User Prompt

Given a d-way tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ such that the data is unaligned (meaning the tensor $\mathcal{T}$ has missing entries), we consider the problem of computing a CP decomposition of rank r where some modes are infinite-dimensional and constrained to be in a Reproducing Kernel Hilbert Space (RKHS). We want to solve this using an alternating optimization approach, and our question is focused on the mode-k subproblem for an infinite-dimensional mode. For the subproblem, then CP factor matrices $A_1, \dots, A_{k-1}, A_{k+1}, \dots, A_d$ are fixed, and we are solving for A_k.
Our notation is as follows. Let $N = \prod_i n_i$ denote the product of all sizes. Let $n \equiv n_k$ be the size of mode k, let $M = \prod_{i \neq k} n_i$ be the product of all dimensions except k, and assume $n \ll M$. Since the data are unaligned, this means only a subset of $\mathcal{T}$'s entries are observed, and we let $q \ll N$ denote the number of observed entries. We let $T \in \mathbb{R}^{n \times M}$ denote the mode-k unfolding of the tensor $\mathcal{T}$ with all missing entries set to zero. The vec operations creates a vector from a matrix by stacking its columns, and we let $S \in \mathbb{R}^{N \times q}$ denote the selection matrix (a subset of the $N \times N$ identity matrix) such that $S^T \operatorname{vec}(T)$ selects the q known entries of the tensor $\mathcal{T}$ from the vectorization of its mode-k unfolding. We let $Z = A_d \odot \cdots \odot A_{k+1} \odot A_{k-1} \odot \cdots \odot A_1 \in \mathbb{R}^{M \times r}$ be the Khatri-Rao product of the factor matrices corresponding to all modes except mode k. We let $B = TZ$ denote the MTTKRP of the tensor $\mathcal{T}$ and Khatri-Rao product Z. We assume $A_k = KW$ where $K \in \mathbb{R}^{n \times n}$ denotes the psd RKHS kernel matrix for mode k. The matrix W of size $n \times r$ is the unknown for which we must solve. The system to be solved is

$$\left[(Z \otimes K)^T S S^T (Z \otimes K) + \lambda (I_r \otimes K)\right] \operatorname{vec}(W) = (I_r \otimes K) \operatorname{vec}(B). \tag{1}$$

Here, I_r denotes the $r \times r$ identity matrix. This is a system of size $nr \times nr$ Using a standard linear solver costs $O(n^3 r^3)$, and explicitly forming the matrix is an additional expense.
Explain how an iterative preconditioned conjugate gradient linear solver can be used to solve this problem more efficiently. Explain the method and choice of preconditioner. Explain in detail how the matrix-vector products are computed and why this works. Provide complexity analysis. We assume $n, r < q \ll N$. Avoid any computation of order N.

System Result

Theorem 1 (PCG for the RKHS mode-k CP subproblem with missing data)**.** *Let $n := n_k$, $M := \prod_{i \neq k} n_i$, and let $\Omega \subset [n] \times [M]$ be the set of observed indices in the mode-k unfolding with $|\Omega| = q$. Let $S \in \mathbb{R}^{(nM) \times q}$ be the associated selection matrix* *and define the masking operator* $P_\Omega : \mathbb{R}^{n\times M} \to \mathbb{R}^{n\times M}$ *by*

$$(P_\Omega(X))_{ij} := \begin{cases} X_{ij}, & (i,j) \in \Omega, \\ 0, & (i,j) \notin \Omega. \end{cases}$$

Let $K \in \mathbb{R}^{n\times n}$ *be symmetric positive definite and let* $\lambda > 0$*. Let* $Z \in \mathbb{R}^{M\times r}$ *be the Khatri–Rao product of the fixed factors in all modes except* k *and let* $B := TZ \in \mathbb{R}^{n\times r}$*. Consider the linear system for* $W \in \mathbb{R}^{n\times r}$*:*

$$\left[(Z \otimes K)^\top SS^\top (Z \otimes K) + \lambda(I_r \otimes K)\right] \operatorname{vec}(W) = (I_r \otimes K) \operatorname{vec}(B). \tag{2}$$

Then:

(i) (2) *is equivalent to the matrix equation*

$$\mathcal{A}(W) = KB, \qquad \mathcal{A}(W) := K\, P_\Omega(KWZ^\top)\, Z + \lambda KW, \tag{3}$$

and the induced linear operator $\mathcal{A} : \mathbb{R}^{n\times r} \to \mathbb{R}^{n\times r}$ *is self-adjoint and positive definite w.r.t. the Frobenius inner product.*

(ii) A preconditioned conjugate-gradient (PCG) method applied to (2) *can be implemented without ever forming matrices of dimension* nM *or* $N = \prod_i n_i$*. Each PCG iteration requires*

$$O(n^2 r) \quad \textit{work for dense kernel multiplies by } K, \textit{ and}$$

$$O(q\, r) \quad \textit{work to handle missingness,}$$

assuming d *is treated as a constant and each observed index provides the corresponding row of* Z *in* $O(r(d-1))$ *time.*

(iii) Under uniform random sampling of Ω *with density* $\rho := q/(nM)$*, the preconditioner*

$$\mathcal{M} := \rho\,(Z^\top Z \otimes K^2) + \lambda(I_r \otimes K) \tag{4}$$

satisfies $\mathbb{E}[(Z\otimes K)^\top SS^\top (Z\otimes K)] = \rho(Z^\top Z \otimes K^2)$ *and can be applied in* $O(n^2 r + nr^2)$ *time after* $O(n^3 + r^3)$ *preprocessing via eigendecompositions.*

Consequently, the total cost to reach an ε*-accurate solution by PCG is*

$$O\left(n^3 + r^3 + \sum_{m\neq k} n_m r^2\right) \;+\; O\big(k_{\text{iter}}\,(n^2 r + qr)\big)\,,$$

where k_{iter} *is the number of PCG iterations, and all computations avoid order-*N *work.*

Proof. **Step 1: Masking as a projection.** By construction of the selection matrix, for any $X \in \mathbb{R}^{n\times M}$ one has

$$SS^\top \operatorname{vec}(X) = \operatorname{vec}(P_\Omega(X)).$$

Indeed, $SS^\top$ is the diagonal projector on the coordinates indexed by Ω.
Step 2: Rewriting the operator in matrix form. Let $W \in \mathbb{R}^{n\times r}$. Using the characteristic Kronecker identity

$$(A \otimes B) \operatorname{vec}(X) = \operatorname{vec}(BXA^\top), \tag{5}$$

(valid for conforming dimensions; see [2, Eq. (S1)]), we obtain

$$(Z \otimes K) \operatorname{vec}(W) = \operatorname{vec}(KWZ^\top) \in \mathbb{R}^{nM}.$$

Applying the mask and then the adjoint gives

$$(Z\otimes K)^\top SS^\top (Z\otimes K) \operatorname{vec}(W) = (Z^\top \otimes K) \operatorname{vec}(P_\Omega(KWZ^\top)) = \operatorname{vec}\big(K\, P_\Omega(KWZ^\top)\, Z\big),$$

where (5) is used again with $A = Z^\top$, $B = K$. Also,

$$(I_r \otimes K) \operatorname{vec}(W) = \operatorname{vec}(KW), \qquad (I_r \otimes K) \operatorname{vec}(B) = \operatorname{vec}(KB)$$

by (5) with $A = I_r$, $B = K$. Therefore (2) is equivalent to (3).
Step 3: Symmetry and positive definiteness. Since K is symmetric and P_Ω is an orthogonal projector in the Frobenius inner product, $\mathcal{A}$ is self-adjoint. For any $W \neq 0$,

$$\langle W, \mathcal{A}(W)\rangle_F = \langle W, KP_\Omega(KWZ^\top)Z\rangle_F + \lambda\langle W, KW\rangle_F.$$

Using (5) and Step 1, the first term equals

$$\langle (Z \otimes K) \operatorname{vec}(W),\, SS^\top (Z \otimes K) \operatorname{vec}(W)\rangle = \|SS^\top (Z \otimes K) \operatorname{vec}(W)\|_2^2 \geq 0.$$

The second term satisfies $\langle W, KW\rangle_F = \operatorname{tr}(W^\top KW) > 0$ because $K \succ 0$ and $W \neq 0$. Hence $\langle W, \mathcal{A}(W)\rangle_F > 0$, so $\mathcal{A}$ (and the matrix in (2)) is positive definite. Therefore CG/PCG is applicable.
Step 4: Matrix-vector products in $O(n^2 r + qr)$ time. PCG requires repeated evaluation of $v \mapsto \mathcal{A}v$, equivalently $V \mapsto \mathcal{A}(V)$ for $V \in \mathbb{R}^{n\times r}$. Let $Y := KV$ (cost $O(n^2 r)$ if K is dense). We must compute $R := P_\Omega(YZ^\top)\, Z \in \mathbb{R}^{n\times r}$ without forming $YZ^\top$ (an $n\times M$ matrix). Write $\Omega = \{(i_t, j_t)\}_{t=1}^q$. Then, for each observed pair (i_t, j_t),

$$(P_\Omega(YZ^\top))_{i_t j_t} = (YZ^\top)_{i_t j_t} = \langle Y_{i_t,:}, Z_{j_t,:}\rangle,$$

and

$$R_{i_t,:} \mathrel{+}= (YZ^\top)_{i_t j_t}\, Z_{j_t,:}.$$

Thus one pass over Ω accumulates R in $O(qr)$ arithmetic given $Z_{j_t,:}$.

Crucially, Z need not be formed explicitly. Each j_t corresponds to a $(d-1)$-tuple of indices $(i_1, \dots, i_{k-1}, i_{k+1}, \dots, i_d)$, and the corresponding row equals the elementwise product

$$Z_{j_t,:} = A_d(i_d, :) \, * \, \cdots \, * \, A_{k+1}(i_{k+1}, :) \, * \, A_{k-1}(i_{k-1}, :) \, * \, \cdots \, * \, A_1(i_1, :),$$

so retrieving $Z_{j_t,:}$ costs $O(r(d-1))$ and does not depend on M or N; cf. the definition of the Khatri–Rao product and its basic properties [1, §2.6]. Finally, compute KR (cost $O(n^2 r)$) and add λY:

$$\mathcal{A}(V) = KR + \lambda Y.$$

Overall, one operator application costs $O(n^2 r + qr)$ (treating d as constant).

Step 5: Computing $B = TZ$ without order-N work (optional completeness). Let the observed entries of the mode-k unfolding be $\{(i_t, j_t, T_{i_t j_t})\}_{t=1}^{q}$. Then

$$B = TZ = \sum_{t=1}^{q} T_{i_t j_t} \, e_{i_t} \, Z_{j_t,:},$$

so B can be accumulated in one pass over observations in $O(qr)$ time, again without forming T or Z.

Step 6: Preconditioner and its efficient application. Assume the observed set Ω is drawn uniformly at random with replacement from $[n] \times [M]$ and write $\rho := q/(nM)$. Then

$$SS^\top = \sum_{t=1}^{q} e_{\ell_t} e_{\ell_t}^\top, \qquad \ell_t \in [nM],$$

and therefore

$$(Z \otimes K)^\top SS^\top (Z \otimes K) = \sum_{t=1}^{q} a_{\ell_t} a_{\ell_t}^\top, \qquad a_\ell := (Z \otimes K)^\top e_\ell.$$

Taking expectations and using independence,

$$\mathbb{E}\big[(Z \otimes K)^\top SS^\top (Z \otimes K)\big] = q \, \mathbb{E}_{\ell \sim \mathrm{Unif}[nM]}[a_\ell a_\ell^\top].$$

A direct computation with $\ell \equiv (i, j)$ and (5) yields $\mathbb{E}[a_\ell a_\ell^\top] = \frac{1}{nM}(Z^\top Z \otimes K^2)$, hence

$$\mathbb{E}\big[(Z \otimes K)^\top SS^\top (Z \otimes K)\big] = \rho(Z^\top Z \otimes K^2),$$

which motivates (4). (If desired, concentration of this random sum around its mean can be obtained from a matrix Bernstein inequality; see [4, Theorem 1.4].)

To apply $\mathcal{M}^{-1}$ efficiently, observe first that

$$Z^\top Z = \mathop{*}_{m \neq k} (A_m^\top A_m), \tag{6}$$

where $*$ denotes the Hadamard product; this is the Khatri–Rao Gram identity [1, §2.6, identity $(A \odot B)^\top (A \odot B) = A^\top A * B^\top B$]. Thus $Z^\top Z$ can be formed in $O(\sum_{m \neq k} n_m r^2)$ time without constructing Z.

Next, take eigendecompositions

$$K = U \Sigma U^\top, \qquad Z^\top Z = Q \Gamma Q^\top,$$

with $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$ and $\Gamma = \mathrm{diag}(\gamma_1, \dots, \gamma_r)$. Using standard Kronecker diagonalization rules (also recorded in [2, Proposition S1, Eq. (S10)]),

$$\mathcal{M} = (Q \otimes U)\, \mathrm{diag}\big(\rho\, \gamma_j \sigma_i^2 + \lambda \sigma_i\big)_{i \in [n],\, j \in [r]}\, (Q^\top \otimes U^\top).$$

Hence, given a residual vector reshaped as $R \in \mathbb{R}^{n \times r}$, the solution $X = \mathcal{M}^{-1} R$ is obtained by

$$\widetilde{R} = U^\top R Q, \qquad \widetilde{X}_{ij} = \frac{\widetilde{R}_{ij}}{\rho\, \gamma_j \sigma_i^2 + \lambda \sigma_i}, \qquad X = U \widetilde{X} Q^\top.$$

This costs $O(n^2 r + n r^2)$ per application after $O(n^3 + r^3)$ preprocessing.

Step 7: PCG convergence and total complexity. Since $\mathcal{A}$ is SPD, PCG produces iterates $W^{(t)}$ in the Krylov subspaces and satisfies standard error bounds in the energy norm in terms of the (preconditioned) condition number; see, e.g., [3, Theorem 2.2]. Combining Step 4 and Step 6 yields per-iteration cost $O(qr + n^2 r)$ (dominant terms), and the stated total complexity follows. All operations depend on (n, r, q) and the mode sizes $\{n_m\}$, and avoid order-N work. □

References

[1] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[2] J. Wenger and P. Hennig. Supplementary material for *Probabilistic linear solvers for machine learning*. NeurIPS 2020. In particular, Eq. (S1) and Proposition S1.

[3] A. Nishimura, D. Z. M. Sussman, and M. A. Stephens. Conjugate gradient convergence for Bayesian linear regression. *arXiv:1810.12437*, 2018. In particular, Theorem 2.2.

[4] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *arXiv:1004.4389*, 2011. In particular, Theorem 1.4 (Matrix Bernstein).

[5] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.