Erratum

Authors' corrigenda/corrections des auteurs on synthetic data method to incorporate external information into a current study

Tian GU[1*], Jeremy M. G. TAYLOR[1], Wenting CHENG[1], and Bhramar MUKHERJEE[1]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48105, U.S.A.*

Key words and phrases: Constrained maximum likelihood; data integration; prediction models; synthetic data.

In Gu, Taylor, Cheng & Mukherjee (2019), four notations used in Appendices A1, A2 and A3, namely $\Omega, \omega^2, \sigma^2$ and η^2, should have been replaced by some of the existing notations. The correct notations that they correspond to are listed as follows:

- $\Omega = \sigma^2_{\mathrm{X}}$, the variance of X;
- $\omega^2 = \sigma^2_{\beta}$, as introduced in equation (5) in the main paper;
- $\sigma^2 = \sigma^2_{\gamma}$, as introduced in equation (6) in the main paper;
- $\eta^2 = \sigma^2_{\theta}$, as introduced in equation (7) in the main paper.

Note that all the results in Appendices A1, A2 and A3 were correct and were not affected by the notation correction. The corrected Appendices A1, A2 and A3 are attached below.

REFERENCE

Gu, T., Taylor, J. M. G., Cheng, W., & Mukherjee, B. (2019). Synthetic data method to incorporate external information into a current study. *The Canadian Journal of Statistics*, 47, 580–603.

APPENDIX

Derivation of asymptotic variances for the special case 1

Appendix A1. Approach 1: Synthetic Data Method

If the synthetic data approach is applied, and under the assumption that the true value of β and σ_β are used to generate the synthetic data, then the combined data will have the same distribution as a dataset of size n+m in which m values of B have been removed. For this particular data structure, it is possible to obtain formulas for the asymptotic variance of the maximum likelihood estimate (MLE) of $\boldsymbol{\gamma}$. In particular, Gourieroux & Monfort (1981) gave the exact expression of the MLE and the corresponding asymptotic covariance in such case. The likelihood for the combined data is $\prod_{\mathrm{i}=1}^{\mathrm{n}} f(\mathrm{Y_i}, \mathrm{B_i}|\mathrm{X_i}) \times \prod_{\mathrm{i}=\mathrm{n}+1}^{\mathrm{n}+\mathrm{m}} f(\mathrm{Y_i}|\mathrm{X_i})$, which can be

[*] *Corresponding author: gutian@umich.edu*

rewritten as $\prod_{\mathrm{i=1}}^{\mathrm{n+m}} f(\mathrm{Y_i}|\mathrm{X_i}) \times \prod_{\mathrm{i=1}}^{\mathrm{n}} f(\mathrm{B_i}|\mathrm{X_i}, \mathrm{Y_i})$. Based on this likelihood, they introduced a set of transformed parameters, and re-parameterized the distributions (5)–(7). They then identified the 1-to-1 relationship among the original parameters and the new set of parameters, which we will explain in the subsequent paragraph.

We obtain the estimators of the original parameters by the re-parameterization method, and then apply the delta method to get the asymptotic variance of $\hat{\gamma}_{\mathrm{B}}$ and $\hat{\gamma}_{\mathrm{X}}$. According to Gourieroux & Monfort (1981), we introduce a set of transformed parameters a, b, c, d and e, and re-parameterize the distributions (5)–(7) as $\mathrm{Y|X} \sim \mathrm{N}(b\mathrm{X}, a^2)$, and $\mathrm{B|Y, X} \sim \mathrm{N}(d\mathrm{Y} + e\mathrm{X}, c^2)$. Then we identify the 1-to-1 relationship among the original parameters and the new set of parameters:

$$
\begin{aligned}
a^2 &= \sigma_\gamma^2 + \gamma_{\mathrm{B}}^2 \sigma_\theta^2 \\
b &= \gamma_{\mathrm{X}} + \theta\gamma_{\mathrm{B}} \\
c^2 &= \frac{\sigma_\gamma^2 \sigma_\theta^2}{a^2} \\
d &= \frac{\gamma_{\mathrm{B}} \sigma_\theta^2}{\mathrm{a}^2} \\
e &= \theta - db.
\end{aligned} \tag{1}
$$

The MLE of a, b and their asymptotic variances are easy to obtain from the linear model $\mathrm{Y_i} = b\mathrm{X_i} + u_i, Var(u_i) = a^2$, where i = 1,…, n+m. Similarly, the MLE of c, d and e and their asymptotic variances are easy to obtain from the linear model $\mathrm{B_i} = d\mathrm{Y_i} + e\mathrm{X_i} + v_i, \mathrm{Var}(v_i) = c^2$ where i = 1,…,n. The estimators of the original parameters are obtained through the relationship derived from Equation (1), where

$$
\begin{aligned}
\theta &= bd + e \\
\sigma_\theta^2 &= a^2 d^2 + c^2 \\
\gamma_{\mathrm{B}} &= \frac{a^2 d}{\sigma_\theta^2} \\
\gamma_{\mathrm{X}} &= \mathrm{b} - \gamma_{\mathrm{B}}\theta \\
\sigma_\gamma^2 &= \frac{a^2 c^2}{\sigma_\theta^2}
\end{aligned}
$$

and the asymptotic variance of $\hat{\gamma}_{\mathrm{X}}$ and $\hat{\gamma}_{\mathrm{B}}$ can be derived using the delta method:

$$
\begin{cases}
\mathrm{Var}(\hat{\gamma}_{\mathrm{B}}) = \frac{1}{\mathrm{n}}\left[\frac{\sigma_\gamma^2}{\sigma_\theta^2} + 2(\lambda - 1)\frac{\gamma_{\mathrm{B}}^2 \sigma_\gamma^4}{\sigma_\beta^4}\right] \\
\mathrm{Var}(\hat{\gamma}_{\mathrm{X}}) = \theta^2 \mathrm{Var}(\hat{\gamma}_{\mathrm{B}}) + \frac{1}{\mathrm{n}} \frac{\sigma_\gamma^2}{\sigma_X^2} \frac{\lambda\sigma_\gamma^2 + \gamma_{\mathrm{B}}^2 \sigma_\theta^2}{\sigma_\beta^2}.
\end{cases}
$$

Therefore, we find the relative efficiency gain of $\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \mathrm{Var}(\hat{\gamma}_{\mathrm{X}}, \hat{\gamma}_{\mathrm{B}})^{\mathrm{T}}$ by adding m synthetic data observations compared to the original dataset of size n is

$$
\mathrm{ARE}[\mathrm{Var}(\hat{\boldsymbol{\gamma}})] = \mathbf{1} - (1 - \lambda)\begin{pmatrix} \frac{2\sigma_\gamma^2 \gamma_{\mathrm{B}}^2 \sigma_\theta^2}{\sigma_\beta^4} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_{\mathrm{X}}^2 \theta^2} \frac{\sigma_\gamma^2 (2\sigma_\gamma^2 - \sigma_\beta^2)}{\sigma_\beta^4} \\ \frac{2\gamma_{\mathrm{B}}^2 \sigma_\theta^2 \sigma_\gamma^2}{\sigma_\beta^4} \end{pmatrix},
$$

where $\theta = \frac{\beta-\gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}}$ and $\sigma_\theta^2 = \frac{\sigma_\beta^2-\sigma_\gamma^2}{\gamma_{\mathrm{B}}^2}$. When m gets very large such that $\lambda \approx 0$, ARE[Var($\hat{\gamma_{\mathrm{X}}}$)] = $1 - \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_{\mathrm{X}}^2\theta^2}\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\gamma_{\mathrm{B}}^2\sigma_\theta^2}\frac{2\sigma_\gamma^2-\sigma_\beta^2}{\sigma_\beta^2} - \frac{2\sigma_\gamma^2\gamma_{\mathrm{B}}^2\sigma_\theta^2}{\sigma_\beta^4}$, and ARE[Var($\hat{\gamma_{\mathrm{B}}}$)] = $1 - \frac{2\sigma_\gamma^2\gamma_{\mathrm{B}}^2\sigma_\theta^2}{\sigma_\beta^4}$. This demonstrates some gain in efficiency for both γ_{X} and γ_{B}.

Appendix A2. Approach 2: Constrained MLE

Depending on the information available from the external model Y|X, two possible situations correspond to two different constraints:

- *Approach 2.1: Only the estimated coefficient β is known from model (5)*
  From model (5)–(7), it is easy to see that the constraint is $\theta = \frac{\beta-\gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}}$, describing the relationship between the unknown variable θ, the known variable β and the target variable $\boldsymbol{\gamma}$. The log-likelihood is given by

$$\begin{aligned} l &= l(\boldsymbol{\gamma}, \theta, \sigma_\gamma^2, \sigma_\theta^2) \\ &= \sum_{\mathrm{i}=1}^{\mathrm{n}} logf(\mathrm{Y_i}, \mathrm{B_i}|\mathrm{X_i}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} logf(\mathrm{Y_i}|\mathrm{X_i}, \mathrm{B_i}; \gamma, \sigma_\gamma^2) + \sum_{\mathrm{i}=1}^{\mathrm{n}} logf(\mathrm{B_i}|\mathrm{X_i}; \theta, \sigma_\theta^2) \\ &= -\frac{\mathrm{n}}{2}log(\sigma_\gamma^2) - \frac{1}{2\sigma_\gamma^2}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{Y_i} - \gamma_{\mathrm{X}}\mathrm{X_i} - \gamma_{\mathrm{B}}\mathrm{B_i}\right)^2 - \frac{\mathrm{n}}{2}log(\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{B_i} - \theta\mathrm{Xi}\right)^2. \qquad (2) \end{aligned}$$

  The goal is to maximize the log-likelihood (2) over $\boldsymbol{\gamma}$, σ_γ and σ_θ subject to the constraint $\theta = \theta^* = \frac{\beta^*-\gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}}$. By replacing θ with θ^*, taking the second derivative over $\boldsymbol{\gamma}$, and taking the inverse of the information matrix, we obtain the asymptotic variance of $\hat{\boldsymbol{\gamma}}$:

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \mathbf{I}^{-1} = \frac{1}{\mathrm{n}}\frac{\sigma_\gamma^2}{\sigma_\theta^2}\begin{pmatrix} \theta^{*2} + \frac{\sigma_\theta^4\gamma_{\mathrm{B}}^2}{\sigma_\gamma^2+\sigma_\theta^2\gamma_{\mathrm{B}}^2}\sigma_{\mathrm{X}}^{-2} & -\theta^* \\ -\theta^* & 1 \end{pmatrix}. \qquad (3)$$

  Thus, the ARE of Var($\hat{\boldsymbol{\gamma}}$) from the constrained MLE compared to the standard MLE is

$$\mathrm{ARE}[\mathrm{Var}(\hat{\boldsymbol{\gamma}})] = \begin{pmatrix} 1 - \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_{\mathrm{X}}^2\theta^{*2}}\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\gamma_{\mathrm{B}}^2\sigma_\theta^2} \\ 1 \end{pmatrix},$$

  where we notice that there is some gain in efficiency for γ_{X} but no gain in efficiency for γ_{B}. We can see that the largest gain in efficiency is when γ_{B}, θ and σ_{X} are small.

- *Approach 2.2: Both of the estimated coefficient β and the standard deviation σ_β are known from model (5)*
  In this situation, knowing the true σ_β gives us more information which is incorporated through an additional constraint. In addition to the constraint $\theta = \theta^* = \frac{\beta-\gamma_{\mathrm{X}}}{\gamma_{\mathrm{B}}^*}$ derived in approach 2.1, we add another constraint Var(Y|X) = $\sigma_\beta^2 = \gamma_{\mathrm{B}}^2\sigma_\theta^2 + \sigma_\gamma^2$, where $\sigma_\theta^2 = \sigma_\theta^{*2} = \frac{\sigma_\beta^{*2}-\sigma_\gamma^2}{\gamma_{\mathrm{B}}^2}$. Then we maximize log-likelihood (2) with respect to $\boldsymbol{\gamma}$ and σ_γ^2 at fixed values $\sigma_\theta^2 = \sigma_\theta^{*2}, \theta = \theta^*$. Note that different from approach 2.1, σ_γ^2 and $\boldsymbol{\gamma}$ are not independent anymore. Thus, we need to

consider σ_γ^2 in the information matrix, and take the inverse of a 3×3 matrix to get the correct asymptotic variance. Let $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \sigma_\gamma^2)^{\mathrm{T}}$,

$$\begin{aligned}\mathbf{I} &= -\mathrm{E}_{\mathrm{XB}}\left(\frac{\partial^2 l}{\partial \boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}}}\right)\\ &= \mathrm{n}\begin{pmatrix} \left(\frac{1}{\sigma_\gamma^2} + \frac{1}{\gamma_{\mathrm{B}}^2\sigma_\theta^{*2}}\right)\sigma_{\mathrm{X}}^2 & \left(\frac{1}{\sigma_\gamma^2} + \frac{1}{\gamma_{\mathrm{B}}^2\sigma_\theta^{*2}}\right)\sigma_{\mathrm{X}}^2\theta^* & 0 \\ \left(\frac{1}{\sigma_\gamma^2} + \frac{1}{\gamma_{\mathrm{B}}^2\sigma_\theta^{*2}}\right)\sigma_{\mathrm{X}}^2\theta^* & \left(\frac{1}{\sigma_\gamma^2} + \frac{1}{\gamma_{\mathrm{B}}^2\sigma_\theta^{*2}}\right)\left(\sigma_\theta^{*2} + \sigma_{\mathrm{X}}^2\theta^{*2}\right) + \frac{1}{\gamma_{\mathrm{B}}^2} & \frac{1}{\sigma_\theta^{*2}\gamma_{\mathrm{B}}^3} \\ 0 & \frac{1}{\sigma_\theta^{*2}\gamma_{\mathrm{B}}^3} & \frac{1}{2}\left(\frac{1}{\sigma_\gamma^4} + \frac{1}{\gamma_{\mathrm{B}}^4\sigma_\theta^{*4}}\right) \end{pmatrix}.\end{aligned}$$

By taking the inverse of $\mathbf{I}$, we can get the asymptotic variance of $\hat{\boldsymbol{\gamma}}$:

$$\begin{cases} \operatorname{Var}(\hat{\gamma}_{\mathrm{B}}) = \frac{1}{\mathrm{n}}\frac{\sigma_\gamma^2}{\sigma_\theta^{*2}}\frac{\sigma_\gamma^4+\gamma_{\mathrm{B}}^4\sigma_\theta^{*4}}{\left(\sigma_\gamma^2+\gamma_{\mathrm{B}}^2\sigma_\theta^{*2}\right)^2} = \frac{1}{\mathrm{n}}\frac{\gamma_{\mathrm{B}}^2\sigma_\gamma^2}{\sigma_\beta^{*2}-\sigma_\gamma^2}\frac{\sigma_\gamma^4+\left(\sigma_\beta^{*2}-\sigma_\gamma^2\right)^2}{\sigma_\beta^{*4}} \\ \operatorname{Var}(\hat{\gamma}_{\mathrm{X}}) = \frac{1}{\mathrm{n}}\frac{\sigma_\gamma^2}{\sigma_\theta^{*2}}\frac{1}{\left(\sigma_\gamma^2+\gamma_{\mathrm{B}}^2\sigma_\theta^{*2}\right)^2}\left[\left(\sigma_\theta^{*2}\sigma_{\mathrm{X}}^{-2} + \theta^{*2}\right)\left(\sigma_\gamma^4 + \gamma_{\mathrm{B}}^4\sigma_\theta^{*4}\right) - \left(\sigma_\gamma^2 - \gamma_{\mathrm{B}}^2\sigma_\theta^{*2}\right)\sigma_\gamma^2\sigma_\theta^{*2}\sigma_{\mathrm{X}}^{-2}\right] \\ \qquad = \left(\sigma_\theta^{*2}\sigma_{\mathrm{X}}^{-2} + \theta^{*2}\right)\operatorname{Var}(\hat{\gamma}_{\mathrm{B}}) - \frac{1}{\mathrm{n}}\sigma_\gamma^4\sigma_{\mathrm{X}}^{-2}\frac{\sigma_\gamma^2-\gamma_{\mathrm{B}}^2\sigma_\theta^{*2}}{\sigma_\beta^{*4}} \end{cases}.$$

Thus, we can obtain the identical ARE to the synthetic data method (approach 1). This demonstrates the asymptotic equivalence of the synthetic data approach with large m compared to the constrained ML approach that uses knowledge of all the parameters in the Y|X distribution.

Appendix A3. Approach 3: Constrained Semiparametric MLE

This approach assumes that β is known, but does not assume that σ_β is known. To calculate the asymptotic variance of $\hat{\boldsymbol{\gamma}}$ in this approach, we need three matrices $\mathbf{I}$, $\mathbf{C}$ and $\mathbf{L}$. After some algebra, it can be shown that $\mathbf{C} = \frac{\sigma_{\mathrm{X}}^2}{\sigma_\beta^2}(1, \theta^*)^{\mathrm{T}}, \mathrm{L} = \frac{\mathrm{n}\gamma_{\mathrm{B}}^2\sigma_\theta^*\sigma_{\mathrm{X}}^2}{\sigma_\beta^{*4}}$. Thus,

$$\operatorname{Cov}(\hat{\boldsymbol{\gamma}}) = (\mathbf{I} + \mathbf{C}\mathbf{L}^{-1}\mathbf{C}^{\mathrm{T}})^{-1} = \frac{1}{\mathrm{n}}\frac{\sigma_\gamma^2}{\sigma_\theta^2}\begin{pmatrix} \theta^{*2} + \frac{\sigma_\theta^4\gamma_{\mathrm{B}}^2}{\sigma_\gamma^2+\sigma_\theta^2\gamma_{\mathrm{B}}^2}\sigma_{\mathrm{X}}^{-2} & -\theta^* \\ -\theta^* & 1 \end{pmatrix},$$

which is identical to approach 2.1.