A MATRIX-FREE PRECONDITIONED CONJUGATE-GRADIENT SOLVER FOR RKHS-CONSTRAINED MODE UPDATES IN INCOMPLETE CP TENSOR DECOMPOSITIONS

Abstract. We study the alternating optimization update for a CP tensor model with missing (unaligned) entries in which one mode is constrained to lie in a reproducing kernel Hilbert space (RKHS). Under the representer parameterization $A_k = KW$, the mode-k update reduces to a linear system of size $nr \times nr$ with $n = n_k$ and CP rank r. Forming the system matrix is infeasible for missing data because it would require objects of size $N = \prod_i n_i$ or $M = \prod_{i \neq k} n_i$, and a direct solve costs $\Theta((nr)^3)$. We derive a rigorously justified matrix-free self-adjoint operator that applies the system matrix using only (i) multiplications by the kernel Gram matrix $K \in \mathbb{R}^{n \times n}$ and (ii) sparse accumulations over the $q \ll N$ observed tensor entries. This enables preconditioned conjugate gradients (PCG) with per-iteration cost $O(qr + n^2 r)$ for dense K, and no computation or storage of order N. We treat the positive semidefinite (psd) kernel subtlety and provide two mathematically sound SPD resolutions: a nugget regularization $K + \varepsilon I$ and an exact range-space reformulation using a factorization $K = LL^\top$. We propose two SPD preconditioners—a robust ridge-term (kernel-block) preconditioner and a Kronecker-spectral preconditioner motivated by a uniform missingness model—and we give a unified, implementation-safe scaling that precludes the common "double-discounting" trap.

1. Introduction

Let $\mathcal{T} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be a d-way data tensor with missing entries. Let $N = \prod_{i=1}^d n_i$ and suppose only $q \ll N$ entries are observed. We consider CP decompositions of rank r with one or more modes constrained to an RKHS. Focusing on a kernelized mode k, a standard ALS/BCD step fixes all factors except mode k and updates the mode-k factor by solving a regularized least-squares problem. Under the representer theorem, the mode-k factor has the finite expansion $A_k = KW$, where $K \in \mathbb{R}^{n \times n}$ is the kernel Gram matrix on the $n = n_k$ mode-k index points and $W \in \mathbb{R}^{n \times r}$ is unknown. The corresponding normal equations form an $nr \times nr$ linear system.

A naive approach forms the dense normal matrix explicitly and applies a dense direct solver, costing $\Theta((nr)^3) = \Theta(n^3 r^3)$. For missing data, explicit formation is additionally prohibitive because the masking operator naturally lives in the ambient dimension N. The purpose of this paper is to give a fully rigorous, self-contained derivation of a matrix-free PCG method that:

- applies the system matrix without forming any object of size N or M,
- uses only the list of q observed indices/values and the fixed CP factors in the other modes,
- admits practical SPD preconditioners whose application is also free of N- or M-scale computation,
- achieves per-iteration cost $O(qr + n^2 r)$ for dense K (and less if K admits faster multiplies or low-rank structure).

2. Notation and the missing-data operator

Fix $k \in \{1, \ldots, d\}$. Define

$$n := n_k, \qquad M := \prod_{i \neq k} n_i, \qquad N := nM.$$

Let $\Omega \subseteq [n_1] \times \cdots \times [n_d]$ denote the set of observed tensor indices and $|\Omega| = q$. We store the data as pairs $\{(\boldsymbol{i}^{(\ell)}, t_\ell)\}_{\ell=1}^q$ with $\boldsymbol{i}^{(\ell)} = (i_1^{(\ell)}, \ldots, i_d^{(\ell)}) \in \Omega$ and $t_\ell = \mathcal{T}_{\boldsymbol{i}^{(\ell)}} \in \mathbb{R}$.

Let $T \in \mathbb{R}^{n \times M}$ denote the mode-k unfolding of $\mathcal{T}$ with all missing entries set to 0. We will never form T explicitly.

Date: February 13, 2026.

2.1. **Selection and projection operators.** We use a selection matrix formalism only for analysis, not computation.

Assumption 1 (Distinct observations)**.** The set Ω contains no repeated tensor indices. Equivalently, the observed unfolding coordinates form a set $\Omega_k \subseteq [n] \times [M]$ with no duplicates.

Definition 1 (Selection matrix and mask projector)**.** Let $\mathrm{vec}(\cdot)$ stack columns. There exists a matrix $S \in \mathbb{R}^{N\times q}$ whose columns are distinct standard basis vectors in $\mathbb{R}^N$ such that $S^\top \mathrm{vec}(T) \in \mathbb{R}^q$ extracts the observed entries of T. Define

$$P_\Omega := SS^\top \in \mathbb{R}^{N\times N}.$$

Then P_Ω is a diagonal orthogonal projector: $P_\Omega^2 = P_\Omega = P_\Omega^\top \succeq 0$.

Remark 1 (If duplicates are present)**.** If Ω is a multiset (repeated observations of the same entry), then one can incorporate multiplicities as weights. In that case the natural "mask" becomes a diagonal weight matrix W_Ω rather than an orthogonal projector. All matrix-free identities in this paper extend verbatim with P_Ω replaced by W_Ω; CG/PCG then applies provided the resulting system remains SPD (e.g. if $\lambda > 0$ and $K \succ 0$). For clarity we work under Assumption 1.

3. CP factors, RKHS parameterization, and the mode-k subproblem

Let $r \in \mathbb{N}$ be the CP rank. For each mode $j \neq k$ we have a fixed factor matrix $A_j \in \mathbb{R}^{n_j\times r}$. Define the Khatri–Rao product over all modes except k:

$$Z := A_d \odot \cdots \odot A_{k+1} \odot A_{k-1} \odot \cdots \odot A_1 \in \mathbb{R}^{M\times r},$$

where $\odot$ is the columnwise Kronecker product. We will *not* form Z.

Let $K \in \mathbb{R}^{n\times n}$ be a symmetric psd kernel Gram matrix on the mode-k indices. The RKHS representer form (columnwise) yields the parameterization

$$A_k = KW, \qquad W \in \mathbb{R}^{n\times r} \text{ unknown.}$$

Set $w := \mathrm{vec}(W) \in \mathbb{R}^{nr}$.

3.1. **The masked least-squares objective.** The standard kernelized, masked ALS subproblem is

$$\min_{W\in\mathbb{R}^{n\times r}} \frac{1}{2}\left\| S^\top \mathrm{vec}(T) - S^\top (Z \otimes K)\,\mathrm{vec}(W)\right\|_2^2 + \frac{\lambda}{2}\,\mathrm{trace}(W^\top KW), \qquad \lambda > 0. \tag{1}$$

This is a finite-dimensional convex problem in W whenever $K \succeq 0$ and $\lambda > 0$.

3.2. **Normal equations: complete derivation.** Define $y := S^\top \mathrm{vec}(T) \in \mathbb{R}^q$ and $X := S^\top (Z \otimes K) \in \mathbb{R}^{q\times nr}$. Then (1) is the ridge-regularized least squares

$$\min_{w\in\mathbb{R}^{nr}} f(w) := \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2} w^\top (I_r \otimes K)\, w.$$

Since $I_r \otimes K$ is symmetric, the gradient is

$$\nabla f(w) = X^\top (Xw - y) + \lambda (I_r \otimes K) w.$$

Setting $\nabla f(w) = 0$ yields

$$\big(X^\top X + \lambda (I_r \otimes K)\big) w = X^\top y. \tag{2}$$

Substituting $X = S^\top (Z \otimes K)$ and $y = S^\top \mathrm{vec}(T)$ gives

$$\Big[(Z \otimes K)^\top SS^\top (Z \otimes K) + \lambda (I_r \otimes K)\Big] w = (Z \otimes K)^\top SS^\top \mathrm{vec}(T). \tag{3}$$

By Definition 1, $P_\Omega = SS^\top$, so this is the stated system matrix.

4. PSD vs. SPD and two rigorous SPD resolutions

CG/PCG requires an SPD linear operator. The normal matrix in (3) is always symmetric and psd, but it need not be SPD if K is singular. We make this precise.

Proposition 1 (Quadratic form and psd property)**.** *Let $K \succeq 0$ and $\lambda \geq 0$. Define*

$$A := (Z \otimes K)^\top P_\Omega (Z \otimes K) + \lambda (I_r \otimes K) \in \mathbb{R}^{nr \times nr}.$$

Then A is symmetric and positive semidefinite. Moreover, for any $W \in \mathbb{R}^{n \times r}$,

$$\operatorname{vec}(W)^\top A \operatorname{vec}(W) = \left\| P_\Omega\big(KWZ^\top\big) \right\|_{\mathrm{F}}^2 + \lambda \operatorname{trace}(W^\top K W). \tag{4}$$

Proof. Symmetry is immediate from $P_\Omega = P_\Omega^\top$ and $K = K^\top$. Let $w = \operatorname{vec}(W)$. Since P_Ω is an orthogonal projector,

$$w^\top (Z \otimes K)^\top P_\Omega (Z \otimes K) w = \|(P_\Omega^{1/2})(Z \otimes K) w\|_2^2 = \|P_\Omega (Z \otimes K) w\|_2^2.$$

Using Lemma 1 below, $(Z \otimes K)\operatorname{vec}(W) = \operatorname{vec}(KWZ^\top)$, hence the first term equals $\|P_\Omega \operatorname{vec}(KWZ^\top)\|_2^2 = \|P_\Omega(KWZ^\top)\|_F^2$. Finally, $w^\top (I_r \otimes K) w = \operatorname{trace}(W^\top K W)$, yielding (4). □

Corollary 1 (SPD condition)**.** *If $K \succ 0$ and $\lambda > 0$, then $A \succ 0$ (SPD).*

Proof. If $K \succ 0$, then $\operatorname{trace}(W^\top K W) = \|K^{1/2} W\|_F^2 > 0$ for any $W \neq 0$. With $\lambda > 0$, (4) gives $\operatorname{vec}(W)^\top A \operatorname{vec}(W) > 0$ for all $W \neq 0$. □

Remark 2 (Why K singular implies A singular)**.** If $K \succeq 0$ is singular, take any nonzero W whose columns lie in $\ker(K)$. Then $KW = 0$, so both terms in (4) vanish and A is singular. This reflects the non-identifiability of W under $A_k = KW$ when K is psd.

We now give two standard, fully rigorous SPD resolutions.

4.1. Resolution I: nugget regularization (modifies the objective).

Assumption 2 (Nugget SPD kernel)**.** Fix $\varepsilon > 0$ and replace K by $\widetilde{K} := K + \varepsilon I_n$, so $\widetilde{K} \succ 0$.

Remark 3 (Effect on the objective)**.** Replacing K by $\widetilde{K}$ changes the regularizer: $\operatorname{trace}(W^\top K W)$ becomes $\operatorname{trace}(W^\top K W) + \varepsilon \|W\|_F^2$. Thus we add a small Euclidean ridge on W, yielding strict convexity and an SPD normal matrix. This "nugget" is standard in numerical kernel methods.

4.2. Resolution II: range-space reformulation (does not modify the objective).

Let $K \succeq 0$ and let $m := \operatorname{rank}(K)$. Let $K = LL^\top$ with $L \in \mathbb{R}^{n \times m}$ full column rank (e.g. using the positive-eigenspace factorization).

Theorem 1 (Exact range-space equivalence)**.** *Assume $K \succeq 0$ and $\lambda > 0$. Let $K = LL^\top$ with $L \in \mathbb{R}^{n \times m}$ full column rank. Define $U := L^\top W \in \mathbb{R}^{m \times r}$ and $A_k := KW = LU$. Then the problem* (1) *is equivalent (same minimal value and same set of achievable factors A_k) to*

$$\min_{U \in \mathbb{R}^{m \times r}} \ \frac{1}{2} \left\| S^\top \operatorname{vec}(T) - S^\top (Z \otimes L) \operatorname{vec}(U) \right\|_2^2 + \frac{\lambda}{2} \ \|U\|_{\mathrm{F}}^2 . \tag{5}$$

Moreover, (5) *has a unique minimizer and its normal equations are SPD of size $mr \times mr$.*

Proof. First, $A_k = KW = LL^\top W = L(L^\top W) = LU$, so any W yields a corresponding U producing the same factor A_k. Conversely, any U defines a factor $A_k = LU$ which lies in $\operatorname{range}(K)$; choosing any W with $L^\top W = U$ (e.g. $W = L(L^\top L)^{-1} U$) recovers $A_k = KW$.

Second, the data term satisfies $KWZ^\top = LUZ^\top$. By Lemma 1, $\operatorname{vec}(LUZ^\top) = (Z \otimes L)\operatorname{vec}(U)$, so the data term in (1) depends only on U. Third, the regularizer satisfies

$$\operatorname{trace}(W^\top K W) = \operatorname{trace}(W^\top LL^\top W) = \operatorname{trace}\big((L^\top W)^\top (L^\top W)\big) = \|U\|_F^2.$$

Thus (1) reduces exactly to (5).

Finally, (5) is a ridge-regularized least squares in $\operatorname{vec}(U)$ with ridge matrix λI_{mr}. Hence its Hessian is $X^\top X + \lambda I_{mr} \succ 0$ and the minimizer is unique. □

Remark 4 (Correct computational interpretation)**.** The range-space formulation does *not* reduce the number of kernel-type multiplications per matvec; it replaces multiplications by the dense $n \times n$ matrix K with multiplications by L and $L^\top$ of cost $O(nmr)$ each. It is faster only when $m \ll n$.

In the remainder we assume an SPD kernel matrix (either $K \succ 0$ or K has been replaced by $\widetilde{K} \succ 0$), and for simplicity we write K for the SPD matrix.

5. Vec–Kronecker identities and the matrix-form operator

We now prove the standard vec–Kronecker identity in a rectangular form sufficient for all applications in this paper.

Lemma 1 (Rectangular vec–Kronecker identity)**.** *Let $A \in \mathbb{R}^{p\times n}$, $X \in \mathbb{R}^{n\times m}$, and $B \in \mathbb{R}^{m\times k}$. Then*

$$(B^\top \otimes A)\,\mathrm{vec}(X) = \mathrm{vec}(AXB). \tag{6}$$

In particular, for $Z \in \mathbb{R}^{M\times r}$ and $K \in \mathbb{R}^{n\times n}$,

$$(Z \otimes K)\,\mathrm{vec}(X) = \mathrm{vec}(KXZ^\top), \qquad (Z \otimes K)^\top \mathrm{vec}(Y) = \mathrm{vec}(KYZ),$$

for all conforming $X \in \mathbb{R}^{n\times r}$ and $Y \in \mathbb{R}^{n\times M}$.

Proof. Let $E_{ij} \in \mathbb{R}^{n\times m}$ be the matrix with a 1 at (i,j) and zeros elsewhere. Then $\{E_{ij}\}$ is a basis of $\mathbb{R}^{n\times m}$ and $\mathrm{vec}(E_{ij}) = e_j \otimes e_i$ where e_i is the i-th standard basis vector. Compute

$$(B^\top \otimes A)\,\mathrm{vec}(E_{ij}) = (B^\top \otimes A)(e_j \otimes e_i) = (B^\top e_j) \otimes (Ae_i).$$

Expanding $B^\top e_j = \sum_{\beta=1}^{k} (B^\top)_{\beta j} e_\beta = \sum_{\beta=1}^{k} B_{j\beta}\, e_\beta$ and $Ae_i = \sum_{\alpha=1}^{p} A_{\alpha i} e_\alpha$ gives

$$(B^\top e_j) \otimes (Ae_i) = \sum_{\beta=1}^{k}\sum_{\alpha=1}^{p} B_{j\beta} A_{\alpha i} (e_\beta \otimes e_\alpha) = \mathrm{vec}\left(\sum_{\alpha=1}^{p}\sum_{\beta=1}^{k} A_{\alpha i} B_{j\beta}\, E_{\alpha\beta}\right),$$

where now $E_{\alpha\beta} \in \mathbb{R}^{p\times k}$. But for all α, β,

$$(AE_{ij}B)_{\alpha\beta} = \sum_{x=1}^{n}\sum_{y=1}^{m} A_{\alpha x} (E_{ij})_{xy} B_{y\beta} = A_{\alpha i} B_{j\beta},$$

so $\sum_{\alpha,\beta} A_{\alpha i} B_{j\beta} E_{\alpha\beta} = AE_{ij}B$ and therefore $(B^\top \otimes A)\,\mathrm{vec}(E_{ij}) = \mathrm{vec}(AE_{ij}B)$. By linearity, (6) holds for all $X = \sum_{i,j} X_{ij} E_{ij}$. The special cases follow by taking $A = K$, X as indicated, and $B = Z^\top$ or $B = Z$. □

5.1. **Normal equations in the problem-statement form.** Since T is defined to be zero on missing entries, $P_\Omega\,\mathrm{vec}(T) = \mathrm{vec}(T)$. Thus the right-hand side of (3) is $(Z \otimes K)^\top \mathrm{vec}(T) = (Z^\top \otimes K)\,\mathrm{vec}(T)$. By Lemma 1,

$$(Z^\top \otimes K)\,\mathrm{vec}(T) = \mathrm{vec}(KTZ) = \mathrm{vec}(KB) = (I_r \otimes K)\,\mathrm{vec}(B),$$

where $B := TZ$. Thus the system is exactly the one given in the problem statement:

$$\Big[(Z \otimes K)^\top P_\Omega (Z \otimes K) + \lambda (I_r \otimes K)\Big]\,\mathrm{vec}(W) = (I_r \otimes K)\,\mathrm{vec}(B). \tag{7}$$

5.2. **Matrix-form linear operator.** Define the linear map $\mathcal{A} : \mathbb{R}^{n\times r} \to \mathbb{R}^{n\times r}$ by

$$\boxed{\mathcal{A}(X) := K\big(P_\Omega(KXZ^\top)\,Z\big) + \lambda KX.} \tag{8}$$

Proposition 2 (Operator equivalence and SPD)**.** *Assume $K \succ 0$ and $\lambda > 0$. Then for all $X \in \mathbb{R}^{n\times r}$,*

$$\mathrm{vec}(\mathcal{A}(X)) = \Big[(Z \otimes K)^\top P_\Omega (Z \otimes K) + \lambda (I_r \otimes K)\Big]\,\mathrm{vec}(X).$$

Moreover, $\mathcal{A}$ is self-adjoint and SPD with respect to the Frobenius inner product:

$$\langle X, \mathcal{A}(X)\rangle_{\mathrm{F}} = \Big\| P_\Omega(KXZ^\top)\Big\|_{\mathrm{F}}^2 + \lambda \|K^{1/2} X\|_F^2 > 0 \quad \text{for } X \neq 0.$$

Proof. Let $x = \mathrm{vec}(X)$. By Lemma 1, $(Z \otimes K)x = \mathrm{vec}(KXZ^\top)$. Apply P_Ω and apply $(Z \otimes K)^\top = (Z^\top \otimes K)$ using Lemma 1:

$$(Z \otimes K)^\top P_\Omega (Z \otimes K)x = (Z^\top \otimes K)\,\mathrm{vec}(P_\Omega(KXZ^\top)) = \mathrm{vec}(K(P_\Omega(KXZ^\top))Z).$$

Also $(I_r \otimes K)\,\mathrm{vec}(X) = \mathrm{vec}(KX)$, yielding (8). Self-adjointness and SPD follow from Corollary 1 and the vec isometry. □

6. Observed-index access to Z: no M- or N-sized objects

The operator (8) appears to involve $KXZ^\top \in \mathbb{R}^{n\times M}$. We show that $P_\Omega(\cdot)$ and multiplication by Z can be executed using only the observed entries.

6.1. Row evaluation of the Khatri–Rao product.

For each observation $\ell \in \{1, \dots, q\}$, define the mode-k row index

$$i^{(\ell)} := i_k^{(\ell)} \in \{1, \dots, n\}.$$

Define the associated *Khatri–Rao row vector* as the column vector $z^{(\ell)} \in \mathbb{R}^r$ with entries

$$z_j^{(\ell)} := \prod_{t \neq k} A_t\big[i_t^{(\ell)}, j\big], \qquad j = 1, \dots, r, \tag{9}$$

i.e. elementwise multiplication of the factor rows. Equivalently,

$$z^{(\ell)} = \Big(A_d[i_d^{(\ell)}, :] * \cdots * A_{k+1}[i_{k+1}^{(\ell)}, :] * A_{k-1}[i_{k-1}^{(\ell)}, :] * \cdots * A_1[i_1^{(\ell)}, :]\Big)^\top \in \mathbb{R}^{r\times 1}.$$

This orientation choice (column vector) avoids row/column ambiguity and keeps the scatter-add updates type-consistent.

Computing $z^{(\ell)}$ costs $O((d-1)r)$ flops.

Remark 5 (Precompute vs. on-the-fly)**.** Since all A_t $(t \neq k)$ are fixed during the mode-k solve, one may precompute and store $\{z^{(\ell)}\}_{\ell=1}^q$ once per outer step: time $O(q(d-1)r)$, memory $O(qr)$. If memory is constrained, one can recompute $z^{(\ell)}$ on-the-fly in each matvec, increasing the per-iteration masked cost by a factor $d-1$.

7. Matrix-free matrix-vector products

We now show how to apply $\mathcal{A}$ in (8) using only: (i) dense kernel multiplications by K (or faster kernel multiplies if available), and (ii) scatter-add over the q observed entries.

7.1. Derivation of the scatter-add formula.

Let $X \in \mathbb{R}^{n\times r}$ and set

$$G := KX \in \mathbb{R}^{n\times r}.$$

Then $KXZ^\top = GZ^\top$. For each observation ℓ, the corresponding observed entry of $GZ^\top$ equals

$$(GZ^\top)_{i^{(\ell)}, m^{(\ell)}} = G[i^{(\ell)}, :]\; z^{(\ell)},$$

where $G[i, :] \in \mathbb{R}^{1\times r}$ is a row and $z^{(\ell)} \in \mathbb{R}^{r\times 1}$ is a column, so this is a scalar.

Define $u_\ell := G[i^{(\ell)}, :]\; z^{(\ell)} \in \mathbb{R}$. The matrix $P_\Omega(GZ^\top)$ is zero except at observed locations, where it equals u_ℓ. Now define

$$H := P_\Omega(GZ^\top)\, Z \in \mathbb{R}^{n\times r}.$$

Only observed entries contribute, yielding the rowwise formula

$$H[i, :] = \sum_{\ell:\, i^{(\ell)} = i} u_\ell\, z^{(\ell)\top}, \qquad i = 1, \dots, n. \tag{10}$$

Finally,

$$\mathcal{A}(X) = KH + \lambda G.$$

Algorithm 1 Matrix-free matvec $Y = \mathcal{A}(X)$

Require: $X \in \mathbb{R}^{n\times r}$, SPD kernel matrix $K \in \mathbb{R}^{n\times n}$, ridge $\lambda > 0$, observations $\{i^{(\ell)}, z^{(\ell)}\}_{\ell=1}^{q}$ (with $z^{(\ell)} \in \mathbb{R}^{r\times 1}$).
Ensure: $Y = \mathcal{A}(X) \in \mathbb{R}^{n\times r}$.
$G \leftarrow KX$ ▷ $O(n^2 r)$ for dense K
$H \leftarrow 0 \in \mathbb{R}^{n\times r}$
for $\ell = 1, \dots, q$ **do**
  $u \leftarrow G[i^{(\ell)}, :]\, z^{(\ell)}$ ▷ scalar; $O(r)$
  $H[i^{(\ell)}, :] \leftarrow H[i^{(\ell)}, :] + u\, z^{(\ell)\top}$ ▷ $O(r)$
end for
$Y \leftarrow KH + \lambda G$ ▷ $O(n^2 r)$ for dense K
return Y

7.2. Correctness proof of the scatter-add implementation.

Proposition 3 (Scatter-add equals masked multiplication)**.** *Let $G \in \mathbb{R}^{n\times r}$ and define H by* (10)*. Then $H = P_\Omega(GZ^\top)\, Z$ without forming Z or P_Ω explicitly.*

Proof. Let $F := GZ^\top \in \mathbb{R}^{n\times M}$, so $F_{i,m} = G[i,:]\, Z[m,:]^\top$. By definition, $(P_\Omega F)_{i,m} = F_{i,m}$ if $(i,m) \in \Omega_k$ and 0 otherwise. Then

$$(P_\Omega F)Z \text{ has row } i \text{ equal to } \sum_{m=1}^{M} (P_\Omega F)_{i,m}\, Z[m,:].$$

Only observed (i,m) contribute, and each such contribution equals $F_{i,m}\, Z[m,:]$. Indexing observed pairs by ℓ and setting $z^{(\ell)} = Z[m^{(\ell)},:]^\top$ yields exactly (10). □

7.3. Matrix-free matvec algorithm and cost.

Proposition 4 (Matvec complexity; no N or M objects)**.** *Assume dense K. Given precomputed $\{z^{(\ell)}\}_{\ell=1}^{q}$, Algorithm 1 computes $Y = \mathcal{A}(X)$ with cost*

$$O(n^2 r) \;+\; O(qr) \;+\; O(n^2 r) \;=\; O(qr + n^2 r).$$

No vector of length N, no matrix of size $n \times M$, and no explicit Z or Kronecker product is formed. If $z^{(\ell)}$ are computed on-the-fly, the masked loop cost becomes $O(q(d-1)r)$.

Proof. Immediate from the two dense kernel multiplications and the q scatter-add updates, each $O(r)$. □

7.4. **Optional preprocessing: row-wise Gram compression.** Define rowwise Gram matrices

$$C_i := \sum_{\ell:\, i^{(\ell)}=i} z^{(\ell)} z^{(\ell)\top} \in \mathbb{R}^{r\times r}, \qquad i = 1, \dots, n.$$

This costs $O(qr^2)$ time and $O(nr^2)$ memory per outer ALS step.

Proposition 5 (Row-wise Gram acceleration)**.** *With $\{C_i\}$ precomputed, the masked accumulation in* (10) *satisfies*

$$H[i,:] = G[i,:]\, C_i, \qquad i = 1, \dots, n.$$

Hence the masked step can be computed in $O(nr^2)$ per matvec (instead of $O(qr)$), and the matvec cost becomes $O(n^2 r + nr^2)$ for dense K.

Proof. For fixed i, substitute $u_\ell = G[i,:]z^{(\ell)}$ into (10) and factor:

$$H[i,:] = \sum_{\ell:i^{(\ell)}=i} (G[i,:]z^{(\ell)})\, z^{(\ell)\top} = G[i,:] \sum_{\ell:i^{(\ell)}=i} z^{(\ell)} z^{(\ell)\top} = G[i,:]C_i.$$

□

Algorithm 2 PCG for $\mathcal{A}(W) = KB$ in matrix form

Require: Initial guess $W_0 \in \mathbb{R}^{n\times r}$, tolerance $\tau > 0$, operator $\mathcal{A}$, RHS KB, SPD preconditioner M.
Ensure: Approximate solution W.
$R_0 \leftarrow KB - \mathcal{A}(W_0)$
$U_0 \leftarrow M^{-1}(R_0)$
$D_0 \leftarrow U_0$
for $j = 0, 1, 2, \dots$ **do**
  $Q_j \leftarrow \mathcal{A}(D_j)$
  $\alpha_j \leftarrow \langle R_j, U_j\rangle_{\mathrm{F}} / \langle D_j, Q_j\rangle_{\mathrm{F}}$
  $W_{j+1} \leftarrow W_j + \alpha_j D_j$
  $R_{j+1} \leftarrow R_j - \alpha_j Q_j$
  if $\|R_{j+1}\|_{\mathrm{F}} / \|KB\|_{\mathrm{F}} \le \tau$ **then**
    return W_{j+1}
  end if
  $U_{j+1} \leftarrow M^{-1}(R_{j+1})$
  $\beta_j \leftarrow \langle R_{j+1}, U_{j+1}\rangle_{\mathrm{F}} / \langle R_j, U_j\rangle_{\mathrm{F}}$
  $D_{j+1} \leftarrow U_{j+1} + \beta_j D_j$
end for

8. Right-hand side assembly without T or Z

The right-hand side of (7) is $(I_r \otimes K)\,\mathrm{vec}(B) = \mathrm{vec}(KB)$ with $B = TZ$. We can assemble B from observations.

Proposition 6 (MTTKRP from observed entries)**.** *Let $B = TZ$ where T is the zero-filled mode-k unfolding. Then*

$$B[i,:] = \sum_{\ell:\, i^{(\ell)}=i} t_\ell\, z^{(\ell)\top}, \qquad i = 1, \dots, n.$$

Thus B can be computed in $O(qr)$ flops given $\{z^{(\ell)}\}$, and KB then costs $O(n^2 r)$ for dense K.

Proof. By definition, $(TZ)[i,:] = \sum_{m=1}^{M} T_{i,m} Z[m,:]$. Since $T_{i,m} = 0$ for unobserved (i,m), only observed entries contribute. Each observation ℓ in row i contributes $t_\ell Z[m^{(\ell)},:] = t_\ell z^{(\ell)\top}$, yielding the formula. □

9. PCG in matrix form and convergence

We solve the SPD linear system

$$\mathcal{A}(W) = KB, \qquad W \in \mathbb{R}^{n\times r}, \tag{11}$$

where $\mathcal{A}$ is given by (8) and KB is assembled via Proposition 6. Because vec is an isometry between $(\mathbb{R}^{n\times r}, \langle\cdot,\cdot\rangle_{\mathrm{F}})$ and $(\mathbb{R}^{nr}, \langle\cdot,\cdot\rangle)$, CG/PCG applied to (11) is exactly CG/PCG on (7), but expressed using Frobenius inner products.

9.1. PCG algorithm.

Let $M : \mathbb{R}^{n\times r} \to \mathbb{R}^{n\times r}$ be an SPD linear preconditioner (self-adjoint and SPD w.r.t. $\langle\cdot,\cdot\rangle_{\mathrm{F}}$). PCG requires one application of $\mathcal{A}$ and one application of M^{-1} per iteration.

9.2. Standard PCG convergence bound.

Theorem 2 (PCG error bound)**.** *Let $\mathcal{A}$ be SPD and M be SPD. Let $W_\star$ solve* (11) *and let W_j be produced by PCG (Algorithm 2) in exact arithmetic. Let $\kappa := \kappa(M^{-1}\mathcal{A})$ be the spectral condition number of the preconditioned operator. Then the energy-norm error satisfies*

$$\|W_j - W_\star\|_{\mathcal{A}} \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{j} \|W_0 - W_\star\|_{\mathcal{A}},$$

where $\|X\|_{\mathcal{A}}^2 := \langle X, \mathcal{A}(X)\rangle_{\mathrm{F}}$.

Proof. This is the classical CG/PCG bound for SPD systems, applied to the vectorized system and transported to matrix form via the vec isometry. □

10. Preconditioners: design, correctness, and costs

We present two SPD preconditioners that respect the "no $O(N)$" requirement.

10.1. **Preconditioner I: kernel-block (ridge-term) preconditioner.** Define

$$M_A := \lambda(I_r \otimes K) \qquad \Longleftrightarrow \qquad M_A(R) = \lambda K R. \tag{12}$$

This is SPD because $K \succ 0$ and $\lambda > 0$. Applying M_A^{-1} reduces to solving

$$\lambda K U = R \quad \text{columnwise.}$$

With a Cholesky factorization $K = LL^\top$, one application costs $O(n^2 r)$ via r triangular solves; the setup cost is $O(n^3)$.

10.2. **Preconditioner II: Kronecker-spectral preconditioner (implementation-safe scaling).** The masked term $(Z \otimes K)^\top P_\Omega (Z \otimes K)$ destroys exact Kronecker structure. Nevertheless, a strong preconditioner can be obtained from an exact full-observation identity and a uniform missingness model. The key implementation point is to *absorb the sampling-rate scaling into a single effective Gram surrogate* to prevent accidental double-scaling.

10.2.1. *Full observation identity.* If all entries are observed, $P_\Omega = I_N$ and

$$(Z \otimes K)^\top (Z \otimes K) = (Z^\top Z) \otimes (K^\top K) = (Z^\top Z) \otimes (K^2), \tag{13}$$

since K is symmetric.

10.2.2. *Uniform missingness model and sampling rate.* Assume a random mask model in which each unfolding coordinate $(i, m) \in [n] \times [M]$ is observed with probability $\rho \in (0, 1]$, so that $\mathbb{E}[P_\Omega] = \rho I_N$. In practice, when the observed set has size q and the ambient tensor has N entries, this sampling rate is computed from the dimensions as

$$\rho := \frac{q}{N}.$$

This is scalar arithmetic in $O(d)$ time (to compute $N = \prod_i n_i$) and does not require any $O(N)$-scale computation.

Proposition 7 (Expected masked data term under uniform sampling)**.** *Assume $\mathbb{E}[P_\Omega] = \rho I_N$. Then*

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

Proof. Linearity of expectation gives $\mathbb{E}[P_\Omega] = \rho I_N$. Hence

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

and (13) completes the proof. □

10.2.3. *Computing $Z^\top Z$ without forming Z.* To maintain strict avoidance of M-sized operations, the exact Gram matrix $G_Z := Z^\top Z$ must not be computed by explicitly forming $Z \in \mathbb{R}^{M \times r}$. Instead, we exploit the structure of the Khatri–Rao product.

Lemma 2 (Khatri–Rao Gram identity)**.** *Let $Z = A_s \odot \cdots \odot A_1$ be a Khatri–Rao product of conforming matrices. Then*

$$Z^\top Z = (A_s^\top A_s) * \cdots * (A_1^\top A_1).$$

Proof. Let z_p denote the p-th column of Z. Then $z_p = \bigotimes_{j=1}^{s} a_j^{(p)}$, where $a_j^{(p)}$ is the p-th column of A_j. For $p, q \in [r]$,

$$(Z^\top Z)_{pq} = z_p^\top z_q = \prod_{j=1}^{s} \big(a_j^{(p)}\big)^\top a_j^{(q)} = \prod_{j=1}^{s} (A_j^\top A_j)_{pq}.$$

Entrywise products are exactly the Hadamard product. □

Thus, if Z is the Khatri–Rao product of the fixed CP factors in all modes $j \neq k$, then $G_Z = Z^\top Z$ can be computed using only the $r \times r$ Gram matrices $A_j^\top A_j$ in $O(\sum_{j\neq k} n_j r^2)$ flops and $O(r^2)$ memory, without forming Z.

10.2.4. *Two Gram surrogates and the correct scaling.* Let $G_Z := Z^\top Z \succeq 0$. An exact-Gram preconditioner matching the expectation in Proposition 7 would use ρG_Z.

If computing G_Z is undesirable, form an observed Gram

$$G_{Z,\Omega} := \sum_{\ell=1}^{q} z^{(\ell)} z^{(\ell)\top} \in \mathbb{R}^{r\times r}.$$

Under uniform sampling, the scaling is:

Proposition 8 (Scaling of the observed Gram)**.** *Assume uniform sampling with rate ρ and recall $N = nM$. Then*

$$\mathbb{E}[G_{Z,\Omega}] = n\rho\, Z^\top Z.$$

Consequently, $\frac{1}{n} G_{Z,\Omega}$ is an unbiased estimator of $\rho\, Z^\top Z$.

Proof. Let $\delta_{i,m} \in \{0,1\}$ indicate whether unfolding coordinate (i,m) is observed. Then $G_{Z,\Omega} = \sum_{i=1}^{n}\sum_{m=1}^{M} \delta_{i,m}\, Z[m,:]^\top Z[m,:]$. Taking expectation and using $\mathbb{E}[\delta_{i,m}] = \rho$ yields

$$\mathbb{E}[G_{Z,\Omega}] = \sum_{i=1}^{n}\sum_{m=1}^{M} \rho\, Z[m,:]^\top Z[m,:] = n\rho \sum_{m=1}^{M} Z[m,:]^\top Z[m,:] = n\rho\, Z^\top Z.$$

Dividing by n gives the unbiasedness statement. □

10.2.5. *Unified, implementation-safe definition.* Define a *single effective Gram surrogate* $\widetilde{G} \succeq 0$ by choosing one of:

$$\widetilde{G} := \begin{cases} \rho\, G_Z, & \text{(exact-Gram, using } G_Z = Z^\top Z\text{)}, \\ \frac{1}{n}\, G_{Z,\Omega}, & \text{(observed-Gram, using } G_{Z,\Omega} = \sum_{\ell=1}^{q} z^{(\ell)} z^{(\ell)\top}\text{)}. \end{cases} \tag{14}$$

We then define the Kronecker-spectral preconditioner by the *single unambiguous formula*

$$\boxed{M_B := \widetilde{G} \otimes K^2 \;+\; \lambda(I_r \otimes K).} \tag{15}$$

This eliminates the risk of "double-discounting" (accidentally multiplying $\frac{1}{n} G_{Z,\Omega}$ by an additional ρ).

10.2.6. *Fast application of M_B^{-1} (with correct variable hygiene).* Let $\widetilde{G} = U_Z \Sigma U_Z^\top$ with $U_Z \in \mathbb{R}^{r\times r}$ orthogonal and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \succeq 0$. Let $K = Q\Lambda Q^\top$ with $Q \in \mathbb{R}^{n\times n}$ orthogonal and $\Lambda = \operatorname{diag}(\kappa_1, \ldots, \kappa_n) \succ 0$. Then $K^2 = Q\Lambda^2 Q^\top$ and

$$M_B = (U_Z \otimes Q)\ \operatorname{diag}\left(\sigma_j\, \kappa_p^2 + \lambda\, \kappa_p\right)_{p=1,\ldots,n;\, j=1,\ldots,r} (U_Z \otimes Q)^\top.$$

For $R \in \mathbb{R}^{n\times r}$, the output $V = M_B^{-1}(R)$ can be computed in matrix form:

$$\widehat{R} := Q^\top R\, U_Z \in \mathbb{R}^{n\times r}, \qquad \widehat{V}_{p,j} := \frac{\widehat{R}_{p,j}}{\sigma_j\, \kappa_p^2 + \lambda\, \kappa_p}, \qquad V := Q\, \widehat{V}\, U_Z^\top. \tag{16}$$

Here U_Z denotes eigenvectors of the $r \times r$ matrix $\widetilde{G}$, while V denotes the $n \times r$ output, avoiding variable shadowing.

Proposition 9 (Cost of M_B^{-1} application)**.** *Assume dense Q and U_Z are applied explicitly. After one-time eigendecompositions costing $O(n^3 + r^3)$, each application of M_B^{-1} via* (16) *costs*

$$O(n^2 r)\ \textit{(for } Q^\top R \textit{ and } Q\widehat{V}\textit{)} \;+\; O(nr^2)\ \textit{(for } RU_Z \textit{ and } \widehat{V}U_Z^\top\textit{)}.$$

No N- or M-sized objects are required.

11. Complexity summary (no $O(N)$ terms)

We summarize one mode-k update under the regime $n, r < q \ll N$ and dense K.

11.1. Per outer ALS/BCD step (setup).

- (Optional) Precompute $z^{(\ell)}$: $O(q(d-1)r)$ time, $O(qr)$ memory (Remark 5).
- Assemble B via Proposition 6: $O(qr)$ time, $O(nr)$ memory.
- Form KB: $O(n^2 r)$ time.
- (Optional) Precompute rowwise Grams C_i: $O(qr^2)$ time, $O(nr^2)$ memory.
- Preconditioner setup:
  - M_A: Cholesky of K: $O(n^3)$ time, $O(n^2)$ memory.
  - M_B:
    - If $\widetilde{G} = \rho G_Z$: compute $\rho = q/N$ and compute $G_Z = Z^\top Z$ via Lemma 2 in $O(\sum_{j\neq k} n_j r^2)$ flops and $O(r^2)$ memory (no M-sized objects).
    - If $\widetilde{G} = \frac{1}{n} G_{Z,\Omega}$: compute $G_{Z,\Omega}$ in $O(qr^2)$ flops and $O(r^2)$ memory.
    - In either case: eigendecompose K and $\widetilde{G}$ in $O(n^3 + r^3)$ time and store $O(n^2 + r^2)$.

11.2. Per PCG iteration.

Each PCG iteration performs:

- One matvec $X \mapsto \mathcal{A}(X)$:
  - without C_i: $O(qr + n^2 r)$ time (Proposition 4);
  - with C_i: $O(nr^2 + n^2 r)$ time (Proposition 5).
- One preconditioner application:
  - M_A^{-1}: $O(n^2 r)$ via triangular solves;
  - M_B^{-1}: $O(n^2 r + nr^2)$ via (16) (Proposition 9).
- Inner products and saxpy updates: $O(nr)$.

Thus, for t PCG iterations, the total cost is

$$O\Big(q(d-1)r \;+\; qr \;+\; n^2 r \;+\; \text{setup}(M) \;+\; t \cdot \big(\text{matvec} + \text{apply}(M^{-1})\big)\Big),$$

with *no* computation or storage of order $N = nM$ and no explicit construction of Z, T, P_Ω, or any Kronecker product.

12. Conclusion

The RKHS-constrained mode-k subproblem in incomplete CP tensor decompositions leads to a large $nr \times nr$ normal system whose direct solution and explicit formation are computationally prohibitive. We derived an exact matrix-form SPD operator whose application is computable from q observed entries and kernel multiplications, yielding a matrix-free PCG solver with per-iteration complexity $O(qr + n^2 r)$ for dense kernels. We treated the psd-kernel singularity rigorously, providing two SPD resolutions: nugget regularization and a range-space reformulation. Finally, we proposed two SPD preconditioners and (crucially) gave a unified, implementation-safe scaling of the Kronecker-spectral variant that precludes double-discounting under uniform missingness, while retaining a strict guarantee that no M- or N-sized objects are required.

References

[1] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, 2003.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, 2013.

[3] T. G. Kolda and B. W. Bader, Tensor decompositions and applications, *SIAM Review* **51** (2009), no. 3, 455–500.

[4] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, Scalable tensor factorizations for incomplete data, *Chemometrics and Intelligent Laboratory Systems* **106** (2011), 41–56.

[5] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.