Bayesian Maximum Likelihood Estimation for CP Decomposition

Edgar Solomonik

December 30, 2025

1 New Model

Let $\mathcal{N}(\mu, M)$ denote a multidimensional normal distribution with mean μ and covariance matrix M. Let $[K] = \{1, \ldots, K\}$ and let $*$ denote the Hadamard/pointwise product. For low-rank matrix factorization or CP decomposition, the Frobenius norm error metric assumes a Gaussian distribution of the error. In the matrix case, this means that for $A \in \mathbb{R}^{n\times n}$, we assume that A is obtained from some U, V, as

$$\text{vec}(A) = \text{vec}(UV^T) + \epsilon\mathcal{N}(0, I),$$

then for any $\epsilon \geq 0$,

$$\arg\min_{U,V} \|A - UV^T\|_F = \arg\max_{U,V} p(U, V|A),$$

since the most likely U, V is given by the value of $A - UV^T$ that is most likely to be taken on by $\epsilon\mathcal{N}(0, I)$, i.e.,

$$\arg\max_{U,V} p(U, V|A) = \arg\max_{U,V} p(\epsilon\mathcal{N}(0, I) = A - UV^T)$$

and

$$p(\epsilon\mathcal{N}(0, I) = A - UV^T) = \frac{(2\pi)^{-n^2}}{\epsilon} \exp\{-\frac{1}{2}\epsilon^2\|A - UV^T\|_F^2\}.$$

Hence, the log of the probability (log likelihood) is proportional to the negation of the Frobenius norm error.

We could instead assume that the distribution of the noise has the same covariance as the signal / quantity of interest. If we assume each column of U and V are samples of some distribution, then their sample covariance matrices are $\frac{1}{R}UU^T$ and $\frac{1}{R}VV^T$. In the matrix case, we could again write a corresponding probabilistic model,

$$\text{vec}(A) = \text{vec}(UV^T) + \epsilon\mathcal{N}(0, (UU^T + (I - UU^+)) \otimes (VV^T + (I - VV^+))),$$

in which we assume noise outside of the span of U and V is independent (identity covariance).

$$\text{vec}(A) = \text{vec}(UV^T) + \epsilon/K \sum_i Ux^i \otimes Vy^i,$$

where x^i and y^i are random vectors whose entries are elementwise independent and normally distributed. In this case, finding the most like U and V corresponds to minimizing the Mahalanobis distance,

$$\arg\max_{U,V} p(U, V|A) = \arg\min_{U,V} \|\text{vec}(A - UV^T)\|_{((UU^T+(I-UU^+))\otimes(VV^T+(I-VV^+)))^{-1}}. \quad (1)$$

$$\arg\max_{U,V} p(U,V|A) = \arg\min_{U,V} \|\operatorname{vec}(A - UV^T)\|_{(UU^T \otimes VV^T)^{-1}}. \tag{2}$$

In the matrix case, the truncated SVD results in $\operatorname{vec}(A - UV^T) \perp \operatorname{span}(U) \otimes \mathbb{R}^n$ and $\operatorname{vec}(A - UV^T) \perp \mathbb{R}^n \otimes \operatorname{span}(V)$, so it also minimizes Mahalanobis distance in this model. In the tensor case, the AMDM algorithm minimizes Mahalanobis distance in an alternating manner and hence, if it convergence successfully, it reaches a locally optimal solution to likelihood maximization.

We now propose a model to justify why the noise term might follow a distribution with the same covariance as the sample covariance obtained by considering the columns of the decomposition factors. The model assumes that each rank-1 term in the decomposition of T is obtained as a sum of identical copies thereof, rescaled and perturbed in a relative way. Specifically, given a tensor of order N, $T \in \mathbb{R}^{n_1 \times \ldots \times n_N}$, consider the inverse problem of reconstructing $U^{(1)}, \ldots, U^{(N)}$ where $U^{(i)} = \begin{bmatrix} u^{(i,1)} & \ldots & u^{(i,R)} \end{bmatrix}$, such that for some $R' > R$,

$$T = \sum_{j=1}^{R'} \alpha_j \bigotimes_{i=1}^{N} (u^{(i,r_j)} + u^{(i,r_j)} * \epsilon \mathcal{N}(0, I)), \text{ where } \forall r \in [R], \sum_{j, r_j = r} \alpha_j = 1.$$

TODO 1: Show that to leading order in ϵ for sufficiently large R' and small α_js, this model is equivalent to (2) generalized to the tensor case.

ANIKET: Fix copy j. Define the notation for simplicity: $a_i^{(j)} := u^{(i,r_j)}$, $h_i^{(j)} := u^{(i,r_j)} * \xi^{(i,j)}$ with $\xi^{(i,j)} \sim \mathcal{N}(0, I_{n_i})$, is a mean-zero perturbation direction, and ϵ is small. Abstractly, for each copy j, the tensor contribution is rank-1:

$$\mathcal{Z}_j(\epsilon) := \bigotimes_{i=1}^{N} \left(a_i^{(j)} + \epsilon h_i^{(j)} \right) \in \mathbb{R}^{n_1 \times \cdots \times n_N}, \quad a_i^{(j)}, h_i^{(j)} \in \mathbb{R}^{n_i}$$

Now we expand $\mathcal{Z}_j(\epsilon)$ in ϵ using multi-linearity of the outer product. For $N = 3$,

$$(a_1 + \epsilon h_1) \otimes (a_2 + \epsilon h_2) \otimes (a_3 + \epsilon h_3).$$

Note that here, the subscripts $1, 2, 3$ *denote modes.* First, expand in the first factor

$$= a_1 \otimes (a_2 + \epsilon h_2) \otimes (a_3 + \epsilon h_3) + \epsilon h_1 \otimes (a_2 + \epsilon h_2) \otimes (a_3 + \epsilon h_3).$$

Now, expand the first term into the second factor:

$$a_1 \otimes a_2 \otimes (a_3 + \epsilon h_3) + \epsilon a_1 \otimes h_2 \otimes (a_3 + \epsilon h_3) + \epsilon h_1 \otimes (a_2 + \epsilon h_2) \otimes (a_3 + \epsilon h_3).$$

Now, expand each $(a_3 + \epsilon h_3)$ and we keep track of orders of ϵ. The ϵ^0 term is $a_1 \otimes a_2 \otimes a_3$. The ϵ^1 terms are $a_1 \otimes a_2 \otimes \epsilon h_3$, $\epsilon a_1 \otimes h_2 \otimes a_3$, $\epsilon h_1 \otimes a_2 \otimes a_3$. Every other term is at least ϵ^2. So, for $N = 3$:

$$\mathcal{Z}_j(\epsilon) = a_1^{(j)} \otimes a_2^{(j)} \otimes a_3^{(j)} + \epsilon \left(h_1^{(j)} \otimes a_2^{(j)} \otimes a_3^{(j)} + a_1^{(j)} \otimes h_2^{(j)} \otimes a_3^{(j)} + a_1^{(j)} \otimes a_2^{(j)} \otimes h_3^{(j)} \right) + O(\epsilon^2).$$

So, for the general N case:

$$\mathcal{Z}_j(\epsilon) = \bigotimes_{i=1}^{N} a_i^{(j)} + \epsilon \sum_{m=1}^{N} \left(h_m^{(j)} \otimes \bigotimes_{i \neq m} a_i^{(j)} \right) + O(\epsilon^2).$$

Suppose the observed tensor is a weighted sum over copies

$$\mathcal{T} = \sum_{j=1}^{R'} \alpha_j \mathcal{Z}_j(\epsilon).$$

Substitute in the expansion:

$$\mathcal{T} = \sum_{j=1}^{R'} \alpha_j \bigotimes_{i=1}^{N} a_i^{(j)} + \epsilon \sum_{j=1}^{R'} \alpha_j \sum_{m=1}^{N} \left(h_m^{(j)} \otimes \bigotimes_{i \neq m} a_i^{(j)} \right) + O(\epsilon^2).$$

Now, define the mean tensor as the ϵ^0 term:

$$\mathcal{Y} := \sum_{j=1}^{R'} \alpha_j \bigotimes_{i=1}^{N} a_i^{(j)}.$$

Then the residual becomes

$$\mathcal{R} := \mathcal{T} - \mathcal{Y} = \epsilon \mathcal{E} + O(\epsilon^2),$$

where

$$\mathcal{E} := \sum_{j=1}^{R'} \alpha_j \sum_{m=1}^{N} \left(h_m^{(j)} \otimes \bigotimes_{i \neq m} a_i^{(j)} \right).$$

So, to leading order, $\mathcal{R}$ is a random tensor scaled by ϵ.

Confusion: The perturbation was previously defined as (roughly) relative entrywise noise:

$$h_i^{(j)} = a_i^{(j)} \odot \xi_i^{(j)}, \quad x_i^{(j)} \sim \mathcal{N}(0, I).$$

Let us compute the covariance explicitly. Let $a \in \mathbb{R}^n$, $\xi \sim \mathcal{N}(0, I_n)$, and $h = a \odot \xi$. We write $h = \mathrm{diag}(a)\xi$. Then we calculate the mean to vanish:

$$\mathbb{E}[h] = \mathrm{diag}(a)\mathbb{E}[\xi] = 0.$$

The covariance is:

$$\begin{aligned} \mathrm{Cov}(h) = \mathbb{E}[hh^T] - \mathbb{E}[h]\mathbb{E}[h]^T = \mathbb{E}[hh^T] &= \mathbb{E}[(\mathrm{diag}(a)\xi)(\mathrm{diag}(a)\xi)^T] \\ &= \mathrm{diag}(a) I \, \mathrm{diag}(a) = \mathrm{diag}(a \circ a). \end{aligned}$$

So the covariance is diagonal in the standard basis. But the covariance we want (and AMDM uses) is span-based, like $U^{(k)}U^{(k)T}$, is generally dense and encodes correlations induced by the column space. So using the Hadamard noise produces the wrong covariance geometry?

Posed Fix: We change the perturbation model. For each mode k, let $U^{(k)} \in \mathbb{R}^{n_k \times R}$ be the factor matrix with columns $u^{(k,r)}$. Draw

$$x^{(k,j)} \sim \mathcal{N}(0, I_R), \quad \textit{independent over } k \textit{ and } j,$$

and set

$$h_k^{(j)} := U^{(k)} x^{(k,j)}.$$

Now compute the mean and covariance:

$$\mathbb{E}[h_k^{(j)}] = U^{(k)}\mathbb{E}[x^{(k,j)}] = 0.$$

Now the covariance:

$$\begin{aligned}\mathrm{Cov}(h_k^{(j)}) &= \mathbb{E}[h_k^{(j)} h_k^{(j)T}] = \mathbb{E}[U^{(k)} x^{(k,j)} x^{(k,j)T} U^{(k)T}] \\ &= U^{(k)} \mathbb{E}[x^{(k,j)} x^{(k,j)T}] U^{(k)T} = U^{(k)} I_R U^{(k)T} = U^{(k)} U^{(k)T}.\end{aligned}$$

This creates the structure we need. Now, we build a rank-1 noise tensor whose vectorized covariance is Kronecker. Define one noise sample tensor:

$$\mathcal{N}_j := \bigotimes_{k=1}^{N} \left(U^{(k)} x^{(k,j)} \right)$$

Now, define $U_\otimes := U^{(N)} \otimes \cdots \otimes U^{(1)}$ and $g_j := x^{(N,j)} \otimes \cdots \otimes x^{(1,j)}$. The key identity is that vectorizing an outer product gives a Kronecker product of the factors. So:

$$\mathrm{vec}\, \mathcal{N}_j = U_\otimes g_j.$$

Now, we compute (g_j).

$$(g_j) = \mathbb{E}[g_j g_j^T] = \bigotimes_{k=1}^{N} \mathbb{E}[x^{(k,j)} x^{(k,j)T}] = \bigotimes_{k=1}^{N} I_R = I_{R^N}.$$

Now, compute the covariance of $\mathrm{vec}(\mathcal{N}_j)$. First, the mean vanishes.

$$\mathbb{E}[\mathrm{vec}(\mathcal{N}_j] = U_\otimes \mathbb{E}[g_j] = 0.$$

Now, the covariance:

$$\begin{aligned}\mathrm{Cov}(\mathrm{vec}(\mathcal{N}_j)) &= \mathbb{E}[\mathrm{vec}\, \mathcal{N}_j\, \mathrm{vec}\, \mathcal{N}_j{}^T] = \mathbb{E}[U_\otimes g_j g_j^T U_\otimes^T] \\ &= U_\otimes \mathbb{E}[g_j g_j^T] U_\otimes^T = U_\otimes I U_\otimes^T = U_\otimes U_\otimes^T \\ &= \left(U^{(N)} \otimes \cdot \otimes U^{(1)} \right) \left(U^{(N)} \otimes \cdot \otimes U^{(1)} \right)^T = \left(U^{(N)} U^{(N)T} \right) \otimes \cdots \otimes \left(U^{(1)} U^{(1)T} \right),\end{aligned}$$

by repeated use of $(A \otimes B)(C \otimes D) = AC \otimes BD$ and $(A \otimes B)^T = A^T \otimes B^T$. So the covariance is shown to be

$$\mathrm{Cov}(\mathrm{vec}(\mathcal{N}_j)) = \bigotimes_{k=1}^{N} \left(U^{(k)} U^{(k)T} \right).$$

Now, we sum many small independent noise tensors and show Gaussian structure by CLT. Define the full residual model

$$\mathcal{R} = \mathcal{T} - \mathcal{Y} = \epsilon \sum_{j=1}^{R'} \alpha_j \mathcal{N}_j \implies r := \mathrm{vec}(\mathcal{R}) = \epsilon \sum_{j=1}^{R'} \alpha_j z_j,$$

where $z_j := \mathrm{vec}(\mathcal{N}_j)$ are i.i.d., mean zero, covariance $\Sigma_0 = \bigotimes_k (U^{(k)} U^{(k)T})$. Now, compute the mean and covariance of r.

$$\mathbb{E}[r] = \epsilon \sum_{j=1}^{R'} \alpha_j \mathbb{E}[z_j] = 0$$

$$\begin{aligned}\mathrm{Cov}(r) &= \mathbb{E}[rr^T] - \mathbb{E}[r]\mathbb{E}[r^T] = \mathbb{E}[rr^T] \\ &= \mathbb{E}\left[\epsilon^2 \left(\sum_{j=1}^{R'} \alpha_j z_j\right)\left(\sum_{j=1}^{R'} \alpha_j z_j\right)^T\right] = \mathbb{E}\left[\epsilon^2 \sum_{j=1}^{R'}\sum_{\ell=1}^{R'} \alpha_j \alpha_\ell z_j z_\ell^T\right] \\ &= \epsilon^2 \sum_{j=1}^{R'}\sum_{\ell=1}^{R'} \alpha_j \alpha_\ell \mathbb{E}[z_j z_\ell^T].\end{aligned}$$

Now, by independence and zero-mean,

$$\mathbb{E}[z_j z_\ell^T] = \begin{cases} \mathbb{E}[z_j]\mathbb{E}[z_\ell]^T = 0, & j \neq l \\ \mathbb{E}[z_j z_j^T] = \mathrm{Cov}(z_j) = \Sigma_0, & j = \ell \end{cases}$$

So only diagonal terms survive.

$$E[rr^T] = \epsilon^2 \sum_{j=1}^{R'} \alpha_j^2 \Sigma_0 = \epsilon^2 \left(\sum_{j=1}^{R'} \alpha_j^2\right) \Sigma_0.$$

Define $s_{R'}^2 := \sum_{j=1}^{R'} \alpha_j^2$. Then

$$\mathrm{Cov}(r) = \epsilon^2 s_{R'}^2 \Sigma_0.$$

Consider the normalized sum:

$$\frac{r}{\epsilon s_{R'}} = \sum_{j=1}^{R'} \frac{\alpha_j}{S_{R'}} z_j$$

Since this is a triangular array, a standard sufficient condition for CLT is z_j i.i.d. with finite covariance and no single weight dominates:

$$\max_j \frac{\alpha_j^2}{s_{R'}^2} \to 0 \text{ as } R' \to \infty.$$

Under those conditions, the CLT for triangular arrays gives

$$\frac{r}{\epsilon s_{R'}} \implies \mathcal{N}(0, \Sigma_0).$$

Equivalently, for large R'

$$r \approx \mathcal{N}(0, \epsilon^2 s_{R'}^2 \Sigma_0).$$

Now, we have

$$\mathrm{vec}(\mathcal{T}) \approx \mathcal{N}(\mathrm{vec}(\mathcal{Y})\epsilon^2 s_{R'}^2 \Sigma_0).$$

Let $r = \mathrm{vec}(\mathcal{T} - \mathcal{Y})$. For a multivariate normal, the negative log-likelihood is

$$\frac{1}{2} r^T (\epsilon^2 s_{R'}^2 \Sigma_0)^{-1} r \propto r^T \Sigma_0^{-1} r.$$

We drop the constants since they do not affect the minimization. Note that the negative log-likelihood is exactly a Mahalanobis distance. Lastly, Σ_0 is a Kronecker product.

$$\Sigma_0 = \bigotimes_{k=1}^{N} \left(U^{(k)} U^{(k)T}\right).$$

When the inverse exists, the precision is also Kronecker:

$$\Sigma_0^{-1} = \bigotimes_{k=1}^{N} \left(U^{(k)} U^{(k)T} \right)^T.$$

This is the structural form AMDM assures for its ground metric: Kronecker product of per-mode metrics.

TODO 2: Actually run some experiments applying AMDM to tensors generated based on this probabilistic model and see if it gives a better approximation of the Us than ALS.

Additional thoughts (Edgar): the introduction of R' terms is needed so that the perturbed rank-1 terms do not themselves minimize the objective, since they do not increase rank, hence the extra terms are needed for averaging. Perhaps there are other / better ways to model this. In particular the averaging seems to correspond to a quadrature rule, so perhaps one can instead say something like

$$T = \sum_{r=1}^{R} T_r, \quad T_r = \frac{1}{|\Omega|} \int_{\Omega} \bigotimes_{i=1}^{N} (u^{(i,r)} + u^{(i,r)} * \chi(\theta)) d\theta,$$

with something like $\chi(\theta) = \epsilon \mathcal{N}(0, I)(\theta)$ (having in mind independent noise for any θ, which is unrealistic). Perhaps there ways to think of the noise in terms of a more general inverse problem with an integral equation / handle a general class of $\chi(\theta)$ and possibly also make u or $du/d\theta$ dependent on θ.

Consider the following perturbation model

$$T = \sum_{j=1}^{R} \bigotimes_{i=1}^{N} u^{(i,r_j)} + \sum_{k=1}^{K} \epsilon / K \bigotimes_{i=1}^{N} U^{(i)} x^{(i,k)},$$

where K is something larger than R, $U^{(i)} = \begin{bmatrix} u^{(i,1)} & \dots & u^{(i,R)} \end{bmatrix}$ and $x^{(i,k)}$ are random vectors whose entries are elementwise independent and normally distributed. We can show that for any given $U^{(i)}$, the noise will actually follow the effect of covariance given by the kind of empirical covariance of $U^{(i)}$, while in the following model the noise follows the true covariance.

1.1 Notes on Application to Probability Density Basis Coefficient Estimation

Suppose we have a Bayesian model to estimate a probability density of the form,

$$p(x, y) = \langle A^T \psi(x), B\phi(y) \rangle = \sum_{i,j,r} a_{ir} b_{jr} \psi_i(x) \phi_j(y)$$

where we fix a basis a prior defined by vector valued functions $\psi, \phi : R \to R^N$. Consider the random variable, $Z = \psi(x)\phi(y)^T$ where $(x, y) \sim p(x, y)$. Then given m independent samples $\{(x_k, y_k)\}_{k=1}^m$ of $p(x, y)$, we have that for

$$M = \frac{1}{m} \sum_{k=1}^{m} \psi(x_k) \phi(y_k)^T$$

that

$$\mathbb{E}[Z] = W_\psi A B^T W_\phi, \text{ where } W_f[i, j] = \int_z f_i(z) f_j(z) dz.$$

If we treat the m samples as random variables, $\mathbb{E}[M] = \mathbb{E}[Z]$ and the variance is $\mathbb{M}[M] = \frac{1}{k}\mathbb{M}[Z]$ may be used to estimate AB^T.

We could further ask which choice of A and B would yield the probabilistic model (define $p(x, y)$) that is most likely to produce M (to solve the maximum likelihood problem given only M). Since M is a sum of iid samples, as $m \to \infty$, M converges to a multidimensional Gaussian distribution. Hence, we know that as $m \to \infty$,

$$p(M|A, B) \propto e^{-\frac{1}{2}x^T V^{-1} x}, x = \text{vec}(M - AB^T), V = \frac{1}{m} \cdot \mathbb{M}_{(x,y)\sim p(x,y)}[\psi(x) \otimes \phi(y)].$$

However, the covariance of $\psi(x) \otimes \phi(y)$ in $p(x, y)$ is a bit difficult to characterize. This random variable is a tensor product of two random variables, but its covariance matrix also has terms that depend on the expectations of the two random variables. Further, the marginal distribution of $p(x, y)$ is of the form,

$$p(x) = \int_y p(x, y)dy = \langle u, \psi(x) \rangle,$$

for some vector u. Hence its covariance appears to be defined by rank-1 terms, as apposed to AA^T, which we would want for the current form of Mahalanobis distance minimization.

We could devise new distance functions based on something like the above, or try to consider other parameterization of the distribution. I was considering models like

$$p(x, y) = \sum_{i,j,r} a_{ir} b_{jr} \psi_{ir}(x) \phi_{jr}(y)$$

where ψ_{ir} and ψ_{js} for each $i \neq j$ have disjoint support (each choice of i index corresponds to a segment of x). Though I am not sure that is the best restriction to apply. The marginal distribution would then involve all of A and B, but its unclear if the resulting formulations make sense in terms of M, etc.

2 Prior Work

2.1 Theory of Expectation Maximization

2.2 Covariance Estimation and Expectation Maximization Applications

Prior works have applied the EM algorithm to estimation of structured covariance matrices. In [1], the covariance matrix is parameterized by a vector and EM is used to derive a tractable iterative update, though the update still requires iterative optimization and varies depending on the underlying model.

3 Maximum Likelihood Estimation with CP Decomposition

3.1 Problem Definition: Empirical Bayes Model for CP Decomposition

Given tensor of order N, $T \in \mathbb{R}^{n_1 \times \ldots \times n_N}$, we model T as a sum of R i.i.d. samples from a distribution of rank-1 tensors, plus noise,

$$T = \mathcal{N} + \frac{1}{R} \sum_{r=1}^{R} \bigotimes_{i=1}^{N} u^{(i,r)},$$

where each $u^{(i,r)}$ is an independent sample of a multivariate normal random variable $u^{(i)}$. We assume the term $\mathcal{N}$, which may represent noise or less significant information, is also a sum of

samples of rank-1 tensors from a rescaled distribution with the same covariance but zero mean, i.e., for some k, ϵ,

$$\mathcal{N} = \epsilon/k \sum_{r=1}^{k} \bigotimes_{i=1}^{N} \tilde{u}^{(i,r)}, \quad \tilde{u}^{(i,r)} \sim u^{(i)} - \mathbb{E}[u^{(i)}].$$

Given T, we seek to recover the most likely samples $u^{(i,r)}$, for $i \in 1, \dots N$, $r \in 1, \dots R$, by estimating the covariance matrices $\mathbb{M}[u^{(i)}]$. Since the samples composing the noise are independent and zero mean,

$$\mathbb{M}[\mathcal{N}] = \epsilon^2/k \mathbb{M}\Big[\bigotimes_{i=1}^{N} \tilde{u}^{(i,r)}\Big] = \epsilon^2/k \bigotimes_{i=1}^{N} \mathbb{M}\Big[u^{(i)}\Big]$$

Are we missing a factor of k due to sum of covariances? Also note that T and $u^{(i)}$ do not have the same covariance as above if their mean is nonzero. We parameterize $\mathbb{M}[u^{(i)}] = \alpha I + (\beta I + (1/R) X X^T)^{-1}$, where the prior distribution of X is

$$p(X) = C \det(\mathcal{V}_\alpha(X))^{-R/2} \prod_{i=1}^{R} \exp\{-\frac{1}{2}\beta \operatorname{Tr}(X^T X)\},$$

for some $\alpha, \beta > 0$, with appropriate choice of constant C. Then, letting $U^{(i)} = \begin{bmatrix} u^{(i,1)} & \cdots & u^{(i,R)} \end{bmatrix}$, $\theta = \{U^{(i)}\}_{i=1}^{N}$, and $\Psi = \{X^{(i)}\}_{i=1}^{N}$, the joint probability density function is

$$p(\theta, \Psi, T) = p(T|\theta, \Psi) p(\theta|\Psi) p(\Psi).$$

We seek to perform maximum likelihood estimation on the samples θ, given T, i.e., we seek to maximize the marginal likelihood,

$$\mathcal{L}(\theta|T) = p(T|\theta) = \int p(T|\theta, \Psi) p(\Psi|\theta) d\Psi.$$

3.2 Expectation Maximization Algorithm

FIXME: EM would usually define Q relative to expectation over the distribution $p(\Psi|\bar{\theta}, T)$. Note that $p(\Psi|\bar{\theta}, T) \neq p(\Psi|\bar{\theta})$, because $\mathcal{N}$ is fixed given $\bar{\theta}$ and T, but the probability of Ψ is dependent on $\mathcal{N}$ if $\mathcal{N}$ is distributed in the same fashion.

At each iteration, given the current estimate of parameters $\bar{\theta}$ the EM algorithm maximizes θ in the expectation (over Ψ conditioned on $\bar{\theta}$) of the log likelihood of Ψ and T, conditioned on θ,

$$Q(\theta, \bar{\theta}) = \int \log(p(\Psi, T|\theta)) p(\Psi|\bar{\theta}, T) d\Psi.$$

FIXME desired form is below, standard form is above

$$\hat{Q}(\theta, \bar{\theta}) = \int \log(p(T|\Psi, \theta)) p(\Psi|\bar{\theta}) d\Psi.$$

Note that, $p(T|\Psi, \theta) = p(\mathcal{N} = T - T(\theta)|\Psi)$ where $T(\theta)$ is the tensor constructed from the CP factors contained in θ. Then, we have that

$$\log p(T|\Psi, \theta) = -\frac{1}{2} \operatorname{vec}(T - T(\theta))^T \mathbb{K}_\Psi \operatorname{vec}(T - T(\theta)), \quad \text{where} \quad \mathbb{K}_\Psi = \bigotimes_{i=1}^{N} \mathbb{K}^{(i)}_{X^{(i)}},$$

So we have, with the desired form of Q, that,

$$Q(\theta, \bar{\theta}) = -\frac{1}{2}\operatorname{vec}(T - T(\theta))^T \left(\int \mathbb{K}_\Psi p(\Psi|\bar{\theta}) d\Psi \right) \operatorname{vec}(T - T(\theta)) = \log p(T|\mathbb{E}_{p(\Psi|\bar{\theta})}[\mathbb{K}_\Psi], \theta).$$

Therefore, the expectation step in the EM algorithm amounts to finding the expected value of $\mathbb{K}_\Psi$ given the current parameters $\bar{\theta}$. Since the samples along each mode of the tensor are independent, $\mathbb{E}_{p(\Psi|\bar{\theta})}[\mathbb{K}_\Psi] = \bigotimes_{i=1}^{N} \mathbb{E}_{p(X^{(i)}|U^{(i)})}[\mathbb{K}^{(i)}_{X^{(i)}}]$. Hence, we next proceed to show that, if $\mathbb{E}[u^{(i)}] = 0$, the regularized form of the inverse of the sample covariance matrix of each $X^{(i)}$, $\mathbb{K}^{(i)}_{X^{(i)}}$, is the conditional expected value of the inverse of the covariance matrix given $U^{(i)}$ and the specified prior distribution of $X^{(i)}$. With that, $Q(\theta, \bar{\theta})$ is shown to be the same objective function as that optimized by the AMDM algorithm [3].

3.3 Sample Covariance Matrix from Conditional Expectation

We now derive a parameterization of covariance matrices such that the conditional expectation of the regularized inverse of the covariance matrix given the observed samples is the regularized inverse of the sample covariance matrix. In the following theorem, we use the shorthand notation,

$$\mathcal{V}_\gamma(W) = \gamma I + \frac{1}{\#\text{cols}(W)} W W^T.$$

Note that $\|\mathcal{V}_\gamma(W)^{-1}\|_2 \leq 1/\gamma$.

Theorem 1. *Given a set of R samples $a_1, \ldots, a_R$ of an m-dimensional normal distribution Y with $\mathbb{E}[Y] = 0$, let $A = \begin{bmatrix} a_1 & \cdots & a_R \end{bmatrix}$, then the covariance matrix $Z = \mathbb{M}[Y]$, satisfies*

$$\mathbb{E}[Z^{-1}|A] = \alpha I + \mathcal{V}_\beta(A)^{-1},$$

if the prior distribution for Z is defined by random matrix $X \in \mathbb{R}^{m \times R}$ so $Z = \mathcal{V}_\alpha(X)^{-1}$ and $p(X) = C \det(\mathcal{V}_\alpha(X))^{-R/2} \prod_{i=1}^{R} \exp\{-\frac{1}{2}\beta \operatorname{Tr}(X^T X)\}$, for any $\alpha, \beta > 0$, with appropriate choice of constant C.

Proof. First, note that for any $\alpha, \beta > 0$, $p(X)$ is Lebesgue integrable (for constant α, β there exists a suitable normalization constant C to make $p(X)$ a valid probability distribution). Using Bayes theorem for continuous random variables [2],

$$\begin{aligned}\mathbb{E}[Z^{-1}|A] = \mathbb{E}[\mathcal{V}_\alpha(X)|A] &= \int_{\mathbb{R}^{m \times R}} \mathcal{V}_\alpha(X) p(X|A) dX \\ &= \int_{\mathbb{R}^{m \times R}} \mathcal{V}_\alpha(X) p(A|X) p(X)/p(A) dX,\end{aligned}$$

where $p(A)$ is defined based on the prior distribution $p(X)$,

$$\begin{aligned}
p(A) &= \int_{\mathbb{R}^{m\times R}} p(A|X)p(X)dX = \int_{\mathbb{R}^{m\times R}} p(X)\prod_{i=1}^{R} p(a_i|X)dX \\
&= (2\pi)^{-mR/2}\int_{\mathbb{R}^{m\times R}} p(X)\prod_{i=1}^{R}\det(\mathcal{V}_\alpha(X))^{1/2}\exp\{-\frac{1}{2}a_i^T\mathcal{V}_\alpha(X)a_i\}dX \\
&= (2\pi)^{-mR/2}\int_{\mathbb{R}^{m\times R}} p(X)\det(\mathcal{V}_\alpha(X))^{R/2}\exp\{-\frac{1}{2}\operatorname{Tr}(A^T\mathcal{V}_\alpha(X)A)\}dX \\
&= (2\pi)^{-mR/2}\int_{\mathbb{R}^{m\times R}} \exp\{-\frac{1}{2}\operatorname{Tr}(A^T\mathcal{V}_\alpha(X)A)\}\prod_{i=1}^{R}\exp\{-\frac{1}{2}\beta x_i^T x_i\}dX \\
&= (2\pi)^{-mR/2}\int_{\mathbb{R}^{m\times R}} \exp\{-\frac{1}{2}\operatorname{Tr}(\alpha A^TA + (1/R)A^TXX^TA + \beta X^TX)\}dX \\
&= (2\pi)^{-mR/2}\exp\{\operatorname{Tr}(-\frac{1}{2}\alpha A^TA)\}\int_{\mathbb{R}^{m\times R}} \exp\{-\frac{1}{2}\operatorname{Tr}(X^T\mathcal{V}_\beta(A)X)\}dX \\
&= (2\pi)^{-mR/2}\exp\{\operatorname{Tr}(-\frac{1}{2}\alpha A^TA)\}\bigg(\int_{\mathbb{R}^m}\exp\{-\frac{1}{2}x^T\mathcal{V}_\beta(A)x\}dx\bigg)^R \\
&= C\exp\{\operatorname{Tr}(-\frac{1}{2}\alpha A^TA)\}/\det(\mathcal{V}_\beta(A))^{R/2}.
\end{aligned}$$

Note the analogy between the forms of $p(X)$ and $p(A)$, which motivates the paramterization of Z by X. We can similarly relate $\mathbb{E}[Z^{-1}|A]$ to the covariance matrix of a Gaussian distribution,

$$\begin{aligned}
\mathbb{E}[Z^{-1}|A] &= \frac{\det(\mathcal{V}_\beta(A))^{R/2}}{C\exp\{\operatorname{Tr}(-\frac{1}{2}\alpha A^TA)\}}\cdot\int_{\mathbb{R}^{m\times R}}\mathcal{V}_\alpha(X)p(A|X)p(X)dX \\
&= \frac{(2\pi)^{-mR/2}\det(\mathcal{V}_\beta(A))^{R/2}}{C\exp\{\operatorname{Tr}(-\frac{1}{2}\alpha A^TA)\}}\cdot\int_{\mathbb{R}^{m\times R}}\mathcal{V}_\alpha(X)p(X)\prod_{i=1}^{R}\det(\mathcal{V}_\alpha(X))^{1/2}\exp\{-\frac{1}{2}a_i^T\mathcal{V}_\alpha(X)a_i\}dX \\
&= \frac{(2\pi)^{-mR/2}\det(\mathcal{V}_\beta(A))^{R/2}}{\exp\{\operatorname{Tr}(-\frac{1}{2}\alpha A^TA)\}}\cdot\int_{\mathbb{R}^{m\times R}}\mathcal{V}_\alpha(X)\prod_{i=1}^{R}\exp\{-\frac{1}{2}\beta x_i^Tx_i\}\exp\{-\frac{1}{2}a_i^T\mathcal{V}_\alpha(X)a_i\}dX \\
&= (2\pi)^{-mR/2}\det(\mathcal{V}_\beta(A))^{R/2}\cdot\int_{\mathbb{R}^{m\times R}}\mathcal{V}_\alpha(X)\prod_{i=1}^{R}\exp\{-\frac{1}{2}x_i^T\mathcal{V}_\beta(A)x_i\}dX \\
&= \alpha I + (2\pi)^{-mR/2}\det(\mathcal{V}_\beta(A))^{R/2}\cdot\int_{\mathbb{R}^{m\times R}}\bigg(\frac{1}{R}\sum_{j=1}^{R}x_jx_j^T\bigg)\prod_{i=1}^{R}\exp\{-\frac{1}{2}x_i^T\mathcal{V}_\beta(A)x_i\}dX \\
&= \alpha I + (2\pi)^{-m/2}\det(\mathcal{V}_\beta(A))^{1/2}\cdot\int_{\mathbb{R}^m}xx^T\exp\{-\frac{1}{2}x^T\mathcal{V}_\beta(A)x\}dx \\
&= \alpha I + \mathcal{V}_\beta(A)^{-1}.
\end{aligned}$$

□

The normalization by $\det(\alpha I + XX^T)^{-R/2}$ in the prior distribution of X in Theorem 1 is motivated by the proof method, but also has the following intuitive interpretation as an uninformative prior. Consider the probability density of the tensor product of the R samples composing A for a fixed X, $p_X(a_1\otimes\cdots\otimes a_R) = p_X(a_1)\cdots p_X(a_R)$. For a fixed X, the generalized variance of

$a_1 \otimes \cdots \otimes a_r$ is then $\det(Z)^R = \det(\alpha I + XX^T)^{-R}$. Hence, in the theorem, the prior distribution probability of the covariance matrices is defined to be proportional to the generalized variance of the joint distribution of the R samples, which means that the distribution of $\log\det(Z)$ is uniform on $\mathbb{R}$ modulo the regularization term (which is needed to ensure a valid probability distribution). The regularization may be avoided altogether by assuming an improper prior distribution for $p(X)$ (the uniform 'distribution' over $\mathbb{R}^{m\times R}$).

3.4 Expectation Estimation

Since $u^{(1)}, \ldots, u^{(N)}$ are independent,

$$\mathbb{E}[\bigotimes_{i=1}^{N} u^{(i)}] = \prod_{i=1}^{N} \mathbb{E}[u^{(i)}].$$

Assuming known covariance matrices, $\mathbb{M}[u^{(i)}]$, for $i = 1, \ldots, N$, we also have $\mathbb{M}[\bigotimes_{i=1}^{N} u^{(i)}]$. Then, we have that $p(\mathbb{E}[u^{(1)}] = \eta^{(1)}, \ldots, \mathbb{E}[u^{(N)}] = \eta^{(N)})$ is maximized whenever the we have minimized the probability of deviation,

$$p\bigg(\mathcal{N} + \frac{1}{R}\sum_{r=1}^{R}\bigotimes_{i=1}^{N} u^{(i,r)} = T - \bigotimes_{i=1}^{N}\eta^{(i)}\bigg)$$
$$= \text{const} \cdot \exp\bigg\{ -\frac{1}{2}\operatorname{vec}\bigg(T - \bigotimes_{i=1}^{N}\eta^{(i)}\bigg)^T \mathbb{M}\bigg[\bigotimes_{i=1}^{N} u^{(i)}\bigg]^{-1} \operatorname{vec}\bigg(T - \bigotimes_{i=1}^{N}\eta^{(i)}\bigg)\bigg\}.$$

The logarithm of the above (log likelihood) implies that finding the best estimation of the expectation of $\{u^{(i)}\}_{i=1}^{N}$, amounts to solving the Mahalanobis distance rank-1 approximation problem,

$$\min_{\eta^{(1)},\ldots,\eta^{(N)}} \bigg\| \operatorname{vec}\bigg(T - \bigotimes_{i=1}^{N}\eta^{(i)}\bigg)\bigg\|_{\mathbb{M}[\bigotimes_{i=1}^{N} u^{(i)}]^{-1}}.$$

References

[1] A. Aubry, A. De Maio, S. Marano, and M. Rosamilia. Structured covariance matrix estimation with missing-(complex) data for radar applications via expectation-maximization. *IEEE Transactions on Signal Processing*, 69:5920–5934, 2021.

[2] A. N. Kolmogorov. *Foundations of the theory of probability.* Chelsea Publishing Company, New York, 1950.

[3] N. Singh and E. Solomonik. Alternating Mahalanobis distance minimization for accurate and well-conditioned CP decomposition. *SIAM Journal on Scientific Computing*, 45(6):A2781–A2812, 2023.