FURTHER EXPLORATIONS OF LIKELIHOOD THEORY FOR MONTE CARLO INTEGRATION

AUGUSTINE KONG, PETER MCCULLAGH, XIAO-LI MENG, AND DAN L. NICOLAE

ABSTRACT. Monte-Carlo estimation of an integral is usually based on the method of moments or on an estimating equation. Recently, Kong, McCullagh, Meng, Nicolae and Tan (2003) proposed a likelihood based theory, which puts Monte-Carlo estimation of integrals on a firmer, less *ad hoc*, basis by formulating the problem as a likelihood inference problem for the baseline measure with simulated observations as data. In this paper, we provide further exploration and development of this theory. After an overview of the likelihood formulation, we first demonstrate the power of the likelihood-based method by presenting a universally improved importance sampling estimator. We then prove that the formal, infinite-dimensional Fisher-information based variance calculation given in Kong et al. (2003) is asymptotically the same as the sampling based "sandwich" variance estimator. Next, we explore the gain in Monte Carlo efficiency when the baseline measure can be parameterized. Furthermore, we show how the Monte Carlo integration problem can also be dealt with by the method of empirical likelihood, and how the baseline measure parameter can be properly profiled out to form a profile likelihood for the integrals only. As a byproduct, we obtain four equivalent conditions for the existence of unique maximum likelihood estimate for mixture models with known components. We also discuss an apparent paradox for Bayesian inference with Monte Carlo integration.

1. INTRODUCTION AND OVERVIEW

1.1. **The need for computing normalizing constants.** Let $\{q_\theta\}$ be a family of unnormalized probability/density functions on some sample space Γ, and let P_θ be the corresponding probability measure, $P_\theta(A) = \int_A q_\theta(x)\mu(dx)/c(\theta)$, for any measurable $A \subset \Gamma$. Here $c(\theta)$ is the normalizing constant, and μ is the dominating measure, typically Lebesgue, counting, or a mixture of the two. Whereas the normalizing constant is not required for many sampling methods, particularly Markov chain Monte Carlo (MCMC) methods (e.g. Gilks et al., 1996), it is a central quantity in many statistical and scientific problems. In physics, it is known as the partition function, and there is a large literature on how to estimate partition functions using MCMC (e.g. Bennett, 1976; Ceperley, 1995; Voter, 1985). In genetics,

Date: April 5, 2006.

Key words and phrases. Empirical likelihood; Exponential family; Fisher information; Generalized inverse; Group theory; Quotient space; Iterative proportional scaling; Invariance; Markov chain Monte Carlo; Normalizing constants; Semi-parametric models; Vector space.

Kong is with deCode genetics, Iceland, McCullagh and Nicolae are with Department of Statistics, The University of Chicago, and Meng is with Department of Statistics, Harvard University. Support for this research was provided in part by NSF Grant DMS-0305009 (for McCullagh) and by NSF Grant DMS-0072510 (for Meng and Nicolae).

many likelihoods are computed using the following identity

$$p(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta) = \frac{L(\theta \mid Y_{\text{obs}}, Y_{\text{mis}})}{L(\theta \mid Y_{\text{obs}})},$$

where $L(\theta|Y_{\text{obs}})$ is the likelihood of interest, and $L(\theta \mid Y_{\text{obs}}, Y_{\text{mis}})$ is the complete-data likelihood, which is typically easier to evaluate because of its simpler structure. With the help of sophisticated MCMC algorithms that can simulate from the conditional distribution of the missing data, $p(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta)$, usually of very high dimension, computing $L(\theta \mid Y_{\text{obs}})$ becomes a problem of estimating the normalizing constant $c(\theta) = L(\theta \mid Y_{\text{obs}})$ of the unnormalized $f(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta)$, namely, $q_\theta(Y_{\text{mis}}) = L(\theta \mid Y_{\text{obs}}, Y_{\text{mis}})$, using our generic notation (here Y_{obs} is fixed). Methods such as importance sampling and bridge sampling are then used to estimate $L(\theta \mid Y_{\text{obs}})$ (e.g. Ott, 1979; Geyer and Thompson, 1992; Irwin et al., 1994; Jensen and Kong, 1999; Stephens and Donnelly, 2001; Thompson, 2000).

In statistics, besides the obvious need for $c(\theta)$ in computing the likelihood, the computation of a Bayes factor is precisely a problem of computing normalizing constants. These likelihood and Bayesian computation problems are the major reasons for the recent interest in this topic, particularly given the increased complexity of Bayesian models (e.g. Geyer, 1994; Gelfand and Dey, 1994; Newton and Raftery, 1994; Chib, 1995; Verdinelli and Wasserman, 1995; Meng and Schilling, 1996; Meng and Wong, 1996; Chen and Shao, 1997b,a; DiCiccio et al., 1997; Gelman and Meng, 1998; Johnson, 1999; Chib and Jeliazkov, 2001; Meng and Schilling, 2002). In all these problems, the quantities of interest are either ratios of normalizing constants (e.g., likelihood ratios) or can be formulated as such. Even for computing the normalizing constant for a single model, for computational efficiency, we can estimate its value relative to the known value of the normalizing constant of a simple approximation to the model (DiCiccio et al., 1997; Meng and Schilling, 2002). Thus we focus on the problem of computing ratios of normalizing constants, or equivalently estimating a set of normalizing constants modulo a common positive multiple. By extending $\{q_\theta\}$ to include integrable but not necessarily non-negative functions, the formulation also covers general Monte Carlo integrations.

1.2. **A review of the likelihood theory.** An intriguing aspect of Monte Carlo (MC) integration is that there is no obvious (non-trivial) lower bound on the variance of the Monte Carlo estimator of the ratios with a *given* simulation size. This is, of course, not surprising, because it is well-known that the variance of the importance sampling estimator approaches zero as the distance between the target density and trial density approaches zero. Also, in MC integration problems, we know all quantities in $c(\theta) = \int_\Gamma q_\theta(x)\,\mu(dx)$, so there appears to be no inference problem to speak of. This has led to a number of quandaries in the general attempts to model MC simulated data just as real data, as discussed in Meng (2005).

To address this problem, Kong et al. (2003) proposed to treat the baseline measure, μ, as the unknown parameter; more arguments on why this is a natural strategy are given in Meng (2005). Specifically, let $q_1, \ldots, q_k$ be real-valued non-negative functions on Γ, and let μ be any non-negative measure on Γ. We are interested in estimating c_r/c_s, where $c_r = \int_\Gamma q_r(x)\,d\mu$ is assumed to be positive and finite. The simulated data are $n_r > 0$ independent samples from the rth weighted distribution

$$P_r(dx) = c_r^{-1} q_r(x)\,\mu(dx). \tag{1.1}$$

There may be additional functions $q_r, r = k+1, \ldots, k+m$ for which (the relative value of) $c_r = \int_\Gamma q_r(x)\,\mu(dx)$ must also be estimated, and these functions need not be non-negative. Also, as a theoretical device, by extending q_r to be a joint density of dependent draws, the formulation covers the practical situation where draws are realizations of a Markov chain.

Under the model of Kong et al. (2003), the parameter space is the set of all non-negative measures on Γ, but our interest lies in the $k+m$ linear functionals

$$c_r = \int_\Gamma q_r(x)\,d\mu, \quad r = 1, \ldots, k+m. \tag{1.2}$$

Since the simulated data are n independent pairs: $(y_1, x_1), \ldots, (y_n, x_n)$, where the labels $y_i \in \{1, \ldots, k\}$ are determined by the simulation design and $x_i \sim P_{y_i}$, the full likelihood for μ is

$$L(\mu; X) = \prod_{i=1}^n P_{y_i}(\{x_i\}) = \prod_{i=1}^n \mu(\{x_i\})\, c_{y_i}^{-1} q_{y_i}(x_i). \tag{1.3}$$

Here we have assumed that Γ is countable; the uncountable case is discussed in Section 1.3. Re-parameterizing in terms of the canonical parameter $\theta(x) = \log \mu(\{x\})$, the log likelihood for θ, except for a constant, is

$$\sum_{i=1}^n \theta(x_i) - \sum_{s=1}^k n_s \log c_s(\theta) = n \int_\Gamma \theta(x)\,d\hat{P} - \sum_{s=1}^k n_s \log c_s(\theta), \tag{1.4}$$

where $\hat{P}$ is the standard empirical measure which puts $1/n$ mass at each observed data point. The maximum likelihood estimate of μ is given by

$$\hat{\mu}(dx) = \frac{n\hat{P}(dx)}{\sum_{s=1}^k n_s \hat{c}_s^{-1} q_s(x)}, \tag{1.5}$$

where $\hat{c}_s$ is the MLE of c_s, which are obtained (up to a proportionality constant) as the solution of the first k equations of

$$\hat{c}_r = \int_\Gamma q_r(x)\,d\hat{\mu} = \sum_{i=1}^n \frac{q_r(x_i)}{\sum_{s=1}^k n_s \hat{c}_s^{-1} q_s(x_i)}, \quad r = 1, \ldots, k+m. \tag{1.6}$$

Note that this set of equations has a unique solution (up to a multiplicative constant) if and only if the set of values $\{q_r(x_i) \geq 0, i = 1, \ldots, n; r = 1, \ldots, k\}$ satisfies the "connected" condition of Vardi (1985), which we assume throughout this paper. We also remark that in the above formulation, the labels $\{y_1, \ldots, y_n\}$ play no role because they are not a part of the minimum sufficient statistic (Vardi, 1985). However, such label information is crucial for the "warp transformation" formulation, as discussed in Section 4.2.

1.3. **Uncountable sample spaces.** While the likelihood theory given in Kong et al. (2003) can be formally extended to cases where Γ is uncountable, the definition of θ becomes problematic because it requires the existence of a dominating measure ν on Γ such that the logarithmic derivative

$$\theta(x) = \log\left(\frac{d\mu}{d\nu}(x)\right)$$

is well-defined on Γ. That is to say, the parameter space is restricted to the set of measures on Γ that are absolutely continuous with respect to ν. This construction

is unsatisfactory when Γ is uncountable. The difficulty is that if μ is Lebesgue measure on $\mathbb{R}$, the "MLE" $\hat{\mu}$ is atomic, and thus not in the parameter space as described. In fact, there does not exist on $\mathbb{R}$ a common dominating measure ν such that $\hat{\mu} \ll \nu$ for all possible estimates $\hat{\mu}$.

The problem is more of a mathematical technicality than a practical obstacle, as equation (1.6) is clearly well-defined whether or not Γ is countable. Nevertheless, it is of some interest to acknowledge the problem, to offer a resolution and to explore the consequences. First, we define $(\Gamma, \mathcal{A})$ as a measure space, in which $\mathcal{A}$ is a σ-algebra of subsets sufficiently rich to include all singletons of Γ. Second, we assume that the functions $q_s(x)$ are $\mathcal{A}$-measurable. Finally, the parameter space $\mathcal{M}$ is taken to be the set of all non-negative measures defined on $\mathcal{A}$. The likelihood at $\mu \in \mathcal{M}$, $L(\mu; X)$, is still given by (1.3). Note here that we define likelihood through its original form using the *probability* of the observed event $\{x_1, \ldots, x_n\}$, not through any *density* function, which is not suitable here as there is no single dominating measure for all elements in $\mathcal{M}$.

From (1.3), it is clear that if $\mu(\{x_i\}) = 0$ or $c_{y_i} = \infty$ for at least one i, then $L(\mu; X) = 0$. Furthermore, for each $\mu \in \mathcal{M}$ such that $\mu(\{x_i\}) > 0$ for all $i = 1, \ldots, n$ and $\mu(\Gamma \backslash \{x_1, \ldots, x_n\}) > 0$, we define a $\tilde{\mu} \in \mathcal{M}$ so that $\tilde{\mu}(\{x_i\}) = \mu(\{x_i\})$ for all $i = 1, \ldots, n$, but $\tilde{\mu}(\Gamma \backslash \{x_1, \ldots, x_n\}) = 0$. Recall that $q_s(x) > 0$ for all $x \in \Gamma$, it is then evident that for each $s \in \{1, \ldots, k\}$,

$$\tilde{c}_s = \int_\Gamma q_s(x) d\tilde{\mu} < \int_\Gamma q_s(x) d\mu = c_s,$$

so that $L(\tilde{\mu}; x) > L(\mu; x)$. Therefore, as far as MLE is concerned, we can concentrate on measures with support on $\{x_1, \ldots, x_n\}$. This effectively implies that we can proceed as if Γ were countable.

Regardless of whether Γ is countable or not, the real power of the likelihood-based method is that any usable knowledge about μ can (and should) be used to form a sub-model for estimating μ, and hence to improve MC efficiency for the resulting estimates of the $c's$. The next section details such an exercise in the context of importance sampling.

2. A universal improvement for importance sampling

2.1. **Symmetrized importance sampling.** While the formulation and results in Section 1 cover the most general bridge sampling (Meng and Wong, 1996), the case with $k = 1$ is of special interest, because it corresponds to the widely used importance sampling approach via

$$\gamma \equiv \frac{c_2}{c_1} = \int_\Gamma \frac{q_2(x)}{q_1(x)} [c_1^{-1} q_1(x)] d\mu = \int_\Gamma \frac{q_2(x)}{q_1(x)} dP_1. \tag{2.1}$$

Here we assume $\mathcal{S}_2 \subset \mathcal{S}_1$, where $\mathcal{S}_r$ is the support of q_r, and we typically set $c_1 = 1$ because the trial density P_1 is completely known. We highlight this common application to emphasize that many current MC integrations can be improved upon because they needlessly ignore usable symmetry properties in the baseline measure (e.g., Lebesgue measure), which can be captured easily by a sub-model under the likelihood formulation.

Specifically, let $\mathcal{G}$ be a compact group acting on Γ in such a way that μ is invariant: $\mu(gA) = \mu(A)$ for each $g \in \mathcal{G}$. The sub-model is the one where the parameter space consists only of measures that are invariant under $\mathcal{G}$. The log

likelihood function (1.4) simplifies because $\theta(x) = \theta(gx)$ for each $g \in \mathcal{G}$. The MLE of μ is still given by (1.5) and (1.6), but with q_s and $\hat{P}$ replaced respectively by their group averages $\bar{q}_s(x) = \text{ave}_{g\in\mathcal{G}}\, q_s(gx)$ and $\hat{P}^{\mathcal{G}}(A) = \text{ave}_{g\in\mathcal{G}}\, \hat{P}(gA)$.

To illustrate, the sub-model shows that (2.1) can be *symmetrized* by group-averaging:

$$\gamma = \frac{c_2}{c_1} = \int_\Gamma \frac{\bar{q}_2(x)}{\bar{q}_1(x)}\, d\bar{P}_1. \tag{2.2}$$

Consequently, the usual importance sampling estimator

$$\hat{\gamma}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{q_2(x_i)}{q_1(x_i)} \equiv \frac{1}{n}\sum_{i=1}^{n} w(x_i), \tag{2.3}$$

where $\{x_1, \ldots, x_n\}$ are draws from P_1, is replaced by

$$\hat{\gamma}_n^{\mathcal{G}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{q}_2(x_i)}{\bar{q}_1(x_i)} \equiv \frac{1}{n}\sum_{i=1}^{n} w^{\mathcal{G}}(x_i). \tag{2.4}$$

Because of (2.2), $\hat{\gamma}_n^{\mathcal{G}}$ is unbiased for γ as long as $\mathcal{S}_2^{\mathcal{G}} \subseteq \mathcal{S}_1^{\mathcal{G}}$ (where $\mathcal{S}_r^{\mathcal{G}}$ is the support of $\bar{q}_r$), which is a weaker requirement than $\mathcal{S}_2 \subseteq \mathcal{S}_1$. Therefore the group average improves (or at least does no harm to) the robustness of importance sampling in the sense of providing more assurance of having enough support in the trial density.

There are several ways to see why (2.4) is more efficient than (2.3). First, (2.3) is an "extreme" special case of (2.4) with $\mathcal{G}$ the identity transformation, while the original analytic integration/summation over Γ in (2.1) can be viewed as the other extreme case of (2.4) with a group rich enough such that each of the $w^{\mathcal{G}}(x_i)$ in (2.4) is exactly the target value γ. The latter can be most easily seen when Γ is finite, where we can take $\mathcal{G}$ be the full permutation group on Γ. By using a suitable, much smaller sub-group, (2.4) takes advantage of our ability to do partial analytic and/or numerical summation, and then uses Monte Carlo to deal with the rest. In contrast, (2.3) relies entirely on MC simulation to estimate γ.

Second, we can view the group transformation as "reparametrizing" the sample space into a set of *orbits* and a *cross-section* that indexes the orbits. Group averaging then analytically integrates over each orbit, and leaves only the integration over the cross-section to MC simulation. In other words, (2.4) uses group averaging to integrate over part of the space and uses simulation to approximate the remainder. For example, suppose that $\Gamma = \mathbb{R}^d$, μ is Lebesgue, and $\mathcal{G}$ is the orthogonal group. Then group averaging is the same as the $d-1$ dimensional integration over all the angles in polar coordinates. This analytic averaging, if feasible, makes (2.4) more efficient than (2.3), because it effectively "Rao-Blackwellizes" out the $d-1$ angle coordinates, and thus (2.4) becomes a one-dimensional MC integral over the radius. Note that the correct "Rao-Blackwellization" carried out by the sub-model, as shown in Section 2.2, averages individual q_r, not the ratios $w = q_2/q_1$. The latter does not in general provide a consistent estimator because P_1 is usually not invariant under $\mathcal{G}$.

Third, group averaging increases the overlap among the underlying "densities" (in quotes as some "densities" can be negative), and thus reduces the variability in the importance-sampling weight w. That is, (2.4) is more efficient than (2.3) for any compact group $\mathcal{G}$. Asymptotically, this is a direct consequence of the Fisher information results for general k and m, obtained in Section 3, because a sub-model necessarily possesses larger Fisher information than the full model for the

same set of parameters. For $k = m = 1$, namely the usual importance sampling, in Section 2.2 we prove this for any finite sample size by showing that (2.4) is the Rao-Blackwell projection of (2.3) given the minimum sufficient statistic, i.e., the cross-section. Consequently, for any finite group $\mathcal{G}$ such that μ is $\mathcal{G}$-invariant,

$$\mathrm{Var}(\hat{\gamma}_n^{\mathcal{G}}) \le \mathrm{Var}(\hat{\gamma}_n), \quad \forall \ n \ge 1, \tag{2.5}$$

where the equality holds if and only if for almost all $x \in \mathcal{S}_1$ (with respect to μ)

$$\frac{\mathrm{ave}_{g\in G}\, q_2(gx)}{\mathrm{ave}_{g\in G}\, q_1(gx)} = \frac{q_2(x)}{q_1(x)}, \quad \forall\ g \in \mathcal{G}. \tag{2.6}$$

2.2. **A theoretical comparison.** The following theorem establishes that averaging over a larger group necessarily reduces the variance of (2.4) for each sample size, under the independence assumption.

Theorem 2.1. *Suppose $\mathcal{G}_2 \subset \mathcal{G}_1$ are two finite groups and μ is $\mathcal{G}_1$-invariant. Let $\mathcal{G}x = \{gx : g \in \mathcal{G}\}$ be the $\mathcal{G}$-orbit of x and*

$$w_j(x) = \frac{\mathrm{ave}_{g\in\mathcal{G}_j}\, q_2(gx)}{\mathrm{ave}_{g\in\mathcal{G}_j}\, q_1(gx)} = \frac{\int_{t\in\mathcal{G}_j x}\ q_2(t)\, d\mu}{\int_{t\in\mathcal{G}_j x}\ q_1(t)\, d\mu}, \quad j = 1, 2.$$

If $\{x_1, \ldots, x_n\}$ is an i.i.d. sample from P_1, then

$$\mathrm{Var}(\hat{\gamma}_n^{\mathcal{G}_1}) \le \mathrm{Var}(\hat{\gamma}_n^{\mathcal{G}_2}), \quad \forall\ n \ge 1, \tag{2.7}$$

where the equality holds if and only if there exists $\Omega \subset \mathcal{S}_1$, the support of P_1, such that $P_1(\Omega) = 1$ and,

$$\forall\ x \in \Omega, \quad w_2(gx) = w_2(x), \quad \text{for all } g \in \mathcal{G}_1. \tag{2.8}$$

Proof. We first prove that for $X \sim P_1$,

$$E\left[w_2(X) |\ \mathcal{G}_1 x\right] = w_1(x). \tag{2.9}$$

The left-hand side of (2.9) is

$$\frac{\int_{t\in\mathcal{G}_1 x} w_2(t) q_1(t)\, d\mu}{\int_{t\in\mathcal{G}_1 x} q_1(t)\, d\mu} = \frac{\int_{t\in\mathcal{G}_1 x} w_2(t)\, [\mathrm{ave}_{g\in\mathcal{G}_2}\, q_1(gt)]\ d\mu}{\int_{t\in\mathcal{G}_1 x} q_1(t)\, d\mu}, \tag{2.10}$$

where the equality holds because $w_2(t)$ and μ are $\mathcal{G}_2$-invariant and $\mathcal{G}_2 \subset \mathcal{G}_1$. The right side of (2.10) is $w_1(x)$ because $w_2(t)\,[\mathrm{ave}_{g\in\mathcal{G}_2}\, q_1(gt)] = \mathrm{ave}_{g\in\mathcal{G}_2}\, q_2(gt)$, and

$$\int_{t\in\mathcal{G}_1 x} \mathrm{ave}_{g\in\mathcal{G}_2}\, q_2(gt)\, d\mu = \mathrm{ave}_{g\in\mathcal{G}_2} \int_{t\in\mathcal{G}_1 x} q_2(gt) d\mu = \int_{t\in\mathcal{G}_1 x} q_2(gt) d\mu,$$

where the last equality follows from the fact that $\int_{t\in\mathcal{G}_1 x} q_2(gt) d\mu$ is $\mathcal{G}_2$-invariant.

It follows from (2.9) that $\hat{\gamma}_n^{\mathcal{G}_1} = \sum_i w_1(x_i)/n$ is the Rao-Blackwell projection of $\hat{\gamma}_n^{\mathcal{G}_2}$ when $\{x_1, \ldots, x_n\}$ are i.i.d. because

$$E\left[\hat{\gamma}_n^{\mathcal{G}_2} \,|\, \mathcal{G}_1 x_1, \ldots, \mathcal{G}_1 x_n\right] = \hat{\gamma}_n^{\mathcal{G}_1}. \tag{2.11}$$

Consequently, (2.7) holds, with equality if and only if there exists $\tilde{\Omega} \subset \mathcal{S}_1$ with $P_1(\tilde{\Omega}) = 1$ such that

$$\forall x \in \tilde{\Omega}, \quad w_2(x) = w_1(x). \tag{2.12}$$

To prove that (2.12) implies (2.8), let $\mathcal{B} = \cup_{g\in\mathcal{G}_1}\{gx : \ x \in \mathcal{S}_1 \backslash \tilde{\Omega}\}$. Then $P_1(\mathcal{B}) = 0$ because $\mathcal{G}_1$ is finite. Let $\Omega = \mathcal{S}_1 \backslash \mathcal{B} \subset \tilde{\Omega}$. Then $P_1(\Omega) = 1$. Furthermore,

if $x \in \Omega$, then $gx \in \Omega \subset \tilde{\Omega}$ for any $g \in \mathcal{G}_1$. Consequently, for any $x \in \Omega$, since $w_1(x)$ is $\mathcal{G}_1$-invariant, (2.12) implies $w_2(gx) = w_1(gx) = w_1(x) = w_2(x)$ for any $g \in \mathcal{G}_1$, which is (2.8).

To prove the converse, we first note that for any x,

$$\text{ave}_{g\in\mathcal{G}_1}\, q_r(gx) = \text{ave}_{g_1\in\mathcal{G}_1}\{\text{ave}_{g_2\in\mathcal{G}_2}\, q_r(g_2g_1x)\}, \quad r = 1, 2, \tag{2.13}$$

which implies

$$\text{ave}_{g\in\mathcal{G}_1}\, q_2(gx) = \text{ave}_{g_1\in\mathcal{G}_1}\{w_2(g_1x)\,\text{ave}_{g_1\in\mathcal{G}_2}\, q_1(g_2g_1x)\}. \tag{2.14}$$

Consequently, if (2.8) holds, then (2.14) becomes $\text{ave}_{g_1\in\mathcal{G}_1}\{w_2(x)\,\text{ave}_{g_2\in\mathcal{G}_2}\, q_1(g_2g_1x)\}$ for any $x \in \Omega$, which together with (2.13) implies

$$\text{ave}_{g\in\mathcal{G}_1}\, q_2(gx) = w_2(x)\,\text{ave}_{g\in\mathcal{G}_1}\, q_1(gx).$$

This establishes (2.12) when we take $\tilde{\Omega} = \Omega$. □

This theorem provides a theoretical confirmation that our ability to carry out the $\text{ave}_{\mathcal{G}_1}$ operator analytically, in addition to our ability to evaluate $\text{ave}_{\mathcal{G}_2}$, always helps to reduce the MC error unless the group averages under $\mathcal{G}_1$ are already invariant under $\mathcal{G}_2$, in the sense of (2.8). The proof makes it clear that the reduction in variance is achieved by the usual Rao-Blackwellization. Taking $\mathcal{G}_1 = \mathcal{G}$ and $\mathcal{G}_2$ identity transformation gives (2.5) and (2.6).

2.3. **Practical implications.** A consequence of (2.5)–(2.6) is that we can improve on the standard MC estimators such as (2.3) by using convenient choices of $\mathcal{G}$ for which (2.4) dominates (2.3). For the most common applications of (2.3) in statistics, where $\Gamma = \mathbb{R}^d$ and μ is the Lebesgue measure, we can always take the two-element group $\mathcal{G}_O = \{I_d, -I_d\}$, where I_d is the $d \times d$ identity matrix. For any problem where (2.3) can be implemented, it is trivial to implement (2.4) with $\mathcal{G} = \mathcal{G}_O$:

$$\hat{\gamma}_n^{\mathcal{G}_O} = \frac{1}{n}\sum_{i=1}^{n} \frac{q_2(x_i) + q_2(-x_i)}{q_1(x_i) + q_1(-x_i)}, \tag{2.15}$$

where $q_r(x) = 0$ if x is outside the support of q_r, $r = 1, 2$. By (2.5)–(2.6), $\text{Var}(\hat{\gamma}_n^{\mathcal{G}_0}) < \text{Var}(\hat{\gamma}_n)$ unless $q_2(-x)/q_1(-x) = q_2(x)/q_1(x)$ for almost all $x \in \mathcal{S}_1$.

In fact, even when (2.3) fails to provide a consistent estimator because $\mathcal{S}_2 \not\subset \mathcal{S}_1$, (2.15) can still be consistent as it requires the weaker assumption $\mathcal{S}_2^{\mathcal{G}} \subseteq \mathcal{S}_1^{\mathcal{G}}$, namely, the support of $q_1(x)+q_1(-x)$ covers that of $q_2(x)+q_2(-x)$. As an extreme example, consider $d = 1$, $c_1 = 1$, $\mathcal{S}_2 = \mathbb{R}$ but $\mathcal{S}_1 = [0, +\infty)$. Then (2.3) will only estimate $\int_0^{+\infty} q_2(x)\, d\mu$. By contrast, because $q_1(-x_i) = 0$ for $x_i \sim P_1$, (2.15) becomes

$$\frac{1}{n}\sum_{i=1}^{n} \frac{q_2(x_i)}{q_1(x_i)} + \frac{1}{n}\sum_{i=1}^{n} \frac{q_2(-x_i)}{q_1(x_i)}, \tag{2.16}$$

which correctly estimates

$$\int_0^{+\infty} q_2(x)\, d\mu + \int_0^{+\infty} q_2(-x)\, d\mu = \int_{-\infty}^{+\infty} q_2(x) d\mu.$$

Upon recognizing the support problem of q_1, one would apply (2.3) twice to form (2.16) to estimate $\int_{\mathbb{R}} q_2(x)\, d\mu$, whereas (2.15) achieves this automatically. This illustrates the robustness of (2.15), or other versions of (2.4), over (2.3) in dealing with the well-known "tail" problem with importance sampling, because it can greatly reduce biases caused by lack of support in the trial density. The requirement

of making more function evaluations by (2.15) or (2.4) is often a negligible premium for its greater efficiency *and* robustness, especially in comparisons with the expense of making the MC draws. In fact, for cases like (2.16) the corrected importance sampling requires the same number of function evaluations as implementing (2.15).

The $\mathcal{G}_O$ group is only one of many that can be used to improve efficiency. For example, one can replace $-I_d$ in $\mathcal{G}_O$ by any of the $2^d - 2$ other diagonal matrices where the diagonal elements are either 1 or -1. Each of these groups represents reflections with respect to some of the d axes, and which one is optimal depends on q_2 and q_1. One advantage of using $\mathcal{G}_O$ is that it automatically symmetrizes $\bar{q}_r$ on each one-dimensional subspace. While the comparisons of (2.4) among non-nested groups can be mathematically complicated, intuitively $\mathcal{G}_O$ is a good "default" choice compared to other two-element reflection groups. If d is not too large, then we can and should consider using the full reflection group consisting of all 2^d diagonal matrices with diagonal elements ± 1, which is superior to any of its sub-group as guaranteed by Theorem 2.1. Further improvement is also possible by using different reflection points/axes for different distributions, as investigated in Meng and Schilling (2002); see Section 4.2.

As briefly mentioned in Kong et al. (2003), there are some similarities between the group averaging method with the *importance link function* (ILF) method of MacEachern and Peruggia (2000). The key of the ILF method is to construct a finite number of 1–1 and onto importance link functions $g_i, i = 1, \ldots, I$ with domain $B_i \subset \Gamma_0$, where Γ_0 is a subset of Γ, such that $\{T_i \equiv g_i(B_i), i = 1, \ldots, I\}$ forms a partition of Γ. Thus, integration on Γ can be estimated from the integral on each T_i via importance sampling using draws from a trial density concentrated on Γ_0. This is a very effective strategy to deal with a common problem in MCMC where the draws "get stuck" in part of the space, say Γ_0, but one needs to estimate integrals on the whole space Γ. Indeed, MacEachern and Peruggia (2000) proposed to use this method for handling reducible chains. In this regard, group averaging achieves the same goal and provides a more systematic way to construct link functions. If Γ_0 is a cross-section of the orbits, then $\{g\Gamma_0, g \in \mathcal{G}\}$ automatically form a partition of Γ and (2.4) will be the same as the ILF estimator using the same $\mathcal{G}$. Estimator (2.16) is such an example with $g_1(x) = x$ and $g_2(x) = -x$. In addition, the group formulation makes it clear that Γ_0 needs to contain at least a cross-section in order for the support of the $\bar{P}_1$ to cover Γ. When Γ_0 is richer than a cross-section, the group averaging estimator (2.4) is more efficient than the ILF estimator using the same $\mathcal{G}$ because the latter is generally not the Rao-Blackwell projection given the cross-section. As a trade-off, the ILF estimator does not require $\{g_i, i = 1, \ldots, I\}$ to be a group, when they are constructed based on information other than symmetries in the baseline measure.

3. Asymptotic covariance matrix

3.1. **Formal Fisher information calculation.** Apart from the special case with $k = 1$ (and $m \geq 1$), the exact calculation of the variance of the MLE of c is not tractable. However, we can obtain the asymptotic covariance matrix via the usual Fisher information calculation, at least formally. The following result (provided in Kong et al., 2003), based on the concept of *Fisher information measure* (e.g. McCullagh, 1999), was introduced to deal with Fisher information "matrix" of infinite order, countably or uncountably.

Specifically, the Fisher information measure for $\theta = \log \mu$ is

$$n\mathcal{I}(A,B) = \sum_{r=1}^{k} n_r \big(P_r(A \cap B) - P_r(A)P_r(B)\big),$$

where P_r is the distribution in (1.1). When Γ is countable, we also use $\mathcal{I}$ (without argument) to denote the $|\Gamma| \times |\Gamma|$ density matrix of the Fisher information measure; thus $\mathcal{I}$ is the usual Fisher information matrix, although of countably infinite order. The asymptotic covariance matrix of $\hat{\theta}$ is the inverse Fisher information matrix, $n^{-1}\mathcal{I}^-$, and the asymptotic covariance matrix of $d\hat{\mu}$ is $n^{-1}d\mu(x)\,d\mu(y)\,\mathcal{I}^-(x,y)$, where $\mathcal{I}^-(x,y)$ is the (x,y) element of $\mathcal{I}^-$, as indexed by $\Gamma \times \Gamma$. From expression (1.6) for $\hat{c}_r$, we find that the asymptotic covariance of $\log \hat{c}$ is given by

$$\operatorname{cov}(\log \hat{c}_r, \log \hat{c}_s) = n^{-1} \int_{\Gamma\times\Gamma} \mathcal{I}^-(x,y)\, dP_r(x)\, dP_s(y), \quad 1 \le r,s \le k+m. \tag{3.1}$$

Before we proceed further, we remark that so far as contrasts of $\log \hat{c}$ are concerned, two variance matrices V, V' are equivalent if $a^\top V b = a^\top V' b$ for all contrast vectors a, b. In other words, $a^\top (V - V')b = 0$, so we may add to V any (symmetric) matrix W such that $a^\top W b = 0$ without affecting the value $a^\top (V+W)b$, the covariance of two contrasts $a^\top \log \hat{c}$ and $b^\top \log \hat{c}$, where $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_{k+m})$. The set $\mathcal{W}$ of such symmetric matrices is the set $1x^\top + x1^\top$, which is a vector subspace of dimension k. The set of symmetric matrices that are equivalent to V is the coset $V + \mathcal{W}$ of symmetric $k \times k$ matrices. Not all elements of this coset need be positive definite. Cosets of this sort arise naturally as the set of symmetric generalized inverses of a non-invertible symmetric matrix A whose kernel is $\mathbf{1}$, the set of constant vectors. In particular, if $A1 = 0$ then $AWA = 0$ for each $W \in \mathcal{W}$. If A^- is a generalized inverse of A, i.e. $AA^-A = A$, then $A^- + W$ is also a generalized inverse for each $W \in \mathcal{W}$. If A^- is symmetric and $\ker(A) = \mathbf{1}$, the coset $A^- + \mathcal{W}$ is the set of symmetric generalized inverses of A. In this paper, whenever A is symmetric, A^- will be restricted to be a symmetric generalized inverse, and any equality between generalized inverses is interpreted in the sense of equivalence. The use of such generalized inverses makes it possible to express the asymptotic covariance matrix of $\log \hat{c}$, which is singular, in a symmetric form without the awkwardness and asymmetry associated with fixing an arbitrary component of $\hat{c}$.

3.2. **Deriving the matrix version.** While expression (3.1) is extendable to cases where Γ is uncountable, for the case where Γ is finite or countably infinite, we can obtain the usual matrix form. Specifically, let $P_{\text{mix}} = \sum_{r=1}^{k} f_r P_r$ be the mixture probability where $f_r = n_r/n$, $r = 1, \ldots, k$. Then the matrix $\mathcal{I}$ is given by

$$\mathcal{I} = D - \mathcal{P}_k F \mathcal{P}_k^\top,$$

where $D = \operatorname{diag}\{P_{\text{mix}}(\{x\})\}$, $\mathcal{P}_k$ is a $|\Gamma| \times k$ matrix with rth column given by $P_r(\{x\})$ for $x \in \Gamma$ with x as the row index, and $F = \operatorname{diag}\{f_1, \ldots, f_k\}$. Provided that P_{mix} is strictly positive on Γ, the matrix $\mathcal{I}$ has kernel equal to $\mathbf{1}$. Let $O_k = \mathcal{P}_k^\top D^{-1} \mathcal{P}_k$ and let $(F^{-1} - O_k)^-$ be a generalized inverse. Then the matrix

$$\mathcal{I}^- = D^{-1} + D^{-1}\mathcal{P}_k (F^{-1} - O_k)^- \mathcal{P}_k^\top D^{-1} \tag{3.2}$$

is a generalized inverse of $\mathcal{I}$. Thus, writing $\hat{c}^{(k)} = (\hat{c}_1, \ldots, \hat{c}_k)$, asymptotically,

$$n \operatorname{cov}(\log \hat{c}^{(k)}) = \mathcal{P}_k^\top \mathcal{I}^- \mathcal{P}_k = O_k + O_k (F^{-1} - O_k)^- O_k, \tag{3.3}$$

which involves only symmetric matrices of order k. The inverse asymptotic variance of $\log \hat{c}^{(k)}$, i.e. the asymptotic precision, is $n(O_k^- - O_k O_k^- F O_k O_k^-)$.

Similarly, for $\hat{c}^{(k+m)} = (\hat{c}_1, \ldots, \hat{c}_{k+m})$, asymptotically we have

$$n \operatorname{cov}(\log \hat{c}^{(k+m)}) = \begin{pmatrix} O_k + O_k L_k O_k & O_{m,k}^\top + O_k L_k O_{m,k}^\top \\ O_{m,k} + O_{m,k} L_k O_k & O_m + O_{m,k} L_k O_{m,k}^\top \end{pmatrix}, \tag{3.4}$$

where $L_k = (F^{-1} - O_k)^-$, $O_m = \mathcal{P}_m^\top D^{-1} \mathcal{P}_m$, $O_{m,k} = \mathcal{P}_m^\top D^{-1} \mathcal{P}_k$, with $\mathcal{P}_m$ the counterpart of $\mathcal{P}_k$ but for $r = k+1, \ldots, k+m$. Note that for $f = (f_1, \ldots, f_k)^\top$, we have $O_k f = 1$ and $(F^{-1} - O_k) f = 0$, so $F^{-1} - O_k$ is singular. For each vector $\alpha \in \mathbb{R}^k$, $(F^{-1} - O_k)^- + \alpha f^\top + f \alpha^\top$ is also a symmetric generalized inverse. But the choice of α has no effect on the variance of any contrast in expression (3.4).

We remark in passing that from (3.2) $\mathcal{I}^- \geq D^{-1}$ in the sense of Löwner ordering, from which we obtain the asymptotic inequality

$$\operatorname{Var}(\log(\hat{c}_r/\hat{c}_s)) \geq n^{-1} \int_\Gamma \Big(\frac{dP_r(x)}{dP_{\mathrm{mix}}(x)} - \frac{dP_s(x)}{dP_{\mathrm{mix}}(x)}\Big)^2 P_{\mathrm{mix}}(dx), \quad \forall\, r, s \in \{1, \ldots, k+m\}.$$

3.3. **Estimating equation "sandwich" version.** A technical difficulty with the Fisher information approach arises when Γ is uncountable. In such cases, the log density estimate $\hat{\theta}$ is generally inconsistent in the sense of pointwise convergence and thus the Fisher information approach presented in Section 3.1 can only be viewed as a formal calculation, suggestive but not rigorous. In this section we show the formula (3.1) or equivalently (3.4) are indeed correct even when Γ is uncountable.

First, equation (1.6) gives an estimating equation for $\log \hat{c}^{(k)}$ via

$$\sum_{i=1}^n \frac{\partial \log[P(x_i|y_i;c)]}{\partial \log c}\bigg|_{c=\hat{c}^{(k)}} = 0, \tag{3.5}$$

where

$$P(x|y;c) = \frac{f_y q_y(x) c_y^{-1}}{\sum_{s=1}^k f_s q_s(x) c_s^{-1}}. \tag{3.6}$$

Applying the standard "sandwich" approach, albeit on the quotient space $\log c \in \mathbb{R}^k/\mathbf{1}$, we obtain, asymptotically

$$n \operatorname{cov}(\log \hat{c}^{(k)}) = \tilde{\mathcal{I}}^- V \tilde{\mathcal{I}}^-,$$

where, denoting by E_r and cov_r, the expectation and the variance under P_r,

$$\tilde{\mathcal{I}} = \sum_{r=1}^k f_r \mathrm{E}_r \left[-\frac{\partial^2 \log[P(x|y;c)]}{(\partial \log c)(\partial \log c)^\top}\bigg|_{c=\hat{c}^{(k)}} \right] = F O_k F - F, \tag{3.7}$$

and

$$V = \sum_{r=1}^k f_r \operatorname{cov}_r \left[\frac{\partial \log[P(x|y;\log c)]}{\partial \log c}\bigg|_{c=\hat{c}^{(k)}} \right] = F O_k F - F O_k F O_k F. \tag{3.8}$$

Therefore, asymptotically,

$$n \operatorname{cov}(\log \hat{c}^{(k)}) = (I - O_k F)^- (O_k - O_k F O_k)(I - O_k F)^{-\top}, \tag{3.9}$$

where $A^{-\top}$ denotes $(A^{-1})^\top$.

To show that (3.3) and (3.9) are equivalent, we only need to show that for a particular choice of the generalized inverse, they are equivalent. A convenient choice is the Moore-Penrose generalized inverse A^+ for any matrix A, which is unique and

satisfies $AA^+A = A$, $A^+AA^+ = A^+$, and A^+A and AA^+ are symmetric. Using this choice, it can be shown that both (3.3) and (3.9) are equivalent to $(I - O_kF)^+O_k$. And thus for computing the variance of any contrast of $\log \hat{c}^{(k)}$, (3.3) and (3.9) are equivalent. The extended version (3.4) can be derived directly in similar way, although the algebra is a bit more involved. In this derivation, a key is to observe that the f_y in (3.6) plays no role in (3.5) because $\log f_y$ is a constant. Thus we can effectively remove f_y from (3.6), which would then allow the y index extended to include $y = k+1, \ldots, k+m$ (for which $f_y = 0$).

3.4. **A numerical example.** For a numerical example, we take $k = 3$, μ unit Poisson, $n_1 = n_2 = n_3$, and $q_r(x) = r^x$ so that P_r is Poisson with mean r. Using (3.3), we find that

$$O_3 = \begin{pmatrix} 1.426 & 0.958 & 0.616 \\ 0.958 & 1.038 & 1.004 \\ 0.616 & 1.004 & 1.380 \end{pmatrix}, \qquad \mathcal{P}_3^\top \mathcal{I}^- \mathcal{P}_3 = \begin{pmatrix} 1.576 & 0.949 & 0.475 \\ 0.949 & 1.039 & 1.012 \\ 0.475 & 1.012 & 1.513 \end{pmatrix}.$$

The variance of any contrast $\log(\hat{c}_r/\hat{c}_s)$ is remarkably little affected by the relative allocation frequencies n_r/n in the design. For example, if the relative frequencies are $f = (0.2, 0.3, 0.5)^\top$ we find

$$O_3 = \begin{pmatrix} 1.733 & 1.085 & 0.656 \\ 1.085 & 1.051 & 0.936 \\ 0.656 & 0.936 & 1.176 \end{pmatrix}, \qquad \mathcal{P}_3^\top \mathcal{I}^- \mathcal{P}_3 = \begin{pmatrix} 2.431 & 1.452 & 0.772 \\ 1.452 & 1.250 & 1.002 \\ 0.772 & 1.002 & 1.200 \end{pmatrix}.$$

The asymptotic variances of the contrasts $\big(\log(\hat{c}_1/\hat{c}_2), \log(\hat{c}_1/\hat{c}_3), \log(\hat{c}_2/\hat{c}_3)\big)$ are thus

$$(0.716, 2.138, 0.528)/n \quad \text{and} \quad (0.775, 2.086, 0.446)/n$$

for the first and second allocation respectively.

In the sense of minimizing the average variance of pairwise contrasts, the relative frequencies in the optimal allocation are approximately $(0.0, 0.8, 0.2)$, with no observations from P_1. But the average variance achieved by this allocation is only 8% less than the average variance in the design with equal weights. Further, the inferior design with equal weights may be superior for interpolation or extrapolation, i.e. for estimating ratios $\log(c_r/c_s)$ with r and/or s in $\{k+1, \ldots, k+m\}$.

The term *bridge sampling* has been used by Meng and Wong (1996) and Gelman and Meng (1998), in connection with the practice of sampling from intermediate distributions $P_2, \ldots, P_{k-1}$ in order to estimate the ratio c_1/c_k more accurately. Since the optimal allocation in the preceding example puts weight zero on P_1, the optimal bridge is in fact a cantilever, supported entirely on P_2, P_3. While the term 'bridge sampling' has a certain metaphoric appeal, this example indicates that it can be misleading to interpret the structural stability of the bridge as evidence of its statistical efficiency.

3.5. **Asymptotic covariance matrix from sub-models.** Since the estimating equation obtained from a sub-model is the same as (1.6) except that q_r is replaced by a group-average $\bar{q}_r$, $r = 1, \ldots, k+m$, the general covariance formula (3.4) is obviously applicable with the same replacement. In notation, this replacement is signified with $\bar{O}$ in place of O in (3.3) and other formulas as necessary. To see the potential gain in efficiency by a sub-model, consider a case where Γ is countable, μ is the counting measure, and $k = 2$. Let A be a finite subset of Γ, and let $\mathcal{G}_A$ be the permutation group on A. Then it is clear that averaging P_1 and P_2 over A

effectively makes both of them uniform on A. The asymptotic variance of $\log \gamma^{\mathcal{G}_A}$, where $\gamma^{\mathcal{G}_A} = \hat{c}_2^{\mathcal{G}_A}/\hat{c}_1^{\mathcal{G}_A}$, from (3.3) is (see Meng and Wong (1996) for a direct proof)

$$\mathrm{Var}(\log \hat{\gamma}^{\mathcal{G}_A}) = \frac{1}{nf_1f_2}(\bar{o}_{12}^{-1} - 1), \tag{3.10}$$

where $\bar{o}_{12}$ is the off-diagonal element of $\bar{O}_2$, that is,

$$\bar{o}_{12} \equiv \int_\Gamma \frac{d\bar{P}_1(x)}{d\bar{P}_{\mathrm{mix}}(x)} \frac{d\bar{P}_2(x)}{d\bar{P}_{\mathrm{mix}}(x)} \bar{P}_{\mathrm{mix}}(dx) \geq \int_A \frac{d\bar{P}_1(x)}{d\bar{P}_{\mathrm{mix}}(x)} \frac{d\bar{P}_2(x)}{d\bar{P}_{\mathrm{mix}}(x)} \bar{P}_{\mathrm{mix}}(dx). \tag{3.11}$$

Using the fact that $\bar{P}_1$ and $\bar{P}_2$ are uniform on A, (3.10)-(3.11) yields

$$\mathrm{Var}(\log \hat{\gamma}^{\mathcal{G}_A}) \leq \frac{1}{n_1}\frac{P_1(A^c)}{P_1(A)} + \frac{1}{n_2}\frac{P_2(A^c)}{P_2(A)}. \tag{3.12}$$

Consequently, the variance decreases with the increase of mass of A under both P_1 and P_2, and it approaches zero as A approaches Γ. This is not surprising because as A approaches Γ, the difficulty of summing over A, as required by the $\mathrm{ave}_{g\in\mathcal{G}_A}$ operator, approaches that of the original summation problem we try to avoid. Putting it differently, the choice and especially the size of A models what we consider to be usable information and computationally feasible. In implementing this sub-model estimator, we do not need to actually perform the permutation because for any $x \in A$, $\bar{P}_r(\{x\}) = \sum_{w\in A} P_r(\{w\})/|A|$, so the computation is linear in $|A|$, not in $|A|!$, which would be the case if we needed to actually permute.

As an illustration, let $\Gamma = \{0, 1, 2, \ldots, \}$, μ counting measure, $q_r(x) = (r\lambda)^x/x!$, $r = 1, 2$, and thus P_r is Poisson with mean $r\lambda$. Take $A = A_k = \{0, 1, \ldots, k\}$. Then $P_r(A_k^c) \leq (r\lambda)^{k+1}/(k+1)!$, and thus the right-hand side of (3.12) goes to zero rapidly as k goes to infinity. Figure 1 gives the relative variance of $\log(\hat{c}_1^{\mathcal{A}_k}/\hat{c}_2^{\mathcal{A}_k})$ verse the same estimator but without permutation (i.e., using $k = 0$), based on the asymptotic variance formula (3.10). It is seen that the size of the set of values of λ, the difference between the means, that show much improvement increases with k. This is expected as when k is suitably large compared to λ and 2λ, A_k will cover a substantial amount of mass under both P_1 and P_2, and thus the group averaging will be significantly better than the original un-permuted one. This also suggests that we can choose other A's, such as the union of two neighborhoods (not necessarily overlapping) of the mean/mode of each distribution, that may lead to even more efficient estimator with the same $|A|$. The key message here is that an effective strategy for increasing overlap between two distributions is to make both of them as close to uniform as possible; see Section 4.2 for more discussion.

4. Modeling with additional Information

4.1. **Parameterizing baseline measure.** An extreme form of additional information arises when we can parameterize the baseline measure μ up to a finite set of unknown parameters. Although this is of little practical interest (as it effectively assumes that we can perform integrations analytically or numerically once we are given the values of the unknown parameters), it provides a framework for examining the maximum possible gains by using additional information on μ. Since it is not possible to give a generic parametric model, we give two examples to illustrate two possible scenarios.

Example 1. Let $k = 2$, $\Gamma = \{0, 1, 2, \ldots, \}$, $q_r(x) = r^x$, and suppose that μ is known to be a distribution in the Poisson family, but with unknown mean λ. Direct

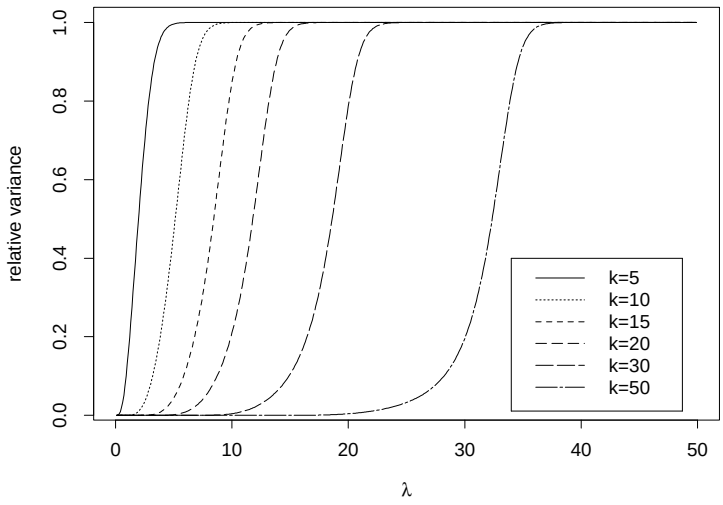

FIGURE 1. Variance of sub-model MLEs relative to that of MLE for the Poisson example of Section 3.5 (here k is the group size).

calculation shows that, for $r = 1, 2$, $c_r = e^{\lambda(r-1)}$, and that $\lambda_r = r\,\lambda$ is the mean of the distribution P_r. Thus $\xi \equiv \log(c_2/c_1) = \lambda_2 - \lambda_1 = \lambda$ may be estimated by the moment estimator $\hat{\xi}_{MNT} = \bar{X}_2 - \bar{X}_1$, or more efficiently by the MLE $\hat{\xi}_{MLE} = \sum_{i=1}^{n} X_i/(n_1 + 2n_2)$. The MLE is minimum-variance, unbiased and with variance

$$\mathrm{Var}(\hat{\xi}_{MLE}) = \frac{\lambda}{n}\frac{1}{f_1 + 2f_2}. \tag{4.1}$$

where $f_i = n_i/n$. Incidentally, the efficiency of $\hat{\xi}_{MNT}$ relative to $\hat{\xi}_{MLE}$ is $[9 + 2(\sqrt{f_2/f_1} - \sqrt{f_1/f_2})^2]^{-1}$, which does not exceed $1/9$, the value achieved when $f_1 = f_2 = 1/2$.

To see the loss of efficiency from not parameterizing μ, we compute the asymptotic variance of the semi-parametric MLE $\hat{\xi}_{SMLE}$ from Section 1, which is also the optimal bridge sampling estimator (Meng and Wong, 1996). This variance is given by (3.10) (with the original o_{12} in the place of $\bar{o}_{12}$), where for our current problem, o_{12} is a function of λ given by

$$o_{12}(\lambda) = \sum_{x=0}^{\infty} \frac{(2\lambda)^x e^{-2\lambda}}{x!(f_1 + f_2 2^x e^{-\lambda})}.$$

Figure 2 plots the asymptotic efficiency (i.e., reciprocal of variance) of $\hat{\xi}_{SMLE}$ relative to $\hat{\xi}_{MLE}$ as a function of λ, where $f_1 = f_2 = 1/2$. It is seen that the relative efficiency is always below one and it approaches zero as $\lambda \to \infty$. This is expected because as the difference in means, $\lambda = \lambda_2 - \lambda_1$, increases, $\mathrm{Var}(\hat{\xi}_{MLE})$ goes up linearly in λ as seen in (4.1), but $\mathrm{Var}(\hat{\xi}_{SMLE})$ goes up exponentially in λ. The latter can be seen by using the inequality (8.4) of Meng and Wong (1996, p. 850),

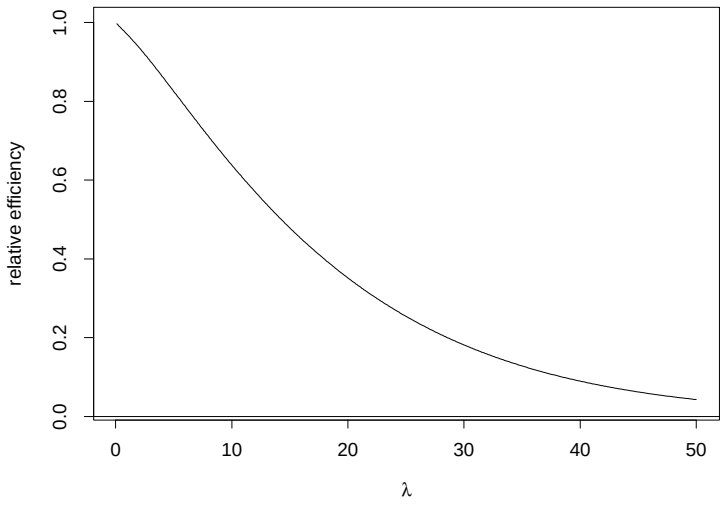

FIGURE 2. Relative efficiency of the semi-parametric MLE versus MLE for the Example 1 of Section 4.1.

which states that when $f_1 = f_2 = 1/2$,

$$H_{12}^2 \leq o_{12} \leq H_{12} \equiv \int_\Gamma \sqrt{\frac{dP_1(x)}{d\mu(x)}\frac{dP_2(x)}{d\mu(x)}}\mu(dx), \tag{4.2}$$

where H_{12} determines the Hellinger distance between P_1 and P_2: $\sqrt{2(1-H_{12})}$. Since for our current problem $H_{12} = e^{-(1.5-\sqrt{2})\lambda}$, we have

$$\frac{4}{n}\left(e^{(3-2\sqrt{2})\lambda} - 1\right) \geq \mathrm{Var}(\hat{\xi}_{SMLE}) \geq \frac{4}{n}\left(e^{(1.5-\sqrt{2})\lambda} - 1\right).$$

The phenomenon that the variance of the optimal bridge sampling estimator (when $k = 2$) goes up exponentially with the difference in means was also reported in Meng and Wong (1996), which suggests that it is crucial to increase the overlap between the P_1 and P_2, using methods such as those given in Meng and Schilling (2002), in order to improve the efficiency of bridge sampling estimators. It is also interesting to note that the parameterized MLE, $\hat{\xi}_{MLE}$, resembles the behavior of some estimators from *path sampling*, which is bridge sampling with infinitely many bridges, in the sense that the latter can also have variances that go up linearly in the difference in the means (Gelman and Meng, 1998).

We note that if we modify the example by setting $d\mu(x) = \zeta\lambda^x e^{-\lambda}/x!$ on the integers, for some unknown scalars $\zeta > 0$ and $\lambda > 0$, then μ is no longer a probability measure, $c_r = \zeta\, e^{\lambda(r-1)}$, but the ratios are unaffected. The value of ζ has no effect on the distribution of the observed data, and the likelihood function is unaffected. Thus ζ is not identifiable, the constants c_r are not identifiable, but the ratios c_r/c_s are identifiable and may be estimated by maximum likelihood as described above.

Example 2. Let $\Gamma = \mathbb{R}^+$, $q_r(x) = e^{-\beta_r x}$, where $0 \le \beta_1 < \beta_2 < \cdots < \beta_k$ are known, and $\mu(dx) = e^{-\rho x}dx$ for some unknown $\rho > 0$. Then $c_r = 1/(\beta_r + \rho)$, $r = 1, \ldots, k$, and P_r is exponential with mean c_r. The model is thus inverse linear (McCullagh and Nelder, 1989, chap. 2) with one unknown parameter ρ which can be estimated by maximum likelihood. The sample ratio, $\bar{X}_2/\bar{X}_1$ is a consistent, but not fully efficient, estimate of the ratio c_2/c_1. The MLE of ρ is $\hat{\rho} = \max(0, \tilde{\rho})$, where

$$\sum_{r=1}^{k} \frac{f_r}{\beta_r + \tilde{\rho}} = \bar{X},$$

which has a unique solution in $(-\beta_1, \infty)$. The Fisher information is

$$I(\rho) = n \sum_{r=1}^{k} \frac{f_r}{(\beta_r + \rho)^2} = n \sum_{r=1}^{k} f_r c_r^2. \tag{4.3}$$

To investigate the gain of efficiency by parameterizing, we consider the case of $k = 2$. From (4.3), the asymptotic variance of $\hat{\xi}_{MLE} = \log(\beta_2 + \hat{\rho}) - \log(\beta_1 + \hat{\rho})$ is

$$\mathrm{Var}(\hat{\xi}_{MLE}) = \frac{1}{n} \frac{(c_1 - c_2)^2}{f_1 c_1^2 + f_2 c_2^2}. \tag{4.4}$$

On the other hand, $H_{12} = 2\sqrt{c_1 c_2}/(c_1 + c_2)$, and thus by (3.10) and (4.2), when $f_1 = f_2 = 1/2$,

$$\frac{1}{n}\left[\frac{2(\sqrt{c_1} - \sqrt{c_2})^2}{\sqrt{c_1 c_2}}\right] \le \mathrm{Var}(\hat{\xi}_{SMLE}) \le \frac{1}{n}\left[\frac{(c_1 - c_2)^2}{c_1 c_2}\right].$$

Consequently, the asymptotic relative efficiency of $\hat{\xi}_{SMLE}$ is bounded by

$$\min\left\{\frac{(\sqrt{c_1} + \sqrt{c_2})^2 \sqrt{c_1 c_2}}{(c_1^2 + c_2^2)},\ 1\right\} \ge \frac{\mathrm{Var}(\hat{\xi}_{MLE})}{\mathrm{Var}(\hat{\xi}_{SMLE})} \ge \frac{2c_1 c_2}{c_1^2 + c_2^2}. \tag{4.5}$$

Consider the case where $\beta_1 = 1$ and $\beta_2 = \beta > 1$. Then the lower bound on the asymptotic efficiency in (4.5) has a minimum value $2\beta/(1+\beta^2)$, achieved when $\rho = 0$. Thus, unlike Example 1 where the asymptotic relative efficiency of $\hat{\xi}_{SMLE}$ can be arbitrarily small when the unknown parameter λ varies, in this example the gain in efficiency from using $\hat{\xi}_{MLE}$ may not be significant. For example, when $\beta = 2$, the asymptotic relative efficiency of $\hat{\xi}_{SMLE}$ is at least 80%, irrespective the value of ρ. This is due to substantial overlap between exponentials with mean $(1 + \rho)^{-1}$ and with mean $(2 + \rho)^{-1}$, regardless of the value of ρ. This again illustrates that the efficiency of $\hat{\xi}_{SMLE}$ is determined by the amount of overlap of the underlying distribution. Of course, when $\beta \to \infty$, the asymptotic efficiency of $\hat{\xi}_{SMLE}$ tends to zero by (4.5), because the exponential with mean $(\beta + \rho)^{-1}$ becomes concentrated at zero, and thus has little overlap with the exponential with mean $(1 + \rho)^{-1}$.

4.2. **Using label information.** As we derived in Section 3, the asymptotic covariance is determined by the design matrix F and the overlap measure matrix $O = \{o_{rs},\ 1 \le r, s \le k + m\}$, where

$$o_{rs} = \int_\Gamma \frac{dP_r(x)}{dP_{\mathrm{mix}}(x)} \frac{dP_s(x)}{dP_{\mathrm{mix}}(x)} P_{\mathrm{mix}}(dx). \tag{4.6}$$

Generally speaking, the more overlap as measured by O, the more accurate the MLE of $\log \hat{c}$, where $c = \{c_1, \ldots, c_{k+m}\}$. When Γ has certain topological structure, we can consider transformations to "warp" P_r's into similar shapes (and locations) in such

a way that the transformations do not alter the normalizing constants. For example, suppose the dominating measure μ is Lebesgue. Then we can consider transforming each $x_r \sim P_r$ via an one-to-one transformation $g_r(x)$. The desired normalizing constant c_r then is the normalizing constant of the transformed distribution, that is, (1.2) is replaced by

$$c_r = \int_\Gamma q_r(g_r^{-1}(x))J_r(x)\,d\mu, \tag{4.7}$$

where J_r is the Jacobian for g_r^{-1}. Since we know both g_r and J_r, (4.7) is simply (1.2) with q_r replaced by $\tilde{q}_r = q_r(g_r^{-1}(x))J_r(x)$, and thus we can proceed as before. However, with appropriate choices of g_r, the corresponding new $\tilde{P}_r$'s can have substantially more overlap than the original P_r's, as measured by O. Therefore, this *warping* technique can help greatly to reduce the Monte Carlo error. Note that we do not need to make new draws from $\tilde{P}_r$ as $\tilde{x}_i = g_{y_i}(x_i)$ is automatically from $\tilde{P}_{y_i}$ by construction. However, the label information is crucial for such a procedure, in contrast to the likelihood formulation given in Section 1, where we do not assume any additional knowledge (e.g., the topological structures of P_r's, including the support Γ).

Meng and Schilling (2002) provide extensive empirical evidence on the effectiveness of this warping strategy. They considered first, second, and third order warping transformations, which correspond to location shift, scale/rotation matching, and symmetrization. The first two orders of warping can be summarized by an affine transformation $g_r(x) = S_r(x - m_r)$, where m_r and S_r can be (i) estimated from the draws from P_r (e.g., sample mean and precision) or (ii) determined analytically from the known q_r (e.g., its mode and the square root of the negative Hessian matrix at the mode). It is evident that (ii) can provide substantially more efficient estimators than (i), at the cost of possibly considerable analytical calculation when the dimension of Γ is high. In contrast, (i) is typically trivial to implement, and as demonstrated in Meng and Schilling (2002), it typically still provides a far more efficient estimator than not using any warping transformation. In particular, using the sample precision matrix in addition to just location shift, which was originally proposed in the physics literature by Voter (1985), can provide substantial additional reduction of MC variance.

For the third order warping, Meng and Schilling (2002) suggested using mixtures of a density with its various reflections to eliminate skewness. This is mathematically equivalent to group averaging when the reflections used in the mixture form a group. Meng and Schilling (2002) also demonstrated that it can achieve dramatic reduction of variance by combining all three orders of warping, namely, to first shift all distributions to a common origin, then reflect with this origin, and then standardize each of the reflected ones to have common covariance matrix. They gave an example where after these warp transformations several rather skewed and very different distributions were effectively transformed to similar multivariate normals, and thus the gain in efficiency was several orders of magnitude.

An unresolved problem with these label-specific transformations is that we currently do not have a model-based way to choose them. In principle, it should be possible to formally include, say, the affine transformation parameters $\{m_r, S_r\}$ in the semi-parametric model and then estimate them by MLE, instead of estimating them by moment methods such as sample mean and precision. The difficult is that

in order to let the model properly estimate $\{m_r, S_r\}$ such that the warped distributions will be close to each other, we need to build such a requirement into the model. While we can estimate the distributional summaries of each P_r using the draws from it, the likelihood function based on these data do not contain direct information on how to transform these sampling distributions together. This appears to be another interesting and challenging problem in modeling our inability, namely the inability to analytically maximize overlap, as measured by O, among the underlying distributions.

5. Using the Profile Likelihood Approach

5.1. **Profiling the empirical likelihood.** In this section, we show that the likelihood (1.4) can be partially maximized to produce the same results as given in Section 2 and Section 3. This empirical likelihood approach not only yields a profile likelihood for c, but also provides another explanation for why the "retrospective likelihood" for c studied in Geyer (1994) is only first-order correct, as demonstrated before (Kong et al., 2003, Section 6).

In the empirical likelihood approach, we treat both $\theta = \{\theta_1, \ldots, \theta_n\}$, where $\theta_i = \theta(x_i)$, and $c = \{c_1, \ldots, c_k\}$ as parameters. Because of (1.4), the profile likelihood of c is defined by

$$l(c) = \max_{\theta \in \Theta(c)} \left(\sum_{i=1}^{n} \theta_i - \sum_{s=1}^{k} n_s \log c_s \right), \tag{5.1}$$

where

$$\Theta(c) = \{\theta : \sum_{i=1}^{n} e^{\theta_i} q_r(x_i) c_r^{-1} \equiv \sum_{i=1}^{n} P_r(\{x_i\}) = 1, \quad r = 1, \ldots, k.\}. \tag{5.2}$$

The equality constraint in (5.2) is motivated by the discussion given in Section 1.3, where it is shown that in maximum likelihood calculations, we can restrict ourselves to measures with support on $\{x_1, , \ldots, x_n\}$. We also note that the summation here is over the entire sample because, as noted in Section 1.2, the labels of the observations are not part of the minimal sufficient statistic.

Before we proceed, we need to set conditions to guarantee that $\Theta(c)$ is not empty. To do so, we let

$$\mathcal{W}(c) = \{W = (w_1, \ldots, w_k) : \sum_{s=1}^{k} w_s = 1 \quad \text{and} \quad p_{\text{mix}}(x_i|W) > 0, \forall i = 1, \ldots, n\}, \tag{5.3}$$

where $p_{\text{mix}}(x|W) \equiv \sum_{s=1}^{k} w_s p_s(x)$, with $p_s = q_s c_s^{-1}$ denoting the normalized density. Note that (5.3) does not require nonnegative w_i's, but only that $p_{\text{mix}}(x_i|W)$ is positive for all i, i.e. we allow negative "weights", as long as the corresponding mixture density is positive. Clearly $\mathcal{W}(c)$ is non-empty because $(k^{-1}, \ldots, k^{-1}) \in \mathcal{W}(c)$ under our sample design. Furthermore, it is trivial to verify that $\mathcal{W}(c)$ is a convex set because $p_{\text{mix}}(x|W)$ is linear in the components of W.

Intuitively, because any two densities with respect to the same dominating measure cannot dominate each other in order to integrate to one, we need the following necessary condition for $\Theta(c)$ to be non-empty:

No Dominance Condition *There does not exist a* $W \in \mathcal{W}(c)$ *and a* $1 \leq t \leq k$ *such that*

$$p_{\text{mix}}(x_i|W) \geq p_t(x_i), \quad \text{for all } i = 1, \ldots, n, \tag{5.4}$$

and where the inequality is strict for at least one i.

It is interesting that this intuitive necessary condition turns out to be also sufficient, as seen from the following theorem, proved in the Appendix. The $\mathcal{A}(c)$ set used in the theorem is the collection of all $(a_1, \ldots, a_k)$ such that $\sum_{s=1}^{k} a_s = 0$ and $p_{\text{mix}}(x_i|A) \geq 0$ for all $i = 1, \ldots, n$ with the inequality being strict for at least one i. Note that the conditions (I), (IV) and (V) were considered before (Tan, 2004), but (II) and (III) appear to be new.

Theorem 5.1. *Assuming $Q_{n\times k} = \{q_j(x_i)\}$ is of rank k, the following five conditions are equivalent:*

(I): *$\Theta(c)$ is non-empty;*
(II): *The "No Dominance Condition" holds;*
(III): *$\mathcal{A}(c)$ is empty;*
(IV): *$\mathcal{W}(c)$ is a bounded convex set;*
(V): *The equations*

$$\sum_{i=1}^{n} \frac{p_r(x_i)}{p_{\text{mix}}(x_i|W)} = n, \quad \text{for all} \quad r = 1, \ldots, k, \tag{5.5}$$

have a unique solution in the interior of $\mathcal{W}(c)$.

We remark here that while conditions (III) and (IV) are geometrically appealing, condition (V) seems to be most convenient for practical purposes, because we can check numerically the existence of the solution to (5.5). We also remark that when $c_r \propto \hat{c}_r$, where $\{\hat{c}_r, r = 1, \ldots, k\}$ is the solution of (1.6), $W = (f_1, \ldots, f_k)$ satisfies (5.5), where $f_r = n_r/n$. Therefore, we know that there exists at least one c (and all its multipliers) such that $\Theta(c)$ is not empty.

Nevertheless, for the case of $k = 2$, the bounds given by (IV) can be established explicitly because condition (II) implies that the following two sets

$$N_1 = \{i : p_1(x_i) - p_2(x_i) < 0\} \quad \text{and} \quad N_2 = \{i : p_1(x_i) - p_2(x_i) > 0\}$$

are non-empty. Consequently, it is easy to verify directly that $\mathcal{W}(c)$ consists of all (w_1, w_2) such that $w_1 + w_2 = 1$ and

$$\max_{i\in N_2} \left\{ \frac{-p_2(x_i)}{p_1(x_i) - p_2(x_i)} \right\} < w_1 < \min_{i\in N_1} \left\{ \frac{-p_2(x_i)}{p_1(x_i) - p_2(x_i)} \right\}.$$

5.2. **The computation of the profile likelihood.** Equipped with Theorem 5.1, we can now proceed to find a computable expression for the profile likelihood $l(c)$ of (5.1). We first observe that the constraint in (5.2) is equivalent to

$$\sum_{s=1}^{k} w_s \left(\sum_{i=1}^{n} e^{\theta_i} q_s(x_i) c_s^{-1} \right) \equiv \sum_{i=1}^{n} e^{\theta_i} p_{\text{mix}}(x_i|W) = 1, \tag{5.6}$$

for any $W \in \mathcal{W}(c)$. Taking logarithms on both sides of (5.6), and then applying Jensen's inequality to the log function, we obtain that

$$\sum_{i=1}^{n} \theta_i \leq -\sum_{i=1}^{n} [\log p_{\text{mix}}(x_i|W) + \log n], \tag{5.7}$$

for any $\theta \in \Theta(c)$ and $W \in \mathcal{W}(c)$. This implies that

$$\max_{\theta\in\Theta(c)} \sum_{i=1}^{n} \theta_i \leq -\max_{W\in\mathcal{W}(c)} \sum_{i=1}^{n} [\log p_{\text{mix}}(x_i|W) + \log n]. \tag{5.8}$$

We now show that the above inequality is actually an equality, and hence the maximization needed to compute the profile likelihood $l(c)$ is equivalent to maximizing the log-likelihood for the unknown weights W under the mixture model as given by $p_{\text{mix}}(x|W)$, with $\mathcal{W}(c)$ as the parameter space. This happens because equation (5.5) is just the normal equation for the maximum likelihood estimate (MLE) of W under this mixture model. Let us denote with $W(c)$ the MLE under this mixture model (recall we assume condition (V) here), namely the unique solution of (5.5), and let

$$\theta_i(c) = -[\log p_{\text{mix}}(x_i|W(c)) + \log n]. \tag{5.9}$$

Then it is clear that this choice of θ makes (5.8) equality. Furthermore, (5.5) implies $\theta(c) \in \Theta(c)$. Consequently, $\theta(c)$ is the maximizer in (5.1), and therefore

$$l(c) = -\sum_{i=1}^{n} \log p_{\text{mix}}(x_i|W(c)) - n \log n - \sum_{s=1}^{k} n_s \log c_s. \tag{5.10}$$

Note that because $W(c)$ is the solution of (5.5), $W(c) \in \mathcal{W}(c)$ and c must satisfy

$$T_r(W(c), c) \equiv \sum_{i=1}^{n} \frac{q_r(x_i)c_r^{-1}}{\sum_{s=1}^{k} w_s(c)q_s(x_i)c_s^{-1}} = n, \quad r = 1, \ldots, k. \tag{5.11}$$

We remark here that the above derivation is similar to the maximization approach used for finding MLE with control variates, as investigated in Tan (2003a, 2004) and Meng (2005). A key advantage of (5.10) is that it provides a direct "marginal likelihood" for c, which can be treated as a likelihood to be used in Bayesian inference for c when we have reliable prior information on it; see Section 6.1.

5.3. **Computing the MLE and the observed Fisher information.** To maximize $l(c)$, we first identify its stationary point(s). We therefore calculate

$$\frac{\partial l(c)}{\partial \log c_r} = c_r \left[-\sum_{i=1}^{n} \frac{\sum_{s=1}^{k} p_s(x_i) \frac{\partial w_s(c)}{\partial c_r} - p_r(x_i)c_r^{-1}w_r(c)}{\sum_{s=1}^{k} p_s(x_i)w_s(c)} - \frac{n_r}{c_r} \right] = nw_r(c) - n_r, \tag{5.12}$$

where the last equality is due to (5.11) and $\sum_{r=1}^{k} w_r(c) = 1$. Consequently, any stationary point c must satisfy $w_r(c) = n_r/n = f_r, r = 1, \ldots, k$. By the uniqueness of the solution $W(c)$ for (5.5), we can conclude that c must be the solution of $T_r(f, c) = n$ for all $r = 1, \ldots, k$, which is exactly (1.6) for $r = 1, \ldots, k$. Under the "connectivity" assumption of Vardi (1985), which we always assume, (1.6) has a unique solution (up to a multiplicative constant), which is the MLE of c under the likelihood (1.4). Since $l(c)$ is the profiled likelihood derived from (1.4), it is clear that the very same $\hat{c}$ also maximizes $l(c)$ in (5.10), as it should.

It is interesting to observe that if we let $W = f$ in (5.7) and (5.9), but *without* realizing that the resulting θ from (5.9) may not satisfy (5.2), we would have arrived at a "profile" log-likelihood (5.10) with $W = f$, as in Geyer (1994). This would be exactly the wrong log-likelihood obtained by the retrospective argument – see (6.1) of Kong et al. (2003). In other words, the wrong likelihood is the same as the incorrectly "profiled" likelihood without realizing the strong compatibility requirement between W and c, as in (5.11). However, because of (5.12), this incorrectly "profiled" log-likelihood does provide the correct MLE as it coincides with the correct profile likelihood when $c = \hat{c}$ from (1.6) since $W(\hat{c}) = f$.

The incorrect "profile" likelihood does not provide the correct second order inference, as demonstrated in Kong et al. (2003). The correct asymptotic covariance for $\log \hat{c}$ can be estimated by the inverse of the observed Fisher information from the profile likelihood (5.10) (Murphy and Van Der Vaart, 1999), namely,

$$\hat{\mathcal{I}} = -\left[\frac{\partial^2 l(c)}{\partial \log c (\log c)^\top}\right]\Bigg|_{c=\hat{c}} = -n \left.\frac{\partial W(c)}{\partial \log c}\right|_{c=\hat{c}}, \tag{5.13}$$

where the last equality is due to (5.12). By (5.11), we have

$$\left[\frac{\partial T(W,c)}{\partial W}\right]\left[\frac{\partial W(c)}{\partial \log c}\right] = -\frac{\partial T(W,c)}{\partial \log c}, \tag{5.14}$$

where $T = (T_1, \ldots, T_k)^\top$. Using $W(\hat{c}) = f$, it is easy to check that for any $r, s \in \{1, \ldots, k\}$,

$$\left.\frac{\partial T_r}{\partial w_s}\right|_{c=\hat{c}} = -\hat{o}_{rs} \quad \text{and} \quad \left.\frac{\partial T_r}{\partial \log c_s}\right|_{c=\hat{c}} = \hat{o}_{rs} f_s - \delta_{\{r=s\}}, \tag{5.15}$$

where

$$\hat{o}_{rs} = \int_\Gamma \left[\frac{dP_r(x)}{dP_{\text{mix}}(x)}\right]\left[\frac{dP_s(x)}{dP_{\text{mix}}(x)}\right] \hat{P}_{\text{mix}}(dx),$$

which is the sample version of (4.6). It follows from (5.13)-(5.15) that $\hat{O}_k \hat{\mathcal{I}}/n = I - \hat{O}_k F$, which implies

$$n\hat{\mathcal{I}}^- = (I - \hat{O}_k F)^+ \hat{O}_k, \tag{5.16}$$

which is equivalent to (3.3) and (3.9), except with $\hat{O}_k$ estimating O_k.

6. A PARADOX AND SOME FUTURE WORK

6.1. **A Bayesian paradox?** Given the success of likelihood based methods, it is natural to ask the question, "what about Bayesian methods?" Indeed, it seems so obvious that Bayesian methods should be particularly useful for dealing with simulated data, since the usual dispute of the correctness of prior information no longer exists. For example, by mathematical inequalities, such as Jensen's inequality or Cauchy-Schwartz, we may know for certain that a normalizing constant c is between two known values, a and b. Surely such prior information can and should be used in our MC integration. But how? Because the model parameter is the baseline measure μ, to put a prior on μ that respects $a \le c = \int_\Gamma q(x)\mu(dx) \le b$ would require similar or even greater analytic effort than what is needed to calculate c analytically. In other words, in order to carry out the Bayesian method, we need more effort than what is needed to solve the original problem.

One, of course, could try to use the profile likelihood as given in (5.10) to conduct Bayesian inference. This is certainly a topic worth investigating, particularly with respect to the question of trading computational efficiency with statistical efficiency because the computation of (5.10) is not cost-free (but at least it is numerically feasible). Nevertheless, from a philosophical point of view, profile likelihood is not an legitimate Bayesian approach, which finds marginal likelihood via integration, not maximization/profiling. Therefore, statistical inference for MC integration appears to be an ultimate paradox for Bayesian inference, because it appears that Bayesian methods can solve (at least in theory) every other inference problem except for their own computational problems (as Bayesian methods rely heavily on MC integration for implementation). Or as Kong et al. (2003) put it "This computational black

hole, an infinite regress of progressively more complicated models, is an unappealing prospect, to say the least."

6.2. **Some future work.** Two of the most important challenges are to extend the models to include the estimation of label-specific transformations, and to use effectively information on dependence structure (e.g., auto-correlation), as in an MCMC setting. There has been no progress regarding the first, but the empirical evidence provided in Meng and Schilling (2002) from the use of "warp transformation" methods suggest that the gain of efficiency by appropriately modeling the label-specific transformations can be substantial. As for the second, several papers (Tan, 2003b,c, 2004) show great promise. In particularly, the use of the kernel functions under the likelihood modeling in a Gibbs sampling setting, or more generally with Metropolis-Hastings algorithms can lead to estimates of normalizing constants with a n^{-1} rate of convergence, instead of the usual $n^{-1/2}$ rate (Kong et al., 2003; Tan, 2003a).

Both of these extensions can therefore have substantial practical consequences, as well as provide theoretical insight into how we should balance analytical, numerical, and simulation efforts in effective use of MC methods. As discussed before (van Dyk and Meng, 2001, rejoinder), model selection with simulated data has a different goal than with real data, because the key question is not which model is approximately true — all models that can link the simulated data to our estimand are known and true. The goal is rather to select a model that provides an effective compromise between computational complexity, human effort, and statistical efficiency. Our sub-modeling via group averaging was guided by this goal, and we look forward to further explorations of this method and to development of other methods that will help to achieve the same goal.

Appendix: Proof of Theorem 5.1

(I)⇒ (II): We prove by contradiction. Suppose (II) is false. Then there exists a $W \in \mathcal{W}(c)$ and t that satisfy (5.4). It follows that for any $\theta \in \Theta(c)$, we will reach the following contradiction:

$$1 = \sum_{i=1}^{n} e^{\theta_i} p_{\text{mix}}(x_i|W) > \sum_{i=1}^{n} e^{\theta_i} p_t(x_i) = 1,$$

where the first and last equality are due to $\theta \in \Theta(c)$ and $W \in \mathcal{W}(c)$, and the inequality is a consequence of (5.4).

(II)⇒ (III): If there exists an $A \in \mathcal{A}(c)$, then it is easy to see that $A + e_1 \in \mathcal{W}(c)$, where $e_1 = (1, 0, \ldots, 0)$, and that (5.4) is satisfied for this element of $\mathcal{W}(c)$, with the inequality being strict for at least one i. This contradicts assumption (II).

(III)⇒ (IV): We again prove by contradiction. Suppose there exists a sequence $W^{(m)} \in \mathcal{W}(c)$ such that it is unbounded as $m \to \infty$. For any m, let r_m be the index such that $|W^{(m)}_{r_m}| = \max\{|W^{(m)}_r|, r = 1, \ldots, k\}$. Because r_m can only take k values, as $m \to \infty$, there is a subsequence such that r_m takes the same value, say, $r_m = 1$.

Along this subsequence $|W^{(m)}_1| \to \infty$, and $a^{(m)}_r \equiv W^{(m)}_r / |W^{(m)}_1| \in [-1, 1]$ for any $r \geq 1$. Therefore we can choose a subsequence such that

the limit of $a_r^{(m)}$ exists for all r. Denote this limit by $a = (a_1, \ldots, a_k)$. Then it is clear from $\sum_r W_r^{(m)} = 1$ that

$$\sum_r a_r = \lim_{m\to\infty} \sum_r \frac{W_r^{(m)}}{|W_1^{(m)}|} = \lim_{m\to\infty} \frac{1}{|W_1^{(m)}|} = 0. \qquad (A.1)$$

Furthermore, from $p_{\text{mix}}(x_i|W^{(m)}) > 0$, we obtain that

$$p_{\text{mix}}(x_i|A) = \lim_{m\to\infty} \frac{p_{\text{mix}}(x_i|W)}{|W_1^{(m)}|} \geq 0, \quad i = 1, \ldots, n. \qquad (A.2)$$

If we can prove that at least one inequality in (A.2) is strict, then (A.1) and (A.2) allow us to conclude that $a \in \mathcal{A}(c)$, which contradicts assumption (III). To prove this, suppose $p_{\text{mix}}(x_i|A) = 0$ for all $i = 1, \ldots, n$. It follows that $P_{n\times k} a^\top = 0$, where $P_{n\times k} = \{p_j(x_i)\}$ is the $n \times k$ matrix of the normalized density values. Since $P_{n\times k} = Q_{n\times k} \operatorname{diag}\{c_1^{-1}, \ldots, c_k^{-1}\}$ and therefore it is of full rank k under our assumption that $Q_{n\times k}$ is of full rank k, we can conclude that $a = 0$. But this is impossible because $|a_1| = 1$ by our construction.

(IV) $\Rightarrow$ (V): Because $\mathcal{W}(c)$ is bounded, all its boundaries are determined by W such that $p_{\text{mix}}(x_i|W) = 0$. Therefore,

$$f(W) = \sum_{i=1}^n \log p_{\text{mix}}(x_i|W),$$

which is a concave function on $\mathcal{W}(c)$, must be maximized at an interior point of $\mathcal{W}(c)$ because at any of these boundaries $f(W) = -\infty$. Since this interior point, labeled by $W(c)$, must be a stationary point, by the method of Lagrange multiplier, it must satisfy (5.5). To prove this solution is unique (and hence $f(W)$ has only one stationary point, the global maximizer), let us suppose there are two solutions, W_1 and W_2, that satisfy (5.5). This implies that, by summing up the left-hand side of (5.5) with respect to either weight and then summing up the two sums,

$$\sum_{i=1}^n \left[\frac{p_{\text{mix}}(x_i|W_1)}{p_{\text{mix}}(x_i|W_2)} + \frac{p_{\text{mix}}(x_i|W_2)}{p_{\text{mix}}(x_i|W_1)} - 2 \right] = 0. \qquad (A.2)$$

Using the fact that $a + a^{-1} \geq 2$ for any $a > 0$, where the equality holds if and only if $a = 1$, we can conclude from (A.2) that $p_{\text{mix}}(x_i|W_1) = p_{\text{mix}}(x_i|W_2)$ for all $i = 1, \ldots, n$, namely $P_{n\times k}(W_1 - W_2) = 0$. It follows immediately that $W_1 = W_2$ because $P_{n\times k}$ is of rank k.

(V) $\Rightarrow$ (I): Let $W(c)$ be the unique solution of (5.5), and define $\theta(c)$ as in (5.9). Then this $\theta(c)$ satisfies the constraints required by (5.2) and hence $\Theta(c)$ must be non-empty.

References

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

[2] Ceperley, D. M. (1995). Path integrals in the theory of condensed helium. *Reviews of Modern Physics* **67** 279–355.

[3] CHEN, M. H. and SHAO, Q. M. (1997a). Estimating ratios of normalizing constants for densities with different dimensions. *Statistica Sinica* **7** 607–630.

[4] CHEN, M. H. and SHAO, Q. M. (1997b). On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics* **25** 1563–1594.

[5] CHIB, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90** 1313–1321.

[6] CHIB, S. and JELIAZKOV, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* **96** 270–281.

[7] 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.

[8] GELFAND, A. E. and DEY, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56** 501–514.

[9] GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* **13** 163–185.

[10] GEYER, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Tech. Rep. 568, School of Statistics, University of Minnesota.

[11] GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society B* **54** 657–699.

[12] GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J. (1996). *Markov chain Monte Carlo in Practice*. Chapman & Hall, London.

[13] IRWIN, M., COX, N. J. and KONG, A. (1994). Sequential imputation for multilocus linkage analysis. *Proc. Natl. Acad. Sci. USA* **91** 11684–11688.

[14] JENSEN, C. S. and KONG, A. (1999). Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. *American Journal of Human Genetics* 885–901.

[15] JOHNSON, V. (1999). Posterior distributions on normalizing constants. Tech. rep., Duke University.

[16] KONG, A., MCCULLAGH, P., MENG, X.-L., NICOLAE, D. and TAN, Z. (2003). A statistical theory for Monte Carlo integration (with discussion). *Journal of the Royal Statistical Society, Series B* **65** 585–618.

[17] MACEACHERN, S. and PERUGGIA, M. (2000). Importance link function estimation for Markov chain Monte Carlo methods. *Journal of Computational and Graphical Statistics* **9** 99–121.

[18] MCCULLAGH, P. (1999). Quotient spaces and statistical models. *Canadian Journal of Statistics* **27** 447–456.

[19] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized linear models (Second edition)*. Chapman & Hall, London.

[20] MENG, X.-L. (2005). Discussion: Computation, survey, and inference (discussion of "Qausi Monte Carlo and control variates" by Hickernell, Lemieux, and Owen). *Statistical Science* .

[21] MENG, X.-L. and SCHILLING, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association* **91** 1254–1267.

[22] MENG, X.-L. and SCHILLING, S. (2002). Warp bridge sampling. *The Journal of Computational and Graphical Statistics* **11** 552–586.

[23] MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical explanation. *Statistica Sinica* **6** 831–860.

[24] MURPHY, S. A. and VAN DER VAART, A. W. (1999). Observed information in semi-parametric models. *Bernoulli* **5** 381–412.

[25] NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference and the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B* **56** 3–48.

[26] OTT, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *American Journal of Human Genetics* **31** 161–175.

[27] STEPHENS, M. and DONNELLY, P. (2001). Inference in molecular population genetics (with discussion). *Journal of the Royal Statistical Society B* **62** 605–655.

[28] TAN, Z. (2003a). *A likelihood approach for Monte Carlo integration.* Ph.D. thesis, The University of Chicago, Dept. of Statistics.

[29] TAN, Z. (2003b). Monte Carlo integration with Markov chain. Tech. Rep. 20, Johns Hopkins Biostatistics.

[30] TAN, Z. (2003c). Monte Carlo integration with acceptance-rejection. Tech. Rep. 21, Johns Hopkins Biostatistics.

[31] TAN, Z. (2004). On a likelihood approach for Monte Carlo integration. *Journal of the American Statistical Association* **99** 1027–1036.

[32] THOMPSON, E. A. (2000). *Statistical Inference from Genetic Data.* NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, California.

[33] VAN DYK, D. A. and MENG, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics (with discussion)* **10** 1–111.

[34] VARDI, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13** 178–203.

[35] 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.

[36] VOTER, A. F. (1985). A Monte Carlo method for determining free-energy differences and transition state theory rate constants. *J. Chem. Phys.* **82(4)** 1890–1899.

A. KONG
DECODE GENETICS
REYKJAVIK, ICELAND

P. MCCULLAGH AND D. NICOLAE
DEPARTMENT OF STATISTICS
THE UNIVERSITY OF CHICAGO
CHICAGO, IL 60637

X.L. MENG
DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
CAMBRIDGE, MA 02138-2901