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 (**??**) 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 $x \neq 0$,

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

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 $O(n^2 r + qr)$ 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) and Su is the sparse $n \times M$ matrix with nonzeros u_t at (i_t, j_t) (scatter). These operations cost $O(q)$ given the index arrays.

Matvec $y = Ax$. Given X:

1. $G \leftarrow KX$ $\qquad (O(n^2 r))$.
2. For each observation t compute a row vector $z_t \equiv Z_{j_t,:} \in \mathbb{R}^r$ and the scalar
$$u_t \leftarrow \langle G_{i_t,:}, z_t\rangle. \tag{3}$$
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 $\operatorname{vec}\big(KH + \lambda G\big)$.

To see correctness: the sparse matrix $\widetilde{U}$ with entries $\widetilde{U}_{i_t,j_t} = u_t$ is exactly the masked prediction $\widetilde{U} = \operatorname{reshape}(P\operatorname{vec}(U))$. Then the adjoint identity $(Z\otimes K)^\top \operatorname{vec}(\widetilde{U}) = \operatorname{vec}(K\widetilde{U}Z)$ gives $\operatorname{vec}(KH)$ since $H = \widetilde{U}Z$ is computed by (**??**). Adding the Tikhonov term yields 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 t_t at (i_t, j_t), do $B_{i_t,:} \mathrel{+}= t_t z_t$ (same sparse accumulation as above), then set $b = \mathrm{vec}(KB)$. 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 with $\alpha \approx q/N$. This is motivated by uniform sampling: $\mathbb{E}[P] = (q/N)I$, hence

$$\mathbb{E}[(Z \otimes K)^\top P (Z \otimes K)] = \alpha\, (Z \otimes K)^\top (Z \otimes K) = \alpha (Z^\top Z) \otimes (K^2).$$

Thus define

$$A_0 \equiv \alpha (Z^\top Z) \otimes (K^2) + \lambda (I_r \otimes K). \tag{5}$$

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

Computing $G = Z^\top Z$ 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

$$G \equiv Z^\top Z = \mathop{*}_{i \neq k} (A_i^\top A_i),$$

with 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 $G = V \Sigma V^\top$. Then A_0 diagonalizes in the Kronecker basis $(V \otimes U)$, so for $x = \mathrm{vec}(X)$,

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

Each application costs $O(n^2 r + n r^2)$ after one-time eigendecompositions ($O(n^3 + r^3)$). A cheaper alternative is the block-diagonal approximation obtained by replacing G by $\mathrm{diag}(G)$, which decouples the r columns.

4. Complexity (no $O(N)$ terms)

Let m be the PCG iteration count. Per iteration:

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

Hence the solve costs $O\big(m(n^2 r + qr + nr^2)\big)$ time and $O(nr + q)$ memory (plus optional $O(qr)$ cache), dramatically improving on the $O((nr)^3) = O(n^3 r^3)$ dense solve and avoiding explicit formation of the $(nr) \times (nr)$ matrix.