The Annals of Statistics
1997, Vol. 25, No. 6, 2410–2432

THE PROPERTIES OF THE CROSS-MATCH ESTIMATE AND SPLIT SAMPLING

BY AUGUSTINE KONG,[1] JUN S. LIU[2] AND WING HUNG WONG[3]

University of Chicago, Stanford University and University of California

By noting the connection with k-sample U-statistics, we find a simple decomposition of the variance of the *cross-match* estimate, which can be regarded as a generalization of Efron and Stein. We apply the decomposition in assessing efficiencies of several plans of using the weighted samples from an importance scheme. The applications of the formula to multiple imputations lead to a method of crossing jointly imputed data to gain more accuracy.

1. Introduction. This paper introduces the *cross-match estimate* and studies its properties with respect to a Monte Carlo simulation scheme called *split sampling*. Let $\mathbf{z}$ be a random vector and of interest is the expectation

$$\mu = E_p[\psi(\mathbf{z})]$$

for some distribution $p(\mathbf{z})$ and some function $\psi(\mathbf{z})$. A standard Monte Carlo procedure for estimating μ is to draw iid samples $\mathbf{z}_1, \ldots, \mathbf{z}_m$ from some distribution $f(\mathbf{z})$, f not necessarily equal to p. This is referred to as the *joint-sampling scheme*. An importance sampling estimate of μ using renormalized weights is

$$\tilde{\mu} = \frac{\sum_{j=1}^{m} w(j)\psi(\mathbf{z}_j)}{\sum_{j=1}^{m} w(j)}, \tag{1.1}$$

where $w(j) = p(\mathbf{z}_j)/f(\mathbf{z}_j)$. If $f(\mathbf{z})$ is not equal to $p(\mathbf{z})$, possible reasons are (a) variance reduction, (b) it is difficult to sample from p directly, (c) samples are originally drawn from $f(\mathbf{z})$, but the current interest is in expectations taken under p. The results in this paper are more relevant for the latter two cases.

Let $\mathbf{z} = (z_1, \ldots, z_k)$ be a decomposition where each component z_i can either be a vector or a scalar. Let $f_i(z_i)$, $i = 1, \ldots, k$, be the marginal distribution of z_i with respect to $f(\mathbf{z})$. With split sampling, m_i iid draws are taken from each of these k marginals independently. The draws are denoted

Received March 1996; revised May 1997.
[1] Supported in part by NIH Grant RO1-GM4600.
[2] Supported in part by NSF Grants DMS-94-04344, DMS-95-01570 and Terman Fellowship from Stanford.
[3] Supported in part by NSF Grant DMS-92-04504.
AMS 1991 *subject classification*. 62G05.
Key words and phrases. Importance sampling, linkage study, multiple imputation, orthogonal representation.

by z_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, m_i$. We will use $f^*(\mathbf{z})$ to denote the product measure $f_1(z_1) \cdots f_k(z_k)$. The *cross-match estimate* of μ using renormalized importance sampling weights is defined as

$$\hat{\mu}^{\otimes} = \frac{\sum_{j_1, \ldots, j_k} w(j_1, \ldots, j_k) \psi(z_{1j_1}, \ldots, z_{kj_k})}{\sum_{j_1, \ldots, j_k} w(j_1, \ldots, j_k)}, \tag{1.2}$$

where

$$w(j_1, \ldots, j_k) = \frac{p(z_{1j_1}, \ldots, z_{kj_k})}{f_1(z_{1j_1}) \cdots f_k(z_{kj_k})} = \frac{p(z_{1j_1}, \ldots, z_{kj_k})}{f^*(z_{1j_1}, \ldots, z_{kj_k})} \tag{1.3}$$

is the importance sampling weight. The purpose of this paper is to investigate the behavior of $\hat{\mu}^{\otimes}$ relative to $\tilde{\mu}$. The results are then qualitatively extrapolated to another cross-match estimate that is constructed with respect to joint sampling.

It is emphasized that the material in this paper is developed under the premise that it is very expensive to draw samples from p, f or f^*. Hence it is desirable to get the most out of the samples generated. A special situation that will be treated in Section 6 is the analysis of multiply imputed complete data sets [Rubin (1987)]. Quite often the imputations were created by the Census Bureau and it is impossible for a user to draw additional samples.

Another application which motivated our research is genetic linkage analysis. Here investigators collect data from k independent pedigrees and z_i corresponds to a statistic which can be computed from the data on pedigree i. The null hypothesis to be tested is that a disease gene is not on a certain chromosome. In parametric analysis, z_i is a log-likelihood ratio, and in nonparametric analysis, z_i is a "Z-score." For calculations of exact p-values and power, one is interested in tail probabilities of the form $P(z_1 + z_2 + \cdots + z_k \geq t)$, which can be written as $E[\psi(\mathbf{z})]$ where $\psi(\mathbf{z})$ is the indicator function $I_{\{z_1 + \cdots + z_k \geq t\}}$. Depending on the pedigree structures and the missing data patterns, the simulation of the pedigree data and the calculation of z_i from the simulated data can be extremely computationally intensive. Hence one can typically afford to simulate only a few values of z_i for each i. The importance sampling aspect is also relevant here. For example, f may correspond to the distribution of $\mathbf{z}$ under some specific alternative hypothesis, while p can be the distribution of $\mathbf{z}$ under the null hypothesis or some other alternative hypothesis. Both f and p are product measures. A single simulation can be used for p-value calculation and power calculation under many different alternatives. Interestingly, the cross-matching idea has been used in practice frequently with an additional twist. For example, for power calculations, Mahtani et al. (1996) simulated $m = 10$ samples each from their $k = 26$ diabetic families. So formally the cross-match estimate involved 10^{26} terms, making its computation infeasible. Instead they used a resampling scheme which draws randomly from the 10^{26} possible combinations and got a Monte Carlo estimate of $\hat{\mu}^{\otimes}$. Following Terwilliger and Ott (1992), they called this resampling scheme "bootstrap," which deviates somewhat from the

standard usage of this term in the statistics literature. More about applications of cross-matching and importance sampling in linkage problems can be found in Kong (1992). More details on tail probability estimation and resampling are given respectively in Example 3.2 and Section 7.

Consider the following three cases which are listed in order of increasing generality.

CASE I. Independent sampling: $p(\mathbf{z}) = f(\mathbf{z}) = f^*(\mathbf{z})$.

CASE II. Independent importance sampling: $f(\mathbf{z}) = f^*(\mathbf{z})$, but $p(\mathbf{z})$ not necessarily equal to $f(\mathbf{z})$.

CASE III. The general case (no independence and importance sampling).

Note that with Cases I and II, if $m_1 = m_2 = \cdots = m_k = m$, split sampling is identical to joint-sampling. Hence both $\tilde{\mu}$ and and $\hat{\mu}^{\otimes}$ can be computed from the same sample. The difference is that $\hat{\mu}^{\otimes}$ uses all m^k combinations of the component samples. With Case I,

$$\tilde{\mu} = \frac{1}{m}\sum_{j=1}^{m}\psi(\mathbf{z}_j) = \frac{1}{m}\sum_{j=1}^{m}\psi(z_{1j},\ldots,z_{kj}) \tag{1.4}$$

and

$$\hat{\mu}^{\otimes} = \frac{1}{m^k}\sum_{j_1,\ldots,j_k}\psi(z_{1j_1},\ldots,z_{kj_k}) \tag{1.5}$$

as all the weights are 1. Here both estimates are obviously unbiased. Noting that $\hat{\mu}^{\otimes}$ is a special case of k-sample U-statistics [a terminology used by Koroljuk and Borovskich (1994); also called generalized U-statistics in Serfling (1980)], we explain in Section 2 that the variance of $\hat{\mu}^{\otimes}$ is always smaller than that of $\tilde{\mu}$. For Case II, we obtain the weaker result that, asymptotically ($m \to \infty$), $\hat{\mu}^{\otimes}$ has smaller mean-squared error than $\tilde{\mu}$.

Case I is studied in detail in Section 2. We derive a simple expression for the variance decomposition of $\hat{\mu}^{\otimes}$, which is different from the one in Koroljuk and Borovskich (1994, Section 1.2) but is a generalization of Efron and Stein (1981). As a consequence, the relative efficiency of $\hat{\mu}^{\otimes}$ over $\tilde{\mu}$ is obtained. Two examples are discussed in Section 3. Cases II and III are studied in Section 4. By applying the results developed in Section 2 in combination with the delta method, we obtain an expression for the asymptotic variance of $\hat{\mu}^{\otimes}$. From a design perspective, this expression is useful for deciding how to decompose $\mathbf{z}$ to reduce the variance of the estimate. The results also suggest how the variance of $\hat{\mu}^{\otimes}$ can be estimated empirically when m_i is large for all i. Section 5 introduces a cross-match estimate that is constructed based on the joint-sampling scheme instead of the split sampling scheme. Properties of this estimate are discussed. Section 6 explores the potential application of the

cross-match estimate in analyzing multiply imputed complete data sets. Section 7 discusses how the cross-match estimate can be approximated if it cannot be computed exactly.

Throughout the paper, we will use subscripts to distinguish expectations and variances taken under different probability distributions. But we will omit the subscripts whenever the meanings are clear from the context.

2. A formula for variance decomposition. Decomposition of variance plays an important role in both applied and theoretical statistics. With the simplest form being $\operatorname{var}(y) = \operatorname{var}\{E(y \mid x)\} + E\{\operatorname{var}(y \mid x)\}$, generalizations of the formula have been made by, say, Hoeffding (1948), Serfling (1980), Efron and Stein (1981), Rubin and Vitale (1980) and Karlin and Rinott (1982), just to start a list. In this section, we assume that $z_1, \ldots, z_k$ are k independent random variables (vectors) with the joint distribution $f^*(z_1, \ldots, z_k) = f_1(z_1) f_2(z_2) \cdots f_k(z_k)$. A major tool for the development in this section is the orthogonal representation of any regular function $g(\cdot)$ of $z_1, \ldots, z_k$, first noticed by Hoeffding in the form

$$g(z_1, \ldots, z_k) = H_0 + \sum_{i=1}^{k} H_i(z_i) + \sum_{i_1 < i_2} H_{i_1 i_2}(z_{i_1}, z_{i_2}) + \cdots + \sum_{i_1 < \cdots < i_l} H_{i_1 \cdots i_l}(z_{i_1}, \ldots, z_{i_l}) + \cdots, \tag{2.1}$$

so that there occur k one-variable functions $H_i(z_i)$, $k(k-1)/2$ two-variable functions $H_{i_1 i_2}(z_{i_1}, z_{i_2})$, and so on. (In this section and the next section, g corresponds to ψ, but will play the role of other functions of ψ in Section 4.) Orthogonality in this context means that all summands in (2.1) are mutually uncorrelated. As a consequence, the following general formula for variance decomposition holds.

$$\operatorname{var}\{g(z_1, \ldots, z_k)\} = \sum_{i=1}^{k} \operatorname{var}(H_i) + \sum_{i_1 < i_2} \operatorname{var}(H_{i_1 i_2}) + \cdots. \tag{2.2}$$

Efron and Stein (1981) specified all the orthogonal terms in the decomposition as

$$\begin{aligned} H_0 &= E(g), \\ H_i(z_i) &= E(g \mid z_i) - E(g), \\ H_{i_1 i_2}(z_{i_1}, z_{i_2}) &= E(g \mid z_{i_1}, z_{i_2}) - E(g \mid z_{i_1}) - E(g \mid z_{i_2}) + E(g), \end{aligned} \tag{2.3}$$

etc., and then used them to show that Tukey's jackknife estimate of variance tends to be biased upwards. The proof of the decomposition identity uses all the conditional expectations $E(g \mid z_{i_1}, \ldots, z_{i_l})$ as primary "basis" and then goes through a Gram–Schmidt-type orthogonalization procedure. From here on, the H's always indicate the orthogonal terms of the Efron–Stein decomposition instead of the more general Hoeffding decomposition. To simplify

notation, (2.1) is rewritten as

$$g(z_1,\ldots,z_k) = \sum_C H_C(\mathbf{z}_C), \tag{2.4}$$

where the summation is over all subsets C of the index set $\{1,\ldots,k\}$, and $\mathbf{z}_C = (z_i; i \in C)$. So $H_\varnothing = H_0$, $H_{\{i\}} = H_i$, etc. Let $h_C = \operatorname{var}(H_C(\mathbf{z}_C))$, so that (2.2) becomes

$$\operatorname{var}\{g(z_1,\ldots,z_k)\} = \sum_{C\neq\varnothing} h_C. \tag{2.5}$$

An important property of the Efron–Stein decomposition is that

$$E(g \mid z_{i_1},\ldots,z_{i_l}) = \sum_{C\subseteq\{i_1,\ldots,i_l\}} H_C(\mathbf{z}_C),$$

that is, the decomposition of a conditional expectation of the function of interest has the same expression as that for the original function.

PROPOSITION 2.1. *For $i = 1,\ldots,k$, let z_{ij}, $j = 1,\ldots,m_i$, be independent draws from the distribution $f_i(z_i)$. Let $g(\cdot)$ be an arbitrary function of the k random variables such that $\operatorname{var}\{g(z_1,\ldots,z_k)\}$ is finite and is decomposed as in* (2.2). *Let*

$$\hat{\mu}^{\otimes} = \frac{1}{m_1\cdots m_k}\sum_{j_1,\ldots,j_k} g(z_{1j_1},\ldots,z_{kj_k});$$

then

$$\begin{aligned}\operatorname{var}(\hat{\mu}^{\otimes}) &= \sum_{i=1}^{k}\frac{\operatorname{var}(H_i)}{m_i} + \sum_{i_1<i_2}\frac{\operatorname{var}(H_{i_1i_2})}{m_{i_1i}m_{i_2}} \\ &\quad + \sum_{i_1<i_2<i_3}\frac{\operatorname{var}(H_{i_1i_2i_3})}{m_{i_1}m_{i_2}m_{i_3}} + \cdots \\ &= \sum_{C\neq\varnothing}\frac{h_C}{\prod_{i\in C} m_i}.\end{aligned} \tag{2.6}$$

PROOF. By definition, we have

$$\operatorname{var}(\hat{\mu}^{\otimes}) = \frac{1}{(m_1\cdots m_k)^2}\operatorname{var}\left\{\sum_{j_1=1}^{m_1}\cdots\sum_{j_k=1}^{m_k} g(z_{1j_1},\ldots,z_{kj_k})\right\}. \tag{2.7}$$

In the expansion of the variance term on the right-hand side, each term is of the form

$$\operatorname{cov}\{g(z_{1j_1},\ldots,z_{kj_k}), g(z_{1j_1'},\ldots,z_{kj_k'})\}.$$

When $j_i \neq j_i'$ for all i, the covariance is zero because of the independence of the z_{ij}. If $j_i = j_i'$ for all $i = 1, \ldots, k$, the above is reduced to $\operatorname{var}\{g(z_1, \ldots, z_k)\}$. In general, if $S = \{i; j_i = j_i'\}$, then

$$\begin{aligned} &\operatorname{cov}\bigl\{g(z_{1j_1}, \ldots, z_{kj_k}), g(z_{1j_1'}, \ldots, z_{kj_k'})\bigr\} \\ &\quad = \operatorname{var}\bigl\{E\bigl(g(z_1, \ldots, z_k) \mid \mathbf{z}_S\bigr)\bigr\} = \sum_{S \supseteq C \neq \varnothing} h_C. \end{aligned} \tag{2.8}$$

Let $C^c = \{1, \ldots, k\} - C$. For any $C \subset \{1, \ldots, k\}$, there are

$$\left[\prod_{i \in C} m_i\right]\left[\prod_{i \in C^c} m_i^2\right]$$

number of ways of choosing an ordered pair of index vectors $(j_1, \ldots, j_k)$ and $(j_1', \ldots, j_k')$ such that $S \supseteq C$. Hence

$$\begin{aligned} \operatorname{var}(\hat{\mu}^{\otimes}) &= \frac{1}{(m_1 \cdots m_k)^2} \sum_{C \neq \varnothing} \left[\prod_{i \in C} m_i \prod_{i \in C^c} m_i^2\right] h_C \\ &= \sum_{C \neq \varnothing} \frac{h_C}{\prod_{i \in C} m_i}. \end{aligned} \tag{2.9}$$

□

REMARK 2.1. Koroljuk and Borovskich (1994, Section 1.2) gave a variance decomposition, expressed in terms of $\operatorname{var}\{E(g \mid z_{i_1}, \ldots, z_{i_l})\}$, for a general k-sample U-statistic. Their formula (1.2.11) can be used to derive Proposition 2.1 but provides little insight. We feel that our proof is more concise and can be generalized in a straightforward fashion to derive a similar variance decomposition for any k-sample U-statistic.

REMARK 2.2. When $m_1 = \cdots = m_k = m$, (2.6) simplifies to

$$\operatorname{var}(\hat{\mu}^{\otimes}) = \frac{u_1}{m} + \frac{u_2}{m^2} + \cdots + \frac{u_k}{m^k}, \tag{2.10}$$

where $u_t = \sum_{|C|=t} h_C$. From the orthogonal decomposition, it is immediate that $\operatorname{var}(g) = u_1 + \cdots + u_k$. Therefore, with $\tilde{\mu}$ as defined in (1.4), $\operatorname{var}(\hat{\mu}^{\otimes}) \leq \operatorname{var}(g)/m = \operatorname{var}(\tilde{\mu})$. Asymptotically ($m \to \infty$), the relative efficiency between $\hat{\mu}^{\otimes}$ and $\tilde{\mu}$ is $\operatorname{var}(g)/u_1$. This indicates that the gain of using $\hat{\mu}^{\otimes}$ over $\tilde{\mu}$ can be small if u_1 is the dominating term of $\operatorname{var}(g)$. For example, if g is an additive function of the z_i, that is, $g(\mathbf{z}) = g_1(z_1) + \cdots + g_k(z_k)$, then $\hat{\mu}^{\otimes} = \tilde{\mu}$ and $\operatorname{var}(g) = u_1$.

REMARK 2.3. Expression (2.6) is useful in deciding the allocation of sample sizes $m_1, \ldots, m_k$. For example, if sampling cost is proportional to $m_1 + \cdots + m_k = M$, then, for a fixed cost, having $m_i \propto h_{\{i\}}$ minimizes the first-order term.

3. Some examples.

EXAMPLE 3.1 (Estimating the moment generating function). Let $\Lambda_i(t) = E(\exp(tz_i)) = \int \exp(tz_i)\,dF_i(z_i) = \int \exp(tz_i) f_i(z_i)\,dz_i$, for $i = 1, \ldots, k$, be the corresponding moment generating functions of the ith distribution, and $M(t) = \prod_{i=1}^{k} \Lambda_i(t)$ be the moment generating function of $\sum_{i=1}^{k} z_i$. Using cross samples to estimate $M(t)$ ends up with

$$\hat{M}^{\otimes}(t) = \frac{1}{m_1 \cdots m_k} \sum_{j_1, \ldots, j_k} \exp\bigl[t(z_{1j_1} + \cdots + z_{kj_k})\bigr] = \prod_{i=1}^{k} \hat{\Lambda}_i(t),$$

where $\hat{\Lambda}_i(t) = (1/m_i)\sum_{j=1}^{m_i} \exp(tz_{ij})$. An easy calculation reveals that

$$\operatorname{var}\bigl(\hat{M}^{\otimes}(t)\bigr) = M^2(t)\left\{\prod_{i=1}^{k}\left[1 + \frac{1}{m_i}\left(\frac{\Lambda_i(2t)}{\Lambda_i^2(t)} - 1\right)\right] - 1\right\}. \tag{3.1}$$

Equation (2.6) of the previous section implies a decomposition:

$$\operatorname{var}\bigl(\hat{M}^{\otimes}(t)\bigr) = \sum_{i=1}^{k} \frac{h_{\{i\}}}{m_i} + \sum_{i_1 < i_2} \frac{h_{\{i_1, i_2\}}}{m_{i_1} m_{i_2}} + \cdots,$$

where the current h's can be obtained from algebraic manipulation of terms. For example,

$$h_{\{i\}} = \operatorname{var}\bigl\{E\bigl(\exp\bigl[t(z_1 + \cdots + z_k)\bigr] \mid z_i\bigr)\bigr\} = M^2(t)\left(\frac{\Lambda_i(2t)}{\Lambda_i^2(t)} - 1\right),$$

$$\begin{aligned} h_{\{i_1, i_2\}} &= \operatorname{var}\Bigl\{E\bigl(\exp\bigl[t(z_1 + \cdots + z_k)\bigr] \mid z_{i_1}, z_{i_2}\bigr)\Bigr\} - h_{\{i_1\}} - h_{\{i_2\}} \\ &= M^2(t)\left(\frac{\Lambda_{i_1}(2t)}{\Lambda_{i_1}^2(t)} - 1\right)\left(\frac{\Lambda_{i_2}(2t)}{\Lambda_{i_2}^2(t)} - 1\right), \end{aligned}$$

and so on. Not surprisingly, it turns out that $h_{\{i_1, \ldots, i_l\}}$ is exactly the term corresponding to $1/(m_{i_1} \cdots m_{i_l})$ in the expansion of the product in (3.1).

EXAMPLE 3.2 (Estimating tail probabilities). Suppose $z_1, \ldots, z_k$ are independent scalars and, for some fixed value of t, we are interested in

$$\mu = P(z_1 + \cdots + z_k \geq t) = E\bigl[g(\mathbf{z})\bigr],$$

where $g(\mathbf{z})$ is the indicator function $I_{\{z_1 + \cdots + z_k \geq t\}}$. Suppose, for each z_i, m independent samples are simulated and denoted by $\{z_{ij}, i = 1, \ldots, k, j = 1, \ldots, m\}$. We now investigate how much improvement we can expect by using the cross-match estimate $\hat{\mu}^{\otimes}$ instead of the simpler estimate $\tilde{\mu}$, both as defined in (1.4) and (1.5). For any set $C \subset \{1, \ldots, k\}$, we let $G_{[C]}(\cdot)$ be the cumulative distribution of $\sum_{i \in C} z_i$; then, in the current setting,

$$\begin{aligned} h_{\{i\}} &= \operatorname{var}\bigl\{P(z_1 + \cdots + z_k \geq t \mid z_i)\bigr\} \\ &= \operatorname{var}\bigl\{1 - G_{[\{i\}^c]}(t - z_i)\bigr\} = \operatorname{var}\bigl\{G_{[\{i\}^c]}(t - z_i)\bigr\}, \\ h_{\{i_1, i_2\}} &= \operatorname{var}\bigl\{G_{[\{i_1, i_2\}^c]}(t - z_{i_1} - z_{i_2})\bigr\} - h_{\{i_1\}} - h_{\{i_2\}}, \\ &\vdots \end{aligned}$$

When no one cdf $G_{\{i\}}$ dominates the others, the convolution $G_{[\{i\}^c]}(t - \cdot)$ will be very smooth compared with the distribution F_i. We observe that $\operatorname{var}(I_{\{z_1+\cdots+z_k\ge t\}}) = \mu(1-\mu) = \mu - \mu^2$, while $h_{\{1\}} \approx E\{P(z_1 + \cdots z_k \ge t \mid z_1)\}^2$ is of the magnitude of μ^2. Typically, we are interested in the tail probability, that is, the case of small μ. In this case, u_1, which is the term corresponding to $1/m$ in (2.10), is much smaller than the total variance $\mu - \mu^2$. Hence, according to formula (2.6), the variance of the cross-match estimate $\hat{\mu}^{\otimes}$ is much smaller than that of the naive estimate $\tilde{\mu}$.

To see how the above argument works, we take the normal distributions for example. Let $S^2 = \sum_{i=1}^k \sigma_i^2$, $S_i^2 = S^2 - \sigma_i^2$ and let $\Phi(x)$ and $\phi(x)$ be respectively the cdf and density of a standard normal random variable. For a fixed number $t > 0$, it is obvious that

$$P(z_1 + \cdots + z_k \ge t) = 1 - \Phi(t/S),$$

and for a given z_1,

$$P(z_1 + \cdots + z_k \ge t \mid z_1) = 1 - \Phi\left(\frac{t - z_1}{S_1}\right).$$

The following heuristic is helpful and can be proved rigorously:

$$h_{\{1\}} = \operatorname{var}\left[\Phi\left(\frac{t-z_1}{S_1}\right)\right] = E\left[\Phi\left(\frac{t-z_1}{S_1}\right) - \Phi\left(\frac{t}{S}\right)\right]^2$$
$$\approx E\left[\phi\left(\frac{t}{S}\right)\left(\frac{t-z_1}{S_1} - \frac{t}{S}\right)\right]^2 = \phi^2\left(\frac{t}{S}\right)\frac{\sigma_1^2}{S_1^2}\left[1 + \frac{t^2\sigma_1^2}{S^2(S_1+S)^2}\right].$$

When all the σ's are assumed to be 1, some algebraic manipulations show that

$$\frac{1}{4\pi(k+1)}\exp\left(-\frac{t^2}{k+1}\right) < h_{\{1\}} \le \frac{1}{\pi k}\exp\left(-\frac{t^2}{k+1}\right)\left(1 + \frac{4t^2}{k^2}\right).$$

On the other hand, $\operatorname{var}(I_{\{z_1+\cdots+z_k\ge t\}}) \approx 1 - \Phi(t/\sqrt{k}) \approx \sqrt{k}\,\phi(t/\sqrt{k})/t$, which is a factor of $\exp(t^2/2k)$ times larger than the upper bound of $h_{\{1\}}$. A similar computation provides that $h_{\{1,2\}} \le \exp[-t^2/(k+3)]\{1 + 4(t^2+1)/k^2\}/k\pi$, which is of the same magnitude as $h_{\{1\}}$.

Binomial distributions can be treated similarly. Let $z_1, \ldots, z_k \sim \operatorname{Bin}(n, p)$, and the tail probability $P(z_1 + \cdots + z_k \ge t)$ be the quantity of interest. Define $\hat{\alpha}(k,t) = (t - nkp)/\sqrt{nkpq}$; then by using the normal approximation to binomial probabilities, we obtain that

$$h_{\{1\}} \approx E\left[\Phi\{\hat{\alpha}(k,t)\} - \Phi\{\hat{\alpha}(k-1, t-z_1)\}\right]^2 \approx \frac{\phi^2(\hat{\alpha}(k,t))}{k-1}\left(1 + \frac{\hat{\alpha}^2(k,t)}{4k}\right).$$

Again, it is much smaller than $\operatorname{var}(I_{\{z_1+\cdots+z_k\ge t\}}) \approx \phi(\hat{\alpha}(k,t))/\hat{\alpha}(k,t)$.

The above discussion gives us some insight into the planning of genetic linkage studies. In the ideal situation in such studies, there is a binomial

observation $y_i \sim \text{Bin}(n_i, \theta)$ from each family i, which yields $z_i = y_i \log\{\theta/(1-\theta)\} + n_i \log(2-2\theta)$ as the log-likelihood ratio (LLR) between a linkage parameter value θ and the unlinked case $\theta = \frac{1}{2}$. In practice, y_i is not directly observable, but the LLR can still be computed from the incomplete observations. In the planning of such studies, one may be interested in the power for detecting a nonnull value θ using k families of roughly equal sizes. Such power considerations essentially require knowledge of tail probabilities of the LLR, which must be obtained by simulations.

4. The importance sampling case. In this section, the samples $\mathbf{z} = (z_1, \ldots, z_k)$ are drawn either from the product measure f^* or from f, depending on whether it is split sampling or joint-sampling. Unless otherwise stated, the components $z_1, \ldots, z_k$ are not assumed to be independent with respect to $p(\cdot)$ or $f(\cdot)$, where p and f are not assumed equal either. The expectations and variances without any subscripts are those taken with respect to f^*.

Rewrite (1.2) as

$$\hat{\mu}^{\otimes} = \frac{(1/m_1 \cdots m_k)\sum_{j_1,\ldots,j_k} w(j_1,\ldots,j_k)\psi(z_{1,j_1},\ldots,z_{k,j_k})}{(1/m_1 \cdots m_k)\sum_{j_1,\ldots,j_k} w(j_1,\ldots,j_k)} \stackrel{\text{def}}{=} \frac{A}{B}. \tag{4.1}$$

Here A is an importance sampling estimate of μ *without* renormalizing the weights. Note that A is an unbiased estimate, but it is not invariant under linear transformation of ψ because the average of the weights, which has expectation 1, will in general not be equal to 1. By contrast, $\hat{\mu}^{\otimes}$ is an invariant estimate. It is a ratio estimate which has a bias of order $(\breve{m})^{-1}$, where $\breve{m} = \min_i[m_i]$. Apart from invariance considerations, another reason for using $\hat{\mu}^{\otimes}$ instead of A is that $\hat{\mu}^{\otimes}$ can be computed even if the importance sampling weights can only be evaluated up to a constant. The latter is often the case in missing data problems where $\mathbf{z}$ corresponds to missing data and $p(\mathbf{z})$ is the conditional distribution given the observed data (see Section 6).

DEFINITION 4.1. We define the following functions that will be used throughout this section:

$$\psi^-(\mathbf{z}) = \psi(\mathbf{z}) - \mu; \qquad \phi(\mathbf{z}) = \frac{p(\mathbf{z})}{f^*(\mathbf{z})}\psi^-(\mathbf{z}); \qquad \varphi(\mathbf{z}) = \frac{p(\mathbf{z})}{f(\mathbf{z})}\psi^-(\mathbf{z}).$$

We then have

$$\hat{\mu}^{\otimes} - \mu = \frac{(1/m_1 \cdots m_k)\sum_{j_1,\ldots,j_k} w(j_1,\ldots,j_k)\psi^-(z_{1,j_1},\ldots,z_{k,j_k})}{(1/m_1 \cdots m_k)\sum_{j_1,\ldots,j_k} w(j_1,\ldots,j_k)} \stackrel{\text{def}}{=} \frac{A^-}{B}.$$

By applying the delta method and noting that $E[A^-] = 0$, $E[B] = 1$,

$$\begin{aligned}\operatorname{var}(\hat{\mu}^{\otimes}) &= \operatorname{var}(\hat{\mu}^{\otimes} - \mu) \\ &= \operatorname{var}\left(\frac{A^-}{B}\right) \\ &\approx \frac{\operatorname{var}(B)(E^2[A^-]/E^2[B]) + \operatorname{var}(A^-) - 2\operatorname{Cov}[A^-, B](E[A^-]/E[B])}{E^2[B]} \\ &= \operatorname{var}(A^-).\end{aligned}$$

The error in the approximation is of order $(\breve{m})^{-2}$.

Let H_i, $i = 0, 1, \ldots, k$, and $H_{i_1 i_2}$, $i_1 < i_2, \ldots,$ etc., be defined as in (2.3), but with their $g(\mathbf{z})$ replaced by $\phi(\mathbf{z})$. Similarly, let $h_{i_1 \cdots i_l} = \operatorname{var}(H_{i_1 \cdots i_l})$. By applying Proposition 2.1 to A^-, with $g(\cdot)$ replaced by $\phi(\cdot)$, we have the following result.

THEOREM 4.1. *We have*

$$\operatorname{var}(A^-) = \sum_{i=1}^{k} \frac{h_i}{m_i} + \sum_{i_1 < i_2} \frac{h_{i_1 i_2}}{m_{i_1} m_{i_2}} + \sum_{i_1 < i_2 < i_3} \frac{h_{i_1 i_2 i_3}}{m_{i_1} m_{i_2} m_{i_3}} + \cdots. \tag{4.2}$$

It follows that

$$\operatorname{var}(\hat{\mu}^{\otimes}) = \operatorname{var}(A^-) + O(\breve{m}^{-2}) = \sum_{i=1}^{k} \frac{h_i}{m_i} + O(\breve{m}^{-2}).$$

LEMMA 4.1. *The conditional expectations $H_i = E[\phi(\mathbf{z}) \mid z_i]$ can be rewritten as $E_f[\varphi(\mathbf{z}) \mid z_i]$. It follows that*

$$\operatorname{var}(\hat{\mu}^{\otimes}) = \sum_{i=1}^{k} \frac{\operatorname{var}_{f_i}\big(E_f[\varphi(\mathbf{z}) \mid z_i]\big)}{m_i} + O(\breve{m}^{-2}). \tag{4.3}$$

PROOF. Note that, with $\mathbf{z}_{(-i)}$ denoting $(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_k)$,

$$\begin{aligned}H_i(z_i) &= E\big[\phi(\mathbf{z}) \mid z_i\big] = E\left[\psi^-(\mathbf{z}) \frac{p(\mathbf{z})}{f^*(\mathbf{z})} \middle| z_i\right] \\ &= \int \psi^-(\mathbf{z}) \frac{p(\mathbf{z})}{f^*(\mathbf{z})} f^*\big(\mathbf{z}_{(-1)} \mid z_i\big)\, d\mathbf{z}_{(-i)} = \int \psi^-(\mathbf{z}) \frac{p(\mathbf{z})}{f_i(z_i)}\, d\mathbf{z}_{(-i)} \\ &= \int \psi^-(\mathbf{z}) \frac{p(\mathbf{z})}{f(\mathbf{z})} f\big(\mathbf{z}_{(-i)} \mid z_i\big)\, d\mathbf{z}_{(-i)} = E_f\big[\varphi(\mathbf{z}) \mid z_i\big].\end{aligned}$$

The lemma then follows from Theorem 4.1 and the fact that $h_i = \operatorname{var}(H_i) = \operatorname{var}_{f_i}(H_i)$. □

LEMMA 4.2. *Suppose* $m_i = m$, $i = 1, \ldots, k$. *For general* $f(\mathbf{z})$,

$$\lim_{m \to \infty} \left[\frac{\operatorname{var}(\hat{\mu}^{\otimes})}{\operatorname{var}_f(\tilde{\mu})} \right] \leq k. \tag{4.4}$$

If $f(\mathbf{z}) = f^*(\mathbf{z})$, *then*

$$\lim_{m \to \infty} \left[\frac{\operatorname{var}(\hat{\mu}^{\otimes})}{\operatorname{var}(\tilde{\mu})} \right] \leq 1. \tag{4.5}$$

The same results apply to the mean-squared errors.

PROOF. By considering $\tilde{\mu}$ as a special case of $\hat{\mu}^{\otimes}$ with $k = 1$ components, Lemma 4.1 gives

$$\begin{aligned} \operatorname{var}_f(\tilde{\mu}) &= \frac{1}{m} \operatorname{var}_f\big(E_f[\varphi(\mathbf{z}) \mid \mathbf{z}]\big) + O(m^{-2}) \\ &= \frac{1}{m} \operatorname{var}_f(\varphi(\mathbf{z})) + O(m^{-2}). \end{aligned} \tag{4.6}$$

Applying Lemma 4.1 to $\hat{\mu}^{\otimes}$, we have

$$\operatorname{var}(\hat{\mu}^{\otimes}) = \frac{1}{m} \sum_{i=1}^{k} \operatorname{var}_{f_i}\big(E_f[\varphi(\mathbf{z}) \mid z_i]\big) + O(m^{-2}). \tag{4.7}$$

For general $f(\mathbf{z})$, $\operatorname{var}_{f_i}(E_f[\varphi(\mathbf{z}) \mid z_i]) \leq \operatorname{var}_f(\varphi(\mathbf{z}))$ for each i. Hence (4.4) follows from (4.6) and (4.7). When $f(\mathbf{z})$ is a product measure, $\sum_{i=1}^{k} \operatorname{var}_{f_i}(E_f[\varphi(\mathbf{z}) \mid z_i]) \leq \operatorname{var}_f(\varphi(\mathbf{z}))$ because (2.2) applies with f and φ substituting for f^* and g, and (4.5) follows. The same results apply to the mean-squared errors because both $\hat{\mu}^{\otimes}$ and $\tilde{\mu}$ have bias of order m^{-1}. □

REMARK 4.1. Both bounds in Lemma 4.2 are tight. To see that, it is adequate to consider the simpler situation where $p(\mathbf{z}) = f(\mathbf{z})$. For the case with general $f(\mathbf{z})$, suppose the z_i are discrete and under f, $z_1 = \cdots = z_k$ with probability 1. Then $\operatorname{var}_{f_i}(E_f[\varphi(\mathbf{z}) \mid z_i]) = \operatorname{var}_f(E_f[\varphi(\mathbf{z}) \mid \mathbf{z}]) = \operatorname{var}_f(\varphi(\mathbf{z}))$. [The same example can also be used to illustrate that the bound (4.4) is indeed asymptotic and the ratio can be bigger than k for small m.] For the case where $f(\mathbf{z}) = f^*(\mathbf{z})$, if $p = f$, the situation reduces to that studied in Section 2. When $\psi(\mathbf{z})$ is additive in the z_i, that is, $\psi(\mathbf{z}) = \psi_1(z_1) + \cdots + \psi_k(z_k)$, then $\tilde{\mu} = \hat{\mu}^{\otimes}$. However, Barnard (1995) showed that in the setting of multiple imputation, even for additive functions, the cross-match idea can lead to a superior approximation of the variances of the multiple imputation estimates.

REMARK 4.2. When $f(\mathbf{z})$ is a product measure (Case II in Section 1), (4.5) in Lemma 4.2 establishes that using the cross-matches will in general lead to a superior estimate asymptotically. While a proof is not available at this moment, we will not be surprised if more careful calculations demonstrate that $\hat{\mu}^{\otimes}$ always has smaller mean-squared error than $\tilde{\mu}$ for any finite m. Indeed, when $f(\mathbf{z})$ is quite different from $p(\mathbf{z})$ and m is only of moderate size,

we believe that $\hat{\mu}^{\otimes}$ can have significantly smaller bias than $\tilde{\mu}$. We also note that a corresponding nonasymptotic result can be easily obtained for the unbiased cross-match estimate A defined in (4.1). Specifically, if $f(\mathbf{z}) = f^*(\mathbf{z})$ and $m_i = m$ for all i, then from Proposition 2.1 and (2.10), we get

$$\operatorname{var}(A) \leq \frac{1}{m} \operatorname{var}\left(\frac{p(\mathbf{z})}{f(\mathbf{z})} \psi(\mathbf{z}) \right)$$

by substituting $p(\mathbf{z})\psi(\mathbf{z})/f(\mathbf{z})$ for their function $g(\mathbf{z})$.

REMARK 4.3. From Lemma 4.1 and (4.6), we see that $\hat{\mu}^{\otimes}$ is asymptotically more efficient than $\tilde{\mu}$ if

$$\sum_{i=1}^{k} \operatorname{var}_{f_i}\left(E_f\left[\varphi(\mathbf{z}) \mid z_i \right] \right) \leq \operatorname{var}_f(\varphi(\mathbf{z})). \tag{4.8}$$

The condition that $f(\mathbf{z})$ factors is sufficient, but by no means necessary, for the inequality to hold. Indeed, if $\varphi(\mathbf{z})$ is highly nonlinear and the z_i are only weakly dependent under f, we can expect $\hat{\mu}^{\otimes}$ to have substantially smaller mean-squared error than $\tilde{\mu}$. This is very important for the type of applications studied in Section 6.

EXAMPLE 4.1. Suppose the z_i are (0–1) binary variables and

$$p(\mathbf{z}) = f(\mathbf{z}) = \frac{\Gamma(\alpha + \beta)\Gamma\left(\alpha + \sum_{i=1}^{k} z_i\right)\Gamma\left(\beta + k - \sum_{i=1}^{k} z_i\right)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + k)}, \tag{4.9}$$

where $\Gamma(\cdot)$ is the gamma function, and $\alpha, \beta > 0$. Note that (4.9) is the joint distribution of the first k steps of a binary Polya sequence and can be constructed from assuming that the z_i are Bernoulli trials given a probability parameter ω, but ω is unknown and is assumed to have a Beta(α, β) distribution. For this example, suppose $\alpha = \beta$ so that (4.9) reduces to

$$p(\mathbf{z}) = f(\mathbf{z}) = \frac{\Gamma(2\alpha)\Gamma\left(\alpha + \sum_{i=1}^{k} z_i\right)\Gamma\left(\alpha + k - \sum_{i=1}^{k} z_i\right)}{\Gamma^2(\alpha)\Gamma(2\alpha + k)}. \tag{4.10}$$

From the construction, it is obvious that the z_i are always positively correlated, the correlation is weak if α is large and the correlation is strong if α is close to zero. First, consider the case with $k = 2$. The joint distribution of z_1 and z_2 is $p(1,1) = p(0,0) = (1/2)(1 + \alpha)/(1 + 2\alpha)$ and $p(1,0) = p(0,1) = (1/2)(\alpha)/(1 + 2\alpha)$. Suppose $\psi(\mathbf{z}) = z_1 + z_2$, then simple calculations show that

$$\operatorname{var}_f(\tilde{\mu}) = \frac{1}{m}\left(\frac{1 + \alpha}{1 + 2\alpha} \right) \tag{4.11}$$

and

$$\begin{aligned}\operatorname{var}(\hat{\mu}^{\otimes}) &= \frac{1}{m}\Big(\operatorname{var}\big(E_f[\varphi(\mathbf{z}) \mid z_1]\big) + \operatorname{var}\big(E_f[\varphi(\mathbf{z}) \mid z_2]\big)\Big) + O(m^{-2}) \\ &= \frac{1}{m}\left[2\left(\frac{1+\alpha}{1+2\alpha}\right)^2\right] + O(m^{-2}).\end{aligned} \tag{4.12}$$

Hence

$$\lim_{m\to\infty}\left[\frac{\operatorname{var}(\hat{\mu}^{\otimes})}{\operatorname{var}_f(\tilde{\mu})}\right] = 2\left(\frac{1+\alpha}{1+2\alpha}\right) > 1, \tag{4.13}$$

which approaches 1 if $\alpha \to \infty$ and approaches 2 if $\alpha \to 0$. Now suppose we change $\psi(\mathbf{z})$ to $z_1 - z_2$. Similar calculations show that

$$\operatorname{var}_f(\tilde{\mu}) = \frac{1}{m}\left(\frac{\alpha}{1+2\alpha}\right), \qquad \operatorname{var}(\hat{\mu}^{\otimes}) = \frac{1}{m}\left[2\left(\frac{\alpha}{1+2\alpha}\right)^2\right] + O(m^{-2}),$$

and

$$\lim_{m\to\infty}\left[\frac{\operatorname{var}(\hat{\mu}^{\otimes})}{\operatorname{var}_f(\tilde{\mu})}\right] = 2\left(\frac{\alpha}{1+2\alpha}\right) < 1,$$

which approaches 1 if $\alpha \to \infty$ and approaches 0 if $\alpha \to 0$! So $\hat{\mu}^{\otimes}$ can sometimes do better than $\tilde{\mu}$ even if ψ is linear and $f = p$ does not factor. Consider now that $\psi(\mathbf{z}) = \prod_{i=1}^{k} z_i$, which is nonlinear, for general k. Here

$$\begin{aligned}P_p(\psi(\mathbf{z}) = 1) &= P_f(\psi(\mathbf{z}) = 1) = P_f(z_1 = \cdots = z_k = 1) \\ &= \frac{\Gamma(2\alpha)\Gamma(\alpha+k)\Gamma(\alpha)}{\Gamma^2(\alpha)\Gamma(2\alpha+k)} = \frac{\Gamma(2\alpha)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(2\alpha+k)} = \mu,\end{aligned}$$

and

$$\begin{aligned}\operatorname{var}_{f_i}\big(E_f[\varphi(\mathbf{z}) \mid z_i]\big) &= \operatorname{var}_{f_i}\big(E_f[\psi(\mathbf{z}) - \mu \mid z_i]\big) \\ &= \operatorname{var}_{f_i}(\psi(\mathbf{z})) = \operatorname{var}_{f_i}(z_i(2\mu)) = (2\mu)^2 \operatorname{var}_{f_i}(z_i) = \mu^2.\end{aligned}$$

Hence

$$\lim_{m\to\infty}\left[\frac{\operatorname{var}(\hat{\mu}^{\otimes})}{\operatorname{var}_f(\tilde{\mu})}\right] = \frac{k\mu^2}{\mu(1-\mu)} = \frac{k\mu}{1-\mu}.$$

It is clear that μ is a decreasing function of α. When $\alpha = 1$, $\mu = 1/(k+1)$ and the above limit is equal to 1. Hence $\hat{\mu}^{\otimes}$ is aymptotically more efficient than $\tilde{\mu}$ if and only if $\alpha > 1$. It is interesting that this cutoff does not depend on k. However, if $\alpha > 1$, then the ratio of the asymptotic variances goes to 0 as $k \to \infty$. For example, if $\alpha = 2$, then $\mu = 6/[(k+3)(k+2)]$.

Up to this point, the role of $p(\mathbf{z})$, specifically the effect of the difference between $p(\mathbf{z})$ and $f(\mathbf{z})$, has been suppressed; it simply got absorbed into $\phi(\mathbf{z})$

or $\varphi(\mathbf{z})$ and the examples used so far assume $p(\mathbf{z}) = f(\mathbf{z})$. The next two lemmas change that by reexpressing the expectations and variances so that they are taken under p.

DEFINITION 4.2. For $i = 1, \ldots, k$, we define

$$\sigma_i^2 = \operatorname{var}_{f_i}\left[\frac{p_i(z_i)}{f_i(z_i)}\right] = E_{p_i}\left[\frac{p_i(z_i)}{f_i(z_i)}\right] - 1,$$

and

$$\sigma^2 = \operatorname{var}_f\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\right] = E_p\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\right] - 1.$$

LEMMA 4.3. *For* $i = 1, \ldots, k$,

$$h_i = \operatorname{var}[H_i(z_i)] = \operatorname{var}_{f_i}[H_i(z_i)] = (1 + \sigma_i^2)\operatorname{var}_{p_i}[G_i(z_i)] + R_i,$$

where $H_i(z_i) = E_{f^*}[\phi(\mathbf{z}) \mid z_i]$, $G_i(z_i) = E_p[\psi(\mathbf{z}) \mid z_i]$ *and*

$$R_i = \operatorname{Cov}_{p_i}\left[\frac{p_i(z_i)}{f_i(z_i)}, (G_i(z_i) - \mu)^2\right].$$

PROOF. Note that, with $\mathbf{z}_{(-i)}$ denoting $(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_k)$,

$$\begin{aligned} H_i(z_i) &= E_{f^*}[\phi(\mathbf{z}) \mid z_i] = E_{f^*}\left[\psi^-(\mathbf{z})\frac{p(\mathbf{z})}{f^*(\mathbf{z})}\middle| z_i\right] \\ &= \int \psi^-(\mathbf{z})\frac{p(\mathbf{z}_{(-i)} \mid z_i)p_i(z_i)}{f_1(z_1)\cdots f_k(z_k)} f^*(\mathbf{z}_{(-i)} \mid z_i)\, d\mathbf{z}_{(-i)} \\ &= \frac{p_i(z_i)}{f_i(z_i)}\int \psi^-(\mathbf{z})\, p(\mathbf{z}_{(-i)} \mid z_i)\, d\mathbf{z}_{(-i)} = \frac{p_i(z_i)}{f_i(z_i)} E_p[\psi^-(\mathbf{z}) \mid z_i]. \end{aligned}$$

So

$$\begin{aligned} E_{f_i}[H_i(z_i)] &= \int \frac{p_i(z_i)}{f_i(z_i)} E_p[\psi^-(\mathbf{z}) \mid z_i] f_i(z_i)\, dz_i \\ &= \int p_i(z_i) E_p[\psi^-(\mathbf{z}) \mid z_i]\, dz_i = E_p[\psi^-(\mathbf{z})] = 0. \end{aligned}$$

Hence

$$\begin{aligned} &\operatorname{var}_{f_i}[H_i(z_i)] \\ &\quad = E_{f_i}\left[\left(\frac{p_i(z_i)}{f_i(z_i)}\right)^2 E_p^2[\psi^-(\mathbf{z}) \mid z_i]\right] \\ &\quad = E_{p_i}\left[\frac{p_i(z_i)}{f_i(z_i)} E_p^2[\psi^-(\mathbf{z}) \mid z_i]\right] \end{aligned}$$

$$= E_{p_i}\left[\frac{p_i(z_i)}{f_i(z_i)}\big(E_p[\psi(\mathbf{z}) \mid z_i] - \mu\big)^2\right]$$

$$= E_{p_i}\left[\frac{p_i(z_i)}{f_i(z_i)}(G_i(z_i) - \mu)^2\right]$$

$$= E_{p_i}\left[\frac{p_i(z_i)}{f_i(z_i)}\right]E_{p_i}\big[(G_i(z_i) - \mu)^2\big] + \mathrm{Cov}_{p_i}\left[\frac{p_i(z_i)}{f_i(z_i)}, (G_i(z_i) - \mu)^2\right]$$

$$= \big(1 + \sigma_i^2\big)\mathrm{var}_{p_i}[G_i(z_i)] + \mathrm{Cov}_{p_i}\left[\frac{p_i(z_i)}{f_i(z_i)}, (G_i(z_i) - \mu)^2\right].$$

The last equality follows from the fact that $E_{p_i}[G_i(z_i)] = \mu$. The lemma follows. □

LEMMA 4.4. *In general,*

$$\sigma_i^2 \le \sigma^2 \tag{4.14}$$

for all i. *If* $f(\mathbf{z}) = f^*(\mathbf{z})$, *then*

$$\sum_{i=1}^{k} \sigma_i^2 \le \sigma^2. \tag{4.15}$$

If $f(\mathbf{z}) = f^*(\mathbf{z})$ *and* $p(\mathbf{z}) = p_1(z_1) \cdots p_k(z_k)$, *then*

$$1 + \sigma^2 = \prod_{i=1}^{k}\big(1 + \sigma_i^2\big). \tag{4.16}$$

PROOF. First note that

$$E_f\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\middle| z_i\right] = \int \frac{p(\mathbf{z})}{f(\mathbf{z})} f\big(\mathbf{z}_{(-i)} \mid z_i\big)\, d\mathbf{z}_{(-i)}$$

$$= \int \frac{p(\mathbf{z})}{f(\mathbf{z})}\frac{f(\mathbf{z})}{f_i(z_i)}\, d\mathbf{z}_{(-i)}$$

$$= \frac{p_i(z_i)}{f_i(z_i)}\int p\big(\mathbf{z}_{(-i)} \mid z_i\big)\, d\mathbf{z}_{(-i)}$$

$$= \frac{p_i(z_i)}{f_i(z_i)}.$$

Hence

$$\sigma^2 = \mathrm{var}_f\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\right] = \mathrm{var}_f\left[E_f\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\middle| z_i\right]\right] + E_f\left[\mathrm{var}_f\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\middle| z_i\right]\right]$$

$$\ge \mathrm{var}_f\left[E_f\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\middle| z_i\right]\right] = \mathrm{var}_{f_i}\left[\frac{p_i(z_i)}{f_i(z_i)}\right] = \sigma_i^2.$$

If $f(\mathbf{z}) = f^*(\mathbf{z})$, which factors, then by applying (2.2) with $g(\mathbf{z}) = [\,p(\mathbf{z})/f(\mathbf{z})]$,

$$\sigma^2 = \operatorname{var}_f\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\right] \geq \sum_i \operatorname{var}_f\left[E_f\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\middle| z_i\right]\right] = \sum_i \operatorname{var}_{f_i}\left[\frac{p_i(z_i)}{f_i(z_i)}\right] = \sum_i \sigma_i^2.$$

If both p and f are product measures, then

$$\begin{aligned} 1 + \sigma^2 &= E_f^2\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\right] + \operatorname{var}_f\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}\right] \\ &= E_f\left[\left(\frac{p(\mathbf{z})}{f(\mathbf{z})}\right)^2\right] = E_f\left[\prod_i\left(\frac{p(z_i)}{f(z_i)}\right)^2\right] = \prod_i E_{f_i}\left[\left(\frac{p(z_i)}{f(z_i)}\right)^2\right] \\ &= \prod_i\left(1 + \sigma_i^2\right). \end{aligned}$$

□

Assuming that $m_1 = \cdots = m_k = m$, Theorem 4.1 and Lemma 4.3 give

$$\operatorname{var}\left[\,\hat{\mu}^{\otimes}\right] = \frac{\sum_i R_i + \sum_i\left(1 + \sigma_i^2\right)\operatorname{var}_{p_i}\left[G_i(z_i)\right]}{m} + O(m^{-2}). \tag{4.17}$$

By regarding $\tilde{\mu}$ as a special version of $\hat{\mu}^{\otimes}$ with $k = 1$, we get

$$\operatorname{var}\left[\,\tilde{\mu}\right] = \frac{R + (1 + \sigma^2)\operatorname{var}_p\left[\,\psi(\mathbf{z})\right]}{m} + O(m^{-2}), \tag{4.18}$$

where

$$R = \operatorname{Cov}_p\left[\frac{p(\mathbf{z})}{f(\mathbf{z})}, (\psi(\mathbf{z}) - \mu)^2\right].$$

Note that σ_i^2 and σ^2 depend only on p and f and not on ψ. They are sometimes referred to as *chi-squared distances* and measure how much the corresponding pair of distributions differ. The terms $\operatorname{var}_{p_i}[G_i(z_i)]$ and $\operatorname{var}_p[\,\psi(\mathbf{z})]$ depend on p and ψ, but not on f. While the G_i is similar to the H_i in (2.3), p is not assumed to be a product measure here and hence (2.2) may not necessarily apply. So, while $\operatorname{var}_{p_i}[G_i(z_i)]$ is always smaller than $\operatorname{var}_p[\,\psi(\mathbf{z})]$, it is not necessary that $\sum_{i=1}^k \operatorname{var}_{p_i}[G_i(z_i)] \leq \operatorname{var}_p[\,\psi(\mathbf{z})]$. The remainder terms R_i and R depend on p, f and ψ, which makes them more difficult to interpret. Note that both R_i and R can be positive or negative, and are not necessarily small since they do not depend on m. In situations where $f(\mathbf{z})$ is deliberately chosen for a specific ψ for the purpose of variance reduction, R and R_i will likely be negative and can, in absolute value, be a large function of $(1 + \sigma_i^2)\operatorname{var}_{p_i}[G_i(z_i)]$. However, as mentioned in Section 1, we are more interested in situations where $f(\mathbf{z})$ just happens to be a convenient distribution to sample from and is not deliberately chosen for a specific ψ. Indeed, for many applications (see Section 6, e.g.), there are many ψ of interest simultaneously. In these instances, we feel that it is often useful to gain an intuitive feeling of what is going on by ignoring the remainder terms.

For example, if $p(\mathbf{z}) = p_1(z_1) \cdots p_k(z_k)$, then $\sum_i \mathrm{var}_{p_i}[G_i(z_i)] \le \mathrm{var}_p[\psi(\mathbf{z})]$. Since Lemma 4.4 states that $\sigma_i^2 \le \sigma^2$ for all i, ignoring the remainder terms R_i and R we get

$$\begin{aligned}\mathrm{var}[\hat{\mu}^{\otimes}] &\approx \frac{1}{m}\sum_{i=1}^{k}(1+\sigma_i^2)\mathrm{var}_{p_i}[G_i(z_i)] \\ &\le \frac{1+\sigma^2}{m}\sum_{i=1}^{k}\mathrm{var}_{p_i}[H_i(z_i)] \le \frac{1+\sigma^2}{m}\mathrm{var}_p[\psi] \approx \mathrm{var}[\tilde{\mu}],\end{aligned}$$

which means that we may expect $\hat{\mu}^{\otimes}$ to be more efficient than $\tilde{\mu}$. When both f and p are product measures of the components, which is the case with the genetics application discussed in the Introduction, (4.18) and (4.17) together with (4.16) in Lemma 4.4 suggest that drastic reduction of variance can be obtained by using $\hat{\mu}^{\otimes}$ instead of $\tilde{\mu}$ when σ^2 is large.

EXAMPLE 4.1 (Continued). Suppose $p(\mathbf{z})$ is (4.10), and f is (4.9) with the same α value as p, but β may be different from α. With $k = 2$ and $\psi(\mathbf{z}) = z_1 + z_2$, (4.11), (4.12) and (4.13) generalize to

$$\begin{aligned}\mathrm{var}_f(\tilde{\mu}) &= \frac{1}{m}\left(\frac{1+\alpha}{1+2\alpha}\right)\frac{1}{2}\left(\frac{(\alpha+\beta)(1+\alpha+\beta)}{2\alpha(1+2\alpha)}\right) \\ &\quad\times\left(1+\frac{\alpha(1+\alpha)}{\beta(1+\beta)}\right)+O(m^{-2}),\end{aligned}$$

$$\mathrm{var}(\hat{\mu}^{\otimes}) = \frac{1}{m}\left(\frac{1+\alpha}{1+2\alpha}\right)^2\frac{(\alpha+\beta)^2}{2\alpha\beta}+O(m^{-2})$$

and

$$\lim_{m\to\infty}\left[\frac{\mathrm{var}(\hat{\mu}^{\otimes})}{\mathrm{var}_f(\tilde{\mu})}\right] = \frac{2(1+\alpha)(1+\beta)(\alpha+\beta)}{[\alpha(1+\alpha)+\beta(1+\beta)](1+\alpha+\beta)}. \tag{4.19}$$

Unlike (4.13), (4.19) is not always bigger than 1. For example, for $\alpha = 5$, $\beta = 15$, (4.19) is 0.677. In general, for a fixed α, if $\beta \to 0$, (4.19) $\to 2/(1+\alpha)$, which is smaller than 1 if $\alpha > 1$, and if $\beta \to \infty$, (4.19) $\to 2(1+\alpha)/\beta \to 0$.

We end this section by pointing out that Theorem 4.1 suggests how the variance of $\hat{\mu}^{\otimes}$ can be estimated from the samples when m_i is large for all i. For $i = 1,\ldots,k$, $j_i = 1,\ldots,m_i$, $E[\phi(\mathbf{z}) \mid z_i = z_{ij_i}]$ can be estimated by

$$\hat{E}\left[\phi(\mathbf{z}) \mid z_i = z_{ij_i}\right] = \frac{\sum_{j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_k} w(j_1,\ldots,j_k)\left(\psi(z_{1j_1},\ldots,z_{kj_k}) - \hat{\mu}^{\otimes}\right)}{m_1 m_2 \cdots m_{i-1} m_{i+1} \cdots m_k}.$$

For each i, $h_i = \mathrm{var}[E\{\phi(\mathbf{z}) \mid z_i\}]$ can be approximated by the sample variance of $\hat{E}(\phi(\mathbf{z}) \mid z_i = z_{ij_i})$, $j_i = 1,\ldots,m_i$.

5. The cross-match estimate with joint sampling. In this section, we want to point out that even if the component samples are drawn jointly instead of independently, a cross-match estimate can still be constructed. For

simplicity, consider the situation with $k = 2$ components. Paired samples $\mathbf{z}_j = (z_{1j}, z_{2j})$, $j = 1, \ldots, m$, are drawn from the joint distribution $f(\mathbf{z})$. A generalization of $\hat{\mu}^{\otimes}$ is

$$\tilde{\mu}^{\otimes} = \frac{\sum_{j,l} w(j,l)\psi(z_{1j}, z_{2l})}{\sum_{j,l} w(j,l)} = \frac{(1/m^2)\sum_{j,l} w(j,l)\psi(z_{1j}, z_{2l})}{(1/m^2)\sum_{j,l} w(j,l)} \stackrel{\text{def}}{=} \frac{A^{\dagger}}{B^{\dagger}}, \tag{5.1}$$

where

$$w(j,l) = \frac{p(z_{1j}, z_{2l})}{(1/m)f(z_{1j}, z_{2l}) + [(m-1)/m]f_1(z_{1j})f_2(z_{2l})}. \tag{5.2}$$

The idea is that we can *pretend* that the pairing information is somehow lost so that each pair (z_{1j}, z_{2l}) can be considered as having the mixture distribution $(1/m)f(z_{1j}, z_{2l}) + [(m-1)/m]f_1(z_{1j})f_2(z_{2l})$. This is because, if the indexes of the z_1 and z_2 samples have been randomly shuffled, then there is $1/m$ chance that z_{1j} and z_{2l} are actually drawn jointly. It is easy to see that $E(A^{\dagger}) = \mu$ and $E(B^{\dagger}) = 1$, and $\check{\mu}^{\otimes}$ is asymptotically $(m \to \infty)$ unbiased. Also, note that $\check{\mu}^{\otimes}$ can be generalized in a natural way to accommodate situations where $\mathbf{z}$ is decomposed into more than two components. In cases of joint sampling, the marginals are often not available. A method for estimating such marginals is discussed in the next section.

As the samples are not drawn independently, the Efron–Stein orthogonal decomposition does not apply. As a consequence, it is rather difficult to get clean and general results for the variance of $\tilde{\mu}^{\otimes}$. However, if z_1 and z_2 are not too highly dependent with respect to f, it seems reasonable to believe that $\check{\mu}^{\otimes}$ and $\hat{\mu}^{\otimes}$ will behave very similarly for large m. Indeed, by qualitatively extrapolating from (4.17) and (4.18), one may use

$$\frac{\operatorname{var}[\tilde{\mu}]}{\operatorname{var}[\check{\mu}^{\otimes}]} \approx \frac{(1+\sigma^2)\operatorname{var}_p[\psi(\mathbf{z})]}{\sum_i (1+\sigma_i^2)\operatorname{var}_{p_i}[G_i(z_i)]} \tag{5.3}$$

as a guideline. The beauty of $\check{\mu}^{\otimes}$ is that the choice of the decomposition can be made after the generation of the samples. Also, with the same set of samples, different decompositions can be chosen for different ψ. In choosing a decomposition to improve on $\tilde{\mu}$, by referring to (5.3), one can either make the σ_i^2 small relative to σ^2 or make $\operatorname{var}_{p_i}[G_i(z_i)]$ small relative to $\operatorname{var}_p[\psi(\mathbf{z})]$. To do the former, in light of Lemma 4.4, one wants a decomposition so that the components are approximately independent with respect to f, and it is even better if the components are also approximately independent with respect to p. This strategy also makes approximation (5.3) more reliable. Another advantage is that it will work for all ψ.

We end this section by noting that, apart from $\check{\mu}^{\otimes}$, there are other ways of utilizing the cross-matches. One natural estimate is

$$s\tilde{\mu} + (1-s)\frac{\sum_{j\neq l} w(j,l)\psi(z_{1j}, z_{2l})}{\sum_{j\neq l} w(j,l)}, \tag{5.4}$$

where $w(j,l) = p(z_{1j}, z_{2l})/(f_1(z_{1j})f_2(z_{2l}))$ and s is some number between 0 and 1. Unlike $\check{\mu}^{\otimes}$, this estimate treats the actual pairs and the cross-matches differently. It has the advantage that, for any decomposition, with an optimally chosen s, it is at least as good as $\tilde{\mu}$. The difficulty is in how to choose s. If the components are not highly dependent with respect to f, then the optimal value of s is probably quite small for large m and the mean-squared error of (5.4) will probably be not too different from that of $\check{\mu}^{\otimes}$. By the same argument as in Section 4, we can approximate $\tilde{\mu}$ by its numerator $\tilde{\mu}_c = \sum_{j=1}^{m} w(j)\psi(\mathbf{z}_j)$ and approximate the latter estimate by $\hat{\mu}_c^{\otimes} = \sum_{j\neq l} w(j,l)\psi(z_{1j}, z_{2l})$. Hence the covariance between the two terms is approximately $\operatorname{cov}(\tilde{\mu}_c, \hat{\mu}_c^{\otimes})$, which is

$$\begin{aligned}\operatorname{cov}(\tilde{\mu}_c, \hat{\mu}_c^{\otimes}) &= \frac{1}{m(m-1)^2}\sum_{j=1}^{m}\sum_{i\neq j}\operatorname{cov}\Big[w(j)\psi(\mathbf{z}_j), \{w(i,j)\psi(z_{1i}, z_{2j}) \\ &\qquad + w(j,i)\psi(z_{1j}, z_{2i})\}\Big] \\ &= \frac{1}{m-1}E_p\left\{G_1^2(z_1)\frac{p_1(z_1)}{f_1(z_1)} + G_2^2(z_2)\frac{p_2(z_2)}{f_2(z_2)}\right\},\end{aligned}$$

assuming that $E_p(\psi) = 0$. By some tedious manipulations, one can also approximate $\operatorname{var}(\hat{\mu}_c^{\otimes})$, accurate up to order $1/m^2$, by

$$\begin{aligned}\operatorname{var}(\hat{\mu}_c^{\otimes}) \approx \frac{1}{m-1}\Bigg[&E_p\left\{G_1^2(z_1)\frac{p_1(z_1)}{f_1(z_1)} + G_2^2(z_2)\frac{p_2(z_2)}{f_2(z_2)}\right\} \\ &+ 2\operatorname{cov}_f\left\{G_1(z_1)\frac{p_1(z_1)}{f_1(z_1)}, G_2(z_2)\frac{p_2(z_2)}{f_2(z_2)}\right\}\Bigg].\end{aligned}$$

The implication here is that, if the last covariance term is negative, one should choose the combination parameter s as zero. Otherwise, one can solve for an optimal s.

6. The cross-match estimate and multiply imputed data sets. One application that motivated our research is the analysis of multiply imputed complete data sets [Rubin (1987)]. Following standard terminology for Bayesian missing data problems, let $\mathbf{y}$ be observed data, let $\mathbf{z}$ be missing data and let θ be the unknown parameter vector. Let $p^+(\mathbf{y}, \mathbf{z}, \theta)$ and $f^+(\mathbf{y}, \mathbf{z}, \theta)$ denote two possibly different joint distributions of $\mathbf{y}$, $\mathbf{z}$ and θ. For example, p^+ and f^+ may be different because they correspond to different prior distributions for θ, or more extremely, they correspond to two different

statistical models for the data [Meng (1994)]. Treating $\mathbf{y}$ as fixed, define

$$f(\mathbf{z}) \overset{\text{def}}{=} f^+(\mathbf{z} \mid \mathbf{y}), \qquad p(\mathbf{z}) \overset{\text{def}}{=} p^+(\mathbf{z} \mid \mathbf{y}).$$

Let

$$\psi(\mathbf{z}) = E_{p^+}[\lambda(\theta) \mid \mathbf{y}, \mathbf{z}]$$

be the complete data posterior mean of some functional $\lambda(\theta)$ with respect to p^+. Then

$$\mu = E_p[\psi(\mathbf{z})] = E_{p^+}\left[E_{p^+}(\lambda \mid \mathbf{y}, \mathbf{z}) \mid \mathbf{y}\right] = E_{p^+}[\lambda \mid \mathbf{y}]$$

is the actual posterior mean of λ under p^+. Assume μ is of interest and that $\mathbf{z}_j$, $j = 1, \ldots, m$, are independent samples of $\mathbf{z}$ drawn from $f(\mathbf{z})$. We can estimate μ by $\tilde{\mu}$ as in (1.1). For comparison, suppose multiple samples of $\mathbf{z}$ are drawn directly from $p(\mathbf{z})$; then the natural estimate of μ is $\bar{\psi}$, which has variance

$$\operatorname{var}\left[\bar{\psi}\right] = \frac{\operatorname{var}_p[\psi(\mathbf{z})]}{\text{sample size}}.$$

Hence, by applying (4.18) and ignoring the remainder term, we may regard

$$\frac{\operatorname{var}_p[\psi(\mathbf{z})]}{\operatorname{var}[\tilde{\mu}]} \approx m\frac{\operatorname{var}_p[\psi(\mathbf{z})]}{(1+\sigma^2)\operatorname{var}_p[\psi(\mathbf{z})]} \approx \frac{m}{(1+\sigma^2)}$$

as the *effective sample size* associated with $\tilde{\mu}$. (Note that the approximation does not depend on ψ.) The factor $(1+\sigma^2)^{-1}$ reflects the loss of efficiency from sampling from f instead of p. If $\mathbf{z}$ is high-dimensional, then even if $f(\mathbf{z})$ and $p(\mathbf{z})$ are only moderately different, σ^2 can be very large and the corresponding effective sample size of $\tilde{\mu}$ can be quite small even for large m. Here is where the cross-match estimate $\check{\mu}^{\otimes}$ can offer very substantial improvement. By decomposing $\mathbf{z}$ in an appropriate fashion, the marginal chi-squared distances σ_i^2 can be a great deal smaller than σ^2. It then follows from (5.3) that the variance of $\check{\mu}^{\otimes}$ can be much smaller than that of $\tilde{\mu}$. It should be pointed out that the best decomposition of $\mathbf{z}$ may require *reparametrizing* it. For example, if $\mathbf{z}$ has (approximately) a multivariate normal distribution under f, then a linear transformation can be applied to make all the scalar components (approximately) independent.

Instead of constructing the cross-match estimate to reduce variance, a natural and simpler alternative is to draw new samples directly from $p(\mathbf{z}) = p^+(\mathbf{z} \mid \mathbf{y})$. However, that can be expensive. In the extreme, consider a situation where the creator of the multiply imputed data sets, for example, the Census Bureau, is not the same as the user. Suppose $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$, and, maybe for confidentiality reasons, $\mathbf{y}_2$ is not available to the user. If it is true that

$$p^+(\theta \mid \mathbf{y}, \mathbf{z}) = p^+(\theta \mid \mathbf{y}_1, \mathbf{z}) \quad \text{and} \quad f^+(\theta \mid \mathbf{y}, \mathbf{z}) = f^+(\theta \mid \mathbf{y}_1, \mathbf{z}),$$

that is, $\mathbf{y}_2$ and θ are conditionally independent given $(\mathbf{y}_1, \mathbf{z})$, then the user can use the provided imputed data sets to estimate μ, but will be unable to create new imputed data sets on his own.

Barnard (1995) applied the cross-match method to the multiple imputations and systematically explored the potentials of the use of $\hat{\mu}^{\otimes}$ in various applications. His simulation studies for both artificial and realistic data sets showed that the cross-match estimate $\hat{\mu}^{\otimes}$ displays consistent gains over the standard multiple imputation estimate $\tilde{\mu}$ in terms of both accuracies of frequency coverages and average lengths of the resulting confidence intervals. He also identified situations when the split-sampling can be approximately achieved (i.e., $p = f = f^*$) so that one can avoid the complications of having to compute the importance weights w as in (5.2). When such a complication is neither avoidable nor straightforward, we propose the following approach.

Again for simplicity, suppose $k = 2$ and rewrite the importance sampling weights (5.2) as

$$
\begin{aligned}
&w(j,l)\\
&\quad= \frac{p^+(z_{1j}, z_{2l} \mid \mathbf{y})}{(1/m) f^+(z_{1j}, z_{2l} \mid \mathbf{y}) + [(m-1)/m] f^+(z_{1j} \mid \mathbf{y}) f^+(z_{2l} \mid \mathbf{y})} \qquad (6.1)\\
&\quad= \frac{f^+(\mathbf{y})}{p^+(\mathbf{y})} \frac{p^+(z_{1j}, z_{2l}, \mathbf{y})}{(1/m) f^+(z_{1j}, z_{2l}, \mathbf{y}) + [(m-1)/m] f^+(z_{1j} \mid \mathbf{y}) f^+(z_{2l}, \mathbf{y})}.
\end{aligned}
$$

Since we only need the weights up to a constant, the factor $f^+(\mathbf{y})/p^+(\mathbf{y})$ can be ignored. Assuming that both $p^+(\theta)$ and $f^+(\theta)$ are conjugate priors to $(\mathbf{y}, \mathbf{z})$, then the complete data predictive probabilities $p^+(z_{1j}, z_{2l}, \mathbf{y})$ and $f^+(z_{1j}, z_{2l}, \mathbf{y})$ can be easily computed [Besag (1989), Kong, Liu and Wong (1994)]. The difficulty is in computing $f^+(z_{1j} \mid \mathbf{y}) f^+(z_{2l}, \mathbf{y})$. If z_1 and z_2 are conditionally independent given $\mathbf{y}$ with respect to f^+, then there is no problem since $f^+(z_{1j} \mid \mathbf{y}) f^+(z_{2l}, \mathbf{y}) = f^+(z_{1j}, z_{2l}, \mathbf{y})$. If not, write

$$
f^+(z_{1,j} \mid \mathbf{y}) = \int f^+(z_{1,j} \mid z_2, \mathbf{y}) f^+(z_2 \mid \mathbf{y})\, dz_2.
$$

Since $z_{2,l}$, $l = 1, \ldots, m$, are samples drawn from $f^+(z_2 \mid \mathbf{y})$, we have the Monte Carlo approximation

$$
f^+(z_{1,j} \mid \mathbf{y}) \approx \frac{1}{m} \sum_{l=1}^{m} f^+(z_{1,j} \mid z_{2,l}, \mathbf{y}) = \frac{1}{m} \sum_{l=1}^{m} \frac{f^+(z_{1,j}, z_{2,l}, \mathbf{y})}{f^+(z_{2,l}, \mathbf{y})} \qquad (6.2)
$$

for $j = 1, \ldots, m$. As mentioned earlier, it is assumed that $f^+(z_{1,j}, z_{2,l}, \mathbf{y})$ can all be evaluated. However, $f^+(z_{2,l}, \mathbf{y})$ cannot be directly computed. Using an argument similar to that above, we obtain the Monte Carlo approximation

$$
\begin{aligned}
f^+(z_{2,l}, \mathbf{y}) &= f^+(z_{2,l} \mid \mathbf{y}) f(\mathbf{y}) \approx \frac{1}{m} \sum_{j=1}^{m} \frac{f^+(z_{1,j}, z_{2,l}, \mathbf{y})}{f^+(z_{1,j}, \mathbf{y})} f^+(\mathbf{y})\\
&= \frac{1}{m} \sum_{j=1}^{m} \frac{f^+(z_{1,j}, z_{2,l}, \mathbf{y})}{f^+(z_{1,j} \mid \mathbf{y})} \qquad (6.3)
\end{aligned}
$$

for $l = 1, \ldots, m$. Combining (6.2) and (6.3), and changing "$\approx$" into "$=$," we get a total of $2m$ equations with $2m$ unknowns. Closer inspection reveals that there are actually only $2m - 1$ free equations and the $2m$ unknowns can only be solved up to a constant. More specifically, consider any solution of $f^+(z_{1,j} \mid \mathbf{y})$ and $f^+(z_{2,l}, \mathbf{y})$, $j = 1, \ldots, m$, $l = 1, \ldots, m$. If we multiply the solution values of $f^+(z_{1,j} \mid \mathbf{y})$, $j = 1, \ldots, m$, by some constant c and multiply the solutions of $f^+(z_{2,l}, \mathbf{y})$, $l = 1, \ldots, m$, by $1/c$, then the results will still be a solution to the equations. This implies that we can get estimates of the ratios $f^+(z_{1,j} \mid \mathbf{y})/f^+(z_{1,j'} \mid \mathbf{y})$ and $f^+(z_{2,l}, \mathbf{y})/f^+(z_{2,l'}, \mathbf{y}) = f^+(z_{2,l} \mid \mathbf{y})/f^+(z_{2,l'} \mid \mathbf{y})$. Most importantly, this is also sufficient for getting estimates of the products $f^+(z_{1,j} \mid \mathbf{y}) f^+(z_{2,l}, \mathbf{y})$, which are what we need.

If we let $F = \{f^+(z_{1,i}, z_{2,j}, \mathbf{y})\}_{m \times m}$ be the complete data matrix, and write $F_1 = (f^+(z_{1,j} \mid \mathbf{y}),\ j = 1, \ldots, m)$ as a row vector, then the two equations (6.2) and (6.3) can be summarized as one fixed-point equation

$$F_1 = \left(F_1^{-1} F\right)^{-1} F^T,$$

where, for a vector $\mathbf{v} = (v_1, \ldots, v_m)$, we define $\mathbf{v}^{-1} = (v_1^{-1}, \ldots, v_m^{-1})$. The solution can be obtained by successive substitutions. The cross products, $f^+(z_{1,j} \mid \mathbf{y}) f^+(z_{2l}, \mathbf{y})$, can be expressed, in the matrix form, as $F\{(F_1^{-1}F)^{-1}\}^T (F_1^{-1}F)/m$.

It is noted that this procedure can be extended to estimate the importance sampling weights for situations where $k > 2$. Having to approximate the importance sampling weights produces extra variation for the cross-match estimate. Although some preliminary simulations showed that the method worked accurately, the effect is not well understood and more work in this direction, both empirical and theoretical, is needed.

7. Approximating the cross-match estimate by resampling. Since the cross-match estimate, either $\hat{\mu}^{\otimes}$ or $\check{\mu}^{\otimes}$, involves $\prod_1^k m_i$ combinations of the data, it may not be feasible to compute it exactly. One solution is to approximate it by resampling. Based on the samples, construct a finite product space $\Omega = \prod_{i=1}^k \Omega_i$, where $\Omega_i = \{z_{i,1}, \ldots, z_{i,m_i}\}$. On Ω, consider the probability measure

$$P\left[\mathbf{z} = (z_{1j_1}, \ldots, z_{kj_k})\right] \propto w(j_1, \ldots, j_k),$$

where $w(j_1, \ldots, j_k)$ is either (1.3) or (5.2). It is easy to see that the cross-match estimate is equal to $E_P[\psi(\mathbf{z})]$. Hence the cross-match estimate can be approximate if we can sample from Ω with respect to P.

With $\hat{\mu}^{\otimes}$, if p factors with respect to the components of the decomposition, then P is a product measure and hence trivial to sample from. In the simplest case when $m_1 = \cdots = m_k$ and $w(j_1, \ldots, j_k) \equiv 1$, a resample of size N will result in a variance of $u_1/m + \mathrm{var}(g)/N$ (ignoring higher order terms) for estimated $\hat{\mu}^{\otimes}$ compared with $\mathrm{var}(\eta)/m$ for $\tilde{\mu}$ [see (2.10)]. This helps one to determine how expensive the Monte Carlo samples have to be for using the cross-match estimate. With $\check{\mu}^{\otimes}$, f also has to factor for P to be a

product measure. Otherwise, depending on the circumstances, methods such as Gibbs sampling, the Metropolis–Hastings algorithm or sequential imputation can be applied to sample from P.

Acknowledgments. We are grateful to Michael Stein, Persi Diaconis, John Barnard, Xiao Li Meng, an Associate Editor and two referees for helpful discussions and suggestions.

REFERENCES

BARNARD, J. (1995). Cross-match procedures for multiple-imputation inference: Bayesian theory and frequentist evaluation. Ph.D. dissertation, Dept. Statistics, Univ. Chicago.

BESAG, J. (1989). A candidate's formula: a curious result in Bayesian prediction. *Biometrika* **76** 183.

EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596.

GRAMS, W. F. and SERFLING, R. J. (1973). Convergence rate for U-statistics and related statistics. *Ann. Statist.* **1** 153–160.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

KARLIN, S. and RINOTT, Y. (1982). Applications of ANOVA type decompositions for comparisons of conditional variance statistics including jackknife estimates. *Ann. Statist.* **10** 495–501.

KONG, A. (1992). A note on Monte Carlo estimation of p-values in linkage analysis with multiple families. Technical Report No. 340, Dept. Statistics, Univ. Chicago.

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

KOROLJUK, V. K. and BOROVSKICH, Y. V. (1994). *Theory of U-Statistics*. Kluwer, Dordrecht.

MAHTANI, M. M., WIDEN, E., LEHTO, M., et al. (1996). Mapping of a gene for type-2 diabetes associated with an insulin-secretion defect by a genome scan in Finnish families. *Nature Genetics* **14** 90–94.

MENG, X. L. (1994). Multiple-imputation with uncongenial sources of input (with discussion). *Statist. Sci.* **9** 538–573.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponses in Surveys*. Wiley, New York.

RUBIN, H. and VITALE, R. A. (1980). Asymptotic distribution of symmetric statistics. *Ann. Statist.* **8** 165–170.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

TERWILLIGER, J. and OTT, J. (1992). A multisample bootstrap approach to the estimation of maximized-over-models lod score distributions. *Cytogenet. Cell Genet.* **59** 142–144.

AUGUSTINE KONG
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637

JUN S. LIU
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-MAIL: jliu@stat.stanford.edu

WING HUNG WONG
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90095