Prediction for mixed models with misspecified random effects distribution

Quan Vu
Joint work with Francis Hui, Samuel Müller, and Alan Welsh

November 2023

Research School of Finance, Actuarial Studies and Statistics
Australian National University

Motivation

Mixed models

- Mixed-effects models are widely used in several areas such as biology, ecology, and environmental sciences, because of their capability to model clustered or longitudinal data.

- Various assumptions are made when using mixed-effects models, including specifying a distribution for the random effects (usually a normal distribution).

Random effects distribution

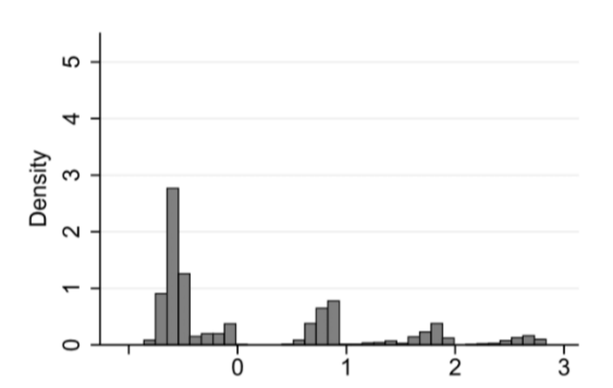
- Question: What happens when the random effects distribution is misspecified?

- Most of the previous works in the literature focus on estimation of fixed effects and variance components under misspecification of random effects (e.g., McCulloch and Neuhaus, 2011; Hui et al., 2021).

- Not much is known about prediction under misspecification of random effects distribution.

Random effects distribution

- Example: Predicted random effects assuming a normal random effects distribution for HERS data (McCulloch and Neuhaus, 2011).



Random effects distribution

- In many applications, the main interest is the prediction of random effects (or functions of random effects), e.g., small area estimation.

- Our goal is to investigate the impact of misspecifying the random effects distribution on prediction and prediction uncertainty.

Prediction for misspecified mixed models

Linear mixed models

- In this talk, we will be focusing on the linear mixed model, in which

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + \epsilon_{ij}.$$

- The true random effects distribution $p(u_i)$ is a mixture of normal distributions.

- The errors ϵ_{ij} is assumed normal with mean 0 and variance τ^2.

Best prediction

- Need to compare prediction under the misspecified distribution

$$p_*(u_i) = \mathcal{N}(0, \sigma_*^2),$$

  against prediction under the true distribution of the random effects

$$p_0(u_i) = \sum_{k=1}^{c} \pi_k \mathcal{N}(\mu_{0k}, \sigma_0^2 \sigma_{0k}^2)$$

  (with constraints for the parameters of the mixture components).

- To predict the random effects u_i, we use the best predictor (e.g., Jiang, 2003), which is given by

$$w_i = \mathrm{E}(u_i \mid \mathbf{y}_i) = \int u_i p(u_i \mid \mathbf{y}_i) \mathrm{d}u_i.$$

Best prediction

- Under the misspecified model, the best predictor w_{i*} of the random effect u_i is given by

$$w_{i*}(\boldsymbol{\theta}_*) = (\sigma_*^{-2} + n_i \tau_*^{-2})^{-1} \tau_*^{-2} \sum_{j=1}^{n_i} (y_{ij} - \mathbf{X}_{ij} \boldsymbol{\beta}_*).$$

- Under the true model, the best predictor w_{i0} of the random effect u_i is given by

$$w_{i0}(\boldsymbol{\theta}_0) = \frac{\sum_{k=1}^{c} \pi_k p_k(\mathbf{y}_i) m_k}{\sum_{k=1}^{c} \pi_k p_k(\mathbf{y}_i)},$$

  where
  $m_k = (\sigma_0^{-2}\sigma_{0k}^{-2} + n_i \tau_0^{-2})^{-1}(\sigma_0^{-2}\sigma_{0k}^{-2}\mu_{0k} + \tau_0^{-2} \sum_{j=1}^{n_i}(y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_0))$,
  and $p_k(\mathbf{y}_i) = \mathcal{N}(\mathbf{X}_i \boldsymbol{\beta}_0 + \sigma_0 \boldsymbol{\mu}_{0k} \mathbf{1}, \sigma_0^2 \sigma_{0k}^2 \mathbf{1} + \tau_0^2 \mathbf{I})$.

Mean squared error of prediction

- Under the misspecified model, the conditional mean squared error of prediction (Booth and Hobert, 1998) is given by

$$\mathrm{E}[(w_i - u_i)^2 \mid \mathbf{y}_i] = v_{i*}(\boldsymbol{\theta}_*) = (\sigma_*^{-2} + n_i \tau_*^{-2})^{-1}.$$

- Under the misspecified model, the unconditional mean squared error of prediction is given by

$$\mathrm{E}[(w_i - u_i)^2] = \mathrm{E}[\mathrm{E}[(w_i - u_i)^2 \mid \mathbf{y}_i]] = (\sigma_*^{-2} + n_i \tau_*^{-2})^{-1}.$$

Mean squared error of prediction

- Under the true model, the conditional mean squared error of prediction is given by

$$\mathrm{E}[(w_i - u_i)^2 \mid \mathbf{y}_i] = v_{i0}(\boldsymbol{\theta}_0) = \frac{\sum_{k=1}^{c} \pi_k p_k(\mathbf{y}_i)(v_k + m_k^2)}{\sum_{k=1}^{c} \pi_k p_k(\mathbf{y}_i)} - w_{i0}^2,$$

  where $v_k = (\sigma_0^{-2}\sigma_{0k}^{-2} + n_i\tau_0^{-2})^{-1}$.

- Under the true model, the unconditional mean squared error of prediction is given by

$$\mathrm{E}[(w_i - u_i)^2] = \mathrm{E}[\mathrm{E}[(w_i - u_i)^2 \mid \mathbf{y}_i]] = \mathrm{E}\left[\frac{\sum_{k=1}^{c} \pi_k p_k(\mathbf{y}_i)(v_k + m_k^2)}{\sum_{k=1}^{c} \pi_k p_k(\mathbf{y}_i)} - w_{i0}^2\right].$$

Prediction interval

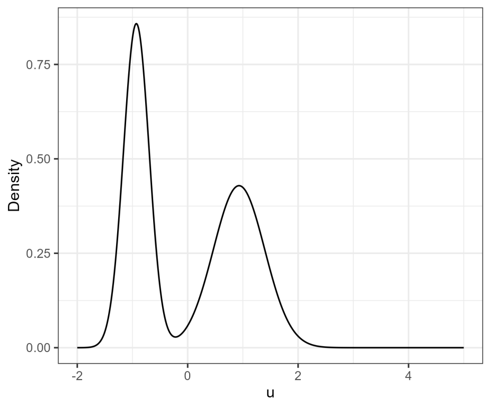
- A prediction interval can be constructed using the predictor and its mean squared error of prediction. For example, a 95% prediction interval is given by

$$[w_i - 1.96\text{msep}(w_i), w_i + 1.96\text{msep}(w_i)]$$

Simulation studies

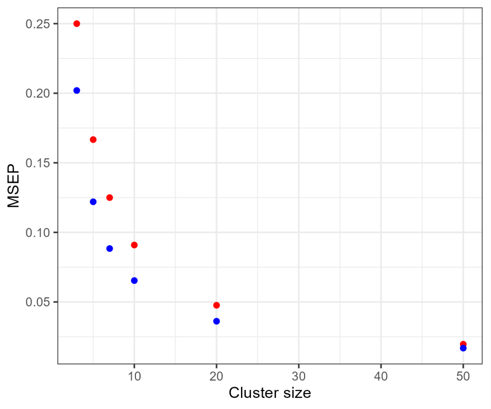
Results

- Example 1: Let the true random effects distribution be the mixture distribution: $p_0(u_i) = 0.5\mathcal{N}(-0.93, 0.23^2) + 0.5\mathcal{N}(0.93, 0.46^2)$. Let $\sigma^2 = \tau^2 = 1$, $\beta = (2, 1)'$, and number of clusters $= 50$.



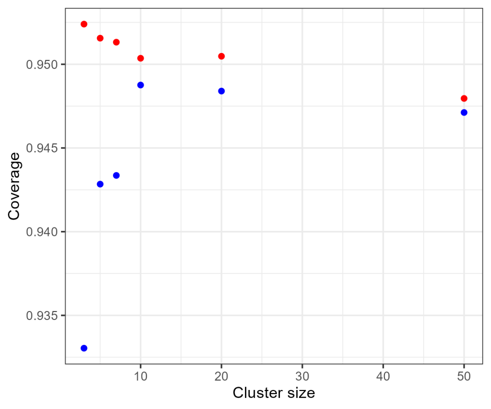
Results

Compare MSEP under the true (blue) and misspecified (red) models.



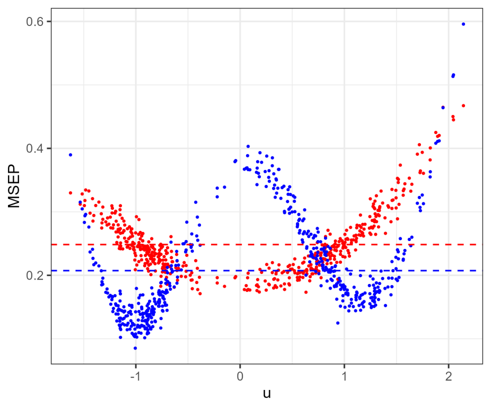
Results

Compare the coverage of the 95% prediction interval under the true (blue) and misspecified (red) models.



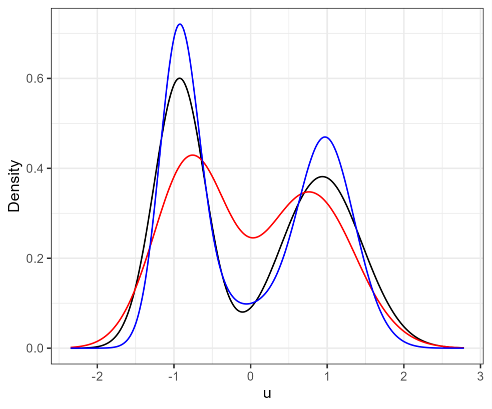
Results

Compare MSEP under the true (blue) and misspecified (red) models conditioned on the random effects u.



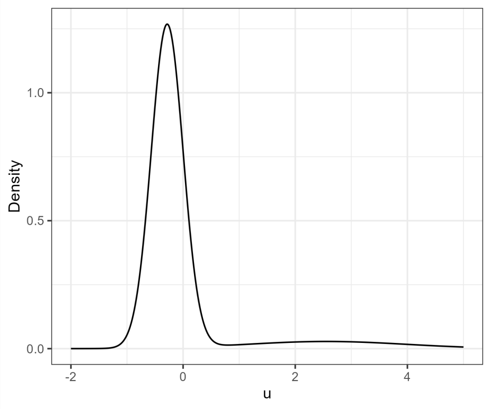
Results

Compare the shape of the distribution of the predicted random effects.



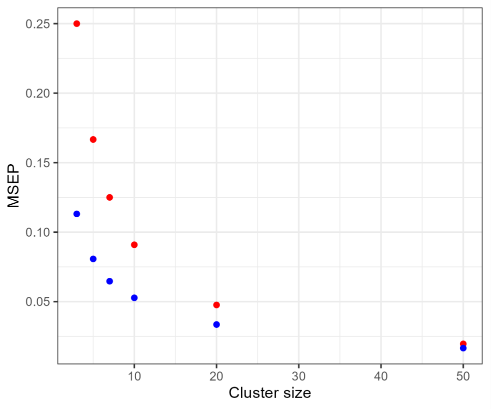
Results

- Example 2: Let the true random effects distribution be the mixture distribution: $p_0(u_i) = 0.9\mathcal{N}(-0.28, 0.28^2) + 0.1\mathcal{N}(2.56, 1.42^2)$. Let $\sigma^2 = \tau^2 = 1$, $\beta = (2, 1)'$, and number of clusters $= 50$.



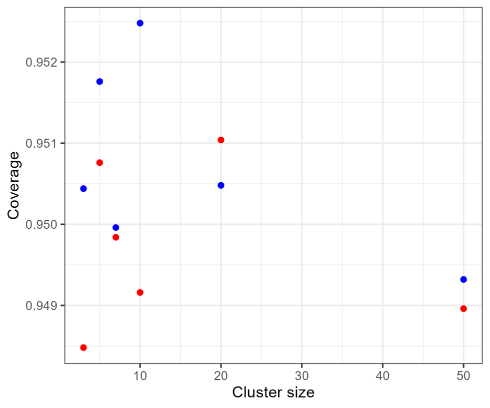
Results

Compare MSEP under the true (blue) and misspecified (red) models.



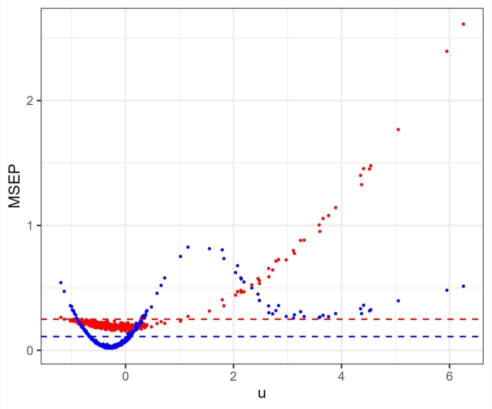
Results

Compare the coverage of the 95% prediction interval under the true (blue) and misspecified (red) models.



Results

Compare MSEP under the true (blue) and misspecified (red) models conditioned on the random effects u.



Discussion

Impact of misspecifying random effects distribution

- Misspecification of random effects distribution can result in larger mean squared error of prediction, especially for small cluster size. Also, MSEPs conditioned on the random effects u can be quite problematic.

- The coverage of prediction intervals under the misspecified random effects distribution is good, but the intervals are wider (as the MSEPs are larger).

Empirical best prediction

- For a real application, we do not know the true values of the parameters.
- Hence, we will need to use the empirical best predictor, which is the best predictor evaluated at the estimated values of the parameters.
- Estimation with a non-normal random effects distribution is more challenging than with the normality assumption.

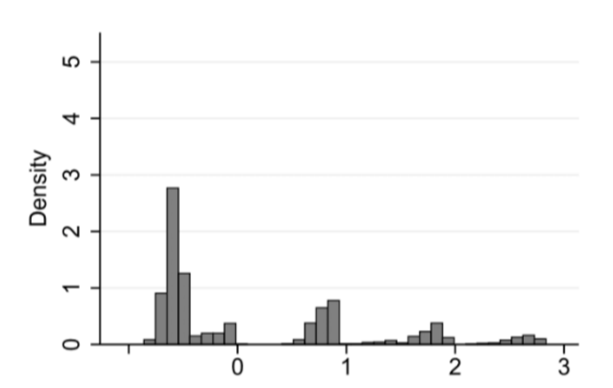
GLMMs

- For future work, we will investigate non-Gaussian responses (such as Poisson, binomial, etc).

- One challenging aspect with GLMMs is that evaluation of the likelihood and the best predictor involve intractable integrals.

Final message

- If the predicted random effects look non-normal, we might want to consider a more general distribution (rather than normal) for the random effects.



Thank you!

References

Booth, J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93:262–272.

Hui, F. K., Müller, S., and Welsh, A. H. (2021). Random effects misspecification can have severe consequences for random effects inference in linear mixed models. *International Statistical Review*, 89:186–206.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111:117–127.

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, 26:388–402.