

Gradient ascent for maximum likelihood estimation in the CSM model

$$p(X | C_{1..K}) = \prod_{n=1}^N p(x_n | C_{1..K}) \quad (1)$$

$$p(X | C_{1..K}) = \prod_{n=1}^N \int \int_{-\infty}^{\infty} p(x_n | z, g) p(g) p(z) dg dz \quad (2)$$

$$(3)$$

approximation of some of the integrals by samples from the priors $p(g_n)$ and $p(z_n)$

$$p(X | C_{1..K}) \approx \prod_{n=1}^N \sum_{l=1}^L p(x_n | z^l, g^l) \quad (4)$$

the integral evaluates as follows, according to Máté Lengyel's intuition, and by the lengthy algebraic manipulations we arrive to a form in which the dependence of covariance matrix on components is simpler, leading to a simpler derivative

$$p(x | z, g) = \int_{-\infty}^{\infty} p(x, v | z, g) dv = \mathcal{N}(x; 0, \sigma_x I + z^2 A \left(\sum_{k=1}^K g_k U_k^T U_k \right) A^T) = \quad (5)$$

$$= \frac{1}{z^{D_v} \sqrt{\det(A^T A)}} \mathcal{N}\left(\frac{1}{z} A^+ x; 0, \frac{\sigma_x}{z^2} (A^T A)^{-1} + \sum_{k=1}^K g_k U_k^T U_k\right) \quad (6)$$

$$f(x, z) \equiv \frac{1}{z} A^+ x, \quad h(z) \equiv \frac{1}{z^{D_v} \sqrt{\det(A^T A)}} \quad (6)$$

$$C(z, g) \equiv \frac{\sigma_x}{z^2} (A^T A)^{-1} + \sum_{k=1}^K g_k U_k^T U_k \quad (7)$$

$$(8)$$

thus, the likelihood can be expressed as

$$p(X | C_{1..K}) \approx \prod_{n=1}^N \sum_{l=1}^L h(z^l) \mathcal{N}(f(x_n, z^l); 0, C(z^l, g^l)) \quad (9)$$

$$\log p(X | C_{1..K}) \approx \sum_{n=1}^N \log \left[\sum_{l=1}^L h(z^l) \mathcal{N}(f(x_n, z^l); 0, C(z^l, g^l)) \right] \quad (10)$$

$$h^l \equiv h(z^l), \quad f_n^l \equiv f(x_n, z^l), \quad C^l \equiv C(z^l, g^l) \quad (11)$$

$$\mathcal{L}_n^l \equiv h^l \mathcal{N}(f_n^l; 0, C^l), \quad \mathcal{L}_n \equiv \sum_{l=1}^L \mathcal{L}_n^l \quad (12)$$

$$\log p(X | C_{1..K}) \approx \sum_{n=1}^N \log \mathcal{L}_n \quad (13)$$

the derivative of the likelihood with respect to a single element of the Cholesky decomposition of one of the covariance components can be decomposed this way

$$\begin{aligned} \frac{\partial \log p(X | C_{1..K})}{\partial [U_k]_{i,j}} &\approx \sum_{n=1}^N \frac{\partial \log \mathcal{L}_n}{\partial [U_k]_{i,j}} = \sum_{n=1}^N \frac{1}{\mathcal{L}_n} \frac{\partial \mathcal{L}_n}{\partial [U_k]_{i,j}} = \sum_{n=1}^N \frac{1}{\mathcal{L}_n} \sum_{l=1}^L \frac{\partial \mathcal{L}_n^l}{\partial [U_k]_{i,j}} = \\ &= \sum_{n=1}^N \frac{1}{\mathcal{L}_n} \sum_{l=1}^L \text{Tr} \left[\frac{\partial \mathcal{L}_n^l}{\partial C^l} \frac{\partial C^l}{\partial [U_k]_{i,j}} \right] \end{aligned} \quad (14)$$

the derivatives in this formula are the following

$$\begin{aligned} \frac{\partial \mathcal{L}_n^l}{\partial C^l} &= h^l \frac{\partial}{\partial C^l} \mathcal{N}(f_n^l; 0, C^l) = h^l \mathcal{N}(f_n^l; 0, C^l) \frac{\partial}{\partial C^l} \log \mathcal{N}(f_n^l; 0, C^l) = \\ &= -\frac{h^l}{2} \mathcal{N}(f_n^l; 0, C^l) [(C^l)^{-1} - (C^l)^{-1} f_n^l (f_n^l)^T (C^l)^{-1}] \end{aligned} \quad (15)$$

$$\frac{\partial C(z, g)}{\partial [U_k]_{i,j}} = g_k \frac{\partial (U_k^T U_k)}{\partial [U_k]_{i,j}} = g_k (U_k^T J^{ij} + J^{ji} U_k) \equiv g_k \hat{U}_k^{ij} \quad (16)$$

substituting back to the derivative

$$\begin{aligned} \frac{\partial \log p(X | C_{1..K})}{\partial [U_k]_{i,j}} &\approx \\ &\approx -\frac{1}{2} \sum_{n=1}^N \frac{1}{\mathcal{L}_n} \sum_{l=1}^L h^l g_k^l \mathcal{N}(f_n^l; 0, C^l) \text{Tr} \left[[(C^l)^{-1} - (C^l)^{-1} f_n^l (f_n^l)^T (C^l)^{-1}] \hat{U}_k^{ij} \right] = \\ &= -\frac{1}{2} \text{Tr} \left[\sum_{n=1}^N \left(\frac{1}{\mathcal{L}_n} \sum_{l=1}^L h^l g_k^l \mathcal{N}(f_n^l; 0, C^l) [(C^l)^{-1} - (C^l)^{-1} f_n^l (f_n^l)^T (C^l)^{-1}] \right) \hat{U}_k^{ij} \right] \end{aligned} \quad (17)$$

The regularities of the \hat{U}_k matrices allow us to replace the trace with a much more efficient computation:

$$M_k = -\frac{1}{2} \sum_{n=1}^N \left(\frac{1}{\mathcal{L}_n} \sum_{l=1}^L h^l g_k^l \mathcal{N}(f_n^l; 0, C^l) [(C^l)^{-1} - (C^l)^{-1} f_n^l (f_n^l)^T (C^l)^{-1}] \right) \quad (18)$$

$$\frac{\partial \log p(X | C_{1..K})}{\partial [U_k]_{i,j}} \approx \text{Tr} [M_k \hat{U}_k^{ij}] = \sum_{a=1}^{Dv} [M_k]_{j,a} [U_k]_{i,a} + [M_k]_{a,j} [U_k]_{i,a} \quad (19)$$

we can move the parameters in the direction of the gradient scaled by a learning rate

$$[U_k]_{i,j} \leftarrow [U_k]_{i,j} + \epsilon \frac{\partial \log p(X | C_{1..K})}{\partial [U_k]_{i,j}} \quad (20)$$