Supplementary Material for: Registration-Free Supervised Tensor-PLS for Unaligned Multiparametric MRI

Axel Faes[a,*], Maryam Amir Haeri[a]

[a]*Department of Learning, Data Analytics and Technology, University of Twente, Enschede, The Netherlands*

S1. Full proofs

This supplement gives the population-level analysis of the estimator. Throughout, the estimand is the *unrolled* transport (Definition S1), matching the implementation; we never invoke uniqueness or smoothness of the nonconvex minimizer of Eq. (2) of the main text.

S1.1. Setup and notation

Let $\Omega \subseteq \mathbb{R}^d$ be the spatial domain and $\mathbb{R}^C$ the feature space. A sample is a finite weighted measure $\mu = \sum_{i=1}^{m} a_i \delta_{(p_i, f_i)}$ with $p_i \in \Omega$, $f_i \in \mathbb{R}^C$, $a \in \Delta_m$. We summarize μ by its feature configuration $F = (f_1, \ldots, f_m)$ with marginal a, and its intra-sample structure $C \in \mathbb{R}^{m \times m}$, $C_{ii'} = \kappa(\|p_i - p_{i'}\|)$ for a fixed 1-Lipschitz, bounded κ; C uses only relative geometry. We observe i.i.d. pairs $(\mu_n, Y_n)_{n=1}^{N}$, $Y_n \in \mathbb{R}^q$. A template is $\Theta = (G, L, b, W)$ with atom features $G \in \mathbb{R}^{K \times C}$, latent coordinates $L \in \mathbb{R}^{K \times r}$ inducing $D_{kl} = \kappa(\|L_k - L_l\|)$, marginal $b \in \Delta_K$, and feature metric $W \succ 0$; $\Theta \in \mathbb{T}$, compact. For $\alpha \in [0, 1]$, $\varepsilon > 0$, the feature cost is $M(F, G)_{ik} = \|f_i - g_k\|_W^2$ and the entropic fused Gromov–Wasserstein transport is Eq. (2) of the main text, with $U(a, b) = \{T \geq 0 : T\mathbf{1} = a, T^\top \mathbf{1} = b\}$ and $H(T) = -\sum T_{ik}(\log T_{ik} - 1)$. Problem (2) of the main text is nonconvex, so its global minimizer is not in general unique or smooth in Θ; we therefore do not take it as the estimand.

*Corresponding author.
Email address: axel.faes@utwente.nl (Axel Faes)

Definition S1 (Unrolled transport)**.** Fix outer steps S and inner Sinkhorn steps J. Let $T^{(0)} = ab^\top$, and for $s = 0, \dots, S-1$ set the linearized cost $C^{(s)} = (1-\alpha)M(F,G) + \alpha(C^2 a\mathbf{1}^\top + \mathbf{1}b^\top D^2 - 2CT^{(s)}D)$, $K^{(s)} = \exp(-C^{(s)}/\varepsilon)$, and let $T^{(s+1)}$ be the result of J Sinkhorn updates $u \leftarrow a \oslash (K^{(s)}v)$, $v \leftarrow b \oslash (K^{(s)\top}u)$ from $u = \mathbf{1}$, $T^{(s+1)} = \mathrm{diag}(u)K^{(s)}\mathrm{diag}(v)$. Write $\mathcal{T}_\varepsilon^{S,J}(\mu, \Theta) := T^{(S)}$.

Definition S2 (Representation and predictor)**.** Let $T = \mathcal{T}_\varepsilon^{S,J}(\mu, \Theta)$. The registration-free representation is

$$\Phi(\mu; \Theta) = \Big(\underbrace{T^\top \mathbf{1}}_{\text{atom mass} \in \mathbb{R}^K} , \underbrace{\mathrm{diag}(T^\top \mathbf{1})^{-1} T^\top F}_{\text{barycentric features} \in \mathbb{R}^{K \times C}} \Big) \in \mathbb{R}^{K(1+C)}.$$

The predictor is $h_{\Theta,\beta}(\mu) = \beta^\top \Phi(\mu; \Theta)$ and the population risk is $\mathcal{R}(\Theta, \beta) = \mathbb{E}\|Y - h_{\Theta,\beta}(\mu)\|^2$.

S1.2. Structural generative model

Assumption S1 (Latent atoms)**.** There are $K^\star \le K$ ground-truth atoms with features $g_1^\star, \dots, g_{K^\star}^\star$ that are δ-separated: $\min_{k \neq l} \|g_k^\star - g_l^\star\|_W \ge \delta > 0$.

Assumption S2 (Sampling)**.** Sample n has nonnegative activations $\alpha_n \in \mathbb{R}_{\ge 0}^{K^\star}$ and latent atom positions $c_{n,k} \in \Omega$. Conditioned on (α_n, c_n), atom k contributes mass $\propto \alpha_{n,k}$ with features i.i.d. $g_k^\star + \xi$, ξ σ-sub-Gaussian, located near $c_{n,k}$. A sample-specific reparametrization $\varphi_n : \Omega \to \Omega$ is then applied to all coordinates.

Assumption S3 (Response)**.** $Y_n = \langle B^\star, \psi(\alpha_n, \Gamma_n) \rangle + e_n$ with $\mathbb{E}[e_n \mid \mu_n] = 0$, $\mathrm{Var}(e_n) < \infty$, where $\Gamma_n = (\kappa(\|c_{n,k} - c_{n,l}\|))_{kl}$ is the inter-atom geometry and ψ is a fixed, Lipschitz, reparametrization-invariant functional (of masses and relative inter-atom distances only).

Assumption S4 (Non-degeneracy)**.** $\mathrm{Cov}(\psi(\alpha, \Gamma)) \succ 0$ and $B^\star$ has no zero block, so every coordinate of ψ that ψ varies in is response-relevant.

Assumption S5 (Regularity)**.** Features are bounded, $\|f\| \le R$; $\mathbb{T}$ is compact; $\varepsilon > 0$ is fixed; each active atom contributes at least $\underline{m}$ points.

Exact parameter recovery (Proposition S2) requires the population risk to grow quadratically away from its minimizer. We prove below that this growth holds automatically on the prediction-relevant block, atom features, atom masses, and head (G, b, β), under the assumptions already in force (Lemma S5). The only directions not controlled by feature-separation are the latent geometry coordinates L, which enter the risk only through the quadratic structural cost $\sum(C_{ii'} - D_{kl})^2 T_{ik} T_{i'l}$ with $D = \kappa(\|L_k - L_l\|)$. We isolate exactly this residual into a single standard, checkable second-order identifiability condition.

Assumption S6 (Second-order geometry identifiability)**.** Let $(\Theta^\star, \beta^\star)$ be a population minimizer (Theorem S1) and let $\mathcal{P}$ denote the finite group of atom permutations together with the global isometries of the latent coordinates L that fix $D = \kappa(\|L_k - L_l\|)$ (the intrinsic, unidentifiable symmetries of the structural block). Quotienting these out, the population Hessian $\nabla^2_L \mathcal{R}(\Theta^\star, \beta^\star)$ of the risk restricted to the latent-geometry coordinates is positive definite on the orthogonal complement of the tangent space of the $\mathcal{P}$-orbit at $\Theta^\star$: $v^\top \nabla^2_L \mathcal{R}(\Theta^\star, \beta^\star) v \geq c_L \|v\|^2$ for all $v \perp T_{\Theta^\star}(\mathcal{P} \cdot \Theta^\star)$, some $c_L > 0$.

Remark S1. Assumption S6 is the textbook local (second-order) identifiability condition for a smooth parametric model: a Hessian nonsingular transverse to the model's intrinsic symmetry orbit. It is mild and generically satisfied. By Assumption S4 the response $\psi(\alpha, \Gamma)$ has nondegenerate covariance and genuinely depends on the inter-atom geometry Γ, so the structural cost is a non-flat function of L near $L^\star$ except along the intrinsic isometry symmetry already quotiented out; a degenerate transverse Hessian is a measure-zero coincidence. It concerns only the latent-geometry block: by Lemma S5 the feature/mass/head block always has positive curvature, so the prediction guarantee never depends on Assumption S6. Because $\mathcal{R}$ is an explicit smooth population functional (Lemma S2), $\nabla^2_L \mathcal{R}$ can be evaluated and its transverse spectrum checked numerically, so the condition is verifiable rather than assumed folklore.

S1.3. Invariance and stability

Lemma S1 (Exact invariance to isometric reparametrization)**.** *Let $\varphi \in \mathrm{Isom}(\Omega)$ be distance-preserving and $\varphi_{\#}\mu$ the pushforward (same features and weights, coordinates $p_i \mapsto \varphi(p_i)$). Then for every Θ and every S, J, $\mathcal{T}_\varepsilon^{S,J}(\varphi_{\#}\mu, \Theta) = \mathcal{T}_\varepsilon^{S,J}(\mu, \Theta)$ and hence $\Phi(\varphi_{\#}\mu; \Theta) = \Phi(\mu; \Theta)$.*

Proof. The unrolled map of Definition S1 reads μ only through (F, a, C): every iterate $C^{(s)}$, $K^{(s)}$, $T^{(s)}$ is an explicit function of (F, a, C) and Θ. Pushforward by φ leaves F, a unchanged and leaves C unchanged because $C_{ii'} = \kappa(\|p_i - p_{i'}\|) = \kappa(\|\varphi(p_i) - \varphi(p_{i'})\|)$ for an isometry. Hence every iterate is identical, and so are $T^{(S)}$ and Φ. □

Remark S2. (F, a) is invariant to any coordinate warp; only C requires isometry. The feature-only part of Φ ($\alpha = 0$) is invariant to arbitrary warps, while the structural part ($\alpha > 0$) requires isometry exactly.

Lemma S2 (Smoothness and Lipschitzness of the unrolled map)**.** *Fix $\varepsilon > 0$ and $S, J < \infty$. On the compact set $\mathbb{T} \times \{\|f\| \le R\}$ with marginals bounded below by $\rho := \min(\min_i a_i, b_{\min}) > 0$, the map $(\Theta, F, C) \mapsto \mathcal{T}_\varepsilon^{S,J}$ is C^∞, and both $\Theta \mapsto \Phi(\mu; \Theta)$ and $(\Theta, \beta) \mapsto h_{\Theta,\beta}(\mu)$ are Lipschitz on $\mathbb{T} \times \{\|\beta\| \le B\}$, with explicit constant $\Lambda_\varepsilon = (c_1 \varepsilon^{-1} \rho^{-2})^{S(J+1)} \mathrm{poly}(R, K, \rho^{-1})$ for a universal c_1.*

Proof. Each constituent map in Definition S1 is C^∞ with derivative bounded on the stated compact domain: (i) the cost assembly $C^{(s)}$ is affine in $M(F, G)$ and in $T^{(s)}$ and polynomial in (C, D), hence smooth with derivative bounded in terms of R and $\sup \kappa$; (ii) $X \mapsto \exp(-X/\varepsilon)$ is smooth with sup-norm derivative $\le \varepsilon^{-1}$ on the bounded cost range; (iii) each Sinkhorn update $u \mapsto a \oslash (Kv)$ is a rational map, C^∞ and Lipschitz where row/column sums exceed $\rho\, e^{-\bar{C}/\varepsilon} > 0$ (guaranteed by the marginal lower bound ρ and bounded costs), with Lipschitz constant $O(\varepsilon^{-1}\rho^{-2})$; (iv) the barycentric projection $T \mapsto (T^\top \mathbf{1}, \mathrm{diag}(T^\top \mathbf{1})^{-1} T^\top F)$ is smooth and $O(\rho^{-1}R)$-Lipschitz where atom mass $\ge \rho$. The estimand is the composition of S outer blocks, each a map of type (i)–(ii) and J maps of type (iii), followed by (iv). Smoothness is preserved under composition, and the Lipschitz constant

is bounded by the product of the per-map constants, giving Λ_ε. Linearity of $h_{\Theta,\beta}$ in β contributes the factor $\|\beta\| \le B$. □

Remark S3. Λ_ε is finite for every fixed (ε, S, J) but grows with the unrolled depth and ε^{-1}; this is the explicit price of dropping any nonconvex-uniqueness assumption. It enters the generalization rate (Theorem S2) only logarithmically.

Proposition S1 (Stability to non-isometric warps)**.** *Let φ be bi-Lipschitz with metric distortion $\eta := \|C^\varphi - C\|_\infty = \sup_{p,p'} |\kappa(\|\varphi(p) - \varphi(p')\|) - \kappa(\|p - p'\|)|$. Then, with Λ_ε as in Lemma S2, $\|\Phi(\varphi_\# \mu; \Theta) - \Phi(\mu; \Theta)\| \le \Lambda_\varepsilon \eta$.*

Proof. A warp changes μ only through $C \mapsto C^\varphi$ (F and a unchanged), with $\|C^\varphi - C\|_\infty = \eta$. By Lemma S2 the map $C \mapsto \Phi$ is Λ_ε-Lipschitz on the compact domain; applying it to a perturbation of size η gives the bound. □

Remark S4. Proposition S1 predicts the observed behavior: exact invariance under rigid warps ($\eta = 0$, Lemma S1), and error growing at most linearly in non-rigid distortion η, i.e. controlled, graceful degradation rather than collapse.

S1.4. Atom recovery

Lemma S3 (Barycentric atom recovery)**.** *Under Assumptions S1,S2,S5, suppose the template features equal the truth up to a permutation, $g_{\Pi(k)} = g_k^\star$, and $\varepsilon \le c\,\delta^2 / \log K$ for a universal $c > 0$. Then with probability $\ge 1 - Ke^{-c' \underline{m} \delta^2/\sigma^2}$, the plan matches each point to the template atom of its generating atom, and the barycentric feature obeys*

$$\|[\Phi(\mu;\Theta)]^{\text{feat}}_{\Pi(k)} - g_k^\star\| \le O(\varepsilon) + O(\sigma/\sqrt{\underline{m}}), \qquad [\Phi(\mu;\Theta)]^{\text{mass}}_{\Pi(k)} = \tfrac{\alpha_k}{\sum_j \alpha_j} + O(\varepsilon).$$

Proof. Fix a point i generated by atom k, so $\|f_i - g_k^\star\|_W \le \tau$ with $\tau = O(\sigma)$. For any other active atom l, separation gives the feature-cost gap $M_{i,\Pi(l)} - M_{i,\Pi(k)} = \|f_i - g_l^\star\|_W^2 - \|f_i - g_k^\star\|_W^2 \ge \delta^2 - 2\delta\tau$. In each entropic layer the plan factors as $T_{ik} = u_i K_{ik} v_k$, so the relative mass point i sends to $\Pi(l)$ versus $\Pi(k)$ is

$$\frac{T_{i,\Pi(l)}}{T_{i,\Pi(k)}} = \frac{v_{\Pi(l)}}{v_{\Pi(k)}} \exp\Big(-\tfrac{C_{i,\Pi(l)} - C_{i,\Pi(k)}}{\varepsilon}\Big) \le \Xi \exp\Big(-\tfrac{(1-\alpha)(\delta^2 - 2\delta\tau) - O(\alpha)}{\varepsilon}\Big),$$

where Ξ bounds the dual-scaling ratio $v_{\Pi(l)}/v_{\Pi(k)}$ and the $O(\alpha)$ term is the structure contribution; both are bounded on the compact domain by the potential bounds of Lemma S2. For $\varepsilon \le c\,\delta^2/\log K$ this ratio is $\le K^{-2}$, so the total off-atom (leaked) mass of point i is $\le K \cdot K^{-2} = K^{-1} = O(\varepsilon)$. Hence each point is matched to $\Pi(k)$ up to $O(\varepsilon)$ leakage. The barycentric feature at $\Pi(k)$ is the leak-weighted average of within-atom features; sub-Gaussian concentration of the within-atom mean over $\ge \underline{m}$ points gives deviation $O(\sigma/\sqrt{\underline{m}})$ with probability $\ge 1 - Ke^{-c'\underline{m}\delta^2/\sigma^2}$, and the $O(\varepsilon)$ leak (times bounded $\|f\| \le R$) plus entropic blur give the $O(\varepsilon)$ bias. Masses follow from $T^\top \mathbf{1} = b$ and per-atom mass conservation up to the same $O(\varepsilon)$ leakage. □

Remark S5. The structure term ($\alpha > 0$) can only sharpen the assignment when intra-sample geometry is consistent with the template; the bound above is the conservative feature-only estimate.

Lemma S4 (Bounded dual scalings of the unrolled map)**.** *Fix $\varepsilon > 0$ and $S, J < \infty$. On the compact domain (costs $\le \bar{C}$, marginals $\ge \rho > 0$) every Sinkhorn scaling $u^{(s,j)}, v^{(s,j)}$ in Definition S1 is strictly positive and bounded: since $K^{(s)} = \exp(-C^{(s)}/\varepsilon)$ has entries in $[e^{-\bar{C}/\varepsilon}, 1]$, each update $u \leftarrow a \oslash (Kv)$, $v \leftarrow b \oslash (K^\top u)$ maps the positive orthant into a box bounded away from 0; a finite composition of $S(J+1)$ such maps takes values in an explicit interval $[m_\varepsilon, M_\varepsilon]$ with $m_\varepsilon \ge \rho\, e^{-(J+1)\bar{C}/\varepsilon}$. Hence the dual-scaling ratio of Lemma S3 satisfies $\Xi = \max_{k,l} v_k/v_l \le M_\varepsilon/m_\varepsilon < \infty$, an explicit constant, so the Ξ used in Lemma S3 is bounded, not assumed.*

Proof. $K^{(s)}$ entrywise $\ge e^{-\bar{C}/\varepsilon}$ gives $(Kv)_i \le K\|v\|_\infty$ and $(Kv)_i \ge e^{-\bar{C}/\varepsilon}\sum_k v_k$, so with $a_i, b_k \in [\rho, 1]$ one update keeps the scalings in $[\rho\, e^{-\bar{C}/\varepsilon}/K,\ \rho^{-1}e^{\bar{C}/\varepsilon}]$; iterating the fixed $S(J+1)$ updates (a finite composition of continuous positive maps on the compact domain) yields the stated $[m_\varepsilon, M_\varepsilon]$. □

S1.5. Identifiability and consistency

Theorem S1 (Population identifiability up to permutation)**.** *Under Assumptions S1–S5, let $(\Theta^\star, \beta^\star) \in \arg\min_{\Theta,\beta} \mathcal{R}(\Theta, \beta)$. Then there is a permutation Π of atoms such*

that the recovered dictionary equals the truth, $g^{\star}_{\Pi^{-1}(k)} = g^{\Theta^\star}_k$ *for active atoms, and the induced predictor equals the best reparametrization-invariant predictor of* Y*; i.e.* $h_{\Theta^\star,\beta^\star}$ *attains* $\inf_{g\in\mathcal{G}_{\mathrm{inv}}} \mathbb{E}\|Y - g(\mu)\|^2$*, where* $\mathcal{G}_{\mathrm{inv}}$ *is the class of predictors measurable with respect to the reparametrization-invariant* σ*-field.*

Proof. By Lemma S1, $\Phi(\cdot;\Theta)$ is measurable with respect to the invariant σ-field, so $h_{\Theta,\beta} \in \mathcal{G}_{\mathrm{inv}}$ for every (Θ,β); thus the achievable risk is bounded below by the invariant Bayes risk. Conversely, by Assumption S3 the regression function depends on μ only through (α,Γ), both invariant; by Lemma S3 a template matching the true features makes Φ an invertible (up to Π and $O(\varepsilon)$) encoding of (α,Γ), so a linear head β reproduces $\langle B^\star, \psi(\cdot)\rangle$ up to $O(\varepsilon)$. Hence the invariant Bayes risk is attained in the limit $\varepsilon \to 0$ and, for fixed ε, up to an explicit $O(\varepsilon)$ bias; Assumption S4 makes the minimizer's active-atom dictionary and head unique up to Π. □

Lemma S5 (Unconditional quadratic growth on the prediction block)**.** *Under Assumptions S1–S5, decompose the parameter as* $\Theta = (G, b, \beta; L, W)$ *into the feature/mass/head block* $\theta_{\mathrm{f}} = (G, b, \beta)$ *and the geometry block* $\theta_{\mathrm{g}} = (L, W)$*. At a population minimizer* $(\Theta^\star, \beta^\star)$ *the risk grows quadratically in the feature/mass/head block, up to permutation: there is* $c_{\mathrm{f}} > 0$ *with*

$$\mathcal{R}(\theta_{\mathrm{f}}, \theta^\star_{\mathrm{g}}) - \mathcal{R}(\theta^\star_{\mathrm{f}}, \theta^\star_{\mathrm{g}}) \geq c_{\mathrm{f}} \operatorname{dist}(\theta_{\mathrm{f}}, \Pi\cdot\theta^\star_{\mathrm{f}})^2,$$

i.e. $\nabla^2_{\theta_{\mathrm{f}}} \mathcal{R}(\Theta^\star, \beta^\star) \geq c_{\mathrm{f}} I$ *transverse to the permutation orbit. This holds with no extra hypothesis beyond those already used for atom recovery.*

Proof. On the saturated-assignment region guaranteed by Lemma S3 (in force whenever $\varepsilon \leq c\,\delta^2/\log K$ and $\Xi < \infty$, Lemma S4), each barycentric feature coordinate is, up to the $O(\varepsilon)$ controlled bias, the identity readout of the corresponding atom feature: $[\Phi]^{\mathrm{feat}}_{\Pi(k)} = g_k + O(\varepsilon)$ and $[\Phi]^{\mathrm{mass}}_{\Pi(k)} = b_k + O(\varepsilon)$, with Jacobian $\partial\Phi/\partial(G,b)$ equal to the identity plus an $O(\varepsilon)$ (and $O(e^{-c'\delta^2/\varepsilon})$ off-diagonal-leakage) perturbation, hence invertible and uniformly well-conditioned for $\varepsilon \leq c\,\delta^2/\log K$ (this is where δ-separation enters). Consequently $\mathcal{R}(\Theta,\beta) = \mathbb{E}\|Y -$

$\beta^\top \Phi(\mu;\Theta)\|^2$ is, in the (G, b, β) coordinates, a quadratic form in Φ composed with a well-conditioned diffeomorphism, whose Hessian factors as $J^\top \mathbb{E}[\Phi\Phi^\top]J$ plus head curvature, with J the invertible readout Jacobian. Assumption S4 ($\mathrm{Cov}(\psi) \succ 0$, $B^\star$ with no zero block) makes $\mathbb{E}[\Phi\Phi^\top] \succ 0$ on the response-relevant coordinates, so the composite Hessian is positive definite transverse to the permutation orbit, with $c_{\mathrm{f}} \geq \sigma_{\min}(J)^2 \lambda_{\min}(\mathbb{E}[\Phi\Phi^\top]) > 0$. The permutation orbit is the only flat direction because atoms are discrete and feature-δ-separated. □

Remark S6. Lemma S5 folds the prediction-relevant part of the former well-separated-minimizer hypothesis into the standing assumptions: feature/mass/head identifiability is now unconditional given the recovery conditions (δ-separation, $\varepsilon \leq c\,\delta^2/\log K$, bounded Ξ). What is not implied by feature-separation alone is curvature in the latent geometry block L, which enters $\mathcal{R}$ only through the structural quadratic $(C - D)^2$ and admits intrinsic isometry symmetries; this single residual is exactly Assumption S6.

Theorem S2 (Excess-risk consistency and rate)**.** *Let* $(\widehat{\Theta}_N, \widehat{\beta}_N) \in \arg\min_{\Theta\in\mathbb{T}, \|\beta\|\leq B} \frac{1}{N}\sum_n \|Y_n - h_{\Theta,\beta}(\mu_n)\|^2$. *Under Assumptions S1–S5 (no minimizer-separation needed),*

$$\mathcal{R}(\widehat{\Theta}_N, \widehat{\beta}_N) - \inf_{\Theta,\beta} \mathcal{R}(\Theta,\beta) = O_p\Big(\sqrt{\tfrac{KC\log(\Lambda_\varepsilon)}{N}}\Big).$$

Proof. By Lemma S2 the loss class $\{(\mu, Y) \mapsto \|Y - h_{\Theta,\beta}(\mu)\|^2\}$ is Lipschitz in the finite-dimensional (Θ, β) with constant $O(\Lambda_\varepsilon)$ and uniformly bounded (Assumption S5); a Lipschitz-parametric class over a compact set is Glivenko–Cantelli and Donsker, giving $\Delta_N := \sup_{\Theta,\beta} |\widehat{\mathcal{R}} - \mathcal{R}| = O_p(\sqrt{KC\log\Lambda_\varepsilon/N})$. Writing $h^\star$ for a population minimizer, ERM optimality and two applications of Δ_N give $\mathcal{R}(\widehat{h}) \leq \widehat{\mathcal{R}}(\widehat{h}) + \Delta_N \leq \widehat{\mathcal{R}}(h^\star) + \Delta_N \leq \mathcal{R}(h^\star) + 2\Delta_N$, i.e. excess risk $\leq 2\Delta_N$. No identifiability or separation is used, only uniform convergence. □

Proposition S2 (Parameter recovery)**.** *Under Assumptions S1–S5, the population risk satisfies the well-separation (quadratic growth) condition* $\mathcal{R}(\Theta,\beta) - \mathcal{R}(\Theta^\star,\beta^\star) \geq c\,\mathrm{dist}((\Theta,\beta), \Pi\cdot(\Theta^\star,\beta^\star))^2$, $c > 0$, *on the feature/mass/head block* unconditionally

(Lemma S5); on the latent geometry block it holds under the standard second-order identifiability Assumption S6, with $c = \min(c_\Gamma, c_L)$ *after a Taylor expansion (the cross block of the Hessian only improves the constant by the Schur-complement bound). Whenever it holds,* $(\widehat{\Theta}_N, \widehat{\beta}_N) \to (\Theta^\star, \beta^\star)$ *in probability up to the permutation* Π*, by the argmin theorem applied to the uniform convergence above. Thus parameter recovery is unconditional for the prediction-relevant parameters (atom features, masses, head) and requires only the mild, checkable Assumption S6 for the latent-geometry coordinates; the excess-risk guarantee of Theorem S2 needs neither.*

Corollary S1 (Bias–smoothness trade-off)**.** *Choosing* $\varepsilon = \varepsilon_N \to 0$ *with* $\varepsilon_N^{-1} = o(N^{1/2}/\sqrt{KC})$ *removes the* $O(\varepsilon)$ *identification bias of Lemma S3 while preserving consistency; the estimator then targets the unregularized invariant Bayes predictor.*

The prediction-relevant guarantees (Lemmas S1–S4, Theorems S1–S2) are unconditional, and Lemma S5 further makes exact recovery of the prediction-relevant parameters (atom features, masses, head) unconditional under the standing recovery conditions. The single remaining conditional element is the second-order identifiability Assumption S6 on the latent-geometry block, used in Proposition S2 only for exact recovery of that block and never for prediction consistency. Identifiability is up to atom permutation, standard for dictionary and factor models, with no rotational ambiguity because the atoms are discrete and feature-separated. The unaligned-CP estimand of Tang, Kolda and Zhang [1] is unsupervised (reconstructing a single tensor via an RKHS mode) and carries no response; our estimand is the supervised, response-identifying object of Theorem S1, recovered through transport, and this supervised estimand is what separates the two settings. A minimax lower bound for invariant supervised recovery, behavior under misspecified K, and a low-rank multilinear head preserving Theorem S2 remain open.

References

[1] R. Tang, T. G. Kolda, A. R. Zhang, Tensor decomposition with unaligned observations, SIAM Journal on Matrix Analysis and Applications 47 (1) (2026) 107–131, arXiv:2410.14046. doi:10.1137/24M1692836.