AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR020167

Key Points:
- We introduce a new estimator of Bayesian model evidence to select the best model among candidate ones
- GAME sampling combines MCMC simulation of the model posterior distribution and bridge sampling
- We demonstrate that GAME sampling returns a robust estimate of the Bayesian model evidence under general conditions

Correspondence to:
E. Volpi,
elena.volpi@uniroma3.it

Citation:
Volpi, E., G. Schoups, G. Firmani, and J. A. Vrugt (2017), Sworn testimony of the model evidence: Gaussian Mixture Importance (GAME) sampling, *Water Resour. Res.*, *53*, 6133–6158, doi:10.1002/2016WR020167.

Received 22 NOV 2016
Accepted 19 JUN 2017
Accepted article online 5 JUL 2017
Published online 28 JUL 2017

Sworn testimony of the model evidence: Gaussian Mixture Importance (GAME) sampling

Elena Volpi[1], Gerrit Schoups[2], Giovanni Firmani[1], and Jasper A. Vrugt[3,4]

[1]Department of Engineering, University of Roma Tre, Rome, Italy, [2]Department of Water Management, Delft University of Technology, Netherlands, [3]Department of Civil and Environmental Engineering, University of California, Irvine, California, USA, [4]Department of Earth System Science, University of California, Irvine, California, USA

Abstract What is the "best" model? The answer to this question lies in part in the eyes of the beholder, nevertheless a good model must blend rigorous theory with redeeming qualities such as parsimony and quality of fit. Model selection is used to make inferences, via weighted averaging, from a set of K candidate models, $\mathcal{M}_k$; $k=(1,\ldots,K)$, and help identify which model is most supported by the observed data, $\tilde{\mathbf{Y}}=(\tilde{y}_1,\ldots,\tilde{y}_n)$. Here, we introduce a new and robust estimator of the model evidence, $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$, which acts as normalizing constant in the denominator of Bayes' theorem and provides a single quantitative measure of relative support for each hypothesis that integrates model accuracy, uncertainty, and complexity. However, $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$ is analytically intractable for most practical modeling problems. Our method, coined GAussian Mixture importancE (GAME) sampling, uses bridge sampling of a mixture distribution fitted to samples of the posterior model parameter distribution derived from MCMC simulation. We benchmark the accuracy and reliability of GAME sampling by application to a diverse set of multivariate target distributions (up to 100 dimensions) with known values of $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$ and to hypothesis testing using numerical modeling of the rainfall-runoff transformation of the Leaf River watershed in Mississippi, USA. These case studies demonstrate that GAME sampling provides robust and unbiased estimates of the evidence at a relatively small computational cost outperforming commonly used estimators. The GAME sampler is implemented in the MATLAB package of DREAM and simplifies considerably scientific inquiry through hypothesis testing and model selection.

Plain Language Summary Science is an iterative process for learning and discovery in which competing ideas about how nature works are evaluated against observations. The translation of each hypothesis to a computational model requires specification of system boundaries, inputs and outputs, state variables, physical/behavioral laws, and material properties; this is difficult and subjective, particularly in the face of incomplete knowledge of the governing spatiotemporal processes and insufficient observed data. To guard against the use of an inadequate model, statisticians advise selecting the "best" model among a set of candidate ones where each might be equally plausible and justifiable a priori. Bayesian model selection uses probability theory to select among competing hypotheses; the key variable is the Bayesian model evidence, which provides a single quantitative measure of relative support for each hypothesis that integrates model accuracy, uncertainty, and complexity. Bayesian model selection has not entered into mainstream use in Earth systems modeling due to the lack of general-purpose methods to reliably estimate the evidence. Here, we introduce a new method, called GAussian Mixture importancE (GAME) sampling. We demonstrate GAME power and usefulness for hypothesis testing using benchmark experiments with known target and numerical modeling of the rainfall-runoff transformation of the Leaf River watershed (Mississippi, USA).

1. Introduction and Scope

Science is an iterative process for learning and discovery in which competing ideas about how nature works are evaluated against observations [*Johnson and Omland*, 2004]. Building upon our perceptual understanding of the real-world system, these ideas can emerge as verbal and pictorial hypotheses but must be translated to mathematical equations or computer models before being fit to data. The capabilities of such computer models typically exceed by far traditional paper-and-pencil calculations and can involve simulations on spatial scales of individual atoms to an entire ecosystem, and temporal scales of nanoseconds to

many millions of years. The translation of one or more hypotheses to a computational model requires specification of (among others) relevant system boundaries, inputs and outputs, state (prognostic) variables, physical and behavioral laws (e.g., conservation of mass, momentum and energy), and material properties. Model building is a complex and intuitive process which is heavily influenced by perception, intuition, and prior knowledge on system functioning and reality and colored by mental concepts (state of mind). From a myriad of countless processes and mechanisms, the modeler seeks to elucidate those key principles, laws, and generalizations, which explain the observed data. Their translation to a computational model is difficult and subjective, particularly in the face of incomplete knowledge of the governing spatiotemporal processes and insufficient data on (spatially distributed) system properties and state variables.

To guard against the use of an inadequate model, statisticians advise selecting the "best" model among a set of plausible candidate models chosen and/or construed by the researcher(s). This approach rules out model selection bias and recognizes explicitly ambiguity in the interpretation and analysis of complex natural systems. The ensemble of models, and their associated hypotheses, constitute a finite sample of possible explanations of the data deemed plausible a priori from the extremely, perhaps even unfathomably, large space of alternatives. This can include black-box, conceptual (empirical), and physically based models and involve widely different variables, mathematical functions that define the spatiotemporal relationships between independent variables and the response variable of interest, computational states, fluxes and parameters, and the initial and boundary conditions that govern system behavior and response. Each of the candidate models might be equally plausible and justifiable a priori [see, e.g., *Neuman*, 2003; *Vrugt and Robinson*, 2007; *Ye et al*., 2008; *Clark et al*., 2011]. Model selection then involves the identification of a single best model by evaluating the relative support for each competing hypothesis. The guiding principle at this step is to avoid generating so many models that spurious findings become likely. *Burnham and Anderson* [2002] argues, on philosophical grounds, that $K = 20$ candidate models are more than sufficient.

The definition of the "best" model is somewhat elusive laying in part in the eyes of the beholder. Empirical findings suggest choosing the simplest explanation of the data, as such hypotheses have led to mathematically rigorous and empirically verifiable theories. This parsimony principle is often attributed to William of Ockham (1287–1347), an English Franciscan friar, scholastic philosopher, and theologian, but traceable to the works of philosophers such as Aristotle (384–322 BC) and Ptolemy (circa AD 90 to circa AD 168). Ockham's believe that ". . .Entities must not be multiplied beyond necessity" is commonly referred to in the literature as Occam's razor, and consistent with requirements of falsifiability in the scientific method [*Popper*, 1992]. Indeed, simpler hypotheses (theories) are preferred as they involve fewer assumptions and are therefore easier testable. Thus, a "good" model selection technique must necessarily balance goodness of fit with complexity (often measured in terms of the number of "free" parameters). Indeed, more complex models may be able to better explain the data, but the additional parameters might have little correspondence with the specific processes and behaviors of the system the model is intended to represent [*Schoups et al*., 2008]. A classic example is polynomial wiggle, wherein the use of a higher degree polynomial hardly improves the approximation error, yet introduces oscillations between observations which magnify at the edges of the data interval. This so-called Runge phenomenon cautions against the use of high-order polynomials for interpolation, let alone out-of-sample prediction. Note, in cases where models have similar levels of support from the data, model averaging can be used to negate statistical bias and improve treatment of conceptual model uncertainty [*Neuman*, 2003; *Refsgaard et al*., 2006; *Vrugt and Robinson*, 2007; *Ye et al*., 2008; *Clark et al*., 2011].

The traditional approach to model selection builds on information theory, and uses principles of entropy maximization to determine which hypothesis, $\mathcal{H}_k$, is most supported by available data, $\tilde{\mathbf{Y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)$. Selection is based on an information criterion, I_k, which quantifies the information that is lost if the hypothesis, $\mathcal{H}_k$, is used to explain the data generating process

$$I_k = -2 \ln \{L(\mathcal{H}_k|\tilde{\mathbf{Y}})\} + C, \tag{1}$$

where $L(\mathcal{H}_k|\tilde{\mathbf{Y}})$ denotes the (maximized) likelihood of $\mathcal{H}_k$ and $C > 0$ is a strictly positive scalar which penalizes for the hypothesis' complexity (say, number of "free" parameters, d), and/or n, the length of the data record, $\tilde{\mathbf{Y}}$, and may account for uncertainty of each hypothesis. Thus, the better a hypothesis explains the data, the larger its likelihood, $L(\mathcal{H}_k|\tilde{\mathbf{Y}})$, and the smaller the value of the information criterion, I_k in equation (1). Thus, among several competing hypothesis, the proposition, $\mathcal{H}_k$, with lowest value of I_k receives most

support by the experimental data, $\tilde{\mathbf{Y}}$. The most popular variants of equation (1) are *Akaike*'s information criterion (AIC, *Akaike* [1998]) with $C=2d$, Bayesian information criterion (BIC, *Schwarz* [1978]) with penalty term $C=d\ln(n)$ and the deviance information criterion (DIC, *Spiegelhalter et al*. [2002]) with $C=2\hat{d}$, wherein $\hat{d}$ is an estimate of the effective number of parameters.

Information criteria are routinely used in many different fields of study to support efforts such as hypothesis testing, scientific inquiry, and model selection. This is explained in part by their parsimony and ease of calculation. Nonetheless, most information criteria consider only the likelihood maximum of each hypothesis without recourse to its underlying statistical uncertainty. This approach suffices if the data honor one particular model, but might not be adequate in situations with nearly equivalent support for the competing hypotheses. The DIC is an exception, and considers explicitly the distribution of the likelihood in determination of the penalty term, $\hat{d}$. What is more, information criteria generally do not allow for informative priors (either assessed from field data or through expert elicitation), and can provide contradictory and biased results, particularly for parameter-rich models [*Ye et al*., 2008; *Lu et al*., 2011; *Schöniger et al*., 2014].

Bayesian model selection provides an attractive alternative to information-theoretic selection, and traditional null hypothesis testing via likelihood ratio tests and metrics such as the adjusted R^2 statistic. This Bayesian approach uses probability theory to select among multiple competing hypotheses. The key variable is the marginal likelihood, or Bayesian model evidence, $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$, which is computed separately for each model, $\mathcal{M}_k$, or hypothesis, $\mathcal{H}_k$, where $k=(1,\ldots,K)$ by *averaging* rather than *maximizing* the likelihood function over the prior parameter distribution. This prior distribution plays a key role in Bayesian epistemology and will affect the support a model receives from the observed data. In fact, with an improper prior the model can be made to fit the data arbitrarily poorly, changing fundamentally our opinion about which model should be favored, a phenomenon known as the Jeffreys-Lindley paradox. Information criteria conveniently ignore this antecedent and use only each model's likelihood as proxy for quality of fit. In general, the larger a model's marginal likelihood the more support it receives from the observed data, simply because this data assigns a relatively high probability to the model output. The marginal likelihood encodes a natural preference for simpler and more constrained models, and combats the selection of overly complex and/or overfitted models by information criteria that incorporate only the likelihood maximum. The marginal likelihood can also be shown to approximate the expected out-of-sample prediction error, and thus implicitly performs a split-sample test without actually setting apart data for model evaluation [see *Bishop*, 2006, p. 32]. Furthermore, the Bayesian model evidence also has theoretical connections with AIC and BIC, as shown by *Schöniger et al*. [2014].

Bayesian model selection is a preferred alternative to null hypothesis testing, yet has not entered into mainstream use across fields in Earth systems modeling. The crux is the lack of general-purpose methods available to reliably estimate the model evidence, $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$. Analytic estimates are available for simulation models whose output depends linearly on its parameters, and for conjugate priors as illustrated by *Schöniger et al*. [2014]. Unfortunately, these conditions are too restrictive to be practically useful for most real-world simulation models. Here, we resort to Monte Carlo simulation to estimate numerically the model evidence via multidimensional integration of the posterior model parameter distribution. This task can be cumbersome and computationally demanding, particularly for CPU-intensive system models and high-dimensional posterior parameter distributions that deviate considerably from normality. Nevertheless, Monte Carlo sampling methods [*Hammersley and Handscom*, 1964] can provide better estimates of the model evidence than information criteria [*Kass and Raftery*, 1995; *Lu et al*., 2011; *Schöniger et al*., 2014]. The most basic and straightforward Monte Carlo approach approximates the evidence by the arithmetic mean of the likelihoods of a large sample of parameter vectors drawn randomly from the prior distribution. This approach, albeit relatively simple, is rather inefficient as a large proportion of the samples might exhibit a negligible density and therefore contribute little to the model evidence. More efficient and viable alternatives are importance sampling and Markov Chain Monte Carlo (MCMC) simulation [see, e.g., *Kass and Raftery*, 1995; *Marshall et al*., 2005].

In this paper, we explore, develop, test, benchmark, and contrast different model evidence estimation methods. We introduce a new method, called GAussian Mixture ImportancE (GAME) sampling, which estimates the evidence via multidimensional numerical integration of the posterior parameter distribution using bridge sampling. This method involves two main steps. First, a large collection of samples is generated from the posterior parameter distribution, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}},\mathcal{M}_k)$, using MCMC simulation with the DREAM algorithm [*Vrugt*

et al., 2008, 2009; *Vrugt and ter Braak*, 2011; *Laloy and Vrugt*, 2012; *Vrugt*, 2016] by conditioning separately each candidate model, $\mathcal{M}_k$, on the observed data, $\tilde{\mathbf{Y}}$. Then, a multivariate mixture distribution is fitted to this MCMC collection of posterior samples using likelihood theory. This distribution serves as our catalyst to estimate the marginal likelihood of each competing hypothesis via bridge sampling [*Meng and Wong*, 1996], a natural generalization of importance sampling. We benchmark the robustness, accuracy, and reliability of GAME sampling by application to a diverse set of multivariate target distributions (up to hundred dimensions) with known values of $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$. We then illustrate the power and usefulness of GAME for hypothesis testing using numerical modeling of the rainfall-runoff transformation of the Leaf River watershed in Mississippi, USA. The GAME sampler is implemented in the MATLAB package of DREAM described by *Vrugt* [2016] and simplifies considerably hypothesis testing and model selection.

The remainder of this paper is organized as follows. Section 2 reviews briefly the theory of Bayesian hypotheses testing, while in section 3 we present the bridge sampling framework for model evidence estimation. This is followed in section 4 by a detailed description of the GAME sampler and its numerical implementation in the MATLAB package of DREAM described by *Vrugt* [2016]. In section 5, we present the results of our multivariate benchmark experiments, and section 6 demonstrates the application of GAME sampling to hypothesis testing using multiple different models of the rainfall-discharge relationship of the Leaf River watershed. Finally, we conclude this paper in section 7 with a summary of our main findings.

2. Hypothesis Testing

The first step in hypothesis selection involves articulation of a reasonable set of competing ideas about the structure and functioning of the real-world system of interest. These ideas may be summarized in drawings, maps, tables, papers, reports, and oral presentations, and depend critically upon an investigator's state of knowledge, process understanding, prior facts, training, and experience. Verbal rendition of these ideas leads to a collection of testable hypotheses. Ideally, the collection of hypotheses is construed before data collection and samples exhaustively and systematically the plausible space of explanations for the experimental data.

Hypothesis formulation can be viewed as a preliminary, informal, stage to model building, wherein an investigator expresses verbally their "perceptual" model. Such model is the result of purely sensory perceptions coupled with qualitative and quantitative interpretations of the data. This interpretive process may be strongly influenced by prior concepts, and may be colored by mental concepts. Unfortunately, perceptual models cannot be subjected to formal analysis as this would require symbolic representation. The resulting computational models may not express faithfully the collection of hypotheses due to lack of knowledge, ideas, or imagination about how to express mathematically (among others) the architecture (extent, structure, and spatial variability), governing processes, state variables, and fluxes of each perceptual model.

In surface hydrology, one can easily envisage multiple working hypotheses that may explain watershed behavior and functioning as catchment behavior is complex and controlled by a myriad of interrelated and spatially distributed physical, chemical, and ecological processes. Each candidate hypothesis may be used to explain the watershed data, $\tilde{\mathbf{Y}}=(\tilde{y}_1,\ldots,\tilde{y}_n)$, observed at discrete times $t=(1,\ldots,n)$ as follows

$$\tilde{\mathbf{Y}} \leftarrow \mathcal{M}_k(\boldsymbol{\theta},\tilde{\mathbf{x}}_0,\tilde{\mathbf{B}})+\mathbf{E}, \tag{2}$$

where $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_d)$ is a d-vector of model parameters, $\tilde{\mathbf{x}}_0$ stores the values of the state variables at the start of simulation, $\tilde{\mathbf{B}}$ signifies the control matrix with temporal measurements of the atmospheric forcing variables, and $\mathbf{E}=(e_1,\ldots,e_n)$ is a n-vector of residuals

$$\mathbf{E}(\mathcal{M}_k)=\tilde{\mathbf{Y}}-\mathcal{M}_k(\boldsymbol{\theta},\tilde{\mathbf{x}}_0,\tilde{\mathbf{B}}). \tag{3}$$

Without further loss of generality, we restrict the model parameters to a closed space, Ω, equivalent to a d-dimensional hypercube, $\boldsymbol{\theta}\in\Omega\in\mathbb{R}^d$, called the feasible parameter space.

We conveniently assume herein that the control variables are observed without error, or $\delta(\mathbf{B},\tilde{\mathbf{B}})=0$, and that a spin-up period of T days suffices to ameliorate the effect of state initialization errors on the model output, $\lim_{t\to T}\delta(y_t(\tilde{\mathbf{x}}_0),y_t(\mathbf{x}_0))\to 0$. These two assumptions simplify considerably hypothesis testing as the support each model receives from the data depends only on the values of its parameters, $\boldsymbol{\theta}$. This approach

may bias model selection and favor hypotheses that do not suffer large measurement errors of their respective (set of) control variables. In principle, we can augment the parameter vector of each model with latent variables that rectify errors in the forcing data. This approach is outside the scope of the present work. Interested readers are referred to the BATEA framework of *Kavetski et al*. [2006a, 2006b], the hydrology backward approach of *Kirchner* [2009] and the parameter augmentation method of *Vrugt et al*. [2008, 2009].

Once a set of suitable candidate models has been specified, we must determine appropriate values of their parameters before we can proceed with model selection. We can estimate each model's posterior parameter distribution, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \mathcal{M}_k)$, via Bayes' theorem using the observed data, $\tilde{\mathbf{Y}}$, as follows

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \mathcal{M}_k) = \frac{p(\boldsymbol{\theta}|\mathcal{M}_k)L(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \mathcal{M}_k)}{p(\tilde{\mathbf{Y}}|\mathcal{M}_k)} \propto p(\boldsymbol{\theta}|\mathcal{M}_k)L(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \mathcal{M}_k), \tag{4}$$

where $p(\boldsymbol{\theta}|\mathcal{M}_k)$ denotes the candidate model's prior parameter distribution, $L(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \mathcal{M}_k)$ signifies the likelihood function. The denominator, $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$, is a normalizing constant which ensures that the posterior parameter distribution integrates to unity over $p(\boldsymbol{\theta}|\mathcal{M}_k)$. This so-called model evidence or marginal likelihood can be ignored for parameter inference as the unnormalized posterior density, $p(\boldsymbol{\theta}|\mathcal{M}_k)L(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \mathcal{M}_k)$, suffices to estimate $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \mathcal{M}_k)$. Knowledge of $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$ is strictly necessary for hypothesis testing to select the best model among a set of plausible alternatives.

The marginal likelihood, $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$, is computed separately for each model $\mathcal{M}_k$ by *averaging* rather than *maximizing* the likelihood function over the prior parameter distribution

$$p(\tilde{\mathbf{Y}}|\mathcal{M}_k) = \int_{\Omega} p(\boldsymbol{\theta}|\mathcal{M}_k)L(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \mathcal{M}_k)\mathrm{d}\boldsymbol{\theta} \tag{5}$$

The candidate model with largest marginal likelihood is preferred statistically, as it assigns the highest probability (density) to the experimental data, $\tilde{\mathbf{Y}}$. Marginalization is used to eliminate from equation (5) the effect that different parameters (their number and prior distribution) have on the data likelihood and thus quality of the model fit. Being the average of the likelihood over the prior distribution, the marginal likelihood is largest for models whose likelihood values are high and uniformly distributed across the parameter space, and smallest for models whose parameter space produces consistently low likelihoods. Models with more parameters have larger output spaces and are often better in fitting the data. Consequently, parameter-rich models may have higher peak likelihoods, nevertheless, in order to increase their marginal likelihood, $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$, the area with maximized likelihood must compensate for other areas in the parameter space where the data fit is much poorer and the likelihood is rather low. Simpler models may yield lower peak likelihoods, but provide larger values of the average likelihood, thus being preferred statistically. Thus, marginalization in the Bayesian framework can be viewed as a formalization of Occam's razor: a simpler theory with compact parameter space will have a larger marginal likelihood than a more complicated model, unless the latter is significantly better at explaining the data. The evidence estimates can also serve as weights for the simulations of the different models, as in Bayesian model averaging [*Hoeting et al*., 1999; *Wasserman*, 2000; *Vrugt and Robinson*, 2007].

The next section discusses theory, concepts, and application of Monte Carlo simulation methods to estimate the marginal likelihood, $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$. For notational simplicity, we suppress the dependence of $p(\tilde{\mathbf{Y}}|\mathcal{M}_k)$ on $\mathcal{M}_k$ from now on.

3. Monte Carlo Approximation of the Marginal Likelihood

Estimation of the marginal likelihood is difficult for nonlinear system models as the integral of their posterior parameter distribution is often high-dimensional and without analytic solution. Monte Carlo simulation methods can be used to approximate the evidence of competing models, but their implementation is not necessarily easy and straightforward.

Before we proceed with bridge sampling as general vehicle for evidence estimation, we illustrate in Figure 1a the definition of the marginal likelihood for a standard normal target distribution (red line). We use *m* Monte Carlo samples of this distribution to approximate the integral. These samples will be distributed exactly according to the underlying target distribution, yet their corresponding densities (blue dots)

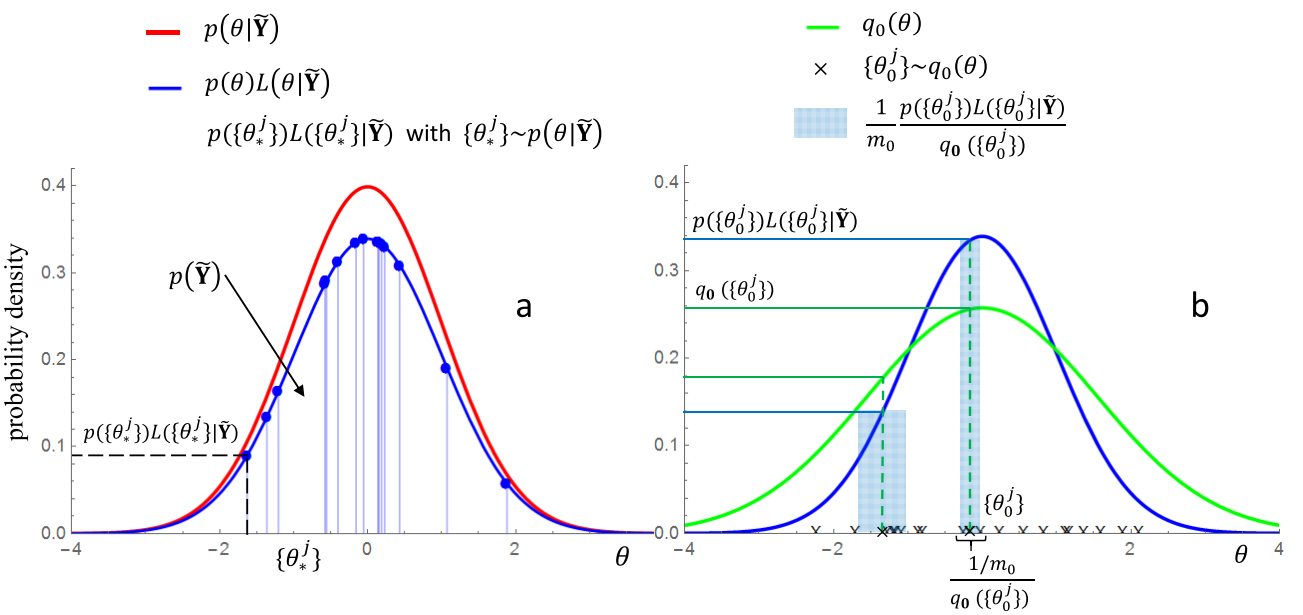

Figure 1. (a) Probability density of standard normal distribution (red curve) and unnormalized target (blue curve). The blue dots signify m different samples, $\{\theta_*^j\} \sim p(\theta|\tilde{\mathbf{Y}})$, where $j=(1,\ldots,m)$. The marginal likelihood, $p(\tilde{\mathbf{Y}})$, is equivalent to the area under the blue curve (see equation (5)). (b) Illustration of importance sampling via equation (10). By drawing m_0 samples at random from the importance distribution, $q_0(\theta)$ (green curve), one can approximate the area/volume under the unnormalized target (blue curve) as mean of the m_0 ratios of the target and importance density. The light-blue rectangles signify the representative area/volume of each sample; this depends on sample density, and is merely for illustrative purposes.

are substantially smaller than their counterparts of the normal distribution (red line). Per equation (5), the marginal likelihood is now equivalent to the area under the blue curve. If we now divide the m unnormalized densities of the posterior samples by this (normalizing) constant, we recover exactly the probability density function of the standard normal target.

The next section discusses bridge sampling as unifying framework to estimate the marginal likelihood (model evidence) from a collection of target samples. We use the standard convention whereby lower case letters are used to denote probability density functions, whereas curly brackets are used to differentiate between random variables and their actual sampled values. Thus, $(\{\boldsymbol{\theta}^1\},\ldots,\{\boldsymbol{\theta}^m\})$ stores a sequence of m different realizations (draws) of the d-vector of model parameters, $\boldsymbol{\theta}$.

3.1. Bridge Sampling for Model Evidence Estimation

Bridge sampling was introduced by *Meng and Wong* [1996] and generalized to thermodynamic integration by *Gelman and Meng* [1998] to estimate the ratio, r, of the normalizing constants, z_0 and z_1, of two unnormalized densities, $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$, with support Ω_0 and Ω_1, respectively. We can write this ratio as follows [*Gelman and Meng*, 1998]

$$r \equiv \frac{z_1}{z_0} = \frac{z_{1/2}/z_0}{z_{1/2}/z_1} = \frac{\mathbb{E}_0\left[q_{1/2}(\boldsymbol{\theta})/q_0(\boldsymbol{\theta})\right]}{\mathbb{E}_1\left[q_{1/2}(\boldsymbol{\theta})/q_1(\boldsymbol{\theta})\right]} \approx \frac{\frac{1}{m_0}\sum_{j=1}^{m_0} q_{1/2}(\{\boldsymbol{\theta}_0^j\})/q_0(\{\boldsymbol{\theta}_0^j\})}{\frac{1}{m_1}\sum_{j=1}^{m_1} q_{1/2}(\{\boldsymbol{\theta}_1^j\})/q_1(\{\boldsymbol{\theta}_1^j\})}, \tag{6}$$

where the numerator and denominator signify the expectations with respect to the density, $p_0(\boldsymbol{\theta})=q_0(\boldsymbol{\theta})/z_0$, and the density, $p_1(\boldsymbol{\theta})=q_1(\boldsymbol{\theta})/z_1$, respectively. The quotient at the right-hand-side of equation (6) expresses the ratio, r, of normalizing constants, z_0 and z_1, as a Monte Carlo approximation using m_0 draws, $\{\Theta_0\}=(\{\boldsymbol{\theta}_0^1\},\ldots,\{\boldsymbol{\theta}_0^{m_0}\})$, from $p_0(\boldsymbol{\theta})$ and m_1 samples, $\{\Theta_1\}=(\{\boldsymbol{\theta}_1^1\},\ldots,\{\boldsymbol{\theta}_1^{m_1}\})$, from $p_1(\boldsymbol{\theta})$. The entity $q_{1/2}(\boldsymbol{\theta})$ is an arbitrary unnormalized density with support $\Omega_0 \cap \Omega_1$ of $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$, respectively, which plays a crucial role in the calculation of r. The subscript "1/2" implies a density that transitions "between" $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$ in the sense of being overlapped by both of them. This density serves as a bridge between $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$, hence the name bridge sampling. The "smoother" and "shorter" the bridge density, $q_{1/2}(\boldsymbol{\theta})$, transitions between $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$ the better it is [*Meng and Wong*, 1996].

To use bridge sampling for model evidence estimation, we take $q_1(\boldsymbol{\theta})$ to be the unnormalized density of the (posterior) target distribution, that is, $q_1(\boldsymbol{\theta})=p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, so that $p_1(\boldsymbol{\theta})$ signifies the normalized density function and $z_1=Z$ is the model evidence, $p(\tilde{\mathbf{Y}})$. In addition, we use for $q_0(\boldsymbol{\theta})$ a normalized density, hence $z_0=1$, and thus the bridge sampling expression for r returns directly the model evidence, Z. In the following, we will substitute Z for r. Given m_1 samples of the posterior distribution, we can now compute the model evidence with equation (6) for different "sensible" choices of $q_0(\boldsymbol{\theta})$ and $q_{1/2}(\boldsymbol{\theta})$. In general, $q_0(\boldsymbol{\theta})$ should satisfy two requirements to be of practical use. First, the density of $q_0(\boldsymbol{\theta})$ should be easy to compute, and, second, the distribution of $q_0(\boldsymbol{\theta})$ should be easy to sample from. Another desirable, but not strictly necessary property of $q_0(\boldsymbol{\theta})$ is, that, it approximates closely the target distribution of interest. Then, the Monte Carlo approximation will be most robust and efficient. In this paper, we evaluate two different choices for $q_0(\boldsymbol{\theta})$, namely (i) the prior distribution, $p(\boldsymbol{\theta})$, and (ii) the posterior distribution, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, approximated with a mixture of normal variates (see section 3.2).

In the remainder of this section, we elaborate on the choice of $q_{1/2}(\boldsymbol{\theta})$, and demonstrate how various commonly used sampling methods originate as special and limiting cases of bridge sampling by setting $q_{1/2}(\boldsymbol{\theta})$ equal to either $q_0(\boldsymbol{\theta})$ or $q_1(\boldsymbol{\theta})$. Last, we generalize the distribution of $q_{1/2}(\boldsymbol{\theta})$ by using its density "in between" $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$.

3.1.1. Methods that Set $q_{1/2}(\boldsymbol{\theta})$ Equal to Either $q_0(\boldsymbol{\theta})$ or $q_1(\boldsymbol{\theta})$

We first study what happens if we set $q_{1/2}(\boldsymbol{\theta})=q_0(\boldsymbol{\theta})$. This reduces the numerator in equation (6) to unity, and, with $r=Z$ this gives

$$Z=1/\mathbb{E}_1\left[\frac{q_0(\boldsymbol{\theta})}{q_1(\boldsymbol{\theta})}\right]=1/\mathbb{E}_1\left[\frac{q_0(\boldsymbol{\theta})}{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}\right]\approx\left[\frac{1}{m_1}\sum_{j=1}^{m_1}\frac{q_0(\{\boldsymbol{\theta}_1^j\})}{p(\{\boldsymbol{\theta}_1^j\})L(\{\boldsymbol{\theta}_1^j\})}\right]^{-1} \tag{7}$$

An advantage of this choice is that we do not need any samples from $q_0(\boldsymbol{\theta})$ as all inferences are made using m_1 realizations, $\{\Theta_1\}$, of the target distribution derived via Monte Carlo simulation. If both $q_0(\boldsymbol{\theta})$ and $q_{1/2}(\boldsymbol{\theta})$ are equivalent to the prior distribution, then we arrive at the harmonic estimator [*Kass and Raftery*, 1995], which is known to be unstable [*Newton and Raftery*, 1994; *Liu et al.*, 2016]. A better choice for both $q_0(\boldsymbol{\theta})$ and $q_{1/2}(\boldsymbol{\theta})$ is a parametric approximation of the posterior distribution, which leads to reciprocal importance sampling [*Gelfand and Dey*, 1994; *Di Ciccio et al.*, 1997]. A further simplification can be made if, instead of averaging over all target samples, one uses only the mode, $\check{\boldsymbol{\theta}}$, of the posterior realizations, $\{\Theta_1\}$. The formula for Z then becomes

$$Z=\frac{p(\check{\boldsymbol{\theta}})L(\check{\boldsymbol{\theta}}|\tilde{\mathbf{Y}})}{q_0(\check{\boldsymbol{\theta}})} \tag{8}$$

If we take $q_0(\boldsymbol{\theta})$ to be the d-variate normal density, $f_d(\boldsymbol{\theta};\check{\boldsymbol{\theta}},\mathbf{C}(\boldsymbol{\theta}))$, with mean $\check{\boldsymbol{\theta}}$ and covariance matrix, $\mathbf{C}(\boldsymbol{\theta})=\mathrm{Cov}(\{\Theta_1\})$, then this reduces to the well-known Laplace-Metropolis (LM) estimator [*Lewis and Raftery*, 1997]

$$Z=p(\check{\boldsymbol{\theta}})L(\check{\boldsymbol{\theta}}|\tilde{\mathbf{Y}})(2\pi)^{d/2}|\mathbf{C}(\boldsymbol{\theta})|^{1/2}, \tag{9}$$

where $|\cdot|$ signifies the determinant operator. If the posterior distribution deviates from normality, then better results can be expected by using for $q_0(\boldsymbol{\theta})$ a mixture approximation of the target density (see section 3.2), turning equation (8) into a "generalized" LM estimator.

We now evaluate what happens if we set $q_{1/2}(\boldsymbol{\theta})=q_1(\boldsymbol{\theta})$. This reduces the denominator in equation (6) to unity, and turns bridge sampling into importance sampling (IS) with $q_0(\boldsymbol{\theta})$ as importance density [*Hammersley and Handscom*, 1964]. This becomes evident if we interpret the simplified formula of r $(=Z)$

$$Z=\mathbb{E}_0\left[\frac{q_1(\boldsymbol{\theta})}{q_0(\boldsymbol{\theta})}\right]=\mathbb{E}_0\left[\frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}{q_0(\boldsymbol{\theta})}\right]\approx\frac{1}{m_0}\sum_{j=1}^{m_0}\frac{p(\{\boldsymbol{\theta}_0^j\})L(\{\boldsymbol{\theta}_0^j\}|\tilde{\mathbf{Y}})}{q_0(\{\boldsymbol{\theta}_0^j\})} \tag{10}$$

While the posterior samples, $\{\Theta_1\}$, do not enter directly in the formula above (samples of $\{\Theta_1\}$ are drawn from $q_0(\boldsymbol{\theta})$), they are still used to construct a good importance distribution. Indeed, the importance distribution should match closely the posterior distribution for importance sampling to be efficient and robust. Ideally, the importance distribution has slightly heavier tails, and is thus more overdispersed than the distribution it is intended to approximate [see, e.g., *Evans and Swartz*, 1995; *Raftery*, 1996; *Tokdar and Kass*,

2010]. The rationale behind importance sampling is that it "allow us to sample from one distribution when we ought to be sampling from another" [*Hammersley and Handscom*, 1964, p. 42]. Specifically, "the object in importance sampling is to concentrate the distribution of the sample points in the parts of the interval that are of most importance instead of spreading them out evenly" [*Hammersley and Handscom*, 1964, p. 58], as in simple Monte Carlo, thus returning a more efficient estimate of the evidence, Z. We graphically illustrate importance sampling in Figure 1b and use it to estimate the area under the posterior distribution, $q_1(\boldsymbol{\theta})=p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ (in blue). We do so by sampling from a distribution (indicated in green) that is biased toward the important regions of the posterior distribution, hence the name importance sampling. This "importance distribution," $q_0(\boldsymbol{\theta})$, has a known integral of unity, and should satisfy that $q_0(\boldsymbol{\theta}) > 0$ whenever $p(\boldsymbol{\theta}) \geq 0$, otherwise the area under $p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ is underestimated. The ratio of the density of the unnormalized posterior (blue curve) and the density of the importance distribution (green curve) now details the contribution (light blue area) of the sample, $\{\boldsymbol{\theta}_0^j\}$, to the marginal likelihood. The integral of the unnormalized posterior distribution is thus equivalent to the expectation of $p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})/q_0(\boldsymbol{\theta})$, which is approximated numerically using equation (10).

A degenerate case occurs when the prior distribution is used as importance distribution, or, $q_0(\boldsymbol{\theta})=p(\boldsymbol{\theta})$. Now the posterior samples are not used at all, and the inference of Z amounts to brute-force Monte Carlo sampling, wherein the model evidence is estimated by averaging of the likelihood over the prior distribution. This method is rather inefficient for high-dimensional targets and/or likelihood functions that are relatively peaked compared to the prior distribution [e.g., *Liu et al*., 2016]. In general, importance sampling may reduce significantly the computational burden of Monte Carlo simulation [*Evans and Swartz*, 1995; *Raftery*, 1996; *Tokdar and Kass*, 2010], yet the exact gains in efficiency and speed of convergence depend critically on the choice of $q_0(\boldsymbol{\theta})$. In practice, however it is not particularly easy to construct an adequate importance distribution [see, e.g., *Neal*, 2001; *Perrakis et al*., 2014], especially when the posterior distribution is high-dimensional and poorly described with a traditional statistical distribution. In section 3.2, we introduce a new and robust method that solves efficiently for equation (10). Our evidence estimator uses as importance density, $q_0(\boldsymbol{\theta})$, a mixture model of a large collection of posterior samples.

3.1.2. Using $q_{1/2}(\boldsymbol{\theta})$ as a "Bridge" Between $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$

From a viewpoint of bridge sampling, the methods in the previous section are at best suboptimal, because they do not exploit the idea of using $q_{1/2}(\boldsymbol{\theta})$ as a bridge to connect $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$. A more efficient evidence estimator would choose a bridge density, $q_{1/2}(\boldsymbol{\theta})$, which lies in between $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$. In this case, the general formula of equation (6) applies. Following *Gelman and Meng* [1998], we consider two generic choices: (i) a geometric bridge, $q_{1/2}(\boldsymbol{\theta})=q_0(\boldsymbol{\theta})^{1-\omega}q_1(\boldsymbol{\theta})^{\omega}$, where each value of ω $(0 < \omega < 1)$ yields a $q_{1/2}(\boldsymbol{\theta})$ in between $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$, and (ii) the optimal bridge, $q_{1/2}(\boldsymbol{\theta})=\left[Z\frac{s_0}{q_1(\boldsymbol{\theta})}+\frac{s_1}{q_0(\boldsymbol{\theta})}\right]^{-1}$, where $s_0=\frac{m_0}{m_0+m_1}$ and $s_1=\frac{m_1}{m_0+m_1}$ [*Meng and Wong*, 1996]. From all possible choices for $q_{1/2}(\boldsymbol{\theta})$, the optimal bridge should, at least in theory, yield the most efficient estimate of Z with smallest relative variance. Indeed, only if the m_0 samples of $\{\Theta_0\}$ and m_1 draws of $\{\Theta_1\}$ are independent we know exactly the values of m_0 and m_1. Serial correlation between the samples within both collections reduces the effective sample size. Further, the optimal bridge depends on the ratio between m_0 and m_1. If $m_0 \ll m_1$, then $q_{1/2}(\boldsymbol{\theta})$ tends to $q_0(\boldsymbol{\theta})$, whereas, on the contrary, if $m_0 \gg m_1$ then the optimal bridge converges toward the posterior (target) distribution, $p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$. Finally, the optimal $q_{1/2}(\boldsymbol{\theta})$ itself depends on Z, necessitating the use of fixed-point iteration to solve for Z in this case (obtained by inserting in equation (6) the expression for the optimal $q_{1/2}(\boldsymbol{\theta})$) [see also *Gelman and Meng*, 1998]

$$Z \leftarrow \frac{\frac{1}{m_0}\sum_{j=1}^{m_0}\frac{l_0^j}{s_0 Z+s_1 l_0^j}}{\frac{1}{m_1}\sum_{j=1}^{m_1}\frac{1}{s_0 Z+s_1 l_1^j}} \tag{11}$$

where $l_u^j=q_1(\{\boldsymbol{\theta}_u^j\})/q_0(\{\boldsymbol{\theta}_u^j\})$ with $u=0$ or $u=1$. Note THAT in this iteration only Z changes. The idea of constructing a bridge between $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$ is illustrated graphically in Figure 2 with a simple one-dimensional example. The choices for the geometric bridge (GB) and optimal bridge (OB) are both presented. Note, that OB (which depends on m_0/m_1) results in a smoother transition between $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$ than GB.

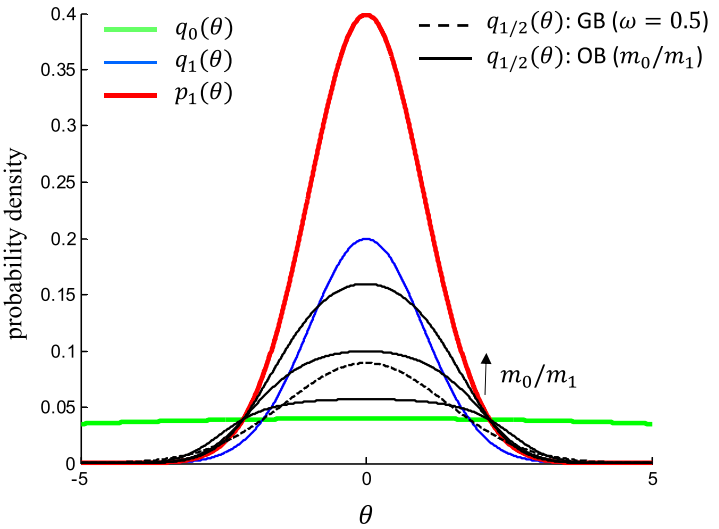

Figure 2. The $d = 1$ bridge density, $q_{1/2}(\theta)$, between $q_0(\theta)=\mathcal{N}(0, 10)$ and $q_1(\theta)=\mathcal{N}(0, 1)$ for a geometric bridge (GB) with $\omega=0.5$ (dashed curve) and optimal bridge (OB) for different values of the ratio $m_0/m_1=0.5$, $m_0/m_1=2$, and $m_0/m_1=5$ (solid black curves). The bridge density, $q_{1/2}(\theta)$, converges on $q_0(\theta)$ when $m_0 \ll m_1$, whereas $q_{1/2}(\theta)$ approaches the target, $q_1(\theta)$ (blue curve) when $m_0 \gg m_1$. The density of the unnormalized target distribution is derived from the standard normal density, $p_1(\theta)$ (red curve), via the identity, $q_1(\theta)=Zp_1(\theta)$ with $Z = 0.5$.

While the methods in this section use the slightly more complex formulation of equation (6), it is important to stress here that they do not necessarily involve a much larger CPU-cost than the evidence estimators discussed in the previous section (apart from reciprocal importance sampling), pending the assumption that all the different methods use the posterior realizations, $\{\Theta_1\}$, as Monte Carlo approximation of $q_1(\boldsymbol{\theta})$. In fact, most computational resources will need to be allocated to additional evaluations of the target density (prior $\times$ likelihood) for samples, $\{\Theta_0\}$ drawn from $q_0(\boldsymbol{\theta})$. Thus, from a computational point of view there is little difference between importance sampling and bridge sampling (with geometric or optimal bridge), yet, bridge sampling may offer substantial improvements of the evidence estimates (lower variance).

The various choices for $q_0(\boldsymbol{\theta})$ and $q_{1/2}(\boldsymbol{\theta})$ and the resulting model evidence estimators are summarized in Table 1. When choosing a suitable method, both accuracy and CPU-cost have to be considered. The closer $q_0(\boldsymbol{\theta})$ is to the posterior distribution, the more accurate and efficient the evidence estimates will be. Therefore, by definition, sampling methods that use for $q_0(\boldsymbol{\theta})$ an approximation of the target distribution are

Table 1. Overview of Model Evidence Estimation Methods Within the Context of Bridge Sampling[a]

$q_0(\boldsymbol{\theta})$	$q_{1/2}(\boldsymbol{\theta})$	Method	Samples
$p(\boldsymbol{\theta})$	$q_0(\boldsymbol{\theta})$	Harmonic estimator[b]	$q_1(\boldsymbol{\theta})$
	$q_1(\boldsymbol{\theta})$	Simple Monte Carlo[c]	$q_0(\boldsymbol{\theta})$
	$q_0(\boldsymbol{\theta})^{1-\omega} q_1(\boldsymbol{\theta})^{\omega}$	Bridge sampling with a geometric bridge[d]	$q_0(\boldsymbol{\theta})$, $q_1(\boldsymbol{\theta})$
	$\left[Z\frac{s_0}{q_1(\boldsymbol{\theta})}+\frac{s_1}{q_0(\boldsymbol{\theta})}\right]^{-1}$	Bridge sampling with the optimal bridge	$q_0(\boldsymbol{\theta})$, $q_1(\boldsymbol{\theta})$
$p_{\text{mix}}(\boldsymbol{\theta})$	$q_0(\boldsymbol{\theta})$	Reciprocal importance sampling[b,e]	$q_1(\boldsymbol{\theta})$
	$q_1(\boldsymbol{\theta})$	Importance sampling[c,f]	$q_0(\boldsymbol{\theta})$, $q_1(\boldsymbol{\theta})$[g]
	$q_0(\boldsymbol{\theta})^{1-\omega} q_1(\boldsymbol{\theta})^{\omega}$	Bridge sampling with a geometric bridge[d]	$q_0(\boldsymbol{\theta})$, $q_1(\boldsymbol{\theta})$
	$\left[Z\frac{s_0}{q_1(\boldsymbol{\theta})}+\frac{s_1}{q_0(\boldsymbol{\theta})}\right]^{-1}$	Bridge sampling with the optimal bridge	$q_0(\boldsymbol{\theta})$, $q_1(\boldsymbol{\theta})$

[a]We summarize possible choices for $q_0(\boldsymbol{\theta})$ and $q_{1/2}(\boldsymbol{\theta})$, wherein $p(\boldsymbol{\theta})$ signifies the prior distribution, $p_{\text{mix}}(\boldsymbol{\theta})$ denotes the mixture approximation of the posterior target distribution, $q_1(\boldsymbol{\theta})$ is the unnormalized posterior, and $q_{1/2}(\boldsymbol{\theta})$ characterizes the bridge density. As discussed in the text, the optimal bridge, $q_{1/2}(\boldsymbol{\theta})$, depends on the model evidence, Z, and on the fractional number of samples, s_0 and s_1, drawn from $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$, respectively.
[b]Limiting case of bridge sampling with a geometric bridge and $\omega = 0$.
[c]Limiting case of bridge sampling with a geometric bridge and $\omega = 1$.
[d]With $0 < \omega < 1$.
[e]Laplace-Metropolis is a special case (see text).
[f]With $q_0(\boldsymbol{\theta})$ as importance distribution
[g]Posterior samples are used to construct $p_{\text{mix}}(\boldsymbol{\theta})$.

expected to outperform those that use for $q_0(\boldsymbol{\theta})$ the prior distribution [*Gelman and Meng*, 1998]. Here, we choose as importance distribution a Gaussian mixture model, $p_{\text{mix}}(\boldsymbol{\theta})$, of a large collection of samples of the target distribution. Thus our method relies heavily on the ability of Monte Carlo sampling methods such as DREAM to characterize adequately the target distribution. The next section discusses the details of this approach.

Before we proceed to the next section, we note that *Gelman and Meng* [1998] have presented a generalization of bridge sampling, called "multi-bridge" sampling. This approach uses an infinite number of intermediate densities to estimate the target's normalizing constant. This method, coined path-sampling, is also known as thermodynamic integration (e.g. *Ogata* [1989]; see also *Xie et al*. [2011] and *Fan et al*. [2011] for related steppingstone estimators). Multiple step integration methods are superior to single-step methods, particularly when the target distribution is initially unknown. The series of intermediate densities then slowly converges to the target distribution, yet at the expense of a much enlarged computational complexity and burden. As our method constructs the mixture distribution, $p_{\text{mix}}(\boldsymbol{\theta})$, from a large sample of posterior draws, it adapts immediately, and in a single-step, to the target, rendering unnecessary multiple step methods.

3.2. Mixture Approximation of the Posterior Distribution

If we choose the importance distribution, $q_0(\boldsymbol{\theta})$, in equation (6) rather loosely and freehand, then evidence estimation may become cumbersome, particularly for parameter-rich models. A much better and more defensible approach would be to modulate the importance distribution after the target distribution. This not only guarantees that we sample the "right" areas of the parameter space, but also make sure that we visit these areas with a frequency approximately equivalent to their underlying posterior density. We choose our importance distribution within the family of normal mixtures [cf., *Di Ciccio et al*., 1997] and coin our approach GAussian Mixture importancE, or GAME, sampling. Normal mixtures are flexible and allow us to approximate as closely and consistently as possible a wide range of target distributions. Indeed, multimodal, truncated, and "quasi-skewed" distributions can be emulated with a mixture distribution if a sufficient number of Gaussian variates is used. In some cases, one may prefer to use a nonparametric importance distribution, yet the efficiency of nonparametric importance sampling relies heavily on the nonparametric estimator being used, thus necessitating development of new estimators that are computationally superior to their kernel-based counterparts [see, e.g., *Neddermeyer*, 2009].

As a precursor to our method, we generate many different (parameter) samples from each candidate model's unnormalized posterior distribution in equation (4) using MCMC simulation with the DREAM algorithm [*Vrugt et al*., 2008, 2009]. We collect these m posterior samples in a $m \times d$ matrix, $\{\Theta_*\}=(\{\boldsymbol{\theta}_*^1\},\ldots,\{\boldsymbol{\theta}_*^m\})$, and store their corresponding unnormalized posterior densities, $p(\{\boldsymbol{\theta}_*^j\})L(\{\boldsymbol{\theta}_*^j\}|\tilde{\mathbf{Y}})$ in a $m\times1$ vector, where $j=(1,\ldots,m)$. We use the DREAM package in MATLAB [*Vrugt*, 2016] because of its demonstrated capabilities and its many built-in options that simplify practical application. Nevertheless, the user is free to select any other Monte Carlo sampling method.

Next, we approximate the m posterior samples, $\{\Theta_*\}$, with a mixture distribution

$$p_{\text{mix}}(\boldsymbol{\theta})=\sum_{j=1}^{J} w_j f_d(\boldsymbol{\theta};\boldsymbol{\mu}_j,\Sigma_j), \tag{12}$$

of $J>0$ different d-variate normal densities, $f_d(\cdot|\boldsymbol{\mu}_j,\Sigma_j)$, where w_j, $\boldsymbol{\mu}_j$, and Σ_j signify the weight, the d-dimensional mean vector, and the $d\times d$-covariance matrix of the jth Gaussian component, respectively, and $j=(1,\ldots,J)$. The weights, or mixing probabilities, must lie on the unit simplex, $\mathcal{S}^d=\{\mathbf{w}\in\mathbb{R}^d : w_j\in[0,1],\sum_{j=1}^{J} w_j=1\}$, and the Σ_j's must be symmetric, $\Sigma_j(a,b)=\Sigma_j(b,a)$, and positive semidefinite.

The Expectation-Maximization (EM) algorithm [see e.g., *McLachlan and Krishnan*, 2007, and references therein] is used to estimate the values of the d_{mix}-variables of the mixture distribution, $\Phi=(w_1,\ldots,w_J,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_J)$, where $\boldsymbol{\beta}_j=(\boldsymbol{\mu}_j,\Sigma_j)$ stores the mean and covariance matrix of the jth normal density of the mixture. This algorithm maximizes the log-likelihood, $\ln\{L(\Phi|\{\Theta_*\},J)\}$, of the mixture distribution

$$\ln\{L(\Phi|\{\Theta_*\},J)\}=\sum_{i=1}^{m}\ln\left\{\sum_{j=1}^{J}w_j f_d(\{\boldsymbol{\theta}_*^i\};\boldsymbol{\mu}_j,\Sigma_j)\right\}, \tag{13}$$

by alternating between an expectation (E) step and a maximization (M) step, until convergence of the values of Φ is achieved for a given number of components, J. The optimum mixture distribution, hereafter referred to as $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$, can be determined from information criteria such as the BIC; see also equation (1)

$$I_{\text{BIC}}(J)=-2\ln\{L(\Phi|\{\Theta_*\},J)\}+d_{\text{mix}}(J)\ln(m). \tag{14}$$

If we treat as unknowns of each mixture component its weight, d-mean vector and $d(d+1)/2$ free elements of its covariance matrix, then $d_{\text{mix}}=J-1+J(d+d(d+1)/2)$. The unit simplex restricts the inference to $J-1$ weights. The BIC strikes a balance between quality of fit (first-term) and the complexity of the mixture distribution (second term). Alternatively, we can select J so that the variance of the ratio between the density of target distribution and the density of its parametric approximation, $p_{\text{mix}}(\boldsymbol{\theta})$, is minimized. This variance can be computed as follows

$$\sigma^2_{\text{mix}}(J)=\frac{1}{m}\sum_{i=1}^{m}\left[\frac{p(\{\boldsymbol{\theta}_*^i\})L(\{\boldsymbol{\theta}_*^i\}|\tilde{\mathbf{Y}})}{p_{\text{mix}}(\{\boldsymbol{\theta}_*^i\})}-\hat{\zeta}\right]^2, \tag{15}$$

where $\hat{\zeta}=\frac{1}{m}\sum_{i=1}^{m}\frac{p(\{\boldsymbol{\theta}_*^i\})L(\{\boldsymbol{\theta}_*^i\}|\tilde{\mathbf{Y}})}{p_{\text{mix}}(\{\boldsymbol{\theta}_*^i\})}$. Unlike BIC, this variance criterion uses all the m posterior samples, $\{\Theta_*\}$, to determine $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$. In practice, we can use different values for J and then select $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$ from within this pool using

$$\hat{J}=\arg\min_{J\in\mathbb{N}_+}\chi(J), \tag{16}$$

where $\mathbb{N}_+$ is the collection of strictly positive integers, and $\chi(J)=I_{\text{BIC}}$ or $\chi(J)=\sigma^2_{\text{mix}}$.

Once the optimal mixture distribution, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$, has been determined, we can set $q_0(\boldsymbol{\theta})=\hat{p}_{\text{mix}}(\boldsymbol{\theta})$ in equation (6) and estimate the normalizing constant, Z, and thus marginal likelihood, $p(\tilde{\mathbf{Y}})$, via reciprocal importance sampling, importance sampling and bridge sampling (see Table 1). This concludes our description of our mixture distribution.

4. GAME Sampling and Implementation in DREAM Package

We now provide an algorithmic outline of GAME sampling for evidence estimation within the context of MCMC simulation [cf., e.g., *Marshall et al.*, 2005]. This recipe will include four different sampling methods that have been detailed in the previous section to approximate the evidence, $\hat{Z}$, and thus integral, of the target distribution.

GAME takes as input a $m \times d$ matrix of posterior samples, $\{\Theta_*\}=(\{\boldsymbol{\theta}_*^1\},\ldots,\{\boldsymbol{\theta}_*^m\})$, and a $m\times 1$ vector of corresponding unnormalized densities, $p(\{\boldsymbol{\theta}_*^i\}|\tilde{\mathbf{Y}})=p(\{\boldsymbol{\theta}_*^i\})L(\{\boldsymbol{\theta}_*^i\}|\tilde{\mathbf{Y}})$, of the $i=(1,\ldots,m)$ realizations. Furthermore, the user also has to specify the value of $\omega\in[0,1]$, which determines the sampling method that will be used to compute $\hat{Z}$.

In words, we first generate m samples, $\{\Theta_*\}$, from each models' posterior parameter distribution using MCMC simulation with the DREAM algorithm. Then J_{max} different mixture distributions, $p_{\text{mix}}(\boldsymbol{\theta})$, with increasing number of normal components, $J=(1,\ldots,J_{\text{max}})$ are fitted to the posterior samples using maximum likelihood estimation with the EM algorithm (step 1). Then, in step 2, we determine the optimum complexity, $\hat{J}$, of the mixture distribution via minimization of I_{BIC} or σ^2_{mix}. This optimal mixture, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$, then serves as our catalyst in step 5 to estimate the evidence, Z, and thus marginal likelihood, $p(\tilde{\mathbf{Y}})$, of the target distribution. Via the identity, $q_{1/2}(\boldsymbol{\theta})=q_0(\boldsymbol{\theta})^{1-\omega}q_1(\boldsymbol{\theta})^{\omega}$, of the bridge density, $q_{1/2}(\boldsymbol{\theta})$, the user can choose among four different sampling methods. A value of $\omega=0$ results in reciprocal importance sampling (RIS), $0<\omega<1$ amounts to bridge sampling with a geometric bridge (GB), and $\omega=1$ equates to importance sampling (IS). As fourth, and last method, the user can activate, at the end of step 5, bridge sampling with an optimal bridge (OB). This method requires as input an initial value of $\hat{Z}$ from IS or GB.

Algorithm 1 GAUSSIAN MIXTURE IMPORTANCE SAMPLING

1. For $J=1:J_{max}$ *Do*

Calibrate mixture distribution, $p_{mix}(\boldsymbol{\theta})$*, of equation* (12) *by maximizing the log-likelihood,* $\ln\{L(\Phi|\{\Theta_*\},J)\}$*, of equation* (13) *with the EM algorithm*

Compute $I_{BIC}(J)$ *and* $\sigma^2_{mix}(J)$ *using equations* (14) *and* (15)

End

2. Select optimal mixture distribution, $\hat{p}_{mix}(\boldsymbol{\theta})$*, via equation* (16) *using* $I_{BIC}(\hat{J})$ *or* $\sigma^2_{mix}(\hat{J})$.

3. Draw m_1 $(m_1 \leq m)$ *samples from* $\{\Theta_*\}$ *and store collection in* $m_1 \times d$ *matrix,* $\{\Theta_1\}$.

4. Evaluate the mixture density, $\hat{p}_{mix}(\{\boldsymbol{\theta}_1^i\})$*, for the* m_1 *target samples;* $i=(1,\ldots,m_1)$

5. If $\omega = 0$ *Then (Reciprocal Importance Sampling)*

Compute $\hat{Z}$ *via equation* (7)

Otherwise ($0 < \omega \leq 1$*)*

Draw m_0 *samples* $\{\Theta_0\}$ *from* $\hat{p}_{mix}(\boldsymbol{\theta})$

Compute $\hat{p}_{mix}(\{\boldsymbol{\theta}_0^j\})$ *and evaluate target density,* $p(\{\boldsymbol{\theta}_0^j\}|\tilde{\mathbf{Y}})$*;* $j=(1,\ldots,m_0)$

If $\omega = 1$ *Then (Importance Sampling)*

Compute $\hat{Z}$ *via equation* (10)

Otherwise ($0 < \omega < 1$*) (Bridge Sampling with Geometric Bridge)*

Determine bridge density, $q_{1/2}(\boldsymbol{\theta}) = q_0(\boldsymbol{\theta})^{1-\omega} q_1(\boldsymbol{\theta})^{\omega}$

Compute $\hat{Z}$ *using* $q_{1/2}(\boldsymbol{\theta})$ *in equation* (6)

End

If (Bridge Sampling with Optimal Bridge) Then

Set $\hat{Z}_{(0)}$ *equal to* $\hat{Z}$ *from previous step*

For $r=1:R$ *Do (fixed point iteration)*

Compute $\hat{Z}_{(r)}$ *with equation* (11) *using* $\hat{Z}_{(r-1)}$

End

Set value of $\hat{Z}$ *equal to* $\hat{Z}_{(R)}$

End

End

6. Return $\hat{Z}$

It should be evident from the algorithmic recipe that the RIS, IS, GB, and OB evidence estimators do not invoke the same computational cost. In general, RIS is most CPU-efficient as the collection of target samples, $\{\Theta_*\}$, suffices to compute the evidence, $\hat{Z}$, in equation (7). The other three methods (IS, GB, and OB) require a second collection, $\{\Theta_0\}$, of m_0 samples drawn randomly from the optimal mixture (= importance) distribution, $\hat{p}_{mix}(\boldsymbol{\theta})$, to calculate the evidence, Z, via equations (10), (6), or (11), respectively. As it takes time to evaluate the (unnormalized) target density, $p(\{\boldsymbol{\theta}_0^j\}|\tilde{\mathbf{Y}}) = p(\{\boldsymbol{\theta}_0^j\})L(\{\boldsymbol{\theta}_0^j\}|\tilde{\mathbf{Y}})$, of each of these $j=(1,\ldots,m_0)$ mixture samples, IS, GB, and OB may require a much larger computational budget, particularly for CPU-demanding forward models. Nevertheless, this second collection of "importance samples" may proof crucial for obtaining unbiased evidence estimates.

The GAME sampler contains several algorithmic parameters that need to be specified by the user. This includes m_0, m_1, J_{max} and R. In our case studies, we used default values of $m_0=1000$, $m_1=1000$, $J_{max}=5$,

and $R = 10$. These values worked well for a range of different targets. Further on, we construct the mixture distribution, $p_{mix}(\boldsymbol{\theta})$, in step 1 using only $h = 2000$ samples from the (much) larger collection of m target realizations (step 0). This approach enhances considerably the overall CPU-efficiency of our method, while still providing a relatively accurate description of the target distribution. To negate sampling bias, we draw the m_1 realizations of $\{\Theta_1\}$ in step 3 from the collection $\{\Theta_*\}$ but without the h posterior samples that were used for mixture calibration (step 1). We report the values of m_1 in our numerical experiments.

Note, we advise to thin the chains of DREAM to reduce, as much as possible, serial correlation between subsequent samples of the target distribution. Thinning is particularly important for parameter-rich models, and explains why the use of a large number of posterior realizations, m_1, in step 3 does not necessarily improve evidence estimates. As discussed before, the culprit is effective sample size.

5. Benchmark Experiments With Known Targets

We conducted a wide range of numerical experiments to benchmark the performance of GAME sampling on target distributions with known normalizing constants. This includes multivariate normal distributions with variable dimensionality (up to $d = 100$), one or two (disconnected) modes, and variably correlated, twisted, and/or truncated dimensions. Each target has a normalizing constant of unity, except for the truncated distributions presented in section 5.1.3. Table 2 summarizes our setup of DREAM for the different benchmark experiments.

Thus, in this section we do not focus on model selection, but rather evaluate the ability of GAME sampling to infer successfully the normalizing constants of a variety of known statistical distributions with diverse and complex problem features. In all these benchmark experiments, the unnormalized posterior density is simply equivalent to the target density. For each target, we use 250 different trials with GAME and report the mean evidence estimates and their associated 95% confidence intervals.

Before we proceed with presentation of the results, we first analyze the impact of the choice of mixture selection criteria, l_{BIC} or σ^2_{mix}, on the selection of the optimal importance distribution. Table 3 reports the results of this analysis for the different case studies discussed in this section. Note that one has to be somewhat careful with interpretation of these results in the absence of detailed knowledge on the optimized values of the weights of the individual normal components of the mixture distribution. The results in this table highlight several important findings. First, notice that the optimal mixture distribution, $\hat{p}_{mix}(\boldsymbol{\theta})$, does not contain more than four normal distributions. This provides support for the claim that the default value of $J_{max} = 5$ is properly chosen. Second, the two selection criteria provide conflicting results for target distributions with fewer than 50 dimensions, but consistently select the same number of mixture components for the most complex targets. Third, the larger the number of dimensions of the target distribution, the lower the optimal number of components of the mixture distribution. In fact, both model selection criteria suggest that a single mode suffices for target distributions with more 50 dimensions. Fourth, the variance criterion, σ^2_{mix}, promotes mixture parsimony. This is particularly evident for low-dimensional targets with fewer than 20 dimensions, for which the BIC almost always selects a more complex mixture distribution with larger number of normal components.

Take home message is that the variance criterion, σ^2_{mix}, guarantees selection of a parsimonious mixture distribution. What is more, σ^2_{mix} and l_{BIC} promote use of a multivariate normal importance density for target distributions with more than 20 dimensions.

Table 2. Dream Settings Used in the Different Benchmark Experiments[a]

d	1	2	5	10	20	50	75	100
T	1,000	2,000	3,000	4,000	8,000	12,000	16,000	20,000
N	10	10	10	10	20	50	75	100
th	1	1	1	1	1	5	5	10
m	5,000	10,000	15,000	20,000	80,000	60,000	120,000	100,000

[a] d = target dimensionality; T = number of samples per chain; N = number of chains; th = thinning rate, and m = size of collection of posterior samples. To give DREAM sufficient opportunity to converge to a point on the target distribution, we discard the samples in the first half of each chain. This equates to a burn-in of 50%.

Table 3. Mixture Model Selection[a]

d	1	2	5	10	20	50	75	100
Normal ($\rho = 0.25$)	1 (1)	1 (2)	1 (1)	1 (2)	1 (1)	1 (1)	1 (1)	1 (1)
Normal ($\rho = 0.5$)	1 (1)	1 (1)	1 (1)	1 (3)	1 (1)	1 (1)	1 (1)	1 (1)
Normal ($\rho = 0.75$)	1 (1)	1 (1)	1 (3)	1 (2)	1 (1)	1 (1)	1 (1)	1 (1)
Banana-shaped		4 (4)	2 (4)	1 (4)	1 (4)	1 (1)	1 (1)	1 (1)
Normal mixture		2 (2)	2 (4)	2 (4)	1 (2)	1 (2)	1 (1)	1 (1)
Truncated normal ($\rho = 0.5$)	4 (4)	4 (4)	1 (2)	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)

[a]The number of normal components, $\hat{J}$, of the optimal mixture (= importance) distribution, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$, according to selection criteria σ^2_{mix} and I_{BIC} (between brackets) as function of target dimensionality, d. We list separately the results for the multivariate normal distribution with variably correlated dimensions using ρ=0.25, ρ=0.5, and $\rho = 0.75$, the banana-shaped distribution with $b = 0.1$, the multivariate mixture distribution with two disconnected modes, and the truncated normal distribution.

Unless stated otherwise, we use the variance criterion, σ^2_{mix}, in equation (15) to select the optimal mixture distribution, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$.

5.1. Evidence Estimation Using End-Member Bridge Densities: IS and RIS

In this section, we discuss the results of IS and RIS, the two extreme bridge densities. These two methods share the same importance density, that is, $q_0(\boldsymbol{\theta}) = \hat{p}_{\text{mix}}(\boldsymbol{\theta})$, but they differ in how they estimate the normalizing constant of the target. To do so, RIS uses only the samples of the target distribution and their respective unnormalized densities, whereas IS (plus GB and OB) requires another sample of points drawn randomly from the optimal mixture distribution, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$. Consequently, IS is computationally more costly. In a later section, we investigate the performance of the intermediate bridge densities using GB and OB.

5.1.1. Multivariate Target With Variably Correlated Dimensions

Our first study considers a zeroth-mean d-variate normal distribution with $d \times d$ covariance matrix, Σ, in $\mathbb{R}^d$

$$f(\boldsymbol{\theta}) = f_d(\boldsymbol{\theta}; \boldsymbol{\theta}, \Sigma). \tag{17}$$

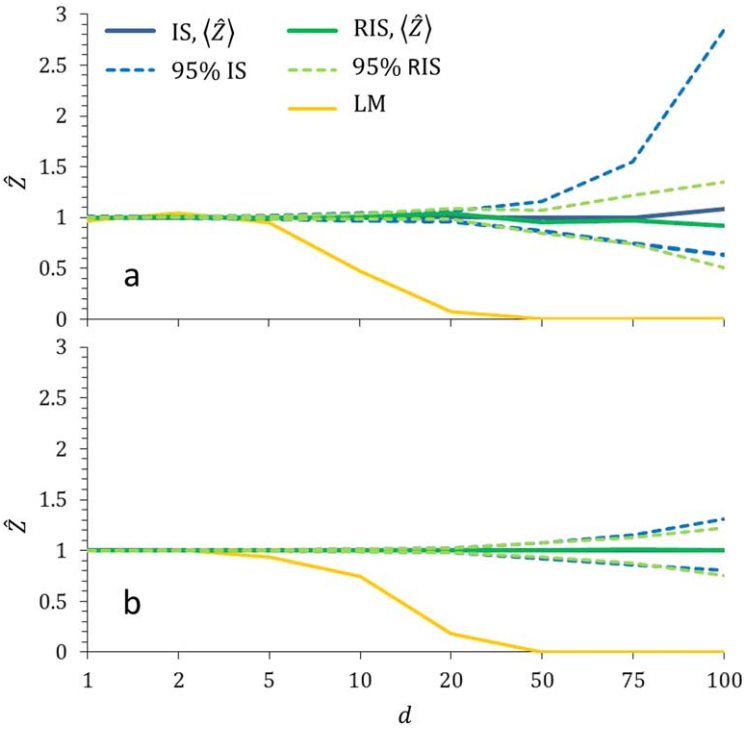

Figure 3. d-variate normal target distribution with correlated dimensions, $\rho = 0.5$: (a) Trace plot of mean evidence estimates, $\hat{Z}$ (solid lines), and their 95% confidence intervals (dashed lines) as function of target dimensionality, d, using 250 independent trials with IS ($m_0 = 1000$) and RIS ($m_1 = 1000$). The mean evidence estimates of the LM method are separately indicated with the yellow lines; (b) Same as Figure 3a, but now using 12m, that is, a collection of MCMC samples that is about 12 times larger. Default values of m are found in Table 2.

The variance of the *jth* variable is set equal to j and all pairwise correlations are set to ρ, where $\rho \in (0.25, 0.50, 0.75)$. Unless stated differently, we first analyze results for $\rho = 0.5$. Note, that the target is within the same family of distributions as the importance distribution.

Figure 3 presents trace plots of the mean evidence estimates, $\hat{Z}$ (solid lines) and their 95% confidence intervals (dashed lines) as function of target dimensionality, d. Color coding is used to differentiate between the results of IS (blue) and RIS (green). The mean evidence estimates of the LM method are separately indicated with the solid yellow lines. The most important results are as follows. First, RIS and IS retrieve correctly the unit normalizing constants of the target distributions. The mean

evidence estimates appear unbiased for all considered target dimensions. Second, the 95% confidence intervals of the evidence estimates of both methods increase with target dimensionality. This result is expected and explained by random between-trial variations in the selection of target (RIS) and importance (IS) samples of both methods. Third, the confidence intervals of IS appear larger than their counterparts derived from RIS. This is most evident at $d = 75$ and $d = 100$, and suggests that the target samples may contain most information to estimate with high confidence the evidence. Finally, the LM evidence estimates are spot on for low-dimensional targets, but gravitate toward values of zero for $d \geq 10$. This may be an unexpected result, certainly because the multivariate normal Laplace approximation satisfies exactly the Gaussian target distribution. Instead, this highlights a problem with the mode of the LM distribution. This mode is derived from the collection of MCMC realizations by locating the sample with largest value of the target density. If this mode deviates only a little bit from the true mode of the target (= zeroth vector), the density will be reduced, and the LM method may underestimate the integral of the target distribution. It is for this reason that *Lewis and Raftery* [1997] suggest using the median of the MCMC realizations instead, as this moment is more robust. Altogether, the results favor RIS.

The bottom plot, Figure 3b, presents the evidence estimates of RIS, IS, and LM, for a much enlarged sample of posterior realizations using a 12× larger value of m. The results are qualitative similar to those presented in the top plot, except, that the 95% confidence intervals of the evidence estimates of IS and RIS have become smaller. This is particularly true for IS. Note, that the evidence estimates of the LM method appear rather unaffected by the value of m.

We next investigate, in Figure 4, the relationship between the dimensionality, d, of the multivariate normal target, and the value of the variance, σ^2_{mix} of the optimal mixture distribution, $p_{\text{mix}}(\boldsymbol{\theta})$, for IS. We present traces for different values of m_0 (solid lines) using the default value of m in Table 2 (and used in Figure 3). We separately also depict a trace (dashed black line) for the much larger collection of $12m$ target realizations using the default value of $m_0 = 1000$. In general, the smaller the value of σ^2_{mix} the closer the mixture distribution is to the target of interest. The solid lines are in qualitative agreement, and demonstrate that the error variance, σ^2_{mix}, of the mixture distribution increases linearly with dimensionality of the target distribution. In fact, σ^2_{mix} increases almost linearly (on a log-scale) with d. This may not be a desirable finding, yet this increase can be countered in part by using a substantially larger number of samples, m_0, from the importance distribution. Note, that the dashed-black line with $12m$ and $m_0 = 1000$ is in close agreement with the solid blue line using $m = m_0 = 10,000$. Figure 4 also highlights that, when comparing models with contrasting number of parameters (different values for d), one may need to use a different number of importance samples, m_0, to achieve evidence estimates with comparable confidence intervals.

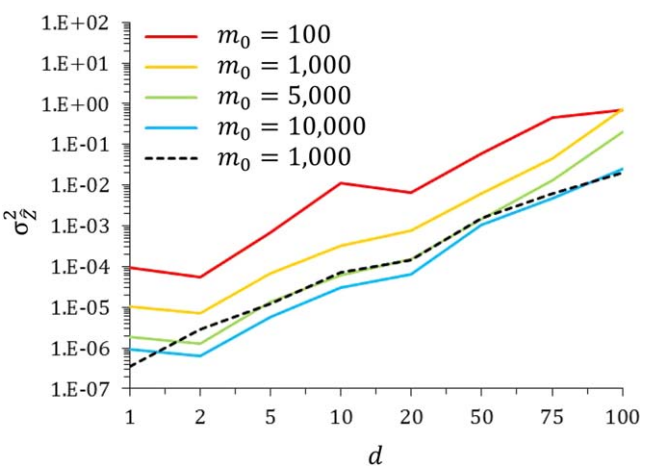

Figure 4. d-variate normal target distribution with correlated dimensions, $\rho = 0.5$: Variance, $\sigma^2_{\hat{z}}$, of IS evidence estimates as function of target dimensionality, d using 250 independent trials. The solid lines (color coded) use differentiate values of m_0 (number of importance samples). The dashed line presents the results of $m_0 = 1000$ but using 12 m, that is, a collection of MCMC samples that is about 12 times larger. Default values of m for the solid lines are found in Table 2.

We next investigate what happens to the performance of IS when the dimensions of the normal target exhibit an increasingly stronger linear correlation. Figure 5 presents the results of this analysis, and displays trace plots of the mean evidence estimates (solid line) and their 95% confidence intervals (dashed lines) using pairwise parameter correlations of (a) $\rho = 0.25$, (b) $\rho = 0.5$, and (c) $\rho = 0.75$, respectively, and $m_0 = 5000$. These results suggest that parameter correlation hardly affects the IS evidence estimates. Indeed, the mean evidence estimates (blue lines) match perfectly their unit

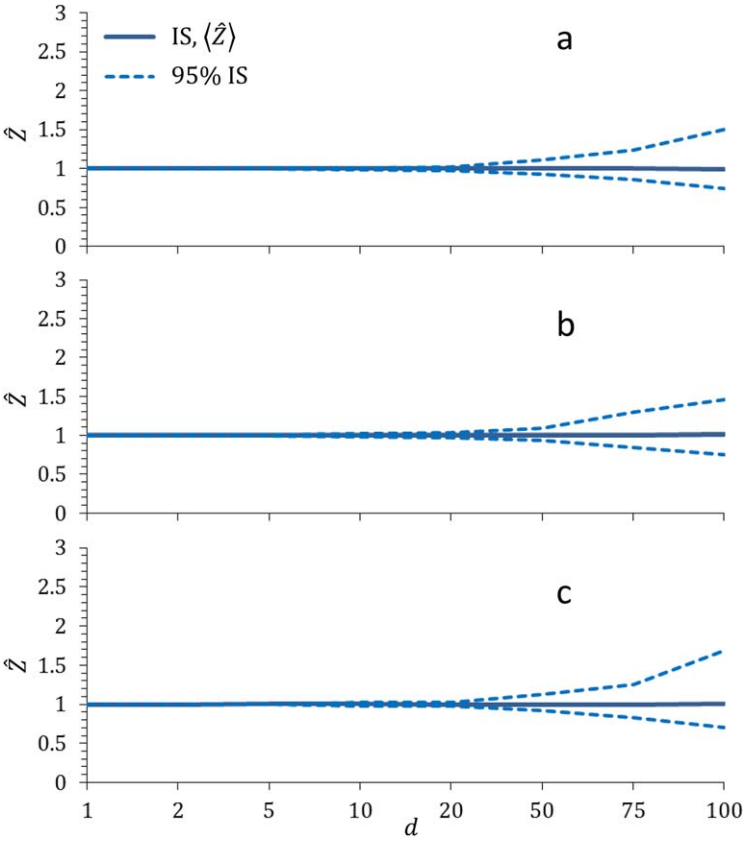

Figure 5. *d*-variate normal target distribution with correlated dimensions, ρ: Trace plot of mean evidence estimates, $\hat{Z}$ (solid blue lines), of IS and their 2.5% and 97.5% percentiles (dashed blue lines) as function of target dimensionality, *d* using a) ρ=0.25, (b) $\rho = 0.5$, and (c) $\rho = 0.75$. Results are based on 250 independent trials with m_0=5000.

values of the target distributions. Stronger linear dependencies among the parameters do enlarge the 95% confidence intervals of the evidence estimates, nevertheless, this increase is relatively small.

5.1.2. Multivariate Target With Twisted Dimensions

The second case study considers a *d*-variate twisted normal distribution

$$f(\boldsymbol{\theta}) = f_d(\phi_b(\boldsymbol{\theta}); \boldsymbol{0}, \Sigma), \quad (18)$$

where $\phi_b(\boldsymbol{\theta}) = (\theta_1, \theta_2 + b\theta_1^2 - 100b, \theta_3, \ldots, \theta_d)$ and $\Sigma = \mathbf{I}_d$, the $d \times d$ identity matrix, except for $\Sigma(1,1) = 100$. We use the value of $b = 0.1$ to yield a strongly twisted, banana-shaped, distribution for the first two dimensions. We consider $d \in [2, 100]$, and use m_0=5000 and m_1=1000 (default).

Figures 6a and 6c (top) displays the evolution of the mean evidence estimates (solid lines) and their 95% confidence intervals (dashed lines) derived from RIS (green) and IS (blue) as function of the dimensionality of the *d*-variate twisted distribution using σ^2_{mix} (left graph) and I_{BIC} (right graph) as selection criteria for the optimal mixture distribution. The solid yellow line in both graphs displays the mean evidence estimates of LM. In general, the results do not seem to depend much on the choice of selection criteria for the optimal mixture distribution. It is evident that the performance of all three methods has deteriorated compared to the multivariate normal target. IS is the only method that correctly retrieves the unit normalizing constant of the target and provides evidence estimates that appear unbiased for $d \in [2, 100]$. Note, however that its 95% confidence intervals have increased compared to the normal target (cf., Figure 6 with Figure 5). The LM method is particularly inferior even for smallest values of *d*, reinforcing the inability of a single multivariate normal to approximate closely a highly nonlinear (twisted) target distribution.

But why then does RIS perform (much) more poorly on this twisted target than IS? RIS uses as bridge density the optimal mixture distribution, and, thus uses only the collection of target samples to compute the evidence. The banana-shaped target is more difficult to sample by DREAM with an acceptance of about 10% which is considerably lower than its counterpart of 14–47% of the normal target. This reduces sample diversity, and, with the use of a fixed *m* (see Table 2), may bias somewhat the description of the target distribution. What is more, the normal mixture model may not be flexible enough to approximate sufficiently the twisted target. This is further illustrated in Figure 7, which displays for $d = 2$ different confidence intervals of (the center of) the target distribution, $(\boldsymbol{\theta})$ (left graph) and the optimal mixture distribution, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$ (right graph). Indeed, the optimal mixture model does not mimic exactly the target distribution. This discrepancy may bias the RIS evidence estimates, but not affect much the results of IS as this method uses as bridge density the target distribution (see section 3.1 and Table 1).

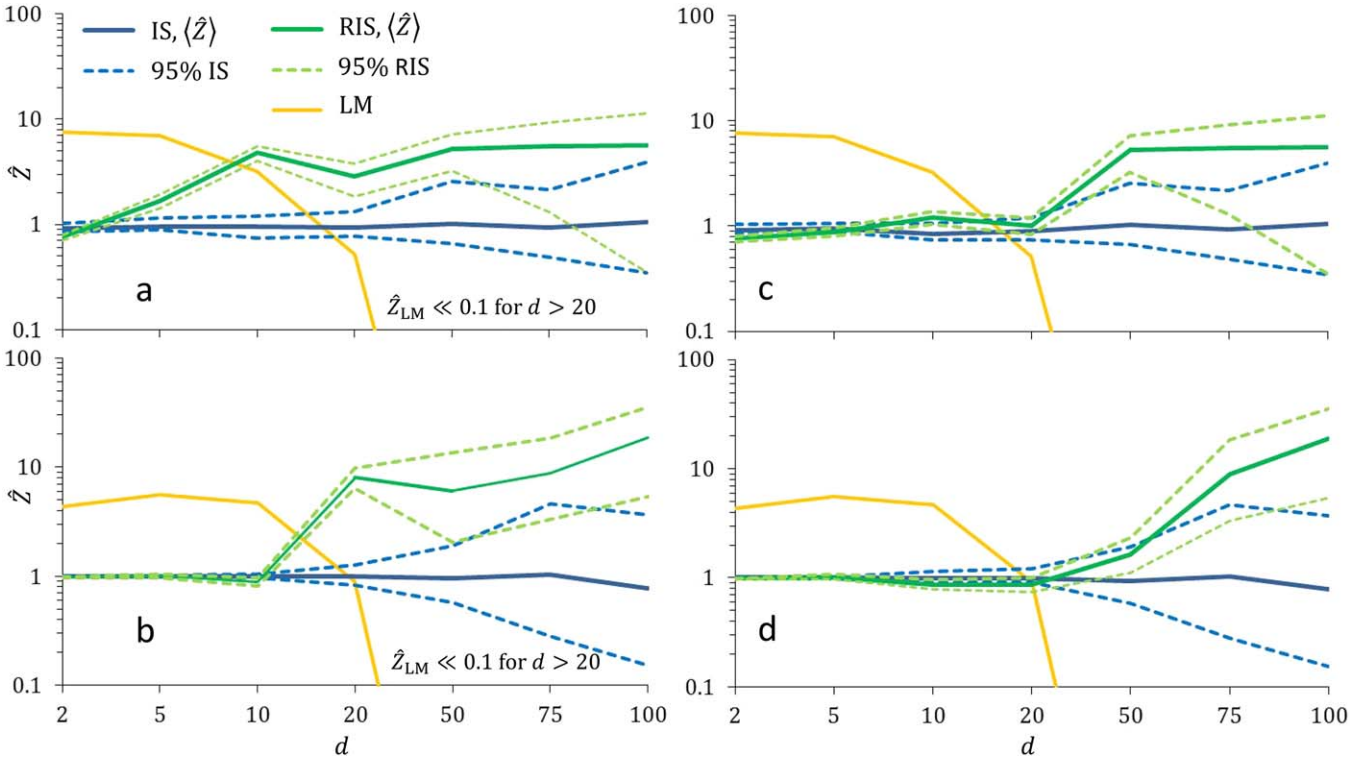

Figure 6. (a, c) *d*-variate twisted normal and(b, d) *d*-variate mixture of two normal variates with disconnected modes: Evolution of mean evidence estimates, $\hat{Z}$ (solid lines), of IS (blue) and RIS (green) and their 2.5% and 97.5% percentiles (dashed lines) as function of target dimensionality, *d* using 250 independent trials with $m_0 = 5000$ and $m_1 = 1000$. The optimum mixture distribution, $\hat{p}_{mix}(\boldsymbol{\theta})$, is determined via minimization of the variance criterion, σ^2_{mix}. The evidence estimates of the LM method are separately indicated with the yellow lines. Note, that for $d > 20$ the LM evidence estimates approach zero (see also Figure 3).

By now, it should be clear that the performance of RIS depends critically on the ability of the mixture distribution to emulate exactly the posterior target distribution. This may suggest using a much larger number of mixture components, *J*, nevertheless, as is shown for I_{BIC} in Figure 6c, this hardly improves the evidence estimates of RIS.

5.1.3. Multivariate Target With Disconnected Modes

The third case study considers a mixture distribution with two disconnected modes. The density of this *d*-variate distribution is given by

$$f(\boldsymbol{\theta}) = 1/3 f_d(\boldsymbol{\theta}; -\mathbf{5}, \Sigma) + 2/3 f_d(\boldsymbol{\theta}; \mathbf{5}, \Sigma), \tag{19}$$

where $-\mathbf{5}$ and $\mathbf{5}$ signify the *d*-means of the first and second normal component, respectively, and the covariance matrix, Σ is set equal to the identity matrix, $\mathbf{I}_d$. This study is relevant for environmental modeling as it portrays a quite common situation in which the posterior distribution is dispersed and concentrated in two or more, disconnected, pockets of the parameter space. This demands separate integration of each posterior mode, complicating tremendously model evidence estimation.

We assume $m_0 = 5000$ and present in the bottom plot of Figure 6 the results of IS, LM, and RIS for $d \in [2, 100]$ using as selection criteria for the optimal mixture distribution σ^2_{mix} (Figure 6b) and I_{BIC} (Figure 6d). Overall, the results are very similar to those presented previously for the twisted target. Indeed, the LM method cannot be relied upon to make accurate estimates of the evidence. The multivariate normal Laplace approximation cannot mimic the two modes and peaks of the target. RIS appears adequate for $d \leq 10$ but does not work well for larger target dimensionalities. This is explained by the selection of a too simplistic, unimodal, optimal mixture distribution, $\hat{p}_{mix}(\boldsymbol{\theta})$, by σ^2_{mix} and I_{BIC} (see Table 3) which cannot capture the two modes of target. As was the case in the previous two studies, IS correctly retrieves the unit normalizing constant of the target and provides evidence estimates that appear unbiased up to $d = 100$ dimensions.

5.1.4. Multivariate Target With Truncated Dimensions

Our last benchmark experiment considers application of GAME sampling to a truncated target distribution. This case study is deliberately included to evaluate the ability of GAME to accurately estimate the evidence

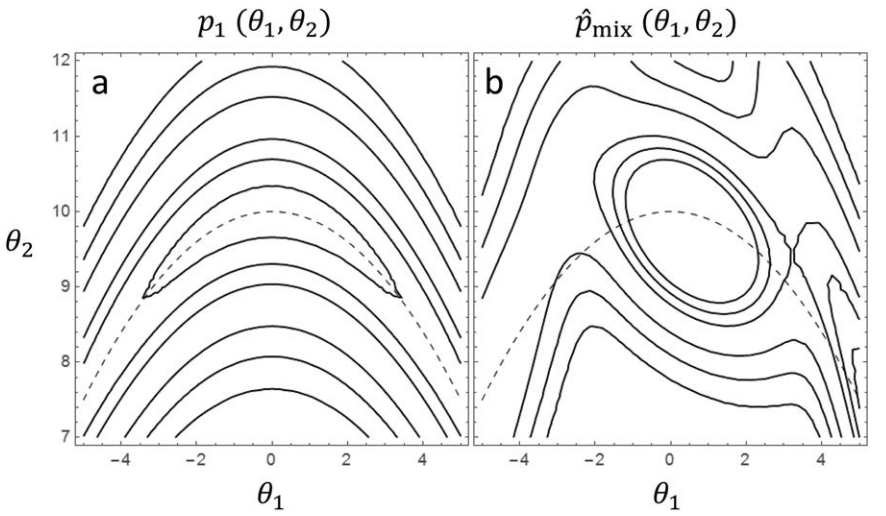

Figure 7. Bivariate contour plot of the density of the (a) $d = 2$ variate, banana-shaped, target distribution, and (b) the "optimal" mixture distribution, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$ with $J = 4$ normal variates (see Table 3). The dashed line in both graphs presents the theoretical relationship between θ_1 and θ_2. The target distribution is described in detail in section 5.1.2.

of models with bounded parameter spaces. This is common in environmental modeling as parameters may represent physical and/or conceptual properties with known upper and lower values. We revisit the *d*-variate Gaussian distribution of equation (17) with $\rho = 0.5$, and truncate each dimension with a box-car prior so that the target's normalizing constant is reduced from unity to 3/4. We approximate this truncated distribution with DREAM by evaluating the target density on a bounded search domain with ranges that are symmetric around zero (= target mean) and increase linearly with *d* in agreement with the covariance matrix. As a consequence, truncation reduces most the ranges of the higher target dimensions as they exhibit the largest variances.

To best emulate the actual target density, we scale the density of the optimal mixture distribution, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$, after step 2 of GAME so that its integral becomes unity within the prior ranges of the parameters of the target distribution. Figure 8 presents the results of our analysis using the default value of $m_0 = 1000$. The mean evidence estimates of IS (solid green line) appear unbiased for all considered target dimensionalities, and its 95% uncertainty ranges (dashed green lines) have increased somewhat compared to the first case study (see Figure 3). RIS (blue lines) works well for $d < 20$, but strongly underestimates the normalizing constant for larger values of *d*. The culprit is the density of the truncated mixture distribution which does not approximate sufficiently closely the target density for $d \geq 20$. This mismatch will affect only RIS and not IS, as this latter method uses directly the target density of the mixture samples. Finally, the LM estimator (solid yellow lines) appears to be deficient and biased for even the smallest values of *d*.

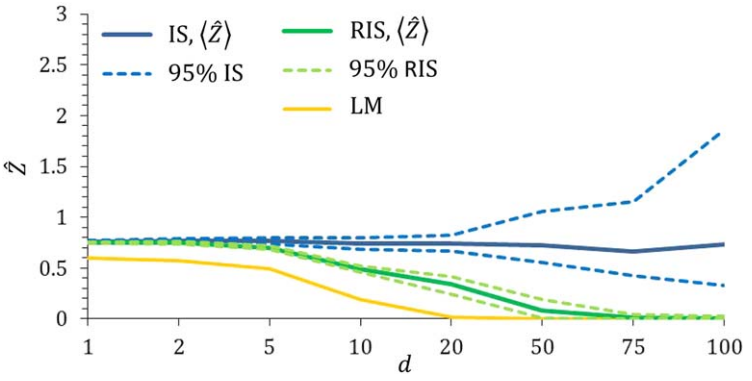

Figure 8. *d*-variate truncated normal target distribution with correlated dimensions, $\rho = 0.5$: Trace plot of mean evidence estimates, $\hat{Z}$ (solid lines), of RIS (green) and IS (blue) and their 2.5% and 97.5% percentiles (dashed lines) as function of target dimensionality, *d* using 250 independent trials with $m_0 = 1000$ and $m_1 = 1000$. The evidence estimates of the LM method are separately indicated with the yellow lines. The theoretical evidence equates to 0.75 for all *d*.

5.2. Model Evidence Estimation Using Geometric and Optimal Bridge Densities

Our benchmark experiments have shown that IS provides accurate estimates of the evidence for target distributions with a host of different problem features. On the other hand, RIS suffers if the

optimal mixture distribution in step 2 of GAME does not accurately portray the target distribution. We can further enhance IS by embedding this method in the bridge sampling framework of section 3.1.2. In this section, we analyze what happens to the evidence estimates if we use a bridge density, $q_{1/2}(\boldsymbol{\theta})$, to traverse between $q_0 = \hat{p}_{\text{mix}}(\boldsymbol{\theta})$ and the unnormalized posterior density, $q_1(\boldsymbol{\theta})$.

Figure 9 displays the evidence estimates of GB as function of ω for a $d = 75$ (top) and $d = 100$ (bottom) variate normal target with $\rho = 0.75$ and $m_0 = 5000$. Note, that if $\omega = 0$ then GB is equivalent to RIS with the optimal mixture distribution, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$, and if $\omega = 1$ then GB becomes equivalent to IS with $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$ as importance distribution. GB appears biased for all values of $\omega \in [0, 1)$ if $d = 75$, but correctly infers the evidence, $\hat{Z}$, of the $d = 100$ variate normal target, irrespective of the value of ω. What is more, if $\omega \simeq 0.6$, then the 95% confidence intervals of the evidence estimates derived from GB are substantially smaller than their counterparts derived from RIS ($\omega = 0$) and IS ($\omega = 1$). Although not further demonstrated herein, we observed similar results for other values of d and target distributions.

If we satisfy the assumption of independent sampling, then among all possible choices for the bridge density, $q_{1/2}(\boldsymbol{\theta})$, OB in equation (11) should yield the most accurate estimates of the model evidence. Figure 10 presents trace plots of the mean evidence estimates (solid red line) and associated 2.5% and 97.5% percentiles (dashed red lines) derived from OB for the d-variate normal target with $\rho = 0.75$ (a: top plot), the d-variate twisted normal target with $b = 0.1$ (b: middle plot) and the d-variate normal mixture with disconnected modes (c: bottom plot). The results of IS are separately displayed using the solid and dashed blue lines, respectively. These results demonstrate that OB provides unbiased estimates of the evidence, and with 95% prediction intervals that are considerably smaller than their IS counterparts, particularly for the higher dimensional targets. This is true even if the m_1 posterior samples from the unnormalized density $q_1(\boldsymbol{\theta})$ violate independence. Indeed, thinning (discussed in section 4) may not remove fully the autocorrelation of the m_1 DREAM samples, certainly for larger values of d. Note, that the RIS evidence estimates for the twisted and bimodal targets appeared particularly poor (see Figure 6).

Finally, we recall that the results of OB depend on the ratio between m_0 and m_1. The simulations in Figure 10 assumed that $m_0 = 5m_1$. In general, the smaller the value of m_0, the larger the variance of the OB evidence estimates. We investigate this further in Figure 11 and display the mean evidence estimates (solid red line) of OB and associated 2.5% and 97.5% percentiles (dashed red lines) as function of m_0 for a $d = 100$ variate normal target with $\rho = 0.75$. The results of IS are separately displayed with the blue lines. The mean evidence estimates of OB appear rather unaffected by the choice of m_0. What is more, the OB evidence estimates also are insensitive to the initial guess of $\hat{Z}$ in step 5 of the algorithmic recipe of GAME. The results in Figures 10 and 11 are obtained using for $\hat{Z}_0$ the evidence estimate from IS. Nevertheless, similar results are obtained if $\hat{Z}_0$ is set equivalent to the evidence estimates of RIS or GB (as in step 5 with $0 \leq \omega \leq 1$).

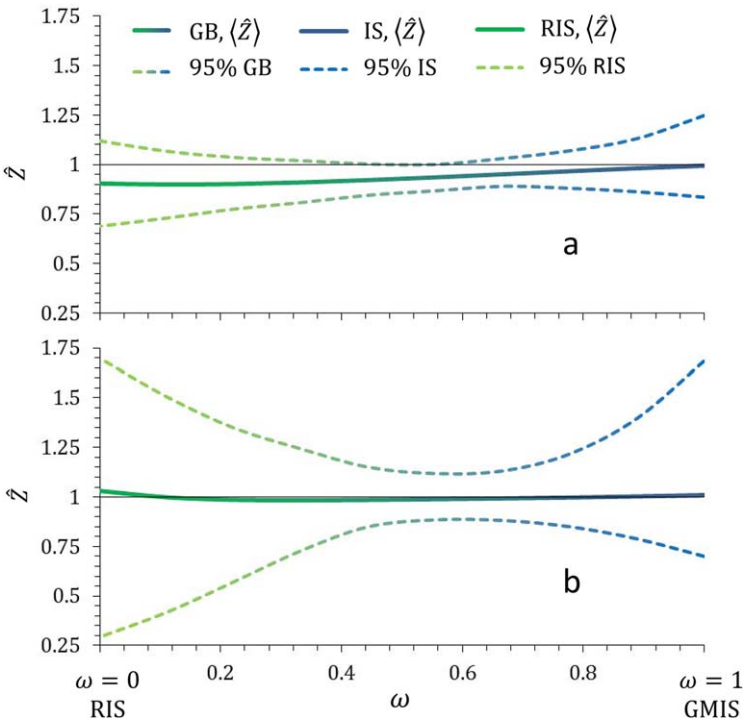

Figure 9. d-variate normal target distribution with $\rho = 0.75$ using (a) $d = 75$, and (b) $d = 100$: Mean evidence estimate, $\hat{Z}$ (solid lines) and its 2.5% and 97.5% percentiles (dashed lines) as function of the scalar ω using 250 different trials with RIS ($\omega = 0$ and $m_1 = 1000$), IS ($\omega = 1$ and $m_0 = 5000$), and GB ($\omega \in [0, 1]$).

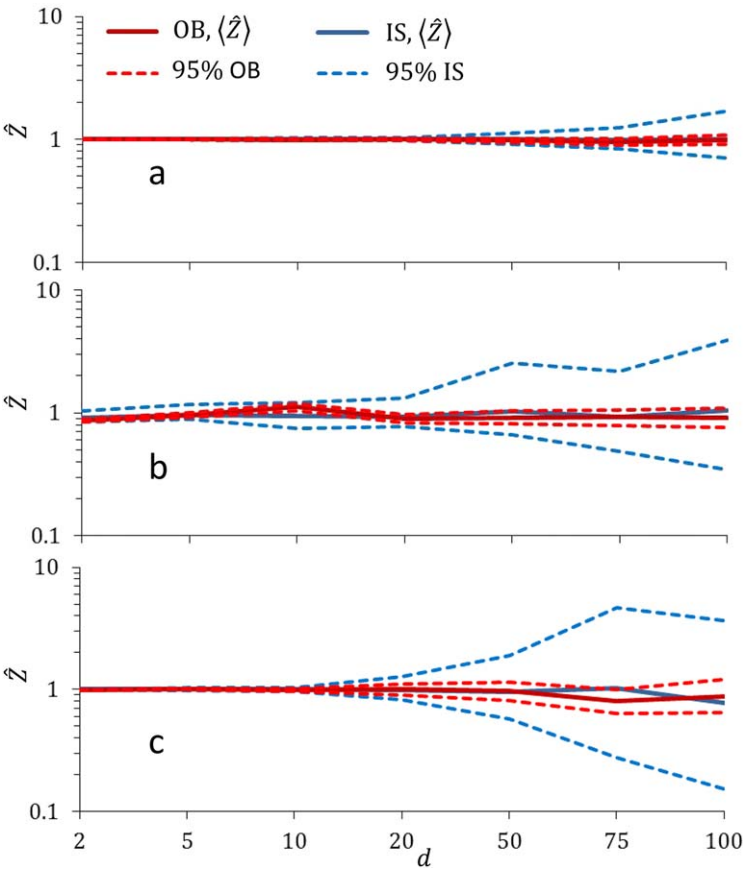

Figure 10. Trace plots of the mean evidence estimate, $\hat{Z}$ (solid lines) and its 2.5% and 97.5% percentiles (dashed lines) as function of the dimensionality, $d \in [2, 100]$ of the (a) d-variate normal target distribution with $\rho = 0.75$, (b) d-variate twisted normal target distribution, and (c) d-variate normal mixture target with disconnected modes, using 250 independent trials with IS (blue; $m_0 = 5000$ as in Figures 6 and 7) and OB (red).

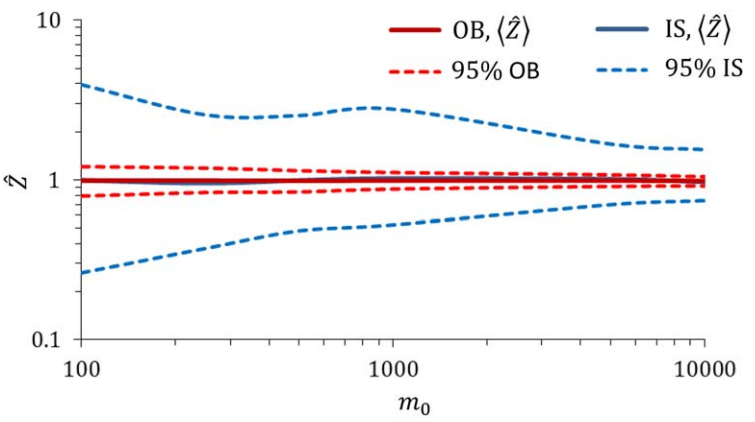

Figure 11. $d = 100$ variate normal target distribution with $\rho = 0.75$: Evolution of mean evidence estimates, $\hat{Z}$ (solid lines) and its 2.5% and 97.5% percentiles (dashed lines) as function of m_0, the number of samples from the importance distribution using 250 independent trials with IS (blue) and OB (red; $m_1 = 1000$). Thus, the ratio, m_0/m_1, will increase from 0.1 to 10 from left to right across the plot.

6. Real-World Case Study: The Rainfall-Runoff Transformation

We now apply GAME sampling to a real-world case study involving the modeling of the rainfall-discharge relationship of the Leaf River watershed in Mississippi, US. This temperate, 1944 km^2, catchment has been studied extensively in the water resources literature, which allows for comparative analysis against published results. Here, we are especially concerned with model selection and evaluate, compare, and contrast the evidence estimates and model rankings derived from GAME sampling with results of information criteria such as AIC and BIC.

6.1. Hydrologic Data and Conceptual Watershed Models

We simulate the rainfall-runoff transformation for a 10 year historical record (1 October 1952 to 30 September 1962) with daily data of discharge (m^3/s), mean areal precipitation (mm/d), and mean areal potential evapotranspiration (mm/d) using the $d = 5$ parameter HYMOD [*Boyle*, 2001], $d = 7$ parameter HMODEL [*Schoups and Vrugt*, 2010], and $d = 14$ parameter SAC-SMA [*Burnash et al.*, 1973] conceptual watershed models. Interested readers are referred to the cited publications for a detailed description of each model.

6.2. Prior, Likelihood, and Posterior Parameter Distribution

We assume the prior distribution, $p(\boldsymbol{\theta})$, of the HYMOD, HMODEL, and SAC-SMA parameters to be d-variate uniform, $\mathcal{U}_d(\mathbf{a}, \mathbf{b})$, on the bounded search domain $\Omega \in \mathbb{R}^d$, with d-vectors of upper, **b**, and lower, **a**, parameter limits listed in

Vrugt et al. [2008], *Schoups and Vrugt* [2010], and *Vrugt et al*. [2009], respectively. We now use the measured daily discharge data to estimate each model's posterior parameter distribution. For the time being, we resort to a simple Gaussian likelihood function, $L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ as we expect inadvertently each model's discharge residuals to be independent, normally distributed, and with a constant variance. Then, MCMC simulation with the DREAM algorithm results in a collection of m posterior samples, $\{\Theta_*\}=(\{\boldsymbol{\theta}_*^1\},\ldots,\{\boldsymbol{\theta}_*^m\})$, and their unnormalized densities, $p(\{\boldsymbol{\theta}_*^j\})L(\{\boldsymbol{\theta}_*^j\}|\tilde{\mathbf{Y}})$, where $j=(1,\ldots,m)$. We can relax the strong assumptions of independence and normality of the streamflow residuals by using a more sophisticated likelihood function with nuisance variables [see *Schoups and Vrugt*, 2010], but this is beyond the scope of the present work.

In our DREAM trials, we use $N=10$ chains for HYMOD and HMODEL and $N=20$ chains for the SAC-SMA model with $T=2000$, $T=4000$ and $T=20{,}000$ samples, respectively, in each chain. Convergence of DREAM to the target distribution is monitored using a variety of built-in diagnostics. The first half of each sampled chain is used as burn-in, resulting in a total of $m=10{,}000$, $m=20{,}000$, and $m=200{,}000$ realizations of the posterior parameter distribution of HYMOD, HMODEL, and the SAC-SMA model, respectively. To assess the relationship between the length of the streamflow data record and the evidence of each model, we consider calibration data sets that vary in length between 40 and 730 days.

6.3. Hydrologic Model Selection Using GAME Sampling and Information Criteria

We express the model evidence on a logarithmic scale to facilitate comparison between their estimates derived from GAME sampling and their values calculated separately with information criteria. For AIC and BIC, the evidence, $\hat{Z}$, or marginal likelihood, $p(\tilde{\mathbf{Y}})$, satisfies $\hat{Z}=\exp(-1/2I)$, which is equivalent to $I=-2\ln\{\hat{Z}\}$ [*Schöniger et al*., 2014]. Per equation (1), we can now separate the evidence into two terms, $2\ln\{\hat{Z}\}=2\ln\{L(\{\check{\boldsymbol{\theta}}_*\}|\tilde{\mathbf{Y}})\}-C$. The first term, $2\ln\{L(\{\check{\boldsymbol{\theta}}_*\}|\tilde{\mathbf{Y}})\}$, summarizes the model's goodness-of-fit via the likelihood maximum. This value is found at the mode, $\check{\boldsymbol{\theta}}_*$, of the likelihood function and does not necessarily maximize the model's posterior density, as information criteria generally preclude the use of prior information. The second term, C, penalizes for model complexity, and varies among the different evidence estimation methods (see Table 4). This penalty term is easy to compute for AIC and BIC and does not demand knowledge of the actual target distribution. The penalty terms of the other three evidence estimation methods (LM, RIS and IS) require command of the likelihood maximum, $L(\check{\boldsymbol{\theta}}_*|\tilde{\mathbf{Y}})$, and the marginal likelihood, $\hat{Z}$ (for RIS and IS).

6.4. Results

In this section, we present and discuss the results of the three watershed models. We report only the mean evidence estimates of IS and RIS. The confidence intervals of these mean estimates are relatively small due to the large number of posterior samples being used (as discussed in section 5.1.1) and the steadily growing values of the absolute log-likelihood (increasing length of calibration data set). As the three watershed models are rather parsimonious, we expect the evidence estimates of LM, RIS, and IS to be rather similar. This is particularly true, if each model's posterior parameter distribution is well described by a multivariate Gaussian, that is, if a large amount of informative data is used for parameter estimation [see, e.g., *Kass and Raftery*, 1995]. Conversely, we may expect the model evidence estimates of AIC and BIC to deviate from their values of LM, RIS, and IS.

Figure 12 summarizes the results of our analysis and depicts graphically the relationship between the length, n, of the discharge calibration data record and the value of the penalty term (left y axis) of each evidence estimation method (colored lines) and the goodness-of-fit of each watershed model (black lines) expressed as $2\ln\{L(\{\check{\boldsymbol{\theta}}_*\})\}$ on the right y axis. The solid, dashed, and dotted line types discriminate between the results of HYMOD (in top graph) and those of the HMODEL and SAC-SMA models (in bottom graph), respectively. In general, an enhanced model complexity is supported by the data if the gain in the goodness-of-fit exceeds increments of the penalty term.

Table 4. Penalty Term for Model Complexity of Different Model Selection Methods

Method	Penalty for Model Complexity
AIC	$2d$[a]
BIC	$d\ln n$
ML using Laplace-Metropolis	$-2\ln\{p(\check{\boldsymbol{\theta}}_*)\}-d\ln(2\pi)-\ln\lvert\boldsymbol{\Sigma}_*\rvert$
ML using RIS	$-2\ln\{\hat{Z}\}+2\ln\{L(\boldsymbol{\theta}_*\|\tilde{\mathbf{Y}})\}$[b]
ML using IS	$-2\ln\{\hat{Z}\}+2\ln\{L(\check{\boldsymbol{\theta}}_*\|\tilde{\mathbf{Y}})\}$[c]

[a] $+2d(d+1)/(n-d-1)$ for $n<40$ d (AICc) [*Burnham and Anderson*, 2004].
[b] $\hat{Z}$ is computed with equation (7).
[c] $\hat{Z}$ is computed with equation (10) where $q_0(\boldsymbol{\theta})=\hat{p}_{\text{mix}}(\boldsymbol{\theta})$ is given by equation (12).

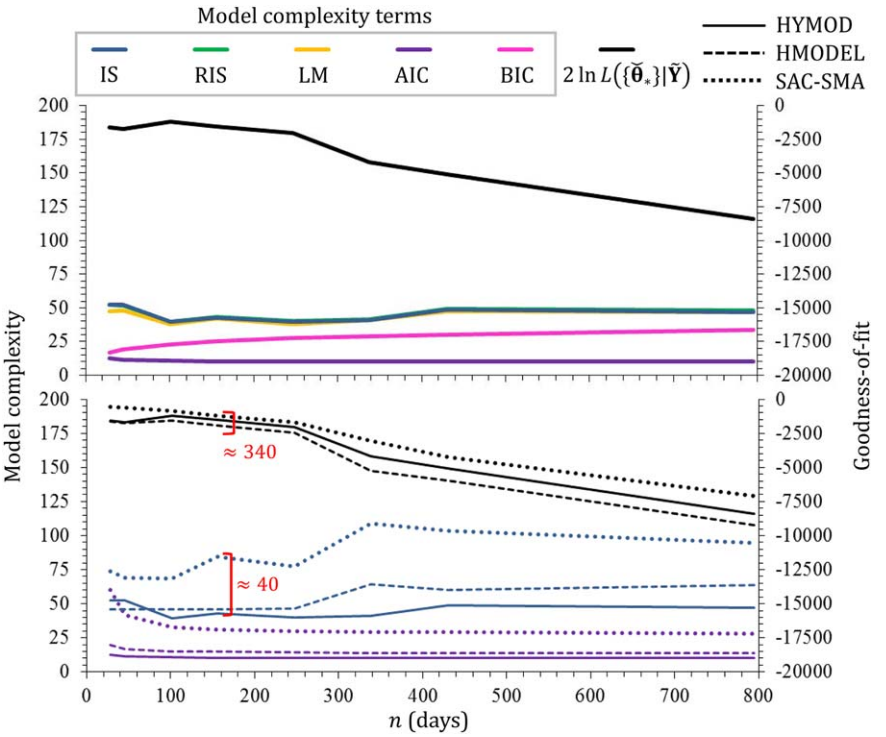

Figure 12. The penalty term (on left *y* axis) of AIC (purple lines), BIC (pink lines), RIS (green lines), IS (blue lines), and LM (yellow lines) and the goodness-of-fit, $2\ln L(\tilde{\boldsymbol{\theta}}_*)$ (on right *y* axis) of each watershed model (black lines) as function of the length, *n*, of the discharge calibration record. Line style differentiates between the results of HYMOD (solid lines in top and bottom graph), HMODEL (dashed lines in bottom graph), and SAC-SMA (dotted lines in bottom graph). As IS and RIS do not possess a separate term to combat model complexity, we estimate their individual penalty terms using Table 4. In general, models with a relatively high goodness-of-fit and small penalty term receive most support from the discharge data.

The results in this figure highlight several important findings. We first summarize the results of HYMOD in the top plot. First, and as expected, the penalty and goodness-of-fit terms decrease with increasing length, *n*, of the calibration data record. The exact shape of this dependency is determined by the choice of discharge observations if the calibration data set is small, but approaches linearity for larger values of *n*. Second, the dispersion (e.g., spread) of each model's posterior parameter distribution, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, decreases with increasing length of the calibration record (not shown). This causes the optimal mixture model, $\hat{p}_{\text{mix}}(\boldsymbol{\theta})$, to collapse to a multivariate normal distribution with a single peak (e.g., $J = 1$). Third, the penalty term of AIC appears constant and unaffected by the length of the calibration record, except if $n < 40$ d (see Table 4). For all other evidence estimation methods, the penalty term for model complexity increases with *n*. Consequently, for a given length of the calibration data record, the penalty term of BIC will be more severe than its counterpart of AIC. This conclusion is readily confirmed by the expressions of their penalty terms in Table 4. Fourth, if, per the findings of our benchmark experiments, we consider IS to be our reference solution, then AIC and BIC condemn insufficiently model complexity [cf., *Schöniger et al*., 2014], whereas the penalty terms of LM and RIS are spot on, especially for larger calibration data sets. We recall here that one of the key assumptions of AIC and BIC is that model complexity depends only on the length of the calibration data set, and not on the actual information content (and thus dynamics) of this data.

We now move on to the bottom graph of Figure 12 and present the results of the HMODEL and SAC-SMA watershed models. These results appear qualitatively similar as those discussed previously for HYMOD but with an enlarged evidence for the HMODEL and SAC-SMA models, and larger discrepancies between the different methods. Indeed, for $n \simeq 150$ we obtain logarithmic values of the evidence, $\ln\{\hat{Z}\}$ that equate to -786, -981, and -635 for HYMOD, HMODEL, and SAC-SMA, respectively. The differences between the model selection methods are irrelevant when comparing the three watershed models. All methods assign the largest evidence to the SAC-SMA model in response to its superior goodness-of-fit and thus strongest ability to describe the observed discharge data. This result generalizes the findings of *Vrugt and Robinson*

[2007] to different lengths of the calibration data record. As expected, the model complexity term increases with the number of "free" parameters, d, except for IS when $n < 100$. For the sake of clarity, we display only the penalty terms of AIC (lowest values) and IS (highest values). However, in this case the differences in goodness-of-fit between the three watershed models are much larger than their differences in model complexity. Under such circumstances, information criteria such as AIC and BIC will single-out correctly the "best" model at a computational cost that is lower than required for more advanced sampling methods (as in *Marshall et al*. [2005]).

Note, that the smaller the length of the calibration data set, the smaller the magnitude of the goodness-of-fit, and the more model selection is governed by model complexity. In this situation with small n, and with the use of parameter-rich models, accurate computation of the evidence becomes of paramount importance. For example, for the limiting case with $n = 1$ calibration measurement, the LM method erroneously ranks highest the HMODEL, whereas RIS and IS lend most support to HYMOD as the most parsimonious model of the ensemble. Information criteria such as AIC and BIC have another weakness, and that is, that they do not allow for the use of prior information. This complicates severely their application to the ranking and selection of parameter-rich models whose ability to describe system behavior can depend crucially on our ability to properly constrain parameters via informative priors.

Before we summarize the main results of this paper, we discuss briefly possible limitations of the proposed GAME sampler. Of course, no evidence estimation method would protect us against the use of an incorrect posterior parameter distribution. Thus, a requirement of GAME is access to an adequate collection of Monte Carlo samples of the target distribution. Otherwise, the performance of RIS will be severely compromised. The use of a second collection samples from the mixture distribution offers IS, GB, and OB more protection against an imperfect description of the target distribution. Nevertheless, these methods also deteriorate rapidly when confronted with an incorrect target distribution. What is more, GAME sampling may suffer if the target distribution is highly irregular with features that are complicated and not easily emulated with a mixture distribution of normal densities. In theory, a normal mixture should be able to approximate closely any target distribution, yet, in practice, this may require a much larger number of components, J, than tolerated in this paper ($J = 5$). Obviously, the results of the GAME sampler also depend on the values of its algorithmic variables. The method may not produce accurate results if, among others, an insufficient number of target samples, m_0, is used, or a too sparse supply, m_1, of mixture realizations.

As a final note, it is important to stress that model support is determined by the choice of prior distribution and likelihood function. An improper choice of these two antecedents will affect the target distribution and thus model evidence estimates. This is by no means unique to GAME sampling, instead, it will affect all model evidence methods. Whereas the choice of the prior distribution will, to some extent, always necessitate subjective decisions regarding each parameter's ranges and distribution, the adequacy of the likelihood function can be carefully scrutinized using residual analysis. If, as in the last case study, parameters evoke physical and/or conceptual entities, then this simplifies the assignment of their ranges.

7. Summary and Conclusions

This paper has presented a new methodology for estimating the marginal likelihood of a model. This so called model evidence provides a single quantitative measure of model support which integrates model accuracy, uncertainty, and complexity, and is of key importance in Bayesian model selection. The proposed approach was coined GAussian Mixture importancE, or GAME, sampling, and uses multidimensional integration of the posterior parameter distribution to efficiently estimate the marginal likelihood. First, we generate a large collection of samples from the target distribution using Markov chain Monte Carlo (MCMC) simulation with the DREAM algorithm. Then, a mixture of normal densities is fitted to the posterior samples using maximum likelihood estimation of the weight, mean, and free elements of the covariance matrix of each mixture component. The optimal mixture distribution can be determined using information-theoretic measures or via minimization of the variance of the distance between the target and mixture density. Finally, the optimal mixture distribution serves as bridge distribution in bridge sampling, and returns estimates of the model evidence.

Bridge sampling is a generalization of importance sampling and uses an intermediate density, or bridge density, to transition smoothly between the importance distribution and the target of interest. The user is

free to select the bridge density, $q_{1/2}(\boldsymbol{\theta})$, via the identity, $q_{1/2}(\boldsymbol{\theta})=q_0(\boldsymbol{\theta})^{1-\omega}q_1(\boldsymbol{\theta})^{\omega}$, using the scalar $\omega \in [0, 1]$. As extreme, or limiting, cases, this results in reciprocal importance sampling (RIS, $\omega = 0$) or importance sampling (IS, $\omega = 1$), and with $\omega \in (0, 1)$ this amounts to a transient case using bridge sampling with a geometric bridge (GB) or optimal bridge (OB). The GAME sampler has been implemented in the DiffeRential Evolution Adaptive Metropolis (DREAM) MATLAB toolbox of *Vrugt* [2016] and simplifies considerably hypothesis testing and model selection.

A wide range of numerical experiments were conducted to benchmark the performance of GAME sampling on target distributions with known normalizing constants. This includes multivariate normal distributions with variable dimensionality (up to $d = 100$), one or two (disconnected) modes, and variably correlated, twisted, and/or truncated dimensions. We also presented a real-world case study involving the application of GAME sampling to Bayesian model selection of the rainfall-discharge response of the Leaf River watershed in the US. This study also compared the results of the GAME sampler with model evidence estimates and rankings derived separately from information theory using Akaike's information criterion (AIC) and Bayesian information criterion (BIC). The main conclusions can be summarized as follows.

1. GAME addresses an important practical problem in importance sampling, namely that of choosing an appropriate scale and orientation of the importance distribution. The importance distribution is construed by fitting a mixture of normal densities to a large collection of target samples derived separately from Markov chain Monte Carlo simulation with the DREAM algorithm.
2. The RIS, IS, GB, and OB evidence estimators do not invoke the same computational cost. In general, RIS is most CPU-efficient as the collection of target samples suffices to compute the model evidence. The other three methods (IS, GB, and OB) necessitate the use of a second collection of samples drawn randomly from the optimal mixture (= importance) distribution. As the unnormalized density of the target distribution must be evaluated for these mixture samples, this enhances computational demands.
3. For all the case studies analyzed herein, IS provides unbiased estimates of the evidence with an estimation uncertainty that increases with complexity and dimensionality of the posterior parameter distribution. The evidence estimates of IS appeared unaffected by the choice of selection criteria for the optimal mixture distribution.
4. The evidence estimates of the Laplace-Metropolis (LM) method have to be interpreted with care, and are particularly suspicious if the target distribution does not satisfy normality. Examples include twisted, truncated, and multimodal posterior distributions.
5. A poor mixture distribution will lead to importance densities of the target samples that deviate considerably from their actual unnormalized values. This mismatch will directly corrupt the evidence estimates of RIS, but not affect much the performance of IS, as this latter method does not use the importance density of the target samples but rather works with the target density of the mixture samples.
6. The use of a second collection of samples from the mixture distribution by IS enhances considerably the diversity of the target approximation, but at the expense of an increased computational cost of GAME sampling.
7. The efficiency of IS can be further enhanced by embedding the method in a bridge sampling framework. OB preserves the accuracy of IS while significantly reducing the variance of the model evidence estimates. This is true even if the posterior realizations sampled by DREAM do not satisfy independence, as required by OB to guarantee, at least in theory, an optimal performance of this evidence estimation method.
8. OB provides robust estimates of the evidence for a range of different quotients between the number of target samples and the number of importance samples. This ratio controls the bridge distribution.
9. One should be particularly careful in using information criteria such as AIC and BIC for model selection purposes. The evidence estimates of these metrics can deviate considerably from their values derived from GAME sampling, and should not be relied upon for hypothesis testing and model selection under general conditions. This is especially true with the use of informative priors and/or parameter-rich models.
10. Information criteria such as AIC and BIC underestimate the value of the penalty term that is used to combat model complexity, even if these metrics yield the same model ranking.
11. GAME sampling provides a method for evidence computation that returns a robust estimate of the Bayesian model evidence under general conditions.

Acknowledgments
The authors would like to thank the associate editor, reviewers, Ming Ye, Ahmed S. Elshall, Marvin Höge, and one anonymous referee for their valuable comments and suggestions that have helped to further improve the quality of the paper. E. V. acknowledges the Italian Ministry of University and Research for partially funding this research through project PRIN 20102AXKAJ. The last author greatly appreciates the support from the UC-Lab Fees Research Program Award 237285. The data of the Leaf River watershed can be found at ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/.

References

Akaike, H. (1998), Information theory and an extension of the maximum likelihood principle, in *Selected Papers of Hirotugu Akaike*, edited by E. Parzen, K. Tanabe, and G. Kitagawa, chap. 13, pp. 199–213, Springer, New York.
Bishop, C. (2006), *Pattern Recognition and Machine Learning*, 738 pp., Springer, Singapore.
Boyle, D. (2001), Multicriteria calibration of hydrologic models, PhD thesis, Univ. of Ariz., Tucson.
Burnash, R., R. Ferral, and R. McGuire (1973), A generalized streamflow simulation system: Conceptual modeling for digital computers, technical report, U.S. Dep. of Commer., Natl. Weather Serv. and State of Calif., Dep. of Water Resour., Sacramento, Calif.
Burnham, K., and D. Anderson (2002), *Model Selection and Multivariate Inference: A Practical Information Theoretic Approach*, 488 pp., Springer, New York.
Burnham, K., and D. Anderson (2004), Multimodel inference: Understanding AIC and BIC in model selection, *Sociol. Methods Res.*, *33*(2), 261–304, doi:10.1177/0049124104268644.
Clark, M., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.
Di Ciccio, T., R. Kass, A. Raftery, and L. Wasserman (1997), Computing Bayes factors by combining simulation and asymptotic approximations, *J. Am. Stat. Assoc.*, *92*(439), 903–915.
Evans, M., and T. Swartz (1995), Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems, *Stat. Sci.*, *10*(3), 254–272.
Fan, Y., R. Wu, M.-H. Chen, L. Kuo, and P. Lewis (2011), Choosing among partition models in Bayesian phylogenetics, *Mol. Biol. Evol.*, *28*(1), 523–532, doi:10.1093/molbev/msq224.
Gelfand, A., and D. Dey (1994), Bayesian model choice, *J. R. Stat. Soc. Ser. C*, *56*(3), 501–514.
Gelman, A., and X.-L. Meng (1998), Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, *Stat. Sci.*, *13*(2), 163–185.
Hammersley, J., and D. Handscom (1964), *Monte Carlo Methods*, Fletcher & Sons Ltd., Norwich.
Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–417.
Johnson, J., and K. Omland (2004), Model selection in ecology and evolution, *Trends Ecol. Evol.*, *19*(2), 101–108.
Kass, R., and A. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, *90*(430), 773–795.
Kavetski, D., G. Kuczera, and S. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resour. Res.*, *42*, W03408, doi:10.1029/2005WR004376.
Kavetski, D., G. Kuczera, and S. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.
Kirchner, J. W. (2009), Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, *Water Resour. Res.*, *45*, W02429, doi:10.1029/2008WR006912.
Laloy, E., and J. Vrugt (2012), High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing, *Water Resour. Res.*, *1*, W01526, doi:10.1029/2011WR010608.
Lewis, S., and A. Raftery (1997), Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator, *J. Am. Stat. Assoc.*, *92*(468), 648–655, doi:10.1080/01621459.1997.10474016.
Liu, P., A. S. Elshall, M. Ye, P. Beerli, X. Zeng, D. Lu, and Y. Tao (2016), Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods, *Water Resour. Res.*, *52*, 734–758, doi:10.1002/2014WR016718.
Lu, D., M. Ye, and S. Neuman (2011), Dependence of Bayesian model selection criteria and Fisher information matrix on sample size, *Math. Geosci.*, *43*, 971–993, 10.1007/s11004–011-9359–0.
Marshall, L., D. Nott, and A. Sharma (2005), Hydrological model selection: A Bayesian alternative, *Water Resour. Res.*, *41*, W10422, doi:10.1029/2004WR003719.
McLachlan, G., and T. Krishnan (2007), *The EM Algorithm and Extensions*, 2nd ed., 359 pp., John Wiley, Hoboken, N. J.
Meng, X.-L., and W. Wong (1996), Simulating ratios of normalizing constants via a simple identity: A theoretical exploration, *Stat. Sin.*, *8*(4), 831–860.
Neal, R. M. (2001), Annealed importance sampling, *Stat. Comput.*, *11*(2), 125–139, doi:10.1023/A:1008923215028.
Neddermeyer, J. C. (2009), Computationally efficient nonparametric importance sampling, *J. Am. Stat. Assoc.*, *104*(486), 788–802, doi:10.1198/jasa.2009.0122.
Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environ. Res. Risk Assess.*, *17*(5), 291–305, doi:10.1007/s00477-003-0151-7.
Newton, M., and A. Raftery (1994), Approximate Bayesian inference with the weighted likelihood bootstrap, *J. R. Stat. Soc. Ser. B*, *56*(1), 3–48.
Ogata, Y. (1989), A Monte Carlo method for high dimensional integration, *Numer. Math.*, *55*, 137–157.
Perrakis, K., I. Ntzoufras, and E. Tsionas (2014), On the use of marginal posteriors in marginal likelihood estimation via importance sampling, *Comput. Stat. Data Anal.*, *77*, 54–69, doi:10.1016/j.csda.2014.03.004.
Popper, K. (1992), Simplicity, in *The Logic of Scientific Discovery*, chap. 7, pp. 121–132, Routledge, London.
Raftery, A. (1996), Hypothesis testing and model selection via posterior simulation, in *Markov Chain Monte Carlo in Practice*, edited by W. Gilks, S. Richardson, and D. Spiegelhalter, chap. 10, pp. 163–188, Chapman and Hall, London.
Refsgaard, J., J. van der Sluijs, J. Brown, and P. van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Adv. Water Resour.*, *29*(11), 1586–1597, doi:10.1016/j.advwatres.2005.11.013.
Schöniger, A., T. Wöhling, L. Samaniego, and W. Nowak (2014), Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence, *Water Resour. Res.*, *60*, 9484–9513, doi:10.1002/2014WR016062.
Schoups, G., and J. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.
Schoups, G., N. Van de Giesen, and H. Savenije (2008), Model complexity control for hydrologic prediction, *Water Resour. Res.*, *44*, W00B03, doi:10.1029/2008WR006836.
Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.
Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002), Bayesian measures of model complexity and fit (with discussion), *J. R. Stat. Soc. Ser. B*, *64*(4), 583–639.
Tokdar, S., and R. Kass (2010), Importance sampling: A review, *WIREs Comput. Stat.*, *2*(1), 54–60, doi:10.1002/wics.56.
Vrugt, J., and B. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*, W01411, doi:10.1029/2005WR004838.

Vrugt, J., and C. ter Braak (2011), DREAM(D): An adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems, *Hydrol. Earth Syst. Sci.*, *15*, 3701–3713, doi:10.5194/hess-15-3701-2011.

Vrugt, J., C. ter Braak, M. Clark, J. Hyman, and B. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Vrugt, J., C. ter Braak, C. Diks, B. Robinson, J. M. Hyman, and D. Higdon (2009), Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), doi:10.1515/IJNSNS.2009.10.3.273.

Vrugt, J. A. (2016), Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, *Environ. Modell. Software*, *75*, 273–316, doi:10.1016/j.envsoft.2015.08.013.

Wasserman, L. (2000), Bayesian model selection and model averaging, *J. Math. Psychol.*, *44*(1), 92–107, doi:10.1006/jmps.1999.1278.

Xie, W., P. Lewis, Y. Fan, L. Kuo, and M.-H. Chen (2011), Improving marginal likelihood estimation for Bayesian phylogenetics model selection, *Syst. Biol.*, *60*(2), 150–160, doi:10.1093/sysbio/syq085.

Ye, M., P. Meyer, and S. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resour. Res.*, *44*, W03428, doi:10.1029/2008WR006803.