We assume that we are in a hierarchical model setting with "fixed" and "random" effects, the fixed effects not being explicitly modeled while the random effects are given a distribution.

For the following, there are N total observations. In addition, there are K "levels" or "grouping factors" in the hierarchy, not necessarily nested. At each of the $k = 1, \ldots, K$ levels, there are J_k different groups, any one of which each of the $i = 1, \ldots, N$ observations can belong. Also at each level, there are Q_k different parameters modeled with one vector of that length for each of the $j = 1, \ldots, J_k$ groups. Finally, there are P unmodeled coefficients that also need to be estimated.

For each observation, associated with the unmodeled parameters is a vector of covariates, x_i. For each pair of an observation and level there is a vector associated with the the modeled parameters, z_{ik}.

If $g_k(i) : \{1, \ldots, N\} \to \{1, \ldots, J_k\}$ is a function that maps the ith observation to its group at the kth level, then our model is:

$$\begin{aligned} y_i \mid \theta', \beta, \sigma^2 &\overset{\text{ind}}{\sim} N\left(x_i^\top \beta + \sum_{k=1}^{K} z_{ik}^\top \theta'_{g_k(i)k}, \sigma^2\right), i = 1, \ldots, N, \\ \theta'_{jk} \mid \Sigma_k, \sigma^2 &\overset{\text{iid}}{\sim} N(0, \sigma^2 \Sigma_k), j = 1, \ldots, J_k, k = 1, \ldots, K. \end{aligned}$$

Furthermore, it is also assumed that the various values of θ' are independent between different levels.

We can write the above more simply in matrix notation:

$$\begin{aligned} \mathbf{Y} \mid \theta', \beta, \sigma^2 &\sim N(\mathbf{X}\beta + \mathbf{Z}\theta', \sigma^2 \mathrm{I}_N), \\ \theta' \mid \Sigma, \sigma^2 &\sim N(0, \sigma^2 \Sigma). \end{aligned}$$

To build the necessary matrices, we call "vec" the vertical concatenation of a sequence of column vectors, "cat" the horizontal concatenation, and "diag" the block diagonal matrix composed of its arguments. Then,

$$\theta' = \mathrm{vec}_{k=1}^{K} \mathrm{vec}_{j=1}^{J_k}(\theta'_{jk}), \Sigma = \mathrm{diag}_{k=1}^{K}(\mathbf{I}_{Q_k} \otimes \Sigma_k), \mathbf{Z} = \mathrm{cat}_{k=1}^{K} \mathrm{cat}_{j=1}^{J_k} \mathrm{vec}_{i=1}^{N}(z_{ik}^{\mathrm{T}} \mathbb{I}\{g_k(i) = j\}),$$

and $\mathbf{X}$ is the standard regression design matrix obtained by vertically stacking each x_i^{T}. Let $\sum_{k=1}^{K} J_k Q_k = Q$. We then have $\dim(\mathbf{X}) = N \times P, \dim(\beta) = P \times 1, \dim(\mathbf{Z}) = N \times \sum_{k=1}^{K} J_k Q_k = N \times Q$, and $\dim(\theta') = Q \times 1$.

The final modification that we make is to define $\Lambda\Lambda^\top = \Sigma$ as the Cholesky factorization of the (unscaled) covariance matrix of the unmodeled coefficients. Note that we can actually compute this for each Σ_k and the same procedure that combines those matrices into Σ can be used on all of the Λ_ks to produce Λ. With this, $\theta = \Lambda^{-1}\theta'$ has a spherical distribution.

We can write the joint density of the modeled coefficients and the observations as:

$$\begin{aligned} p(\mathbf{Y}, \theta \mid \Lambda, \beta, \sigma^2) &\propto (\sigma^2)^{-(N+Q)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\|\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\Lambda\theta\|^2 + \|\theta\|^2\right]\right\}, \\ &= (\sigma^2)^{-(N+Q)/2} \exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix}\mathbf{Y} \\ 0\end{bmatrix} - \begin{bmatrix}\mathbf{Z}\Lambda & \mathbf{X} \\ \mathbf{I} & 0\end{bmatrix}\begin{bmatrix}\theta \\ \beta\end{bmatrix}\right\|^2\right\} \end{aligned}$$

In this sense, the joint density can be seen as a single Gaussian with diagonal covariance. If the design matrix is $\mathbf{A}$, then the mode would be given by $\left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{Y}$. To facilitate this calculation, we compute the block-wise Cholesky factorization of inner product of the "augmented" design matrix above, i.e. $\mathbf{A}^\top \mathbf{A}$.

$$\begin{bmatrix} \Lambda^\top \mathbf{Z}^\top \mathbf{Z} \Lambda + \mathbf{I} & \Lambda^\top \mathbf{Z}^\top \mathbf{X} \\ \mathbf{X}^\top \mathbf{Z} \Lambda & \mathbf{X}^\top \mathbf{X} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_Z & 0 \\ \mathbf{L}_{ZX} & \mathbf{L}_X \end{bmatrix} \begin{bmatrix} \mathbf{L}_Z^\top & \mathbf{L}_{ZX}^\top \\ 0 & \mathbf{L}_X^\top \end{bmatrix},$$
$$\mathbf{L}_Z \mathbf{L}_Z^\top = \Lambda^\top \mathbf{Z}^\top \mathbf{Z} \Lambda + \mathbf{I},$$
$$\mathbf{L}_{ZX} = \mathbf{X}^\top \mathbf{Z} \Lambda \mathbf{L}_Z^{-\top},$$
$$\mathbf{L}_X \mathbf{L}_X^\top = \mathbf{X}^\top \mathbf{X} - \mathbf{L}_{ZX} \mathbf{L}_{ZX}^\top.$$

The inverse of a block-wise triagonal matrix is given by:

$$\begin{bmatrix} \mathbf{L}_Z & 0 \\ \mathbf{L}_{ZX} & \mathbf{L}_X \end{bmatrix} \begin{bmatrix} \mathbf{L}_Z^{-1} & 0 \\ -\mathbf{L}_X^{-1} \mathbf{L}_{ZX} \mathbf{L}_Z^{-1} & \mathbf{L}_X^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{bmatrix}.$$

Thus the modes of the joint distribution are given by:

$$\begin{aligned}
\begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} &= \begin{bmatrix} \mathbf{L}_Z^{-\top} & -\mathbf{L}_Z^{-\top} \mathbf{L}_{ZX}^\top \mathbf{L}_X^{-\top} \\ 0 & \mathbf{L}_X^{-\top} \end{bmatrix} \begin{bmatrix} \mathbf{L}_Z^{-1} & 0 \\ -\mathbf{L}_X^{-1} \mathbf{L}_{ZX} \mathbf{L}_Z^{-1} & \mathbf{L}_X^{-1} \end{bmatrix} \begin{bmatrix} \Lambda^\top \mathbf{Z}^\top & \mathbf{I} \\ \mathbf{X}^\top & 0 \end{bmatrix} \begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix}, \\
&= \begin{bmatrix} \mathbf{L}_Z^{-\top} & -\mathbf{L}_Z^{-\top} \mathbf{L}_{ZX}^\top \mathbf{L}_X^{-\top} \\ 0 & \mathbf{L}_X^{-\top} \end{bmatrix} \begin{bmatrix} \mathbf{L}_Z^{-1} & 0 \\ -\mathbf{L}_X^{-1} \mathbf{L}_{ZX} \mathbf{L}_Z^{-1} & \mathbf{L}_X^{-1} \end{bmatrix} \begin{bmatrix} \Lambda^\top \mathbf{Z}^\top \mathbf{Y} \\ \mathbf{X}^\top \mathbf{Y} \end{bmatrix}, \\
&= \begin{bmatrix} \mathbf{L}_Z^{-\top} & -\mathbf{L}_Z^{-\top} \mathbf{L}_{ZX}^\top \mathbf{L}_X^{-\top} \\ 0 & \mathbf{L}_X^{-\top} \end{bmatrix} \begin{bmatrix} \mathbf{L}_Z^{-1} \Lambda^\top \mathbf{Z}^\top \mathbf{Y} \\ \mathbf{L}_X^{-1} \mathbf{X}^\top \mathbf{Y} - \mathbf{L}_X^{-1} \mathbf{L}_{ZX} \mathbf{L}_Z^{-1} \Lambda^\top \mathbf{Z}^\top \mathbf{Y} \end{bmatrix}, \\
&= \begin{bmatrix} \mathbf{L}_Z^{-\top} & -\mathbf{L}_Z^{-\top} \mathbf{L}_{ZX}^\top \mathbf{L}_X^{-\top} \\ 0 & \mathbf{L}_X^{-\top} \end{bmatrix} \begin{bmatrix} \underset{\sim}{\theta} \\ \mathbf{L}_X^{-1} \left(\mathbf{X}^\top \mathbf{Y} - \mathbf{L}_{ZX} \theta \right) \end{bmatrix}, \\
&= \begin{bmatrix} \mathbf{L}_Z^{-\top} & -\mathbf{L}_Z^{-\top} \mathbf{L}_{ZX}^\top \mathbf{L}_X^{-\top} \\ 0 & \mathbf{L}_X^{-\top} \end{bmatrix} \begin{bmatrix} \underset{\sim}{\theta} \\ \underset{\sim}{\beta} \end{bmatrix}, \\
&= \begin{bmatrix} \mathbf{L}_Z^{-\top} \left(\underset{\sim}{\theta} - \mathbf{L}_{ZX}^\top \mathbf{L}_X^{-\top} \underset{\sim}{\beta} \right) \\ \mathbf{L}_X^{-T} \underset{\sim}{\beta} \end{bmatrix}, \\
&= \begin{bmatrix} \mathbf{L}_Z^{-\top} \left(\underset{\sim}{\theta} - \mathbf{L}_{ZX}^\top \tilde{\beta} \right) \\ \mathbf{L}_X^{-T} \underset{\sim}{\beta} \end{bmatrix}.
\end{aligned}$$

where, $\underset{\sim}{\theta}$ and $\underset{\sim}{\beta}$ are intermediate calculations that we can use to compute the penalized residual sum of squares (needed to profile out $\hat{\sigma}$). Noting that:

$$\underset{\sim}{\theta}^\top \underset{\sim}{\theta} + \underset{\sim}{\beta}^\top \underset{\sim}{\beta} = \begin{bmatrix} \mathbf{Y}^\top & 0 \end{bmatrix} \begin{bmatrix} \mathbf{Z} \Lambda & \mathbf{X} \\ \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \Lambda^\top \mathbf{Z}^\top \mathbf{Z} \Lambda + \mathbf{I} & \Lambda \mathbf{Z}^\top \mathbf{X} \\ \mathbf{X}^\top \mathbf{Z} \Lambda & \mathbf{X} \top \mathbf{X} \end{bmatrix}^{-1} \begin{bmatrix} \Lambda^\top \mathbf{Z}^\top & \mathbf{I} \\ \mathbf{X}^\top & 0 \end{bmatrix} \begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix}$$

.

If this was a simple linear regression, we would write, $\mathbf{Y}^\top \mathbf{A} \left(\mathbf{A}^\top \mathbf{A} \right)^{-1} \mathbf{A}^\top \mathbf{Y} = \mathbf{Y}^\top \mathbf{A} \hat{\beta}$. But,

$$\begin{aligned}
\mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{A} \hat{\beta} &= \mathbf{Y}^\top \left(\mathbf{Y} - \mathbf{A} \hat{\beta} \right), \\
&= \left(\mathbf{Y} - \mathbf{A} \hat{\beta} \right)^\top \left(\mathbf{Y} - \mathbf{A} \hat{\beta} \right), \\
&= \left\| \mathbf{Y} - \mathbf{A} \hat{\beta} \right\|^2.
\end{aligned}$$

So that in our full model,

$$\left\| \begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{Z} \Lambda & \mathbf{X} \\ \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2 = \mathbf{Y}^\top \mathbf{Y} - \underset{\sim}{\theta}^\top \underset{\sim}{\theta} - \underset{\sim}{\beta}^\top \underset{\sim}{\beta}.$$

Once we have obtained the modes of the joint distribution, we can proceed to integrate out the modeled coefficients.

$$p(\mathbf{Y}, \theta \mid \Lambda, \beta, \sigma^2) \propto (\sigma^2)^{-(N+Q)/2} \exp\left\{ -\frac{1}{2\sigma^2} \left\| \begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda & \mathbf{X} \\ \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \beta \end{bmatrix} \right\|^2 \right\},$$

$$= (\sigma^2)^{-(N+Q)/2} \exp\left\{ -\frac{1}{2\sigma^2} \left[\begin{bmatrix} \theta - \tilde{\theta} \\ \beta - \tilde{\beta} \end{bmatrix}^\top \begin{bmatrix} \mathbf{L}_Z & 0 \\ \mathbf{L}_{ZX} & \mathbf{L}_X \end{bmatrix} \begin{bmatrix} \mathbf{L}_Z^\top & \mathbf{L}_{ZX}^\top \\ 0 & \mathbf{L}_X^\top \end{bmatrix} \begin{bmatrix} \theta - \tilde{\theta} \\ \beta - \tilde{\beta} \end{bmatrix} + \left\| \begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda & \mathbf{X} \\ \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2 \right] \right\}.$$

Considering just the parts that involve θ and rotating the covariance with β into the mean, we have:

$$(\sigma^2)^{-Q/2} \exp\left\{ -\frac{1}{2\sigma^2} \left[\begin{bmatrix} \theta - \tilde{\theta} + \mathbf{L}_Z^{-\top} \mathbf{L}_{ZX}^\top (\beta - \tilde{\beta}) \\ \beta - \tilde{\beta} \end{bmatrix}^\top \begin{bmatrix} \mathbf{L}_Z & 0 \\ 0 & \mathbf{L}_X \end{bmatrix} \begin{bmatrix} \mathbf{L}_Z^\top & 0 \\ 0 & \mathbf{L}_X^\top \end{bmatrix} \begin{bmatrix} \theta - \tilde{\theta} + \mathbf{L}_Z^{-\top} \mathbf{L}_{ZX}^\top (\beta - \tilde{\beta}) \\ \beta - \tilde{\beta} \end{bmatrix} \right] \right\}.$$

When integrated out, we obtain:

$$p(\mathbf{Y} \mid \Lambda, \beta, \sigma^2) \propto (\sigma^2)^{-N/2} |\mathbf{L}_Z|^{-1} \exp\left\{ -\frac{1}{2\sigma^2} \left[(\beta - \tilde{\beta})^\top \mathbf{L}_X \mathbf{L}_X^\top (\beta - \tilde{\beta}) + \left\| \begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda & \mathbf{X} \\ \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2 \right] \right\}.$$

From this, the MLE for β is the joint mode, $\tilde{\beta}$. Profiling out β gives us that the mode of σ^2 is $\frac{1}{N}\left[\|\mathbf{Y} - \mathbf{Z}\Lambda\tilde{\theta} - \mathbf{X}\tilde{\beta}\|^2 + \|\tilde{\theta}\|^2\right]$, or the penalized residual sum of squares divided by the sample size. Finally, the fully profiled deviance is given by:

$$d(\Lambda) = N\left(1 + \log(2\pi\hat{\sigma}^2)\right) + 2\log|\mathbf{L}_Z|.$$

If we had wanted the REML, we can further take the likelihood and integrate out β with a flat prior, leaving:

$$p(\mathbf{Y} \mid \Lambda, \sigma^2) = (\sigma^2)^{-(N-P)/2} |\mathbf{L}_Z|^{-1} |\mathbf{L}_X|^{-1} \exp\left\{ -\frac{1}{2\sigma^2} \left\| \begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda & \mathbf{X} \\ \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \tilde{\beta} \end{bmatrix} \right\|^2 \right\}.$$

Consequently, the REML estimate of σ^2 is the penalized, weighted residual sum of squares divided by $N-P$. The profiled deviance is

$$d(\Lambda) = (N - P)\left(1 + \log(2\pi\hat{\sigma}^2)\right) + 2\log|\mathbf{L}_Z| + 2\log|\mathbf{L}_X|.$$

For the following, we now impose a Gaussian prior on β with a known variance Σ_β (with decomposition $\mathbf{L}_\beta \mathbf{L}_\beta^\top = \Sigma_\beta$), so that the full model is given by:

$$\begin{aligned}
\mathbf{Y} \mid \theta, \beta, \Lambda, \sigma &\overset{\text{ind}}{\sim} N\left(\mathbf{X}\beta + \mathbf{Z}\Lambda\theta, \sigma^2 \mathbf{I}_N\right), \\
\theta \mid \sigma &\overset{\text{iid}}{\sim} N\left(0, \sigma^2 \mathbf{I}_Q\right), \\
\beta &\overset{\text{iid}}{\sim} N\left(0, \Sigma_\beta\right).
\end{aligned}$$

Consequently, the joint distribution of the observations and the coefficients is proportional to:

$$\begin{aligned}
p\left(\mathbf{Y}, \theta, \beta \mid \Lambda, \sigma\right) &\propto (\sigma^2)^{-(N+Q)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\|\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\Lambda\theta\|^2 + \|\theta\|^2 + \left\|\sigma \mathbf{L}_\beta^{-1}\beta\right\|^2\right]\right\}, \\
&\propto (\sigma^2)^{-(N+Q)/2} \exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix}\mathbf{Y} \\ 0 \\ 0\end{bmatrix} - \begin{bmatrix}\mathbf{Z}\Lambda & \mathbf{X} \\ 0 & \sigma\mathbf{L}_\beta^{-1} \\ \mathbf{I}_Q & 0\end{bmatrix}\begin{bmatrix}\theta \\ \beta\end{bmatrix}\right\|^2\right\}.
\end{aligned}$$

As before, we can take a block-wise Cholesky factorization of the augmented design matrix:

$$\begin{aligned}
\mathbf{L}_Z \mathbf{L}_Z^\top &= \Lambda^\top \mathbf{Z}^\top \mathbf{Z}\Lambda + \mathbf{I}_Q, \\
\mathbf{L}_{ZX} &= \mathbf{X}^\top \mathbf{Z}\Lambda \mathbf{L}_Z^{-\top}, \\
\mathbf{L}_X(\sigma^2)\mathbf{L}_X^\top(\sigma^2) &= \mathbf{X}^\top\mathbf{X} + \sigma^2\Sigma_\beta^{-1} - \mathbf{L}_{ZX}\mathbf{L}_{ZX}^\top.
\end{aligned}$$

Proceeding as before by calculating the joint mode and integrating with respect to θ produces:

$$\begin{aligned}
p(\mathbf{Y}, \beta \mid \Lambda, \sigma) \propto (\sigma^2)^{-N/2}|\mathbf{L}_Z|^{-1} \exp\left\{-\frac{1}{2\sigma^2}\left[(\beta - \tilde{\beta}(\sigma))^\top \mathbf{L}_X(\sigma)\mathbf{L}_X^\top(\sigma)(\beta - \tilde{\beta}(\sigma))\right]\right\} \times \\
\exp\left\{-\frac{1}{2\sigma^2}\left[\left\|\begin{bmatrix}\mathbf{Y} \\ 0 \\ 0\end{bmatrix} - \begin{bmatrix}\mathbf{Z}\Lambda & \mathbf{X} \\ 0 & \sigma\mathbf{L}_\beta^{-1} \\ \mathbf{I}_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{\theta}(\sigma) \\ \tilde{\beta}(\sigma)\end{bmatrix}\right\|^2\right]\right\}.
\end{aligned}$$

If we are interested in maximizing the posterior mode, $\beta \mid \mathbf{Y}, \Lambda, \sigma^2$, we can see that for any fixed value of σ^2 and Λ, the mode in β will be the same as the joint. Consequently, the profiled posterior is:

$$p(\hat{\beta} \mid \mathbf{Y}, \Lambda, \sigma) \propto (\sigma^2)^{-N/2}|\mathbf{L}_Z|^{-1} \exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix}\mathbf{Y} \\ 0 \\ 0\end{bmatrix} - \begin{bmatrix}\mathbf{Z}\Lambda & \mathbf{X} \\ 0 & \sigma\mathbf{L}_\beta^{-1} \\ \mathbf{I}_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{\theta}(\sigma) \\ \tilde{\beta}(\sigma)\end{bmatrix}\right\|^2\right\}.$$

If, however, we are interested in the likelihood, we can obtain it by integrating out β from the joint. The result is:

$$p(\mathbf{Y} \mid \Lambda, \sigma) = (\sigma^2)^{-(N-P)/2}|\mathbf{L}_Z|^{-1}|\mathbf{L}_X(\sigma)|^{-1} \exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix}\mathbf{Y} \\ 0 \\ 0\end{bmatrix} - \begin{bmatrix}\mathbf{Z}\Lambda & \mathbf{X} \\ 0 & \sigma\mathbf{L}_\beta^{-1} \\ \mathbf{I}_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{\theta}(\sigma) \\ \tilde{\beta}(\sigma)\end{bmatrix}\right\|^2\right\}.$$

I have chosen to highlight the dependencies on σ above as we typically numerically optimize over Λ, and had previously been able to profile out σ. Since we can no longer do that, the goal will be to be able to brute-force optimize over σ, conditioned on the hyper-parameters. For this, we use Newton's method, requiring the first and second derivatives of the objective function.

Also as before, we can write the sum of squared residuals in terms of the intermediate calculations - $\mathbf{Y}^\top\mathbf{Y} - \underset{\sim}{\theta}^\top\underset{\sim}{\theta} - \underset{\sim}{\beta}^\top(\sigma)\underset{\sim}{\beta}(\sigma)$ - which simplifies calculating the derivative of the log-posterior or log-likelihood. Consequently,

$$\frac{\partial}{\partial\sigma}l(\hat{\beta}\mid\mathbf{Y},\Lambda,\sigma)=-N\frac{1}{\sigma}+\frac{1}{\sigma^3}\left(\mathbf{Y}^\top\mathbf{Y}-\underset{\sim}{\theta}^\top\underset{\sim}{\theta}-\underset{\sim}{\beta}^\top(\sigma)\underset{\sim}{\beta}(\sigma)\right)+\frac{1}{2\sigma^2}\frac{\partial}{\partial\sigma}\underset{\sim}{\beta}^\top(\sigma)\underset{\sim}{\beta}(\sigma).$$

Where $\underset{\sim}{\beta}(\sigma)=\mathbf{L}_X^{-1}(\sigma)\left(\mathbf{X}^\top\mathbf{Y}-\mathbf{L}_{ZX}\underset{\sim}{\theta}\right)$ and $\underset{\sim}{\theta}=\mathbf{L}_Z^{-1}\Lambda^\top\mathbf{Z}^\top\mathbf{Y}$. From this, we have

$$\underset{\sim}{\beta}^\top(\sigma)\underset{\sim}{\beta}(\sigma)=\left(\mathbf{X}^\top\mathbf{Y}-\mathbf{L}_{ZX}\underset{\sim}{\theta}\right)^\top\mathbf{L}_X^{-\top}(\sigma)\mathbf{L}_X^{-1}(\sigma)\left(\mathbf{X}^\top\mathbf{Y}-\mathbf{L}_{ZX}\underset{\sim}{\theta}\right).$$

Noting that $\mathbf{L}_X^{-\top}\mathbf{L}_X^{-1}=\left(\mathbf{L}_X\mathbf{L}_X^\top\right)^{-1}=\left(\mathbf{X}^\top\mathbf{X}+\sigma^2\Sigma_\beta^{-1}-\mathbf{L}_{ZX}\mathbf{L}_{ZX}^\top\right)^{-1}$, we let $a=\mathbf{X}^\top\mathbf{Y}-\mathbf{L}_{ZX}\underset{\sim}{\theta}$ and $\mathbf{S}=\mathbf{X}^\top\mathbf{X}-\mathbf{L}_{ZX}\mathbf{L}_{ZX}^\top$. We can write

$$\begin{aligned}\underset{\sim}{\beta}^\top(\sigma)\underset{\sim}{\beta}(\sigma)&=a^\top\left(\mathbf{S}+\sigma^2\Sigma_\beta^{-1}\right)^{-1}a,\\&=a^\top\mathbf{L}_\beta\left(\mathbf{L}_\beta^\top\mathbf{S}\mathbf{L}_\beta+\sigma^2\mathbf{I}_P\right)^{-1}\mathbf{L}_\beta^\top a.\end{aligned}$$

For the sake of computing the derivative, we further make the notational simplifications of $b=\mathbf{L}_\beta^\top a$ and $\mathbf{R}=\mathbf{L}_\beta^\top\mathbf{S}\mathbf{L}_\beta$.

$$\begin{aligned}H(\sigma)&=G(F(\sigma)),\\G(\mathbf{D})&=b^\top\mathbf{D}^{-1}b,\\F(\sigma)&=\mathbf{R}+\sigma^2\mathbf{I}_P,\\\frac{\mathrm{d}}{\mathrm{d}\sigma}H(\sigma)&=\frac{\mathrm{d}\,\mathrm{vec}(G(\mathbf{D}))}{\mathrm{d}\,\mathrm{vec}(\mathbf{D})^\top}\frac{\mathrm{d}\,\mathrm{vec}(F(\sigma))}{\mathrm{d}\sigma},\\\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathrm{vec}(F(\sigma))&=2\sigma\mathrm{vec}(\mathbf{I}_P),\\\frac{\mathrm{d}}{\mathrm{d}\,\mathrm{vec}(\mathbf{D})^\top}\mathrm{vec}(G(\mathbf{D}))&=-\mathrm{vec}\left(\mathbf{D}^{-\top}bb^\top\mathbf{D}^{-\top}\right)^\top.\end{aligned}$$

To help clarify this a bit, as F maps a scalar to a $Q\times Q$ matrix, its derivative is a $Q^2\times 1$ matrix. As G maps $Q\times Q$ matrix to a scalar, its derivative is a $1\times Q^2$ matrix. When we multiply the two in the chain rule, we get the desired scalar derivative.

Furthermore, as $\mathrm{d}F/\mathrm{d}\sigma$ involves vectorizing the identity matrix, we are going to add from the derivative of G the elements that correspond to the diagonal. As such, we can express the derivative as:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\sigma}H(\sigma)&=-2\sigma\times\mathrm{tr}\left(\left(\mathbf{R}+\sigma^2\mathbf{I}_P\right)^{-\top}bb^\top\left(\mathbf{R}+\sigma^2\mathbf{I}_P\right)^{-\top}\right),\\&=-2\sigma b^\top\left(\mathbf{R}+\sigma^2\mathbf{I}_P\right)^{-1}\left(\mathbf{R}+\sigma^2\mathbf{I}_P\right)^{-1}b,\\&=-2\sigma a^\top\mathbf{L}_\beta\left(\mathbf{L}_\beta^\top\mathbf{S}\mathbf{L}_\beta+\sigma^2\mathbf{I}_P\right)^{-1}\left(\mathbf{L}_\beta^\top\mathbf{S}\mathbf{L}_\beta+\sigma^2\mathbf{I}_P\right)^{-1}\mathbf{L}_\beta^\top a,\\&=-2\sigma a^\top\left(\mathbf{S}+\sigma^2\Sigma_\beta^{-1}\right)^{-1}\mathbf{L}_\beta^{-\top}\mathbf{L}_\beta^{-1}\left(\mathbf{S}+\sigma^2\Sigma_\beta^{-1}\right)^{-1}a,\\&=-2\sigma\underset{\sim}{\beta}^\top(\sigma)\mathbf{L}_X^{-1}(\sigma)\mathbf{L}_\beta^{-\top}\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\underset{\sim}{\beta}(\sigma),\\&=-2\sigma\left\|\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\underset{\sim}{\beta}(\sigma)\right\|^2.\end{aligned}$$

Summing up, the first derivative is given by:

$$\frac{\partial}{\partial\sigma}l(\hat{\beta}\mid\mathbf{Y},\Lambda,\sigma)=-N\frac{1}{\sigma}+\frac{1}{\sigma^3}\left(\mathbf{Y}^\top\mathbf{Y}-\underset{\sim}{\theta}^\top\underset{\sim}{\theta}-\left\|\underset{\sim}{\beta}(\sigma)\right\|^2\right)-\frac{1}{\sigma}\left\|\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\underset{\sim}{\beta}(\sigma)\right\|^2.$$

Furthermore, the second derivative is given by:

$$\frac{\partial^2}{(\partial\sigma)^2} l(\hat{\beta} \mid \mathbf{Y}, \Lambda, \sigma) = N\frac{1}{\sigma^2} - \frac{3}{\sigma^4}\left(\mathbf{Y}^\top\mathbf{Y} - \underset{\sim}{\theta}^\top\underset{\sim}{\theta} - \left\|\underset{\sim}{\beta}(\sigma)\right\|^2\right) + \frac{3}{\sigma^2}\left\|\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\underset{\sim}{\beta}(\sigma)\right\|^2 + 4\left\|\mathbf{L}_X^{-1}(\sigma)\mathbf{L}_\beta^{-\top}\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\underset{\sim}{\beta}(\sigma)\right\|^2.$$

We have written in this fashion to highlight how might compute the various derivatives efficiently. If one caches $\mathbf{X}^\top\mathbf{X} - \mathbf{L}_{ZX}\mathbf{L}_{ZX}^\top$, for a new value of σ one can efficiently compute the new $\mathbf{L}_X(\sigma)$. With this value, and having also cached $\mathbf{X}^\top\mathbf{Y} - \mathbf{L}_{ZX}\underset{\sim}{\theta}$, the new $\underset{\sim}{\beta}$ is, as before, $\mathbf{L}_X^{-1}(\sigma)\left(\mathbf{X}^\top\mathbf{Y} - \mathbf{L}_{ZX}\underset{\sim}{\theta}\right)$.

Then, one only needs to compute $\|\underset{\sim}{\beta}\|^2$, $\|\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}\underset{\sim}{\beta}\|^2$, and $\|\mathbf{L}_X^{-1}\mathbf{L}_\beta^{-\top}\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}\underset{\sim}{\beta}\|^2$.

To obtain the derivative for the case in which the unmodeled coefficients are integrated out, we have to be able to take the derivative of $|\mathbf{L}_X(\sigma)|$. Noting that, for an arbitrary matrix $\mathbf{A}$,

$$\frac{\mathrm{d}|\mathbf{A}|}{\mathrm{d}\mathrm{vec}(\mathbf{A})^\top} = |\mathbf{A}|\mathrm{vec}(\mathbf{A}^{-\top})^\top,$$
$$\frac{\mathrm{d}|\mathbf{A}(\sigma)|}{\mathrm{d}\sigma} = |\mathbf{A}|\mathrm{vec}(\mathbf{A}^{-\top})^\top\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\sigma}.$$

Consequently, using our previous definition of $\mathbf{R}$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\sigma}|\mathbf{L}_X(\sigma)| &= \frac{\mathrm{d}}{\mathrm{d}\sigma}\left|\mathbf{L}_\beta^{-\top}\left(\mathbf{R} + \sigma^2\mathbf{I}_P\right)\mathbf{L}_\beta^{-1}\right|, \\
&= \left|\Sigma_\beta^{-1}\right|\frac{\mathrm{d}}{\mathrm{d}\sigma}\left|\mathbf{R} + \sigma^2\mathbf{I}_P\right|, \\
&= \left|\Sigma_\beta^{-1}\right|\left|\mathbf{R} + \sigma^2\mathbf{I}_P\right|\mathrm{vec}\left(\left(\mathbf{R} + \sigma^2\mathbf{I}_P\right)^{-1}\right)^\top \times 2\sigma\mathrm{vec}(\mathbf{I}_P), \\
&= 2\sigma\left|\Sigma_\beta^{-1}\right|\left|\mathbf{R} + \sigma^2\mathbf{I}_P\right|\mathrm{tr}\left(\mathbf{R} + \sigma^2\mathbf{I}_P\right)^{-1}, \\
&= 2\sigma\left|\mathbf{L}_X(\sigma)\right|\mathrm{tr}\left(\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\mathbf{L}_X^{-1}(\sigma)\mathbf{L}_\beta^{-T}\right).
\end{aligned}$$

The first derivative of the log-likelihood is then:

$$\frac{\partial}{\partial\sigma} l(\mathbf{Y} \mid \Lambda, \sigma) = -(N-P)\frac{1}{\sigma} + \frac{1}{\sigma^3}\left(\mathbf{Y}^\top\mathbf{Y} - \underset{\sim}{\theta}^\top\underset{\sim}{\theta} - \left\|\underset{\sim}{\beta}(\sigma)\right\|^2\right) - \frac{1}{\sigma}\left\|\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\underset{\sim}{\beta}(\sigma)\right\|^2 - 2\sigma \times \mathrm{tr}\left(\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\mathbf{L}_X^{-1}(\sigma)\mathbf{L}_\beta^{-T}\right).$$

Now we utilize the fact that $\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathrm{tr}(\mathbf{A}) = \mathrm{tr}\left(\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathbf{A}\right)$ to compute:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathrm{tr}\left(\mathbf{R} + \sigma^2\mathbf{I}_P\right)^{-1} &= -2\sigma \times \mathrm{tr}\left(\left(\mathbf{R} + \sigma^2\mathbf{I}_P\right)^{-1}\left(\mathbf{R} + \sigma^2\mathbf{I}_P\right)^{-1}\right), \\
&= -2\sigma \times \mathrm{tr}\left(\left(\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\mathbf{L}_X^{-1}(\sigma)\mathbf{L}_\beta^{-T}\right)^2\right).
\end{aligned}$$

Putting this together, we have

$$\frac{\partial^2}{(\partial\sigma)^2} l(\mathbf{Y} \mid \Lambda, \sigma) = (N-P)\frac{1}{\sigma^2} - \frac{3}{\sigma^4}\left(\mathbf{Y}^\top\mathbf{Y} - \underset{\sim}{\theta}^\top\underset{\sim}{\theta} - \left\|\underset{\sim}{\beta}(\sigma)\right\|^2\right) + \frac{3}{\sigma^2}\left\|\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\underset{\sim}{\beta}(\sigma)\right\|^2 +$$
$$4\left\|\mathbf{L}_X^{-1}(\sigma)\mathbf{L}_\beta^{-\top}\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\underset{\sim}{\beta}(\sigma)\right\|^2 - 2\times\mathrm{tr}\left(\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\mathbf{L}_X^{-1}(\sigma)\mathbf{L}_\beta^{-T}\right) + 4\sigma^2\times\mathrm{tr}\left(\left(\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}(\sigma)\mathbf{L}_X^{-1}(\sigma)\mathbf{L}_\beta^{-T}\right)^2\right).$$

To compute these traces, we will already have access to the the matrix product $\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}$. The trace of $\mathbf{A}\mathbf{A}^\top$ is just the sum of the squares of the elements of that matrix, but it seems unavoidable for us to at least consider the product $\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}\mathbf{L}_X^{-1}\mathbf{L}_\beta^{-\top}$. With this, we can compute the first order trace by summing down the diagonal, and the second by summing the squares of the elements. For many models, $\mathbf{L}_\beta^{-1}$ is diagonal, so that $\mathbf{L}_\beta^{-1}\mathbf{L}_X^{-\top}$ is triangular and the crossproduct can be computed efficiently.