Simple Review of Gaussian Variational Approximate Inference

Daeyoung Lim*
Department of Statistics
Korea University

January 24, 2016

1 Figure 2

This section reviews the figure plotted in p.13 of the associated paper.

Variational Coefficients					
β_0	β_1	β_2	β_3	β_4	β_5
-1.325	0.883	-0.933	0.481	-0.160	0.339

GVA Standard Errors					
β_0	β_1	β_2	β_3	β_4	β_5
1.179	0.131	0.400	0.346	0.055	0.203

- Coefficient estimates via GVA seem to align well with the plot given in the paper.
- The caption under the *confidence interval* plots should be a misnomer since interval estimation is called the *credible interval* in Bayesian frameworks. Requires revision.
- The credible intervals plotted in box and whisker plots are consistent with the numerical results calculated with the results above.
- *Adaptive Gauss-Hermite quadrature* and *penalized quasi-likelihood* methods were unavailable.

2 Figure 3

This section reviews the code for figure 3 in p.14.

- The code returned an error that the number of *adaptive Gauss-Hermite quadrature* greater than 1 is only available for models with a single, scalar random-effects term.
- The line of code `glmer(formula=y Base*trt+Age+Visit+(Visit|subject), data=epil2, family = poisson, nAGQ = 20)` is the source of this error.

*Prof. Taeryon Choi

3 Figure 4

This section reviews the code for figure 4 in p.15.

Variational Coefficients			
β_0	β_1	β_2	β_3
-1.434	-0.134	-0.377	-0.130

GVA Standard Errors			
β_0	β_1	β_2	β_3
0.386	0.526	0.043	0.065

- R package {glmmAK} is deprecated under R version 3.2.3 (2015-12-10) – "Wooden Christmas-Tree".
- The required dataset '`toenail`' in R package {glmmAK} was downloaded from https://github.com/cran/glm
- The results all match what is written in the paper.

4 Model review

Model Specifications	
Likelihood	$\boldsymbol{y}_i \vert \boldsymbol{u}_i \sim \exp\left\{\boldsymbol{y}_i^T\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)-\mathbf{1}_i^T b\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)+\mathbf{1}_i^T c(\boldsymbol{y}_i)\right\}$
Prior	$\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$

The log-marginal likelihood $p(\boldsymbol{y})$ is, with the abuse of notation, $\int \ell(\boldsymbol{u}, \boldsymbol{\beta}, \Sigma)\, d\boldsymbol{u}$ where

$$\begin{aligned}\ell(\boldsymbol{u}, \boldsymbol{\beta}, \Sigma) &= \log\left(\prod_{i=1}^m \exp\left\{\boldsymbol{y}_i^T\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)-\mathbf{1}_i^T b\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)+\mathbf{1}_i^T c(\boldsymbol{y}_i)\right\}\cdot\frac{1}{\sqrt{|2\pi\Sigma|}}e^{-\frac{1}{2}\boldsymbol{u}_i^T\Sigma^{-1}\boldsymbol{u}_i}\right)\\ &= \sum_{i=1}^m\left[\boldsymbol{y}_i^T\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)-\mathbf{1}_i^T b\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)+\mathbf{1}_i^T c(\boldsymbol{y}_i)-\frac{1}{2}\boldsymbol{u}_i^T\Sigma^{-1}\boldsymbol{u}_i-\frac{1}{2}\log|2\pi\Sigma|\right]\end{aligned}$$

Therefore,

$$\begin{aligned}\ell(\boldsymbol{\beta}, \Sigma) = &\sum_{i=1}^m\left[\boldsymbol{y}_i^T X_i\boldsymbol{\beta}+\mathbf{1}_i^T c(\boldsymbol{y}_i)-\frac{1}{2}\log|2\pi\Sigma|\right]\\ &+\sum_{i=1}^m\log\int_{\mathbb{R}^k}\exp\left\{\boldsymbol{y}_i^T Z_i\boldsymbol{u}_i-\mathbf{1}_i^T b\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)-\frac{1}{2}\boldsymbol{u}_i^T\Sigma^{-1}\boldsymbol{u}_i\right\}d\boldsymbol{u}\end{aligned}$$

The integral does not change even if we multiply the integrand by 1:

$$\int_{\mathbb{R}^k}\exp\left\{\boldsymbol{y}_i^T Z_i\boldsymbol{u}_i-\mathbf{1}_i^T b\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)-\frac{1}{2}\boldsymbol{u}_i^T\Sigma^{-1}\boldsymbol{u}_i\right\}\frac{\varphi_{\Lambda_i}(\boldsymbol{u}-\mu_i)}{\varphi_{\Lambda_i}(\boldsymbol{u}-\mu_i)}\,d\boldsymbol{u} \tag{1}$$

where $\varphi_{\Lambda_i}(\boldsymbol{u}-\mu_i)$ is the multivariate Gaussian p.d.f. of a random vector $\boldsymbol{u}$ with a mean vector μ_i and a covariance matrix Λ_i. Then (1) is rewritten in terms of the expectation with respect to $\boldsymbol{u}$:

$$(1) = \mathrm{E}_{\boldsymbol{u}}\left[\frac{\exp\left\{\boldsymbol{y}_i^T Z_i \boldsymbol{u}_i - \mathbf{1}_i^T b\left(X_i\boldsymbol{\beta} + Z_i\boldsymbol{u}_i\right) - \frac{1}{2}\boldsymbol{u}_i^T\Sigma^{-1}\boldsymbol{u}_i\right\}}{\varphi_{\Lambda_i}(\boldsymbol{u}-\mu_i)}\right]$$

According to *Jensen's inequality*, $\log\left(\mathrm{E}(X)\right) \geq \mathrm{E}\left(\log X\right)$, which thus yields

$$\begin{aligned}
&\log\left(\mathrm{E}_{\boldsymbol{u}}\left[\frac{\exp\left\{\boldsymbol{y}_i^T Z_i \boldsymbol{u}_i - \mathbf{1}_i^T b\left(X_i\boldsymbol{\beta} + Z_i\boldsymbol{u}_i\right) - \frac{1}{2}\boldsymbol{u}_i^T\Sigma^{-1}\boldsymbol{u}_i\right\}}{\varphi_{\Lambda_i}(\boldsymbol{u}-\mu_i)}\right]\right) \\
&\geq \mathrm{E}_{\boldsymbol{u}}\left[\boldsymbol{y}_i^T Z_i \boldsymbol{u}_i - \mathbf{1}_i^T b\left(X_i\boldsymbol{\beta} + Z_i\boldsymbol{u}_i\right) - \frac{1}{2}\boldsymbol{u}_i^T\Sigma^{-1}\boldsymbol{u}_i - \log\left(\varphi_{\Lambda_i}(\boldsymbol{u}-\mu_i)\right)\right]
\end{aligned}$$

The inequality gives us the lower bound of the variational approximation for the generalized linear mixed model.

$$\begin{aligned}
\underline{\ell}\left(\boldsymbol{\beta}, \Sigma, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right) = &\sum_{i=1}^{m}\left[\boldsymbol{y}_i^T X_i\boldsymbol{\beta} + \mathbf{1}_i^T c\left(\boldsymbol{y}_i\right) - \frac{1}{2}\log\left|2\pi\Sigma\right|\right] \\
&+ \sum_{i=1}^{m}\mathrm{E}_{\boldsymbol{u}}\left[\boldsymbol{y}_i^T Z_i \boldsymbol{u}_i - \mathbf{1}_i^T b\left(X_i\boldsymbol{\beta} + Z_i\boldsymbol{u}_i\right) - \frac{1}{2}\boldsymbol{u}_i^T\Sigma^{-1}\boldsymbol{u}_i - \log\left(\varphi_{\Lambda_i}(\boldsymbol{u}-\mu_i)\right)\right]
\end{aligned}$$

Computing the expectation is not hard except for the unspecified function $b(\cdot)$ which can be easily approximated via *Gauss-Hermite quadrature*. Therefore, we have a scalar function as the objective function and are only left with optimizing the function over variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ along with the original parameters $\boldsymbol{\beta}$ and Σ. The compuational formula for *Newton-Raphson* algorithm is given in the supplementary document. However, the real problem is making sense of why the formula works and how the original authors of the paper computed the necessary elements such as the Hessian matrix and gradient vector for each parameter.

Stuff yet to be resolved

- How to compute the abscissa of Hermite polynomials that are used in *Gauss-Hermite quadrature*.
- How to determine the number of nodes of *Gauss-Hermite quadrature*.
- How to compute the Hessian of $\mathrm{vech}\left(\Sigma\right)$ or other terms with $\mathrm{vech}\left(\cdot\right)$.
- Why the *Newton-Raphson* formula is written as what is given in the supplementary paper.

5 Examples

We will from now on follow the examples and try to calculate the gradient vectors and Hessian matrices whose general forms are given in the supplementary material. To achieve this, we first need to compute $\mathcal{B}^{(1)}\left(\boldsymbol{\beta}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i\right)$.

5.1 Poisson Mixed

Specifying the parametric family of the GLM adds clarity to how we should compute the derivatives necessary for Newton's method. As a gentle reminder, the general model was

$$\boldsymbol{y}_i|\boldsymbol{u}_i \sim \exp\left\{\boldsymbol{y}_i^T\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)-\mathbf{1}_i^T b\left(X_i\boldsymbol{\beta}+Z_i\boldsymbol{u}_i\right)+\mathbf{1}_i^T c(\boldsymbol{y}_i)\right\}.$$

In Poisson mixed model, $b(x)=e^x$ and $c(x)=-\log(x!)$.

$$B(\mu,\sigma^2)=\exp\left(\mu+\frac{1}{2}\sigma^2\right)$$
$$\mathcal{B}\left(\boldsymbol{\beta},\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i\right)=B\left(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}_i\boldsymbol{\mu}_i,\operatorname{dg}\left(\boldsymbol{Z}_i\boldsymbol{\Lambda}_i\boldsymbol{Z}_i^T\right)\right)$$

$\frac{\partial\mathcal{B}}{\partial\boldsymbol{\beta}}$ is our goal. For notational convenience, let's replace $\mathcal{B}$ with simple f.

$$\begin{aligned}df &= d\left(\exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right)\circ\exp\left\{\boldsymbol{Z}_i\boldsymbol{\mu}_i+\frac{1}{2}\operatorname{dg}\left(\boldsymbol{Z}_i\Lambda_i\boldsymbol{Z}_i^T\right)\right\}\right)\\ &= d\exp\left(\boldsymbol{X}_i\boldsymbol{\beta}\right)\circ\exp\left\{\boldsymbol{Z}_i\boldsymbol{\mu}_i+\frac{1}{2}\operatorname{dg}\left(\boldsymbol{Z}_i\Lambda_i\boldsymbol{Z}_i^T\right)\right\}\end{aligned}$$

where $\circ$ indicates *Hadamard product(or Schur product)* which is simply the elementwise product between two matrices(or vectors) of the same size. We differentiate the notation between Diag and dg. The former is applied to vectors and converts it into a diagonal matrix whose diagonal elements correspond to the elements of the vector. On the other hand, dg is applied to matrices and collects the diagonal entries and creates a column vector with them.

5.1.1 $\mathrm{D}_{\boldsymbol{\beta}}\underline{\ell}$

Let

$$\begin{aligned}s &= \boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}_i\boldsymbol{\mu}_i+\frac{1}{2}\operatorname{dg}\left(\boldsymbol{Z}_i\boldsymbol{\Lambda}_i\boldsymbol{Z}_i^T\right)\\ e &= \exp(s)\\ f &= \mathbf{1}^T e\end{aligned}$$

Then,

$$ds=\boldsymbol{X}_i d\boldsymbol{\beta}+\boldsymbol{Z}_i d\boldsymbol{\mu}_i+\frac{1}{2}\operatorname{dg}\left(\boldsymbol{Z}_i d\boldsymbol{\Lambda}_i\boldsymbol{Z}_i^T\right).$$

But since we are only interested in $d\boldsymbol{\beta}$, we set $d\boldsymbol{\mu}_i=0$, $d\boldsymbol{\Lambda}_i=0$.

$$\begin{aligned}de &= d\exp(s)\\ &= e\circ ds\\ &= e\circ\boldsymbol{X}_i\,d\boldsymbol{\beta}\end{aligned}$$

Therefore, $df=\mathbf{1}^T de=e^T\boldsymbol{X}_i d\boldsymbol{\beta}$.

$$\frac{\partial\underline{\ell}}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{m}\left(\boldsymbol{y}_i^T-e^T\right)\boldsymbol{X}_i$$

5.1.2 $\mathrm{D}_{\mathrm{vech}(\boldsymbol{\Sigma})}\, \underline{\ell}$

Note that $A : B$ where A and B are matrices of the same size, is the *Frobenius product*. Then we can rewrite $d\underline{\ell}$ as follows:

$$d\underline{\ell} = \frac{1}{2}\sum_{i=1}^{m}\left[d\left(\log\det\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i\right)\right) - d\left(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T : \boldsymbol{\Sigma}^{-1}\right) - d\left(\boldsymbol{\Lambda_i} : \boldsymbol{\Sigma}^{-1}\right)\right] \tag{2}$$

$$= \frac{1}{2}\sum_{i=1}^{m}\left[d\left(\mathrm{Tr}\log\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i\right)\right) - d\left(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T : \boldsymbol{\Sigma}^{-1}\right) - d\left(\boldsymbol{\Lambda}_i : \boldsymbol{\Sigma}^{-1}\right)\right] \tag{3}$$

$$= \frac{1}{2}\sum_{i=1}^{m}\left[\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i\right)^{-T} : \left(d\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i\right) - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T : d\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Lambda}_i : d\boldsymbol{\Sigma}^{-1}\right] \tag{4}$$

$$= \frac{1}{2}\sum_{i=1}^{m}\left[\left(\boldsymbol{\Sigma}^T\boldsymbol{\Lambda}_i^{-T}\right)\boldsymbol{\Lambda}_i^T - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T - \boldsymbol{\Lambda}_i\right] : d\boldsymbol{\Sigma}^{-1} \tag{5}$$

$$= \frac{1}{2}\sum_{i=1}^{m}\left[-\boldsymbol{\Sigma}^T + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Lambda}_i\right] : \boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} \tag{6}$$

$$= \frac{1}{2}\sum_{i=1}^{m}\left[\Sigma^{-T}\left(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \boldsymbol{\Lambda}_i\right)\boldsymbol{\Sigma}^{-T} - \boldsymbol{\Sigma}^{-T}\right] : d\boldsymbol{\Sigma} \tag{7}$$

- From equations (1) to (2), $\log\det(A) = \mathrm{Tr}\log(A)$ where $\log(A)$ is NOT the elementwise logarithm but rather the matrix logarithm defined in terms of the Taylor series of e^A.
- Let $f(X) = \log\det(X)$ where X is a positive definite matrix.

  $$\frac{\partial}{\partial X} f(X) = X^{-1}.$$

  This should not be surprising since $(\log x)' = 1/x$. Positive-definiteness is important since there should be a guarantee that there is no eigenvalue whose value is 0. (The determinant should not be 0 since log(0) is not defined.)
- In relation to the previous point,

  $$\frac{\partial}{\partial X}\mathrm{Tr}\, f(X) = \left(\left.\frac{d}{dx} f(x)\right|_{x=X}\right)^T$$

  We can rewrite $\log\det(X) = \mathrm{Tr}\log(X)$ whose derivatives are consistent with each other provided that X is positive definite.
- It is a well-proven fact that the differential of the inverse of a matrix is as follows:

  $$d\Sigma^{-1} = -\Sigma^{-1}d\Sigma\Sigma^{-1}.$$

- The *Frobenius product* $A : B$ can be described differently (in reference to the transition from eqn (4) to (5)):

  $$A : B = \mathrm{Tr}\left(B^T A\right).$$

  Since trace operator is not affected by transposition,

  $$A : BC = \mathrm{Tr}\left(C^T B^T A\right) = \mathrm{Tr}\left(A^T BC\right) = B^T A : C.$$

- Another variant of the *Frobenius product* is

$$A : B = \operatorname{vec}(A)^T \operatorname{vec}(B).$$

Using $\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}$ and $\operatorname{vec}(X) = D_p \operatorname{vech}(X)$, we can recast the last line as follows:

$$d\ell = \frac{1}{2} \sum_{i=1}^{m} \operatorname{vec}\left[\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i \boldsymbol{\mu}_i^T + \boldsymbol{\Lambda}_i\right)\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right]^T \left(D_K \operatorname{vech}(d\boldsymbol{\Sigma})\right)$$

$$\therefore \frac{\partial \underline{\ell}}{\partial \operatorname{vech}(\boldsymbol{\Sigma})} = \frac{1}{2} \sum_{i=1}^{m} \operatorname{vec}\left\{\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i \boldsymbol{\mu}_i^T + \boldsymbol{\Lambda}_i\right)\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right\}^T D_K$$

5.1.3 $\mathrm{D}_{\boldsymbol{\mu}_i} \underline{\ell}$

This is similar to section 5.1.1 $\mathrm{D}_{\boldsymbol{\beta}} \underline{\ell}$.

$$\frac{\partial f}{\partial \boldsymbol{\mu}_i} = e^T \boldsymbol{Z}_i d\boldsymbol{\mu}_i$$

$$\therefore \frac{\partial \underline{\ell}}{\partial \boldsymbol{\mu}_i} = (y_i - e)^T \boldsymbol{Z}_i - \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}$$

5.1.4 $\mathrm{D}_{\operatorname{vech}(\boldsymbol{\Lambda}_i)} \underline{\ell}$

Collecting just the terms that contain $\boldsymbol{\Lambda}_i$,

$$d\underline{\ell} = -d\left(\mathbf{1}_i^T \exp\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_i + \frac{1}{2}\operatorname{dg}\left(\boldsymbol{Z}_i\boldsymbol{\Lambda}_i\boldsymbol{Z}_i^T\right)\right)\right) + \frac{1}{2}\left\{d\log\det\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i\right) - d\left(\boldsymbol{\Sigma}^{-1} : \boldsymbol{\Lambda}_i\right)\right\}$$

As per the first term, for convenience, let's write

$$\begin{aligned} s &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\mu}_i + \frac{1}{2}\operatorname{dg}\left(\boldsymbol{Z}_i\boldsymbol{\Lambda}_i\boldsymbol{Z}_i\right) \\ e &= \exp(s) \\ f &= \mathbf{1}^T e \end{aligned}$$

Then,

$$\begin{aligned} ds &= \boldsymbol{X}_i d\boldsymbol{\beta} + \boldsymbol{Z}_i d\boldsymbol{\mu}_i + \frac{1}{2}\operatorname{dg}\left(\boldsymbol{Z}_i d\boldsymbol{\Lambda}_i \boldsymbol{Z}_i^T\right) \\ de &= e \circ ds \\ df &= \mathbf{1}^T de \\ &= \mathbf{1}^T (e \circ ds) \\ &= e^T ds \\ &= e^T\left(\boldsymbol{X}_i d\boldsymbol{\beta} + \boldsymbol{Z}_i d\boldsymbol{\mu}_i + \frac{1}{2}\operatorname{dg}\left(\boldsymbol{Z}_i d\boldsymbol{\Lambda}_i \boldsymbol{Z}_i^T\right)\right) \end{aligned}$$

Since we are working on the gradient with respect to $\boldsymbol{\Lambda}_i$, set $d\boldsymbol{\beta} = 0$ and $d\boldsymbol{\mu}_i = 0$.

$$\begin{aligned} df &= \frac{1}{2} e^T \operatorname{dg}\left(\boldsymbol{Z}_i d\boldsymbol{\Lambda}_i \boldsymbol{Z}_i^T\right) \\ &= \frac{1}{2} \operatorname{Diag}\left(e^T\right) : \boldsymbol{Z}_i d\boldsymbol{\Lambda}_i \boldsymbol{Z}_i^T \\ &= \frac{1}{2} E : \boldsymbol{Z}_i d\boldsymbol{\Lambda}_i \boldsymbol{Z}_i^T \\ &= \frac{1}{2} \boldsymbol{Z}_i^T E \boldsymbol{Z}_i : d\boldsymbol{\Lambda}_i \end{aligned}$$

The second term is rather easy.

$$\begin{aligned} d \log\det\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i\right) &= d \operatorname{Tr}\log\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i\right) \\ &= \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i\right)^{-T} : d\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i \\ &= \boldsymbol{\Sigma}^{-T}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_i\right)^{-T} : d\boldsymbol{\Lambda}_i \\ &= \boldsymbol{\Lambda}_i^{-T} : d\boldsymbol{\Lambda}_i \end{aligned}$$

Combining all three terms,

$$\begin{aligned} d\underline{\ell} &= -\frac{1}{2}\boldsymbol{Z}_i^T E \boldsymbol{Z}_i : d\boldsymbol{\Lambda}_i + \frac{1}{2}\left[\boldsymbol{\Lambda}_i^{-T} : d\boldsymbol{\Lambda}_i - \boldsymbol{\Sigma}^{-1} : d\boldsymbol{\Lambda}_i\right] \\ &= \frac{1}{2}\left(\boldsymbol{\Lambda}_i^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{Z}_i^T E \boldsymbol{Z}_i\right) : d\boldsymbol{\Lambda}_i \end{aligned}$$

Using the vec operator, the differential form translates to

$$d\underline{\ell} = \frac{1}{2}\operatorname{vec}\left(\boldsymbol{\Lambda}_i^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{Z}_i^T E \boldsymbol{Z}_i\right)^T D_K \operatorname{vech}\left(d\boldsymbol{\Lambda}_i\right)$$

$$\therefore \frac{\partial \underline{\ell}}{\partial \operatorname{vech}\left(\boldsymbol{\Lambda}_i\right)} = \frac{1}{2}\operatorname{vec}\left(\boldsymbol{\Lambda}_i^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{Z}_i^T E \boldsymbol{Z}_i\right)^T D_K$$

5.1.5 $\mathrm{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}\,\underline{\ell}$

Recall that

$$\frac{\partial \underline{\ell}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m} (\boldsymbol{y} - e)^T \boldsymbol{X}_i.$$

We set $f = \frac{\partial \underline{\ell}}{\partial \boldsymbol{\beta}}$.

$$\begin{aligned} df &= -\sum_{i=1}^{m} (de)^T \boldsymbol{X}_i \\ &= -\sum_{i=1}^{m} (e \circ \boldsymbol{X}_i d\boldsymbol{\beta})^T \boldsymbol{X}_i \end{aligned}$$

Here we should note that between two vectors x, y,

$$x \circ y = \operatorname{Diag}(x)\, y = \operatorname{Diag}(y)\, x.$$

Therefore,

$$\begin{aligned} df &= -\sum_{i=1}^{m} \left(\operatorname{Diag}\left(e\right) \boldsymbol{X}_i d\boldsymbol{\beta}\right)^T \boldsymbol{X}_i \\ &= -\sum_{i=1}^{m} \left(d\boldsymbol{\beta}\right)^T \boldsymbol{X}_i^T \operatorname{Diag}\left(e\right) \boldsymbol{X}_i \end{aligned}$$

Finally, we can recast the Hessian as

$$\begin{aligned} \frac{\partial^2 \underline{\ell}}{\partial \boldsymbol{\beta}^2} &= -\sum_{i=1}^{m} \boldsymbol{X}_i^T \operatorname{Diag}\left(e\right) \boldsymbol{X}_i \\ &= -\sum_{i=1}^{m} \boldsymbol{X}_i^T E \boldsymbol{X}_i. \end{aligned}$$

5.1.6 $\mathrm{H}_{\operatorname{vech}(\boldsymbol{\Sigma})\,\operatorname{vech}(\Sigma)\underline{\ell}}$

Recall that

$$\frac{\partial \underline{\ell}}{\partial \operatorname{vech}\left(\boldsymbol{\Sigma}\right)} = \frac{1}{2} \sum_{i=1}^{m} \operatorname{vec}\left\{\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i \boldsymbol{\mu}_i^T + \boldsymbol{\Lambda}_i\right) \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right\}^T D_K.$$

Let $f = \frac{\partial \underline{\ell}}{\partial \operatorname{vech}(\boldsymbol{\Sigma})}$ this time. Before we jump in, let this be a gentle reminder of what we will be using:

- $\operatorname{vec}\left(\boldsymbol{\Sigma}\right) = D_K \operatorname{vech}\left(\boldsymbol{\Sigma}\right)$
- $\partial X^{-1} = -X^{-1} \partial X X^{-1}$
- $\operatorname{vec}\left(A \pm B\right) = \operatorname{vec}\left(A\right) \pm \operatorname{vec}\left(B\right)$
- $\operatorname{vec}\left(AXB\right) = \left(B^T \otimes A\right) \operatorname{vec}\left(X\right)$
- $\left(A \otimes B\right)^T = A^T \otimes B^T$
- Lastly, let's write

$$M_i := \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i \boldsymbol{\mu}_i^T + \boldsymbol{\Lambda}_i\right) \boldsymbol{\Sigma}^{-1} \quad \left(M_i^T = M_i\right).$$

Then,

$$
\begin{aligned}
df &= \frac{1}{2}\sum_{i=1}^{m}\left[\operatorname{vec}\left(-\boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma}M_i - M_i d\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\right) + \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\right)\right]^T D_K \\
&= \frac{1}{2}\sum_{i=1}^{m}\left[-\operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma}M_i\right) - \operatorname{vec}\left(M_i d\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\right) + \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\right)\right]^T D_K \\
&= \frac{1}{2}\sum_{i=1}^{m}\left[-\left(M_i \otimes \boldsymbol{\Sigma}^{-1}\right)\operatorname{vec}\left(d\boldsymbol{\Sigma}\right) - \left(\boldsymbol{\Sigma}^{-1} \otimes M_i\right)\operatorname{vec}\left(d\boldsymbol{\Sigma}\right) + \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right)\operatorname{vec}\left(d\boldsymbol{\Sigma}\right)\right]^T D_K \\
&= \frac{1}{2}\sum_{i=1}^{m}\left\{-\operatorname{vec}\left(d\boldsymbol{\Sigma}\right)^T\left(M_i \otimes \boldsymbol{\Sigma}^{-1}\right) - \operatorname{vec}\left(d\boldsymbol{\Sigma}\right)^T\left(\boldsymbol{\Sigma}^{-1} \otimes M_i\right) + \operatorname{vec}\left(d\boldsymbol{\Sigma}\right)^T\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right)\right\} D_K \\
&= \frac{1}{2}\operatorname{vec}\left(d\boldsymbol{\Sigma}\right)^T \sum_{i=1}^{m}\left\{-\left(M_i \otimes \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \otimes M_i\right) + \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right\} \\
&= \frac{1}{2}\operatorname{vech}\left(d\boldsymbol{\Sigma}\right)^T D_K^T \sum_{i=1}^{m}\left\{-\left(M_i \otimes \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \otimes M_i\right) + \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right\} D_K \\
&= \operatorname{vech}\left(d\boldsymbol{\Sigma}\right)^T \frac{1}{2} D_K^T \left\{-\sum_{i=1}^{m}\left(M_i \otimes \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \otimes M_i\right) + m\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right)\right\} D_K
\end{aligned}
$$

Therefore,

$$
\frac{\partial^2 \underline{\ell}}{\partial \operatorname{vech}\left(d\boldsymbol{\Sigma}\right) \partial \operatorname{vech}\left(d\boldsymbol{\Sigma}\right)} = \frac{1}{2} D_K^T \left\{-\sum_{i=1}^{m}\left(M_i \otimes \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \otimes M_i\right) + m\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right)\right\} D_K
$$

5.1.7 $\mathrm{H}_{\boldsymbol{\beta}\boldsymbol{\mu}_i}\,\underline{\ell}$

Recall that

$$
\frac{\partial \underline{\ell}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m}\left(\boldsymbol{y}_i^T - e^T\right)\boldsymbol{X}_i.
$$

We need to go through the same thing again. Differentiate $\frac{\partial \underline{\ell}}{\partial \boldsymbol{\beta}}$ with respect to $\boldsymbol{\mu}_i$. However, for the time being, we will only consider the i^{th} element of $\boldsymbol{\mu}$, which rules out the need for summation.

$$
df = -\left(de\right)^T \boldsymbol{X}_i.
$$

$$
\begin{aligned}
de &= e \circ ds \\
ds &= \boldsymbol{X}_i d\boldsymbol{\beta} + \boldsymbol{Z}_i d\boldsymbol{\mu}_i + \frac{1}{2}\operatorname{dg}\left(\boldsymbol{Z}_i d\boldsymbol{\Lambda}_i \boldsymbol{Z}_i^T\right)
\end{aligned}
$$

Setting $d\boldsymbol{\beta} = 0$, $d\boldsymbol{\Lambda}_i = 0$,

$$
\begin{aligned}
de &= e \circ ds \\
&= e \circ \boldsymbol{Z}_i d\boldsymbol{\mu}_i \\
&= E\boldsymbol{Z}_i d\boldsymbol{\mu}_i
\end{aligned}
$$

where again $E = \operatorname{Diag}\left(e\right)$. (Readers who are not familiar with the notations are referred to section 5.1.1 and 5.1.4.) Therefore,

$$
df = -\left(d\boldsymbol{\mu}_i\right)^T \boldsymbol{Z}_i^T E \boldsymbol{X}_i
$$

which is then yields

$$\frac{\partial^2 \underline{\ell}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\mu}_i} = -\boldsymbol{Z}_i^T E \boldsymbol{X}_i$$

5.1.8 $\mathrm{H}_{\mathrm{vech}(\boldsymbol{\Lambda}_i)\mu_i}$

We start from

$$\frac{\partial \ell}{\partial \operatorname{vech}(\boldsymbol{\Lambda}_i)} = \frac{1}{2}\left(\operatorname{vec}\left(\boldsymbol{\Lambda}_i^{-1} - \Sigma^{-1} - \boldsymbol{Z}_i^T E \boldsymbol{Z}_i\right)\right)^T D_K$$

Now setting $f = \frac{\partial \ell}{\partial \operatorname{vech}(\boldsymbol{\Lambda}_i)}$,

$$\begin{aligned} df &= -\frac{1}{2}\left(\operatorname{vec}\left(\boldsymbol{Z}_i^T E \boldsymbol{Z}_i\right)\right)^T D_K \\ dE &= \operatorname{Diag}(de). \end{aligned}$$

We now use the following theorem:

$$\operatorname{vec}\left(Z^T \operatorname{Diag}(de)\, Z\right) = \mathcal{Q}(Z)^T de$$

where $\mathcal{Q}(Z) = \left(Z \otimes \mathbf{1}^T\right) \circ \left(\mathbf{1}^T \otimes Z\right)$. $\mathbf{1}$ is a column vector filled with ones of the same dimension as the columns of Z. Then,

$$\begin{aligned} df &= -\frac{1}{2}\left(\operatorname{vec}\left(\boldsymbol{Z}_i^T \operatorname{Diag}(de)\, \boldsymbol{Z}_i\right)\right)^T D_K \\ &= -\frac{1}{2}(de)^T \mathcal{Q}(\boldsymbol{Z}_i) D_K \\ &= -\frac{1}{2}(e \circ ds)^T \mathcal{Q}(\boldsymbol{Z}_i) D_K \\ &= -\frac{1}{2}\left(\operatorname{Diag}(e)\, \boldsymbol{Z}_i d\mu_i\right)^T \mathcal{Q}(\boldsymbol{Z}_i) D_K \\ &= -\frac{1}{2}(d\mu_i)^T \boldsymbol{Z}_i^T E \mathcal{Q}(\boldsymbol{Z}_i) D_K \end{aligned}$$

Thus,

$$\frac{\partial^2 \ell}{\partial \operatorname{vech}(\boldsymbol{\Lambda}_i)\, \partial \mu_i} = -\frac{1}{2} \boldsymbol{Z}_i^T E \mathcal{Q}(\boldsymbol{Z}_i) D_K.$$