RELEASING MULTIPLY-IMPUTED SYNTHETIC DATA GENERATED IN TWO STAGES TO PROTECT CONFIDENTIALITY

Jerome P. Reiter and Jörg Drechsler

Duke University and Institute for Employment Research

Supplementary Material

This supplement contains the derivations of the inferential methods for two stage synthetic data. Section S1 describes the notation used in the derivations. Section S2 presents the derivation of the inferential methods for two-stage, fully synthetic data, which are described in Section 3.1 of the main text. Section S3 presents the derivation of the inferential methods for two-stage, partially synthetic data, which are described in Section 3.2 of the main text. Section S4 presents the derivations of the degrees of freedom for the t-distributions used in inferences for two-stage synthetic data.

S1. Notation

To begin, we repeat some notation used in both the full and partial synthesis cases. For a finite population of size N, let $I_l = 1$ if unit l is included in the survey, and $I_l = 0$ otherwise, where $l = 1, \ldots, N$. Let $I = (I_1, \ldots, I_N)$, and let the sample size $s = \sum I_l$. Let X be the $N \times d$ matrix of sampling design variables, e.g. stratum or cluster indicators or size measures. We assume that X is known approximately for the entire population, for example from census records or the sampling frame(s). Let Y be the $N \times p$ matrix of survey data for the population. Let $Y_{inc} = (Y_{obs}, Y_{mis})$ be the $s \times p$ sub-matrix of Y for all units with $I_l = 1$, where Y_{obs} is the portion of Y_{inc} that is observed and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. Let R be an $N \times p$ matrix of indicators such that $R_{lk} = 1$ if the response for unit l to item k is recorded, and $R_{lk} = 0$ otherwise. The observed data is thus $D_{obs} = (X, Y_{obs}, I, R)$. Let Y_a be the values simulated in stage 1, and let Y_b be the values simulated in stage 2.

S2. Inferences for two stage full synthesis

We suppose that the agency has generated m partially completed populations as described in Section 3.1 of the main text. Let $P^{(i,j)} = (D_{obs}, Y_a^{(i)}, Y_b^{(i,j)})$ be a completed population, where $i = 1, \ldots, m$ and $j = 1, \ldots, r$. For each (i, j), let $D^{(i,j)}$ be a simple random sample from $P^{(i,j)}$. These M samples, $D_{syn} = \{D^{(i,j)} : i = 1, \ldots, m; j = 1, \ldots, r\}$, are released to the public. Each released $D^{(i,j)}$ includes a label indicating its value of i, i.e. an indicator for its nest.

Let Q be the estimand of interest, such as a population mean or a regression coefficient. The analyst of synthetic data seeks $f(Q|D_{syn})$. The three-step process for creating D_{syn} described in Section 3.1 of the

main text suggests that

$$f(Q|D_{syn}) = \int f(Q|D_{obs}, P_{syn}, D_{syn}) f(D_{obs}|P_{syn}, D_{syn}) f(P_{syn}|D_{syn}) dD_{obs} dP_{syn}, \quad \text{(S2.1)}$$

where $P_{syn} = \{P^{(i,j)} : i = 1, \ldots, m; j = 1, \ldots, r\}$. For all derivations in this section, we assume that the analyst's distributions are identical to those used by the agency for creating D_{syn}. We also assume that the sample sizes are large enough to permit normal approximations for these distributions. Thus, we require only the first two moments for each distribution, which we derive using standard large sample Bayesian arguments. Diffuse priors are assumed for all parameters.

To begin, the synthetic data are irrelevant for inference about Q given the observed data, so that $f(Q|D_{obs}, P_{syn}, D_{syn}) = f(Q|D_{obs})$. We assume that

$$(Q|D_{obs}) \sim N(Q_{obs}, U_{obs}), \quad \text{(S2.2)}$$

where Q_{obs} and U_{obs} are the estimates of the mean and variance computed from D_{obs} if it were released. We note that only Q needs to have a normal distribution, not the data Y itself.

The D_{syn} is irrelevant given P_{syn}, so that $f(D_{obs}|P_{syn}, D_{syn}) = f(D_{obs}|P_{syn})$. Because inferences for Q depend only on Q_{obs} and U_{obs}, it is sufficient to determine $f(Q_{obs}, U_{obs}|P_{syn})$. Let $Q^{(i,j)}$ be the estimate of Q in population $P^{(i,j)}$. Let $\bar{Q}_r^{(i)} = \sum_j Q^{(i,j)}/r$, and $\bar{Q}_M = \sum_i \bar{Q}_r^{(i)}/m$. Let $B_M = \sum_i (\bar{Q}_r^{(i)} - \bar{Q}_M)^2/(m-1)$, and $W_r^{(i)} = \sum_j (Q^{(i,j)} - \bar{Q}_r^{(i)})^2/(r-1)$. We assume the following sampling distributions:

$$\left(\bar{Q}_\infty^{(i)}|D_{obs}, B_\infty\right) \sim N(Q_{obs}, B_\infty) \quad \text{(S2.3)}$$

$$\left(Q^{(i,j)}|\bar{Q}_\infty^{(i)}, W_\infty^{(i)}\right) \sim N(\bar{Q}_\infty^{(i)}, W_\infty^{(i)}) \quad \text{(S2.4)}$$

where $\bar{Q}_\infty^{(i)}$, $W_\infty^{(i)}$, and B_∞ are the limits of the corresponding finite-sum quantities as $m \to \infty$ and $r \to \infty$. The process of repeatedly completing populations and estimating Q in this nested manner is equivalent to simulating the posterior distribution of Q. Hence, $U_{obs} = B_\infty + \bar{W}_\infty$, where $\bar{W}_\infty = \lim \sum_i W_\infty^{(i)}/m$ as $m \to \infty$. From (S2.2), (S2.3), and (S2.4), for finite m and r we have

$$\left(Q|P_{syn}, B_\infty, W_\infty^{(1)}, \ldots, W_\infty^{(m)}\right) \sim N(\bar{Q}_M, (1 + 1/m)B_\infty + (1 + 1/(mr))\bar{W}_\infty). \quad \text{(S2.5)}$$

We also have

$$\left((m-1)B_M/(B_\infty + \bar{W}_\infty/r)|P_{syn}, \bar{W}_\infty\right) \sim \chi^2_{m-1} \quad \text{(S2.6)}$$

$$\left((r-1)W_r^{(i)}/W_\infty^{(i)}|P_{syn}\right) \sim \chi^2_{r-1}. \quad \text{(S2.7)}$$

The posterior distribution of Q conditioning on P_{syn} alone is found by integrating (S2.5) over the distributions in (S2.6) and (S2.7).

In general, releasing P_{syn} is impractical for agencies, as it could require releasing M data files of very large size N. We therefore take random samples of size n_{syn} from each population, i.e. $D^{(i,j)}$. We require the distributions of $\bar{Q}_M, B_\infty$, and each $W_\infty^{(i)}$ conditional on D_{syn}. For all (i,j), let $q^{(i,j)}$ be the estimate of $Q^{(i,j)}$, and let $u^{(i,j)}$ be the estimate of the variance associated with $q^{(i,j)}$. The $q^{(i,j)}$ and $u^{(i,j)}$

are computed based on the design used to sample from $P^{(i,j)}$. Note that when $n_{syn} = N$, $u^{(i,j)} = 0$. Let $\bar{q}_r^{(i)} = \sum_j q^{(i,j)}/r$, and $\bar{q}_M = \sum_i \bar{q}_r^{(i)}/m$. Let $b_M = \sum_i (\bar{q}_r^{(i)} - \bar{q}_M)^2/(m-1)$, and $w_r^{(i)} = \sum_j (q^{(i,j)} - \bar{q}_r^{(i)})^2/(r-1)$. Finally, let $\bar{u}_M = \sum_{i,j} u^{(i,j)}/(mr)$.

For large n_{syn}, we assume the sampling distribution of each $(q^{(i,j)}|P_{syn}^{(i)})$ is $N(Q^{(i,j)}, U^{(i)})$, where $U^{(i)}$ is an implied sampling variance. We further assume that the sampling variability in the $u^{(i,j)}$ is negligible, so that $u^{(i,j)} \approx U^{(i)}$. We also make the simplifying assumption that the variability in $U^{(i)}$ across nests is small, so that $U^{(i)} \approx \sum U^{(i)}/m$. Thus, we have

$$(q^{(i,j)}|P_{syn}^{(i)}) \sim N(Q^{(i,j)}, \bar{u}_M). \tag{S2.8}$$

Using the standard Bayesian arguments based on these sampling distributions, we have

$$\left(\bar{Q}^{(i)}|\bar{q}_r^{(i)}, \bar{u}_M\right) \sim N(\bar{q}_r^{(i)}, \bar{u}_M/r) \tag{S2.9}$$

and

$$\left(\bar{Q}_M|D_{syn}\right) \sim N(\bar{q}_M, \bar{u}_M/(mr)). \tag{S2.10}$$

To obtain the conditional distributions of B_∞ and each $W_\infty^{(i)}$, we use an analysis of variance setup. From (S2.4) and (S2.8), we have

$$\left(\frac{(r-1)w_r^{(i)}}{W_\infty^{(i)} + \bar{u}_M}|D_{syn}\right) \sim \chi^2_{r-1}. \tag{S2.11}$$

From (S2.3), (S2.4), and (S2.8), and making the simplifying assumption that $W_\infty^{(i)} = \bar{W}_\infty$ for all i, we have

$$\left(\frac{(m-1)b_M}{B_\infty + \bar{W}_\infty/r + \bar{u}_M/r}|D_{syn}, \bar{W}_\infty\right) \sim \chi^2_{m-1} \tag{S2.12}$$

$$\left(\frac{m(r-1)\bar{w}_M}{\bar{W}_\infty + \bar{u}_M}|D_{syn}\right) \sim \chi^2_{m(r-1)} \tag{S2.13}$$

where $\bar{w}_M = \sum_i w_r^{(i)}/m$.

To obtain the conditional distribution of Q given D_{syn}, we integrate the distribution in (S2.5) with respect to the distributions of $\bar{Q}_M$, B_∞, and the $W_\infty^{(i)}$ in (S2.10), (S2.12), and (S2.13). This can be done via numerical integration. For example, analysts can take the following steps after computing $\bar{q}_M$, b_M, $\bar{w}_M$, and $\bar{u}_M$ from the released datasets.

1. Draw a value of $\bar{W}_\infty$, say $\bar{W}_\infty^*$, from the chi-squared distribution in (S2.13). That is, draw a value c from a chi-squared distribution with $m(r-1)$ degrees of freedom, and take $\bar{W}_\infty^* = (m(r-1)\bar{w}_M - c\bar{u}_M)/c$.

2. Given $\bar{W}_\infty^*$, draw a value of B_∞, say B_∞^*, from the chi-squared distribution in (S2.12). That is, draw a value d from a chi-squared distribution with $(m-1)$ degrees of freedom, and take $\bar{B}_\infty^* = ((m-1)b_M - d(\bar{W}_\infty^*/r + \bar{u}_M/r))/d$.

3. Draw a value of $\bar{Q}_M$, say $\bar{Q}_M^*$, from the normal distribution in (S2.10).

4. Given $\bar{W}_\infty^*$, B_∞^* and $\bar{Q}_M^*$, draw a value of Q, say Q^*, from the normal distribution in (S2.5).

5. Store the value of Q^*, and repeat steps 1 - 4 independently a large number of times, say 10000 times.

The resulting draws of Q are samples from the posterior distribution in (S2.1), subject to the conditions needed for (S2.2) – (S2.13). We note that, for scalar Q, this integral can be computed even if m or r is small. The normality assumptions underpinning (S2.2) – (S2.13) come from large sample Bayesian arguments, where the relevant sample sizes are n and n_{syn}, not m or M.

Instead of direct simulation, some analysts may desire a straightforward approximation using D_{syn}. For large m and r, we can approximate $f(Q|D_{syn})$ by a normal distribution with mean $E(Q|D_{syn})$ and variance $Var(Q|D_{syn})$. Using (S2.5) and (S2.10), we have

$$E(Q|D_{syn}) = E[E(Q|\bar{Q}_M)|D_{syn}] = E(\bar{Q}_M|D_{syn}) = \bar{q}_M. \tag{S2.14}$$

Similarly,

$$\begin{aligned} Var(Q|D_{syn}) &= E[Var(Q|P_{syn}, B_\infty, \bar{W}_\infty)|D_{syn}] + Var[E(Q|P_{syn}, B_\infty, \bar{W}_\infty)|D_{syn}] \\ &= (1 + m^{-1})E(B_\infty|D_{syn}) + (1 + 1/(mr))E(\bar{W}_\infty|D_{syn}) + \bar{u}_M/(mr). \end{aligned} \tag{S2.15}$$

Based on (S2.12) and (S2.13), we approximate the expectations in (S2.15) as $E(\bar{W}_\infty|D_{syn}) \approx \bar{w}_M - \bar{u}_M$ and $E(B_\infty|D_{syn}) \approx b_M - \bar{w}_M/r$. Substituting these approximate expectations in (S2.15), we have

$$\begin{aligned} Var(Q|D_{syn}) &\approx (1 + m^{-1})(b_M - \bar{w}_M/r) + (1 + 1/(mr))(\bar{w}_M - \bar{u}_M) + \bar{u}_M/(mr) \\ &= (1 + m^{-1})b_M + (1 - 1/r)\bar{w}_M - \bar{u}_M = T_f. \end{aligned} \tag{S2.16}$$

For modest m and r, we obtain inferences by using a t-distribution, $(\bar{q}_M - Q) \sim t_{\nu_f}(0, T_f)$. The degrees of freedom, ν_f, equal

$$\nu_f = \left(\frac{((1 + 1/m)b_M)^2}{(m - 1)T_f^2} + \frac{((1 - 1/r)\bar{w}_M)^2}{(m(r - 1))T_f^2} \right)^{-1}.$$

The degrees of freedom is derived by matching the first two moments of $(\nu_f T_f)/(\bar{u}_M/(mr) + (1 + 1/m)B_\infty + (1 + 1/(mr))\bar{W}_\infty)$ to an inverse chi-squared distribution with ν_f degrees of freedom. The derivation is presented in Section S4.1 of this supplement.

When the normality assumptions underpinning (S2.2) – (S2.13), specifically those in (S2.2) – (S2.4) and (S2.8), do not hold, neither the posterior simulation approach nor the t-approximation result in valid inferences. In general, multiple imputation approaches with modest M (or m in one stage multiple imputation) are not adequate for inference about quantities with asymmetric posterior distributions. Fortunately, with large sample sizes n and n_{syn}, as is expected to be the case for multiple imputation for disclosure limitation, these normality assumptions are usually reasonable.

When (S2.2) – (S2.4) and (S2.8) do not hold, the analyst need either: (i) apply for special access to the genuine data, or (ii) simulate more synthetic datasets if the agency provides sufficient meta-data to do so, e.g., the agency provides a program that simulates datasets in the same way used to generate D_{syn}. This type of detailed meta-data may not be available in practice. Agencies that provide a data simulator essentially release the exact values of the parameters of the imputation model, which could represent a disclosure risk.

For example, if categorical data are simulated from a log-linear model, the exact values of the parameters could be used to determine the counts in the cells of the observed data contingency table. A more likely scenario is that the agency releases descriptions of the synthesis model, or the synthesis code itself, without the distributions of the parameters of the synthesis models (which are estimated with D_{obs}). Given this meta-data, it may be possible for the analyst to estimate the parameters of the synthesis model from D_{syn} and then simulate additional datasets. The properties of this approach have not been investigated. However, these inferences clearly would have larger uncertainty than those based on a data simulator, because of the additional variance due to estimating the imputation model parameters with modest m.

Even when (S2.2) – (S2.13) are reasonable, it is possible that $T_f < 0$, which causes problems for inferences. This occurs because of high variability in the estimates of B_∞ and $\bar{W}_\infty$. Problems tend to arise more with estimation of B_∞ than with $\bar{W}_\infty$. We have $m(r-1)$ degrees of freedom to estimate $\bar{W}_\infty$, and only $(m-1)$ degrees of freedom to estimate B_∞. We note that negative variance estimates are not a problem with direct posterior simulation.

S3. Inference for two stage partial synthesis

To obtain inferences from nested partially synthetic data, we assume the analyst acts as if each $D^{(i,j)}$ is a sample according to the original design. We require the integral,

$$f(Q|D_{syn}) = \int f(Q|D_{obs}, D_{syn}) f(D_{obs}|D_{syn}) dD_{obs}. \tag{S3.1}$$

Unlike in fully synthetic data, there is no intermediate step of completing populations. Let $q^{(i,j)}$, $\bar{q}_r^{(i)}$, $\bar{q}_M$, b_M, and the $w_r^{(i)}$ be defined as in the previous section. Define $\bar{q}_\infty^{(i)} = \lim \bar{q}_r^{(i)}$, $b_\infty = \lim b_M$, and $w_\infty^{(i)} = \lim w_r^{(i)}$ as $m \to \infty$ and $r \to \infty$.

With large samples, we assume again that $f(Q|D_{obs}) = N(Q_{obs}, U_{obs})$. We assume that the sampling distributions of the synthetic data point estimators are

$$\left(\bar{q}_\infty^{(i)}|D_{obs}, b_\infty\right) \sim N(Q_{obs}, b_\infty) \tag{S3.2}$$

$$\left(q^{(i,j)}|D_{obs}, \bar{q}_\infty^{(i)}, w_\infty^{(i)}\right) \sim N(\bar{q}_\infty^{(i)}, w_\infty^{(i)}). \tag{S3.3}$$

When coupled with (S2.2) and diffuse priors on all parameters, (S3.2) and (S3.3) imply that

$$\left(Q|D_{syn}, b_\infty, w_\infty^{(1)}, \ldots, w_\infty^{(m)}\right) \sim N(\bar{q}_M, U_{obs} + b_\infty/m + \bar{w}_\infty/(mr)). \tag{S3.4}$$

Since the Y_a and Y_b are simulated from their conditional distributions, each $u^{(i,j)}$ approximates U_{obs}. We assume that each $u^{(i,j)}$ has low variability, so that $u^{(i,j)} \approx \bar{u}_M \approx U_{obs}$.

The posterior distributions of b_∞ and each $w_\infty^{(i)}$ are obtained from an analysis of variance setup. From (S3.3), we have

$$\left(\frac{(r-1)w_r^{(i)}}{w_\infty^{(i)}}|D_{syn}\right) \sim \chi^2_{r-1}. \tag{S3.5}$$

From (S3.2), (S3.3), and (S3.5), and making the simplifying assumption that $w_\infty^{(i)} = \bar{w}_\infty$ for all i, we have

$$\left(\frac{(m-1)b_M}{b_\infty + \bar{w}_\infty/r}|D_{syn}, \bar{w}_\infty\right) \sim \chi^2_{m-1} \tag{S3.6}$$

$$\left(\frac{m(r-1)\bar{w}_M}{\bar{w}_\infty}|D_{syn}\right) \sim \chi^2_{m(r-1)}. \tag{S3.7}$$

To obtain the conditional distribution of Q, we integrate (S3.4) over the distributions in (S3.6) and (S3.7).

For large m and r, we can approximate this integral with a normal distribution, substituting the approximate expected values of b_∞ and $\bar{w}_\infty$ into the variance in (S3.4). For large m and r, this variance simplifies to $T_p = \bar{u}_M + b_M/m$, so that the approximate normal distribution is $(\bar{q}_M - Q) \sim N(0, T_p)$.

For small m and r, we can use a t-distribution for inferences, $(\bar{q}_M - Q) \sim t_{\nu_p}(0, T_p)$. The degrees of freedom $\nu_p = (m-1)(1 + m\bar{u}_M/b_M)^2$. The degrees of freedom is derived by matching the first two moments of $(\nu_p(\bar{u}_M + b_M/m))/(\bar{u}_M + b_\infty/m + \bar{w}_\infty/(mr))$ to an inverse chi-squared distribution with ν_p degrees of freedom. The derivation is presented in Section S4.2 of this supplement.

S4. Derivation of approximate degrees of freedom

Here we derive the degrees of freedom for the approximate t-distributions for two stage fully and partially synthetic data.

S4.1. Fully synthetic data

The key step is to approximate the distribution of

$$\left(\frac{\nu_f T_f}{\bar{u}_M/(mr) + (1+1/m)B_\infty + (1+1/(mr))\bar{W}_\infty} \mid D_{syn}\right) \tag{S4.1}$$

as a chi-squared distribution with ν_f degrees of freedom. The ν_f is determined by matching the mean and variance of the inverted χ^2 distribution to the mean and variance of (S4.1).

Let $\gamma = (B_\infty + \bar{W}_\infty/r + \bar{u}_M/r)/b_M$, and let $\delta = (\bar{W}_\infty + \bar{u}_M)/\bar{w}_M$. Making the approximation that $W_\infty^{(i)} = \bar{W}_\infty$ for all i, $(\gamma^{-1} \mid b_M)$ and $(\delta^{-1} \mid \bar{w}_M)$ have mean square distributions with degrees of freedom $m-1$ and $m(r-1)$, respectively. Substituting γ and δ into (S4.1), the random variable is

$$\frac{T_f}{\bar{u}_M/(mr) + (1+1/m)(\gamma b_M - \delta\bar{w}_M/r) + (1+1/(mr))(\delta\bar{w}_M - \bar{u}_M)}. \tag{S4.2}$$

We need to approximate the expectation and variance of (S4.2) and match them to a mean square random variable with ν_f degrees of freedom. We write the expectation as

$$E\left(E\left(\frac{T_f}{\bar{u}_M/(mr) + (1+1/m)(\gamma b_M - \delta\bar{w}_M/r) + (1+1/(mr))(\delta\bar{w}_M - \bar{u}_M)} \mid \delta\right)\right), \tag{S4.3}$$

where D_{syn} is suppressed from both expectations for brevity. We approximate the expectations using first order Taylor series expansions in γ^{-1} and δ^{-1} around their expectations, which equal one. The approximation boils down to substituting ones for γ and δ. After substitution, the denominator in (S4.2) approximately equals T_f, and the expectation approximately equals one.

For the variance, we use the conditional variance representation

$$Var\left(E\left(\frac{T_f}{\bar{u}_M/(mr)+(1+1/m)(\gamma b_M-\delta\bar{w}_M/r)+(1+1/(mr))(\delta\bar{w}_M-\bar{u}_M)} \mid \delta\right)\right)$$
$$+E\left(Var\left(\frac{T_f}{\bar{u}_M/(mr)+(1+1/m)(\gamma b_M-\delta\bar{w}_M/r)+(1+1/(mr))(\delta\bar{w}_M-\bar{u}_M)} \mid \delta\right)\right). \quad \text{(S4.4)}$$

For the interior expectation and variance, we use first order Taylor series expansions in γ^{-1} around its expectation. The first term in (S4.4) approximately equals

$$Var\left(\frac{T_f}{\bar{u}_M/(mr)+(1+1/m)(b_M-\delta\bar{w}_M/r)+(1+1/(mr))(\delta\bar{w}_M-\bar{u}_M)}\right). \quad \text{(S4.5)}$$

Since $Var(\gamma^{-1} \mid D_{syn}, \delta) = 2/(m-1)$, the second term in (S4.4) approximately equals

$$E\left(\frac{(2/(m-1))T_f^2((1+1/m)b_M)^2}{(\bar{u}_M/(mr)+(1+1/m)(b_M-\delta\bar{w}_M/r)+(1+1/(mr))(\delta\bar{w}_M-\bar{u}_M))^4}\right). \quad \text{(S4.6)}$$

We next approximate the variance in (S4.5) and the expectation in (S4.6) using first order Taylor series expansions in δ^{-1} around its expectation. Since $Var(\delta^{-1} \mid D_{syn}) = 2/(m(r-1))$, the variance in (S4.5) approximately equals

$$\frac{2/(m(r-1))T_f^2((1-1/r)\bar{w}_M)^2}{T_f^4}. \quad \text{(S4.7)}$$

The expectation in (S4.6) approximately equals

$$\frac{(2/(m-1))T_f^2((1+1/m)b_M)^2}{T_f^4}. \quad \text{(S4.8)}$$

The variance in (S4.4) is approximately the sum of (S4.7) and (S4.8). Since a mean square random variable has variance equal to 2 divided by its degrees of freedom, we conclude that the

$$\nu_f = \left(\frac{((1+1/m)b_M)^2}{(m-1)T_f^2}+\frac{((1-1/r)\bar{w}_M)^2}{(m(r-1))T_f^2}\right)^{-1}. \quad \text{(S4.9)}$$

S4.2. Partially synthetic data

We approximate the distribution of

$$\left(\frac{\nu_p T_p}{\bar{u}_M+b_\infty/m+\bar{w}_\infty/(mr)} \mid D_{syn}\right) \quad \text{(S4.10)}$$

as a chi-squared distribution with ν_p degrees of freedom. The ν_p is determined by matching the mean and variance of the inverted χ^2 distribution to the mean and variance of (S4.10).

Let $\phi = (b_\infty + \bar{w}_\infty/r)/b_M$, and let $\psi = \bar{w}_\infty/\bar{w}_M$. Making the approximation that $w_\infty^{(i)} = \bar{w}_\infty$ for all i, $(\phi^{-1} \mid D_{syn}, \bar{w}_\infty)$ and $(\psi^{-1} \mid D_{syn})$ have mean square distributions with degrees of freedom $m-1$ and $m(r-1)$, respectively. We write the random variable in (S4.10) as

$$\frac{T_p}{\bar{u}_M+\phi b_M/m}. \quad \text{(S4.11)}$$

To match moments, we need to approximate the expectation and variance of (S4.11) and match them to a mean square random variable with ν_p degrees of freedom.

We write the expectation of (S4.11) as

$$E\left(E\left(\frac{T_p}{\bar{u}_M + \phi b_M/m} \mid D_{syn}, \bar{w}_\infty\right) \mid D_{syn}\right). \tag{S4.12}$$

We approximate these expectations using first order Taylor series expansions in ψ^{-1} and ϕ^{-1} around their expectations, which equal one. The approximation boils down to substituting one for ϕ, as the ψ never enters the computations except in the conditioning arguments for ϕ. After substitution, the denominator in (S4.11) approximately equals T_p, and the expectation approximately equals one.

For the variance, we use the conditional variance representation

$$E\left(Var\left(\frac{T_p}{\bar{u}_M + \phi b_M/m} \mid d^M, \bar{w}_\infty\right) \mid D_{syn}\right) + Var\left(E\left(\frac{T_p}{\bar{u}_M + \phi b_M/m} \mid d^M, \bar{w}_\infty\right) \mid D_{syn}\right). \tag{S4.13}$$

For the interior expectation and variance, we use first order Taylor series expansions in ϕ^{-1} and ψ^{-1} around their expectations. The interior expectation equals approximately one, so that the variance in the second term equals zero. Since $Var(\phi^{-1} \mid D_{syn}, \bar{w}_\infty) = 2/(m-1)$, the interior variance in (S4.13) approximately equals

$$E\left(\frac{2T_p^2(b_M/m)^2}{(m-1)(\bar{u}_M + b_M/m)^4} \mid D_{syn}\right) = \frac{2(b_M/m)^2}{(m-1)T_p^2}. \tag{S4.14}$$

Since a mean square random variable has variance equal to 2 divided by its degrees of freedom, we conclude that

$$\nu_p = (m-1)(T_p/(b_M/m))^2 = (m-1)(1 + m\bar{u}_M/b_M)^2. \tag{S4.15}$$