EXPLORING HYPOTHESIS-TESTING PROCEDURES WITH MULTIPLY-IMPUTED DATA UNDER UNEQUAL FRACTIONS OF MISSING INFORMATION

Steven Pedlow and Xiao-Li Meng, The University of Chicago

Steven Pedlow, Dept. of Statistics, Univ. of Chicago, 5734 University Ave., Chicago, IL 60637 U.S.A.

Key Words: Bayesian p-value, Incomplete Data, Missing Data, Singular Wishart

1. Introduction

Multiple imputation (Rubin, 1987) is a general and efficient method for statistical analyses with incomplete data, and is especially suited for handling nonresponse in large sample surveys that produce public-use data files. Its development is guided by Bayesian principles and calculations with sensible frequentist evaluations, especially under the randomization perspective, to ensure its general applicability and validity. Recent development and overview of multiple imputation techniques can be found in Meng (1994) and Rubin (1995), both of which also contain extensive citations of literature on studies and applications of multiple imputation.

Briefly speaking, the major task of multiple imputation is to construct a sensible imputation model, implicit or explicit, to describe the predictive distribution of the missing values given all of the available data and information; this work is best accomplished by the data collectors (e.g., U.S. Census Bureau). Once such a model is built, the imputer creates $m(\geq 2)$ sets of imputations by making m draws from the imputation model. Each set of imputed values is then added to the set of observed values to form a "completed" data set. Given m such completed-data sets, a user applies the standard complete-data procedure that he would have used if there were no missing data to each of them. He then combines these m analyses to form one multiple-imputation inference.

Forming multiple-imputation point estimates is straightforward, as reviewed in Section 2. Hypothesis-testing with multiply-imputed data sets is somewhat more complicated because we cannot directly combine p-values from the complete-data testing procedure as it does not take into account the extra variability due to the imputations. Various hypothesis-testing procedures have been proposed in the literature (e.g. Li, 1985; Li, Meng, Raghunathan, and Rubin, 1991; and Meng and Rubin, 1992). One key assumption made under these procedures, which is necessary to simplify the computations, is that the fractions of missing information, to be defined in Section 3, are the same across all of the different components of the parameter vector being tested. This is obviously a strong assumption that will rarely hold in practice. Fortunately, however, simulations as well as empirical studies have shown that the resulting procedures are not too sensitive to this assumption when the variability among the fractions of missing information is not too high (e.g., Li, Raghunathan, and Rubin, 1991; hereafter LRR). Nevertheless, the performance of these procedures, in terms of both level and power, decays as the variability increases.

The purpose of this paper is to report some initial efforts and findings in attempting to establish testing procedures that do not rely on the assumption of equal fractions of missing information. After presenting the necessary background in Section 2, we review in Section 3 the procedure proposed by LRR, which is an approximation to a Bayesian p-value. We then describe in Section 4 a potential extension of LRR's procedure. In Section 5, we discuss the computation of the exact Bayesian p-value underlying LRR's procedure, as well as the interesting finding that LRR's approximation has better frequentist properties than the Bayesian p-value it approximates. Finally, Section 6 discusses a difficult problem our extensions have to face — the possibility that the dimensionality of the parameter being tested is at least as great as the number of imputations.

2. Background

Suppose $\hat{\theta} = \hat{\theta}(X)$ and $U = U(X)$ are an efficient estimate of a k-dimensional population parameter, θ , and its associated variance-covariance matrix, respectively, in which X is an $n \times k$ complete-data matrix. In the presence of missing data, we write $X = (X_{obs}, X_{mis})$ in a convenient, but possibly imprecise notation. With multiple imputation, let $X_*^{(\ell)} = (X_{obs}, X_{mis}^{(\ell)}), \ell = 1, \ldots, m$, be the m completed-data sets. We then compute $\hat{\theta}_{*\ell} = \hat{\theta}(X_*^{(\ell)})$ and $U_{*\ell} = U(X_*^{(\ell)})$, for each $\ell = 1, \ldots, m$, and we write $\mathcal{S}_m = \{\hat{\theta}_{*\ell}, U_{*\ell}, \ell = 1, 2, \ldots, m\}$ for notational simplicity.

Given $\mathcal{S}_m$, the multiple-imputation estimate of θ is a simple average

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

The variance associated with $\bar{\theta}_m$ is

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

where

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

is the within-imputation variance, and

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

is the between-imputation variance. The Bayesian derivations and frequentist evaluations of these procedures can be found in Rubin (1987, Ch. 3, 4).

For hypothesis testing, Rubin (1987, Ch. 3) derived a Bayesian p-value, which induced a testing procedure that was studied in detail by LRR. Here the Bayesian p-value for a null hypothesis $H_0 : \theta = \theta_0$ is defined as the posterior probability of all θ whose posterior density values are no less than that of θ_0. Under the assumption that the imputation model is "proper", Rubin (1987, Ch. 3) shows that the Bayesian p-value, under the constant prior on $\bar{\theta}_\infty (\equiv \lim_m \bar{\theta}_m)$ given $B_\infty (\equiv \lim_m B_m)$ and the assumption $\bar{U}_\infty \approx \bar{U}_m$, is

$$P(\theta_0 | \mathcal{S}_m) = \int Pr\{\chi_k^2 \geq (\bar{\theta}_m - \theta_0)^{\mathsf{T}} [\bar{U}_m + (1 + \frac{1}{m}) B_\infty]^{-1} \times (\bar{\theta}_m - \theta_0)\} Pr(B_\infty | \mathcal{S}_m) dB_\infty \qquad (2.1)$$

in which $Pr(B_\infty | \mathcal{S}_m)$ is a posterior density of B_∞ given $\mathcal{S}_m$. The computational difficulty of (2.1) is caused by the averaging over the posterior density $Pr(B_\infty | \mathcal{S}_m)$, especially when $k > 1$.

3. The Current Best Procedure, P_m

The procedure of LRR is obtained by considering the simplest case for (2.1), when $B_\infty = \bar{\lambda} \bar{U}_\infty$, where $\bar{\lambda}$ is a scalar quantity. This is equivalent to all eigenvalues of $\bar{U}_\infty^{-\frac{1}{2}} B_\infty \bar{U}_\infty^{-\frac{1}{2}}$ being equal, $\lambda_1 = \cdots = \lambda_k \equiv \bar{\lambda}$. In other words, LRR assumed that the relative increase in variance due to missing data is the same for any component of θ, an assumption that greatly simplifies the computation because it reduces a k-dimensional problem into a one-dimensional one.

Specifically, under the non-informative prior $\pi(\bar{\lambda}) \propto \bar{\lambda}^{-1}$, the posterior distribution of $(1 + \frac{1}{m})\bar{\lambda}$ is $r_m k(m-1) \chi_{k(m-1)}^{-2}$, where

$$r_m = (1 + \frac{1}{m}) trace(B_m \bar{U}_m^{-1})/k \qquad (3.1)$$

is a consistent estimator of $\bar{\lambda}$ (in general, r_m estimates the average of $\{\lambda_1, \ldots, \lambda_k\}$). Consequently, (2.1) is simplified to

$$P(\theta_0 | \mathcal{S}_m, B_\infty = \bar{\lambda} \bar{U}_\infty) = Pr\{\chi_k^2 \geq \left[1 + \frac{k(m-1)}{\chi_{k(m-1)}^2} r_m\right]^{-1} \times (\bar{\theta}_m - \theta_0)^{\mathsf{T}} \bar{U}_m^{-1} (\bar{\theta}_m - \theta_0)\}.$$

By approximating $(1+r_m)(1+r_m k(m-1)/\chi_{k(m-1)}^2)^{-1}$ with an mean-squared random variable, Rubin (1987, Ch. 3) constructed the following statistic for testing $H_0 : \theta = \theta_0$:

$$D_m = \frac{(\bar{\theta}_m - \theta_0) \bar{U}_m^{-1} (\bar{\theta}_m - \theta_0)^{\mathsf{T}}}{k[1 + r_m]},$$

where r_m is given in (3.1). A good approximate frequentist reference distribution for D_m was given in LRR, which yields an approximate p-value

$$P_m = Pr[F_{k,w} > D_m], \qquad (3.2)$$

where

$$w = \begin{cases} \frac{v}{2}(1 + \frac{1}{k})[1 + r_m^{-1}]^2, & \text{if } v = k(m-1) \leq 4; \\ 4 + (v - 4)[1 + (1 - \frac{2}{v}) r_m^{-1}]^2, & \text{otherwise.} \end{cases}$$

A problem with LRR's procedure is that as $m \to \infty$, unless $B_\infty = \bar{\lambda}\bar{U}_\infty$ holds exactly, D_m does not converge to the ideal test statistic

$$D_{ideal} = \frac{(\bar{\theta}_\infty - \theta_0)^\top T_\infty^{-1}(\bar{\theta}_\infty - \theta_0)}{k},$$

where $T_\infty = \lim_m T_m$. The corresponding ideal p-value is

$$P_{ideal} = Pr[\chi_k^2/k > D_{ideal}]. \qquad (3.3)$$

This is called ideal because it is based on an infinite number of imputations. This P_{ideal} is taken in LRR as the ideal level that P_m approximates (the subscript *ideal* is used here instead of the original *obs* used by LRR to include the "uncongenial" cases; see Meng, 1994, for detail).

The difference between D_∞ and D_{ideal} was studied by LRR. For example, they found that, under $H_0 : \theta = \theta_0$,

$$Var(D_\infty) = Var(D_{ideal})(1 + C_\xi^2),$$

and

$$Corr(D_\infty, D_{ideal}) = (1 + C_\xi^2)^{-\frac{1}{2}},$$

in which C_ξ^2 is the coefficient of variation for the $\xi_j = 1 + \lambda_j, j = 1, \ldots k$ (the fractions of missing information are actually $1 - 1/\xi_j$), that is ,

$$1 + C_\xi^2 = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1 + \lambda_i}{1 + \bar{\lambda}}\right)^2.$$

LRR also gave similar expressions under alternative hypotheses. Based on both theoretical and simulation studies of level and power, LRR concluded that D_∞ is satisfactory as long as C_ξ is not too big (e.g, $C_\xi \le 40\%$). The loss of power, however, does increase with the value of C_ξ. For example, LRR's simulations show that the maximum relative loss of power is 6% when $C_\xi \le 20\%$, but the loss doubles when $C_\xi = 40\%$. Our attempt here is to see if this loss of power can be recovered by relaxing the restrictive assumption $B_\infty = \bar{\lambda}\bar{U}_\infty$.

4. A Potential Extension of P_m

An obvious approach in relaxing the assumption $B_\infty = \bar{\lambda}\bar{U}_\infty$ is to estimate the individual eigenvalues of $\bar{U}_\infty^{-\frac{1}{2}} B_\infty \bar{U}_\infty^{-\frac{1}{2}}$. This leads to the following procedure, which estimates the eigenvalues of $\bar{U}_\infty^{-\frac{1}{2}} B_\infty \bar{U}_\infty^{-\frac{1}{2}}, \lambda_j, j = 1, \ldots, k$ by the eigenvalues of $\bar{U}_m^{-\frac{1}{2}} B_m \bar{U}_m^{-\frac{1}{2}}, \hat{\lambda}_j, j = 1, \ldots, k$. The four steps of this procedure are:

Step 1. Find $\bar{U}_m^{-\frac{1}{2}}$ and compute

$$\tilde{B}_m \equiv \bar{U}_m^{-\frac{1}{2}} B_m \bar{U}_m^{-\frac{1}{2}}$$

Step 2. Find Γ_m such that

$$\tilde{B}_m = \Gamma_m diag(\hat{\lambda}_1, \ldots, \hat{\lambda}_k)\Gamma_m^\top$$

Step 3. Compute

$$\tilde{\theta}_m = \Gamma_m^\top \bar{U}_m^{-\frac{1}{2}}(\bar{\theta}_m - \theta_0) = (\tilde{\theta}_{m,1}, \ldots, \tilde{\theta}_{m,k})^\top$$

Step 4. Compute (via simulation)

$$P_m^{(e)} = Pr\Big\{\chi_k^2 \ge \sum_{j=1}^{k} \frac{\tilde{\theta}_{m,j}^2}{1 + (1 + \frac{1}{m})\frac{m-1}{\chi_{m-1,j}^2}\hat{\lambda}_j}\Big\}, \quad (4.1)$$

where all χ^2 variables are mutually independent.

As expected, this procedure is theoretically superior to P_m because it converges to the ideal procedure as $m \to \infty$:

Result. $\lim_{m\to\infty} P_m^{(e)} = P_{ideal}$.

Proof: Referring to (2.1) and (4.1), we need only show that

$$\sum_{j=1}^{k} \frac{\tilde{\theta}_{m,j}^2}{1 + (1 + \frac{1}{m})\frac{m-1}{\chi_{m-1,j}^2}\hat{\lambda}_j} \to kD_{ideal}, \qquad \text{as } m \to \infty$$

Because $\frac{m-1}{\chi_{m-1,j}^2} \to 1$ as $m \to \infty$, the left side converges to

$$(\bar{\theta}_\infty - \theta_0)^\top \bar{U}_\infty^{-\frac{1}{2}}[I_k + \bar{U}_\infty^{-\frac{1}{2}} B_\infty \bar{U}_\infty^{-\frac{1}{2}}]^{-1} \bar{U}_\infty^{-\frac{1}{2}}(\bar{\theta}_\infty - \theta_0)$$

as $m \to \infty$. This is kD_{ideal}, because $\bar{U}_\infty + B_\infty = T_\infty$. ∎

We do, however, see several shortcomings of this procedure. First, while $P_m^{(e)}$ is theoretically superior to P_m for large m, it may be less stable for small m because it requires estimating all k eigenvalues instead of just their average. In fact, if $m \le k$, then we do not have enough degrees of freedom to estimate all k eigenvalues, as further discussed in Section 6. Secondly, the computation of $P_m^{(e)}$ involves much more work than the computation of P_m does. We need to calculate $P_m^{(e)}$ by Monte Carlo simulation. However, because we only need to draw k independent χ_{m-1}^2 variates, the simulation is straightforward. Thirdly, $P_m^{(e)}$ is still not the exact Bayesian p-value of (2.1) because the posterior of $(\lambda_1, \ldots, \lambda_k)$ is generally not $(\frac{\hat{\lambda}_1(m-1)}{\chi_{m-1,1}^2}, \ldots, \frac{\hat{\lambda}_k(m-1)}{\chi_{m-1,k}^2})$. In fact, this is only true if $\Gamma_m = \Gamma_\infty$. Nevertheless, this procedure has the virtue of having the correct limit, and at the same time avoids the full integration over $Pr(B_\infty|\mathcal{S}_m)$ required by (2.1). The frequentist properties (e.g., level and power) of this procedure are under investigation.

5. The Exact Bayesian p-value

Since the procedures in Sections 3 and 4 were motivated as an approximation to the Bayesian p-value given in (2.1), we have also considered the direct computation of (2.1). This will not only serve as a standard against which we can check our approximations, but can also be used as a procedure itself.

The exact value of (2.1) depends on the prior for B_∞. Under the standard non-informative prior $\pi(B_\infty) \propto |B_\infty|^{-(k+1)/2}$, (2.1) becomes

$$P(\theta_0|\mathcal{S}_m) = Pr\{\chi_k^2 \geq (\bar{\theta}_m - \theta_0)^\top [\bar{U}_m + (1 + \frac{1}{m})\times (m-1)B_m{}^{\frac{1}{2}} W^{-1} B_m{}^{\frac{1}{2}}]^{-1} (\bar{\theta}_m - \theta_0)\} \qquad (5.1)$$

where $W \sim Wishart_k(m-1, I_k)$. This p-value can be simulated by drawing a large number of Wishart variates, calculating the quantity to the right of the "$\geq$" sign each time. Then, the average tail area of the chi-square distribution with k degrees of freedom is an estimate of the p-value. This is, of course, even more demanding computationally than $D_m^{(e)}$, but is still feasible, especially if it becomes part of a software package.

Our simulation results reveal an interesting phenomenon, that is, for small m, this Bayesian p-value does not perform as well in terms of level and power as P_m of (3.2), which was derived as a simplification and approximation of (5.1)! The reason for this, we believe, is that the "non-informative prior" for B_∞ used in (5.1) is in fact informative, and for small m, there is not enough "data" to correct the artifact in the prior. In contrast, in deriving P_m, LRR directly considered the sampling distribution of D_m when constructing its reference distribution, and thus lessened the impact of the prior. This finding suggests that we need to expand the family of priors in constructing our extensions, and also reinforces the idea that the ultimate criterion for our procedures must be direct frequentist evaluations, like those presented in LRR.

6. A Complication when $m \leq k$

A major difficulty in attempting to drop the assumption of equal fractions of missing information is when the dimension of θ, k, is not less than the number of imputations, m. The problem is that we do not then have enough data or degrees of freedom to estimate the k eigenvalues. In other words, the Wishart distribution in (5.1) would be singular. Thus, the singular Wishart distribution, usually only considered theoretically, becomes important in this application. We are currently looking into this problem.

For $P_m^{(e)}$, when $m \leq k$, we can only estimate the $m-1$ largest eigenvalues of $\tilde{B}_\infty$; at least $k-m+1$ of the $\hat{\lambda}_j$'s are zero because the rank of $\tilde{B}_m$ is at most $m-1$. The question then is how to estimate the remaining $k-m+1$ eigenvalues, and the choice of the prior becomes critical. In deriving $P_m^{(e)}$, we chose a uniform prior on the logarithms of the λ's. In this case, the remaining $k-m+1$ eigenvalues would be set to zero due to the spike in the prior at zero. This would obviously result in a liberal procedure, and to what degree this approximation is acceptable needs to be investigated. One alternative to create a conservative procedure is to set the unestimable eigenvalues to the smallest non-zero eigenvalue of $\tilde{B}_m$. A more reasonable approach would perhaps be to assign a uniform prior for $\lambda_m, \ldots, \lambda_k$ on $\lambda_{m-1} \geq \lambda_m \geq \ldots \geq \lambda_k$.

Acknowledgements

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

References

Li, K. H. (1985) Hypothesis testing in multiple imputation with emphasis on mixed-up frequencies in contingency tables. *Ph. D. Thesis, Department of Statistics, University of Chicago* .

Li, K. H., Meng, X. L., Raghunathan, T. E. & Rubin, D. B. (1991) Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica* **1,** 65-92.

Li, K. H., Raghunathan, T. E. & Rubin, D. B. (1991) Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association* **86,** 1065-1073.

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

Meng, X.L. and Rubin, D.B. (1992) Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79,** 103-111.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D.B. (1995) Multiple imputation after 18 years. *Journal of the American Statistical Association*, to appear.