Vecchia's Approximation for Gaussian Processes

June 19, 2020

Contents

0.1 Likelihood Estimation for Gaussian Processes

Let $A \subset \mathbb{R}^d$ and let $\{Y(\mathbf{x}) : \mathbf{x} \in A\}$ be a Gaussian process with mean function $\mu_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{m}(\mathbf{x})^T\boldsymbol{\beta}$, where $\boldsymbol{m} : \mathbb{R}^d \to \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, and covariance function $K_{\boldsymbol{\theta}}$. Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in A$ and suppose we observe $Y_1 = Y(\mathbf{x}_1), \dots, Y_n = Y(\mathbf{x}_n)$. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\mathbf{X} = (\boldsymbol{m}(\mathbf{x}_1), \dots, \boldsymbol{m}(\mathbf{x}_n))^T$ 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

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

The density of $\mathbf{Y}$ is

$$p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-1/2} \exp\big(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\big) \tag{2}$$

which is called the likelihood of $\boldsymbol{\beta}, \boldsymbol{\theta}$ when $\mathbf{y}$ is fixed and it is viewed as a function of the parameters. Maximum likelihood estimates of the parameters can be obtained by maximizing the log of the likelihood function

$$\log p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3}$$

From now one we will suppress the dependence of $\boldsymbol{\Sigma}$ on $\boldsymbol{\theta}$. If $\boldsymbol{\Sigma}$ has no exploitable structure, the the standard way of calculating $\log p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta})$ is to first compute the lower Cholesky factor $\mathbf{L}$ of $\boldsymbol{\Sigma}$. Then $|\boldsymbol{\Sigma}| = \prod_{i=1}^n \mathbf{L}_{ii}^2$. For the quadratic form, we first solve $\mathbf{L}\mathbf{z} = \mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}$ and then compute $\boldsymbol{z}^T\boldsymbol{z}$. Computing $\mathbf{L}$ requires $\mathcal{O}(n^3)$ operations, which can be prohibitive for large large n.

0.2 Vecchia's Approximate Likelihood

For $j = 2, \dots, n$, let $\mathbf{Y}_{(j-1)} := (Y_1, \dots, Y_{j-1})^T$. Then we can write (3) as

$$\log p(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\theta}) = p(y_1; \boldsymbol{\beta}, \boldsymbol{\theta}) + \sum_{j=1}^n \log p(y_j|\mathbf{y}_{(j-1)}; \boldsymbol{\beta}, \boldsymbol{\theta}). \tag{4}$$

Let $m \ll n$. For $j = 2, \dots, m+1$, let $\mathbf{S}_{(j-1)} = \mathbf{Y}_{(j-1)}$. For $j > m+1$ let $\mathbf{S}_{(j-1)} \subset \mathbf{Y}_{(j-1)}$ such that $\mathbf{S}_{(j-1)}$ has m entries. Then $\log p(y_j|\mathbf{y}_{(j-1)}; \boldsymbol{\beta}, \boldsymbol{\theta}) \approx \log p(y_j|\mathbf{s}_{(j-1)}; \boldsymbol{\beta}, \boldsymbol{\theta})$, and with $p(y_1|\mathbf{s}_{(0)}; \boldsymbol{\beta}, \boldsymbol{\theta}) := p(y_1; \boldsymbol{\beta}, \boldsymbol{\theta})$ we have

$$\log p(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\theta}) \approx \sum_{j=1}^n \log p(y_j|\mathbf{s}_{(j-1)}; \boldsymbol{\beta}, \boldsymbol{\theta}). \tag{5}$$

Let us write down the the formula for $p(y_j|\mathbf{s}_{(j-1)}; \boldsymbol{\beta}, \boldsymbol{\theta})$:

1. Let $\boldsymbol{\Sigma}_{(j-1)}$ denote the covariance matrix for $\mathbf{S}_{(j-1)}$.

2. Let $\mathbf{X}_{(j-1)}$ denote the design matrix for for $\mathbf{S}_{(j-1)}$.

3. Let $\mathbf{k}_j$ denote the vector of covariances between $\mathbf{S}_{(j-1)}$ and Y_j.

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

5. Let σ_j^2 denote the variance of Y_j.

The conditional density is given by

$$\begin{aligned} p(y_j|\mathbf{s}_{(j-1)};\boldsymbol{\beta},\boldsymbol{\theta}) = & -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_j^2 - \mathbf{k}_j^\top \boldsymbol{\Sigma}_{j-1}^{-1}\mathbf{k}_j\right) \\ & -\frac{1}{2}\frac{\left(\left(y_j - \mathbf{k}_j^\top \boldsymbol{\Sigma}_{j-1}^{-1}(\mathbf{y}_{(j-1)} - \mathbf{X}_{(j-1)}\boldsymbol{\beta})\right) - \mathbf{X}_j\boldsymbol{\beta}\right)^2}{\sigma_j^2 - \mathbf{k}_j^\top \boldsymbol{\Sigma}_{j-1}^{-1}\mathbf{k}_j}. \end{aligned}$$

Let m_j denote the number of entries in $(\mathbf{S}_{(j-1)}, Y_j)^T$ (note that $m_j = m+1$ when $j > m+1$).

Lemma 1. Let the $m_j \times m_j$ matrix $\boldsymbol{\Gamma}^j$ be the inverse of the lower Cholesky factor of

$$\mathrm{Cov}\begin{pmatrix}\mathbf{S}_{(j-1)} \\ Y_j\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{(j-1)} & \mathbf{k} \\ \mathbf{k}^\top & \sigma_j^2\end{pmatrix}$$

Then the last row of $\boldsymbol{\Gamma}^j$ is given by

$$\boldsymbol{\Gamma}^j_{m_j} = \begin{pmatrix}-\mathbf{k}_j^\top \boldsymbol{\Sigma}_{j-1}\left(\sigma_j^2 - \mathbf{k}_j^\top \boldsymbol{\Sigma}_{j-1}^{-1}\mathbf{k}_j\right)^{-1/2} & \left(\sigma_j^2 - \mathbf{k}_j^\top \boldsymbol{\Sigma}_{j-1}^{-1}\mathbf{k}_j\right)^{-1/2}\end{pmatrix}$$

Proof. □

The last lemma implies

$$\log p(y_j|\mathbf{s}_{(j-1)};\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2}\left(\log(2\pi) + 2\log \mathbf{L}^j_{m_j,m_j} + \left(\boldsymbol{\Gamma}^j_{m_j}\left((\mathbf{S}_{(j-1}, Y_j)^\top - \mathbf{X}_{(j)}\boldsymbol{\beta}\right)\right)^2\right).$$

0.3 Vecchia's Approximation of the Restricted Likelihood

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

$$\log r(\mathbf{w};\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W} \tag{6}$$

$$= \sum_{j=1}^{n-p}\frac{1}{2}\Big(-\log(2\pi) - \log \mathbf{V}_{jj} - \mathbf{V}_{jj}^{-1}\mathbf{W}_j^2\Big) \tag{7}$$

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

$$\log r(\mathbf{w}; \boldsymbol{\beta}, \boldsymbol{\theta}) \approx \sum_{j=1}^{n-p} -\frac{1}{2}\Big(\log(2\pi) + \log \mathbf{V}_{jj} + \mathbf{V}_{jj}^{-1}\mathbf{W}_j^2\Big) \tag{8}$$

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

1. Let $\boldsymbol{\Sigma}_{(p+j-1)}$ denote the covariance matrix for $\mathbf{S}_{(p+j-1)}$.
2. Let $\mathbf{X}_{(p+j-1)}$ denote the design matrix for for $\mathbf{S}_{(p+j-1)}$.
3. Let $\mathbf{k}_j$ 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 $\boldsymbol{\lambda}_j$ be the first b entries of the vector

$$\begin{pmatrix} \boldsymbol{\Sigma}_{(p+j-1)} & \mathbf{X}_{(p+j-1)} \\ \mathbf{X}_{(p+j-1)}^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{k}_j \\ \mathbf{X}_j \end{pmatrix}$$

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

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

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

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

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

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

The restricted likelihood, gradient and Fisher Information can be computed in one pass through the data.

0.4 Bordered Cholesky

Theorem 1. Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and $B \in \mathbb{R}^{n\times p}$ is full rank. Then

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & -L_{22} \end{pmatrix}\begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}^T \tag{9}$$

Proof. We can write (9) as

$$\begin{pmatrix} L_{11} & 0 \\ L_{21} & -L_{22} \end{pmatrix}\begin{pmatrix} L_{11}^T & L_{21}^T \\ 0 & L_{22}^T \end{pmatrix} = \begin{pmatrix} L_{11}L_{11}^T & L_{11}L_{21}^T \\ L_{21}L_{11}^T & L_{21}L_{21}^T - L_{22}L_{22}^T \end{pmatrix} \tag{10}$$

Let $L_{11}L_{11}^T$ be the Cholesky decomposition of A. Define $L_{21} = (L_{11}^{-1}B)^T$. Then $B^T = L_{21}L_{11}^T$. Furthermore, $L_{21}L_{21}^T = B^TL^{-T}L^{-1}B = B^TA^{-1}B$ is symmetric. Let z be a nonzero vector in $\mathbb{R}^p$. Since B is full rank, $Bz \neq 0$, and then $z^TB^TA^{-1}Bz > 0$ since A^{-1} is positive definite. Thus, $B^TA^{-1}B$ has a Cholesky decomposition $L_{22}L_{22}^T$, and $L_{21}L_{21}^T - L_{22}L_{22}^T - L_{22}L_{22}^T = 0$.

□