Warp Bridge Sampling
Author(s): Xiao-Li Meng and Stephen Schilling
Source: *Journal of Computational and Graphical Statistics*, Vol. 11, No. 3 (Sep., 2002), pp. 552-586
Published by: American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America
Stable URL: http://www.jstor.org/stable/1391113
Accessed: 08/03/2011 11:27

Warp Bridge Sampling

Xiao-Li MENG and Stephen SCHILLING

Bridge sampling, a general formulation of the acceptance ratio method in physics for computing free-energy difference, is an effective Monte Carlo method for computing normalizing constants of probability models. The method was originally proposed for cases where the probability models have overlapping support. Voter proposed the idea of shifting physical systems before applying the acceptance ratio method to calculate free-energy differences between systems that are highly separated in a configuration space. The purpose of this article is to push Voter's idea further by applying more general transformations, including stochastic transformations resulting from mixing over transformation groups, to the underlying variables before performing bridge sampling. We term such methods *warp bridge sampling* to highlight the fact that in addition to location shifting (i.e., centering) one can further reduce the difference/distance between two densities by warping their shapes without changing the normalizing constants. Real data-based empirical studies using the full information item factor model and a nonlinear mixed model are provided to demonstrate the potentially substantial gains in Monte Carlo efficiency by going beyond centering and by using efficient bridge sampling estimators. Our general method is also applicable to a couple of recent proposals for computing marginal likelihoods and Bayes factors because these methods turn out to be covered by the general bridge sampling framework.

Key Words: Bayes factors; Latent variables; Likelihood ratio; Markov chain Monte Carlo; Mixture; Normalizing constants; Orthogonal group; Orthogonal transformation.

1. INTRODUCTION AND BACKGROUND

1.1 BRIDGE SAMPLING

Computing the normalizing constant of a probability model, or more generally a definite integration, is a common problem in statistical and scientific studies. It is also a very difficult one when the model or integrand is complex and high dimensional, as in genetic linkage analysis and in theoretical physics. A brief review of these and other applications can be found in Meng and Wong (1996), which is one of a number of recent works in statistical

Xiao-Li Meng is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: meng@stat.harvard.edu). Stephen Schilling is Assistant Professor, School of Education, The University of Michigan, Ann Arbor, MI 48109 (E-mail: schillsg@umich.edu).

Journal of Computational and Graphical Statistics, Volume 11, Number 3, Pages 552–586
DOI: 10.1198/106186002457

literature concerning the search for effective methods for computing normalizing constants, either for specific models or in general (e.g., Chen and Shao 1997a, 1997b; Chib 1995; Chib and Jeliazkov 2001; DiCiccio, Kass, Raftery, and Wasserman 1997; Gelfand and Dey 1994; Gelman and Meng 1998; Geyer 1994; Johnson, 1999; Meng and Schilling 1996; Newton and Raftery 1994; Verdinelli and Wasserman 1995). In most applications, the objective is to compute a (log) ratio of two normalizing constants (e.g., a likelihood ratio, a free-energy difference) instead of just a single normalizing constant. Even if the interest is on a single normalizing constant, it is often beneficial to construct a convenient *matching* density with known normalizing constant and then implement the following method, as in DiCiccio et al. (1997) and Chib and Jeliazkov (2001; see Sections 1.2–1.3 for explanation).

To fix the idea, let $p_i(w), w \in \Omega_i, i = 1, 2$, be two densities with respect to a common measure μ, with each density known up to a normalizing constant: $p_i(w) = q_i(w)/c_i$. We have draws from each of the densities, and we want to use these *available* draws to estimate $r = c_1/c_2$ or $\lambda = \log r$. Bridge sampling (Bennett 1976; Meng and Wong 1996) is an effective method for addressing this problem. We first assume that the two densities have overlapping support; that is, $\mu(\Omega_1 \cap \Omega_2) > 0$. It is then trivial to verify that for any $\alpha(w)$ such that

$$0 < \left| \int_{\Omega_1 \cap \Omega_2} \alpha(w) p_1(w) p_2(w) \mu(dw) \right| < \infty, \tag{1.1}$$

the following identity holds (where E_i denotes the expectation with respect to $p_i, i = 1, 2$):

$$r \equiv \frac{c_1}{c_2} = \frac{E_2[q_1(w)\alpha(w)]}{E_1[q_2(w)\alpha(w)]}. \tag{1.2}$$

Consequently, for any given α, the corresponding bridge sampling estimate for r is given by

$$\hat{r}_\alpha = \frac{\frac{1}{n_2} \sum_{j=1}^{n_2} q_1(w_{2j})\alpha(w_{2j})}{\frac{1}{n_1} \sum_{j=1}^{n_1} q_2(w_{1j})\alpha(w_{1j})}, \tag{1.3}$$

where $\{w_{i1}, \ldots, w_{in_i}\}$ are (possibly dependent) draws from $p_i(w), i = 1, 2$.

Under the assumption that all the draws are *independent*, it has been shown (e.g., Bennett 1976; Meng and Wong 1996) that the optimal choice of α in the sense of minimizing the asymptotic variance of $\hat{\lambda}_\alpha = \log \hat{r}_\alpha$, or equivalently the asymptotic relative variance $E(\hat{r}_\alpha - r)^2/r^2$, is

$$\alpha_O(w) \propto \frac{1}{s_1 q_1 + s_2 r q_2}, \quad w \in \Omega_1 \cap \Omega_2, \tag{1.4}$$

where $s_i = n_i/(n_1 + n_2), i = 1, 2$. When draws are not independent, which typically is the case in practice, $\alpha_O(w)$ of (1.4) is no longer optimal, and it is still an open problem to find the optimal choice of α in the general dependent setting. Nevertheless, empirical evidence (e.g., see Meng and Schilling 1996; Servidea 2002) shows that the use of $\alpha_O(w)$ is still quite good in many situations unless the dependences among the draws are very strong and uneven between p_1 and p_2. A first-order adjustment for the dependence is to use the "effective size," such as $\tilde{n}_i = n_i(1-\rho_i)/(1+\rho_i)$ with an appropriately estimated autocorrelation $\rho_i(i = 1,2)$ in defining the weight s_i in (1.4), as in the application of Section 4.

Since $\alpha_O(w)$ depends on the unknown r, Meng and Wong (1996) suggested the following iterative sequence

$$\hat{r}_O^{(t+1)} = \frac{\frac{1}{n_2}\sum_{j=1}^{n_2}\left[\frac{l_{2j}}{s_1 l_{2j}+s_2\hat{r}_O^{(t)}}\right]}{\frac{1}{n_1}\sum_{j=1}^{n_1}\left[\frac{1}{s_1 l_{1j}+s_2\hat{r}_O^{(t)}}\right]}, \qquad t = 0,1,2,\ldots, \tag{1.5}$$

where $l_{ij} = q_1(w_{ij})/q_2(w_{ij}), j = 1,\ldots,n_i, i = 1,2$, are calculated before iterating. Meng and Wong (1996) showed that this sequence has a unique limit $\hat{r}_O$ and that $|\hat{r}_O^{(t)} - \hat{r}_O|$ converges to zero monotonically in t; this convergence is also typically very rapid (e.g., less than five iterations), as demonstrated empirically in Meng and Schilling (1996). Furthermore, the asymptotic error of $\hat{\lambda}_O = \log \hat{r}_O$ is the same as that of the bridge sampling estimator using the unknown optimal α_O, namely,

$$\mathrm{V}(\hat{\lambda}_O) = \frac{1}{n}\left[\int_{\Omega_1\cap\Omega_2}((s_1p_1)^{-1}+(s_2p_2)^{-1})^{-1}dw\right]^{-1} - \frac{1}{n_1} - \frac{1}{n_2} + O\left(\frac{1}{n^2}\right). \tag{1.6}$$

1.2 INDIVIDUAL CASES

Although identity (1.2) is trivial, it covers many individual cases that are either well-known or are being made known. The most obvious case, of course, is the much-used importance sampling identity (e.g., Ott 1979; Geyer and Thompson 1992)

$$\frac{c_1}{c_2} = E_2\left[\frac{q_1(w)}{q_2(w)}\right], \qquad \text{assuming} \quad \Omega_1 \subseteq \Omega_2, \tag{1.7}$$

which is (1.2) with $\alpha = 1/q_2$. Other individual cases, such as *geometric bridge* with $\alpha = (q_1 q_2)^{-1/2}$, were studied and applied by Meng and Wong (1996), Meng and Schilling (1996), DiCiccio et al. (1997), Gelman and Meng (1998), Jensen and Kong (1999), among others. While in these articles the role of bridge sampling was clear, some recent proposals suggest that the generality and power of bridge sampling is yet to be fully recognized. For example, Johnson (1999) proposed to estimate r by the value of r that minimizes, in our notation

$$\left[\frac{1}{n_2}\sum_{j=1}^{n_2}\min\left(\frac{q_1(w_{2j})}{rq_2(w_{2j})},1\right) - \frac{1}{n_1}\sum_{j=1}^{n_1}\min\left(\frac{rq_2(w_{1j})}{q_1(w_{1j})},1\right)\right]^2$$
$$\equiv \left[\frac{1}{n_2}\sum_{j=1}^{n_2}\min\left(\frac{l_{2j}}{r},1\right) - \frac{1}{n_1}\sum_{j=1}^{n_1}\min\left(\frac{r}{l_{1j}},1\right)\right]^2. \tag{1.8}$$

This turns out to be the same as taking $\alpha = \min\left(q_1^{-1}, (rq_2)^{-1}\right)$ in (1.3) and then iterating in the same way as with (1.5). In fact, this choice of α is the limiting case of the *power family* of Meng and Wong (1996) as the power approaches zero.

As another example, Chib and Jeliazkov (2001) proposed the following generalization of Chib (1995) for computing a marginal likelihood. Suppose under a model $\mathcal{M}$ the sampling distribution is $f(y|\theta, \mathcal{M})$ and the prior on the parameter θ is $\pi(\theta|\mathcal{M})$, where both of them are assumed to be easy to evaluate as functions of θ. Noting that the posterior of θ given data y is

$$\pi(\theta|y, \mathcal{M}) = \frac{f(y|\theta, \mathcal{M})\pi(\theta|\mathcal{M})}{m(y|\mathcal{M})}, \tag{1.9}$$

Chib and Jeliazkov (2001) suggested to compute the *marginal likelihood* of model $\mathcal{M}$, $m(y|\mathcal{M})$ via

$$\log m(y|\mathcal{M}) = \log f(y|\theta^*, \mathcal{M}) + \log \pi(\theta^*|\mathcal{M}) - \log \pi(\theta^*|y, \mathcal{M}), \tag{1.10}$$

where θ^* is some fixed point of θ (e.g., a mode of $\pi(\theta|y, \mathcal{M})$). Suppose we have n_1 draws $\{\theta_{11}, \ldots, \theta_{1n_1}\}$ from $\pi(\theta|y, \mathcal{M})$ via a Metropolis–Hastings algorithm with the proposal density $\tilde{\pi}(\theta|\theta', y)$, then Chib and Jeliazkov's proposal is to estimate $\pi(\theta^*|y, \mathcal{M})$ by

$$\hat{\pi}(\theta^*|y, \mathcal{M}) = \frac{\frac{1}{n_1}\sum_{j=1}^{n_1} a(\theta_{1j}, \theta^*|y)\tilde{\pi}(\theta^*|\theta_{1j}, y)}{\frac{1}{n_2}\sum_{j=1}^{n_2} a(\theta^*, \theta_{2j})}, \tag{1.11}$$

where the n_2 draws $\{\theta_{2j}, \ldots, \theta_{2n_2}\}$ are from $\tilde{\pi}(\theta|\theta^*, y)$, and $a(\theta, \theta'|y)$ is the Metropolis–Hastings accepting probability

$$a(\theta, \theta'|y) = \min\left\{1, \frac{f(y|\theta', \mathcal{M})\pi(\theta'|\mathcal{M})}{f(y|\theta, \mathcal{M})\pi(\theta|\mathcal{M})} \frac{\tilde{\pi}(\theta|\theta', y)}{\tilde{\pi}(\theta'|\theta, y)}\right\}. \tag{1.12}$$

Once again it turns out that there is a simpler bridge-sampling derivation of Chib and Jeliazkov's (2001) method. Observing from (1.9) that the desired $m(y|\mathcal{M})(= c_1)$ is simply the normalizing constant of $p_1(\theta) \equiv \pi(\theta|y, \mathcal{M})$ with $q_1(\theta) \equiv f(y|\theta, \mathcal{M})\pi(\theta|\mathcal{M})$ as the unnormalized density, we can directly implement the bridge sampling by choosing $p_2(\theta) \equiv \tilde{\pi}(\theta|\theta^*, y)$ as the *matching* density. Since p_2 is completely known, we have $c_2 = 1$, and thus the ratio of constant $r = c_1/c_2$ is $m(y|\mathcal{M})$. A little algebra then shows that Chib and Jeliazkov's (2001) estimate of $m(y|\mathcal{M})$ is the bridge-sampling estimate (1.3) with (after equating θ with w)

$$\begin{aligned} \alpha(\theta) &= \min\left\{\frac{\tilde{\pi}(\theta^*|\theta, y)}{\tilde{\pi}(\theta|\theta^*, y)}, \frac{f(y|\theta^*, \mathcal{M})\pi(\theta^*|\mathcal{M})}{f(y|\theta, \mathcal{M})\pi(\theta|\mathcal{M})}\right\} \\ &= \min\left\{\frac{\tilde{\pi}(\theta^*|\theta, y)}{\tilde{\pi}(\theta|\theta^*, y)}, \frac{\pi(\theta^*|y, \mathcal{M})}{\pi(\theta|y, \mathcal{M})}\right\}. \end{aligned} \tag{1.13}$$

Other variations proposed in Chib and Jeliazkov (2001) are also individual cases of bridge sampling, as shown by Mira and Nicholls (2000).

1.3 Bridge Sampling Is Not Just For Ratios

It is sometime said that bridge sampling is useful only for estimating ratios of normalizing constants, not individual constants, and hence methods such as Chib and Jeliazkov's (2001) appear to be more general. It is indeed true that any application of bridge sampling requires draws from (at least) two densities, and that the resulting estimator is in the form of a ratio (or ratios). However, in applying bridge sampling we have great freedom in choosing the second density $p_2 = q_2/c_2$. If we choose it in such a way that $c_2 = 1$, then any bridge sample estimator will deliver a direct estimator of $c_1 \equiv c_1/1$. This is exactly what underlies Chib and Jeliazkov's (2001) method, because it corresponds to choosing p_2 as the completely known Metropolis–Hastings proposal density $\tilde{\pi}(\theta|\theta^*, y)$, and then choosing the bridge function given by (1.12).

We emphasize this point not to discount the contribution of Chib and Jeliazkov (2001), for using Metropolis–Hastings proposal distribution, which includes the full conditionals for Gibbs sampler (Chib 1995), as a matching density for implementing bridge sampling is a good idea. Indeed, as applied by DiCiccio et al. (1997), when estimating the ratio of normalizing constants of two distributions with very different structures (e.g., with different dimensions), as often occur is the context of computing Bayes factors, a good strategy is to bridge each density with a convenient approximation of itself and then apply bridge sampling to estimate each individual normalizing constant separately; a sensible Metropolis–Hastings proposal will serve this purpose quite well. This is typically much more effective than to artificially bridge the original two densities by, for example, augmenting the dimension of the lower one to match the higher one (Chen and Shao 1997b). The benefit of recasting a method in the general bridge sampling framework is that not only can we use the general identity (1.2) to avoid sometime tedious algebraic derivations with specific cases, but more importantly we can use the general bridge sampling theory to guide us in choosing better and often also simpler bridge function α (see Section 4.4 for an empirical demonstration). As users, we all want methods that are efficient and simple at the same time, but often in practice we have to make a compromise. This is one of those happy situations where the efficient choice is typically much simpler than many nonefficient choices guided by distracting details of individual cases. Furthermore, the bridge sampling perspective also implies that any techniques for improving the bridge sampling in general, such as the warping transformation methods of this article, are immediately applicable to these individual cases.

2. WARP BRIDGE SAMPLING

2.1 Bridge Sampling After Transformations

A cursory examination of (1.1) and (1.6) reveals the intuitive finding that the precision of the bridge sampling estimates depends on the overlap of p_1 and p_2. Specifically, Meng and Wong (1996) found that the Hellinger distance

$$\begin{aligned} H(p_1,p_2) &= \left[\int (\sqrt{p_1(w)} - \sqrt{p_2(w)})^2 \mu(dw)\right]^{\frac{1}{2}} \\ &= \left[2(1 - \int \sqrt{p_1(w)p_2(w)}\mu(dw))\right]^{\frac{1}{2}}, \end{aligned} \tag{2.1}$$

governs the variance of both $\hat{\lambda}_O$ and geometric bridge sampling estimator $\hat{\lambda}_G$, because, under the independence assumption, asymptotically we have

$$\begin{aligned} \frac{1}{s_1 s_2 n}\left[\frac{2\sqrt{s_1 s_2}}{\int \sqrt{p_1(w)p_2(w)}\mu(dw)} - 1\right] &\leq \mathrm{V}(\hat{\lambda}_O) \leq \mathrm{V}(\hat{\lambda}_G) \\ &\leq \frac{1}{s_1 s_2 n}\left[\frac{1}{(\int \sqrt{p_1(w)p_2(w)}\mu(dw))^2} - 1\right]. \end{aligned} \tag{2.2}$$

In the extreme case when $H(p_1,p_2)$ reaches its maximum value $\sqrt{2}$, that is, when $\mu(\Omega_1 \cap \Omega_2) = 0$, no α can satisfy (1.1). There are important applications in statistics and other fields where p_1 and p_2 are completely separated and thus (1.2) is not applicable or where the overlap is so small that the resulting bridge sampling estimate with a single bridge is too variable to be useful (e.g., Voter 1985; Servidea 2002). One way of addressing this problem is to use multiple bridges to link p_1 and p_2; using infinitely many bridges yields what Gelman and Meng (1998) called "path sampling" (see also Bennett 1976 and Neal 1993, Section 3.2). While such methods are generally quite efficient, they require draws other than those from p_1 and p_2.

A different method, which *uses only the available draws* from p_1 and p_2, was proposed by Voter (1985) in the context of estimating free-energy difference for systems whose ensembles (i.e., distributions) are highly separated. Voter's method is extremely simple and intuitive—if the two densities are far apart, *move* them together; here "move" means to apply a simple location shift to one of the distributions. Specifically, Voter (1985) proposed to use the following identity to construct estimators of r:

$$r = \frac{E_2[q_1(w + D)\alpha(w)]}{E_1[q_2(w - D)\alpha(w - D)]}, \tag{2.3}$$

where D is a constant (vector), called "displacement vector" by Voter (1985). Identity (2.3) can be obtained by first transforming w_1 to $w_1^{(D)} = w_1 - D$, where $w_1 \sim p_1$, then applying (1.2) with $q_1^{(D)}(w) = q_1(w + D)$ in place of $q_1(w)$, and finally transforming back to q_1, that is, the E_1 in the denominator of (2.3) is still with respect to the original p_1. Hence, no new draws are needed for implementing Voter's estimator, once D is chosen. Voter (1985) found that using $D = m_1 - m_2$ worked well in his applications, where m_i is the mode of q_i $(i = 1, 2)$.

By viewing Voter's location shift method as a special case of a random variable transformation, we can easily generalize the method further by applying more general transformations, aiming to further reduce the Hellinger distance between the two densities. The more general transformations will generally warp the original geometric shapes of the underlying densities but will not alter the normalizing constants we wish to compute. There are two general types of transformations, *deterministic* transformations, discussed in the following, and *stochastic* transformations, discussed in Section 2.4.

Suppose Ω_1 and Ω_2 are two (not necessarily overlapping) subsets of R^d, and T_1 and T_2 are two one-to-one (and thus deterministic) transformations on R^d that aim to warp p_1 and p_2 into similar shapes. For $w_i \sim p_i$, let $w_i^{(T_i)} = T_i(w_i)$, which has density/probability function

$$p_i^{(T_i)}\left(w_i^{(T_i)}\right) = \frac{q_i\left(T_i^{-1}\left(w_i^{(T_i)}\right)\right)|J_i\left(w_i^{(T_i)}\right)|}{c_i} \equiv \frac{q_i^{(T_i)}\left(w_i^{(T_i)}\right)}{c_i}, \qquad i = 1, 2, \quad (2.4)$$

where T_i^{-1} is the inverse transformation of T_i and $J_i(w)$ is its Jacobian, which is 1 if μ is a counting measure. Because the Jacobian $J_i(w_i^{(T_i)})$ is known to us (and we assume it is easy to compute), we have constructed a new known unnormalized density $q_i^{(T_i)}$ that has the same normalizing constant, c_i, as the original q_i. Since (1.2) holds with $q_i = q_i^{(T_i)}$ and $\{w_{ij}^{(T_i)} = T_i(w_{ij}), j = 1, \ldots, n_i\}$ are draws from $p_i^{(T_i)}$ $(i = 1, 2)$, we can implement (1.3) without making any new draws. Namely, we can compute

$$\hat{r}_\alpha^{(T_1,T_2)} = \frac{\frac{1}{n_2}\sum_{j=1}^{n_2} q_1\left(T_1^{-1}(T_2(w_{2j}))\right) J_1\left(T_2(w_{2j})\right) \alpha\left(T_2(w_{2j})\right)}{\frac{1}{n_1}\sum_{j=1}^{n_1} q_2\left(T_2^{-1}(T_1(w_{1j}))\right) J_2\left(T_1(w_{1j})\right) \alpha\left(T_1(w_{1j})\right)}, \qquad (2.5)$$

which makes it clear how $\hat{r}_\alpha^{(T_1,T_2)}$ depends on T_i $(i = 1, 2)$ and the original draws.

Since the estimate (2.5) is straightforward to implement once $T_i (i = 1, 2)$ are chosen, the key issue centers on the choice of the warp transformations $T_i (i = 1, 2)$. The discussions given in Section 2.1 make it clear that our goal is to warp the two densities into similar shapes. While in principle we may even find T_1 and T_2 such that the warped densities are identical, in practice this is typically neither feasible nor desirable. This is because our goal is to improve the efficiency of the original bridge sampling at a computational cost that is substantially below the level that would offset the gain in efficiency. In Monte Carlo simulation there is almost always a trade-off between computational and statistical efficiency. Indeed, Voter's (1985) original location shift method has almost no additional cost, but can produce very substantial gains. The next section proposes a more general class of deterministic transformations that typically maintain this important property of Voter's method.

2.2 Matching Center and Spread: Warp-II Transformations

Given the success of the location shift approach, one naturally wants to consider the next order transformation, namely, scaling and rotation. In fact, we shall empirically demonstrate in Sections 3 and 4 that such second-order warping can lead to dramatic gain in efficiency over the first-order warping (i.e., Voter's method), without unduly increasing the computational load. For brevity, we will call any estimator from the second-order warping a

Warp-II estimator; accordingly, with the obvious abuse of the meaning of "warp," we label Voter's estimator based on (2.3) Warp-I estimator and the unwarped estimator (1.3) Warp-0 estimator. The third-order warping, Warp-III, will be presented in Section 2.4.

Specifically, a Warp-II estimator is obtained by using $\mathcal{T}_i(w_i) = \mathcal{S}_i^{-1}(w_i - \mu_i), i = 1, 2$, in (2.5), where μ_i and $\mathcal{S}_i$ are, respectively, some measure of the "center" and "spread" of $p_i, i = 1, 2$. For given μ_i and $\mathcal{S}_i$ $(i = 1, 2)$, the iterative estimator corresponding to (1.5) is obtained by iterating (1.5) with l_{ij} replaced by

$$\tilde{l}_{1j} = \frac{q_1(w_{1j})}{q_2\left(\mathcal{S}_2\mathcal{S}_1^{-1}(w_{1j} - \mu_1) + \mu_2\right)} J_{12}$$

$$\text{and} \quad \tilde{l}_{2j} = \frac{q_1\left(\mathcal{S}_1\mathcal{S}_2^{-1}(w_{2j} - \mu_2) + \mu_1\right)}{q_2(w_{2j})} J_{12}, \quad (2.6)$$

where $J_{12} = |\mathcal{S}_1|/|\mathcal{S}_2|$ can be set to 1 in the iterations, as long as we multiply the final limit by the correct J_{12}. A computationally convenient way of implementing a warping method, especially when several ratios involved, is to first transform all the draws and then use the standard bridge sampling estimators with the warped unnormalized densities defined in (2.4).

The warping parameters μ_i and $\mathcal{S}_i$ $(i = 1, 2)$ can be calculated from the known unnormalized density q_i or estimated from the available draws. For example, since the calculation of a mode and the corresponding curvature (i.e., the observed Fisher information when w is viewed as a parameter) does not require the knowledge of the normalizing constant, we can use them for μ_i and $\mathcal{S}_i$ $(i = 1, 2)$. Such methods, which directly use the knowledge of the unnormalized densities, are effective only when the required calculations are easy to perform. A more practical approach in general is to use the sample mean and sample variance-covariance of $\{w_{ij}, j = 1, \ldots, n_i\}$, respectively, for μ_i and $\mathcal{S}_i, i = 1, 2$. If there are more sophisticated estimates of μ_i and $\mathcal{S}_i$ available, they can and should be used [see, e.g., DiCiccio et al. (1997) for better estimates of $\mathcal{S}$ with high dimension], especially when the underlying distributions do not possess any moments. For obvious reasons, using draw-dependent warping parameters will generally lead to a less, sometimes much less, efficient estimate of r compared to using warp parameters that are directly calculated from the unnormalized densities. However, compared to Warp-0 estimators, Warp-II (and Warp-I) estimators are still superior even with draw-dependent warping parameters. We emphasize that precise estimates of the "center" and "spread" of p_i $(i = 1, 2)$ are not necessary in order to have substantial gains over Warp-0 estimators. This is precisely because bridge sampling is more effective than common importance sampling estimators [e.g., estimators based on (1.7)] when the two densities are not very close to each other, as demonstrated in the literature (e.g., Meng and Wong 1996; DiCiccio et al. 1997) and in Sections 3 and 4.

However, the above methods have their limitations. For example, it is not difficult to construct two bivariate distributions (e.g., two mixtures of bivariate normal distributions) that essentially have no overlap yet with the same mean zero and covariance matrix I_2. For such a pair of distributions, the mean-variance matching obviously will not improve the efficiency, and a more carefully designed rotation may be needed; for example, by matching a direction of major mass under both distribution. Indeed, we can consider further transformations to equalize the skewness in both distributions, as detailed in the next section.

2.3 ELIMINATING SKEWNESS: WARP-III TRANSFORMATIONS VIA MIXTURE

Unlike the Voter's (1985) centering transformation, a Warp-II transformation generally alters the physical shape of the underlying distribution, except for special cases (e.g., when the original covariance matrix is proportional to the identity matrix). To push this shape-warping idea further, consider an extreme case where q_1 is symmetric about $w = 0$ but q_2 is highly skewed and in fact has support only on $w \geq 0$. This suggests that we should use a "symmetrizing" transformation to extend q_2 onto the entire real line. This can be done easily by letting $\tilde{q}_2(w) = q_2(|w|)/2, w \in R^1$, which clearly has the same normalizing constant as q_2. Applying bridge sampling to $\{q_1, \tilde{q}_2\}$ will substantially improve the Monte Carlo efficiency because p_1 has significantly more overlap with $\tilde{p}_2$ than with p_2—it is easy to show that $\int \sqrt{p_1(w)\tilde{p}_2(w)}dw = \sqrt{2}\int \sqrt{p_1(w)p_2(w)}dw$.

This symmetrizing transformation is a case of Warp-III transformation because it matches the skewness of the two distributions. It can easily be extended to multidimensional w, and it should be used in conjunction with a Warp-II transformation to increase the overlap. In our limited exploration, we have found the following general strategy quite effective, as in the application of Section 4. After identifying the center μ and spread $\mathcal{S}$ for an unnormalized $q(w)$, as in Section 2.3, we can construct the following mixture

$$\tilde{q}(w|\mu, \mathcal{S}) = \frac{|\mathcal{S}|}{2}\left[q(\mu - \mathcal{S}w) + q(\mu + \mathcal{S}w)\right], \quad w \in R^d. \tag{2.7}$$

Clearly, $\tilde{q}$ centers at $w = 0$ with standard spread I_d, and has the same normalizing constant as q. More importantly, being a point reflection, $\tilde{q}$ is symmetric about zero in any univariate projection $\ell^{\top}w$. This can also be seen by observing that the random variable corresponding to $\tilde{q}(w|\mu, \mathcal{S})$ can be written as $\tilde{w} = b\mathcal{S}^{-1}(w - \mu)$, where b is a Bernoulli variable on $\{1, -1\}$ with equal probability and it is independent of $w \sim q(w)$. That is, the symmetrization warping in (2.7) is a case of *stochastic* transformation. It follows then that for any $\ell \in R^d$, $\ell^{\top}\tilde{w} = b[\ell^{\top}\mathcal{S}^{-1}(w - \mu)]$, which is symmetric about zero because b is and it is independent of w.

Implementing (2.7) is trivial because we can regard $\{\tilde{w}_{ij} = \mathcal{S}_j^{-1}(w_{ij} - \mu_j), i = 1, \ldots, n_j\}$ and $\{-\tilde{w}_{ij}, i = 1, \ldots, n_j\}$ as $2n_j$ (stratified) draws from $\tilde{q}_j(w|\mu, \mathcal{S}), j = 1, 2$. In fact, for $\alpha(\tilde{w})$ depending on $\tilde{w}$ only through $\tilde{q}_j(\tilde{w}|\mu, \mathcal{S}), j = 1, 2$, as with $\hat{r}_O$ and $\hat{r}_G$, we do not need to consider $\{-\tilde{w}_{ij}, i = 1, \ldots, n_j\}$ as draws because of the symmetry of $\tilde{q}_j(w|\mu, \mathcal{S}), j = 1, 2$. Indeed, because of this symmetry, in computing the the weight s_j of (1.4), the sample size from $\tilde{q}_j(w|\mu, \mathcal{S})$ should be still counted as n_j, not $2n_j$.

The mixture (2.7) is the simplest symmetrization warping via mixing over the point-reflection group $\mathcal{G}_2 = \{I_d, -I_d\}$. We can, of course, consider mixing over more sophisticated groups, such as the 2^d-element group $\mathcal{G}_{2^d}$, consisting of all reflections with respect to the d axis in the Cartesian coordinates as well as all their (distinct) compositions. This mixing also leads to symmetry for any linear projection. Furthermore, the Hessian matrix of the log of the mixture density at $w = 0$ is the diagonal matrix from the Hessian matrix of the log of the original density at $w = 0$ (assuming $w = 0$ is a mode for the original density). This is because for $q_{\text{mix}}(w) = \sum_{R \in \mathcal{G}} q(Rw)$, where $\mathcal{G}$ is a finite orthogonal matrix group,

$$\left.\frac{\partial^2 \log q_{\text{mix}}(w)}{\partial w \partial w^\top}\right|_{w=0} = \frac{1}{|\mathcal{G}|} \sum_{R \in \mathcal{G}} R \left[\left.\frac{\partial^2 \log q(w)}{\partial w \partial w^\top}\right|_{w=0} \right] R^\top, \tag{2.8}$$

if $w = 0$ is a stationary point of $q(w)$. This implies that when mixing over the $\mathcal{G}_{2^d}$ group ($d > 1$), the mixture renders independence among all components of w locally around $w = 0$, a property that is not shared by mixing over the $\mathcal{G}_2$, which leaves the Hessian matrix unchanged. However, this does not necessarily imply that the former mixture will produce better approximation to $N(0, I_d)$ in terms of Hellinger distance. Furthermore, mixing over the $\mathcal{G}_{2^d}$ group is not practical when d is large, while mixing over $\mathcal{G}_2$ is essentially trivial regardless of the value of d. Indeed, if it is not for practicality, we could even consider mixing over all possible orthogonal transformations (with respect to the corresponding Haar measure), which would provide a mixture that is invariant under orthogonal transformation, a well-known property of $N(0, I_d)$.

2.4 OPTIMIZING WARP TRANSFORMATIONS

To further increase the overlap of the two underlying distributions, we can consider optimizing any of the aforementioned warp transformations over the warping parameters, mostly the center μ and spread $\mathcal{S}$, by minimizing the corresponding Hellinger distance. Since this involves additional computation, it is wise to optimize over the more powerful Warp-III transformations, such as the one given in (2.7), than spending similar effort for Warp-I or for Warp-II transformations.

As a specific but important example, consider optimizing over the choice of μ and/or $\mathcal{S}$ in the $\tilde{q}(w|\mu, \mathcal{S})$ of (2.7) by minimizing the Hellinger distance between $\tilde{q}(w|\mu, \mathcal{S})$ and the standard normal $p_0 := N(0, I_d)$—recall our aim of using (2.7) is to match it with $N(0, I_d)$ in terms of the first three moments. An important observation is that this minimization does not require the knowledge of the unknown normalizing constant of $\tilde{q}(w|\mu, \mathcal{S})$, even though the Hellinger distance itself does depend on it. This is because, as can be seen clearly from (2.1), minimizing the Hellinger distance between $\tilde{p}(w|\mu, \mathcal{S})$ and p_0 is the same as maximizing the *overlap measure* (known as Bhattacharrya's measure of affinity)

$$\mathcal{O}(\tilde{p}, p_0) = \int \sqrt{\tilde{p} p_0} \mu(dw) = \sqrt{\frac{c_0}{\tilde{c}}} E_0 \left[\sqrt{\frac{\tilde{q}}{q_0}} \right], \tag{2.9}$$

where the expectation is with respect to p_0, and c_0 can be chosen for convenience (e.g., $c_0 = 1$). Since $c_0/\tilde{c}$ does not depend on $(\mu, \mathcal{S})$, we only need to maximize the *unnormalized* overlap measure $o(\mu, \mathcal{S}) \equiv E_0[\sqrt{\tilde{q}/q_0}]$. This optimization can be done by numerically maximizing the exact $o(\mu, \mathcal{S})$ in simple problems.

For example, for the univariate example of Section 2.6, $p_0 := N(0, 1)$, and $p(w) \propto we^{-w/2} 1_{(w \geq 0)}$, namely, the χ^2_4 distribution. A little calculation shows that the unnormalized overlap measure $o(\mu, \mathcal{S})$ between $N(0,1)$ and $\tilde{q}(w|\mu,\mathcal{S}) \propto \mathcal{S}[p(\mu - \mathcal{S}w) + p(\mu + \mathcal{S}w)]$ is given by

$$o(\mu,\mathcal{S}) \propto e^{-\frac{\mu}{4}}\mathcal{S}^{-1/2}\left[\int_0^{\mu} e^{-\frac{y^2}{4\mathcal{S}^2}}\sqrt{(\mu-y)e^{\frac{y}{2}}+(\mu+y)e^{-\frac{y}{2}}}\,dy + \int_{\mu}^{\infty} e^{-\frac{y^2}{4\mathcal{S}^2}}\sqrt{(\mu+y)e^{-\frac{y}{2}}}\,dy\right]. \quad (2.10)$$

While it is not feasible to analytically maximize $o(\mu, \mathcal{S})$, we can use many numerical routines, including taking a grid of $\{\mu, \mathcal{S}\}$ and using numerical integrations, as we did for the illustration in Section 2.6.

For most real applications, especially the high dimensional ones, numerical evaluation of $o(\mu, \mathcal{S})$ is out of question. However, it can be estimated by the sample average

$$\hat{o}_m(\mu,\mathcal{S}) = \frac{1}{m}\left[\sum_{i=1}^{m}\sqrt{\frac{\tilde{q}(w_i|\mu,\mathcal{S})}{q_0(w_i)}}\right], \quad (2.11)$$

where $\{w_1, \ldots, w_m\}$ are iid draws from $p_0 := N(0, I_d)$. Because our goal is to match $\tilde{q}$ and $N(0, I_d)$, with a reasonable initial search area for optimal $(\mu, \mathcal{S})$, $\hat{o}_m(\mu, \mathcal{S})$ would be accurate enough for empirically optimizing $o(\mu, \mathcal{S})$ via numerical procedures such as quasi Newton–Raphson. We emphasize again that it is not necessary to find the exact optimal μ and $\mathcal{S}$ values in order to have dramatic gain; see Section 4.4 for an illustration.

2.5 A GRAPHICAL APPETIZER

Here we present a graphical illustration of Warp-I to Warp-III transformations as well as their impacts on the root mean square error (RMSE) of three estimators: (1) the importance-sampling estimator (1.7), $\hat{\lambda}_S = \log \hat{r}_S$; (2) the bridge-sampling estimator with geometric bridge, $\hat{\lambda}_G = \log \hat{r}_G$, and (3) the (iterative) optimal bridge-sampling estimator, $\hat{\lambda}_O = \log \hat{r}_O$. Consider $p_1 = N(0,1)$ and $p_2 = \chi^2_4$, as displayed in Panel 1 of Figure 1. Panels 2–9 display, respectively, the positions of the two distributions after eight different warp transformations to χ^2_4, in order of (nearly) decreasing Hellinger distance, denoted by H in each panel (the RMSE value is from $\hat{\lambda}_O$). The true value of the estimand λ is zero. Panel 0 plots the RMSE of the three estimators as a function of the log Hellinger distance, where the log is used for better visualization. The RMSE is obtained via simulation with 1,000 replications, each of which makes 250 independent draws, respectively, from $N(0,1)$ and χ^2_4 for bridge sampling estimators and 500 independent draws from χ^2_4 for the importance sampling estimator. This explains the "off-the-chart" value of the RMSE for $\hat{\lambda}_S$ under Panel 1, because χ^2_4 is a terrible choice as a trial density when the target density is $N(0,1)$.

The second row of Figure 1 corresponds to Warp-I transformations with mode matching (Panel 2) and mean matching (Panel 3). Since the mode and mean of χ^2_4 are, respectively, 2 and 4, this example illustrates the possibility of having very different centering with nearly identical bridge-sampling efficiency—note that the two Hellinger distances are almost the same. However, the two centerings will have very different impacts on the importance

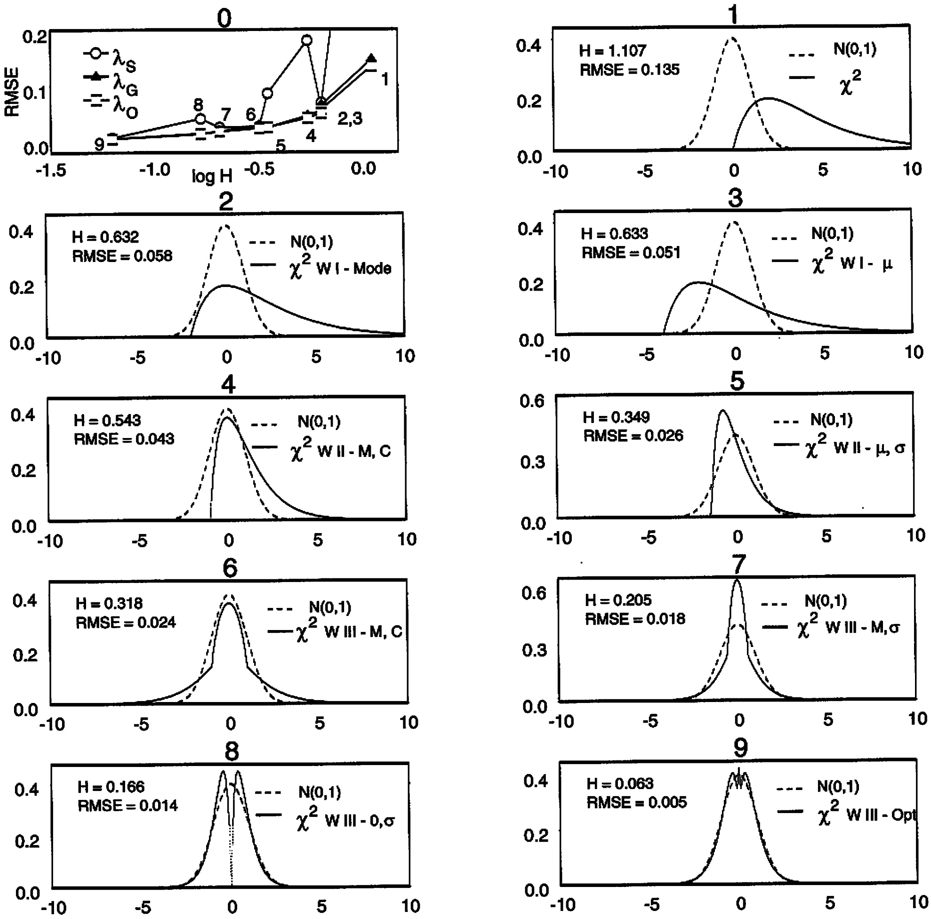

Figure 1. Warping the χ^2_4 Distribution to More Closely Match the Normal Distribution.

sampling estimator $\hat{\lambda}_S$ when we use $N(0, 1)$ as the trial density—the mean-matching one will have a lot larger RMSE because $N(0, 1)$ has a tail that is much too light in the negative area of the mean-shifted χ^2_4. This well-known problem of importance sampling is also seen by comparing the second row with the third row, which corresponds to Warp-II transformations with mode-curvature matching (Panel 4) and with mean-variance matching (Panel 5). Although the Hellinger distance is decreased quite a bit from Warp-I transformations to Warp-II transformations, and thus the RMSE of either $\hat{\lambda}_G$ and $\hat{\lambda}_O$ is also decreased, the RMSE of $\hat{\lambda}_S$ is greatly increased, especially for Panel 4, because the curvature (and variance) matching has made the left-tail of the transformed χ^2_4 much shorter than that of $N(0, 1)$. This illustrates the fact that Hellinger distance is not the right metric for controlling the RMSE of an importance sampling estimator. It is well known that the variance of an importance sampling estimator is the χ^2-distance between the target and trial density.

The fourth row displays Warp-III transformation reflecting about the mode followed by mode-curvature matching (Panel 6) and followed by mode-variance matching (Panel 7)—

note that after reflection (i.e., symmetrizing) the mode is the same as mean. These two symmetrizing transformations not only help to further reduce the Hellinger distance, but also the aforementioned χ^2-distance and thus the RMSEs for all three estimators are reduced. Instead of reflecting around the mode of χ_4^2, we can also reflect around zero and then match the variance. This further reduces the Hellinger distance, as displayed in Panel 8. Note that because of the "deep dip" in the middle of the transformed χ_4^2, the χ^2-distance is actually increased by this transformation, which leads to the increase in the RMSE of $\hat{\lambda}_S$, as seen in Panel 0. The last panel, Panel 9, shows the result from the optimal Warp-III transformation as detailed in Section 2.5, where the optimal reflection point was found to be approximately $\mu = 0.5$, which is closer to the origin than to the mode of χ_4^2, and the corresponding optimal scaling was approximately $\mathcal{S} = 4.4$. The most interesting finding here is that the best match to the unimodal $N(0, 1)$ among the mixture class (2.7) is a tri-mode density, where the middle mode is the result of "merging" the two modes in Panel 8, but not all the way as in Panel 7.

Examining Panel 0 of Figure 1, we see that $\hat{\lambda}_O$ and $\hat{\lambda}_G$ outperform $\hat{\lambda}_S$ for all transformations. Moving from Warp-0 through Warp-III estimators causes the RMSE of $\hat{\lambda}_O$ and $\hat{\lambda}_G$ to be roughly halved for each successive transformation. The RMSE resulting from the optimal Warp-III transformation is about 4% of the original RMSE under no transformation, and about 10% of the one from Voter's centering transformation. We will see similar dramatic gains in efficiency in the next two real-data multivariate applications.

3. EMPIRICAL STUDY I: THE FULL INFORMATION ITEM FACTOR MODEL

3.1 MODEL AND DATA DESCRIPTION

Our first empirical study focuses on the problem of computing the log-likelihood ratio from Bock and Aitken's (1981) full information item factor (FIIF) model. A detailed derivation of the model was given by Meng and Schilling (1996), so here we provide only a brief description. Suppose there are J test items given to n subjects, and we let $u_{ij} = 1$ if the ith subject gives the correct answer to the jth item and 0 otherwise. The FIIF model hypothesizes that given the ith subject's d latent (ability) factors $\mathbf{z}_i = (z_{i1}, \ldots, z_{id})^\top$, the probability of the ith subject's response pattern $\mathbf{u}_i = (u_{i1}, \ldots, u_{iJ})^\top$ is given by

$$\Pr(\mathbf{u}_i \mid \mathbf{z}_i, \theta) = \prod_{j=1}^{J} \left[\Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j)\right]^{u_{ij}} \left[1 - \Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j)\right]^{1-u_{ij}}. \tag{3.1}$$

Here Φ is the cdf of $N(0, 1)$, b_j is the *item intercept* for the jth item, and $\mathbf{a}_j = (a_{1j}, \ldots, a_{dj})^\top$ with a_{mj} being the *item slope* for factor m. Let $\mathbf{b} = (b_1, \ldots, b_J)^\top$, and let $\mathbf{A}$ be a $d \times J$ matrix whose jth column is $\mathbf{a}_j, j = 1, \ldots, J$. We then call $\theta \equiv \{\mathbf{A}, \mathbf{b}\}$ the set of item parameters. Taking the product of (3.1) over i and making the assumption of independence between subjects, we obtain the likelihood function of θ given the observed score matrix

$\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)^\top$ and the unobserved latent factors $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$,

$$L(\theta \mid \mathbf{Z}, \mathbf{U}) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left[\Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j)\right]^{u_{ij}} \left[1 - \Phi(\mathbf{z}_i^\top \mathbf{a}_j + b_j)\right]^{1-u_{ij}}. \qquad (3.2)$$

The complication in computing the likelihood from a FIIF model arises from the fact that $\mathbf{Z}$ is unobserved, and thus needs to be integrated out. As a part of the model assumption, the FIIF model assumes $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are iid $N_d(0, \mathbf{I})$, and thus the actual likelihood from the FIIF model is

$$L(\theta \mid \mathbf{U}) = \prod_{i=1}^{n_0} \left\{ E_{\mathbf{z}} \left[\prod_{j=1}^{J} \left[\Phi(\mathbf{z}^\top \mathbf{a}_j + b_j)\right]^{u_{ij}} \left[1 - \Phi(\mathbf{z}^\top \mathbf{a}_j + b_j)\right]^{1-u_{ij}} \right] \right\}^{s_i}, \qquad (3.3)$$

where $E_{\mathbf{z}}$ is with respect to $\mathbf{z} \sim N_d(0, \mathbf{I})$, n_0 ($\leq \min\{n, 2^J\}$) is the number of the distinct response patterns, and s_i is the number of the subjects who share $\mathbf{u}_i$.

Meng and Schilling (1996) described how to implement Monte Carlo EM (MCEM) for finding the MLE of θ treating $\mathbf{Z}$ as missing data, using the Gibbs sampler to carry out the E-step. The draws from $p(\mathbf{Z}|\mathbf{U}, \theta) = \prod_{i=1}^{n_0} p(\mathbf{z}_i|\mathbf{u}_i, \theta)$ are thus available for various values of θ, particularly at the MLE. Therefore, because $p(\mathbf{z}_i|\mathbf{u}_i, \theta) = p(\mathbf{z}_i, \mathbf{u}_i|\theta)/p(\mathbf{u}_i|\theta)$, we can apply bridge sampling to compute

$$\log \frac{L(\theta_2 \mid \mathbf{U})}{L(\theta_1 \mid \mathbf{U})} = \sum_{i=1}^{n_0} s_i \log \frac{p(\mathbf{u}_i|\theta_1)}{p(\mathbf{u}_i|\theta_2)} \qquad (3.4)$$

by treating $p(\mathbf{u}_i|\theta)$ as the normalizing constant of $p(\mathbf{z}_i|\mathbf{u}_i, \theta)$, with $p(\mathbf{z}_i, \mathbf{u}_i|\theta)$ as the unnormalized density. That is, we compute the desired log-likelihood ratio as a weighted sum of the log ratios of normalizing constants. Such likelihood ratios are very useful for monitoring the convergence of MCEM (Meng and Schilling 1996). The simulations we provide here use a five-factor model. We choose a response pattern and two sets of item parameters to create a "worst case" practical scenario. One set of the item parameters is composed of the maximum likelihood estimates from a real dataset, specifically, 25 selected items from a 100-item spelling test administered to 660 undergraduate psychology students at the University of Kansas in 1987. The other set of item parameters represents a null model that might be tested in practice (i.e., with all intercepts zero, and slopes one or zero chosen for different items). The response pattern chosen was the pattern with the largest Hellinger distance between the two conditional densities corresponding to the two sets of item parameters; the resulting Hellinger distance was $H = 0.818$. The item parameters were also constructed so that, conditional on $\mathbf{u}_i$ and θ, the five factors are mutually independent. This independence allows us to simplify the computation of the normalizing constants into a series of one-dimensional integrations, which can then be computed accurately through the use of numerical integration routines. The aim is to provide an objective gold standard against which our simulation results can be checked. The PQUAD program (Wichura 1989) was used for computing all requisite one-dimensional integrations to a relative accuracy of 10^{-12}.

3.2 EMPIRICAL COMPARISONS

As in Section 2.6, three estimators were compared: $\hat{\lambda}_S$, $\hat{\lambda}_G$, and $\hat{\lambda}_O$. The Gibbs sampler was used to generate 100 draws for each $p_i, i = 1, 2$ corresponding to each set of item parameters. To conduct a (nearly) fair comparison, we used 50 of the draws from each distribution to compute $\hat{\lambda}_O$ and $\hat{\lambda}_G$, and the entire 100 draws from p_2 to compute $\hat{\lambda}_S$. We repeated this process 1,000 times, yielding 1,000 estimates for each method. In addition to the unwarped case, we consider four warping transformations: (1) Warp-I using modes, (2) Warp-II using modes and curvatures, (3) Warp-II using sample means and covariances, and (4) Warp-II using true means and covariances. The last warping is included for the purpose of comparison, as it should, with reasonable distributional shapes, yield near optimal Warp-II estimators. We leave Warp-III transformations to Section 4, since with the current application Warp-II transformations have already produced very accurate estimators for practical purposes.

As Table 1 shows, matching even just the location (i.e., mode) leads to a great increase in efficiency for all the estimators. For example, the MSE of $\hat{\lambda}_O$ decreases about 70% when the modes of both distributions were matched. Matching both the modes and their curvatures leads to the MSE that is just 1.3% of the MSE with only the mode matched. The estimator using the sample mean and variance is, as expected, less accurate than one that uses the exact modes and their curvatures, but it is still far superior to the original one (about 5% MSE) as well as the one that only matches the modes (about 17% MSE). The estimate obtained by matching the true mean/variance has a MSE about 2% of that using sample mean/variances. This severe loss is expected since we are using only 50 draws and thus also face the serious issue of "losing degrees of freedom." Indeed, with very small simulation sizes, using sample means and covariances could lead to serious bias—an issue that fortunately need not be of much concern as the sample sizes of the simulated data are typically under the direct control of the investigator. It is noteworthy that, for the current problem, using sample estimates does not induce bias for $\hat{\lambda}_O$ or $\hat{\lambda}_G$, but produces relatively large bias for $\hat{\lambda}_S$.

The results in Table 1 also illustrate that the performance of $\hat{\lambda}_G$ is nearly the same as that of $\hat{\lambda}_O$, particularly when the two distributions are closely matched. In fact, when two distributions are closely matched, any reasonable estimator (e.g., $\hat{\lambda}_S$) should work well; this is evident in the last comparisons (i.e., when matching with the true mean and variances). The bridge sampling estimators are designed to handle cases where accurate matching is not feasible. It is clear from Table 1 that the relative gains of bridge sampling estimators over the importance sampling estimator (i.e., $\hat{\lambda}_S$) generally increase with the distance between the two underlying densities. For the untransformed cases, the MSE of $\hat{\lambda}_O$ is only about 16% of the MSE of $\hat{\lambda}_S$, despite the fact that both require essentially the same amount of computation.

Table 1. Effects of Warping for Three Estimators (true $\lambda \equiv \log r = 3.6226$)

	$\hat{\lambda}_S$	$\hat{\lambda}_O$	$\hat{\lambda}_G$
	Warp-0		
Mean	3.3909	3.6348	3.6238
Var	0.1935	0.0390	0.0471
MSE	0.2472	0.0391	0.0471
	Warp-I with Mode		
Mean	3.5779	3.6287	3.6256
Var	0.0448	0.0112	0.0124
MSE	0.0468	0.0112	0.0124
	Warp-II with Mode and Curvature		
Mean	3.6221	3.6225	3.6225
Var	0.000148	0.000147	0.000150
MSE	0.000148	0.000147	0.000150
	Warp-II with Sample Mean and Covariance		
Mean	3.7213	3.6245	3.6238
Var	0.00402	0.00190	0.00204
MSE	0.01377	0.00190	0.00204
	Warp-II with True Mean and Covariance		
Mean	3.6221	3.6225	3.6225
Var	0.000041	0.000041	0.000042
MSE	0.000041	0.000041	0.000042

4. EMPIRICAL STUDY II: A NONLINEAR MIXED-EFFECT MODEL

4.1 Model and Data Description

Computing likelihood ratios for nonlinear mixed-effects model is typically a difficult problem due to the analytically intractable integrations over the random effects. The illustration we provide here uses data from the Fort Bragg evaluation project (Bickman et al. 1995) where three military bases were chosen to participate in a quasi-experiment. The experimental treatment, an integrated system of children's mental health services, was offered to personnel at Fort Bragg, North Carolina. As experimental control, child outcome data were also collected at two other military bases, Fort Campbell, Kentucky, and Fort Stewart, Georgia, where typical mental health care services were available. The main hypothesis of the study was that patients receiving the experimental treatment would show more improvement, with a smaller relapse effect. Data were collected at intake, 6 months, 12 months, 18 months, 36 months, and 48 months for 510 patients at the demonstration site and 379 patients at the two control sites. However, missing data were common, with many of the

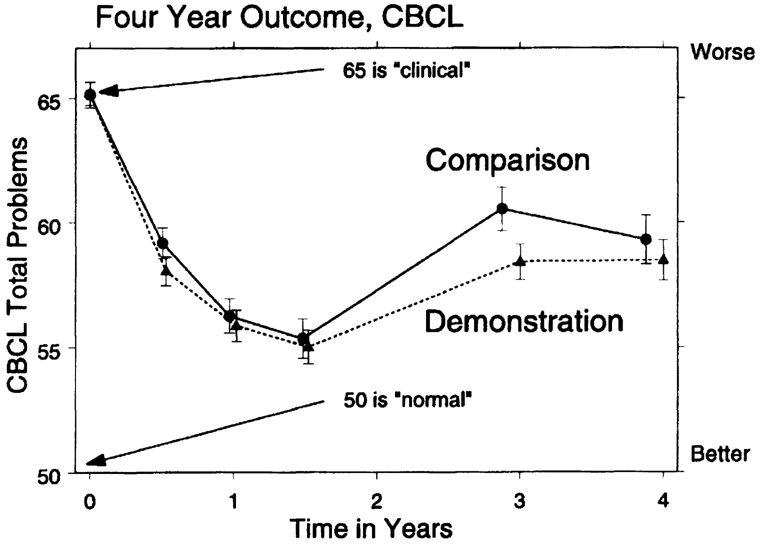

Figure 2. Mean CBCL Scores for Demonstration and Control Sites by Time.

subjects having fewer than the full complement of six observations. The dependent measure is the Child Behavioral Check-List (CBCL), a standard parental report psychiatric rating scale commonly used in child psychiatric research. Mean values for the six time periods for the demonstration site and the comparison site (aggregated over the two control sites) are plotted in Figure 2. The change at both sites appears to reasonably follow a pattern of exponential decline, followed by a decay of the effects of psychiatric treatment at 36 and 48 months. For each child, there was generally a relapse effect after the continuous integrated treatment, which usually lasts from one to two years, was over. More details of this dataset can be found in Schilling (1998).

A realistic modeling of the data for estimating the treatment effect is rather involved, particularly because the relapse onset time is unobserved. For our computational purpose, we adopted the following two-level model:

Level 1 Model:

$$y_{ijk} = \gamma_{0ij} + \gamma_{1ij} I(k \geq 5) + \gamma_{2ij}\left[g_{ij}(t_{ik}, \gamma_{3ij}) - \bar{g}_{ij}(\gamma_{3ij})\right] + \epsilon_{ijk}, \qquad \epsilon_{ijk} \sim N(0, \sigma^2),$$

where $g(t,\gamma) = \gamma^{-1}\exp(\gamma(t-1.5))$, $\bar{g}(\gamma) = \sum_{k=1}^{n_{ij}} g(t_{ik},\gamma)/n_{ij}$, n_{ij} is the number of observations for subject i at site j, and $I(A)$ is the indicator function of set A.

Level 2 Model:

$$\gamma_{ij} \equiv \begin{pmatrix} \gamma_{0ij} \\ \gamma_{1ij} \\ \gamma_{2ij} \\ \gamma_{3ij} \end{pmatrix} \sim \mathrm{N}(\mu_j, \Sigma).$$

Here y_{ijk} is the CBCL measure for the ith child at the jth site ($j = 1, 2$), and k indexes the six time points of data collection. The Level 1 model parameter γ_{ij} represents a *reparameterized* vector parameter describing person-level baseline, relapse effect, treatment effect, and declining rate. The reparameterization was adopted, following a suggestion by Ross (1990) to deal with a practical near nonidentifiability problem due to relatively small exponential declining rates for some children. It also helps to eliminate unintended adverse consequences of the somewhat arbitrary normality assumption in the Level 2 model, since our main goal is to compare μ_1 and μ_2. Using a Bayesian analogy, the ellipsoidal shape of the prior (i.e., Level 2 model) would be in serious conflict of the severely banana shaped likelihood (i.e., the Level 1 model) with the original parameterization, and thus the normality can no longer be treated as a convenient "non-informative" assumption. See Box and Tiao (1973, chap. 1) for the closely related idea of "data-translated likelihood" in the content of constructing noninformative priors. The more harmonized Level 1 and Level 2 modeling also makes it easier to construct more reasonably behaved Metropolis algorithms, which was important for the simulation study we report here.

As a common computational task that arises in such problems, we consider the problem of computing the log-likelihood ratio for testing $\mu_1 = \mu_2$, namely, $\log f(Y|\hat{\theta}) - \log f(Y|\hat{\theta}_0)$, where $\hat{\theta}$ and $\hat{\theta}_0$ are, respectively, the MLE and the constrained MLE under $\mu_1 = \mu_2$ for $\theta = \{\mu_1, \mu_2, \Sigma, \sigma^2\}$. Similar to the case for the FIIF model, that is, (3.3) and (3.4), this log-likelihood ratio can be expressed as a sum of 889 log ratios of normalizing constants for $p(\gamma_{ij}|Y_{ij}, \theta) = p(\gamma_{ij}, Y_{ij}|\theta)/p(Y_{ij}|\theta)$, where $Y_{ij} = \{Y_{ijk}, k = 1, \ldots n_{ij}.\}$. We used a Metropolis algorithm to simulate from $p(\gamma_{ij}|Y_{ij}, \theta)$ with $N(\gamma_{ij}^{(t-1)}, I^{-1}(\tilde{\gamma}_{ij}))$ as the proposal distribution at the tth iteration, where $\gamma_{ij}^{(t-1)}$ is the output from the previous iteration and $I(\tilde{\gamma}_{ij})$ is the observed Fisher information matrix at the mode $\tilde{\gamma}_{ij}$. This yielded draws that were highly correlated, with a median lag 1 correlation of 0.86 for $\gamma_{0i\cdot}$ through $\gamma_{3i\cdot}$ and an acceptance rate of 0.23. This high correlation allows us to investigate the performance of $\hat{r}_O$ of (1.5) when the assumption of independence, under which $\hat{r}_O$ is asymptotically optimal, is seriously violated. Another key feature of this study is that there are 889 log ratios to be examined and the Hellinger distances between these 889 pairs of distributions, under various warping transformations, cover virtually the entire range of the possible values, $[0, \sqrt{2}]$, thereby providing an ideal setting for examining the performance of the bridge sampling estimators with different orders of warping.

As before, three estimators were compared: $\hat{\lambda}_S$, $\hat{\lambda}_G$, and $\hat{\lambda}_O$. The Metropolis sampler was used to generate 500 draws for each $p_i, i = 1, 2$ corresponding to each set of Level 2 parameters, for each of the 889 sets of the Level 1 distributions. To conduct a (nearly) fair comparison, we used 250 of the draws from each distribution to compute $\hat{\lambda}_O$ and $\hat{\lambda}_G$, and the entire 500 draws from p_2 to compute $\hat{\lambda}_S$. We start our comparisons among various Warp-I and Warp-II estimators in Section 4.2. In Section 4.3, we investigate the performance of Chib and Jeliazkov's (2001) method as a case of the bridge sampling. In Section 4.4, we demonstrate the use of Warp-III transformations to eliminate asymmetry and thus further improve upon the best Warp-II estimator found in Sections 4.1–4.2. For each estimator, the simulation is repeated 1,000 times to yield 1,000 estimates, which are then compared to the exact value of the estimand. The exact values were calculated again using PQUAD with relative accuracy of 10^{-12}; only one-dimensional numerical integrations were needed because given γ_{3i}, the other Level 1 parameters are conditionally linear and can be integrated out analytically.

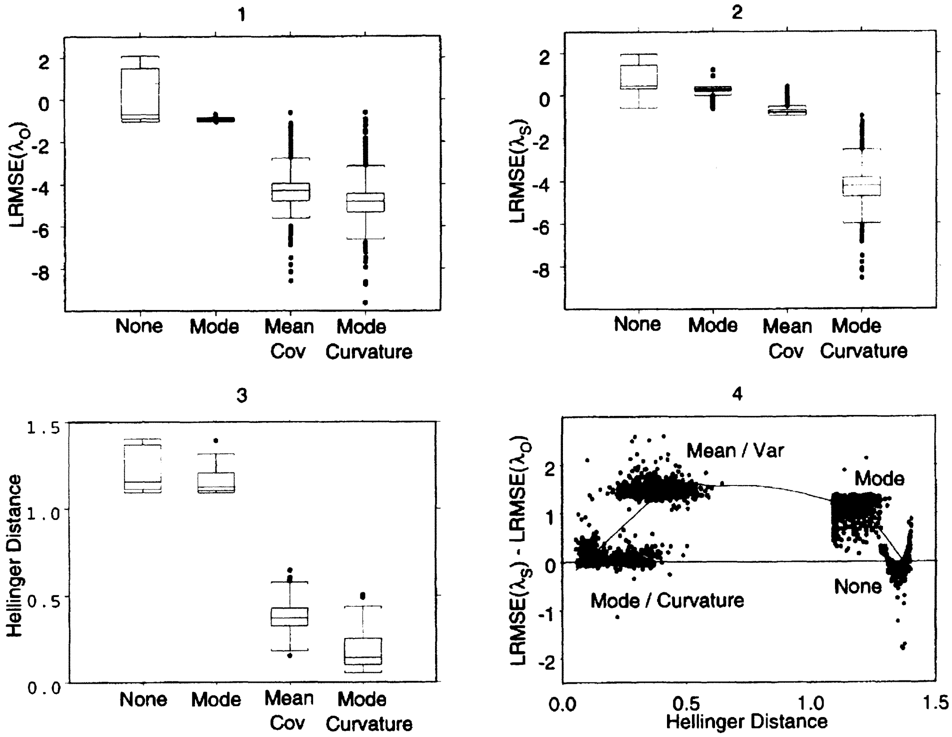

Figure 3. The Impact of Different Warping Transformations on Hellinger Distances and LRMSE.

4.2 Empirical Comparisons for Warp-I and Warp-II Estimators

The first two panels of Figure 3 present box plots of the log of the root mean squared error (LRMSE) of $\hat{\lambda}_O$ and $\hat{\lambda}_S$ for each of the four warping transformations, while the third panel presents box plots of the Hellinger distances, across the 889 subjects. Both $\hat{\lambda}_O$ and $\hat{\lambda}_S$ perform poorly under Warp-0 because of the very large Hellinger distances, as seen in the third panel corresponding to the label "None." Warp-I with mode has a very small effect on the median Hellinger distances, reducing the median from 1.15 to 1.12. But it dramatically reduces the right skewness in the distance distribution: with no warping, 40.6% of the pairs of Level 1 distributions have Hellinger distances greater the 1.3, compared to 0.3% after matching the modes. Consequently, Warp-I with mode matching produces small but noticeable reductions in the median of LRMSE (from 0.45 to 0.27 for $\hat{\lambda}_S$ and from -0.69 to -0.95 for $\hat{\lambda}_O$), but greatly reduces the percentage of large LRMSEs for both $\hat{\lambda}_O$ and $\hat{\lambda}_S$. Applying the second-order warping has a profound effect on the Hellinger distances and on LRMSE. For instance, Warp-II using sample means/covariances and using modes/curvatures reduce 95% of the Hellinger distances below 0.5 and 0.35, respectively. Therefore, the RMSEs were reduced dramatically, as seen from the large reduction on the log scale in the first two panels.

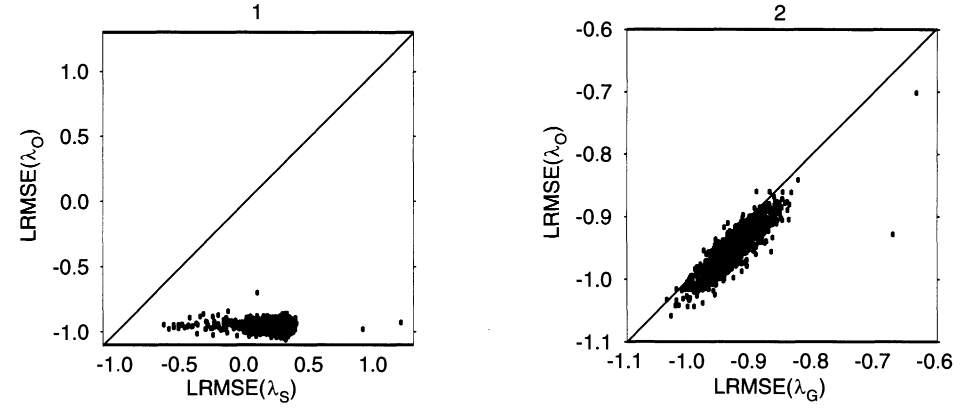

Figure 4. Performance of Warp-I Estimators with Mode Transformation.

The fourth panel of Figure 3 plots LRMSE($\hat{\lambda}_S$)− LRMSE($\hat{\lambda}_O$) for all 889 cases across the four warping transformations as functions of the Hellinger distance. It is clear that $\hat{\lambda}_O$ dominates $\hat{\lambda}_S$ especially with the Warp-II using the easily implemented sample mean/covariance matching, with all points are far above the reference line at zero. The interesting parabola shape (fitted by *lowess*) deserves some comment. Intuitively, the improvement of $\hat{\lambda}_O$ over $\hat{\lambda}_S$ generally increases with the Hellinger distance. However, when the distance is very large, both estimators become dominated by numerical errors because of the large instability in computing ratios. The moral is that one should use neither $\hat{r}_S$ nor $\hat{r}_O$ when the two densities are far apart—the exact comparison between them is not relevant when both are unusable.

To make more detailed comparisons, Figure 4 and Figure 5 display the scatterplots of LRMSE of $\hat{\lambda}_O$ versus that of $\hat{\lambda}_S$ and of $\hat{\lambda}_G$ under Warp-I with mode and Warp-II with sample mean and covariance matching, respectively. In both cases the advantage of $\hat{\lambda}_O$ compared to $\hat{\lambda}_S$ is considerable, with the maximum LRMSE for $\hat{\lambda}_O$ less than the minimum

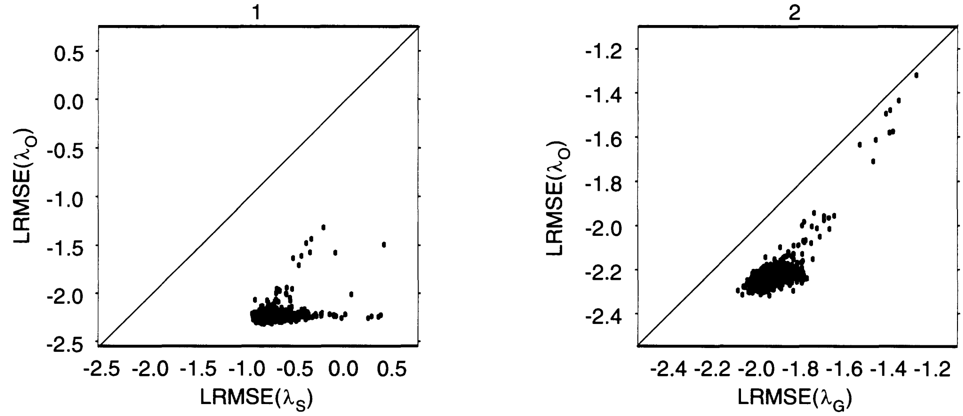

Figure 5. Performance of Warp-II Estimators with Sample Mean/Covariance Transformation.

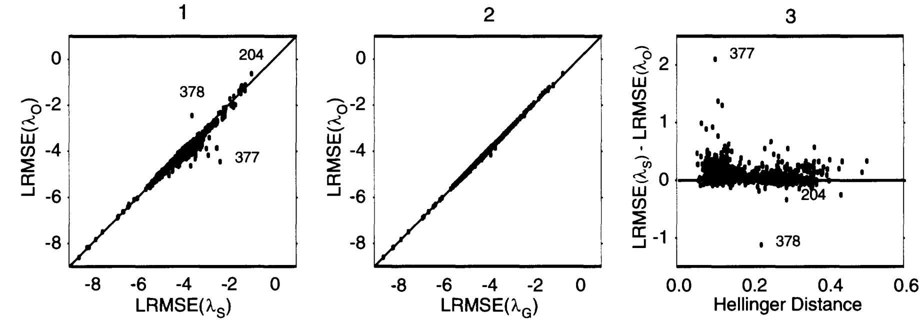

Figure 6. Performance of Warp-II Estimators with Mode/Curvature Transformation.

LRMSE for $\hat{\lambda}_S$. The performance of $\hat{\lambda}_G$ is closer to that of $\hat{\lambda}_O$, more so for Warp-I with mode than Warp-II with sample mean and covariance.

Figure 6 compares the three estimators when the underlying Hellinger distance is small. Thanks to the powerful warping with mode/curvatures nearly all of the 889 pairs have Hellinger distances less than 0.4 with more than half less than 0.2. When the two densities are this close, $\hat{\lambda}_S$ can also work rather well. Nevertheless, the first panel shows a small but noticeable advantage for $\hat{\lambda}_O$ compared to $\hat{\lambda}_S$, in that 76.3 percent of the points fall below the 45-degree line. The second panel shows that $\hat{\lambda}_G$ is essentially the same as $\hat{\lambda}_O$. The third panel plots the difference in the LRMSE between $\hat{\lambda}_S$ and $\hat{\lambda}_O$ as a function of Hellinger distance. It shows that even when the distributions are very close to each other, there are a number of instances where $\hat{\lambda}_O$ substantially outperforms $\hat{\lambda}_S$, most noticeably for subject 377. On the other hand, we note for subjects 204 and 378, $\hat{\lambda}_S$ outperforms $\hat{\lambda}_O$ by a good margin. (Three cases consistently produced large differences in additional confirmatory simulations.) Examining these three cases more closely, the first three panels of Figure 7 compare the box-plots of distributions of the two estimators, obtained from the 1,000 replications, for each of the three subjects, respectively. The panel for subject 377 shows the reason why that the LRMSE of $\hat{\lambda}_S$ is much larger than that of $\hat{\lambda}_O$: the existence of a very large outlier, reinforcing the old concern of instability in the tails of the importance sampling estimator. Even when distributions are closely matched, they may exhibit rather different tail behavior in some extreme tail regions, and thus produce large outliers in the importance weights. In contrast, all the weights used by $\hat{r}_O^{(t+1)}$ are bounded by $\max\{1/s_1, 1/(s_2\hat{r}_O^{(t)})\}$; see (1.5).

However, it is possible for $\hat{\lambda}_S$ to outperform $\hat{\lambda}_O$ when the densities are closely matched because the optimality of $\hat{\lambda}_O$ is only guaranteed with independent draws, and when the number of draws used for $\hat{\lambda}_S$ does not exceed $\max\{n_1, n_2\}$. This is seen in the first two panels of Figure 7, where the distributions of $\hat{\lambda}_O$ have larger spread than those of $\hat{\lambda}_S$. To confirm that such relatively larger variability was mainly caused by the high autocorrelations, for $j = 0, 1, \ldots, 15$, we computed $\hat{\lambda}_O$ and $\hat{\lambda}_S$ for subject 204, using the same mode/curvature warping but with draws obtained by skipping j draws from the Metropolis sampler between consecutive draws (i.e., we only use every $(j + 1)$st draw). For each j, labeled *Thinning*,

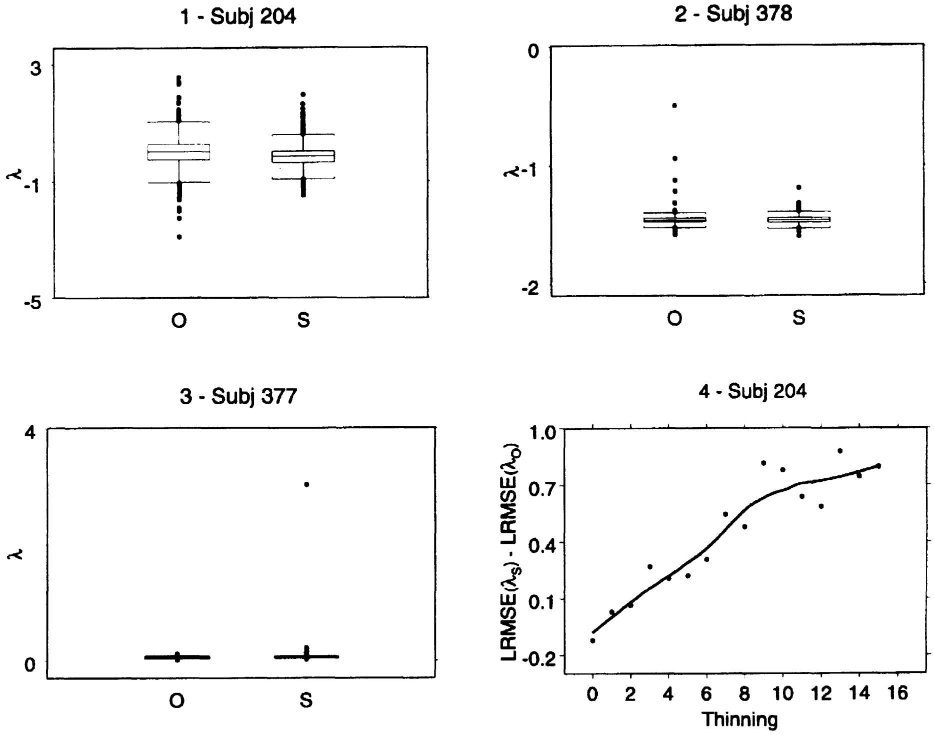

Figure 7. Three Extreme Cases under Mode/Curvature Warping with an Illustration of the Effect of Thinning.

500 draws were used for each estimator, and the process was repeated 1,000 times. The fourth panel of Figure 7 plots the resulting LRMSE($\hat{\lambda}_S$) - LRMSE($\hat{\lambda}_O$) as a function of thinning. We see that $\hat{\lambda}_O$ outperforms $\hat{\lambda}_S$ as soon as $j = 1$, and the improvement continues until it stabilizes at about $j = 15$, when the draws become essentially uncorrelated. Subject 204 was chosen because of the unusually high (estimated) lag-1 autocorrelations, which were respectively 0.89, 0.92, 0.84, and 0.97 for γ_{0i} through γ_{3i}, compared to the median lag-1 autocorrelation 0.86 for all 889 subjects. To investigate more broadly, we repeated the above process at $j = 4$ for each of the 889 subjects. The median lag-1 autocorrelation for the 889 cases was reduced to 0.48. The third panel of Figure 8 indicates that there is no longer a case where $\hat{\lambda}_S$ shows any noticeable advantage over $\hat{\lambda}_O$, while the fourth panel shows that the performance of $\hat{\lambda}_G$ compared to $\hat{\lambda}_O$ is essentially unchanged. The first two panels in Figure 8 again show that when $\hat{\lambda}_S$ performs poorly, most noticeably for subject 36 identified in the third panel, it is because of its tendency to produce more outliers and outliers of greater magnitude than those produced by $\hat{\lambda}_O$.

4.3 COMPARISON WITH CHIB AND JELIAZKOV'S PROPOSAL

Because the Metropolis sampler with a normal proposal was used to generate draws from the target density $p_i (i = 1, 2)$, we can easily implement Chib and Jeliazkov's (CJ)

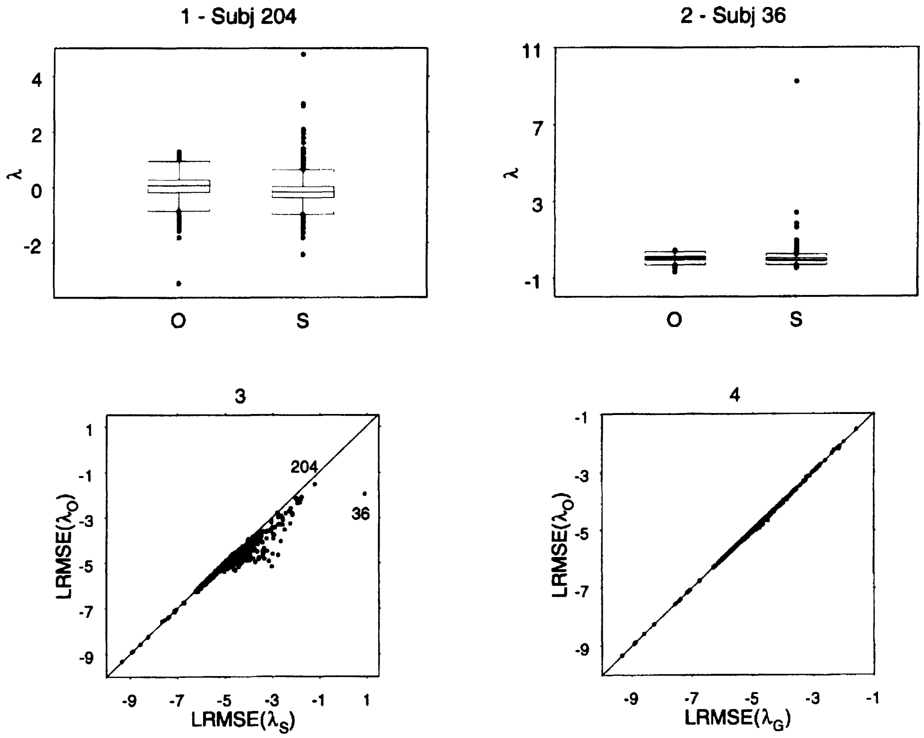

Figure 8. Performance of Warp-II Estimators with Mode/Curvature Transformation when Thinning = 4.

proposal of bridging p_i with its normal proposal density and then using the $\alpha(\theta)$ given in (1.13), with $\theta = \gamma$. As we emphasized and explained in Section 1.3, because the normalizing constants of the proposal densities are chosen to be known, we are able to directly estimate the normalizing constant c_i, $i = 1, 2$. To compare the performance of CJ's choice of α to the three choices we have investigated, we repeated the simulations as described previously, except this time we also generated 250 draws from each of the two normal proposal densities. Four estimators were thus compared for estimating each of c_1 and c_2: importance sampling with the proposal density as the trial density ($\hat{c}_S$), geometric ($\hat{c}_G$), optimal ($\hat{c}_O$), and CJ estimator ($\hat{c}_C$). Note that since the proposal distributions were chosen to have the same mode and curvature as the $p_i, i = 1, 2$ (see Section 4.1), these bridge-sampling estimators are automatically Warp-II mode/curvature matched estimators.

Figure 9 gives the scatterplots, across the 889 individuals, of log RMSEs of four paired comparisons for estimating c_1; the comparisons for c_2 were virtually identical to those for c_1 and are thus omitted. The first panel shows an order of magnitude advantage of $\hat{c}_O$ compared to $\hat{c}_C$, which has a relative RMSE 13 times larger than that of $\hat{c}_O$. (The relative RMSE is asymptotically the same as the RMSE of the log of the estimator.) The second panel shows nearly equivalent performance for $\log \hat{c}_G$ and $\log \hat{c}_O$, with $\log \hat{c}_O$ being slightly more accurate, as expected. The third panel shows that the simplest importance sampling estimator $\log \hat{c}_S$ performs really well, in fact slightly outperforming $\log \hat{c}_O$. Consequently,

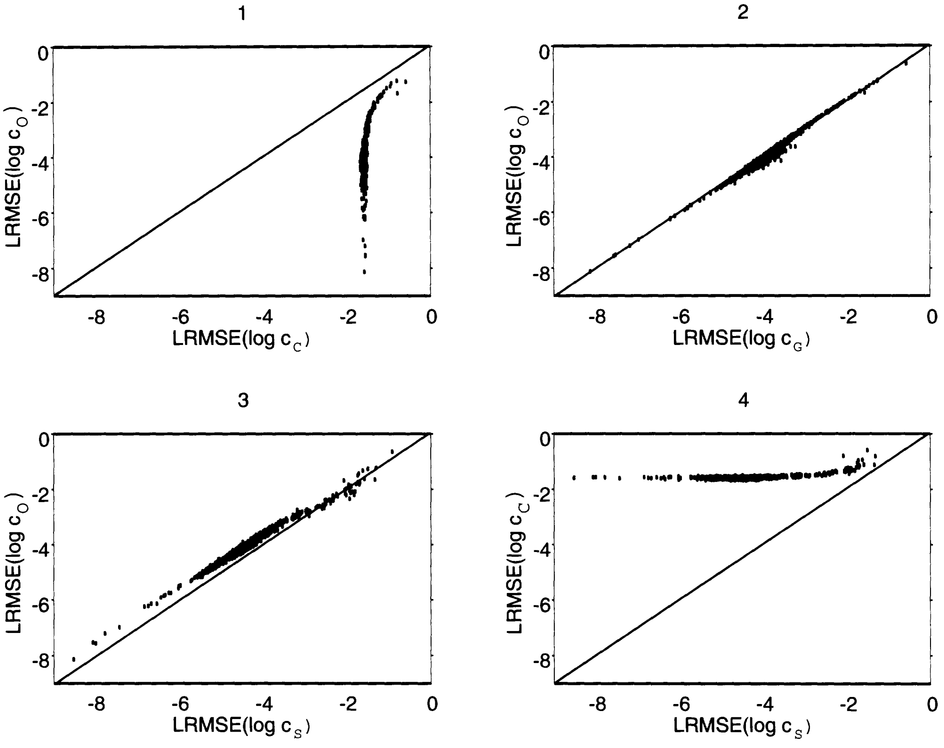

Figure 9. A Comparison of the Chib-Jeliazkov and Other Bridge Sampling Estimators.

$\log \hat{c}_S$ is also much better than $\log \hat{c}_C$, as clearly seen in the fourth panel. We found that altering the choice of the proposal distribution could produce improved performance for the the CJ's proposal in specific cases. However, even then $\hat{c}_O$ and $\hat{c}_S$ outperformed $\hat{c}_C$.

The reason that $\hat{c}_S$ slightly outperforms $\hat{c}_O$ is because $\hat{c}_O$ is (asymptotically) optimal only when the two sets of draws are independent, as we discussed after equation (1.4). In our current setting, the draws from the normal proposal density are independent, but the draws from the target density are highly correlated. Consequently, in terms of the "effective size" as we discussed in Section 1.1, the size from the target density is much smaller than 250. This suggests that the equal-weighted $\hat{c}_O$ is further from the actual optimal estimator than $\hat{c}_S$, which puts all weight on the normal proposal density. As we discussed in Section 1.1, a first-order correction is $\tilde{n}_1 = n_1(1-\hat{\rho})/(1+\hat{\rho})$, where $\hat{\rho}$ is an estimated lag 1 correlation. Figure 10 compares $\log \hat{c}_S$ with three correlation-adjusted $\log \hat{c}_O$'s. The first $\hat{\rho}$ was simply set to 0.86, the average lag-1 autocorrelation reported in Section 4.1. The second and third $\hat{\rho}$ values were respectively the sample lag-1 autocorrelation of l_{1j} and of $1/(l_{1j}+r)$, where l_{1j} is the one given in (1.5), whose form suggested us to use these two lag-1 autocorrelations. It is seen that all three $\hat{c}_O$'s are essentially equivalent, and they all slightly outperform $\hat{c}_S$. Note that, as we discussed in Section 4.2, even under the assumption of independent draws, $\hat{c}_O$ dominates $\hat{c}_S$ only when the size of draws used for $\hat{c}_S$ does not exceed the maximum of the two sizes used for $\hat{c}_O$, which may be viewed as an unfair comparison. (Recall in the

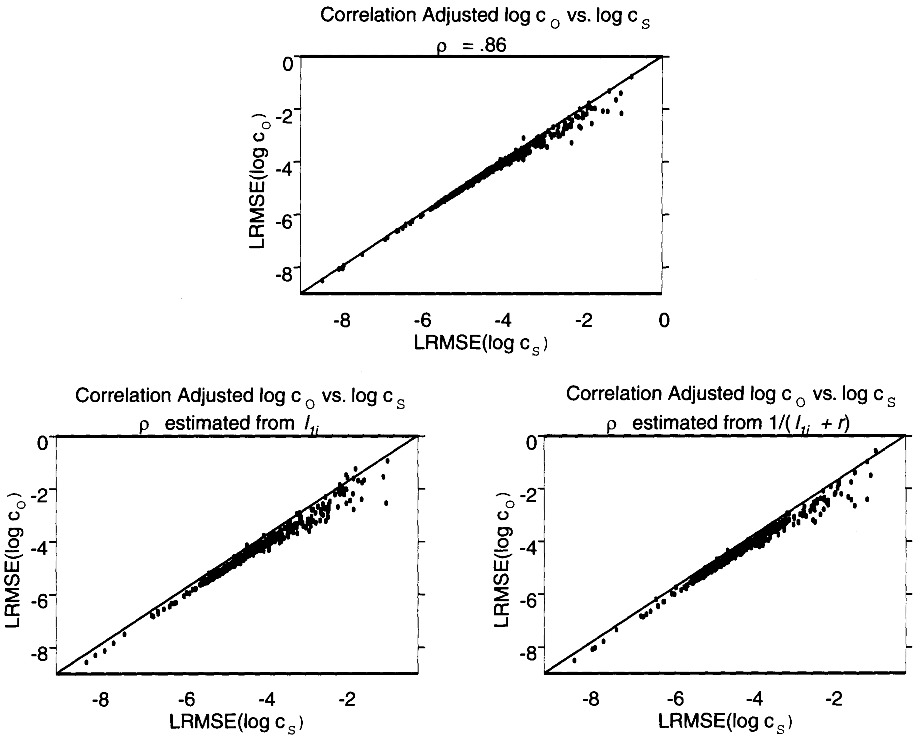

Figure 10. The Effect of Correlation Adjustments.

examples of Section 2.6 and Section 3, we have used twice as many draws for computing $\hat{\lambda}_S$ than for computing $\hat{\lambda}_O$ to make a fair comparison.) However, one must recognize that in CJ's setting, the draws from the proposal density (e.g., normal) are essentially free compared to that from the target density. Consequently, a more sensible comparison is to include all draws that have already made from the target distribution, because the key practical question here is how to best utilize these expensive draws.

From the above comparison, one might conclude that the draws from the target densities do not help because the simple $\hat{c}_S$ using draws from the proposal density worked quite well, even when compared to the correlation-adjusted $\hat{c}_O$. However, we have to keep in mind that this comparison was based on CJ's method which estimates c_1 and c_2 separately and then computes the estimate of λ via $\lambda = \log(c_1/c_2)$. Let us label the resulting four estimates of λ by $\hat{\lambda}_{SP}$, $\hat{\lambda}_{GP}$, $\hat{\lambda}_{OP}$, and $\hat{\lambda}_{CP}$, corresponding respectively to $\hat{c}_S$, $\hat{c}_G$, $\hat{c}_O$, and $\hat{c}_C$ (for estimating both c_1 and c_2). Here the subscript "P" stands for "proposal," as the key of CJ's method is to bridge the target density with the proposal density. The first two panels of Figure 11 compare, respectively, the LRMSE of $\hat{\lambda}_{OP}$ and of $\hat{\lambda}_{SP}$ with that of $\hat{\lambda}_{CP}$, which reinforce our previous conclusion that the CJ's choice of α can be quite inefficient. The LRMSE of $\hat{\lambda}_{GP}$ is very similar to that of LRMSE of $\hat{\lambda}_{OP}$ and thus not shown. The third panel compares the LRMSE of the Warp-II mode-curvature matched $\hat{\lambda}_O$ of Section 4.2, with that of $\hat{\lambda}_{CP}$. The fourth panel compares the LRMSE of $\hat{\lambda}_O$ with that of $\hat{\lambda}_{OP}$, which clearly shows

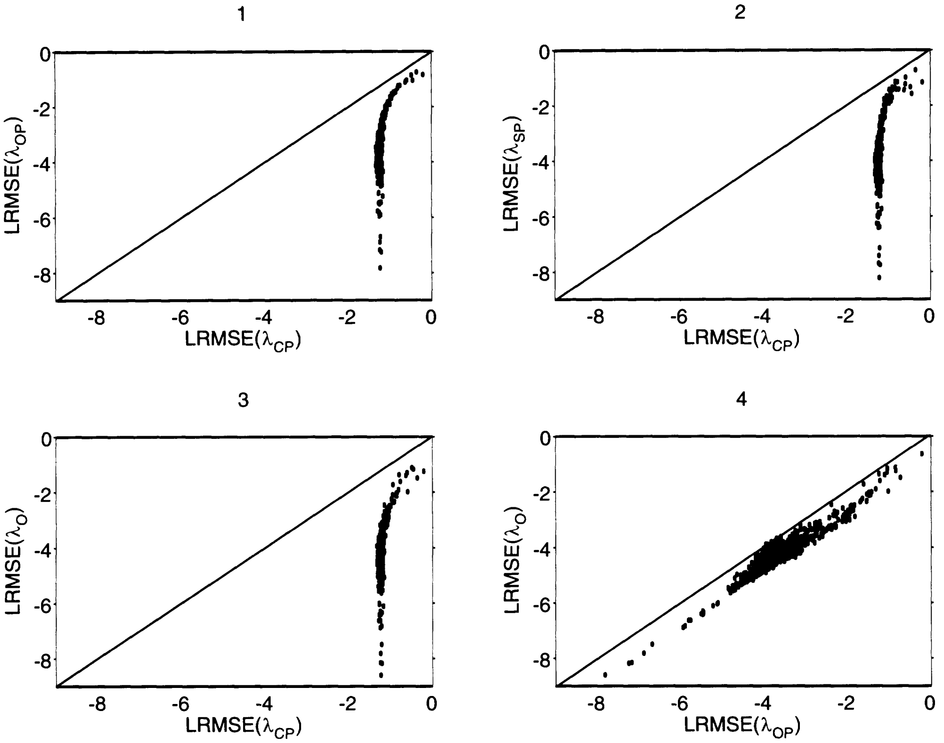

Figure 11. A Comparison of Direct and Indirect Procedures for Estimating λ.

that, in the current application, directly bridging the two target densities (namely, directly estimating the ratio) is better than separately estimating the two normalizing constants by bridging each target density with its corresponding proposal density and then taking the ratio. In other words, those draws from the target densities are indeed useful as it allows us to improve upon $\hat{\lambda}_{SP}$ (which is very similar to $\hat{\lambda}_{OP}$ for this example). The median RMSE for $\hat{\lambda}_{OP}$ is 1.9 times larger than that for $\hat{\lambda}_{O}$. The advantage for $\hat{\lambda}_{O}$ compared to $\hat{\lambda}_{CP}$ is proportionally greater than before: the median RMSE for $\hat{\lambda}_{CP}$ is 21 times that of $\hat{\lambda}_{O}$. One explanation for the advantage of bridging p_1 and p_2 directly is that the two distribution shapes, after Warp-II transformation, are more close to each other than to the symmetrical normal proposal distributions. This also suggests that we can use Warp-III symmetrizing transformations to further improve precision, as in the next section.

4.4 Empirical Investigations of Warp-III Estimators

The rationale for Warp-III transformations can be best understood by looking at their effects on the shape of the individual target distributions. Their effects on p_1 (corresponding to the unconstrained MLE) for subject 36 are illustrated in the contour plots given in Figure 12 and Figure 13. The three columns of these figures present the two-dimensional marginal distributions respectively of γ_0, γ_1, and γ_2 with γ_3: γ_3 being chosen because that is the dimension exhibiting the greatest amount of skewness. The first row of Figure 12 presents the original marginal distributions, where the H value is the Hellinger distance between the corresponding four-dimensional joint density (with mode recentered to the origin) and $N(0, I_4)$. The second row of Figure 12 displays the results after the Warp-III transformation with mode reflection (and recentering to the origin). The third row is the same as the second row except the reflection point is the estimated optimal μ obtained by maximizing (2.11) over *both* μ and S using 1,000,000 draws from $q_0 := N(0, I_4)$. The final row presents the results of the optimal μ reflection followed by the optimal S rescaling.

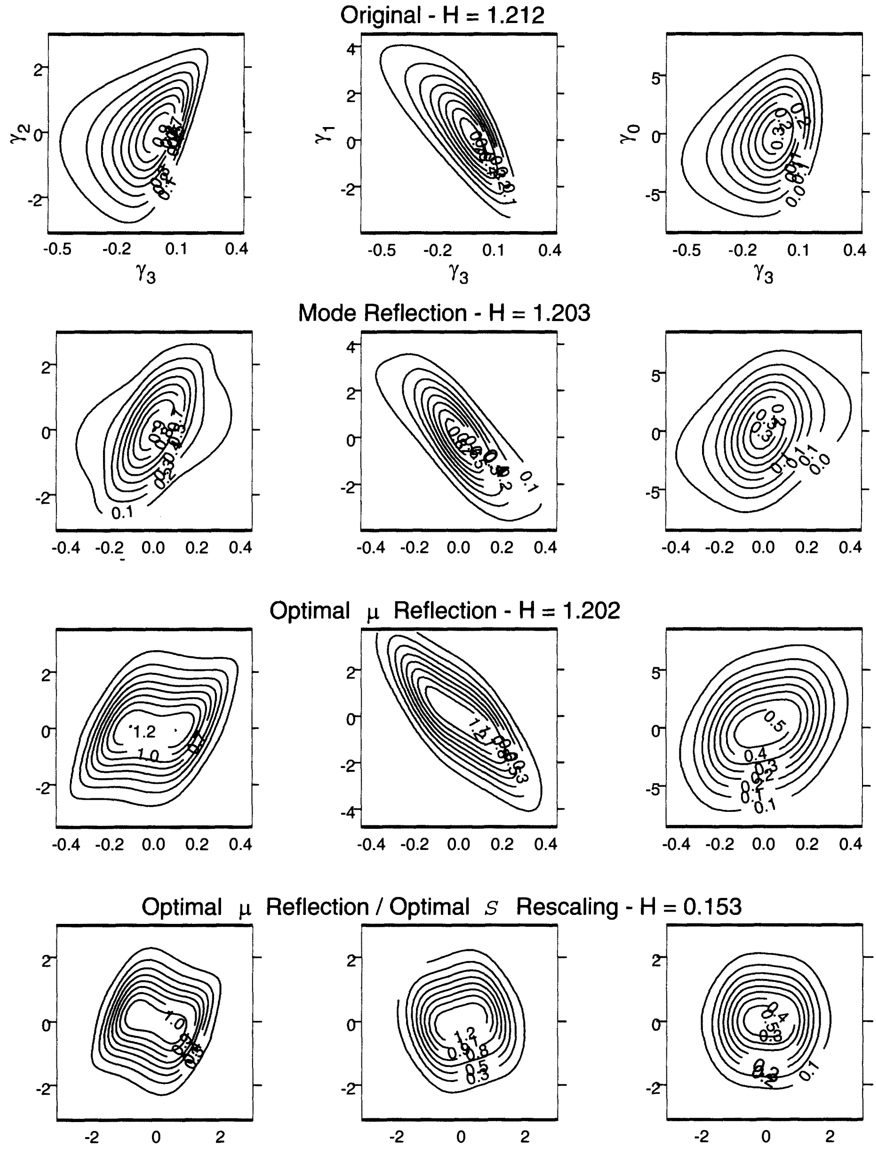

Figure 12. The Effect of Optimal Warping on Posterior Distributions.

The Warp-III transformation about the mode symmetrizes the originally very skewed distributions, but the resulting distributions exhibit varying degrees of heavy tails. Warping about the optimal μ reduces the tails, but leaves the distributions with a flatter center with possible multiple modes. Because the resulting distributions still differ greatly from $N(0, I_4)$ in terms of spread, neither warping achieves useful reduction of the H value. However, Warp-III using both the optimal (μ, S) dramatically reduces H value to 0.15, with the three bivariate marginal distributions much closer in appearance to a $N(0, I_2)$. We emphasize here

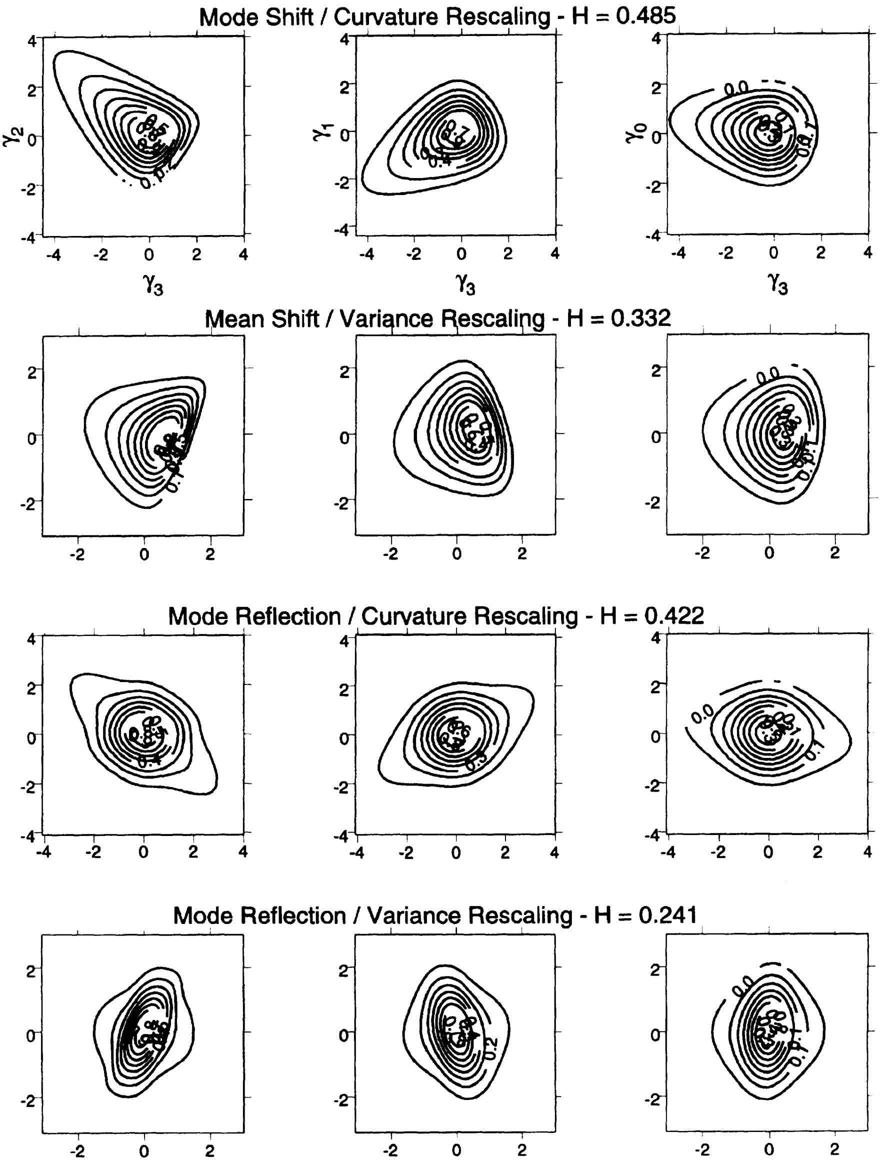

Figure 13. The Effect of Alternative Warp II and Warp III Transformations on Posterior Distributions.

Table 2. Comparison of Warping Transformations for Subject 36

		RMSE			*STD*		
Transformation	*H*	*log* $\hat{c}_S$	*log* $\hat{c}_G$	*log* $\hat{c}_O$	*log* $\hat{c}_S$	*log* $\hat{c}_G$	*log* $\hat{c}_O$
Warp-II Mode/Curvature	0.485	0.269	0.252	0.213	0.251	0.252	0.213
Warp-III Mode/Curvature	0.422	0.241	0.231	0.200	0.223	0.230	0.200
Warp-II Mean/Variance	0.331	0.063	0.075	0.046	0.063	0.075	0.046
Warp-III Mode/Variance	0.241	0.051	0.071	0.039	0.051	0.071	0.039
Warp-III Sample Optimal	0.282	0.227	0.084	0.156	0.191	0.069	0.050
Warp-III Optimal	0.153	0.051	0.040	0.031	0.051	0.040	0.031

that it is not necessary to use optimal $(\mu, \mathcal{S})$ in (2.11) in order for a Warp-III transformation to produce noticeable reductions in Hellinger distance over Warp-II transformation. As an example, the first two rows of Figure 13 present the effects of Warp-II transformations with mode/curvature matched and mode/variance matched, respectively. The next two rows presents the corresponding results under the Warp-III transformations using the same μ (i.e., mode) and scaling matrices, which resulted in, respectively, 13% and 28% reductions in the H value. The avoidance of optimizing (2.11) over μ and especially over $\mathcal{S}$ is particularly appealing for large dimensional problems.

To examine the effects of Warp-III transformations for computing ratio estimates between skewed distributions and symmetric distributions, we repeated the simulations described in Section 4.3 using 250 draws from each of p_1 and $N(0, I_4)$ for subject 36. This process was repeated 1,000 times for each of the warp transformations illustrated in the previous two figures, as well as for a Warp-III *sample* optimal estimator, which uses the optimal μ and $\mathcal{S}$ estimated via maximizing (2.11) based on the 250 draws already made from $N(0, I_4)$ (in contrast to the Warp-III optimal estimator, where the optimal values of $\{\mu, \mathcal{S}\}$ were estimated based on 1,000,000 draws external to our simulation study). The $\hat{c}_O$ estimator was adjusted for auto-correlation $\hat{\rho}$ with $\hat{\rho}$ set at 0.86, the median lag-1 correlation reported in Section 4.1. The simulation results are summarized in Table 2 and plotted in the first row of Figure 14. It is seen that Warp-III transformations dominate the Warp-II transformations with the same scaling matrix. Except for the Warp-III sample optimal transformation, both RMSE and the standard deviations (STD) increase as the Hellinger distance increases, and the RMSE's and STD's for $\log \hat{c}_O$ dominate those for $\log \hat{c}_S$ and $\log \hat{c}_G$.

The performance of the Warp-III sample optimal transformation deserves some comments. First, since the estimated optimal value μ and $\mathcal{S}$ vary with the sample, the reported Hellinger distance is the median of the Hellinger distances over the 1,000 replications. The Hellinger distances for those 1,000 replications ranged from 0.152 to 0.487, with 98% of them less than 0.422, the Hellinger distance for the Warp-III mode/curvature transformation. Thus, we would expect the RMSE's and the STD's to be smaller than those under Warp-III model/curvature transformation, but larger than expected given the median Hellinger distance. As expected, the STD's of $\log \hat{c}_O$ is less than that for both $\log \hat{c}_S$ and $\log \hat{c}_G$. Second, while all other transformations led to essentially unbiased estimators, Warp-III sample optimal transformation results in relatively large biases for all three estimators, with

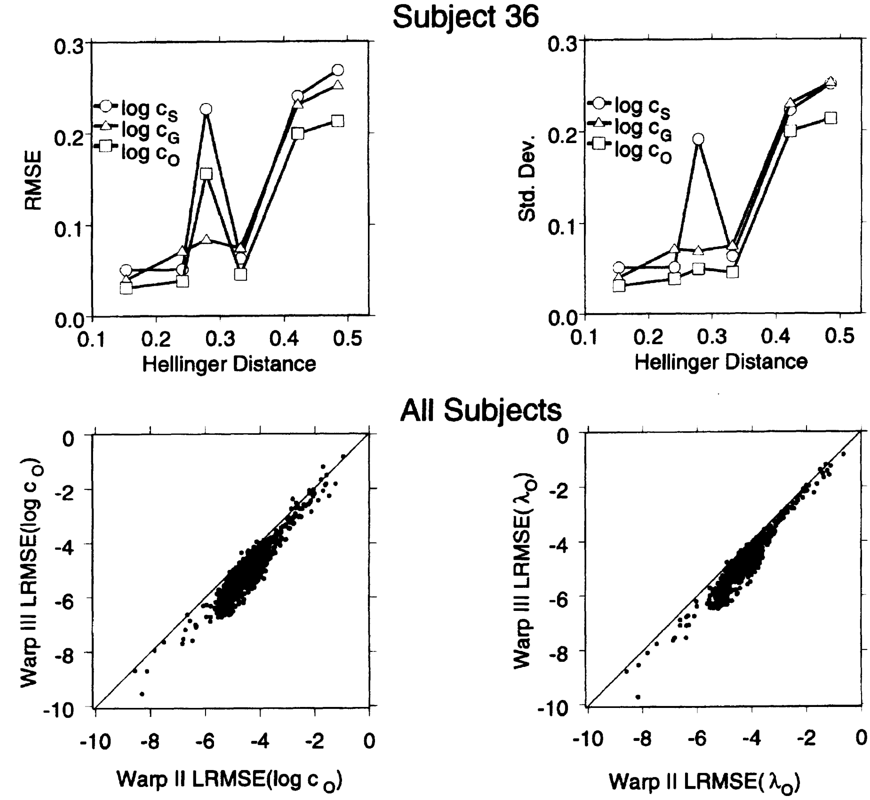

Figure 14. A Detailed Comparison of Warp II and Warp III Estimators.

$\log \hat{c}_G$ having smaller bias than that of $\log \hat{c}_O$. A possible explanation for this interesting phenomenon that $\hat{c}_G$ outperforms $\hat{c}_O$ may lie in the estimating equation underlying c_G:

$$c = \frac{E_0\left(\sqrt{\frac{\tilde{q}}{p_0}}\right)}{E_1\left(\sqrt{\frac{p_0}{\tilde{q}}}\right)}. \tag{4.1}$$

The numerator of (4.1) is proportional to the quantity (2.11) attempts to estimate. It is thus possible that $\log \hat{c}_G$ exhibits the smallest RMSE because the sample optimization is "tuned" for that estimator. This finding might have important practical consequences, because $\hat{\lambda}_G$ is (slightly) easier to implement and generally has practically comparable RMSE compared to $\hat{\lambda}_O$ with a fixed warp transformation. It may be preferred if it generally outperforms $\hat{\lambda}_O$ under the sample optimization procedure, a possibility for further investigation,

Even without optimization, the Warp-III mode-reflection/curvature-rescaling can produce significant reductions in RMSE for both c_1 and r estimates. To demonstrate the range of reductions that can be achieved, we repeated the simulations described in Sections 4.1 and 4.3 for the Warp-III mode/curvature transformation. The resulting comparisons on log of RMSE for $\log \hat{c}_O$ and $\hat{\lambda}_O$ across all 889 subjects, are presented in the second row of Figure 14. The Warp-III transformation is clearly superior with a median reduction in RMSE for $\log \hat{c}_O$ about 40% and an interquartile range from 27% to 56%. Similar results are observed for $\hat{\lambda}_O$, with a median reduction about 39% and an interquartile range from 28% to 51%.

5. LIMITATIONS, POSSIBLE REMEDIES, AND CONCLUDING REMARKS

5.1 DEALING WITH MULTIMODALITY: STOCHASTIC WARPING

It should be clear that all the transformations we discussed so far work best when the underlying distributions have only one (major) mode. (In fact, it is easy to construct multi-mode examples where a Warp-I or Warp-II transformation can actually increase the Hellinger distance.) As is well known, the problem of multimodality generally poses difficulties for MCMC and related problems. The most fundamental difficulty, of course, is to identify the (major) modes as well as their curvatures—a task that needs to be completed before one can produce trustworthy draws from p_1. Without such trustworthy draws, any method based on draws from p_1 is fundamentally problematic. But if the modes and their curvatures are approximately known, then one can construct $\tilde{p}_1$ as a mixture of normal or t distributions for example (e.g., West 1993), and then bridging p_1 with $\tilde{p}_1$, as with the unimodal cases. We emphasize that, thanks to the use of the bridge density, it is not crucial if we have missed a few modes when constructing $\tilde{p}_1$, as long as these modes are not too dominant. Indeed, even if we have missed some major modes when constructing $\tilde{p}_1$, the bridge sampling estimators are still generally more efficient than the corresponding importance sampling estimator using draws only from $\tilde{p}_1$ or p_1. With some complicated system/model p_1, such as those in theoretical physics, convenient theoretical approximations of p_1 are not available. On the other hand, one can make draws from various systems that are of interest (e.g., p_1, p_2, and some "in-between" systems; see Voter 1985). In such cases, Voter (1985) suggested the use of the following extension of (2.3)

$$r = \frac{\int E_2[q_1(w+D)\alpha(w)]\pi(dD)}{\int E_1[q_2(w-D)\alpha(w-D)]\pi(dD)}, \tag{5.1}$$

where $\pi(dD)$ should be chosen to effectively address the issue of multimodality. This is, in some sense, in the spirit of path sampling (e.g., Gelman and Meng 1998), for it attempts to create a "path," indexed by D, to link p_1 to p_2. (Indeed, the warping idea is also useful for path sampling.)

Estimator (5.1) is a case of stochastic warping, as it mixes the location shift parameter D over a distribution π, possibly continuous. Therefore, this is also in the spirit of the Warp-III warping transformations described in Section 2.4, where we discussed the possibility of mixing over all possible orthogonal transformations. However, the practicality of any such method should not be assumed without careful investigation, especially given our desire to achieve better Monte Carlo efficiency without unduly increasing the computational load.

5.2 DEALING WITH DISCRETE DISTRIBUTIONS: SWAPPING AND PERMUTATION

Another limitation of the warp transformations discussed so far is that they are not effective, even if possible to implement, when the underlying distributions are discrete, or more generally when they have "loose" neighborhood structures (i.e., not concentrated around a few modes). For discrete distributions, a different kind of warping transformations may be more effective.

For example, suppose both $q_1(w)$ and $q_2(w)$ are discrete and have the same support Ω. Suppose we know that $q_1(a) > q_1(b)$ for two states $a, b \in \Omega$, and the order is reversed for q_2. Then by picturing the two densities having opposite bump and dip at these two locations, we immediately see that we should swap the two probabilities for, say, q_2, in order to make the two distributions more similar. Mathematically, we can define a new unnormalized density $\tilde{q}_2$ such that $\tilde{q}_2(a) = q_2(b)$, $\tilde{q}_2(b) = q_2(a)$, and $\tilde{q}_2(w) = q_2(w)$ for $w \in \Omega \setminus \{a, b\}$. Clearly, $\tilde{q}_2$ and q_2 have the same normalizing constant. We thus can apply bridge sampling to $\{q_1, \tilde{q}_2\}$ to compute the same ratio r; in the notation of (2.5), this is the same as using $T_1(w) = w$ and T_2 the swap transformation between $w = a$ and $w = b$. However, the Monte Carlo efficiency of bridge sampling is improved because q_1 has more overlap with $\tilde{q}_2$ than with q_2. For example, simple algebra shows for the Hellinger distance critical for (2.2), $H(p_1, \tilde{p}_2) < H(p_1, p_2)$. It can also be directly verified that the asymptotic optimal relative error given in (1.6) is reduced by the swapping transformation.

Of course, in practice swapping a single pair of states may not produce significant improvement unless $p_1(a)$ and $p_2(b)$ represent a major portion of the probability mass, respectively, for p_1 and p_2. One may need to use a more general permutation transformation to achieve significant improvement in practical applications, and the exact permutation will depend on our knowledge of the two underlying distributions (e.g., we may need to swap two clusters of equal size)—in general, specific knowledge of the underlying densities are most important for finding helpful transformations. Finding effective warping transformations for discrete distributions is an important problem in genetic linkage analysis, as discussed in Meng (1999).

5.3 CONCLUDING REMARKS

The development of the bridge sampling was a direct consequence of our ability to simulate from densities with analytically intractable normalizing constants, a key advantage of MCMC. In other words, the gain in efficiency of the bridge sampling estimators, compared to commonly used importance sampling estimators based on a trial density, is an inherent byproduct of the efforts that we have already made in producing reliable MCMC draws. Once these draws are made, constructing efficient estimators for normalizing constants becomes an inference problem, that is, how to extract the most efficiency from the available data and information for inferring the "unknown" normalizing constants.

From an inferential point of view, this is a fascinating problem because there are many ways of linking the estimand with the simulated data, depending on how much shape information of the unnormalized densities we allow ourselves to use. This article demonstrates that, with more and more sophisticated warping transformations, we can achieve better and better estimation efficiency based on the same set of draws, and it seems there is no lower bound on the Monte Carlo error. This should not come as a surprise, nor is a contradiction to the well-known efficiency lower bounds such as the Cramér-Rao lower bound or Fisher information, because the problem of estimating the normalizing constant given an unnormalized density is not a usual inference problem. Indeed, if it were not for the computational limitation, there would be no inference problem to speak of as the normalizing constant is completely determined by the unnormalized density. Thus, as emphasized in the rejoinder of van Dyk and Meng (2000), in the context of analyzing simulated data, the issue of model selection is not which model is (approximately) true, but rather which model represents a more sensible compromise among human efforts, computational load and statistical efficiency.

Like any statistical or computational method, the applicability and advantages of the bridge sampling method, with or without warping, depends on particular applications. Nevertheless, there has been much empirical as well as theoretical evidence, in and outside the statistical literature, that demonstrate the effectiveness of bridge sampling for a variety of problems (e.g., Bennett 1976; Voter 1985; Meng and Wong 1996; Meng and Schilling 1996; DiCiccio et al. 1997; Jensen and Kong 1999; Servidea 2002). After comparing various theoretical approximations and simulation methods for computing Bayes factors (which are ratios of normalizing constants), DiCiccio et al. (1997) recommended the use of bridge sampling because it often achieves an order of magnitude improvement in accuracy compared to other methods they have investigated. We further recommend the use of warp bridge sampling estimators, particularly Warp-III estimators, whenever feasible and appropriate, because they can lead to an additional order of magnitude improvement in accuracy without unduly increasing the computational load.

ACKNOWLEDGMENTS

The authorship is alphabetic. We thank V. Dukic, A. Gelman, P. McCullagh, A. Mira, G.K. Nicholls, and J. Servidea for helpful exchanges, and the Editors, A. Buja and D. Scott, and reviewers for constructive comments that led to great improvement of the paper. We also thank Michele Schilling, a *Star Trek* fan, for inspiring us to adopt the very fitting term "warp." We were also pleased to learn, during our revision for the final version, that physicist C. Jarzynski of Los Alamos had an article in press that also generalizes Voter's (1985) idea to more general transformations in the context of free-energy calculations, and thereby independently confirming the usefulness and applicability of the key idea we propose in this article. Meng's research was supported in part by NSF grants DMS-9505043, DMS-9626691, and DMS-0072510, and in part by NSA grant MDA 904-96-1-0007.

[Received May 1999. Revised October 2000.]

REFERENCES

Bennett, C. H. (1976), "Efficient Estimation of Free Energy Differences from Monte Carlo Data," *Journal of Computational Physics*, 22, 245–268.

Bickman, L., Guthrie, P. R., Foster, E. M., Lambert, E. W., Summerfelt, W. T., Breda, C. S., and Heflinger, C. A. (1995), *Evaluating Managed Mental Health Services: Fort Bragg Experiment*, New York: Plenum, pp. 245–268.

Bock, R. D., and Aitken, M. (1981), "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm," *Psychometrika*, 37, 29–51.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, New York: Wiley.

Chen, M.-H., and Shao, Q. M. (1997a), "On Monte Carlo Methods for Estimating Ratios of Normalizing Constants," *The Annals of Statistics*, 25, 1563–1594.

——— (1997b), "Estimating Ratios of Normalizing Constants for Densities With Different Dimensions," *Statistica Sinica*, 7 607–630.

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

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

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, 92, 903–915.

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

Gelman, A., and Meng, X. L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

Geyer, C. J. (1994), "Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo, Technical Report 568, School of Statistics, University of Minnesota.

Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 657–699.

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

Johnson, V. (1999), "Posterior Distributions on Normalizing Constants," Technical Report, Duke University.

Meng, X. L. (1999), Invited discussion of Matthew Stephens's and Simon Tavaré's papers on statistical and computational approaches to genetic evolution. In *Bulletin of the International Statistical Institute; 52nd Session, Helsinki*, Vol. 3, pp. 111–112.

Meng, X. L., and Schilling, S. (1996), "Fitting Full-Information Factor Models and an Empirical Investigation of Bridge Sampling," *Journal of the American Statistical Association*, 91, 1254–1267.

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

Mira, A., and Nicholls, G. K. (2000), "Bridge Estimation of the Probability Density at a Point," Technical Report 456, Mathematics Department, University of Auckland.

Neal, R. M. (1993), "Probabilistic Inference Using Markov Chain Monte Carlo Methods," Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

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

Ott, J. (1979), "Maximum Likelihood Estimation by Counting Methods Under Polygenic and Mixed Models in Human Pedigrees," *American Journal of Humand Genetics*, 31, 161–175.

Ross, G. S. J. (1990), *Nonlinear Estimation*, New York: Springer-Verlag.

Schilling, S. (1998), "Estimating Decay in Treatment Effects: An Application of Nonlinear Mixed Models," Technical Report, Department of Psychology and Human Development, Vanderbilt University.

Servidea, J. D. (2002), "Bridge Sampling with Dependent Random Draws: Techniques and Strategy," Ph.D. Thesis, Department of Statistics, The University of Chicago.

van Dyk, D. A., and Meng, X. L. (2000), "The Art of Data Augmentation " (with discussion), *Journal of Compu-*

tational and Graphic Statistics, 10, 1–111.

Verdinelli, I., and Wasserman, L. (1995), "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio," *Journal of the American Statistical Association*, 90, 614–618.

Voter, A. F. (1985), "A Monte Carlo Method for Determining Free-Energy Differences and Transition State Theory Rate Constants," *Journal of Chemical Physics*, 82, 1890–1899.

West, M. (1993), "Approximating Posterior Distributions by Mixtures," *Journal of the Royal Statistical Society*, Ser. B, 55, 409–422.

Wichura, M. J. (1989), "An Algorithm for Patterson Gaussian Quadrature," Technical Report 257, Department of Statistics, University of Chicago.