Supplementary Web Appendix to the article: Pooling Designs for Outcomes Under a Gaussian Random Effects Model

Yaakov Malinovsky, Paul S. Albert, and Enrique F. Schisterman

Section 3: Random Effects Model

We denote $Y = (Y_{11}, \ldots, Y_{1n}, \ldots, Y_{l1}, \ldots, Y_{ln})$. Maximum likelihood estimation of the mean and the variance components of Model (1) depends on distribution assumptions. Using the normality assumptions, we can write

$$Y \sim MVN(\mu 1_N, V)\,,$$

where MVN denotes a multivariate normal distribution, 1_N is a vector of size N with every element equal to unity, $N = ln$, and V is a block-diagonal matrix of size $ln \times ln$ with l identical matrices $V_0 = \sigma^2 I_n + \sigma_A^2 J_n$ comprising the main diagonal, where $J_n = 1_n 1_n^{'}$ and I_n is the identity matrix of size n. In addition, we denote $Y = (Y_{11}, \ldots, Y_{1n}, \ldots, Y_{l1}, \ldots, Y_{ln})$, and the log-likelihood function is then defined as:

$$\begin{aligned}
&logL\left(Y; \mu, \sigma_e^2, \sigma_A^2\right) \\
&= -\frac{N}{2}(2\pi) - \frac{1}{2}log|V| - \frac{1}{2}(Y - \mu 1_N)' V^{-1}(Y - \mu 1_N) \\
&= -\frac{N}{2}(2\pi) - \frac{l(n-1)}{2}log(\sigma_e^2) - \frac{l}{2}log(\sigma_e^2 + n\sigma_A^2) \\
&-\frac{1}{2\sigma_e^2}\left(SSE + \frac{\sigma_e^2}{\sigma_e^2 + n\sigma_A^2}\left\{SSA + ln\,(Y_{\cdot\cdot} - \mu)^2\right\}\right),
\end{aligned}$$

where $SSE = \sum_{i=1}^{l}\sum_{j=1}^{n}(Y_{ij} - Y_{i\cdot})^2$, $SSA = n\sum_{i=1}^{l}(Y_{i\cdot} - Y_{\cdot\cdot})^2$, $Y_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$, and $Y_{\cdot\cdot} = \frac{1}{ln}\sum_{i=1}^{l}\sum_{j=1}^{n} Y_{ij}$.

Mean and variance component estimation as well as random effects prediction for Model (1) can be found in Searle et al. (1992). Using the above likelihood function and its first and second derivatives, we calculate the maximum likelihood estimators (MLE) of parameters. We define $\tau^2 = \sigma_e^2 + n\sigma_A^2$. The maximum likelihood equations are those equations obtained by equating to zero the partial derivatives of log-likelihood with respect to μ, σ_e^2, and τ^2. Let the symbols $\dot{\mu}, \dot{\sigma}_e^2$, and $\dot{\tau}^2$ represent solutions to the maximum likelihood equations. Direct calculations show that

$$\dot{\mu} = Y_{..}, \;\; \dot{\sigma}_e^2 = \frac{SSE}{l(n-1)}, \quad \dot{\tau}^2 = \frac{SSA}{l}, \;\text{ and }\; \dot{\sigma}_A^2 = \frac{\dot{\tau}^2 - \dot{\sigma}_e^2}{n},$$

where $SSE = \sum_{i=1}^{l}\sum_{j=1}^{n}(Y_{ij} - Y_{i\cdot})^2$, $SSA = n\sum_{i=1}^{l}(Y_{i\cdot} - Y_{..})^2$, $Y_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$, and $Y_{..} = \frac{1}{ln}\sum_{i=1}^{l}\sum_{j=1}^{n} Y_{ij}$.

We denote the MLE of $\mu, \sigma_e^2, \sigma_A^2$, and τ^2 by $\widetilde{\mu}$, $\widetilde{\sigma}_e^2$, $\widetilde{\sigma}_A^2$, and $\widetilde{\tau}^2$, respectively (MLE of the variance components defined in the $(0, \infty)$ parameter space). The MLE's of the three parameters are $\widetilde{\mu} = \dot{\mu}$ and

$$\begin{aligned}
&\text{if } \dot{\sigma}_A^2 \geq 0, \;\text{ then } \widetilde{\sigma}_A^2 = \dot{\sigma}_A^2, \;\text{ and } \widetilde{\sigma}_e^2 = \dot{\sigma}_e^2 \\
&\text{if } \dot{\sigma}_A^2 < 0, \;\text{ then } \widetilde{\sigma}_A^2 = 0, \;\text{ and } \widetilde{\sigma}_e^2 = \frac{SST}{ln},
\end{aligned}$$

where $SST = \sum_{i=1}^{l}\sum_{j=1}^{n}(Y_{ij} - Y_{..})^2$. (Searle et al. (1992), Herbach (1959)).

The joint asymptotic distributions (when $l \to \infty$) of $\widetilde{\mu}$, $\widetilde{\sigma}_e^2$ and $\widetilde{\tau}$ can be obtained directly from the distributions of $Y_{..}$, SSE, and SSA by using the multivariate Central Limit Theorem (see, for example, Lehmann and Casella (1998)).

The asymptotic variance matrix is the inverse of the information matrix and can be written as

$$Var\Big[\left(\widetilde{\mu}, \widetilde{\sigma}_e^2, \widetilde{\tau}^2\right)'\Big] \simeq \begin{pmatrix} \frac{\tau^2}{ln} & 0 & 0 \\ 0 & \frac{2\sigma_e^4}{l(n-1)} & 0 \\ 0 & 0 & \frac{2\tau^4}{l} \end{pmatrix}.$$

Section 6: Robustness to the Additive Gaussian Assumption

We consider the following model

$$Y_{ij}\big|\,|A_i| \sim Gamma(\alpha,\beta),\ \ \alpha=\mu+|A_i|,\ \ \beta=1,\ \ A_i\sim N(0,\sigma_A^2),\ \ \mu\geq 0.$$

Here,

$$\begin{aligned}
&E(Y_{ij})=\mu+E\left(|A_i|\right)=\mu+\sigma_A\sqrt{2/\pi},\\
&Var(Y_{ij})=Var\left(E(Y_{ij}\big||A_i|)\right)+E\left(Var(Y_{ij}\big||A_i|)\right)=Var(|A_i|)+\mu+E(|A_i|)\\
&\ =\sigma_A^2\left(1-2/\pi\right)+\mu+\sigma_A\sqrt{2/\pi},\\
&cov(Y_{11},Y_{12})=cov\left(E(Y_{11}|A_1),E(Y_{12}|A_1)\right)+E\left(cov(Y_{11},Y_{12}\big|A_1)\right)\\
&\ =Var(|A_1|)=\sigma_A^2\left(1-2/\pi\right),\ \ \text{and}\\
&\ ICC=\gamma=corr(Y_{11},Y_{12})=\frac{(1-2/\pi)\sigma_A^2}{(1-2/\pi)\sigma_A^2+\mu+\sqrt{2/\pi}\sigma_A}. \qquad (S0)
\end{aligned}$$

Section 8: Discussion

Technical Variation

The designs considered in Section 2 of the paper assume that there is no technical variation. In some studies, technical variation may be sizable (for example, if the laboratory assay process results in substantial measurement error). This error should be distinguished from the biological error of measuring repeated samples across time on the same subject. In this section, we explore the implications of assuming no technical variation when this variation could be sizable. Here, we assume that we observe the vector $Y^{**}=Y^{*}+e^{M}$, where the technical variation e^M is assumed to be normally distributed with zero mean and variance σ_M^2. In this case, the ICC is $\gamma=\dfrac{\sigma_A^2}{\sigma_A^2+\sigma_e^2+\sigma_M^2}$.

In the above model the parameters are identifiable only for a non-symmetric pooling design or for a symmetric design when we assume that σ_M^2 is known or estimated from a small pilot study. In practice, σ_M^2 could be estimated from a small pilot study where an assay is repeated multiple times on the same sample or from the literature. We could then plug in the estimator for σ_M^2 into the appropriate likelihood to estimate the other parameters.

Through simulations we compare Designs I, T, and 2 (Figure 1) for estimating γ in the case where σ_M^2 is assumed known. We fix σ_e^2 to be one. We chose γ to be $0.5, 0.7, 0.9$ and $\sigma_M^2 = 0.01, 0.25, 0.5$. These nine different combinations of the couple γ and σ_M^2 define parameter σ_A^2. We set $l = 1000, n = 4$. For Design T we consider designs where we pool every 2 repeated measurements, and for Design I we consider designs where we pool every 2 individuals (see Figure 1). Further, we consider the non-symmetric design 2 where some pooling is done over individuals, while other pooling is done across time points (Figure 1). Assuming that σ_M^2 is known, we estimate σ_e^2 and σ_A^2 using maximum-likelihood estimation for each Design T and I. We then use a known or plug-in estimator of σ_M^2 to estimate γ. For each γ and σ_M^2 combination we repeat this process 1000 times. Table 1R present these simulation results. We present the MLE of the ICC, relative bias, $\dfrac{E(\widetilde{\gamma}) - \gamma}{\gamma}$ in percentages, and R_{dF} for each Design d. From Table 1R, as expected, we can see that the relative biases of all three designs are close to zero and that Design T is more efficient than Design I. Further, similar to our results where technical variation was assumed negligible, the efficiency of Design 2 is between those of Designs T and I.

Table 1R: **Technical Variation**, $l = 1000, n = 4, \sigma_e^2 = 1$ and 1000 Monte-Carlo simulations. In this case, the ICC is $\gamma = \sigma_A^2/(\sigma_A^2 + \sigma_e^2 + \sigma_M^2)$. It follows from Result 1 and Result 2 that under normal assumptions, $R_{TF} = 1.17, R_{IF} = 2$.

			$\widetilde{\gamma}$			$\frac{E(\widetilde{\gamma}) - \gamma}{\gamma}100\%$					
γ	σ_M^2	σ_A^2	T	I	2	T	I	2	R_{TF}	R_{IF}	R_{2F}
0.5	0.1	1.1	0.500	0.499	0.499	-0.055	-0.261	-0.149	1.69	2.06	2.11
	0.25	1.25	0.501	0.500	0.500	0.268	0.039	-0.028	1.99	2.15	2.03
	0.5	1.50	0.500	0.500	0.500	0.055	-0.057	-0.015	2.35	2.82	2.59
0.7	0.1	2.57	0.701	0.698	0.699	0.117	-0.257	-0.088	1.63	2.18	1.86
	0.25	2.92	0.700	0.699	0.699	-0.040	-0.113	-0.079	1.82	2.44	1.86
	0.5	3.50	0.699	0.699	0.699	-0.179	-0.100	-0.153	2.29	2.58	2.56
0.9	0.1	9.9	0.900	0.900	0.900	-0.046	-0.009	0.018	1.53	1.98	1.54
	0.25	11.25	0.900	0.900	0.900	-0.007	-0.008	0.005	1.81	2.08	1.87
	0.5	13.5	0.900	0.900	0.900	-0.024	-0.003	-0.004	2.08	2.36	2.34

Model with the Center Effect

We extend Model (1) to the following 2-way random nested model:

(i) $Y_{ijk} = \mu + a_i + b_{ij} + e_{ijk}$,

where $i = 1, \ldots, L$ (center), $j = 1, \ldots, l$ (individuals) and $k = 1, \ldots, n$ (repeated measurements).

Model (i) corresponds to the situation where individual specimens come from different centers and interest is on estimating the intraclass correlation coefficient. The ICC for this nested model is $\gamma = corr(Y_{111}, Y_{112}) = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma_e^2}$, where σ_a^2 is the between-centers variation, σ_b^2 is the between-subjects variation, and σ_e^2 is the within-subjects variation. Using the Delta method it is possible to calculate the variance of the MLE of the ICC under the full model and under balanced Designs I and T. However, it is impossible to compare them as in the simple case of Model (1). We therefore performed a simulation study to examine this situation. The results are presented in Table 1R. For this nested model we generated a_i, b_{ij}, e_{ijk} from the a normal distributions with expectation 0. We fix $\sigma_e^2 = 1$, $\gamma = 0.75$ and set $l = 200, n = 8$. For Design T we consider designs where we pool every 4 repeated measurements, and for Design I we consider designs where we pool every 4 individuals. We calculated the relative efficiencies of Designs F, T, and I. For each γ we repeated this process 1000 times. Table 2R presents the simulation results, where we compare the relative efficiency between all pairs of designs (e.g., T vs F, I vs F, and I vs T). Under the normal assumptions without the center effect ($\sigma_a^2 = 0$), it follows from Result 1 and Result 2 that $Var_T(\widetilde{\gamma})/Var_F(\widetilde{\gamma}) = 1.75$, $Var_I(\widetilde{\gamma})/Var_F(\widetilde{\gamma}) = 4$ and $Var_I(\widetilde{\gamma})/Var_T(\widetilde{\gamma}) = 2.286$. The practical implications of these results are as follows.

- The T and I designs are more efficient than the full design under Model (i). These relative efficiencies increase with σ_a^2.

Table 2R:

σ_a^2	lab	$Var_T(\widetilde{\gamma})/Var_F(\widetilde{\gamma})$	$Var_I(\widetilde{\gamma})/Var_F(\widetilde{\gamma})$	$Var_I(\widetilde{\gamma})/Var_T(\widetilde{\gamma})$
0.01	3	1.82	3.45	1.90
	25	1.53	3.24	2.24
	1000	1.51	3.46	2.29
0.3	3	1.24	1.91	1.52
	25	1.26	2.04	1.62
	1000	1.45	1.95	1.35

- Although the T design is always better than the I design, the efficiency of gain T vs I decreases as σ_a^2 increases.

Although it is impossible to fully examine this question for all hierarchical structures, these conclusions appear to hold for additional sources of variation.

Proofs

Proof of Result 1 Recall that Model (1) becomes:

$$Y_{pj}^* = \mu + A_p + e_{pj},$$

with $A_p \sim N(0, \sigma_A^2/k)$ and $e_{pj} \sim N(0, \sigma^2/k)$, where $p = 1, \ldots, P_I$ and $j = 1, \ldots, n$. We then apply Equation (2) with with appropriate re-parameterizations, (with within-subject variance σ_e^2/k, between-subject variance σ_A^2/k, number of individuals is P_I, and number of repeated measurements is n) to obtain Equation (5). From the first-order Taylor series expansion of a scalar-valued function of two variables, we have:

$$Var\left(\widetilde{\gamma}\right) = Var\left(\frac{\widetilde{\sigma}_A^2}{\widetilde{\sigma}_A^2 + \widetilde{\sigma}_e^2}\right)$$
$$\simeq \frac{\left(\sigma_e^2\right)^2}{\left(\sigma_A^2 + \sigma_e^2\right)^4} Var\left(\widetilde{\sigma}_A^2\right) + \frac{\left(\sigma_A^2\right)^2}{\left(\sigma_A^2 + \sigma_e^2\right)^4} Var\left(\widetilde{\sigma}_e^2\right) - 2\frac{\sigma_e^2\sigma_A^2}{\left(\sigma_A^2 + \sigma_e^2\right)^4} Cov\left(\widetilde{\sigma}_A^2, \widetilde{\sigma}_e^2\right). \qquad (S1)$$

Equation (6) follows from combining Equation ($S1$) with Equation (5). □

Proof of Result 2 Recall that Model (1) becomes:

$$Y_{ip}^* = \mu + A_i + e_{ip},$$

with $A_i \sim N(0, \sigma_A^2)$ and $e_{ip} \sim N(0, \sigma_e^2/k)$, where $p = 1, \ldots, P_T$ and $i = 1, \ldots, l$. We apply Equation (2) with with appropriate re-parameterizations, (with within-subject variance σ_e^2/k, between-subject variance σ_A^2, number of initials is l, and number of repeated measurements is P_T) to obtain Equation (7). Equation (8) follows from combining Equation ($S1$) with Equation (7). □

Proof of Result 3 It is easy to see that Equation (10) is a quadratic equation in x, $Ax^2 + Bx + C = 0$ with $A < 0, B > 0, C > 0$. From Descartes' rule of signs, it follows that Equation (10) has one positive root. □

Proof of Result 4 Let Y_{ip}^* be the average value of measurements of individual $i, i = 1, \ldots, l$ in p-th pooling ($p = 1, \ldots, P$) of size k_{ip} ($k_{ip} \geq 1$), $\sum_{p=1}^{P} k_{ip} = n_i$. Let $Y^* = (Y_1^*, \ldots, Y_l^*)$, where $Y_i^* = (Y_{i1}^*, \ldots, Y_{iP}^*)$, $i = 1, \ldots, l$. From Equation (4) with the given pooling design, it follows that $Y_i^* \sim MVN(\mu 1_P, V_i^*)$, $i = 1, \ldots, l$ and $Y_1^*, \ldots, Y_l^*$ are independent with $V_i^* = \sigma_e^2 diag\left(\frac{1}{k_{i1}}, \ldots, \frac{1}{k_{iP}}\right) + \sigma_A^2 J_P$, where $J_P = 1_P 1_P'$. Applying a standard linear algebra calculation (see, for example, Rao (1973)), we have

$$V_i^{*-1} = 1/\sigma_e^2 diag\left(k_{i1}, \ldots, k_{iP}\right) - \frac{1}{\sigma_e^2} \frac{\sigma_A^2}{\sigma_e^2 + n_i \sigma_A^2} \left(k_{i1}, \ldots, k_{iP}\right)' \left(k_{i1}, \ldots, k_{iP}\right)$$

$$\text{and } |V_i^*| = \frac{\left(\sigma_e^2\right)^{P-1}}{\prod_{p=1}^{P} k_{ip}} \left(\sigma_e^2 + n_i \sigma_A^2\right).$$

Therefore,

$$\begin{aligned}
L^* &= logL\left(Y_1^*,\dots,Y_l^*;\mu,\sigma_e^2,\sigma_A^2\right)\\
&= -\frac{lP}{2}\left(2\pi\right) - \frac{1}{2}\sum_{i=1}^{l} log|V_i^*| - \frac{1}{2}\sum_{i=1}^{l}\left(Y_i^* - \mu 1_P\right)' V_i^{*\,-1}\left(Y_i^* - \mu 1_P\right)\\
&= -\frac{l(P-1)}{2} log(\sigma_e^2) - \frac{1}{2}\sum_{i=1}^{l} log(\sigma_e^2 + n_i\sigma_A^2)\\
&\quad -\frac{1}{2\sigma_e^2}\left(\sum_{i=1}^{l}\sum_{p=1}^{P} k_{ip}(Y_{ip}^* - \mu)^2 - \sum_{i=1}^{l}\frac{n_i\sigma_A^2}{\sigma_e^2 + n_i\sigma_A^2} n_i(Y_{i\cdot} - \mu)^2\right) + Cons\\
&= -\frac{l(P-1)}{2} log(\sigma_e^2) - \frac{1}{2}\sum_{i=1}^{l} log(\sigma_e^2 + n_i\sigma_A^2)\\
&\quad -\frac{1}{2\sigma_e^2}\left(\sum_{i=1}^{l}\sum_{p=1}^{P} k_{ip}(Y_{ip}^* - Y_{i\cdot})^2 + \sum_{i=1}^{l}\left(1 - \frac{n_i\sigma_A^2}{\sigma_e^2 + n_i\sigma_A^2}\right)\left\{n_i(Y_{ip}^* - Y_{\cdot\cdot})^2 + n_i(Y_{\cdot\cdot} - \mu^2)\right\}\right) + Cons\\
&= -\frac{l(P-1)}{2} log(\sigma_e^2) - \frac{1}{2}\sum_{i=1}^{l} log(\tau_i^2) - \frac{SSE^*}{2\sigma_e^2} - \frac{1}{2}\sum_{i=1}^{l}\frac{SSA_i^*}{\tau_i^2} - \frac{1}{2}\sum_{i=1}^{l}\frac{n_i(Y_{\cdot\cdot} - \mu)^2}{\tau_i^2} + Cons,
\end{aligned}$$

where $\tau_i^2 = \sigma_e^2 + n_i\sigma_A^2$, $SSE^* = \sum_{i=1}^{l}\sum_{p=1}^{P} k_{ip}(Y_{ip}^* - Y_{i\cdot})^2$ and $SSA_i^* = n_i(Y_{ip}^* - Y_{\cdot\cdot})^2$. Calculating the information matrix using the second derivatives of log-likelihood L^*, we obtain the information matrix under design, which we will call Design T*:

$$I_{T^*}\left(\widetilde{\mu},\widetilde{\sigma}_e^2,\widetilde{\sigma}_A^2\right) = \frac{1}{2}\begin{pmatrix} 2\sum_{i=1}^{l}\frac{n_i}{\tau_i^2}, & 0, & 0\\ 0, & \frac{l(P-1)}{\sigma_e^4} + \sum_{i=1}^{l}\frac{1}{\tau_i^4}, & \sum_{i=1}^{l}\frac{n_i}{\tau_i^4}\\ 0, & \sum_{i=1}^{l}\frac{n_i}{\tau_i^4}, & \sum_{i=1}^{l}\frac{n_i^2}{\tau_i^4}\end{pmatrix},\ \ \tau_i^2 = \sigma_e^2 + n_i\sigma_A^2,\ \ i = 1,\dots,l.$$

Result 4 follows by noticing that the above information matrix does not depend on the values of pooled group sizes, k_{ip}, $i = 1,\dots,l$; $p = 1,\dots,P$, and depends only on the number of pooled groups P for each individual $i = 1,\dots l$. □

Proof of Result 5 From direct calculations (see, for example, Searle et al. (1992)) it follows that under the full data design, which we call Design F*, the information matrix under the unbalanced fixed design is,

$$I_{F^*}\left(\widetilde{\mu},\widetilde{\sigma}_e^2,\widetilde{\sigma}_A^2\right)=\frac{1}{2}\begin{pmatrix} 2\sum_{i=1}^{l}\frac{n_i}{\tau_i^2}, & 0, & 0 \\ 0, & \frac{\sum_{i=1}^{l}(n_i-1)}{\sigma_e^4}+\sum_{i=1}^{l}\frac{1}{\tau_i^4}, & \sum_{i=1}^{l}\frac{n_i}{\tau_i^4} \\ 0, & \sum_{i=1}^{l}\frac{n_i}{\tau_i^4}, & \sum_{i=1}^{l}\frac{n_i^2}{\tau_i^4} \end{pmatrix}.$$

We have,

$$I_{F^*}\left(\widetilde{\mu},\widetilde{\sigma}_e^2,\widetilde{\sigma}_A^2\right)-I_{T^*}\left(\widetilde{\mu},\widetilde{\sigma}_e^2,\widetilde{\sigma}_A^2\right)=\begin{pmatrix} 0, & 0, & 0 \\ 0, & \frac{1}{2}\frac{\sum_{i=1}^{l}(n_i-1)-l(P-1)}{\sigma_e^4} & 0 \\ 0, & 0, & 0 \end{pmatrix}.$$

From the fact that $E_G(n_i)=n$ for all i, Result 5 follows. □

References

Herbach, L.H. (1959). Properties of Model II type analysis of variance tests, A: optimum nature of the F-test for Model II in the balanced case. *The Annals of Mathematical Statistics* **30,** 939–959.

Lehmann, E.L., and Casella, G. (2nd ed. 1998). *Theory of Point Estimation*. New York: Springer-Verlag.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications* . New York: John Wiley & Sons.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components*. New York: John Wiley & Sons.