Statistical Science
2014, Vol. 29, No. 3, 397–419
DOI: 10.1214/13-STS465

Recursive Pathways to Marginal Likelihood Estimation with Prior-Sensitivity Analysis

Ewan Cameron and Anthony Pettitt

Abstract. We investigate the utility to computational Bayesian analyses of a particular family of recursive marginal likelihood estimators characterized by the (equivalent) algorithms known as "biased sampling" or "reverse logistic regression" in the statistics literature and "the density of states" in physics. Through a pair of numerical examples (including mixture modeling of the well-known galaxy data set) we highlight the remarkable diversity of sampling schemes amenable to such recursive normalization, as well as the notable efficiency of the resulting pseudo-mixture distributions for gauging prior sensitivity in the Bayesian model selection context. Our key theoretical contributions are to introduce a novel heuristic ("thermodynamic integration via importance sampling") for qualifying the role of the bridging sequence in this procedure and to reveal various connections between these recursive estimators and the nested sampling technique.

Key words and phrases: Bayes factor, Bayesian model selection, importance sampling, marginal likelihood, Metropolis-coupled Markov Chain Monte Carlo, nested sampling, normalizing constant, path sampling, reverse logistic regression, thermodynamic integration.

1. INTRODUCTION

Though typically unnecessary for computational parameter inference in the Bayesian framework, the factor, Z, required to normalize the product of prior and likelihood nevertheless plays a vital role in Bayesian model selection and model averaging (Kass and Raftery, 1995; Hoeting et al., 1999). For priors admitting an "ordinary" density, $\pi(\theta)$, with respect to the Lebesgue measure (a "Λ-density"), we write for the posterior

$$\pi(\theta|y) = L(y|\theta)\pi(\theta)/Z \quad \text{with} \tag{1}$$
$$Z = \int_\Omega L(y|\theta)\pi(\theta)\,d\theta,$$

and, more generally (e.g., for stochastic process priors) we write

$$dP_{\theta|y}(\theta) = L(y|\theta)\,dP_\theta(\theta)/Z \quad \text{with} \tag{2}$$
$$Z = \int_\Omega L(y|\theta)\{dP_\theta(\theta)\},$$

with the likelihood, $L(y|\theta)$, a non-negative, real-valued function supposed integrable with respect to the prior. In this context Z is generally referred to as either the *marginal likelihood* (i.e., the likelihood of the observed data marginalized [averaged] over the prior) or the *evidence*. With the latter term though, one risks the impression of overstating the value of this statistic in the case of limited prior knowledge (cf. Gelman et al., 2004, Chapter 6).

Problematically, few complex statistical problems admit an analytical solution to Equations (1) or (2), or span such low-dimensional spaces [$D(\theta) \lesssim 5$–10] that direct numerical integration presents a viable alternative. With errors (at least in principle) independent of dimension, Monte Carlo-based integration methods have thus become the mode of choice for marginal

Ewan Cameron is Research Associate, Science and Engineering Faculty, School of Mathematical Sciences (Statistical Science), Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia (e-mail: dr.ewan.cameron@gmail.com). Anthony Pettitt is Professor, Science and Engineering Faculty, School of Mathematical Sciences (Statistical Science), Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia.

likelihood estimation across a diverse range of scientific disciplines, from evolutionary biology (Xie et al., 2011; Arima and Tardella, 2012; Baele et al., 2012) and cosmology (Mukherjee, Parkinson and Liddle, 2006; Kilbinger et al., 2010) to quantitative finance (Li, Ni and Lin, 2011) and sociology (Caimo and Friel, 2013).

1.1 Monte Carlo-Based Integration Methods

With the posterior most often "thinner-tailed" than the prior and/or constrained within a much diminished sub-volume of the given parameter space, the simplest marginal likelihood estimators drawing solely from $\pi(\theta)$ or $\pi(\theta|y)$ cannot be relied upon for model selection purposes. In the first case—strictly, that $\int_\Omega [1/L(y|\theta)]\pi(\theta)\,d\theta$ diverges—the harmonic mean estimator (HME; Newton and Raftery, 1994),

$$\hat{Z}^H = \left[\sum_{i=1}^{n} 1/n/L(y|\theta_i)\right]^{-1} \quad \text{for } \theta_i \sim \pi(\theta|y),$$

suffers *theoretically* from an infinite variance, meaning *in practice* that its convergence toward the true Z as a one-sided α-stable limit law can be incredibly slow (Wolpert and Schmidler, 2012). Even when "robustified" as per Gelfand and Dey (1994) or Raftery et al. (2007), however, the HME remains notably insensitive to changes in $\pi(\theta)$, whereas Z itself is characteristically sensitive (Robert and Wraith, 2009; Friel and Wyse, 2012). [See also Weinberg (2012) for yet another approach to robustifying the HME.] Though assuredly finite by default, the variance of the prior arithmetic mean estimator (AME),

$$\hat{Z}^A = \sum_{i=1}^{n} L(y|\theta_i)/n \quad \text{for } \theta_i \sim \pi(\theta),$$

on the other hand, will remain impractically large whenever there exists a substantial difference in "volume" between the regions of greatest concentration in prior and posterior mass, with huge sample sizes necessary to achieve reasonable accuracy (e.g., Neal, 1999).

A wealth of more sophisticated integration methods have thus lately been developed for generating improved estimates of the marginal likelihood, as reviewed in depth by Chen, Shao and Ibrahim (2000), Robert and Wraith (2009) and Friel and Wyse (2012). Notable examples include the following: adaptive multiple importance sampling (Cornuet et al., 2012), annealed importance sampling (Neal, 2001), bridge sampling (Meng and Wong, 1996), [ordinary] importance sampling (cf. Liu, 2001), path sampling/thermodynamic integration (Gelman and Meng, 1998; Lartillot and Phillipe, 2006; Friel and Pettitt, 2008; Calderhead and Girolami, 2009), nested sampling (Skilling, 2006; Feroz and Hobson, 2008), nested importance sampling (Chopin and Robert, 2010), reverse logistic regression (Geyer, 1994), sequential Monte Carlo (SMC; Cappé et al., 2004; Del Moral, Doucet and Jasra, 2006), the Savage–Dickey density ratio (Marin and Robert, 2010) and the density of states (Habeck, 2012; Tan et al., 2012). A common thread running through almost all these schemes is the aim for a superior exploration of the relevant parameter space via "guided" transitions across a sequence of intermediate distributions, typically following a bridging path between the $\pi(\theta)$ and $\pi(\theta|y)$ extremes. [Or, more generally, the $h(\theta)$ and $\pi(\theta|y)$ extremes if a suitable auxiliary/reference density, $h(\theta)$, is available to facilitate the integration; cf. Lefebvre, Steele and Vandal (2010).] However, the nature of this bridging path differs significantly between algorithms. Nested sampling, for instance, evolves its "live point set" over a sequence of constrained-likelihood distributions, $f(\theta) \propto \pi(\theta)I(L(y|\theta) \geq L_{\lim})$, transitioning from the prior ($L_{\lim} = 0$) through to the vicinity of peak likelihood ($L_{\lim} \approx L_{\max} - \varepsilon$), while thermodynamic integration, on the other hand, draws progressively (via Markov Chain Monte Carlo [MCMC]; Tierney, 1994) from the family of "power posteriors,"

$$\pi_t(\theta|y) \propto \pi(\theta)L(y|\theta)^t, \tag{3}$$

explicitly connecting the prior at $t = 0$ to the posterior at $t = 1$.

Another key point of comparison between these rival Monte Carlo techniques lies in their choice of identity by which the evidence is ultimately computed. The (geometric) path sampling identity,

$$\log Z = \int_0^1 \mathrm{E}_{\pi_t}\{\log L(y|\theta)\}\,dt,$$

for example, is shared across both thermodynamic integration and SMC, in addition to its namesake. However, SMC can also be run with the "stepping-stone" solution (cf. Xie et al., 2011),

$$Z = \prod_{j=2}^{m} Z_{t_j}/Z_{t_{j-1}}, \quad \text{where } t_1 = 0 \text{ and } t_m = 1,$$

with $\{t_j : j = 1, \ldots, m\}$ indexing a sequence of ("tempered") bridging densities, and, indeed, this is the mode preferred by experienced practitioners (e.g., Del Moral, Doucet and Jasra, 2006). Yet another identity for computing the marginal likelihood is that of the recursive pathway explored here.

First introduced within the "biased sampling" paradigm (Vardi, 1985), the recursive pathway is shared by the popular techniques of "reverse logistic regression" (RLR) and "the density of states" (DoS). By *recursive* we mean that, algorithmically, each may be run such that the desired Z is obtained through backward induction of the complete set of intermediate normalizing constants corresponding to the sequence of distributions in the given bridging path by supposing these to be already known. That is, a stable solution may be found in a Gauss–Seidel-type manner (Ortega and Rheinboldt, 1967) by starting with a guess for each normalizing constant as input to a convex system of equations for updating these guesses, returning the new output as input to the same equations, and iterating until convergence. In fact, although the RLR and the DoS approaches differ vastly in concept and derivation—the former emerging from considerations of the reweighting mixtures problem in applied statistics (Geyer and Thompson, 1992; Geyer, 1994; Chen and Shao, 1997; Kong et al., 2003) and the latter from computational strategies for free energy estimation in physics/chemistry/biology (Ferrenberg and Swendsen, 1989; Kumar et al., 1992; Shirts and Chodera, 2008; Habeck, 2012; Tan et al., 2012)—both may be seen to recover the same algorithmic form in practice. To illustrate this equivalence, and to explain further the recursive pathway to marginal likelihood estimation, we describe each in detail below (Sections 2.1 and 2.2), though we begin with the more general biased sampling algorithm (Section 2).

Following this review of the recursive family (which includes our theoretical contributions concerning the link between the DoS and nested sampling in Section 2.2.1), we highlight the potential for efficient prior-sensitivity analysis when using these marginal likelihood estimators (Section 2.3) and discuss issues regarding design and sampling of the bridging sequence (Section 2.4). We then introduce a novel heuristic to help inform the latter by characterizing the connection between the bridging sequences of biased sampling and thermodynamic integration (Section 3). Finally, we present two numerical case studies illustrating the issues and techniques discussed in the previous sections: the first concerns a "mock" banana-shaped likelihood function (Section 4) and includes the demonstration of a novel synthesis of the recursive pathway with nested sampling (Section 4.2), while the second concerns mixture modeling of the galaxy data set (Section 5) and includes a demonstration of importance sample reweighting of an infinite-dimensional mixture posterior to recover its finite-dimensional counterparts (Section 5.4.3).

2. BIASED SAMPLING

The archetypal recursive marginal likelihood estimator—from which both the RLR and DoS methods may be directly recovered—is that of biased sampling, introduced by Vardi (1985) for finite-dimensional parameter spaces and extended to general sample spaces by Gill, Vardi and Wellner (1988). The basic premise of biased sampling is that one has available m sets of n_j i.i.d. draws, $\{\theta_i\}_j$, from a series of $w_j(\theta)$-weighted versions of a common, unknown measure, F, that is,

$$\{\theta_i\}_j \sim F_j, \quad \text{where } dF_j(\theta) = w_j(\theta)/W_j\, dF(\theta).$$

The W_j term here represents the normalization constant of the jth weighted distribution, typically unknown. As Vardi (1985) demonstrates, provided the drawn $\{\theta_i\}_j$ obey a certain graphical condition (discussed later), then there exists a unique nonparametric maximum likelihood estimator (NPMLE) for F, which as a by-product produces consistent estimates of all unknown W_j. If the common measure, F, is in fact the parameter prior, P_θ, then the choices $w_1(\theta) = 1$ and $w_m(\theta) = L(y|\theta)$ describe sampling from the prior and posterior, respectively. Hence, we switch to the notation $W_j = Z_j$ with $Z_1 = 1$ (for a proper prior) and $Z_m = Z$ for the above choices of w_1 and w_m.

For a given bridging scheme to be amenable to normalization via biased sampling, it is of course necessary that each intermediate sampling distribution be absolutely continuous with respect to the prior (i.e., $P_j \ll P_\theta$) such that the weight function corresponds to the Radon–Nikodym derivative, $w_j(\theta) = \frac{dP_j}{dP_\theta}(\theta)$. It is easy to verify then the applicability of biased sampling to, for example, (I) importance sampling from a sequence of bridging densities, $f_j(\theta)$, with (at least the union of their) supports matching but not exceeding that of a Λ-density prior, $w_j(\theta) = f_j(\theta)/\pi(\theta)$; and (II) thermodynamic integration over tempered likelihoods, $w_j(\theta) = L(y|\theta)^{\beta_j}$, for both the Λ-density and general case. In fact, if we view the likelihood function as defining a transformation of the prior, P_θ, to the measure P_L in univariate "likelihood space," $0 \leq L \leq \infty$, then such tempering may be seen as directly analogous to Vardi's example of "length biased sampling." Accordingly, Vardi's case study of $m = 2$ with $w_1 = 1$ and $w_2 = x$ (read L) equates to marginal likelihood estimation via defensive importance sampling

from the prior and posterior (Newton and Raftery, 1994; Hesterberg, 1995), while his one sample study with $w_1 = x$ (L) matches the HME.

For Bayesian analysis problems in which the prior measure is explicitly known (as opposed to being "known" only implicitly as the induced measure belonging to a well-defined stochastic process), the application of the biased sampling paradigm to the task of marginal likelihood estimation is arguably paradoxical since we make the pretence to estimate P_θ (known) in order to recover an estimate for Z (unknown). However, we would propose that an adequate justification for the use of Vardi's method in this context is already provided by the same pragmatic reasoning used to adopt *any* statistical estimator for the task of marginal likelihood computation in place of the direct approach of numerical integration (quadrature)—namely, that although Z is defined exactly by our known prior and likelihood function, we choose to treat it as if it were an unknown variable simply because the MC integration techniques this brings into play are more computationally efficient (being relatively insensitive to the dimension of the problem; cf. Liu, 2001).

Vardi's derivation of the NMPLE for the unknown F (i.e., P_θ) in biased sampling involves two key steps. The first is the observation that, as is typical of the NMPLE method in general, the resulting estimator, $d\hat{F}(\theta)$, will be strictly atomic with point masses assigned to each of the sampled θ_i (also called a histogram estimate of F). The second is that the normalization constants for each W_j corresponding to the atomic $d\hat{F}(\theta)$ can then be learned via an appropriately weighted summation over *all* the observed θ_i (not just those from the corresponding jth distribution). In the notation for our marginal likelihood estimation scenario, Vardi (1985) shows that the estimation problem for $d\hat{P}_\theta(\theta) = \{p_i\}_j$ can ultimately be reduced to the maximization of the following log-likelihood function,

$$\log \mathcal{L}(p) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log\bigl(w_j(\{\theta_i\}_j)\{p_i\}_j / \hat{Z}_j\bigr),$$

subject to the constraints, $\sum_{j=1}^{m} \sum_{i=1}^{n_j} \{p_i\}_j = 1$ and all $\{p_i\}_j > 0$ [see Vardi's Equation (2.2), where we avoid his explicit treatment of matching θ_i draws, implicitly allowing multiple point mass contributions at the same θ_i to give a summed contribution to the atomic $d\hat{P}_\theta(\theta)$].

Importantly, the resulting biased sampling estimator for the unknown Z_k allows for a recursive solution via the iterative updating of initial guesses ($\hat{Z}_k > 0$) as follows:

$$\hat{Z}_k = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(w_k(\{\theta_i\}_j) \Big/ \left(\sum_{s=1}^{m} n_s w_s(\{\theta_i\}_j) / \hat{Z}_s \right) \right) \tag{4}$$

(adapted from Gill, Vardi and Wellner's 1988 Proposition 1.1c). As discussed by Vardi (1985) and Geyer (1994), the above system of $(m-1)$ equations in $(m-1)$ unknowns (given $Z_1 = 1$) with Gauss–Seidel type iterative updates is globally convergent, although the gradient and Hessian of the likelihood function are also accessible, meaning that alternative maximization strategies harnessing this information may prove more efficient within a restricted domain.

The convergence properties of the biased sampling estimator for the unknown F (i.e., P_θ) and its associated W_j (Z_j) in general state spaces (possibly infinite-dimensional) have been thoroughly characterized by Gill, Vardi and Wellner (1988) using the theory of empirical processes indexed by sets and functions (cf. Dudley and Philipp, 1983). In particular, Gill, Vardi and Wellner (1988) demonstrate a central limit theorem (CLT) for convergence of the vector of normalization estimates, $\hat{\mathbf{W}}$, to the truth, $\mathbf{W}$, as $\sqrt{N}(\hat{\mathbf{W}} - \mathbf{W}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$, where the covariance matrix, Σ, takes the form given in their Proposition 2.3 [for the case here of $Z_1 = 1$ known, otherwise their Equation (2.24)]. The sample-based estimate of this error matrix, $\hat{\Sigma}$, is easily computed from the output of a standard biased sampling simulation, and in our numerical experiments with the banana-shaped pseudo-likelihood function of Section 4 it was observed to give (on average, with an approximate transformation via Slutsky's lemma) a satisfactory, though slightly conservative, match to the sample variance of $\log \hat{Z}$ under repeat simulation, even at relatively small sample sizes.

However, as noted by Christian Robert in his discussion of Kong et al.'s (2003) "read" paper, the availability of such formulae (for the asymptotic covariance matrix) can sometimes "give a false confidence in estimators that should not be used." A canonical example is that of the HME, for which the usual importance sampling variance formula applied to the pos-

terior draws may well give a finite result, though in fact the theoretical variance is infinite (meaning that the convergence of the HME is no longer obeying the assumed CLT). In particular, for finite theoretical variance of the HME (cf. Section 1) we require that the prior is fatter tailed than the posterior such that $\int_\Omega [1/L(y|\theta)]\pi(\theta)\,d\theta < \infty$. As was recognized by Vardi (1985) and Gill, Vardi and Wellner (1988), the same condition effectively holds for the validity of the CLT for biased sampling and may be expressed as an inverse mean bias-weighted integrability requirement over the indexing class of functions or sets in its empirical process construction. Important to note in the context of marginal likelihood estimation is that provided the prior itself is contained within the weighting scheme [e.g., $w_1(\theta) = 1$], then the above condition is automatically satisfied; this of course parallels the strategy of defensive importance sampling (Newton and Raftery, 1994; Hesterberg, 1995).

Finally, we observe here the other key prerequisite for successful biased sampling: that the bridging sequence of weighting functions and the random draws from them are such that a unique NPMLE for F (P_θ) actually exists. To ensure the *asymptotic* existence of a unique NPMLE (i.e., with an unlimited number of draws from each weighted distribution), Vardi (1985) gives the following condition on the supports, $\mathrm{Supp}(w_j)$, of the bridging sequence: that there does not exist a proper subset, B, of $\{1,\dots,m\}$ such that

$$\Bigl(\bigcup_{j\in B}\mathrm{Supp}(w_j)\Bigr)\cap\Bigl(\bigcup_{j\notin B}\mathrm{Supp}(w_j)\Bigr)=\varnothing.$$

In effect, the set of bridging distributions must overlap in such a way that the relative normalization of each with respect to all others will be inevitably constrained by the data. This condition is again satisfied automatically if the support of at least one of the bridging distributions encompasses all others, such as that of the prior or an equivalent reference density. In the finite sample sizes of real-world simulation the above must be strengthened to specify that the drawn $\{\theta_i\}_j$ do in fact cover each critical region of overlap. Formally, Vardi (1985) introduces a requirement of strong connectivity on the directed graph, G, with m vertices and edges h to j for each (h, j)-pairing, such that $w_h(\theta_k) > 0$ for some $\theta_k \in \{\theta_i\}_j$. This is equivalent to the finite sample "inseparability" condition given by Geyer (1994).

2.1 Reverse Logistic Regression

In the reweighting mixtures problem (cf. Geyer and Thompson, 1992 and Geyer, 1994) the aim is to discover an efficient proposal density for use in the importance sampling of an arbitrary target about which little is known a priori. Geyer's solution was to suggest sampling not from a single density of standard form, but rather from an *ensemble* of different densities, $f_j(\theta) = q_j(\theta)/Q_j$, for $j = 1,\dots,m$ with $q_j(\theta)$ known and Q_j typically unknown. The pooled draws, $\{\{\theta_i\}_j : i = 1,\dots,n_j;\ j = 1,\dots,m\}$, are then to be treated as if from a single mixture density, with each free normalizing constant—and hence the appropriate weighting scheme—to be derived recursively. As with biased sampling, if we suppose $q_1(\theta)$ to be the Bayesian prior (with $Q_1 = 1$) and $q_m(\theta)$ the (unnormalized) posterior (with $Q_j = Z$), the relevance of this approach to marginal likelihood estimation becomes readily apparent. In this context we write the imagined (i.e., pseudo-) mixture density, $p(\theta)$, in the form

$$p(\theta)=\sum_{j=1}^{m}[n_j/n]\bigl[q_j(\theta)/Z_j\bigr], \tag{5}$$

where $n = \sum_{j=1}^m n_j$.

The recursive normalization scheme introduced by Geyer (1994) for this purpose is based on maximization in $\{Z_2,\dots,Z_m\}$ (i.e., $[\mathbb{R}_+]^{m-1}$) of the following *quasi*-log-likelihood function representing the likelihood of each set of $\{\theta_i\}_j$ having been drawn from its true $f_j(\theta)$ rather than some other $f_{k[\neq j]}(\theta)$ in the pseudo-mixture:

$$\begin{aligned}&\log L\bigl(\{\{\theta_i\}_j : i = 1,\dots,n_j;\ j = 1,\dots,m\}\,|\\ &\qquad\qquad\{Z_1,\dots,Z_m\}\bigr)\\ &\quad=\sum_{j=1}^{m}\sum_{i=1}^{n_j}\log\bigl(q_j(\{\theta_i\}_j)/Z_j/p(\{\theta_i\}_j)\bigr).\end{aligned} \tag{6}$$

Owing to the arithmetic equivalence between Equation (6) and the objective function of logistic regression in the generalized linear modeling framework—but with the "predictor" here random and the "response" nonrandom—Geyer (1994) has dubbed this method "reverse logistic regression." Setting the partial derivative in each unknown Z_k to zero yields the series of convex equations defining the RLR marginal likelihood estimator:

$$\hat{Z}_k=\sum_{j=1}^{m}\sum_{i=1}^{n_j}q_k(\{\theta_i\}_j)/p(\{\theta_i\}_j)/n, \tag{7}$$

which, with reference to our definition of the pseudo-mixture density above, may be confirmed equivalent to biased sampling [Equation (4)] in the Λ-density case

for $w_j(\theta) = q_j(\theta)/\pi(\theta)$. [The $\pi(\theta)$ term ultimately cancels out from both the numerator and denominator of Equation (4), but serves here to establish our connection with the notion of a common unknown distribution, F or P_θ.]

As Kong et al. (2003) explore in detail, the fact that Geyer's RLR derivation via the quasi-log-likelihood function of Equation (6) leads to the same set of recursive update equations as Vardi's biased sampling hides a certain weakness of this "retrospective formulation": that the Hessian of the *quasi*-log-likelihood does not provide the correct asymptotic covariance matrix for the output $\hat{Z}_k$. (Though the difference in practice is almost negligible; cf. Section 4.) The same applies to a "naïve," alternative derivation of the RLR estimator—relevant to the thermodynamic integration via importance sampling methodology we describe in Section 3—given by Evans et al. (2003) in their discussion of Kong et al.'s "read" paper. That is, treat the pooled $\{\theta_i\}_j$ as if drawn from the pseudo-mixture density, $p(\theta)$, with Z_k $(k = 2, \ldots, m)$ unknown, and apply the ordinary importance sampling estimator—based on the identity, $Z_k = \int_\Omega \frac{q_k(\theta)}{p(\theta)} p(\theta)\, d\theta$—to recover the recursive update scheme of Equation (4) (but again without a corresponding argument to arrive at the correct variance).

An interesting observation often made in connection with RLR is that Equation (7) can in fact be applied without knowledge of which $f_j(\theta)$ each θ_i was drawn from, such that we may rewrite the recursive update scheme,

$$\hat{Z}_k = \sum_{i=1}^{n} q_k(\theta_i)/p(\theta_i)/n, \tag{8}$$

where we have taken the step of "losing the labels," j, on our $\{\theta_i\}_j$. This is made possible, as Kong et al. (2003) explain, because "under the model as specified ... the association of draws with distribution labels is uninformative. The reason for this is that all the information in the labels for estimating the ratios is contained in the design constants, $\{n_1, \ldots, n_m\}$."

2.2 The Density of States

Yet another construction of the convex series of $\hat{Z}_k$ updates characterizing the recursive appoach [cf. Equation (4)] has recently been demonstrated in the context of free energy estimation for molecular interactions by Habeck (2012) and Tan et al. (2012). In this framework rather than aiming directly for estimation of the marginal likelihood one aims instead to reconstruct a closely-related distribution, namely, "the density of states" (DoS), $g(e)$, defined in the physics literature in terms of a composition of the Dirac delta "function," $\delta(\cdot)$, as

$$g(e) = \int_\Omega \pi(\theta)\delta\big(e + \log L(y|\theta)\big)\, d\theta.$$

Important to note from a mathematical perspective, however, is that the composition of the Dirac delta "function"—which is itself not strictly a function, being definable only as a measure or a generalized function—lacks an intrinsic definition. Hörmander (1983) proposes a version in $\mathbb{R}^n$ valid only when the composing function, here $v(\theta) = e + \log L(y|\theta)$, is continuously differentiable and $dv(\theta)/d\theta$ nowhere zero, clearly problematic whenever the likelihood function holds constant over a set of nonzero measures with respect to P_θ! We therefore begin by suggesting a robust, alternative definition of the DoS as a transformation of the likelihood through the prior, an exercise that also serves to elucidate its connections with Skilling's nested sampling.

As briefly noted earlier with respect to characterization of the HME as Vardi's "length biased sampling," the likelihood function can serve as the basis for construction of a number of measure theoretic transformations of the prior. Most notably, the mapping $L(y|\theta) : \Omega \mapsto \mathbb{R}^+$ gives the prior in likelihood space $(0 \le L \le \infty)$,

$$P_L : \mathrm{P}_L\{B\} = \int_{L^{-1}B} \{dP_\theta(\theta)\}$$

for $B \in \mathcal{B}(\mathbb{R}^+)$ (the Borel sets on the extended reals) following Halmos [(1950), page 163], with the notation $L^{-1}B$ denoting the (assumed P_θ-measurable) set of all θ transformed through $L(y|\theta)$ into B. If the domain of θ is a metric space, then continuity (or at least discontinuity on no more than a countable set) of $L(y|\theta)$ is sufficient to ensure the P_θ-measurability of B (i.e., the validity of the above), while the continuity of the logarithm in $e(\theta) = -\log L(y|\theta)$ ensures the same for the corresponding transformation of the prior to "energy" space $(-\infty \le e \le \infty)$,

$$P_e : \mathrm{P}_e\{C\} = \int_{e^{-1}C} \{dP_\theta(\theta)\},$$

with $C \in \mathcal{B}(\mathbb{R}^+)$. In each case the appropriate version of the marginal likelihood shares equality with the original [Equation (2)] wherever Z is itself finite, owing to the P_L- and P_e-measurability of L and $\exp(-e)$,

respectively:

$$Z = \int_0^\infty L\{dP_L(L)\} \quad \text{and}$$
$$Z = \int_{-\infty}^\infty \exp(-e)\{dP_e(e)\}$$

(cf. Halmos, 1950, page 164).

Although unnecessary for a straightforward application of biased sampling, one might choose to further require that P_e admit a Λ-density, equivalent to the requirement that its distribution function, $G_e(e') = \int_{-\infty}^{e'}\{dP_e(e)\}$, be everywhere differentiable. For a continuous likelihood function we can be assured of this provided that $L(y|\theta)$ at no place holds constant over a set of nonzero measures with respect to P_θ—the same limitation on its δ "function" definition. If so, we may write the marginal likelihood integral as Habeck (2012),

$$Z = \int_{-\infty}^\infty \exp(-e) g_e(e)\, de. \tag{9}$$

Estimation of $g_e(e)$ (or in fact the general measure, P_e) can of course be accomplished via biased sampling given i.i.d.'s draws from a series of $w(e)$-weighted versions of g_e, and, indeed, this is the justification of the DoS algorithm—seen as the limiting case of the weighted histogram analysis method (Ferrenberg and Swendsen, 1989) with bin size approaching zero—given by Tan et al. (2012). The derivation of the recursive update formula [Equation (4)] presented by Habeck (2012) for the DoS is alternatively via a novel functional analysis procedure for optimization of the log-likelihood of an empirical energy histogram; however, as with Geyer's RLR derivation, this approach does not lead to an uncertainty estimate or CLT for the output $\hat{Z}_k$.

2.2.1 *Relation to nested sampling.* The nested sampling identity (Skilling, 2006),

$$Z = \int_0^1 L(X)\, dX, \tag{10}$$

where $L(X)$ represents the inverse of the survival function of likelihood with respect to the prior—that is, $X(L') = 1 - \mathrm{P}_L\{L \le L'\}$—and dX denotes Riemann integration over the "prior mass cumulant," may best be understood by reference to a well-known relation between the expectation of a non-negative random variable and its distribution function, namely, that for $y \sim P_Y$ with $y \ge 0$,

$$\mathrm{E}_Y\{Y\} = \int_0^\infty Y\{dP_Y(Y)\} = \int_0^\infty \big(1 - \mathrm{P}_Y\{Y \le y\}\big)\, dy$$

(cf. Billingsley, 1968, page 223). Importantly, this relation (which follows from integration by parts) holds irrespective of whether or not P_Y admits a Λ-density, and in the marginal likelihood context becomes $Z = \int_0^\infty 1 - \mathrm{P}_L\{L \le L'\}\, dL'$. If $\mathrm{P}_L\{L = \infty\} = 0$, then this monotonically decreasing, *cadlag* function on $\mathbb{R}_+$ with bounded range (between zero and one) is (perhaps improper) Riemann integrable, and we may simply "switch axes" to obtain Equation (10). While the uniqueness of the inverse survival function, $L(X)$, can be ensured by requiring $L(y|\theta)$ to be continuous with connected support (Chopin and Robert, 2010), the weaker condition of $L(y|\theta)$ discontinuous on a set of measure zero with respect to P_L suffices to ensure an $L(X)$ defined uniquely on all but a corresponding set of Lebesgue measure zero, negligible also for our Riemann integration.

Now for differentiable $G_e(e') = \mathrm{P}_\theta\{e(\theta) < e'\}$, such that $g(e)$ might be defined without our earlier measure theoretic considerations as $g(e) = dG_e(e)/de$, the DoS version of the marginal likelihood [Equation (9)] can nevertheless be recovered using the nested sampling identity. Observing that $G_e(e) = X(\exp[-e]) = X(L)$, we have

$$\begin{aligned} g(e) &= dX\big(\exp[-e]\big)/de = dX(L)/dL \times dL/de \\ &= dX(L)/dL \times -\exp[-e]. \end{aligned}$$

Substitution of $X(L)$ into Equation (10) yields

$$\begin{aligned} Z &= \int_{X(\infty)}^{X(0)} L(X)\, dX \\ &= \int_\infty^0 L\big(X(L')\big) \times dX\big(L'\big)/dL' \times dL', \end{aligned}$$

and then by substitution of e we recover

$$\begin{aligned} Z &= \int_{e(-\infty)}^{e(\infty)} L' \times dX\big(L'\big)/dL' \times dL' \\ &= \int_{-\infty}^{\infty} \exp[-e] \times -g(e)\exp[e] \times -\exp[-e] \times de. \end{aligned}$$

That is, consistent with the requirements of Habeck (2012) and Tan et al. (2012), this alternative DoS formulation returns the identity

$$Z = \int_{-\infty}^\infty g(e)\exp(-e)\, de.$$

Interestingly, the above relationship between the DoS and nested sampling identities is mirrored by the existence of a measure theoretic construction for the latter (cf. Appendix C of Feroz et al., 2013). If we take

the survival function, $X(L) = 1 - \int_0^L \{dP_L\}$, as defining yet another transformation of the prior through the likelihood—a transformation ensured P_L-measurable, and hence P_θ-measurable, by the right continuity of $X(L)$—we recover the following distribution in prior cumulant space ($0 \leq X \leq 1$):

$$P_X : \mathrm{P}_X\{D\} = \int_{X^{-1}D} \{dP_\theta(\theta)\}.$$

Similarly, the marginal likelihood formula equivalent to the nested sampling identity becomes

$$Z = \int_0^1 L(X)\{dX\}$$

for $X(L)$ invertible, that is, $L(y|\theta)$ continuous with connected support (Chopin and Robert, 2010). More generally, though, we can view $L(X)$ as the conditional probability function of likelihood given prior mass cumulant defined modulo P_θ by the relation

$$\int_{X^{-1}D} L(y|\theta)\{dP_\theta(\theta)\} = \int_D e_\theta(L|X)\{dP_X(X)\} \tag{11}$$

(cf. Halmos and Savage, 1949). For statistical problems on a complete separable metric space there will always exist a unique local version of $e_\theta(L|X)$ defined as a weak limit such that $e_\theta(L|X = x)$ is meaningful even for atomic x (Pfanzagl, 1979).

The value of this insight becomes apparent when we examine the nested sampling estimator for posterior functionals (cf. Chopin and Robert, 2010),

$$\mathrm{E}_{\pi(\theta|y)}\{f(\theta)\} \approx \sum_{i=1}^n \tilde{w}_i L(\theta_i|y) f(\theta_i),$$

where $\tilde{w}_i$ here represents the nested sampling posterior weight for θ_i, $d\hat{P}_X(X(\theta_i))$—typically $\tilde{w}_i = (\hat{X}_{i-1} - \hat{X}_i)$ (Skilling, 2006). This estimator relies on the relation given by Equation (11) with $L(y|\theta)$ replaced by $L(y|\theta)f(\theta)$, which holds for $f(\theta)$ measurable—a more general condition than that of $e_\theta(f|L)$ absolutely continuous given by Chopin and Robert (2010). Importantly, this ensures the validity of prior-sensitivity analysis via computation of the posterior functional of $\pi_{\mathrm{alt}}(\theta)/\pi(\theta)$ in nested sampling—a powerful technique not previously exploited in nested sampling analyses—as we shall discuss for the case of biased sampling below.

2.3 Importance Sample Reweighting for Prior-Sensitivity Analysis

In the Bayesian framework (Jeffreys, 1961; Jaynes, 2003) the ratio of marginal likelihoods under rival hypotheses (i.e., the Bayes factor) operates directly on the prior odds ratio for model selection to produce the posterior odds ratio as

$$\begin{aligned} &\mathrm{P}\{M_1|y\}/\mathrm{P}\{M_2|y\} \\ &\quad = \big[\mathrm{P}\{y|M_1\}/\mathrm{P}\{y|M_2\}\big]\big[\mathrm{P}\{M_1\}/\mathrm{P}\{M_2\}\big] \\ &\quad = [Z_{M_1}/Z_{M_2}]\big[\mathrm{P}\{M_1\}/\mathrm{P}\{M_2\}\big]. \end{aligned} \tag{12}$$

A much maligned feature of the marginal likelihood in this context is its possible sensitivity to the choice of the parameter priors, $\mathrm{P}\{\theta|M_1\}$ and $\mathrm{P}\{\theta|M_2\}$, through Z_{M_1} and Z_{M_2}. When limited information is available to inform (or justify) this choice, the resulting Bayes factor can appear almost arbitrary. [On the other hand, viewed as a quantitative implementation of Ockham's Razor, the key role of prior precision may well serve as strong justification for the use of Bayesian model selection in the scientific context; cf. Jeffreys and Berger (1991).] In their influential treatise on this topic Kass and Raftery (1995) thus argue that some form of prior-sensitivity analysis be conducted as a routine part of all Bayesian model choice experiments, their default recommendation being the recomputation of the Bayes factor under a doubling and halving of key hyperparameters.

If the original marginal likelihoods have been estimated under an amenable simulation scheme, then, as Chopin and Robert (2010) point out for the case of nested importance sampling, alternative Bayes factors under (moderate) prior rescalings may be easily recovered by appropriately reweighting the existing draws without the need to incur further (computationally expensive) likelihood function calls; and, indeed, the RLR method was conceived specifically to facilitate such computations (though in the reweighting mixtures context; Geyer and Thompson, 1992; Geyer, 1994). Using the $\hat{Z}_k$ from biased sampling under our nominal prior for a given model, the pseudo-mixture density, $p(\theta)$, of Equation (5) now serves as an efficient "proposal" for pseudo-importance sampling of various other targets with mass concentrated near that of the posterior. In particular, for the alternative marginal likelihood, $\hat{Z}_{\mathrm{alt}}$, under some alternative prior density, $\pi_{\mathrm{alt}}(\theta)$, we have

$$\hat{Z}_{\mathrm{alt}} = \sum_{i=1}^n L(y|\theta_i)\pi_{\mathrm{alt}}(\theta_i)/p(\theta_i)/n. \tag{13}$$

The stability of this importance sample reweighting procedure may be monitored via the effective sample size, $\mathrm{ESS} = n/[1 + \mathrm{var}_p\{\pi_{\mathrm{alt}}(\theta)/p(\theta)\}]$, following Kong, Liu and Wong (1994), and its asymptotic variance estimated via recomputation of Equation (13)

under perturbations to the original $\hat{Z}_k$ drawn from the biased sampling covariance matrix with bootstrap resampling of the pooled θ_i.

For the general case of biased sampling from $w_j(\theta)$-weighted versions of a prior distribution, P_θ, not necessarily admitting a Λ-density, the equivalent formula takes the Radon–Nikodym derivative of the alternative prior with respect to the original, $\frac{dP_{\theta,\text{alt}}}{dP_\theta}(\theta)$ (for $P_{\theta,\text{alt}} \ll P_\theta$), such that

$$\hat{Z}_{\text{alt}} = \sum_{i=1}^{n} L(y|\theta_i) \frac{dP_{\theta,\text{alt}}}{dP_\theta}(\theta) \Big/ \left[\sum_{j=1}^{m} n_j/n \times w_j(\theta_i) \right] \Big/ n. \tag{14}$$

We demonstrate the utility of this approach to prior-sensitivity analysis in our finite and infinite mixture modeling of the well-known galaxy data set in Section 5—and we refer the interested reader to our other recent astronomical application concerning a semiparameteric mixed effects model presented in Cameron and Pettitt (2013). Though both these examples are based on the Dirichlet process prior, one can envisage application of the same technique to investigate prior sensitivity in many other problems of applied statistics—for example, Gaussian or Ornstein–Uhlenbeck process modeling of astronomical time series (Brewer and Stello, 2009; Bailer-Jones, 2012).

2.4 Designing and Sampling the Bridging Sequence

Although the recursive update scheme of biased sampling provides a powerful technique for estimating the marginal likelihood given i.i.d. draws from a prespecified sequence of $w_j(\theta)$-weighted distributions, the design of this bridging sequence and the choice of an algorithm to sample from it are left to the user. While it is possible from theoretical principles to identify the optimal choice of $w_j(\theta)$ with respect to the asymptotic variance under perfect sampling for a limited range of problems—for example, Gill, Vardi and Wellner (1988) show the optimality of $w_1(\theta) = |L - Z|$ (requiring Z known!) for the one sample case with $F = P_L$ (in our marginal likelihood notation)—the design problem cannot easily be solved in general. Moreover, even where a theoretically optimal sequence can be identified, it will not necessarily be computationally feasible to sample from such a sequence. Of more practical value therefore are heuristic guides for the pragmatic choice of $w_j(\theta)$: strategies that will in a wide variety of applied problems produce adequate bridging sequences to ensure manageable uncertainty in the output $\hat{Z}$ while remaining accessible to existing posterior sampling techniques. This topic in various guises is the focus for the remainder of this paper, including our numerical examples.

Perhaps the most natural family of bridging sequence for use on the recursive pathway is that of the power posteriors method [Equation (3); Lartillot and Phillipe, 2006; Friel and Pettitt, 2008]: this being both the favored approach for past DoS-based applications (Habeck, 2012; Tan et al., 2012)—where the parameter, t, has a physical interpretation as the inverse system temperature—and in Geyer's formulation of RLR—where this particular sampling strategy ties in neatly with his parallel tempering MCMC algorithm (MC^3; Geyer, 1992). And, indeed, in Section 3 below we will describe yet another conceptual connection between these two methods, providing a heuristic justification for the borrowing of thermodynamic integration strategies to this end. Importantly, simulation from the power posterior at an arbitrary t_j is typically no more difficult than simulation from the full posterior ($t_m = 1$), the required modifications to a standard MCMC and/or Gibbs sampling code being often quite trivial (e.g., Cameron and Pettitt, 2013). With biased sampling devised for i.i.d. draws, though, it is important to thin the resulting chains (Tan et al., 2012) so as not to bias the corresponding asymptotic covariance estimates. Experience has shown that prior-focused temperature schedules, such as $t = \{0, 1/(m-1), 2/(m-1), \ldots, 1\}^c$ with $c \sim 3$–5, tend to work well for thermodynamic integration (Friel and Pettitt, 2008), and we confirm this also for biased sampling of our banana-shaped likelihood case study in Section 4. [Likewise for tempering from a normalized auxiliary density, $h(\theta)$, closer in Kullback–Leibler divergence to the posterior than the prior; Lefebvre, Steele and Vandal (2010) and see our Section 4.1.]

Another effective choice of bridging sequence for biased sampling, which we demonstrate in our galaxy data set case study of Section 5, is that of partial data posteriors (cf. Chopin, 2002): that is, $L(y^{(r_j)}|\theta)\pi(\theta)$ where $y^{(r_j)}$ represents a subset of r_j elements of the full data set with $r_1 = 0$ the prior and $r_m = n_{\text{tot}}$ the full posterior. For i.i.d. y, with an expected contribution of r_j times the unit Fisher information, the "volume" of highest posterior mass should shrink as roughly $\sqrt{r_j}$, suggesting an automatic choice of roughly $r_j = \lfloor n_{\text{tot}} \times \{0, 1/(m-1), 2/(m-1), \ldots, 1\}^c \rfloor$ with $c = 2$ for this method. (In practice though, the first nonzero

r_j may well be limited by sampling/identifiability constraints on the model; for our mixture model, for instance, we must specify $r_2 = k$, the number of mixture components.)

Finally, as observed by Habeck (2012), the constrained-likelihood bridging sequence of nested sampling can also be represented within the DoS framework via $w_j(e) = I(e < e_j)$ with $e_j < e_{j-1}$, although in practice (as we explore in Section 4) the non-i.i.d. nature of the resulting draws (with each draw from e_{j-1} influencing the placement of the next e_j and its successors) violates the assumptions of the biased sampling paradigm and ultimately limits the utility of this approach by biasing its asymptotic covariance estimate. In fact, this issue more generally remains an open problem for recursive marginal likelihood estimation theory: how can we best design effective strategies for *adaptively* choosing our bridging sequence, and how can such modifications to the biased sampling paradigm be accounted for theoretically? Given the effectiveness of empirical process theory for characterizing the asymptotics of Vardi's biased sampling, it seems likely that a solution to the above will require extensive work in this area (with a focus on the impact of long-range dependencies). A similar problem arises in describing the asymptotics of adaptive multiple importance sampling (Cornuet et al., 2012), which without its adaptive behavior could be considered a version of biased sampling with known W_j; Marin, Pudlo and Sedki (2012) were recently able to provide a consistency proof for a modified version of this algorithm, but with a CLT remaining elusive.

3. THERMODYNAMIC INTEGRATION VIA IMPORTANCE SAMPLING

Inspired by the recursive pathway of biased sampling, RLR and the DoS, we present here yet another such strategy for marginal likelihood estimation, which we name "thermodynamic integration via importance sampling" (TIVIS). Although quite novel at face value, it is easily shown to be a direct transformation of the recursive update methodology; yet by effectively recasting this as a thermodynamic integration procedure we attain insight into the relationship between its error budget and bridging sequence. Specifically, the error in the estimation of each Z_k may be thought of as dependent on both the J-divergence (Lefebvre, Steele and Vandal, 2010) between it and the remainder of the ensemble (via the thermodynamic identity) and on the accuracy of our estimates for those other Z_j ($j \neq k$).

To construct the TIVIS estimator, we once again assume the availability of pooled draws, $\{\{\theta_i\}_j : i = 1, \ldots, n_j;\ j = 1, \ldots, m\}$, from a sequence of bridging densities, $f_j(\theta) = q_j(\theta)/Z_j$ ($j = 1, \ldots, m$), with each $g_k(\theta)$ exactly known. Moreover, we suppose that $j = 1$ indexes a normalized reference/auxiliary, $\pi(\theta)$ or $h(\theta)$, such that $Z_1 = 1$ is known, but with the remaining Z_k typically unknown. Despite our subsequent use of the thermodynamic identity, however, we do not necessarily require here that the bridging densities follow the geometric path between these two extremes. Now, rather than seek each $\hat{Z}_k$ via direct importance sampling from $p(\theta)$ as per the RLR, the TIVIS method is to instead seek each normalization constant via thermodynamic integration from its preceding density in the ensemble, $q_{k-1}(\theta)$, using the identity,

$$\log Z_k = \int_0^1 E_{\pi_t^k}\{\log(q_k(\theta)/f_{k-1}(\theta))\}\,dt, \tag{15}$$

where $\pi_t^k(\theta) \propto [f_k(\theta)]^t [f_{k-1}(\theta)]^{1-t} \propto [g_k(\theta)]^t \times [g_{k-1}(\theta)]^{1-t}$. For existence of the log-ratio in Equation (15) we must impose the strict condition (*not* necessary for ordinary RLR) that all $f_k(\theta)$ share matching supports. Pseudo-importance sampling from $p(\theta)$—that is, importance sample reweighting of the drawn $\{\theta_i\}_j$—allows construction of the appropriate (but unnormalized) weighting function,

$$u(\theta, t) = \big[g_k(\theta)\big]^t \big[g_{k-1}(\theta)\big]^{1-t} / p(\theta),$$

which in substitution to Equation (15) yields the TIVIS estimator,

$$\begin{aligned}\log(\hat{Z}_k/\hat{Z}_{k-1}) &= \int_0^1 \left[\sum_{i=1}^n \log\big(g_k(\theta_i)/g_{k-1}(\theta_i)\big) u(\theta_i, t)\right] \\ &\qquad \Big/ \left[\sum_{i=1}^n u(\theta_i, t)\right] dt.\end{aligned} \tag{16}$$

In computational terms, numerical solution of the one-dimensional integral in the above may be achieved to arbitrary accuracy by simply evaluating the integrand at sufficiently many t_j on the unit interval, followed by summation with Simpson's rule. If the sequence of bridging densities is well chosen (and suitably ordered), the J-divergence between each $f_k(\theta)$ and $f_{k-1}(\theta)$ pairing should be far less than that between prior and posterior, such that a naïve regular spacing of the t_j will suffice.

To show the equivalence between this estimator and that of the recursive update scheme defined by Equation (4), we simply observe that the derivative of the

denominator in Equation (16) equals the numerator and, thus, by analogy to $\int_0^1 s'(x)/s(x)\,dx = \log s(1) - \log s(0)$, we have

$$\log(\hat{Z}_k/\hat{Z}_{k-1}) = \log\left[\sum_{i=1}^{n} g_k(\theta_i)/p(\theta_i)\right] - \log\left[\sum_{i=1}^{n} g_{k-1}(\theta_i)/p(\theta_i)\right],$$

and, thus,

$$\log \hat{Z}_k = \log\left[\sum_{i=1}^{n} g_k(\theta_i)/p(\theta_i)/n\right].$$

In the following two case studies we further explore by numerical example various issues concerning the design of the bridging sequence [with particular reference to the efficiency in $L(y|\theta)$ calls; Section 4], and we highlight the utility of the *normalized* bridging sequence for prior-sensitivity analysis (Section 5).

4. CASE STUDY: BANANA-SHAPED LIKELIHOOD FUNCTION

For our first case study we consider a ("mock," that is, data independent) banana-shaped likelihood function, defined in two dimensions ($\theta = \{\theta_1, \theta_2\}$) as

$$L(\theta) = \exp\left(-\left(10 \times (0.45 - \theta_1)\right)^2/4 - \left(20 \times \left(\theta_2/2 - \theta_1^4\right)\right)^2\right), \tag{17}$$

with a Uniform prior density of $\pi(\theta) = 1/4$ on the rectangular domain, $[-0.5, 1.5] \times [-0.5, 1.5]$. A simple illustration of this likelihood function as a logarithmically-spaced contour plot is presented in the left-hand panel of Figure 1. Brute-force numerical integration via quadrature returns the "exact" solution, $Z = 0.01569[6]$ (or $\log Z = -4.154[3]$).

As a benchmark of the method we first apply the biased sampling estimator to draws from a sequence of bridging densities following the standard power posteriors path. Though even a cursory inspection of the likelihood function for this simple case study is sufficient to confirm its unimodality and to motivate a family of suitable proposal densities for straightforward importance sampling of $\pi_t(\theta) \propto \pi(\theta)L(\theta)^t$, for demonstrative purposes we have chosen to implement an MC^3 (Geyer, 1992) approach here instead, the latter being ultimately amenable to more complex posteriors than the former. Following standard practice for thermodynamic integration—as per our motivation from Sections 2.4 and 3 above—we adopt a prespecified tempering schedule spaced geometrically as $t = \{0, 1/(m-1), 2/(m-1), \ldots, 1\}^c$ with $c = 5$ and $m = 5$. To illustrate the $1/\sqrt{n}$ convergence of biased sampling, we run this procedure 100 times at each of five total sample sizes ($n_{\text{tot}} = \{125, 500, 1250, 5000, 12{,}500\}$; distributed equally across all five temperatures) thinned at a rate of 0.25 from their parent MC^3 chains. The resulting mean and standard error (SE) at each n_{tot} are marked in the right-hand panel of Figure 1.

Overlaid are (the means of) the corresponding "per simulation" estimates of this standard error computed from the rival asymptotic covariance matrix forms of Gill, Vardi and Wellner (1988)/Kong et al. (2003) and Geyer (1994): the former being originally derived from

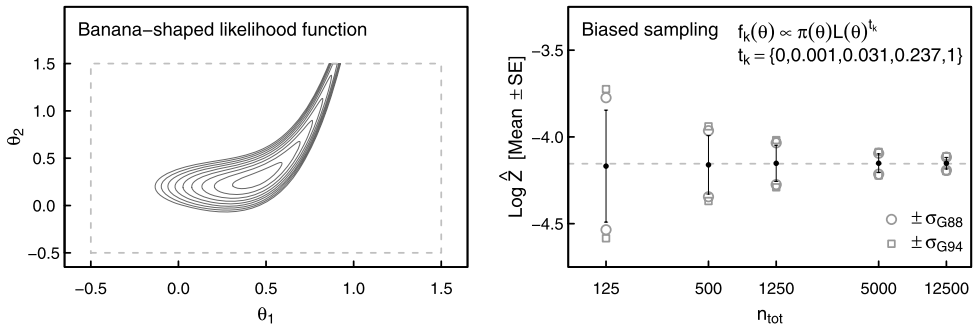

FIG. 1. *The banana-shaped likelihood function of our first case study* [*Equation* (17) *of Section* 4] *illustrated graphically as a logarithmically-spaced contour plot on the domain of our Uniform prior,* $[-0.5, 1.5] \times [-0.5, 1.5]$ (*left-hand panel*). *Convergence of the biased sampling estimator for the corresponding marginal likelihood under MC*^{3} *sampling of the power posterior* (*at five prespecified temperatures*) *as a function of the total sample size is shown in the right-hand panel. The marked points and error bars on this figure indicate respectively the recovered mean and standard error* (*SE*) *in* $\log \hat{Z}$ *for* 100 *trials at each* n_{tot}. *The dashed, light grey line indicates the* "*exact*" $\log Z$ *for this example derived via brute-force quadrature, and the open, light grey symbols indicate the mean* "*per simulation*" *estimate of the SE from the asymptotic covariance matrix formulae of Gill, Vardi and Wellner* (1988) *and Geyer* (1994) *alternately.*

the empirical process CLT applicable to biased sampling and the latter from maximum likelihood theory using the Hessian of the quasi-likelihood function for reverse logistic regression. As noted in Section 2, Kong et al. (2003) have previously discussed the inadequacy of Geyer's covariance estimator—though for the present design the difference is negligible. It is worth noting that both estimates are a little conservative at low n_{tot} but give an excellent agreement with the repeat simulation SE by $n_{\text{tot}} = 1250$.

With this power posteriors version of biased sampling as benchmark, we now consider the merits of two alternative schemes for defining, and sampling from, the required sequence of bridging densities, $f_k(\theta)$, in Sections 4.1 and 4.2 below.

4.1 Thermodynamic Integration from a Reference/Auxiliary Density

As highlighted by Lefebvre, Steele and Vandal (2010), the error budget of thermodynamic integration over the geometric path depends to first-order upon the J-divergence between the reference/auxiliary density, $h(\theta)$, and the target, $\pi(\theta|y)$. Thus, it will generally be more efficient to set a "data-driven" $h(\theta)$—such as may be recovered from the position and local curvature of the posterior mode—than to integrate "naïvely" from the prior, that is, $h(\theta) = \pi(\theta)$. Here we demonstrate the corresponding improvement to the performance of the biased sampling estimator resulting from the choices, $h(\theta) \sim \mathcal{N}_{\text{Trunc.}}(\mu_{\text{mode}}, \Sigma^{-1}_{\text{mode}})$ and $h(\theta) \sim \mathcal{T}_{\text{Trunc.}}(\mu_{\text{mode}}, \Sigma^{-1}_{\text{mode}})$. Here $\mathcal{N}_{\text{Trunc.}}$ and $\mathcal{T}_{\text{Trunc.}}$ denote the two-dimensional Normal and Student's t ($\nu = 1$) distributions (truncated to our prior support), respectively, while μ_{mode} denotes the posterior mode and Σ_{mode} its local curvature (recovered here analytically, but estimable at minimal cost in many Bayesian analysis problems via standard numerical methods). As before, we apply MC^3 to explore the tempered posterior and repeat both experiments 100 times at each of our five n_{tot}. In contrast to the power posteriors case, we adopt here a regular temperature grid, $t = \{0, 0.25, 0.5, 0.75, 1\}$, to allow for the imposed/intended similarity between $\pi(\theta|y)$ and $h(\theta)$. Our results are presented in Figure 2 and discussed below.

As expected from both theoretical considerations (Gelman and Meng, 1998; Lefebvre, Steele and Vandal, 2010) and reports of practical experience with other marginal likelihood estimators (Fan et al., 2012), use of a "data-driven" auxiliary in this example has indeed reduced markedly the standard error of the biased sampling scheme (at fixed n_{tot}) with respect to that of the naïve (power posteriors) path, that is, $h(\theta) = \pi(\theta)$. In this instance the (thinner-tailed) Normal auxiliary has outperformed the (fatter-tailed) Student's t (with one d.o.f.); however, although this result is again consistent with theoretical expectations—as a quick computation using the "exact" $\log Z$ confirms $J[\mathcal{N}(\mu_{\text{mode}}, \Sigma^{-1}_{\text{mode}}), h(\theta)] \ll J[\mathcal{T}(\mu_{\text{mode}}, \Sigma^{-1}_{\text{mode}}), h(\theta)]$—it should be remembered that the optimal choice of auxiliary from within a standard parametric family depends on the likelihood function itself, and so will vary from problem to problem. Moreover, without knowledge of the desired Z it is not possible to optimize $h(\theta)$ a priori; and even a crude estimator of the J-divergence run with, for example, the Laplace approximation to the marginal likelihood will neverthe-

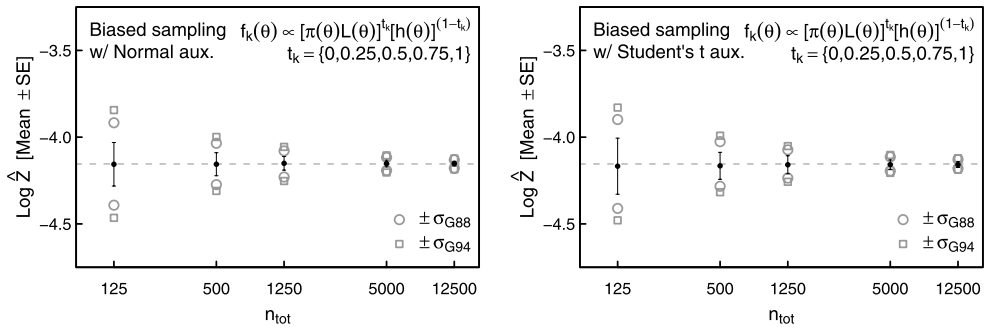

FIG. 2. *Convergence of the biased sampling estimator for the marginal likelihood of our banana-shaped likelihood function under MC^3 sampling (at five prespecified temperatures) on the geometric path between a "data-driven" reference/auxiliary density, $h(\theta)$, and the posterior, shown as a function of the total sample size. The adopted $h(\theta)$ takes a two-dimensional Student's t form in the left-hand panel and a Normal form in the right-hand, with its controlling parameters (μ_{mode} and Σ_{mode}) in each case set to the location and curvature of the posterior mode. A marked reduction in standard error (at fixed n_{tot}) with respect to that of the naïve (power posteriors) path, that is, $h(\theta) = \pi(\theta)$, is evident from comparison with Figure 1.*

less add numerous extra likelihood evaluations to the computational budget. Although "fatter-tailed" than a typical likelihood function, the Student's t may well prove a superior choice for some multimodel posterior problems in practice by better facilitating mixing during the MC^3 sampling stage.

4.2 Ellipse/Ellipsoid-Based Nested Sampling

Recalling the connections between the DoS derivation of the recursive pathway and the nested sampling algorithm described in Section 2.2, it is of some interest to compare directly the performance of these rival techniques. The present case study with its Uniform prior density is in fact well suited to this purpose since in the field of cosmological model selection, where nested sampling has been most extensively used of late (Mukherjee, Parkinson and Liddle, 2006; Feroz and Hobson, 2008), it is standard practice to adopt separable priors from which a Uniform sample space may be easily constructed under the quantile function transformation, which, for the discussion below, we assume has been done such that $\pi(\theta)$ may be taken as strictly Uniform on $[0,1]^N$ (in the transformed coordinate space). Given these conditions, Mukherjee, Parkinson and Liddle (2006) outline a crude-but-effective scheme for exploring the constrained-likelihood shells of nested sampling, in which the new "live" particle for each update must be drawn with density proportional to $\pi(\theta)I(L(\theta) > L(\theta_{i-1}))$.

Under the Mukherjee, Parkinson and Liddle (2006) scheme, to draw the required θ_i, one simply identifies the minimum bounding ellipse [or with $D(\theta) > 2$, the minimum bounding *ellipsoid*] for the present set of "live" particles, expands this ellipse by a small factor $\sim$1.5–2 with the aim of enclosing the full support of $I(L(\theta) > L(\theta_{i-1}))$, and then draws randomly from its interior until a valid $\{\theta_i, L(\theta_i)\}$ is discovered. Supposing the elliptical sampling window thus defined has been enlarged sufficiently to fully enclose the desired likelihood surface [which it must do to ensure unbiased sampling of $\{\theta_i, L(\theta_i)\}$, although we can rarely be *sure* that it has], it remains unlikely to match its shape exactly, leading to an overhead of n_{oh} discarded draws, $\{\theta_i^{(j)} : L(\theta_i^{(j)}) < L(\theta_{i-1}), j = 1, \ldots, n_{\text{oh}}\}$. At each θ_i the incurred n_{oh} may be thought of as a single realization of the negative binomial distribution with p equal to the fraction of the bounded ellipse for which $L(\theta) < L(\theta_{i-1})$, hence, $E(n_{\text{oh}}) = 1/p - 1$. The magnitude of this overhead can in general be expected to scale with the geometric volume of the parameter space, potentially limiting the utility of this otherwise dimensionally-insensitive Monte Carlo-based estimator. However, where applicable, the Mukherjee, Parkinson and Liddle (2006) scheme may nevertheless prove more efficient than the alternative of constrained-MCMC-sampling to find the new θ_i (cf. Friel and Wyse, 2012) in which one must discard at least $\sim$10–20 burn-in moves [each with a necessary $L(\theta)$ call] per step to achieve approximate stationarity.

Applying the ellipse-based approach to nested sampling of the banana-shaped likelihood function of Equation (17) with $N_{\text{live}} = \{12, 25, 50, 125\}$ live particles evolved over $10 \times N_{\text{live}}$ steps in each case [and a small extrapolation of the mean L_{live} times $\exp(-10)$ at the final step; cf. Skilling, 2006], we recover a convergence to the true $\log Z$ as shown in the left-hand panel of Figure 3. Important to note is that with the ellipse scale factor of 1.5 used here the result is an overhead of $n_{\text{oh}} \approx 2.3$ likelihood calls per accepted θ_i,

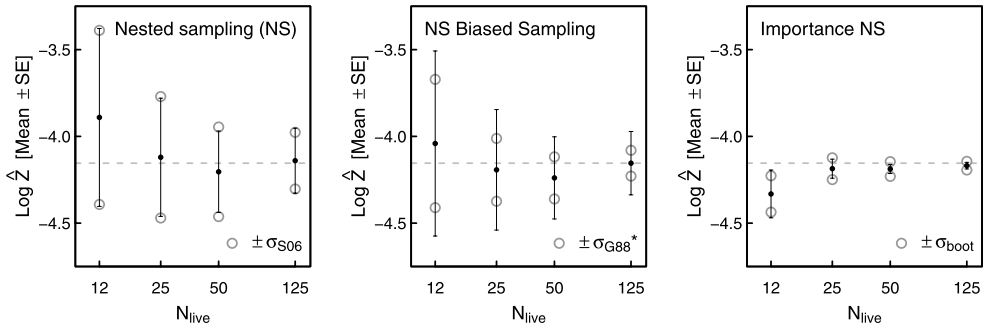

FIG. 3. *The performance of nested sampling* (*left-hand panel*) *as a marginal likelihood estimator for our banana-shaped pseudo-likelihood function, run under the ellipse-based strategy for exploring the sequence of constrained-likelihood densities proposed by Mukherjee, Parkinson and Liddle* (2006); *compared with that of biased sampling* (*middle panel*) *and* "*importance nested sampling*" (*right-hand panel*) *with the same bridging sequence. The first two schemes converge to the true* $\log Z$ *at a similar rate in* N_{live}, *while the third is faster since it harnesses the information content of draws otherwise discarded from nested sampling in its constrained-likelihood search.*

such that nested sampling at $N_{\text{live}} = 125$ corresponds to $n_{\text{tot}} \approx 2875$ in the previous examples. An overhead of this magnitude should be a concern for "real world" applications of nested sampling in which the likelihood function may be genuinely expensive to evaluate; indeed, for modern cosmological simulations MCMC exploration of the $D(\theta) \lesssim 12$ posterior is effectively a super-computer-only exercise due solely to the cost of solving for $L(y|\theta)$. [At this point the skeptical reader might object that the distinctly nonelliptical $L(\theta)$ considered in this example be considered a particularly unfair case for testing the Mukherjee, Parkinson and Liddle (2006) method, but such banana-shaped likelihoods are in fact quite common in higher-order cosmological models; see, for instance, Davis et al. (2007).] We therefore suggest that one might improve upon the efficiency of ellipse-based nested sampling by co-opting its bridging sequence into the biased sampling framework in some manner.

As Habeck (2012) has pointed out, the nested sampling pathway can be accommodated roughly within the DoS (and hence biased sampling) framework, for example, by treating the accepted θ_i (pooled with the surviving N_{live} live particles) as drawn from the series of weighted distributions, $w_j(\theta)\, dF(\theta) = I(L(\theta) > L_j)\, dF(\theta)$. However, with each $w_j(\theta)$ $(j > 1)$ now dependent on past draws—and hence the $\{\theta_i\}_j$ no longer i.i.d.—although we can apply the recursive update scheme of Equation (4) to normalize the bridging sequence and then importance sample reweight to Z, the biased sampling CLT no longer holds. To demonstrate this, we apply the above procedure to the draws from our previous nested sampling runs and plot the mean and repeat simulation SE at each N_{live} in the middle panel of Figure 3. While the efficiency of this estimator is almost identical to that of ordinary nested sampling, the "naïve" application of Gill et al.'s asymptotic covariance matrix does not yield an SE estimate matching that of repeat simulation.

A more interesting alternative is to observe that for the ellipse-based nested sampling method (given uniform priors) the normalization of each $f_k(\theta)$ is in fact easily computed from the area/volume of the corresponding ellipse/ellipsoid. That is, we can simply pool our draws—*including* the θ_i with $L(\{\theta_i\}_j) < L_j$ otherwise discarded from nested sampling—and apply the importance sample reweighting procedure of Equation (13) with $\pi_{\text{alt}}(\theta) = \pi(\theta)$ and

$$p(\theta) = \sum_{k=1}^{m} [n_k / n_{\text{tot}}] \big[I\big(\theta \in Ell[E_{\text{live}(k)}]\big) / V_k \big]$$

(with V_k the volume of the kth ellipse and $Ell[E_{\text{live}(k)}]$ its interior). Application of this strategy—which we dub "importance nested sampling" (INS)—to the present example yields $\log Z$ estimates with much smaller repeat simulation SE than either of the previous summations as shown in the right-hand panel of Figure 3. Bootstrap resampling of the drawn $\{\theta_i, L_i\}$ gives a reasonable estimator of this SE, though we note that INS does not appear to be unbiased in $\log Z$, with a slight tendency toward underestimation at small n_{tot}. Further computational experiments are now underway to better quantify the advantages offered by this approach to harnessing the information content of these otherwise discarded draws in the ellipse-based nested sampling paradigm (presented in Feroz et al., 2013).

5. CASE STUDY: NORMAL MIXTURE MODELING OF THE GALAXY DATA SET

The well-known galaxy data set, first proposed as a test case for kernel density estimation by Roeder (1990), consists of precise recession velocity measurements (in units of 1000 km s^{-1}) for 82 galaxies in the Corona Borealis region of the Northern sky reported by Postman, Huchra and Geller (1986). The purpose of the original astronomical study was to search—in light of a then recently discovered void in the neighboring Boötes field (Kirshner et al., 1981)—for further large-scale inhomogeneities in the distribution of galaxies. Given the well-defined selection function of their survey, Postman, Huchra and Geller (1986) were easily able to compute as a benchmark the recession velocity density function expected under the null hypothesis of a uniform distribution of galaxies throughout space, and by visual comparison of this density against a histogram of their observed velocities the astronomers were able to establish strong evidence against the null, thereby boosting support for the (now canonical) hierarchical clustering model of cosmological mass assembly (Gunn, 1972). However, under the latter hypothesis, as Roeder (1990) insightfully observed, one can then ask the more challenging statistical question of "*how many distinct clustering components are in fact present in the recession velocity data set?*"

Many authors have since attempted to answer this question (posed for simplicity as a univariate Normal mixture modeling problem) as a means to demonstrate the utility of their preferred marginal likelihood estimation or model space exploration strategy. Notable such contributions to this end include the following: the infinite mixture model (Dirichlet process prior) analyses

of Escobar and West (1995) and Phillips and Smith (1996); Chib's exposition of marginal likelihood estimation from Gibbs sampling output (Chib, 1995); the reversible jump MCMC approach of Richardson and Green (1997); and the label switching studies of Stephens (2000) and Jasra, Holmes and Stephens (2005). The earliest of these efforts are well summarized by Aitkin (2001), who highlights a marked dependence of the inferred number of mixture components on the chosen priors. For this reason, as much as its historical significance, the galaxy data set provides a most interesting case study with which to illustrate the potential of prior-sensitivity analysis under the recursive pathway.

The outline of our presentation is as follows. In Section 5.1 we set forth the finite and infinite mixture models to be examined here and in Section 5.2 we describe the MCMC strategies we use to explore their complete and partial data posteriors. In Section 5.3 we discuss various astronomical motivations for our default hyperprior choices and, finally, in Section 5.4 we present the results of a biased sampling run on this problem with importance sample reweighting-based transformations between alternative priors.

5.1 Normal Mixture Model

5.1.1 *Finite mixture model*. Following Diebolt and Robert (1994) and Lee et al. (2008), we write the k-component Normal mixture model with component weights, ϕ, in the latent allocation variable form for data vector, y, and (unobserved) allocation vector, z, such that

$$\pi(z_i = j) = \phi_j \quad \text{and} \quad \pi(y_i|z_i = j) = f_{\mathcal{N}}(y_i|\theta_j).$$

Here $f_{\mathcal{N}}(\cdot|\theta_j)$ represents the one-dimensional Normal density, which we will reference in mean–precision syntax as $\mathcal{N}(\mu_j, \tau_j^{-1})$, that is, $\theta_j = \{\mu_j, \tau_j\}$.

Given priors for the number of components in the mixture, the distribution of weights at a given k and the vector of mean precisions—that is, $\pi(k)$, $\pi(\phi|k)$ and $\pi(\theta|\phi, k)$, respectively—the posterior for the number of mixture components in the *finite* mixture case may be recovered by integration over $\{\phi, \theta\}$ at each k,

$$\begin{aligned}\pi(k|y) &= \pi(k)/Z \\ &\quad \times \int_\Omega \pi(\phi|k)\pi(\theta|\phi, k)L(y|\theta, \phi, k)\,d\phi\,d\theta \\ &= \pi(k)Z^{(k)}/Z.\end{aligned}$$

Here the likelihood, $L(y|\theta, \phi, k)$, is given by a summation over the n_{tot} unobserved, z_i, as

$$L(y|\theta, \phi, k) = \prod_{i=1}^{n_{\text{tot}}} \sum_{j=1}^{k} \phi_j f_{\mathcal{N}}(y_i|\theta_j). \tag{18}$$

That is, for a $\pi(k)$ assigning mass to only a small set of elements, one approach to recovering $\pi(k|y)$ is simply to estimate the "per component" marginal likelihood, $\hat{Z}^{(k)}$, at each of these k and then reweight by $\pi(k)$. The full marginal likelihood of the model can then of course be estimated from the sum, $\hat{Z} = \sum_k \hat{Z}^{(k)}$. While this is indeed the strategy adopted here for exposition purposes, it is worth noting that such direct marginal likelihood estimation to recover $\pi(k|y)$ for this model can in fact be entirely avoided via either the reversible jump MCMC algorithm (Richardson and Green, 1997) or Gibbs sampling over the infinite mixture version described below.

5.1.2 *Infinite mixture model*. Rather than specify a maximum number of mixture components a priori, Escobar and West (1995) and Phillips and Smith (1996) (among others) have advocated an infinite-dimensional solution based on the Dirichet process prior. In particular, one may suppose the data to have been drawn from an infinite mixture of Normals with means, variances and weights drawn as the realization, Q, of a Dirichlet process (DP), $\text{DP}(M, G_0)$, on $\mathbb{R} \times \mathbb{R}_+$, the characterization of the DP being via a concentration index, M, and reference density, G_0, and with all Q being both normalized and strictly atomic. For small M ($\lesssim 10$) the tendency is for these Q to be dominated by only a few (mixture) components, while for large M the number of significant components inevitably increases, with the typical Q thereby becoming closer (in the metric of weak convergence) to G_0. The likelihood of i.i.d. y for a given Q requires (in theory) an infinite sum over the contribution from each of its components,

$$L(y|Q) = \prod_{i=1}^{n_{\text{tot}}} \sum_{j=1}^{\infty} \phi_j f_{\mathcal{N}}(y_i|\theta_j),$$

where each ϕ_j represents the limiting fraction of points in the realization assigned to a particular θ_j. (In practice, however, this summation can generally be truncated with negligible loss of accuracy after accounting for the contributions of only the most dominant components.) Computation of the marginal likelihood for the above model is thus nominally by integration over the infinite-dimensional space of Q. In particular, if we suppose a hyperprior density for the hyperparameters, ψ, of the DP (i.e., for M and the controlling parameters of G_0), we have $Z = \int_{\Omega(\psi)} \int_{\Omega(Q)} L(y|Q)\{dP_{Q|\psi}(Q)\}\pi(\psi)\,d\psi$.

As per the finite mixture case, we can simplify our posterior exploration and relevant computations by introducing latent variables, z and θ, for allocation of the

y and the corresponding mean-precision vectors of the parent components in Q. In this version the likelihood takes the form

$$L(y|\{z,\theta\}) = \prod_{i=1}^{n_{\text{tot}}} f_{\mathcal{N}}(y_i|\theta_{z_i}),$$

and the marginal likelihood becomes

$$Z = \int_{\Omega(\psi)} \int_{\Omega(Q)} \int_{\Omega(\{z,\theta\})} L(y|\{z,\theta\}) \cdot \{dP_{\{z,\theta\}|Q}(\{z,\theta\})\} \cdot \{dP_{Q|\psi}(Q)\}\pi(\psi)\,d\psi. \tag{19}$$

Importantly, existing Gibbs sampling methods for the DP allow for collapsed sampling from the posterior for $\{z,\theta,\psi\}$ and Equation (19) can be reduced to $\int_{\Omega(\{z,\theta,\psi\})} L(y|\{z,\theta\})\{dP_{\{z,\theta,\psi\}}(\{z,\theta,\psi\})\}$. In one further twist, however, we note that since the reduced expression is degenerate across component labelings, it is in fact more computationally efficient to estimate Z from

$$\int_{\Omega(\{z,\theta,\psi\})} \int_{\Omega(\hat{Q})} L(y|\hat{Q}) \cdot \{dP_{\hat{Q}|\{z,\theta,\psi\}}\} \cdot \{dP_{\{z,\theta,\psi\}}(\{z,\theta,\psi\})\}, \tag{20}$$

where $P_{\hat{Q}|\{z,\theta,\psi\}}$ takes a particularly simple analytic form by the nature of the DP (cf. Escobar and West, 1995).

Finally, it is important to note that since each realization of the DP has always an infinite number of components with probability one (though usually only a few with significant mass), the usual interpretation for the posterior, $\pi(k|y)$, in this context is the posterior distribution of the number of unique label assignments *among the observed data set* (i.e., the dimension of θ in $\{z,\theta\}$). However, although pragmatically useful for such modeling problems as that exhibited by the galaxy data set, as Miller and Harrison (2013) note, this estimate is not consistent.

5.2 MC Exploration of the Mixture Model Posterior

5.2.1 *Finite mixture model*. Exploration of the posterior for $\{\theta,\phi\}$ at fixed k in this finite mixture model can be accomplished rather efficiently (modulo the well-known problem of mixing *between* modes; cf. Neal, 1999) via Gibbs sampling given conjugate prior choices, as explained in detail by Richardson and Green (1997). To this end, we suppose

$$\mu_j \sim \mathcal{N}(\kappa,\xi^{-1}), \quad \tau_j \sim \Gamma(\alpha,\beta), \quad \text{and} \quad \beta \sim \Gamma(\beta_1,\beta_2),$$

where $\Gamma(a,b)$ represents the Gamma distribution with rate a and shape b. To simulate from the resulting posterior, we use the purpose-built code provided by `BMMmodel` and `JAGSrun` in the `BayesMix` package (Grün and Leisch, 2010) for `R`. No modifications to this code are necessary for sampling the partial data posterior, and both the partial and full data likelihoods given partial likelihood draws (at fixed k) may be recovered with Equation (18). The range of k for which we compute marginal likelihoods is here limited by the range of a truncated Poisson prior on k.

5.2.2 *Infinite mixture model*. As noted earlier, exploration of the infinite mixture model posterior can also be facilitated through Gibbs sampling with the appropriate choice of priors (Escobar and West, 1995); and although contemporary codes typically use the (more efficient) alternative algorithm of Neal (2000), the prior forms dictated by the conjugacy necessary for Gibbs sampling remain the default. Hence, to this end, we suppose a fixed concentration index of $M = 1$ and a Normal-Gamma reference density,

$$G_0 : \tau_j \sim \Gamma(s/2, S/2), \quad \mu_j|\tau_j \sim \mathcal{N}(m, \tau_j h),$$

assigning hyperpriors of $h \sim \Gamma(h_1/2, h_2/2)$ and $1/S \sim \Gamma(\nu_1/2, \nu_2/2)$. Here we use the `DPdensity` function in the `DPpackage` (Jara et al., 2011) for `R` to explore this posterior. While no modifications to this code are required for sampling the partial likelihood posteriors, the computation of full data likelihoods given the partial likelihood posterior requires that we sample a series of dummy components from the current posterior until some appropriate truncation point, $k' : \sum_{j=1}^{k'} \phi_j \approx 1$, before applying (the k'-truncated version of) Equation (20).

5.3 Astronomical Motivations for our Priors

5.3.1 *Finite mixture model*. As noted earlier, by considering the well-defined selection function of their observational campaign, the authors of the original astronomical study were able to construct the expected probability density function of recession velocities for their survey under the null hypothesis of a uniform distribution of galaxies throughout space. In particular, Postman, Huchra and Geller (1986) recognized that the

strict *apparent* magnitude limit of their spectroscopic targeting strategy ($m_r < 15.7$ mag) would act as a luminosity (or *absolute* magnitude) limit evolving with recession velocity (distance) according to

$$M_{r,\mathrm{lim}}(v) \approx m_r - 5\log_{10}(v) - 30,$$

where we have assumed units of 1000 $\mathrm{km\,s^{-1}}$ for v and a "Hubble constant" of $H_0 = 100\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$. To estimate the form of the resulting selection function, $S_{\mathrm{mag}}(v)$, Postman, Huchra and Geller (1986) considered how the relative number of galaxies per unit volume brighter than this limit would vary with distance given the absolute magnitude distribution function, $F_{\mathrm{mag}}(\cdot)$, for galaxies in the local Universe, that is, $S_{\mathrm{mag}}(v) \propto 1 - F_{\mathrm{mag}}(M_{r,\mathrm{lim}}(v))$. To approximate the latter, the astronomers simply integrated over a previous estimate of the local luminosity density parameterized as a Schechter function (Schechter, 1976) with characteristic magnitude, $M_r^* \approx -19.40 - 1.5$ mag, and faint-end slope, $\alpha_r^* \approx -1.3$, such that

$$f(M) \propto \left[10^{2/5(M_r^* - M)}\right]^{\alpha_r^*+1} \exp\left[-10^{2/5(M_r^* - M)}\right]$$

and

$$S_{\mathrm{mag}}(v) \propto \int_{-\infty}^{M_{r,\mathrm{lim}}(v)} f(M)\, dM.$$

An interesting feature of magnitude-limited astronomical surveys is that, although with increasing recession velocity this $S_{\mathrm{mag}}(v)$ selection function restricts their sampling to the decreasing fraction of galaxies above $M_{r,\mathrm{lim}}(v)$, the volume of the Universe probed by (the projection into three-dimensional space of) their two-dimensional angular viewing window is, in contrast, rapidly increasing. Hence, there exists an important additional selection effect, $S_{\mathrm{vol}}(v)$, operating in competition with, and initially dominating, that on magnitude, and scaling with (roughly) the third power of recession velocity such that

$$S_{\mathrm{vol}}(v) \propto v^3.$$

The product of these two effects therefore returns the net selection function of the galaxy data set, which we illustrate (along with each effect in isolation) in Figure 4 (see also Figure 4b from Postman, Huchra and Geller, 1986); the point being that there do exist informative astronomical considerations for choosing at least some of the hyperparameters of our priors in this mixture modeling case study, though past analyses have tended to ignore this context (contributing somewhat to the apparent "failure" of Bayesian mixture modeling for this data set; Aitkin, 2001). In particular, the shape of the selection function in velocity

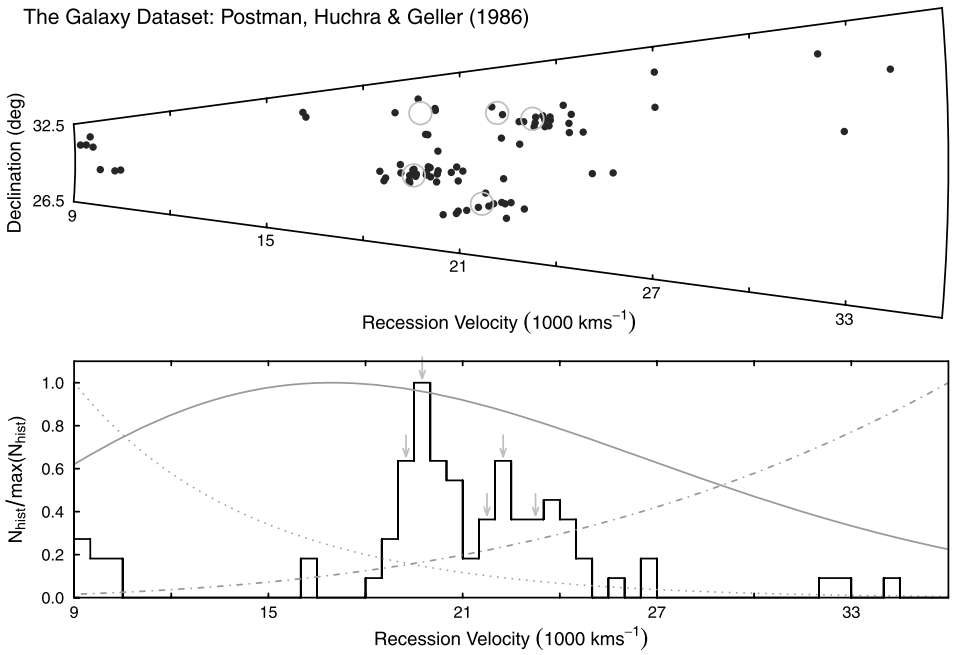

FIG. 4. *Visualization of the galaxy data set, including its Abell clusters and selection function. The clustering of galaxies in recession velocity–declination space is illustrated by way of the "cone diagram" shown in the top panel and its projection to a recession velocity histogram shown in the bottom panel. The positions of five Abell clusters targeted by the original survey are also highlighted here (open circles and arrows in light grey), along with the survey's magnitude-dependent, volume-dependent and net selection functions (shown as the dotted, dash-dotted and solid curves, respectively, in the bottom panel).*

space suggests the form for our prior on the component means: a choice of $\{\kappa = 17, \xi = 0.008\}$ gives a reasonable match to the shape of $S_{\text{mag}}(v)S_{\text{vol}}(v)$. Perhaps surprisingly, as we will demonstrate later, the choice of prior on the component means has a substantial influence on the resulting $\pi(k|y)$; changing only these of our hyperparameters to "data-driven" values chosen as $\{\kappa = \bar{y}\ (20.8), \xi = 1/\text{var}\{y\}\ (0.048)\}$ results in a drastic shift of the posterior.

Likewise, we can inform our prior choice for the number of components in the mixture with regard to the original survey design, which featured five separate observational windows placed so as to cover five previously identified galaxy clusters from the Abell catalog. (The positions of these clusters in bivariate recession velocity–declination space, and its projection to univariate velocity space, are also marked on Figure 4 for reference.) Hence, we select a mode of $\lambda = 5$ for our truncated Poisson prior for $\pi(k)$. With the $k = 1$ and $k = 2$ mixture models already well excluded by previous analyses, and $k > 10$ a pragmatic upper bound for exploration given $\lambda = 5$, we therefore truncate our prior to the range $3 \leq k \leq 10$. This contrasts somewhat with the Uniform priors on $k \leq 10$ and $k \leq 30$ assumed by Roeder and Wasserman (1997) and Richardson and Green (1997), respectively—though reweighting for alternative $\pi(k)$ (on this support) is trivial in any case.

Only the precisions of the Normal mixture components are not well constrained from astronomical considerations—although we can at least be confident that any large-scale clustering should occur above the scale of individual galaxy clusters (∼1 Mpc or $\Delta v \approx 0.1$) and (unless the uniform space-filling hypothesis were correct) well below the width of our selection function. Thus, we simply adopt a fixed shape hyperparameter of $\alpha = 2$ for our Gamma prior on the τ_j and allow the rate hyperparameter to vary according to its Gamma hyperprior form $\beta_1 = 1$ and $\beta_2 = 0.05$. Our choice here is thus comparable to that of Richardson and Green (1997) who suppose $\pi(\beta) \sim \Gamma(0.2, 0.016)$—not $\Gamma(0.2, 0.573)$ as misquoted by Aitkin (2001)—though we evidently place far less prior weight on exceedingly large precisions (small variances).

5.3.2 *Infinite mixture model.* The same considerations can also help shape our hyperparameter choices for the priors on our infinite mixture model. In particular, we take $\{m = 17, s = 4, h_1 = 2, h_2 = 8, \nu_1 = 1, \nu_2 = 1\}$ for the hyperparameters shaping the Normal-Gamma reference density, G_0, with the aim of matching as closely as possible to the priors of our finite-dimensional model. With the scale parameter of our prior on the component precisions taking an inverse-Gamma hyperprior form in the infinite case and a Gamma form in the finite case, it was not possible to exactly match these distributions: our choice of $\{\nu_1 = 1, \nu_2 = 1\}$ is intended to at least give comparable 5% and 95% quantiles. Finally, we adopt a fixed value for the concentration parameter of $M = 1$; this choice coincidentally gives a similar effective prior for the number of unique components among the 82 observed galaxies to that of the $\pi(k)$ adopted for our finite mixture model (see Escobar and West, 1995, for instance).

5.4 Numerical Results

5.4.1 *Chib example.* As an initial verification of our code, we first run the Gibbs sampling procedure outlined above (Section 5.2.1) to explore the partial data posteriors of a three-component (unequal variance) mixture model using the priors from Chib (1995), with the biased sampling algorithm then applied for marginal likelihood estimation. Neal (1999) has made public the results of a 10^8 draw AME calculation providing a precise benchmark for the marginal likelihood under these priors of -226.791 (± 0.089) (SE), though it should be noted that the galaxy data set used for this purpose is that *with* Chib's transcription error in the 78th observation (which we insert explicitly into the public `R` version for the present application only). Given just 200 saved draws from Gibbs sampling (at a thinning rate of 0.9) of the partial data posterior at each of 10 steps spaced as $r_j = \lfloor n_{\text{tot}} \times \{0, 1/9, 2/9, \ldots, 1, \}^{c=2} \rfloor$ (with r_2 reset to 3 to facilitate sampling), we can confirm the recovery of this benchmark as -226.79 (± 0.15) (SE). Estimation of the (single run) standard error (SE) was for this purpose conducted via 1000 repeat simulations. Further repeats of this procedure with both more posterior focused ($c = 0.5, 1$) and more prior focused ($c = 4$) partial data schedules confirm the optimality of the $c = 2$ choice anticipated from Fisher information principles (Section 2.4). The results of this experiment are illustrated in Figure 5.

5.4.2 *Finite mixture model.* To estimate "per component" marginal likelihoods for each k ($3 \leq k \leq 10$) in our finite mixture model, we run the same procedure of partial data posterior exploration followed by biased sampling with 4000 draws from each of ten steps on the $c = 2$ bridging sequence. The results of this computation are illustrated in Figure 6; the uncertainties indicated are gauged from the asymptotic covariance ma-

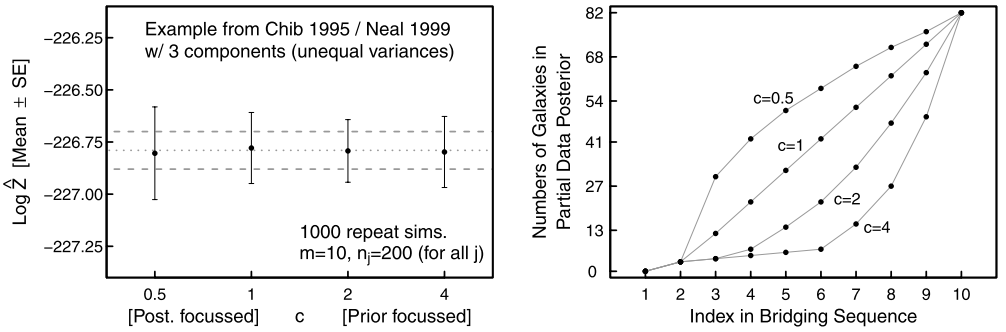

FIG. 5. *Relationship between the standard error (SE) of* $\log Z$ *estimation and the choice of a partial data posterior bridging sequence for biased sampling of the* 3*-component (unequal variance) mixture model under the Chib* (1995) *priors. The data points in the left-hand panel represent the mean* $\log \hat{Z}$ *and the error bars its (single run) SE, computed from* 1000 *repeat simulations with* 200 *draws from each of* 10 *steps in the bridging sequence. The dashed grey lines indicate the benchmark estimate (*$\pm$ *SE) from Neal* (1999)*, and the c values of the horizontal axis refer to the design of the partial data posterior bridging sequence as* $r_j = n_{\text{tot}} \times \{0, 1/9, 2/9, \ldots, 1\}^c$*. These sequences are also illustrated graphically for clarity in the right-hand panel.*

trix of the biased sampling estimator (as per Gill, Vardi and Wellner, 1988). We recover a posterior mode of $k = 7$ components, the recession velocity density belonging to which at the corresponding mode in $\{\phi, \theta\}$ is also illustrated in Figure 6 for reference. To the eye, it appears that $k = 7$ may be a slight overestimate since the third and fourth components (in order of increasing recession velocity) are more or less on top of each other, suggesting that one is being used to account for a slight non-Normality in the shape of this peak.

To demonstrate the potential for efficient prior-sensitivity analysis via importance sample reweighting of the pseudo-mixture density of partial data posteriors normalized by biased sampling (Section 2.3), we begin by recovering the Richardson and Green (1997) result from the above simulation output. The results of this reweighting procedure are shown in Figure 7. Since the Richardson and Green (1997) priors are significantly different to those chosen here from astronomical considerations (as discussed in Section 5.3), the effective sample sizes provided by our pseudo-mixture of 4000×10 draws range from just 13 to 928, yet the resulting approximation to the former benchmark is actually rather good. Moreover, the corresponding 95% credible intervals [recovered via bootstrap resampling from our pseudo-mixture plus estimates of the asymptotic covariance matrix for each $\log \hat{Z}^{(k)}$] indeed enclose all eight $\pi(k|y)$ reference points.

As a second demonstration we also show in Figure 7 the results of reweighting for alternative "data-

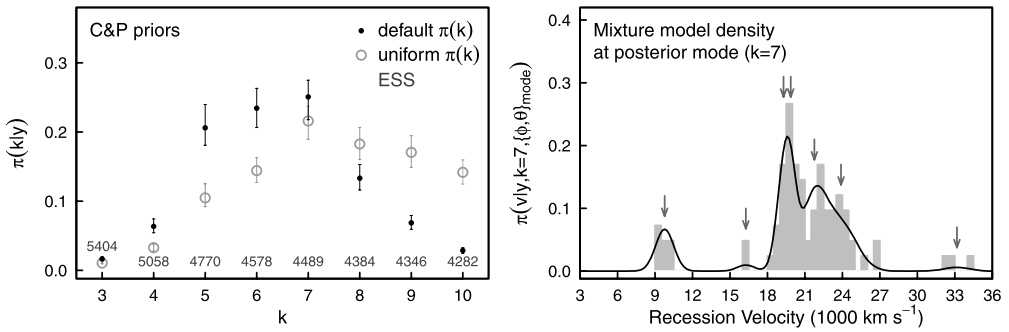

FIG. 6. *Posterior probabilities for the number of Normal mixture components in the galaxy data set,* $\pi(k|y)$*, under our astronomically motivated priors (left-hand panel). The solid, dark grey symbols here denote the true posterior, while the open, light grey symbols indicate for reference the raw, "per component" marginal likelihood-based result, that is, before application of our truncated Poisson* $\pi(k)$*. In each case the relevant uncertainties [recovered from estimates of the asymptotic covariance matrix for each* $\log \hat{Z}^{(k)}$*] are illustrated as* 95% *credible interval error bars. The effective sample size (ESS) provided by the pseudo-mixture of* 10 *partial data posteriors sampled for* 4000 *draws each is noted in grey for each* k*. The inferred probability density (in velocity space) at the maximum a posteriori parameterization of our Normal mixture model (*$k = 7$*) is then illustrated for reference against a scaled histogram of the galaxy data set in the right-hand panel.*

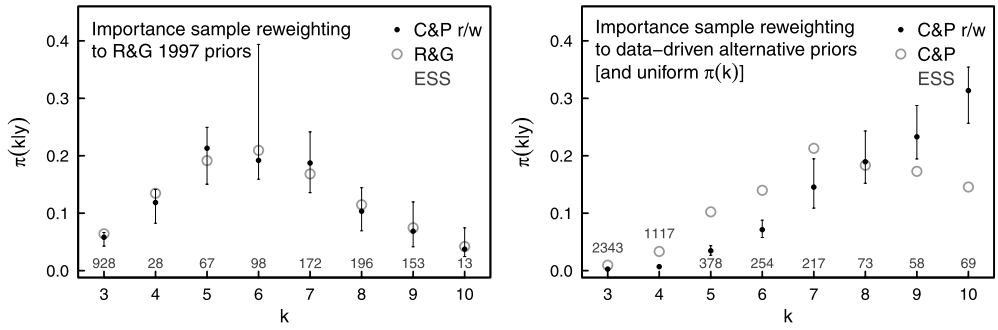

FIG. 7. *Importance sample reweighting of our draws from the pseudo-mixture of partial data posteriors used to estimate* $\pi(k|y)$ *under the Richardson and Green* (1997) *priors* (*left-hand panel*). *The solid, dark grey symbols here denote the reweighted posterior estimate, while the open, light grey symbols indicate for reference the Richardson and Green* (1997) *benchmark. The results of the equivalent procedure to approximate the effect of using alternative "data-driven" priors are shown in the right-hand panel; here the solid, dark grey symbols again represent the reweighted estimate, with the open, light grey symbols illustrating the reference point provided by our default priors. In both instances we treat* $\pi(k)$ *as uniform to emphasize the difference to the "per component" marginal likelihoods made by this modest change of prior. In each panel the relevant uncertainties* [*recovered via bootstrap resampling from our pseudo-mixture plus estimates of the asymptotic covariance matrix for each* $\log \hat{Z}^{(k)}$] *are illustrated as* 95% *credible interval error bars. The effective sample size* (*ESS*) *provided by the pseudo-mixture of* 10 *partial data posteriors sampled for* 4000 *draws each is noted in grey for each* k.

driven" choices for the hyperparameters of our prior on the component means: $\{\kappa = \bar{y}\ (20.8), \xi = 1/\operatorname{var}\{y\}\ (0.048)\}$. To emphasize the large difference this small change in $\pi(\theta)$ makes to the "per component" $\log Z^{(k)}$ values, the comparison presented is between our default and "data-driven" priors with $\pi(k)$ removed (i.e., treated as uniform). This investigation clearly confirms the remarkable prior sensitivity of $\pi(k|y)$ in the galaxy data set example. Interestingly, the preference under our "data-driven" priors is for an even greater number of mixture components ($k > 7$), despite the $k = 7$ solution already seeming (visually) to be an overfitting of the available data.

5.4.3 *Infinite mixture model.* In Figure 8 we present the results of Gibbs sampling the posterior of our infinite mixture model. In particular, we show in the left-hand panel of this figure the posterior for the number of unique label assignments among the galaxy data set, which, as we have noted earlier, is typically used as a proxy for the number of mixture components present (although under the Dirichet process prior this is formally always infinite). In the right-hand panel we demonstrate again the power of importance sample

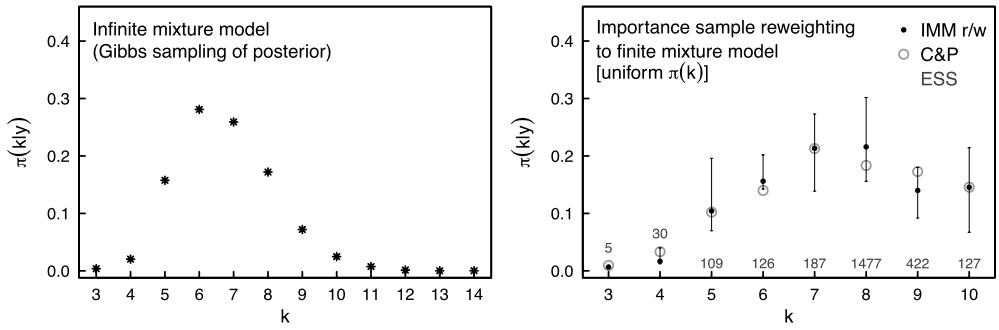

FIG. 8. *Posterior for the number of unique label assignments among the galaxy data set recovered from Gibbs sampling of our* (*Dirichlet process-based*) *infinite mixture model* (*left-hand panel*). *The results of importance sample reweighting of these draws, combined as a pseudo-mixture with those simulated under our bridging sequence of partial data posteriors, are shown via the solid, dark grey symbols in the right-hand panel. The target of this reweighting procedure is the "per component"* [*i.e., uniform* $\pi(k)$] *posterior for the number of mixture model components in our benchmark finite mixture model* (*shown as the open, light grey symbols*). *The relevant uncertainties* [*recovered via bootstrap resampling from our pseudo-mixture plus estimates of the asymptotic covariance matrix for each* $\log \hat{Z}^{(k)}$] *are illustrated as* 95% *credible interval error bars. The effective sample size* (*ESS*) *provided by the pseudo-mixture of* 10 *partial data posteriors sampled for* 4000 *draws each is noted in grey for each* k.

reweighting for prior-sensitivity analysis, though for this particular case the stochastic process prior used requires that we apply the appropriate Radon–Nikodym derivative version given by Equation (14).

The Radon–Nikodym derivative, $\frac{dP_{\{z,\theta,\psi\},\text{alt}}}{dP_{\{z,\theta,\psi\}}}(\{z,\theta,\psi\})$, of the measure on $\{z,\theta,\psi\}$ assigned by a k-component finite mixture model with respect to that assigned by the Dirichlet process prior of our infinite mixture model may be computed as follows. First, we observe that the Radon–Nikodym derivative between two Dirichlet process priors on the (equivalent) space of $\{\{\theta_1,\ldots,\theta_{n_{\text{tot}}}\},\psi\}$ (with the θ_i possibly nonunique) has been previously derived by Doss (2012), thereby providing a direct formula for computing $\frac{dP_{\{z,\theta,\psi\},\text{int}}}{dP_{\{z,\theta,\psi\}}}(\{z,\theta,\psi\})$, where $P_{\{z,\theta,\psi\},\text{int}}$ represents a Dirichlet process prior with hyperpriors on the ψ of its reference density chosen to be identical to those on the $\{\mu_j,\tau_j\}$ and β of our finite mixture model. That is, we choose $P_{\{z,\theta,\psi\},\text{int}}$ such that its projection to $P_{\{z,\theta,\psi\},\text{int}}$ for z with k unique elements is equivalent (a.e.) to that of $P_{\{z,\theta\},\text{alt}}$ with our hyperparameter on β integrated out, allowing $P_{\{z,\theta,\psi\},\text{alt}}$ to be defined identical to $f(z)P_{\{z,\theta,\psi\},\text{int}}$. The necessary $f(z)$ to ensure that $\frac{dP_{\{z,\theta,\psi\},\text{alt}}}{dP_{\{z,\theta,\psi\},\text{int}}}\frac{dP_{\{z,\theta,\psi\},\text{int}}}{dP_{\{z,\theta,\psi\}}} = \frac{dP_{\{z,\theta,\psi\},\text{alt}}}{dP_{\{z,\theta,\psi\}}}$ is then simply the ratio of the labeling probabilities under our finite mixture model and the intermediate version of our infinite mixture model [with $f(z)\neq 0$ *only* where the number of unique elements in z equals k].

A formula for the desired $f(z)$ can be derived by combining standard properties of the Dirichlet-Multinomial distribution (our finite-dimensional model prior on z) with results from the work of Antoniak (1974) on the marginals of the Dirichlet process. In each case the probability of a given labeling sequence depends not on its ordering, but rather on its vector of per-label counts. Using Antoniak's system of writing $C(m_1,m_2,\ldots,m_{n_{\text{tot}}})$ as the set of labelings with m_1 unique elements, m_2 pairs, etc., we have wherever $\sum_{i=1}^{n_{\text{tot}}} m_i \le k$,

$$f(z\in C) = \left(\frac{n_{\text{tot}}!}{\prod_{i=1}^{n_{\text{tot}}}(i!)^{m_i}}\frac{\Gamma(k\alpha)}{\Gamma(n_{\text{tot}}+k\alpha)} \cdot\prod_{i=1}^{n_{\text{tot}}}\left(\frac{\Gamma(i+\alpha)}{\Gamma(\alpha)}\right)^{m_i}\right) \Big/\left(\frac{(k-\sum_{i=1}^{n_{\text{tot}}} m_i)!(\prod m_i)!}{k!} \cdot\frac{n_{\text{tot}}!}{\prod_{i=1}^{n_{\text{tot}}} i^{m_i}(m_i!)}\frac{M^{\sum_{i=1}^{n_{\text{tot}}} m_i}}{M^{[n_{\text{tot}}]}}\right),$$

where $x^{[j]}$ denotes the rising factorial function as per Proposition 3 of Antoniak (1974). For our case of $\alpha = 1$ and $M = 1$ this reduces to

$$f(z\in C) = \left(\frac{n_{\text{tot}}!(k-1)!}{(n_{\text{tot}}+k-1)!}\right) \Big/\left(\frac{(k-\sum_{i=1}^{n_{\text{tot}}} m_i)!(\prod m_i)!}{k!\prod_{i=1}^{n_{\text{tot}}} i^{m_i}(m_i!)}\right).$$

6. CONCLUSIONS

In this paper we have presented an extensive review of the recursive pathway to marginal likelihood estimation as characterized by biased sampling, reverse logistic regression and the density of states; in particular, we have highlighted the diversity of bridging sequences amenable to recursive normalization and the utility of the resulting pseudo-mixtures for prior-sensitivity analysis (in the Bayesian context). Our key theoretical contributions have included the introduction of a novel heuristic ("thermodynamic integration via importance sampling") for guiding design of the bridging sequence and an elucidation of various connections between these recursive estimators and the nested sampling technique. Our two numerical case studies illustrate in depth the practical implementation of these ideas using both "ordinary" and stochastic process priors.

REFERENCES

AITKIN, M. (2001). Likelihood and Bayesian analysis of mixtures. *Statist. Model.* **1** 287–304.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

ARIMA, S. and TARDELLA, L. (2012). Improved harmonic mean estimator for phylogenetic model evidence. *J. Comput. Biol.* **19** 418–438. MR2913981

BAELE, G., LEMEY, P., BEDFORD, T., RAMBAUT, A., SUCHARD, M. A. and ALEKSEYENKO, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29** 2157–2167.

BAILER-JONES, C. A. L. (2012). A Bayesian method for the analysis of deterministic and stochastic time series. *Astron. Astrophys.* **546** A89.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York. MR0233396

BREWER, B. J. and STELLO, D. (2009). Gaussian process modelling of asteroseismic data. *Mon. Not. R. Astron. Soc.* **395** 2226–2233.

CAIMO, A. and FRIEL, N. (2013). Bayesian model selection for exponential random graph models. *Social Networks* **35** 11–24.

CALDERHEAD, B. and GIROLAMI, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. *Comput. Statist. Data Anal.* **53** 4028–4045. MR2744303

CAMERON, E. and PETTITT, A. N. (2013). On the evidence for cosmic variation of the fine structure constant (II): A semi-parametric Bayesian model selection analysis of the quasar dataset. Preprint. Available at arXiv:1309.2737.

CAPPÉ, O., GUILLIN, A., MARIN, J. M. and ROBERT, C. P. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.* **13** 907–929. MR2109057

CHEN, M.-H. and SHAO, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25** 1563–1594. MR1463565

CHEN, M.-H., SHAO, Q.-M. and IBRAHIM, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York. MR1742311

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321. MR1379473

CHOPIN, N. (2002). A sequential particle filter method for static models. *Biometrika* **89** 539–551. MR1929161

CHOPIN, N. and ROBERT, C. P. (2010). Properties of nested sampling. *Biometrika* **97** 741–755. MR2672495

CORNUET, J.-M., MARIN, J.-M., MIRA, A. and ROBERT, C. P. (2012). Adaptive multiple importance sampling. *Scand. J. Stat.* **39** 798–812. MR3000850

DAVIS, T. M. et al. (2007). Scrutinizing exotic cosmological models using ESSENCE supernova data combined with other cosmological probes. *Astrophys. J.* **666** 716–725.

DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 411–436. MR2278333

DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56** 363–375. MR1281940

DOSS, H. (2012). Hyperparameter and model selection for nonparametric Bayes problems via Radon–Nikodym derivatives. *Statist. Sinica* **22** 1–26. MR2933165

DUDLEY, R. M. and PHILIPP, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrsch. Verw. Gebiete* **62** 509–552. MR0690575

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

EVANS, M., ROBERT, C. P., DAVISON, A. C., JIANG, W., TANNER, M. A., DOSS, H., QIN, J., FOKIANOS, K., MACEACHERN, S. N., PERUGGIA, M., GUHA, S., CHIB, S., RITOV, Y., ROBINS, J. M. and VARDI, Y. (2003). Discussion on the paper by Kong, McCullagh, Meng, Nicolas and Tan. *J. Roy. Statist. Soc. B* **65** 604–618.

FAN, Y., RUI, W., CHEN, M.-H., KUO, L. and LEWIS, P. O. (2012). Choosing among partition models in Bayesian phylogenetics. *Mol. Biol. Evol.* **28** 523–532.

FEROZ, F. and HOBSON, M. P. (2008). Multimodal nested sampling: An efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Mon. Not. R. Astron. Soc.* **384** 449–463.

FEROZ, F., HOBSON, M. P., CAMERON, E. and PETTITT, A. N. (2013). Importance nested sampling and the multinest algorithm. Preprint. Available at arXiv:1306.2144.

FERRENBERG, A. M. and SWENDSEN, R. H. (1989). Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63** 1195–1198.

FRIEL, N. and PETTITT, A. N. (2008). Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 589–607. MR2420416

FRIEL, N. and WYSE, J. (2012). Estimating the evidence—A review. *Stat. Neerl.* **66** 288–308. MR2955421

GELFAND, A. E. and DEY, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514. MR1278223

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2027492

GEYER, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7** 473–483.

GEYER, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report 568, School of Statistics, Univ. Minnesota, Minneapolis, MN.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B* **54** 657–699. MR1185217

GILL, R. D., VARDI, Y. and WELLNER, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16** 1069–1112. MR0959189

GRÜN, B. and LEISCH, F. (2010). BayesMix: An R package for Bayesian mixture modeling. Technical report.

GUNN, J. E. and GOTT, J. R. III (1972). On the infall of matter into clusters of galaxies and some effects on their evolution. *Astrophys. J.* **176** 1–19.

HABECK, M. (2012). Evaluation of marginal likelihoods via the density of states. *J. Mach. Learn. Res.* **22** 486–494.

HALMOS, P. R. (1950). *Measure Theory*. Van Nostrand, New York. MR0033869

HALMOS, P. R. and SAVAGE, L. J. (1949). Application of the Radon–Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.* **20** 225–241. MR0030730

HESTERBERG, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37** 185–194.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176

HÖRMANDER, L. (1983). *The Analysis of Linear Partial Differential Operators. I: Distribution Theory and Fourier Analysis. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **256**. Springer, Berlin. MR0717035

JARA, A., HANSON, T. E., QUINTANA, F. A., MÜLLER, P. and ROSNER, G. L. (2011). DPpackage: Bayesian semi- and nonparametric modelling in R. *J. Statist. Softw.* **40** 1–30.

JASRA, A., HOLMES, C. C. and STEPHENS, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20** 50–67. MR2182987

JAYNES, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge Univ. Press, Cambridge. MR1992316

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. MR0187257

JEFFREYS, W. H. and BERGER, J. O. (1991). Sharpening Ockham's razor on a Bayesian strop. Technical Report 91-44C, Dept. Statistics, Purdue Univ., West Lafayette, IN.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

KILBINGER, M., WRAITH, D., ROBERT, C. P., BENABED, K., CAPPÉ, O., CARDOSO, J.-F., FORT, G., PRUNET, S. and BOUCHET, F. R. (2010). Bayesian model comparison in cosmology with population Monte Carlo. *Mon. Not. R. Astron. Soc.* **405** 2381–2390.

KIRSHNER, R. P., OEMLER, A. JR., SCHECHTER, P. L. and SHECTMAN, S. A. (1981). A million cubic megaparsec void in Boötes? *Astrophys. J.* **248** 57–60.

KONG, A., LIU, J. S. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.

KONG, A., MCCULLAGH, P., MENG, X.-L., NICOLAE, D. and TAN, Z. (2003). A theory of statistical models for Monte Carlo integration. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 585–618. MR1998624

KUMAR, S., ROSENBERG, J. M., BOUZIDA, D., SWENDSEN, R. H. and KOLLMAN, P. A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13** 1011–1021.

LARTILLOT, N. and PHILLIPE, H. (2006). Computing Bayes factors using thermodynamic integration. *Syst. Biol.* **55** 195–207.

LEE, K., MARIN, J.-M., MENGERSEN, K. and ROBERT, C. P. (2008). Bayesian inference on mixtures of distributions. Preprint. Available at arXiv:0804.2413.

LEFEBVRE, G., STEELE, R. and VANDAL, A. C. (2010). A path sampling identity for computing the Kullback–Leibler and J divergences. *Comput. Statist. Data Anal.* **54** 1719–1731. MR2608968

LI, Y., NI, Z.-X. and LIN, J.-G. (2011). A stochastic simulation approach to model selection for stochastic volatility models. *Comm. Statist. Simulation Comput.* **40** 1043–1056. MR2792481

LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR1842342

MARIN, J.-M., PUDLO, P. and SEDKI, M. (2012). Consistency of the adaptive multiple importance sampling. Preprint. Available at arXiv:1301.2548.

MARIN, J.-M. and ROBERT, C. P. (2010). On resolving the Savage–Dickey paradox. *Electron. J. Stat.* **4** 643–654. MR2660536

MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406

MILLER, J. W. and HARRISON, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. Preprint. Available at arXiv:1301.2708v1.

MUKHERJEE, P., PARKINSON, D. and LIDDLE, A. R. (2006). A nested sampling algorithm for cosmological model selection. *Astrophys. J.* **638** L51–L54.

NEAL, R. (1999). Erroneous results in "Marginal likelihood from the Gibbs output." Available at http://www.cs.toronto.edu/~radford/ftp/chib-letter.pdf.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

NEAL, R. M. (2001). Annealed importance sampling. *Stat. Comput.* **11** 125–139. MR1837132

NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B* **56** 3–48. MR1257793

ORTEGA, J. M. and RHEINBOLDT, W. C. (1967). Monotone iterations for nonlinear equations with application to Gauss–Seidel methods. *SIAM J. Numer. Anal.* **4** 171–190. MR0215487

PFANZAGL, J. (1979). Conditional distributions as derivatives. *Ann. Probab.* **7** 1046–1050. MR0548898

PHILLIPS, D. B. and SMITH, A. F. M. (1996). Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice* 215–239. Chapman & Hall, London. MR1397970

POSTMAN, M., HUCHRA, J. P. and GELLER, M. J. (1986). Probes of large-scale structure in the Corona Borealis region. *Astrophys. J.* **92** 1238–1247.

RAFTERY, A. E., NEWTON, M. A., SATAGOPAN, J. M. and KRIVITSKY, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian Statistics* 8 371–416. Oxford Univ. Press, Oxford. MR2433201

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213

ROBERT, C. P. and WRAITH, D. (2009). Computational methods for Bayesian model choice. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*: *The* 29*th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. *AIP Conference Proceedings* **1193** 251–262. American Institute of Physics, New York.

ROEDER, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* **85** 617–624.

ROEDER, K. and WASSERMAN, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92** 894–902. MR1482121

SCHECHTER, P. (1976). An analytic expression for the luminosity function of galaxies. *Astrophys. J.* **203** 297–306.

SHIRTS, M. R. and CHODERA, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129** 124105.

SKILLING, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Anal.* **1** 833–859 (electronic). MR2282208

STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. MR1762903

TAN, Z., GALLICCHIO, E., LAPELOSA, M. and LEVY, R. M. (2012). Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *J. Chem. Phys.* **136** 144102.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762. MR1329166

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13** 178–205. MR0773161

WEINBERG, M. D. (2012). Computing the Bayes factor from a Markov chain Monte Carlo simulation of the posterior distribution. *Bayesian Anal.* **7** 737–769. MR2981634

WOLPERT, R. L. and SCHMIDLER, S. C. (2012). α-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statist. Sinica* **22** 1233–1251. MR2987490

XIE, W., LEWIS, P., FAN, Y., KUO, L. and CHEN, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **18** 1001–1013.