

Gradient ascent for maximum likelihood estimation in the CSM model

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

$$p(X | C_{1..K}) = \prod_{n=1}^N \int \int \int_{-\infty}^{\infty} p(x_n, v_n | z_n, g_n) p(g_n) p(z_n) dv_n dg_n dz_n \quad (2)$$

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

$$p(X | C_{1..K}) \approx \prod_{n=1}^N \sum_{l=1}^L \int_{-\infty}^{\infty} p(x_n, v_n | z_n^l, g_n^l) dv_n \quad (3)$$

the integral evaluates as follows

$$\int_{-\infty}^{\infty} p(x_n, v_n | z_n, g_n) dv_n = \frac{1}{z^{D_v} \sqrt{\det(A^T A)}} \mathcal{N}\left(\frac{1}{z} A^+ x; 0, \frac{\sigma_x}{z^2} (A^T A)^{-1} + \sum_{k=1}^K g_k U_k^T U_k\right) \quad (4)$$

$$f(x, z) \equiv \frac{1}{z} A^+ x \quad (5)$$

$$C(x, z, g) \equiv \frac{\sigma_x}{z^2} (A^T A)^{-1} + \sum_{k=1}^K g_k U_k^T U_k \quad (6)$$

$$h(z) \equiv \frac{1}{z^{D_v} \sqrt{\det(A^T A)}} \quad (7)$$

thus, the likelihood can be expressed as

$$p(X | C_{1..K}) \approx \prod_{n=1}^N \sum_{l=1}^L h(z_n^l) \mathcal{N}(f(x_n, z_n^l); 0, C(x_n, z_n^l, g_n^l)) \quad (8)$$

$$\log p(X | C_{1..K}) \approx \sum_{n=1}^N \log \left[\sum_{l=1}^L h(z_n^l) \mathcal{N}(f(x_n, z_n^l); 0, C(x_n, z_n^l, g_n^l)) \right] \quad (9)$$

$$h_n^l \equiv h(z_n^l), \quad f_n^l \equiv f(x_n, z_n^l), \quad C_n^l \equiv C(x_n, z_n^l, g_n^l) \quad (10)$$

$$\mathcal{L}_n^l \equiv h_n^l \mathcal{N}(f_n^l; 0, C_n^l), \quad \mathcal{L}_n \equiv \sum_{l=1}^L \mathcal{L}_n^l \quad (11)$$

$$\log p(X | C_{1..K}) \approx \sum_{n=1}^N \log \mathcal{L}_n \quad (12)$$

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

$$\begin{aligned} \frac{\partial \log p(X | C_{1..K})}{\partial [U_k]_{i,j}} &\approx \sum_{n=1}^N \frac{\partial \log \mathcal{L}_n}{\partial [U_k]_{i,j}} = \sum_{n=1}^N \frac{1}{\mathcal{L}_n} \frac{\partial \mathcal{L}_n}{\partial [U_k]_{i,j}} = \sum_{n=1}^N \frac{1}{\mathcal{L}_n} \sum_{l=1}^L \frac{\partial \mathcal{L}_n^l}{\partial [U_k]_{i,j}} = \\ &= \sum_{n=1}^N \frac{1}{\mathcal{L}_n} \sum_{l=1}^L \text{Tr} \left[\frac{\partial \mathcal{L}_n^l}{\partial C_n^l} \frac{\partial C_n^l}{\partial [U_k]_{i,j}} \right] \end{aligned} \quad (13)$$

the derivatives in this formula are the following

$$\frac{\partial \mathcal{L}_n^l}{\partial C_n^l} = h_n^l \frac{\partial}{\partial C_n^l} \mathcal{N}(f_n^l; 0, C_n^l) = h_n^l \mathcal{N}(f_n^l; 0, C_n^l) \frac{\partial}{\partial C_n^l} \log \mathcal{N}(f_n^l; 0, C_n^l) = \quad (14)$$

$$= -\frac{h_n^l}{2} \mathcal{N}(f_n^l; 0, C_n^l) [(C_n^l)^{-1} - (C_n^l)^{-1} f_n^l (f_n^l)^T (C_n^l)^{-1}]$$

$$\frac{\partial C(x, z, g)}{\partial [U_k]_{i,j}} = g_k \frac{\partial (U_k^T U_k)}{\partial [U_k]_{i,j}} = g_k (U_k^T J^{ij} + J^{ji} U_k) \equiv g_k \hat{U}_k^{ij} \quad (15)$$

substituting back to the derivative

$$\begin{aligned} \frac{\partial \log p(X | C_{1..K})}{\partial [U_k]_{i,j}} &\approx \\ &\approx -\frac{1}{2} \sum_{n=1}^N \frac{1}{\mathcal{L}_n} \sum_{l=1}^L h_n^l (g_n^l)_k \mathcal{N}(f_n^l; 0, C_n^l) \text{Tr} \left[[(C_n^l)^{-1} - (C_n^l)^{-1} f_n^l (f_n^l)^T (C_n^l)^{-1}] \hat{U}_k^{ij} \right] = \\ &= -\frac{1}{2} \text{Tr} \left[\sum_{n=1}^N \left(\frac{1}{\mathcal{L}_n} \sum_{l=1}^L h_n^l (g_n^l)_k \mathcal{N}(f_n^l; 0, C_n^l) [(C_n^l)^{-1} - (C_n^l)^{-1} f_n^l (f_n^l)^T (C_n^l)^{-1}] \right) \hat{U}_k^{ij} \right] \end{aligned} \quad (16)$$

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

$$M = -\frac{1}{2} \sum_{n=1}^N \left(\frac{1}{\mathcal{L}_n} \sum_{l=1}^L h_n^l (g_n^l)_k \mathcal{N}(f_n^l; 0, C_n^l) [(C_n^l)^{-1} - (C_n^l)^{-1} f_n^l (f_n^l)^T (C_n^l)^{-1}] \right) \quad (17)$$

$$\frac{\partial \log p(X | C_{1..K})}{\partial [U_k]_{i,j}} \approx \sum_{a=1}^{Dv} [M]_{j,a} [U_k]_{i,a} + [M]_{a,j} [U_k]_{i,a} \quad (18)$$

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

$$[U_k]_{i,j} \leftarrow [U_k]_{i,j} + \epsilon \frac{\partial \log p(X | C_{1..K})}{\partial [U_k]_{i,j}} \quad (19)$$