

Proper and improper multiple imputation

Søren Feodor Nielsen

Department of Statistics and Operations Research

Abstract

Multiple imputation has become viewed as a general solution to missing data problems in statistics. However, in order to lead to consistent asymptotically normal estimators, correct variance estimators and valid tests, the imputations must be *proper*. So far it seems that only Bayesian multiple imputation, i.e. using a Bayesian predictive distribution to generate the imputations, or approximately Bayesian multiple imputations has been shown to lead to proper imputations. In this paper, we shall see that Bayesian multiple imputation does not generally lead to proper multiple imputations. Furthermore, it will be argued that for general statistical use, Bayesian multiple imputation is inefficient even when it is proper.

Keywords: Missing data, multiple imputation, congeniality, efficiency

1 Multiple imputation

Missing –or incomplete– data is a problem to most applications of statistics. It is well-known that we have to be very careful when analysing incomplete data; using *ad hoc* methods often gives the wrong answer. However, many sophisticated methods for handling missing data correctly exist. Lately, multiple imputation has become viewed as a general answer to missing data problems; see for instance the book by Schafer (1997) or the paper by Rubin (1996).

Multiple imputation was originally introduced as a method for handling missing data in surveys. As many other types of data, survey data is prone to missingness. It differs from most other kinds of data, however, in that it often ends up in large databases, where many different users have access to and analyse the data. Naturally, it would be preferred that different users get the same answer if they perform the same statistical analysis on the data. This is ensured when the missing data is imputed. Furthermore, by performing the imputation “at the data base”, the imputer may be able to generate better imputations than the individual user would be able to, using additional information which is not revealed to the public either for confidentiality reasons or simply because it is deemed irrelevant to the user. (Even if irrelevant, it may be very useful for imputing missing data). Finally, by imputing the missing data, the data base will appear to be complete and allow the users to analyse the data using complete data methods. Unfortunately, analysing imputed data as if it was real data generally leads to variance estimates that are too low, confidence intervals which are too narrow, and wrong tests (real significance level above nominal level). Intuitively, this is due to the fact that imputation does not generate information which is not already present in the data; hence the sample size is the same for the incomplete data and for the imputed (“pseudo-complete”) data. Multiple imputation has been suggested as a way of overcoming this problem. By multiply imputing the missing data –i.e. by imputing $m > 1$ values for the missing data– the analyser can perform m complete data analyses and use the results to correct for the variability in the imputations, which differs from the variability in the observed data.

The key to the success of multiple imputation is the concept of *proper* multiple imputation. If the multiple imputations are proper then the average of the estimators is a consistent, asymptotically normal estimator, and an estimator of its asymptotic variance is given by a simple combination of the average of the complete data variance estimators and the empirical variance of the m estimators (the “between imputation variance”) according to “Rubin’s rule” (see Section 2). Rubin (1987) gives a precise definition of proper multiple imputation but for practical purposes it suffices to define proper multiple imputation to be multiple imputations for which “Rubin’s rule” yields a consistent asymptotically normal estimator

of the unknown parameter and a weakly unbiased (see Section 2) estimator of its asymptotic variance in sufficiently regular models.

It seems to be generally believed that imputations drawn from a *Bayesian predictive distribution* are proper when the model used for the imputations and the model used for the analysis are “compatible”. In this paper we will show that such Bayesian imputations are not generally proper even if the two models are “the same”. Furthermore, we will argue that even when the imputations are proper, multiple imputation is a highly inefficient way of handling missing data outside the world of survey data.

2 Large sample results

We start by outlining some large sample results for Bayesian multiple imputation in a case where it is proper. To get the desired asymptotic results we shall use the following result repeatedly:

Lemma 1 *Let $Y = (Y_n)_{n \in \mathbb{N}}$ be a sequence of random variables. Suppose that for a sequence of functions, $(f_n)_{n \in \mathbb{N}}$ and a sequence of d -dimensional real random variables $(Z_n)_{n \in \mathbb{N}}$*

$$\sqrt{n}(Z_n - f_n(Y_1, \dots, Y_n)) \xrightarrow{D} N(0, \Sigma_1) \quad \text{given } Y \text{ for almost every } Y$$

and

$$\sqrt{n}(f_n(Y_1, \dots, Y_n) - \xi_n) \xrightarrow{D} N(0, \Sigma_2)$$

for some $\xi_n \in \mathbb{R}^d$ and $d \times d$ -matrices Σ_1 and Σ_2 . Then

$$\sqrt{n}(Z_n - \xi_n) \xrightarrow{D} N(0, \Sigma_1 + \Sigma_2)$$

Proof: It suffices to show convergence of the characteristic function of $\sqrt{n}(Z_n - \xi_n)$. For any $s \in \mathbb{R}^d$

$$\begin{aligned} & E[\exp(is^t \sqrt{n}(Z_n - \xi_n))] \\ &= E \left[E[\exp(is^t \sqrt{n}(Z_n - f_n(Y_1, \dots, Y_n))) | Y] \right. \\ & \quad \left. \cdot \exp(is^t \sqrt{n}(f_n(Y_1, \dots, Y_n) - \xi_n)) \right] \\ &= \exp(-s^t \Sigma_1 s) E[\exp(is^t \sqrt{n}(f_n(Y_1, \dots, Y_n) - \xi_n))] \\ & \quad + E \left[(E[\exp(is^t \sqrt{n}(Z_n - f_n(Y_1, \dots, Y_n))) | Y] - \exp(-s^t \Sigma_1 s)) \right. \\ & \quad \left. \cdot \exp(is^t \sqrt{n}(f_n(Y_1, \dots, Y_n) - \xi_n)) \right] \\ & \rightarrow \exp(-s^t \Sigma_1 s) \exp(-s^t \Sigma_2 s) + 0 = \exp(-s^t (\Sigma_1 + \Sigma_2) s) \text{ as } n \rightarrow \infty \end{aligned}$$

proving the result. \square

Remark. Lemma 1 is a special case of (a part of) Lemma 1 in Schenker and Welsh (1988); the proof is simpler. It is included to make this paper self-contained.

Let X_1, \dots, X_n be iid random variables (the complete data) with a distribution depending on an unknown parameter $\theta \subseteq \Theta \in \mathbb{R}^d$; the true value is denoted θ_0 . We let $s_X(\theta)$ denote the corresponding score function (the derivative of the log-likelihood) and let $V(\theta) = E[s_X(\theta)^{\otimes 2}]$ denote the expected Fisher information.

Rather than observing X_1, \dots, X_n we assume we observe Y_1, \dots, Y_n . Typically, Y_i is X_i if X_i is observed and an indicator of missingness if X_i is missing. More generally, Y_i could be a random subset of X_i 's sample space, i.e. a coarsening of X_i (see Heitjan and Rubin 1991, Nielsen 2000). For instance, with right censored data Y_i would be $\{X_i\}$ if X_i is observed and $[C_i; \infty[$ if X_i is censored at C_i . We assume that the Y_i s are independent, and that the missing data (or coarsening) mechanism does not depend on θ . Hence the distribution (and the likelihood) of the observed data will in general depend on θ and on an additional (possibly high-dimensional or even non-Euclidean) parameter. Under missing (or coarsening) at random, the observed data likelihood is a product of a factor only dependent on θ and a factor depending on the parameter of the missing data (or coarsening) mechanism. As a consequence, the score function also has a θ -part and a “non- θ ”-part. Thus for estimating θ the “non- θ ”-part may be ignored. We let $s_Y(\theta)$ denote the θ -part of the score function corresponding to the observed data and $I(\theta)$ the expected Fisher information. Without missing (or coarsening) at random, the “non- θ ”-part cannot be ignored; in this case we let θ denote the concatenation of the interest parameter (the parameter indicing the distribution of X_i) and the nuisance parameter from the missing data mechanism.

Finally, we let $s_{X|Y}(\theta) = s_X(\theta) - s_Y(\theta)$ and put $I_Y(\theta) = E[s_{X|Y}(\theta)^{\otimes 2} | Y]$ so that $E I_Y(\theta) = V(\theta) - I(\theta)$.

We assume standard regularity assumptions on both the complete data model and the observed data model. In particular, the observed data MLE, $\hat{\theta}_n$, is a solution to the likelihood equations, $1/n \sum_{i=1}^n s_{y_i}(\theta) = 0$, and it is asymptotically normal with mean θ_0 and variance $I(\theta_0)^{-1}/n$.

By the Bernstein-von Mises theorem, the posterior distribution given the observed data is asymptotically normal with mean equal to the observed data MLE and the inverse Fisher information, $I(\theta_0)^{-1}/n$, as asymptotic variance for any reasonable prior under weak regularity assumptions (see e.g. van der Vaart 1998, Theorem 10.1).

Now, for $j = 1, \dots, m$, let the imputations $(\tilde{X}_{ij})_{i=1, \dots, n}$ be drawn from the Bayesian predictive distribution. In practice, this is usually done by first sampling $\tilde{\theta}_{nj}$ from the posterior distribution and then sampling

\tilde{X}_{ij} independently from the distribution $\mathcal{L}_{\tilde{\theta}_{nj}}(X_i|Y_i = y_i)$ for $i = 1, \dots, n$, $j = 1, \dots, m$; from a theoretical point of view it is not a restriction to assume that this is indeed the case. These m sets of imputations are then used to form the m complete data MLEs, $\hat{\theta}_{nj}$, $j = 1, \dots, m$. Then for each $j = 1, \dots, m$ (under suitable regularity conditions; see appendix for a discussion)

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n s_{\tilde{X}_{ij}}(\hat{\theta}_{nj}) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n s_{\tilde{X}_{ij}}(\hat{\theta}_n) + \sqrt{n} (\hat{\theta}_{nj} - \hat{\theta}_n) \frac{1}{n} \sum_{i=1}^n D_{\theta} s_{\tilde{X}_{ij}}(\hat{\theta}_{nj}^*) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n s_{\tilde{X}_{ij}|y_i}(\hat{\theta}_n) + \frac{1}{\sqrt{n}} \sum_{i=1}^n s_{y_i}(\hat{\theta}_n) \\
&\quad + \sqrt{n} (\hat{\theta}_{nj} - \hat{\theta}_n) \frac{1}{n} \sum_{i=1}^n D_{\theta} s_{\tilde{X}_{ij}}(\hat{\theta}_{nj}^*) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n s_{\tilde{X}_{ij}|y_i}(\tilde{\theta}_{nj}) + \sqrt{n} (\hat{\theta}_n - \tilde{\theta}_{nj}) \frac{1}{n} \sum_{i=1}^n D_{\theta} s_{\tilde{X}_{ij}|y_i}(\tilde{\theta}_{nj}^*) \\
&\quad + \sqrt{n} (\hat{\theta}_{nj} - \hat{\theta}_n) \left(\frac{1}{n} \sum_{i=1}^n D_{\theta} s_{\tilde{X}_{ij}|y_i}(\hat{\theta}_{nj}^*) + \frac{1}{n} \sum_{i=1}^n D_{\theta} s_{y_i}(\hat{\theta}_{nj}^*) \right)
\end{aligned}$$

where $\hat{\theta}_{nj}^*$ ($\tilde{\theta}_{nj}^*$) is a point on the line from $\hat{\theta}_n$ to $\hat{\theta}_{nj}$ ($\tilde{\theta}_{nj}$).

It now follows that given the observed data and $\tilde{\theta}_{nj}$

$$\sqrt{n} (\hat{\theta}_{nj} - \hat{\theta}_n) - F(\theta_0)^t \sqrt{n} (\tilde{\theta}_{nj} - \hat{\theta}_n) \xrightarrow{\mathcal{D}} N(0, V(\theta_0)^{-1} E I_Y(\theta_0) V(\theta_0)^{-1})$$

for each $j = 1, \dots, m$. Hence, given the observed data only,

$$\sqrt{n} (\hat{\theta}_{nj} - \hat{\theta}_n) \xrightarrow{\mathcal{D}} N(0, F(\theta_0)^t I(\theta_0)^{-1} F(\theta_0) + V(\theta_0)^{-1} E I_Y(\theta_0) V(\theta_0)^{-1})$$

by Lemma 1. Moreover, given the observed data, $\hat{\theta}_{n1}, \dots, \hat{\theta}_{nm}$ are independent. Straightforward manipulations (multiply by $V(\theta_0)$ from both sides) yield

$$F(\theta_0)^t I(\theta_0)^{-1} F(\theta_0) + V(\theta_0)^{-1} E I_Y(\theta_0) V(\theta_0)^{-1} = I(\theta_0)^{-1} - V(\theta_0)^{-1}$$

Putting $\bar{\theta}_n = \frac{1}{m} \sum_{j=1}^m \hat{\theta}_{nj}$, continuous mapping and the partitioning theorem give us (conditional on the observed data)

$$\sqrt{n} (\bar{\theta}_n - \hat{\theta}_n) \xrightarrow{\mathcal{D}} N\left(0, \frac{1}{m} (I(\theta_0)^{-1} - V(\theta_0)^{-1})\right) \quad (1)$$

and

$$\frac{n}{m-1} \sum_{j=1}^m (\hat{\theta}_{nj} - \bar{\theta}_n)^{\otimes 2} \xrightarrow{\mathcal{D}} \text{Wishart}\left(m-1, \frac{1}{m-1} (I(\theta_0)^{-1} - V(\theta_0)^{-1})\right) \quad (2)$$

which are asymptotically independent. Notice also that the asymptotic distribution of $\frac{1}{m-1} \sum_{j=1}^m n \left(\hat{\theta}_{nj} - \bar{\theta}_n \right)^2$ does not depend on the observed data. Thus (2) also holds unconditionally. Finally, (1) and Lemma 1 gives us

$$\sqrt{n} (\bar{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N \left(\mathbf{0}, I(\theta_0)^{-1} + \frac{1}{m} (I(\theta_0)^{-1} - V(\theta_0)^{-1}) \right) \quad (3)$$

Let \hat{V}_n be a (consistent) estimator of the complete data Fisher information based on the complete data. Let \hat{V}_{nj} be the same estimator based on the observed data and the j th set of imputations. Then

$$\begin{aligned} \hat{\Sigma}_n &= \frac{1}{m} \sum_{j=1}^m \hat{V}_{nj}^{-1} + \left(1 + \frac{1}{m} \right) \frac{n}{m-1} \sum_{j=1}^m \left(\hat{\theta}_{nj} - \bar{\theta}_n \right)^{\otimes 2} \\ &\rightsquigarrow V(\theta_0)^{-1} + \left(1 + \frac{1}{m} \right) (I(\theta_0)^{-1} - V(\theta_0)^{-1}) \quad \text{as } n \rightarrow \infty \\ &= I(\theta_0)^{-1} + \frac{1}{m} (I(\theta_0)^{-1} - V(\theta_0)^{-1}) \end{aligned} \quad (4)$$

estimates the asymptotic variance of $\bar{\theta}_n$. The notation “ $Z_n \rightsquigarrow \alpha$ ” is used to denote that Z_n converges in distribution and that the limiting distribution has mean α . We will call such an estimator *weakly unbiased* for α . Indeed it should be noted that $\hat{\Sigma}_n$ is not consistent (as $n \rightarrow \infty$); rather its asymptotic distribution is the Wishart distribution given in (2) shifted (by $I(\theta_0)^{-1}$) so that it has the correct mean. To account for the variability in the variance estimator it is generally recommended that Wald type tests and confidence intervals are based on a t- (or F-) distribution with degrees of freedom based on a Satterthwaite approximation (see e.g. Schafer 1997) rather than a normal distribution. One should be aware that the asymptotic distribution of the Wald test statistic is non-standard and non-similar: As $I(\theta_0) \rightarrow V(\theta_0)$, corresponding to no missing information, the distribution of the Wald test statistic tends to a χ^2 -distribution.

Basically “Rubin’s Rule” boils down to this: Impute m datasets and perform m complete data analyses to get m estimators of θ and m corresponding variance estimators. Average the estimators of the unknown parameters to get the point estimate (cf. (3)) and estimate the variance matrix by the average of the variance estimators from the complete data analyses and $1 + 1/m$ times the empirical variance of the m estimators of θ (cf. (4)).

If the imputations are proper, then “Rubin’s rule” yields a weakly unbiased estimator of the variance. The argument above shows that Bayesian multiple imputations are proper, when the complete data estimator is the complete data MLE.

Remark. The argument given above leading to the large sample justification of Bayesian multiple imputation may be seen as a more explicit

formulation of the argument given by Rubin (1987, Section 4.5). Robins and Wang (2000) give a more general argument.

In many practical examples, one would be interested in less than the full complete data model. For instance, X_i could be both outcome and covariates; here the interest would be on the conditional (regression) model of the outcome given the covariates. If the covariates are missing, we need the full model to impute but in practice we would only estimate the conditional model. If θ can be written as (θ_1, θ_2) , where θ_1 is the parameter of the conditional model of interest and θ_2 the parameter of the marginal model of the covariates, and the parameters are variation independent, then the result above shows that Bayesian multiple imputation are still proper for estimating θ_1 (or θ_2) using a conditional (or a marginal) maximum likelihood estimator. That this is not the case when the parameters are not variation independent will be indicated in Section 3.2 below.

Also it is worth noting that the iid structure assumed above is not necessary (nor used) in the derivations. Thus the result also holds for non-iid observations (given sufficient regularity of course).

3 Counter examples

If a conclusion is to be drawn from the results in Section 2, clearly it is that multiple imputation works nicely when we stay within a maximum likelihood framework. We do not, however, have to look very far for an example where Bayesian multiple imputations fail at being proper.

3.1 First example

Let X_1, \dots, X_n be independent identically Gamma distributed with shape parameter λ and intensity (inverse scale) θ . Suppose that λ is known and larger than 2. Each X_i is missing completely at random with probability $1 - p$; i.e. each X_i is observed with probability p independently of X_1, \dots, X_n . Suppose that only N of the n X_i s are observed; without loss of generality we assume it to be the first N . Note that $N/n \xrightarrow{P} p$.

The MLE of θ based on the observed data is $\hat{\theta}_n = \lambda/\bar{X}$, where \bar{X} is the average of the observed X_i s. The asymptotic distribution of $\hat{\theta}_n$ is $N(\theta_0, \frac{1}{pn} \frac{\theta_0^2}{\lambda})$.

Assuming a Gamma prior distribution with parameters (α, β) , the posterior is again a Gamma distribution with parameters $(\alpha + N\lambda, \beta + N\bar{X})$. We now let $\tilde{\theta}_{nj}, j = 1, \dots, m$ be drawn from the posterior; note that (given the observed data)

$$\sqrt{n} \left(\tilde{\theta}_{nj} - \hat{\theta}_n \right) \rightarrow N \left(0, \frac{1}{p} \frac{\theta_0^2}{\lambda} \right). \quad (5)$$

Finally, let the imputations \tilde{X}_{ij} be drawn independently for $i = N+1, \dots, n$ from the Gamma distribution with shape λ and intensity $\tilde{\theta}_{nj}$ for $j = 1, \dots, m$.

Suppose now that rather than using the MLE as our complete data estimator we use $1/n \sum_{i=1}^n (\lambda - 1)/X_i$. In the complete data model, it is unbiased and asymptotically normal with variance $1/n \cdot \theta_0^2/(\lambda - 2)$. Hence

$$\theta_{nj}^* = \frac{1}{n} \sum_{i=1}^N (\lambda - 1)/X_i + \frac{1}{n} \sum_{i=N+1}^n (\lambda - 1)/\tilde{X}_{ij}.$$

Let $\theta_n^* = 1/N \sum_{i=1}^N (\lambda - 1)/X_i$ be the corresponding estimator based on the observed data only (the complete case estimator). Then

$$\begin{aligned} \sqrt{n}(\theta_{nj}^* - \theta_n^*) &= \frac{1}{\sqrt{n}} \sum_{i=N+1}^n \left(\frac{\lambda - 1}{\tilde{X}_{ij}} - \theta_n^* \right) \\ &= \sqrt{\frac{n-N}{n}} \frac{1}{\sqrt{n-N}} \sum_{i=N+1}^n \left(\frac{\lambda - 1}{\tilde{X}_{ij}} - \tilde{\theta}_{nj} \right) \\ &\quad + \frac{n-N}{n} \sqrt{n} (\tilde{\theta}_{nj} - \hat{\theta}_n) + \frac{n-N}{n} \sqrt{n} (\hat{\theta}_n - \theta_n^*) \end{aligned}$$

Going through arguments similar to those given in Section 2 (using first the Lindeberg central limit theorem conditional on the data and $\tilde{\theta}_{nj}$ and then (5)), we see that

$$\sqrt{n}(\theta_{nj}^* - \hat{\theta}_n) + \frac{N}{n} \sqrt{n}(\hat{\theta}_n - \theta_n^*) \xrightarrow{\mathcal{D}} N \left(0, \frac{(1-p)^2 \theta_0^2}{p \lambda} + (1-p) \frac{\theta_0^2}{\lambda - 2} \right)$$

conditional on the observed data. Using the central limit theorem and the delta method, it is straightforward to show that

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}_n - \theta_0) \\ \sqrt{n}(\hat{\theta}_n - \theta_n^*) \end{bmatrix} \xrightarrow{\mathcal{D}} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{p} \frac{\theta_0^2}{\lambda} & 0 \\ 0 & \frac{1}{p} \left(\frac{\theta_0^2}{\lambda - 2} - \frac{\theta_0^2}{\lambda} \right) \end{bmatrix} \right)$$

Thus the multiple imputation estimator $\bar{\theta}_n = \frac{1}{m} \sum_{j=1}^m \theta_{nj}^*$ is asymptotically normal with mean θ_0 and variance

$$\frac{1}{n} \left(\frac{1}{p} \frac{\theta_0^2}{\lambda} + p \left(\frac{\theta_0^2}{\lambda - 2} - \frac{\theta_0^2}{\lambda} \right) + \frac{1}{m} \left(\frac{(1-p)^2 \theta_0^2}{p \lambda} + (1-p) \frac{\theta_0^2}{\lambda - 2} \right) \right) \quad (6)$$

The variance estimator is

$$\begin{aligned} &\frac{1}{m} \sum_{j=1}^m \frac{\theta_{nj}^{*2}}{\lambda - 2} + \left(1 + \frac{1}{m} \right) \frac{n}{m-1} \sum_{j=1}^m (\theta_{nj}^* - \bar{\theta}_n)^{\otimes 2} \\ &\rightsquigarrow \frac{\theta_0^2}{\lambda - 2} + \left(1 + \frac{1}{m} \right) \left(\frac{(1-p)^2 \theta_0^2}{p \lambda} + (1-p) \frac{\theta_0^2}{\lambda - 2} \right) \end{aligned} \quad (7)$$

and the difference between (n times) (6) and (7) is

$$2(1-p) \left(\frac{\theta_0^2}{\lambda-2} - \frac{\theta_0^2}{\lambda} \right) > 0 \quad (8)$$

Thus the imputations are not proper in this example. Or rather, they are not proper for this estimator. Clearly, if we had used the complete data MLE as our estimator rather than the inefficient but unbiased estimator, the imputations would have been proper by the results of Section 2. Thus “being proper” is not a property, which is independent of the subsequent analysis.

Though the imputations are improper in the example above, at least the estimator of the variance is *larger* than the asymptotic variance. Hence Wald-type tests will be conservative, and confidence intervals will be too large, but at least they will be valid. Meng (1994, MAIN RESULT) gives sufficient conditions for this to be the case. However, as the following example shows we might as well get a variance estimator, which is too small.

3.2 Second example

Let Y_1, \dots, Y_n be a random sample from a Gaussian distribution with mean θ and variance 1. Suppose that a second sample, X_1, \dots, X_n from a Gaussian distribution with the same unknown mean but a different, yet known, variance σ^2 is not observed, but that we (rather foolishly) wish to incorporate it in our estimation of θ . Here our imputations are $\tilde{X}_{ij} = \tilde{\theta}_{nj} + \sigma \varepsilon_{ij}$ where the ε_{ij} s are iid standard normal random variables and $\tilde{\theta}_{nj}$ are drawn independently for $j = 1, \dots, m$ from a normal distribution with a mean equal to $\sum_{i=1}^n Y_i / (1+n)$ and variance equal to $1/(1+n)$; this corresponds to using a standard normal prior for θ . Taking the average of the observed Y_i s and the imputed \tilde{X}_{ij} s (for each $j = 1, \dots, m$) as our estimator, we get

$$\begin{aligned} \sqrt{n}(\bar{\theta}_n - \theta_0) &\sim N\left(\frac{\sqrt{n}}{n+1}, \left(\frac{n}{n+1}\right)^2 + \frac{1}{4m} \left(\frac{n}{n+1} + \sigma^2\right)\right) \\ &\xrightarrow{\mathcal{D}} N\left(0, 1 + \frac{1}{m} \frac{1 + \sigma^2}{4}\right) \end{aligned} \quad (9)$$

But the estimator of the variance using “Rubin’s rule” has expectation

$$\frac{1 + \sigma^2}{4} + \left(1 + \frac{1}{m}\right) \frac{1 + \sigma^2}{4} + O\left(\frac{1}{n}\right), \quad (10)$$

so that for n large the difference is $(1 - \sigma^2)/2$. In particular, if $\sigma^2 < 1$, then the estimator of the variance is systematically smaller than the true variance. Hence, if $\sigma^2 < 1$, we do not obtain asymptotically valid tests,

and our confidence intervals will be too short. For $\sigma^2 > 1$ the variance is overestimated. Only for $\sigma^2 = 1$ do we get a weakly unbiased estimator; of course, in this case our complete data estimator is maximum likelihood.

We get similar results if we use a marginal maximum likelihood estimator, the average of the X_i s, as our complete data estimator, indicating that we cannot extend the large sample results from Section 2 to estimating a sub-parameter which is not variation independent of the rest of θ .

Remark. It is not essential for the result (merely convenient for the calculations) to assume the variances of X_i and Y_i to be known; we get exactly the same (asymptotic) result if we use, say, independent inverse χ^2 -priors for the variances. Of course, if the variance of X_i is unknown we would be very foolish indeed to try to impute the missing data.

3.3 Simulations

To get a slightly more interesting example we generalise the complete data model in Subsection 3.2 to have an unknown variance matrix

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \sim N \left(\begin{bmatrix} \theta \\ \theta \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho \\ \rho & \sigma_y^2 \end{bmatrix} \right)$$

Thus we have two measurements on each individual with different precision, and we wish to estimate the population mean. Furthermore, we let X_i be missing at random, rather than completely at random. Thus whether X_i is observed or missing is allowed to depend on Y_i . If the variance of Y_i is large compared to the variance of X_i it will be highly relevant to incorporate the observed X_i s in the estimation of θ . However, since X_i is not missing completely at random we have to do something intelligent. Let us try imputing from a predictive distribution.

To specify a Bayesian imputation model, we re-parametrise $(\theta, \sigma_x^2, \rho, \sigma_y^2)$ to $(\theta, \sigma^2, \beta, \tau^2)$ where $\beta = \rho/\sigma_y^2$ is the regression coefficient in the conditional model of X_i given Y_i , $\tau^2 = \sigma_x^2 - \rho^2/\sigma_y^2$ is the conditional variance, and $\sigma^2 = \sigma_y^2$. A prior distribution is chosen as standard normals for θ and β and inverse χ^2 distributions with parameters $(2, 3/2)$ for σ^2 and τ^2 ; all components are a priori independent. The posterior (and hence the predictive distribution) becomes non-standard, so we simulate instead of giving exact results.

For illustration we simulate 2500 data sets of sample size 200 with parameters $(\theta, \sigma^2, \rho, \sigma_x^2) = (1, 1.1, 0.99, 1)$. X_i is missing if $|Y_i - 1|$ is large so that the overall probability of missingness is approximately 0.2. The imputations are generated by $m = 5$ independent Gibbs samplers run for 1000 iterations. The results show that for any practical purpose the multiple imputation estimator is normally distributed with mean 1 and variance 0.0072; see Figure 1. A Kolmogorov-Smirnoff test accepts the hypothesis of normality with a p-value of 77.2%. The variance estimator

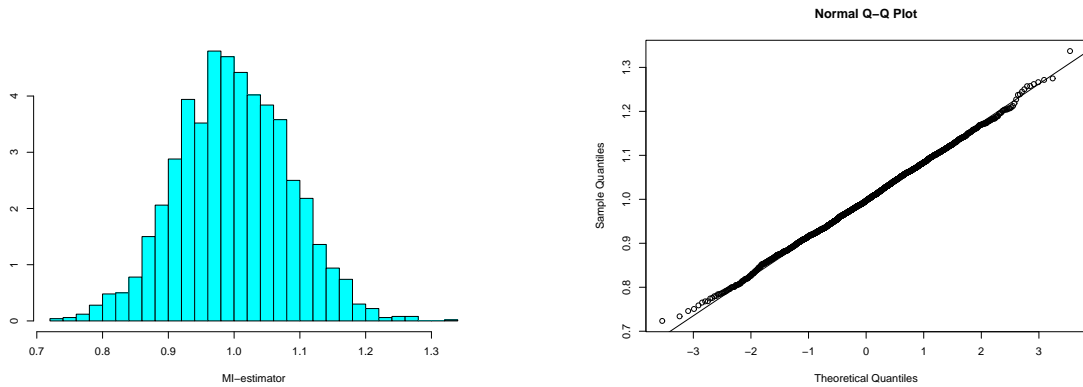


Figure 1: Simulated distribution of the MI-estimator

is however too small as seen from Figure 2: The true variance (indicated by the vertical line in Figure 2) is the 95% quantile of the (simulated) distribution of the variance estimator.

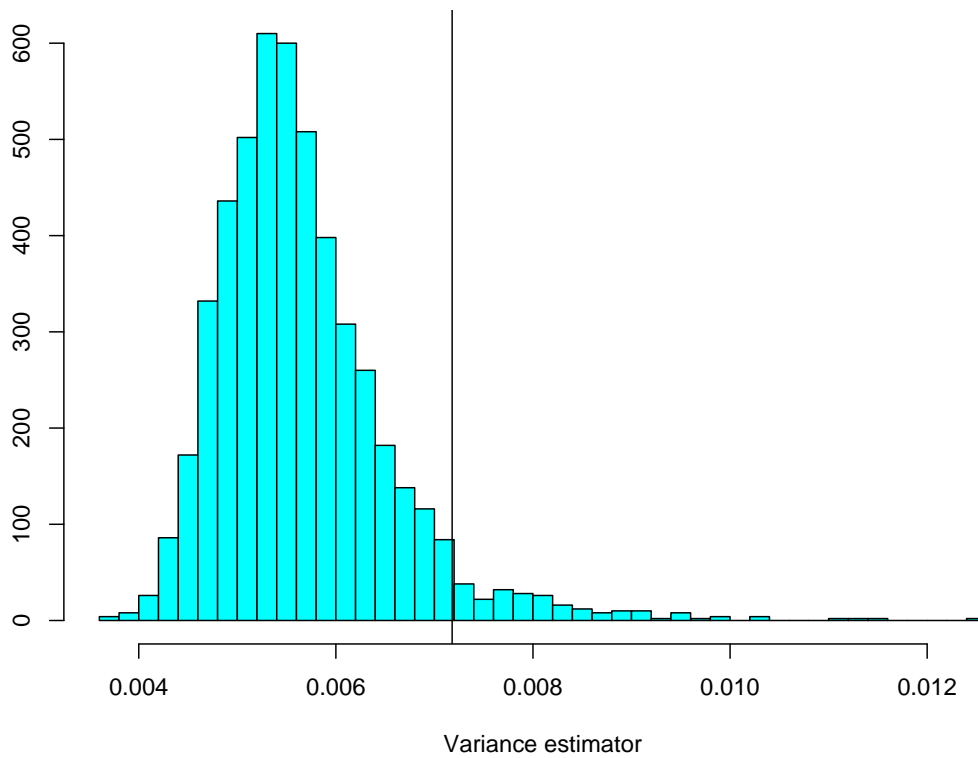


Figure 2: Simulated distribution of the variance estimator. The true variance is indicated by the vertical line.

To be fair it must be added that these results are highly dependent on the chosen values of the parameters. For other parameter values the result is the reverse: The variance estimator overestimates the true variance. Keeping the simple example of Section 3.2 in mind, this is not unexpected. The conclusion from this observation is not heartening though: Since we cannot in general see if the procedure we use over- or under-estimates the variance, we really cannot trust it at all.

It is worth noticing that the confidence intervals based on the variance estimator and the t-distribution approximation are, though wrong, not as bad as the variance results would suggest: A nominal 95% confidence interval has 91.3% coverage. This appears to be due to the fact that when the variance is underestimated, then the approximate degrees of freedom are underestimated too resulting in longer confidence intervals. However, “two wrongs do not make a right”; the confidence intervals are still too short.

Remark. In practice it will be tempting (and simpler) to fit a Bayesian model to the conditional distribution of X given Y since this is what we need to impute the missing data. This corresponds to imputing from a Bayesian predictive distribution derived using an unrestricted bivariate Gaussian distribution for the data (“imputing from a larger model”). As one would expect this gives very similar results as the full model used above. In particular, the variance of the estimator of θ is underestimated.

4 A note on uncongeniality

Meng (1994) discusses a concept he calls “(un-)congeniality”. He defines an “analysis procedure” consisting of an estimator of the unknown parameter θ based on the observed data and an estimator of θ based on the complete data both with associated variance estimators to be *congenial* to an imputation procedure if

1. the imputations come from a Bayesian predictive distribution
2. the observed (complete) data estimator asymptotically equals the posterior mean of θ given the observed (complete) data and the associated variance estimator asymptotically equals the posterior variance of θ given the observed (complete) data.

Meng (1994) shows that congeniality implies properness in the sense that the variance estimator derived from “Rubin’s rule” is consistent as $m \rightarrow \infty$. He does not explicitly consider the asymptotic distribution of the estimator, and his treatment of congeniality is restricted to multiple imputation with $m = \infty$.

In order to discuss asymptotic results it seems fruitful to extend the asymptotic equivalence of the estimators of θ to be such that two estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ are asymptotically equal if $\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) \xrightarrow{P} 0$. Since the

posterior mean and variance asymptotically equals the MLE and inverse Fisher information in sufficiently regular models, the argument given in Section 2 actually verifies Meng's (1994) result for fixed finite m , if consistency of the variance estimator is replaced by it being weakly unbiased. Moreover, we see that (with our strengthening of Meng's (1994) concept) the following result is true in sufficiently regular models:

Result 1 *An analysis procedure is congenial to an imputation procedure if and only if the complete data and observed data estimators are efficient (i.e. maximum likelihood) and their associated variance estimators estimate the corresponding inverse Fisher informations.*

Hence, if our complete and observed data estimators are efficient, our Bayesian multiple imputations are proper in the sense adopted for this paper. Of course, they may be proper even if the analysis procedure is not congenial to the imputation procedure. As the asymptotic results of Section 2 show, if the complete data estimator is efficient, the imputations derived from a corresponding Bayesian predictive distribution are proper, regardless of whatever observed data estimator we have in mind. Meng's (1994) insistence on an observed data estimator seems to be based on his intention of comparing the multiple imputation estimator to an estimator based on the observed data. It makes a further discussion of the results reported in this paper difficult to relate to his concepts. However, we note that in our examples, the analysis procedure is uncongenial to the imputation procedure.

Schenker and Welsh (1988) discuss multiple imputation in a linear regression setting with missing outcomes. The noise is only restricted to be iid zero mean random variables with finite variance (so that their model is really semi-parametric), but they impute using a Bayesian predictive distribution calculated under the assumption that the noise is indeed normal. Their complete data estimator is the usual least squares estimator. Clearly, the distribution of the noise could be chosen such that the complete data estimator is not efficient but their results indicate that the Bayesian multiple imputations are proper in our sense. Of course, one might argue that since their result is due to the fact that they rely on first and second moments and their calculations are exactly as they would be in the case of Gaussian errors, where the least squares estimator is maximum likelihood, this does not really give much of a counterexample to the need for efficient complete data estimators. Furthermore, if the imputations were made assuming a non-Gaussian distribution of the errors, the variance estimator of the least squares estimator obtained using "Rubin's rule" would typically be wrong. Alternatively, one might argue that their estimator is efficient in the semi-parametric model Schenker and Welsh (1988) consider, but that would be missing the point: Their imputations are proper even if the errors are assumed to be double exponential as long as we use the least squares estimator and the Gaussian

imputation model. We see again that being proper is as much related to the complete data estimator as to the model for the data.

Though there may be examples – the example considered by Schenker and Welsh (1988) could be considered one such example– where Bayesian multiple imputations are proper even though the complete data estimator is not efficient, the examples given in Section 3 suggest that this would be the exception rather than the rule. Thus, to be on the safe side we must insist that our complete data estimators are efficient.

5 Inefficiency

The insistence on efficiency is sound statistical advice with or without missing data. But a multiple imputation estimator is clearly not efficient: It is a simulation based estimator, and simulation introduces noise. Obviously, there is little point in using a multiple imputation estimator if an efficient estimator (based on the observed data) can be found. Thus criticising multiple imputation estimators for being inefficient when compared to the observed data MLE is hardly fair. Moreover, (3) tells us that we can make the multiple imputation estimator based on the complete data MLE almost efficient by picking m sufficiently large.

However, as Wang and Robins (1998) discuss in detail, the inconsistent variance estimator leads to much broader confidence intervals (weaker tests) than a consistent variance estimator would do. Thus even though the estimator is almost efficient (for m large), the inference is inefficient. Robins and Wang (2000) suggest a consistent variance estimator; it is, however, more complicated to calculate than the one obtained from “Rubin’s rule”. When using the complete data MLE as the complete data estimator, a simple estimator is available (Wang and Robins 1998): Choose $m' \in \{1, \dots, m-1\}$; then (assuming sufficient regularity)

$$\hat{I} = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{m'} \sum_{j=1}^{m'} s_{\tilde{X}_{ij}}(\tilde{\theta}_{nj}) \right)^t \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{m-m'} \sum_{j=1+m'}^m s_{\tilde{X}_{ij}}(\tilde{\theta}_{nj}) \right)$$

is a consistent estimator of the observed data information, and

$$\hat{I}^{-1} + \frac{1}{m} \left(\hat{I}^{-1} - \frac{1}{m} \sum_{j=1}^m \hat{V}_{nj}^{-1} \right) \quad (11)$$

is a consistent estimator of the variance of $\bar{\theta}_n$ (cf (3)).

In practice the multiple imputation estimator is often inefficient in a much more direct sense. To be able to impute from a Bayesian predictive distribution we –more or less– need to be able to simulate from the posterior distribution of θ . In all but trivial examples, to draw from the

posterior we need a Markov chain method such as a Gibbs sampler. In order to get m independent imputations, we need to run our Gibbs sampler m times independently to convergence. However, the average of the last k iterations of the m chains will provide a much better estimator than multiple imputation; here the additional variance due to simulations goes down as $1/(mk)$. Actually, to find the posterior mean only one chain is necessary, though m independent chains may be useful for checking convergence (Gelman 1996). This single chain may then be run m times as long making more efficient use of the simulation budget. Alternatively, the simulation noise can be reduced to whatever level desired by running the m chains further (i.e. increasing k). In practice, this would typically be a limited extension of the simulation budget. To reduce the simulation noise of the multiple imputation estimator, more chains run until convergence are needed. In practical implementation of multiple imputation, one may choose to run just one Markov chain and choose the $\tilde{\theta}_{nj}$ s far apart, so that they are approximately independent. Clearly, a more efficient use of this single chain would be to find the posterior mean, i.e. the average of the chain. Consequently, in most practical examples there will be an estimator more efficient than the multiple imputation estimator which may be obtained by the same amount of simulation (or at worst negligible more).

If we can draw directly from the posterior distribution, multiple imputation might be quite efficient. If our complete data estimator is maximum likelihood, the multiple imputation estimator will be more efficient than the average of the simulated $\tilde{\theta}_{nj}$ s, since $1/m \sum_{j=1}^m \tilde{\theta}_{nj}$ is asymptotically normal with mean θ_0 and variance $(1 + 1/m)I(\theta_0)^{-1}/n$ (use Lemma 1), which is larger than the asymptotic variance found in (3). However, if the aim is to reduce the simulation part of the variance, it may well be faster to use the posterior mean estimator (with a slightly larger m) than to do the extra work necessary to simulate from the predictive distribution and find a complete data estimator, especially if an iterative procedure is needed in order to find this.

The final case, where drawing directly from the predictive distribution is possible, seems to be rare in practice. Especially if we rule out examples so simple that we can find an efficient estimator directly. However if such a problem should occur in practice, multiple imputation may be the solution.

6 Practical implementation

For practical application of the methods discussed in the previous section it is important that software implementing them exist. Indeed, the wealth of free and commercial software available for multiple imputation (see www.multiple-imputation.com for a list) makes multiple imputation easy and tempting to use. Keeping our examples in mind, one should be

careful when using this software that it actually imputes from the model one intends to analyse. Much of the software only allows a very limited choice of imputation models even if the models offered are flexible. Imputing from an incorrect model will generally lead to wrong (asymptotically biased) estimators: Restricting attention to estimators derived from an estimating equation, we need the imputations to be such that the estimating equation is approximately unbiased to obtain consistent estimators. Of course, as the examples clearly indicate, for the variance estimator to be reasonable, we need much more.

On the other hand, free and flexible software for implementing Bayesian models (BUGS/WinBUGS, www.mrc-bsu.cam.ac.uk/bugs/) exists making the posterior mean estimator a feasible and attractive alternative.

7 Conclusion

When it comes to the examples given in Section 3, it is important to notice that all the methods used are “correct”: The complete data estimators would lead to consistent asymptotically normal estimators with complete data, and the Bayesian models are based on the correct likelihood and sensible priors. Hence, our examples are –unlike most examples on how multiple imputation may fail found in the literature– not examples of the imputer assuming/knowing more or less than the data analyst. Indeed we think of the imputer and the analyst as being the same person as would be common in most statistical applications outside analysis of large public-use databases.

The examples are very simple. Indeed, when it comes to the first two examples the observed data MLE can be written down explicitly and hopefully nobody would even consider using multiple imputation or anything else to estimate the unknown parameter. The third example is slightly more complicated. Though a consistent normally distributed estimator –the average of the Y_i s– is easily obtained, the facts that only 20% of the X_i s are missing and that the X_i s have smaller variance, makes it tempting to improve on the simple but very inefficient estimator. Missing at random rather than completely at random rules out simple solutions such as complete case analysis. Finally, the unrestricted variance matrix makes exact maximum likelihood difficult both in the observed data case and the complete data case, leading us to an inefficient complete data estimator. This being said, it is not too difficult to guess a better complete data estimator, to find an efficient two-step estimator, or even to find the MLE by iterative means. On the other hand, inefficient estimators (for instance based on estimating equations or quasi-likelihood) are frequently used in practice when models get more complicated than the ones considered in the examples here.

Whereas it may be argued that none of the examples are “realistic”, the implications of these simple examples cannot be denied: Unless we

use an efficient complete data estimator, the variance estimator derived from “Rubin’s rule” will be asymptotically biased. Furthermore, the bias may be upwards, leading to inefficient but correct inference, as well as downwards, leading to incorrect inference. Moreover, the third example indicates that in practice it may be impossible to judge whether the bias is upwards or downwards.

As discussed in Section 5, even if we use an efficient complete data estimator, the multiple imputation procedure is inefficient: The inconsistency of the variance estimator leads to weaker tests than a consistent variance estimator (which in principle can be obtained) would. Even though a consistent variance estimator can be constructed, we should keep in mind that a more efficient estimator based on approximately the same amount of (simulation) work can be found.

The discussion so far has been given from a frequentist point of view. From a Bayesian point of view it would seem strange first to find the posterior distribution by simulation and then impute the missing data to find a point estimate and forget about the posterior. Clearly to be able to find the posterior distribution answers our Bayesian needs.

The situation is more complicated when it comes to large public-use data bases. Here multiple imputation may be the best strategy: The imputer uses her or his knowledge to create better imputations than the analyser is able to, and an analyser, using a correct model and an efficient complete data estimator, obtains consistent asymptotically normal estimators and –using the variance estimator given by (11)– acceptable confidence intervals and tests. Or the analyser ignores the imputations provided and creates his or her own; multiple imputations (unlike single imputation) makes it obvious which observations are actually imputed and which are observed. Also the multiple imputations carries a little information on what the imputer thought was important to correct for when creating the imputations even if this information may be difficult to quantify. It must also be acknowledged that for public-use data bases imputation is necessary; it cannot be expected that all users are able to treat missing data correctly. Here multiple imputation allows the user to perform correct analyses as long as (s)he stays within certain limits.

The findings discussed above suggest that for general statistical practice multiple imputation is not the solution to missing data problems. Of course, we should keep in mind that we have only considered imputations derived from a Bayesian predictive distribution. Thus there may be other ways of generating proper multiple imputations and some of these may yield more satisfying results. However, so far only Bayesian or approximately (cf Rubin 1987, p. 124) Bayesian imputation schemes have been shown to be proper in some settings. We close with an exam-

ple of an approximately Bayesian imputation method: Find the observed data MLE, $\hat{\theta}_n$, and a consistent estimator of the Fisher information, \hat{I} . Construct multiple imputations by drawing \tilde{X}_{ij} from the conditional distribution of X_i given Y_i using $\tilde{\theta}_{nj}$ drawn from $N(\hat{\theta}_n, \hat{I}^{-1}/n)$ as the parameter. The argument given in Section 2 shows that this gives proper imputations when the complete data estimator is the MLE. But would anyone use this estimator instead of the observed data MLE? Hopefully not. Do we?

References

- Gelman, A. (1996) Inference and monitoring convergence in W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov chain Monte Carlo in practice* Chapman & Hall.
- Heitjan, D. F. and Rubin, D. B. (1991) Ignorability and coarse data *Annals of Statistics* **19**, 2244–2253.
- Meng, X.-L. (1994) Multiple-imputation inferences with uncongenial sources of input *Statistical Science* **9**, 538–558.
- Nielsen, S. F. (2000) Relative coarsening at random *Statistica Neerlandica* **54**, 79–99.
- Robins, J. M. and Wang, N. (2000) Inference for imputation estimators *Biometrika* **87**, 113–124.
- Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys* Wiley.
- Rubin, D. B. (1996) Multiple imputation after 18+ years *Journal of the American Statistical Association* **91**, 473–489.
- Schafer, J. (1997) *Analysis of Incomplete Multivariate Data* Chapman & Hall.
- Schenker, N. and Welsh, A. H. (1988) Asymptotic results for multiple imputation *Annals of Statistics* **16**, 1550–1566.
- van der Vaart, A. W. (1998) *Asymptotic Statistics* Cambridge University Press.
- Wang, N. and Robins, J. M. (1998) Large-sample theory for parametric multiple imputation procedures *Biometrika* **85**, 935–948.

A Regularity assumptions

Regularity assumptions necessary for the calculations given in Section 2 may take many forms. The regularity conditions suggested here are not supposed to be technically exhaustive but merely suggestive of what may be needed:

1. “Standard regularity assumptions” on the complete and observed data models, so that the MLE is asymptotically linear and efficient in both models.
2. The draws from the posterior are asymptotically normal, i.e. given the observed data

$$\sqrt{n} \left(\tilde{\theta}_{nj} - \hat{\theta}_n \right) \xrightarrow{\mathcal{D}} N \left(0, I(\theta_0)^{-1} \right).$$

Often this follows from the same regularity conditions used to obtain the asymptotic distribution of the observed data MLE.

3. With $\tilde{X}_i \sim \mathcal{L}_{\theta_n}(X_i | Y_i = y_i)$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n s_{\tilde{X}_i | y_i}(\theta_n) \xrightarrow{\mathcal{D}} N \left(0, EI_Y(\theta_0)^{-1} \right)$$

for any sequence θ_n with $\sqrt{n}(\theta_n - \theta_0)$ bounded and (almost) every sequence of y_i s. In practice, a Lindeberg condition should be verified.

4. Uniform laws of large numbers hold for

$$\frac{1}{n} \sum_{i=1}^n D_{\theta} s_{\tilde{X}_i | y_i}(\theta) \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n D_{\theta} s_{Y_i}(\theta)$$

on a compact set of θ with θ_0 as an inner point; here $\tilde{X}_i \sim \mathcal{L}_{\theta_n}(X_i | Y_i = y_i)$ as above. In practice this could be verified from additional smoothness (a Lipschitz condition) or empirical process techniques.