Appendix for CVEK-boot
Wenying Deng

Contents

1 Derivation of Derivative wrt λ **2**
1.1 **Derivative of REML** 2
1.2 **Derivative of GCV** 4
1.3 **Derivative of AIC** 5

2 Derivation of the REML based Test Statistic **5**
2.1 **Derivation of the Score Test Statistic** 5
2.2 **The Null Distribution of the Test Statistic** 6

1 Derivation of Derivative wrt λ

Corresponding to (2.2.3),

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{h} + \boldsymbol{\epsilon} \quad where \quad \mathbf{h} \sim N(\mathbf{0}, \tau \mathbf{K}_\delta) \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

where $\mathbf{K}_\delta$ is the kernel matrix generated by $k_\delta(\mathbf{z}, \mathbf{z}')$.
the general model may be expressed in matrix form as [1]:

$$\mathbf{y} = \boldsymbol{\mu} + \Phi(\mathbf{X})^T \boldsymbol{\beta} + \boldsymbol{\epsilon} \quad where\ \Phi(\mathbf{X})^T\ is\ n \times p \tag{1.1}$$

where $\Phi(\mathbf{X})$ is the aggregation of columns $\phi(\mathbf{x})$ for all cases in the training set, and $\phi(\mathbf{x})$ is a function mapping a D-dimensional input vector $\mathbf{x}$ into an p-dimensional feature space. This model is fitted by penalized least squares, i.e., our estimate is

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \underset{\mu, \boldsymbol{\beta}}{argmin} \left(\| \mathbf{y} - \boldsymbol{\mu} - \Phi(\mathbf{X})^T \boldsymbol{\beta} \|^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}\right) \tag{1.2}$$

The development that follows depends on the following Assumptions:

1. $\mathbf{1^T}\Phi(\mathbf{X})^T = \mathbf{0}$.

2. $\mathbf{y}$ is not in the column space of $\mathbf{1}$.

where $\mathbf{1}$ is a $n \times 1$ vector.

1.1 Derivative of REML

As our choice of matrix notation suggests, model (1.1) can be seen as equivalent to a linear mixed model, in the following sense. The criterion in (1.2) is proportional to the log likelihood for the partly observed "data" $(\mathbf{y}, \boldsymbol{\beta})$ with respect to the unknowns $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$, i.e., the best linear unbiased prediction (BLUP) criterion, for the mixed model

$$\mathbf{y}|\boldsymbol{\beta} \sim N(\boldsymbol{\mu} + \Phi(\mathbf{X})^T \boldsymbol{\beta}, \sigma^2 \mathbf{I}), \quad \boldsymbol{\beta} \sim N(0, (\sigma^2/\lambda)\mathbf{I})$$

Under this model, $Var(\mathbf{y}) = \sigma^2 \mathbf{V}_\lambda$ where

$$\mathbf{V}_\lambda = \mathbf{I} + \lambda^{-1} \Phi(\mathbf{X})^T \Phi(\mathbf{X}) = \mathbf{I} + \lambda^{-1} \mathbf{K}_\delta \tag{1.3}$$

The mixed model formulation motivates treating λ as a variance parameter to be estimated by maximizing the log likelihood

$$l(\mu, \lambda, \sigma^2 | \mathbf{y}) = -\frac{1}{2}\left[\log | \sigma^2 \mathbf{V}_\lambda | + (\mathbf{y} - \boldsymbol{\mu})^T (\sigma^2 \mathbf{V}_\lambda)^{-1} (\mathbf{y} - \boldsymbol{\mu})\right]$$

Maximizing this log likelihood results in estimating σ^2 with a downward bias, which is removed if we instead maximize the restricted log likelihood

$$l_R(\mu, \lambda, \sigma^2 | \mathbf{y}) = -\frac{1}{2}\left[\log | \sigma^2 \mathbf{V}_\lambda | + (\mathbf{y} - \boldsymbol{\mu})^T (\sigma^2 \mathbf{V}_\lambda)^{-1} (\mathbf{y} - \boldsymbol{\mu}) + \log | \sigma^{-2} \mathbf{1}^T \mathbf{V}_\lambda^{-1} \mathbf{1} |\right] \tag{1.4}$$

We shall refer to the resulting estimate of λ as the REML choice of the parameter.
For given μ and λ, the value of σ^2 maximizing the restricted log likelihood (1.4) is

$$\hat{\sigma}^2_{\mu, \lambda} = (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}_\lambda^{-1} (\mathbf{y} - \boldsymbol{\mu}) / (n - 1) \tag{1.5}$$

substituting in this value and ignoring an additive constant leads to the profile restricted log likelihood

$$l_R(\mu, \lambda|\mathbf{y}) = -\frac{1}{2}\Big[\log|\mathbf{V}_\lambda| + \log|\mathbf{1}^T\mathbf{V}_\lambda^{-1}\mathbf{1}| + (n-1)\log\{(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{V}_\lambda^{-1}(\mathbf{y}-\boldsymbol{\mu})\}\Big] \tag{1.6}$$

For given λ, the value of μ maximizing this last expression is the generalized least square fit $\hat{\mu}_\lambda = (\mathbf{1}^T\mathbf{V}_\lambda^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{V}_\lambda^{-1}\mathbf{y}$.
Using the readily verified equality $\mathbf{V}_\lambda^{-1} = \mathbf{I} - \mathbf{A}_\lambda$, the following key facts about $\mathbf{P}_\lambda$ can be shown to hold under Assumptions 1-2:

$$\mathbf{P}_\lambda = \mathbf{I} - \mathbf{H}_\lambda \tag{1.7}$$

where $\mathbf{H}_\lambda$ is the hat matrix defined by $\hat{\mathbf{y}} = \mathbf{H}_\lambda\mathbf{y}$ and given by

$$\mathbf{H}_\lambda = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T + \mathbf{A}_\lambda \tag{1.8}$$

$$\mathbf{V}_\lambda^{-1}\mathbf{1} = \mathbf{1} \tag{1.9}$$

$$\mathbf{P}_\lambda^k = \mathbf{V}_\lambda^{-k} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T \text{ for } k = 1,\ 2, \ldots \tag{1.10}$$

Under Assumptions 1-2, repeated application of (1.9) gives $\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda = [\mathbf{I} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T]\mathbf{y}$, and hence

$$(\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda)^T\mathbf{V}_\lambda^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda) = \mathbf{y}^T\mathbf{P}_\lambda\mathbf{y} \tag{1.11}$$

Substituting (1.11) into (1.6) yields the profile restricted log likelihood for λ alone:

$$l_R(\lambda|\mathbf{y}) = -\frac{1}{2}\Big[\log|\mathbf{V}_\lambda| + \log|\mathbf{1}^T\mathbf{V}_\lambda^{-1}\mathbf{1}| + (n-1)\log(\mathbf{y}^T\mathbf{P}_\lambda\mathbf{y})\Big] \tag{1.12}$$

Setting the derivative of (1.12) with respect of λ to zero will yield an equation for the REML estimate of λ. By (1.9) again, $\log|\mathbf{1}^T\mathbf{V}_\lambda^{-1}\mathbf{1}| = \log|\mathbf{1}^T\mathbf{1}|$, which does not depend on λ, so the differentiation reduces to finding the derivatives of $\log|\mathbf{V}_\lambda|$ and $\log(\mathbf{y}^T\mathbf{P}_\lambda\mathbf{y})$. To that end we shall need the (component-wise) derivatives of $\mathbf{V}_\lambda$ and $\mathbf{P}_\lambda$ with respect to λ; these can be shown to be:

$$\frac{\partial\mathbf{V}_\lambda}{\partial\lambda} = \lambda^{-1}(\mathbf{I} - \mathbf{V}_\lambda) \tag{1.13}$$

$$\frac{\partial\mathbf{P}_\lambda}{\partial\lambda} = \lambda^{-1}(\mathbf{P}_\lambda - \mathbf{P}_\lambda^2) \tag{1.14}$$

A formula in [2](p. 1016), together with (1.13), leads to

$$\frac{\partial}{\partial\lambda}\log|\mathbf{V}_\lambda| = \lambda^{-1}\mathrm{tr}(\mathbf{V}_\lambda^{-1} - \mathbf{I})$$

By (1.10), $\mathrm{tr}(\mathbf{V}_\lambda^{-1}) = \mathrm{tr}(\mathbf{P}_\lambda) + \mathrm{tr}[\mathbf{I} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T] = \mathrm{tr}(\mathbf{P}_\lambda) + 1$, so we conclude that

$$\frac{\partial}{\partial\lambda}\log|\mathbf{V}_\lambda| = \lambda^{-1}[\mathrm{tr}(\mathbf{P}_\lambda) - (n-1)] \tag{1.15}$$

By Assumption 2, $\mathbf{y}^T\mathbf{P}_\lambda\mathbf{y} > 0$. Thus, using (1.14), we obtain

$$\frac{\partial}{\partial\lambda}\log(\mathbf{y}^T\mathbf{P}_\lambda\mathbf{y}) = \lambda^{-1}\Big[1 - \frac{\mathbf{y}^T\mathbf{P}_\lambda^2\mathbf{y}}{\mathbf{y}^T\mathbf{P}_\lambda\mathbf{y}}\Big] \tag{1.16}$$

Under our Assumptions, the matrix

$$\mathbf{P}_\lambda = \mathbf{V}_\lambda^{-1} - \mathbf{V}_\lambda^{-1}\mathbf{1}(\mathbf{1}^\mathsf{T}\mathbf{V}_\lambda^{-1}\mathbf{1})^{-1}\mathbf{1}^\mathsf{T}\mathbf{V}_\lambda^{-1}$$

which plays a role in some treatments of mixed model theory, turns out to be important for both the REML and the GCV approach to choosing λ.
By (1.12), (1.16) and (**??**), we obtain

$$\frac{\partial l_R(\lambda|\mathbf{y})}{\partial\lambda} = \frac{1}{2\lambda}\left[(n-1)\frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^2\mathbf{y}}{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda\mathbf{y}} - \mathrm{tr}(\mathbf{P}_\lambda)\right] \tag{1.17}$$

Thus by (1.7), (1.11) and (1.17), $\frac{\partial l_R(\lambda|\mathbf{y})}{\partial\lambda} = 0$ implies

$$\frac{(\mathbf{y}-\hat{\mu}_\lambda)^\mathsf{T}\mathbf{V}_\lambda^{-1}(\mathbf{y}-\hat{\mu}_\lambda)}{n-1} = \frac{\mathbf{y}^\mathsf{T}(\mathbf{I}-\mathbf{H}_\lambda)^2\mathbf{y}}{\mathrm{tr}(\mathbf{I}-\mathbf{H}_\lambda)} \tag{1.18}$$

which is also

$$\frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda\mathbf{y}}{n-1} = \frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^2\mathbf{y}}{\mathrm{tr}(\mathbf{P}_\lambda)} \quad \text{or} \quad \frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda\mathbf{y}}{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^2\mathbf{y}} = \frac{n-1}{\mathrm{tr}(\mathbf{P}_\lambda)} \tag{1.19}$$

where $\hat{\mu}_\lambda$ and $\mathbf{H}_\lambda$ are the parameter estimate and hat matrix, respectively, obtained with smoothing parameter value λ. The left side of (1.18) is the REML estimate of σ^2 [3]. The right side equals $\| \mathbf{y}-\hat{\mathbf{y}} \|^2 /[n-\mathrm{tr}(\mathbf{H}_\lambda)]$, an estimate of σ^2 based on viewing $\mathrm{tr}(\mathbf{H}_\lambda)$ as the degrees of freedom of the smoother [4](p. 487) and [5](p. 279). In other words, when λ is estimated by REML, the REML error variance estimate agrees with the "smoothing-theoretic" variance estimate.

1.2 Derivative of GCV

The GCV criterion is given by

$$\mathrm{GCV}(\lambda) = \frac{\| \mathbf{y}-\hat{\mathbf{y}} \|^2}{[1-\mathrm{tr}(\mathbf{H}_\lambda)/n]^2} = \frac{\mathbf{y}^\mathsf{T}(\mathbf{I}-\mathbf{H}_\lambda)^2\mathbf{y}}{[\mathrm{tr}(\mathbf{I}-\mathbf{H}_\lambda)]^2} = \frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^2\mathbf{y}}{[\mathrm{tr}(\mathbf{P}_\lambda)]^2}$$

with the last equality following from (1.7). This criterion, originally proposed by [6], is an approximation to $\frac{1}{n}\sum_{i=1}^n \frac{(y_i-\hat{y}_i)^2}{(1-h_{\lambda[ii]})^2}$, where $h_{\lambda[11]}, ..., h_{\lambda[nn]}$ are the diagonal elements of $\mathbf{H}_\lambda$. The latter expression can be shown (at least in some smoothing problems) to be equal to the leave-one-out cross-validation criterion, but lacks an invariance-under-reparametrization property that is gained by instead using GCV [3](pp. 52-53). Using (1.14), we can obtain

$$\frac{\partial \mathrm{GCV}(\lambda)}{\partial\lambda} = \frac{2}{\lambda[\mathrm{tr}(\mathbf{P}_\lambda)]^3}\left[\mathrm{tr}(\mathbf{P}_\lambda^2)\mathbf{y}^\mathsf{T}P_\lambda^2\mathbf{y} - \mathrm{tr}(\mathbf{P}_\lambda)\mathbf{y}^\mathsf{T}P_\lambda^3\mathbf{y}\right] \tag{1.20}$$

Thus at the GCV-minimizing λ we have

$$\frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^3\mathbf{y}}{\mathrm{tr}(\mathbf{P}_\lambda^2)} = \frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^2\mathbf{y}}{\mathrm{tr}(\mathbf{P}_\lambda)} \quad \text{or} \quad \frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^3\mathbf{y}}{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^2\mathbf{y}} = \frac{\mathrm{tr}(\mathbf{P}_\lambda^2)}{\mathrm{tr}(\mathbf{P}_\lambda)}$$

1.3 Derivative of AIC

The AIC criterion is given by

$$\mathrm{AIC}(\lambda) = \log(\| \mathbf{y} - \hat{\mathbf{y}} \|^2) + \frac{2}{n}[\mathrm{tr}(\mathbf{H}_\lambda) + 1] = \log(\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^2\mathbf{y}) + \frac{2}{n}\mathrm{tr}(\mathbf{I} - \mathbf{P}_\lambda) + \frac{2}{n}$$

Using (1.14), we can obtain

$$\frac{\partial \mathrm{AIC}(\lambda)}{\partial \lambda} = \frac{2}{\lambda}\Big[1 - \frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^3\mathbf{y}}{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^2\mathbf{y}} - \frac{1}{n}\mathrm{tr}(\mathbf{P}_\lambda - \mathbf{P}_\lambda^2)\Big] \tag{1.21}$$

Thus at the AIC-minimizing λ we have

$$\frac{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^3\mathbf{y}}{\mathbf{y}^\mathsf{T}\mathbf{P}_\lambda^2\mathbf{y}} = \frac{\mathrm{tr}(\mathbf{I} - \mathbf{P}_\lambda + \mathbf{P}_\lambda^2)}{n}$$

2 Derivation of the REML based Test Statistic

2.1 Derivation of the Score Test Statistic

In this section, we derive the score test statistic based on REML [7].
Denote $\mathbf{V}(\theta) = \sigma^2\mathbf{V}_\lambda = \sigma^2\mathbf{I} + \tau\mathbf{K}_\delta$, where $\theta = (\delta, \tau, \sigma^2)$. The REML given in (1.4) can be rewritten as

$$l_R = -\frac{1}{2}\Big[\log \mid \mathbf{V}(\theta) \mid + \log \mid \mathbf{1}^\mathsf{T}\mathbf{V}(\theta)^{-1}\mathbf{1} \mid + (\mathbf{y} - \boldsymbol{\mu})^\mathsf{T}\mathbf{V}(\theta)^{-1}(\mathbf{y} - \boldsymbol{\mu})\Big] \tag{2.1}$$

Under $H_0 : \delta = 0$ (2.2.2), we set $\theta_0 = (0, \tau, \sigma^2)$ and

$$\mathbf{P}_0(\theta_0) = \mathbf{V}(\theta_0)^{-1} - \mathbf{V}(\theta_0)^{-1}\mathbf{1}[\mathbf{1}^\mathsf{T}\mathbf{V}(\theta_0)^{-1}\mathbf{1}]^{-1}\mathbf{1}^\mathsf{T}\mathbf{V}(\theta_0)^{-1}$$

Take the derivative of (2.1) with respect to δ,

$$\begin{aligned}
\frac{\partial l_R}{\partial \delta} =& -\frac{1}{2}\Big[\frac{\partial \log \mid \mathbf{V}(\theta) \mid}{\partial \delta} + \frac{\partial \log \mid \mathbf{1}^\mathsf{T}\mathbf{V}(\theta)^{-1}\mathbf{1} \mid}{\partial \delta} + \frac{\partial (\mathbf{y} - \boldsymbol{\mu})^\mathsf{T}\mathbf{V}(\theta)^{-1}(\mathbf{y} - \boldsymbol{\mu})}{\partial \delta}\Big] \\
=& -\frac{1}{2}\Big[\mathrm{tr}\big(\mathbf{V}(\theta)^{-1}\frac{\partial \mathbf{V}(\theta)}{\partial \delta}\big) + \mathrm{tr}\big([\mathbf{1}^\mathsf{T}\mathbf{V}(\theta)^{-1}\mathbf{1}]^{-1}\mathbf{1}^\mathsf{T}\frac{\partial \mathbf{V}(\theta)^{-1}}{\partial \delta}\mathbf{1}\big) \\
& + (\mathbf{y} - \boldsymbol{\mu})^\mathsf{T}\frac{\partial \mathbf{V}(\theta)^{-1}}{\partial \delta}(\mathbf{y} - \boldsymbol{\mu})\Big] \\
=& -\frac{1}{2}\Big[\mathrm{tr}\big(\mathbf{V}(\theta)^{-1}\tau(\partial \mathbf{K}_\delta)\big) - \mathrm{tr}\big(\tau(\partial \mathbf{K}_\delta)\mathbf{V}(\theta)^{-1}\mathbf{1}[\mathbf{1}^\mathsf{T}\mathbf{V}(\theta)^{-1}\mathbf{1}]^{-1}\mathbf{1}^\mathsf{T}\mathbf{V}(\theta)^{-1}\big) \\
& - (\mathbf{y} - \boldsymbol{\mu})^\mathsf{T}\mathbf{V}(\theta)^{-1}\tau(\partial \mathbf{K}_\delta)\mathbf{V}(\theta)^{-1}(\mathbf{y} - \boldsymbol{\mu})\Big] \\
=& \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\mathsf{T}\mathbf{V}(\theta)^{-1}\tau(\partial \mathbf{K}_\delta)\mathbf{V}(\theta)^{-1}(\mathbf{y} - \boldsymbol{\mu}) \\
& - \frac{1}{2}\mathrm{tr}\Big[\tau(\partial \mathbf{K}_\delta)\big[\mathbf{V}(\theta)^{-1} - \mathbf{V}(\theta)^{-1}\mathbf{1}[\mathbf{1}^\mathsf{T}\mathbf{V}(\theta)^{-1}\mathbf{1}]^{-1}\mathbf{1}^\mathsf{T}\mathbf{V}(\theta)^{-1}\big]\Big]
\end{aligned} \tag{2.2}$$

where $\partial \mathbf{K}_\delta$ is the derivative kernel matrix whose $(i,j)^{th}$ entry is $\frac{\partial k_\delta(\mathbf{x}, \mathbf{x}')}{\partial \delta}$. If we further denote $\mathbf{K}_0 = \mathbf{K}_\delta \mid_{\delta=0}$ and $\partial \mathbf{K}_0 = (\partial \mathbf{K}_\delta) \mid_{\delta=0}$, we get the REML based score function of δ evaluated at H_0

$$S_{\delta=0} = \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\mathsf{T}\mathbf{V}(\theta_0)^{-1}\tau(\partial \mathbf{K}_0)\mathbf{V}(\theta_0)^{-1}(\mathbf{y} - \boldsymbol{\mu}) - \frac{1}{2}\mathrm{tr}[\tau(\partial \mathbf{K}_0)\mathbf{P}_0]$$

To test for $H_0 : \delta = 0$, we propose to use the score-based test statistic

$$\hat{T}_0 = \hat{\tau}(\mathbf{y} - \hat{\mu})^T \mathbf{V}_0^{-1}(\partial \mathbf{K}_0)\mathbf{V}_0^{-1}(\mathbf{y} - \hat{\mu}) \tag{2.3}$$

where $\mathbf{V}_0 = \hat{\sigma}^2 \mathbf{I} + \hat{\tau}\mathbf{K}_0$.

2.2 The Null Distribution of the Test Statistic

For simplicity, we denote

$$\mathbf{V} = \mathbf{V}(\theta)$$
$$\mathbf{P} = \mathbf{P}(\theta) = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{1}[\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}]^{-1}\mathbf{1}^T\mathbf{V}^{-1}$$

With similar derivation as (2.2), for each $\theta_i \in \boldsymbol{\theta} = (\delta, \tau, \sigma^2)$, we have

$$\frac{\partial l_R}{\partial \theta_i} = -\frac{1}{2}\Big[\mathrm{tr}\big(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_i}\big) - (\mathbf{y} - \mu)^T\mathbf{V}^{-1}\big(\frac{\partial \mathbf{V}}{\partial \theta_i}\big)\mathbf{V}^{-1}(\mathbf{y} - \mu)\Big] \tag{2.4}$$

From [8] we know $\hat{\mu} = [\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}]^{-1}\mathbf{1}^T\mathbf{V}^{-1}\mathbf{y}$, plug it in [9], we obtain

$$(\mathbf{y} - \mu)^T\mathbf{V}^{-1} = \mathbf{y}^T\big(\mathbf{I} - \mathbf{1}[\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}]^{-1}\mathbf{1}^T\mathbf{V}^{-1}\big)^T\mathbf{y}^{-1} = \mathbf{y}^T\mathbf{P}$$

(2.4) becomes

$$\frac{\partial l_R}{\partial \theta_i} = -\frac{1}{2}\Big[\mathrm{tr}\big(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_i}\big) - \mathbf{y}^T\mathbf{P}\big(\frac{\partial \mathbf{V}}{\partial \theta_i}\big)\mathbf{P}\mathbf{y}\Big]$$

The second-order partial derivatives with respect to θ_i and θ_j is

$$\begin{aligned}\frac{\partial^2 l_R}{\partial \theta_i \partial \theta_j} =& -\frac{1}{2}\Big[\mathrm{tr}\big(\frac{\partial \mathbf{P}}{\partial \theta_j}\frac{\partial \mathbf{V}}{\partial \theta_i}\big) + \mathrm{tr}\big(\mathbf{P}\frac{\partial^2 \mathbf{V}}{\partial \theta_i \partial \theta_j}\big) + \mathbf{y}^T\mathbf{P}\big(\frac{\partial \mathbf{V}}{\partial \theta_i}\big)\mathbf{P}\big(\frac{\partial \mathbf{V}}{\partial \theta_j}\big)\mathbf{P}\mathbf{y} \\ &+ \mathbf{y}^T\mathbf{P}\big(\frac{\partial \mathbf{V}}{\partial \theta_j}\big)\mathbf{P}\big(\frac{\partial \mathbf{V}}{\partial \theta_i}\big)\mathbf{P}\mathbf{y} - \mathbf{y}^T\mathbf{P}\frac{\partial^2 \mathbf{V}}{\partial \theta_i \partial \theta_j}\mathbf{P}\mathbf{y}\Big]\end{aligned} \tag{2.5}$$

where we have used the fact that

$$\begin{aligned}\frac{\partial \mathbf{P}}{\partial \theta_j} =& -\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{V}^{-1} + \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{V}^{-1}\mathbf{1}[\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}]^{-1}\mathbf{1}^T\mathbf{V}^{-1} \\ &+ \mathbf{V}^{-1}\mathbf{1}[\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}]^{-1}\mathbf{1}^T\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{V}^{-1} \\ &- \mathbf{V}^{-1}\mathbf{1}\big([\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}]^{-1}\mathbf{1}^T\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{V}^{-1}\mathbf{1}[\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}]^{-1}\big)\mathbf{1}^T\mathbf{V}^{-1} \\ =& -\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{P}\end{aligned}$$

Then (2.6) turns into

$$\begin{aligned}\frac{\partial^2 l_R}{\partial \theta_i \partial \theta_j} =& -\frac{1}{2}\Big[-\mathrm{tr}\big(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_i}\big) + \mathrm{tr}\big(\mathbf{P}\frac{\partial^2 \mathbf{V}}{\partial \theta_i \partial \theta_j}\big) + \mathbf{y}^T\mathbf{P}\big(\frac{\partial \mathbf{V}}{\partial \theta_i}\big)\mathbf{P}\big(\frac{\partial \mathbf{V}}{\partial \theta_j}\big)\mathbf{P}\mathbf{y} \\ &+ \mathbf{y}^T\mathbf{P}\big(\frac{\partial \mathbf{V}}{\partial \theta_j}\big)\mathbf{P}\big(\frac{\partial \mathbf{V}}{\partial \theta_i}\big)\mathbf{P}\mathbf{y} - \mathbf{y}^T\mathbf{P}\frac{\partial^2 \mathbf{V}}{\partial \theta_i \partial \theta_j}\mathbf{P}\mathbf{y}\Big]\end{aligned} \tag{2.6}$$

Since

$$\begin{aligned} \mathsf{E}(\mathbf{P}\mathbf{y}\mathbf{y}^\mathsf{T}) &= \mathbf{P}[\mathsf{Var}(\mathbf{y}) + (\mathsf{E}\mathbf{y})(\mathsf{E}\mathbf{y})^\mathsf{T}] = \mathbf{P}[\mathbf{V} + \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}] = \mathbf{P}\mathbf{V} \\ \mathbf{P}\mathbf{V}\mathbf{P} &= \mathbf{P}[\mathbf{I} - \mathbf{1}[\mathbf{1}^\mathsf{T}\mathbf{V}^{-1}\mathbf{1}]^{-1}\mathbf{1}^\mathsf{T}\mathbf{V}^{-1}] = \mathbf{P} \end{aligned}$$

we get

$$\begin{aligned} \mathsf{E}\Big[\mathbf{y}^\mathsf{T}\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_j})\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_i})\mathbf{P}\mathbf{y}\Big] =& \mathrm{tr}\Big(\mathsf{E}\Big[\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_j})\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_i})\mathbf{P}\mathbf{y}\mathbf{y}^\mathsf{T}\Big]\Big) \\ =& \mathrm{tr}\Big(\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_j})\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_i})\mathbf{P}\mathbf{V}\Big) \\ =& \mathrm{tr}\Big(\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_j})\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_i})\Big) \\ \mathsf{E}\Big[\mathbf{y}^\mathsf{T}\mathbf{P}\frac{\partial^2 \mathbf{V}}{\partial \theta_i \partial \theta_j}\mathbf{P}\mathbf{y}\Big] =& \mathrm{tr}\Big(\mathbf{P}\frac{\partial^2 \mathbf{V}}{\partial \theta_i \partial \theta_j}\Big) \end{aligned}$$

Therefore,

$$\mathbf{I}_{\theta_i,\theta_j} = -\mathsf{E}\Big[\frac{\partial^2 l_R}{\partial \theta_i \partial \theta_j}\Big] = \frac{1}{2}\mathrm{tr}\Big(\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_j})\mathbf{P}(\frac{\partial \mathbf{V}}{\partial \theta_i})\Big)$$

References

[1] Philip T. Reiss and R. Todd Ogden. Smoothing Parameter Selection for a Class of Semiparametric Linear Models. *JRSS-B*, 71(2), 2009.

[2] Mary J. Lindstrom and Douglas M. Bates. Newton—Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association*, 83(404):1014–1022, December 1988.

[3] G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, January 1990.

[4] Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford, New York, August 2001.

[5] Lee, Nelder, and Pawitan. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood, July 2006.

[6] P. Craven and G. Wahba. Smoothing noisy data with spline functions | SpringerLink, 1979.

[7] Arnab Maity and Xihong Lin. Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. *Biometrics*, 67(4):1271–1284, December 2011.

[8] Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, December 2007.

[9] X. Lin and D. Zhang. Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):381–400, 1999.