\hat{r}_{RI}$.
2. Approach 2 (splitting): p_2 is the density of $N(\bar{\theta}_1^s, (\hat{\sigma}_1^s)^2)$, where $\bar{\theta}_1^s = \frac{\sum_{i=1}^{kN} \theta_1^i}{kN}$ and $(\hat{\sigma}_1^s)^2 = \frac{\sum_{i=1}^{kN} (\theta_1^i - \bar{\theta}_1^s)^2}{kN-1}$. Hence, the reciprocal importance sampling estimator is evaluated at the remaining samples from p_1, $\{\theta_1^{(kN+1):N}\}$, whence $N_1 = (1-k)N$, producing $\hat{r}^s_{RI}$.

Again, we set $q_2 = p_2$ so that the true value of r is one. We consider $N \in \{100, 200, \ldots, 10000\}$, with each computation replicated $R = 10\,000$ times. We also examine six different splitting proportions, $k = 1/10, 1/5, 1/3, 1/2, 4/5, 9/10$.

Figure 6 illustrates how the mean estimates and the associated 95% intervals of $\hat{r}_{RI}$ and $\hat{r}^s_{RI}$ (for various splitting proportions) vary against N. As expected, the naive approach systematically underestimates the true value, where the bias is a decreasing function of N (we hypothesize that it is of order $1/N$ but this remains to be proven). On the other hand, the splitting approach produces unbiased estimates for all k considered. This approach also yields comparatively wider (but symmetric) intervals than the naive approach (which yields asymmetric intervals), as consistent with our mathematical derivation above. Among the different splitting proportions, $k = 1/2$ appears to be the best by producing narrowest intervals. Interestingly, $\hat{r}^s_{RI}$ with $k = 1/10$ and $k = 9/10$ produce intervals of similar width, while those produced with $k = 1/5$ and $k = 4/5$ are similar too. This signifies that the variance of $\hat{r}^s_{RI}$ decreases as k increases until $k = 1/2$, and then begin to increase again until $k = 1$ in a symmetrical manner.

With reference to the bottom panel of Figure 6, $\hat{r}^s_{RI}$ with $k = 1/2$ possesses the lowest RMSE at each N, but is exactly the same as $\hat{r}_{RI}$, implying that the trade-off between the bias and variance does not particularly favour either of them in this context. Hence, it is a matter of preference, whether one prefers to deal with biased estimates, or estimates with larger uncertainty. Despite having the same RMSE, $\hat{r}^s_{RI}$ with $k = 1/2$ is arguably better than the naive approach as it alleviates the bias completely and has lower computational cost since the estimator is only evaluated at a smaller portion of the sample (this could be beneficial in scenario where the evaluation of p_1 is slow).

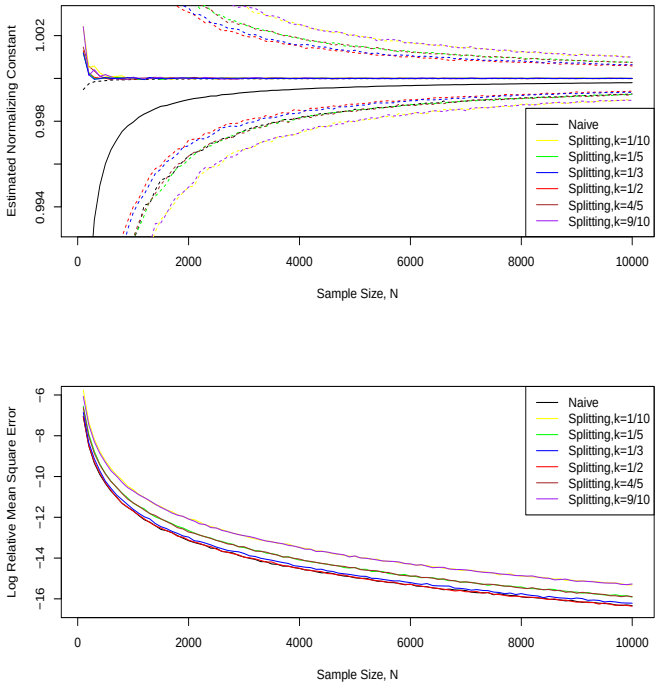

Fig. 6: Top panel shows the mean estimates (solid lines) of $\hat{r}_{RI}$ and $\hat{r}^s_{RI}$ with various k plotted against N, accompanied by the associated 95% intervals (dotted lines). The bottom panel shows the corresponding log RMSE.

As mentioned before, the negative bias of the naive approach worsens as dimensionality increases. To explicitly explore the behaviour of $\hat{r}_{RI}$ and $\hat{r}^s_{RI}$ with respect to the dimensionality involved, we consider the case when p_1 and q_1 are both densities of a 10-dimensional standard normal distribution: $N(0, I_{10})$. A sample of size N is generated from this distribution to form $\{\theta_1^{1:N}\}$. Then let p_2 and q_2 be the densities of normal distributions with mean and variance derived from $\{\theta_1^{1:N}\}$ in a similar set up as described previously, except each computation is only replicated $R = 1000$ times for computational feasibility (producing more erratic curves).

According to Figure 7, similar patterns are observed: that $\hat{r}_{RI}$ systematically underestimates the true r, while $\hat{r}^s_{RI}$ yields unbiased estimates at the expense of larger standard errors. A closer inspection revealed that the underestimation of $\hat{r}_{RI}$ is much more apparent than the uni-dimensional case, confirming that the bias of $\hat{r}_{RI}$ is amplified by the dimensionality involved. The ranking performance of k is preserved, in that the closer k is to $1/2$, the better the resulting estimate. Notice also that $\hat{r}^s_{RI}$ now outperforms $\hat{r}_{RI}$ for all k

considered in terms of the RMSE (the bottom panel of Figure 7). This is primarily because the splitting approach still manages to alleviate the bias despite the higher dimensionality, whereas the bias produced by the naive approach scales up substantially with increasing dimensionality. The increase in standard errors due to the splitting approach no longer offsets the reduction in the bias as in the uni-dimensional case, implying that there is an overall gain in efficiency by performing the splitting for higher dimensional problems. This renders the idea of splitting more valuable, since it corrects for the correlation induced bias irrespective of the dimensionality involved.

However, one has to be cautious when N is relatively small because there appears to be a threshold before the bias elimination by $\hat{r}^s_{RI}$ operates for all k, as indicated by the idiosyncratic behaviour in Figure 7 for all k (which is also discernible in Figure 6). Intuitively, this is because a minimum number of samples is required to effectively learn about p_1 for the estimation procedure to work properly. Knowing that k determines the amount of samples used to derive moments, it is clear that the threshold sample size for the bias elimination to take effect is larger for smaller k, as evident in Figure 7.

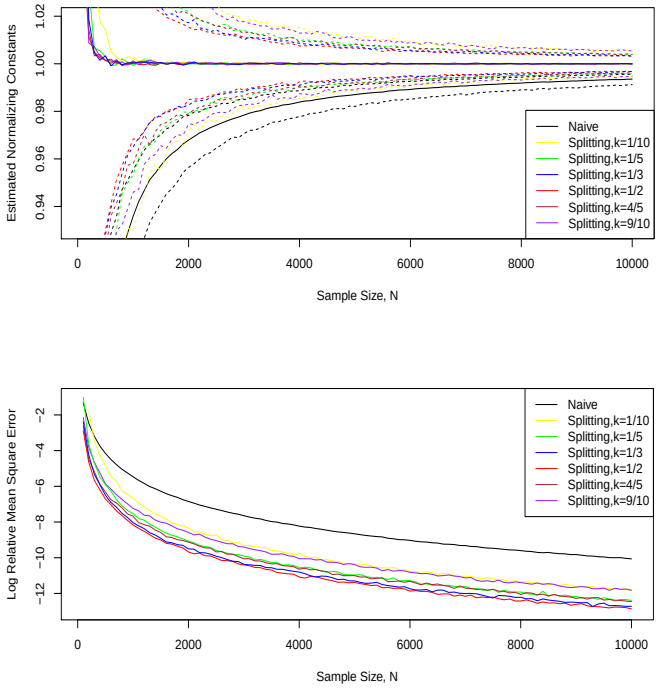

Fig. 7: Top panel shows the mean estimates (solid lines) of $\hat{r}_{RI}$ and $\hat{r}^s_{RI}$ with various k plotted against N, accompanied by the associated 95% intervals (dotted lines), for the 10-dimensional case. The bottom panel shows the corresponding log RMSE.

4.1 Applying the Idea of Splitting on $\hat{r}_O$

Since it was demonstrated that $\hat{r}_{RI}$ is the key component leading to the bias of $\hat{r}_O$, it is anticipated that implementing the splitting approach also eliminates the bias of $\hat{r}_O$. Motivated by the bias analysis in Section 2.2, we write $\hat{r}^s_O = \frac{\bar{\eta}^s_2}{\bar{\eta}^s_1}$, where

$$\bar{\eta}^s_1 = \frac{1}{N_1}\sum_{i=N-N_1+1}^{N} \frac{q^s_2(\theta^i_1)}{N_1 q_1(\theta^i_1) + rN_2 q^s_2(\theta^i_1)},$$

and

$$\bar{\eta}^s_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} \frac{q_1(\theta^i_2)}{N_1 q_1(\theta^i_2) + rN_2 q^s_2(\theta^i_2)},$$

with $N_1 = (1-k)N$ and $N_2 = N$. When draws are random and independent, and conditional on the samples for moments estimation, we obtain

$$\eta^s_1 \equiv \mathbb{E}_{\theta_1^{(kN+1):N}|\theta_1^{1:kN}}[\bar{\eta}^s_1] = \frac{c_2}{c_1}\int \frac{p_1(\theta)p^s_2(\theta)}{N_1 p_1(\theta) + rN_2 p^s_2(\theta)}\mathrm{d}\theta,$$

and

$$\eta^s_2 \equiv \mathbb{E}_{\theta_2^{1:N}|\theta_1^{1:kN}}[\bar{\eta}^s_2] = \int \frac{p_1(\theta)p^s_2(\theta)}{N_1 p_1(\theta) + rN_2 p^s_2(\theta)}\mathrm{d}\theta.$$

Thus, it can be shown that (see Appendix F)

$$\begin{aligned}\mathbb{E}[\hat{r}^s_O] &= \mathbb{E}_{\theta_1^{1:kN}}\left[\mathbb{E}_{\theta_1^{(kN+1):N},\theta_2^{1:N}|\theta_1^{1:kN}}\left(\frac{\bar{\eta}^s_2}{\bar{\eta}^s_1}\right)\right] \\ &\approx \frac{c_1}{c_2} + \frac{c_1}{c_2(\eta^s_1)^2}\times \mathbb{E}_{\theta_1^{1:kN}}[\mathrm{Var}(\bar{\eta}^s_1)]. \end{aligned} \tag{12}$$

As before, the second term in Equation (12) is going to be small when p_2 is constructed to be close to p_1 implying that $\mathbb{E}[\hat{r}^s_O] \approx \frac{c_1}{c_2}$.

The simulation study in Section 2.1 is revisited, where we now set $p^s_2 = q^s_2$ as the density of $N(\bar{\theta}^s_1, (\hat{\sigma}^s_1)^2)$ with $\bar{\theta}^s_1 = \frac{\sum_{i=1}^{kN}\theta^i_1}{kN}$ and $(\hat{\sigma}^s_1)^2 = \frac{\sum_{i=1}^{kN}(\theta^i_1 - \bar{\theta}^s_1)^2}{kN-1}$. The bridge sampler in (7) is then evaluated at the remaining samples from p_1, $\{\theta_1^{(kN+1):N}\}$, and the entire sample from p_2, $\{\theta_2^{1:N}\}$. The iterative formulae is now

$$\hat{r}^s_O{}^{(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\left[\frac{l^s(\theta^i_2)}{N_1 l^s(\theta^i_2) + N_2 \hat{r}^s_O{}^{(t)}}\right]}{\frac{1}{N_1}\sum_{j=kN+1}^{N}\left[\frac{1}{N_1 l^s(\theta^j_1) + N_2 \hat{r}^s_O{}^{(t)}}\right]}, \tag{13}$$

where $\hat{r}^s_O{}^{(t)}$ is the t^{th} iteration of the estimate, $N_1 = (1-k)N$ and $N_2 = N$, while $l^s(\theta) = \frac{q_1(\theta)}{q^s_2(\theta)}$. Equation (13) is then iterated until convergence to yield the estimate, $\hat{r}^s_O$. Figure 1 is reconstructed, including $\hat{r}^s_O$ for various $k \in \{1/10, 1/5, 1/3, 1/2, 4/5, 9/10\}$, forming

Figure 8. Note that a similar simulation study has been performed by Overstall and Forster (2010), but only $k = 1/2$ was considered and they focused mainly on the bias correction. They also simulated a shorter sample size from p_2 ($N_2 = \frac{1}{2}N$), which can be improved with no substantial additional computational cost (relative to the naive approach) based on the results from Section 3, where the increase in standard error due to the splitting approach is slightly compensated by a reduction due to allocating a larger N_2.

As expected, similar phenomena are observed from Figure 8, that $\hat{r}_O^s$ alleviates the bias at the expense of having a larger standard error for all k. In terms of the RMSE, the performance of $\hat{r}_O^s$ with $k = 1/2$ is similar to $\hat{r}_O$, as before. Interestingly, the ranking of the performance of $\hat{r}_O^s$ with respect to k is altered slightly when compared with that for $\hat{r}_{RI}^s$. This is because the computation of optimal bridge sampling estimates requires samples from both p_1 and p_2, whereas for reciprocal importance sampling estimates, samples from p_1 are only involved in the construction of p_2, but not in the evaluation of the associated estimator. In other words, there is an extra influence on the overall efficiency of $\hat{r}_O^s$ by using different k through the allocation of N_1 and N_2 (see Section 3). This highlights the difference between $\hat{r}_{RI}^s$ and $\hat{r}_O^s$, that both samples from p_1 and p_2 play a direct role in the evaluation of the estimator for the latter, but not for the former. It appears that the best performing proportion is no longer $k = 1/2$, but rather $k = 9/10$, closely followed by $k = 4/5$. Or more specifically, the larger the value of k, the better the resulting estimate in this particular instance.

For the 10-dimensional case (see Figure 9), the bias of $\hat{r}_O$ is considerably larger than the uni-dimensional case as expected, while $\hat{r}_O^s$ is unbiased for all values of k. The ranking of $\hat{r}_O^s$ in terms of k is preserved, such that the closer the value of k is to one, the better the resulting estimates. $\hat{r}_O^s$ also outperforms $\hat{r}_O$ for all values of k, as the former alleviates the bias irrespective of the dimensionality involved while the latter has an amplified bias, confirming that the splitting approach is more advantageous in higher-dimensional problems.

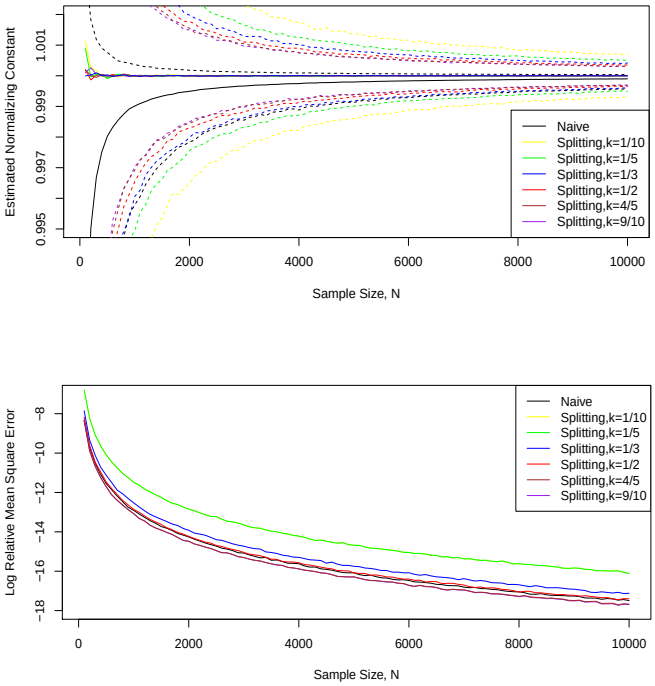

Fig. 8: Top panel shows the mean estimates (solid lines) of $\hat{r}_O$ and $\hat{r}_O^s$ with various k plotted against N, accompanied by the associated 95% intervals (dotted lines). The bottom panel shows the corresponding log RMSE.

4.2 The Optimal Choice of k for $\hat{r}_O^s$

The choice of $0 < k \leq 1$ has a two-fold effect: it determines the amount of samples used for computing sample moments (the larger the k, then more samples are used to estimate the moments of p_1, the more the p_2 constructed resembles p_1, resulting in a higher accuracy for $\hat{r}_O^s$); and the amount used to evaluate the estimator (the larger the k, the smaller the number of samples used to evaluate $\hat{r}_O^s$ and thus a less precise estimate). An interesting question then is whether it is more important to obtain a more accurate estimate of the moments (larger k), or it is more important to prioritize the evaluation of $\hat{r}_O^s$ (smaller k). The experiment in Section 4.1 may indicate that it is more crucial to maximize the area of overlap between p_1 and p_2, than having more samples from p_1 to evaluate the estimator. However, this may not be true in general because if the experiment is repeated, but with p_1 replaced by the density of (t_3), then the ranking of k is reversed, that smaller k results in better estimates (see Appendix G).

Recall from Equation (8) that the relative sizes of N_1 and N_2 determine the resulting behaviour of $\hat{r}_O$ (see Section 3 for a detailed description). In our experiments, $N_2 = N$ is fixed so k is inversely related to N_1. For instance, using a larger k effectively means a smaller N_1 is allocated for evaluating the estimator, yielding $\hat{r}_O^s$ that behaves more similarly to the importance sampling estimate. On the contrary, using a smaller k corresponds to a larger N_1 for evaluating the estimator, with N_1 approaching $N_2 = N$ as k tends to 0, producing estimates that behave like the original bridge sampler (with increasingly poorer estimate of the moments of p_1). The boundary value of $k = 1$ is equivalent to set-

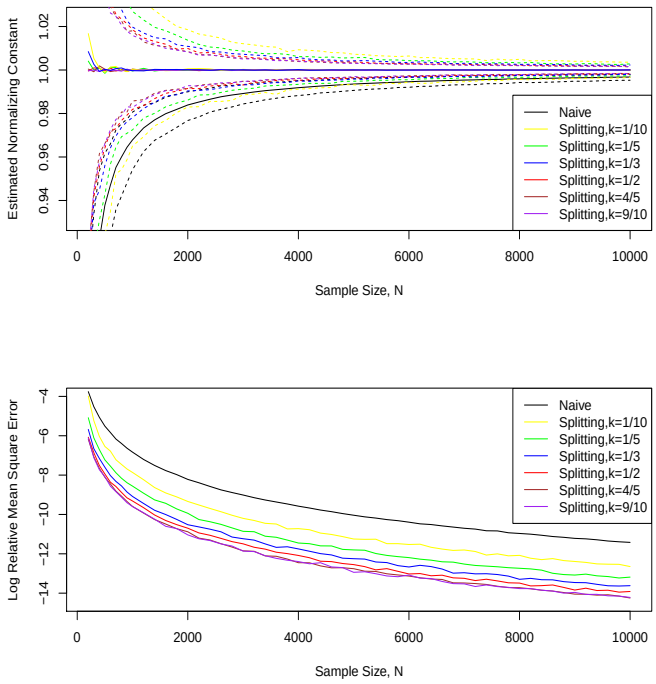

Fig. 9: Top panel shows the mean estimates (solid lines) of $\hat{r}_O$ and $\hat{r}_O^s$ for various k plotted against N, accompanied by the associated 95% intervals (dotted lines), for the 10-dimensional case. The bottom panel shows the corresponding log RMSE.

ting $N_1 = 0$, which leads to $\hat{r}_I$ using moments-matched normal distribution.

To study the influence of k, the simulation study in Section 4.1 is repeated, but with log RMSE plotted against $k \in \{0.01, 0.02, \ldots, 0.99\}$, fixing $N = 1000$, and with computation at each k replicated 10 000 times. We exclude $k = 0$ in our investigation because the resulting estimates become aggressively volatile when none of the samples from p_1 is used for constructing p_2, i.e. a normal density with any parameters could be used as p_2. We have also included several cases, each with different p_1 and p_2:

1. Case 1: p_1 is the density of $N(0, 1)$, p_2 is the sample moments-matched normal density;
2. Case 2: p_1 is the density of t_3, p_2 is the sample moments-matched normal density;
3. Case 3: p_1 is the density of a Laplace distribution with location and scale parameters given as 0 and 1 respectively, i.e. $p_1(\theta) = \frac{1}{2}\exp(-|\theta|)$, p_2 is the sample moments-matched normal density;
4. Case 4: p_1 is the density of $N(0, 1)$, while p_2 is the density of a non-standardized t_3 (Johnson et al. 1995, Chapter 28), constructed using samples of p_1 through method of moments;
5. Case 5: p_1 is the density of t_3, while p_2 is the density of a non-standardized t_3, constructed using samples of p_1 through method of moments.

According to Figure 10, the RMSE of cases 1 and 5 appear to be decreasing functions of k generally. The reason why a larger k is beneficial in cases 1 and 5 is perhaps not so surprising since p_1 and p_2 belong to the same family of distributions, and hence, prioritizing the importance sampling component due to using a large k (see above) will not be problematic. It is then more crucial to obtain a more accurate estimate of the moments to maximize the overlap between p_1 and p_2, which is achieved by using large k. In other words, when p_1 and p_2 are from the same family of distributions, the gain in statistical efficiency of $\hat{r}_O^s$ through maximizing the overlap between p_1 and p_2 outweighs the loss due to having less samples to evaluate the estimator. The reverse is true for cases 2-4, where prioritizing the importance sampling component is detrimental when p_2 is lighter-tailed (see Frühwirth-Schnatter 2004 for explanation), which is especially apparent when the RMSE is observed to increase drastically as k approaches 1 for cases 2 and 3. The same phenomenon is not observed for case 4, as p_2 is more heavy-tailed there. Notice also that similar characteristics are observed at small k for all cases, that the RMSE increases sharply as k approaches 0. This is an indication that there are insufficient samples to learn about p_1 through the moments estimated, prohibiting the bridge sampling procedure from operating efficiently. Once the threshold value is exceeded (around $k = 0.05$ according to Figure 10), then the behaviour of $\hat{r}_O^s$ begin to show consistent patterns. To conclude, the optimal value of k for $\hat{r}_O^s$ depends on the nature of p_1 and p_2: if p_1 and p_2 belong to the same family of distributions, then it is more favourable to prioritize accurate moments estimation (large k); whereas if p_1 and p_2 are from different families (which is more common in practice), it is more crucial to have more samples for evaluating the estimator, correspondingly using less samples for moments estimation (small k), provided that the minimum threshold of having sufficient samples for moments estimation is surpassed.

5 The Cross-Splitting Approach

In this section, we investigate a method of further reducing the RMSE of $\hat{r}_O^s$, given a sample of size N from p_1. Notice that while computing $\hat{r}_O^s$, the estimator in (13) is only evaluated at a portion of the samples from p_1 because the first subset is required for constructing p_2. This results in an increased variance for $\hat{r}_O^s$ due to having less samples to evaluate the estimator (Section

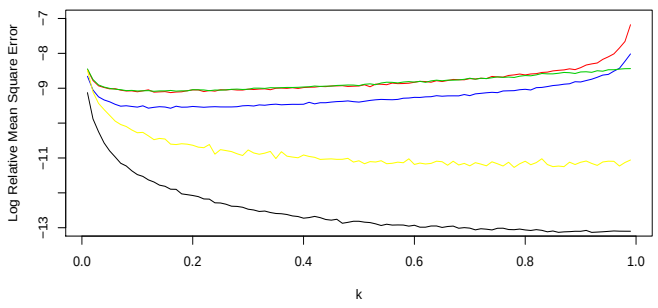

Fig. 10: Plot of the log RMSE of $\hat{r}_O^s$ against k under Case 1 (black), Case 2 (red), Case 3 (blue), Case 4 (green), and Case 5 (yellow).

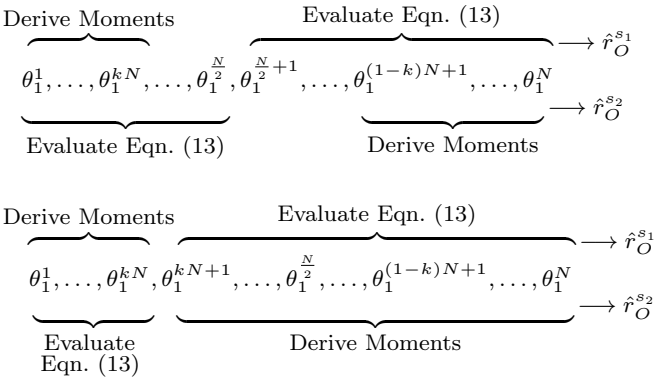

Fig. 11: A diagram to summarise and distinguish between the two cross splitting approaches (estimates computed by $\hat{r}_O^{cs} = \frac{1}{2}(\hat{r}_O^{s_1} + \hat{r}_O^{s_2})$), where the approach by Wang et al. (2019) is presented at the top panel and our suggested approach at the bottom.

4). Wang et al. (2019) suggested a sub-sampling strategy, which makes full use of the entire set of sample from p_1 to ensure statistical efficiency. In particular, a proportion, k, of the samples from p_1 up to the first half (where they have $0 < k \leq 1/2$), is used to compute sample moments, while Equation (13) is evaluated at the remaining half of the samples, $\{\theta_1^{(N/2+1):N}\}$, producing $\hat{r}_O^{s_1}$. The above procedure is then repeated in the reverse direction, where sample moments are derived from a portion of the second half of the sample (of proportion k), $\{\theta_1^{((1-k)N+1):N}\}$, while Equation (13) is evaluated at the first half of the samples, producing $\hat{r}_O^{s_2}$. The cross-splitting optimal bridge sampling estimate, $\hat{r}_O^{cs}$, is then formed by averaging between the two, i.e. $\hat{r}_O^{cs} = \frac{1}{2}(\hat{r}_O^{s_1} + \hat{r}_O^{s_2})$. Given that Equation (13) is evaluated at half of the samples from p_1 (i.e. N_1 is fixed at $N/2$) each time, it is obvious that the larger the k the better the estimate since a larger k ensures a better estimation of the underlying moments (which increases the overlap between p_1 and p_2). Moreover, when $k < 1/2$ is implemented, part of the samples from p_1 is not involved during the individual calculation of $\hat{r}_O^{s_i} (i = 1, 2)$. Even though this is no longer an issue when they are averaged to form $\hat{r}_O^{cs}$, the remaining samples from p_1 could have been better utilized in the calculation of each of the $\hat{r}_O^{s_i}$.

Therefore, we propose to use a proportion, k, of the samples from p_1 to compute the sample moments, while Equation (13) is evaluated at all of the remaining samples, yielding $\hat{r}_O^{s_1}$. Then, the above procedure is repeated in the reverse direction while maintaining the partitioning of the samples, yielding $\hat{r}_O^{s_2}$. See Figure 11 for a graphical summary of our approach. That way, we ensure that the samples from p_1 are fully utilized during the cross computation, while avoiding samples from being used twice for the evaluation of the estimator. It is then of interest to investigate the impact of k on the efficiency of $\hat{r}_O^{cs}$, where it is sufficient to consider the range $0 < k \leq 1/2$ (the behaviour for $1/2 \leq k < 1$ would be identical). Note that our approach is the same with that of Wang et al. (2019) when $k = 1/2$. We focus on our proposed approach for the remaining part of this paper.

It is not possible to completely nullify the correlation between $\hat{r}_O^{s_1}$ and $\hat{r}_O^{s_2}$ since the samples used to construct p_2 (through moment estimation) technically still appear in the estimator in the form of the parameters for p_2. Wang et al. (2019) claimed that the correlation between $\hat{r}_O^{s_1}$ and $\hat{r}_O^{s_2}$ is empirically found to be rather small, implying that the cross-splitting approach is capable of reducing the variance of $\hat{r}_O^{s_i}$ $(i = 1, 2)$ by almost half. We demonstrate that this is not always true, especially when the two densities involved have similar functional forms. Returning to our case study in Section 4.1, the correlation between $\hat{r}_O^{s_1}$ and $\hat{r}_O^{s_2}$ is computed for $k = 1/10, 1/5, 1/3, 1/2$, with varying sample sizes $N \in \{100, 200, \ldots, 10000\}$ (see Figure 12). The estimated correlations seem to have converged to the true underlying values for all N (subject to fluctuations), where the converged values (mostly non-zero) vary across k such that the larger the k the larger the correlation. For example, the correlation between $\hat{r}_O^{s_1}$ and $\hat{r}_O^{s_2}$ is around 0.33 at $k = 1/2$ (as implemented by Wang et al. 2019), which is most likely due to p_1 and p_2 both being normal densities. However, $\hat{r}_O^{cs}$ with $k = 1/2$ also appears to result in the lowest RMSE despite having the highest correlation.

To pin down situations where the magnitude of the correlation between $\hat{r}_O^{s_1}$ and $\hat{r}_O^{s_2}$ is negligible, we repeat the above study with various p_1 and p_2 under four different scenarios:

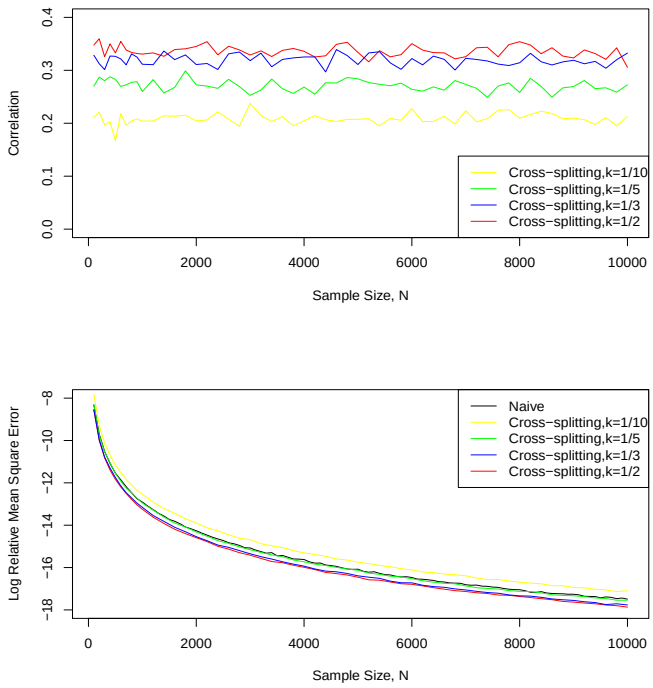

Fig. 12: Plot of the correlation between $\hat{r}_O^{s_1}$ and $\hat{r}_O^{s_2}$ against N for various k (top panel), and plot of the log RMSE of the resulting $\hat{r}_O^{cs}$ (bottom panel), when p_1 is the density of $N(0,1)$ and p_2 is the moments-matched normal distribution.

1. Case 1: p_1 is the density of Exp(λ_1), where $\lambda_1 = 0.5$ and p_2 is the density of Exp(λ_2), where λ_2 is derived from the samples from p_1 by method of moments;
2. Case 2: p_1 is the density of t_3, while p_2 is the density of the sample moments-matched normal distribution;
3. Case 3: p_1 is the density of a standard normal distribution, while p_2 is the density of a non-standardized t_3, constructed using samples from p_1 using method of moments;
4. Case 4: p_1 is the density of t_3, while p_2 is the density of a non-standardized t_3, constructed to possess moments that equate to those of the sample moments from p_1.
   [Note that in cases where p_1 is multi-modal, it may be more efficient to use a Gaussian (mixture) approximation in conjunction with Warp-U transformation (Wang et al. 2019), but will not be considered here as the influence of relative tail behaviours is of interest.]

Knowing that the correlation does not vary against N, we fix $N = 1000$ as an illustration and the results are summarized in Table 1. As consistent with Wang et al. (2019), there are certain situations where the correlation is close to being negligible, i.e. cases 2 and 3 here for all k. This implies that the cross splitting procedure is able to yield $\hat{r}_O^{cs}$ with half the variance of $\hat{r}_O^{s_i}$. We hypothesize that this is because p_1 and p_2 have different tail weights due to them being densities of different functional forms. In case 4, the correlations are slightly larger than zero, again, likely due to p_1 and p_2 both being densities of t_3. It is also evident from Table 1 that $\hat{r}_O^{cs}$ with $k = 1/2$ is consistently outperforming other k across all four cases considered (which is in agreement with Wang et al. 2019), despite consistently having the highest correlation between $\hat{r}_O^{s_1}$ and $\hat{r}_O^{s_2}$. This could be linked to the result in Section 4.1, where we note that $\hat{r}_O^{cs}$ for a given k is essentially the average of $\hat{r}_O^{s}$ given k and $1-k$. From Table 1, we learn that averaging between both $\hat{r}_O^{s}$ with $k = 0.50$ is better than averaging between $k = 1/10$ (the worst) and $k = 9/10$ (the best). Therefore, this suggests that the cross-splitting approach not only reduces the large standard error caused by applying the splitting approach, but also negates the need to determine the optimal splitting proportion (as studied in Section 4.2), given that $\hat{r}_O^{cs}$ with $k = 1/2$ is the most optimal irrespective of the nature of p_1 and p_2.

Even though $\hat{r}_O^{cs}$ with $k = 1/2$ proved to be the best choice from our experiments, technically, $\hat{r}_O^{cs}$ with any $0 < k \leq 1/2$ is guaranteed to further improve the bridge sampling estimator in the sense of RMSE at a cost of a slightly higher computational effort relative to the splitting approach. This is particularly beneficial in scenarios where the main concern is to optimize the efficiency of bridge sampling estimates given a fixed number of samples from p_1, since no sample is wasted purely for the construction of p_2 and the resulting estimate is unbiased with low standard error.

Table 1: The correlation between $\hat{r}_O^{s_1}$ and $\hat{r}_O^{s_2}$ for various cases considered, with the corresponding log RMSE shown in parentheses.

	Case 1	Case 2	Case 3	Case 4
$k = 1/10$	0.1890	0.0199	-0.0047	0.0623
	(-13.2204)	(-9.3313)	(-9.4601)	(-11.2439)
$k = 1/5$	0.2696	0.0073	0.0072	0.0726
	(-13.6600)	(-9.4672)	(-9.5395)	(-11.5274)
$k = 1/3$	0.3005	-0.0058	0.0122	0.0740
	(-13.8829)	(-9.5771)	(-9.5711)	(-11.6442)
$k = 1/2$	0.3364	-0.0044	0.0260	0.0785
	(-13.9197)	(-9.5984)	(-9.5693)	(-11.7372)

6 Extending the Idea of Cross-Splitting

The cross-splitting approach can be extended by borrowing the idea from the n-fold cross-validation (Kohavi 1995). In particular, the original sample is partitioned into n subsets, where a single subsample is retained for moments estimation, while the remaining subsamples are then used to evaluate the optimal bridge sampling estimator, together with using a large N_2. This process is then repeated n times, with each of the subsamples used exactly once for moments estimation, producing $\hat{r}_O^{s_1}, \ldots, \hat{r}_O^{s_n}$. The resulting n-fold cross-splitting estimate, $\hat{r}_O^{ncs}$, can then be formed by taking the average of the n estimates produced in each individual repetition, i.e. $\hat{r}_O^{ncs} = \frac{\sum_{i=1}^{n} \hat{r}_O^{s_i}}{n}$. A graphical example when $n = 3$ is provided in Figure 13. Technically speaking, the n-fold cross splitting approach is anticipated to further improve the bridge sampling estimate, the justification of which can be formulated based on our previous findings. Firstly, partitioning the samples from p_1 into multiple smaller subsets before evaluating the estimator corresponds to the splitting approach with small k, which was discovered to be optimal (since typically p_1 and p_2 are of different functional forms) in Section 4.2. In particular, repeatedly evaluating the optimal bridge sampling estimator separately for the individual subsamples and then averaging (as in $\hat{r}_O^{ncs}$) is a way to evaluate the estimator at a larger number of samples while avoiding using a larger N_2 relative to N_1 in a single computation. Secondly, averaging across all the estimates produced from each of the subsets corresponds to a multiple application of the cross-splitting approach. The expectation is that the standard error of the resulting bridge sampling estimate is further reduced due to evaluating Equation (13) at more posterior samples (around $n - 1$ times more evaluations than the cross-splitting approach). Indeed, this implies that the n-fold cross-splitting approach is computationally more costly to execute than the cross-splitting approach because the bridge sampler is required to be evaluated at more posterior samples. However, the additional computation is close to negligible since posterior samples are evaluated at the Gaussian approximation densities, which is cheap to carry out. The combination of these two features should be capable of improving the efficiency of $\hat{r}_O^{ncs}$.

As an illustration, we set $n = 3$ and consider a similar simulation study as in Section 5, where p_1 is now the density of t_3, while p_2 is the moments-matched normal density, with moments computed from different subset of the samples from p_1. The algorithm for computing the 3-fold cross-splitting estimate is in accordance with that depicted in Figure 13. Computationally speaking, $\hat{r}_O^{ncs}$ is surely going to yield better estimates than $\hat{r}_O^{s_i}$

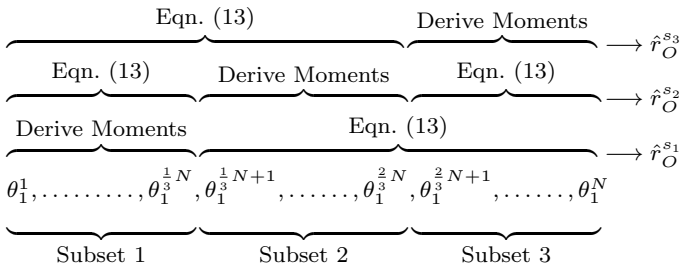

Fig. 13: A diagram to summarise the 3-fold cross-splitting approach (estimates computed by $\hat{r}_O^{ncs} = \frac{1}{3}(\hat{r}_O^{s_1} + \hat{r}_O^{s_2} + \hat{r}_O^{s_3})$).

(for any $i = 1, \ldots, n$) if we only simulate $N_2 = N$ samples from p_2 for each of the repetitions, since this implies that we effectively have more samples (three times when $n = 3$) from p_2 to work with collectively. To demonstrate that the improvement of $\hat{r}_O^{ncs}$ is not entirely due to using a larger N_2 and to ensure similar computational costs are incurred, we fix the total number of samples generated from the Gaussian approximation density to be $3N$ for a fair comparison. Specifically, we simulate $N_2 = 3N$ samples from p_2 and p_2^s respectively for the naive approach and the splitting approach with $k = 1/3$. On the other hand, for the cross-splitting approach with $k = 1/3$, we simulate $N_2 = 1.5N$ samples from each of $p_2^{s_1}$ and $p_2^{s_2}$ (so in total we have $3N$ samples from the Gaussian approximation densities) for the purpose of comparison. Finally, we also include the cross-splitting approach with $k = 1/2$ (the optimal k from Section 5), all of which are depicted in Figure 14. It is evident that $\hat{r}_O^{ncs}$ for $n = 3$ is the best estimate out of its counterparts, with an apparent outperformance observed over the naive and splitting approaches. The improvement of $\hat{r}_O^{ncs}$ for $n = 3$ over $\hat{r}_O^{cs}$ with $k = 1/3$ and $k = 1/2$ is not substantial, but is still discernible. Intuitively, this is because the reduction in standard error (since Equation (13) is evaluated at more posterior samples in total) is less apparent due to the correlations among $\hat{r}_O^{s_i} (i = 1, \ldots, n)$. However, it can still be deduced that the gain in efficiency of the 3-fold cross-splitting approach over other approaches is recognizable, and is not entirely due to evaluating the estimator at a larger N_2. Nevertheless, it is ill-advised to use a very large n in high-dimensional problems. This is because the samples are segregated into subsets which contain limited amount of samples, the moments derived from each subset would then be inadequate to summarize p_1 (each p_2 constructed has small overlap with p_1), resulting in poor estimates.

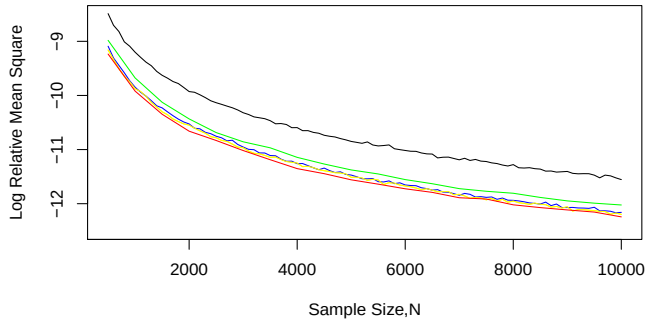

Fig. 14: Plot of the log RMSE of $\hat{r}_O^{ncs}$ for $n = 3$ (red), against N. Also included are the log RMSE of $\hat{r}_O$ (black), $\hat{r}_O^s$ with $k = 1/3$ (green), and $\hat{r}_O^{cs}$ with $k = 1/2$ (yellow) as well as with $k = 1/3$ (blue).

7 Implication of Our Results on Bayesian Computations

As mentioned previously, marginal likelihoods and Bayes factors are quantities of interest in Bayesian model selection. One can use the bridge sampling approach to estimate the marginal likelihood by setting $q_1(\theta) = f_M(X|\theta)f_M(\theta)$, with $q_2(\theta) = p_2(\theta)$ conveniently chosen as the moments-matched normal distribution. Assuming that a fixed number of samples, $\{\theta_1^{1:N}\}$, is available from the posterior distribution, the aim is to formulate an algorithm of computing bridge sampling estimates with maximal statistical efficiency. We have demonstrated that naively evaluating the bridge sampling estimator leads to biased estimates. Here, we describe some plausible strategies of achieving the aim based on our preceding studies.

Our recommended algorithm for efficiently applying the bridge sampling approach is as follows:

1. Partition the posterior samples into n (the choice of which will be commented later) subsets:
$$\overbrace{\theta_1^1,\ldots,\theta_1^{\frac{N}{n}}}^{\text{Subset 1}},\overbrace{\theta_1^{\frac{N}{n}+1},\ldots\theta_1^{2\frac{N}{n}}}^{\text{Subset 2}},\ldots\ldots,\overbrace{\theta_1^{\frac{(n-1)N}{n}+1},\ldots,\theta_1^N}^{\text{Subset } n}.$$
2. Set $m = 1$.
3. Compute the mean and covariance of subset m of the posterior samples to form the moments-matched normal distribution, i.e. $q_2^{s_m} = p_2^{s_m}$ is the density of $N(\bar{\theta}_1^{s_m}, (\hat{\sigma}_1^{s_m})^2)$, where $\bar{\theta}_1^{s_m} = \frac{\sum_{i=(m-1)N/n+1}^{mN/n}\theta_1^i}{N/n}$ and $(\hat{\sigma}_1^{s_m})^2 = \frac{\sum_{i=(m-1)N/n+1}^{mN/n}(\theta_1^i-\bar{\theta}_1^{s_m})^2}{N/n-1}$.
4. Generate a sample of size N_2 from $N(\bar{\theta}_1^{s_m}, (\hat{\sigma}_1^{s_m})^2)$ to form $\{\theta_2^{1:N_2}\}$, where N_2 is chosen to be moderately large.
5. Using the remaining subsets and $\{\theta_2^{1:N_2}\}$, evaluate the optimal bridge sampling estimator. Specifically, Equation (13) is modified to form
$$\hat{r}_O^{s_m(t+1)} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\left[\frac{l^{s_m}(\theta_2^i)}{N_1 l^{s_m}(\theta_2^i)+N_2\hat{r}_O^{s_m(t)}}\right]}{\frac{1}{N_1}\sum_{j\neq(m-1)N/n+1,\ldots,mN/n}\left[\frac{1}{N_1 l^{s_m}(\theta_1^j)+N_2\hat{r}_O^{s_m(t)}}\right]},$$
where $N_1 = N - N/n$, $N_2 = N_2$, $l^{s_m}(\theta) = \frac{f_M(X|\theta)f_M(\theta)}{q_2^{s_m}(\theta)}$, and is iterated until convergence to yield $\hat{r}_O^{s_m}$.
6. Set $m = m + 1$ and repeat steps 3-5 until $m = n$.
7. Calculate the estimate $\hat{r}_O^{ncs} = \frac{\sum_{m=1}^n \hat{r}_O^{s_m}}{n}$.

It is desirable to use a large n. However, users should also be warned that partitioning the posterior samples into more subsets may not be ideal when the number of parameters is large, as more samples are required to estimate their moments (particularly when serially correlated MCMC samples are used). A general strategy is to choose n such that it is reasonable in the context of the problem, properly considering the sample size relative to the dimensionality of the problem. A rule of thumb to check if the n chosen is reasonable is to repeat the above algorithm for $n - 2$ or even $n/2$, a large discrepancy in their values indicates that smaller n should be used. If it is believed that the posterior samples are not sufficiently long, then it is still advisable to use $n = 2$, corresponding to the cross-splitting approach with $k = 1/2$, which has been shown in Section 5 to produce estimates with good properties and is the most optimal among various k.

The use of a large N_2 is recommended primarily for the purpose of reducing the overall standard error (as demonstrated in Section 3) and the ease with which samples from normal distributions can be generated. However, using an astronomically large N_2 has the inherent risk of producing estimates that behave similarly to the importance sampling estimates, meaning that we do not benefit as much from the additional computation devoted, particularly when the posterior density has a heavy tail or is costly to evaluate. In some Bayesian applications though, especially when data size is large, this is not a major problem since the posterior distribution is approximately normal (Gelman et al. 1995, Chapter 13). Also note that the mixture coefficients in Equation (8) are in fact N_1 and rN_2, where r (the marginal likelihood here) also plays a role in determining the estimate's behaviour. So far, we have only considered $r = 1$ in our simulations for simplicity so the values of N_1 and N_2 directly reflect the mixture proportion used. In typical Bayesian applications, r is numerically small, so using $N_1 = N_2$ does not imply that an

equal mixture proportion is assigned as in previously, rather, rN_2 would be considerably smaller than N_1 (implicitly favouring the biased reciprocal importance sampling behaviour). Using a larger N_2 could potentially counterbalance this effect, even though this can alternatively be circumvented by multiplying r with a large constant (to be readjusted once $\hat{r}_O^{ncs}$ is computed). The sequence with which the posterior subsets are used (for moments estimation and evaluation of the estimators) could also be permutated if one prefers to further minimize the effect of the serial correlations of MCMC samples.

7.1 A Practical Example

We illustrate our computational strategies for marginal likelihoods on the Natural Selection Study Data (see for example Sinharay and Stern 2005, Overstall and Forster 2010). The data contain the survival indicator (0: died, 1: survived), birth-weight (in grams) and clutch (family) membership of $n = 244$ newborn turtles from $G = 31$ different clutches. The scientific interest is to determine the effects (if any) of birth-weight and clutch membership on the survival of newborn turtles (response variable). Let y_{ij} and x_{ij} denote respectively the survival indicator and birth-weight of the jth turtle in the ith clutch, $i = 1, 2, \ldots, G = 31$, $j = 1, 2, \ldots, n_i$, where n_i is the number of turtles in clutch i with $\sum_{i=1}^{31} n_i = n$. We fit the probit regression model, i.e. $y_{ij}|p_{ij} \sim \text{Bernoulli}(p_{ij})$, where $p_{ij} = \Phi(\eta_{ij})$, and consider two models with latent variables as follows:

- Model A: $\eta_{ij} = \beta_0 + \beta_1 x_{ij} + u_i$, where $u_i \overset{\text{iid}}{\sim} (0, \sigma^2)$;
- Model B: $\eta_{ij} = \beta_0 + (\beta_1 + v_i)x_{ij} + u_i$, where $(u_i, v_i)^\top \overset{\text{iid}}{\sim} (\mathbf{0}, \boldsymbol{D})$.

Here, β_0 and β_1 are regression parameters, while u_i's and v_i's are random effects for clutch membership. The default priors as proposed by Overstall and Forster (2010) are adopted, i.e.

$$\begin{cases} (\beta_0, \beta_1)^\top \sim N(\mathbf{0}, \frac{n\pi}{2}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}), \\ \sigma^{-2} \sim \text{Gamma}(\frac{1}{2}, \frac{\pi}{4}), \\ \boldsymbol{D} \sim \text{Inverse Wishart}(2, G\pi \times \left[\sum_{i=1}^{G} \frac{1}{n_i} \boldsymbol{Z}_i^\top \boldsymbol{Z}_i\right]^{-1}), \end{cases}$$

where $\boldsymbol{X} = (\boldsymbol{Z}_1^\top \ \ldots \ \boldsymbol{Z}_G^\top)^\top$ and $\boldsymbol{Z}_i^\top = \begin{pmatrix} 1 & \ldots & 1 \\ x_{i1} & \ldots & x_{in_i} \end{pmatrix}$ for $i = 1, \ldots, G$.

To obtain posterior samples from both models, we run OpenBUGS (an open source variant of WinBUGS (Lunn et al. 2000)) from the statistical software R using the package "R2OpenBUGS" (Sturtz et al. 2010). In particular, the supplementary R codes provided by Overstall and Forster (2010) are modified as appropriate to produce a total of 200000 posterior samples for both models, after a burn-in phase of 20000 iterations. The posterior samples generated are then used to compute the "true" log marginal likelihoods (assuming that the corresponding MCMC schemes have achieved convergence) of Models A and B, given respectively as -156.70 and -158.44. For more technical details on the computation of marginal likelihoods of both models, readers are referred to Sinharay and Stern (2005) and Overstall and Forster (2010).

Posterior samples of sizes 1000 are then used to estimate the marginal likelihoods under each of the models to illustrate the efficiencies of various bridge sampling computational strategies under situation with "limited" posterior samples. Specifically, we compare the following strategies:

1. Naive: Derive sample moments from the entire posterior samples to form the Gaussian approximation density p_2, and generate 3000 samples from p_2. Then, evaluate the bridge sampler at the entire posterior samples and the samples from p_2 to form $\hat{r}_O$, where $N_1 = 1000$ and $N_2 = 3000$.
2. Splitting: Derive sample moments from the first half of posterior samples to form the Gaussian approximation density p_2^s, and generate 3000 samples from p_2^s. Then, evaluate the bridge sampler at the second half of the posterior samples and the samples from p_2^s to form $\hat{r}_O^s$, where $N_1 = 500$ and $N_2 = 3000$.
3. Cross-splitting: Derive sample moments from the first half of posterior samples to form the Gaussian approximation density $p_2^{s_1}$, and generate 1500 samples from $p_2^{s_1}$. Then, evaluate the bridge sampler at the second half of posterior samples and the samples from $p_2^{s_1}$ to form $\hat{r}_O^{s_1}$, where $N_1 = 500$ and $N_2 = 1500$. Repeat in the reverse order to form $\hat{r}_O^{s_2}$, from which we obtain $\hat{r}_O^{cs} = \frac{1}{2}(\hat{r}_O^{s_1} + \hat{r}_O^{s_2})$.
4. n-fold cross-splitting: As an illustration, we let $n = 3$. Divide the posterior samples into three subsets of sizes 333, 333, and 334 respectively. Derive sample moments from the first subset to form the Gaussian approximation density $p_2^{s_1}$, and generate 1000 samples from $p_2^{s_1}$. Then, evaluate the bridge sampler at the remaining posterior samples and the samples from $p_2^{s_1}$ to form $\hat{r}_O^{s_1}$, where $N_1 = 666$ and $N_2 = 1000$. Repeat the procedure for the second and third subsets to form $\hat{r}_O^{s_2}$ and $\hat{r}_O^{s_3}$, from which we obtain $\hat{r}_O^{ncs} = \frac{1}{3}(\hat{r}_O^{s_1} + \hat{r}_O^{s_2} + \hat{r}_O^{s_3})$.

Note that we fix the total number of samples from the Gaussian approximation densities to be 3000 for all of the above strategies to ensure approximately similar computational costs are incurred, as mentioned in Section 6. Each computation is replicated 100 times to ob-

tain empirical properties of the estimates. Table 2 shows the estimated log RMSE of the competing strategies under Models A and B, with relative biases (see Section 2.1) included in parentheses as percentages.

Table 2: The log RMSE for various bridge sampling computational strategies under Models A and B, with the corresponding relative biases (in percentages) shown in parentheses.

Strategies	Model A	Model B
Naive	−3.48	−0.50
	(−17.52%)	(−77.68%)
Splitting	−6.89	−2.63
	(0.21%)	(2.12%)
Cross-splitting	−7.18	−2.80
	(0.26%)	(1.81%)
3-fold cross-splitting	−7.22	−2.86
	(0.35%)	(3.13%)

According to Table 2, it is evident that the naive approach underestimates the marginal likelihoods by a considerable margin, resulting in the largest RMSE relative to other strategies for both models. Crucially, the magnitude of the bias increases for the higher-dimensional Model B. The splitting, cross-splitting, and 3-fold cross splitting approaches seemingly produce unbiased estimates of marginal likelihoods (subject to Monte Carlo error). The 3-fold cross-splitting approach also appears to be the best strategy in terms of having the lowest RMSE under both models, although the improvement over the cross-splitting approach is not substantial. Overall, the result from this practical example is as expected from our preceding investigations, that all three splitting approaches manage to mitigate the negative bias of the naive approach regardless of the dimension of the problem. Among the three splitting approaches, the n-fold cross-splitting approach is the best computational strategy, with a slight edge over the cross-splitting approach.

8 Conclusion

This paper investigates the bridge sampling estimator developed by Meng and Wong (1996) for estimating normalizing constants. Specifically, we highlight its potential in Bayesian computation, where marginal likelihoods/Bayes factor are core model selection quantities. First, it was illustrated that naively applying the bridge sampling estimator leads to biased estimates. Theoretical calculations are then presented to identify the sources of bias. We classify the bias of bridge sampling estimator into two categories: one originates from the distance between p_1 and p_2, while another is induced from the correlation due to the moments-matching procedure, where we proceeded to focus on the latter. The effect of sample size allocation was discovered to have an impact on both the bias and standard error of bridge sampling estimator. We demonstrated how the splitting approach (partitioning posterior samples for moments estimation and evaluation of estimator separately) eliminates the correlation induced bias at the expense of a larger standard error. The optimal way of partitioning the posterior samples for splitting (controlled by k) was also examined and found to be dependent on the nature of p_1 and p_2. Next, the cross-splitting approach was described as a method capable of lowering the larger standard error due to splitting. We also shed light on the fact that it is not crucial to determine the optimal k for splitting when the cross-splitting approach is implemented because the influence of k is diminished through averaging. The cross-splitting approach was then extended to form the n-fold cross-splitting to further improve the bridge sampling estimator. Finally, we presented an algorithm for efficiently implementing the bridge sampling approach to estimate marginal likelihoods in a Bayesian context based on our findings (where posterior samples are expensive to generate), and concluded with a practical illustration.

Acknowledgements

The authors greatly appreciate the comments from the anonymous referees involved, who helped in improving the quality of the paper.

Appendix: Supplementary Material

Supplementary material related to this article can be found online at **[to be included if accepted]**.
R software: The file containing codes to perform the bridge sampling procedures described in the article is in "optimal_bridge_MCMC.R".

References

Augustine Kong, A., Liu, J., Wong, W.H.: Sequential Imputations and Bayesian Missing Data Problems. Journal of the American Statistical Association **89**(425), 278–288 (1994)

Bartolucci, F., Scaccia, L., Mira, A.: Efficient Bayes factor estimation from the reversible jump output. Biometrika, **93**(1), 41–52 (2006)

Bennett, C.H.: Efficient estimation of free energy differences from Monte Carlo data. Journal of Computational Physics **22**(2), 245–268 (1976)

Carlin, B.P., Chib, S.: Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society, Ser. B, **57**, 473–484 (1995)

Carlin, B.P., Louis, T.A.: Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall CRC Press, 2nd ed. (2000)

Chen, M.H., Shao, Q.M., Ibrahim, J.G.: Monte Carlo Methods in Bayesian Computation. Springer-Verlag New York, Inc (2000)

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

Frenkel, D.: Free-Energy Computation and First-Order Phase Transitions. In Molecular-Dynamics Simulation of Statistical Mechanical Systems, edited by G. Ciccotti and W. G. Hoover, (North-Holland, Amsterdam) pp. 151–188 (1986)

Frühwirth-Schnatter, S.: Estimating marginal likelihoods for mixture and markov switching models using bridge sampling techniques. The Econometrics Journal **7**(1), 143–167 (2004)

Gelfand, A.E., Dey, D.K.: Bayesian Model Choice: Asymptotic and Exact Calculations. Journal of the Royal Statistical Society. Series B (Mehodological), **56**(3), 501–514 (1994)

Gelman, A., Meng, X.L.: Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical Science, **13**(2), 163–185 (1998)

Gelman, A., Rubin, D.B., Carlin, J.B., Stern, H.S.: Bayesian Data Analysis. Chapman and Hall Ltd, 1st ed. (1995)

Geweke, J.: Bayesian inference in econometric models using monte carlo integration. Econometrica **57**(6), 1317–1339 (1989)

Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, **82**, 711–732 (1995)

Gronau, Q.F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D.S., Forster, J.J., Wagenmakers, E.J., Steingroever, H.: A tutorial on bridge sampling. Journal of Mathematical Psychology **81**, 80 – 97 (2017a)

Gronau, Q.F., Singmann, H., Wagenmakers, E.J.: Bridgesampling: Bridge sampling for marginal likelihoods and Bayes factors (2017b). https://github.com/quentingronau/bridgesampling (R package version 0.2-2)

Guy, J.A., Bijak, J., Forster, J.J., Raymer, J., Smith, P.W.F., Wong, J.S.T.: Integrating Uncertainty in Time Series Population Forecasts: An Illustration Using a Simple Projection Model. Demographic Research, **29**(43), 1187–1226 (2013)

Irwin, M., Cox, N., Kong, A.: Sequential imputation for multilocus linkage analysis. Proceedings of the National Academy of Sciences of the United States of America, **91**(24), 11684–11688 (1994)

Jensen, C.S., Kong, A.: Blocking Gibbs Sampling for Linkage Analysis in Large Pedigrees with Many Loops. The American Journal of Human Genetics **65**(3), 885 – 901 (1999)

Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, vol. 2. Wiley Series in Probability and Mathematical Statistics, 2nd ed. (1995)

Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association, **90**(430), 773–795 (1995)

Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, Que., Morgan Kaufmann, Los Altos, CA pp. 1137–1143 (1995)

Kong, A., McCullagh, P., Meng, X.L., Nicolae, D., Tan, Z.: A theory of statistical models for Monte Carlo integration. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **65**(3), 585–604 (2003)

Kullback, S., Leibler, R.A.: On Information and Sufficiency. The Annals of Mathematical Statistics, **22**(1), 79–86 (1951)

Lee, S.Y., Song, X.Y., Lee, J.C.K.: Maximum Likelihood Estimation of Nonlinear Structural Equation Models with Ignorable Missing Data. Journal of Educational and Behavioral Statistics, **28**(2), 111–134 (2003)

Lopes, H.F., West, M.: Bayesian model assessment in factor analysis. Statistica Sinica, **14**(1), 41–67 (2004)

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D.: WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing **10**(2000), 325–337 (2000)

Meng, X.L., Schilling, S.: Fitting Full-Information Item Factor Models and an Empirical Investigation of Bridge Sampling. Journal of the American Statistical Association, **91**(435), 1254–1264 (1996)

Meng, X.L., Wong, W.H.: Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration. Statistica Sinica, **6**(4), 831–860 (1996)

Mira, A., Nicholls, G.: Bridge Estimation of The Probability Density at A Point. Statistica Sinica, **14**(2), 603–612 (2004)

Neal, R.M.: Probabilistic Inference using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, Univ. Toronto (1993)

Neal, R.M.: Contribution to the discussion of "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap" by Michael A. Newton and Adrian E. Raftery. Journal of the Royal Statistical Society. Series B (Mehodological), **56**(1), 41–42 (1994)

Newton, M.A., Raftery, A.E.: Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society. Series B (Mehodological), **56**(1), 3–48 (1994)

Overstall, A.M., Forster, J.J.: Default bayesian model determination methods for generalised linear mixed models. Computational Statistics and Data Analysis, **54**(12), 3269–3288 (2010)

Sinharay, S., Stern, H.S.: An Empirical Comparison of Methods for Computing Bayes Factor in Generalized Linear Mixed Models. Journal of Computational and Graphical Statistics, **14**(2), 415–435 (2005)

Sturtz, S., Ligges, U., Gelman, A.: R2OpenBUGS: A Package for Running OpenBUGS from R. http://cran.r-project.org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS.pdf (2010). Accessed 25 September 2019

Tan, Z.: Calibrated path sampling and stepwise bridge sampling. Journal of Statistical Planning and Inference, **143**(4), 675–690 (2013)

Wang, L., Jones, D.E., Meng, X.L.: Warp bridge sampling: The next generation (2019). arXiv preprint arXiv:1609.07690

Wong, J.S.T.: Bayesian estimation and model comparison for mortality forecasting. Ph.D. thesis, University of Southampton (2017)

Wong, J.S.T., Forster, J.J., Smith, P.W.F.: Bayesian mortality forecasting with overdispersion. Insurance: Mathematics and Economics **83**(2018), 206–221 (2018)