Usable and precise asymptotics for generalized linear mixed model analysis and design

Jiming Jiang,
University of California, Davis, U.S.A.

Matt P. Wand and Aishwarya Bhaskaran
University of Technology Sydney, Australia

Summary. We derive precise asymptotic results that are directly usable for confidence intervals and Wald hypothesis tests for likelihood-based generalized linear mixed model analysis. The essence of our approach is to derive the exact leading term behaviour of the Fisher information matrix when both the number of groups and number of observations within each group diverge. This leads to asymptotic normality results with simple studentizable forms. Similar analyses result in tractable leading term forms for the determination of approximate locally D-optimal designs.

Keywords. D-optimality, longitudinal data analysis, maximum likelihood estimation, studentization.

1. Introduction

We derive simple and usable theorems concerning the exact leading term behaviour of likelihood-based generalized linear mixed model estimators and D-optimality criteria. The theorems allow for straightforward construction of asymptotically valid confidence intervals, Wald hypothesis tests and locally D-optimal designs. The main theorem concerns the joint asymptotic normality of all model parameters and elegantly shows faster rates of convergence of fixed effects that are not accompanied by a random effect compared with fixed effects that have a partnering random effect. Maximum likelihood estimation of random effects covariance matrix parameters are also included in our results.

Since the early 1990s generalized linear mixed models have been a mainstay of regression-type statistical analyses in areas such as longitudinal data analysis, multilevel modelling, panel data analysis and small area estimation. Overviews of generalized linear mixed models, and access to their large literature that includes methodology, theoretical results and software, are provided by books such as Faraway (2016), Jiang (2017), McCulloch et al. (2008) and Stroup (2013). Both frequentist and Bayesian approaches are common throughout the generalized linear mixed models literature. In this article we focus on maximum likelihood estimation and frequentist inference. Our results allow for the quasi-likelihood extension for which a dispersion parameter is present.

Despite the large volume of research concerning generalized linear mixed models, there is very little theory concerning the statistical properties of maximum likelihood estimators. Nie (2007) contains some asymptotic normality results for the setting considered in this article, but they do not have a ready-to-use form for practical tasks such as obtaining studentized confidence intervals and optimal design determination. For example, Theorem

3 of Nie (2007) is such that the asymptotic covariance matrix of the fixed effects involves limits and expectations over the response distribution. In contrast, our Theorem 1 in Section 3 is such that the asymptotic covariance matrix is devoid of limits and involves expectation over the random effects distribution. As we explain in Section 4, studentization based on Theorem 1 is often quite simple for the construction of confidence intervals and carrying out Wald hypothesis tests. For the design setting, Theorem 2 in Section 5 facilitates approximate locally D-optimal design determination in a simpler and more direct manner compared with proposals given in, for example, Waite and Woods (2015) and Zhang et al. (2017).

Key aspects of our approach to obtaining precise asymptotics for generalized linear mixed models include: allowing both the number of groups and the within-group sample sizes to diverge, use of multi-term Laplace's method expansions for the ratios of sample size-dependent integrals as given in Tierney et al. (1989) and Miyata (2004), working with population limits of predictor-dependent sample mean quantities and establishing matrix norm asymptotic negligibility between matrix square roots of asymptotic inverse Fisher information matrices and simpler block diagonal forms.

Section 2 lays out the class of generalized linear mixed models treated in our theoretical analysis. The main result, Theorem 1, concerning usable asymptotic normality of maximum likelihood estimators, is given in Section 3. In Section 4 we explain how Theorem 1 aids practical and asymptotically valid statistical inference in generalized linear mixed model analysis and investigate finite sample performance via a simulation study. Section 5 is concerned with the design consequences of our theory, and evidence of practical D-optimal design construction is provided. In Section 6 we briefly discuss ramifications for the Gaussian variational approximation approach to generalized linear mixed model analysis in light of our main theorem. Some concluding remarks are given in Section 7. The proofs are in appendices.

2. Model description and maximum likelihood estimation

Consider the class of one-parameter exponential family of density, or probability mass, functions with generic form

$$p(y;\eta) = \exp\{y\eta - b(\eta) + c(y)\}\, h(y) \tag{1}$$

where η is the *natural parameter*. Common examples include the Bernoulli probability mass function for which $b(x) = \log(1+e^x)$, $c(x) = 0$ and $h(x) = I(x \in \{0,1\})$ and the Poisson probability mass function for which $b(x) = e^x$, $c(x) = -\log(x!)$ and $h(x) = I(x \in \{0\} \cup \mathbb{N})$. Here $I(\mathcal{P}) = 1$ if the condition $\mathcal{P}$ is true and $I(\mathcal{P}) = 0$ if $\mathcal{P}$ is false. If the random variable Y has density function (1) then $E(Y) = b'(\eta)$ and $\text{Var}(Y) = b''(\eta)$. A common modelling extension, usually to account for overdispersion, is $\text{Var}(Y) = \phi\, b''(\eta)$ where $\phi > 0$ is a dispersion parameter. This involves replacement of $\log\{p(y;\eta)\}$ by a quasi-likelihood function

$$\{y\eta - b(\eta) + c(y)\}/\phi + d(y,\phi) \tag{2}$$

where $d(y,\phi)$ is a function of y and ϕ only. Note that ϕ is fixed at 1 for ordinary Binomial and Poisson response models. For the Gaussian and Gamma response models (2) corresponds to $\log\{p(y;\eta,\phi)\}$ for a two-parameter density function $p(y;\eta,\phi)$ and ordinary likelihood applies. We study generalized linear mixed models of the form, for observations of the random triples $(\boldsymbol{X}_{\mathrm{A}ij},\boldsymbol{X}_{\mathrm{B}ij},Y_{ij})$, $1\le i\le m$, $1\le j\le n_i$,

$$\begin{aligned}&Y_{ij}|\boldsymbol{X}_{\mathrm{A}ij},\boldsymbol{X}_{\mathrm{B}ij},\boldsymbol{U}_i \text{ independent having quasi-likelihood function (2) with}\\ &\text{natural parameter } (\boldsymbol{\beta}^0_{\mathrm{A}}+\boldsymbol{U}_i)^T\boldsymbol{X}_{\mathrm{A}ij}+(\boldsymbol{\beta}^0_{\mathrm{B}})^T\boldsymbol{X}_{\mathrm{B}ij} \text{ such that the } \boldsymbol{U}_i \text{ are}\\ &\text{independent } N(\boldsymbol{0},\boldsymbol{\Sigma}^0) \text{ random vectors.}\end{aligned}\tag{3}$$

The $\boldsymbol{U}_i$ are $d_{\mathrm{A}}\times 1$ unobserved random effects vectors. The $\boldsymbol{X}_{\mathrm{A}ij}$ are $d_{\mathrm{A}}\times 1$ random vectors corresponding to predictors that are partnered by a random effect. The $\boldsymbol{X}_{\mathrm{B}ij}$ are $d_{\mathrm{B}}\times 1$ random vectors are predictors that have a fixed effect only. Let $\boldsymbol{X}_{ij}\equiv(\boldsymbol{X}_{\mathrm{A}ij}^T,\boldsymbol{X}_{\mathrm{B}ij}^T)^T$ denote the combined predictor vectors. We assume that the $\boldsymbol{X}_{ij}$ and $\boldsymbol{U}_i$, for $1\le i\le m$ and $1\le j\le n_i$, are totally independent, with the $\boldsymbol{X}_{ij}$ each having the same distribution as the $(d_{\mathrm{A}}+d_{\mathrm{B}})\times 1$ random vector $\boldsymbol{X}=(\boldsymbol{X}_{\mathrm{A}}^T,\boldsymbol{X}_{\mathrm{B}}^T)^T$ and the $\boldsymbol{U}_i$ each having the same distribution as the random vector $\boldsymbol{U}$.

For any $\boldsymbol{\beta}_{\mathrm{A}}$ $(d_{\mathrm{A}}\times 1)$, $\boldsymbol{\beta}_{\mathrm{B}}$ $(d_{\mathrm{B}}\times 1)$ and $\boldsymbol{\Sigma}$ $(d_{\mathrm{A}}\times d_{\mathrm{A}})$ that is symmetric and positive definite and conditional on the $\boldsymbol{X}_{ij}$ data, the quasi-likelihood is

$$\begin{aligned}\ell(\boldsymbol{\beta}_{\mathrm{A}},\boldsymbol{\beta}_{\mathrm{B}},\boldsymbol{\Sigma})&=\sum_{i=1}^m\sum_{j=1}^{n_i}[\{Y_{ij}(\boldsymbol{\beta}_{\mathrm{A}}^T\boldsymbol{X}_{\mathrm{A}ij}+\boldsymbol{\beta}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij})+c(Y_{ij})\}/\phi+d(Y_{ij},\phi)]-\frac{m}{2}\log|2\pi\boldsymbol{\Sigma}|\\ &+\sum_{i=1}^m\log\int_{\mathbb{R}^{d_{\mathrm{A}}}}\exp\left[\sum_{j=1}^{n_i}\{Y_{ij}\boldsymbol{u}^T\boldsymbol{X}_{\mathrm{A}ij}-b\big((\boldsymbol{\beta}_{\mathrm{A}}+\boldsymbol{u})^T\boldsymbol{X}_{\mathrm{A}ij}+\boldsymbol{\beta}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\big)\}/\phi-\tfrac{1}{2}\boldsymbol{u}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{u}\right]d\boldsymbol{u}.\end{aligned}$$

The maximum quasi-likelihood estimator of $(\boldsymbol{\beta}^0_{\mathrm{A}},\beta^0_{\mathrm{B}},\boldsymbol{\Sigma}^0)$ is

$$(\widehat{\boldsymbol{\beta}}_{\mathrm{A}},\widehat{\boldsymbol{\beta}}_{\mathrm{B}},\widehat{\boldsymbol{\Sigma}})=\underset{\boldsymbol{\beta}_{\mathrm{A}},\boldsymbol{\beta}_{\mathrm{B}},\boldsymbol{\Sigma}}{\operatorname{argmax}}\,\ell(\boldsymbol{\beta}_{\mathrm{A}},\boldsymbol{\beta}_{\mathrm{B}},\boldsymbol{\Sigma}).$$

Despite the ubiquity of model (3) and availability of established software such as the function `glmer()` in the package lme4 (Bates et al., 2015) within the R computing environment (R Core Team, 2020), asymptotic normality results that lend themselves to confidence interval construction and Wald hypothesis tests via studentization are not present in the existing generalized linear mixed model literature. We address this issue in the next section.

3. Asymptotic normality theorem

The main theoretical contribution of this article is an asymptotic normality theorem for the maximum quasi-likelihood estimators. Define

$$n\equiv\frac{1}{m}\sum_{i=1}^m n_i=\text{average of the within-group sample sizes,}$$

$$\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{B}}}(\boldsymbol{U})\equiv E\left\{b''\Big((\boldsymbol{\beta}^0_{\mathrm{A}}+\boldsymbol{U})^T\boldsymbol{X}_{\mathrm{A}}+(\boldsymbol{\beta}^0_{\mathrm{B}})^T\boldsymbol{X}_{\mathrm{B}}\Big)\begin{bmatrix}\boldsymbol{X}_{\mathrm{A}}\boldsymbol{X}_{\mathrm{A}}^T & \boldsymbol{X}_{\mathrm{A}}\boldsymbol{X}_{\mathrm{B}}^T\\ \boldsymbol{X}_{\mathrm{B}}\boldsymbol{X}_{\mathrm{A}}^T & \boldsymbol{X}_{\mathrm{B}}\boldsymbol{X}_{\mathrm{B}}^T\end{bmatrix}\middle|\,\boldsymbol{U}\right\}$$

and

$$\boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\rm B}} \equiv \left(E\left[\left\{\text{lower right } d_{\rm B}\times d_{\rm B} \text{ block of } \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\rm B}}(\boldsymbol{U})^{-1}\right\}^{-1}\right]\right)^{-1}.$$

Let $\|\boldsymbol{v}\| \equiv (\boldsymbol{v}^T\boldsymbol{v})^{1/2}$ denote the Euclidean norm of a column vector $\boldsymbol{v}$. For a symmetric matrix $\boldsymbol{M}$ let $\lambda_{\min}(\boldsymbol{M})$ denote the smallest eigenvalue of $\boldsymbol{M}$. Also, let $\boldsymbol{D}_d$ denote the matrix of zeroes and ones such that $\boldsymbol{D}_d\text{vech}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A})$ for all $d\times d$ symmetric matrices $\boldsymbol{A}$. The Moore-Penrose inverse of $\boldsymbol{D}_d$ is $\boldsymbol{D}_d^+ = (\boldsymbol{D}_d^T\boldsymbol{D}_d)^{-1}\boldsymbol{D}_d^T$.

The theorem relies on the following assumptions:

(A1) The number of groups m diverges to ∞.

(A2) The within-group sample sizes n_i diverge to ∞ in such a way that $n_i/n \to C_i$ for constants $0 < C_i < \infty$, $1 \le i \le m$. Also, $n/m \to 0$ as m and n diverge.

(A3) The distribution of $\boldsymbol{X}$ is such that

$$E\left[\frac{E\Big[\max\{1,\|\boldsymbol{X}\|\}^8\,\max\{1,b''((\boldsymbol{\beta}_{\rm A}+\boldsymbol{U})^T\boldsymbol{X}_{\rm A}+\boldsymbol{\beta}_{\rm B}^T\boldsymbol{X}_{\rm B})\}^2\Big|\boldsymbol{U}\Big]}{\min\{1,\lambda_{\min}\big(E\{\boldsymbol{X}_{\rm A}\boldsymbol{X}_{\rm A}^T b''((\boldsymbol{\beta}_{\rm A}+\boldsymbol{U})^T\boldsymbol{X}_{\rm A}+\boldsymbol{\beta}_{\rm B}^T\boldsymbol{X}_{\rm B})|\boldsymbol{U}\}\big)\}^2}\right] < \infty$$

for all $\boldsymbol{\beta}_{\rm A}\in\mathbb{R}^{d_{\rm A}}$, $\boldsymbol{\beta}_{\rm B}\in\mathbb{R}^{d_{\rm B}}$ and $\boldsymbol{\Sigma}$ a $d_{\rm A}\times d_{\rm A}$ symmetric and positive definite matrix.

Theorem 1. *Assume that conditions (A1)–(A3) hold. Then*

$$\sqrt{m}\begin{bmatrix}\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_A^0\\ \sqrt{n}\big(\widehat{\boldsymbol{\beta}}_B-\boldsymbol{\beta}_B^0\big)\\ \text{vech}(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^0)\end{bmatrix}\xrightarrow{\mathcal{D}} N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}^0 & \boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \phi\boldsymbol{\Lambda}_{\boldsymbol{\beta}_B} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O} & 2\boldsymbol{D}_{d_A}^+(\boldsymbol{\Sigma}^0\otimes\boldsymbol{\Sigma}^0)\boldsymbol{D}_{d_A}^{+T}\end{bmatrix}\right).$$

Some remarks concerning Theorem 1 are:

1. The asymptotic variances of estimators of fixed effects that are not partnered by random effects have order $(mn)^{-1}$, which are superior to the order m^{-1} asymptotic variances of estimators of fixed effects that are attached to random effects. The random effects variance and covariance parameters also have order m^{-1} asymptotic variances. For the $d_{\rm A}=1$ case, usually corresponding to random intercept models, results of this type are given in Nie (2007) and, in the Gaussian case, follow from results such as (3.60) and (3.61) of McCulloch et al. (2008). In the literature to date, we are not aware of such results at Theorem 1's multivariate $\widehat{\boldsymbol{\beta}}_{\rm A}$ level of generality. We are also not aware of other results for generalized linear mixed models that provide the precise leading term and joint behaviour of all maximum likelihood estimators under (A1)–(A2) asymptotics.

2. Asymptotic orthogonality between $\boldsymbol{\beta}_{\rm A}$, $\boldsymbol{\beta}_{\rm B}$ and $\boldsymbol{\Sigma}$ is apparent from Theorem 1.

3. Results in the existing literature, such as Theorem 3 of Nie (2007) and equation (6) of Waite and Woods (2015) involve expectations over the response distribution in their Fisher information approximations. In contrast, the matrix $\boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\mathrm{B}}}$ involves expectation over the simpler random effects distribution. This simplification is due to careful asymptotic analysis of the response distribution expectations in the derivation of Theorem 1.

4. Theorem 1 treats the dispersion parameter ϕ as fixed. For the Gaussian and Gamma response cases all parameters in (3), including ϕ, can be estimated using ordinary maximum likelihood. Exact orthogonality between ϕ and $(\boldsymbol{\beta}_{\mathrm{A}}, \boldsymbol{\beta}_{\mathrm{B}})$ and asymptotic orthogonality between ϕ and $\boldsymbol{\Sigma}$ means that the asymptotic covariance matrices of Theorem 1 still hold for $\widehat{\boldsymbol{\beta}}_{\mathrm{A}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{B}}$ and $\widehat{\boldsymbol{\Sigma}}$. The extension of Theorem 1 for maximum likelihood estimation of ϕ^0 involves the addition of $\sqrt{mn}\big(\widehat{\phi} - \phi^0\big) \xrightarrow{\mathcal{D}} N\big(0, v(\phi^0)\big)$ where $v(x) \equiv 2x^2$ for Gaussian responses and $v(x) \equiv x^4/\{\text{trigamma}(1/x) - x\}$ for Gamma responses; with details given in Bhaskaran (2022). For the quasi-likelihood extension of the Binomial and Poisson response cases ϕ cannot be estimated via maximum quasi-likelihood and, typically, is estimated via a method of moments approach following the quasi-likelihood estimation phase. The values of the maximum quasi-likelihood estimates of $\boldsymbol{\beta}_{\mathrm{A}}$, $\boldsymbol{\beta}_{\mathrm{B}}$ and $\boldsymbol{\Sigma}$ do not depend on ϕ. Hence, Theorem 1 is unaffected by estimation of ϕ for these response cases too.

4. Asymptotically valid inference

The asymptotic normality results for maximum quasi-likelihood estimators given in Theorem 1 also hold when the quantities appearing in the asymptotic variances are replaced by consistent estimators. This process is often referred to as *studentization*. Since, for $\widehat{\boldsymbol{\beta}}_{\mathrm{A}}$ and $\widehat{\boldsymbol{\Sigma}}$, the asymptotic covariance matrices only involve $\boldsymbol{\Sigma}^0$, studentization simply involves its replacement with $\widehat{\boldsymbol{\Sigma}}$ and we have the asymptotic normality results

$$\sqrt{m}\,\widehat{\boldsymbol{\Sigma}}^{-1/2}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{A}} - \boldsymbol{\beta}_{\mathrm{A}}^0\Big) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{I}) \text{ and}$$

$$\sqrt{m}\{2\boldsymbol{D}_{d_{\mathrm{A}}}^{+}(\widehat{\boldsymbol{\Sigma}} \otimes \widehat{\boldsymbol{\Sigma}})\boldsymbol{D}_{d_{\mathrm{A}}}^{+T}\}^{-1/2}\text{vech}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{I}).$$

If $\widehat{\sigma}_k^2$, $1 \le k \le d_{\mathrm{A}}$, denotes the kth diagonal entry of $\widehat{\boldsymbol{\Sigma}}$ then it follows that the intervals

$$(\widehat{\boldsymbol{\beta}}_{\mathrm{A}})_k \pm \Phi^{-1}\big(1 - \tfrac{1}{2}\,\alpha\big)\sqrt{\frac{\widehat{\sigma}_k^2}{m}} \quad \text{and} \quad \widehat{\sigma}_k^2 \pm \Phi^{-1}\big(1 - \tfrac{1}{2}\,\alpha\big)\sqrt{\frac{2(\widehat{\sigma}_k^2)^2}{m}}, \tag{4}$$

where Φ is the $N(0,1)$ cumulative distribution function, are asymptotically valid $100(1-\alpha)\%$ confidence intervals for the kth entry of $\boldsymbol{\beta}_{\mathrm{A}}^0$ and the (k,k) entry of $\boldsymbol{\Sigma}^0$, respectively.

Practical asymptotically valid inference for the entries of $\boldsymbol{\beta}_{\mathrm{B}}^0$ is more intricate. Studentization of the Theorem 1 results for $\boldsymbol{\beta}_{\mathrm{B}}^0$ leads to the following $100(1-\alpha)\%$ confidence interval for the kth entry of $\boldsymbol{\beta}_{\mathrm{B}}^0$ $(1 \le k \le d_{\mathrm{B}})$:

$$\big(\widehat{\boldsymbol{\beta}}_{\mathrm{B}}\big)_k \pm \Phi^{-1}\big(1 - \tfrac{1}{2}\,\alpha\big)\sqrt{\frac{\phi\big(\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{\beta}_{\mathrm{B}}}\big)_{kk}}{mn}}. \tag{5}$$

Here $(\widehat{\mathbf{\Lambda}}_{\boldsymbol{\beta}_{\mathrm{B}}})_{kk}$ is the (k,k) entry of

$$\widehat{\mathbf{\Lambda}}_{\boldsymbol{\beta}_{\mathrm{B}}}\equiv\Bigg[|2\pi\widehat{\mathbf{\Sigma}}|^{-1/2}\int_{\mathbb{R}^{d_{\mathrm{A}}}}\Big\{\text{lower right } d_{\mathrm{B}}\times d_{\mathrm{B}} \text{ block of } \widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}_{\mathrm{B}}}(\boldsymbol{u})^{-1}\Big\}^{-1} \times\exp\left(-\tfrac{1}{2}\boldsymbol{u}^T\widehat{\mathbf{\Sigma}}^{-1}\boldsymbol{u}\right)\,d\boldsymbol{u}\Bigg]^{-1} \tag{6}$$

where, for each $\boldsymbol{u}\in\mathbb{R}^{d_{\mathrm{A}}}$,

$$\widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}_{\mathrm{B}}}(\boldsymbol{u})\equiv\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n_i}b''\Big((\widehat{\boldsymbol{\beta}}_{\mathrm{A}}+\boldsymbol{u})^T\boldsymbol{X}_{\mathrm{A}ij}+\widehat{\boldsymbol{\beta}}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\Big)\begin{bmatrix}\boldsymbol{X}_{\mathrm{A}ij}\boldsymbol{X}_{\mathrm{A}ij}^T & \boldsymbol{X}_{\mathrm{A}ij}\boldsymbol{X}_{\mathrm{B}ij}^T\\ \boldsymbol{X}_{\mathrm{B}ij}\boldsymbol{X}_{\mathrm{A}ij}^T & \boldsymbol{X}_{\mathrm{B}ij}\boldsymbol{X}_{\mathrm{B}ij}^T\end{bmatrix}.$$

In (6) integration is applied element-wise to each entry of the matrix inside the integral. In some circumstances the integrals in (6) can be evaluated exactly. The most obvious case is the Gaussian response situation for which $b''(x)=1$, implying that $\widehat{\mathbf{\Omega}}_{\boldsymbol{\beta}_{\mathrm{B}}}(\boldsymbol{u})$ is constant as a function of $\boldsymbol{u}$. A less obvious one is Poisson random intercept models for which $b''(x)=e^x$, $\boldsymbol{X}_{\mathrm{A}}=1$ and $\boldsymbol{\beta}_{\mathrm{A}}$ set to the scalar fixed effects intercept parameter β_0. In this special case

$$\widehat{\mathbf{\Lambda}}_{\boldsymbol{\beta}_{\mathrm{B}}}=\text{lower right } d_{\mathrm{B}}\times d_{\mathrm{B}} \text{ block of}$$
$$\left\{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\exp\left(\widehat{\beta}_0+\tfrac{1}{2}\widehat{\sigma}^2+\widehat{\boldsymbol{\beta}}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\right)\begin{bmatrix}1 & \boldsymbol{X}_{\mathrm{B}ij}^T\\ \boldsymbol{X}_{\mathrm{B}ij} & \boldsymbol{X}_{\mathrm{B}ij}\boldsymbol{X}_{\mathrm{B}ij}^T\end{bmatrix}\right\}^{-1}.$$

In more general cases numerical integration is required to evaluate the entries of $\widehat{\mathbf{\Lambda}}_{\boldsymbol{\beta}_{\mathrm{B}}}$. Investigations for the logistic and Poisson cases have revealed that the integrands are well-behaved with exponentiated quadratic decaying tails. In the $d_{\mathrm{A}}=1$ case we found that the simple strategy of a line search for each integrand's effective support and then application of either trapezoidal integration or the R function `integrate()` provided effective and stable solutions. This strategy was used in the logistic mixed model simulation study described in the next paragraph. The study entailed 100,000 numerical integrations of this type and all of them were achieved in a stable manner. The relevant R functions are available in the supplementary material.

We ran a simulation study to assess the efficacy of Theorem 1-based confidence intervals for the $d_{\mathrm{A}}=d_{\mathrm{B}}=1$ logistic mixed model with $\phi=1$. In this case $\boldsymbol{\beta}_{\mathrm{A}}$, $\boldsymbol{\beta}_{\mathrm{B}}$ and $\mathbf{\Sigma}$ are replaced by the scalar parameter symbols β_0, β_{B} and σ^2. The true parameter vector $(\beta_0^0,\beta_{\mathrm{B}}^0,(\sigma^2)^0)$ varied over the set

$$\{(0.2,-0.1,0.25),(-0.3,1.2,0.6),(1.6,0.2,0.7),(0.15,-0.5,1),(-1.3,0.1,0.8)\}$$

and the distribution of the X_{ij} was taken to be either $N(0,1)$ or Uniform$(-1,1)$, the uniform distribution over the interval $(-1,1)$. The number of groups m varied over the set $\{100,200,\ldots,1000\}$ and the sample size within each group was n fixed at $m/10$. For each of the possible combinations of the true parameter vector, the X_{ij} distribution and

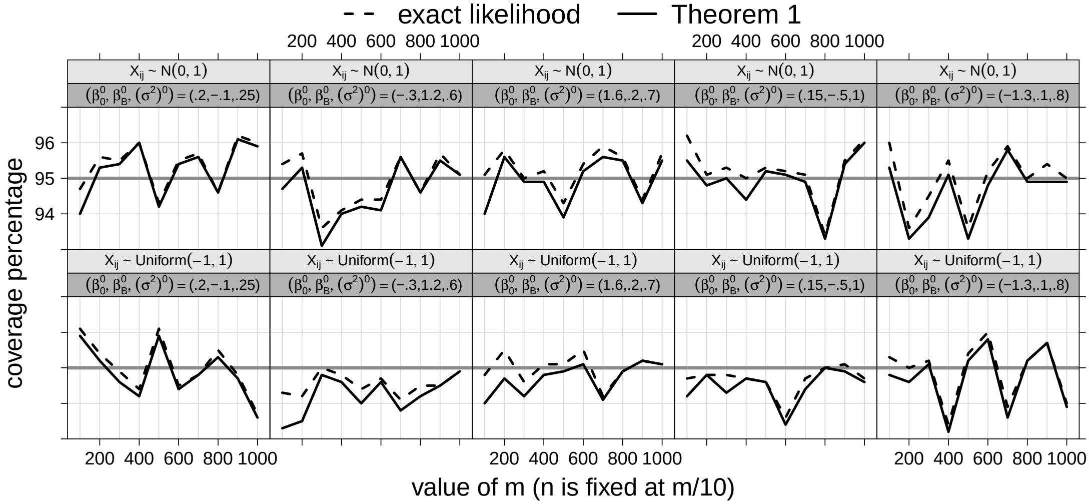

Fig. 1. Actual coverage percentage of nominally 95% confidence intervals for β_{B}^0 in a $d_{\text{A}} = d_{\text{B}} = 1$ logistic mixed model. The confidence intervals are obtained using the exact observed Fisher information computations provided by the function `glmer()` in the R package lme4 (dashed lines) and Theorem 1 with studentization according to (5) (solid lines). The nominal percentage is shown as a thick grey horizontal line. The percentages are based on 1000 replications. The values of m are $100, 200, \ldots, 1000$. The value of n is fixed at $m/10$.

the sample size pair we simulated $1,000$ replications. For each sample, the maximum likelihood estimates of β_0^0, β_{B}^0 and $(\sigma^2)^0$ were obtained using the function `glmer()` in the R package lme4 (Bates et al., 2015). Using these estimates, we computed 95% confidence intervals based on (5) with $\alpha = 0.05$. For comparison, the 95% confidence interval based on the exact observed Fisher information, as provided by `glmer()`, was also computed.

Figure 1 shows the actual coverage percentages for the advertized 95% confidence intervals. It is seen that the two approaches give almost identical coverage percentages for all one hundred truth, response distribution and sample size combinations. This is suggestive of $(\widehat{\mathbf{\Lambda}}_{\beta_{\text{B}}})_{11}/(mn)$ providing a very good approximation to the variance of $\widehat{\beta}_{\text{B}}$ that arises from the exact observed Fisher information, and subsequent investigations show this to be the case. In this logistic case the numerical integration is simpler for the approach involving Theorem 1 and studentization. As mentioned above, the Poisson case requires no numerical integration.

Inspection of Figure 2 reveals that, for inference concerning β_0^0, the very simple confidence interval given by the first expression in (4) performs well when m is above about 500. For lower m the order leading term asymptotics, involving the order m^{-1} asymptotic variance, are seen to be rather crude. Of course, the exact observed Fisher information approach leads to better coverage. However, if m is in the several hundreds or thousands then the closed form confidence interval arising from Theorem 1 and studentization is an attractive alternative to the numerical integration-based exact approach. For $d_{\text{A}} > 1$ mul-

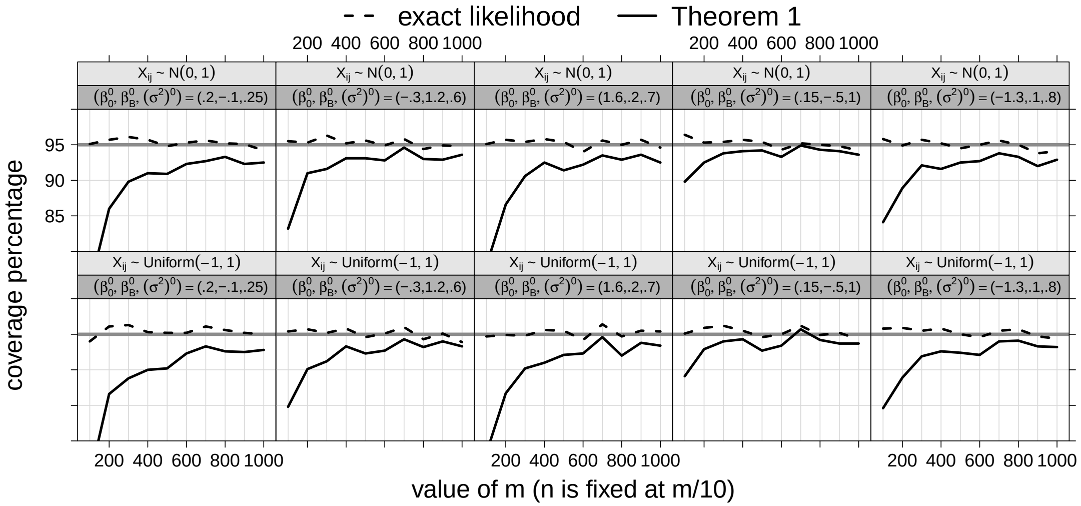

Fig. 2. Actual coverage percentage of nominally 95% confidence intervals for β_0^0 in a $d_{\text{A}} = d_{\text{B}} = 1$ logistic mixed model. The confidence intervals are obtained using the exact observed Fisher information computations provided by the function `glmer()` in the R package lme4 (dashed lines) and Theorem 1 with studentization according to (5) (solid lines). The nominal percentage is shown as a thick grey horizontal line. The percentages are based on 1000 replications. The values of m are $100, 200, \ldots, 1000$. The value of n is fixed at $m/10$.

tivariate numerical integration is needed for the exact approach, whereas the studentized alternative is trivial. Theorem 1 provides the analyst with this quicker and simpler option for large m.

An interesting problem for future research is the development of *second order* asymptotics for quantification of the variability of $\widehat{\boldsymbol{\beta}}_{\text{B}}$ and facilitating more accurate studentization. For the Gaussian response $d_{\text{A}} = d_{\text{B}} = 1$ case, equation (3.60) of McCulloch et al. (2008) indicates that the asymptotic variance of $\widehat{\beta}_0$ is $m^{-1}(\boldsymbol{\Sigma}^0)_{11} + (mn)^{-1}K\{1 + o_P(1)\}$ where $K > 0$ depends on ϕ and moments of the predictor distribution. Therefore, assuming that such behaviour also holds for generalized responses, the simple studentization used in (4) under-approximates the variability of $\widehat{\beta}_0$ and explains the lower empirical coverage values manifest in Figure 2.

In this section we have focussed on asymptotically valid inference based on confidence intervals. Similar discussion applies to Wald hypothesis tests concerning the model parameters.

5. Approximate optimal design

Theorem 1 and its derivation involve large sample expressions for the Fisher information for the class of generalized linear mixed models defined in Section 2. Here we explain how the same type of approximation applies to the design setting. The analogous large sample approximation of the Fisher information has a tractable form which allows for approximate locally optimal design determination. We restrict attention to *D-optimality*, which corresponds to maximising the *determinant* of the Fisher information, and to optimal design for random intercept generalized linear mixed models. Also, we only consider designed experiments for which large sample sizes are feasible. Common practical situations, such as resource-driven restrictions to incomplete designs, are not covered by our theory. Throughout this section we follow the nomenclature of Russell (2018)'s Chapter 3 on optimal design theory.

In Sections 2–4 we assumed that the data have been observed according to model (3). In this section model (3) applies with $d_{\text{A}} = 1$, $\boldsymbol{\beta}_{\text{A}} = \beta_0$ and $\boldsymbol{\Sigma} = \sigma^2$, but the data are yet to be observed. The unique values of the predictor variables is a finite set of points in $\mathbb{R}^{d_{\text{B}}}$ denoted by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$ and labelled the *support points*. Let n_k denote the number of observations made at $\boldsymbol{x}_k$ and then let $\delta_k \equiv n_k/(n_1 + \ldots + n_k)$ be the proportion of the data placed at support point $\boldsymbol{x}_k$. The δ_k are known as the *design weights*. Let $\mathcal{X} \subseteq \mathbb{R}^{d_{\text{B}}}$ denote the set to which the support points are restricted. For example, if $d_{\text{B}} = 2$ with the first predictor being binary and the second predictor being a proportion then $\mathcal{X} = \{(x_1, x_2) : x_1 \in \{0, 1\},\ 0 \le x_2 \le 1\}$. In non-Gaussian generalized response regression models, the Fisher information matrix depends on model parameters and designs that maximise its determinant for fixed values of the parameters are labelled *locally* D-optimal designs.

Define

$$n \equiv \frac{1}{s}\sum_{k=1}^{s} n_k = \text{average of the support point replication sizes within each group.}$$

The theorem involves the following assumption:

(A4) The design sample sizes n_k diverge to ∞ in such a way that $n_k/(sn) \to \delta_k$ for constants $0 < \delta_k < 1$, $1 \le k \le s$.

Theorem 2. *Consider the $d_{\text{A}} = 1$ random intercept generalized linear mixed model with design weights δ_k and corresponding support points $\boldsymbol{x}_k \in \mathcal{X} \subseteq \mathbb{R}^{d_B}$, $1 \le k \le s$. Assume that condition (A4) holds. Then, based on the exact leading term behaviour of the determinant of the Fisher information matrix, approximate locally D-optimal designs at the parameter vector $(\beta_0, \boldsymbol{\beta}_B, \sigma^2)$ are those for which*

$$\left| \int_{-\infty}^{\infty} \left\{ \textit{lower right } d_{\text{B}} \times d_{\text{B}} \textit{ block of } \left(\sum_{k=1}^{s} \delta_k b''(\beta_0 + \boldsymbol{\beta}_B^T \boldsymbol{x}_k + u) \begin{bmatrix} 1 & \boldsymbol{x}_k^T \\ \boldsymbol{x}_k & \boldsymbol{x}_k \boldsymbol{x}_k^T \end{bmatrix} \right)^{-1} \right\}^{-1} \times \exp\{-u^2/(2\sigma^2)\}\, du \right| \tag{7}$$

is maximal over $\left\{ \delta_k : \delta_k \ge 0,\ \sum_{k=1}^{s} \delta_k = 1,\ 1 \le k \le s \right\}$ *and* $\{\boldsymbol{x}_k \in \mathcal{X} : 1 \le k \le s\}$.

We have the following remarks about Theorem 2:

1. For Poisson mixed models $b''(x) = \exp(x)$ and Theorem 2 simplifies considerably. The D-optimality criterion reduces to

$$\left|\sum_{k=1}^{s} \delta_k \exp(\boldsymbol{\beta}_{\mathrm{B}}^T \boldsymbol{x}_k) \begin{bmatrix} 1 & \boldsymbol{x}_k^T \\ \boldsymbol{x}_k & \boldsymbol{x}_k\boldsymbol{x}_k^T \end{bmatrix}\right| \Bigg/ \sum_{k=1}^{s} \delta_k \exp(\boldsymbol{\beta}_{\mathrm{B}}^T \boldsymbol{x}_k).$$

The numerator of this quantity is the approximate locally D-optimality criterion for Poisson regression models (e.g. equation (5.4) of Russell (2018)). Moreover, as in the generalized linear model situation, approximate locally D-optimal designs for Poisson mixed models only are not impacted by β_0 or σ^2 and only depend on $\boldsymbol{\beta}_{\mathrm{B}}$.

2. In the case of logistic mixed models $b''(x) = 1/[2\{1 + \cosh(x)\}]$ and there is no further simplification of (7). Hence, approximate locally D-optimal designs depend on each of β_0, $\boldsymbol{\beta}_{\mathrm{B}}$ and σ^2. Even though (7) does not admit an explicit form for the logistic case, each of the entries of the approximate Fisher information matrix can be computed using univariate numerical integration. An illustration is given later in this section.

3. Waite and Woods (2015) and Zhang et al. (2017) are examples of recent articles that consider D-optimality for classes of generalized linear mixed models similar to those considered here. However, they use approximations to the Fisher information matrix based on paradigms such as generalized estimating equations and Laplace's method. In contrast, (7) is based on the precise leading term behaviour of the Fisher information matrix.

4. With succinctness in mind, we have restricted our theory and discussion to D-optimality. Other optimality criteria, such as *A-optimality*, also benefit from our precise asymptotics for generalized linear mixed models.

5. Whilst this article is concerned chiefly with situations where the response variables are non-Gaussian, it should be mentioned briefly that for the Gaussian special case $b''(x) = 1$ and the determinant in Theorem 2 is proportional to

$$\left|\sum_{k=1}^{s} \delta_k \begin{bmatrix} 1 & \boldsymbol{x}_k^T \\ \boldsymbol{x}_k & \boldsymbol{x}_k\boldsymbol{x}_k^T \end{bmatrix}\right|. \tag{8}$$

Since it does not depend on any model parameters, designs that (8) maximise are *globally* D-optimal.

6. Approximate locally D-optimal designs for the $d_{\mathrm{A}} > 1$ extension is an interesting challenge that is not met by the theory presented here. A case in point is $d_{\mathrm{A}} = 2$ and $d_{\mathrm{B}} = 0$, so that the only non-intercept predictor has both a fixed effect and random effect. The leading term of the Fisher information matrix is a function of $\boldsymbol{\Sigma}$ only. Therefore second order Fisher information asymptotic behaviour, not covered by theory given here, is required for approximate locally D-optimal design.

To illustrate use of Theorem 2, consider the case where $d_{\mathrm{B}} = 2$ and both predictors are binary. In this case, the only possible support points are $\boldsymbol{x}_k \in \{(0,0),(0,1),(1,0),(1,1)\}$

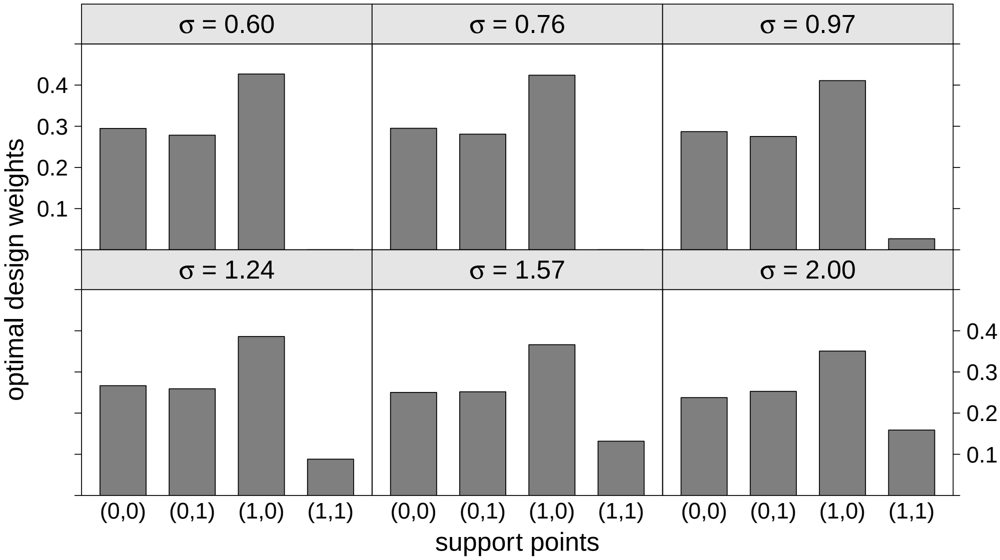

Fig. 3. Approximate locally D-optimal designs for logistic mixed models with two binary predictors when $\beta_0 = -0.3$, $\boldsymbol{\beta}_{\text{B}} = [1.7 \;\; 2.1]^T$ and the values of σ are a geometric sequence of length six between $\sigma = 0.6$ and $\sigma = 2$.

and the number of support points is at most $s = 4$. Therefore, we only need to maximise (7) over the design weights. Figure 3 shows the approximate locally D-optimal designs for the situation where $\beta_0 = -0.3$, $\boldsymbol{\beta}_{\text{B}} = (1.7, 2.1)$ and the values of σ are a geometric sequence of length six between $\sigma = 0.6$ and $\sigma = 2$. To obtain Figure 3 we used code in the R language (R Core Team, 2020), similar to that provided in Section 4.5 of Russell (2018), based on the function `optim()` and Nelder-Mead searches with 100 random initial values. The results were insensitive to the initial value choices.

We see from Figure 3 that for the two lowest values of σ the optimal designs have only three support points, with $(1, 1)$ excluded from the design. As σ increases the design weight for $(1, 1)$ becomes positive and larger and is 0.159 when $\sigma = 2$.

6. Ramifications for Gaussian variational approximation

The *Gaussian variational approximation* approach to fitting and inference for generalized linear mixed models was proposed and developed by Ormerod and Wand (2012). Extensions to related models have been developed by, for example, Hui et al. (2011) and Jeon et al. (2017). The gist of the Gaussian variational approach is to replace the log-likelihood by a Gaussian-type lower bound containing so-called variational parameters. The lower bound has tractability advantages over the exact log-likelihood and, in some sense, replaces difficult numerical integration problems by enlarged optimization problems.

Hall et al. (2011) provided a deep theoretical analysis of the statistical properties of Gaussian variational approximation. Focussing on the $d_{\text{A}} = d_{\text{B}} = 1$ Poisson response special case of (3), which is devoid of the need for numerical integration, they derived the precise asymptotic distributional behaviour of the Gaussian variational approximation estimators of the model parameters, and this is summarized in their Theorem 3.1. However, comparison of these results with those of Theorem 1 reveals that, at least in the $d_{\mathrm{A}} = d_{\mathrm{B}} = 1$ Poisson case, the asymptotic variances of Gaussian variational approximation match those of maximum likelihood and the simpler approach is asymptotically fully efficient. The full efficiency claim was not made in Hall et al. (2011) since the results in the current article were not known at the time. In light of this article's Theorem 1, and Theorem 3.1 of Hall et al. (2011), it is conjectured that Gaussian variational approximation delivers asymptotic fully efficient inference for a wide range of generalized response settings. Other variational inference approaches, such as mean field variational Bayes, are susceptible to under-approximation of the variability of parameter estimates (e.g. Wang and Titterington (2005)). It appears that Gaussian variational approximation does not suffer from this drawback for the class of models considered here.

7. Concluding remarks

Since the emergence of generalized linear mixed models about thirty years ago as a major vehicle for analysis of grouped data with non-Gaussian responses, the asymptotic properties of maximum likelihood estimators has received relatively little attention. Our main theorem provides the definitive, interpretable and usable state of affairs concerning the joint large sample behaviour of the maximum likelihood estimators of all model parameters. The adaptation of our theory to the design context leads to a second theorem that is demonstrably usable for construction of precise leading term-based approximate locally D-optimal designs for generalized linear mixed models. Bhaskaran (2022) provides further details and extensions of the results presented here.

A. Proof of Theorem 1

A.1. Notation and preliminary results

For a generic $d \times 1$ vector $\boldsymbol{v}$ we define $\boldsymbol{v}^{\otimes 0} \equiv 1$, $\boldsymbol{v}^{\otimes 1} \equiv \boldsymbol{v}$ and $\boldsymbol{v}^{\otimes 2} \equiv \boldsymbol{v}\boldsymbol{v}^T$. We also let $\mathrm{diag}(\boldsymbol{v})$ denote the $d \times d$ diagonal matrix with the entries of $\boldsymbol{v}$ along the diagonal. For a matrix $\boldsymbol{M}$ let $\|\boldsymbol{M}\|_{F} = \{\mathrm{tr}(\boldsymbol{M}^T\boldsymbol{M})\}^{1/2}$ denote the Frobenius norm of $\boldsymbol{M}$ and $\|\boldsymbol{M}\|_{S} = \{\text{largest eigenvalue of } \boldsymbol{M}^T\boldsymbol{M}\}^{1/2}$ denote the spectral norm of $\boldsymbol{M}$.

For f a smooth real-valued function of the d-variate argument $\boldsymbol{x} \equiv (x_1, \ldots, x_d)$, let $\nabla f(\boldsymbol{x})$ denote the $d \times 1$ vector with ith entry $\partial f(\boldsymbol{x})/\partial x_i$, $\nabla^2 f(\boldsymbol{x})$ denote the $d \times d$ matrix with (i,j) entry $\partial^2 f(\boldsymbol{x})/(\partial x_i \partial x_j)$ and $\nabla^3 f(\boldsymbol{x})$ denote the $d \times d \times d$ array with (i,j,k) entry $\partial^3 f(\boldsymbol{x})/(\partial x_i \partial x_j \partial x_k)$. Then the multivariate extension of (2.6) of Tierney et al. (1989), and which follows from results in Appendix A of Miyata (2004), for smooth real-valued d-variate functions g, c and h, is

$$\begin{aligned}\frac{\int_{\mathbb{R}^d} g(\boldsymbol{x})c(\boldsymbol{x})\exp\{-nh(\boldsymbol{x})\}\,d\boldsymbol{x}}{\int_{\mathbb{R}^d} c(\boldsymbol{x})\exp\{-nh(\boldsymbol{x})\}\,d\boldsymbol{x}} &= g(\boldsymbol{x}^*) + \frac{\nabla g(\boldsymbol{x}^*)^T\{\nabla^2 h(\boldsymbol{x}^*)\}^{-1}\nabla c(\boldsymbol{x}^*)}{nc(\boldsymbol{x}^*)} \\ &+ \frac{\mathrm{tr}[\{\nabla^2 h(\boldsymbol{x}^*)\}^{-1}\nabla^2 g(\boldsymbol{x}^*)]}{2n} - \frac{\nabla g(\boldsymbol{x}^*)^T\{\nabla^2 h(\boldsymbol{x}^*)\}^{-1}\boldsymbol{a}(\boldsymbol{x}^*)}{2n} + O(n^{-2})\end{aligned} \qquad (9)$$

where $\boldsymbol{x}^*$ is the argument that minimises $h(\boldsymbol{x})$ and $\boldsymbol{a}(\boldsymbol{x})$ is the $d \times 1$ vector having kth entry equal to $\mathrm{tr}\big[\{\nabla^2 h(\boldsymbol{x})\}^{-1}\nabla^3 h(\boldsymbol{x})_{[k]}\big]$ and $\nabla^3 h(\boldsymbol{x})_{[k]}$ is the $d\times d$ matrix with (i,j) entry equal to the (i,j,k) entry of $\nabla^3 h(\boldsymbol{x})$.

Next define,

$$\mathcal{G}_{\mathrm{A}i} \equiv \sum_{j=1}^{n_i}\{Y_{ij} - b'\big((\boldsymbol{\beta}_{\mathrm{A}} + \boldsymbol{U}_i)^T\boldsymbol{X}_{\mathrm{A}ij} + \boldsymbol{\beta}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\big)\}\boldsymbol{X}_{\mathrm{A}ij},$$

and

$$\mathcal{H}_{\mathrm{AA}i} \equiv \sum_{j=1}^{n_i} b''\big((\boldsymbol{\beta}_{\mathrm{A}} + \boldsymbol{U}_i)^T\boldsymbol{X}_{\mathrm{A}ij} + \boldsymbol{\beta}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\big)\boldsymbol{X}_{\mathrm{A}ij}\boldsymbol{X}_{\mathrm{A}ij}^T.$$

Define $\mathcal{G}_{\mathrm{B}i}$ to be the same as $\mathcal{G}_{\mathrm{A}i}$ but with $\boldsymbol{X}_{\mathrm{A}ij}$ replaced by $\boldsymbol{X}_{\mathrm{B}ij}$. Also, define $\mathcal{H}_{\mathrm{AB}i}$ and $\mathcal{H}_{\mathrm{BB}i}$ to be the same as $\mathcal{H}_{\mathrm{AA}i}$ but with $\boldsymbol{X}_{\mathrm{A}ij}\boldsymbol{X}_{\mathrm{A}ij}^T$ replaced by, respectively, $\boldsymbol{X}_{\mathrm{A}ij}\boldsymbol{X}_{\mathrm{B}ij}^T$ and $\boldsymbol{X}_{\mathrm{B}ij}\boldsymbol{X}_{\mathrm{B}ij}^T$. In view of assumption (A3), the orders of magnitude of the these quantities are

$$\begin{gathered}\mathcal{G}_{\mathrm{A}i} = O_P(n^{1/2})\mathbf{1}_{d_{\mathrm{A}}}, \quad \mathcal{G}_{\mathrm{B}i} = O_P(n^{1/2})\mathbf{1}_{d_{\mathrm{B}}}, \\ \mathcal{H}_{\mathrm{AA}i} = O_P(n)\mathbf{1}_{d_{\mathrm{A}}}^{\otimes 2}, \quad \mathcal{H}_{\mathrm{AB}i} = O_P(n)\mathbf{1}_{d_{\mathrm{A}}}\mathbf{1}_{d_{\mathrm{B}}}^T \quad \text{and} \quad \mathcal{H}_{\mathrm{BB}i} = O_P(n)\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2}.\end{gathered} \tag{10}$$

Let $\boldsymbol{X}_{ij} \equiv (\boldsymbol{X}_{\mathrm{A}ij}^T, \boldsymbol{X}_{\mathrm{B}ij}^T)^T$ and $\boldsymbol{X}_i \equiv (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i})$. Key results, which can be obtained using conditional moment calculations, are:

$$E(\mathcal{G}_{\mathrm{A}i}^{\otimes 2}|\boldsymbol{X}_i, U_i) = \phi\mathcal{H}_{\mathrm{AA}i},\ E(\mathcal{G}_{\mathrm{A}i}\mathcal{G}_{\mathrm{B}i}^T|\boldsymbol{X}_i, U_i) = \phi\mathcal{H}_{\mathrm{AB}i},\ E(\mathcal{G}_{\mathrm{B}i}^{\otimes 2}|\boldsymbol{X}_i, U_i) = \phi\mathcal{H}_{\mathrm{BB}i}. \tag{11}$$

A.2. Score exact expressions

For $1 \le i \le m$, let $p_{\boldsymbol{Y}_i|\boldsymbol{X}_i}$ denote the conditional density function, or probability mass function, of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$. Then let

$$\boldsymbol{S}_{\mathrm{A}i} \equiv \nabla_{\boldsymbol{\beta}_{\mathrm{A}}} \log p_{\boldsymbol{Y}_i|\boldsymbol{X}_i}(\boldsymbol{Y}_i|\boldsymbol{X}_i), \quad \boldsymbol{S}_{\mathrm{B}i} \equiv \nabla_{\boldsymbol{\beta}_{\mathrm{B}}} \log p_{\boldsymbol{Y}_i|\boldsymbol{X}_i}(\boldsymbol{Y}_i|\boldsymbol{X}_i)$$

and

$$\boldsymbol{S}_{\mathrm{C}i} \equiv \nabla_{\mathrm{vech}(\boldsymbol{\Sigma})} \log p_{\boldsymbol{Y}_i|\boldsymbol{X}_i}(\boldsymbol{Y}_i|\boldsymbol{X}_i)$$

denote the ith contribution to the scores with respect to each of $\boldsymbol{\beta}_{\mathrm{A}}$, $\boldsymbol{\beta}_{\mathrm{B}}$ and $\mathrm{vech}(\boldsymbol{\Sigma})$. Then standard algebraic manipulations lead to

$$\boldsymbol{S}_{\mathrm{A}i} = \frac{\int_{\mathbb{R}^{d_{\mathrm{A}}}} \boldsymbol{g}_{iA}(\boldsymbol{u})c_D(\boldsymbol{u})\exp\{-nh_i(\boldsymbol{u})\}\,d\boldsymbol{u}}{\int_{\mathbb{R}^{d_{\mathrm{A}}}} c_D(\boldsymbol{u})\exp\{-nh_i(\boldsymbol{u})\}\,d\boldsymbol{u}}, \ \boldsymbol{S}_{\mathrm{B}i} = \frac{\int_{\mathbb{R}^{d_{\mathrm{A}}}} \boldsymbol{g}_{iB}(\boldsymbol{u})c_D(\boldsymbol{u})\exp\{-nh_i(\boldsymbol{u})\}\,d\boldsymbol{u}}{\int_{\mathbb{R}^{d_{\mathrm{A}}}} c_D(\boldsymbol{u})\exp\{-nh_i(\boldsymbol{u})\}\,d\boldsymbol{u}} \tag{12}$$

and

$$\boldsymbol{S}_{\mathrm{C}i} = \frac{\int_{\mathbb{R}^{d_{\mathrm{A}}}} \boldsymbol{g}_{iC}(\boldsymbol{u})c_D(\boldsymbol{u})\exp\{-nh_i(\boldsymbol{u})\}\,d\boldsymbol{u}}{\int_{\mathbb{R}^{d_{\mathrm{A}}}} c_D(\boldsymbol{u})\exp\{-nh_i(\boldsymbol{u})\}\,d\boldsymbol{u}} - \tfrac{1}{2}\boldsymbol{D}_{d_{\mathrm{A}}}^T\mathrm{vec}(\boldsymbol{\Sigma}^{-1}) \tag{13}$$

where $c_D(\boldsymbol{u}) \equiv \exp(-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{u})$, $\boldsymbol{g}_{iA}(\boldsymbol{u}) \equiv \boldsymbol{\Sigma}^{-1}\boldsymbol{u}$, $\boldsymbol{g}_{iC}(\boldsymbol{u}) \equiv \frac{1}{2}\boldsymbol{D}_{d_{\mathrm{A}}}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathrm{vec}(\boldsymbol{u}\boldsymbol{u}^T)$,

$$\boldsymbol{g}_{iB}(\boldsymbol{u}) \equiv \frac{1}{\phi}\sum_{j=1}^{n_i}[\boldsymbol{X}_{\mathrm{B}ij}\{Y_{ij} - b'\big((\boldsymbol{\beta}_{\mathrm{A}} + \boldsymbol{u})^T\boldsymbol{X}_{\mathrm{A}ij} + \boldsymbol{\beta}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\big)\}]$$

and

$$h_i(\boldsymbol{u}) \equiv -\frac{1}{n\phi}\sum_{j=1}^{n_i}\left\{Y_{ij}\boldsymbol{u}^T\boldsymbol{X}_{\mathrm{A}ij} - b\big((\boldsymbol{\beta}_{\mathrm{A}}+\boldsymbol{u})^T\boldsymbol{X}_{\mathrm{A}ij}+\boldsymbol{\beta}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\big)\right\}.$$

Note that an integration by parts step is used to obtain the $\boldsymbol{S}_{\mathrm{A}i}$ expression.

A.3. Score asymptotic expansions

To deal with the integral ratios apparent in (12) and (13) we appeal to (9). Approximations of $\boldsymbol{S}_{\mathrm{A}i}$, $\boldsymbol{S}_{\mathrm{B}i}$ and $\boldsymbol{S}_{\mathrm{C}i}$ that use this equation depend on the random vector

$$\boldsymbol{U}_i^* \equiv \underset{\boldsymbol{u}\in\mathbb{R}^{d_{\mathrm{A}}}}{\mathrm{argmin}}\, h_i(\boldsymbol{u}).$$

It is easily verified that $\boldsymbol{U}_i^*$ is the unique solution of $\nabla h_i(\boldsymbol{u}) = \boldsymbol{0}$.

A.3.1. Asymptotic expansion of $\boldsymbol{U}_i^*$

Note that

$$\begin{aligned}\boldsymbol{0} &= \sum_{j=1}^{n_i}\Big\{Y_{ij} - b'\big((\boldsymbol{\beta}_{\mathrm{A}}+\boldsymbol{U}_i^*)^T\boldsymbol{X}_{\mathrm{A}ij}+\boldsymbol{\beta}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\big)\Big\}\boldsymbol{X}_{\mathrm{A}ij}\\
&= \sum_{j=1}^{n_i}\big\{Y_{ij} - b'\big((\boldsymbol{\beta}_{\mathrm{A}}^0+\boldsymbol{U}_i)^T\boldsymbol{X}_{\mathrm{A}ij}+(\boldsymbol{\beta}_{\mathrm{B}}^0)^T\boldsymbol{X}_{\mathrm{B}ij}\big)\big\}\boldsymbol{X}_{\mathrm{A}ij}\\
&\qquad -\sum_{j=1}^{n_i} b''\big((\boldsymbol{\beta}_{\mathrm{A}}+\boldsymbol{U}_i)^T\boldsymbol{X}_{\mathrm{A}ij}+(\boldsymbol{\beta}_{\mathrm{B}})^T\boldsymbol{X}_{\mathrm{B}ij}\big)\boldsymbol{X}_{\mathrm{A}ij}\boldsymbol{X}_{\mathrm{A}ij}^T(\boldsymbol{U}_i^*-\boldsymbol{U}_i)+\boldsymbol{r}_{it}\\
&= \mathcal{G}_{\mathrm{A}i} - \mathcal{H}_{\mathrm{AA}i}(\boldsymbol{U}_i^*-\boldsymbol{U}_i)+\boldsymbol{r}_{it}\end{aligned}$$

where $\boldsymbol{r}_{it}$ is the Lagrange form of the remainder and is a quadratic form in $\boldsymbol{U}_i^*-\boldsymbol{U}_i$ and a smooth function of $\boldsymbol{U}_{it}^\dagger \equiv (1-t)\boldsymbol{U}_i + t\,\boldsymbol{U}_i^*$ for some $t\in[0,1]$. Inversion of this asymptotic series leads to

$$\boldsymbol{U}_i^* = \boldsymbol{U}_i + \mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{G}_{\mathrm{A}i} + O_P(n^{-1})\boldsymbol{1}_{d_{\mathrm{A}}}.$$

A.3.2. Asymptotic expansion of $\boldsymbol{S}_{\mathrm{A}i}$

If (9) is applied to each entry of the expression for $\boldsymbol{S}_{\mathrm{A}i}$ at (12) then the first three terms on the right-hand side are

$$\boldsymbol{\Sigma}^{-1}\Big(\boldsymbol{U}_i+\mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{G}_{\mathrm{A}i}\Big)+O_P(n^{-1})\boldsymbol{1}_{d_{\mathrm{A}}},\quad -\phi\boldsymbol{\Sigma}^{-1}\mathcal{H}_{\mathrm{AA}i}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_i+O_P(n^{-3/2})\boldsymbol{1}_{d_{\mathrm{A}}}\quad\text{and}\quad \boldsymbol{0}_{d_{\mathrm{A}}}.$$

Note that $\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_i$ is $O_P(1)\boldsymbol{1}_{d_{\mathrm{A}}}$, $\boldsymbol{\Sigma}^{-1}\mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{G}_{\mathrm{A}i}$ is $O_P(n^{-1/2})\boldsymbol{1}_{d_{\mathrm{A}}}$ and $\boldsymbol{\Sigma}^{-1}\mathcal{H}_{\mathrm{AA}i}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_i$ is $O_P(n^{-1})\boldsymbol{1}_{d_{\mathrm{A}}}$. The contribution from the fourth term of (9) is also $O_P(n^{-1})\boldsymbol{1}_{d_{\mathrm{A}}}$ but does not have a succinct matrix algebraic representation. Putting these together we can assert that

$$\boldsymbol{S}_{\mathrm{A}i}=\boldsymbol{\Sigma}^{-1}\Big(\boldsymbol{U}_i+\mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{G}_{\mathrm{A}i}\Big)+O_P(n^{-1})\mathbf{1}_{d_{\mathrm{A}}}.$$

A.3.3. Asymptotic expansion of $\boldsymbol{S}_{\mathrm{B}i}$

For each $1\le k\le d_{\mathrm{B}}$, consider application of (9) to the kth entry of $\boldsymbol{S}_{\mathrm{B}i}$. The first term on the right-hand side of (9) is

$$\text{the } k\text{th entry of } \frac{1}{\phi}\sum_{j=1}^{n_i}\boldsymbol{X}_{\mathrm{B}ij}\big\{Y_{ij}-b'\big((\boldsymbol{\beta}_{\mathrm{A}}+\boldsymbol{U}_i^*)^T\boldsymbol{X}_{\mathrm{A}ij}+\boldsymbol{\beta}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\big)\big\}. \tag{14}$$

Next note that

$$\begin{aligned}b'\big((\boldsymbol{\beta}_{\mathrm{A}}+\boldsymbol{U}_i^*)^T\boldsymbol{X}_{\mathrm{A}ij}+\boldsymbol{\beta}_{\mathrm{B}}^T\boldsymbol{X}_{\mathrm{B}ij}\big)&=b'\big((\boldsymbol{\beta}_{\mathrm{A}}+\boldsymbol{U}_i)^T\boldsymbol{X}_{\mathrm{A}ij}+(\boldsymbol{\beta}_{\mathrm{B}})^T\boldsymbol{X}_{\mathrm{B}ij}\big)\\&\quad+\boldsymbol{X}_{\mathrm{A}ij}^T(\boldsymbol{U}_i^*-\boldsymbol{U}_i)b''\big((\boldsymbol{\beta}_{\mathrm{A}}+\boldsymbol{U}_i)^T\boldsymbol{X}_{\mathrm{A}ij}+(\boldsymbol{\beta}_{\mathrm{B}})^T\boldsymbol{X}_{\mathrm{B}ij}\big)+O_P(n^{-1})\mathbf{1}_{d_{\mathrm{B}}}.\end{aligned}$$

Plugging this into (14) we obtain the first term of $\boldsymbol{S}_{\mathrm{B}i}$ taking the form

$$\frac{1}{\phi}\Big(\mathcal{G}_{\mathrm{B}i}-\mathcal{H}_{\mathrm{AB}i}^T\mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{G}_{\mathrm{A}i}\Big)+O_P(1)\mathbf{1}_{d_{\mathrm{A}}}=O_P(n^{1/2})\mathbf{1}_{d_{\mathrm{A}}}$$

The contribution to $\boldsymbol{S}_{\mathrm{B}i}$ from the second term on the right-hand side of (9) is

$$\mathcal{H}_{\mathrm{AB}i}^T\mathcal{H}_{\mathrm{AA}i}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_i+O_P(n^{-1/2})\mathbf{1}_{d_{\mathrm{B}}}=O_P(1)\mathbf{1}_{d_{\mathrm{B}}}.$$

which is $O_P(1)\mathbf{1}_{d_{\mathrm{B}}}$. The contributions to $\boldsymbol{S}_{\mathrm{B}i}$ from the third and fourth terms on the right-hand side of (9) are also $O_P(1)\mathbf{1}_{d_{\mathrm{B}}}$ but do not admit succinct matrix algebraic forms. Combining all four asymptotic approximations, we are able to declare:

$$\boldsymbol{S}_{\mathrm{B}i}=\frac{1}{\phi}\Big(\mathcal{G}_{\mathrm{B}i}-\mathcal{H}_{\mathrm{AB}i}^T\mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{G}_{\mathrm{A}i}\Big)+O_P(1)\mathbf{1}_{d_{\mathrm{B}}}.$$

A.3.4. Asymptotic expansion of $\boldsymbol{S}_{\mathrm{C}i}$

Application of (9) to the integral ratio component of $\boldsymbol{S}_{\mathrm{C}i}$ leads to the first, second and third terms equalling

$$\begin{aligned}&\tfrac{1}{2}\boldsymbol{D}_{d_{\mathrm{A}}}^T\mathrm{vec}\Big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{U}_i\boldsymbol{U}_i^T+2\mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{G}_{\mathrm{A}i}\boldsymbol{U}_i^T)\boldsymbol{\Sigma}^{-1}\Big)+O_P(n^{-1})\mathbf{1}_{d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2},\\&-\phi\boldsymbol{D}_{d_{\mathrm{A}}}^T\mathrm{vec}\Big(\boldsymbol{\Sigma}^{-1}\mathcal{H}_{\mathrm{AA}i}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_i\boldsymbol{U}_i^T\boldsymbol{\Sigma}^{-1}\Big)+O_P(n^{-3/2})\mathbf{1}_{d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2}\end{aligned} \tag{15}$$

and

$$\frac{\phi}{2}\boldsymbol{D}_{d_{\mathrm{A}}}^T\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\mathcal{H}_{\mathrm{AA}i}^{-1}\boldsymbol{\Sigma}^{-1})+O_P(n^{-3/2})\mathbf{1}_{d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2}. \tag{16}$$

Since $\mathcal{H}_{\mathrm{AA}i}=O_P(n)\mathbf{1}_{d_{\mathrm{A}}}^{\otimes 2}$, (15) and (16) are $O_P(n^{-1})\mathbf{1}_{d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2}$. The fourth term arising from (9) is also $O_P(n^{-1})\mathbf{1}_{d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2}$ but does not have a simple matrix algebraic form. Combining these results we have

$$\boldsymbol{S}_{\mathrm{C}i}=\tfrac{1}{2}\boldsymbol{D}_{d_{\mathrm{A}}}^T\Big\{\mathrm{vec}\Big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{U}_i\boldsymbol{U}_i^T+2\mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{G}_{\mathrm{A}i}\boldsymbol{U}_i^T)\boldsymbol{\Sigma}^{-1}\Big)-\mathrm{vec}(\boldsymbol{\Sigma}^{-1})\Big\}+O_P(n^{-1})\mathbf{1}_{d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2}.$$

A.4. Lemma for the population leading term of the main Fisher information block

The population quantity $\boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\mathrm{B}}}$ appearing in Theorem 1 corresponds to the convergence in probability limit of a particular random form involving the $\mathcal{H}_{ri}$s. In this section we isolate the problem of deriving this population leading term in the form of Lemma 1.

Lemma 1. *Let* $\boldsymbol{X} \equiv (\boldsymbol{X}_A^T, \boldsymbol{X}_B^T)^T$ *and* $\boldsymbol{X}_{ij} \equiv (\boldsymbol{X}_{Aij}^T, \boldsymbol{X}_{Bij}^T)^T$, $1 \le i \le m$, $1 \le j \le n_i$ *be independent and identically distributed* $(d_{\scriptscriptstyle A} + d_{\scriptscriptstyle B}) \times 1$ *random vectors, with* $d_{\scriptscriptstyle A} \ge 1$ *being the number of entries of* $\boldsymbol{X}_A$ *and the* $\boldsymbol{X}_{Aij}$*s. Also, let* $\boldsymbol{U}$ *and* $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m$ *be independent and identically distributed random vectors, distributed independently of* $\boldsymbol{X}$ *and the* $\boldsymbol{X}_{ij}$*s. Let* f *be a Borel measurable, positive real-valued function on* $\mathbb{R}^{d_A+d_B}$ *and assume that*

$$E\left[\frac{E\Big[\max\{1, \|\boldsymbol{X}\|\}^8 \max\{1, f(\boldsymbol{X}, \boldsymbol{U})\}^2 \Big| \boldsymbol{U}\Big]}{\min\{1, \lambda_{\min}(E\{\boldsymbol{X}_A\boldsymbol{X}_A^T f(\boldsymbol{X}, \boldsymbol{U})|\boldsymbol{U}\})\}^2}\right] < \infty. \tag{17}$$

If m *and the* n_i *satisfy assumptions (A1) and (A2) then*

$$\begin{aligned} &E\left[\frac{1}{mn}\sum_{i=1}^m \left\{\sum_{j=1}^{n_i} \boldsymbol{X}_{Bij}\boldsymbol{X}_{Aij}^T f(\boldsymbol{X}_{ij}, \boldsymbol{U}_i)\right\}\left\{\sum_{j=1}^{n_i} \boldsymbol{X}_{Aij}\boldsymbol{X}_{Aij}^T f(\boldsymbol{X}_{ij}, \boldsymbol{U}_i)\right\}^{-1}\right. \\ &\qquad\qquad \left. \times \left\{\sum_{j=1}^{n_i} \boldsymbol{X}_{Bij}\boldsymbol{X}_{Aij}^T f(\boldsymbol{X}_{ij}, \boldsymbol{U}_i)\right\}^T \middle| \boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{mn_m}\right] \\ &\xrightarrow{P} E\Big(E\{\boldsymbol{X}_B\boldsymbol{X}_A^T f(\boldsymbol{X}, \boldsymbol{U})|\boldsymbol{U}\}\big[E\{\boldsymbol{X}_A\boldsymbol{X}_A^T f(\boldsymbol{X}, \boldsymbol{U})|\boldsymbol{U}\}\big]^{-1} \\ &\qquad\qquad \times E\{\boldsymbol{X}_B\boldsymbol{X}_A^T f(\boldsymbol{X}, \boldsymbol{U})|\boldsymbol{U}\}^T\Big). \end{aligned} \tag{18}$$

A.4.1. Proof of Lemma 1

Definitions

Let

$$\widehat{\mathcal{N}}_i(\boldsymbol{U}) \equiv \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{X}_{\mathrm{B}ij}\boldsymbol{X}_{\mathrm{A}ij}^T f(\boldsymbol{X}_{ij}, \boldsymbol{U}), \quad \widehat{\mathcal{D}}_i(\boldsymbol{U}) \equiv \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{X}_{\mathrm{A}ij}\boldsymbol{X}_{\mathrm{A}ij}^T f(\boldsymbol{X}_{ij}, \boldsymbol{U}),$$

$$\mathcal{N}(\boldsymbol{U}) \equiv E\{\boldsymbol{X}_{\mathrm{B}}\boldsymbol{X}_{\mathrm{A}}^T f(\boldsymbol{X}, \boldsymbol{U})|\boldsymbol{U}\} \quad \text{and} \quad \mathcal{D}(\boldsymbol{U}) \equiv E\{\boldsymbol{X}_{\mathrm{A}}\boldsymbol{X}_{\mathrm{A}}^T f(\boldsymbol{X}, \boldsymbol{U})|\boldsymbol{U}\}.$$

Next, for $t \in [0, 1]$, let

$$\mathcal{N}_{it}^{\dagger}(\boldsymbol{U}) \equiv (1-t)\mathcal{N}(\boldsymbol{U}) + t\widehat{\mathcal{N}}_i(\boldsymbol{U}) \quad \text{and} \quad \mathcal{D}_{it}^{\dagger}(\boldsymbol{U}) \equiv (1-t)\mathcal{D}(\boldsymbol{U}) + t\widehat{\mathcal{D}}_i(\boldsymbol{U}).$$

If $\boldsymbol{S}$ is a $d_{\scriptscriptstyle \mathrm{B}} \times d_{\scriptscriptstyle \mathrm{A}}$ matrix and $\boldsymbol{T}$ is a $d_{\scriptscriptstyle \mathrm{A}} \times d_{\scriptscriptstyle \mathrm{A}}$ symmetric matrix define

$$\mathcal{R}\left(\begin{bmatrix} \boldsymbol{S} \\ \boldsymbol{T} \end{bmatrix}\right) = \mathrm{vec}(\boldsymbol{S}\boldsymbol{T}^{-1}\boldsymbol{S}^T)^T.$$

Throughout this proof we let $\boldsymbol{X}_i \equiv \{\boldsymbol{X}_{i1},\ldots,\boldsymbol{X}_{in_i}\}$. For each $1\le i\le m$, define the event

$$\mathcal{A}_i \equiv \Big\{\|\widehat{\mathcal{N}}_i(\boldsymbol{U}_i)-\mathcal{N}(\boldsymbol{U}_i)\|_{S}\le 1,\ \lambda_{\min}\big(\widehat{\mathcal{D}}_i(\boldsymbol{U}_i)\big)\ge \tfrac{1}{2}\lambda_{\min}\big(\mathcal{D}(\boldsymbol{U}_i)\big)\Big\}.$$

Family of Intermediate Moment Conditions

The proof involves moment conditions of the form

$$E\left[\frac{\{E\big(\|\boldsymbol{X}_{\mathrm{A}}\|^{p_1}\,\|\boldsymbol{X}_{\mathrm{B}}\|^{p_2}f(\boldsymbol{X},\boldsymbol{U})|\boldsymbol{U}\big)\}^{p_3}}{\big[\min\big\{1,\lambda_{\min}\big(E\{\boldsymbol{X}_{\mathrm{A}}\boldsymbol{X}_{\mathrm{A}}^T f(\boldsymbol{X},\boldsymbol{U})|\boldsymbol{U}\}\big)\big\}\big]^2}\right]<\infty \tag{19}$$

for various values of the triple (p_1,p_2,p_3).

Proof Strategy

The required result follows from

$$\frac{1}{mn}\sum_{i=1}^m n_i E\left\|E\left[\left\{\mathcal{R}\left(\left[\begin{array}{c}\widehat{\mathcal{N}}_i(\boldsymbol{U}_i)\\ \widehat{\mathcal{D}}_i(\boldsymbol{U}_i)\end{array}\right]\right)-\mathcal{R}\left(\left[\begin{array}{c}\mathcal{N}(\boldsymbol{U}_i)\\ \mathcal{D}(\boldsymbol{U}_i)\end{array}\right]\right)\right\}I(\mathcal{A}_i)\middle|\boldsymbol{X}_i\right]\right\|_{S}\to 0 \tag{20}$$

and

$$\frac{1}{mn}\sum_{i=1}^m n_i E\left\|E\left[\left\{\mathcal{R}\left(\left[\begin{array}{c}\widehat{\mathcal{N}}_i(\boldsymbol{U}_i)\\ \widehat{\mathcal{D}}_i(\boldsymbol{U}_i)\end{array}\right]\right)-\mathcal{R}\left(\left[\begin{array}{c}\mathcal{N}(\boldsymbol{U}_i)\\ \mathcal{D}(\boldsymbol{U}_i)\end{array}\right]\right)\right\}I(\mathcal{A}_i^C)\middle|\boldsymbol{X}_i\right]\right\|_{S}\to 0. \tag{21}$$

as $m,n\to\infty$. Our strategy involves proving each of (20) and (21) separately.

Proof of (20)

A Taylor series expansion of $\mathcal{R}$ with the Lagrange form of the remainder is

$$\mathcal{R}\left(\left[\begin{array}{c}\boldsymbol{S}\\ \boldsymbol{T}\end{array}\right]\right)=\mathcal{R}\left(\left[\begin{array}{c}\boldsymbol{S}_0\\ \boldsymbol{T}_0\end{array}\right]\right)+\left[\begin{array}{c}\mathrm{vec}(\boldsymbol{S}-\boldsymbol{S}_0)\\ \mathrm{vec}(\boldsymbol{T}-\boldsymbol{T}_0)\end{array}\right]^T\left[\begin{array}{c}[\{(\boldsymbol{T}_t^\dagger)^{-1}(\boldsymbol{S}_t^\dagger)^T\}\otimes\boldsymbol{I}_{d_{\mathrm{B}}}](\boldsymbol{I}_{d_{\mathrm{B}}^2}+\boldsymbol{K}_{d_{\mathrm{B}}})\\ -\{(\boldsymbol{T}_t^\dagger)^{-1}(\boldsymbol{S}_t^\dagger)^T\}\otimes\{(\boldsymbol{T}_t^\dagger)^{-1}(\boldsymbol{S}_t^\dagger)^T\}\end{array}\right]$$

where $\boldsymbol{K}_{d_{\mathrm{B}}}$ is the commutation matrix of order d_{B} (Magnus and Neudecker, 1979), $\boldsymbol{S}_t^\dagger\equiv(1-t)\boldsymbol{S}_0+t\boldsymbol{S}$, $\boldsymbol{T}_t^\dagger\equiv(1-t)\boldsymbol{T}_0+t\boldsymbol{T}$, and $t\in[0,1]$. Using $\|\boldsymbol{K}_{d_{\mathrm{B}}}\|_{S}=1$ (Magnus and Neudecker, 1979) and $\|\boldsymbol{A}\otimes\boldsymbol{B}\|_{S}=\|\boldsymbol{A}\|_{S}\,\|\boldsymbol{B}\|_{S}$ (e.g. Section 12.3.1 of Golub and Van Loan (2013)) we obtain

$$\begin{aligned}&\left\|\mathcal{R}\left(\left[\begin{array}{c}\boldsymbol{S}\\ \boldsymbol{T}\end{array}\right]\right)-\mathcal{R}\left(\left[\begin{array}{c}\boldsymbol{S}_0\\ \boldsymbol{T}_0\end{array}\right]\right)\right\|_{S}\\ &\qquad\le 2\|\boldsymbol{S}_t^\dagger\|_{S}\,\|(\boldsymbol{T}_t^\dagger)^{-1}\|_{S}\|\boldsymbol{S}-\boldsymbol{S}_0\|_{F}+\big(\|\boldsymbol{S}_t^\dagger\|_{S}\,\|(\boldsymbol{T}_t^\dagger)^{-1}\|_{S}\big)^2\|\boldsymbol{T}-\boldsymbol{T}_0\|_{F}.\end{aligned} \tag{22}$$

Suppose that, for $1\le i\le m$, $(\boldsymbol{U}_i,\boldsymbol{X}_i)$ are such that $\mathcal{A}_i$ occurs. Then standard arguments lead to the bounds

$$\|\mathcal{N}_{it}^\dagger(\boldsymbol{U}_i)\|_{S}\le\|\mathcal{N}(\boldsymbol{U}_i)\|_{S}+1\quad\text{and}\quad\|\mathcal{D}_{it}^\dagger(\boldsymbol{U}_i)^{-1}\|_{S}\le\frac{2}{\lambda_{\min}\big(\mathcal{D}(\boldsymbol{U}_i)\big)}. \tag{23}$$

On application of (22) and (23) and routine arguments the left-hand side of (20) is bounded above by

$$\frac{4}{mn}\sum_{i=1}^{m} n_i \left[E\left\{ \mathcal{W}(\boldsymbol{U}_i)\|\widehat{\mathcal{N}}_i(\boldsymbol{U}_i)-\mathcal{N}(\boldsymbol{U}_i)\|_{\scriptscriptstyle F}\right\} + E\left\{\mathcal{W}(\boldsymbol{U}_i)^2\|\widehat{\mathcal{D}}_i(\boldsymbol{U}_i)-\mathcal{D}(\boldsymbol{U}_i)\|_{\scriptscriptstyle F}\right\}\right] \tag{24}$$

where $\mathcal{W}(\boldsymbol{U}) \equiv \{\|\mathcal{N}(\boldsymbol{U})\|_{\scriptscriptstyle S}+1\}/\lambda_{\min}\big(\mathcal{D}(\boldsymbol{U})\big)$. Convergence of (24) to zero under (A1) and (A2) is readily established under the assumption that

$$E\left[\frac{\{E\{\|\boldsymbol{X}_{\rm A}\|^{p_1}\,\|\boldsymbol{X}_{\rm B}\|^{p_2}\,f(\boldsymbol{X},\boldsymbol{U})|\boldsymbol{U}\}+1\}\left\{E\Big(\|\boldsymbol{X}_{\rm A}\|^{p_3}\,\|\boldsymbol{X}_{\rm B}\|^{p_4}f(\boldsymbol{X},\boldsymbol{U})|\boldsymbol{U}\Big)\right\}^{1/2}}{\lambda_{\min}\Big(E\{\boldsymbol{X}_{\rm A}\boldsymbol{X}_{\rm A}^T\,f(\boldsymbol{X},\boldsymbol{U})|\boldsymbol{U}\}\Big)^{p_5}}\right] \tag{25}$$

is finite for each of $(p_1,p_2,p_3,p_4,p_5)\in\{(1,1,2,2,1),(2,0,4,0,2)\}$. Using the inequalities $(x+1)y<1+x^2+y^2$ for all $x,y\in\mathbb{R}$ and $\max(1/x,1/x^2)\le 1/\{\min(1,x)\}^2$ for all $x>0$ we can replace finiteness of (25) by that of (19) for $(p_1,p_2,p_3)\in\{(0,0,0),(1,1,2),(2,0,2)\}$.

Proof of (21)

First note that the left-hand side of (21) is bounded above by

$$\begin{aligned}\frac{1}{mn}\sum_{i=1}^{m} n_i\Bigg(&\left[E\left\{\left\|\mathcal{R}\left(\left[\begin{array}{c}\widehat{\mathcal{N}}_i(\boldsymbol{U}_i)\\ \widehat{\mathcal{D}}_i(\boldsymbol{U}_i)\end{array}\right]\right)\right\|_{\scriptscriptstyle S}^2\right\}\right]^{1/2}\\ &+\left[E\left\{\left\|\mathcal{R}\left(\left[\begin{array}{c}\mathcal{N}(\boldsymbol{U}_i)\\ \mathcal{D}(\boldsymbol{U}_i)\end{array}\right]\right)\right\|_{\scriptscriptstyle S}^2\right\}\right]^{1/2}\Bigg)P(\mathcal{A}_i^C)^{1/2}.\end{aligned} \tag{26}$$

Making use of the generalized Cauchy-Schwartz inequality on page 1093 of Chipman (1964) we obtain

$$E\left\{\left\|\mathcal{R}\left(\left[\begin{array}{c}\widehat{\mathcal{N}}_i(\boldsymbol{U}_i)\\ \widehat{\mathcal{D}}_i(\boldsymbol{U}_i)\end{array}\right]\right)\right\|_{\scriptscriptstyle S}^2\right\}\le E\left\{\left\|\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{X}_{{\rm B}ij}\boldsymbol{X}_{{\rm B}ij}^T f(\boldsymbol{X}_{ij},\boldsymbol{U}_i)\right\|_{\scriptscriptstyle S}^2\right\}\le E\{\|\boldsymbol{X}_{\rm B}\|^4 f(\boldsymbol{X},U)^2\}$$

which is finite if (17) holds. Similar arguments lead to

$$E\left\{\left\|\mathcal{R}\left(\left[\begin{array}{c}\mathcal{N}(\boldsymbol{U}_i)\\ \mathcal{D}(\boldsymbol{U}_i)\end{array}\right]\right)\right\|_{\scriptscriptstyle S}^2\right\}\le E\left\{\frac{E\{\|\boldsymbol{X}_{\rm A}\|^4\,\|\boldsymbol{X}_{\rm B}\|^4\,f(\boldsymbol{X},\boldsymbol{U})|\boldsymbol{U}\}}{\lambda_{\min}\big(E\{\boldsymbol{X}_{\rm A}\boldsymbol{X}_{\rm A}^T f(\boldsymbol{X},\boldsymbol{U})|\boldsymbol{U}\}\big)^2}\right\}$$

which is finite if (19) holds for $(p_1,p_2,p_3)=(4,4,1)$. Lastly, note that

$$\begin{aligned}P(\mathcal{A}_i^C)&\le P\big(\|\widehat{\mathcal{N}}_i(\boldsymbol{U}_i)-\mathcal{N}(\boldsymbol{U}_i)\|_{\scriptscriptstyle S}>1\big)\\ &\quad+P\Big(\big|\lambda_{\min}\big(\widehat{\mathcal{D}}_i(\boldsymbol{U}_i)\big)-\lambda_{\min}\big(\mathcal{D}(\boldsymbol{U}_i)\big)\big|>\tfrac{1}{2}\lambda_{\min}\big(\mathcal{D}(\boldsymbol{U}_i)\big)\Big).\end{aligned}$$

Application of Markov's inequality and Theorem 8.1.4 (Wielandt-Hoffman) of Golub and Van Loan (2013) leads to $P(\mathcal{A}_i^C)\le Bn_i^{-1}$, for some constant $0<B<\infty$, assuming that

(19) is true for all $(p_1, p_2, p_3) \in \{(2,2,1),(4,0,1)\}$. Substitution of each of these bounds into (26) leads to (21) holding under sample size assumptions (A1) and (A2).

Condensation of the Moment Assumptions

The full list of moment assumptions involving the form (19) are such that (p_1, p_2, p_3) takes values over the set $\{(0,0,0),(4,0,1),(2,2,1),(4,4,1),(1,1,2),(2,0,2)\}$. Inequalities such as $\max\{\|\boldsymbol{X}_{\mathrm{A}}\|, \|\boldsymbol{X}_{\mathrm{B}}\|\} \le \|\boldsymbol{X}\|$ lead to each of these moments of the form in (19) being dominated by the left-hand side of (17). Therefore (17) is sufficient for all moment assumptions appearing in this proof.

A.5. Fisher information matrix

The asymptotic expansions of $\boldsymbol{S}_{\mathrm{A}i}$, $\boldsymbol{S}_{\mathrm{B}i}$ and $\boldsymbol{S}_{\mathrm{C}i}$, as well as results (10) and (11) and Theorem 4.3(iv) of Magnus and Neudecker (1979), lead to

$$\begin{aligned}
&E(\boldsymbol{S}_{\mathrm{A}i}^{\otimes 2}|\boldsymbol{X}_i) = \boldsymbol{\Sigma}^{-1} + O_P(n^{-1})\mathbf{1}_{d_{\mathrm{A}}}^{\otimes 2},\\
&E(\boldsymbol{S}_{\mathrm{B}i}^{\otimes 2}|\boldsymbol{X}_i) = \tfrac{1}{\phi}E\Big(\mathcal{H}_{\mathrm{BB}i} - \mathcal{H}_{\mathrm{AB}i}^T\mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{H}_{\mathrm{AB}i}\Big|\boldsymbol{X}_i\Big) + O_P(1)\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2},\\
&E(\boldsymbol{S}_{\mathrm{C}i}^{\otimes 2}|\boldsymbol{X}_i) = \tfrac{1}{2}\boldsymbol{D}_{d_{\mathrm{A}}}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{D}_{d_{\mathrm{A}}} + O_P(n^{-1})\mathbf{1}_{d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2}^{\otimes 2},\\
&E(\boldsymbol{S}_{\mathrm{A}i}\boldsymbol{S}_{\mathrm{B}i}^T|\boldsymbol{X}_i) = O_P(1)\mathbf{1}_{d_{\mathrm{A}}}\mathbf{1}_{d_{\mathrm{B}}}^T,\ E(\boldsymbol{S}_{\mathrm{A}i}\boldsymbol{S}_{\mathrm{C}i}^T|\boldsymbol{X}_i) = O_P(n^{-1})\mathbf{1}_{d_{\mathrm{A}}}\mathbf{1}_{d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2}^T\\
&\quad\text{and}\quad E(\boldsymbol{S}_{\mathrm{B}i}\boldsymbol{S}_{\mathrm{C}i}^T|\boldsymbol{X}_i) = O_P(1)\mathbf{1}_{d_{\mathrm{B}}}\mathbf{1}_{d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2}^T.
\end{aligned}\tag{27}$$

Under assumptions (A1)–(A3), we have from Lemma 1,

$$\frac{1}{mn}\sum_{i=1}^{m} E\Big(\mathcal{H}_{\mathrm{BB}i} - \mathcal{H}_{\mathrm{AB}i}^T\mathcal{H}_{\mathrm{AA}i}^{-1}\mathcal{H}_{\mathrm{AB}i}\Big|\boldsymbol{X}_i\Big) = \boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\mathrm{B}}}^{-1} + o_P(1)\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2}.$$

Therefore, we have following expression for the Fisher information matrix of the parameter vector $(\boldsymbol{\beta}_{\mathrm{A}}, \boldsymbol{\beta}_{\mathrm{B}}, \mathrm{vech}(\boldsymbol{\Sigma}))$:

$$\begin{aligned}
&I\Big(\boldsymbol{\beta}_{\mathrm{A}}, \boldsymbol{\beta}_{\mathrm{B}}, \mathrm{vech}(\boldsymbol{\Sigma})\Big)\\
&= \begin{bmatrix}
m\boldsymbol{\Sigma}^{-1} + O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}}^{\otimes 2} & O_P(m)\mathbf{1}_{d_{\mathrm{A}}}\mathbf{1}_{d_{\mathrm{B}}}^T & O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}}\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^T\\
O_P(m)\mathbf{1}_{d_{\mathrm{B}}}\mathbf{1}_{d_{\mathrm{A}}}^T & \dfrac{mn\boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\mathrm{B}}}^{-1}}{\phi} + o_P(mn)\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2} & O_P(m)\mathbf{1}_{d_{\mathrm{B}}}\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^T\\
O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}\mathbf{1}_{d_{\mathrm{A}}}^T & O_P(m)\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}\mathbf{1}_{d_{\mathrm{B}}}^T & \begin{array}{c}\dfrac{m\boldsymbol{D}_{d_{\mathrm{A}}}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{D}_{d_{\mathrm{A}}}}{2}\\ +O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^{\otimes 2}\end{array}
\end{bmatrix}.
\end{aligned}$$

where $d_{\mathrm{A}}^{\boxplus} \equiv d_{\mathrm{A}}(d_{\mathrm{A}}+1)/2$. Strictly speaking, the term *Fisher information* applies to the $\phi = 1$ ordinary likelihood situation. In this derivation, we use the same term for the $\phi \neq 1$ quasi-likelihood adjustment for responses such as Binomial and Poisson.

A.6. Inverse Fisher information matrix

For inversion of the Fisher information matrix we work with the ordering $(\boldsymbol{\beta}_{\mathrm{A}}, \mathrm{vech}(\boldsymbol{\Sigma}), \boldsymbol{\beta}_{\mathrm{B}})$ rather than $(\boldsymbol{\beta}_{\mathrm{A}}, \boldsymbol{\beta}_{\mathrm{B}}, \mathrm{vech}(\boldsymbol{\Sigma}))$. A trivial rearrangement of matrix entries leads to

$$
\begin{aligned}
&I\Big(\boldsymbol{\beta}_{\mathrm{A}}, \mathrm{vech}(\boldsymbol{\Sigma}), \boldsymbol{\beta}_{\mathrm{B}}\Big)\\
&= \left[\begin{array}{ccc}
m\boldsymbol{\Sigma}^{-1} + O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}}^{\otimes 2} & O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}}\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^T & O_P(m)\mathbf{1}_{d_{\mathrm{A}}}\mathbf{1}_{d_{\mathrm{B}}}^T\\
O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}\mathbf{1}_{d_{\mathrm{A}}}^T & \begin{array}{c}\dfrac{m\boldsymbol{D}_{d_{\mathrm{A}}}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{D}_{d_{\mathrm{A}}}}{2}\\ +O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^{\otimes 2}\end{array} & O_P(m)\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}\mathbf{1}_{d_{\mathrm{B}}}^T\\
O_P(m)\mathbf{1}_{d_{\mathrm{B}}}\mathbf{1}_{d_{\mathrm{A}}}^T & O_P(m)\mathbf{1}_{d_{\mathrm{B}}}\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^T & \dfrac{mn\boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\mathrm{B}}}^{-1}}{\phi} + o_P(mn)\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2}
\end{array}\right].
\end{aligned}
$$

Then partition $I\big(\boldsymbol{\beta}_{\mathrm{A}}, \mathrm{vech}(\boldsymbol{\Sigma}), \boldsymbol{\beta}_{\mathrm{B}}\big)$ and $I\big(\boldsymbol{\beta}_{\mathrm{A}}, \mathrm{vech}(\boldsymbol{\Sigma}), \boldsymbol{\beta}_{\mathrm{B}}\big)^{-1}$ according to

$$
I\big(\boldsymbol{\beta}_{\mathrm{A}}, \mathrm{vech}(\boldsymbol{\Sigma}), \boldsymbol{\beta}_{\mathrm{B}}\big) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{12}^T & \boldsymbol{A}_{22}\end{bmatrix} \quad \text{and} \quad I\big(\boldsymbol{\beta}_{\mathrm{A}}, \mathrm{vech}(\boldsymbol{\Sigma}), \boldsymbol{\beta}_{\mathrm{B}}\big)^{-1} = \begin{bmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12}\\ (\boldsymbol{A}^{12})^T & \boldsymbol{A}^{22}\end{bmatrix}
$$

where

$$
\begin{aligned}
\boldsymbol{A}_{11} &\equiv \begin{bmatrix} m\boldsymbol{\Sigma}^{-1} + O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}}^{\otimes 2} & O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}}\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^T\\ O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}\mathbf{1}_{d_{\mathrm{A}}}^T & \dfrac{m\boldsymbol{D}_{d_{\mathrm{A}}}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{D}_{d_{\mathrm{A}}}}{2} + O_P(mn^{-1})\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^{\otimes 2}\end{bmatrix},\\
\boldsymbol{A}_{12} &\equiv O_P(m)\big[\mathbf{1}_{d_{\mathrm{B}}}\mathbf{1}_{d_{\mathrm{A}}}^T \quad \mathbf{1}_{d_{\mathrm{B}}}\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^T\big]^T,
\end{aligned}
\tag{28}
$$

$\boldsymbol{A}_{22} \equiv (mn/\phi)\boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\mathrm{B}}}^{-1} + o_P(mn)\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2}$ and $\boldsymbol{A}^{11}$ has dimension $(d_{\mathrm{A}} + d_{\mathrm{A}}^{\boxplus}) \times (d_{\mathrm{A}} + d_{\mathrm{A}}^{\boxplus})$.

The upper left block of $I\big(\boldsymbol{\beta}_{\mathrm{A}}, \mathrm{vech}(\boldsymbol{\Sigma}), \boldsymbol{\beta}_{\mathrm{B}}\big)^{-1}$ is

$$
\boldsymbol{A}^{11} = \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}(\boldsymbol{A}_{22} - \boldsymbol{A}_{12}^T\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1}\boldsymbol{A}_{12}^T\boldsymbol{A}_{11}^{-1} \tag{29}
$$

where, using Theorem 13(d) in Chapter 3 of Magnus and Neudecker (1999),

$$
\boldsymbol{A}_{11}^{-1} = \begin{bmatrix} \dfrac{\boldsymbol{\Sigma}}{m} + O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\mathrm{A}}}^{\otimes 2} & O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\mathrm{A}}}\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^T\\ O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}\mathbf{1}_{d_{\mathrm{A}}}^T & \dfrac{2\boldsymbol{D}_{d_{\mathrm{A}}}^{+}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{D}_{d_{\mathrm{A}}}^{+T}}{m} + O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\mathrm{A}}^{\boxplus}}^{\otimes 2}\end{bmatrix}. \tag{30}
$$

It follows that $\boldsymbol{A}_{12}^T\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} = O_P(m)\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2}$ and so $\boldsymbol{A}_{22} - \boldsymbol{A}_{12}^T\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} = O_P(mn)\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2}$ which then leads to $(\boldsymbol{A}_{22} - \boldsymbol{A}_{12}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}^T)^{-1} = O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2}$. Consequently,

$$
\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}(\boldsymbol{A}_{22} - \boldsymbol{A}_{12}^T\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1}\boldsymbol{A}_{12}^T\boldsymbol{A}_{11}^{-1} = O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\mathrm{B}}}^{\otimes 2}. \tag{31}
$$

Results (29), (30) and (31) imply that

$$\boldsymbol{A}^{11}=\left[\begin{array}{cc}\dfrac{\boldsymbol{\Sigma}}{m}+O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\rm A}}^{\otimes 2} & O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\rm A}}\mathbf{1}_{d_{\rm A}^{\boxplus}}^T\\ O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\rm A}^{\boxplus}}\mathbf{1}_{d_{\rm A}}^T & \dfrac{2\boldsymbol{D}_{d_{\rm A}}^{+}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{D}_{d_{\rm A}}^{+T}}{m}+O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\rm A}^{\boxplus}}^{\otimes 2}\end{array}\right].$$

Next note that

$$\boldsymbol{A}^{22}=\boldsymbol{A}_{22}^{-1}+\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^T(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^T)^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \tag{32}$$

and

$$\boldsymbol{A}_{22}^{-1}=\frac{\phi}{mn}\boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\rm B}}+o_P(m^{-1}n^{-1})\mathbf{1}_{d_{\rm B}}^{\otimes 2}. \tag{33}$$

Hence $\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^T=O_P(mn^{-1})\mathbf{1}_{d_{\rm A}+d_{\rm A}^{\boxplus}}^{\otimes 2}$ implying that $\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^T=\{O_P(m)+O_P(mn^{-1})\}\mathbf{1}_{d_{\rm A}+d_{\rm A}^{\boxplus}}^{\otimes 2}$ and then $(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^T)^{-1}=\{O_P(m^{-1})+O_P(m^{-1}n^{-1})\}\mathbf{1}_{d_{\rm A}+d_{\rm A}^{\boxplus}}^{\otimes 2}$. Continuing in this fashion we get

$$\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^T(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^T)^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}=O_P(m^{-1}n^{-2})\mathbf{1}_{d_{\rm B}}^{\otimes 2}. \tag{34}$$

Combining (32), (33) and (34) we then have

$$\boldsymbol{A}^{22}=\frac{\phi}{mn}\boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\rm B}}+o_P(m^{-1}n^{-1})\mathbf{1}_{d_{\rm B}}^{\otimes 2}.$$

The upper right off-diagonal block of $I\big(\boldsymbol{\beta}_{\rm A},\text{vech}(\boldsymbol{\Sigma}),\boldsymbol{\beta}_{\rm B}\big)$ is

$$\boldsymbol{A}^{12}=-(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^T)^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}=O_P(m^{-1}n^{-1})\mathbf{1}_{d_{\rm A}+d_{\rm A}^{\boxplus}}\mathbf{1}_{d_{\rm B}}^T.$$

The resultant expression for the inverse Fisher information matrix of $\big(\boldsymbol{\beta}_{\rm A},\text{vech}(\boldsymbol{\Sigma}),\boldsymbol{\beta}_{\rm B}\big)$ is

$$\begin{aligned}I\big(\boldsymbol{\beta}_{\rm A},\text{vech}(\boldsymbol{\Sigma}),\boldsymbol{\beta}_{\rm B}\big)^{-1}&=I\big(\boldsymbol{\beta}_{\rm A},\text{vech}(\boldsymbol{\Sigma}),\boldsymbol{\beta}_{\rm B}\big)_{\infty}^{-1}\\ &\quad+\frac{1}{mn}\left[\begin{array}{ccc}O_P(1)\mathbf{1}_{d_{\rm A}}^{\otimes 2} & O_P(1)\mathbf{1}_{d_{\rm A}}\mathbf{1}_{d_{\rm A}^{\boxplus}}^T & O_P(1)\mathbf{1}_{d_{\rm A}}\mathbf{1}_{d_{\rm B}}^T\\ O_P(1)\mathbf{1}_{d_{\rm A}^{\boxplus}}\mathbf{1}_{d_{\rm A}}^T & O_P(1)\mathbf{1}_{d_{\rm A}^{\boxplus}}^{\otimes 2} & O_P(1)\mathbf{1}_{d_{\rm A}^{\boxplus}}\mathbf{1}_{d_{\rm B}}^T\\ O_P(1)\mathbf{1}_{d_{\rm B}}\mathbf{1}_{d_{\rm A}}^T & O_P(1)\mathbf{1}_{d_{\rm B}}\mathbf{1}_{d_{\rm A}^{\boxplus}}^T & o_P(1)\mathbf{1}_{d_{\rm B}}^{\otimes 2}\end{array}\right]\end{aligned} \tag{35}$$

where

$$I\big(\boldsymbol{\beta}_{\rm A},\text{vech}(\boldsymbol{\Sigma}),\boldsymbol{\beta}_{\rm B}\big)_{\infty}^{-1}\equiv\left[\begin{array}{ccc}\dfrac{\boldsymbol{\Sigma}}{m} & \boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \dfrac{2\boldsymbol{D}_{d_{\rm A}}^{+}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{D}_{d_{\rm A}}^{+T}}{m} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O} & \dfrac{\phi\boldsymbol{\Lambda}_{\boldsymbol{\beta}_{\rm B}}}{mn}\end{array}\right].$$

A.7. Lemma for the asymptotic equivalence of $\{I(\boldsymbol{\beta}_{A},\mathrm{vech}(\boldsymbol{\Sigma}),\boldsymbol{\beta}_{B})^{-1}\}^{1/2}$ and $\{I(\boldsymbol{\beta}_{A},\mathrm{vech}(\boldsymbol{\Sigma}),\boldsymbol{\beta}_{B})^{-1}_{\infty}\}^{1/2}$

Theorem 1 involves replacement of the matrix $\{I(\boldsymbol{\beta}_{\mathrm{A}},\mathrm{vech}(\boldsymbol{\Sigma}),\boldsymbol{\beta}_{\mathrm{B}})^{-1}\}^{1/2}$ by the matrix $\{I(\boldsymbol{\beta}_{\mathrm{A}},\mathrm{vech}(\boldsymbol{\Sigma}),\boldsymbol{\beta}_{\mathrm{B}})^{-1}_{\infty}\}^{1/2}$, due to the remainder terms in (35) having asymptotically negligible effect on the relevant matrix square roots. Lemma 2 provides a formalisation of this state of affairs, which is used in the final steps of the derivation in Section A.8.

Lemma 2. *Define the sequences of matrices*

$$\boldsymbol{M}_n \equiv \begin{bmatrix} \boldsymbol{K}+Q_n\boldsymbol{1}_p^{\otimes 2} & R_n\boldsymbol{1}_p\boldsymbol{1}_q^T \\ R_n\boldsymbol{1}_q\boldsymbol{1}_p^T & \frac{1}{n}\boldsymbol{L}+T_n\boldsymbol{1}_q^{\otimes 2} \end{bmatrix} \quad \textit{and} \quad \boldsymbol{M}_{n,\infty} \equiv \begin{bmatrix} \boldsymbol{K} & \boldsymbol{O} \\ \boldsymbol{O} & \frac{1}{n}\boldsymbol{L} \end{bmatrix}$$

where $\boldsymbol{K}$ $(p\times p)$ and $\boldsymbol{L}$ $(q\times q)$ are symmetric positive definite matrices and Q_n, R_n and T_n are sequences of random variables satisfying $Q_n=o_P(1)$, $R_n=O_P(n^{-1})$ and $T_n=o_P(n^{-1})$. Then, as $n\to\infty$,

$$\|\boldsymbol{M}_{n,\infty}^{-1/2}\boldsymbol{M}_n^{1/2}-\boldsymbol{I}\|_{F}\xrightarrow{P}0.$$

A.7.1. Proof of Lemma 2

Our proof uses the integral form of the square root of a matrix which, for a matrix $\boldsymbol{A}$ having no eigenvalues on $\mathbb{R}_{-}$ is given by

$$\boldsymbol{A}^{1/2}=\frac{2}{\pi}\int_0^\infty \boldsymbol{A}(\boldsymbol{A}+t^2\boldsymbol{I})^{-1}\,dt \tag{36}$$

(e.g. Higham (2008)). For all n sufficiently large so that negative eigenvalues are avoided, application of (36) to each of $\boldsymbol{M}_{n,\infty}^{-1}$ and $\boldsymbol{M}_n$ leads to

$$\begin{aligned}\boldsymbol{M}_{n,\infty}^{-1/2}\boldsymbol{M}_n^{1/2}=\frac{4}{\pi^2}\int_0^\infty\int_0^\infty &\begin{bmatrix} (\boldsymbol{I}_p+t^2\boldsymbol{K})^{-1}\boldsymbol{K} & R_n(\boldsymbol{I}_p+t^2\boldsymbol{K})^{-1}\boldsymbol{1}_p\boldsymbol{1}_q^T \\ R_n\{\boldsymbol{I}_q+t^2(\frac{1}{n}\boldsymbol{L})\}^{-1}\boldsymbol{1}_q\boldsymbol{1}_p^T & \frac{1}{n}\{\boldsymbol{I}_q+t^2(\frac{1}{n}\boldsymbol{L})\}^{-1}\boldsymbol{L} \end{bmatrix}\\ &\times\begin{bmatrix} \boldsymbol{K}+u^2\boldsymbol{I}_p+Q_n\boldsymbol{1}_p^{\otimes 2} & R_n\boldsymbol{1}_p\boldsymbol{1}_q^T \\ R_n\boldsymbol{1}_q\boldsymbol{1}_p^T & \frac{1}{n}\boldsymbol{L}+u^2\boldsymbol{I}_q+T_n\boldsymbol{1}_q^{\otimes 2} \end{bmatrix}^{-1} dt\,du.\end{aligned} \tag{37}$$

Note that, as consequences of (36), we have

$$\boldsymbol{I}_p=\frac{4}{\pi^2}\int_0^\infty\int_0^\infty(\boldsymbol{I}_p+t^2\boldsymbol{K})^{-1}\boldsymbol{K}\,(\boldsymbol{K}+u^2\boldsymbol{I}_p)^{-1}\,dt\,du$$

and

$$\boldsymbol{I}_q=\frac{4}{\pi^2}\int_0^\infty\int_0^\infty\{\boldsymbol{I}_q+t^2(\tfrac{1}{n}\boldsymbol{L})\}^{-1}(\tfrac{1}{n}\boldsymbol{L})\{(\tfrac{1}{n}\boldsymbol{L})+u^2\boldsymbol{I}_q\}^{-1}\,dt\,du.$$

Then straightforward, albeit long-winded, matrix algebra leads to

$$\begin{aligned}
&\frac{\pi}{2}\big(\boldsymbol{M}_{n,\infty}^{-1/2}\boldsymbol{M}_n^{1/2}-\boldsymbol{I}_{p+q}\big)\\
&\qquad=\begin{bmatrix}\boldsymbol{K}^{-1/2}\displaystyle\int_0^\infty \boldsymbol{F}_{11n}(u;\boldsymbol{K},\boldsymbol{L})\,du & \boldsymbol{K}^{-1/2}\displaystyle\int_0^\infty \boldsymbol{F}_{12n}(u;\boldsymbol{K},\boldsymbol{L})\,du\\ \boldsymbol{L}^{-1/2}\displaystyle\int_0^\infty \boldsymbol{F}_{21n}(u;\boldsymbol{K},\boldsymbol{L})\,du & \boldsymbol{L}^{-1/2}\displaystyle\int_0^\infty \boldsymbol{F}_{22n}(u;\boldsymbol{K},\boldsymbol{L})\,du\end{bmatrix}
\end{aligned}\tag{38}$$

where, for example,

$$\begin{aligned}
\boldsymbol{F}_{11n}(u;\boldsymbol{K},\boldsymbol{L})&\equiv\boldsymbol{K}\,\boldsymbol{\Gamma}_{4n}(u)\boldsymbol{\Gamma}_{2n}(u)(\boldsymbol{K}+u^2\boldsymbol{I}_p)^{-1}-\boldsymbol{\Gamma}_{2n}(u)\boldsymbol{\Gamma}_{4n}(u),\\
\boldsymbol{F}_{21n}(u;\boldsymbol{K},\boldsymbol{L})&\equiv n^{1/2}\,R_n\,u^2\,\boldsymbol{\Gamma}_{1n}(u)\boldsymbol{1}_q\boldsymbol{1}_p^T\boldsymbol{\Gamma}_{4n}(u),\\
\boldsymbol{F}_{22n}(u;\boldsymbol{K},\boldsymbol{L})&\equiv n^{1/2}\Big[(\tfrac{1}{n}\boldsymbol{L})\boldsymbol{\Gamma}_{5n}(u)\boldsymbol{\Gamma}_{3n}(u)\{(\tfrac{1}{n}\boldsymbol{L})+u^2\boldsymbol{I}_q\}^{-1}+T_n\boldsymbol{1}_q\boldsymbol{1}_q^T\boldsymbol{\Gamma}_{5n}(u)\\
&\qquad\qquad -R_n^2\boldsymbol{1}_q\boldsymbol{1}_p^T\boldsymbol{\Gamma}_{4n}(u)\boldsymbol{1}_p\boldsymbol{1}_q^T\boldsymbol{\Gamma}_{1n}(u)\Big]
\end{aligned}$$

with

$$\begin{aligned}
&\boldsymbol{\Gamma}_{1n}(u)\equiv\big\{(\tfrac{1}{n}\boldsymbol{L})+u^2\boldsymbol{I}_q+T_n\boldsymbol{1}_q\boldsymbol{1}_q^T\big\}^{-1},\quad \boldsymbol{\Gamma}_{2n}(u)\equiv\{R_n^2\boldsymbol{1}_q^T\boldsymbol{\Gamma}_{1n}(u)\boldsymbol{1}_q-Q_n\}\boldsymbol{1}_p\boldsymbol{1}_p^T,\\
&\boldsymbol{\Gamma}_{3n}(u)\equiv\{R_n^2\boldsymbol{1}_p^T(\boldsymbol{K}+u^2\boldsymbol{I}_p+Q_n\boldsymbol{1}_p\boldsymbol{1}_p^T)^{-1}\boldsymbol{1}_p-T_n\}\boldsymbol{1}_q\boldsymbol{1}_q^T,\\
&\boldsymbol{\Gamma}_{4n}(u)\equiv\Big\{\boldsymbol{K}+u^2\boldsymbol{I}_p-\boldsymbol{\Gamma}_{2n}(u)\Big\}^{-1}\quad\text{and}\quad\boldsymbol{\Gamma}_{5n}(u)\equiv\Big\{(\tfrac{1}{n}\boldsymbol{L})+u^2\boldsymbol{I}_q-\boldsymbol{\Gamma}_{3n}(u)\Big\}^{-1}.
\end{aligned}$$

Let $\lambda_{\min}(\boldsymbol{K})$ and $\lambda_{\max}(\boldsymbol{K})$ denote, respectively, the smallest and largest eigenvalues of $\boldsymbol{K}$ and let $\lambda_{\min}(\boldsymbol{L})$ and $\lambda_{\max}(\boldsymbol{L})$ be defined similarly for $\boldsymbol{L}$. Since $Q_n=o_P(1)$ and $T_n=o_P(n^{-1})$ for every $0<\varepsilon\le 1$ we can choose n large enough so that $|Q_n|<\lambda_{\min}(\boldsymbol{K})/2$ and $|T_n|<\lambda_{\min}(\boldsymbol{L})/(2n)$ with probability exceeding $1-\varepsilon$. Standard steps then lead to the following spectral norm bounds for all sufficiently large n:

$$\begin{aligned}
&\|\boldsymbol{\Gamma}_{1n}(u)\|_S<\frac{1}{\frac{1}{2n}\lambda_{\min}(\boldsymbol{L})+u^2},\quad \|\boldsymbol{\Gamma}_{2n}(u)\|_S<p\left\{\frac{2qnR_n^2}{\lambda_{\min}(\boldsymbol{L})}+|Q_n|\right\},\\
&\|\boldsymbol{\Gamma}_{3n}(u)\|_S<\frac{pqR_n^2}{\frac{1}{2}\lambda_{\min}(\boldsymbol{K})+u^2}+q|T_n|,\quad \|\boldsymbol{\Gamma}_{4n}(u)\|_S<\frac{1}{\frac{1}{2}\lambda_{\min}(\boldsymbol{K})+u^2}\\
&\text{and}\quad \|\boldsymbol{\Gamma}_{5n}(u)\|_S<\frac{1}{\frac{1}{2n}\lambda_{\min}(\boldsymbol{L})+u^2}\quad\text{for all } u>0.
\end{aligned}$$

Then for all n large enough and $u > 0$ we have

$$\begin{aligned}\|\boldsymbol{F}_{11n}(u;\boldsymbol{K},\boldsymbol{L})\|_s &\leq \|\boldsymbol{K}\|_s\|\boldsymbol{\Gamma}_{4n}(u)\|_s\|\boldsymbol{\Gamma}_{2n}(u)\|_s\|(\boldsymbol{K}+u^2\boldsymbol{I}_p)^{-1}\|_s \\ &\quad +\|\boldsymbol{\Gamma}_{2n}(u)\|_s\|\boldsymbol{\Gamma}_{4n}(u)\|_s \\ &< p\left\{\frac{2qnR_n^2}{\lambda_{\min}(\boldsymbol{L})}+|Q_n|\right\}\left\{\frac{\lambda_{\max}(\boldsymbol{K})}{\lambda_{\min}(\boldsymbol{K})+u^2}+1\right\}\left\{\frac{1}{\frac{1}{2}\lambda_{\min}(\boldsymbol{K})+u^2}\right\}.\end{aligned}$$

It follows that, for all sufficiently large n,

$$\begin{aligned}\left\|\int_0^\infty \boldsymbol{F}_{11n}(u;\boldsymbol{K},\boldsymbol{L})\,du\right\|_s &\leq \int_0^\infty \|\boldsymbol{F}_{11n}(u;\boldsymbol{K},\boldsymbol{L})\|_s\,du \\ &< p\left\{\frac{2qnR_n^2}{\lambda_{\min}(\boldsymbol{L})}+|Q_n|\right\}\int_0^\infty\left\{\frac{\lambda_{\max}(\boldsymbol{K})}{\lambda_{\min}(\boldsymbol{K})+u^2}+1\right\}\left\{\frac{1}{\frac{1}{2}\lambda_{\min}(\boldsymbol{K})+u^2}\right\}du\end{aligned}$$

with probability exceeding $1-\varepsilon$. Since $R_n = O_P(n^{-1})$, $Q_n = o_p(1)$ and ε is arbitrary we then have $\|\int_0^\infty \boldsymbol{F}_{11n}(u;\boldsymbol{K},\boldsymbol{L})\,du\|_s \xrightarrow{P} 0$ as $n\to\infty$. Similar steps lead to $\|\int_0^\infty \boldsymbol{F}_{12n}(u;\boldsymbol{K},\boldsymbol{L})\,du\|_s \xrightarrow{P} 0$, $\|\int_0^\infty \boldsymbol{F}_{22n}(u;\boldsymbol{K},\boldsymbol{L})\,du\|_s \xrightarrow{P} 0$ and $\|\int_0^\infty \boldsymbol{F}_{21n}(u;\boldsymbol{K},\boldsymbol{L})\,du\|_s \xrightarrow{P} 0$ and the lemma is proven.

A.8. Final steps

For likelihood situations, standard results concerning asymptotic normality of maximum likelihood estimators give

$$\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)^{-1}\}^{-1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0) \xrightarrow{\mathcal{D}} N(\boldsymbol{0},\boldsymbol{I}) \tag{39}$$

where $\widehat{\boldsymbol{\theta}} = [(\widehat{\boldsymbol{\beta}}_{\mathrm{A}})^T\ \mathrm{vech}(\widehat{\boldsymbol{\Sigma}})^T\ (\widehat{\boldsymbol{\beta}}_{\mathrm{B}})^T]^T$ and $\boldsymbol{\theta}^0 = [(\boldsymbol{\beta}_{\mathrm{A}}^0)^T\ \mathrm{vech}(\boldsymbol{\Sigma}^0)^T\ (\beta_{\mathrm{B}}^0)^T]^T$. The general quasi-likelihood situation, requires asymptotic normality theory for M-estimators as treated in, for example, Section 5.3 of van der Vaart (1998). It follows from (39) that, for all $(d_{\mathrm{A}}+d_{\mathrm{A}}^{\boxplus}+d_{\mathrm{B}})\times 1$ vectors $\boldsymbol{a}\neq\boldsymbol{0}$, we have

$$\boldsymbol{a}^T\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)^{-1}\}^{-1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0) \xrightarrow{\mathcal{D}} N(0,\boldsymbol{a}^T\boldsymbol{a}).$$

As a consequence

$$\boldsymbol{a}^T\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)_\infty^{-1}\}^{-1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)+r_{mn}(\boldsymbol{a}) \xrightarrow{\mathcal{D}} N(0,\boldsymbol{a}^T\boldsymbol{a}) \tag{40}$$

where

$$\begin{aligned}r_{mn}(\boldsymbol{a}) &\equiv \boldsymbol{a}^T[\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)^{-1}\}^{-1/2}-\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)_\infty^{-1}\}^{-1/2}](\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0) \\ &= \boldsymbol{a}^T[\boldsymbol{I}-\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)_\infty^{-1}\}^{-1/2}\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)^{-1}\}^{1/2}] \\ &\quad \times\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)^{-1}\}^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0) \\ &= \Big([\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)_\infty^{-1}\}^{-1/2}\{I(\boldsymbol{\beta}_{\mathrm{A}}^0,\mathrm{vech}(\boldsymbol{\Sigma}^0),\beta_{\mathrm{B}}^0)^{-1}\}^{1/2}-\boldsymbol{I}]^T\boldsymbol{a}\Big)^T\boldsymbol{Z}\end{aligned}$$

and $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_{d_{\rm A}+d^{\boxplus}_{\rm A}+d_{\rm B}})$. Then note that

$$\begin{aligned}&\Big\| \big[\{I\big(\boldsymbol{\beta}^0_{\rm A}, \text{vech}(\boldsymbol{\Sigma}^0), \beta^0_{\rm B}\big)^{-1}_{\infty}\}^{-1/2}\{I\big(\boldsymbol{\beta}^0_{\rm A}, \text{vech}(\boldsymbol{\Sigma}^0), \beta^0_{\rm B}\big)^{-1}\}^{1/2} - \boldsymbol{I}\big]^T \boldsymbol{a}\Big\|_{F} \\ &\qquad \leq \big\|\{I\big(\boldsymbol{\beta}^0_{\rm A}, \text{vech}(\boldsymbol{\Sigma}^0), \beta^0_{\rm B}\big)^{-1}_{\infty}\}^{-1/2}\{I\big(\boldsymbol{\beta}^0_{\rm A}, \text{vech}(\boldsymbol{\Sigma}^0), \beta^0_{\rm B}\big)^{-1}\}^{1/2} - \boldsymbol{I}\big\|_{F} \, \|\boldsymbol{a}\|_{F}.\end{aligned}$$

From Lemma 2, as $n \to \infty$,

$$\|\{I\big(\boldsymbol{\beta}^0_{\rm A}, \text{vech}(\boldsymbol{\Sigma}^0), \beta^0_{\rm B}\big)^{-1}_{\infty}\}^{-1/2}\{I\big(\boldsymbol{\beta}^0_{\rm A}, \text{vech}(\boldsymbol{\Sigma}^0), \beta^0_{\rm B}\big)^{-1}\}^{1/2} - \boldsymbol{I}\|_{F} \xrightarrow{P} 0 \tag{41}$$

and so

$$\big[\{I\big(\boldsymbol{\beta}^0_{\rm A}, \text{vech}(\boldsymbol{\Sigma}^0), \beta^0_{\rm B}\big)^{-1}_{\infty}\}^{-1/2}\{I\big(\boldsymbol{\beta}^0_{\rm A}, \text{vech}(\boldsymbol{\Sigma}^0), \beta^0_{\rm B}\big)^{-1}\}^{1/2} - \boldsymbol{I}\big]\boldsymbol{a} \xrightarrow{P} 0.$$

Application of Slutsky's Theorem then gives $r_{mn}(\boldsymbol{a}) \xrightarrow{P} 0$. From (40) and another application of Slutsky's Theorem we have

$$\boldsymbol{a}^T\,\{I\big(\boldsymbol{\beta}^0_{\rm A}, \text{vech}(\boldsymbol{\Sigma}^0), \beta^0_{\rm B}\big)^{-1}_{\infty}\}^{-1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{a}^T\boldsymbol{a}).$$

Theorem 1 then follows from the Cramér-Wold Device and the Continuous Mapping Theorem.

B. Proof of Theorem 2

Let $\boldsymbol{x}_{ik}$, $1 \leq k \leq s$, be the support points for the ith group, $1 \leq i \leq m$. Symmetry arguments lead to the restriction $\boldsymbol{x}_{ik} = \boldsymbol{x}_k$, $1 \leq i \leq m$, for optimal designs. For each of the m groups let

$$n_k \equiv \text{number of } \boldsymbol{x}_k \text{ values in the design}, \ 1 \leq k \leq s,$$

and let $Y^{[k]}_{ij}$ be the jth response within the ith group at support point $\boldsymbol{x}_k$. Then

$$\begin{aligned}&Y^{[k]}_{ij}|U_i \text{ independent having quasi-likelihood function (2) with natural parameter} \\ &\beta^0_0 + (\boldsymbol{\beta}^0_1)^T\boldsymbol{x}_k + U_i \text{ such that the } U_i \text{ are independent } N(0, (\sigma^2)^0).\end{aligned} \tag{42}$$

for $1 \leq i \leq m$, $1 \leq j \leq n_k$ and $1 \leq k \leq s$. The log-quasi-likelihood is

$$\begin{aligned}\ell(\beta_0, \boldsymbol{\beta}_{\rm B}, \sigma^2) = &\sum_{i=1}^{m}\sum_{k=1}^{s}\sum_{j=1}^{n_k}\{Y^{[k]}_{ij}(\beta_0 + \boldsymbol{\beta}^T_{\rm B}\boldsymbol{x}_k)/\phi + d(Y^{[k]}_{ij}, \phi)\} - \frac{m}{2}\log(2\pi\sigma^2) \\ &+ \sum_{i=1}^{m}\log\int_{-\infty}^{\infty}\exp\left[\sum_{k=1}^{s}\sum_{j=1}^{n_k}\frac{1}{\phi}\left\{Y^{[k]}_{ij}u - b(\beta_0 + \boldsymbol{\beta}^T_{\rm B}\boldsymbol{x}_k + u)\right\} - \frac{u^2}{2\sigma^2}\right]du.\end{aligned}$$

Let

$$\widetilde{\mathcal{H}}_{ri} \equiv \sum_{k=1}^{s} n_k \boldsymbol{x}_k^{\otimes r} b''(\beta_0 + \boldsymbol{\beta}^T_{\rm B}\boldsymbol{x}_k + U_i).$$

Then arguments analogous to those given in the proof of Theorem 1 lead to the Fisher information matrix having the following form as $m, n \to \infty$:

$$I(\beta_0, \boldsymbol{\beta}_{\text{B}}, \sigma^2) = \begin{bmatrix} \dfrac{m}{\sigma^2} + O(mn^{-1}) & O(m)\mathbf{1}_{d_{\text{B}}}^T & O(mn^{-1}) \\[2ex] O(m)\mathbf{1}_{d_{\text{B}}} & \dfrac{m}{\phi} E\left(\widetilde{\mathcal{H}}_{21} - \dfrac{\widetilde{\mathcal{H}}_{11}^{\otimes 2}}{\widetilde{\mathcal{H}}_{01}}\right) + O(m)\mathbf{1}_{d_{\text{B}}}^{\otimes 2} & O(m)\mathbf{1}_{d_{\text{B}}} \\[2ex] O(mn^{-1}) & O(m)\mathbf{1}_{d_{\text{B}}}^T & \dfrac{m}{2\sigma^4} + O(mn^{-1}) \end{bmatrix}.$$

Next, change the ordering of the parameters from $(\beta_0, \boldsymbol{\beta}_{\text{B}}, \sigma^2)$ to $(\beta_0, \sigma^2, \boldsymbol{\beta}_{\text{B}})$ and apply a standard result concerning the determinant of a 2 × 2 block-partitioned matrix (e.g. Theorem 13.3.8 of Harville (2008)) to obtain

$$\left|I(\beta_0, \boldsymbol{\beta}_{\text{B}}, \sigma^2)\right| = \left|\widetilde{\boldsymbol{A}}_{11}\right|\left|\widetilde{\boldsymbol{A}}_{22} - \widetilde{\boldsymbol{A}}_{12}^T \widetilde{\boldsymbol{A}}_{11}^{-1} \widetilde{\boldsymbol{A}}_{12}\right|$$

where $\widetilde{\boldsymbol{A}}_{11}$ and $\widetilde{\boldsymbol{A}}_{12}$ have forms analogous to those of $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{12}$ in (28) and

$$\widetilde{\boldsymbol{A}}_{22} \equiv \frac{m}{\phi} E\left(\widetilde{\mathcal{H}}_{21} - \frac{\widetilde{\mathcal{H}}_{11}^{\otimes 2}}{\widetilde{\mathcal{H}}_{01}}\right) + O(m)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}.$$

It is easily verified that $\left|\widetilde{\boldsymbol{A}}_{11}\right| = m^2/(2\sigma^6) + O(m^2 n^{-1})$ and $\widetilde{\boldsymbol{A}}_{12}^T \widetilde{\boldsymbol{A}}_{11}^{-1} \widetilde{\boldsymbol{A}}_{12} = O(m)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}$. It follows that

$$\frac{2\phi^{d_{\text{B}}}\sigma^6\{1 + O(n^{-1})\}}{m^{d_{\text{B}}+2}} \left|I(\beta_0, \boldsymbol{\beta}_{\text{B}}, \sigma^2)\right| = \left|\boldsymbol{\Psi}_n + O(1)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}\right| \tag{43}$$

where $\boldsymbol{\Psi}_n \equiv E\left(\widetilde{\mathcal{H}}_{21} - \widetilde{\mathcal{H}}_{11}^{\otimes 2}/\widetilde{\mathcal{H}}_{01}\right)$. Since $\widetilde{\mathcal{H}}_{01} = O_P(n)$, $\widetilde{\mathcal{H}}_{11} = O_P(n)\mathbf{1}$ and $\widetilde{\mathcal{H}}_{21} = O_P(n)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}$, we have $\boldsymbol{\Psi}_n = O(n)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}$. Let $\lambda_1(\boldsymbol{M}), \ldots, \lambda_{d_{\text{B}}}(\boldsymbol{M})$ denote the eigenvalues of a generic $d_{\text{B}} \times d_{\text{B}}$ matrix $\boldsymbol{M}$. Then $\left|\boldsymbol{\Psi}_n + O(1)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}\right| = \prod_{j=1}^{d_{\text{B}}} \lambda_j\left(\boldsymbol{\Psi}_n + O(1)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}\right)$. As a consequence of Theorem 8.1.4 (Wielandt-Hoffman) of Golub and Van Loan (2013) , $\lambda_j\left(\boldsymbol{\Psi}_n + O(1)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}\right) = \lambda_j\left(\boldsymbol{\Psi}_n\right) + O(1)$ for each $1 \le j \le d_{\text{B}}$. Hence

$$\left|\boldsymbol{\Psi}_n + O(1)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}\right| = \left|\boldsymbol{\Psi}_n\right| + O(1)\sum_{j=1}^{d_{\text{B}}} \left|\boldsymbol{\Psi}_n\right|/\lambda_j(\boldsymbol{\Psi}_n). \tag{44}$$

To obtain the order of magnitude of the $\lambda_j(\boldsymbol{\Psi}_n)$ we appeal to Theorem 8.1.3 (Gershgorin) of Golub and Van Loan (2013). Since all entries of $\boldsymbol{\Psi}_n$ are $O(n)$, the same is true for the lower and upper limits of each of the Gershgorin discs of $\boldsymbol{\Psi}_n$. Since each eigenvalue of $\boldsymbol{\Psi}_n$ is inside at least one Gershgorin disc, we have $\lambda_j(\boldsymbol{\Psi}_n) = O(n)$, $1 \le j \le d_{\text{B}}$. It follows from this fact and (44) that $\left|\boldsymbol{\Psi}_n + O(1)\mathbf{1}_{d_{\text{B}}}^{\otimes 2}\right| = \left|\boldsymbol{\Psi}_n\right|\{1 + o(1)\}$. In view of (43), the determinant of $I(\beta_0, \boldsymbol{\beta}_{\text{B}}, \sigma^2)$ is proportional to a quantity with leading term $\left|\boldsymbol{\Psi}_n\right|$ as $n \to \infty$. Recalling that $n_k = ns\delta_k$ and dividing through by ns we can assert that approximate locally D-optimal designs, based on the exact leading term behaviour of the determinant of the

Fisher information matrix, are those which maximise

$$\left| E\left[\sum_{k=1}^{s} \delta_k \boldsymbol{x}_k^{\otimes 2} b''(\beta_0 + \boldsymbol{\beta}_{\mathrm{B}}^T \boldsymbol{x}_k + U) - \frac{\left\{ \sum_{k=1}^{s} \delta_k \boldsymbol{x}_k b''(\beta_0 + \boldsymbol{\beta}_{\mathrm{B}}^T \boldsymbol{x}_k + U) \right\}^{\otimes 2}}{\sum_{k=1}^{s} \delta_k b''(\beta_0 + \boldsymbol{\beta}_{\mathrm{B}}^T \boldsymbol{x}_k + U)} \right] \right|, \qquad (45)$$

$U \sim N(0, \sigma^2)$, over the design weights δ_k and support points $\boldsymbol{x}_k$, $1 \leq k \leq s$. Except for an innocuous normalizing factor, the integral appearing in (7) can be shown to equal the expectation appearing in (45) and the theorem is proven.

Acknowledgements

We are grateful to Iain Johnstone, Luca Maestrini, Song Mei, Tung Pham, Andy Wood, an associate editor and two referees for advice related to this research. This research was partially supported by the U.S. National Science Foundation grants DMS-1510219 and DMS-1713120 and the Australian Research Council Discovery grants CE140100049 and DP140100441.

References

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67(1)**, 1–48.

Bhaskaran, A. (2022). *Likelihood Theory and Methods for Generalized Linear Mixed Models.* Doctor of Philosophy thesis, University of Technology Sydney, Australia.

Chipman, J.S. (1964). On least squares with insufficient observations. *Journal of the American Statistical Association*, **59**, 1078–1111.

Faraway, J.J. (2016). *Extending the Linear Model with R. Second Edition.* Boca Raton, Florida: CRC Press.

Golub, G.H. and Van Loan, C.F. (2013). *Matrix Computation, Fourth Edition.* Baltimore: The Johns Hopkins University Press.

Hall, P., Pham, T., Wand, M.P. and Wang, S.S.J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *The Annals of Statistics*, **39**, 2502–2532.

Harville, D.A. (2008). *Matrix Algebra from a Statistician's Perspective.* New York: Springer.

Higham, N.J. (2008). *Functions of Matrices.* Philadelphia: Society for Industrial and Applied Mathematics.

Hui, F.K.C., Warton, D.I., Ormerod, J.T., Haapaniemi, V. and Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, **26**, 35–43.

Jeon, M., Rijmen, F. and Rabe-Hesketh, S. (2017). A variational maximization-maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika*, **3**, 693–716.

Jiang, J. (2017). *Asymptotic Analysis of Mixed Effects Models.* Boca Raton, Florida: CRC Press.

Magnus, J.R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *The Annals of Statistics*, **7**, 381–394.

Magnus, J.R. and Neudecker, H. (1999). *Matrix Differential Calculus. Revised Edition.* Chichester, U.K.: John Wiley & Sons.

McCulloch, C.E., Searle, S.R. and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models. Second Edition.* New York: John Wiley & Sons.

Miyata, Y. (2004). Fully exponential Laplace approximation using asymptotic modes. *Journal of the American Statistical Association*, **99**, 1037–1049.

Nie, L. (2007). Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: theory and applications. *Journal of Statistical Planning and Inference*, **137**, 1787–1804.

Ormerod, J.T. and Wand, M.P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, **21**, 2–17.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Russell, K.G. (2018). *Design of Experiments for Generalized Linear Models.* Boca Raton, Florida: CRC Press.

Stroup, W.W. (2013). *Generalized Linear Mixed Models.* Boca Raton, Florida: CRC Press.

Tierney, L., Kass, R.E. and Kadane, J.B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, **84**, 710–716.

van der Vaart, A.W. (2013). *Asymptotic Statistics.* Cambridge, U.K.: Cambridge University Press.

Waite, T.W. and Woods, D.C. (2015). Designs for generalized linear models with random block effects via information matrix approximations. *Biometrika*, **102**, 677–693.

Wang, B. and Titterington, D.M. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Proceedings of the 10th International Workshop of Artificial Intelligence and Statistics* (eds. R.G. Cowell & Z. Ghahramani), 373–380.

Zhang, W., Mandal, A. and Stufken, J. (2017). Approximations of the Fisher information matrix for a panel mixed logit model. *Journal of Statistical Theory and Practice*, **39**, 269–295.