EXPLORING CROSS-MATCH ESTIMATORS WITH MULTIPLY-IMPUTED DATA SETS

John Barnard, Xiao-Li Meng, The University of Chicago
John Barnard, Department of Statistics, The University of Chicago, 5734 University Avenue, Chicago, IL 60637

Key Words: Splitting Scheme, Crossed Imputation.

Abstract

Kong, Liu, and Wong (1994) show through theoretical derivations that cross-match estimators could be very useful for reducing Monte-Carlo error. They indicate that a potentially fruitful arena for cross-match estimators is multiple imputation. We explore this use of cross-match estimators under restrictive conditions that facilitate analytical calculations essential for initial theoretical insight.

1 Introduction

Multiple imputation (MI) (Rubin, 1987) has proven to be a useful mode of inference in the presence of missing data, especially in the context of public-use data files (e.g., Meng, 1994; Rubin, 1995). The basic aim of multiple imputation is to allow general users, who typically have little information about the nonresponse mechanism, to reach valid statistical inferences in the presence of nonresponse using only valid complete-data analysis techniques. The standard MI methodology reviewed in Section 2 meets this aim under general conditions (see Rubin, 1987; Meng, 1994; Rubin, 1995). In this paper, we investigate the possibility of improving the efficiency of standard MI estimators when there exists (approximate) independence among imputed values within each imputation.

Coupling Bayesian derivations with frequentist evaluation is a powerful means of obtaining statistical procedures with general applicability. This coupling is central to the multiple-imputation approach. Rubin (1987), in his seminal book on multiple imputation, first gives the Bayesian underpinnings of the MI methodology, and then proceeds to evaluate the methodology from a randomization-based perspective, demonstrating the frequentist validity of MI inference under general settings. Further development along this line is given in Meng (1994) and Rubin (1995). We followed this path in developing our extensions to the standard MI estimators. Here we present some of the underlying Bayesian theory and leave the randomization-based evaluation to subsequent research. Also, we will focus only on cases where the estimand of interest is a scalar quantity. Inference about multicomponent estimands, the problem in which our extension has potential to provide great gains, will be presented in later work.

2 Standard MI Methodology

Let Y be the complete data, i.e., what we would observe in the absence of any missing data; let Y_{obs} be the observed segment of Y, and Y_{mis} be the missing segment, i.e., $Y = (Y_{\text{mis}}, Y_{\text{obs}})$. We assume that with complete data, valid inference about a quantity Q, possibly a model parameter or a finite population characteristic, would be based on the standard large sample statement

$$(Q - \hat{Q}) \sim N(0, U), \tag{1}$$

where $\hat{Q} = \hat{Q}(Y_{\text{mis}}, Y_{\text{obs}})$ is an efficient estimate of Q, and $U = U(Y_{\text{mis}}, Y_{\text{obs}})$ is its associated variance. Statement (1) has a dual interpretation: Frequentists can interpret it as saying that the sampling distribution of $\hat{Q}$ is approximately normal with mean Q and variance U, while Bayesians can view it as the usual large-sample normal approximation to the posterior of Q, with posterior mean $\hat{Q}$ and posterior variance U.

The basic idea of multiple imputation is to fill in the missing data *multiple* times with values drawn from some distribution that predicts the missing values given Y_{obs} and other available information. Each draw of Y_{mis} is called an imputation (or imputation vector). We denote an imputation by $\mathbf{z}$. Given m independent (conditional on Y_{obs}) imputations of Y_{mis}, $\mathbf{z}^1$ through $\mathbf{z}^m$, an analyst calculates the completed-data statistics $\hat{Q}$ and U for each of the m completed-data sets, $Y^\ell = (\mathbf{z}^\ell, Y_{\text{obs}})$, $\ell = 1, \ldots, m$. Let $\mathbf{S}_m = \left\{\hat{Q}_{*\ell}, U_{*\ell},\ \ell = 1, \ldots, m\right\}$ be the set of the resulting $2m$ completed-data statistics, where $\hat{Q}_{*\ell} = \hat{Q}(\mathbf{z}^\ell, Y_{\text{obs}})$ and $U_{*\ell} = U(\mathbf{z}^\ell, Y_{\text{obs}})$. Secondly, the analyst computes the combined statistic

$$\bar{Q}_m = \frac{1}{m}\sum_{\ell=1}^{m} \hat{Q}_{*\ell},$$

and its associated variance

$$T_m = \bar{U}_m + (1 + \frac{1}{m})B_m,$$

where

$$\bar{U}_m = \frac{1}{m}\sum_{\ell=1}^{m} U_{*\ell}$$

measures the within-imputation variability,

$$B_m = \frac{1}{m-1}\sum_{\ell=1}^{m}(\hat{Q}_{*\ell} - \bar{Q}_m)^2$$

measures the between-imputation variability, and the adjustment $(1 + \frac{1}{m})$ is due to the finite number of imputations. Inference about Q is based on the statement

$$(2) \qquad (Q - \bar{Q}_m) \sim t_\nu(0, T_m),$$

where $\nu = (m-1)(1 + r_m^{-1})^2$, and $r_m = (1 + m^{-1})B_m/\bar{U}_m$ estimates the odds of the fraction of missing information.

The justification of these procedures is given in Rubin (1987) and is most easily established using Bayesian calculations. When the imputations are draws from a posterior predictive distribution of Y_{mis}, i.e., the imputations are generated under an explicit Bayesian model, the resulting inference is called repeated-imputation inference (distinguishing it from the more general multiple-imputation inference; see Meng, 1994). Assuming the imputations are repeated imputations under a "proper" model, Rubin (1987) uses (1) and large-sample approximations to obtain

$$(3) \qquad (Q \mid Y_{\text{obs}}) \sim N(\bar{Q}_\infty, \bar{U}_\infty + B_\infty),$$

where $\bar{Q}_\infty = \lim_{m\to\infty} \bar{Q}_m$, $B_\infty = \lim_{m\to\infty} B_m$, and $\bar{U}_\infty = \lim_{m\to\infty} \bar{U}_m$. The key observation is that the approximate posterior of Q depends only on $(\bar{Q}_\infty, \bar{U}_\infty, B_\infty)$. In other words, (3) is equivalent to

$$(4) \qquad (Q \mid \mathbf{S}_\infty) \sim N(\bar{Q}_\infty, \bar{U}_\infty + B_\infty),$$

where $\mathbf{S}_\infty$ is $\mathbf{S}_m$ when m is infinite. However, when the number of imputations is finite (typically m is less than ten), one observes $\mathbf{S}_m$. The loss of information from replacing $\mathbf{S}_\infty$ by $\mathbf{S}_m$ can be substantial when m is very small (e.g., $m \leq 3$), particularly when the fraction of missing information, $B_\infty/(\bar{U}_\infty + B_\infty)$, is large. The loss of information about the between-imputation variance B_∞ is a particularly hard problem to handle, just as it is generally difficult to estimate a variance based on very few observations.

Rubin (1987) derived the conditional distribution of Q given $\mathbf{S}_m$ and used the t-approximation (2) to this distribution to make inference about Q. In the next two sections we present an alternative procedure that attempts to reduce the information loss that results from replacing $\mathbf{S}_\infty$ with $\mathbf{S}_m$. The gain in information comes from assuming a certain kind of conditional independence among sub-vectors within each imputation vector $\mathbf{z}$.

3 Cross-Match Estimators

In this section we examine a class of estimators called cross-match (CM) estimators by Kong, Liu, and Wong (1994). In the multiple-imputation setting, a CM estimator is constructed from the given imputations and a splitting scheme $\mathcal{S}_k$.

Splitting Scheme $\mathcal{S}_k$

A *splitting scheme* $\mathcal{S}_k$ is a scheme for dividing a vector $\mathbf{z}$ into k mutually exclusive (but not necessarily contiguous) sub-vectors, denoting the i^{th} sub-vector by z_i, such that

$$\mathbf{z}(\mathcal{S}_k) = (z_1(\mathcal{S}_k), \ldots, z_k(\mathcal{S}_k)),$$

where $\mathbf{z}(\mathcal{S}_k)$ possibly differs from the original $\mathbf{z}$ by a permutation of its components. The definition and number of the sub-vectors depends upon the choice of splitting scheme. For example, if we have two imputations of four ages where the sex of the four subjects is known, we could split each imputation into two sub-vectors ($k = 2$), z_1 and z_2, according to the sex of the subjects.

subjects	$\mathbf{z}^1$	$\mathbf{z}^2$	sex		$\mathbf{z}^1$	$\mathbf{z}^2$	sex
1	61	59	M	z_1	61	59	M
2	69	65	F		58	60	M
3	58	60	M	z_2	69	65	F
4	72	71	F		72	71	F

Crossing Sub-Vectors

Originally we started with m imputation vectors, indexed from 1 to m. After splitting each imputation vector into k sub-vectors according to the splitting scheme $\mathcal{S}_k$, we are left with a total of $m \times k$ sub-vectors. By crossing the sub-vectors, i.e., exchanging a sub-vector from an imputation with the corresponding sub-vector from a different imputation, we can form m^k crossed imputations, which include

the original m independent imputations. Denoting the i^{th} sub-vector from the ℓ_i^{th} imputation by $z_i^{\ell_i}$, a *crossed imputation* $\mathbf{z}^{\boldsymbol{\ell}}$ can be written as

$$\mathbf{z}^{\boldsymbol{\ell}} = (z_1^{\ell_1}, \ldots, z_k^{\ell_k}) = \mathbf{z}^{(\ell_1,\ldots,\ell_k)},$$

where $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k) \in \mathbb{K}$, and

$$\mathbb{K} = \{\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k) : 1 \le \ell_i \le m, 1 \le i \le k\}$$

is the index set. When the k indices in $\boldsymbol{\ell}$ are the same, the corresponding imputation is one of the original m imputations (e.g., $\boldsymbol{\ell} = (1, 1, \ldots, 1)$). Crossing the four sub-vectors obtained in the previous examples gives

crossed imputations

	$\mathbf{z}^{(1,1)}$	$\mathbf{z}^{(1,2)}$	$\mathbf{z}^{(2,1)}$	$\mathbf{z}^{(2,2)}$	sex
z_1	61	61	59	59	M
	58	58	60	60	M
z_2	69	65	69	65	F
	72	71	72	71	F

Calculating a Cross-Match Estimate

A cross-match estimator for Q is defined as

(5)
$$\tilde{Q}_{\mathbb{K}} = \frac{1}{|\mathbb{K}|} \sum_{\boldsymbol{\ell} \in \mathbb{K}} \hat{Q}_{*\boldsymbol{\ell}} = \frac{1}{m^k} \sum_{\ell_1,\ldots,\ell_k}^{m} \hat{Q}_{*(\ell_1,\ldots,\ell_k)},$$

where $|\mathbb{K}| = m^k$ is the cardinality of $\mathbb{K}$, and

$$\hat{Q}_{*\boldsymbol{\ell}} = \hat{Q}(\mathbf{z}^{\boldsymbol{\ell}}, Y_{\text{obs}}) = \hat{Q}((z_1^{\ell_1}, \ldots, z_k^{\ell_k}), Y_{\text{obs}}).$$

The procedure for calculating a CM estimate $\tilde{Q}_{\mathbb{K}}$ given a splitting scheme $\mathcal{S}_k$ and m imputations is as follows:

1. Split each imputation vector $\mathbf{z}$ into k sub-vectors $(z_1, \ldots, z_k)$ according to the splitting scheme $\mathcal{S}_k$;
2. For each $\boldsymbol{\ell} \in \mathbb{K}$, treat $Y^{\boldsymbol{\ell}} = (\mathbf{z}^{\boldsymbol{\ell}}, Y_{\text{obs}})$ as the complete data and compute $\hat{Q}_{*\boldsymbol{\ell}} = \hat{Q}(Y^{\boldsymbol{\ell}})$;
3. Calculate $\tilde{Q}_{\mathbb{K}}$ according to (5).

Similarly, we can get a CM estimate $\tilde{U}_{\mathbb{K}}$ of the within-imputation variance. These computations are easily performed using a computer with a simple routine that keeps track of the crossed imputations. Note that $\bar{Q}_m$ and $\bar{U}_m$ are a special case of $\tilde{Q}_{\mathbb{K}}$ and $\tilde{U}_{\mathbb{K}}$, respectively, with $k = 1$ (i.e., no splitting).

Variance Under the Imputation Model

Conditional on Y_{obs}, $\hat{Q}$ is a function of the random variable $\mathbf{z}$, the imputation variable. We make the vital assumption that a splitting scheme can be found such that the components or sub-vectors of $\mathbf{z}$ are conditionally independent (conditional on Y_{obs}), i.e., the z_i's are conditionally independent. Under this independence assumption, following the Efron-Stein decomposition for a function of independent variables (Efron and Stein, 1981), $\hat{Q}$ can be expressed as

(6)
$$\begin{aligned}\hat{Q}(\mathbf{z}) &= H_0 + \sum_{i=1}^{k} H_i(z_i) + \sum_{i_1<i_2} H_{i_1,i_2}(z_{i_1}, z_{i_2}) + \ldots \\ &= \sum_{\mathfrak{C}} H_{\mathfrak{C}}(\mathbf{z}_{\mathfrak{C}}),\end{aligned}$$

where $\mathfrak{C} \subseteq \{1, \ldots, k\}$, $\mathbf{z}_{\mathfrak{C}} = (z_i; i \in \mathfrak{C})$, and

$$\begin{aligned}H_0 &= \mathrm{E}(\hat{Q}(\mathbf{z}) \mid Y_{\text{obs}}) \\ H_i(z_i) &= \mathrm{E}(\hat{Q}(\mathbf{z}) \mid z_i, Y_{\text{obs}}) - H_0, \\ H_{i_1,i_2}(z_{i_1}, z_{i_2}) &= \mathrm{E}(\hat{Q}(\mathbf{z}) \mid z_{i_1}, z_{i_2}, Y_{\text{obs}}) - H_{i_1}(z_{i_1}) - \\ &\quad - H_{i_2}(z_{i_2}) + H_0,\end{aligned}$$

etc. The notation $\sum_{\mathfrak{C}}$ means to sum over all subsets $\mathfrak{C}$ of $\{1, \ldots, k\}$. Since all of the H's are uncorrelated, using (6) we get the variance expression

(7)
$$\begin{aligned}\mathrm{V}(\hat{Q}(\mathbf{z}) \mid Y_{\text{obs}}) &= \sum_{i=1}^{k} \mathrm{V}(H_i \mid Y_{\text{obs}}) + \\ &\quad \sum_{i_1<i_2} \mathrm{V}(H_{i_1,i_2} \mid Y_{\text{obs}}) + \ldots \\ &= \sum_{\mathfrak{C}\neq\emptyset} h_{\mathfrak{C}},\end{aligned}$$

where $h_{\mathfrak{C}} = \mathrm{V}(H_{\mathfrak{C}}(\mathbf{z}_{\mathfrak{C}}) \mid Y_{\text{obs}})$. Kong *et al* (1994) proved using (7) that

(8)
$$\mathrm{V}(\tilde{Q}_{\mathbb{K}} \mid Y_{\text{obs}}) = \sum_{\mathfrak{C}\neq\emptyset} \frac{h_{\mathfrak{C}}}{m^{|\mathfrak{C}|}} \le \mathrm{V}(\bar{Q}_m \mid Y_{\text{obs}}) = \sum_{\mathfrak{C}\neq\emptyset} \frac{h_{\mathfrak{C}}}{m}.$$

This result led us to pursue the use of cross-match estimators in the multiple-imputation setting with the hope of reducing the information loss that results from replacing $\mathbf{S}_\infty$ by $\mathbf{S}_m$. The crucial issue here is whether we can find a splitting scheme that satisfies the independence assumption. We do not expect such an assumption to hold exactly in practice, but we do believe that it is often possible to find a splitting scheme that achieves approximate

independence, especially when the size of the Y_{obs} is large, the typical case with public-use data files. For example, splitting on some demographic variables, such as sex (as in our example), seems likely to produce approximate independence in many data sets. When the independence holds only approximately, choosing between $\tilde{Q}_{\mathbb{K}}$ and $\bar{Q}_m$ becomes a bias-variance trade off. We expect that the mean-squared error of $\tilde{Q}_{\mathbb{K}}$ is substantially less than that of $\bar{Q}_m$ when m is small, the fraction of missing information is high, and $\hat{Q}$ is highly non-linear in $\mathbf{z}$ (e.g., p-values). Evidence of this sort will be presented in later work. Here we only develop some of the theory in order to gain insight into our extensions.

4 Underlying Bayesian Theory

Under the Bayesian paradigm, to make inference about Q we need to derive the posterior of Q given the data. In Section 2, the "data" consisted of the m pairs of completed-data estimates and their associated variances. In the CM setting, the "data" consist of the m^k pairs of completed-data estimates and variances, $\mathbf{S}_{\mathbb{K}} = \{\hat{Q}_{*\boldsymbol{\ell}}, U_{*\boldsymbol{\ell}}, \boldsymbol{\ell} \in \mathbb{K}\}$. We wish to calculate the conditional distribution of Q given $\mathbf{S}_{\mathbb{K}}$.

In this section we derive the conditional distribution of Q given $\mathbf{S}_{\mathbb{K}}$ assuming that $\hat{Q}$ is an additive function of the k sub-vectors, $z_1, \ldots, z_k$, i.e., assuming

$$\hat{Q}(\mathbf{z}) = \sum_{i=1}^{k} \hat{Q}^{\{i\}}(z_i), \tag{9}$$

treating Y_{obs} as fixed. This additivity assumption is made for theoretical simplification and will be relaxed in subsequent work, as it restricts the utility of the CM estimators (i.e., $\tilde{Q}_{\mathbb{K}} \equiv \bar{Q}_m$). Both the additivity of $\hat{Q}$ and the independence of the sub-vectors depend upon the choice of splitting scheme $\mathcal{S}_k$; $\hat{Q}$ may be additive under one splitting scheme and not additive under another. Note that additive functions are more general than linear functions (in $\mathbf{z}$). For example, the difference in quantiles between males and females is an additive function if we split by sex, but it is not linear. We call the functions $\hat{Q}^{\{i\}}(\cdot)$ sub-functions, as $\hat{Q}^{\{i\}}(\cdot)$ only depends upon the i^{th} sub-vector z_i (again treating Y_{obs} as fixed).

Our plan of attack follows that of Rubin's (1987) for approximating $[Q \mid \mathbf{S}_m]$ (using $[\,]$ to denote distribution). First derive the conditional distribution of $(\bar{Q}_\infty, \bar{U}_\infty)$ given $\mathbf{S}_{\mathbb{K}}$ (instead of $\mathbf{S}_m$) and B_∞, then combine that result with (4) and the conditional distribution of B_∞ given $\mathbf{S}_{\mathbb{K}}$. As in Rubin (1987), we treat $(\bar{Q}_\infty, B_\infty, \bar{U}_\infty)$ as the set of parameters (thus assigning them priors), and treat $\mathbf{S}_{\mathbb{K}}$ as the data.

Sampling Distribution of $\mathbf{S}_{\mathbb{K}}$ Given Y_{obs}

The first task is to specify the sampling distribution of the m^k pairs $(\hat{Q}_{*\boldsymbol{\ell}}, U_{*\boldsymbol{\ell}})$ given Y_{obs}. In the standard MI setting, each pair is an i.i.d. draw of $\bar{Q}_\infty$ and $\bar{U}_\infty$. In the CM setting the m^k pairs are identically distributed but not independent. The marginal distributions of $\hat{Q}_{*\boldsymbol{\ell}}$ and $U_{*\boldsymbol{\ell}}$ given Y_{obs} are the same as those given in Rubin (1987, Chapter 3),

$$(\hat{Q}_{*\boldsymbol{\ell}} \mid Y_{\text{obs}}) \sim N(\bar{Q}_\infty, B_\infty), \tag{10}$$

$$(U_{*\boldsymbol{\ell}} \mid Y_{\text{obs}}) \sim [\bar{U}_\infty, \ll B_\infty], \tag{11}$$

where $[A, \ll C]$ means the distribution is centered around A with lower order of variability than C.

Consider the variance and covariance structure of the m^k $\hat{Q}_{*\boldsymbol{\ell}}$'s, which follows easily from the additivity assumption (9) and the variance decomposition (7):

$$\mathrm{V}(\hat{Q}_{*\boldsymbol{\ell}} \mid Y_{\text{obs}}) = B_\infty = \sum_{i=1}^{k} h_i, \tag{12}$$

and

$$\mathrm{Cov}(\hat{Q}_{*\boldsymbol{\ell}}, \hat{Q}_{*\boldsymbol{\ell}'} \mid Y_{\text{obs}}) = \sum_{i \in \mathfrak{S}} h_i, \tag{13}$$

where $\mathfrak{S} = \{i : \ell_i = \ell'_i\}$. The last equality in (12) follows because under the additivity assumption, h_i is the between-imputation variance of the i^{th} sub-function, $\hat{Q}^{\{i\}}$.

To specify the joint distribution of the $\hat{Q}_{*\boldsymbol{\ell}}, \boldsymbol{\ell} \in \mathbb{K}$, we use

$$\hat{Q}_{*\boldsymbol{\ell}} = H_0 + \sum_{i=1}^{k} H_i^{\ell_i}, \tag{14}$$

where,

$$H_i^{\ell_i} = \hat{Q}^{\{i\}}(z_i^{\ell_i}) - \mathrm{E}[\hat{Q}^{\{i\}}(z_i^{\ell_i}) \mid Y_{\text{obs}}].$$

The representation follows from the additivity assumption and the Efron-Stein decomposition. Then, in parallel to Rubin's assumption about the conditional distribution of $\hat{Q}_{*\ell}$, we assume that $H_i^{\ell_i} \sim N(0, h_i)$, which is reasonable when k is not too large. Combining the normality and the independence of the $H_i^{\ell_i}$'s, we see that the joint distribution of the $\hat{Q}_{*\boldsymbol{\ell}}$'s is multivariate normal with mean $\bar{Q}_\infty \mathbf{1}_{m^k}$ and variance-covariance structure given by (12) and (13).

The joint distribution of the $U_{*\boldsymbol{\ell}}$'s is more difficult to specify precisely (or even imprecisely), but fortunately, a joint specification is not required because of the lower order variability assumption in (11).

Conditional Distribution of $(\bar{Q}_\infty, \bar{U}_\infty)$ Given $\mathbf{S}_\mathbb{K}$ and B_∞

Accepting the sampling distributions given in the previous section, it is straightforward to obtain the conditional distribution for $(\bar{Q}_\infty, \bar{U}_\infty)$ given $\mathbf{S}_\mathbb{K}$ and B_∞. First, if the prior distribution of $\bar{Q}_\infty$ given B_∞ is proportional to a constant, then

$$(15) \qquad (\bar{Q}_\infty \mid \mathbf{S}_\mathbb{K}, B_\infty) \sim N(\bar{Q}_m, \frac{B_\infty}{m}).$$

Note that this is the same as $[\bar{Q}_\infty \mid \mathbf{S}_m, B_\infty]$ given in (3.3.5) of Rubin (1987). This is not surprising since $\tilde{Q}_\mathbb{K} = \bar{Q}_m$ under the additivity assumption.

Following Rubin's (1987) assumption (11), we also have

$$(16) \qquad (\bar{U}_\infty \mid \mathbf{S}_\mathbb{K}, B_\infty) \sim [\tilde{U}_\mathbb{K}, \ll B_\infty/m].$$

Conditional Distribution of Q Given $\mathbf{S}_\mathbb{K}$ and B_∞

Combining the results of the previous section and (4) yields an approximation to the conditional distribution of Q given $\mathbf{S}_\mathbb{K}$ and B_∞. Expression (16) implies that in (4) $\bar{U}_\infty$ can be replaced by $\tilde{U}_\mathbb{K}$ to give

$$(17) \quad (Q \mid \mathbf{S}_\mathbb{K}, \bar{Q}_\infty, B_\infty) \sim N(\bar{Q}_\infty, \tilde{U}_\mathbb{K} + B_\infty).$$

Combining expression (15) and (17) yields

$$(18) \quad (Q \mid \mathbf{S}_\mathbb{K}, B_\infty) \sim N\left(\bar{Q}_m, \tilde{U}_\mathbb{K} + B_\infty\left(1 + m^{-1}\right)\right).$$

This is the same result as (3.3.7) in Rubin (1987) for $[Q \mid \mathbf{S}_m, B_\infty]$ if $\bar{U}_m = \tilde{U}_\mathbb{K}$ ($\tilde{U}_\mathbb{K}$ is slightly more efficient than $\bar{U}_m$). Hence, conditional on B_∞, because of the additivity assumption, using $\mathbf{S}_\mathbb{K}$ has not helped us obtain sharper inferences about Q relative to the standard Bayesian MI inference. However, as we shall see in the next section, the crossed data $\mathbf{S}_\mathbb{K}$ does contain extra information beyond that in $\mathbf{S}_m$ about the unknown B_∞.

Conditional Distribution of B_∞ Given $\mathbf{S}_\mathbb{K}$

Since under the additivity assumption $B_\infty = \sum_{i=1}^k h_i$, the conditional distribution of B_∞ given $\mathbf{S}_\mathbb{K}$ depends on the joint conditional distribution of the h_i's. If we let the h_i, $i = 1, \ldots, k$, be *apriori* independent and be distributed $a_i b_i \chi_{b_i}^{-2}$, then the h_i's are *aposteriori* independent and

$$(19) \quad (h_i \mid \mathbf{S}_\mathbb{K}) \sim [(m-1)\hat{h}_i + a_i b_i]\, \chi_{m-1+b_i}^{-2},$$

where $\hat{h}_i = [m^{k-1}(m-1)]^{-1} SS_i$,

$$SS_i = m^{k-1} \sum_{\ell=1}^{m} \left(\hat{Q}_{*\ell}^{\{i\}} - \bar{Q}_m^{\{i\}}\right)^2,$$

and

$$\bar{Q}_m^{\{i\}} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{Q}_{*\ell}^{\{i\}}.$$

The sum of squares $SS_i, i = 1, \ldots, k$, can be computed by creating an ANOVA table using the $\hat{Q}_{*\ell}$'s as the response and the ℓ's as the explanatory variables. By doing this, we avoid having to perform calculations directly on the $m \times k$ sub-functions. Taking b_i as zero in the density of $a_i b_i \chi_{b_i}^{-2}$ results in the noninformative and improper prior h_i^{-1} and makes $(h_i \mid \mathbf{S}_\mathbb{K}) \sim (m-1)\hat{h}_i\, \chi_{m-1}^{-2}$, which is a direct generalization of Rubin's (1987) result for $[B_\infty \mid \mathbf{S}_m]$. We consider this class of priors for the h_i's because it is conjugate (making calculations easier) and quite flexible. The use of informative priors (i.e., $b_i > 0$) is important in calibrating our Bayesian procedures. We will later adjust the b_i's to construct procedures with good frequentist properties. From (19) and the additivity assumption we get

$$(20) \quad (B_\infty \mid \mathbf{S}_\mathbb{K}) \sim \sum_{i=1}^{k} [(m-1)\hat{h}_i + a_i b_i]\, \chi_{m-1+b_i}^{-2},$$

where all of the inverse-chi-squared variables are mutually independent.

Conditional Distribution of Q Given $\mathbf{S}_\mathbb{K}$

Combining the results in (18) and in (20) gives

$$(21) \quad (Q \mid \mathbf{S}_\mathbb{K}) \sim \bar{Q}_m + \left[\tilde{U}_\mathbb{K} + (1 + \frac{1}{m}) \sum_{i=1}^{k} \frac{(m-1)\hat{h}_i + a_i b_i}{\chi_{m-1+b_i}^2}\right]^{1/2} \mathcal{N},$$

where $\mathcal{N}$ is a standard normal random variable, and all of the variables are mutually independent. An alternative expression for $(Q \mid \mathbf{S}_\mathbb{K})$ exists that is less direct than (21) but more intuitive. Following (18) and the fact that a sum of independent normal random variables is also a normal random variable, we have the random variable decomposition

$$(22) \quad (Q \mid \mathbf{S}_\mathbb{K}, h_1, \ldots, h_k) \sim \bar{Q}_m + \mathcal{N}_0 \sqrt{\tilde{U}_\mathbb{K}} + \sqrt{1 + \frac{1}{m}} \sum_{i=1}^{k} \mathcal{N}_i \sqrt{h_i},$$

where all of the $\mathcal{N}_i$'s are independent. Combining this result with (19) gives

$$(23) \quad (Q \mid \mathbf{S}_\mathbb{K}) \sim \tilde{Q}_\mathbb{K} + \mathcal{N}_0 \sqrt{\bar{U}_m} + \sqrt{1 + \frac{1}{m}} \sum_{i=1}^{k} \mathcal{T}_{m-1+b_i} \sqrt{\frac{(m-1)\hat{h}_i + a_i b_i}{m-1+b_i}},$$

where $\mathcal{T}_d$ is a standard t random variable with d degrees of freedom, and all variables are independent. Expression (23) is more useful for calculating the moments of Q given $\mathbf{S}_{\mathbb{K}}$ while (21) is more useful for calculating the conditional density of Q.

When $k = 1$, the case of no splitting, (21) becomes

$$(Q \mid \mathbf{S}_m) \sim \bar{Q}_m + \left[\bar{U}_m + (1 + \frac{1}{m}) \frac{(m-1)B_m + a_1 b_1}{\chi^2_{m-1+b_1}}\right]^{1/2} \mathcal{N},$$

or alternatively, (23) becomes

$$(Q \mid \mathbf{S}_m) \sim \bar{Q}_m + \mathcal{N}_0 \sqrt{\bar{U}_m} + \sqrt{1 + \frac{1}{m}}\, \mathcal{T}_{m-1+b_1} \sqrt{\frac{(m-1)B_m + a_1 b_1}{m - 1 + b_1}}.$$

Taking $b_1 = 0$ leads to the conditional distribution found in Rubin (1987, Chapter 3),

$$(24) \qquad (Q \mid \mathbf{S}_m) \sim \bar{Q}_m + \mathcal{N}_0 \sqrt{\bar{U}_m} + \sqrt{1 + \frac{1}{m}}\, \mathcal{T}_{m-1} \sqrt{B_m},$$

which was derived under $\pi(B_\infty) \propto 1/B_\infty$. Direct comparisons of (24) with (23), however, are problematic because the assignment of priors is not compatible. For example, even when taking $b_i = 0$, $i = 1, \ldots, k$, in (23), the priors $\pi(h_i) \propto h_i^{-1}$ used there do not imply $\pi(\sum_{i=1}^k h_i) \propto (\sum_{i=1}^k h_i)^{-1}$, the prior used in (24). However, our ultimate goal is to compare the frequentist properties of these Bayesianly derived procedures. We hope that the extra information in $\mathbf{S}_{\mathbb{K}}$ relative to $\mathbf{S}_m$, when coupled with the flexibility in our formulation introduced via the additional prior parameters (i.e., the a_i's and b_i's), can lead to procedures with better frequentist properties, at least in some cases. We also expect that the basic results established will be useful for the more general situation when $\hat{Q}$ depends on $\mathbf{z}$ through $\hat{Q}^{\{i\}}(z_i)$, $i = 1, \ldots, k$ (e.g., p-values).

Acknowledgements

This manuscript was prepared using computer facilities supported in part by the National Science Foundation Grants DMS 89-05292, DMS 87-87-03942, and DMS 86-01732 awarded to the Department of Statistics at the University of Chicago, and by The University of Chicago Block Fund. The research was supported in part by the National Science Foundation Grant DMS 92-04504. It was also supported in part by the U.S. Census Bureau through a contract with the National Opinion Research Center at the University of Chicago. We thank A. Kong, D.B. Rubin, and D.L. Wallace for helpful conversations.

References

[1] Efron, B. and Stein, C. (1981) The jackknife estimate of variance. *Annals of Statistics*, **9**, 586-596.

[2] Kong, A., Liu, J. S., and Wong, W. H. (1994) The properties of the cross-match estimate and split sampling. Unpublished paper.

[3] Meng, X. L. (1994) Multiple-imputation inferences under uncongenial sources of input (with discussion). *Statistical Science*, to appear.

[4] Rubin, D. B. (1987) *Multiple imputation for nonresponse in surveys.* John Wiley & Sons: New York.

[5] Rubin, D. B. (1995) Multiple imputation after 18 years. *JASA*, to appear.