Multivariate Regression with Gross Errors on Manifold-valued Data – Supplemental Material

Xiaowei Zhang, *Member, IEEE,* Xudong Shi, Yu Sun, *Member, IEEE,* Li Cheng, *Senior Member, IEEE*

CONTENTS

1 OVERVIEW

In this supplementary material, we present supplemental material to the main text. The rest of this material is organized as follows. In Section 2, we provide convergence analysis of the proposed algorithm PALMR. For this purpose, we first present some necessary results for smooth functions and K-L functions on Hadamard manifolds, then derive the detailed proof of the convergence results in Theorem 1 of the main text. In Section 3, we show the derivation of partial derivatives of the loss function $\partial_{\boldsymbol{p}} E$, $\partial_{\boldsymbol{v}_j} E$, and $\partial_{\boldsymbol{g}_i} E$. These partial derivatives are crucial components of Algorithm 2 in the main text. In Section 4, we provide several operations on the manifold of symmetric positive definite matrices, including the Riemannian metric, geodesic distance, exponential map, inverse exponential map, and parallel transport. All of these operations have been implemented in the codes availabel on our project websit. In Section 5, we show more experimental results on the real DTI data.

2 CONVERGENCE ANALYSIS OF PALMR

In this section we provide convergence analysis of PALMR (Algorithm 1 in the main paper), including preliminary results about nonsmooth analysis and Kurdyka–Łojasiewicz (K-L) property on Hadamard manifolds in Subsection 2.1 and detailed proof of Theorem 1 in Subsection 2.2.

2.1 Preliminary Results

First, we provide some results for affine functions on Hadamard manifold $\mathcal{M}$.

Lemma 1 ([1]). *Let $\boldsymbol{x} \in \mathcal{M}$ and $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$ be given, and define function $f : \mathcal{M} \to \mathbb{R}$ as*

$$f(\boldsymbol{y}) := \left\langle \boldsymbol{v},\ Exp_{\boldsymbol{x}}^{-1}\boldsymbol{y} \right\rangle,\ \forall \boldsymbol{y} \in \mathcal{M}.$$

Then, $grad f(\boldsymbol{y}) = P_{\gamma(0)\gamma(1)}(\boldsymbol{v})$, where $\gamma : [0,1] \to \mathcal{M}$ is the geodesic curve such that $\gamma(0) = \boldsymbol{x}$ and $\gamma(1) = \boldsymbol{y}$, and $P_{\gamma(0)\gamma(1)}$ denotes the parallel transport along γ.

For smooth functions with Lipschitz gradient on Hadamard manifolds, we have the following descent lemma which resembles the one in the Euclidean spaces, see Proposition A.24 of [2].

Lemma 2. *Let $h : \mathcal{M} \to \mathbb{R}$ be a continuously differentiable function that has L-Lipschitz gradient, then*

$$\boxed{h(\boldsymbol{y}) \leq h(\boldsymbol{x}) + \left\langle \partial h(\boldsymbol{x}),\ Exp_{\boldsymbol{x}}^{-1}\boldsymbol{y} \right\rangle + \frac{L}{2} d^2(\boldsymbol{x}, \boldsymbol{y}),\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}.}$$

Proof. Let $\gamma : [0, 1] \to \mathcal{M}$ be the unique geodesic joining $\boldsymbol{x}$ and $\boldsymbol{y}$ such that $\gamma(0) = \boldsymbol{x}$, $\gamma(1) = \boldsymbol{y}$ and $\gamma'(0) = \mathrm{Exp}_{\boldsymbol{x}}^{-1}\boldsymbol{y}$. It is easy to show that

$$d(\boldsymbol{x}, \gamma(t)) = t d(\boldsymbol{x}, \boldsymbol{y}),\ \forall t \in [0, 1].$$

Let $\tilde{h} = h \circ \gamma : [0, 1] \to \mathbb{R}$, then the chain rule implies that

$$\tilde{h}'(t) = \left\langle \partial h(\gamma(t)),\ P_{\gamma(0)\gamma(t)}(\gamma'(0)) \right\rangle_{\gamma(t)}.$$

Applying the fundamental theorem of calculus to $\tilde{h}$, we get

$$\begin{aligned}
h(\boldsymbol{y}) - h(\boldsymbol{x}) =& \tilde{h}(1) - \tilde{h}(0) = \int_0^1 \tilde{h}'(t) dt \\
=& \int_0^1 \left\langle \partial h(\gamma(t)),\ P_{\gamma(0)\gamma(t)}(\gamma'(0)) \right\rangle_{\gamma(t)} dt \\
\overset{①}{=}& \left\langle \partial h(\boldsymbol{x}),\ \gamma'(0) \right\rangle_{\boldsymbol{x}} + \int_0^1 \left\langle \partial h(\gamma(t)) - P_{\gamma(0)\gamma(t)}(\partial h(\boldsymbol{x})),\ P_{\gamma(0)\gamma(t)}(\gamma'(0)) \right\rangle_{\gamma(t)} dt \\
\overset{②}{\leq}& \left\langle \partial h(\boldsymbol{x}),\ \gamma'(0) \right\rangle_{\boldsymbol{x}} + \int_0^1 L d^2(\boldsymbol{x}, \boldsymbol{y}) t dt \\
=& \left\langle \partial h(x),\ \mathrm{Exp}_{\boldsymbol{x}}^{-1}\boldsymbol{y} \right\rangle + \frac{L}{2} d^2(\boldsymbol{x}, \boldsymbol{y}),
\end{aligned}$$

where to get ① we used $\left\langle \partial h(\boldsymbol{x}),\ \gamma'(0) \right\rangle_{\boldsymbol{x}} = \left\langle P_{\gamma(0)\gamma(t)}(\partial h(\boldsymbol{x})),\ P_{\gamma(0)\gamma(t)}(\gamma'(0)) \right\rangle_{\gamma(t)}$ since the inner product does not change under the parallel transport along geodesics, and to get ② we used the Cauchy-Schwartz inequality and the fact that h has L-Lipschitz gradient. □

With the above descent lemma in hand, we can prove that sufficient decrease of the objective function value is guaranteed after a proximal step like (10) or (11) in the main text.

Lemma 3. *Let $h : \mathcal{M} \to \mathbb{R}$ be a continuously differentiable function that has L-Lipschitz gradient and let $\sigma : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ be a PLS function with $\inf_{\boldsymbol{x} \in \mathcal{M}} \sigma(\boldsymbol{x}) > -\infty$. For any fixed $\alpha \in \mathbb{R}$ and $\boldsymbol{y} \in$ dom σ, let*

$$\boldsymbol{y}^+ \in arg \min_{\boldsymbol{x} \in \mathcal{M}} \sigma(\boldsymbol{x}) + \left\langle Exp_{\boldsymbol{y}}^{-1}\boldsymbol{x},\ \partial h(\boldsymbol{y}) \right\rangle + \frac{\alpha}{2} d^2(\boldsymbol{y}, \boldsymbol{x}),$$

we have

$$\boxed{\sigma(\boldsymbol{y}^+) + h(\boldsymbol{y}^+) \leq \sigma(\boldsymbol{y}) + h(\boldsymbol{y}) - \frac{\alpha - L}{2} d^2(\boldsymbol{y}^+, \boldsymbol{y}).}$$

Proof. Since $\boldsymbol{y}^+$ is a minimizer, let $\boldsymbol{x} = \boldsymbol{y}$ in the objective function, we have

$$\sigma(\boldsymbol{y}^+) + \left\langle \mathrm{Exp}_{\boldsymbol{y}}^{-1}\boldsymbol{y}^+,\ \partial h(\boldsymbol{y}) \right\rangle + \frac{\alpha}{2} d^2(\boldsymbol{y}, \boldsymbol{y}^+) \leq \sigma(\boldsymbol{y}).$$

Moreover, since h has L-Lipschitz gradient, it follows from Lemma 2 that

$$h(\boldsymbol{y}^+) \leq h(\boldsymbol{y}) + \left\langle \mathrm{Exp}_{\boldsymbol{y}}^{-1}\boldsymbol{y}^+,\ \partial h(\boldsymbol{y}) \right\rangle + \frac{L}{2} d^2(\boldsymbol{y}, \boldsymbol{y}^+).$$

Adding the above two inequalities yields the inequality we require. □

Regarding the subdifferential of the objective function Ψ defined in (9) of the main text, we have the following result whose derivation is similar to that in [3].

Proposition 4 ([4]). *Let Ψ be as in (9), then for all $(\boldsymbol{x}, \boldsymbol{y}) \in$ dom $\Psi =$ dom $f +$ dom g, we have*

$$\boxed{\partial \Psi(\boldsymbol{x}, \boldsymbol{y}) = \{\partial f(\boldsymbol{x}) + \partial_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{y})\} \times \{\partial g(\boldsymbol{y}) + \partial_{\boldsymbol{y}} h(\boldsymbol{x}, \boldsymbol{y})\} = \partial_{\boldsymbol{x}} \Psi(\boldsymbol{x}, \boldsymbol{y}) \times \partial_{\boldsymbol{y}} \Psi(\boldsymbol{x}, \boldsymbol{y}).}$$

All the above results will be used to prove Theorem 1 in the next subsection. The next result is related to the K-L property. Although not needed in our proof of Theorem 1, it is helpful to understand the K-L property. In particular, we show that if $\bar{\boldsymbol{x}} \in$ dom σ is not a critical point of σ (i.e. $\bar{\boldsymbol{x}} \notin$ crit σ) then K-L inequality holds at $\bar{\boldsymbol{x}}$.

Proposition 5 ([4]). *Let $\sigma : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ be a PLS function and $\bar{\boldsymbol{x}} \in$ dom $\partial \sigma$ such that $0 \notin \partial \sigma(\bar{\boldsymbol{x}})$. Then the K-L inequality holds at $\bar{\boldsymbol{x}}$.*

Proof. For any $\delta > 0$, take $\phi(t) = t/\delta$, $U = B(\bar{\boldsymbol{x}}, \delta/2)$, $\eta = \delta/2$, then for each $\boldsymbol{x} \in \text{dom}\ \sigma$, we have

$$\phi'(\sigma(\boldsymbol{x}) - \sigma(\bar{\boldsymbol{x}}))\text{dist}(0, \partial\sigma(\boldsymbol{x})) = \text{dist}(0, \partial\sigma(\boldsymbol{x}))/\delta. \tag{1}$$

Notice that $\boldsymbol{x} \in U \cap [\sigma(\bar{\boldsymbol{x}}) < \sigma < \sigma(\bar{\boldsymbol{x}}) + \eta]$ implies

$$d(\boldsymbol{x}, \bar{\boldsymbol{x}}) + |\sigma(\boldsymbol{x}) - \sigma(\bar{\boldsymbol{x}})| < \delta. \tag{2}$$

We further claim that for each $\boldsymbol{x}$ satisfying (2), it holds

$$\text{dist}(0, \partial\sigma(\boldsymbol{x})) \geq \delta. \tag{3}$$

Otherwise, there exist sequences $\{(\boldsymbol{x}^k, \boldsymbol{u}^k) : \boldsymbol{u}^k \in \partial\sigma(\boldsymbol{x}^k)\}$ and $\{\delta_k\} \subset \mathbb{R}_{++}$ such that $\delta_k \to 0$ as $k \to \infty$ and

$$d(\boldsymbol{x}^k, \bar{\boldsymbol{x}}) + |\sigma(\boldsymbol{x}^k) - \sigma(\bar{\boldsymbol{x}})| < \delta_k, \quad \|\boldsymbol{u}^k\| < \delta_k,$$

which implies

$$\boldsymbol{x}^k \to \bar{\boldsymbol{x}},\ \sigma(\boldsymbol{x}^k) \to \sigma(\bar{\boldsymbol{x}}),\ \boldsymbol{u}^k \in \partial\sigma(\boldsymbol{x}^k),\ \text{and}\ \|\boldsymbol{u}^k\| \to 0.$$

Thus, $0 \in \partial\sigma(\bar{\boldsymbol{x}})$, which is a contradiction with the assumption.

Therefore, combining (1), (2) and (3), we get

$$\phi'(\sigma(\boldsymbol{x}) - \sigma(\bar{\boldsymbol{x}}))\text{dist}(0, \partial\sigma(\boldsymbol{x})) \geq 1.$$

So, the K-L inequality holds at $\bar{\boldsymbol{x}}$. □

2.2 Proof of Theorem 1

Following the idea of [5], [6], we outline the proof in three steps: (1) sufficient decrease of objective function value; (2) a subdifferential lower bound for the iterates gap; (3) using the K-L property.

We first show the sufficient decrease of objective function value under Assumption 1. For simplicity of notations, we use $\Psi^k := \Psi(\boldsymbol{x}^k, \boldsymbol{y}^k)$ for $k \geq 0$ in the sequel.

Lemma 6. *Suppose Assumption 1 holds. Let $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ be the sequence generated by PALMR. The following assertions hold.*

(i) *The sequence $\{\Psi^k\}_{k\in\mathbb{N}}$ is nonincreasing and satisfies*

$$\frac{\tau}{2}(d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k)) \leq \Psi^k - \Psi^{k+1}, \tag{4}$$

where $\tau := \min\{(\mu_1 - 1)\lambda_1^-, (\mu_2 - 1)\lambda_2^-\} > 0$.

(ii) *We have*

$$\sum_{k=0}^{\infty}\left(d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k)\right) < \infty, \tag{5}$$

which implies

$$\lim_{k\to\infty} d_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) = \lim_{k\to\infty} d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k) = 0. \tag{6}$$

Proof. (i) Applying Lemma 3 to subproblems (10) and (11), we get

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) + h(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k) \leq & f(\boldsymbol{x}^k) + h(\boldsymbol{x}^k, \boldsymbol{y}^k) - \frac{c_k - L_1(\boldsymbol{y}^k)}{2} d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k), \\ g(\boldsymbol{y}^{k+1}) + h(\boldsymbol{x}^{k+1}, \boldsymbol{y}^{k+1}) \leq & g(\boldsymbol{y}^k) + h(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k) - \frac{d_k - L_2(\boldsymbol{x}^{k+1})}{2} d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k). \end{aligned}$$

Adding the above two inequalities, we obtain for $k \geq 0$ that

$$\begin{aligned} \Psi^k - \Psi^{k+1} \overset{①}{\geq} & \frac{(\mu_1 - 1)L_1(\boldsymbol{y}^k)}{2} d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + \frac{(\mu_2 - 1)L_2(\boldsymbol{x}^{k+1})}{2} d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k) \\ \overset{②}{\geq} & \frac{(\mu_1 - 1)\lambda_1^-}{2} d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + \frac{(\mu_2 - 1)\lambda_2^-}{2} d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k) \\ \geq & \frac{\tau}{2}(d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k)), \end{aligned} \tag{7}$$

where in ① we used $c_k = \mu_1 L_1(\boldsymbol{y}^k)$, $d_k = \mu_2 L_2(\boldsymbol{x}^{k+1})$ and $\mu_1, \mu_2 > 1$, and in ② we used the assumption $\inf\{L_1(\boldsymbol{y}^k) : k \in \mathbb{N}\} \geq \lambda_1^-$ and $\inf\{L_2(\boldsymbol{x}^k) : k \in \mathbb{N}\} \geq \lambda_2^-$.

(ii) Inequality (4) shows that $\{\Psi^k\}_{k\in\mathbb{N}}$ is nonincreasing sequence, and we know from Assumption 1(i) that $\inf_{\boldsymbol{x},\boldsymbol{y}} \Psi(\boldsymbol{x}, \boldsymbol{y}) > -\infty$, it follows that $\{\Psi^k\}_{k\in\mathbb{N}}$ converges. Denote the limit by Ψ^*, we have

$$\boxed{\lim_{k\to\infty} \Psi^k = \Psi^*.} \tag{8}$$

Let K be a positive integer and sum (7) from $k=0$ to $k=K$, we get

$$\Psi^0 - \Psi^{K+1} \geq \frac{\tau}{2}\sum_{k=0}^{K} d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1},\boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k).$$

Taking $K\to\infty$ in the above inequality leads to

$$\sum_{k=0}^{\infty} d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1},\boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k) \leq \frac{2(\Psi^0-\Psi^*)}{\tau} < +\infty,$$

which completes the proof. □

Next, we derive a subdifferential lower bound for the gap between two consecutive iterates, and study some properties of the limiting points of the sequence of iterates under the assumption of boundedness.

Lemma 7. *Suppose Assumption 1 holds. Let $\{(\boldsymbol{x}^k,\boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ be the sequence generated by PALMR which is assumed to be bounded. For each $k\geq 0$, define*

$$\begin{aligned} A_x^{k+1} =& c_k Exp^{-1}_{\boldsymbol{x}^{k+1}}\boldsymbol{x}^k + \partial_{\boldsymbol{x}} h(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1}) - P_{\gamma_x^k(0)\gamma_x^k(1)}\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^k)\\ A_y^{k+1} =& d_k Exp^{-1}_{\boldsymbol{y}^{k+1}}\boldsymbol{y}^k + \partial_{\boldsymbol{y}} h(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1}) - P_{\gamma_y^k(0)\gamma_y^k(1)}\partial_{\boldsymbol{y}}, h(\boldsymbol{x}^{k+1},\boldsymbol{y}^k). \end{aligned}$$

where $\gamma_{\boldsymbol{x}}^k:[0,1]\to\mathcal{M}_1$ is the geodesic joining $\boldsymbol{x}^k$ and $\boldsymbol{x}^{k+1}$ such that $\gamma_{\boldsymbol{x}}^k(0)=\boldsymbol{x}^k$ and $\gamma_{\boldsymbol{x}}^k(1)=\boldsymbol{x}^{k+1}$, and $\gamma_{\boldsymbol{y}}^k$ is similarly defined. Then

$$(A_x^{k+1},A_y^{k+1}) \in \partial\Psi(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1}) \subseteq T_{\boldsymbol{x}^{k+1}}\mathcal{M}_1 \times T_{\boldsymbol{y}^{k+1}}\mathcal{M}_2. \tag{9}$$

Moreover, we have

$$\boxed{\|A_x^{k+1}\|_{\boldsymbol{x}^{k+1}} \leq (1+\mu_1)\lambda_1^+ d_{\mathcal{M}_1}(\boldsymbol{x}^{k+1},\boldsymbol{x}^k) + L d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k), \quad \|A_y^{k+1}\|_{\boldsymbol{y}^{k+1}} \leq (1+\mu_2)\lambda_2^+ d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k).} \tag{10}$$

Proof. Since $\boldsymbol{x}^{k+1}$ is a minimizer of subproblem (10), by the Fermat's rule, we have

$$\begin{aligned} 0 \in& \partial\left(f(\boldsymbol{x}) + \left\langle \mathrm{Exp}_{\boldsymbol{x}^k}^{-1}\boldsymbol{x},\ \partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^k)\right\rangle + \frac{c_k}{2} d^2_{\mathcal{M}_1}(\boldsymbol{x}^k,\boldsymbol{x})\right)\Big|_{\boldsymbol{x}=\boldsymbol{x}^{k+1}}\\ =& \partial f(\boldsymbol{x}^{k+1}) + P_{\gamma_x^k(0)\gamma_x^k(1)}\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^k) - c_k \mathrm{Exp}^{-1}_{\boldsymbol{x}^{k+1}}\boldsymbol{x}^k, \end{aligned} \tag{11}$$

where we used Lemma 1, Proposition 4 and the fact [4] that $\partial_{\boldsymbol{x}} d^2_{\mathcal{M}}(\boldsymbol{x}',\boldsymbol{x}) = -2\mathrm{Exp}_{\boldsymbol{x}}^{-1}\boldsymbol{x}'$ for Hadamard manifold $\mathcal{M}$ to get the equality. Similarly, we also have

$$0 \in \partial g(\boldsymbol{y}^{k+1}) + P_{\gamma_y^k(0)\gamma_y^k(1)}\partial_{\boldsymbol{y}} h(\boldsymbol{x}^{k+1},\boldsymbol{y}^k) - c_k \mathrm{Exp}^{-1}_{\boldsymbol{y}^{k+1}}\boldsymbol{y}^k. \tag{12}$$

Combining (11) and (12) , we get

$$A_x^{k+1} \in \partial f(\boldsymbol{x}^{k+1}) + \partial_{\boldsymbol{x}} h(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1}), \quad A_y^{k+1} \in \partial g(\boldsymbol{y}^{k+1}) + \partial_{\boldsymbol{y}} h(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1}),$$

which, together with Proposition 4, results in the assertion in (9).

Now, we estimate the norms of A_x^{k+1} and A_y^{k+1} as follows.

$$\begin{aligned} \|A_x^{k+1}\|_{\boldsymbol{x}^{k+1}} \leq& c_k d(\boldsymbol{x}^{k+1},\boldsymbol{x}^k) + \|\partial_{\boldsymbol{x}} h(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1}) - P_{\gamma_x^k(0)\gamma_x^k(1)}\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^k)\|_{\boldsymbol{x}^{k+1}}\\ \overset{①}{\leq}& c_k d(\boldsymbol{x}^{k+1},\boldsymbol{x}^k) + \|\partial_{\boldsymbol{x}} h(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1}) - P_{\gamma_x^k(0)\gamma_x^k(1)}\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^{k+1})\|_{\boldsymbol{x}^{k+1}} + \|\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^{k+1}) - \partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^k)\|_{\boldsymbol{x}^k}\\ \overset{②}{\leq}& (c_k + L_1(\boldsymbol{y}^{k+1}))d_{\mathcal{M}_1}(\boldsymbol{x}^{k+1},\boldsymbol{x}^k) + L d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k)\\ \leq& (1+\mu_1)\lambda_1^+ d_{\mathcal{M}_1}(\boldsymbol{x}^{k+1},\boldsymbol{x}^k) + L d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k), \end{aligned}$$

where the first two inequalities are obtained from triangular inequality and in ① we used the fact $\|P_{\gamma_x^k(0)\gamma_x^k(1)}\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^{k+1}) - P_{\gamma_x^k(0)\gamma_x^k(1)}\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^k)\|_{\boldsymbol{x}^{k+1}} = \|\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^{k+1}) - \partial_{\boldsymbol{x}} h(\boldsymbol{x}^k,\boldsymbol{y}^k)\|_{\boldsymbol{x}^k}$, and in ② we used the Lipschitz continuity in Assumption 1 (ii)-(iii) since we assume that $\{(\boldsymbol{x}^k,\boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ is bounded.

Similarly, from the Lipschitz continuity of $\partial_{\boldsymbol{y}} h$, we get

$$\begin{aligned} \|A_y^{k+1}\|_{\boldsymbol{y}^{k+1}} \leq& d_k d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k) + \|\partial_{\boldsymbol{y}} h(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1}) - P_{\gamma_y^k(0)\gamma_y^k(1)}\partial_{\boldsymbol{y}} h(\boldsymbol{x}^{k+1},\boldsymbol{y}^k)\|_{\boldsymbol{y}^{k+1}}\\ \leq& d_k d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k) + L_2(\boldsymbol{x}^{k+1}) d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k)\\ \leq& (1+\mu_2)\lambda_2^+ d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1},\boldsymbol{y}^k). \end{aligned}$$

Therefore, we obtain the assertions in (10).

□

It follows from Lemma 6 (ii) and Lemma 7 that

$$\lim_{k\to\infty} \|A_x^k\|_{\boldsymbol{x}^k} = \lim_{k\to\infty} \|A_y^k\|_{\boldsymbol{y}^k} = 0. \tag{13}$$

In addition, for $k \geq 1$ we have

$$\begin{aligned}
\operatorname{dist}(0, \partial\Psi(\boldsymbol{x}^k, \boldsymbol{y}^k)) \leq & \|(A_x^k, A_y^k)\|_{(\boldsymbol{x}^k, \boldsymbol{y}^k)} \leq \beta_0(\|A_x^k\|_{\boldsymbol{x}^k}^2 + \|A_y^k\|_{\boldsymbol{y}^k}^2)^{1/2} \\
\leq & \beta_0[2((1+\mu_1)\lambda_1^+)^2 d_{\mathcal{M}_1}^2(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) + (2L^2 + ((1+\mu_2)\lambda_2^+)^2) d_{\mathcal{M}_2}^2(\boldsymbol{y}^k, \boldsymbol{y}^{k-1})]^{1/2} \\
\leq & \beta(d_{\mathcal{M}_1}^2(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) + d_{\mathcal{M}_2}^2(\boldsymbol{y}^k, \boldsymbol{y}^{k-1}))^{1/2} \to 0, \text{ as } k \to \infty,
\end{aligned} \tag{14}$$

where β_0 is a universal constant [1] and $\beta := \beta_0 \max\left\{\sqrt{2}(1+\mu_1)\lambda_1^+, \sqrt{(2L^2 + ((1+\mu_2)\lambda_2^+)^2)}\right\}$.

Under the assumption that $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ is bounded, there exists at least one limit point. We denote the set of all limiting points of $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ as $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$, that is

$$\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0) := \{(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathcal{M}_1 \times \mathcal{M}_2 : \exists \text{ subsequence } (\boldsymbol{x}^{k_j}, \boldsymbol{y}^{k_j}) \to (\boldsymbol{x}^*, \boldsymbol{y}^*) \text{ as } j \to \infty\}.$$

Some properties of $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$ are presented in Lemma 8.

Lemma 8. *Suppose Assumption 1 holds. Let $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ be the sequence generated by PALMR which is assumed to be bounded. Then the following assertions hold.*

(i) $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0) \subseteq$ *crit* Ψ *is a nonempty, compact and connected set and*

$$\lim_{k\to\infty} \operatorname{dist}((\boldsymbol{x}^k, \boldsymbol{y}^k), \Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)) = 0. \tag{15}$$

(ii) Ψ *is finite and constant on* $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$*, which equals to* Ψ^*.

Proof. (i) Since the sequence $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ is bounded, there exists at least one limit point, so $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$ is a nonempty bounded closed set, which, together with the Hopf-Rinow's Theorem [7], implies that it is compact. It is easy to show

$$\lim_{k\to\infty} \operatorname{dist}((\boldsymbol{x}^k, \boldsymbol{y}^k), \Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)) = 0$$

by the definition of limit point. The connectedness can be proved by contradiction. Suppose $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$ is not connected, which means there exist two nonempty and closed disjoint subsets C_0 and C_1 of $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$ such that $C_0 \cup C_1 = \Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$. According to the smooth Urysohn Lemma [8], there exists a smooth function $\zeta : \mathcal{M}_1 \times \mathcal{M}_2 \to [0, 1]$ such that $C_0 = \zeta^{-1}(0)$ and $C_1 = \zeta^{-1}(1)$. Setting $U_0 := [\zeta < 1/4]$ and $U_1 := [\zeta < 3/4]$, we obtain two open neighborhoods of C_0 and C_1, respectively. Since $\lim_{k\to\infty} \operatorname{dist}((\boldsymbol{x}^k, \boldsymbol{y}^k), \Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)) = 0$, there exists some integer k_0 such that $(\boldsymbol{x}^k, \boldsymbol{y}^k) \in U_0 \cup U_1$ for all $k \geq k_0$. Let $\delta_k = \zeta(\boldsymbol{x}^k, \boldsymbol{y}^k)$, the sequence $\{\delta_k\}_{k\in\mathbb{N}}$ satisfies

(1) $\delta_k \notin [1/4, 3/4]$ for all $k \geq k_0$,
(2) there exists infinitely many k such that $\delta_k < 1/4$,
(3) there exists infinitely many k such that $\delta_k > 3/4$,
(4) ζ is uniformly continuous on compact sets which, together with the boundedness of $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}$, $d_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) \to 0$ and $d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k) \to 0$, implies that $|\delta^{k+1} - \delta_k| \to 0$ as $k \to \infty$,

simultaneously. However, there exists no sequence satisfying all the above four requirements. Therefore, $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$ is connected.

Now, we show that every point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ in $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$ is a critical point of Ψ. Suppose $(\boldsymbol{x}^{k_j}, \boldsymbol{y}^{k_j}) \to (\boldsymbol{x}^*, \boldsymbol{y}^*)$ as $j \to \infty$, since both f and g are PLS functions, we have

$$\liminf_{j\to\infty} f(\boldsymbol{x}^{k_j}) \geq f(\boldsymbol{x}^*) \quad \text{and} \quad \liminf_{j\to\infty} g(\boldsymbol{y}^{k_j}) \geq g(\boldsymbol{y}^*). \tag{16}$$

From subproblem (10), we have

$$\begin{aligned}
& f(\boldsymbol{x}^{k_j}) + \left\langle \operatorname{Exp}_{\boldsymbol{x}^{k_j-1}}^{-1} \boldsymbol{x}^{k_j},\ \partial_{\boldsymbol{x}} h(\boldsymbol{x}^{k_j-1}, \boldsymbol{y}^{k_j-1}) \right\rangle + \frac{c_{k_j-1}}{2} d_{\mathcal{M}_1}^2(\boldsymbol{x}^{k_j}, \boldsymbol{x}^{k_j-1}) \\
& \leq f(\boldsymbol{x}^*) + \left\langle \operatorname{Exp}_{\boldsymbol{x}^{k_j-1}}^{-1} \boldsymbol{x}^*,\ \partial_{\boldsymbol{x}} h(\boldsymbol{x}^{k_j-1}, \boldsymbol{y}^{k_j-1}) \right\rangle + \frac{c_{k_j-1}}{2} d_{\mathcal{M}_1}^2(\boldsymbol{x}^*, \boldsymbol{x}^{k_j-1}),
\end{aligned}$$

which yields

$$\begin{aligned}
f(\boldsymbol{x}^{k_j}) \leq & f(\boldsymbol{x}^*) + \left\langle \operatorname{Exp}_{\boldsymbol{x}^{k_j-1}}^{-1} \boldsymbol{x}^*,\ \partial_{\boldsymbol{x}} h(\boldsymbol{x}^{k_j-1}, \boldsymbol{y}^{k_j-1}) \right\rangle - \left\langle \operatorname{Exp}_{\boldsymbol{x}^{k_j-1}}^{-1} \boldsymbol{x}^{k_j},\ \partial_{\boldsymbol{x}} h(\boldsymbol{x}^{k_j-1}, \boldsymbol{y}^{k_j-1}) \right\rangle + \frac{c_{k_j-1}}{2} d_{\mathcal{M}_1}^2(\boldsymbol{x}^*, \boldsymbol{x}^{k_j-1}) \\
\leq & f(\boldsymbol{x}^*) + (d_{\mathcal{M}_1}(\boldsymbol{x}^*, \boldsymbol{x}^{k_j-1}) + d_{\mathcal{M}_1}(\boldsymbol{x}^{k_j}, \boldsymbol{x}^{k_j-1})) \|\partial_{\boldsymbol{x}} h(\boldsymbol{x}^{k_j-1}, \boldsymbol{y}^{k_j-1})\|_{\boldsymbol{x}^{k_j-1}} + \frac{c_{k_j-1}}{2} d_{\mathcal{M}_1}^2(\boldsymbol{x}^*, \boldsymbol{x}^{k_j-1}),
\end{aligned}$$

1. For any $(\boldsymbol{u}, \boldsymbol{v}) \in T_{\boldsymbol{x}}\mathcal{M}_1 \times T_{\boldsymbol{y}}\mathcal{M}_2$, the function $(\boldsymbol{u}, \boldsymbol{v}) \to (\|\boldsymbol{u}\|_{\boldsymbol{x}}^2 + \|\boldsymbol{v}\|_{\boldsymbol{y}}^2)^{-1/2}$ defines a norm on the linear space $T_{\boldsymbol{x}}\mathcal{M}_1 \times T_{\boldsymbol{y}}\mathcal{M}_2$. On the other hand, $\|(\boldsymbol{u}, \boldsymbol{v})\|_{(\boldsymbol{x},\boldsymbol{y})}$ also defines a norm. Due to the equivalence of norms, there exists a universal constant β_0 such that $\|(\boldsymbol{u}, \boldsymbol{v})\|_{(\boldsymbol{x},\boldsymbol{y})} \leq \beta_0(\|\boldsymbol{u}\|_{\boldsymbol{x}}^2 + \|\boldsymbol{v}\|_{\boldsymbol{y}}^2)^{-1/2}$

where we used the Cauchy-Schwarz inequality in the last inequality. From Lemma 6 (ii), we know $\lim_{j\to\infty} d_{\mathcal{M}_1}(\boldsymbol{x}^{k_j}, \boldsymbol{x}^{k_j-1}) = 0$, which implies $\lim_{j\to\infty} \boldsymbol{x}^{k_j-1} = \boldsymbol{x}^*$. Note that c_{k_j-1} is bounded and $\partial_{\boldsymbol{x}} h$ is continuous and thus bounded. Taking $\limsup_{j\to\infty}$ on both sides of the above inequality yields $\limsup_{j\to\infty} f(\boldsymbol{x}^{k_j}) \le f(\boldsymbol{x}^*)$. By a similar argument we obtain $\limsup_{j\to\infty} g(\boldsymbol{y}^{k_j}) \le g(\boldsymbol{y}^*)$. Thus, in view of (16), we have

$$\lim_{j\to\infty} f(\boldsymbol{x}^{k_j}) = f(\boldsymbol{x}^*) \quad \text{and} \quad \lim_{j\to\infty} g(\boldsymbol{y}^{k_j}) = g(\boldsymbol{y}^*),$$

which leads to

$$\lim_{j\to\infty} \Psi^{k_j} \to \Psi(\boldsymbol{x}^*, \boldsymbol{y}^*). \tag{17}$$

In addition, from Lemma 7 we know that $(A_x^{k_j}, A_y^{k_j}) \in \partial\Psi(\boldsymbol{x}^{k_j}, \boldsymbol{y}^{k_j})$ which, together with equation (13) and the closedness of $\partial\Psi$, implies that $0 \in \partial\Psi(\boldsymbol{x}^*, \boldsymbol{y}^*)$. From the definition of critical point, we get $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a critical point of Ψ. Hence $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0) \subseteq \text{crit } \Psi$.

(ii) From equation (17) and the fact that $\{\Psi^k\}_{k\in\mathbb{N}}$ converges to Ψ^*, we get $\Psi(\boldsymbol{x}^*, \boldsymbol{y}^*) = \Psi^*$ for any $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$. So assertion (ii) holds.

□

Next, we prove that the sequence generated by PALMR converges to a critical point of the objective function $\Psi(\boldsymbol{x}, \boldsymbol{y})$ in (9) of the main text. For this purpose, we need the K-L property of Ψ. With this property, we have the following result, which is adapted from Lemma 6 of [6].

Lemma 9. *Let $\Gamma \subseteq \mathcal{M}$ be a compact set and $\sigma : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ be a PLS function such that it is constant on Γ and has K-L property at each point of Γ. Then, there exist $\epsilon > 0$, $\eta > 0$ and a continuous concave ϕ satisfying Definition 4 (i), such that for all $\bar{\boldsymbol{p}} \in \Gamma$ and all $\boldsymbol{p} \in \{\boldsymbol{p} \in \mathcal{M} : dist(\boldsymbol{p}, \Gamma) < \epsilon\} \cap [\sigma(\bar{\boldsymbol{p}}) < \sigma(\boldsymbol{p}) < \sigma(\bar{\boldsymbol{p}}) + \eta]$, we have*

$$\boxed{\phi'(\sigma(\boldsymbol{p}) - \sigma(\bar{\boldsymbol{p}}))dist(0, \partial\sigma(\boldsymbol{p})) \ge 1.}$$

Equipped with all these results, we are ready to prove Theorem 1.

Proof. Suppose the sequence $\{d_{\mathcal{M}_1\times\mathcal{M}_2}((\boldsymbol{x}^0, \boldsymbol{y}^0), (\boldsymbol{x}^k, \boldsymbol{y}^k))\}_{k\in\mathbb{N}}$ is bounded, or equivalently, the sequence $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ is bounded. Then Lemma 8 (i) shows that the limit set $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$ is a nonempty compact set and equalities (8) and (15) hold. If there exists integer $k_0 \ge 0$ such that $\Psi^{k_0} = \Psi^*$, then the decreasing property (4) in Lemma 6 shows that $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\ge k_0}$ is stationary and the claimed results are trivial to prove. In the rest of the proof, we assume $\Psi^k > \Psi^*$ for all $k \ge 0$.

(i) Lemma 8 (ii) shows that Ψ is constant on $\Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$ and equals to Ψ^*. Applying Lemma 9 with $\Gamma = \Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)$ and $\sigma = \Psi$, there exist $\epsilon > 0$, $\eta > 0$ and a continuous concave ϕ satisfying Definition 4 (i), such that for all $(\boldsymbol{x}, \boldsymbol{y})$ in the intersection

$$\{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{M}_1 \times \mathcal{M}_2 : \text{dist}((\boldsymbol{x}, \boldsymbol{y}), \Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)) < \epsilon\} \cap [\Psi^* < \Psi(\boldsymbol{x}, \boldsymbol{y}) < \Psi^* + \eta],$$

we have

$$\phi'(\Psi(\boldsymbol{x}, \boldsymbol{y}) - \Psi^*)\text{dist}(0, \partial\Psi(\boldsymbol{x}, \boldsymbol{y})) \ge 1. \tag{18}$$

For the above ϵ and η, equalities (8) and (15) imply that there exists integer k_0 such that

$$\text{dist}((\boldsymbol{x}^k, \boldsymbol{y}^k), \Gamma(\boldsymbol{x}^0, \boldsymbol{y}^0)) < \epsilon \quad \text{and} \quad \Psi^* < \Psi^k < \Psi^* + \eta, \ \forall k \ge k_0,$$

which further implies

$$\phi'(\Psi^k - \Psi^*)\text{dist}(0, \partial\Psi(\boldsymbol{x}^k, \boldsymbol{y}^k)) \ge 1$$

holds for $k \ge k_0$. Substituting inequality (14) into the above inequality, we get

$$\phi'(\Psi^k - \Psi^*) \ge \beta^{-1}(d^2_{\mathcal{M}_1}(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^k, \boldsymbol{y}^{k-1}))^{-1/2}. \tag{19}$$

Define $\Delta_{s,t} := \phi(\Psi^s - \Psi^*) - \phi(\Psi^t - \Psi^*)$, then the concavity of ϕ yields that for $k \ge k_0$

$$\begin{aligned}\Delta_{k,k+1} \ge& \phi'(\Psi^k - \Psi^*)(\Psi^k - \Psi^{k+1}) \\ \ge& \frac{\tau}{2\beta}(d^2_{\mathcal{M}_1}(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^k, \boldsymbol{y}^{k-1}))^{-1/2}(d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k)),\end{aligned}$$

or equivalently,

$$\begin{aligned}(d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k))^{1/2} \le& \sqrt{\frac{2\beta}{\tau}\Delta_{k,k+1}}(d^2_{\mathcal{M}_1}(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^k, \boldsymbol{y}^{k-1}))^{1/2} \\ \le& \frac{\beta}{\tau}\Delta_{k,k+1} + \frac{1}{2}(d^2_{\mathcal{M}_1}(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^k, \boldsymbol{y}^{k-1}))^{1/2},\end{aligned}$$

where we used the fact that $\sqrt{ab} \leq (a+b)/2$ for all $a, b \geq 0$ in the second inequality. Summing up the above inequality from $k = k_0$ to $k = K$, we get

$$\begin{aligned}\sum_{k=k_0}^{K} (d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k))^{-1/2} \leq & \frac{\beta}{\tau}\Delta_{k_0,K+1} + \frac{1}{2}\sum_{k=k_0}^{K} (d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k}, \boldsymbol{x}^{k-1}) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k}, \boldsymbol{y}^{k-1}))^{1/2} \\ \leq & \frac{\beta}{\tau}\phi(\Psi^{k_0} - \Psi^*) + \frac{1}{2}\sum_{k=k_0-1}^{K-1} (d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^{k}) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^{k}))^{1/2},\end{aligned}$$

where the first inequality follows from the fact that $\Delta_{s,r} + \Delta_{r,t} = \Delta_{s,t}$ and the second inequality follows from $\phi \geq 0$. A simple manipulation of the above inequality yields

$$\sum_{k=k_0}^{K} (d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k))^{-1/2} \leq \frac{2\beta}{\tau}\phi(\Psi^{k_0} - \Psi^*) + (d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k_0}, \boldsymbol{x}^{k_0-1}) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k_0}, \boldsymbol{y}^{k_0-1}))^{1/2}.$$

Taking $K \to \infty$ leads to

$$\sum_{k=k_0}^{\infty} (d^2_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + d^2_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k))^{-1/2} < \infty,$$

which implies that

$$\boxed{\sum_{k=0}^{\infty} d_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) < \infty \quad \text{and} \quad \sum_{k=0}^{\infty} d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k) < +\infty.}$$

(ii) From assertion (i) we know that both sequences $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{y}^k\}_{k\in\mathbb{N}}$ are Cauchy sequences on Hadamard manifolds. Thus, by the Hopf-Rinow's Theorem we infer that both sequences converge. Denote the limit points by $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$, then we conclude from Lemma 8 that $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a critical point of Ψ. □

3 Partial Derivatives of E

In this section, we derive partial derivatives of the loss function E in the main text via the caculus of variation approach. For this purpose, we denote $\boldsymbol{p}^\epsilon = \mathrm{Exp}_{\boldsymbol{p}}(\epsilon\boldsymbol{v})$ the neighboring point of $\boldsymbol{p}$ along tangent vector direction $\boldsymbol{v}$, where $\epsilon \in \mathbb{R}$ and $\boldsymbol{v}$ is an arbitrary tangent vector of $\boldsymbol{p}$. We also let $\hat{\boldsymbol{v}}_j$ denote the parallel transport of $\boldsymbol{v}_j$ from $\boldsymbol{p}$ to $\boldsymbol{p}^\epsilon$.

By taking derivative of $E\big(\boldsymbol{p}^\epsilon, \{\hat{\boldsymbol{v}}_j\}, \{\boldsymbol{g}_i\}\big)$ at $\epsilon = 0$ we have

$$\begin{aligned}&\partial_\epsilon E|_{\epsilon=0} \\ &= \frac{1}{2}\sum_i \partial_\epsilon d^2\Big(\boldsymbol{y}_i^c,\ \mathrm{Exp}_{\boldsymbol{p}^\epsilon}\big(\sum_j x_i^j \hat{\boldsymbol{v}}_j\big)\Big)\Big|_{\epsilon=0} \\ &= \sum_i \left\langle -\mathrm{Exp}^{-1}_{\mathrm{Exp}_{\boldsymbol{p}^\epsilon}\left(\boldsymbol{x}_i^j \hat{\boldsymbol{v}}_j\right)} \boldsymbol{y}_i^c,\ \partial_\epsilon \mathrm{Exp}_{\boldsymbol{p}^\epsilon}\big(\sum_j x_i^j \hat{\boldsymbol{v}}_j\big) \right\rangle\Big|_{\epsilon=0} \\ &= \sum_i \left\langle -\mathrm{Exp}^{-1}_{\hat{\boldsymbol{y}}_i} \boldsymbol{y}_i^c,\ \partial_\epsilon \mathrm{Exp}_{\boldsymbol{p}^\epsilon}\big(\sum_j x_i^j \hat{\boldsymbol{v}}_j\big)|_{\epsilon=0} \right\rangle_{\hat{\boldsymbol{y}}_i} \\ &= \sum_i \left\langle -\mathrm{Exp}^{-1}_{\hat{\boldsymbol{y}}_i} \boldsymbol{y}_i^c,\ d_{\boldsymbol{p}}\mathrm{Exp}_{\boldsymbol{p}}\big(\sum_j x_i^j \boldsymbol{v}_j\big)\boldsymbol{v} \right\rangle_{\hat{y}_i} \\ &= \sum_i \left\langle -\Big(d_{\boldsymbol{p}}\mathrm{Exp}_{\boldsymbol{p}}\big(\sum_j x_i^j \boldsymbol{v}_j\big)\Big)^\dagger \mathrm{Exp}^{-1}_{\hat{\boldsymbol{y}}_i} \boldsymbol{y}_i^c,\ \boldsymbol{v} \right\rangle_{\boldsymbol{p}},\end{aligned}$$

where in the second equality we used the fact that $\partial_{\boldsymbol{p}} d^2_{\mathcal{M}}(\boldsymbol{p}', \boldsymbol{p}) = -2\mathrm{Exp}^{-1}_{\boldsymbol{p}} \boldsymbol{p}'$ for Hadamard manifold $\mathcal{M}$ and in the last equality we used the definition of adjoint derivative. As a result, we have

$$\boxed{\partial_{\boldsymbol{p}} E = -\sum_i \Big(d_{\boldsymbol{p}}\mathrm{Exp}_{\boldsymbol{p}}\big(\sum_j x_i^j \boldsymbol{v}_j\big)\Big)^\dagger \mathrm{Exp}^{-1}_{\hat{\boldsymbol{y}}_i} \boldsymbol{y}_i^c.}$$

For any fixed j_0 satisfying $1 \leq j_0 \leq d$ and arbitrary $\boldsymbol{v}$ from $T_{\boldsymbol{p}}\mathcal{M}$, let $\boldsymbol{v}_{j_0}^{\epsilon} = \boldsymbol{v}_{j_0} + \epsilon \boldsymbol{v}$ and take derivative of $E\big(\boldsymbol{p}, \boldsymbol{v}_{j_0}^{\epsilon}, \{\boldsymbol{v}_j\}_{j\neq j_0}, \{\boldsymbol{g}_i\}\big)$ at $\epsilon = 0$, we have

$$
\begin{aligned}
&\partial_\epsilon E|_{\epsilon=0} \\
&= \sum_i \left\langle -\mathrm{Exp}_{\hat{\boldsymbol{y}}_i}^{-1}\boldsymbol{y}_i^c,\ \partial_\epsilon \mathrm{Exp}_{\boldsymbol{p}}\big(\sum_{j\neq j_0} x_i^j \boldsymbol{v}_j + x_i^{j_0}\boldsymbol{v}_{j_0}^{\epsilon}\big)\Big|_{\epsilon=0} \right\rangle_{\hat{\boldsymbol{y}}_i} \\
&= \sum_i \left\langle -\mathrm{Exp}_{\hat{\boldsymbol{y}}_i}^{-1}\boldsymbol{y}_i^c,\ d_{\boldsymbol{v}}\mathrm{Exp}_{\boldsymbol{p}}\big(\sum_j x_i^j \boldsymbol{v}_j\big)(x_i^{j_0}\boldsymbol{v}) \right\rangle_{\boldsymbol{y}_i} \\
&= \sum_i \left\langle -x_i^{j_0}\Big(d_{\boldsymbol{v}}\mathrm{Exp}_{\boldsymbol{p}}\big(\sum_j x_i^j \boldsymbol{v}_j\big)\Big)^{\dagger}\mathrm{Exp}_{\hat{\boldsymbol{y}}_i}^{-1}\boldsymbol{y}_i^c,\ \boldsymbol{v} \right\rangle_{\boldsymbol{p}}.
\end{aligned}
$$

Therefore,

$$
\boxed{\partial_{\boldsymbol{v}_j} E = -\sum_i x_i^j \Big(d_{\boldsymbol{v}}\mathrm{Exp}_{\boldsymbol{p}}\big(\sum_{j'} x_i^{j'}\boldsymbol{v}_{j'}\big)\Big)^{\dagger}\mathrm{Exp}_{\hat{\boldsymbol{y}}_i}^{-1}\boldsymbol{y}_i^c}
$$

for $j = 1, \cdots, d$. By a similar procedure, we get

$$
\boxed{\partial_{\boldsymbol{g}_i} E = -\left(d_{\boldsymbol{v}}\mathrm{Exp}_{\boldsymbol{y}_i}(\boldsymbol{g}_i)\right)^{\dagger}\mathrm{Exp}_{\boldsymbol{y}_i^c}^{-1}\hat{\boldsymbol{y}}_i}
$$

for $i = 1, \cdots, N$.

4 Manifold of Symmetric Positive Definite Matrices

It is well known that the set of $n \times n$ SPD matrices $\mathcal{S}_{++}(n)$ is a Riemannian symmetric space with nonpositive sectional curvature [9]. Hence, $\mathcal{S}_{++}(n)$ is a Hadamard manifold. For any $\boldsymbol{p} \in \mathcal{S}_{++}(n)$, the tangent space at $\boldsymbol{p}$, which we denote as $T_{\boldsymbol{p}}\mathcal{M}$, is the space of $n \times n$ symmetric matrices $\mathcal{S}(n)$. The inner product of two tangent vectors $\boldsymbol{u}, \boldsymbol{v} \in T_{\boldsymbol{p}}\mathcal{M}$ is given by

$$
\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{p}} = \mathrm{Tr}(\boldsymbol{p}^{-1/2}\boldsymbol{u}\boldsymbol{p}^{-1}\boldsymbol{v}\boldsymbol{p}^{-1/2}), \tag{20}
$$

where $\mathrm{Tr}(\cdot)$ is the trace of matrices. The exponential map and the inverse exponential map are given by

$$
\mathrm{Exp}_{\boldsymbol{p}}(\boldsymbol{v}) = \boldsymbol{p}^{1/2}\mathrm{exp}(\boldsymbol{p}^{-1/2}\boldsymbol{v}\boldsymbol{p}^{-1/2})\boldsymbol{p}^{1/2}, \tag{21}
$$

and

$$
\mathrm{Exp}_{\boldsymbol{p}}^{-1}\boldsymbol{q} = \boldsymbol{p}^{1/2}\mathrm{log}(\boldsymbol{p}^{-1/2}\boldsymbol{q}\boldsymbol{p}^{-1/2})\boldsymbol{p}^{1/2}, \tag{22}
$$

respectively, where $\mathrm{exp}(\cdot)$ and $\mathrm{log}(\cdot)$ denote matrix exponential and logarithm [10], respectively. The geodesic distance between any two points $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{S}_{++}(n)$ is given by

$$
d(\boldsymbol{p}, \boldsymbol{q}) = \mathrm{Tr}(\mathrm{log}^2(\boldsymbol{p}^{-1/2}\boldsymbol{q}\boldsymbol{p}^{-1/2})). \tag{23}
$$

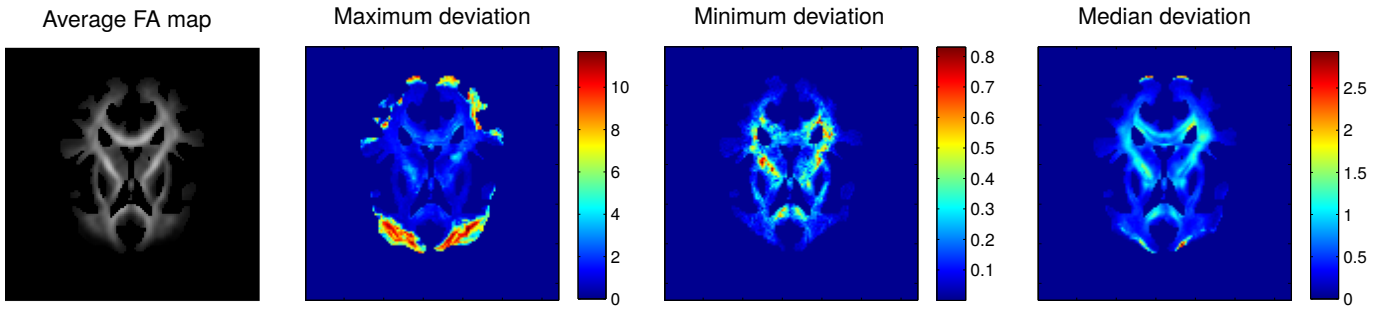
Let $\boldsymbol{v} \in T_{\boldsymbol{p}}\mathcal{M}$, the parallel transport of $\boldsymbol{v}$ along the geodesic from $\boldsymbol{p}$ to $\boldsymbol{q}$ is given by

$$
P_{\boldsymbol{p}\boldsymbol{q}}(\boldsymbol{v}) = \boldsymbol{p}^{1/2}\boldsymbol{u}\boldsymbol{p}^{-1/2}\boldsymbol{v}\boldsymbol{p}^{-1/2}\boldsymbol{u}\boldsymbol{p}^{1/2}, \tag{24}
$$

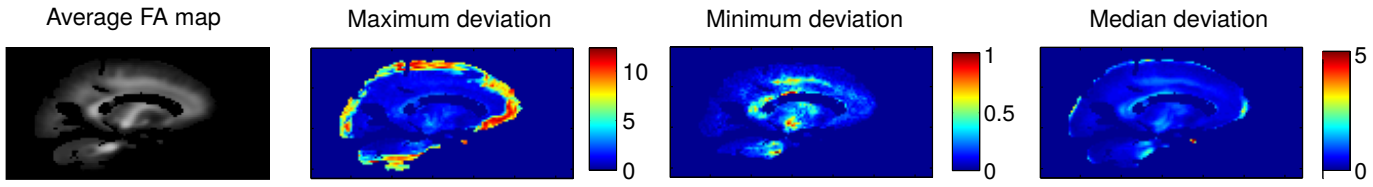
where $\boldsymbol{u} = \mathrm{exp}(\boldsymbol{p}^{-1/2} \cdot \mathrm{Exp}_{\boldsymbol{p}}^{-1}\boldsymbol{q} \cdot \boldsymbol{p}^{-1/2}/2)$.

5 Additional Experiments on Real DTI Data

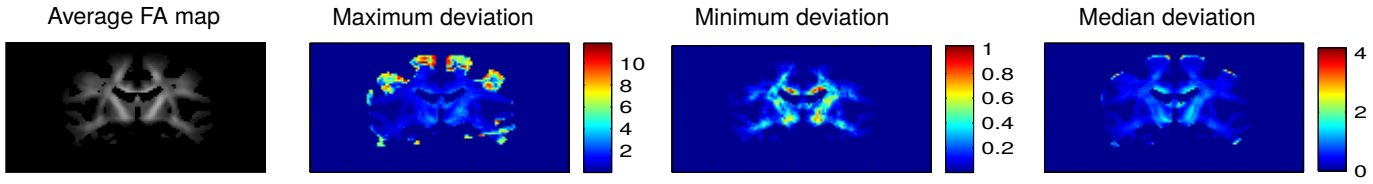
In Subsection 4.2 of the main paper, we conducted experiments on a real DTI data with three different settings: no gross error, 20% manual gross error, and 20% registration error. In this section, we provide more information and addtional experiments on the DTI data.



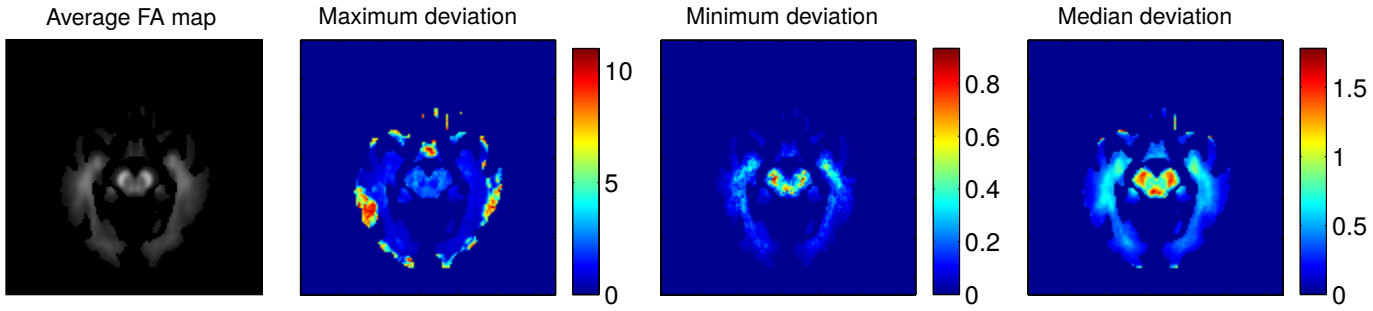
(a) Slice $z = 32$



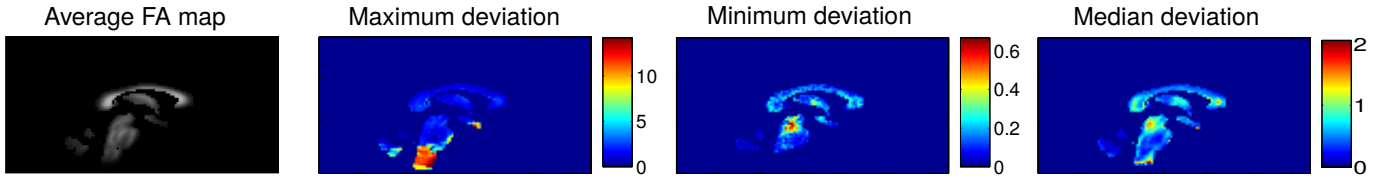
(b) Slice $x = 55$



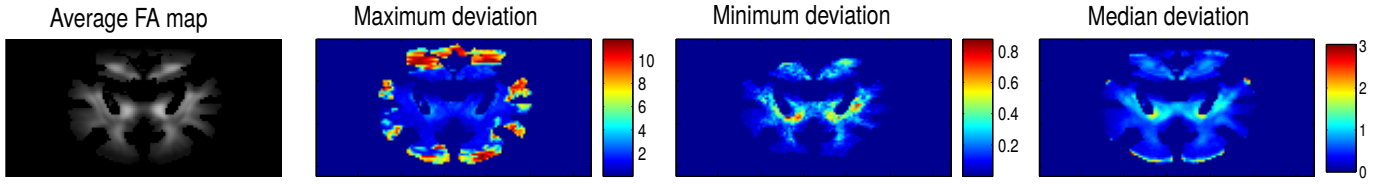
(c) Slice $y = 64$



(d) Slice $z = 24$



(e) Slice $x = 64$

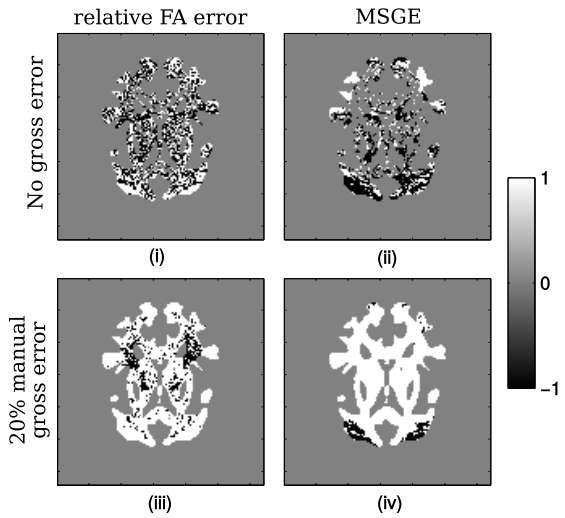


(f) Slice $y = 45$

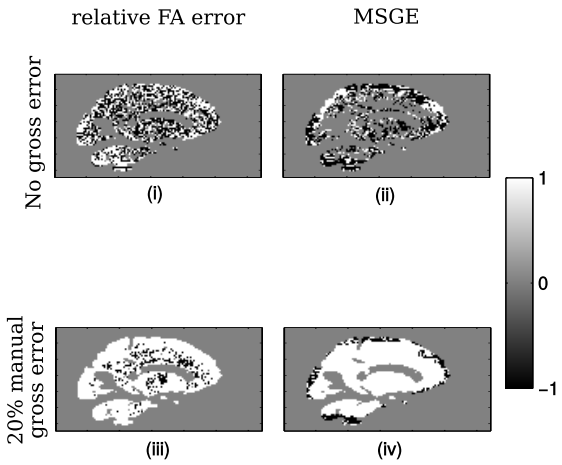
Fig. 1. Tensor deviation between training data without gross error and with 20% registration error. On each voxel, we compute an average FA value over all 58 patients and plot the FA map over the whole slice in the first column. Similarly, we compute a deviation vector of length 58 (the number of patients in the dataset) whose maximum, minimum, and median values are shown in three heat maps, respectively.

5.1 Additional Information on DTI Data with Registration Error

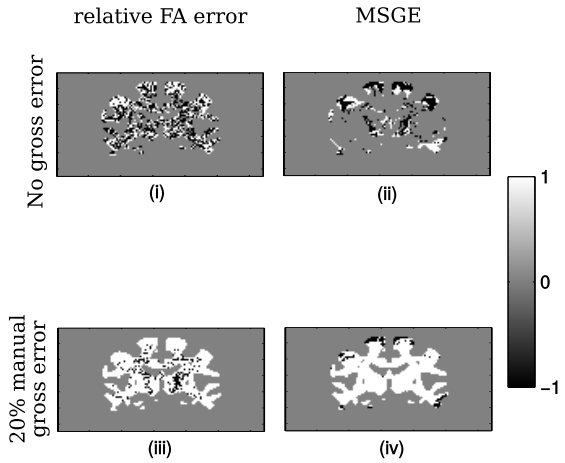
In Fig. 1, we show the tensor deviation between training data without gross error and training data with 20% registration error. On each voxel, we compute a deviation vector consisting of 58 (the number of patients in the dataset) entries, where each entry records the geodesic distance between the tensor without gross error and the tensor with 20% registration error of the same patient. The maximum,



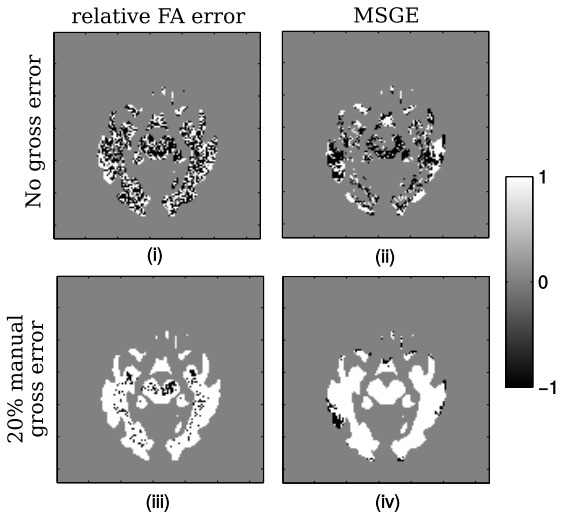
(a) Slice $z = 32$



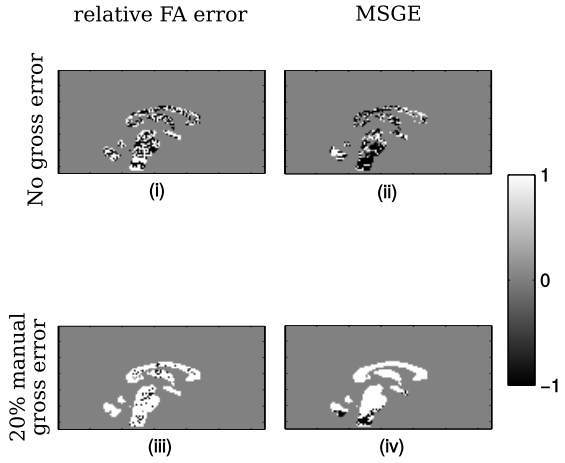
(b) Slice $x = 55$



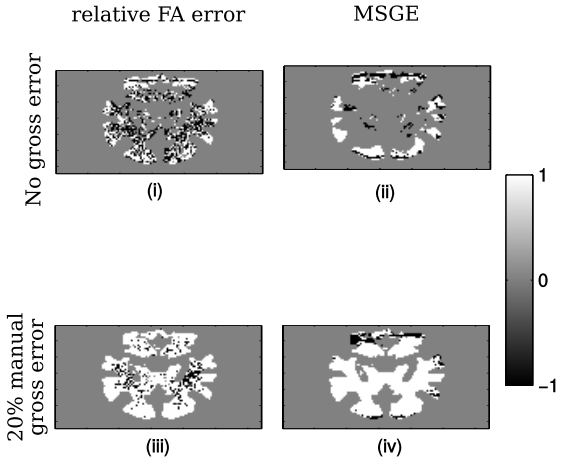
(c) Slice $y = 64$



(d) Slice $z = 24$



(e) Slice $x = 64$



(f) Slice $y = 45$

Fig. 2. Performance comaprison of MGLM and PALMR on six slices. On each voxel, we compare the median prediction error (measured in both the relative FA error and MSGE) on the testing data of MGLM and PALMR. If MGLM achieves smaller error than PALMR, we assign -1 to the voxel; if MGLM achieves larger error than PALMR, we assign 1 to the voxel; if the voxel is outside the mask or the error difference bwtween MGLM and PALMR is less than 1e-3, we assign 0 to the voxel. In each subfigure, there are four subplots corresponding to two experimental settings under two error metrics.

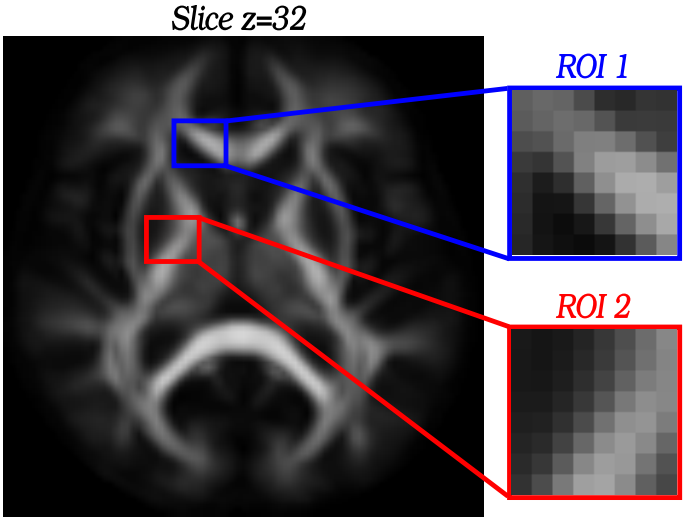


Fig. 3. Slice $z = 32$ and two ROIs. For each voxel in the ROIs, the underlying intensity value indicates FA value.

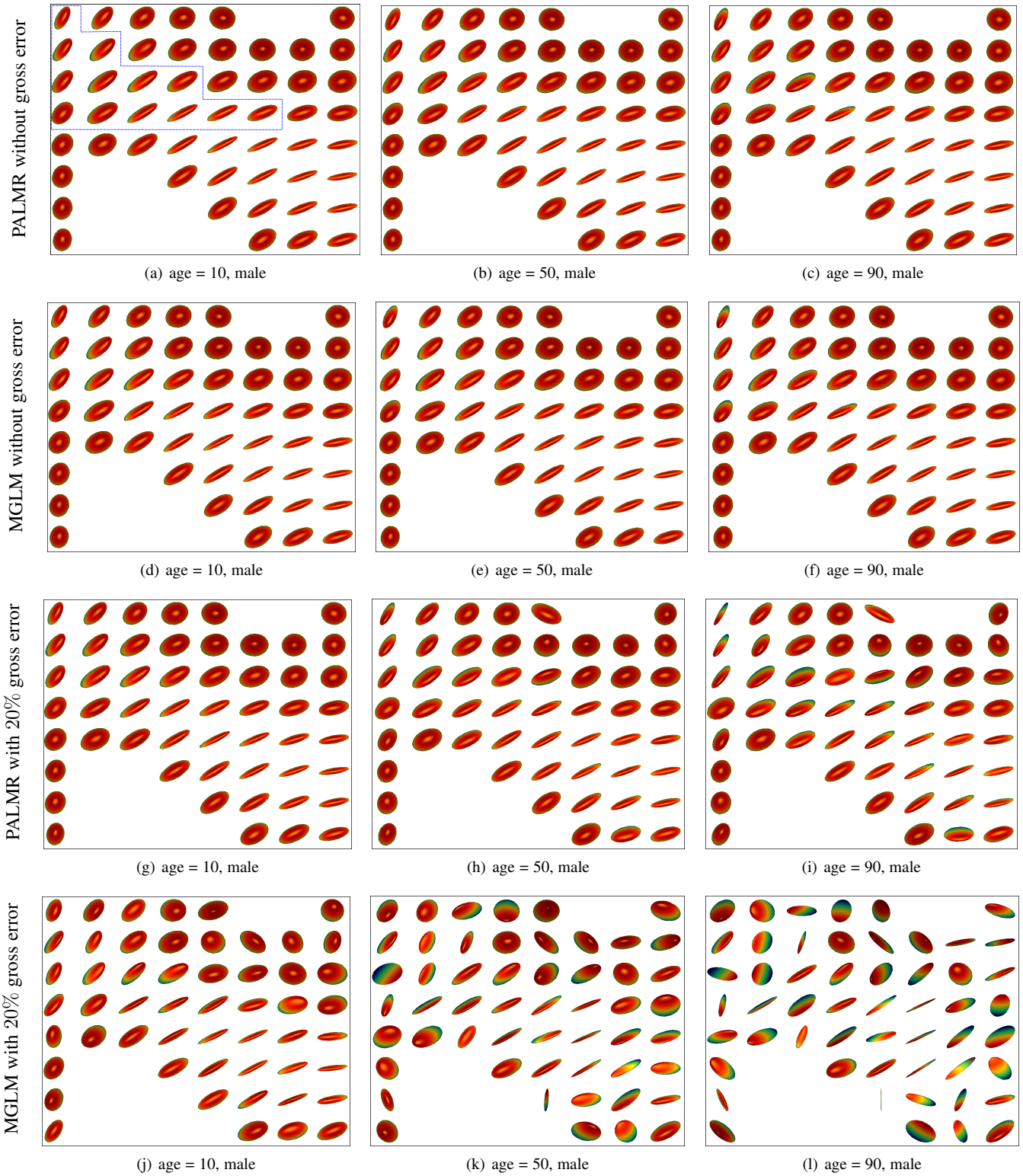
minimum, and median values of the deviation vector are shown in three heat maps, respectively. As a reference, for each voxel we also compute the population average FA value on the origianl clean data, and plot the average FA map for all slices which are shown in the first column of Fig. 1. From Fig. 1, we observe that the magnitude of registration errors varies a lot among different patients and different voxels. In particular, on the boundary region of the mask the registration error can be very high, as shown in the second column of Fig. 1. However, such high registration error on the boundary happens to only a small fraction of patients, since in the same regions the median deviation is usually small, as shown in the last column of Fig. 1. In addition, the last two columns of Fig. 1 demonstrate that registration errors appear in a structured manner, namely, registration errors with moderate median deviation appears in the interior of white matter while only a tiny fraction of errors with extreme magnitude appears on the boundary. Based on this observation, in our experiments, we focus on those voxels whose minimum registration errors are larger than certain threshold, that is, the interior regions of the white matter.

In the main paper, we presented the distribution of prediction errors on each slice for all comparison methods. In Fig. 2, we show the comparison results of MGLM and PALMR one each voxel of all six slices, separately. On each voxel, we compare the median prediction error (measured in both the relative FA error and MSGE) on the testing data of MGLM and PALMR. If MGLM achieves smaller error than PALMR, we assign -1 to the voxel; if MGLM achieves larger error than PALMR, we assign 1 to the voxel; if the voxel is outside the mask or the error difference between MGLM and PALMR is less than 1e-3, we assign 0 to the voxel. Since for training data with 20% registration error, experiments were not conducted on the whole slice, we did not show the comparison result in this way. From Fig. 2 we have two observations: (1) When the training data contain no gross error, MGLM and PALMR have similar prediction errors with negligible difference on many voxels, while for the rest of voxels MGLM is better on half of them and PALMR is better on the other half. Such observation is from the first row of each subfigure. (2) When 20% of the training data contain gross error, PALMR outperforms MGLM on most of the voxels (i.e., a large portion of white region). This can be observed from the second row of each subfigure.

5.2 Region of Interest Analysis

To further test the predictability and robustness of our model, we conducted a region of interest (ROI) analysis on the real DTI data. First, we used all 58 instances as training data and focused on white matter tensors in slice $z = 32$. Then, we selected two ROIs [11]: the genu of the corpus callosum (ROI 1) and posterior limb of internal capsule (ROI 2) as shown in Fig. 3. Both ROIs are believed to be affected by age or gender, and each ROI contains 8×8 voxels. After that, for each voxel in each ROI we trained our model and MGLM using all 58 training instances either without or with $\beta = 20\%$ manual gross errors with magnitude $\sigma_g = 5$, and applied the trained models to predict DTI tensors for different ages. Visualization results of MGLM and our model in both ROIs using different ages (age = 10, 50 and 90) are shown in Fig. 4 and Fig. 5, respectively. In subfigure (a) of each figure, we use a box with blue-dotted boundary to highlight those voxels where DTI tensors have large visual variation in both size and orientation with respect to age. To investigate the effect of gender, we apply our model to predict tensors for both male and female on the ROIs. For the sake of comparison, in each voxel we apply our model to predict two tensors, one for male and the other for female, using the same age, then compute the geodesic distance between these two tensors. Visualization results are shown in Fig. 6.

In Fig. 4 and Fig. 5, we can observe that both MGLM and PALMR trained on training data without gross errors can identify a few voxels (highlighted with blue box) where the shape and orientation of tensors change significantly as age increases. Moreover, the voxels identified are contained in the region of high white matter intensity as shown in Fig. 3. This observation is consistent with findings in neuroscience [12] that with increasing age, FA values increase in the internal capsule and the corpus callosum. The posterior limb of the internal capsule and the corpus callosum show the most significant overlaps between white matter intensity and FA changes with age. We also observe that PALMR has similar predictions as MGLM when both are trained on data without gross error, but outperforms MGLM when gross errors are present in the training data in the sense that when there are 20% manual gross errors, PALMR can still predict tensors showing meaningful diffusion trends while the predictions of MGLM are nearly random.



(a) age = 10, male (b) age = 50, male (c) age = 90, male

(d) age = 10, male (e) age = 50, male (f) age = 90, male

(g) age = 10, male (h) age = 50, male (i) age = 90, male

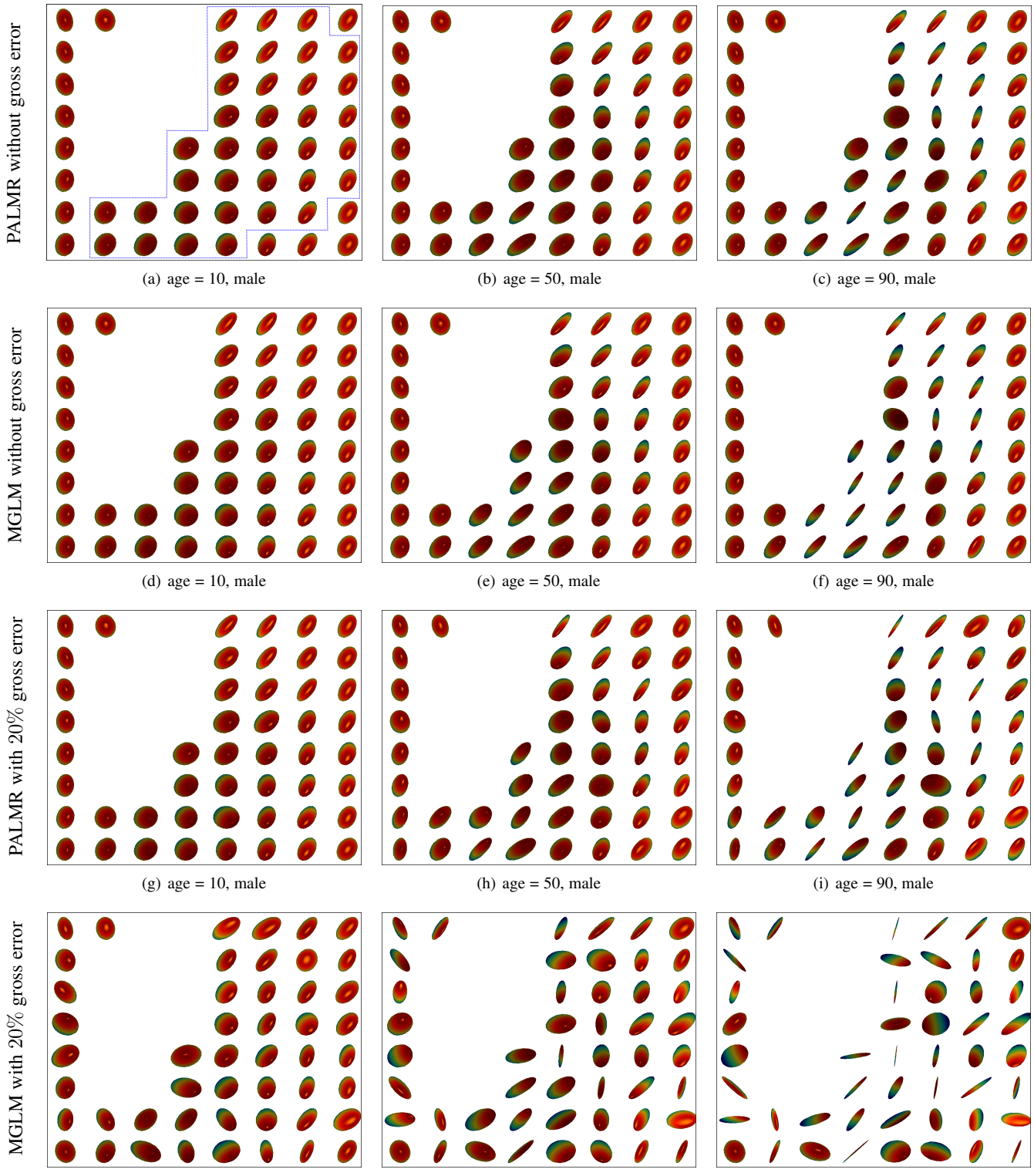
(j) age = 10, male (k) age = 50, male (l) age = 90, male

Fig. 4. Visualization of aging effect on ROI 1: First two rows show results by models trained on data without gross error. Last two rows show results by models trained on data with 20% manual gross errors. Only tensors in the white matter mask are illustrated. Better viewed in color.

In Fig. 6, we observe that for some regions (e.g. ROI 1 and ROI 2) the diffusion tensors of male and female are quite different in the shape and orientation. Moreover, in the same region of interest the voxels significantly affected by gender are contained in the region of high FA values. All these observations imply that gender plays an important role in deciding the diffusion trend in the brain and is worth further study.

REFERENCES

[1] E. A. Papa Quiroz, "An extension of the proximal point algorithm with Bregman distances on Hadamard manifolds," *J. Glob. Optim.*, vol. 56, no. 1, pp. 43–59, 2013.



(a) age = 10, male (b) age = 50, male (c) age = 90, male

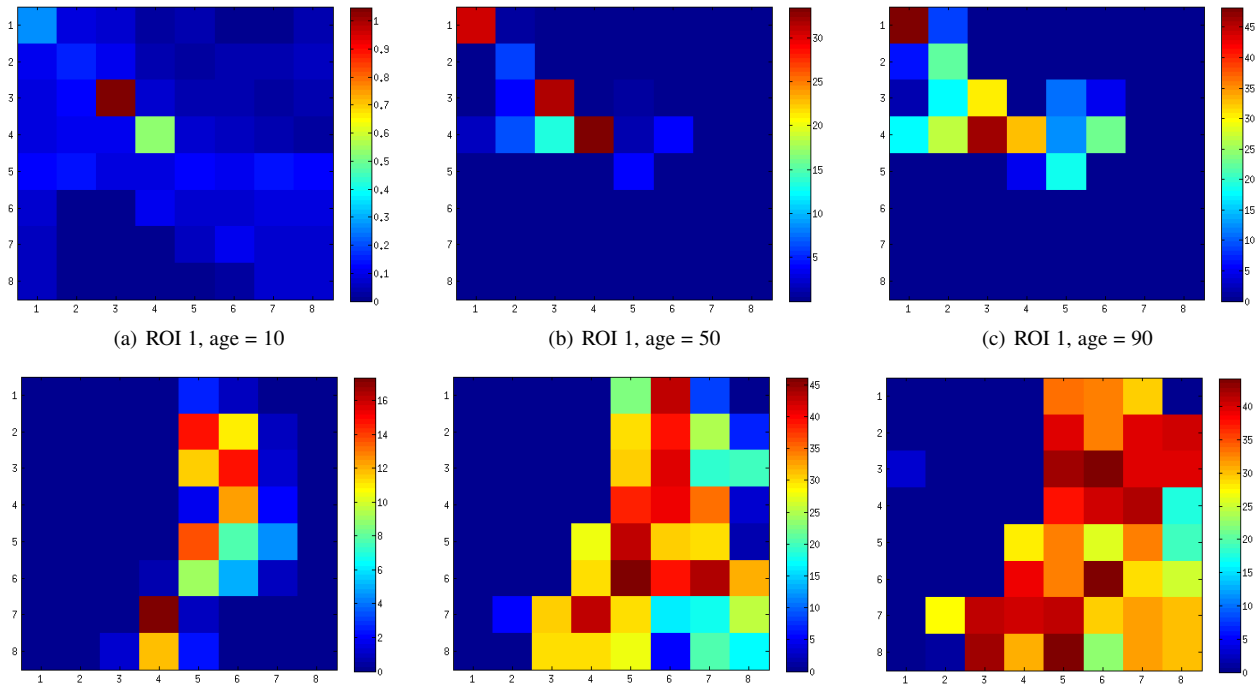
(d) age = 10, male (e) age = 50, male (f) age = 90, male

(g) age = 10, male (h) age = 50, male (i) age = 90, male

(j) age = 10, male (k) age = 50, male (l) age = 90, male

Fig. 5. Visualization of aging effect on ROI 2: First two rows show results by models trained on data without gross error. Last two rows show results by models trained on data with 20% manual gross errors. Only tensors in the white matter mask are illustrated. Better viewed in color.

[2] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.
[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality," *Math. Oper. Res.*, vol. 35, no. 2, pp. 438–457, 2010.
[4] J. X. da Cruz Neto, P. R. Oliveira, P. A. S. Jr, and A. Soubeyran, "Learning how to play Nash, potential games and alternating minimization method for structured nonconvex problems on Riemannian manifolds," *J. Convex Anal.*, vol. 20, no. 2, pp. 395 – 438, 2013.
[5] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods." *Math. Program.*, vol. 137, pp. 91–129, 2013.
[6] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, no. 1-2, pp. 459–94, 2014.



(a) ROI 1, age = 10 (b) ROI 1, age = 50 (c) ROI 1, age = 90

(d) ROI 2, age = 10 (e) ROI 2, age = 50 (f) ROI 2, age = 90

Fig. 6. Visualization of gender effect on different ROIs using predicted tensors by PALMR. The value in each voxel denotes the geodesic distance between tensors of male and female obtained by our model using training samples without gross error. Better viewed in color.

[7] M. P. do Carmo, *Riemannian Geometry*. Birkhäuser, 1992.

[8] P. Peterson, *Manifold Theory*, http://www.math.ucla.edu/~petersen/manifolds.pdf.

[9] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces*. American Mathematical Society, 2001.

[10] N. Higham, *Functions of Matrices: Theory and Computation*. SIAM, 2008.

[11] A. Toga, P. Thompson, S. Mori, K. Amunts, and K. Zilles, "Towards multimodal atlases of the human brain," *Nat. Rev. Neurosci.*, vol. 7, no. 12, pp. 952–966, 2006.

[12] N. Barnea-Goraly, V. Menon, M. Eckert, L. Tamm, R. Bammer, A. Karchemskiy, C. Dant, and A. Reiss, "White matter development during childhood and adolescence: a cross-sectional diffusion tensor imaging study," *Cereb. Cortex*, vol. 15, no. 12, pp. 1848–1854, 2005.