PCG for the RKHS CP-ALS mode-k subproblem (missing data)

Fix all CP factors except the (possibly infinite-dimensional) RKHS mode k. Write the mode-k unfolding as $T \in \mathbb{R}^{n\times M}$ with missing entries set to zero, and let $S \in \mathbb{R}^{N\times q}$ (with $N = nM$) be the selection matrix so that $S^\top \operatorname{vec}(T)$ extracts the q observed entries. Let $Z \in \mathbb{R}^{M\times r}$ be the Khatri–Rao product of the other CP factors and $B = TZ$. Assume the RKHS representer form $A_k = KW$, where $K \in \mathbb{R}^{n\times n}$ is symmetric psd. The ALS subproblem in $W \in \mathbb{R}^{n\times r}$ is the linear system

$$\Big[(Z \otimes K)^\top SS^\top (Z \otimes K) + \lambda(I_r \otimes K)\Big] \operatorname{vec}(W) = (I_r \otimes K)\operatorname{vec}(B) = \operatorname{vec}(KB), \tag{1}$$

of size $nr \times nr$. The goal is to solve (1) without forming the dense matrix and without any $O(N)$ work, assuming $n, r < q \ll N$.

1. SPD and why PCG applies. Let $P \equiv SS^\top$ (a diagonal mask, $P = P^\top = P^2$) and define

$$A \equiv (Z \otimes K)^\top P(Z \otimes K) + \lambda(I_r \otimes K), \qquad b \equiv \operatorname{vec}(KB).$$

Then A is symmetric. If $K \succ 0$ and $\lambda > 0$, for any $x \neq 0$,

$$x^\top Ax = \|P^{1/2}(Z \otimes K)x\|_2^2 + \lambda\, x^\top (I_r \otimes K)x > 0,$$

so $A \succ 0$ and (preconditioned) conjugate gradients (PCG) is applicable. If K is only psd, add a small nugget εI to K (standard in kernel ridge regression) or reduce to the rank-m eigenspace of K to obtain an SPD system of size mr.

2. Matvecs in $O(n^2r + qr)$ using gather/scatter

Write $x \in \mathbb{R}^{nr}$ as $x = \operatorname{vec}(X)$ with $X \in \mathbb{R}^{n\times r}$ (column-stacked). Use the identity

$$(Z \otimes K)\operatorname{vec}(X) = \operatorname{vec}(KXZ^\top), \tag{2}$$

so that the action of $(Z \otimes K)$ is “form the prediction matrix” $U \equiv KXZ^\top \in \mathbb{R}^{n\times M}$.

Observed index list. Store the q observed indices in unfolding coordinates as pairs (i_t, j_t), $t = 1, \dots, q$. Then $S^\top \operatorname{vec}(U) = (U_{i_t,j_t})_{t=1}^q$ (gather). Conversely, for $u \in \mathbb{R}^q$, the vector $Su \in \mathbb{R}^N$ has entries $(Su)_{(i_t,j_t)} = u_t$ and zeros elsewhere; reshaping Su into $n \times M$ via vec^{-1} gives the corresponding sparse matrix (scatter). These operations cost $O(q)$ given the index arrays.

Matvec $y = Ax$. Given X:

1. $\Gamma \leftarrow KX$ $\qquad (O(n^2r))$.
2. For each observation t compute a row vector $z_t \equiv Z_{j_t,:} \in \mathbb{R}^r$ and the scalar
$$u_t \leftarrow \langle \Gamma_{i_t,:}, z_t\rangle. \tag{3}$$
3. Accumulate $H \in \mathbb{R}^{n\times r}$ via
$$H_{i_t,:} \mathrel{+}= u_t\, z_t, \qquad t = 1, \dots, q. \tag{4}$$

4. Output $\operatorname{vec}\big(KH + \lambda\Gamma\big)$.

To see correctness: $U = KXZ^\top$ implies $P\operatorname{vec}(U) = S(S^\top \operatorname{vec}(U)) = Su$ where $u_t = U_{i_t,j_t}$. Let $\widetilde{U} \in \mathbb{R}^{n\times M}$ be the reshape of Su, so $\operatorname{vec}(\widetilde{U}) = P\operatorname{vec}(U)$ and $\widetilde{U}_{i_t,j_t} = u_t$. The adjoint Kronecker identity gives

$$(Z \otimes K)^\top \operatorname{vec}(\widetilde{U}) = \operatorname{vec}(K\widetilde{U}Z),$$

and the sparse accumulation (4) computes $H = \widetilde{U}Z$ without ever materializing $\widetilde{U}$. Thus the first term equals $\operatorname{vec}(KH)$, and adding $\lambda(I_r \otimes K)x = \operatorname{vec}(\lambda KX) = \operatorname{vec}(\lambda\Gamma)$ yields $y = Ax$.

Avoiding explicit Z (avoiding M and N). Although Z is of size $M \times r$, we never form it. Given an observed tensor multi-index $(i_1^{(t)}, \dots, i_d^{(t)})$, the required row is

$$z_t = A_d(i_d^{(t)}, :) \odot \cdots \odot A_{k+1}(i_{k+1}^{(t)}, :) \odot A_{k-1}(i_{k-1}^{(t)}, :) \odot \cdots \odot A_1(i_1^{(t)}, :),$$

computable on the fly in $O((d-1)r)$ time. If memory allows, cache all z_t once in a $q \times r$ array to make each PCG iteration cost $O(qr)$ for the sparse part.

RHS. Compute $B = TZ$ without forming T: for each observed value y_t at (i_t, j_t), do $B_{i_t,:} \mathrel{+}= y_t z_t$ (same sparse accumulation as above), then set $b = \operatorname{vec}(KB)$. Cost: $O(qr + n^2 r)$ (or $O(qdr + n^2 r)$ if computing z_t on the fly).

3. A Kronecker preconditioner and fast application

A convenient SPD preconditioner replaces P by a scaled identity αI. Under uniform sampling, $\mathbb{E}[P] = (q/N)I$ so a natural default is $\alpha = q/N$. More generally one can choose α by matching traces,

$$\alpha := \frac{\operatorname{Tr}\big((Z\otimes K)^\top P (Z \otimes K)\big)}{\operatorname{Tr}\big((Z^\top Z)\otimes K^2\big)} = \frac{\sum_{t=1}^{q} \|K_{:,i_t}\|_2^2 \, \|z_t\|_2^2}{\operatorname{Tr}(K^2)\,\operatorname{Tr}(Z^\top Z)}, \tag{5}$$

which is computable from the observed indices in $O(qr + n^2)$ time (or $O(qr)$ if $\|K_{:,i_t}\|_2^2$ are precomputed).

With this scalar approximation,

$$A_0 \equiv \alpha(Z^\top Z)\otimes(K^2) + \lambda(I_r \otimes K). \tag{6}$$

When there are no missing entries ($P = I$) and $\alpha = 1$, A_0 equals A exactly.

Computing $\Phi = Z^\top Z$ without forming Z. With $Z = A_d \odot \cdots \odot A_{k+1} \odot A_{k-1} \odot \cdots \odot A_1$, the Gram matrix satisfies the standard Khatri–Rao identity

$$\Phi \equiv Z^\top Z = \underset{i\neq k}{\circ} (A_i^\top A_i),$$

where $\circ$ denotes the entrywise (Hadamard) product. This costs $O(\sum_{i\neq k} n_i r^2)$ and is already computed in many CP-ALS implementations.

Applying A_0^{-1}. Let $K = U\Lambda U^\top$ and $\Phi = V\Sigma V^\top$. Then A_0 diagonalizes in the Kronecker basis $(V \otimes U)$, so for $x = \operatorname{vec}(X)$,

$$\widehat{X} \leftarrow U^\top X V, \qquad \widehat{X}_{b,a} \leftarrow \widehat{X}_{b,a}/(\alpha\sigma_a\lambda_b^2 + \lambda\lambda_b), \qquad X \leftarrow U\widehat{X}V^\top.$$

Each application costs $O(n^2 r + nr^2)$ after one-time eigendecompositions ($O(n^3 + r^3)$).

Cheaper block-diagonal alternative. Replacing Φ by $\mathrm{diag}(\Phi)$ yields r independent column-wise preconditioners: if $X = [x_1, \ldots, x_r]$ then $(A_0^{\mathrm{bd}})^{-1}$ applies

$$x_a \;\mapsto\; \big(\alpha\,\Phi_{aa}K^2 + \lambda K\big)^{-1}x_a = \big(K(\alpha\,\Phi_{aa}K + \lambda I)\big)^{-1}x_a,$$

which can be done using only the eigendecomposition of K in $O(n^2)$ per column.

4. Complexity and iteration count (no $O(N)$ terms)

Let m be the PCG iteration count. Per iteration:

$$\text{matvec } Ax:\; O(n^2r + qr) \text{ (or } O(n^2r + qdr)), \qquad \text{preconditioner } A_0^{-1}:\; O(n^2r + nr^2).$$

Thus the PCG solve costs $O\big(m(n^2r + qr + nr^2)\big)$ time and $O(nr + q)$ memory (plus optional $O(qr)$ cache). One-time setup (per outer ALS sweep) includes forming Φ in $O(\sum_{i\neq k} n_i r^2)$ and eigendecompositions in $O(n^3 + r^3)$, which are amortized over the m iterations.

How m depends on the preconditioner. Standard PCG theory gives for the mth iterate x_m:

$$\frac{\|x_m - x_\star\|_A}{\|x_0 - x_\star\|_A} \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m, \qquad \kappa \equiv \kappa(A_0^{-1}A),$$

so $m = O\big(\sqrt{\kappa}\,\log(1/\varepsilon)\big)$ iterations suffice for relative A-norm error ε. Moreover, if $\|A_0^{-1/2}(A - A_0)A_0^{-1/2}\|_2 \le \eta < 1$ then the eigenvalues of $A_0^{-1}A$ lie in $[1-\eta, 1+\eta]$ and

$$\kappa(A_0^{-1}A) \le \frac{1+\eta}{1-\eta}.$$

Under near-uniform sampling and bounded “leverage” of the observed rank-one terms, one can bound η with high probability using matrix Bernstein/Chernoff inequalities, implying κ (and hence m) is $O(1)$ once q is moderately larger than nr (up to log factors).