Vecchia REML Estimation

June 15, 2020

Let $A \subset \mathbb{R}^d$ and let $\{Y(\mathbf{x}) : \mathbf{x} \in A\}$ be a Gaussian process with mean function $\mu_{\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{m}(\mathbf{x})^T\boldsymbol{\beta}$, where $\mathbf{m} : \mathbb{R}^d \to \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, and covariance function $K_{\boldsymbol{\theta}}$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in A$ and suppose we observe $Y_1 = Y(\mathbf{x}_1), \ldots, Y_n = Y(\mathbf{x}_n)$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (\mathbf{m}(\mathbf{x}_1), \ldots, \mathbf{m}(\mathbf{x}_n))^T$ and let $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be an $n \times n$ matrix with i, jth entry given by $K_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)$. Then we have

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta})). \tag{1}$$

Suppose that the $n \times p$ matrix $\mathbf{X}$ is full rank. To carry out REML estimation, we need to first write down the joint density of a set of contrasts $\mathbf{KY}$ where $\mathbf{K}$ any $n - p \times n$ full rank matrix such that $\mathbb{E}[\mathbf{KY}] = 0$. Suppose that the first p rows of $\mathbf{X}$ are linearly independent, and let $\mathbf{Y}_{[j]}$ denote $(Y_1, \ldots, Y_j)^T$. Then the BLUP of Y_{p+j} given $\mathbf{Y}_{[p+j-1]}$ exists for $j = 1, \ldots, n - p$.

1. Let $\boldsymbol{\Sigma}_{[p+j-1]}$ denote the covariance matrix for $\mathbf{Y}_{[p+j-1]}$.
2. Let $\mathbf{X}_{[p+j-1]}$ denote the design matrix for for $\mathbf{Y}_{[p+j-1]}$.
3. Let $\mathbf{k}_j$ denote the vector of covariances between $\mathbf{Y}_{[p+j-1]}$ and Y_{p+j}.
4. Let $\mathbf{X}_j$ denote the jth row of $\mathbf{X}$.

For $j = 1, \ldots, n - p$, let $\boldsymbol{\lambda}_j$ be the first $p + j - 1$ entries of the vector

$$\begin{pmatrix} \boldsymbol{\Sigma}_{[p+j-1]} & \mathbf{X}_{[p+j-1]} \\ \mathbf{X}^T_{[p+j-1]} & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{k}_j \\ \mathbf{X}_j \end{pmatrix}$$

Let $\mathbf{K}$ be an $n - p \times n$ where the jth row is given by $(-\boldsymbol{\lambda}_j^T, 1, 0, \ldots, 0)$. Then $\mathbf{K}$ is full rank and $\mathbb{E}[\mathbf{KY}] = 0$, so $\mathbf{W}$ is a suitable set of contrasts. The jth entry of $\mathbf{W} = \mathbf{KY}$ is just the error of the BLUP of Y_{p+j} based on $\mathbf{Y}_{[p+j-1]}$. Consequently, the entries of $\mathbf{W}$ are uncorrelated with each other. Since they are also jointly normal, they are independent. Let $\mathbf{V} = \mathbf{K\Sigma K}^T$. If $r(\mathbf{w}; \boldsymbol{\beta}, \boldsymbol{\theta})$ denotes the joint density of $\mathbf{W}$, then

$$\log r(\mathbf{w}; \boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W} \tag{2}$$

$$= \sum_{j=1}^{n-p} \frac{1}{2}\Big(-\log(2\pi) - \log \mathbf{V}_{jj} - \mathbf{V}_{jj}^{-1}\mathbf{W}_j^2 \Big) \tag{3}$$

Note that $\mathbf{V}_{jj}$ is just the variance of the error of the BLUP of Y_{p+j} based on $\mathbf{Y}_{[p+j-1]}$, or equivalently, the mse of the BLUP . Now let $\mathbf{S}_{[p+j-1]} \subset \mathbf{Y}_{[p+j-1]}$ have b $= \min(p+j-1, m)$ entries for $j = 1, \ldots, n - p$ where $m << n - p$ (b corresponds to bsize-1 in the code). Vecchia's approximation of (6) is

$$\log r(\mathbf{w}; \boldsymbol{\beta}, \boldsymbol{\theta}) \approx \sum_{j=1}^{n-p} \frac{1}{2}\Big(-\log(2\pi) - \log \tilde{\mathbf{V}}_{jj} - \tilde{\mathbf{V}}_{jj}^{-1}\tilde{\mathbf{W}}_j^2 \Big) \tag{4}$$

where $\tilde{\mathbf{W}}_j$ is the error of the BLUP of Y_{p+j} based on $\mathbf{S}_{[p+j-1]}$ and $\tilde{\mathbf{V}}_{jj}$ is the variance of this error. We can obtain $\tilde{\mathbf{W}}_j$ and $\tilde{\mathbf{V}}_{jj}$ as follows:

1. Let $\tilde{\boldsymbol{\Sigma}}_{[p+j-1]}$ denote the covariance matrix for $\mathbf{S}_{[p+j-1]}$.
2. Let $\tilde{\mathbf{X}}_{[p+j-1]}$ denote the design matrix for for $\mathbf{S}_{[p+j-1]}$.
3. Let $\tilde{\mathbf{k}}_j$ denote the vector of covariances between $\mathbf{S}_{[p+j-1]}$ and Y_{p+j}.

4. Let $\mathbf{X}_j$ denote the jth row of $\mathbf{X}$.

For $j = 1, \ldots, n-p$, let $\tilde{\boldsymbol{\lambda}}_j$ be the first b entries of the vector

$$\begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_{[p+j-1]} & \tilde{\mathbf{X}}_{[p+j-1]} \\ \tilde{\mathbf{X}}^T_{[p+j-1]} & 0 \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\mathbf{k}}_j \\ \mathbf{X}_j \end{pmatrix}$$

Then $\tilde{\mathbf{W}}_j = (-\tilde{\boldsymbol{\lambda}}_j^T, 1)\mathbf{S}_{[p+j]}$ and $\tilde{\mathbf{V}}_{jj} = (-\tilde{\boldsymbol{\lambda}}_j^T, 1)\boldsymbol{\Sigma}_{[p+j]}(-\tilde{\boldsymbol{\lambda}}_j^T, 1)^T$.

We can embed $(-\tilde{\boldsymbol{\lambda}}_j^T, 1)$ in an n-row-vector of zeros $\mathbf{C}_j$ to make $\tilde{\mathbf{W}}_j = \mathbf{C}_j\mathbf{Y}$. Let $\mathbf{C}$ be an $n \times (n-p)$ matrix with rows $\mathbf{C}_j$. Then $\tilde{\mathbf{W}} := (\tilde{\mathbf{W}}_{11}, \ldots, \tilde{\mathbf{W}}_{n-p,n-p})^T = \mathbf{C}\mathbf{Y} \sim \mathcal{N}(0, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T)$ where $\tilde{\mathbf{V}} := \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T$ is a diagonal matrix with diagonal entries $\tilde{\mathbf{V}}_{jj}$. There are formulas for obtaining $\tilde{\mathbf{V}}_{jj}$ directly.

The formula for the gradient is contained in the Stein et al. paper :

$$\frac{\partial}{\partial\theta_k} rl(\theta, w) = -\frac{1}{2}\Big[\Big(\tilde{\mathbf{V}}_{jj}^{-1} \cdot \frac{\partial}{\partial\theta_k}\tilde{\mathbf{V}}_{jj}\Big) + \Big(2\tilde{\mathbf{W}}_j \cdot \tilde{\mathbf{V}}_{jj}^{-1} \cdot \frac{\partial}{\partial\theta_k}\tilde{\mathbf{W}}_j\Big) - \Big(\tilde{\mathbf{W}}_j^2 \cdot \tilde{\mathbf{V}}_j^{-2} \cdot \frac{\partial}{\partial\theta_k}\tilde{\mathbf{V}}_j)\Big]$$

The Fisher Information matrix can be obtained using the fact that $\tilde{\mathbf{W}} \sim \mathcal{N}(0, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T)$:

$$\boldsymbol{\mathcal{I}}_{kl} = \frac{1}{2}\sum_{j=1}^{n-p}\Big[\tilde{\mathbf{V}}_{jj}^{-1} \cdot \frac{\partial}{\partial\theta_k}\tilde{\mathbf{V}}_{jj} \cdot \tilde{\mathbf{V}}_{jj}^{-1} \cdot \frac{\partial}{\partial\theta_l}\tilde{\mathbf{V}}_{jj}\Big].$$

The restricted likelihood, gradient and Fisher Information can be computed in one pass through the data, if the formulas above are correct.