Asymptotic Results for Penalized Quasi-Likelihood Estimation in Generalized Linear Mixed Models

The Australian National University

Supplementary Material

The supplementary material is organised as follows. In Section S1, we prove consistency results for the PQL estimator separately for the conditional and unconditional regimes. Section S2 focuses on the distributional results, and Section S3 treats the remainder term in the Taylor expansion. Section S4 contains results for unpartnered fixed effects, for some special cases. Section S5 provides extra simulation results, such as for the conditional regime and different choices of $\hat{\boldsymbol{G}}$.

In the developments, we prove all results below assuming the working dispersion parameter $\hat{\phi}$ is equal to the true dispersion parameter $\dot{\phi}$. Then for the general result using any $O_p(1)$ working $\hat{\phi}$, we note that solving

$$\nabla Q(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \hat{\phi}^{-1}\boldsymbol{X}^\top\{\boldsymbol{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})\} \\ \hat{\phi}^{-1}\boldsymbol{Z}^\top\{\boldsymbol{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})\} - (\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1})\hat{\boldsymbol{b}} \end{bmatrix} = \boldsymbol{0}_{(m+1)p}$$

for $\hat{\boldsymbol{\theta}}$ is equivalent to solving

$$\nabla Q(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \dot{\phi}^{-1}\boldsymbol{X}^\top\{\boldsymbol{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})\} \\ \dot{\phi}^{-1}\boldsymbol{Z}^\top\{\boldsymbol{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})\} - (\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}_s^{-1})\hat{\boldsymbol{b}} \end{bmatrix} = \mathbf{0}_{(m+1)p},$$

where $\hat{\boldsymbol{G}}_s = \dot{\phi}\hat{\phi}^{-1}\hat{\boldsymbol{G}}$, whose inverse is still $O_p(1)$ and positive definite. This is equivalent to setting $\hat{\phi}$ to $\dot{\phi}$ and scaling $\hat{\boldsymbol{G}}$ by $\dot{\phi}\hat{\phi}^{-1}$. The general result then follows since the results proved under $\hat{\phi} = \dot{\phi}$ hold for any $\hat{\boldsymbol{G}}$ that has an $O_p(1)$, positive definite inverse.

S0.1 Bias and Identifiability in the Conditional Regime

By differentiating (2.2), we see that the PQL estimators satisfy $\sum_{i=1}^m \hat{\phi}^{-1}\boldsymbol{X}_i^\top\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\theta}})\} = \mathbf{0}$ and $\hat{\phi}^{-1}\boldsymbol{Z}_i^\top\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\theta}})\} - \boldsymbol{G}^{-1}\hat{\boldsymbol{b}}_i = \mathbf{0}$, $i = 1, \ldots, m$. Summing both sides of the second equation across all i, since $\boldsymbol{X}_i = \boldsymbol{Z}_i$, it follows that $\sum_{i=1}^m \hat{\boldsymbol{b}}_i = \mathbf{0}_p$. That is, the PQL estimators of the random effects must satisfy a sum-to-zero constraint regardless of the underlying true parameter values. Under a conditional regime, this induces an asymptotic bias as captured by the term $\mathbf{1}_m^* \otimes (m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i)$ in Theorem 1, which can be interpreted as shifting the mean of the random effects into the corresponding fixed effects. We can deal with the bias by reparametrising the model *a priori* to satisfy a sum-to-zero constraint. That is, we can define a reparametrized vector of true values $\dot{\boldsymbol{\theta}}^*$ which

satisfy $\mathbf{1}_m^* \otimes (m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i^*) = \mathbf{0}_{(m+1)p}$, and the PQL estimator will then be asymptotically normally distributed centered around $\dot{\boldsymbol{\theta}}^*$. Furthermore, Theorem 1 remains practically useful as, for any given sample size, we can always reparameterise the GLMM to satisfy this identifiability constraint.

The asymptotic bias discussed above is analogous to that seen in a over-parametrized one-way analysis of variance (ANOVA) model. That is, in the ANOVA model one can always reparametrise to satisfy a sum-to-zero constraint, and the corresponding estimator is consistent for this vector of the reparametrized true values. Note however that when we work unconditionally (Section 4), reparametrising in this way will lead to a different model to the original, since the clusters are no longer independent.

S1 Proofs for Consistency

To establish our large sample distributional results, we first require the following consistency result.

Lemma 1. *Suppose Conditions (C1)-(C5) hold and* $mn_L^{-2} \to 0$. *Then, as* $m, n_L \to \infty$ *and unconditional on the random effects* $\dot{\boldsymbol{b}}$, $\|\hat{\dot{\boldsymbol{\theta}}} - \dot{\boldsymbol{\theta}}\|_\infty = o_p(1)$.

These results are required to control the remainder term in the Taylor expansions we use to derive the distributional results in Section S2. To prove the result, we wish to show that for any given $\epsilon > 0$, there exists a

large enough constant $C > 0$ such that, for large m, n_L, we have

$$P\left\{\sup_{\|\boldsymbol{u}\|_\infty = C} Q(\dot{\boldsymbol{\theta}} + \delta_{m,n_L}^{-1}\boldsymbol{u}) < Q(\dot{\boldsymbol{\theta}})\right\} \geq 1 - \epsilon,$$

for some positive, unbounded, monotonically increasing sequence δ_{m,n_L}. The above result implies that with probability tending to one, there exists a local maximum $\hat{\boldsymbol{\theta}}$ in the ball $\{\dot{\boldsymbol{\theta}} + \delta_{m,n_L}^{-1}\boldsymbol{u} : \|\boldsymbol{u}\|_\infty \leq C\}$ so that $\|\delta_{m,n_L}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})\|_\infty = O_p(1)$, and thus $\|\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}\|_\infty = o_p(1)$.

Consider the difference $Q(\dot{\boldsymbol{\theta}} + \boldsymbol{u}) - Q(\dot{\boldsymbol{\theta}})$. By a Taylor expansion, we obtain

$$Q(\dot{\boldsymbol{\theta}} + \boldsymbol{u}) - Q(\dot{\boldsymbol{\theta}}) = \boldsymbol{u}^\top\{\nabla Q(\dot{\boldsymbol{\theta}})\} - 0.5\boldsymbol{u}^\top\{-\nabla^2 Q(\bar{\boldsymbol{\theta}})\}\boldsymbol{u}. \tag{S1.1}$$

where $\bar{\boldsymbol{\theta}}$ lies on the line segment joining $\dot{\boldsymbol{\theta}}$ and $\dot{\boldsymbol{\theta}} + \boldsymbol{u}$. If we can prove that (S1.1) is negative as $m, n_L \to \infty$ for any choice of C, then there must exist some δ_{m,n_L} such that $Q(\dot{\boldsymbol{\theta}} + \delta_{m,n_L}^{-1}\boldsymbol{u}) - Q(\dot{\boldsymbol{\theta}})$ is negative for large enough C, and the required result follows. We have

$$\begin{aligned}\nabla Q(\dot{\boldsymbol{\theta}}) &= \begin{bmatrix} \dot{\phi}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}}) \\ \dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}}) - (\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1})\dot{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \dot{\phi}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}}) \\ \dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}}) \end{bmatrix} + \begin{bmatrix} \boldsymbol{0}_p \\ -(\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1})\dot{\boldsymbol{b}} \end{bmatrix} \\ &\triangleq \boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2,\end{aligned}$$

and

$$
\begin{aligned}
-\nabla^2 Q(\bar{\boldsymbol{\theta}}) &= \begin{bmatrix}
\boldsymbol{X}^\top \bar{\boldsymbol{W}} \boldsymbol{X} & \boldsymbol{X}_1^\top \bar{\boldsymbol{W}}_1 \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_m^\top \bar{\boldsymbol{W}}_m \boldsymbol{X}_m \\
\boldsymbol{X}_1^\top \bar{\boldsymbol{W}}_1 \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \bar{\boldsymbol{W}}_1 \boldsymbol{X}_1 + \hat{\boldsymbol{G}}^{-1} & & \boldsymbol{0} \\
\vdots & \vdots & \ddots & \vdots \\
\boldsymbol{X}_m^\top \bar{\boldsymbol{W}}_m \boldsymbol{X}_m & \boldsymbol{0} & & \boldsymbol{X}_m^\top \bar{\boldsymbol{W}}_m \boldsymbol{X}_m + \hat{\boldsymbol{G}}^{-1}
\end{bmatrix} \\
&= \begin{bmatrix}
\boldsymbol{X}^\top \bar{\boldsymbol{W}} \boldsymbol{X} & \boldsymbol{X}_1^\top \bar{\boldsymbol{W}}_1 \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_m^\top \bar{\boldsymbol{W}}_m \boldsymbol{X}_m \\
\boldsymbol{X}_1^\top \bar{\boldsymbol{W}}_1 \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \bar{\boldsymbol{W}}_1 \boldsymbol{X}_1 & & \boldsymbol{0} \\
\vdots & \vdots & \ddots & \vdots \\
\boldsymbol{X}_m^\top \bar{\boldsymbol{W}}_m \boldsymbol{X}_m & \boldsymbol{0} & & \boldsymbol{X}_m^\top \bar{\boldsymbol{W}}_m \boldsymbol{X}_m
\end{bmatrix} + \text{blockdiag}(\boldsymbol{0}_p, \boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1}) \\
&\triangleq \boldsymbol{\Gamma}_1(\bar{\boldsymbol{\theta}}) + \boldsymbol{\Gamma}_2,
\end{aligned}
$$

where $\bar{\boldsymbol{W}}_i = \dot{\phi}^{-1}\text{diag}\{a''(\bar{\eta}_{i1}), \ldots, a''(\bar{\eta}_{in_i})\}$ and $\bar{\boldsymbol{W}} = \dot{\phi}^{-1}\text{diag}\{a''(\bar{\eta}_{11}), \ldots, a''(\bar{\eta}_{mn_m})\}$. Also, let $\boldsymbol{\Gamma}_1(\dot{\boldsymbol{\theta}}) + \boldsymbol{\Gamma}_2$ denote the analogous decomposition of $-\nabla^2 Q(\dot{\boldsymbol{\theta}})$. For both the conditional and unconditional regimes, we will prove that the second term is positive and dominates the first. However, the treatment of the terms differs between the two cases, and as such the proofs will need to be dealt with separately. In the following three sections, we will first treat the Poisson pure random intercept example, followed by the more general conditional and unconditional regimes.

Before proceeding, we demonstrate an inequality that is used in the

proofs below. Write $\boldsymbol{u} = (\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top)^\top$, $\boldsymbol{u}_2 = (\boldsymbol{u}_{21}^\top, \ldots, \boldsymbol{u}_{2m}^\top)^\top$. First, for any $\boldsymbol{\theta}$ we have

$$\boldsymbol{u}^\top \boldsymbol{\Gamma}_1(\boldsymbol{\theta})\boldsymbol{u} = \boldsymbol{u}_1^\top \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} \boldsymbol{u}_1 + 2\boldsymbol{u}_1^\top \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{Z} \boldsymbol{u}_2 + \boldsymbol{u}_2^\top \boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} \boldsymbol{u}_2^\top \geq 0.$$

Next, we have

$$\begin{aligned}
&\boldsymbol{u}^\top \boldsymbol{\Gamma}_1(\bar{\boldsymbol{\theta}})\boldsymbol{u} - c_0^2 \boldsymbol{u}^\top \boldsymbol{\Gamma}_1(\boldsymbol{\theta})\boldsymbol{u} \\
&= \boldsymbol{u}_1^\top \boldsymbol{X}^\top (\bar{\boldsymbol{W}} - c_0^2 \boldsymbol{W}) \boldsymbol{X} \boldsymbol{u}_1 + 2\boldsymbol{u}_1^\top \boldsymbol{X}^\top (\bar{\boldsymbol{W}} - c_0^2 \boldsymbol{W}) \boldsymbol{Z} \boldsymbol{u}_2 + \boldsymbol{u}_2^\top \boldsymbol{Z}^\top (\bar{\boldsymbol{W}} - c_0^2 \boldsymbol{W}) \boldsymbol{Z} \boldsymbol{u}_2^\top.
\end{aligned}$$

If we denote $\boldsymbol{W}^* = \bar{\boldsymbol{W}} - c_0^2 \boldsymbol{W}$, then by Condition (C1) $\boldsymbol{W}^*$ is a diagonal matrix with non-negative entries as the entries of $c_0^2 \boldsymbol{W}$ are upper bounded by the smallest component in $\bar{\boldsymbol{W}}$. Therefore

$$\boldsymbol{u}^\top \boldsymbol{\Gamma}_1(\bar{\boldsymbol{\theta}})\boldsymbol{u} - c_0^2 \boldsymbol{u}^\top \boldsymbol{\Gamma}_1(\boldsymbol{\theta})\boldsymbol{u} = \boldsymbol{u}_1^\top \boldsymbol{X}^\top \boldsymbol{W}^* \boldsymbol{X} \boldsymbol{u}_1 + 2\boldsymbol{u}_1^\top \boldsymbol{X}^\top \boldsymbol{W}^* \boldsymbol{Z} \boldsymbol{u}_2 + \boldsymbol{u}_2^\top \boldsymbol{Z}^\top \boldsymbol{W}^* \boldsymbol{Z} \boldsymbol{u}_2^\top \geq 0,$$

so that $\boldsymbol{u}^\top \boldsymbol{\Gamma}_1(\bar{\boldsymbol{\theta}})\boldsymbol{u} \geq c_0^2 \boldsymbol{u}^\top \boldsymbol{\Gamma}_1(\boldsymbol{\theta})\boldsymbol{u}$. Finally, note that we can choose $\boldsymbol{\theta} = \dot{\boldsymbol{\theta}}$ or $\boldsymbol{\theta} = E(\dot{\boldsymbol{\theta}})$ without altering the above argument.

S1.1 Poisson pure random intercept example

We begin with the Poisson pure random intercept example, which gives insight and covers a case where $\boldsymbol{X}_i \neq \boldsymbol{Z}_i$. The following result is unconditional on the random effects $\dot{\boldsymbol{b}}$.

Lemma 2. *Assume Conditions (C1)-(C5) hold, and let* $mn^{-2} \to 0$. *Then for the Poisson pure random intercept model, as* $m, n \to \infty$ *and unconditional on the random effects* $\dot{\boldsymbol{b}}$, *it holds that* $\|\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}\|_\infty = o_p(1)$.

Proof. Let $\boldsymbol{u} = \boldsymbol{u}_2 = (u_{21}, \ldots, u_{2m})^\top$, $\boldsymbol{\theta} = \boldsymbol{b} = (b_1, \ldots, b_m)^\top$, $\hat{\boldsymbol{G}} = \hat{\sigma}_b^2$ (a scalar),

$-\nabla^2 Q(\bar{\boldsymbol{\theta}}) = \mathrm{diag}(ne^{\bar{b}_1} + \hat{\sigma}_b^{-2}, \ldots, ne^{\bar{b}_m} + \hat{\sigma}_b^{-2}) \equiv \boldsymbol{\Gamma}_1(\bar{\boldsymbol{\theta}}) + \boldsymbol{\Gamma}_2$, and

$$\nabla Q(\dot{\boldsymbol{\theta}}) = \begin{bmatrix} \sum_{j=1}^n (y_{1j} - e^{\dot{b}_1}) - \hat{\sigma}_b^{-2}\dot{b}_1 \\ \vdots \\ \sum_{j=1}^n (y_{mj} - e^{\dot{b}_m}) - \hat{\sigma}_b^{-2}\dot{b}_m \end{bmatrix} \equiv \boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2.$$

Let $\boldsymbol{M} = E\{\mathrm{diag}(ne^{\dot{b}_1}, \ldots, ne^{\dot{b}_m})\}$. Then $\boldsymbol{M} = \mathrm{Var}\{\dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}})\}$. By Condition (C1), $c_0^2\boldsymbol{u}^\top\boldsymbol{M}\boldsymbol{u} \le \boldsymbol{u}^\top\boldsymbol{\Gamma}_1(\bar{\boldsymbol{\theta}})\boldsymbol{u}$. Next, let $\lambda = \dot{\hat{\sigma}}_b^2\hat{\sigma}_b^{-2}$. Then $\mathrm{Var}(\boldsymbol{\lambda}_2) = \dot{\hat{\sigma}}_b^2\hat{\sigma}_b^{-4}\boldsymbol{I}_m$ and

$$\lambda^{-1}\boldsymbol{u}_2^\top(\dot{\hat{\sigma}}_b^2\hat{\sigma}_b^{-4}\boldsymbol{I}_m)\boldsymbol{u}_2 = \boldsymbol{u}_2^\top(\hat{\sigma}_b^{-2}\boldsymbol{I}_m)\boldsymbol{u}_2 = \boldsymbol{u}^\top\boldsymbol{\Gamma}_2\boldsymbol{u}.$$

Finally, by the laws of iterated expectation and variance we have

$$\begin{aligned} E\{\nabla Q(\dot{\boldsymbol{\theta}})\nabla Q(\dot{\boldsymbol{\theta}})^\top\} &= \mathrm{Var}\{\nabla Q(\dot{\boldsymbol{\theta}})\} \\ &= E[\mathrm{Var}\{\nabla Q(\dot{\boldsymbol{\theta}})|\dot{\boldsymbol{b}}\}] + \mathrm{Var}[E\{\nabla Q(\dot{\boldsymbol{\theta}})|\dot{\boldsymbol{b}}\}] \\ &= E\{\mathrm{Var}(\boldsymbol{\lambda}_1|\dot{\boldsymbol{b}})\} + \mathrm{Var}(\boldsymbol{\lambda}_2) \end{aligned}$$

$$= \text{Var}(\boldsymbol{\lambda}_1) + \text{Var}(\boldsymbol{\lambda}_2).$$

Therefore, we have that

$$\begin{aligned}\boldsymbol{u}^\top\{-\nabla^2 Q(\bar{\boldsymbol{\theta}})\}\boldsymbol{u} &\geq \min(\lambda^{-1}, c_0^2)\{\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u} + \boldsymbol{u}_2^\top(\dot{\hat{\sigma}}_b^2\hat{\sigma}_b^{-4}\boldsymbol{I}_m)\boldsymbol{u}_2\} \\ &= \min(\lambda^{-1}, c_0^2)E\{\nabla Q(\dot{\boldsymbol{\theta}})\nabla Q(\dot{\boldsymbol{\theta}})^\top\},\end{aligned}$$

where the latter and hence former term grows at the same rate as $\{\boldsymbol{u}^\top\nabla Q(\dot{\boldsymbol{\theta}})\}^2$. Since at least one component of $\boldsymbol{u}$ equals $\pm C$, for any given $\boldsymbol{u}$, $\boldsymbol{u}^\top\{-\nabla^2 Q(\bar{\boldsymbol{\theta}})\}\boldsymbol{u}$ is at least of order $O_p(m)$ in probability and hence always dominates.

Since the choice of which $|u_{2i}| = C$ is arbitrary however, we also need to make sure that the mth order statistic $\max_{i\in\{1,\ldots,m\}}[\{\sum_{j=1}^n(y_{ij} - e^{\dot{b}_i}) - \hat{\sigma}_b^{-2}\dot{b}_i\}/(ne^{\dot{b}_i} + \hat{\sigma}_b^{-2})]$, which grows with the dimension, is of order $o_p(1)$. We know that the leading term in (4.3b) is $(\boldsymbol{Z}^\top\dot{\boldsymbol{W}}\boldsymbol{Z})^{-1}\{\dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}})\}$ when $mn^{-1} \to \infty$; for this Poisson random intercept example, up to some smaller order terms, this simplifies to the ratio $\{\sum_{j=1}^n(y_{ij} - e^{\dot{b}_i}) - \hat{\sigma}_b^{-2}\dot{b}_i\}/(ne^{\dot{b}_i} + \hat{\sigma}_b^{-2})$. Intuitively then, proving a result for $\|\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}\|_\infty$ should involve studying $\max_{i\in\{1,\ldots,m\}}[\{\sum_{j=1}^n(y_{ij} - e^{\dot{b}_i}) - \hat{\sigma}_b^{-2}\dot{b}_i\}/(ne^{\dot{b}_i} + \hat{\sigma}_b^{-2})]$.

Put another way, consider the set of $\boldsymbol{u}$ such that one component of $\boldsymbol{u}$ equals $\pm C$ and zero elsewhere. When C is the ith component of $\boldsymbol{u}$, this corresponds to deviating away from $\dot{\boldsymbol{\theta}}$ in the ith direction. In this case, we

need $C\{\sum_{j=1}^{n}(y_{ij}-e^{\dot{b}_i})-\hat{\sigma}_b^{-2}\dot{b}_i\}$ to be dominated by $C^2ne^{\dot{b}_i}$ for any C and all m, n large enough, i.e., $\{\sum_{j=1}^{n}(y_{ij}-e^{\dot{b}_i})-\hat{\sigma}_b^{-2}\dot{b}_i\}/ne^{\dot{b}_i} = o_p(1)$. This is indeed true as this ratio is $O_p(n^{-1/2})$, since $\sum_{j=1}^{n}(y_{ij}-e^{\dot{b}_i})-\hat{\sigma}_b^{-2}\dot{b}_i = O_p(n^{1/2})$ due to conditional independence and Chebyshev's inequality, and $e^{\dot{b}_i} = O_p(1)$. However, although the ratio is of order $O_p(n^{-1/2})$, for any given m, n there is still a positive probability that the ratio (a random variable) is greater than one in magnitude. On the other hand, for the consistency argument to hold we need to make sure the ratio is smaller than one in magnitude for all m directions with probability tending to one, as $m, n \to \infty$. In particular, it is sufficient for the maximum of m of these ratios to be $o_p(1)$: this maximum grows with m, corresponding to the number of directions we need to bound. Intuitively, this should hold if m does not grow too fast relative to n.

Now, Downey (1990) proves that the maximum over m realisations of independently and identically distributed random variables with a finite qth moment is $o_p(m^{1/q})$. By Condition (C5), the ratio $n^{1/2}\{\sum_{j=1}^{n}(y_{ij}-e^{\dot{b}_i})-\hat{\sigma}_b^{-2}\dot{b}_i\}/(ne^{\dot{b}_i}+\hat{\sigma}_b^{-2})$ has finite fourth moments for all i and n. Thus, the maximum of these (normalised) ratios over m clusters is of order $o_p(m^{1/4})$. As a result, the maximum ratio of interest is $o_p(m^{1/4}n^{-1/2})$. Therefore, when $mn^{-2} \to 0$, there exists $\delta_{m,n}$ such that we can always choose a large

enough C for $\delta_{m,n}^{-1}\boldsymbol{u}^\top \nabla Q(\dot{\boldsymbol{\theta}})$ to be dominated by $\delta_{m,n}^{-2}\boldsymbol{u}^\top\{-\nabla^2 Q(\bar{\boldsymbol{\theta}})\}\boldsymbol{u}$, and hence $\|\delta_{m,n}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})\|_\infty = O_p(1)$ as required.

□

To conclude this section, we remark that although $mn^{-2} \to 0$ is needed for the consistency and thus distributional result, this is a sufficient condition. Intuitively, in the Poisson pure random effects model there are no fixed parameters to estimate, and the estimate of the random effects for each cluster only depends on observations in that cluster. Thus, the relative rates of m and n should not matter for a distributional result concerning a finite subset of the random effects.

S1.2 Conditional on the Random Effects

In this section, we prove the consistency result under the conditional regime. In the conditional regime, we assume without loss of generality that $\sum_{i=1}^m \dot{\boldsymbol{b}}_i = \boldsymbol{0}_p$, recalling that we can always reparametrise the random effect coefficients so this holds.

Let $\boldsymbol{M} = \boldsymbol{\Gamma}_1(\dot{\boldsymbol{\theta}})$. Then $\boldsymbol{M} = \mathrm{Var}(\boldsymbol{\lambda}_1|\dot{\boldsymbol{b}}) = E(\boldsymbol{\lambda}_1\boldsymbol{\lambda}_1^\top|\dot{\boldsymbol{b}})$ since $E(\boldsymbol{\lambda}_1|\dot{\boldsymbol{b}}) = \boldsymbol{0}_{(m+1)p}$. By Condition (C1), we have $c_0^2\boldsymbol{u}^\top\boldsymbol{M}\boldsymbol{u} \leq \boldsymbol{u}^\top\boldsymbol{\Gamma}_1(\bar{\boldsymbol{\theta}})\boldsymbol{u}$.

We now consider two cases: the special case when $\boldsymbol{u}_1 = -\boldsymbol{u}_{2i}$ for all i, and when this is not the case. For the former, we have $\boldsymbol{u}^\top\boldsymbol{\lambda}_1 = \boldsymbol{u}^\top\boldsymbol{M}\boldsymbol{u} = 0$.

Then we must examine $\boldsymbol{u}^\top\boldsymbol{\lambda}_2$ and $\boldsymbol{u}^\top\boldsymbol{\Gamma}_2\boldsymbol{u}$. In this case, we have $\boldsymbol{u}^\top\boldsymbol{\lambda}_2 = \sum_{i=1}^m \boldsymbol{u}_{2i}^\top\hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i = -\boldsymbol{u}_1^\top\hat{\boldsymbol{G}}^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i = 0$, and $\boldsymbol{u}^\top\boldsymbol{\Gamma}_2\boldsymbol{u} = m\boldsymbol{u}_1^\top\hat{\boldsymbol{G}}^{-1}\boldsymbol{u}_1 > 0$ since $\hat{\boldsymbol{G}}$ is a positive definite matrix. Thus the difference (S1.1) is negative for large enough m, n_L and any choice of constant C.

Next, consider the case when $\boldsymbol{u}_1 = -\boldsymbol{u}_{2i}$ for all i does not hold. Under this setting, as $\boldsymbol{\Gamma}_2$ is a positive semi-definite matrix, we still have $\boldsymbol{u}^\top\{-\nabla^2 Q(\bar{\boldsymbol{\theta}})\}\boldsymbol{u} \geq \boldsymbol{u}^\top\boldsymbol{\Gamma}_1(\bar{\boldsymbol{\theta}})\boldsymbol{u} \geq c_0^2\boldsymbol{u}^\top\boldsymbol{M}\boldsymbol{u}$, where the last and hence former terms grow at the same rate as $(\boldsymbol{u}^\top\boldsymbol{\lambda}_1)^2$. Since at least one component of $\boldsymbol{u}$ equals $\pm C$, by Conditions (C1)-(C3) we have that $\boldsymbol{u}^\top\{-\nabla^2 Q(\bar{\boldsymbol{\theta}})\}\boldsymbol{u}$ is at least of order $O_p(n_L)$, and always dominates since $\boldsymbol{u}^\top\boldsymbol{\lambda}_2 = O_p(m)$ at most.

Since the choice of $\boldsymbol{u}$ is arbitrary, we must take into account the growth rate of the mth order statistic. That is, for any $1 \leq k \leq p$, we need $\max_{i\in\{1,\ldots,m\}}[(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\{\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) - \hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i\}]_{[k]} = o_p(1)$, as per the argument for the Poisson pure random intercept model. Since the responses y_{ij} are from the exponential family and thus the moment generating function always exists, the maximum is of order $o_p(m^{1/r}n_L^{-1/2})$ for any $r \in \mathbb{N}$ (Downey, 1990), and hence $o_p(1)$ since $mn_L^{-1} \to 0$ by taking $r = 2$, for example. Note that the first p components of $\nabla Q(\dot{\boldsymbol{\theta}})$, which are associated with the fixed effects, do not need to be bounded in this way

because the dimension is fixed.

S1.3 Unconditional on the Random Effects

In this section, we prove the consistency result under the unconditional regime. The main differences to the derivation under the conditional regime arise from the treatment of $\boldsymbol{\lambda}_2$, and the distribution of $\boldsymbol{y}$. In the unconditional regime it holds that $\sum_{i=1}^{m} \dot{\boldsymbol{b}}_i = O_p(m^{1/2})$, while in the conditional regime we impose a sum to zero constraint. Furthermore, in the unconditional regime we bound $\boldsymbol{u}^\top \boldsymbol{\lambda}_2$ using its variance, while in the conditional regime this is not possible because $\boldsymbol{\lambda}_2$ is not a random variable. Finally, in the unconditional regime we cannot use the properties of the exponential family to bound the mth order statistic, instead requiring Condition (C5).

Let $\boldsymbol{M} = E\{\boldsymbol{\Gamma}_1(\dot{\boldsymbol{\theta}})\}$. Then $\boldsymbol{M} = \mathrm{Var}(\boldsymbol{\lambda}_1) = E(\boldsymbol{\lambda}_1 \boldsymbol{\lambda}_1^\top)$ since $E(\boldsymbol{\lambda}_1) = \boldsymbol{0}_{(m+1)p}$. By Condition (C1), $c_0^2 \boldsymbol{u}^\top \boldsymbol{M} \boldsymbol{u} \leq \boldsymbol{u}^\top \boldsymbol{\Gamma}_1(\bar{\boldsymbol{\theta}}) \boldsymbol{u}$.

We consider two cases: the special case when $\boldsymbol{u}_1 = -\boldsymbol{u}_{2i}$ for all i, and when this is not the case. In the former, we have $\boldsymbol{u}^\top \boldsymbol{\lambda}_1 = \boldsymbol{u}^\top \boldsymbol{M} \boldsymbol{u} = 0$. Thus we must examine $\boldsymbol{u}^\top \boldsymbol{\lambda}_2$ and $\boldsymbol{u}^\top \boldsymbol{\Gamma}_2 \boldsymbol{u}$. In this case, we have $\boldsymbol{u}^\top \boldsymbol{\lambda}_2 = \sum_{i=1}^{m} \boldsymbol{u}_{2i}^\top \hat{\boldsymbol{G}}^{-1} \dot{\boldsymbol{b}}_i = -\boldsymbol{u}_1^\top \hat{\boldsymbol{G}}^{-1} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i = O_p(m^{1/2})$, and $\boldsymbol{u}^\top \boldsymbol{\Gamma}_2 \boldsymbol{u} = m \boldsymbol{u}_1^\top \hat{\boldsymbol{G}}^{-1} \boldsymbol{u}_1 > 0$ since $\hat{\boldsymbol{G}}$ is a positive definite matrix. Hence the difference (S1.1) is negative for large enough m, n_L, and any choice of constant C.

Next, consider the case when $\boldsymbol{u}_1 = -\boldsymbol{u}_{2i}$ for all i does not hold. Then we still have $\boldsymbol{u}^\top\{-\nabla^2 Q(\bar{\boldsymbol{\theta}})\}\boldsymbol{u} \geq c_0^2 \boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u}$. Letting $\lambda = \lambda_{\max}(\hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{G}}\hat{\boldsymbol{G}}^{-1})/\lambda_{\min}(\hat{\boldsymbol{G}}^{-1})$, we have

$$\mathrm{Var}(\boldsymbol{\lambda}_2) = \boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{G}}\hat{\boldsymbol{G}}^{-1}$$

and

$$\lambda^{-1}\boldsymbol{u}_2^\top(\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{G}}\hat{\boldsymbol{G}}^{-1})\boldsymbol{u}_2 \leq \boldsymbol{u}_2^\top(\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1})\boldsymbol{u}_2 = \boldsymbol{u}^\top\boldsymbol{\Gamma}_2\boldsymbol{u}.$$

Now, by the laws of iterated expectation and variance,

$$\begin{aligned}
E\{\nabla Q(\dot{\boldsymbol{\theta}})\nabla Q(\dot{\boldsymbol{\theta}})^\top\} &= \mathrm{Var}\{\nabla Q(\dot{\boldsymbol{\theta}})\} \\
&= E[\mathrm{Var}\{\nabla Q(\dot{\boldsymbol{\theta}})|\dot{\boldsymbol{b}}\}] + \mathrm{Var}[E\{\nabla Q(\dot{\boldsymbol{\theta}})|\dot{\boldsymbol{b}}\}] \\
&= E\{\mathrm{Var}(\boldsymbol{\lambda}_1|\dot{\boldsymbol{b}})\} + \mathrm{Var}(\boldsymbol{\lambda}_2) \\
&= \mathrm{Var}(\boldsymbol{\lambda}_1) + \mathrm{Var}(\boldsymbol{\lambda}_2).
\end{aligned}$$

Thus we have that

$$\begin{aligned}
\boldsymbol{u}^\top\{-\nabla^2 Q(\bar{\boldsymbol{\theta}})\}\boldsymbol{u} &\geq \min(\lambda^{-1}, c_0^2)\{\boldsymbol{u}^\top\boldsymbol{M}\boldsymbol{u} + \boldsymbol{u}_2^\top(\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{G}}\hat{\boldsymbol{G}}^{-1})\boldsymbol{u}_2\} \\
&= \min(\lambda^{-1}, c_0^2)\boldsymbol{u}^\top E\{\nabla Q(\dot{\boldsymbol{\theta}})\nabla Q(\dot{\boldsymbol{\theta}})^\top\}\boldsymbol{u},
\end{aligned}$$

where the latter and hence former term grows at the same rate as $\{\boldsymbol{u}^\top\nabla Q(\dot{\boldsymbol{\theta}})\}^2$.

Since at least one component of $\boldsymbol{u}$ equals $\pm C$, for any given $\boldsymbol{u}$ we have that $\boldsymbol{u}^\top\{-\nabla^2 Q(\bar{\boldsymbol{\theta}})\}\boldsymbol{u}$ is at least of order $O_p(n_L)$ and always dominates.

Since the choice of $\boldsymbol{u}$ is arbitrary, we must take into account the growth rate of the nth order statistic. That is, for any $1 \leq k \leq p$, we require $\max_{i\in\{1,\ldots,m\}}[(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\{\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) - \hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i\}]_{[k]} = o_p(1)$, as per the argument for the Poisson pure random intercept model. By Condition (C5), this term is of order $o_p(m^{1/4}n_L^{-1/2})$, and hence the result follows. Note that the first p components of $\nabla Q(\dot{\boldsymbol{\theta}})$, which are associated with the fixed effects, do not need to be bounded in this way because the dimension is fixed.

S2 Proofs of Distributional Results

For both the conditional and unconditional regimes, our proof relies on examining the behaviour of the leading term in the Taylor expansion of the estimating function. Under Conditions (C1) and (C3), we take the Taylor expansion of $\nabla Q(\hat{\boldsymbol{\theta}})$ around $\dot{\boldsymbol{\theta}}$ and obtain, as $m, n_L \to \infty$,

$$\nabla Q(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}_{(m+1)p} = \nabla Q(\dot{\boldsymbol{\theta}}) + \nabla^2 Q(\dot{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) + \frac{1}{2}\boldsymbol{R}(\tilde{\boldsymbol{\theta}}), \tag{S2.1}$$

where $\tilde{\boldsymbol{\theta}}$ is a $(m+1)p \times (m+1)p$ matrix with each row lying on the line segment between $\dot{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{R}(\tilde{\boldsymbol{\theta}})$ is the remainder term. Rearranging,

we have

$$\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}} = -\{\nabla^2 Q(\dot{\boldsymbol{\theta}})\}^{-1} \nabla Q(\dot{\boldsymbol{\theta}}) - \frac{1}{2}\{\nabla^2 Q(\dot{\boldsymbol{\theta}})\}^{-1} \boldsymbol{R}(\tilde{\boldsymbol{\theta}}). \tag{S2.2}$$

We show in Section S3 that the remainder term is of smaller order than the leading term and thus negligible in the limit, in both the conditional and unconditional regimes.

From (S2.2), to study the asymptotic behaviour of the PQL estimator we will first apply the blockwise matrix inversion formula to obtain an expression for $-\{\nabla^2 Q(\dot{\boldsymbol{\theta}})\}^{-1}$. Using this result, we will then obtain an expression for $-\{\nabla^2 Q(\dot{\boldsymbol{\theta}})\}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$, and subsequently study the asymptotic behaviour of each constituent term. Note that since $\nabla Q(\dot{\boldsymbol{\theta}})$ is a $(m+1)p$-vector and $-\{\nabla^2 Q(\dot{\boldsymbol{\theta}})\}^{-1}$ is a $(m+1)p \times (m+1)p$ matrix, we cannot simply take their limits as per standard fixed dimension asymptotics. Instead, we must evaluate $-\{\nabla^2 Q(\dot{\boldsymbol{\theta}})\}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ as a whole.

We can write

$$\nabla Q(\dot{\boldsymbol{\theta}}) = \begin{bmatrix} \dot{\phi}^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \dot{\boldsymbol{\mu}}) \\ \dot{\phi}^{-1} \boldsymbol{Z}^\top (\boldsymbol{y} - \dot{\boldsymbol{\mu}}) - (\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1}) \dot{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \dot{\phi}^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} (y_{ij} - \dot{\mu}_{ij}) \\ \dot{\phi}^{-1} \sum_{j=1}^{n_1} \boldsymbol{x}_{1j} (y_{1j} - \dot{\mu}_{1j}) - \hat{\boldsymbol{G}}^{-1} \dot{\boldsymbol{b}}_1 \\ \vdots \\ \dot{\phi}^{-1} \sum_{j=1}^{n_m} \boldsymbol{x}_{mj} (y_{mj} - \dot{\mu}_{mj}) - \hat{\boldsymbol{G}}^{-1} \dot{\boldsymbol{b}}_m \end{bmatrix}$$

$$\triangleq \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_{21} + \boldsymbol{S}_{31} \\ \vdots \\ \boldsymbol{S}_{2m} + \boldsymbol{S}_{3m} \end{bmatrix} \triangleq \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_4 + \boldsymbol{S}_5 \end{bmatrix} \triangleq \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_6 \end{bmatrix},$$

$$\boldsymbol{B}(\dot{\boldsymbol{\theta}}) = -\nabla^2 Q(\dot{\boldsymbol{\theta}}) = \begin{bmatrix} \boldsymbol{X}^\top \dot{\boldsymbol{W}} \boldsymbol{X} & \boldsymbol{X}^\top \dot{\boldsymbol{W}} \boldsymbol{Z} \\ \boldsymbol{Z}^\top \dot{\boldsymbol{W}} \boldsymbol{X} & \boldsymbol{Z}^\top \dot{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1} \end{bmatrix} \triangleq \begin{bmatrix} \underset{p\times p}{\boldsymbol{B}_1} & \underset{p\times mp}{\boldsymbol{B}_2} \\ \underset{mp\times p}{\boldsymbol{B}_2^\top} & \underset{mp\times mp}{\boldsymbol{B}_3 + \boldsymbol{B}_4} \end{bmatrix}.$$

Letting $\boldsymbol{C} = \boldsymbol{B}_1 - \boldsymbol{B}_2(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top$, by the matrix block inversion formula

we have

$$\boldsymbol{B}^{-1} = \begin{bmatrix} \boldsymbol{C}^{-1} & -\boldsymbol{C}^{-1}\boldsymbol{B}_2(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} \\ -(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top \boldsymbol{C}^{-1} & (\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} + (\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top \boldsymbol{C}^{-1}\boldsymbol{B}_2(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} \end{bmatrix}. \tag{S2.3}$$

Next, based on the forms of $\boldsymbol{B}_2$ and $(\boldsymbol{B}_3 + \boldsymbol{B}_4)$, we obtain

$$\boldsymbol{B}_2(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} = [\boldsymbol{I}_p - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1 + \hat{\boldsymbol{G}}^{-1})^{-1}, \ldots, \boldsymbol{I}_p - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m + \hat{\boldsymbol{G}}^{-1})^{-1}].$$

Then since $\boldsymbol{Z}_i = \boldsymbol{X}_i$ for all i, we can show that

$$\begin{aligned} \boldsymbol{B}_2(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top &= \sum_{i=1}^m \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i \\ &= \sum_{i=1}^m (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1} - \hat{\boldsymbol{G}}^{-1})(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i \end{aligned}$$

$$= \sum_{i=1}^{m} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i$$
$$= \boldsymbol{B}_1 - \sum_{i=1}^{m} \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i.$$

It follows that

$$\boldsymbol{C} = \sum_{i=1}^{m} \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i = \sum_{i=1}^{m} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \hat{\boldsymbol{G}}^{-1}, \tag{S2.4}$$

where the second equality arises from the fact that as a covariance matrix, $\boldsymbol{C}$ must be symmetric. We may also write $\boldsymbol{C}$ as

$$\begin{aligned} \sum_{i=1}^{m} \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i &= \sum_{i=1}^{m} \{\boldsymbol{I}_p - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\} \hat{\boldsymbol{G}}^{-1} \\ &= \hat{\boldsymbol{G}}^{-1} \sum_{i=1}^{m} \{\boldsymbol{I}_p - (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \hat{\boldsymbol{G}}^{-1}\}. \end{aligned} \tag{S2.5}$$

Note that $\boldsymbol{C}$ is of order $O_p(m)$ component-wise in probability in both the conditional and unconditional regimes. Using the fact that $\boldsymbol{C}^{-1}$ must also be symmetric, we obtain

$$\boldsymbol{C}^{-1} = \left\{ \sum_{i=1}^{m} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i \right\}^{-1} \hat{\boldsymbol{G}} = \hat{\boldsymbol{G}} \left\{ \sum_{i=1}^{m} \boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \right\}^{-1} \tag{S2.6}$$

or equivalently

$$\boldsymbol{C}^{-1} = \boldsymbol{C}^{-1\top} = \left[\sum_{i=1}^{m}\{\boldsymbol{I}_p - (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1}\}\right]^{-1}\hat{\boldsymbol{G}}$$
$$= \left\{m^{-1}\boldsymbol{I}_p + m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1}\boldsymbol{C}^{-1}\right\}\hat{\boldsymbol{G}}, \qquad \text{(S2.7)}$$

where the last line is derived from (a special case of) the Woodbury identity, given by $(\boldsymbol{Q}-\boldsymbol{R})^{-1} = \boldsymbol{Q}^{-1} + \boldsymbol{Q}^{-1}\boldsymbol{R}(\boldsymbol{Q}-\boldsymbol{R})^{-1}$ for arbitrary matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ such that $\boldsymbol{Q}$ and $(\boldsymbol{Q}-\boldsymbol{R})$ are invertible. The first term in (S2.7) is the dominating term, being of order $O(m^{-1})$, while the second term is $O_p(m^{-1}n_L^{-1})$ in both the conditional and unconditional regimes. We will use all the above forms of $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ in subsequent developments. Similarly, we can apply the Woodbury identity to $(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}$ and $(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}$ to obtain $n_L(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1} = n_L\boldsymbol{B}_3^{-1} - n_L\boldsymbol{B}_3^{-1}\boldsymbol{B}_4(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1} = O_p(1) + O_p(n_L^{-1})$ and $n_i(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} = n_i(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} - n_i(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} = O_p(1) + O_p(n_i^{-1})$, where the order results hold component-wise. These hold irrespective of whether we are conditioning on the random effects.

To further simplify expressions, for the rest of this article we will only use order results when representing quantities associated with these smaller order terms. Furthermore, as we want the derivations for the remainder of

this section to be applicable to both the conditional and unconditional regime, we will not distinguish between $O(\cdot)$ and $O_p(\cdot)$ in the following developments, and simply use $O_p()$ to represent both as appropriate. The terms we use "big-O notation" for will have the same order under both the conditional and unconditional regime. To simplify expressions, we will also drop the dependence on $\boldsymbol{\theta}$, unless stated otherwise.

Finally, it is worth emphasising that

$$[-\boldsymbol{I}_p, \boldsymbol{I}_p, \ldots, \boldsymbol{I}_p] \begin{bmatrix} \dot{\phi}^{-1}\boldsymbol{X}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}}) \\ \dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}}) \end{bmatrix} = -\boldsymbol{S}_1 + \sum_{i=1}^{m} \boldsymbol{S}_{2i} = \boldsymbol{S}_1 - \sum_{i=1}^{m} \boldsymbol{S}_{2i} = \boldsymbol{0}_p, \tag{S2.8}$$

due to the $\boldsymbol{X}_i = \boldsymbol{Z}_i$ assumption. This is a key identity that is critical to the proofs throughout this article.

We now use the expressions above to multiply out $-\{\nabla^2 Q(\dot{\boldsymbol{\theta}})\}^{-1}\nabla Q(\dot{\boldsymbol{\theta}})$ and obtain expressions for $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}$. From equation (S2.2), the first p components of $\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}$ are

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}} &= \left[\boldsymbol{C}^{-1} \quad -\boldsymbol{C}^{-1}\boldsymbol{B}_2(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\right]\nabla Q + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]} \\
&= \boldsymbol{C}^{-1}\left[\boldsymbol{I}_p \quad -[\boldsymbol{I}_p - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_1^\top\dot{\boldsymbol{W}}_1\boldsymbol{X}_1 + \hat{\boldsymbol{G}}^{-1})^{-1}, \ldots, \boldsymbol{I}_p - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_m^\top\dot{\boldsymbol{W}}_m\boldsymbol{X}_m + \hat{\boldsymbol{G}}^{-1})^{-1}]\right]\nabla Q \\
&\quad + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]}
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{C}^{-1}\left(\boldsymbol{S}_1 - \sum_{i=1}^{m}\boldsymbol{S}_{2i} - \sum_{i=1}^{m}\boldsymbol{S}_{3i}\right) + \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}\Bigg\{\sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{S}_{2i} \\
&+ \sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{S}_{3i}\Bigg\} + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]} \\
&= \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}\Bigg\{\sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{S}_{2i} - \hat{\boldsymbol{G}}\sum_{i=1}^{m}\boldsymbol{S}_{3i} \\
&+ \sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{S}_{3i}\Bigg\} + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]},
\end{aligned}
$$

where the final equality uses equation (S2.8). Thus, letting $\boldsymbol{V}_1 = \sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{S}_{2i} - \hat{\boldsymbol{G}}\sum_{i=1}^{m}\boldsymbol{S}_{3i} + \sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{S}_{3i}$ and applying equation (S2.7), we obtain

$$
\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}} = m^{-1}\boldsymbol{V}_1 + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]} + O_p(n_L^{-1}) \times m^{-1}\boldsymbol{V}_1.
$$

Finally, using the Woodbury identity for $(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}$, we have that $\sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{S}_{2i} = \sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i)^{-1}\boldsymbol{S}_{2i} + \sum_{i=1}^{m}O_p(n_L^{-2})\boldsymbol{S}_{2i}$. Letting $\boldsymbol{V}_2 = \sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i)^{-1}\boldsymbol{S}_{2i} - \hat{\boldsymbol{G}}\sum_{i=1}^{m}\boldsymbol{S}_{3i} + \sum_{i=1}^{m}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{S}_{3i}$, we obtain

$$
\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}} = m^{-1}\boldsymbol{V}_2 + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]} + O_p(n_L^{-1}) \times m^{-1}\boldsymbol{V}_1 + m^{-1}\sum_{i=1}^{m}O_p(n_L^{-2})\boldsymbol{S}_{2i}.
$$

Next, the last mp components of $\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}$ are

$$
\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} = [-(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top\boldsymbol{C}^{-1} \quad (\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} + (\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top\boldsymbol{C}^{-1}\boldsymbol{B}_2(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}]\nabla Q
$$

$$
\begin{aligned}
&+ \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[p+1:(m+1)p]} \\
&= [-(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top\boldsymbol{C}^{-1} \quad (\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top\boldsymbol{C}^{-1}\boldsymbol{B}_2(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}]\nabla Q \\
&+ [\boldsymbol{0}_{mp\times p} \quad (\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}]\nabla Q + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[p+1:(m+1)p]} \\
&= -(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top[\boldsymbol{C}^{-1} \quad -\boldsymbol{C}^{-1}\boldsymbol{B}_2(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}]\nabla Q \\
&+ [\boldsymbol{0}_{mp\times p} \quad (\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}]\nabla Q + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[p+1:(m+1)p]}.
\end{aligned}
$$

Notice that we already have an expression for $[\boldsymbol{C}^{-1} \quad -\boldsymbol{C}^{-1}\boldsymbol{B}_2(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}]\nabla Q$ from the fixed effects above. Namely, it is $m^{-1}\boldsymbol{V}_1 + O_p(n_L^{-1}) \times m^{-1}\boldsymbol{V}_1$. Thus we have

$$
\begin{aligned}
\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} &= -(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top(m^{-1}\boldsymbol{V}_1 + O_p(n_L^{-1}) \times m^{-1}\boldsymbol{V}_1) \\
&\quad + (\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{S}_6 + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[p+1:(m+1)p]}.
\end{aligned}
$$

Applying the Woodbury identity for $(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}$, we obtain

$$
\begin{aligned}
\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} &= -\boldsymbol{1}_m \otimes (m^{-1}\boldsymbol{V}_1 + O_p(n_L^{-1}) \times m^{-1}\boldsymbol{V}_1) + O_p(n_L^{-1})(m^{-1}\boldsymbol{V}_1 + O_p(n_L^{-1}) \times m^{-1}\boldsymbol{V}_1) \\
&\quad + \boldsymbol{B}_3^{-1}\boldsymbol{S}_6 + O_p(n_L^{-2})\boldsymbol{S}_6 + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[p+1:(m+1)p]} \\
&= -\boldsymbol{1}_m \otimes m^{-1}\boldsymbol{V}_1 + O_p(n_L^{-1}) \times m^{-1}\boldsymbol{V}_1 + O_p(n_L^{-2}) \times m^{-1}\boldsymbol{V}_1 \\
&\quad + \boldsymbol{B}_3^{-1}\boldsymbol{S}_4 + \boldsymbol{B}_3^{-1}\boldsymbol{S}_5 + O_p(n_L^{-2})\boldsymbol{S}_6 + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[p+1:(m+1)p]}.
\end{aligned}
$$

Replacing all the $\boldsymbol{V}_\cdot$ and $\boldsymbol{S}_\cdot$ terms in the above with their definitions, we

finally obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}} &= m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) + m^{-1}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i \\ &\quad - m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]} \\ &\quad + O_p(n_L^{-1})\Bigg\{ m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) + m^{-1}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i \\ &\quad - m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i \Bigg\} + m^{-1}\sum_{i=1}^{m} O_p(n_L^{-2})\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i),\end{aligned} \tag{S2.9}$$

and

$$\begin{aligned}\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} &= -\mathbf{1}_m \otimes \Bigg\{ m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) \\ &\quad + m^{-1}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i - m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i \Bigg\} \\ &\quad + \boldsymbol{B}_3^{-1}\{\dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}})\} - \boldsymbol{B}_3^{-1}\{(\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1})\dot{\boldsymbol{b}}\} + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[p+1:(m+1)p]} \\ &\quad + O_p(n_L^{-1})\Bigg\{ m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) \\ &\quad + m^{-1}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i - m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i \Bigg\} \\ &\quad + O_p(n_L^{-2})\Bigg\{ m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) \\ &\quad + m^{-1}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i - m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i \Bigg\}\end{aligned}$$

$$+ O_p(n_L^{-2})\{\dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}}) - (\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1})\dot{\boldsymbol{b}}\}. \tag{S2.10}$$

The expressions for $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}$ above underlie our proofs. We use these same expressions in both the conditional and unconditional regimes, but the asymptotic behaviours of the terms on the right hand side, and the way we treat them, will differ greatly between the two cases.

As we will show later, the key leading terms for the fixed effects are $m^{-1}\sum_{i=1}^m (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i)$ and $m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i$. The key leading terms for the random effects are $-\mathbf{1}_m \otimes m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i$ and $\boldsymbol{B}_3^{-1}\{\dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}})\}$. When conditioning on the random effects $\dot{\boldsymbol{b}}$, we have $m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i = O(1)$, while in the unconditional regime the same quantity is of order $O_p(m^{-1/2})$ in probability. In both the conditional and unconditional regimes, we have that $m^{-1}\sum_{i=1}^m (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i)$ is of order $O_p(N^{-1/2})$ component-wise, while the quantity $\boldsymbol{B}_3^{-1}\{\dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}})\}$ is of order $O_p(n_L^{-1/2})$ component-wise.

S2.1 Proof of Theorem 1

The dominating terms on the right hand sides of equations (S2.9) and (S2.10) are $m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i$ and $\mathbf{1}_m \otimes m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i$ for the fixed and random effects, respectively. Conditional on the random effects $\dot{\boldsymbol{b}}_i$, these dominat-

ing terms are deterministic and of order $O(1)$. Thus we treat them as bias terms and move them to the left hand side. Next, by Conditions (C1)-(C2), $\boldsymbol{B}_3^{-1}$ is a component-wise $O(n_L^{-1})$ block-diagonal matrix, while we also have $\boldsymbol{B}_2^\top = O(n_U)$, $\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i^\top = O(n_i)$, and $\boldsymbol{C}^{-1} = O(m^{-1})$ component-wise. Since $E\{\boldsymbol{Z}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}})|\dot{\boldsymbol{b}}\} = \boldsymbol{0}_{mp}$ and $\mathrm{Var}\{\boldsymbol{Z}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}})|\dot{\boldsymbol{b}}\} = \boldsymbol{Z}^\top \dot{\boldsymbol{W}} \boldsymbol{Z}$, we obtain $\dot{\phi}^{-1}\boldsymbol{D}_r^{-1}\boldsymbol{Z}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}}) = O_p(1)$ using Chebyshev's inequality and the conditional independence.

Multiplying both sides of (S2.9) and (S2.10) by $N^{1/2}$ and $\boldsymbol{D}_r$ respectively, and applying the order results for the remainder term in Section S3.1, we obtain

$$N^{1/2}\left(\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}} - m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i\right) = m^{-1/2}\sum_{i=1}^m n^{1/2} n_i^{-1/2}(n_i^{-1}\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} n_i^{-1/2}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) + O_p(m^{1/2}n_L^{-1/2}),$$

and

$$\boldsymbol{D}_r\left(\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} + \boldsymbol{1}_m \otimes m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i\right) = \boldsymbol{D}_r\boldsymbol{B}_3^{-1}\boldsymbol{D}_r\boldsymbol{D}_r^{-1}\{\dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}})\} + O_p(n_L^{-1/2}).$$

Recalling that $\boldsymbol{X}_i = \boldsymbol{Z}_i$, to prove Theorem 1 we will show a Lindeberg

condition for

$$\boldsymbol{A}\begin{bmatrix} m^{-1/2}\sum_{i=1}^{m} n^{1/2} n_i^{-1/2}(n_i^{-1}\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} n_i^{-1/2}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) \\ (n_1^{-1}\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1)^{-1}\{n_1^{-1/2}\dot{\phi}^{-1}\boldsymbol{X}_1^\top(\boldsymbol{y}_1 - \dot{\boldsymbol{\mu}}_1)\} \\ \vdots \\ (n_m^{-1}\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m)^{-1}\{n_m^{-1/2}\dot{\phi}^{-1}\boldsymbol{X}_m^\top(\boldsymbol{y}_m - \dot{\boldsymbol{\mu}}_m)\} \end{bmatrix} =: S,$$

and thus apply the Lindeberg-Feller central limit theorem, from which the result follows from Slutsky's theorem.

To prove the condition, first define $\boldsymbol{U} = [\boldsymbol{Z}\boldsymbol{B}_3^{-1}(\boldsymbol{1}_m \otimes \boldsymbol{I}_p), \boldsymbol{Z}\boldsymbol{B}_3^{-1}]$, and $\boldsymbol{U}_k$ as the kth row of $\boldsymbol{U}$, noting it only has $2p$ non-zero components. Then we can write $S = \sum_{k=1}^{N} \boldsymbol{A}\boldsymbol{D}\boldsymbol{U}_k\dot{\phi}^{-1}\{y_k - \mu_k(\dot{\boldsymbol{\theta}})\} \triangleq \sum_{k=1}^{N}\boldsymbol{\xi}_k$, where y_k is the kth component in $(y_{11}, y_{12}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{mn_m})^\top$, and similarly for $\mu_k(\dot{\boldsymbol{\theta}})$.

Conditional on $\dot{\boldsymbol{b}}$, the quantities $\{\boldsymbol{\xi}_k\}_{k=1}^N$ are independent q-vectors with expectation zero and covariance $\mathrm{Var}(\boldsymbol{\xi}_k|\dot{\boldsymbol{b}}) = \boldsymbol{A}\boldsymbol{D}\boldsymbol{U}_k W_k \boldsymbol{U}_k^\top \boldsymbol{D}\boldsymbol{A}^\top$, where W_k is the kth diagonal component in $\dot{\boldsymbol{W}}$. Therefore, we have that

$$\sum_{k=1}^{N}\mathrm{Var}(\boldsymbol{\xi}_k|\dot{\boldsymbol{b}}) = \sum_{k=1}^{N}\boldsymbol{A}\boldsymbol{D}\boldsymbol{U}_k W_k \boldsymbol{U}_k^\top \boldsymbol{D}\boldsymbol{A}^\top$$

$$= \boldsymbol{A} \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m} \frac{n}{n_i}\left(\frac{\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i}{n_i}\right)^{-1} & \frac{1}{\sqrt{m}}\sqrt{\frac{n}{n_1}}\left(\frac{\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1}{n_1}\right)^{-1} & \cdots & \frac{1}{\sqrt{m}}\sqrt{\frac{n}{n_m}}\left(\frac{\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m}{n_m}\right)^{-1} \\ \frac{1}{\sqrt{m}}\sqrt{\frac{n}{n_1}}\left(\frac{\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1}{n_1}\right)^{-1} & \left(\frac{\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1}{n_1}\right)^{-1} & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \frac{1}{\sqrt{m}}\sqrt{\frac{n}{n_m}}\left(\frac{\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m}{n_m}\right)^{-1} & \boldsymbol{0} & \boldsymbol{0} & \left(\frac{\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m}{n_m}\right)^{-1} \end{bmatrix} \boldsymbol{A}^\top.$$

Hence using the finite selection property of $\boldsymbol{A}$, and the fact that $m^{-1/2}n^{1/2}n_i^{-1/2}\left(n_i^{-1}\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i\right)^{-1} = o(1)$ component-wise, we obtain

$$\lim_{m,n_L\to\infty} \sum_{k=1}^{N} \mathrm{Cov}(\boldsymbol{\xi}_k|\dot{\boldsymbol{b}})$$
$$= \lim_{m,n_L\to\infty} \boldsymbol{A}\,\mathrm{bdiag}\left\{\frac{1}{m}\sum_{i=1}^{m}\frac{n}{n_i}\left(\frac{\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i}{n_i}\right)^{-1}, \left(\frac{\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1}{n_1}\right)^{-1}, \ldots, \left(\frac{\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m}{n_m}\right)^{-1}\right\}\boldsymbol{A}^\top$$
$$= \boldsymbol{\Omega}.$$

Next, by the Cauchy-Schwarz inequality, we have

$$E\{\|\boldsymbol{\xi}_k\|^2 I(\|\boldsymbol{\xi}_k\| > \epsilon)|\dot{\boldsymbol{b}}\} \leq E(\|\boldsymbol{\xi}_k\|^4|\dot{\boldsymbol{b}})^{1/2} P(\|\boldsymbol{\xi}_k\| > \epsilon|\dot{\boldsymbol{b}})^{1/2}.$$

Finally, we make a note about the form of $\mathrm{Cov}[\boldsymbol{D}\boldsymbol{U}_k\{y_k - \mu_k(\dot{\boldsymbol{\theta}})\}]$. Without loss of generality, suppose $k = 1$. Then

$$\mathrm{Cov}[\boldsymbol{D}\boldsymbol{U}_1\{y_1 - \mu_1(\dot{\boldsymbol{\theta}})\}] =$$

$$\begin{bmatrix} n(mn_1^2)^{-1}\boldsymbol{H}_1\boldsymbol{x}_{11}W_1\boldsymbol{x}_{11}^\top\boldsymbol{H}_1^\top & n_1^{-1}m^{-1/2}(nn_1^{-1})^{1/2}\boldsymbol{H}_1\boldsymbol{x}_{11}W_1\boldsymbol{x}_{11}^\top\boldsymbol{H}_1^\top & \mathbf{0} \\ n_1^{-1}m^{-1/2}(nn_1^{-1})^{1/2}\boldsymbol{H}_1\boldsymbol{x}_{11}W_1\boldsymbol{x}_{11}^\top\boldsymbol{H}_1^\top & n_1^{-1}\boldsymbol{H}_1\boldsymbol{x}_{11}W_1\boldsymbol{x}_{11}^\top\boldsymbol{H}_1^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{S2.11}$$

Again without loss of generality, consider the case $\boldsymbol{A} = [\boldsymbol{I}_{2p}, \mathbf{0}_{(p+p)\times(m-1)p}]$. Then by equation (S2.11) and Chebyshev's inequality, when $k \in \{1, 2, \ldots, n_1\}$ we have that $P(\|\boldsymbol{\xi}_k\| > \epsilon|\dot{\boldsymbol{b}}) \leq \mathrm{tr}\{\mathrm{Cov}(\boldsymbol{\xi}_k|\dot{\boldsymbol{b}})\}/\epsilon^2 = O(n_1^{-1})$. Thus, given $\dot{\boldsymbol{b}}$, we obtain $\|\boldsymbol{\xi}_k\| = O_p(n_1^{-1/2})$ and $E(\|\boldsymbol{\xi}_k\|^4|\dot{\boldsymbol{b}}) = O(n_1^{-2})$ by Conditions (C1)-(C3) and the properties of the exponential family. However when $k > n_1$, by equation (S2.11) and Chebyshev's inequality, we have that $P(\|\boldsymbol{\xi}_k\| > \epsilon|\dot{\boldsymbol{b}}) \leq \mathrm{tr}\{\mathrm{Cov}(\boldsymbol{\xi}_k|\dot{\boldsymbol{b}})\}/\epsilon^2 = O(N^{-1})$ since $n(mn_1^2)^{-1} = O(N^{-1})$. Thus given $\dot{\boldsymbol{b}}$, it holds that $\|\boldsymbol{\xi}_k\| = O_p(N^{-1/2})$ and $E(\|\boldsymbol{\xi}_k\|^4|\dot{\boldsymbol{b}}) = O(N^{-2})$. Therefore

$$\begin{aligned}
\sum_{k=1}^{N} E\{\|\boldsymbol{\xi}_k\|^2 I(\|\boldsymbol{\xi}_k\| > \epsilon)|\dot{\boldsymbol{b}}\} &\leq \sum_{k=1}^{N} E(\|\boldsymbol{\xi}_k\|^4|\dot{\boldsymbol{b}})^{1/2} P(\|\boldsymbol{\xi}_k\| > \epsilon|\dot{\boldsymbol{b}})^{1/2} \\
&= \sum_{k=1}^{n_1} E(\|\boldsymbol{\xi}_k\|^4|\dot{\boldsymbol{b}})^{1/2} P(\|\boldsymbol{\xi}_k\| > \epsilon|\dot{\boldsymbol{b}})^{1/2} \\
&\quad + \sum_{k=n_1+1}^{N} E(\|\boldsymbol{\xi}_k\|^4|\dot{\boldsymbol{b}})^{1/2} P(\|\boldsymbol{\xi}_k\| > \epsilon|\dot{\boldsymbol{b}})^{1/2} \\
&\leq n_1 \max_{1\leq k\leq n_1} \{E(\|\boldsymbol{\xi}_k\|^4|\dot{\boldsymbol{b}})^{1/2} P(\|\boldsymbol{\xi}_k\| > \epsilon|\dot{\boldsymbol{b}})^{1/2}\}
\end{aligned}$$

$$
\begin{aligned}
&+ (N - n_1) \sup_{k > n_1} \{E(\|\boldsymbol{\xi}_k\|^4 | \dot{\boldsymbol{b}})^{1/2} P(\|\boldsymbol{\xi}_k\| > \epsilon | \dot{\boldsymbol{b}})^{1/2}\} \\
&= n_1 \times O(n_1^{-3/2}) + (N - n_1) \times O(N^{-3/2}) \\
&= O(n_1^{-1/2}) + O(N^{-1/2}) \\
&= o(1).
\end{aligned}
$$

The required result follows by Conditions (C1)-(C2) and the Lindeberg-Feller Central Limit Theorem. Furthermore, the general result holds straightforwardly by replacing n_1 with $O(n_L)$ in the above argument, noting that any row of $\boldsymbol{A}$ can only select a fixed number of clusters.

S2.2 Proof of Equation (4)

For the Poisson pure random intercept model, we have $\boldsymbol{B} = \text{diag}(ne^{\dot{b}_1} + \hat{\sigma}_b^{-2}, \ldots, ne^{\dot{b}_m} + \hat{\sigma}_b^{-2})$ and $\boldsymbol{R}(\tilde{\boldsymbol{\theta}}) = \{ne^{\tilde{b}_1}(\hat{b}_1 - \dot{b}_1)^2, \ldots, ne^{\tilde{b}_m}(\hat{b}_m - \dot{b}_m)^2\}^\top$. Next, suppose that $\boldsymbol{A}$ picks out the first random intercept, i.e., $\boldsymbol{A} = [1, \boldsymbol{0}_{m-1}^\top]$. Then we have

$$
\begin{aligned}
n^{1/2}(\hat{b}_1 - \dot{b}_1) &= n^{1/2} \boldsymbol{A} \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}}) + \frac{1}{2} n^{1/2} \boldsymbol{A} \boldsymbol{B}^{-1} \boldsymbol{R}(\tilde{\boldsymbol{\theta}}) \\
&= n^{-1/2} \left\{ \left(\sum_{j=1}^{n} y_{1j} - e^{\dot{b}_1} \right) - \dot{b}_1 / \hat{\sigma}_b^2 \right\} \Big/ \left\{ e^{\dot{b}_1} + 1/(\hat{\sigma}_b^2 n) \right\} \\
&\quad - \frac{1}{2} \left\{ n^{1/2} e^{\tilde{b}_1} (\hat{b}_1 - \dot{b}_1)^2 \right\} \Big/ \left\{ e^{\dot{b}_1} + 1/(\hat{\sigma}_b^2 n) \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \left[n^{-1/2} \left\{ \left(\sum_{j=1}^{n} y_{1j} - e^{\dot{b}_1} \right) - \dot{b}_1 / \hat{\sigma}_b^2 \right\} \Big/ \left\{ e^{\dot{b}_1} + 1/(\hat{\sigma}_b^2 n) \right\} \right] \Big/ \\
&\left[1 + \left\{ \frac{1}{2} e^{\tilde{b}_1} (\hat{b}_1 - \dot{b}_1) \right\} \Big/ \left\{ e^{\dot{b}_1} + 1/(\hat{\sigma}_b^2 n) \right\} \right] \\
&= n^{-1/2} \sum_{j=1}^{n} (y_{1j} e^{-\dot{b}_1} - 1) + o_p(1),
\end{aligned}
$$

where $\tilde{b}_1$ lies between $\hat{b}_1$ and $\dot{b}_1$, and for the last line we have used the fact that $\hat{b}_1 - \dot{b}_1 = o_p(1)$.

Now, $\{y_{1j} e^{-\dot{b}_1} - 1\}_{j=1}^{m}$ is an exchangeable collection of uncorrelated random variables with mean zero and finite non-zero variance. Furthermore, we have for $k \neq l$

$$
\begin{aligned}
\mathrm{Cov}\{(y_{1k} e^{-\dot{b}_1} - 1)^2, (y_{1l} e^{-\dot{b}_1} - 1)^2\} &= E[\mathrm{Cov}\{(y_{1k} e^{-\dot{b}_1} - 1)^2, (y_{1l} e^{-\dot{b}_1} - 1)^2 | \dot{b}_1\}] \\
&\quad + \mathrm{Cov}[E\{(y_{1k} e^{-\dot{b}_1} - 1)^2 | \dot{b}_1\}, E\{(y_{1l} e^{-\dot{b}_1} - 1)^2 | \dot{b}_1\}] \\
&= 0 + \mathrm{Cov}(e^{-\dot{b}_1}, e^{-\dot{b}_1}) \\
&= e^{\dot{\sigma}_b^2} (e^{\dot{\sigma}_b^2} - 1) \neq 0.
\end{aligned}
$$

Thus by the Central Limit Theorem for exchangeable random variables (Blum et al., 1958), it holds that $n^{-1/2} \sum_{j=1}^{n} (y_{1j} e^{-\dot{b}_1} - 1) \overset{D}{\nrightarrow} N(0, e^{\dot{\sigma}_b^2})$. Since we know $\mathrm{Var}\{n^{-1/2} \sum_{j=1}^{n} (y_{1j} e^{-\dot{b}_1} - 1)\} = e^{\dot{\sigma}_b^2/2}$ and also that $n^{-1/2} \sum_{j=1}^{n} (y_{1j} e^{-\dot{b}_1} - 1) = O_p(1)$ by Chebyshev's inequality, there is no other normalization possible for an asymptotic normality result to hold.

Finally, we also have

$$\begin{aligned} n^{1/2}(\hat{b}_1 - \dot{b}_1) &= n^{-1/2}\sum_{j=1}^{n}(y_{1j}e^{-\dot{b}_1} - 1) + O_p(n^{-1/2}) \\ \implies \hat{b}_1 &= \dot{b}_1 + n^{-1}\sum_{j=1}^{n}(y_{1j}e^{-\dot{b}_1} - 1) + O_p(n^{-1}) \\ &= \dot{b}_1 + o_p(1), \quad \text{by the Weak Law of Large Numbers.} \end{aligned}$$

S2.3 Proof of Theorem 2

We begin by developing two key equations, (S2.12) and (S2.13), that will be used throughout the unconditional regime. These are derived from equations (S2.9) and (S2.10) and are used in the proofs of Theorems 2-5 as well as Corollary 1. Under Conditions (C1)-(C2), the following order results are used: $\boldsymbol{B}_3^{-1}$ is a component-wise $O_p(n_L^{-1})$ block-diagonal matrix, $\boldsymbol{B}_2 = O_p(n_U)$ component-wise, $\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i^\top = O_p(n_i)$ component-wise, and $\boldsymbol{C}^{-1} = O_p(m^{-1})$ component-wise. Also, by the conditional independence, we have

$$E\{\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}})\} = E[E\{\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}})|\dot{\boldsymbol{b}}\}] = \boldsymbol{0}_{mp},$$

$$\text{Var}\{\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}})\} = E[\text{Var}\{\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}})|\dot{\boldsymbol{b}}\}] + \text{Var}[E\{\boldsymbol{Z}^\top(\boldsymbol{y} - \dot{\boldsymbol{\mu}})|\dot{\boldsymbol{b}}\}] = E(\boldsymbol{Z}^\top \dot{\boldsymbol{W}} \boldsymbol{Z}),$$

so that $\dot{\phi}^{-1}\boldsymbol{D}_r^{-1}\boldsymbol{Z}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}}) = O_p(1)$ using Chebyshev's inequality. Therefore we have the key equations

$$\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}} = m^{-1}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i + O_p(N^{-1/2}) + O_p(n_L^{-1}) + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]} \quad \text{(S2.12)}$$

and

$$\begin{aligned}\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} = -\mathbf{1}_m \otimes m^{-1}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i + \boldsymbol{B}_3^{-1}\{\dot{\phi}^{-1}\boldsymbol{Z}^\top(\boldsymbol{y}-\dot{\boldsymbol{\mu}})\} \\ + O_p(N^{-1/2}) + O_p(n_L^{-1}) + \frac{1}{2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[p+1:(m+1)p]}. \quad \text{(S2.13)}\end{aligned}$$

By equation (S2.12), we have

$$m^{1/2}(\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}) = m^{-1/2}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i + O_p(n_L^{-1/2}) + O_p(m^{1/2}n_L^{-1}) + \frac{1}{2}m^{1/2}\{\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]}.$$

Next, we consider two separate scenarios. First, suppose that $mn_U^{-1} \to \infty$. Then by the order results for the remainder term in Section S3.2, the first p components of $\boldsymbol{D}^*\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})$ are of order $O_p(m^{1/2}n_L^{-1})$, and so the first p components of $\boldsymbol{D}^*(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})$ can be shown to be

$$m^{1/2}(\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}) = m^{-1/2}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i + o_p(1).$$

The required result then follows from the independence of the random effects and the normal assumption on the $\dot{\boldsymbol{b}}_i$; note the $mn_L^{-2} \to 0$ assumption

is required for the remainder term to be smaller order than the linear term.

On the other hand, when $mn_L^{-1} \to 0$, the only difference from the $mn_L^{-1} \to \infty$ case is that the first p components of $\boldsymbol{D}^+\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})$ are now of order $O_p(m^{-1/2})$ due to the different convergence rate of the prediction gap. The result however follows along similar lines as above.

S2.4 Proof of Theorem 3

Again we consider two different scenarios. First, suppose $mn_L^{-1} \to \infty$. Then from equation (S2.13) and the order results for the remainder term in Section S3.2, we have that

$$\boldsymbol{D}_r(\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}) = O_p(n_U^{1/2}m^{-1/2}) + O_p(1) + O_p(n_L^{-1/2}).$$

Based on the above, we obtain $\boldsymbol{D}_r\hat{\boldsymbol{b}} = \boldsymbol{D}_r\dot{\boldsymbol{b}} + O_p(1)$, and thus $\hat{\boldsymbol{b}} = \dot{\boldsymbol{b}} + O_p(n_L^{-1/2})$. The required result follows by multiplying both sides by $\boldsymbol{A}_r$.

On the other hand, suppose now $mn_L^{-1} \to 0$. Then a normalization by $m^{1/2}$ is needed instead, and the third derivative term is consequently of order $O_p(m^{-1/2})$ in probability. We thus obtain

$$m^{1/2}(\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}) = O_p(1) + O_p(m^{1/2}n_L^{-1/2}) + O_p(m^{-1/2}),$$

and the result follows.

As an side remark, note from the above proof when $mn_L^{-1} \to 0$, it holds that $\|\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}\|_2 = O_p(1)$, where $\|\cdot\|_2$ denotes the l_2-norm. But if $mn_U^{-1} \to \infty$ then we instead obtain $\|\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}\|_2 = O_p(m^{1/2} n_U^{-1/2})$. This implies that, under the unconditional regime, a consistency result based on the l_2-norm cannot hold for the entire vector of random effects when there is a partnered fixed effect. If there is no partnered fixed effect though, consistency of the entire vector is sometimes possible. For example, in the Poisson counterexample, we demonstrate that $\|\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}\|_2 = O_p(m^{1/2} n^{-1/2}) = o_p(1)$ when $mn^{-1} \to 0$.

S2.5 Proof of Theorem 4 and Corollary 1

We will prove each of the three parts of the theorem separately. The proof of part (a) also proves Corollary 1.

Part (a): When $mn_U^{-1} \to \infty$, we have from equation (S2.13) and the order results for the remainder term in Section S3.2 that

$$\boldsymbol{D}_r(\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}) = \boldsymbol{D}_r \boldsymbol{B}_3^{-1} \boldsymbol{D}_r \boldsymbol{D}_r^{-1} \dot{\phi}^{-1} \boldsymbol{Z}^\top (\boldsymbol{y} - \dot{\boldsymbol{\mu}}) + o_p(1).$$

This is identical to the proof of Theorem 3. Next, without loss of generality, suppose $\boldsymbol{A}_r$ selects the first cluster only. Then we have

$$n_1^{1/2}(\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1) = (n_1^{-1} \boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1)^{-1} n_1^{-1/2} \{\dot{\phi}^{-1} \boldsymbol{X}_1^\top (\boldsymbol{y}_1 - \dot{\boldsymbol{\mu}}_1)\} + o_p(1)$$

$$\stackrel{\Delta}{=} \boldsymbol{P}_{n_1} + o_p(1).$$

We wish to study the distribution of $\boldsymbol{P}_{n_1}$ as $m, n_L \to \infty$. By definition,

$$\lim_{m,n_L\to\infty} F_{\boldsymbol{P}_{n_1}}(\boldsymbol{x}) = \lim_{m,n_L\to\infty} \int F_{\boldsymbol{P}_{n_1}|\dot{\boldsymbol{b}}_1}(\boldsymbol{x}) f(\dot{\boldsymbol{b}}_1) d\dot{\boldsymbol{b}}_1.$$

Since $F_{\boldsymbol{P}_{n_1}|\dot{\boldsymbol{b}}_1}(\boldsymbol{x})$ is a cdf, then $F_{\boldsymbol{P}_{n_1}|\dot{\boldsymbol{b}}_1}(\boldsymbol{x}) f(\dot{\boldsymbol{b}}_1)$ is bounded by $f(\dot{\boldsymbol{b}}_1)$. Hence applying $\int f(\dot{\boldsymbol{b}}_1) d\dot{\boldsymbol{b}}_1 = 1$ and the dominated convergence theorem, we obtain

$$\lim_{m,n_L\to\infty} F_{\boldsymbol{P}_{n_1}}(\boldsymbol{x}) = \int \lim_{m,n_L\to\infty} F_{\boldsymbol{P}_{n_1}|\dot{\boldsymbol{b}}_1}(\boldsymbol{x}) f(\dot{\boldsymbol{b}}_1) d\dot{\boldsymbol{b}}_1 = \int \Psi_{\boldsymbol{P}_{n_1}|\dot{\boldsymbol{b}}_1}(\boldsymbol{x}) f(\dot{\boldsymbol{b}}_1) d\dot{\boldsymbol{b}}_1,$$

where $\Psi_{\boldsymbol{P}_{n_1}|\dot{\boldsymbol{b}}_1}(\cdot)$ is the cdf associated with $N(\boldsymbol{0}, \boldsymbol{K}_1)$, a result which follows from conditional independence and the Lindeberg-Feller Central Limit Theorem used in Theorem 1. The general result follows by noting that the same argument can be applied to any finite subset of the random effects. Note also that the result holds regardless of the true distribution of $\dot{\boldsymbol{b}}_i$.

Part (b): When $mn_i^{-1} \to \gamma_i \in (0, \infty)$, we have from (S2.13) and the order results for the remainder term in Section S3.2 that

$$n_i^{1/2}(\hat{\dot{\boldsymbol{b}}}_i - \dot{\boldsymbol{b}}_i) = (n_i^{-1}\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} n_i^{-1/2} \{\dot{\phi}^{-1} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i)\} - (\gamma_i m)^{-1/2} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i + O_p(n_L^{-1/2}),$$

from the same development as in the proof of Part (a). Letting $\boldsymbol{E}_1 = (n_i^{-1}\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} n_i^{-1/2} \dot{\phi}^{-1} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i)$ and $\boldsymbol{E}_2 = m^{-1/2} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i$, then

since $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are independent given $\dot{\boldsymbol{b}}_i$, we obtain for any i,

$$\begin{aligned}
\lim_{m,n_L\to\infty} F_{\boldsymbol{E}_1,\boldsymbol{E}_2}(\boldsymbol{x},\boldsymbol{y}) &= \lim_{m,n_L\to\infty} \int F_{\boldsymbol{E}_1,\boldsymbol{E}_2|\dot{\boldsymbol{b}}_i}(\boldsymbol{x},\boldsymbol{y}) f(\dot{\boldsymbol{b}}_i) d\dot{\boldsymbol{b}}_i \\
&= \lim_{m,n_L\to\infty} \int F_{\boldsymbol{E}_1|\dot{\boldsymbol{b}}_i}(\boldsymbol{x}) F_{\boldsymbol{E}_2|\dot{\boldsymbol{b}}_i}(\boldsymbol{y}) f(\dot{\boldsymbol{b}}_i) d\dot{\boldsymbol{b}}_i \\
&= \int \lim_{m,n_L\to\infty} F_{\boldsymbol{E}_1|\dot{\boldsymbol{b}}_i}(\boldsymbol{x}) F_{\boldsymbol{E}_2|\dot{\boldsymbol{b}}_i}(\boldsymbol{y}) f(\dot{\boldsymbol{b}}_i) d\dot{\boldsymbol{b}}_i \\
&= \int \lim_{n_L\to\infty} F_{\boldsymbol{E}_1|\dot{\boldsymbol{b}}_i}(\boldsymbol{x}) \lim_{m\to\infty} F_{\boldsymbol{E}_2|\dot{\boldsymbol{b}}_i}(\boldsymbol{y}) f(\dot{\boldsymbol{b}}_i) d\dot{\boldsymbol{b}}_i \\
&= \Psi_{\boldsymbol{E}_2}(\boldsymbol{y}) \int \lim_{n_L\to\infty} F_{\boldsymbol{E}_1|\dot{\boldsymbol{b}}_i}(\boldsymbol{x}) f(\dot{\boldsymbol{b}}_i) d\dot{\boldsymbol{b}}_i,
\end{aligned}$$

where $\Psi_{\boldsymbol{E}_2}(\cdot)$ is the cdf of $N(\boldsymbol{0}, \dot{\boldsymbol{G}})$. The third line follows from the Dominated Convergence Theorem since $F_{\boldsymbol{E}_1|\dot{\boldsymbol{b}}_i}(\boldsymbol{x})$ and $F_{\boldsymbol{E}_2|\dot{\boldsymbol{b}}_i}(\boldsymbol{y})$ are cdfs and $\int f(\dot{\boldsymbol{b}}_i) d\dot{\boldsymbol{b}}_i = 1$. Thus $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are asymptotically independent. The result follows from this asymptotic independence.

Part (c): When $mn_L^{-1} \to 0$, we have from (S2.13) and the order results for the remainder term in Section S3.2 that

$$m^{1/2}(\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}) = -\boldsymbol{1}_m \otimes \boldsymbol{I}_p m^{-1/2} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i + o_p(1).$$

The result then follows immediately from the normality assumption on $\dot{\boldsymbol{b}}_i$.

S2.6 Proof of Theorem 5

Given $mn_L^{-2} \to 0$ and $mn_U^{-1/2} \to \infty$, by summing equations (S2.12) and (S2.13) we see that the $m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i$ terms cancel. Therefore, we are left with

$$\begin{aligned} n_i^{1/2}(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b}}_i - \dot{\boldsymbol{\beta}} - \dot{\boldsymbol{b}}_i) &= n_i(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} n_i^{-1/2}\{\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i)\} \\ &\quad + O_p(m^{-1/2}) + O_p(n_L^{-1/2}) + O_p(m^{-1}n_U^{1/2}) \\ &= n_i(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} n_i^{-1/2}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) + o_p(1). \end{aligned}$$

The required result follows from the Dominated Convergence Theorem.

S2.7 Result for Difference Between the Prediction Gaps of Two Clusters

Assume Conditions (C1)-(C5) are satisfied, $mn_L^{-2} \to 0$, $mn_U^{-1/2} \to \infty$, *and* $n_i n_{i'}^{-1} \to \gamma \in (0, \infty)$. *Then as* $m, n_L \to \infty$ *and unconditional on the random effects* $\dot{\boldsymbol{b}}$, *for each* $i \neq i' \in \{1, \ldots, m\}$ *we have*

$$n_i^{1/2}\{(\hat{\boldsymbol{b}}_i - \dot{\boldsymbol{b}}_i) - (\hat{\boldsymbol{b}}_{i'} - \dot{\boldsymbol{b}}_{i'})\} \xrightarrow{D} \text{mixN}(\boldsymbol{0}, \dot{\boldsymbol{K}}_i, F_{\dot{\boldsymbol{b}}_i}) * \text{mixN}(\boldsymbol{0}, \gamma\dot{\boldsymbol{K}}_{i'}, F_{\dot{\boldsymbol{b}}_{i'}}).$$

Proof: Theorem 4 implies that, given $mn_L^{-2} \to 0$, $mn_U^{-1/2} \to \infty$, and

$n_i n_{i'}^{-1} \to \gamma \in (0, \infty)$, we have

$$\begin{aligned} n_i^{1/2}(\hat{\boldsymbol{b}}_i - \dot{\boldsymbol{b}}_i - \hat{\boldsymbol{b}}_{i'} + \dot{\boldsymbol{b}}_{i'}) &= (n_i^{-1}\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} n_i^{-1/2} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) \\ &\quad + \gamma^{1/2}(n_{i'}^{-1}\boldsymbol{X}_{i'}^\top \dot{\boldsymbol{W}}_{i'} \boldsymbol{X}_{i'})^{-1} n_{i'}^{-1/2} \boldsymbol{X}_{i'}^\top (\boldsymbol{y}_{i'} - \dot{\boldsymbol{\mu}}_{i'}) \\ &\quad + O_p(m^{-1/2}) + O_p(n_L^{-1/2}) + O_p(m^{-1} n_U^{1/2}), \end{aligned}$$

and the result follows by the independence of $\dot{\boldsymbol{b}}_i$ and $\dot{\boldsymbol{b}}_{i'}$.

S3 Remainder Term in the Taylor Expansion

In this section, we show that in the Taylor expansion (S2.2), the remainder term $-\frac{1}{2}\{\nabla^2 Q(\dot{\boldsymbol{\theta}})\}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})$ is of smaller order component-wise than $-\{\nabla^2 Q(\dot{\boldsymbol{\theta}})\}^{-1}\nabla Q(\dot{\boldsymbol{\theta}})$. To deal with this remainder term, we have the following from equation (S2.2)

$$\begin{aligned} &\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}} = \boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}}) + \frac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}}) \\ \Rightarrow &\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}} - \frac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}}) \\ \Rightarrow &(\boldsymbol{I}_{(m+1)p} - \boldsymbol{\Lambda})(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) = \boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}}) \\ \Rightarrow &\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}} = (\boldsymbol{I}_{(m+1)p} - \boldsymbol{\Lambda})^{-1}\boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}}) \\ &\quad = \boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}}) + \left(\sum_{s=1}^{\infty}\boldsymbol{\Lambda}^s\right)\boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}}), \end{aligned}$$

where the last line is derived from repeated application of the Woodbury identity, and $\boldsymbol{\Lambda}$ is the appropriate $(m+1)p \times (m+1)p$ matrix defined in detail later on. The convergence of the geometric sum and thus invertibility of $(\boldsymbol{I}_{(m+1)p} - \boldsymbol{\Lambda})$ is shown in Lemma 4. We will show, using the consistency result $\|\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}\|_\infty = o_p(1)$, that $\sum_{s=1}^{\infty} \boldsymbol{\Lambda}^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ is of smaller order component-wise than $\boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}})$. This is equivalent to $0.5\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})$ being smaller order component-wise than $\boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}})$ in (S2.2).

Let $\boldsymbol{T}_1$ denote the first p components of $\boldsymbol{R}(\tilde{\boldsymbol{\theta}})$, $\boldsymbol{T}_2$ its remaining mp components, and $\boldsymbol{T}_{2i}$ denote the $\{(i-1)p+1\}$-th to (ip)-th components of $\boldsymbol{T}_2$. We first prove a result needed for later developments.

Lemma 3. *Assume Conditions (C1) and (C3) are satisfied. Then irrespective of whether $\dot{\boldsymbol{b}}$ is conditioned on, it holds that $\boldsymbol{R}(\tilde{\boldsymbol{\theta}})_{[1:p]} = \sum_{i=1}^{m} \boldsymbol{R}(\tilde{\boldsymbol{\theta}})_{[ip+1:(i+1)p]}$.*

Proof. Recall the Taylor expansion $\nabla Q(\hat{\boldsymbol{\theta}}) = \boldsymbol{0} = \nabla Q(\dot{\boldsymbol{\theta}}) + \nabla^2 Q(\dot{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) + \boldsymbol{R}(\tilde{\boldsymbol{\theta}})$. Then

$$
\begin{aligned}
\boldsymbol{0}_{p\times 1} &= \nabla Q(\hat{\boldsymbol{\theta}})_{[1:p]} \\
&= \{\nabla Q(\dot{\boldsymbol{\theta}}) + \nabla^2 Q(\dot{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) + \boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]} \\
&= \sum_{i=1}^{m} \nabla Q(\hat{\boldsymbol{\theta}})_{[ip+1:(i+1)p]} \\
&= \sum_{i=1}^{m} \{\nabla Q(\dot{\boldsymbol{\theta}}) + \nabla^2 Q(\dot{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) + \boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[ip+1:(i+1)p]}.
\end{aligned}
$$

Since $\boldsymbol{Z}_i = \boldsymbol{X}_i$ for all $i = 1, \ldots, m$ under our simplifying assumption, and $\sum_{i=1}^m \hat{\boldsymbol{b}}_i = \boldsymbol{0}$, then we obtain

$$\{\nabla Q(\dot{\boldsymbol{\theta}}) + \nabla^2 Q(\dot{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) + \boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]} = \sum_{i=1}^m \{\nabla Q(\dot{\boldsymbol{\theta}}) + \nabla^2 Q(\dot{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) + \boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[ip+1:(i+1)p]}.$$

Therefore, we have $\boldsymbol{T}_1 = \boldsymbol{R}(\tilde{\boldsymbol{\theta}})_{[1:p]} = \sum_{i=1}^m \boldsymbol{R}(\tilde{\boldsymbol{\theta}})_{[ip+1:(i+1)p]} = \sum_{i=1}^m \boldsymbol{T}_{2i}$, which follows from the fact that $\sum_{i=1}^m \nabla Q(\dot{\boldsymbol{\theta}})_{[ip+1:(i+1)p]} = \nabla Q(\dot{\boldsymbol{\theta}})_{[1:p]} - \sum_{i=1}^m \hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i$ and $\sum_{i=1}^m \{\nabla^2 Q(\dot{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})\}_{[ip+1:(i+1)p]} = \{\nabla^2 Q(\dot{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})\}_{[1:p]} + \sum_{i=1}^m \hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i - \sum_{i=1}^m \hat{\boldsymbol{G}}^{-1}\hat{\boldsymbol{b}}_i$. □

Next, let $\boldsymbol{S}(\boldsymbol{\theta}) = \nabla Q(\boldsymbol{\theta})$, $\tilde{\boldsymbol{W}}' = \dot{\phi}^{-1}\mathrm{diag}\{a'''(\tilde{\eta}_{11}), \ldots, a'''(\tilde{\eta}_{1n_1}), \ldots, a'''(\tilde{\eta}_{mn_m})\}$. Then the remainder term can be written as

$$\boldsymbol{R}(\tilde{\boldsymbol{\theta}}) = \begin{bmatrix} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})^\top \frac{\partial^2 \boldsymbol{S}_{[1]}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \\ \vdots \\ (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})^\top \frac{\partial^2 \boldsymbol{S}_{[(m+1)p]}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \end{bmatrix}.$$

Now, for $1 \le j \le p$, we have $\boldsymbol{S}_{[j]}(\boldsymbol{\theta}) = \dot{\phi}^{-1}\boldsymbol{X}_{[,j]}^\top\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\} = \dot{\phi}^{-1}\sum_{i=1}^m \sum_{l=1}^{n_i} x_{il[j]}\{y_{il} - a'(\eta_{il})\}$, noting this is a scalar. Thus

$$\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{S}_{[j]}(\boldsymbol{\theta}) = -\dot{\phi}^{-1} \sum_{i=1}^m \sum_{l=1}^{n_i} \begin{bmatrix} \boldsymbol{x}_{il} \\ \frac{\partial}{\partial \boldsymbol{b}} \eta_{il} \end{bmatrix} a''(\eta_{il}) x_{il[j]} = - \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}_{[,j]} \\ \boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{X}_{[,j]} \end{bmatrix},$$

which is an $(m+1)p$-vector. Hence the $(m+1)p \times (m+1)p$ matrix can be

written as

$$\frac{\partial^2 \boldsymbol{S}_{[j]}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = -\dot{\phi}^{-1} \sum_{i=1}^{m} \sum_{l=1}^{n_i} \begin{bmatrix} \boldsymbol{x}_{il} \\ \frac{\partial}{\partial \boldsymbol{b}} \eta_{il} \end{bmatrix} a'''(\tilde{\eta}_{il}) x_{il[j]} \begin{bmatrix} \boldsymbol{x}_{il} \\ \frac{\partial}{\partial \boldsymbol{b}} \eta_{il} \end{bmatrix}^\top$$
$$= - \begin{bmatrix} \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{X}_{[,j]}) \tilde{\boldsymbol{W}}' \boldsymbol{X} & \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{X}_{[,j]}) \tilde{\boldsymbol{W}}' \boldsymbol{Z} \\ \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{X}_{[,j]}) \tilde{\boldsymbol{W}}' \boldsymbol{X} & \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{X}_{[,j]}) \tilde{\boldsymbol{W}}' \boldsymbol{Z} \end{bmatrix}, \quad 1 \leq j \leq p.$$

Similarly, for $1 \leq k \leq mp$, $\boldsymbol{S}_{[p+k]}(\boldsymbol{\theta}) = \dot{\phi}^{-1} \boldsymbol{Z}_{[,k]}^\top \{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\} - \{(\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}) \boldsymbol{b}\}_{[k]}$, such that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{S}_{[p+k]}(\boldsymbol{\theta}) = - \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{Z}_{[,k]} \\ \boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z}_{[,k]} + \frac{\partial}{\partial \boldsymbol{b}} \{(\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}) \boldsymbol{b}\}_{[k]} \end{bmatrix},$$

where $\partial / \partial \boldsymbol{b} \{(\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}) \boldsymbol{b}\}_{[k]}$ is not a function of $\boldsymbol{\theta}$. Thus

$$\frac{\partial^2 \boldsymbol{S}_{[p+k]}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = - \begin{bmatrix} \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,k]}) \tilde{\boldsymbol{W}}' \boldsymbol{X} & \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,k]}) \tilde{\boldsymbol{W}}' \boldsymbol{Z} \\ \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,k]}) \tilde{\boldsymbol{W}}' \boldsymbol{X} & \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,k]}) \tilde{\boldsymbol{W}}' \boldsymbol{Z} \end{bmatrix}, \quad 1 \leq k \leq mp.$$

Next, recall that $\boldsymbol{B}_2(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} = [\boldsymbol{I}_p - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1 + \hat{\boldsymbol{G}}^{-1})^{-1}, \ldots, \boldsymbol{I}_p - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m + \hat{\boldsymbol{G}}^{-1})^{-1}]$. By Lemma 1 and the blockwise inversion formula for $\boldsymbol{B}^{-1}$, the first p components of $\boldsymbol{B}^{-1} \boldsymbol{R}(\tilde{\boldsymbol{\theta}})$ are given by

$$\begin{bmatrix} \boldsymbol{C}^{-1} & -\boldsymbol{C}^{-1} \boldsymbol{B}_2 (\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} \end{bmatrix} \boldsymbol{R}(\tilde{\boldsymbol{\theta}})$$
$$= \boldsymbol{C}^{-1} \begin{bmatrix} \boldsymbol{I}_p & -[\boldsymbol{I}_p - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1 + \hat{\boldsymbol{G}}^{-1})^{-1}, \ldots, \boldsymbol{I}_p - \hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m + \hat{\boldsymbol{G}}^{-1})^{-1}] \end{bmatrix} \boldsymbol{R}(\tilde{\boldsymbol{\theta}})$$

$$= \boldsymbol{C}^{-1}\left\{\boldsymbol{T}_1 - \sum_{i=1}^{m}\boldsymbol{T}_{2i} + \sum_{i=1}^{m}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{T}_{2i}\right\}. \tag{S3.1}$$

Similarly, the last mp components of $\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})$ are

$$\begin{aligned}&\left[-(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top\boldsymbol{C}^{-1} \quad (\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1} + (\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top\boldsymbol{C}^{-1}\boldsymbol{B}_2(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\right]\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\\ &= -(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top\left[\boldsymbol{C}^{-1} \quad -\boldsymbol{C}^{-1}\boldsymbol{B}_2(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\right]\boldsymbol{R}(\tilde{\boldsymbol{\theta}}) + (\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{T}_2.\end{aligned} \tag{S3.2}$$

Hence the first p components of $\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})$ are given by

$$\boldsymbol{F}_1 = \boldsymbol{C}^{-1}\sum_{i=1}^{m}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{T}_{2i},$$

and the last mp components of $\boldsymbol{B}^{-1}\boldsymbol{R}(\tilde{\boldsymbol{\theta}})$ are given by

$$\boldsymbol{F}_2 = -(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top\boldsymbol{F}_1 + (\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{T}_2.$$

Next, we have

$$\boldsymbol{T}_2 = \begin{bmatrix}(\hat{\boldsymbol{\theta}}-\dot{\boldsymbol{\theta}})^\top\frac{\partial^2\boldsymbol{S}_{[p+1]}(\tilde{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}(\hat{\boldsymbol{\theta}}-\dot{\boldsymbol{\theta}})\\ \vdots\\ (\hat{\boldsymbol{\theta}}-\dot{\boldsymbol{\theta}})^\top\frac{\partial^2\boldsymbol{S}_{[(m+1)p]}(\tilde{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}(\hat{\boldsymbol{\theta}}-\dot{\boldsymbol{\theta}})\end{bmatrix} = \begin{bmatrix}(\hat{\boldsymbol{\theta}}-\dot{\boldsymbol{\theta}})^\top\frac{\partial^2\boldsymbol{S}_{[p+1]}(\tilde{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\\ \vdots\\ (\hat{\boldsymbol{\theta}}-\dot{\boldsymbol{\theta}})^\top\frac{\partial^2\boldsymbol{S}_{[(m+1)p]}(\tilde{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\end{bmatrix}(\hat{\boldsymbol{\theta}}-\dot{\boldsymbol{\theta}}) \triangleq \boldsymbol{F}_3(\hat{\boldsymbol{\theta}}-\dot{\boldsymbol{\theta}})$$

and

$$\boldsymbol{T}_{2i} = \begin{bmatrix} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})^\top \frac{\partial^2 \boldsymbol{S}_{[(i-1)p+1]}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \\ \vdots \\ (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})^\top \frac{\partial^2 \boldsymbol{S}_{[ip]}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \end{bmatrix} = \begin{bmatrix} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})^\top \frac{\partial^2 \boldsymbol{S}_{[(i-1)p+1]}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \\ \vdots \\ (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})^\top \frac{\partial^2 \boldsymbol{S}_{[ip]}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \end{bmatrix} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \triangleq \boldsymbol{F}_{3i} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}).$$

Here, $\boldsymbol{F}_3$ is a $mp \times (m+1)p$ matrix and $\boldsymbol{F}_{3i}$ is $p \times (m+1)p$. Notice that $\boldsymbol{F}_3 = [\boldsymbol{F}_{31}^\top, \ldots, \boldsymbol{F}_{3n}^\top]^\top$. Furthermore,

$$\begin{aligned} \boldsymbol{B}^{-1} \boldsymbol{R}(\tilde{\boldsymbol{\theta}}) &= \begin{bmatrix} \boldsymbol{F}_1 \\ \boldsymbol{F}_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m \boldsymbol{C}^{-1} \hat{\boldsymbol{G}}^{-1} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{T}_{2i} \\ -(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} \boldsymbol{B}_2^\top \boldsymbol{F}_1 + (\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} \boldsymbol{T}_2 \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^m \boldsymbol{C}^{-1} \hat{\boldsymbol{G}}^{-1} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{F}_{3i} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \\ -(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} \boldsymbol{B}_2^\top \sum_{i=1}^m \boldsymbol{C}^{-1} \hat{\boldsymbol{G}}^{-1} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{F}_{3i} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) + (\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} \boldsymbol{F}_3 (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^m \boldsymbol{C}^{-1} \hat{\boldsymbol{G}}^{-1} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{F}_{3i} \\ -(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} \boldsymbol{B}_2^\top \sum_{i=1}^m \boldsymbol{C}^{-1} \hat{\boldsymbol{G}}^{-1} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{F}_{3i} + (\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1} \boldsymbol{F}_3 \end{bmatrix} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \\ &= 2\boldsymbol{\Lambda} (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}). \end{aligned}$$

The kth row of $\boldsymbol{F}_{3i}$ for $1 \leq k \leq p$ is given by

$$\begin{aligned} &- (\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}})^\top \begin{bmatrix} \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]}) \tilde{\boldsymbol{W}}' \boldsymbol{X} & \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]}) \tilde{\boldsymbol{W}}' \boldsymbol{Z} \\ \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]}) \tilde{\boldsymbol{W}}' \boldsymbol{X} & \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]}) \tilde{\boldsymbol{W}}' \boldsymbol{Z} \end{bmatrix} \\ &= -\delta_{m,n_L}^{-1} \big[\delta_{m,n_L} (\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]}) \tilde{\boldsymbol{W}}' \boldsymbol{X} + \delta_{m,n_L} (\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}})^\top \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]}) \tilde{\boldsymbol{W}}' \boldsymbol{X}, \end{aligned}$$

$$\delta_{m,n_L}(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]})\tilde{\boldsymbol{W}'}\boldsymbol{Z} + \delta_{m,n_L}(\hat{\boldsymbol{b}}-\dot{\boldsymbol{b}})^\top \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]})\tilde{\boldsymbol{W}'}\boldsymbol{Z}\big], \tag{S3.3}$$

where δ_{m,n_L} is a positive unbounded monotonically increasing sequence such that $\delta_{m,n_L}\|\hat{\boldsymbol{\theta}}-\dot{\boldsymbol{\theta}}\|_\infty = O_p(1)$. The consistency results proved in Section S1 ensure that such a δ_{m,n_L} must exist; this is true for both the conditional and unconditional regimes.

Observe that only the $\{(\sum_{l=0}^{i-1} n_l)+1\}$th to $(\sum_{l=0}^{i} n_l)$th components of $\boldsymbol{Z}_{[,(i-1)p+k]}$ are non-zero, where we define $n_0 := 0$. This means that, for any $1 \le k \le p$, only the $\{(i-1)p+1\}$th to (ip)th columns of both $\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]})\tilde{\boldsymbol{W}'}\boldsymbol{Z}$ and $\boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,(i-1)p+k]})\tilde{\boldsymbol{W}'}\boldsymbol{Z}$ will be non-zero. In other words, other than the first p columns, only the $(ip+1)$th to $\{(i+1)p\}$th columns of $\boldsymbol{F}_{3i}$ are non-zero. Thus $\boldsymbol{F}_3$, disregarding its first p columns, is an $mp \times mp$ block-diagonal matrix.

The non-zero components of $\delta_{m,n_L}\boldsymbol{F}_3$ and $\delta_{m,n_L}\boldsymbol{F}_{3i}$ are all $O_p(n_U)$ component-wise, again because at most n_i components of $\boldsymbol{Z}_{[,(i-1)p+k]}$ are non-zero. For ease of notation and understanding, we now represent all terms using their orders only. Since $\boldsymbol{C}^{-1} = O_p(m^{-1})$ and $\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} = O_p(n_i^{-1})$, from the above discussion we have that $\sum_{i=1}^m \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{F}_{3i}$ is a $p \times (m+1)p$ matrix of the form $\delta_{m,n_L}^{-1}[O_p(1), O_p(m^{-1}), \ldots, O_p(m^{-1})]$. Next, $(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top = [\boldsymbol{I}_p + O_p(n_1^{-1}), \ldots, \boldsymbol{I}_p +$

$O_p(n_m^{-1})]^\top$ and $(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}$ is a block-diagonal $O_p(n_L^{-1})$ matrix component-wise. Therefore, we find that $\boldsymbol{\Lambda}$ is of the form

$$0.5\begin{bmatrix}\sum_{i=1}^m \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i+\hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{F}_{3i}\\ -(\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{B}_2^\top\sum_{i=1}^m \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top\dot{\boldsymbol{W}}_i\boldsymbol{X}_i+\hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{F}_{3i}\end{bmatrix}+0.5\begin{bmatrix}\mathbf{0}_{p\times(m+1)p}\\ (\boldsymbol{B}_3+\boldsymbol{B}_4)^{-1}\boldsymbol{F}_3\end{bmatrix}$$

$$\triangleq \boldsymbol{\Lambda}_1+\boldsymbol{\Lambda}_2$$

$$=\frac{1}{\delta_{m,n_L}}\begin{bmatrix}O_p(1) & O_p(m^{-1}) & \cdots & O_p(m^{-1})\\ \vdots & \vdots & \vdots & \vdots\\ O_p(1) & O_p(m^{-1}) & \cdots & O_p(m^{-1})\end{bmatrix}+\frac{1}{\delta_{m,n_L}}\begin{bmatrix}\multicolumn{5}{c}{\mathbf{0}_{p\times(m+1)p}}\\ O_p(1) & O_p(1) & \mathbf{0} & \cdots & \mathbf{0}\\ O_p(1) & \mathbf{0} & O_p(1) & \cdots & \mathbf{0}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ O_p(1) & \mathbf{0} & \cdots & \mathbf{0} & O_p(1)\end{bmatrix}$$

$$=\frac{1}{\delta_{m,n_L}}\begin{bmatrix}O_p(1) & O_p(m^{-1}) & O_p(m^{-1}) & \cdots & O_p(m^{-1})\\ O_p(1) & O_p(1) & O_p(m^{-1}) & \cdots & O_p(m^{-1})\\ O_p(1) & O_p(m^{-1}) & O_p(1) & \cdots & O_p(m^{-1})\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ O_p(1) & O_p(m^{-1}) & \cdots & O_p(m^{-1}) & O_p(1)\end{bmatrix}. \tag{S3.4}$$

Writing $\boldsymbol{\Lambda}=\delta_{m,n_L}^{-1}\boldsymbol{\Lambda}_\delta$, we see that the component-wise order of $\boldsymbol{\Lambda}_\delta$ remains the same no matter how many times it is multiplied by itself. Furthermore, each row of $\boldsymbol{\Lambda}_\delta^s$ is $O_p(1)$ for only a finite number of components, and $O_p(m^{-1})$ for the others. We will use these facts to examine the behaviour

of $\sum_{s=1}^{\infty} \boldsymbol{\Lambda}^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}}) = \sum_{s=1}^{\infty} \delta_{m,n_L}^{-s} \boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$, and we will do so separately for the conditional and unconditional regimes. Before proceeding, we first confirm the convergence of $\sum_{s=1}^{\infty} \boldsymbol{\Lambda}^s$.

Lemma 4. *Assume Conditions (C1)-(C5) are satisfied. Then with probability tending to one as $m, n_L \to \infty$, the geometric sum $\sum_{s=1}^{\infty} \boldsymbol{\Lambda}^s$ converges.*

Proof. To prove the result we will show that, with probability tending to one as $m, n_L \to \infty$, $\|\boldsymbol{\Lambda}\| < 1$ for some sub-multiplicative matrix norm $\|\cdot\|$. In particular, we will consider the maximum absolute row sum of $\boldsymbol{\Lambda}$, denoted by $\|\cdot\|_\infty$ i.e., the operator norm induced by the vector infinity norm.

From (S3.4), we have $\|\boldsymbol{\Lambda}\|_\infty \leq \|\boldsymbol{\Lambda}_1\|_\infty + \|\boldsymbol{\Lambda}_2\|_\infty$. We first examine $\|\boldsymbol{\Lambda}_1\|_\infty$. We may break up $\boldsymbol{\Lambda}_1$ into

$$0.5 \begin{bmatrix} \sum_{i=1}^{m} \boldsymbol{C}^{-1} \hat{\boldsymbol{G}}^{-1} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{F}_{3i} \\ -\sum_{i=1}^{m} \boldsymbol{C}^{-1} \hat{\boldsymbol{G}}^{-1} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{F}_{3i} \\ -\sum_{i=1}^{m} \boldsymbol{C}^{-1} \hat{\boldsymbol{G}}^{-1} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \boldsymbol{F}_{3i} \\ \vdots \end{bmatrix} +$$

$$0.5\begin{bmatrix} \mathbf{0}_p \\ (\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1 + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1} \sum_{i=1}^m \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{F}_{3i} \\ (\boldsymbol{X}_2^\top \dot{\boldsymbol{W}}_2 \boldsymbol{X}_2 + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1} \sum_{i=1}^m \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{F}_{3i} \\ \vdots \\ (\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1} \sum_{i=1}^m \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{F}_{3i} \end{bmatrix}$$

$$\triangleq \boldsymbol{\Lambda}_3 + \boldsymbol{\Lambda}_4\boldsymbol{\Lambda}_5,$$

where $\boldsymbol{\Lambda}_4 = \text{bdiag}(\mathbf{0}_{p\times p}, (\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1 + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1}, \ldots, (\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m + \hat{\boldsymbol{G}}^{-1})^{-1}\hat{\boldsymbol{G}}^{-1})$, and $\boldsymbol{\Lambda}_5 = (0, \mathbf{1}_m^\top)^\top \otimes \sum_{i=1}^m \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{F}_{3i}$. We can also write

$\boldsymbol{\Lambda}_3 = -0.5(\mathbf{1}_m^* \otimes \boldsymbol{I}_p) \sum_{i=1}^m \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{F}_{3i}$ and use the (component-wise) order results as used in (S3.4) to see that $\|\boldsymbol{\Lambda}_3\|_\infty \leq \|-0.5(\mathbf{1}_m^* \otimes \boldsymbol{I}_p)\|_\infty \| \sum_{i=1}^m \boldsymbol{C}^{-1}\hat{\boldsymbol{G}}^{-1}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{F}_{3i}\|_\infty = o_p(1)$. Next, we have $\|\boldsymbol{\Lambda}_4\boldsymbol{\Lambda}_5\|_\infty \leq \|\boldsymbol{\Lambda}_4\|_\infty \|\boldsymbol{\Lambda}_5\|_\infty$. We know $\|\boldsymbol{\Lambda}_5\|_\infty = o_p(1)$, and under conditions (C1)-(C2), we have $\|\boldsymbol{\Lambda}_4\|_\infty = O_p(1)$. Thus we obtain $\|\boldsymbol{\Lambda}_1\|_\infty = o_p(1)$.

Turning to $\boldsymbol{\Lambda}_2$, we examine each row of $(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\boldsymbol{F}_3$. First, $\|(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\boldsymbol{F}_3\|_\infty \leq \|(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\|_\infty \|\boldsymbol{F}_3\|_\infty$ and by conditions (C1)-(C2), we have $\|(\boldsymbol{B}_3 + \boldsymbol{B}_4)^{-1}\|_\infty = O_p(n_L^{-1})$. Now, without loss of generality consider the

first row of $\boldsymbol{F}_3$. This is given by

$$
\begin{aligned}
&-\Big[(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X} + (\hat{\boldsymbol{b}}-\dot{\boldsymbol{b}})^\top \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}, \\
&(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{Z} + (\hat{\boldsymbol{b}}-\dot{\boldsymbol{b}})^\top \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{Z}\Big] \\
&= -\Big[(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X} + (\hat{\boldsymbol{b}}-\dot{\boldsymbol{b}})^\top \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}, \\
&(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X} + (\hat{\boldsymbol{b}}-\dot{\boldsymbol{b}})^\top \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}, \boldsymbol{0}_{(m-1)p}^\top\Big],
\end{aligned}
$$

since $\mathrm{diag}(\boldsymbol{Z}_{[,1]})$ selects for the first cluster. Let $\bar{\mathbf{1}}_p$ be a p-vector whose entries consist of the (component-wise) signs of $(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X} + (\hat{\boldsymbol{b}}-\dot{\boldsymbol{b}})^\top \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}$. Then the absolute row sum of the first row of $\boldsymbol{F}_3$ is given by

$$
\begin{aligned}
&2|\{(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X} + (\hat{\boldsymbol{b}}-\dot{\boldsymbol{b}})^\top \boldsymbol{Z}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\bar{\mathbf{1}}_p| \\
=&2|\{(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X} + (\hat{\boldsymbol{b}}_1-\dot{\boldsymbol{b}}_1)^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\bar{\mathbf{1}}_p| \\
=&2|\{(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}}+\hat{\boldsymbol{b}}_1-\dot{\boldsymbol{b}}_1)^\top \boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\bar{\mathbf{1}}_p| \\
\leq&2p\|\{\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}}+\hat{\boldsymbol{b}}_1-\dot{\boldsymbol{b}}_1)\}\|_\infty \\
\leq&2p\|\{\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\|_\infty\|\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}}+\hat{\boldsymbol{b}}_1-\dot{\boldsymbol{b}}_1\|_\infty \\
\leq&2p\|\{\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,1]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\|_\infty(\|\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}}\|_\infty+\|\hat{\boldsymbol{b}}_1-\dot{\boldsymbol{b}}_1\|_\infty) \\
\leq&2p\max_{k\in\{1,\ldots,mp\}}\|\{\boldsymbol{X}^\top \mathrm{diag}(\boldsymbol{Z}_{[,k]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\|_\infty(\|\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}}\|_\infty+\|\hat{\boldsymbol{b}}_1-\dot{\boldsymbol{b}}_1\|_\infty)
\end{aligned}
$$

$$
\begin{aligned}
=&2p \max_{k\in\{1,\ldots,mp\}} \|\{\boldsymbol{X}^\top \text{diag}(\boldsymbol{Z}_{[,k]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\|_\infty \|\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}\|_\infty \\
&+ 2p \max_{k\in\{1,\ldots,mp\}} \|\{\boldsymbol{X}^\top \text{diag}(\boldsymbol{Z}_{[,k]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\|_\infty \|\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1\|_\infty \\
&\triangleq \alpha + \alpha_1 \triangleq \omega_1,
\end{aligned}
$$

where the second equality follows from $\text{diag}(\boldsymbol{Z}_{[,1]})$ selecting for only the first cluster, and $\boldsymbol{X}_i = \boldsymbol{Z}_i$. The first inequality is due to Hölder's inequality. Again, using Conditions (C1)-(C2) we have $\max_{k\in\{1,\ldots,mp\}} \|\{\boldsymbol{X}^\top \text{diag}(\boldsymbol{Z}_{[,k]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\|_\infty = O_p(n_U)$.

Now, p is a constant and the absolute row sum of any row of $\boldsymbol{F}_3$ can be bounded analogously in the above way, the only difference being that for the kth row, then the quantity $(\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1)$ changes to the prediction gap for the cluster that $\text{diag}(\boldsymbol{Z}_{[,k]})$ selects for. This means that the absolute row sums for the first p rows of $\boldsymbol{F}_3$ are bounded by ω_1, the next p rows by ω_2, and so on. Hence, to ensure $\|\boldsymbol{\Lambda}_2\|_\infty = o_p(1)$ it suffices to ensure that $\|\boldsymbol{\omega} \otimes \mathbf{1}_p\|_\infty = \|\boldsymbol{\omega}\|_\infty = o_p(n_L)$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)^\top$.

To show this, define $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$. Then $\|\boldsymbol{\omega}\|_\infty \leq \|\alpha \mathbf{1}_m\|_\infty + \|\boldsymbol{\alpha}\|_\infty$. By Conditions (C1)-(C2), we have $\|\alpha \mathbf{1}_m\|_\infty = \alpha = O_p(n_U) \times o_p(1) = o_p(n_U) = o_p(n_L)$. We also have

$$
\|\boldsymbol{\alpha}\|_\infty = 2p \max_{k\in\{1,\ldots,mp\}} \|\{\boldsymbol{X}^\top \text{diag}(\boldsymbol{Z}_{[,k]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\|_\infty \max_{i\in\{1,\ldots,m\}} \|\hat{\boldsymbol{b}}_i - \dot{\boldsymbol{b}}_i\|_\infty
$$

$$= 2p \max_{k \in \{1,\ldots,mp\}} \|\{\boldsymbol{X}^\top \text{diag}(\boldsymbol{Z}_{[,k]})\tilde{\boldsymbol{W}}'\boldsymbol{X}\}\|_\infty \|\hat{\dot{\boldsymbol{b}}} - \dot{\boldsymbol{b}}\|_\infty$$

$$= O_p(n_U) \times o_p(1) = o_p(n_U) = o_p(n_L),$$

where the last line follows from conditions (C1)-(C2), and the fact that $\|\hat{\dot{\boldsymbol{\theta}}} - \dot{\boldsymbol{\theta}}\|_\infty = o_p(1)$. The result follows since $\|\boldsymbol{\Lambda}\|_\infty$ is therefore of order $o_p(1)$, and for any $\epsilon > 0$ we have $\|\boldsymbol{\Lambda}\|_\infty < \epsilon$ with probability tending to one as $m, n_L \to \infty$. The argument above holds for both the conditional and unconditional regime, and the required result follows. □

S3.1 Conditional Regime

In the conditional regime, we assume without loss of generality that $\sum_{i=1}^m \dot{\boldsymbol{b}}_i = \boldsymbol{0}_p$, recalling that we can always reparametrise the random effects to satisfy this. From previous derivations, we know that when $mn_L^{-1} \to 0$, the quantity $\boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}})$ is of order $O_p(N^{-1/2})$ for the first p components and $O_p(n_L^{-1/2})$ for the last mp components. By the two properties of $\boldsymbol{\Lambda}_\delta^s$ noted above, we therefore know that $\boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}})$ is at most $O_p(n_L^{-1/2})$ component-wise for any s. Hence $\sum_{s=1}^\infty \delta_{m,n_L}^{-s} \boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}}) = \delta_{m,n_L}^{-1} O_p(n_L^{-1/2}) = o_p(n_L^{-1/2})$ for sufficiently large m, n_L by the properties of a geometric sum. This is sufficient to show that the last mp components of $\sum_{s=1}^\infty \boldsymbol{\Lambda}^s \boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}})$ are of smaller order component-wise than $\boldsymbol{B}^{-1}\nabla Q(\dot{\boldsymbol{\theta}})$, so that the result

for the prediction gap holds. In particular, we thus know that $\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} = O_p(n_L^{-1/2})$. Furthermore, we also know that the convergence rate of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ is at least of order $O_p(n_L^{-1/2})$. As a result, we can choose $\delta_{m,n_L} = n_L^{1/2}$ without affecting the component-wise order properties of $\boldsymbol{\Lambda}_\delta$. Applying $\delta_{m,n_L} = n_L^{1/2}$, we thus have that $\sum_{s=1}^{\infty} \delta_{m,n_L}^{-s} \boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ is at most of order $O_p(n_L^{-1})$ component-wise. This is smaller than $O_p(N^{-1/2})$ when $mn_L^{-1} \to 0$, and the required result follows.

S3.2 Unconditional Regime

For the unconditional regime, we consider two cases: when $mn_L^{-1} \to 0$, and when $mn_U^{-1} \to \infty$ but $mn_L^{-2} \to 0$.

First, consider the case when $mn_L^{-1} \to 0$. From previous derivations, we know that when $mn_L^{-1} \to 0$, the quantity $\boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ is of order $O_p(m^{-1/2})$ for the first p components and $O_p(m^{-1/2})$ for the last mp components. By the two properties of $\boldsymbol{\Lambda}_\delta^s$ noted above, we therefore know that $\boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ is at most $O_p(m^{-1/2})$ component-wise for any s. Hence $\sum_{s=1}^{\infty} \delta_{m,n_L}^{-s} \boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}}) = \delta_{m,n_L}^{-1} O_p(m^{-1/2}) = o_p(m^{-1/2})$ for sufficiently large m, n_L, by the properties of a geometric sum. The required result follows from this. Furthermore, this implies we may set $\delta_{m,n_L} = m^{1/2}$ without affecting the component-wise order properties of $\boldsymbol{\Lambda}_\delta$. Applying

$\delta_{m,n_L} = m^{1/2}$, we thus have that $\sum_{s=1}^{\infty} \delta_{m,n_L}^{-s} \boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ is at most of order $O_p(m^{-1})$ component-wise.

Next, consider the case when $mn_U^{-1} \to \infty$ and $mn_L^{-2} \to 0$. From previous derivations, we know in this setting it holds that $\boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ is of order $O_p(m^{-1/2})$ for the first p components and $O_p(n_L^{-1/2})$ for the last mp components. By the two properties of $\boldsymbol{\Lambda}_\delta^s$ noted above, we therefore obtain that $\boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ is at most $O_p(n_L^{-1/2})$ component-wise for any s. Hence $\sum_{s=1}^{\infty} \delta_{m,n_L}^{-s} \boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}}) = \delta_{m,n_L}^{-1} O_p(n_L^{-1/2}) = o_p(n_L^{-1/2})$ for sufficiently large m, n_L, by the properties of a geometric sum. This is sufficient to show that the last mp components of $\sum_{s=1}^{\infty} \boldsymbol{\Lambda}^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ are of smaller order component-wise than $\boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$, so that the result for the prediction gap holds. In particular, we thus know that $\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} = O_p(n_L^{-1/2})$. Furthermore, we also know that the convergence rate of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ is at least $O_p(n_L^{-1/2})$. As a result, we can set $\delta_{m,n_L} = n_L^{1/2}$ without affecting the component-wise order properties of $\boldsymbol{\Lambda}_\delta$. Applying $\delta_{m,n_L} = n_L^{1/2}$, we thus have that $\sum_{s=1}^{\infty} \delta_{m,n_L}^{-s} \boldsymbol{\Lambda}_\delta^s \boldsymbol{B}^{-1} \nabla Q(\dot{\boldsymbol{\theta}})$ is at most of order $O_p(n_L^{-1})$ component-wise. This is smaller than $O_p(m^{-1/2})$ when $mn_U^{-1} \to \infty$, $mn_L^{-2} \to 0$ and the result follows.

S4 Unpartnered Fixed Effects

S4.1 Generalised Linear Models

In the special case when $\dot{\boldsymbol{G}} = \mathbf{0}_{p\times p}$, i.e., all fixed effects are unpartnered in the true data generating process, the GLMM reduces to a GLM. We may then obtain a result based on a special case of our results in the conditional case, when all the true random effects are equal to zero. The result is as follows.

Corollary A1. Assume Conditions (C1) - (C5) are satisfied and $mn_L^{-1} \to 0$. Then as $m, n_L \to \infty$ and when the true vector of random effects $\dot{\boldsymbol{b}} = \mathbf{0}_{mp}$, it holds that $\boldsymbol{AD}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Omega})$.

S4.2 Linear Mixed Models

Suppose for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$ we observe data from the model $y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i + \boldsymbol{x}_{ij}^{(O)\top} \boldsymbol{\beta}^{(O)} + \epsilon_{ij}$, where $\boldsymbol{x}_{ij} = \boldsymbol{z}_{ij}$ for all (i, j), $\boldsymbol{b}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \hat{\boldsymbol{G}})$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \phi)$. Note that this is part of the exponential family. Partition $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(P)\top}, \boldsymbol{\beta}^{(U)\top})^\top$, corresponding to the p_P partnered and p_U unpartnered fixed effects ($p_P + p_U = p$). That is, if we partition $\boldsymbol{b}_i = (\boldsymbol{b}_i^{(P)\top}, \boldsymbol{b}_i^{(U)\top})^\top$, then $\boldsymbol{b}_i^{(U)} = \mathbf{0}_{p_U}$ for all i, and the corresponding elements in $\dot{\boldsymbol{G}}$ are zero. Let $\boldsymbol{\theta}^\times = (\boldsymbol{\beta}^\top, \boldsymbol{b}_1^{(P)\top}, \ldots, \boldsymbol{b}_m^{(P)\top}, \boldsymbol{b}_1^{(U)\top}, \ldots, \boldsymbol{b}_m^{(U)\top}, \boldsymbol{\beta}^{(O)\top})^\top$,

$\boldsymbol{\theta}^{-} = (\boldsymbol{\beta}^{\top}, \boldsymbol{b}^{\top}, \boldsymbol{\beta}^{(O)\top})^{\top}$,

$\boldsymbol{D}^{\times} = \mathrm{diag}(m^{1/2}\mathbf{1}_{p_P}, N^{1/2}\mathbf{1}_{p_U}, m^{1/2}\mathbf{1}_{mp_P}, n_1^{1/2}\mathbf{1}_{p_U}, \ldots, n_m^{1/2}\mathbf{1}_{p_U}, N^{1/2}\mathbf{1}_{p_O})$, and $\boldsymbol{D}^{-} = \mathrm{diag}(m^{1/2}\mathbf{1}_{p_P}, N^{1/2}\mathbf{1}_{p_U}, n_1^{1/2}\mathbf{1}_{p}, \ldots, n_m^{1/2}\mathbf{1}_{p}, N^{1/2}\mathbf{1}_{p_O})$. Also let $\boldsymbol{X}_i^{(O)} = [\boldsymbol{x}_{i1}^{(O)}, \ldots, \boldsymbol{x}_{in_i}^{(O)}]^{\top}$ and $\boldsymbol{X}^{(O)} = [\boldsymbol{X}_1^{(O)\top}, \ldots, \boldsymbol{X}_m^{(O)\top}]^{\top}$. The p_O orthogonal fixed effects $\boldsymbol{x}_{ij}^{(O)}$ satisfy $\boldsymbol{X}^{(O)\top}\boldsymbol{Z} = \mathbf{0}_{p_O \times mp}$, for example orthogonal polynomials of $\boldsymbol{x}_{ij}$. This implies $\boldsymbol{X}_i^{(O)\top}\boldsymbol{X}_i = \mathbf{0}_{p_O \times p}$ for all i. For a $q \times \{(m+1)p + p_O\}$ matrix $\boldsymbol{A}^*$ with the finite selection property, we have the following.

Corollary A2. Assume Conditions (C1) - (C4) are satisfied. Then as $m, n_L \to \infty$ and unconditional on the random effects $\dot{\boldsymbol{b}}$, it holds that

1. $\boldsymbol{A}^*\boldsymbol{D}^{\times}(\hat{\boldsymbol{\theta}}^{\times} - \dot{\boldsymbol{\theta}}^{\times}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Omega}_a)$ if $mn_L^{-1} \to 0$, and

2. $\boldsymbol{A}^*\boldsymbol{D}^{-}(\hat{\boldsymbol{\theta}}^{-} - \dot{\boldsymbol{\theta}}^{-}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Omega}_b)$ if $mn_U^{-1} \to \infty$,

where

$$\boldsymbol{\Omega}_a = \lim_{m, n_L \to \infty} \boldsymbol{A}^* \begin{bmatrix} \dot{\boldsymbol{G}}_{[1:p_P, 1:p_P]} & \mathbf{0}_{p_P \times p_U} & \mathbf{1}_m^{\top} \otimes \dot{\boldsymbol{G}}_{[1:p_P, 1:p_P]} & \mathbf{0}_{p_P \times mp_U} & \mathbf{0}_{p_P \times p_O} \\ \mathbf{0}_{p_U \times p_P} & \boldsymbol{\Omega}_1 & \mathbf{0}_{p_U \times mp_P} & \mathbf{0}_{p_U \times mp_U} & \mathbf{0}_{p_U \times p_O} \\ \mathbf{1}_m \otimes \dot{\boldsymbol{G}}_{[1:p_P, 1:p_P]} & \mathbf{0}_{mp_P \times p_U} & \mathbf{1}_{m \times m} \otimes \dot{\boldsymbol{G}}_{[1:p_P, 1:p_P]} & \mathbf{0}_{mp_P \times mp_U} & \mathbf{0}_{mp_P \times p_O} \\ \mathbf{0}_{mp_U \times p_P} & \mathbf{0}_{mp_U \times p_U} & \mathbf{0}_{mp_U \times mp_P} & \boldsymbol{\Omega}_2 & \mathbf{0}_{mp_U \times p_O} \\ \mathbf{0}_{p_O \times p_P} & \mathbf{0}_{p_O \times p_U} & \mathbf{0}_{p_O \times mp_P} & \mathbf{0}_{p_O \times mp_U} & \boldsymbol{\Omega}_3 \end{bmatrix} \boldsymbol{A}^{*\top}$$

$$\boldsymbol{\Omega}_b = \lim_{m,n_L\to\infty} \boldsymbol{A}^* \begin{bmatrix} \dot{\boldsymbol{G}}_{[1:p_P,1:p_P]} & \boldsymbol{0}_{p_P\times p_U} & \boldsymbol{0}_{p_P\times mp} & \boldsymbol{0}_{p_P\times p_O} \\ \boldsymbol{0}_{p_U\times p_P} & \boldsymbol{\Omega}_1 & \boldsymbol{0}_{p_U\times mp} & \boldsymbol{0}_{p_U\times p_O} \\ \boldsymbol{0}_{mp\times p_P} & \boldsymbol{0}_{mp\times p_U} & \boldsymbol{\Omega}_4 & \boldsymbol{0}_{mp\times p_O} \\ \boldsymbol{0}_{p_O\times p_P} & \boldsymbol{0}_{p_O\times p_U} & \boldsymbol{0}_{p_O\times mp} & \boldsymbol{\Omega}_3 \end{bmatrix} \boldsymbol{A}^{*\top}$$

$$\boldsymbol{\Omega}_1 = \left\{ \frac{\dot{\phi}}{m} \sum_{i=1}^{m} \frac{n}{n_i} \left(\frac{\boldsymbol{X}_i^\top \boldsymbol{X}_i}{n_i} \right)^{-1} \right\}_{[(p-p_U+1):p,(p-p_U+1):p]}$$

$$\boldsymbol{\Omega}_2 = \text{bdiag} \left[\left\{ \dot{\phi} \left(\frac{\boldsymbol{X}_1^\top \boldsymbol{X}_1}{n_1} \right)^{-1} \right\}_{[(p-p_U+1):p,(p-p_U+1):p]}, \ldots, \left\{ \dot{\phi} \left(\frac{\boldsymbol{X}_m^\top \boldsymbol{X}_m}{n_m} \right)^{-1} \right\}_{[(p-p_U+1):p,(p-p_U+1):p]} \right]$$

$$\boldsymbol{\Omega}_3 = \dot{\phi} \left(\frac{\boldsymbol{X}^{(O)\top} \boldsymbol{X}^{(O)}}{N} \right)^{-1}$$

$$\boldsymbol{\Omega}_4 = \text{bdiag} \left[\left\{ \dot{\phi} \left(\frac{\boldsymbol{X}_1^\top \boldsymbol{X}_1}{n_1} \right)^{-1} \right\}, \ldots, \left\{ \dot{\phi} \left(\frac{\boldsymbol{X}_m^\top \boldsymbol{X}_m}{n_m} \right)^{-1} \right\} \right].$$

Proof. We use the same approach as previous proofs and examine the Taylor expansion (S2.1). In this case, we have the expressions

$$\nabla Q(\dot{\boldsymbol{\theta}}) = \begin{bmatrix} \dot{\phi}^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \dot{\boldsymbol{\mu}}) \\ \dot{\phi}^{-1} \boldsymbol{Z}^\top (\boldsymbol{y} - \dot{\boldsymbol{\mu}}) - (\boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1}) \dot{\boldsymbol{b}} \\ \dot{\phi}^{-1} \boldsymbol{X}^{(O)\top} (\boldsymbol{y} - \dot{\boldsymbol{\mu}}) \end{bmatrix},$$

$$\boldsymbol{B}(\dot{\boldsymbol{\theta}}) = -\nabla^2 Q(\dot{\boldsymbol{\theta}}) = \begin{bmatrix} \boldsymbol{X}^\top \dot{\boldsymbol{W}} \boldsymbol{X} & \boldsymbol{X}^\top \dot{\boldsymbol{W}} \boldsymbol{Z} & \boldsymbol{X}^\top \dot{\boldsymbol{W}} \boldsymbol{X}^{(O)} \\ \boldsymbol{Z}^\top \dot{\boldsymbol{W}} \boldsymbol{X} & \boldsymbol{Z}^\top \dot{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1} & \boldsymbol{Z}^\top \dot{\boldsymbol{W}} \boldsymbol{X}^{(O)} \\ \boldsymbol{X}^{(O)\top} \dot{\boldsymbol{W}} \boldsymbol{X} & \boldsymbol{X}^{(O)\top} \dot{\boldsymbol{W}} \boldsymbol{Z} & \boldsymbol{X}^{(O)\top} \dot{\boldsymbol{W}} \boldsymbol{X}^{(O)} \end{bmatrix}$$

$$= \dot{\phi}^{-1} \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{Z} & \boldsymbol{0}_{p \times p_O} \\ \boldsymbol{Z}^\top \boldsymbol{X} & \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{I}_m \otimes \hat{\boldsymbol{G}}^{-1} & \boldsymbol{0}_{mp \times p_O} \\ \boldsymbol{0}_{p_O \times p} & \boldsymbol{0}_{p_O \times mp} & \boldsymbol{X}^{(O)\top} \boldsymbol{X}^{(O)} \end{bmatrix},$$

where the last equality follows from the fact that $\dot{\boldsymbol{W}} = \dot{\phi}^{-1} \boldsymbol{I}_N$ and $\boldsymbol{X}^{(O)\top} \boldsymbol{Z} = \boldsymbol{0}_{p_O \times mp}$. Since $\boldsymbol{B}(\dot{\boldsymbol{\theta}})$ is block diagonal, we thus know that expressions (S2.9) and (S2.10) still hold. Recall

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}} = {} & m^{-1} \sum_{i=1}^{m} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} \dot{\phi}^{-1} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) + m^{-1} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i \\ & - m^{-1} \sum_{i=1}^{m} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \hat{\boldsymbol{G}}^{-1} \dot{\boldsymbol{b}}_i + \frac{1}{2} \{\boldsymbol{B}^{-1} \boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[1:p]} \\ & + O_p(n_L^{-1}) \Bigg\{ m^{-1} \sum_{i=1}^{m} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \dot{\phi}^{-1} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) + m^{-1} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i \\ & - m^{-1} \sum_{i=1}^{m} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1} \hat{\boldsymbol{G}}^{-1} \dot{\boldsymbol{b}}_i \Bigg\} + m^{-1} \sum_{i=1}^{m} O_p(n_L^{-2}) \dot{\phi}^{-1} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) \end{aligned}$$

and

$$\begin{aligned} \hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} = {} & -\boldsymbol{1}_m \otimes m^{-1} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i + \boldsymbol{B}_3^{-1} \{\dot{\phi}^{-1} \boldsymbol{Z}^\top (\boldsymbol{y} - \dot{\boldsymbol{\mu}})\} \\ & + O_p(N^{-1/2}) + O_p(n_L^{-1}) + \frac{1}{2} \{\boldsymbol{B}^{-1} \boldsymbol{R}(\tilde{\boldsymbol{\theta}})\}_{[p+1:(m+1)p]}. \end{aligned}$$

In the LMM case, the remainder term in the Taylor expansion is zero. Thus the dominating term on the right hand side for $\hat{\boldsymbol{\beta}}^{(U)} - \dot{\boldsymbol{\beta}}^{(U)}$ are the last p_U components of $m^{-1} \sum_{i=1}^{m} (\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1} \dot{\phi}^{-1} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i)$, since

the last p_U components of $m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i$ are zero. Noting that $\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^\top =: \boldsymbol{\epsilon}_i$, the result for the unpartnered fixed effects follows after normalising by $N^{1/2}$.

Next, again from the Taylor expansion we have from the block-diagonal structure of $\boldsymbol{B}(\dot{\boldsymbol{\theta}})$ that $\hat{\boldsymbol{\beta}}^{(O)} - \dot{\boldsymbol{\beta}}^{(O)} = (\boldsymbol{X}^{(O)\top}\boldsymbol{X}^{(O)})^{-1}\boldsymbol{X}^{(O)\top}(\boldsymbol{y} - \dot{\boldsymbol{\mu}})$ and the result follows after normalising by $N^{1/2}$ since $\boldsymbol{y} - \dot{\boldsymbol{\mu}} = (\epsilon_{11}, \ldots, \epsilon_{mn_m})^\top =: \boldsymbol{\epsilon}$.

Finally, the result for the unpartnered random effects follows from the fact that the last p_U components of $m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i$ are zero so that the dominating term on the right hand side is $(\boldsymbol{X}_i^\top \boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i)$, and normalising by n_i.

The proofs for the partnered fixed and random effects are analogous to the proofs of Theorems 2 and 4, based on examining the leading term in the Taylor expansion.

For the joint behaviour of the estimator, we examine the joint behaviour of the leading terms in the Taylor Expansion. Note that $\boldsymbol{\epsilon}$ is multivariate normal with covariance matrix $\dot{\phi}\boldsymbol{I}_N$, $\dot{\boldsymbol{b}}$ is multivariate normal with covariance matrix $\boldsymbol{I}_m \otimes \dot{\boldsymbol{G}}$, $\boldsymbol{\epsilon}$ and $\dot{\boldsymbol{b}}$ are independent, and all the leading terms in the Taylor expansion are linear functions of $\boldsymbol{\epsilon}$ and $\dot{\boldsymbol{b}}$. To determine the joint behaviour of the estimator it is thus sufficient to derive the limiting covariance between the normalised leading terms, as we see (from the lead-

ing terms) that the estimator itself is also (asymptotically) multivariate normal. For example,

$$\begin{aligned}
&\mathrm{Cov}\left\{ N^{1/2}m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i-\dot{\boldsymbol{\mu}}_i), N^{1/2}(\boldsymbol{X}^{(O)\top}\boldsymbol{X}^{(O)})^{-1}\boldsymbol{X}^{(O)\top}(\boldsymbol{y}-\dot{\boldsymbol{\mu}})\right\} \\
&= n\dot{\phi}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^\top\boldsymbol{X}_i^{(O)}(\boldsymbol{X}^{(O)\top}\boldsymbol{X}^{(O)})^{-1} \\
&= \mathbf{0}_{p\times p_O}
\end{aligned}$$

due to the mutual independence of the ϵ_{ij} and orthogonality condition of $\boldsymbol{X}^{(O)}$. The pairwise limiting covariances between the leading terms can all be derived in a similar way and the result follows. Notice here that quantities with different convergence rates are always asymptotically uncorrelated and independent in this case.

□

Note that the results hold by the Lindeberg-Feller Central Limit Theorem even if the true distribution of ϵ_{ij} is not normal, as long as it is mean zero with finite variance. Also note that condition (C5) is no longer required, and that there is no restriction on the relative rates of m and n_L, since there is no remainder term to deal with. Our result is consistent with the results derived in Lyu and Welsh (2021a,b) who also derive a $N^{1/2}$ convergence rate for unpartnered fixed effects that are time-varying.

In practice, we do not know if a fixed effect is truly partnered with a random effect or not, and therefore the correct asymptotic distribution and convergence rate is also unknown. In this case, an appropriate finite sample approximation, given consistent estimators $\tilde{\boldsymbol{G}}$ and $\tilde{\phi}$ of $\dot{\boldsymbol{G}}$ and $\dot{\phi}$ respectively, is

$$\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}} \sim N\left\{\mathbf{0}, m^{-1}\tilde{\boldsymbol{G}} + N^{-1}\frac{\tilde{\phi}}{m}\sum_{i=1}^{m}\frac{n}{n_i}\left(\frac{\boldsymbol{X}_i^\top \boldsymbol{X}_i}{n_i}\right)^{-1}\right\},$$

which is based on the distribution of $m^{-1}\sum_{i=1}^{m}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1}\dot{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) + m^{-1}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i$, noting that the two terms are independent.

S5 Additional Simulation Results

S5.1 Main Results for the Conditional Regime

Figures 1, 2 and 3 display the empirical coverage probabilities and results from applying the Shapiro-Wilk test, respectively, under the conditional regime and for the 25 combinations of (m, n). Although our coverage intervals often undercovered or overcovered for small cluster sizes e.g., $n = 25$, especially for the Bernoulli case, they all moved toward nominal coverage as n becomes larger than m. This is consistent with Theorem 1. The fact the empirical coverage probabilities were slow in tending towards the nominal 95% level was also not overly surprising, as the third derivative term in the corresponding Taylor expansion is $O_p(m^{1/2}n_L^{-1/2})$. The Shapiro-Wilk tests overall did not indicate any evidence of deviations away from normality when $m < n$, although there were occasionally a few p-values less than 0.05. Overall, these results strongly support the use of Theorem 1 for inference under the conditional regime.

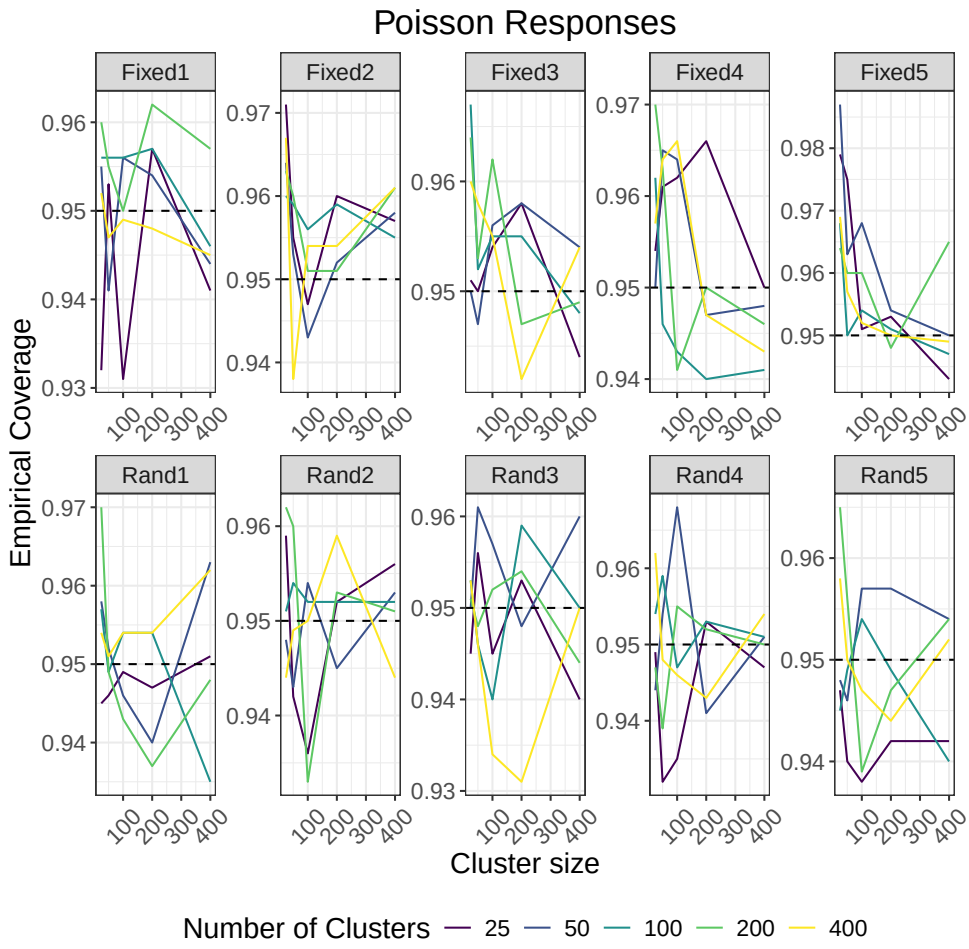

Figure 1: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Poisson responses.

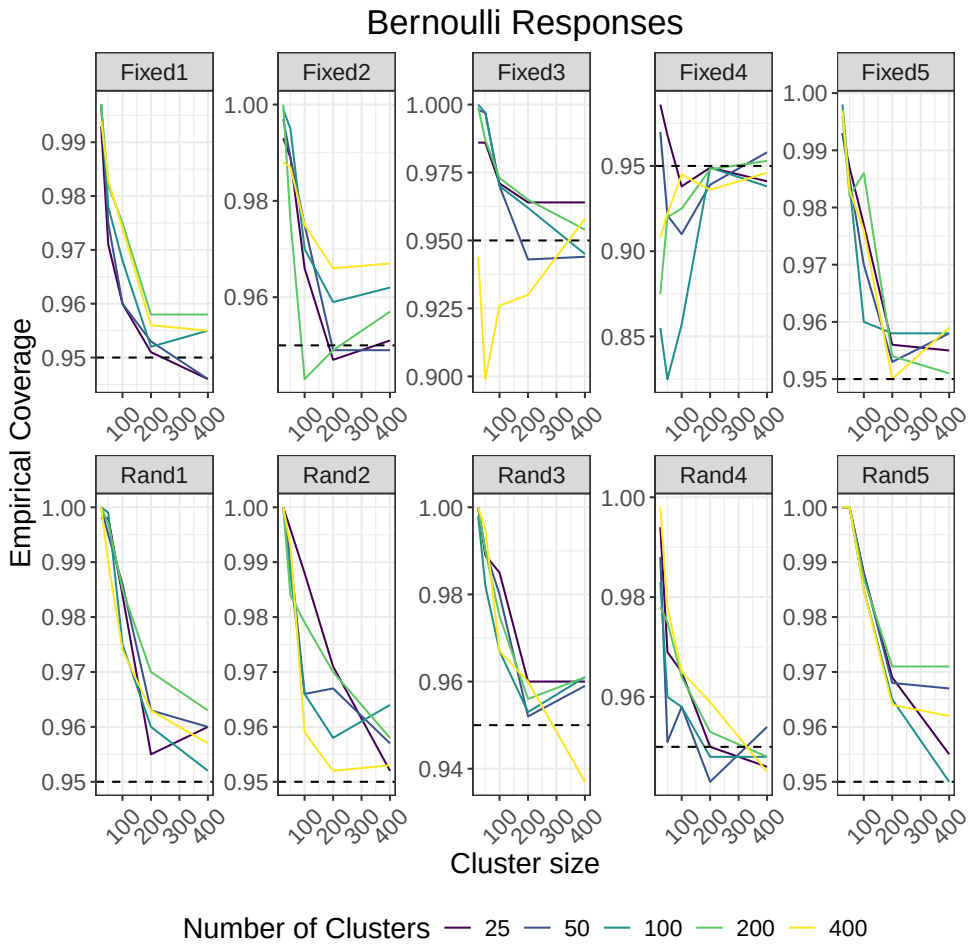

Figure 2: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Bernoulli responses.

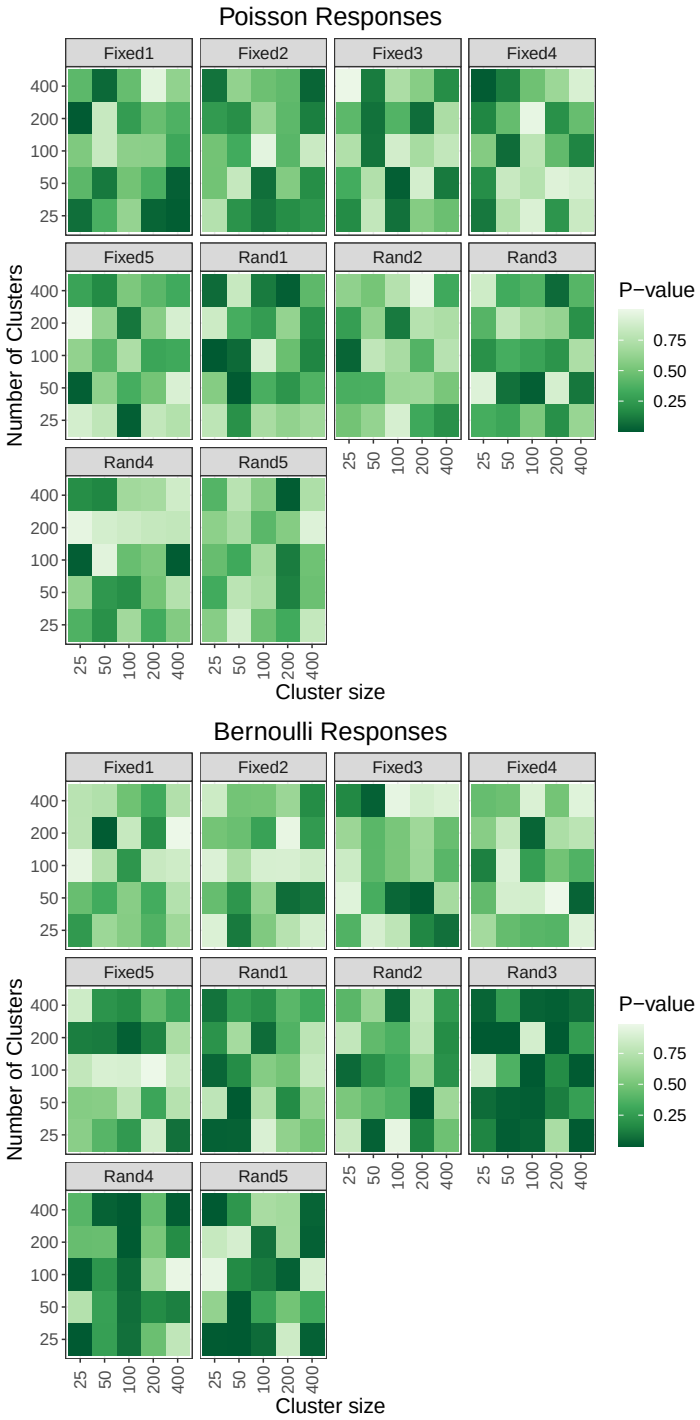

Figure 3: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the conditional regime.

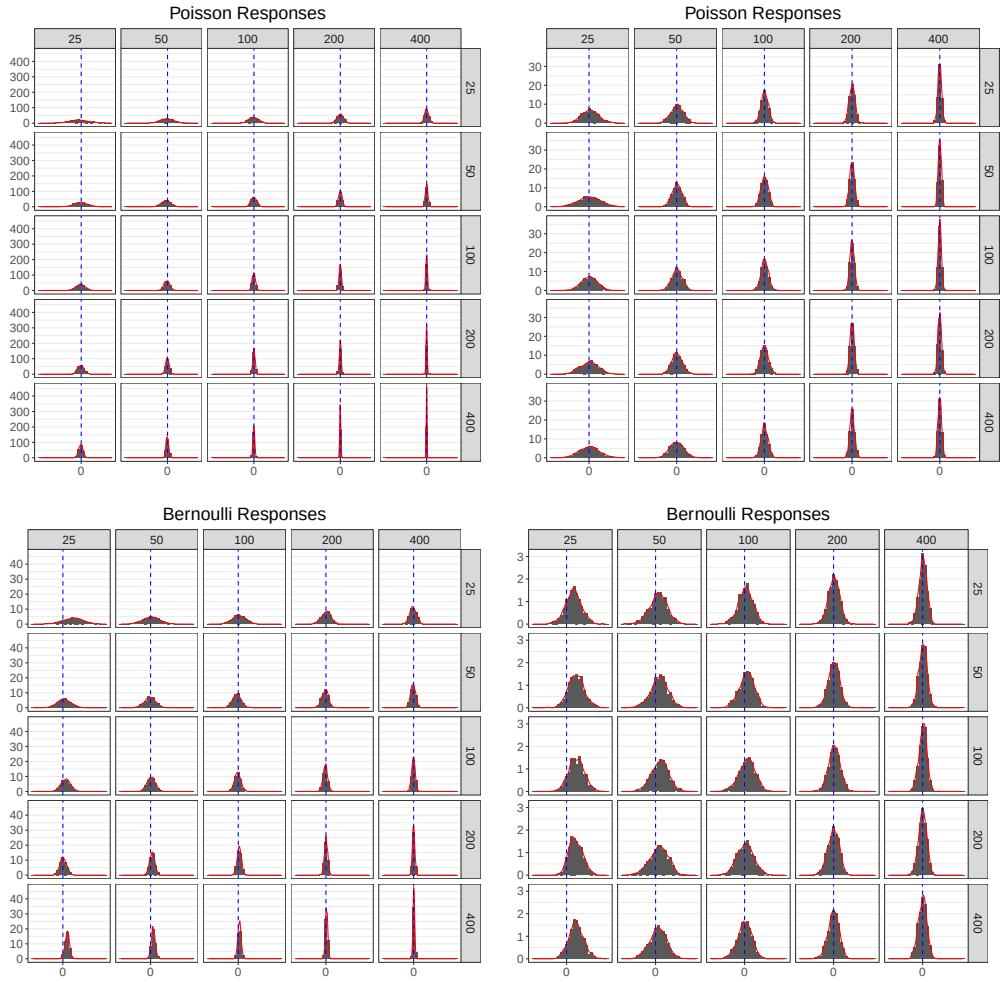

Figure 4: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the conditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

S5.2 Frobenius Norm

Table 1: Empirical mean Frobenius norm of the difference between estimated and true random effects covariance matrix.

		Poisson					Bernoulli				
	m	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
$\hat{\boldsymbol{G}} = m^{-1}\sum_{i=1}^{m}\hat{\boldsymbol{b}}_i$	25	1.06	1.06	1.06	1.05	1.06	1.76	1.47	1.24	1.12	1.09
	50	0.77	0.75	0.75	0.76	0.75	1.79	1.40	1.03	0.83	0.77
	100	0.54	0.54	0.54	0.54	0.54	1.80	1.38	0.89	0.63	0.56
	200	0.39	0.38	0.38	0.38	0.38	1.80	1.35	0.82	0.51	0.41
	400	0.27	0.27	0.27	0.27	0.27	1.81	1.35	0.78	0.43	0.32
$\hat{\boldsymbol{G}} = 0.25\boldsymbol{I}_2$	25	1.02	1.01	1.03	1.05	1.04	1.90	1.71	1.47	1.23	1.04
	50	0.73	0.73	0.74	0.74	0.76	1.90	1.70	1.44	1.15	0.91
	100	0.56	0.53	0.52	0.53	0.54	1.90	1.69	1.42	1.11	0.84
	200	0.44	0.39	0.37	0.38	0.38	1.90	1.68	1.41	1.09	0.79
	400	0.38	0.29	0.27	0.27	0.27	1.89	1.68	1.40	1.08	0.77
$\hat{\boldsymbol{G}} = 0.5\boldsymbol{I}_2$	25	1.02	1.03	1.04	1.05	1.05	1.61	1.39	1.16	1.01	0.96
	50	0.74	0.75	0.75	0.75	0.74	1.61	1.34	1.07	0.86	0.75
	100	0.53	0.52	0.54	0.54	0.54	1.60	1.32	1.03	0.77	0.61
	200	0.39	0.38	0.38	0.38	0.38	1.59	1.31	1.01	0.73	0.52
	400	0.30	0.27	0.27	0.27	0.27	1.59	1.30	0.99	0.70	0.47
$\hat{\boldsymbol{G}} = \boldsymbol{I}_2$	25	1.06	1.05	1.04	1.04	1.06	1.21	1.06	0.98	0.97	1.00
	50	0.74	0.75	0.75	0.75	0.76	1.17	0.93	0.78	0.73	0.71
	100	0.53	0.53	0.54	0.53	0.54	1.13	0.86	0.65	0.56	0.53
	200	0.38	0.38	0.38	0.38	0.38	1.12	0.82	0.58	0.44	0.39
	400	0.27	0.27	0.27	0.27	0.27	1.10	0.80	0.55	0.38	0.30
$\hat{\boldsymbol{G}} = 2\boldsymbol{I}_2$	25	1.06	1.06	1.06	1.05	1.06	0.84	0.98	1.03	1.04	1.05
	50	0.75	0.74	0.75	0.76	0.75	0.71	0.71	0.74	0.74	0.75
	100	0.54	0.54	0.54	0.53	0.53	0.56	0.51	0.53	0.53	0.54
	200	0.38	0.38	0.38	0.38	0.38	0.47	0.38	0.37	0.38	0.38
	400	0.27	0.27	0.27	0.27	0.27	0.42	0.29	0.27	0.27	0.27
$\hat{\boldsymbol{G}} = 4\boldsymbol{I}_2$	25	1.06	1.06	1.06	1.06	1.06	1.33	1.42	1.25	1.16	1.11
	50	0.77	0.76	0.76	0.75	0.76	1.18	1.07	0.93	0.83	0.80
	100	0.55	0.54	0.54	0.54	0.54	0.97	0.86	0.70	0.61	0.57
	200	0.39	0.38	0.38	0.38	0.38	0.86	0.72	0.55	0.45	0.41
	400	0.27	0.27	0.27	0.27	0.27	0.80	0.65	0.46	0.34	0.30

S5.3 $\hat{\boldsymbol{G}} = 0.25\ \boldsymbol{I}_2$

Using a large $\hat{\boldsymbol{G}}$ of $4\boldsymbol{I}_2$ had the least impact on the results, while a small $\hat{\boldsymbol{G}}$, e.g., $0.25\boldsymbol{I}_2$ had more of a noticeable impact at small sample sizes. This is not surprising since the latter corresponds to more shrinkage, such that larger sample sizes are needed before asymptotic results apply.

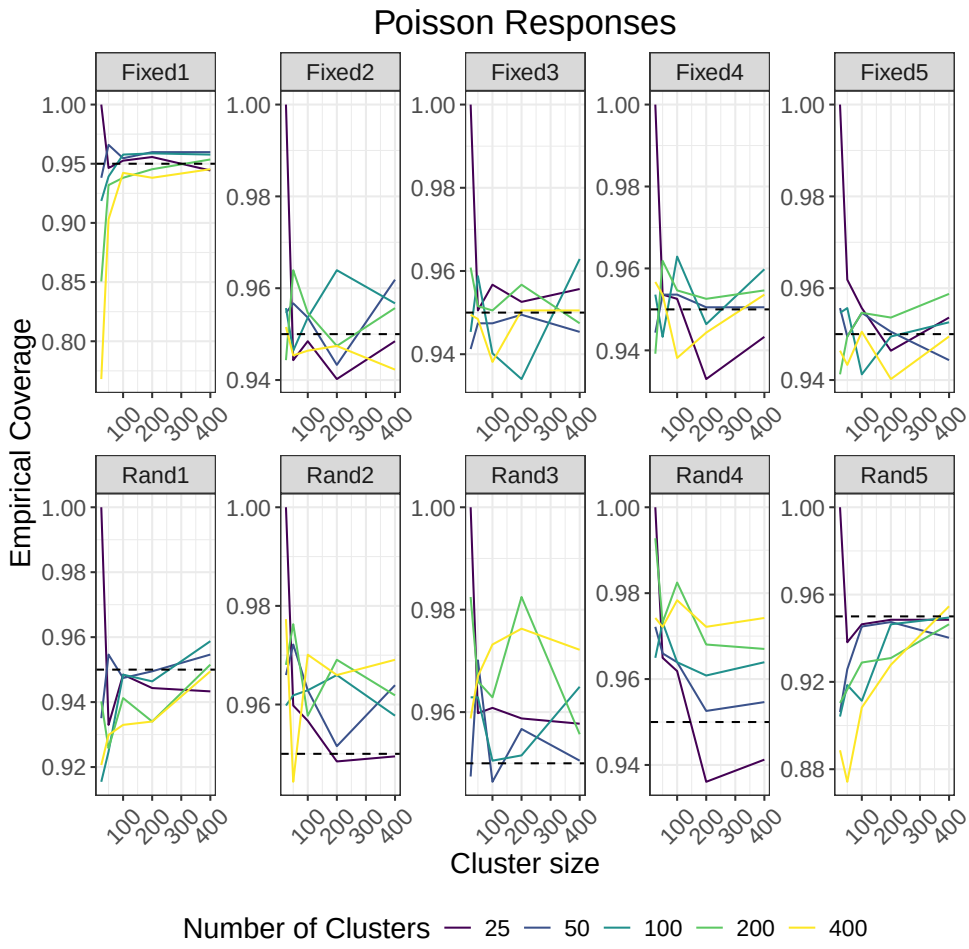

Figure 5: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Poisson responses.

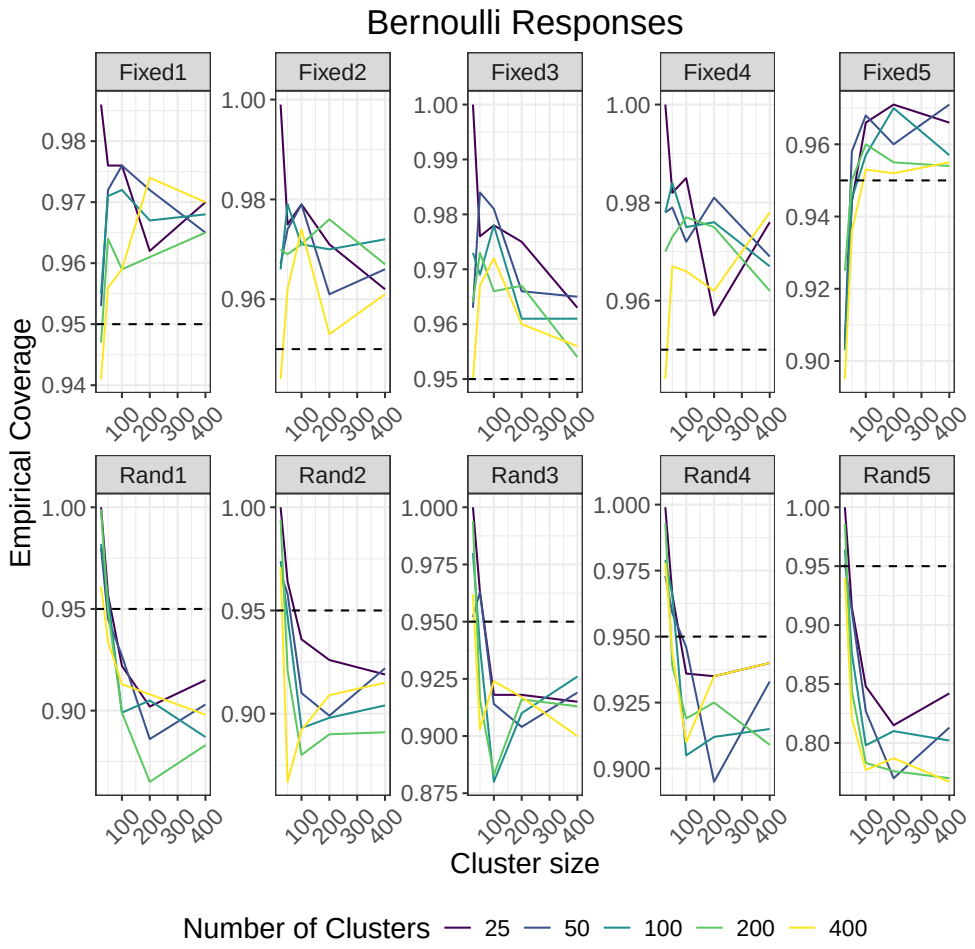

Figure 6: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Bernoulli responses.

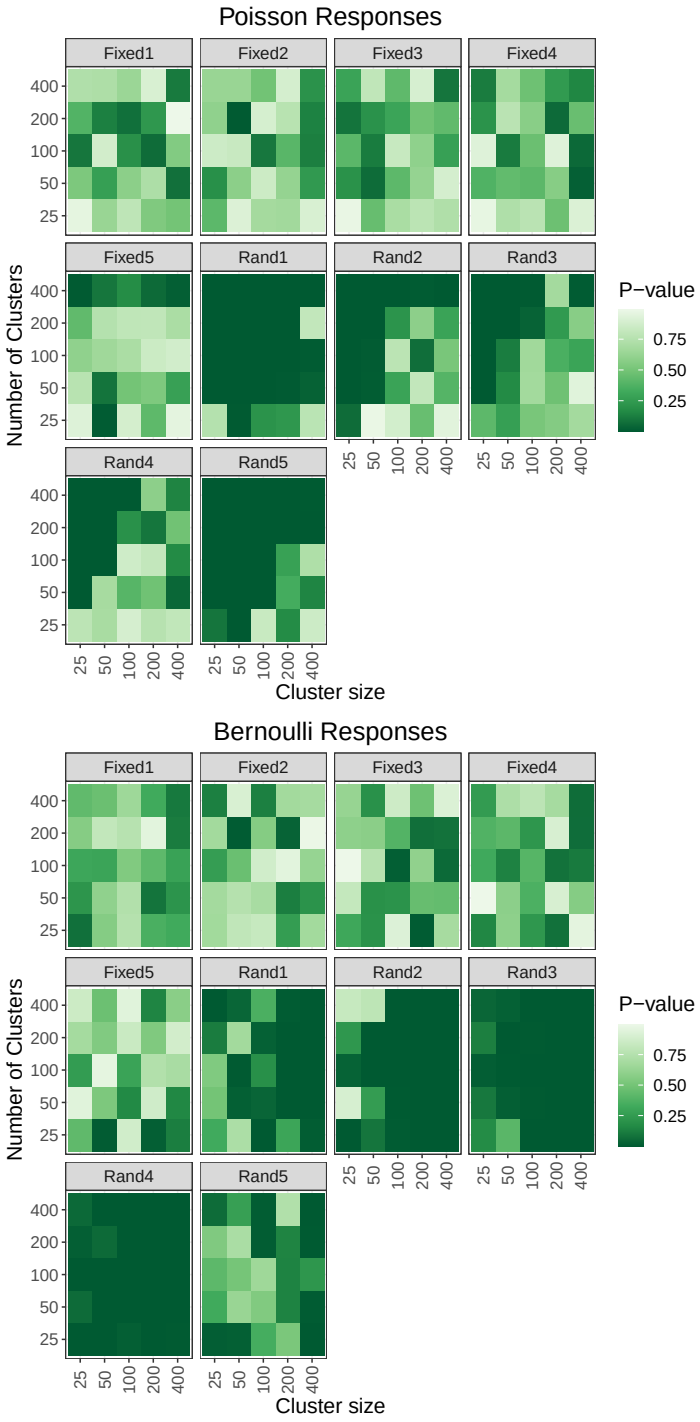

Figure 7: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the unconditional regime.

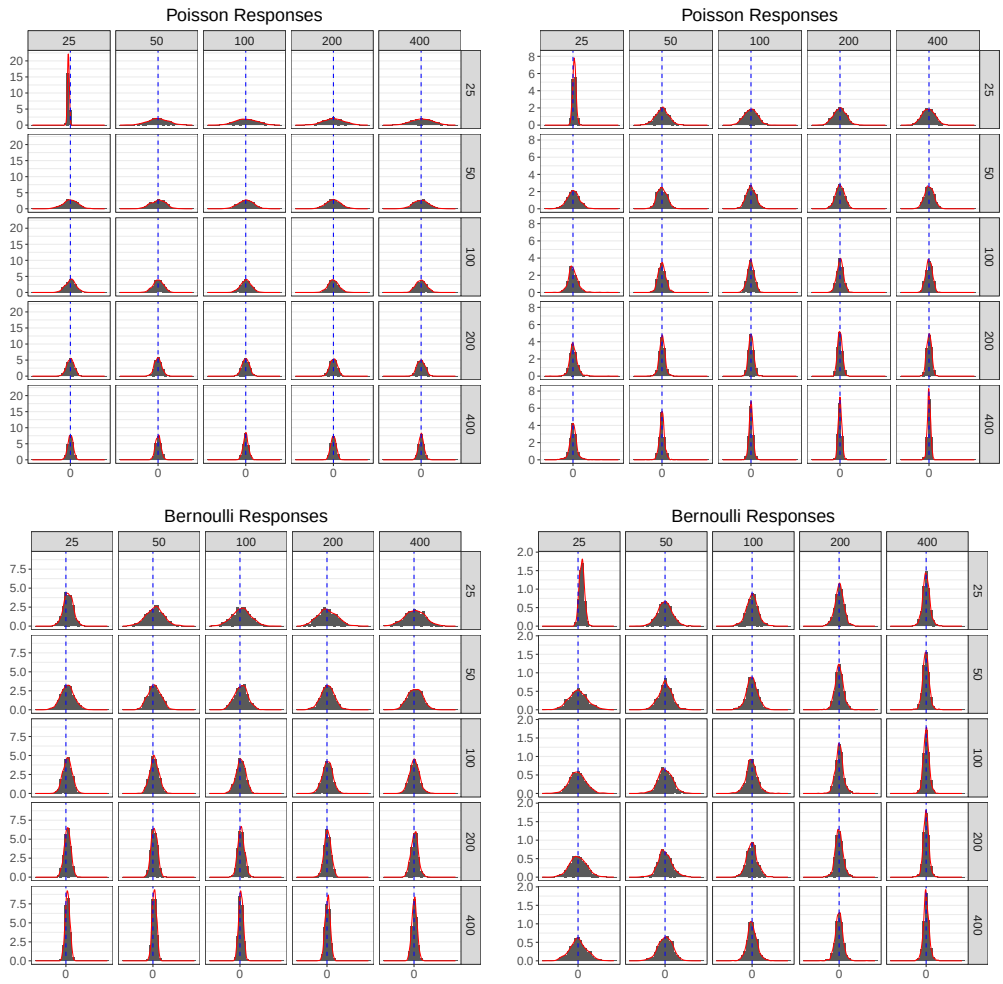

Figure 8: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

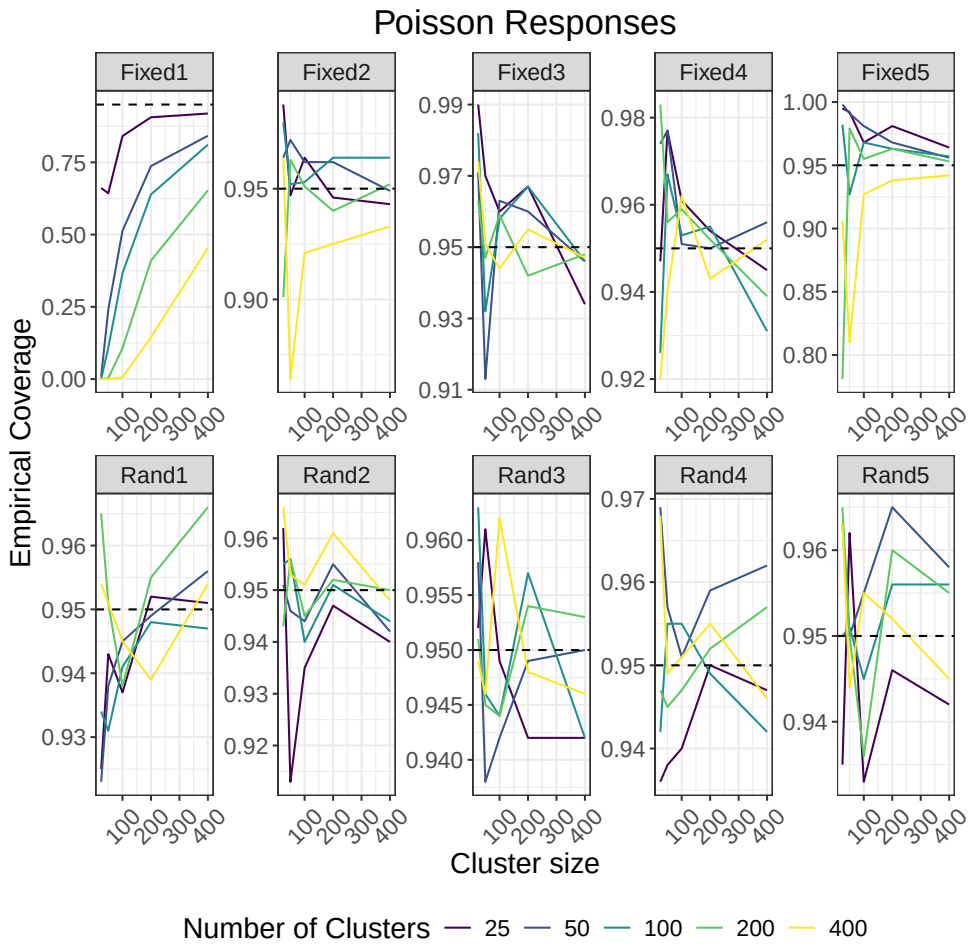

Figure 9: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Poisson responses.

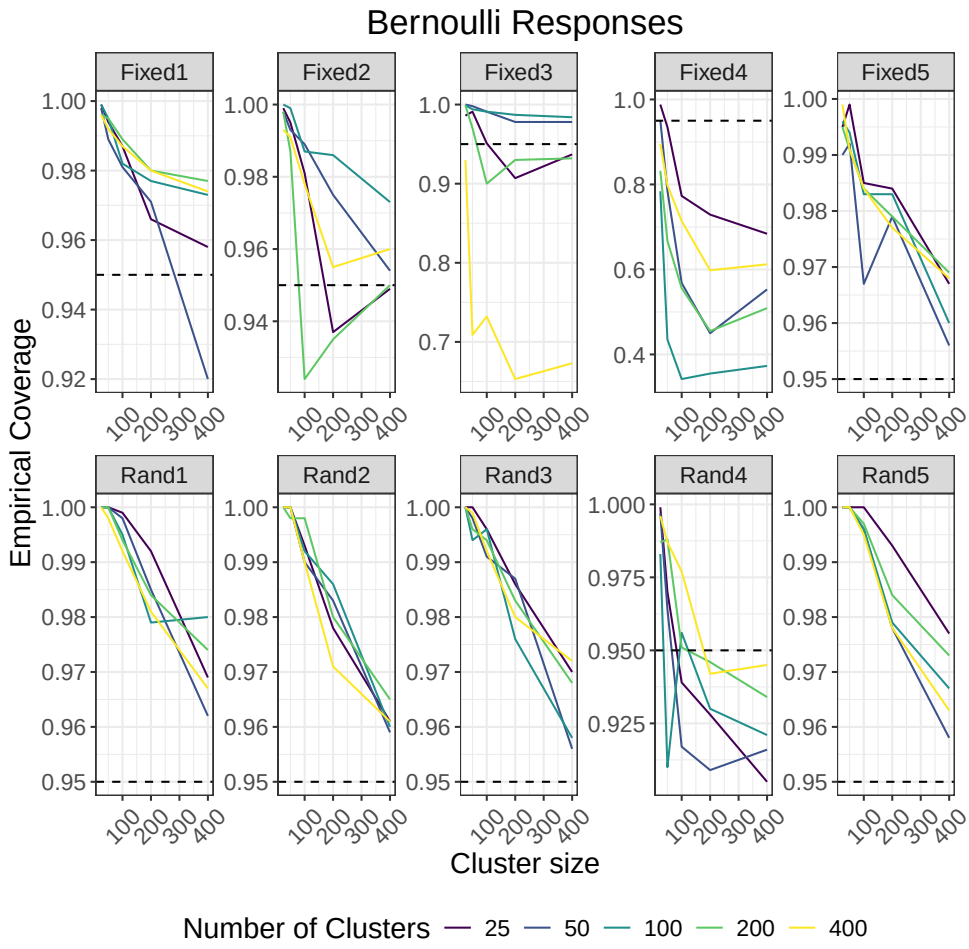

Figure 10: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Bernoulli responses.

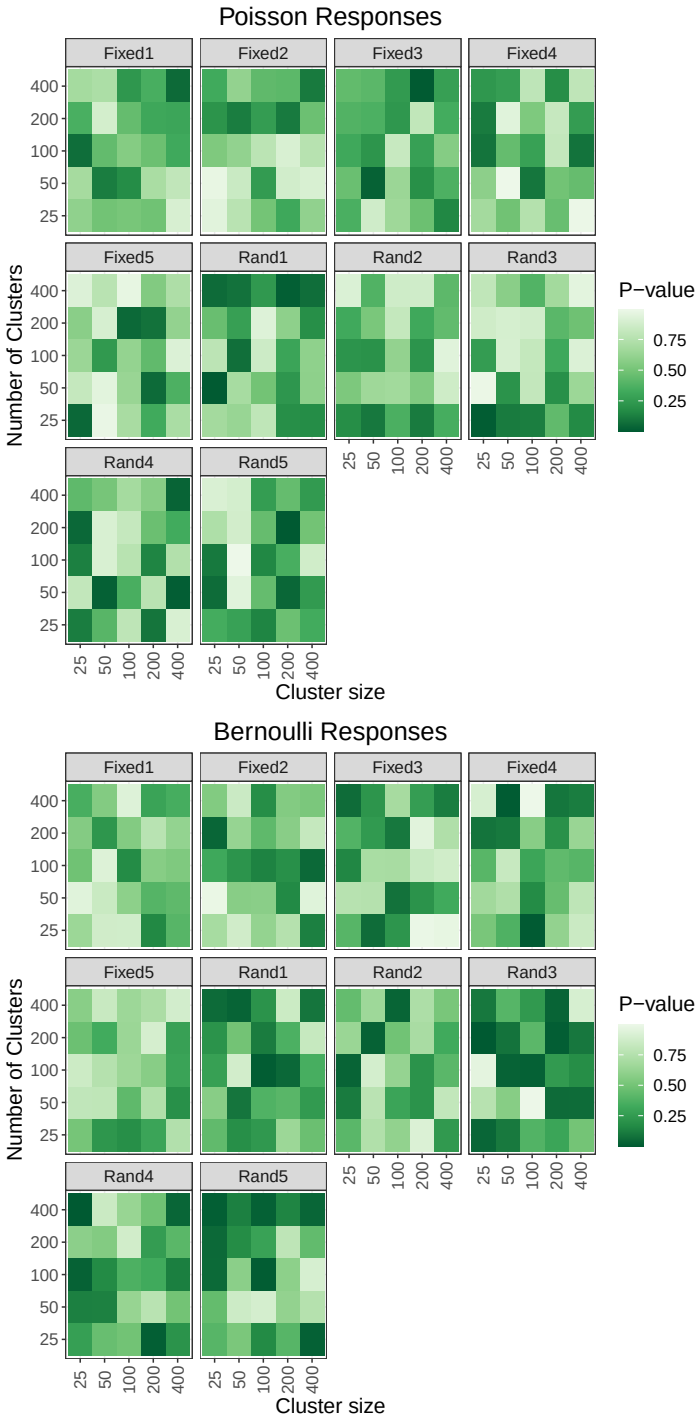

Figure 11: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the conditional regime.

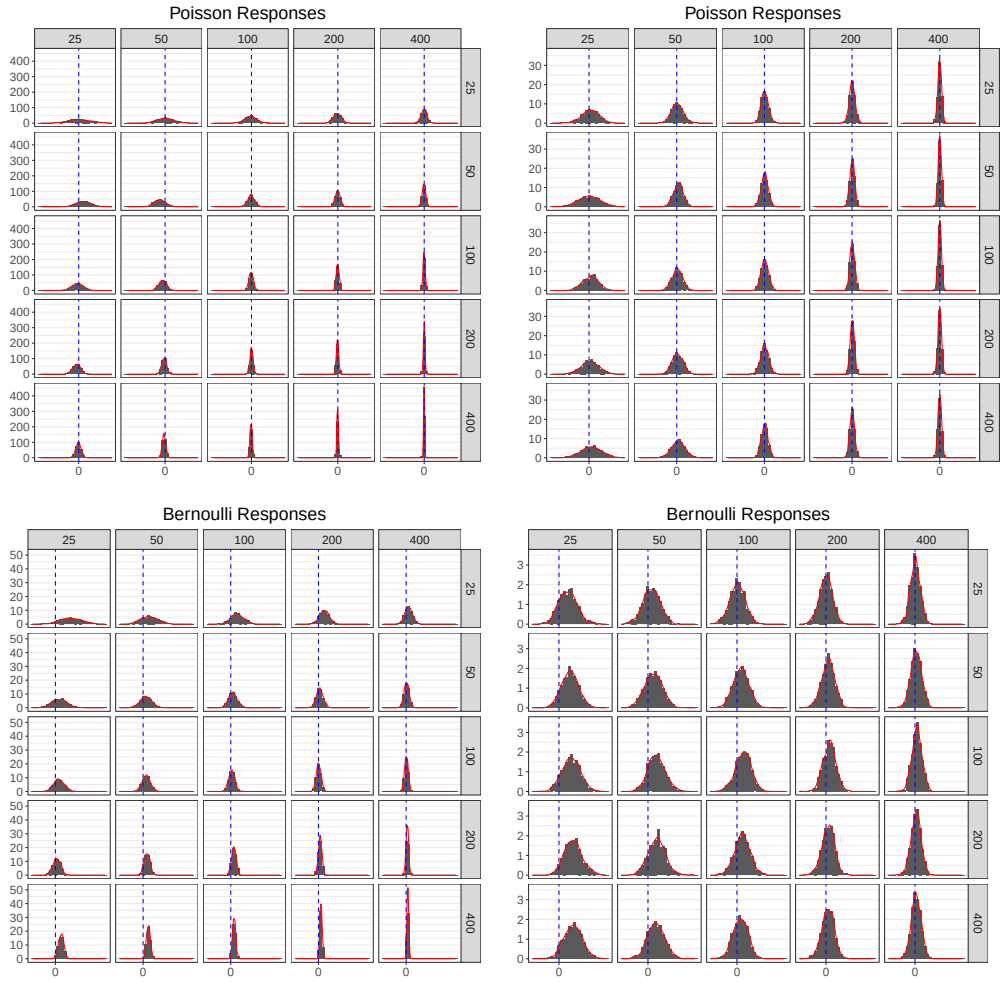

Figure 12: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

S5.4 $\hat{G} = 0.50\ I_2$

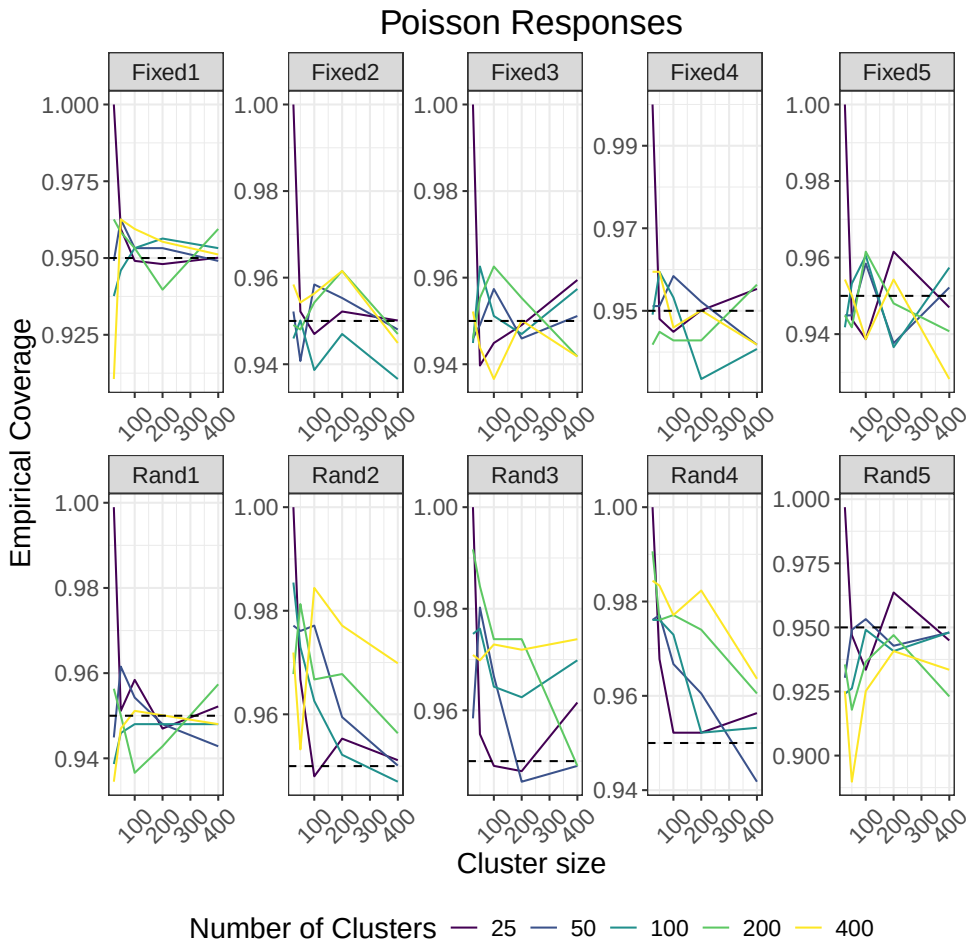

Figure 13: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Poisson responses.

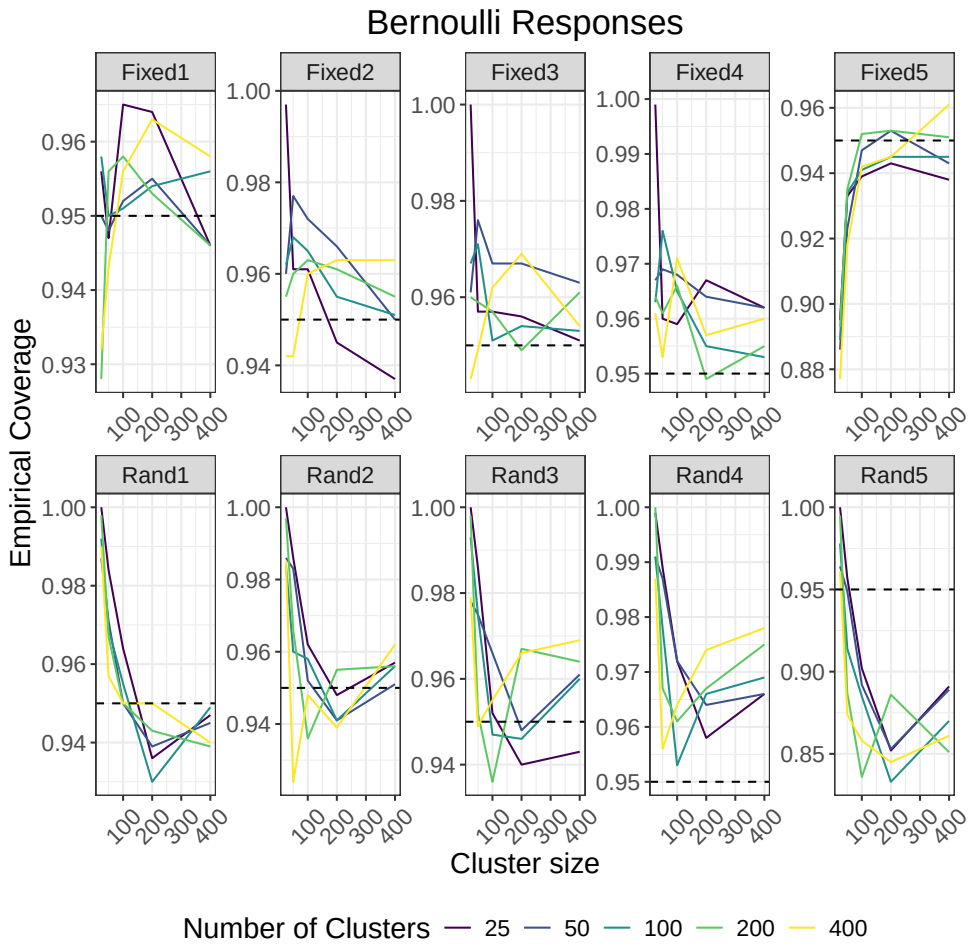

Figure 14: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Bernoulli responses.

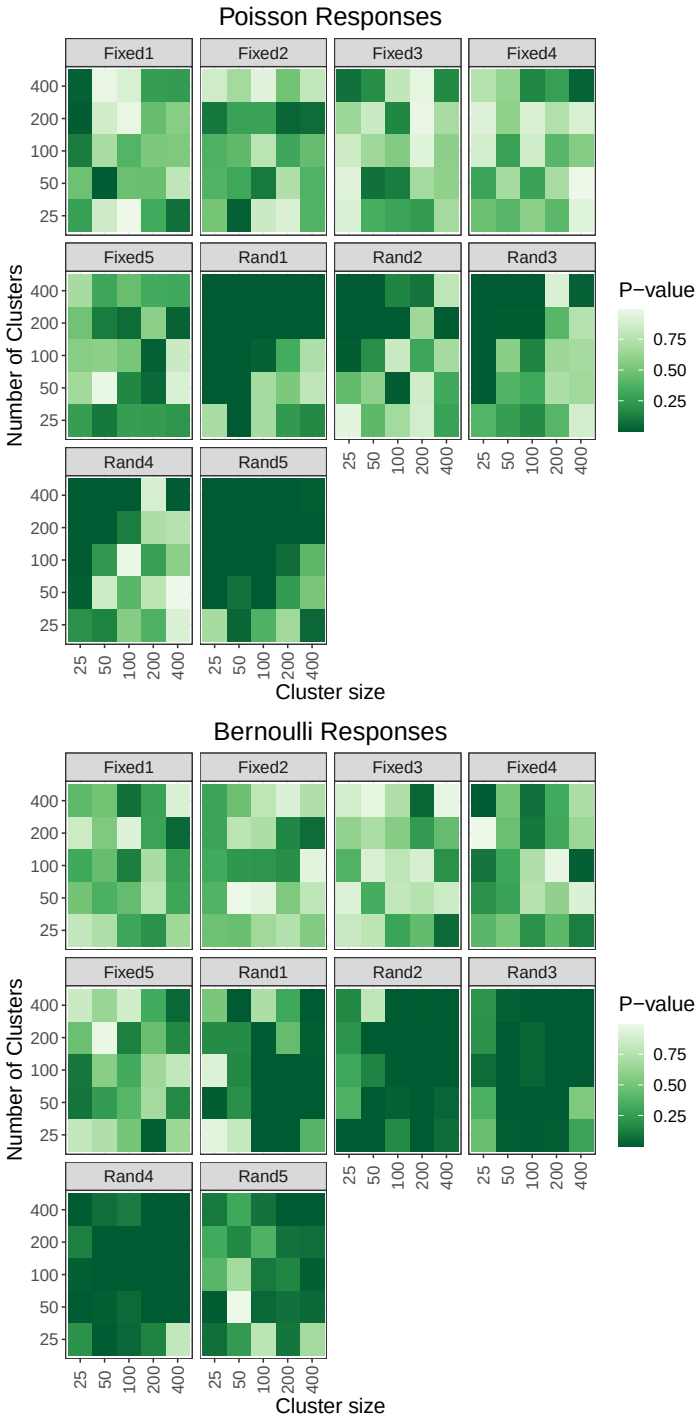

Figure 15: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the unconditional regime.

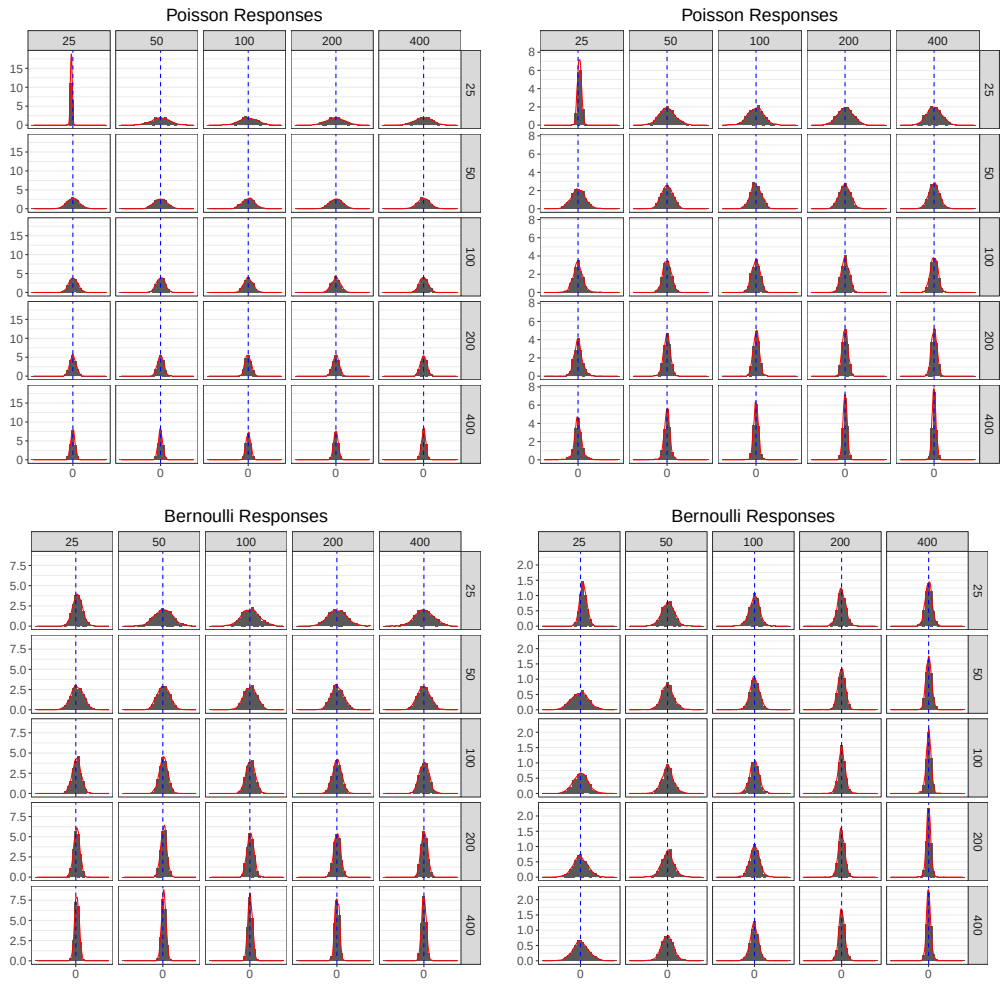

Figure 16: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

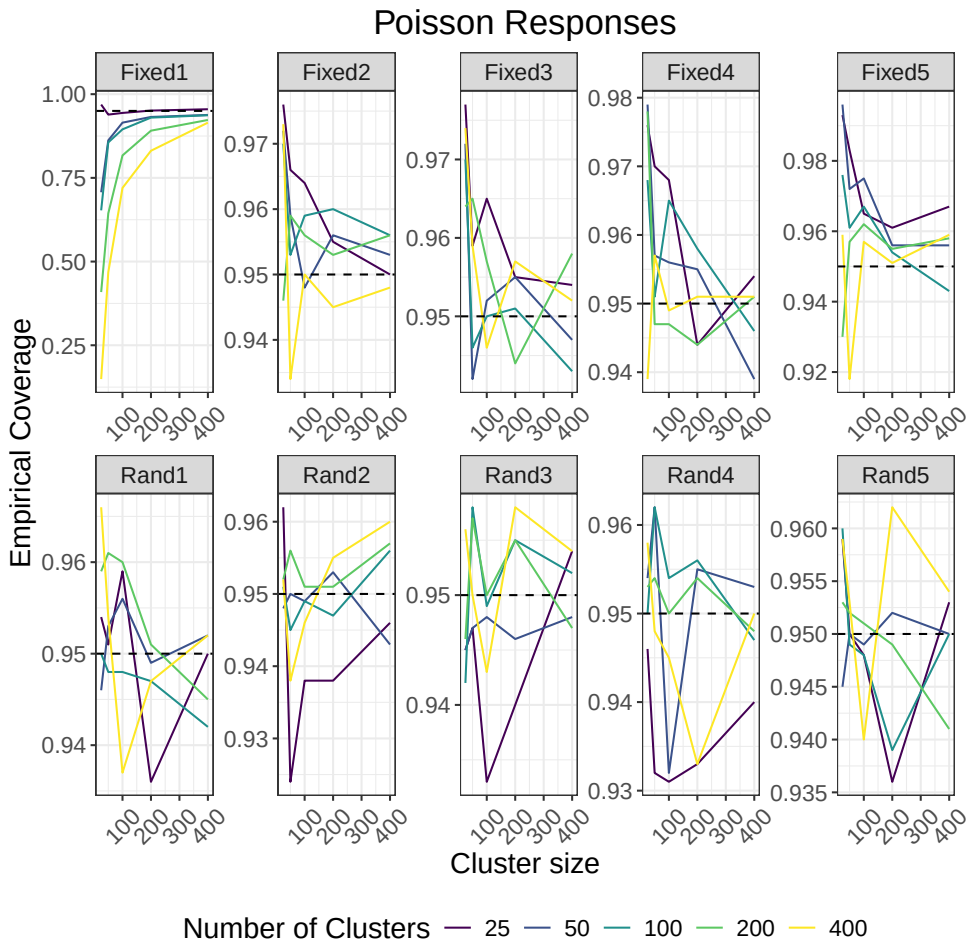

Figure 17: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Poisson responses.

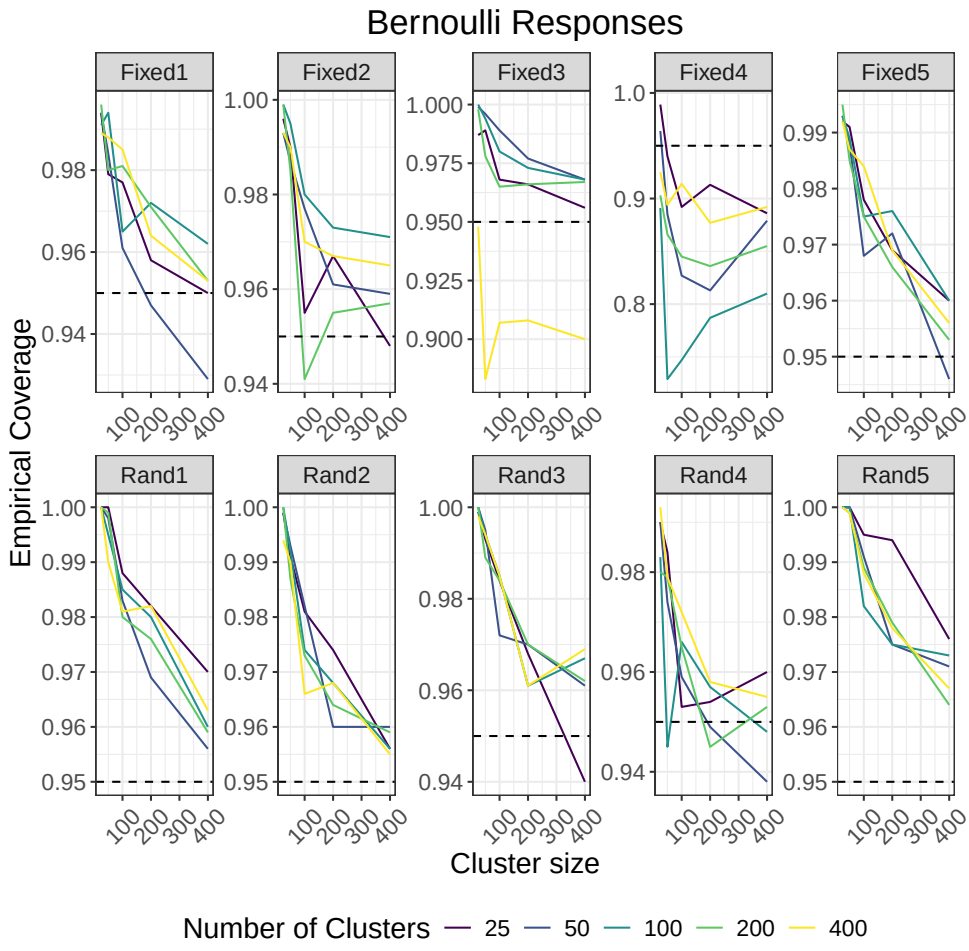

Figure 18: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Bernoulli responses.

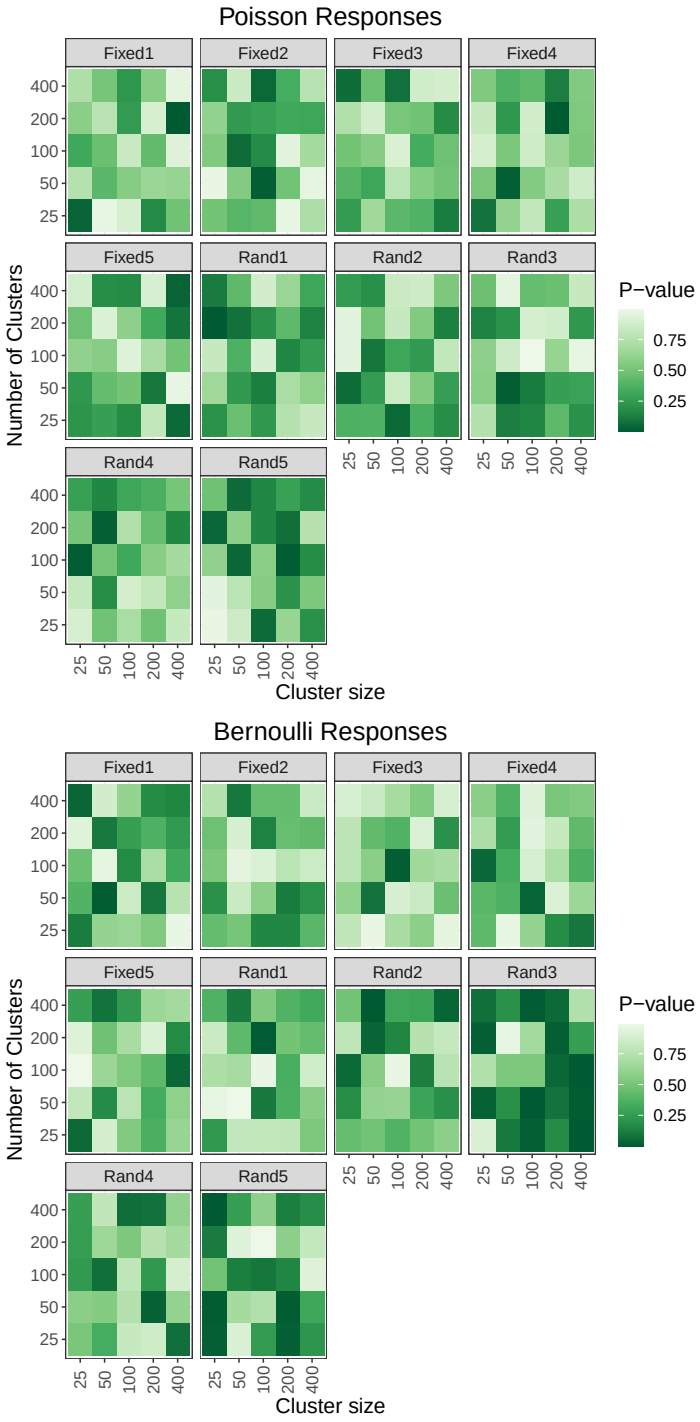

Figure 19: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the conditional regime.

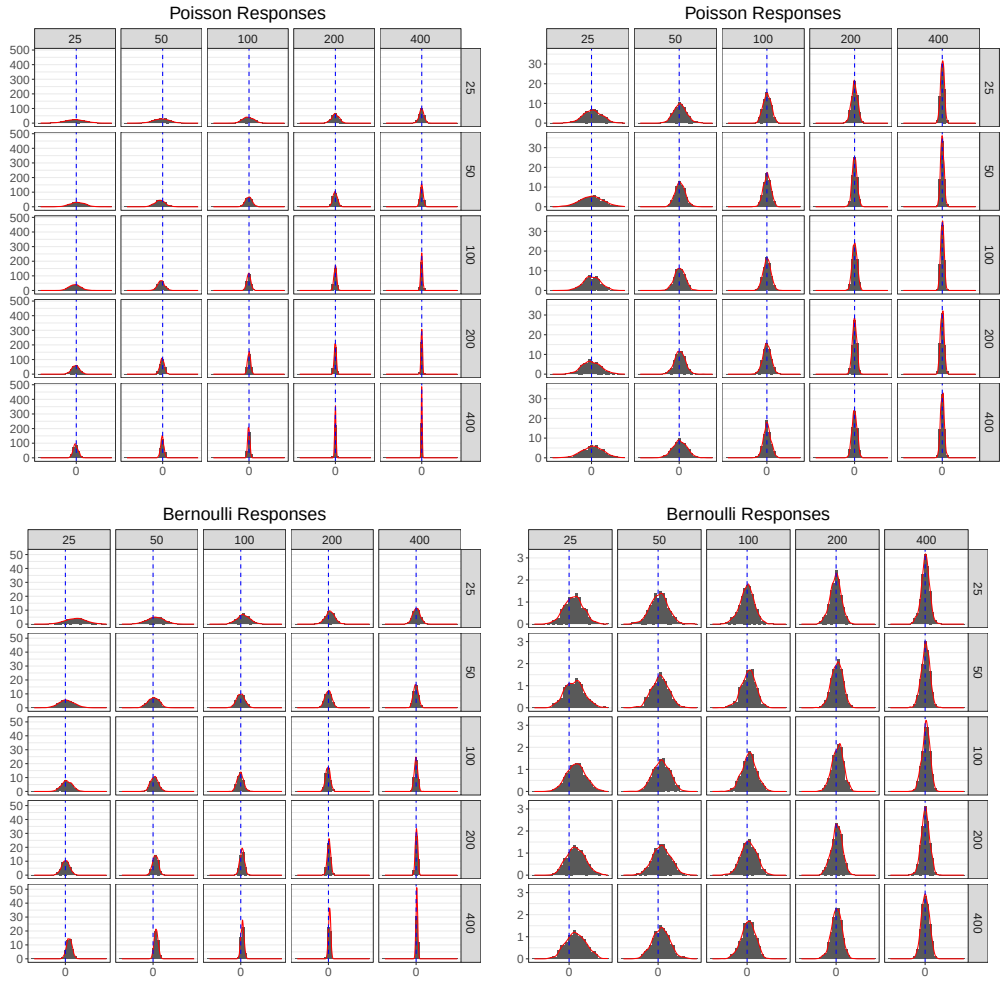

Figure 20: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

S5.5 $\hat{G} = I_2$

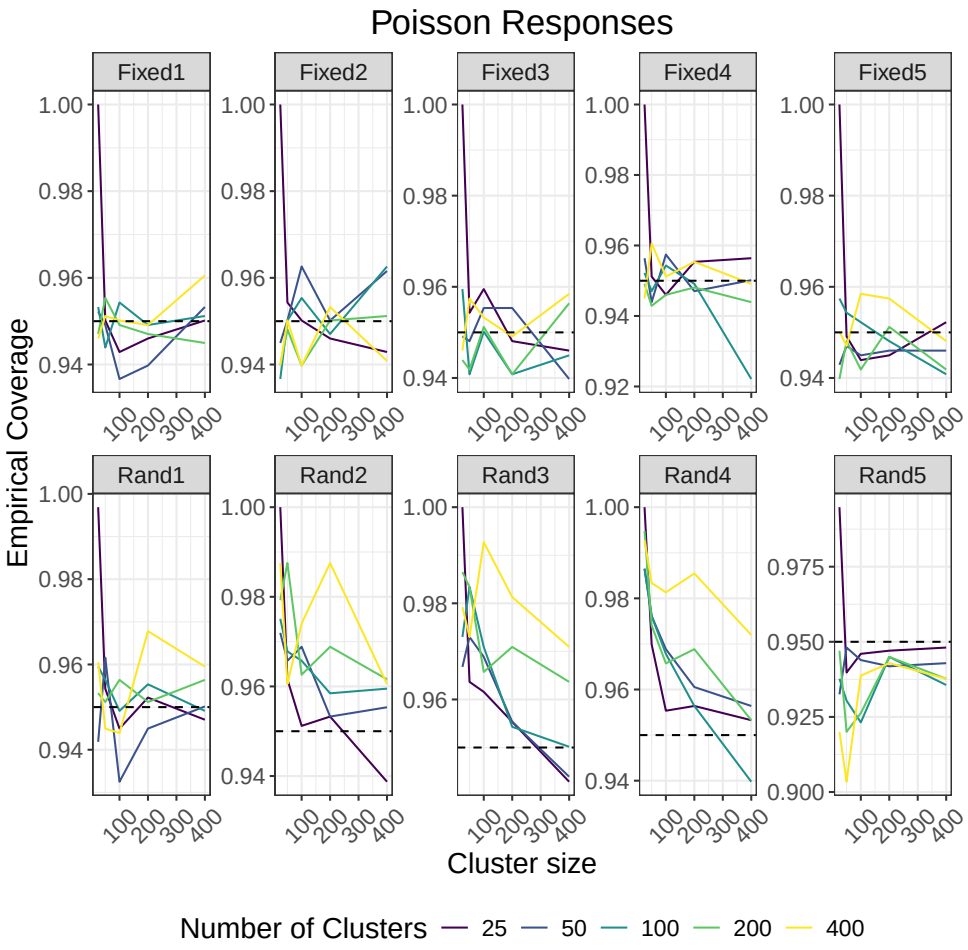

Figure 21: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Poisson responses.

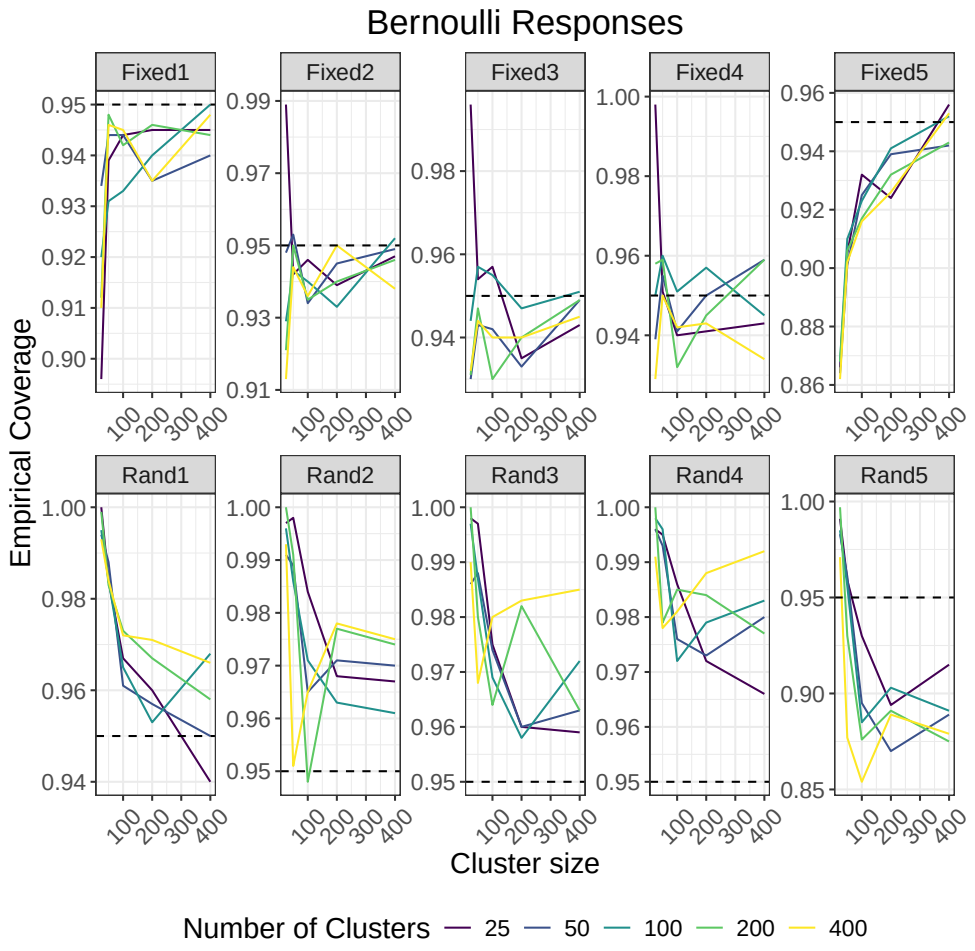

Figure 22: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Bernoulli responses.

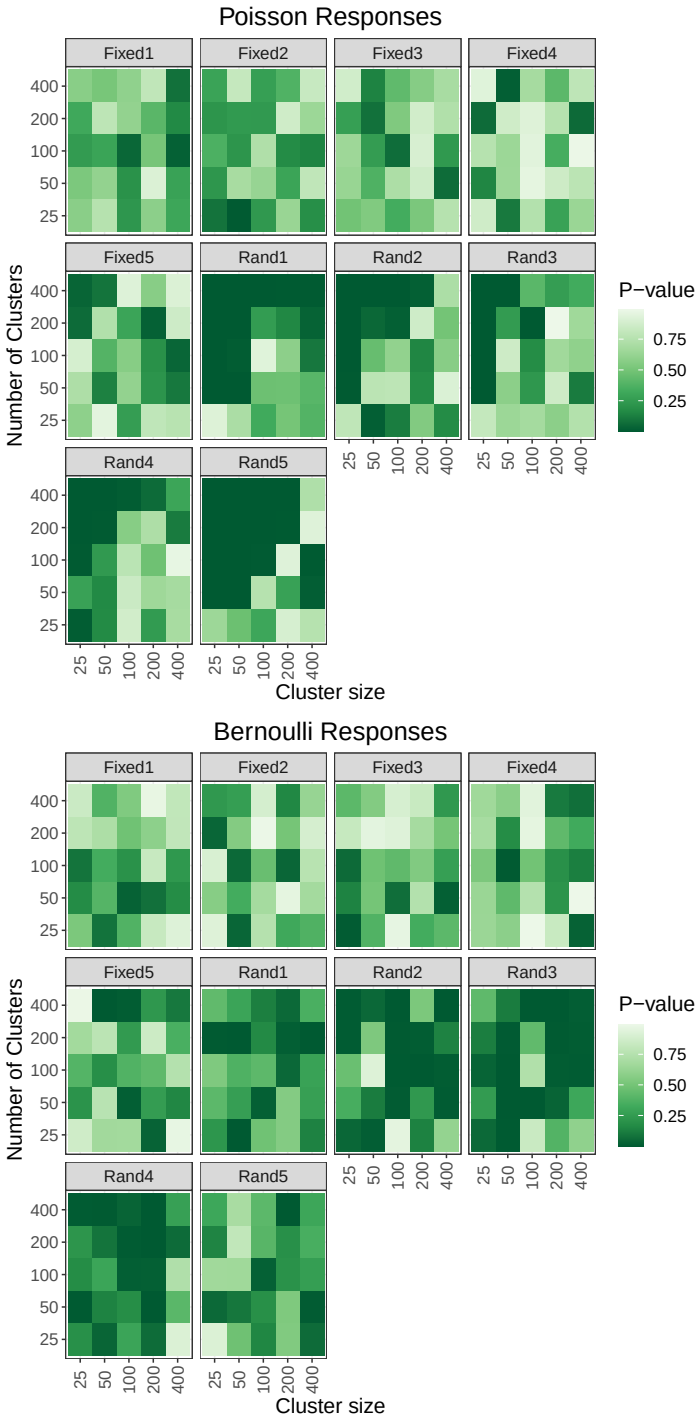

Figure 23: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the unconditional regime.

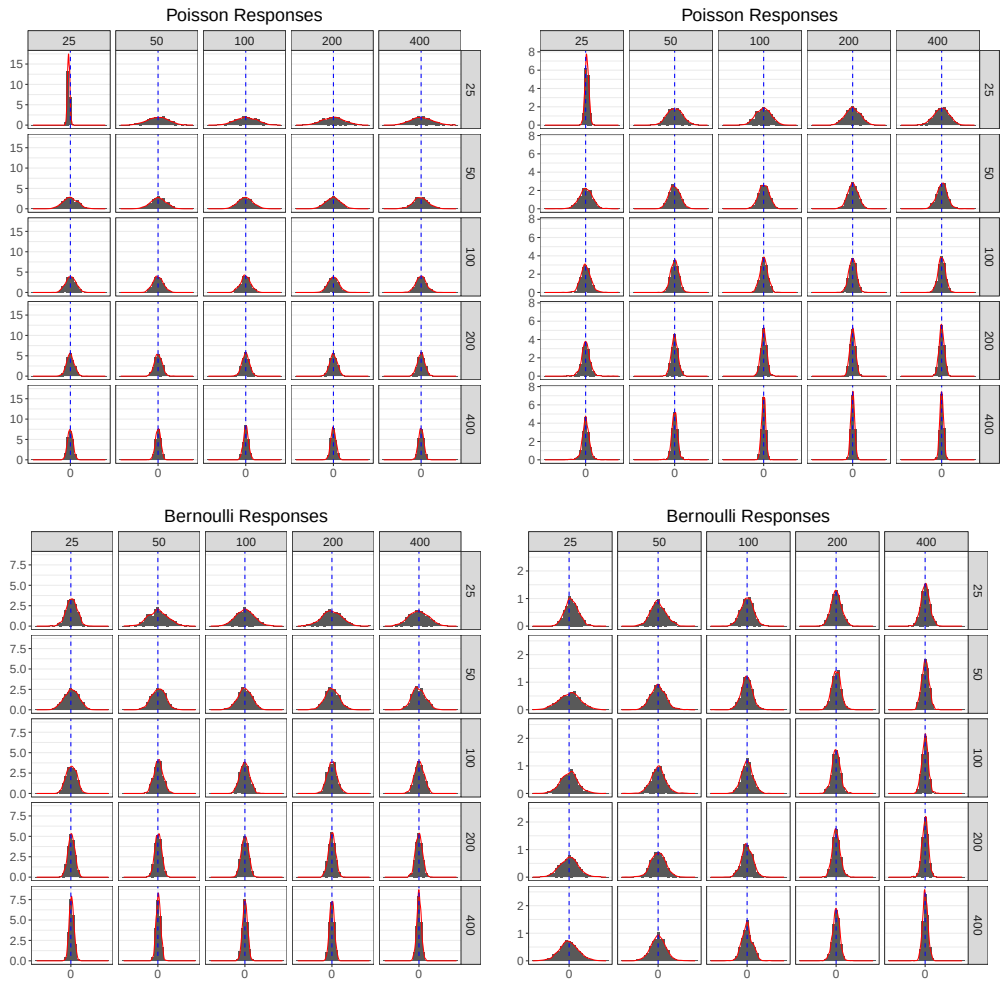

Figure 24: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

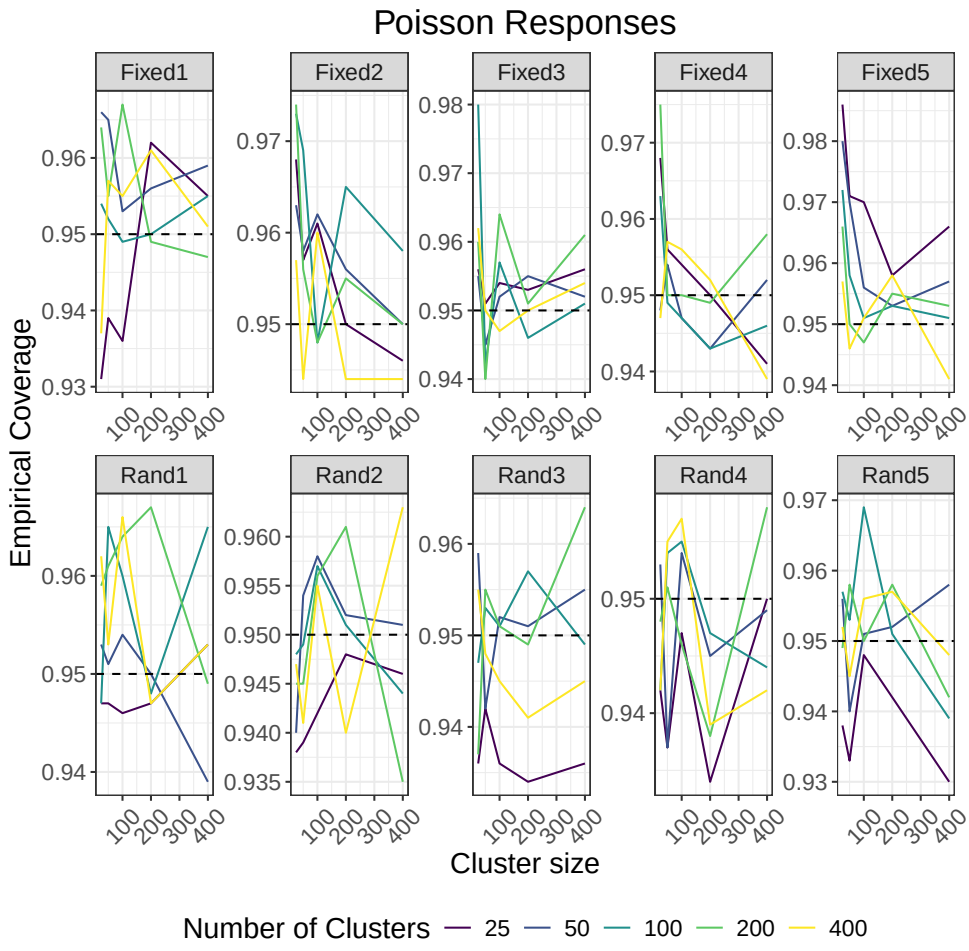

Figure 25: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Poisson responses.

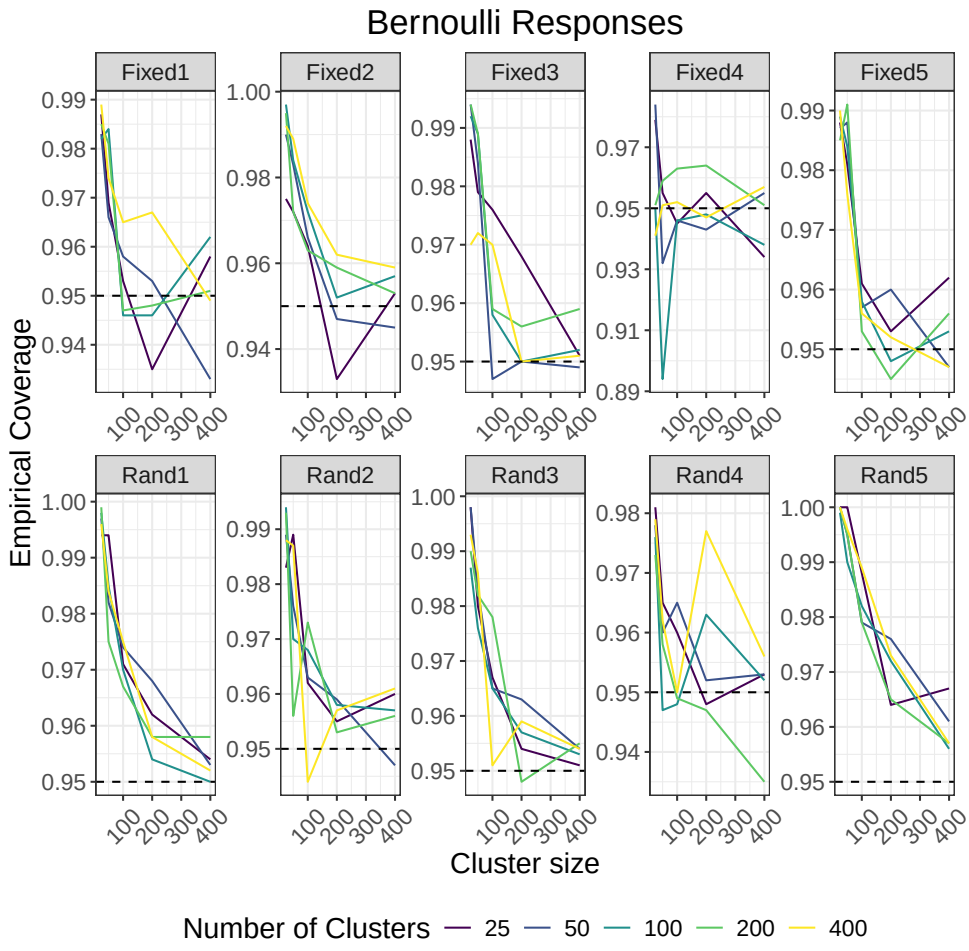

Figure 26: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Bernoulli responses.

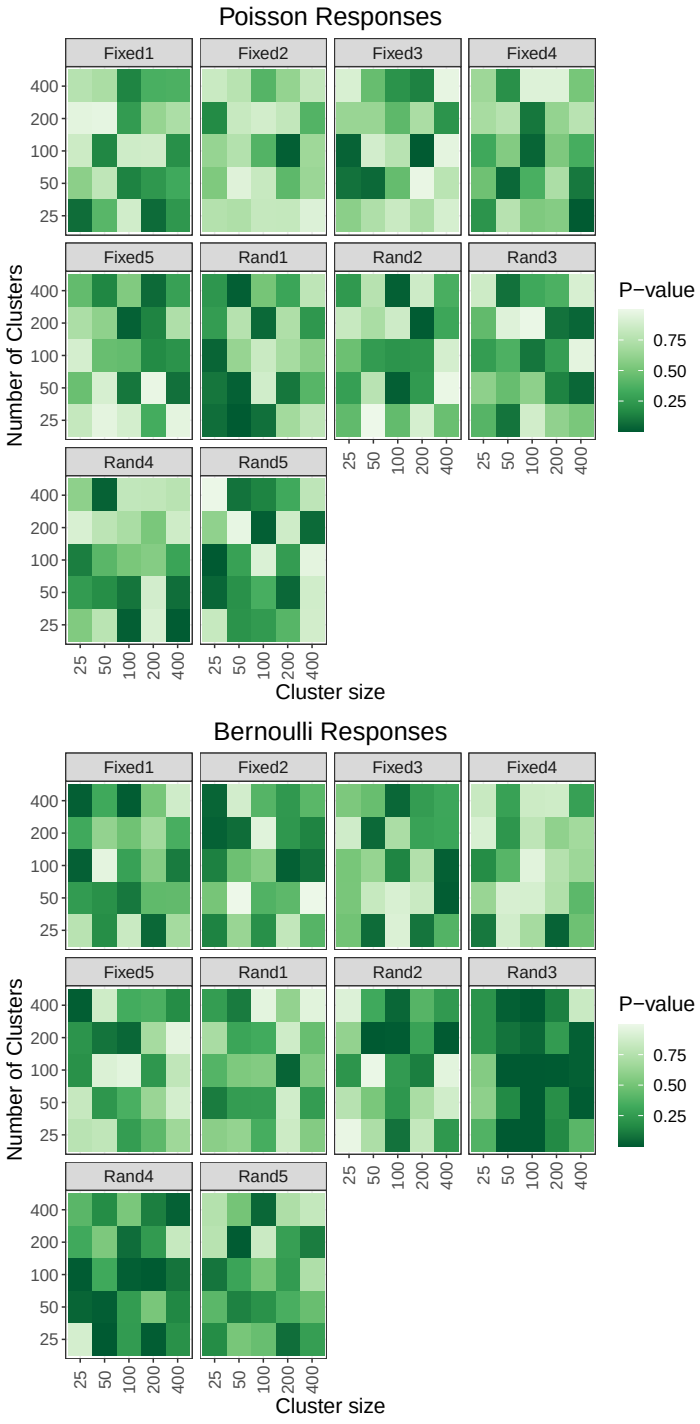

Figure 27: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the conditional regime.

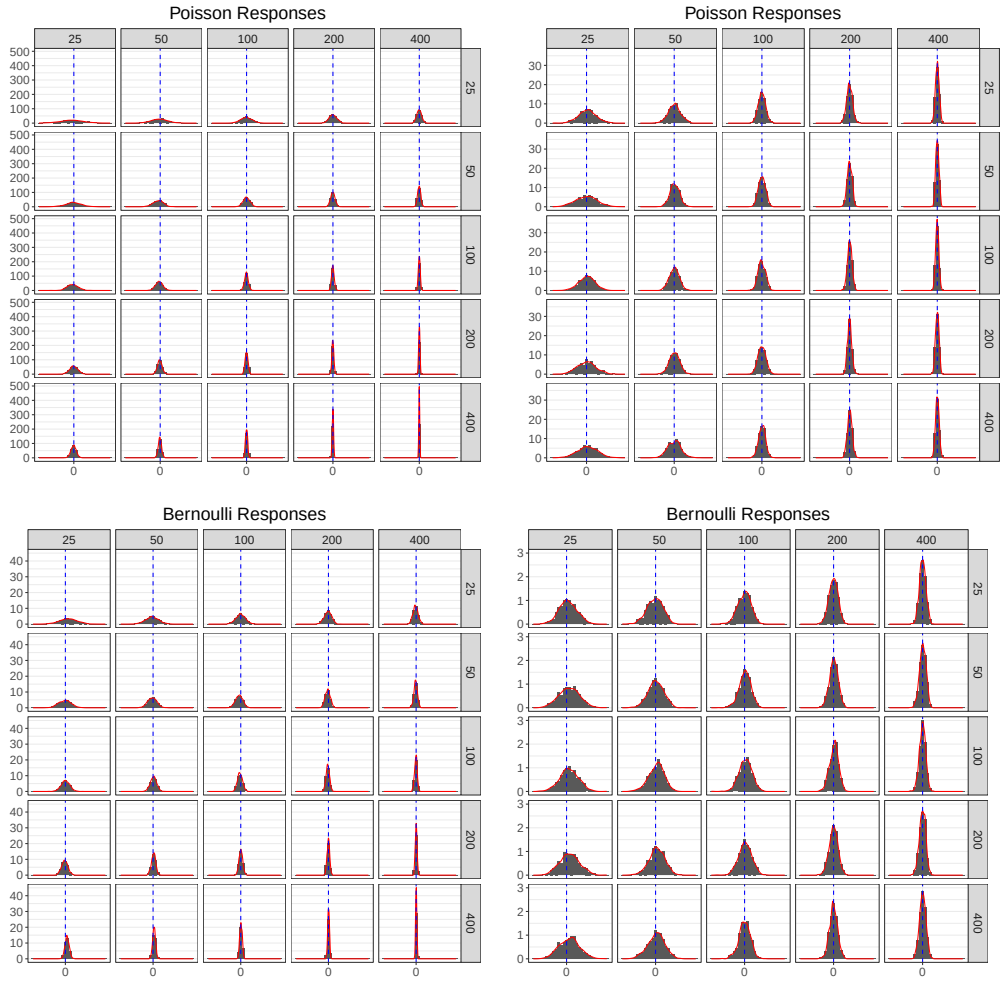

Figure 28: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

S5.6 $\hat{G} = 2\ I_2$

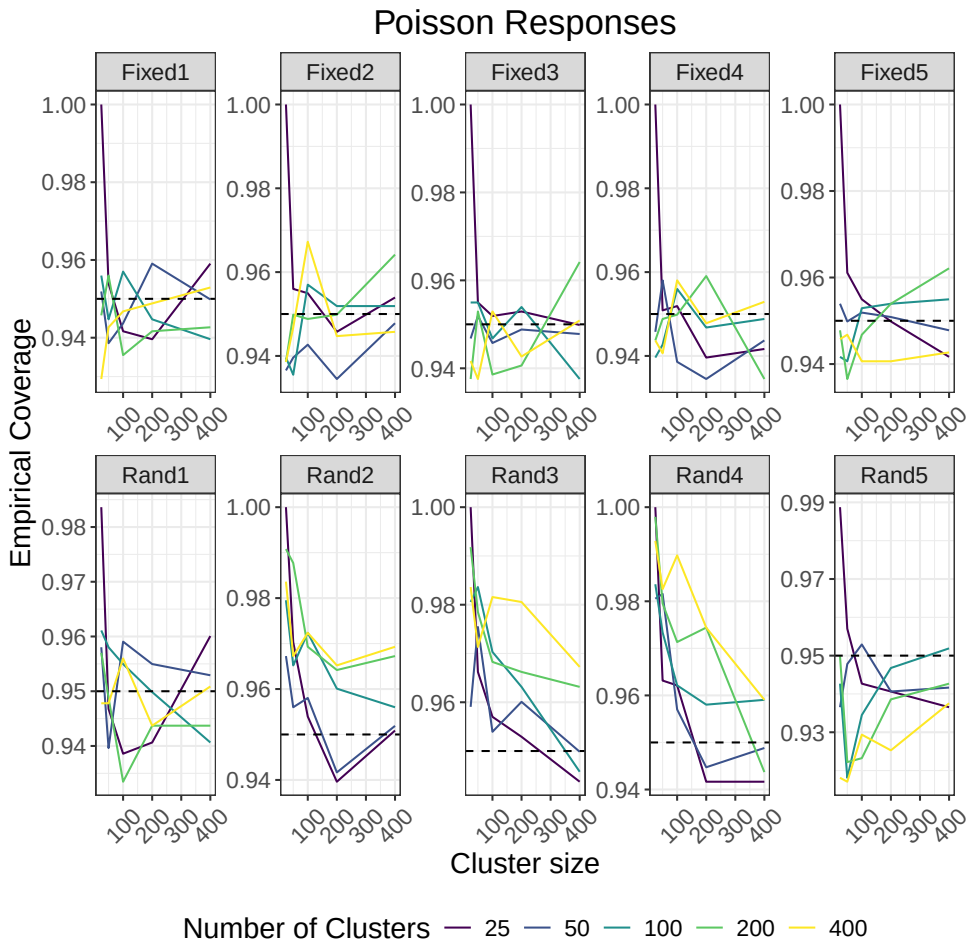

Figure 29: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Poisson responses.

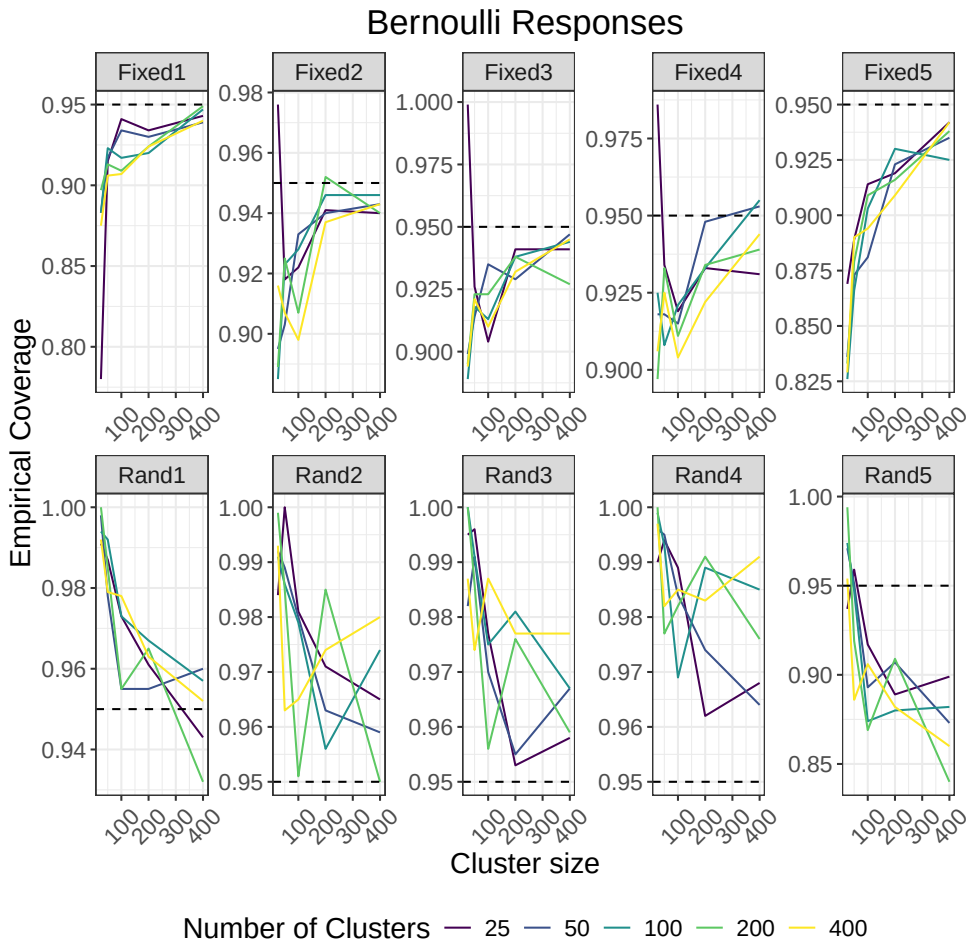

Figure 30: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Bernoulli responses.

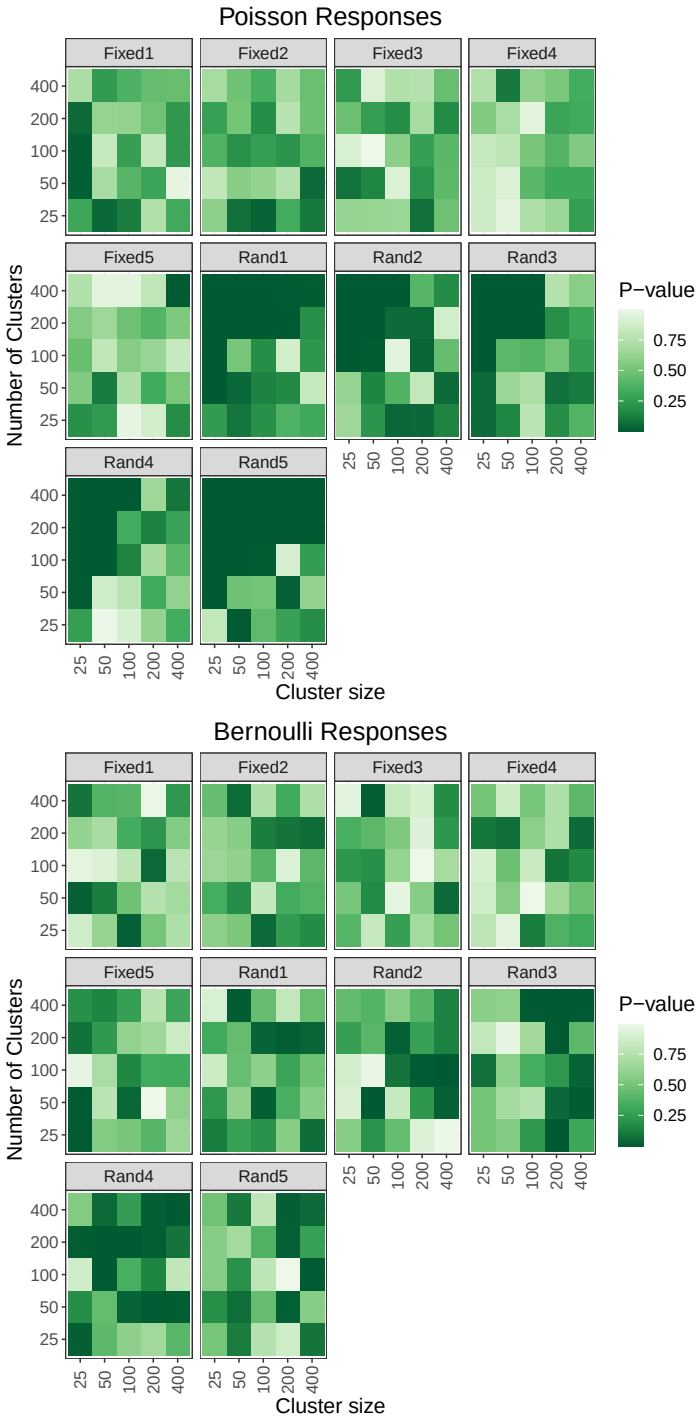

Figure 31: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the unconditional regime.

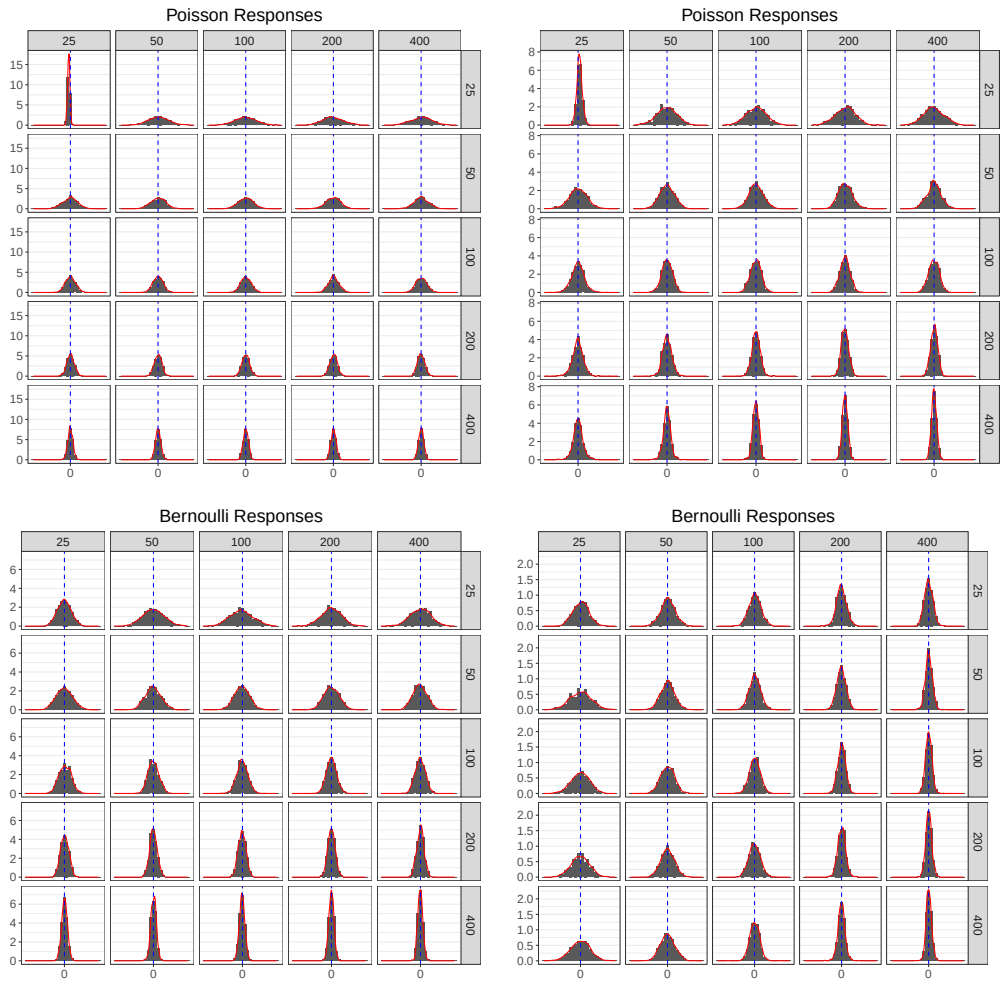

Figure 32: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

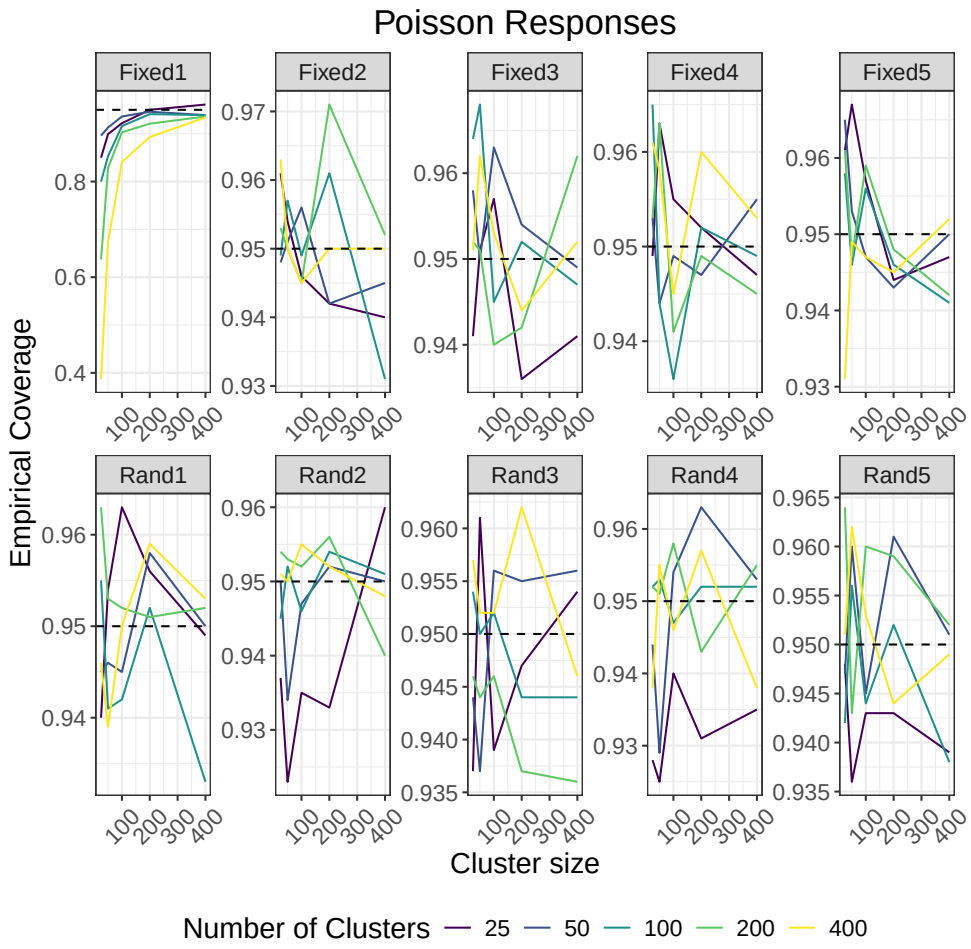

Figure 33: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Poisson responses.

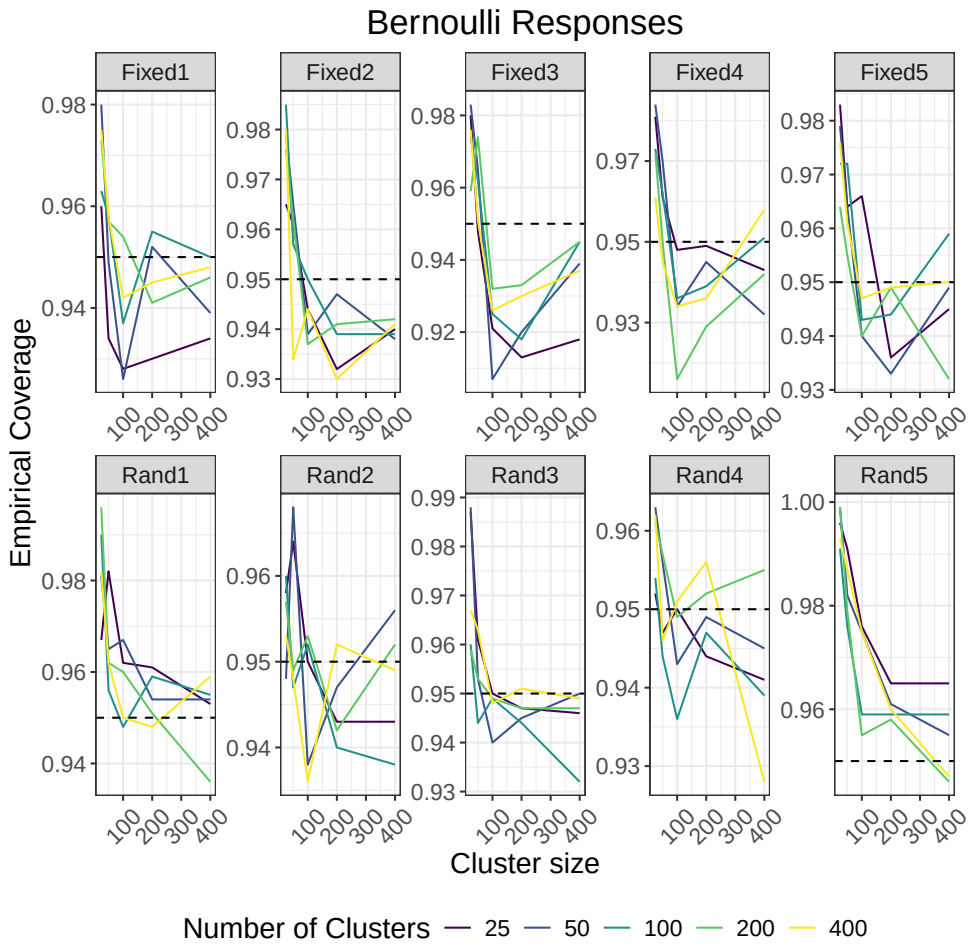

Figure 34: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Bernoulli responses.

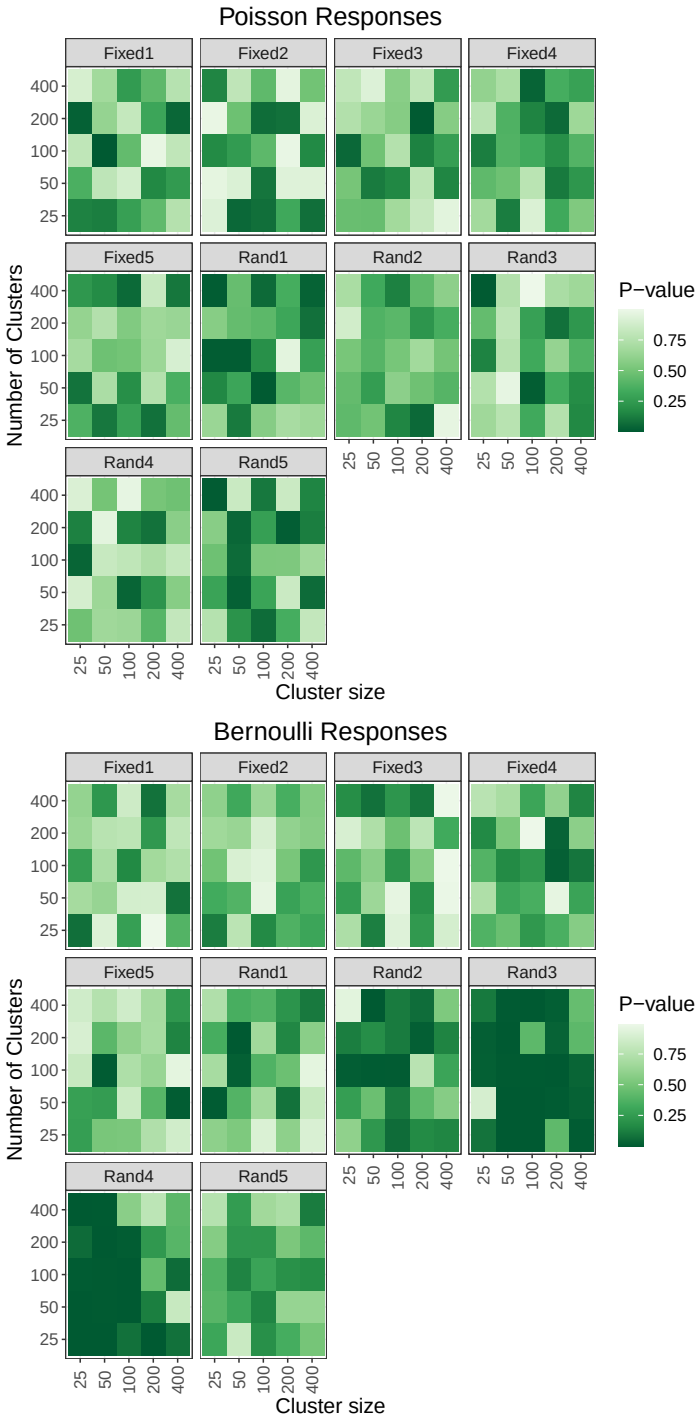

Figure 35: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the conditional regime.

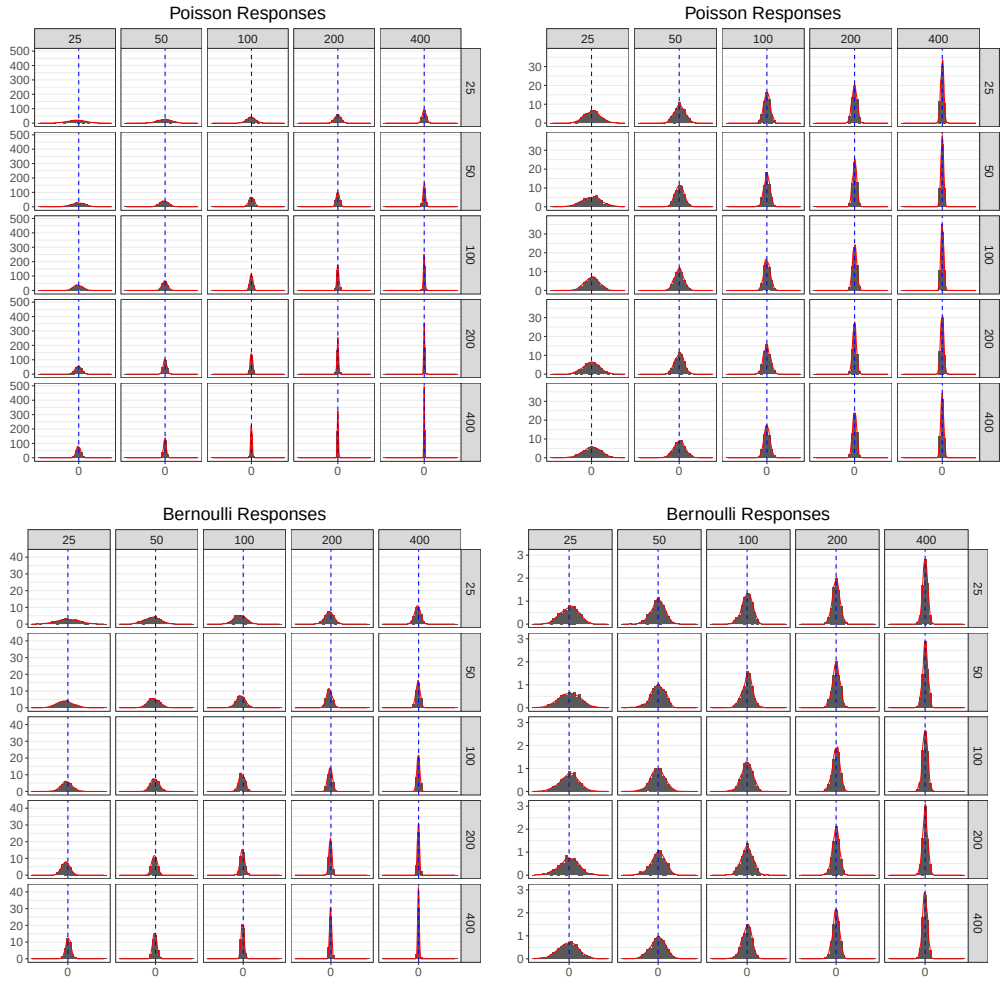

Figure 36: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

S5.7 $\hat{G} = 4\ I_2$

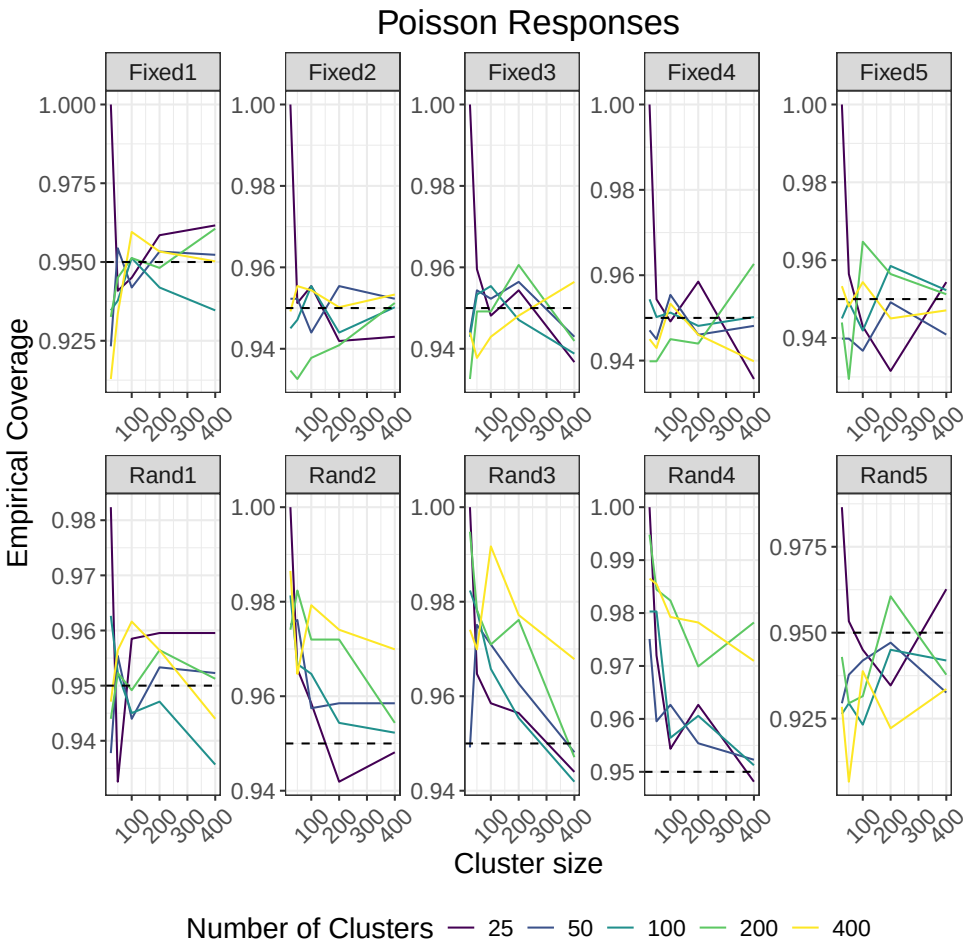

Figure 37: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Poisson responses.

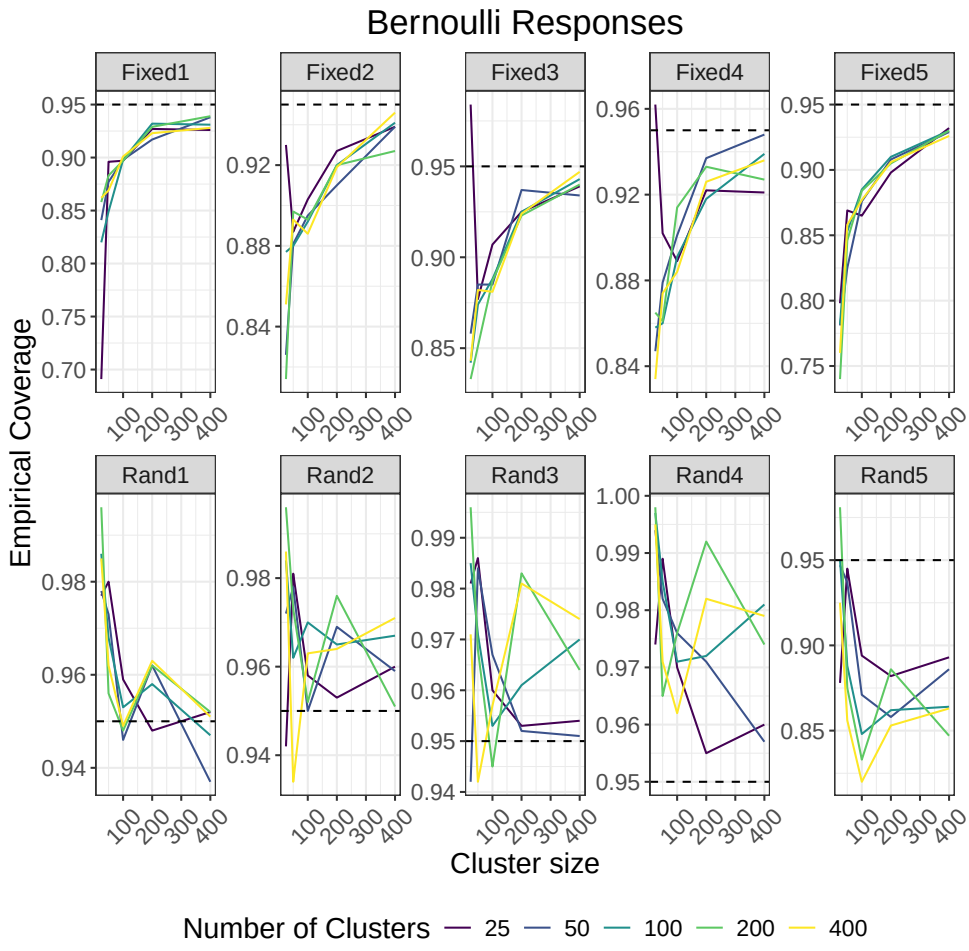

Figure 38: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the unconditional regime with Bernoulli responses.

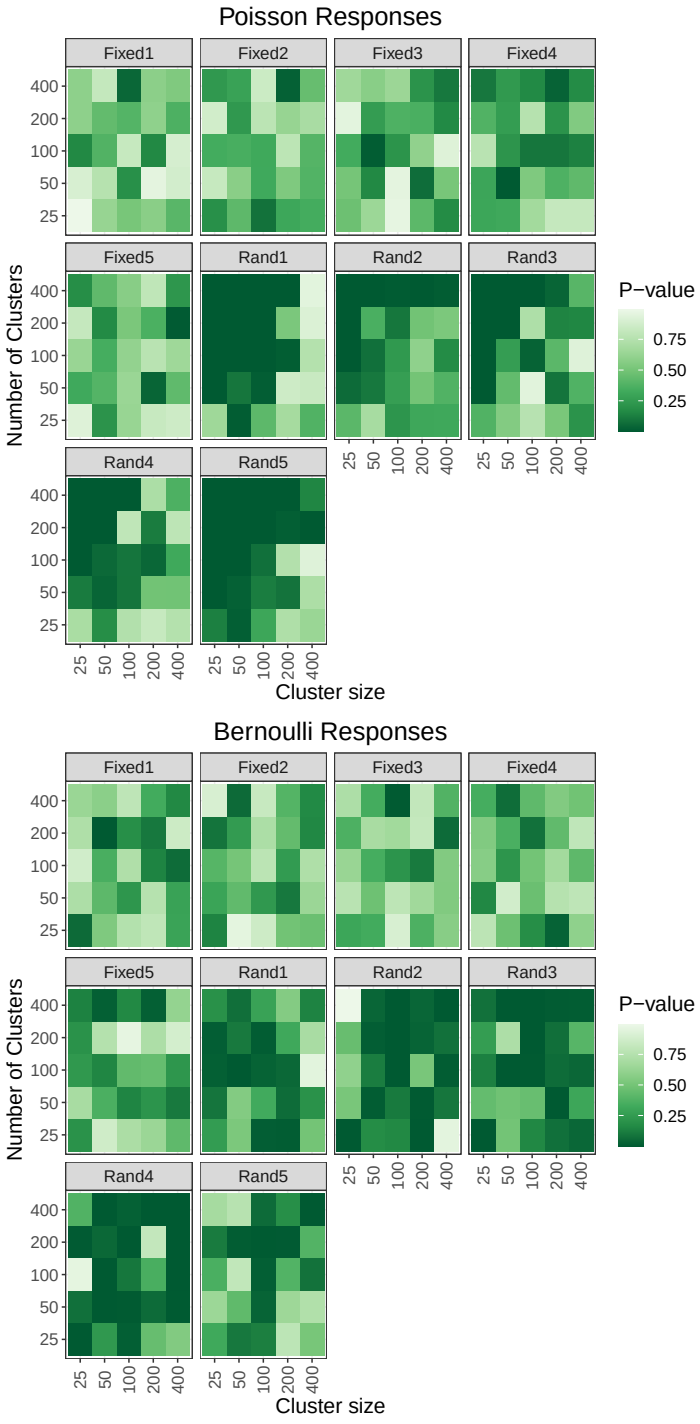

Figure 39: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the unconditional regime.

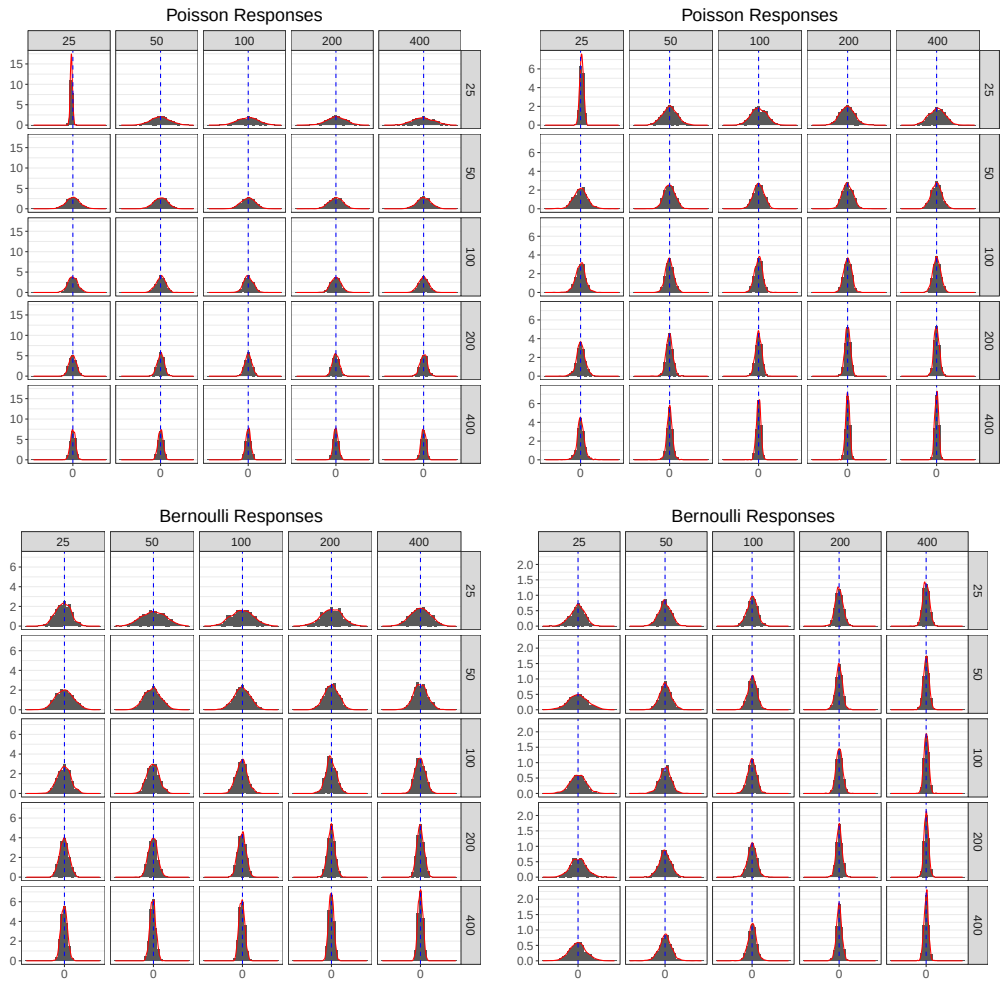

Figure 40: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

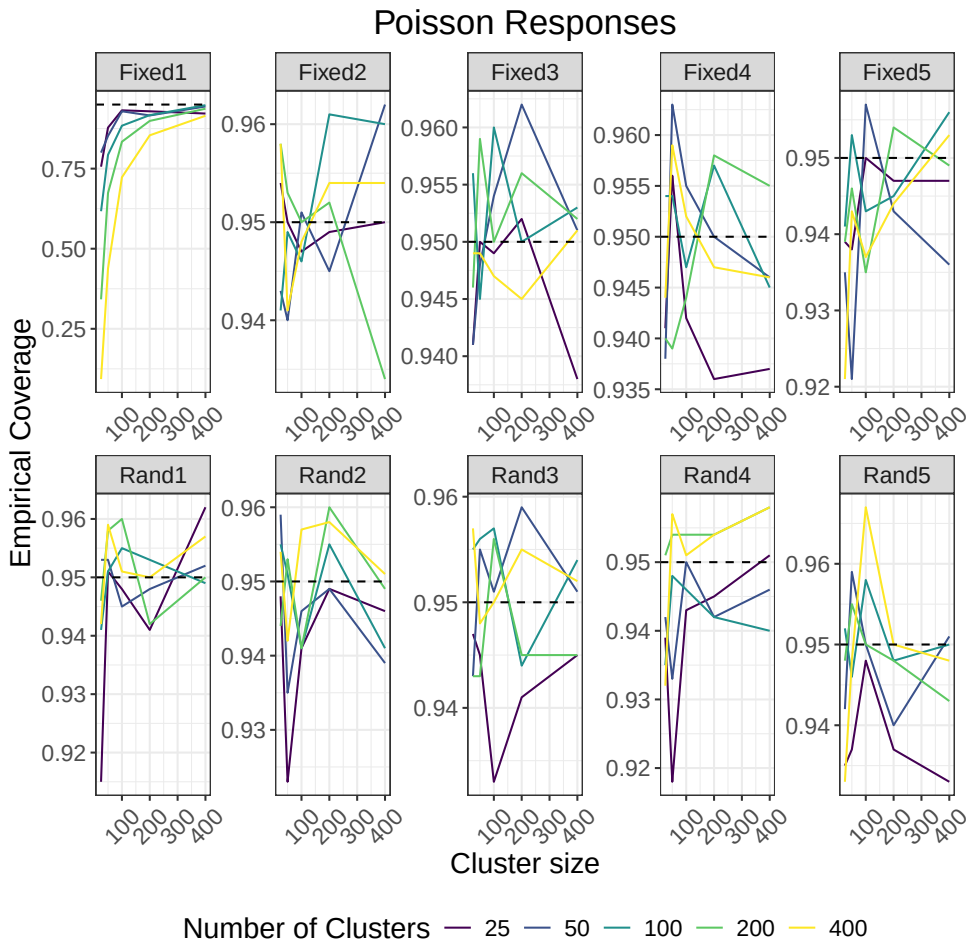

Figure 41: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Poisson responses.

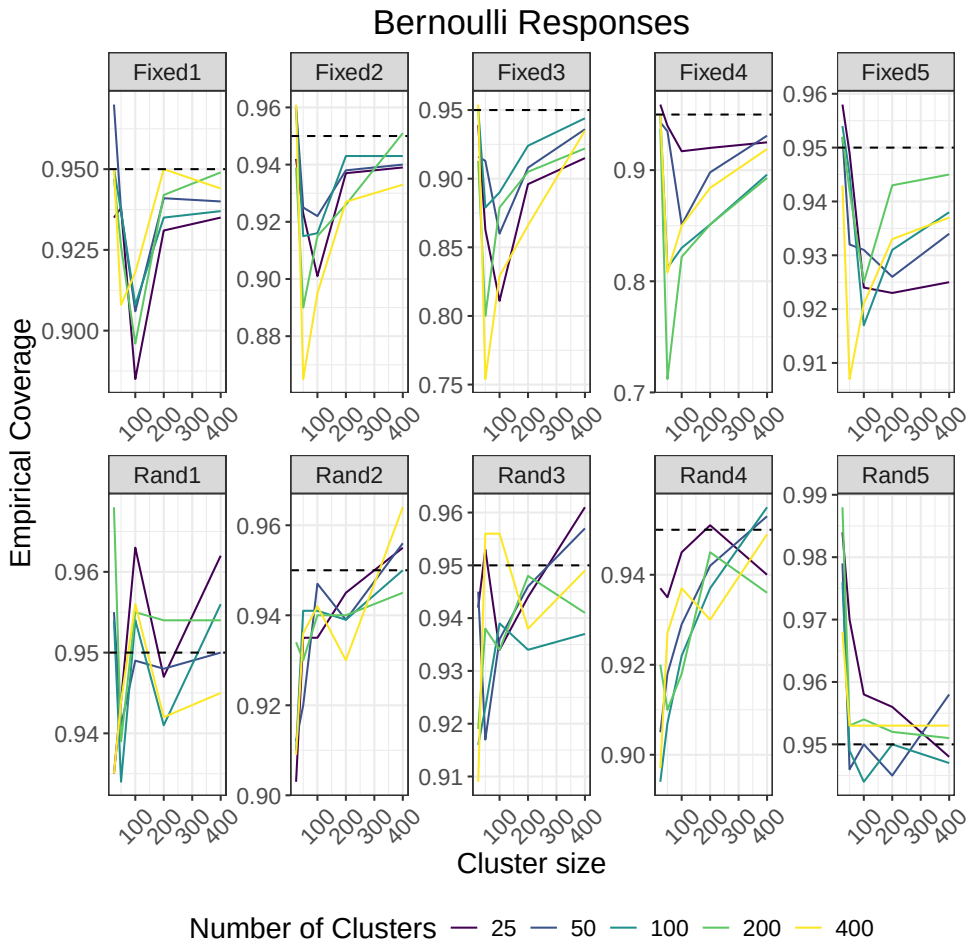

Figure 42: Empirical coverage probability of 95% coverage intervals for the five fixed and random effects estimates, obtained under the conditional regime with Bernoulli responses.

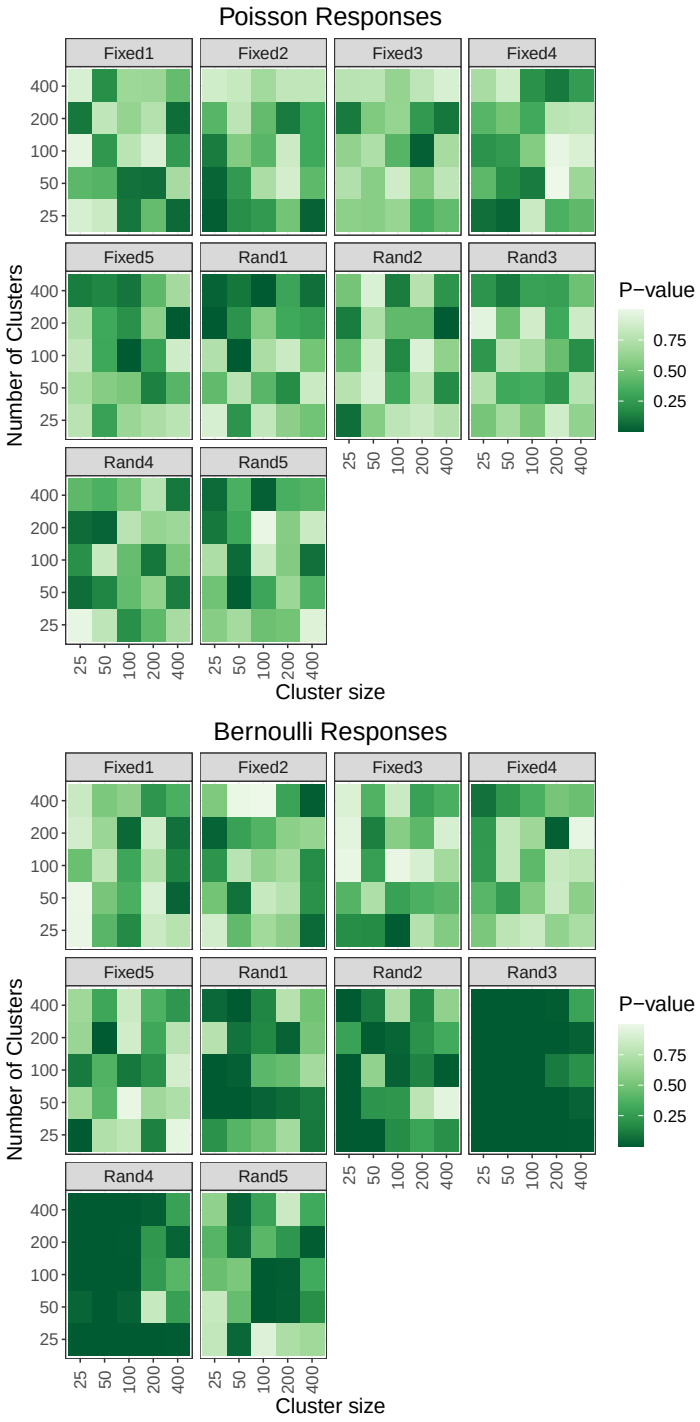

Figure 43: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the conditional regime.

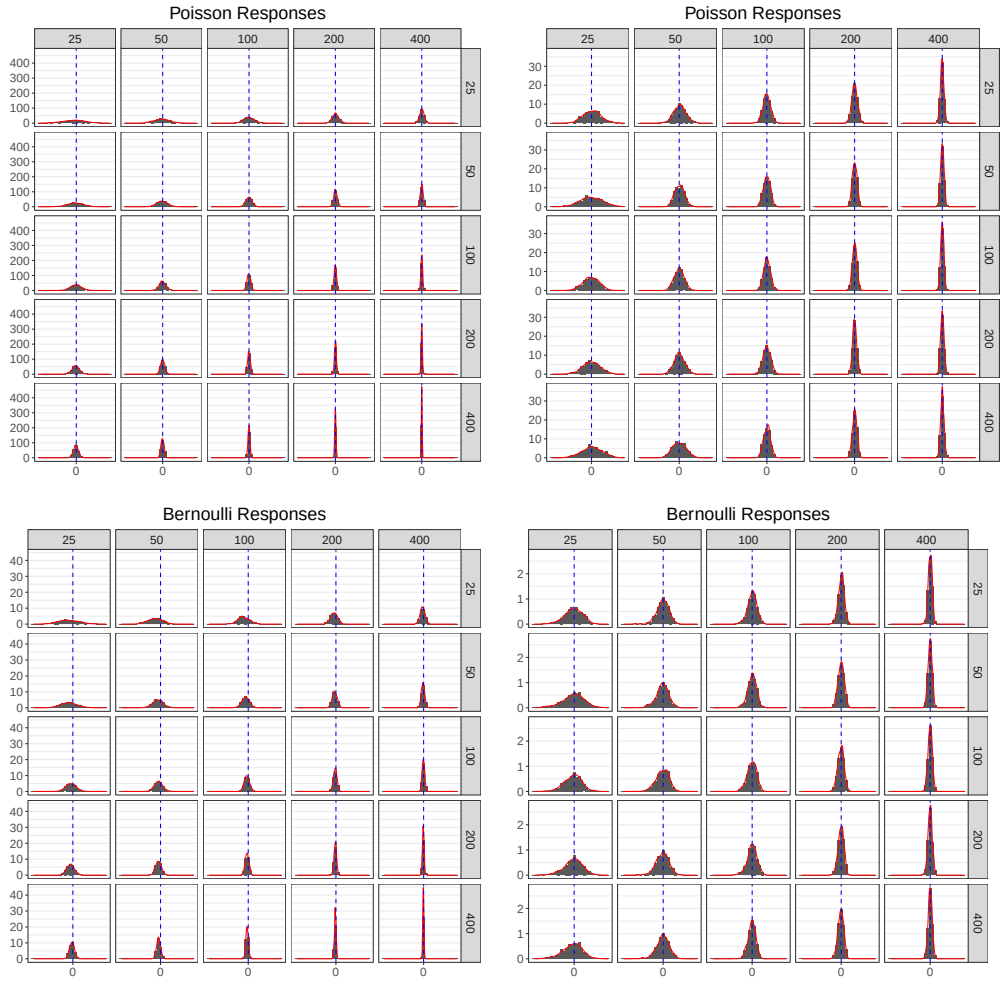

Figure 44: Histograms for the third components of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ (left panels) and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$ (right panels), under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted blue line indicates zero, and the red curve is a kernel density smoother.

Bibliography

Blum, J., H. Chernoff, M. Rosenblatt, and H. Teicher (1958). Central limit theorems for interchangeable processes. *Canadian Journal of Mathemat-*

ics 10, 222–229.

Downey, P. J. (1990). Distribution-free bounds on the expectation of the maximum with scheduling applications. *Operations Research Letters 9*, 189–201.

Lyu, Z. and A. H. Welsh (2021a). Asymptotics for EBLUPs: Nested error regression models. *Journal of the American Statistical Association 117*, 1–15.

Lyu, Z. and A. H. Welsh (2021b). Increasing cluster size asymptotics for nested error regression models. *Journal of Statistical Planning and Inference 217*, 52–68.