Gradient ascent for maximum likelihood estimation in the CSM model

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

$$p(X \mid C_{1..K}) = \prod_{n=1}^{N} \iint_{-\infty}^{\infty} p(x_n \mid z, g) p(g) p(z) \mathrm{d}g \mathrm{d}z \tag{2}$$

$$\tag{3}$$

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

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

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

$$\begin{aligned} p(x \mid z, g) = \int_{-\infty}^{\infty} p(x, v \mid z, g) \mathrm{d}v = \mathcal{N}(x; 0, \sigma_x I + z^2 A \left(\sum_{k=1}^{K} g_k U_k^T U_k \right) A^T) = \\ = \frac{1}{z^{D_v} \sqrt{\det(A^T A)}} \mathcal{N}(\frac{1}{z} A^+ x; 0, \frac{\sigma_x}{z^2} (A^T A)^{-1} + \sum_{k=1}^{K} g_k U_k^T U_k) \end{aligned} \tag{5}$$

$$f(x, z) \equiv \frac{1}{z} A^+ x, \; h(z) \equiv \frac{1}{z^{D_v}} \tag{6}$$

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

$$\tag{8}$$

thus, the likelihood can be expressed as

$$p(X \mid C_{1..K}) \approx \det(A^TA)^{-\frac{N}{2}} \prod_{n=1}^{N} \sum_{l=1}^{L} h(z^l)\mathcal{N}(f(x_n, z^l); 0, C(z^l, g^l)) \tag{9}$$

$$\log p(X \mid C_{1..K}) \approx -\frac{N}{2}\log(\det(A^TA)) + \sum_{n=1}^{N} \log\left[\sum_{l=1}^{L} h(z^l)\mathcal{N}(f(x_n, z^l); 0, C(z^l, g^l))\right] \tag{10}$$

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

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

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

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

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

the derivatives in this formula are the following

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

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

substituting back to the derivative

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

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

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

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

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

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