

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 |

1 Derivation of Derivative wrt λ

Corresponding to (2.2.3),

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

where \mathbf{K}_δ is the kernel matrix generated by $k_\delta(\mathbf{z}, \mathbf{z}')$.

the general model may be expressed in matrix form as [1]:

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

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

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\mu}, \boldsymbol{\beta}}{\operatorname{argmin}} (\| \mathbf{y} - \boldsymbol{\mu} - \Phi(\mathbf{X})^\top \boldsymbol{\beta} \|^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}) \quad (1.2)$$

The development that follows depends on the following Assumptions:

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

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

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

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

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

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

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

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

We shall refer to the resulting estimate of λ as the REML choice of the parameter.

For given $\boldsymbol{\mu}$ and λ , the value of σ^2 maximizing the restricted log likelihood (1.4) is

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

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

$$l_R(\mu, \lambda | \mathbf{y}) = -\frac{1}{2} \left[\log |\mathbf{V}_\lambda| + \log |\mathbf{1}^\top \mathbf{V}_\lambda^{-1} \mathbf{1}| + (n-1) \log \{(\mathbf{y} - \mu)^\top \mathbf{V}_\lambda^{-1} (\mathbf{y} - \mu)\} \right] \quad (1.6)$$

For given λ , the value of μ maximizing this last expression is the generalized least square fit $\hat{\mu}_\lambda = (\mathbf{1}^\top \mathbf{V}_\lambda^{-1} \mathbf{1})^{-1} \mathbf{1}^\top \mathbf{V}_\lambda^{-1} \mathbf{y}$.

Using the readily verified equality $\mathbf{V}_\lambda^{-1} = \mathbf{I} - \mathbf{A}_\lambda$, the following key facts about \mathbf{P}_λ can be shown to hold under Assumptions 1-2:

$$\mathbf{P}_\lambda = \mathbf{I} - \mathbf{H}_\lambda \quad (1.7)$$

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

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

$$\mathbf{V}_\lambda^{-1} \mathbf{1} = \mathbf{1} \quad (1.9)$$

$$\mathbf{P}_\lambda^k = \mathbf{V}_\lambda^{-k} - \mathbf{1}(\mathbf{1}^\top \mathbf{1})^{-1} \mathbf{1}^\top \text{ for } k = 1, 2, \dots \quad (1.10)$$

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

$$(\mathbf{y} - \hat{\mu}_\lambda)^\top \mathbf{V}_\lambda^{-1} (\mathbf{y} - \hat{\mu}_\lambda) = \mathbf{y}^\top \mathbf{P}_\lambda \mathbf{y} \quad (1.11)$$

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

$$l_R(\lambda | \mathbf{y}) = -\frac{1}{2} \left[\log |\mathbf{V}_\lambda| + \log |\mathbf{1}^\top \mathbf{V}_\lambda^{-1} \mathbf{1}| + (n-1) \log (\mathbf{y}^\top \mathbf{P}_\lambda \mathbf{y}) \right] \quad (1.12)$$

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, $\log |\mathbf{1}^\top \mathbf{V}_\lambda^{-1} \mathbf{1}| = \log |\mathbf{1}^\top \mathbf{1}|$, which does not depend on λ , so the differentiation reduces to finding the derivatives of $\log |\mathbf{V}_\lambda|$ and $\log (\mathbf{y}^\top \mathbf{P}_\lambda \mathbf{y})$. To that end we shall need the (component-wise) derivatives of \mathbf{V}_λ and \mathbf{P}_λ with respect to λ ; these can be shown to be:

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

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

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

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

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

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

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

$$\frac{\partial}{\partial \lambda} \log (\mathbf{y}^\top \mathbf{P}_\lambda \mathbf{y}) = \lambda^{-1} \left[1 - \frac{\mathbf{y}^\top \mathbf{P}_\lambda^2 \mathbf{y}}{\mathbf{y}^\top \mathbf{P}_\lambda \mathbf{y}} \right] \quad (1.16)$$

Under our Assumptions, the matrix

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

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

$$\frac{\partial l_R(\lambda|\mathbf{y})}{\partial \lambda} = \frac{1}{2\lambda} \left[(n-1) \frac{\mathbf{y}^\top \mathbf{P}_\lambda^2 \mathbf{y}}{\mathbf{y}^\top \mathbf{P}_\lambda \mathbf{y}} - \text{tr}(\mathbf{P}_\lambda) \right] \quad (1.17)$$

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

$$\frac{(\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda)^\top \mathbf{V}_\lambda^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda)}{n-1} = \frac{\mathbf{y}^\top (\mathbf{I} - \mathbf{H}_\lambda)^2 \mathbf{y}}{\text{tr}(\mathbf{I} - \mathbf{H}_\lambda)} \quad (1.18)$$

which is also

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

where $\hat{\boldsymbol{\mu}}_\lambda$ and \mathbf{H}_λ 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 $\| \mathbf{y} - \hat{\mathbf{y}} \|^2 / [n - \text{tr}(\mathbf{H}_\lambda)]$, an estimate of σ^2 based on viewing $\text{tr}(\mathbf{H}_\lambda)$ 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

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

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 $h_{\lambda[11]}, \dots, h_{\lambda[nn]}$ are the diagonal elements of \mathbf{H}_λ . 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

$$\frac{\partial \text{GCV}(\lambda)}{\partial \lambda} = \frac{2}{\lambda [\text{tr}(\mathbf{P}_\lambda)]^3} \left[\text{tr}(\mathbf{P}_\lambda^2) \mathbf{y}^\top \mathbf{P}_\lambda^2 \mathbf{y} - \text{tr}(\mathbf{P}_\lambda) \mathbf{y}^\top \mathbf{P}_\lambda^3 \mathbf{y} \right] \quad (1.20)$$

Thus at the GCV-minimizing λ we have

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

1.3 Derivative of AIC

The AIC criterion is given by

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

Using (1.14), we can obtain

$$\frac{\partial \text{AIC}(\lambda)}{\partial \lambda} = \frac{2}{\lambda} \left[1 - \frac{\mathbf{y}^\top \mathbf{P}_\lambda^3 \mathbf{y}}{\mathbf{y}^\top \mathbf{P}_\lambda^2 \mathbf{y}} - \frac{1}{n} \text{tr}(\mathbf{P}_\lambda - \mathbf{P}_\lambda^2) \right] \quad (1.21)$$

Thus at the AIC-minimizing λ we have

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

2 Derivation of the REML based Test Statistic

In this section, we derive the score test statistic based on REML [7].

Denote $\mathbf{V}(\boldsymbol{\theta}) = \sigma^2 \mathbf{V}_\lambda = \sigma^2 \mathbf{I} + \tau \mathbf{K}_\delta$, where $\boldsymbol{\theta} = (\delta, \lambda, \sigma^2)$. The REML given in (1.4) can be rewritten as

$$l_R = -\frac{1}{2} \left[\log |\mathbf{V}(\boldsymbol{\theta})| + \log |\mathbf{1}^\top \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{1}| + (\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right] \quad (2.1)$$

Under $H_0 : \delta = 0$ (2.2.2), we set $\boldsymbol{\theta}_0 = (0, \lambda, \sigma^2)$ and

$$\mathbf{P}_0(\boldsymbol{\theta}_0) = \mathbf{V}(\boldsymbol{\theta}_0)^{-1} - \mathbf{V}(\boldsymbol{\theta}_0)^{-1} \mathbf{1} [\mathbf{1}^\top \mathbf{V}(\boldsymbol{\theta}_0)^{-1} \mathbf{1}]^{-1} \mathbf{1}^\top \mathbf{V}(\boldsymbol{\theta}_0)^{-1}$$

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

$$\begin{aligned} \frac{\partial l_R}{\partial \delta} &= -\frac{1}{2} \left[\frac{\partial \log |\mathbf{V}(\boldsymbol{\theta})|}{\partial \delta} + \frac{\partial \log |\mathbf{1}^\top \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{1}|}{\partial \delta} + \frac{\partial (\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \boldsymbol{\mu})}{\partial \delta} \right] \\ &= -\frac{1}{2} \left[\text{tr}(\mathbf{V}(\boldsymbol{\theta})^{-1} \frac{\partial \mathbf{V}(\boldsymbol{\theta})}{\partial \delta}) + \text{tr}([\mathbf{1}^\top \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{1}]^{-1} \mathbf{1}^\top \frac{\partial \mathbf{V}(\boldsymbol{\theta})^{-1}}{\partial \delta} \mathbf{1}) \right. \\ &\quad \left. + (\mathbf{y} - \boldsymbol{\mu})^\top \frac{\partial \mathbf{V}(\boldsymbol{\theta})^{-1}}{\partial \delta} (\mathbf{y} - \boldsymbol{\mu}) \right] \\ &= -\frac{1}{2} \left[\text{tr}(\mathbf{V}(\boldsymbol{\theta})^{-1} \tau (\partial \mathbf{K}_\delta)) - \text{tr}(\tau (\partial \mathbf{K}_\delta) \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{1} [\mathbf{1}^\top \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{1}]^{-1} \mathbf{1}^\top \mathbf{V}(\boldsymbol{\theta})^{-1}) \right. \\ &\quad \left. - (\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{V}(\boldsymbol{\theta})^{-1} \tau (\partial \mathbf{K}_\delta) \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right] \\ &= \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{V}(\boldsymbol{\theta})^{-1} \tau (\partial \mathbf{K}_\delta) \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \boldsymbol{\mu}) \\ &\quad - \frac{1}{2} \text{tr} \left[\tau (\partial \mathbf{K}_\delta) [\mathbf{V}(\boldsymbol{\theta})^{-1} - \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{1} [\mathbf{1}^\top \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{1}]^{-1} \mathbf{1}^\top \mathbf{V}(\boldsymbol{\theta})^{-1}] \right] \end{aligned}$$

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

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

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

$$\hat{\tau}_0 = \hat{\tau}(\mathbf{y} - \hat{\boldsymbol{\mu}})^\top \mathbf{V}_0^{-1} (\partial \mathbf{K}_0) \mathbf{V}_0^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}})$$

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

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. ISSN 0162-1459. doi: 10.1080/01621459.1988.10478693. URL <https://doi.org/10.1080/01621459.1988.10478693>.
- [3] G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, January 1990. ISBN 978-0-89871-244-5. doi: 10.1137/1.9781611970128. URL <https://epubs.siam.org/doi/book/10.1137/1.9781611970128>.
- [4] Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford, New York, August 2001. ISBN 978-0-19-850765-9.
- [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. URL <https://link.springer.com/article/10.1007/BF01404567>.
- [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. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2011.01598.x.