Multivariate Regression with Gross Errors on Manifold-Valued Data

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

Abstract—We consider the topic of multivariate regression on manifold-valued output, that is, for a multivariate observation, its output response lies on a manifold. Moreover, we propose a new regression model to deal with the presence of grossly corrupted manifold-valued responses, a bottleneck issue commonly encountered in practical scenarios. Our model first takes a correction step on the grossly corrupted responses via geodesic curves on the manifold, then performs multivariate linear regression on the corrected data. This results in a nonconvex and nonsmooth optimization problem on Riemannian manifolds. To this end, we propose a dedicated approach named PALMR, by utilizing and extending the proximal alternating linearized minimization techniques for optimization problems on euclidean spaces. Theoretically, we investigate its convergence property, where it is shown to converge to a critical point under mild conditions. Empirically, we test our model on both synthetic and real diffusion tensor imaging data, and show that our model outperforms other multivariate regression models when manifold-valued responses contain gross errors, and is effective in identifying gross errors.

Index Terms—Manifold-valued data, multivariate linear regression, gross error, nonsmooth optimization on manifolds, diffusion tensor imaging

1 Introduction

This paper focuses on multivariate regression on manifolds [1], [2], [3], [4], where given a multivariate observation $\boldsymbol{x} \in \mathbb{R}^d$, the output response $\boldsymbol{y}$ lies on a Riemannian manifold $\mathcal{M}$. This line of work has many applications. For example, research evidence in diffusion tensor imaging (DTI) (e.g., [5]) indicates that the shape and orientation of diffusion tensors are profoundly affected by age, gender and handedness (i.e., left- or right-handed). In particular, we consider noisy manifold-valued output scenarios where data are subject to sporadic contamination by gross errors of large or even unbounded magnitude. Such grossly corrupted data are often encountered in practice due to unreliable data collection or data with missing values: For example, errors in DTI data can be introduced by Echo-Planar Imaging (EPI) distortion [6] or inter-subject registration [7], where practical measurement errors such as Rician noise or other sensor noise have a significant impact on the shape and orientation of tensors [8], [9]. Although the problem of learning from data with possible gross error in euclidean spaces has gained increasing interest [10], [11], [12], [13], [14], [15], to our best knowledge, there exists no prior work in dealing with manifold-valued response with gross errors.

- *X. Zhang and L. Cheng are with the Bioinformatics Institute, A*STAR, Singapore 138632. E-mail: zxwtroy87@gmail.com, chengli@bii.a-star.edu.sg.*
- *X. Shi is with the School of Computing, National University of Singapore, Singapore 117456. E-mail: shixudongleo@gmail.com.*
- *Y. Sun is with the Singapore Institute for Neurotechnology, National University of Singapore, Singapore 117456. E-mail: lsisu@nus.edu.sg..*

Manuscript received 24 Feb. 2017; revised 10 Sept. 2017; accepted 17 Nov. 2017. Date of publication 3 Jan. 2018; date of current version 16 Jan. 2019. (Corresponding author: Li Cheng.)
Recommended for acceptance by H. Li.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2776260

Our main idea can be summarized as follows: For each manifold-valued response $\boldsymbol{y} \in \mathcal{M}$, we explicitly model its possible gross error (in $\boldsymbol{y}$). This gives rise to a *corrected* manifold-valued data $\boldsymbol{y}^c$ by removing the identified gross error component from $\boldsymbol{y}$, which is realized via geodesic curves on $\mathcal{M}$. Note that $\boldsymbol{y}^c$ could be the same as $\boldsymbol{y}$, corresponding to no gross error in $\boldsymbol{y}$. Then the corrected manifold-valued data can be utilized as the responses in multivariate geodesic regression, which boils down to a known problem [2]. More details are illustrated in Fig. 1 and are fully described in Section 3. Unfortunately, the induced optimization problem becomes rather challenging as it contains nonconvex and nonsmooth functions on manifolds. Inspired by the recent development of proximal alternating linearized minimization (PALM) methods in euclidean spaces, in this paper we propose to generalize this technique onto Riemannian manifolds [16], which we have named as *PALMR*.

The main contributions of this paper are three-fold. First, we propose to address a novel problem of multivariate regression on manifolds where the manifold-valued responses are subject to possible contamination of gross errors. Second, a new algorithm named PALMR is proposed to tackle the induced nonconvex and nonsmooth optimization on manifolds, for which we also provide the convergence analysis. The algorithm and analysis is applicable to a class of nonconvex and nonsmooth optimization problem on manifolds. Empirically our algorithm has been evaluated on both synthetic and real DTI data, where results suggest the algorithm is effective in identifying gross errors and recovering corrupted data, and it produces better predictive results than regression models that do not consider gross errors. Third, our approach makes connections to two established research areas, namely learning from grossly corrupted data and multivariate regression on manifolds:

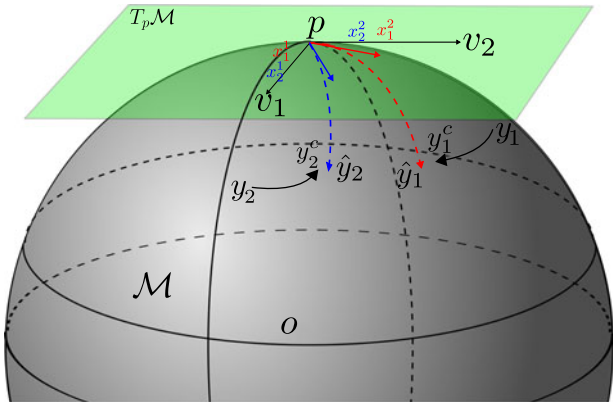


Fig. 1. An illustration of the proposed approach for multivariate regression on grossly corrupted manifold-valued data, which contains two main ingredients: The first is to obtain the corrected response $\boldsymbol{y}_i^c$ by removing its possible gross error, as illustrated by the directed curves on the manifold; The second one involves the manifold-valued regression process using $\{\boldsymbol{x}_i, \boldsymbol{y}_i^c\}$. Here x_1^1 and x_1^2 are the components of input $\boldsymbol{x}_1$ along tangent vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ of point $\boldsymbol{p} \in \mathcal{M}$. Red solid arrow denotes a tangent vector $x_1^1\boldsymbol{v}_1 + x_1^2\boldsymbol{v}_2$ at $\boldsymbol{p}$, and red dash arrow is its corresponding geodesic path. x_2^1 and x_2^2 are also defined similarly using blue color. See Section 3.1 for details.

When we restrict ourselves to the special case of euclidean space, our approach reduces to robust regression considered in e.g., [13], [14]; On the other hand, when there is no gross error, the problem boils down to that of multivariate regression on manifolds as considered in [2], where the method of [2] can be regarded as a special case of our approach. Our code is also made publicly available.[1]

1.1 Related Work

Manifold-valued data arise from a wide range of application domains including neural imaging [17], shape modelling [18], [19], [20], robotics [21], graphics [22], and symmetric positive matrices [23], [24], [25], [26]. One prominent example is DTI [24] where data lie in the Riemannian manifold of 3×3 symmetric positive definite (SPD) matrices. In this work, we use $\mathcal{S}(n)$ and $\mathcal{S}_{++}(n)$ to denote the set of $n \times n$ symmetric matrices and $n \times n$ SPD matrices, respectively. Other examples include higher angular resolution diffusion imaging where data can be modelled as the square root of orientation distribution functions lying on the unit sphere [2], [27], as well as group-valued data such as $SO(3)$ and $SE(3)$ in shape analysis [19] and robotics [21]. It is well known that for such scenarios, it is in general much better to conduct statistical analysis directly on the manifold (i.e., curved space) instead of in the ambient euclidean space (i.e., flat space), which we also verify empirically.

Unsurprisingly, there exists plenty of prior work studying statistics on manifolds [19], [24], [28], [29], [30]. This is to be distinguished from the well-known topic of *manifold learning* [31], where the data are assumed to be sampled from certain manifold embedded in a usually much higher dimensional euclidean space and one is supposed to extract intrinsic geometric properties of the manifold from observations. Instead here the manifold is usually known in priori, and the task is to engage appropriate statistical models in the analysis of the manifold-valued data.

In the area of regression on manifolds, Fletcher [28] proposes *geodesic regression* that generalizes univariate linear regression on flat spaces to manifolds by regressing a manifold-valued response from a real-valued independent data with a geodesic curve. [27] adapts the idea of geodesic regression for regressing sphere-valued data against real scalar. [20] investigates parametric polynomial regression on Riemannian manifolds, while [1] studies regression on the group of diffeomorphisms for detecting longitudinal anatomical shape changes. Banerjee et al. [32] propose a nonlinear kernel-based regression method for manifold-valued data. Hong et al. [33] propose a shooting spline-based regression technique specifically designed for the Grassmannian manifold. [34], [35] investigate a family of nonparametric regression models for data on manifolds. The closest work might be [2], which extends the idea of geodesic regression [28] to multivariate regression on manifolds, and applies it to analyze diffusion weighted imaging data. In [3], the authors investigate multivariate regression models on Riemannian symmetric spaces from a statistical perspective and develop several test statistics for evaluating linear hypotheses of the regression coefficients. In the area of learning with grossly corrupted data, there have been various methods [10], [13], [14], [15], [36] proposed for linear regression with gross errors in the euclidean space, among which robust lasso in [13] and robust multi-task regression in [14] can be considered as special cases of our approach when restricted to euclidean spaces.

A recent trend in manifold data analysis is kernel methods on manifolds which aim at embedding the manifold to a reproducing kernel Hilbert space (RKHS). In [37] and [38], kernel methods are developed for sparse coding and dictionary learning on SPD and Grassmann manifolds, respectively. In [39], kernels on SPD and Grassmann manifolds are considered for classification. As it is important for such kernels on manifolds to satisfy the positive definite constraint, significant efforts [40], [41], [42], [43], [44] have been made in this regard. Meanwhile, as shown in [44], these kernels tend to either disregard the original Riemannian structure due to linearization requirement, or violates the positive definiteness constraint. In particular, a geodesic Gaussian kernel is positive definite only if the underlying manifold is euclidean. Moreover, a geodesic Laplacian kernel is positive definite if and only if conditionally negative definite conditions are satisfied, which is in general not true for curved Riemannian manifolds. These results suggest that the application of kernel methods in curved manifolds has its limitation. On the other hand, it is also of interest for the community to investigate on approaches other than kernel based methods. This motivates us to consider in this work a manifold-valued geodesic regression approach by directly considering the intrinsic Riemannian metric.

2 Background

We first briefly review some concepts in Riemannian manifolds in Section 2.1, nonsmooth analysis and Kurdyka–Łojasiewicz property on Riemannian manifolds in Sections 2.2 and 2.3, respectively, which are necessary for the derivation of

1. Our implementation is available at the project website http://web.bii.a-star.edu.sg/~zhangxw/palmr-SPD/.

our algorithm and the proof of convergence. We then review some models regarding multivariate linear regression with gross errors in euclidean space, whose ideas are utilized to design our new model.

2.1 Riemannian Manifolds

Let $(\mathcal{M}, \varrho)$ denote a smooth manifold $\mathcal{M}$ endowed with a Riemannian metric ϱ. Moreover, $T_{\boldsymbol{p}}\mathcal{M}$ denotes the tangent space at point $\boldsymbol{p}$ and $T\mathcal{M} := \cup_{\boldsymbol{p}\in\mathcal{M}} T_{\boldsymbol{p}}\mathcal{M}$ denotes the tangent bundle. Notation $(\boldsymbol{p}, \boldsymbol{v}) \in \mathcal{M} \times T\mathcal{M}$ refers to $\boldsymbol{p}$ being a point of $\mathcal{M}$ and $\boldsymbol{v}$ being a tangent vector at $\boldsymbol{p}$. $\langle \boldsymbol{u},\ \boldsymbol{v}\rangle_{\boldsymbol{p}} := \varrho_{\boldsymbol{p}}(\boldsymbol{u}, \boldsymbol{v})$ is the inner product between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in $T_{\boldsymbol{p}}\mathcal{M}$, with $\varrho_{\boldsymbol{p}}$ being the metric at $\boldsymbol{p}$. The induced norm thus becomes $\|\boldsymbol{u}\|_{\boldsymbol{p}} := \langle \boldsymbol{u},\ \boldsymbol{u}\rangle_{\boldsymbol{p}}^{1/2}$. Let $\gamma : [a, b] \to \mathcal{M}$ be a piecewise smooth curve such that $\gamma(a) = \boldsymbol{p}$ and $\gamma(b) = \boldsymbol{q}$, with the curve length as $\int_a^b \|\gamma'(t)\|_{\gamma(t)} dt$ where $\gamma'(t)$ denotes derivative. The Riemannian distance $d_{\mathcal{M}}(\boldsymbol{p}, \boldsymbol{q})$ between $\boldsymbol{p}$ and $\boldsymbol{q}$ is defined as the infimum of the length over all piecewise smooth curves joining these two points. Let ∇ be the Levi-Civita connection[2] associated with $(\mathcal{M}, \varrho)$. Curve γ is called a *geodesic* if $\nabla_{\gamma'}\gamma' = 0$. A Riemannian manifold is *complete* if its geodesics $\gamma(t)$ are defined for any value of $t \in \mathbb{R}$. The *parallel transport* along γ from $\boldsymbol{p} = \gamma(a)$ to $\boldsymbol{q} = \gamma(b)$ is a mapping $P_{\gamma(a)\gamma(b)} : T_{\boldsymbol{p}}\mathcal{M} \to T_{\boldsymbol{q}}\mathcal{M}$ defined by $P_{\gamma(a)\gamma(b)}(\boldsymbol{v}) = V(b)$, where V is the unique vector field satisfying $\nabla_{\gamma'}V = 0$ and $V(a) = \boldsymbol{v}$. The exponential map at point $\boldsymbol{p}$ is a mapping $\mathrm{Exp}_{\boldsymbol{p}} : T_{\boldsymbol{p}}\mathcal{M} \to \mathcal{M}$ defined as $\mathrm{Exp}_{\boldsymbol{p}}(\boldsymbol{v}) = \gamma(1)$, where $\gamma : [0, 1] \to \mathcal{M}$ is the geodesic such that $\gamma(0) = \boldsymbol{p}$ and $\gamma'(0) = \boldsymbol{v}$. The inverse of the exponential map, if exists, is denoted by $\mathrm{Exp}_{\boldsymbol{p}}^{-1}$. To simplify notations, we also use $\langle,\ \rangle$, $\|\cdot\|$, $d(\cdot, \cdot)$, and $\mathrm{Exp}(\boldsymbol{p}, \boldsymbol{v})$ to denote inner product, norm, Riemannian distance, and exponential map respectively, when there is no confusion. We focus on *Hadamard manifold* $\mathcal{M}$, which is a complete and simply connected finite dimensional Riemannian manifold with nonpositive sectional curvature. The class of Hadamard manifolds possesses many nice properties: For example, any two points in $\mathcal{M}$ can be joined by a *unique* geodesic. In this case, the exponential map is a global diffeomorphism and $d(\boldsymbol{p}, \boldsymbol{q}) = \|\mathrm{Exp}_{\boldsymbol{p}}^{-1}\boldsymbol{q}\|_{\boldsymbol{p}}$. One example of Hadamard manifold is the manifold of symmetric positive definite matrices. Motivated readers can consult [45] for further details of manifolds and differential geometry.

2.2 Nonsmooth Analysis on Riemannian Manifolds

Given an extended real-valued function $\sigma : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ we define its domain by $\mathrm{dom}\ \sigma := \{\boldsymbol{p} \in \mathcal{M} : \sigma(\boldsymbol{p}) < +\infty\}$ and its epigraph by $\mathrm{epi}\sigma := \{(\boldsymbol{p}, \beta) \in \mathcal{M} \times \mathbb{R} : \sigma(\boldsymbol{p}) \le \beta\}$. We say that σ is a lower semicontinuous function if $\mathrm{epi}\sigma$ is closed, and is proper if $\mathrm{dom}\ \sigma \neq \emptyset$ and $\sigma(\boldsymbol{p}) > -\infty$ for all $\boldsymbol{p} \in \mathrm{dom}\ \sigma$. Proper and lower semicontinuous (PLS) functions play important roles in optimization, since it guarantees the well-definedness of the proximal operator. In particular, given $\boldsymbol{p}$ and $\lambda > 0$, the proximal map defined as

$$\mathrm{prox}_{\lambda}^{\sigma}(\boldsymbol{p}) := \underset{z}{\mathrm{argmin}}\left\{\sigma(z) + \frac{\lambda}{2}\|z - \boldsymbol{p}\|\right\},$$

is well-defined when σ is PLS and $\inf \sigma(z) > 0$. In Section 3, we will see that the objective function in our approach is a PLS. Moreover, we have the following definition of (sub)differential of PLS functions on manifolds.

Definition 1 ([46]). *Let σ be a PLS function, then*

- *the Fréchet subdifferential of σ at any $\boldsymbol{p} \in \mathrm{dom}\ \sigma$, denoted as $\hat{\partial}\sigma(\boldsymbol{p})$, is defined as the set of all $\boldsymbol{v} \in T_{\boldsymbol{p}}\mathcal{M}$ which satisfies*

$$\liminf_{\boldsymbol{q}\neq\boldsymbol{p}\ \boldsymbol{q}\to\boldsymbol{p}} \frac{\sigma(\boldsymbol{q}) - \sigma(\boldsymbol{p}) - \langle \boldsymbol{v},\ \gamma'(0)\rangle}{d(\boldsymbol{p}, \boldsymbol{q})} \ge 0,$$

  for geodesic γ joining $\gamma(0) = \boldsymbol{p}$ and $\gamma(1) = \boldsymbol{q}$. When $\boldsymbol{p} \notin \mathrm{dom}\ \sigma$, we set $\hat{\partial}\sigma(\boldsymbol{p}) = \emptyset$.
- *the (limiting) subdifferential of σ at any $\boldsymbol{p} \in \mathcal{M}$, denoted as $\partial\sigma(\boldsymbol{p})$, is defined as*

$$\begin{aligned}\partial\sigma(\boldsymbol{p}) = \{\boldsymbol{v} \in T_{\boldsymbol{p}}\mathcal{M} : \exists(\boldsymbol{p}^k, \sigma(\boldsymbol{p}^k)) \to (\boldsymbol{p}, \sigma(\boldsymbol{p})),\\ \exists \boldsymbol{v}^k \in \hat{\partial}\sigma(\boldsymbol{p}^k)\ \text{s.t.}\ P_{\gamma^k(0)\gamma^k(1)}(\boldsymbol{v}^k) \to \boldsymbol{v}\},\end{aligned}$$

  where γ^k is the geodesic joining $\boldsymbol{p}^k$ and $\boldsymbol{p}$.
- *$\boldsymbol{p} \in \mathcal{M}$ is a critical point of σ if $0 \in \partial\sigma(\boldsymbol{p})$. We denote the set of critical points of σ by $\mathrm{crit}\ \sigma$. That is*

$$\mathrm{crit}\ \sigma = \{\boldsymbol{x} \in \mathcal{M} :\ 0 \in \partial\sigma(\boldsymbol{x})\}.$$

If $\boldsymbol{p}$ is a local minimizer of σ then by the Fermat's rule $0 \in \partial\sigma(\boldsymbol{p})$. If σ is differentiable, then its subdifferential reduces to a unique gradient, denoted as $\mathrm{grad}\sigma$, which is a vector field satisfying $\langle \mathrm{grad}\sigma(\boldsymbol{p}),\ \boldsymbol{v}\rangle_{\boldsymbol{p}} = \boldsymbol{v}(\sigma)$ for all $\boldsymbol{v} \in T_{\boldsymbol{p}}\mathcal{M}$ and $\boldsymbol{p} \in \mathcal{M}$. Here $\boldsymbol{v}(\sigma)$ denotes the directional derivative of σ in the direction $\boldsymbol{v}$. In this case $\partial\sigma(\boldsymbol{p}) = \{\mathrm{grad}\sigma(\boldsymbol{p})\}$. Moreover, we have the following definition of Lipschitz gradients for smooth functions on manifolds:

Definition 2 ([47]). *Let $\sigma : \mathcal{M} \to \mathbb{R}$ be a continuously differentiable function and $L > 0$. σ is said to have L-Lipschitz gradient if, for any $\boldsymbol{p},\ \boldsymbol{q} \in \mathcal{M}$ and any geodesic segment $\gamma : [0, r] \to \mathcal{M}$ joining $\boldsymbol{p}$ and $\boldsymbol{q}$, then*

$$\|\partial\sigma(\gamma(t)) - P_{\gamma(0)\gamma(t)}\partial\sigma(\boldsymbol{p})\|_{\gamma(t)} \le Ll(t),\ \forall t \in [0, r],$$

where $\gamma(0) = \boldsymbol{p}$, $P_{\gamma(0)\gamma(t)}$ is the parallel transport along γ from $\boldsymbol{p}$ to $\gamma(t)$, and $l(t)$ denotes the length of the segment between $\boldsymbol{p}$ and $\gamma(t)$. In addition, if $\mathcal{M}$ is a Hadamard manifold, then the last inequality becomes

$$\|\partial\sigma(\gamma(t)) - P_{\gamma(0)\gamma(t)}\partial\sigma(\boldsymbol{p})\|_{\gamma(t)} \le Ld(\boldsymbol{p}, \gamma(t)).$$

Since σ is continuously differentiable, $\partial\sigma(\gamma(t))$ and $\partial\sigma(\boldsymbol{p})$ are the unique tangent vectors at $\gamma(t)$ and $\boldsymbol{p}$, respectively. Parallel transport thus becomes necessary to move them onto the same tangent space. Note in general, the right hand sides of the two inequalities above are different. This is due to the fact that for non-Hadamard manifolds, geodesic between two points is usually not unique. Since $d(\boldsymbol{p}, \gamma(t))$ is defined as the infimum length of geodesic segments between $\boldsymbol{p}$ and $\gamma(t)$, it could be smaller than $l(t)$, which is

2. Roughly speaking, a connection acts as a generalization of directional derivative that connects tangent spaces of nearby points and provides a consistent manner of transporting tangent vectors from one point to another along geodesic curves. A manifold may have many connections. Levi-Civita connection, also called Riemannian connection, is a unique connection that is symmetric and compatible with the Riemannian metric.

the length of the segment between $\boldsymbol{p}$ and $\gamma(t)$ along a given geodesic γ. For Hadamard manifolds on the other hand, there exists a unique geodesic between any two points, hence $d(\boldsymbol{p}, \gamma(t)) = l(t)$ always holds.

2.3 Kurdyka–Łojasiewicz (K-L) Property on Riemannian Manifolds

The Kurdyka–Łojasiewicz (K-L) property plays a crucial role in nonsmooth analysis [48], [49]. In this section we extend the K-L property from euclidean spaces to Riemannian manifolds. To do this, we need to introduce some basic notations. If A is a subset of $\mathcal{M}$, then the distance between $\boldsymbol{x} \in \mathcal{M}$ and A is defined by

$$\operatorname{dist}(\boldsymbol{x}, A) := \inf\{d(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{y} \in A\},$$

where A is nonempty, and $\operatorname{dist}(\boldsymbol{x}, A) = +\infty$ for all $\boldsymbol{x} \in \mathcal{M}$ when A is empty. For a fixed $\boldsymbol{x} \in \mathcal{M}$, the open ball neighborhood of $\boldsymbol{x}$ with radius η is defined as $B(\boldsymbol{x}, \eta) := \{\boldsymbol{y} \in \mathcal{M} : d(\boldsymbol{x}, \boldsymbol{y}) < \eta\}$.

Definition 3. *Given real scalars α, β, and PLS function σ, we define*

$$[\alpha \leq \sigma \leq \beta] := \{\boldsymbol{x} \in \mathcal{M} : \alpha \leq \sigma(\boldsymbol{x}) \leq \beta\}.$$

We define similarly $[\alpha < \sigma < \beta]$.

Now, we define the K-L property.

Definition 4 ([49]). *Let $\sigma : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ be a PLS function. The function σ is said to have K-L property at $\bar{\boldsymbol{x}} \in \operatorname{dom} \sigma$ if there exists $\eta \in (0, \infty]$, a neighborhood U of $\bar{\boldsymbol{x}}$ and a continuous concave function $\phi : [0, \eta) \to \mathbb{R}_+$ such that*

(i) $\phi(0) = 0$, ϕ is continuously differentiable on $(0, \eta)$ and $\phi'(s) > 0$ for all $s \in (0, \eta)$;
(ii) the following K-L inequality holds

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

$$\forall \boldsymbol{x} \in U \cap [\sigma(\bar{\boldsymbol{x}}) < \sigma < \sigma(\bar{\boldsymbol{x}}) + \eta].$$

We call σ a K-L function if it has K-L property at each point of $\operatorname{dom} \sigma$.

The K-L property basically asserts that function σ can be made sharp by a reparameterization of its values using ϕ. In particular, when σ is differentiable and $\bar{\boldsymbol{x}}$ is critical, i.e., $\partial\sigma(\bar{\boldsymbol{x}}) = 0$, we can define reparameterization $f(\boldsymbol{x}) := \phi(\sigma(\boldsymbol{x}) - \sigma(\bar{\boldsymbol{x}}))$, then the K-L inequality becomes $\|\partial f(\boldsymbol{x})\| \geq 1$, which avoids flatness around $\bar{\boldsymbol{x}}$. This geometrical feature plays a critical role in proving that the sequence generated by our algorithm converges to a critical point. In Proposition 4 of the supplementary, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2017.2776260, we also establish K-L property in the neighborhood of non-critical points. K-L functions are ubiquitous in a wide range of applications, including for example semi-algebraic, subanalytic, semiconvex, uniformly convex, and log-exp functions [48], [49].

2.4 Multivariate Linear Regression with Gross Errors

Given a matrix representation of N observations $X \in \mathbb{R}^{N \times d}$, and the corresponding m-dimensional response matrix $Y \in \mathbb{R}^{N \times m}$, one of the central problems in linear regression is to accurately estimate the regression matrix $V \in \mathbb{R}^{d \times m}$ from

$$Y = XV^* + Z, \tag{1}$$

with $Z \in \mathbb{R}^{N \times m}$ being the stochastic noise. In most of existing work regarding linear regression, Z is assumed to be composed of entries following normal distribution with zero mean. However, when the response Y is subject to possible gross error, the estimated regression matrix deviates significantly from the true value and becomes unreliable. To deal with this problem, several recent works [12], [13], [14] suggest to consider model

$$Y = XV^* + G^* + Z, \tag{2}$$

where $G^* \in \mathbb{R}^{N \times m}$ is used to explicitly characterize the gross error component. As in practice only a subset of responses are corrupted by gross error, G^* is a sparse matrix whose nonzero entries are unknown and magnitudes can be arbitrarily large. Moreover, this model can as well be applied to deal with the case where some entries of Y are missing. A commonly used paradigm of estimating (V^*, G^*) is by solving convex optimization problem

$$\min_{V,G} \frac{1}{2}\|Y - XV - G\|_F^2 + \lambda R_v(V) + \rho R_g(G), \tag{3}$$

where $\lambda > 0$ and $\rho > 0$ are tuning parameters, and R_v and R_g are regularization terms of V and G, respectively. Some frequently used regularization norms include ℓ_1 norm $\|\cdot\|_1$ which is the summation of the absolute value of all entries, and $\ell_{1,2}$ norm which is the summation of ℓ_2 norm of rows of a matrix. For example, in [14] the authors propose to use $R_v(V) = \|V\|_{1,2}$ and $R_g(G) = \|G\|_1$.

3 Our Approach

Consider a set of training examples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$, where $\boldsymbol{y}_i$ lies on Riemannian manifold $\mathcal{M}$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ is the associated independent variable. We propose a novel extension of the modeling approach of Eq. (2) for euclidean spaces to deal with the more general curved spaces, as follows.

3.1 From Euclidean Spaces to Manifolds

The Model of Eq. (2) can be reformulated as $Y - G^* = XV^* + Z$. Denote $Y^c := Y - G^*$, which can be interpreted as *corrected* response after removing the gross error. Now the model of Eq. (2) can be reformulated as standard linear regression in Eq. (1) with response Y^c. With this in mind, we proceed to extend the aforementioned idea to regression on manifolds. For each manifold-valued response $\boldsymbol{y}_i$, denote as $\boldsymbol{y}_i^c$ its corrected version. Different from the euclidean space setting where Y^c can be obtained from Y simply by a translation, we need to ensure that $\boldsymbol{y}_i^c$ remains on the manifold. This is accomplished by the exponential map $\boldsymbol{y}_i^c = \operatorname{Exp}_{\boldsymbol{y}_i}(\boldsymbol{g}_i)$ with the gross error $\boldsymbol{g}_i \in T_{\boldsymbol{y}_i}\mathcal{M}$ over each of the training examples, $i \in \{1, \ldots, N\}$. Note that when $\mathcal{M}$ is an euclidean space, the exponential map reduces to addition, as $\operatorname{Exp}_{\boldsymbol{y}_i}(\boldsymbol{g}_i) = \boldsymbol{y}_i + \boldsymbol{g}_i$. In other words, translation in the affine space is a special case of exponential map in the more general curved space.

As illustrated in Fig. 1, we first obtain the corrected manifold-valued response $\boldsymbol{y}_i^c = \mathrm{Exp}_{\boldsymbol{y}_i}(\boldsymbol{g}_i)$. Then the relationship between $\boldsymbol{x}_i$ and $\boldsymbol{y}_i^c$ can be modeled as

$$\mathrm{Exp}_{\boldsymbol{y}_i}(\boldsymbol{g}_i) = \mathrm{Exp}\left(\mathrm{Exp}\left(\boldsymbol{p}, \sum_{j=1}^{d} x_i^j \boldsymbol{v}_j\right),\ z_i\right), \tag{4}$$

where $\boldsymbol{p} \in \mathcal{M}$ and $\{\boldsymbol{v}_j\}_{j=1}^d \in T_p\mathcal{M}$ is a set of tangent vectors at $\boldsymbol{p}$, x_i^j is the jth component of $\boldsymbol{x}_i$, and z_i is a tangent vector at $\mathrm{Exp}(\boldsymbol{p}, \sum_{j=1}^d x_i^j \boldsymbol{v}_j)$. Our model can be viewed as a generalization of linear regression model of Eq. (1) from flat spaces to manifolds, where $\boldsymbol{p}$ denotes the intercept that is in analogy to the origin 0 in the flat space as in Eq. (1), and exponential map corresponds to the addition operator in Eq. (1).

To measure the training loss, we use

$$E\big(\boldsymbol{p}, \{\boldsymbol{v}_j\}, \{\boldsymbol{g}_i\}\big) := \frac{1}{2}\sum_i d^2\left(\mathrm{Exp}_{\boldsymbol{y}_i}(\boldsymbol{g}_i), \mathrm{Exp}_{\boldsymbol{p}}\left(\sum_j x_i^j \boldsymbol{v}_j\right)\right),$$

to denote the sum-of-squared Riemannian distance between the corrected data $\boldsymbol{y}_i^c = \mathrm{Exp}_{\boldsymbol{y}_i}(\boldsymbol{g}_i)$ and the prediction $\hat{\boldsymbol{y}}_i = \mathrm{Exp}_{\boldsymbol{p}}(\sum_j x_i^j \boldsymbol{v}_j)$, and let R_v and R_g denote two regularization terms controlling the magnitude of $\{\boldsymbol{v}_j\}$ and $\{\boldsymbol{g}_i\}$, respectively. The problem considered in our paper can now be formulated as the following optimization problem

$$\begin{aligned}(\tilde{\boldsymbol{p}}, \{\tilde{\boldsymbol{v}}_j\}, \{\tilde{\boldsymbol{g}}\}_i) = \underset{\substack{(\boldsymbol{p},\{\boldsymbol{v}_j\})\in\mathcal{M}\times T\mathcal{M}\\ \{\boldsymbol{g}_i\}\in T_{\boldsymbol{y}_i}\mathcal{M}}}{\operatorname{argmin}} & E\big(\boldsymbol{p}, \{\boldsymbol{v}_j\}, \{\boldsymbol{g}_i\}\big)\\ & + \lambda R_v\big(\{\boldsymbol{v}_j\}\big) + \rho R_g(\{\boldsymbol{g}_i\}),\end{aligned} \tag{5}$$

where $\lambda \geq 0$ and $\rho \geq 0$ are regularization parameters. Without loss of generality, we consider regularization terms $R_v\big(\{\boldsymbol{v}_j\}\big) := \sum_{j=1}^d \big\|\boldsymbol{v}_j\big\|_{\boldsymbol{p}}$ and $R_g(\{\boldsymbol{g}_i\}) := \sum_{i=1}^N \|\boldsymbol{g}_i\|_{\boldsymbol{y}_i}$, with $\|\cdot\|_{\boldsymbol{p}}$ and $\|\cdot\|_{\boldsymbol{y}_i}$ being the norm of tangent vectors at $\boldsymbol{p}$ and $\boldsymbol{y}_i$, respectively. There are two reasons for the choice of R_v: First, it enables problem of Eq. (5) to contain the multivariate linear regression problems with feature selection in euclidean spaces as special cases, as shown in Examples 1 and 2 below; Second, in many applications one may collect a large set of possible variables $\{\boldsymbol{x}^j\}$ for each response, and want to find a compact subset of base tangent vectors from $\{\boldsymbol{v}_j\}$ and the corresponding $\{\boldsymbol{x}^j\}$ that are significant to the manifold-valued output $\boldsymbol{y}$. The choice of R_g is based on the assumption that gross errors are usually sporadically spread among data. Now, the optimization problem becomes

$$\min_{\substack{(\boldsymbol{p},\{\boldsymbol{v}_j\})\in\mathcal{M}\times T\mathcal{M}\\ \boldsymbol{g}_i\in T_{\boldsymbol{y}_i}\mathcal{M}}} E\big(\boldsymbol{p}, \{\boldsymbol{v}_j\}, \{\boldsymbol{g}_i\}\big) + \lambda\sum_{j=1}^{d}\big\|\boldsymbol{v}_j\big\|_{\boldsymbol{p}} + \rho\sum_{i=1}^{N}\|\boldsymbol{g}_i\|_{\boldsymbol{y}_i}. \tag{6}$$

3.2 Connections to Existing Works

We would like to point out that model in Eq. (6) includes as special cases a number of related research works on gross error or on manifold-valued regression. In this section, we provide three such examples.

Example 1. When $\mathcal{M} = \mathbb{R}^m$, we can establish a connection between the model of Eq. (6) and the robust multi-task regression studied in [14]. Specifically, instead of optimizing Eq. (6) over $\boldsymbol{p} \in \mathbb{R}^m$ we select $\boldsymbol{p} = 0$, resulting in

$$\min_{\boldsymbol{v}_j,\boldsymbol{g}_i\in\mathbb{R}^m} \frac{1}{2}\sum_{i=1}^{N}\left\|\boldsymbol{y}_i - \sum_{j=1}^{d}\boldsymbol{x}_i^j\boldsymbol{v}_j + \boldsymbol{g}_i\right\|^2 + \lambda\sum_{j=1}^{d}\big\|\boldsymbol{v}_j\big\| + \rho\sum_{i=1}^{N}\|\boldsymbol{g}_i\|,$$

which can be rewritten as

$$\min_{V,G} \frac{1}{2}\|Y - XV - G\|_F^2 + \lambda\|V\|_{1,2} + \rho\|G\|_{1,2}, \tag{7}$$

where $\|\cdot\|$ becomes the usual euclidean norm, $Y = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]^\top \in \mathbb{R}^{N\times m}$, $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N\times d}$, $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d]^\top \in \mathbb{R}^{d\times m}$ and $G = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_N]^\top \in \mathbb{R}^{N\times m}$. The resulting model of Eq. (7) is exactly the one considered in [14] except that regularization term $\|G\|_1$ in [14] is replaced by $\|G\|_{1,2}$ here.

Example 2. If $\mathcal{M} = \mathbb{R}^m$, we can show by Fermat's rule that the optimal solution $\tilde{\boldsymbol{p}}$ is given by $\tilde{\boldsymbol{p}} = \frac{1}{N}\sum_{i=1}^N(\boldsymbol{y}_i + \boldsymbol{g}_i - \sum_j \boldsymbol{x}_i^j\boldsymbol{v}_j)$. By substituting $\tilde{\boldsymbol{p}}$ into problem of Eq. (6) and assuming $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ has empirical mean 0, that is, $\sum_{i=1}^N \boldsymbol{x}_i = 0$ and $\sum_{i=1}^N \boldsymbol{y}_i = 0$, the optimization problem of Eq. (6) reduces to

$$\begin{aligned}\min_{\boldsymbol{v}_j,\boldsymbol{g}_i\in\mathbb{R}^m} & \frac{1}{2}\sum_{i=1}^{N}\left\|\boldsymbol{y}_i - \sum_{j=1}^{d}\boldsymbol{x}_i^j\boldsymbol{v}_j + \boldsymbol{g}_i - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{g}_i\right\|^2\\ & + \lambda\sum_{j=1}^{d}\big\|\boldsymbol{v}_j\big\| + \rho\sum_{i=1}^{N}\|\boldsymbol{g}_i\|,\end{aligned}$$

which can be reformulated as

$$\begin{aligned}&\min_{V,G} \frac{1}{2}\|Y - XV - \bar{G}\|_F^2 + \lambda\|V\|_{1,2} + \rho\|G\|_{1,2}\\ &s.t.\ \bar{G} = \left(I - \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^\top\right)G,\end{aligned} \tag{8}$$

where $\mathbb{1}_N \in \mathbb{R}^N$ is a column vector with all entries being 1.

The difference between Examples 1 and 2 lies in that the former is obtained from selecting $\boldsymbol{p} = 0$ while the latter is from optimizing $\boldsymbol{p}$ which exactly follows model of Eq. (6). The resulting models are quite similar except that model of Eq. (8) needs to center variable G.

Example 3. If we let $\lambda = 0$ and $\rho = +\infty$, then optimization problem of Eq. (6) reduces to

$$\min_{(\boldsymbol{p},\{\boldsymbol{v}_j\})\in\mathcal{M}\times T\mathcal{M}} \frac{1}{2}\sum_{i=1}^{N} d^2\left(\boldsymbol{y}_i, \mathrm{Exp}_{\boldsymbol{p}}\Big(\sum_{j=1}^{d} x_i^j\boldsymbol{v}_j\Big)\right),$$

which recovers exactly the model considered in [2]. In this regard, the MGLM model in [2] is a special case of our model.

3.3 PALM for Optimization on Hadamard Manifolds

In this section, we propose a new algorithm to solve optimization problem of Eq. (6), which is actually a nonsmooth optimization problem on Hadamard manifolds. As explained in details in Section 3.4, problem of Eq. (6) admits the form

$$\min_{\boldsymbol{x}\in\mathcal{M}_1,\boldsymbol{y}\in\mathcal{M}_2} \Psi(\boldsymbol{x}, \boldsymbol{y}) := f(\boldsymbol{x}) + g(\boldsymbol{y}) + h(\boldsymbol{x}, \boldsymbol{y}), \tag{9}$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are Hadamard manifolds, $f : \mathcal{M}_1 \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathcal{M}_2 \to \mathbb{R} \cup \{+\infty\}$ are PLS functions, and $h : \mathcal{M}_1 \times \mathcal{M}_2 \to \mathbb{R}$ is a smooth function.

Many existing optimization techniques are developed to work with euclidean spaces, thus not directly applicable to curved manifolds. Meanwhile, an increasing amount of attention has been drawn to the field of optimization on manifolds [50]. For smooth optimization, classical optimization techniques, such as gradient, conjugate gradients, and trust-region methods, have been generalized to the manifold setting [50], [51], [52], [53], which are however not suitable for the nonconvex and nonsmooth optimization manifold-based problem of Eq. (6). For nonsmooth optimization, there exist many prior works [54], [55], [56], [57]. Unfortunately they either cannot exploit the composition structure in Eq. (9) (e.g., [54], [55], [57]), or fail to guarantee convergence (e.g., [56]).

Recently, a proximal alternating linearized minimization algorithm has been proposed in [16] for optimization problem of Eq. (9) with $\mathcal{M}_1 = \mathbb{R}^n$ and $\mathcal{M}_2 = \mathbb{R}^m$. Inspired by the success of PALM in the euclidean setting, in what follows we propose PALMR, an inexact proximal alternating minimization algorithm for problem of Eq. (9).

We alternately solve the following two proximally linearized subproblems

$$\begin{aligned}\boldsymbol{x}^{k+1} \in \underset{\boldsymbol{x}\in\mathcal{M}_1}{\operatorname{argmin}}\ & f(\boldsymbol{x}) + \left\langle \operatorname{Exp}_{\boldsymbol{x}^k}^{-1}\boldsymbol{x},\ \partial_{\boldsymbol{x}} h(\boldsymbol{x}^k, \boldsymbol{y}^k)\right\rangle \\ & + \frac{c_k}{2} d^2_{\mathcal{M}_1}(\boldsymbol{x}^k, \boldsymbol{x}),\end{aligned} \tag{10}$$

$$\begin{aligned}\boldsymbol{y}^{k+1} \in \underset{\boldsymbol{y}\in\mathcal{M}_2}{\operatorname{argmin}}\ & g(\boldsymbol{y}) + \left\langle \operatorname{Exp}_{\boldsymbol{y}^k}^{-1}\boldsymbol{y},\ \partial_{\boldsymbol{y}} h(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k)\right\rangle \\ & + \frac{d_k}{2} d^2_{\mathcal{M}_2}(\boldsymbol{y}^k, \boldsymbol{y}),\end{aligned} \tag{11}$$

where $c_k = \mu_1 L_1(\boldsymbol{y}^k)$ and $d_k = \mu_2 L_2(\boldsymbol{x}^{k+1})$ with $\mu_1 > 1$, $\mu_2 > 1$ and $L_1(\boldsymbol{y}^k)$, $L_2(\boldsymbol{x}^{k+1})$ being the Lipschitz constants of $\partial_{\boldsymbol{x}} h$ and $\partial_{\boldsymbol{y}} h$, respectively, as to be explained in Assumption 1. In particular, by exploiting the fact that $\mathcal{M}_1$ is a Hadamard manifold on which any two points can be joined by a unique geodesic, we have a one-to-one mapping between $\boldsymbol{v} \in T_{\boldsymbol{x}^k}\mathcal{M}_1$ and $\boldsymbol{x} \in \mathcal{M}_1$ such that $\boldsymbol{x} = \operatorname{Exp}_{\boldsymbol{x}^k}(\boldsymbol{v})$, $\boldsymbol{v} = \operatorname{Exp}^{-1}_{\boldsymbol{x}^k}\boldsymbol{x}$ and $d_{\mathcal{M}_1}(\boldsymbol{x}^k, \boldsymbol{x}) = \|\boldsymbol{v}\|$. Thus, a simple substitution reformulates Eq. (10) as

$$\boldsymbol{v}^k \in \underset{\boldsymbol{v}\in T_{\boldsymbol{x}^k}\mathcal{M}_1}{\operatorname{argmin}} (f \circ \operatorname{Exp}_{\boldsymbol{x}^k})(\boldsymbol{v}) + \left\langle \boldsymbol{v},\ \partial_{\boldsymbol{x}} h(\boldsymbol{x}^k, \boldsymbol{y}^k)\right\rangle + \frac{c_k}{2}\|\boldsymbol{v}\|^2,$$

or equivalently

$$\boldsymbol{v}^k \in \underset{\boldsymbol{v}\in T_{\boldsymbol{x}^k}\mathcal{M}_1}{\operatorname{argmin}} (f \circ \operatorname{Exp}_{\boldsymbol{x}^k})(\boldsymbol{v}) + \frac{c_k}{2}\|\boldsymbol{v} + \frac{1}{c_k}\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k, \boldsymbol{y}^k)\|^2,$$

which becomes an optimization problem in linear space $T_{\boldsymbol{x}^k}\mathcal{M}_1$, and as a result, we have $\boldsymbol{x}^{k+1} = \operatorname{Exp}_{\boldsymbol{x}^k}(\boldsymbol{v}^k)$. Since f is PLS satisfying $\inf_{\boldsymbol{x}\in\mathcal{M}} f(\boldsymbol{x}) > -\infty$ and $\operatorname{Exp}_{\boldsymbol{x}^k}$ is smooth, it follows that the composite function $f \circ \operatorname{Exp}_{\boldsymbol{x}^k}$ is PLS and $\inf_{\boldsymbol{v}\in T_{\boldsymbol{x}^k}\mathcal{M}} f \circ \operatorname{Exp}_{\boldsymbol{x}^k}(\boldsymbol{v}) > -\infty$ which, together with Theorem 1.25 of [58], implies that $\boldsymbol{v}^k$ is well-defined. Moreover, the above optimization problem for $\boldsymbol{v}^k$ is called proximity operator [59], denoted as

$$\boldsymbol{v}^k = \operatorname{prox}_{c_k}^{f\circ\operatorname{Exp}_{\boldsymbol{x}^k}}\left(-\frac{1}{c_k}\partial_{\boldsymbol{x}} h(\boldsymbol{x}^k, \boldsymbol{y}^k)\right).$$

Similar claims apply to problem of Eq. (11), implying the well-definiteness of $\boldsymbol{x}^{k+1}$ and $\boldsymbol{y}^{k+1}$. Solving Eqs. (10) and (11) alternately yields the algorithm PALMR outlined in Algorithm 1.

Algorithm 1. (PALMR): PALM on Riemannian Manifolds

Input: $\mu_1 > 1$ and $\mu_2 > 1$.
Output: the sequence $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$.
1: Initialization: $(\boldsymbol{x}^0, \boldsymbol{y}^0)$ and $k = 0$.
2: **while** stopping criterion not satisfied **do**
3: Set $c_k = \mu_1 L_1(\boldsymbol{y}^k)$ and compute $\boldsymbol{x}^{k+1}$ as in Eq. (10).
4: Set $d_k = \mu_2 L_2(\boldsymbol{x}^{k+1})$ and compute $\boldsymbol{y}^{k+1}$ as in Eq. (11).
5: **end while**

To analyze the convergence of PALMR, we need the following assumptions.

Assumption 1. *$\Psi(\boldsymbol{x}, \boldsymbol{y})$ satisfies the following conditions:*

(i) $\inf f > -\infty$, $\inf g > -\infty$ *and* $\inf \Psi > -\infty$.

(ii) *For any fixed $\boldsymbol{y}$, the function $\boldsymbol{x} \to h(\boldsymbol{x}, \boldsymbol{y})$ has $L_1(\boldsymbol{y})$-Lipschitz gradient. Likewise, for any fixed $\boldsymbol{x}$, the function $\boldsymbol{y} \to h(\boldsymbol{x}, \boldsymbol{y})$ has $L_2(\boldsymbol{x})$-Lipschitz gradient. Moreover, there exist real scalars $\lambda_i^-, \lambda_i^+ > 0$ for $i = 1, 2$, such that*

$$\begin{aligned}&\inf_{k\in\mathbb{N}}\{L_1(\boldsymbol{y}^k)\} \geq \lambda_1^-, \quad \inf_{k\in\mathbb{N}}\{L_2(\boldsymbol{x}^k)\} \geq \lambda_2^-,\\ &\sup_{k\in\mathbb{N}}\{L_1(\boldsymbol{y}^k)\} \leq \lambda_1^+, \quad \sup\nolimits_{k\in\mathbb{N}}\{L_2(\boldsymbol{x}^k)\} \geq \lambda_2^+.\end{aligned}$$

(iii) *∂h is Lipschitz continuous on bounded subset of $\mathcal{M}_1 \times \mathcal{M}_2$. More specifically, for bounded subset $A_1 \times A_2 \in \mathcal{M}_1 \times \mathcal{M}_2$, there exists constant $L > 0$ such that for all $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in A_1 \times A_2$, $i = 1, 2$, we have*

$$\begin{aligned}&\|\partial_{\boldsymbol{x}} h(\boldsymbol{x}_1, \boldsymbol{y}_1) - \partial_{\boldsymbol{x}} h(\boldsymbol{x}_1, \boldsymbol{y}_2)\| \leq L d_{\mathcal{M}_2}(\boldsymbol{y}_1, \boldsymbol{y}_2),\\ &\|\partial_{\boldsymbol{y}} h(\boldsymbol{x}_1, \boldsymbol{y}_1) - \partial_{\boldsymbol{y}} h(\boldsymbol{x}_2, \boldsymbol{y}_1)\| \leq L d_{\mathcal{M}_1}(\boldsymbol{x}_1, \boldsymbol{x}_2).\end{aligned}$$

(iv) *$\Psi(\boldsymbol{x}, \boldsymbol{y})$ has the Kurdyka–Łojasiewicz (K-L) property on Hadamard manifolds.*

Assumption (i) establishes that proximal operators in Eqs. (10) and (11) are well-defined, leading to the well-definedness of algorithm PALMR. Assumption (ii) provides that h is locally block-Lipschitz continuous, and the boundedness of Lipschitz constants are to ensure sufficient decrease of objective function value over iterations. Assumption (iii) considers the partial gradients of h being Lipschitz continuous, which would be used to derive lower bound for the iteration gap $d(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) + d(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k)$. Assumption (iv) guarantees that $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}$ form a Cauchy sequence.

Under Assumption 1 we have the following theorem, whose proof is provided in the supplementary, available online.

Theorem 1. *Suppose Assumption 1 holds. Let $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ be a sequence generated by PALMR. Then either the sequence $\{d_{\mathcal{M}_1\times\mathcal{M}_2}((\boldsymbol{x}^0, \boldsymbol{y}^0), (\boldsymbol{x}^k, \boldsymbol{y}^k))\}$ is unbounded or the following assertions hold:*

1) *The sequence $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ has finite length, i.e., $\sum_k d_{\mathcal{M}_1}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) < \infty$ and $\sum_k d_{\mathcal{M}_2}(\boldsymbol{y}^{k+1}, \boldsymbol{y}^k) < \infty$.*

2) *The sequence* $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k\in\mathbb{N}}$ *converges to a critical point* $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ *of* Ψ.

Based on Theorem 1, we know that the sequence $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}$ generated by PALMR converges to a critical point of Ψ, provided the boundedness of the sequence. As shown in [16], there are many scenarios where such assumption holds. For example, when functions f and g are convex and $h(\boldsymbol{x}, \boldsymbol{y}) = \|A\boldsymbol{x} - B\boldsymbol{y}\|$ where A and B are matrices, then the sequence $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}$ is bounded.

In what follows, we specifically investigate the dedicated realization of PALMR to solve the optimization problem of Eq. (6). To simplify the notation, the resulting algorithm is also referred to as PALMR when there is no confusion.

3.4 Applying PALMR to Optimization Problem of Eq. (6)

Optimization problem of Eq. (6) in our context can be reformulated as

$$\min_{\substack{(\boldsymbol{p},\{\boldsymbol{v}_j\})\in\mathcal{M}_1\\ \{\boldsymbol{g}_i\}\in\mathcal{M}_2}} \underbrace{E(\boldsymbol{p}, \{\boldsymbol{v}_j\}, \{\boldsymbol{g}_i\})}_{h(\boldsymbol{p},\{\boldsymbol{v}_j\},\{\boldsymbol{g}_i\})} + \underbrace{\lambda \sum_{j=1}^{d} \big\|\boldsymbol{v}_j\big\|_{\boldsymbol{p}}}_{f(\boldsymbol{p},\{\boldsymbol{v}_j\})} + \underbrace{\rho \sum_{i=1}^{N} \|\boldsymbol{g}_i\|_{\boldsymbol{y}_i}}_{g(\{\boldsymbol{g}_i\})},$$

which is of the form in Eq. (9) with $\mathcal{M}_1 = \mathcal{M} \times T\mathcal{M}$ and $\mathcal{M}_2 = T_{\boldsymbol{y}_1}\mathcal{M} \times \cdots \times T_{\boldsymbol{y}_N}\mathcal{M}$. To apply PALMR to solve problem of Eq. (6), we need to evaluate the gradients of $E(\boldsymbol{p}, \{\boldsymbol{v}_j\}, \{\boldsymbol{g}_i\})$. To simplify the notation, we further denote the prediction $\hat{\boldsymbol{y}}_i := \mathrm{Exp}_{\boldsymbol{p}}(\sum_j x_i^j \boldsymbol{v}_j)$, as well as the derivatives of the exponential map with respect to $\boldsymbol{p}$ and $\boldsymbol{v}$ as $d_{\boldsymbol{p}}\mathrm{Exp}_{\boldsymbol{p}}(v)$ and $d_{\boldsymbol{v}}\mathrm{Exp}_{\boldsymbol{p}}(v)$, respectively. Now, the partial gradient of E with respect to $\boldsymbol{p}$ amounts to

$$\partial_{\boldsymbol{p}} E = -\sum_i \left(d_{\boldsymbol{p}}\mathrm{Exp}_{\boldsymbol{p}}\left(\sum_j x_i^j \boldsymbol{v}_j\right)\right)^{\dagger} \mathrm{Exp}_{\hat{\boldsymbol{y}}_i}^{-1} \boldsymbol{y}_i^c \in T_{\boldsymbol{p}}\mathcal{M}, \quad (12)$$

where $(\cdot)^{\dagger}$ is the adjoint derivative of the exponential map [28] defined by $\big\langle \boldsymbol{\mu}, d_{\boldsymbol{p}}\mathrm{Exp}_{\boldsymbol{p}}(v)\boldsymbol{w}\big\rangle_{\mathrm{Exp}_{\boldsymbol{p}}(v)} = \big\langle \big(d_{\boldsymbol{p}}\mathrm{Exp}_{\boldsymbol{p}}(v)\big)^{\dagger}\boldsymbol{\mu}, \boldsymbol{w}\big\rangle_{\boldsymbol{p}}$ with $\boldsymbol{\mu} \in T_{\mathrm{Exp}_{\boldsymbol{p}}(v)}\mathcal{M}$, $\boldsymbol{w} \in T_{\boldsymbol{p}}\mathcal{M}$. The adjoint derivative operator maps $\mathrm{Exp}_{\hat{\boldsymbol{y}}_i}^{-1}(\boldsymbol{y}_i^c)$ from the tangent space of $\hat{\boldsymbol{y}}_i$ to the tangent space of $\boldsymbol{p}$. Thus $\partial_{\boldsymbol{p}} E \in T_{\boldsymbol{p}}\mathcal{M}$. Similarly, the partial gradient of E with respect to $\boldsymbol{v}_j$ and $\boldsymbol{g}_i$ are given by

$$\partial_{\boldsymbol{v}_j} E = -\sum_i x_i^j \left(d_{\boldsymbol{v}}\mathrm{Exp}_{\boldsymbol{p}}\left(\sum_{j'} x_i^{j'} \boldsymbol{v}_{j'}\right)\right)^{\dagger} \mathrm{Exp}_{\hat{\boldsymbol{y}}_i}^{-1} \boldsymbol{y}_i^c \in T_{\boldsymbol{p}}\mathcal{M}, \quad (13)$$

and

$$\partial_{\boldsymbol{g}_i} E = -\big(d_{\boldsymbol{v}}\mathrm{Exp}_{\boldsymbol{y}_i}(\boldsymbol{g}_i)\big)^{\dagger} \mathrm{Exp}_{\boldsymbol{y}_i^c}^{-1} \hat{\boldsymbol{y}}_i \in T_{\boldsymbol{y}_i}\mathcal{M}, \quad (14)$$

respectively.

The PALMR algorithm for problem of Eq. (6) proceeds as follows: To update $(\boldsymbol{p}, \{\boldsymbol{v}_j\})$, we let $\partial_{\boldsymbol{p}} E^k := \partial_{\boldsymbol{p}} E(\boldsymbol{p}^k, \{\boldsymbol{v}_j^k\}, \{\boldsymbol{g}_i^k\})$ and $\partial_{\boldsymbol{v}_j} E^k := \partial_{\boldsymbol{v}_j} E(\boldsymbol{p}^k, \{\boldsymbol{v}_j^k\}, \{\boldsymbol{g}_i^k\})$ and solve

$$\begin{aligned}(\boldsymbol{p}^{k+1}, \{\boldsymbol{v}_j^{k+1}\}) = \underset{\boldsymbol{p},\{\boldsymbol{v}_j\}}{\mathrm{argmin}} \Big\langle \mathrm{Exp}_{\boldsymbol{p}^k}^{-1}\boldsymbol{p}, \partial_{\boldsymbol{p}} E^k\Big\rangle + \frac{c_k}{2} d^2(\boldsymbol{p}, \boldsymbol{p}^k)\\ + \sum_{j=1}^{d}\Big(\Big\langle P_{\boldsymbol{p}\boldsymbol{p}^k}(\boldsymbol{v}_j) - \boldsymbol{v}_j^k, \partial_{\boldsymbol{v}_j} E^k\Big\rangle\\ + \lambda\big\|\boldsymbol{v}_j\big\|_{\boldsymbol{p}} + \frac{c_k}{2}\|P_{\boldsymbol{p}\boldsymbol{p}^k}(\boldsymbol{v}_j) - \boldsymbol{v}_j^k\|^2\Big),\end{aligned}$$

where $P_{\boldsymbol{p}\boldsymbol{p}^k}$ is the parallel transport from $\boldsymbol{p}$ to $\boldsymbol{p}^k$ along the unique geodesic between them. Due to the constraint $\boldsymbol{v} \in T_{\boldsymbol{p}}\mathcal{M}$, it is difficult to solve $\boldsymbol{p}$ and $\boldsymbol{v}_j$ together. Instead, the above subproblem is solved by alternating minimization over $\boldsymbol{p}$ and $\boldsymbol{v}_j$. Specifically, to update $\boldsymbol{p}$, we solve

$$\boldsymbol{p}^{k+1} = \arg\min_{\boldsymbol{p}\in\mathcal{M}} \Big\langle \mathrm{Exp}_{\boldsymbol{p}^k}^{-1}\boldsymbol{p}, \partial_{\boldsymbol{p}} E^k\Big\rangle + \frac{c_k}{2} d^2(\boldsymbol{p}, \boldsymbol{p}^k),$$

which, by a change of variable $\boldsymbol{u} = \mathrm{Exp}_{\boldsymbol{p}^k}^{-1}\boldsymbol{p}$, is equivalent to solving

$$\boldsymbol{u}^k = \underset{\boldsymbol{u}\in T_{\boldsymbol{p}^k}\mathcal{M}}{\mathrm{argmin}} \big\langle \boldsymbol{u}, \partial_{\boldsymbol{p}} E^k\big\rangle + \frac{c_k}{2}\|\boldsymbol{u}\|_{\boldsymbol{p}^k}^2 = -\frac{1}{c_k}\partial_{\boldsymbol{p}} E^k,$$

and $\boldsymbol{p}^{k+1} = \mathrm{Exp}_{\boldsymbol{p}^k}(\boldsymbol{u}^k)$.

To update $\{\boldsymbol{v}_j\}$, we need to first obtain $\hat{\boldsymbol{v}}_j^k$ by

$$\begin{aligned}\hat{\boldsymbol{v}}_j^k &= \underset{\boldsymbol{v}_j\in T_{\boldsymbol{p}^k}\mathcal{M}}{\mathrm{argmin}} \Big\langle \boldsymbol{v}_j - \boldsymbol{v}_j^k, \partial_{\boldsymbol{v}_j} E^k\Big\rangle + \frac{c_k}{2}\Big\|\boldsymbol{v}_j - \boldsymbol{v}_j^k\Big\|_{\boldsymbol{p}^k}^2 + \lambda\big\|\boldsymbol{v}_j\big\|_{\boldsymbol{p}^k}\\ &= \underset{\boldsymbol{v}_j\in T_{\boldsymbol{p}^k}\mathcal{M}}{\mathrm{argmin}} \frac{1}{2}\Big\|\boldsymbol{v}_j - \boldsymbol{s}_j^k\Big\|_{\boldsymbol{p}^k}^2 + \frac{\lambda}{c_k}\big\|\boldsymbol{v}_j\big\|_{\boldsymbol{p}^k},\end{aligned}$$

where $\boldsymbol{s}_j^k = \boldsymbol{v}_j^k - \frac{1}{c_k}\partial_{\boldsymbol{v}_j} E^k$. Notice that the above optimization problem have closed form solution of

$$\hat{\boldsymbol{v}}_j^k = \mathrm{prox}_{\frac{\lambda}{c_k}}^{\|\cdot\|_{\boldsymbol{p}^k}}(\boldsymbol{s}_j^k) := \left(1 - \frac{\lambda}{c_k\Big\|\boldsymbol{s}_j^k\Big\|_{\boldsymbol{p}^{k+1}}}\right)_+ \boldsymbol{s}_j^k,$$

where $(\alpha)_+ = \alpha$ if $\alpha > 0$ and 0 otherwise. Since $\{\hat{\boldsymbol{v}}_j^k\}$ lie on the tangent space at $\boldsymbol{p}^k$, we need to parallel transport them to $T_{\boldsymbol{p}^{k+1}}\mathcal{M}$ by $\boldsymbol{v}_j^{k+1} = P_{\boldsymbol{p}^k\boldsymbol{p}^{k+1}}(\hat{\boldsymbol{v}}_j^k)$ along the unique geodesic between $\boldsymbol{p}^k$ and $\boldsymbol{p}^{k+1}$.

Similarly, update $\{\boldsymbol{g}_i\}$ by

$$\begin{aligned}\boldsymbol{g}_i^{k+1} &= \arg\min_{\boldsymbol{g}_i\in T_{\boldsymbol{y}_i}\mathcal{M}} \frac{1}{2}\big\|\boldsymbol{g}_i - \boldsymbol{t}_i^k\big\|_{\boldsymbol{y}_i}^2 + \frac{\rho}{e_k}\|\boldsymbol{g}_i\|_{\boldsymbol{y}_i}\\ &= \left(1 - \frac{\rho}{e_k\big\|\boldsymbol{t}_i^k\big\|_{\boldsymbol{y}_i}}\right)_+ \boldsymbol{t}_i^k,\end{aligned}$$

where $\boldsymbol{t}_i^k = \boldsymbol{g}_i^k - \frac{1}{e_k}\partial_{\boldsymbol{g}_i} E(\boldsymbol{p}^{k+1}, \{\boldsymbol{v}_j^{k+1}\}, \{\boldsymbol{g}_i^k\})$.

Now, we are ready to present our algorithm for multivariate regression with grossly corrupted manifold-valued data, as shown in Algorithm 2. Notice that when letting $\lambda = 0$ and $\rho = +\infty$, Algorithm 2 alternately updates the values of $\boldsymbol{p}$ and $\boldsymbol{v}_j$ via three steps: (1) $\boldsymbol{p}^{k+1} = \mathrm{Exp}_{\boldsymbol{p}^k}(-\frac{1}{c_k}\partial_{\boldsymbol{p}} E^k)$, (2) $\hat{\boldsymbol{v}}_j^k = \boldsymbol{v}_j^k - \frac{1}{c_k}\partial_{\boldsymbol{v}_j} E^k$, (3) $\boldsymbol{v}_j^{k+1} = P_{\boldsymbol{p}^k\boldsymbol{p}^{k+1}}(\hat{\boldsymbol{v}}_j^k)$, which recovers the gradient descent method proposed in [2].

3.5 Implementation of Algorithm 2

During each iteration of Algorithm 2, the partial derivatives $\partial_{\boldsymbol{p}} E$, $\partial_{\boldsymbol{v}_j} E$ and $\partial_{\boldsymbol{g}_i} E$ of Eqs. (12), (13), and (14) are evaluated. Their detailed derivations are provided in Section 3 of the supplementary file, available online. Nevertheless, these terms could be practically intractable to compute for some manifolds, due to the presence of adjoint derivatives of the exponential map. As a remedy to this issue, we adopt the variational technique of [2], [60] for computing derivatives, which basically replaces the adjoint derivative operators by parallel transports

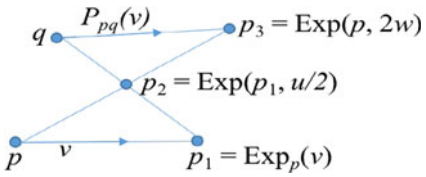


Fig. 2. An illustration of the Schild's ladder approximation of parallel transport of the tangent vector v from p to q. It consists of four steps: (1) Obtain p_1; (2) Compute tangent vector $u = \mathrm{Exp}_{p_1}^{-1}(q)$ and take half step along u to arrive at p_2; (3) Compute tangent vector $w = \mathrm{Exp}_{p}^{-1}(p_2)$ and take two steps along w to have p_3; (4) Compute tangent vector joining q and p_3 $P_{pq}(v) = \mathrm{Exp}_{q}^{-1}(p_3)$. If the distance between p and q is large, the above process can be iterated over points along the geodesic path joining p and q.

$$\partial_{\boldsymbol{p}} E \approx -\sum_i P_{\hat{\boldsymbol{y}}_i \boldsymbol{p}}(\mathrm{Exp}_{\hat{\boldsymbol{y}}_i}^{-1} \boldsymbol{y}_i^c), \tag{15}$$

$$\partial_{\boldsymbol{v}_j} E \approx -\sum_i x_i^j P_{\hat{\boldsymbol{y}}_i \boldsymbol{p}}(\mathrm{Exp}_{\hat{\boldsymbol{y}}_i}^{-1} \boldsymbol{y}_i^c), \tag{16}$$

$$\partial_{\boldsymbol{g}_i} E \approx -P_{\boldsymbol{y}_i^c \boldsymbol{y}_i}(\mathrm{Exp}_{\boldsymbol{y}_i^c}^{-1} \hat{\boldsymbol{y}}_i). \tag{17}$$

One advantage of such approximation is that for some special manifolds, including manifold of SPD matrices $\mathcal{S}_{++}(n)$, parallel transports have analytical expressions and can be computed directly. For general manifolds that have no analytical expressions for parallel transports, approximation approaches such as Schild's ladder approximation [61], [62] can be used. The method approximates parallel transport by constructing geodesic parallelograms, which requires three exponential maps and two inverse exponential maps, as shown in Fig. 2.

Algorithm 2. PALMR for Multivariate Regression with Gross Error on Manifolds

Input: $\{(x_i, y_i)\}$, $\lambda \geq 0$, $\rho \geq 0$, $\mu_1 > 1$, $\mu_2 > 1$, and $k = 0$.
Output: $\tilde{\boldsymbol{p}}$, $\{\tilde{\boldsymbol{v}}_j\}$, and $\{\tilde{\boldsymbol{g}}\}_i$.
1: Initialize $\boldsymbol{p}$, $\{\boldsymbol{v}_j\}$, and $\{\boldsymbol{g}\}_i$.
2: **while** stopping criterion not satisfied **do**
3: $\boldsymbol{p}^{k+1} = \mathrm{Exp}_{\boldsymbol{p}^k}(-\frac{1}{c_k}\partial_{\boldsymbol{p}} E^k)$.
4: $\hat{\boldsymbol{v}}_j^k = \mathrm{prox}_{\frac{\lambda}{d_k}}^{R_v}(\boldsymbol{s}_j^k)$.
5: $\boldsymbol{v}_j^{k+1} = P_{\boldsymbol{p}^k \boldsymbol{p}^{k+1}}(\hat{\boldsymbol{v}}_j^k)$.
6: $\boldsymbol{g}_i^{k+1} = \mathrm{prox}_{\frac{\rho}{e_k}}^{R_g}(\boldsymbol{t}_i^k)$.
7: **end while**
8: **return** $\tilde{\boldsymbol{p}} \leftarrow \boldsymbol{p}^{k+1}$, $\tilde{\boldsymbol{v}}_j \leftarrow \boldsymbol{v}_j^{k+1}$ and $\tilde{\boldsymbol{g}}_i \leftarrow \boldsymbol{g}_i^{k+1}$.

4 Experiments

In this section, we empirically evaluate the performance of the proposed approach (i.e., *PALMR*) in working with synthetic and real DTI data sets, which lies in the $\mathcal{S}_{++}(3)$ manifold of SPD matrices. Throughout all experiments, we fix $\lambda = 0.1$ and choose the optimal ρ from set $\{0.05, 0.1, \ldots, 0.95, 1\}$ by a validation process using a validation data set consisting of the same number of data points as the testing data. As our algorithm is iterative by nature, in practice it stops if either of the two stopping criteria is met: (1) the difference between consecutive objective function values is below 1e-5, or (2) maximum number of iterations (100) is reached.

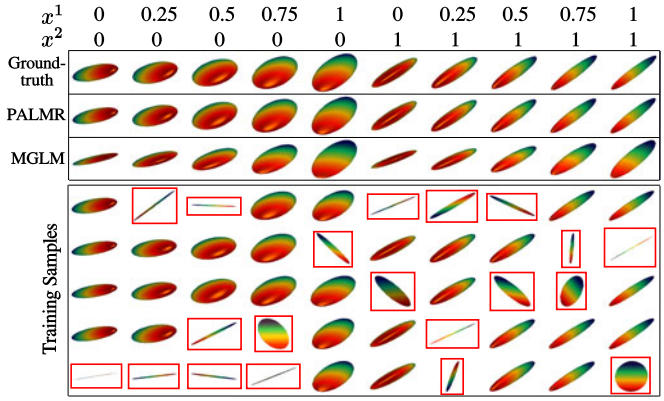


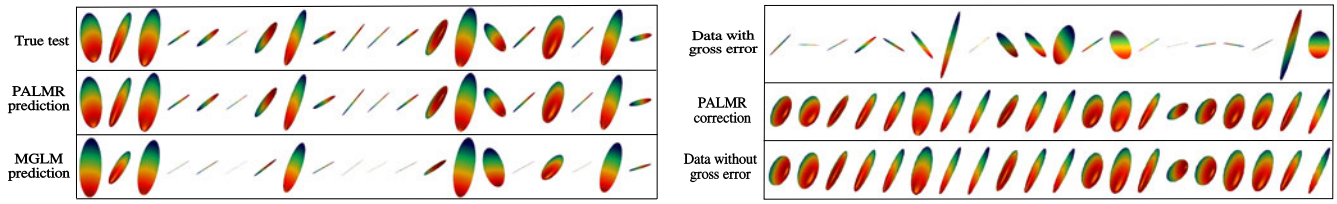
Fig. 3. Visualization of the synthesized training samples and the predictions of PALMR and MGLM. The two row vectors on the top give the values of X generating the data, red boxes identify the samples with gross error. The rows indexed by PALMR and MGLM display the predictions of corresponding method on the training data. All objects are viewed directly from overhead. Best viewed in color.

4.1 Synthetic DTI Data

Synthetic DTI data sets are constructed with known ground-truths and gross errors as follows: First, we randomly generate $\boldsymbol{p} \in \mathcal{S}_{++}(3)$, symmetric matrices $\{\boldsymbol{v}_j\}_{j=1}^d \subseteq \mathcal{S}(3)$ and $\{\boldsymbol{x}_i\}_{i=1}^N \subseteq \mathbb{R}^d$ where entries of $\boldsymbol{x}_i$ are sampled from standard normal distribution $\mathcal{N}(0, 1)$. Then the ground-truth DTI data is obtained as $\boldsymbol{y}_i^t := \mathrm{Exp}_{\boldsymbol{p}}(\sum_{j=1}^d x_i^j \boldsymbol{v}_j)$. This is followed by DTI data with stochastic noise as $\boldsymbol{y}_i^s := \mathrm{Exp}_{\boldsymbol{y}_i^t}(z_i)$, where z_i is a random matrix in $\mathcal{S}(3)$ with its entries being sampled from $\mathcal{N}(0, 1)$ and satisfies $\|z_i\|_{\boldsymbol{y}_i^t} \leq 0.1$. Meanwhile, the gross errors are generated by a two-step process: (a) Randomly select an index subset I_g from $\{1, 2, \ldots, N\}$, such that $|I_g| = \beta * N$ with $0 \leq \beta \leq 1$. (b) For $i \in I_g$, its grossly corrupted response is attained by $\boldsymbol{y}_i = \mathrm{Exp}_{\boldsymbol{y}_i^s}(\boldsymbol{g}_i)$, where $\boldsymbol{g}_i$ is a random matrix in $\mathcal{S}(3)$ satisfying $\|\boldsymbol{g}_i\|_{\boldsymbol{y}_i^s} = \sigma_g$. The rest of the training data remain unchanged, i.e., $\boldsymbol{y}_i = \boldsymbol{y}_i^s$ for $i \notin I_g$. Thus, among all N manifold-valued data, the percentage of grossly corrupted data is β. With the same $\boldsymbol{p}$ and $\{\boldsymbol{v}_j\}$, we also generate N_t pairs of testing data $\{(\boldsymbol{x}_i^{test}, \boldsymbol{y}_i^{test})\}$ and validation data.

We first conduct experiments on a data set with $d = 2$, $N = 50$, $\beta = 40\%$ and $\sigma_g = 5$, and compared with the multivariate general linear model (MGLM) of [2] which has not considered gross error. Training samples are displayed in Fig. 3, where we also show the predictions of PALMR and MGLM on the training data. Visual results of PALMR and MGLM on 20 testing data and training data correction by PALMR are presented in Figs. 4a and 4b, respectively. Collectively, the results suggest that PALMR indeed is capable of correctly identifying the gross errors during training. This enables the delivery of a better-behaved model. Fig. 4b shows that PALMR can effectively recover the original data (i.e., true data without gross error). It also produces improved regression results on testing data as displayed in Fig. 4a.

Next we quantitatively evaluate the effect of varying the internal parameters of PALMR, which include the number of independent variables d, the number of training data N, magnitude of gross error σ_g, and percentage of grossly corrupted training data β. To see the effect of one specific parameter, synthetic DTI data are generated by varying this parameter value while keeping rest parameters at their

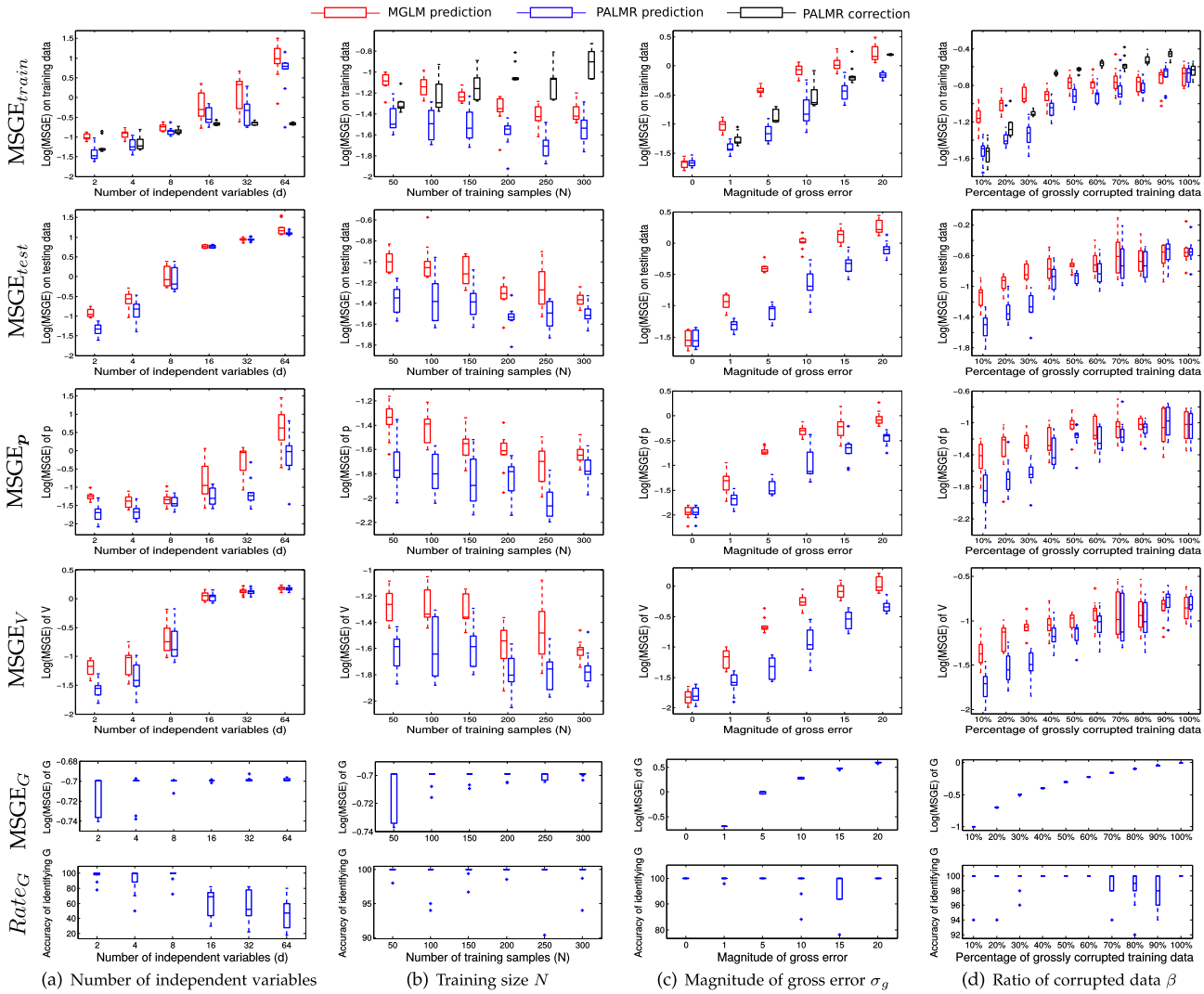


(a) Predictions of PALMR and MGLM on the testing data (b) Training data correction by our model

Fig. 4. Visual results of PALMR and MGLM. (a) Predictions for 20 testing data. (b) From top to bottom: Training samples corrupted by gross error (i.e., samples marked by red boxes in Fig. 3), correction results of PALMR, and the true data without gross error. Best viewed in color.

default values. The following default values are used: $d = 2$, $N = 50$, $\beta = 20\%$, and $\sigma_g = 1$. To evaluate performance of PALMR, the following mean squared geodesic error (MSGE) metrics are considered: $\text{MSGE}_{train} := \frac{1}{N}\sum_i d^2(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)$, $\text{MSGE}_{test} := \frac{1}{N_t}\sum_i d^2(\boldsymbol{y}_i^{test}, \hat{\boldsymbol{y}}_i^{test})$, $\text{MSGE}_{\boldsymbol{p}} := d^2(\boldsymbol{p}, \tilde{\boldsymbol{p}})$, $\text{MSGE}_V := \frac{1}{d}\sum_j \|\boldsymbol{v}_j - P_{\tilde{\boldsymbol{p}}\boldsymbol{p}}(\tilde{\boldsymbol{v}}_j)\|^2_{\boldsymbol{p}'}$, and $\text{MSGE}_G := \frac{1}{N}\sum_i \|\text{Exp}^{-1}_{\boldsymbol{y}_i}(\boldsymbol{y}_i^s) - \tilde{\boldsymbol{g}}_i\|^2_{\boldsymbol{y}_i'}$, where $\tilde{\boldsymbol{p}}$, $\tilde{\boldsymbol{v}}_j$ and $\tilde{\boldsymbol{g}}_i$ are the outputs of Algorithm 2. The data correction error is measured as $\frac{1}{N}\sum_i d(\boldsymbol{y}_i^s, \boldsymbol{y}_i^c)^2$. In addition, we say that gross error $\boldsymbol{g}_i$ is correctly identified if both $\boldsymbol{g}_i$ and $\tilde{\boldsymbol{g}}_i$ are either zero or nonzero, and compute the rate $Rate_G :=$ $number\ of\ correctly\ identified\ gross\ errors/N$. Results averaged over 10 repetitions are presented in Fig. 5, where each column corresponds to the effect of one parameter and each row corresponds to the results using one metric.

From Fig. 5, we have four observations: (1) PALMR has lower MSGE for all values of d, and our correction performs well on training data, cf. column Fig. 5a. (2) PALMR has large advantage over MGLM for all values of training size (N) and magnitude of gross error (σ_g), cf. columns Figs. 5b and 5c. (3) PALMR can handle training data with up to 80 percent being grossly corrupted, and delivers better



(a) Number of independent variables d (b) Training size N (c) Magnitude of gross error σ_g (d) Ratio of corrupted data β

Fig. 5. Box plots showing the effect of four different parameters: The number of tangent basis d, the number of training data N, the magnitude of gross error σ_g, and the percentage of grossly corrupted training data β, corresponding to four columns accordingly. Plots in each row show results using the same metric. Since MGLM does not consider gross error, the last two rows only show results of PALMR. See text for details. Best viewed in color.

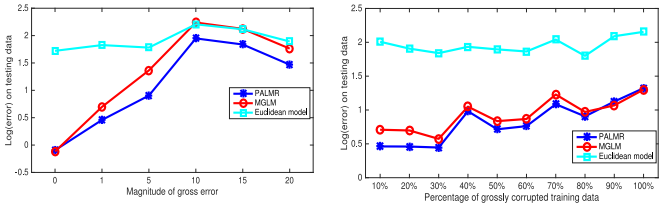


Fig. 6. Results of comparing PALMR and MGLM with euclidean model (7) under different magnitude of gross error (left) and different ratio of gross error in the training data (right). For each plot, the y-axis denotes the log-scale of median error over 10 reptitions measured by Frobenious norm.

result than MGLM. On the other hand, the performance is slightly worse if more than 80 percent of training data are corrupted, cf. column Fig. 5d. (4) PALMR can reliably identify most of the gross errors. Still it may not always correctly recover the true value of the error. This is evidenced in the last row of Fig. 5, where the MSGE on G increases as σ_g or β increases, and our correction error starts to stand out (i.e., being larger than both prediction errors of PALMR and MGLM) when over $\beta = 30\%$ of the training samples are grossly corrupted. We believe this is acceptable as in most practical situations, only small fraction of the training examples would be contaminated by gross errors.

Finally, we compare the proposed method PALMR and MGLM with an euclidean multivariate linear regression model with gross errors described in Eq. (7) of Example 1. All experimental settings are the same as above except three aspects: (i) Since the euclidean model can not deal with DTI tensors directly, for each tensor $\boldsymbol{y}$, we vectorize its upper triangle part into a 6-dimensional vector. Therefore, $X \in \mathbb{R}^{50\times 2}$ and $Y \in \mathbb{R}^{50\times 6}$ in model (7). (ii) Since predictions of the euclidean model are not guaranteed to lie on the SPD manifold, the geodesic metrics are not applicable. As alternate, we adopt Frobenious norm distance $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_F$ to measure the distance between prediction $\hat{\boldsymbol{y}}$ and ground-truth $\boldsymbol{y}$. (iii) We only investigate the effect of the magnitude of gross errors and the ratio of gross errors in the training data. Results are shown in Fig. 6, where the y-axis in each plot denotes the log-scale of median error over 10 reptitions measured by Frobenious norm. We observe that PALMR achieves the best performance and outporforms the euclidean model by a large margin under various settings. MGLM also performs better than the euclidean model, but when there are large gross errors in the training data, its advantage disappears, as can be seen in the left plot. These observations are within our expectation, since the euclidean model does not respect the intrinsic structure of the DTI data.

4.2 Real DTI Data

In this section, we apply PALMR to examine the effect of age and gender on human brain white matter. We experiment with the C-MIND database[3] released by Cincinnati Children's Hospital Medical Center (CCHMC) with the purpose of investigating brain development in children from infants and toddlers ($0 \sim 3$ years) through adolescence (18 years). We use the imaging data of participants who were scanned at CCHMC at year one and whose age were between 8 and 18 (2,947 to 6,885 days), consisting of 27 female and 31 male. The DTI data of each subject are first manually inspected and corrected for subject movements and eddy current distortions using FSL's eddy tool [63], then passed to FSL's brain extraction tool to delete non-brain tissue.[4] After the pre-processing, we use FSL's DTIFIT tool to reconstruct DTI tensors. Finally, all DTIs are registered to a population specific template constructed using DTI-TK.[5] We investigate six exemplar slices that have been identified as typical slices by domain experts and have been also similarly used by many existing works such as [2], [27]. And in particular, we are interested in the white matter region. At each voxel within the white matter region, the following multivariate regression model

$$\boldsymbol{y} = \mathrm{Exp}_{\boldsymbol{p}}(\boldsymbol{v}_1 \times \mathrm{age} + \boldsymbol{v}_2 \times \mathrm{gender}), \tag{18}$$

is adopted to describe the relation between the DTI data $\boldsymbol{y}$ and variables 'age' and 'gender'.

In DTI studies, another frequently used measure of a tensor is fractional anisotropy (FA) [64], [65] defined as

$$FA = \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_1 - \lambda_3)^2}{2(\lambda_1^2 + \lambda_2^2 + \lambda_2^2)}},$$

where λ_1, λ_2 and λ_3 are eigenvalues of the tensor. FA is an important measurement of diffusion asymmetry within a voxel and reflects fiber density, axonal diameter, and myelination in white matter. In our experiments, we also compared three models: two geodesic regression models, MGLM and PALMR, and the FA regression model which uses FA value to replace tensor $\boldsymbol{y}$ in Eq. (18). The relative FA error metric is employed to compare the results of geodesic regressions and FA regression, as follows: Since the responses of geodesic regression are tensors, the FA values of the tensors can be computed. The relative FA error metric is then evaluated on testing data, which is defined as the mean relative error between the FA values of the predicted tensors and the true tensors. Besides this relative FA error metric, the aforementioned mean squared geodesic error on testing data as in Section 4.1 is still engaged to compare the performance of MGLM and PALMR.

4.2.1 Model Significance

To examine the significance of the statistical model of Eq. (18) considered in our approach, the following hypothesis test is performed. The null hypothesis is $H_0 : \boldsymbol{v}_1 = 0$, which means, under this hypothesis, age has no effect on the DTI data. We randomly permute the values of age[6] among all samples and fix the DTI data, then apply model of Eq. (18) to the permuted data and compute the mean squared geodesic error $MSGE_{perm} = \frac{1}{N}\sum_i \mathrm{dist}(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i^p)^2$, where $\hat{\boldsymbol{y}}_i^p$ is the prediction of PALMR on the permuted data. Repeat the permutation $T = 1000$ times, we get a sequence of errors $\{MSGE_{perm}^i\}_{i=1}^T$ and calculate a p-value at each voxel using $p\text{-value} := \frac{|\{i \,|\, MSGE_{perm}^i < MSGE_{train}\}|}{T}$. Fig. 7 presents the maps of voxel-wise p-values for three models using six

3. https://cmind.research.cchmc.org

4. http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/

5. http://dti-tk.sourceforge.net/pmwiki/pmwiki.php

6. Empirical results investigating the effect of 'gender' are provided in Section 5 of the supplementary file, available online.

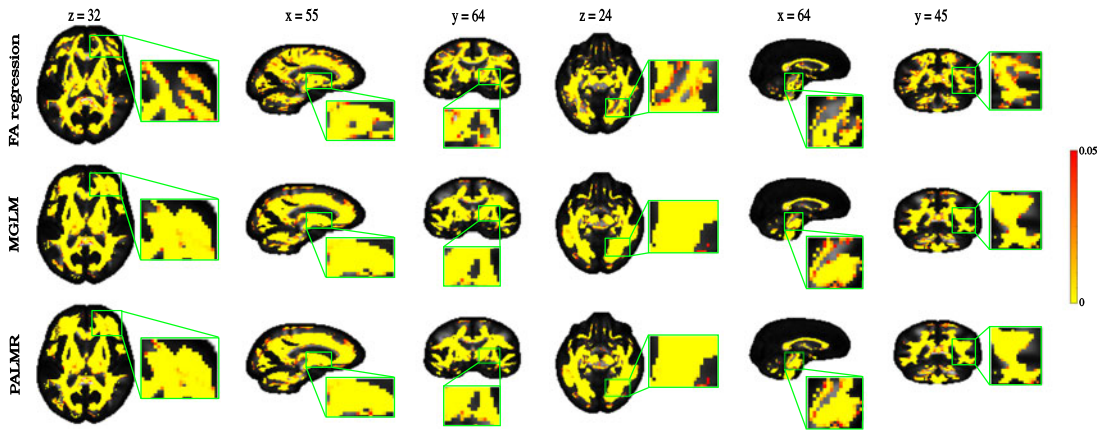


Fig. 7. p-value maps obtained by three methods: FA regression (top), MGLM (middle) and PALMR (bottom). p-value is only illustrated for voxels with p-value ≤ 0.05. Best viewed in color.

typical slices, and Fig. 8 displays the distribution of p-values for all six slices collectively.

As shown in Figs. 7 and 8, geodesic regression models are able to capture more white matter regions with aging effects than FA regression model. In addition, voxels satisfying p-value ≤ 0.05 are more spatially contiguous when geodesic regression models are used, as can be seen from the zoom-in plot for each slice in Fig. 7. This may be attributed to the fact that geodesic regression models preserve more geometric information of tensor images than that of FA regression. We also observe that PALMR and MGLM obtain very similar results. This is to be expected, as both methods use model of Eq. (18) and adopt geodesic regression on manifolds. The main difference is that PALMR considers gross error while MGLM does not, and in this experiment, there is no gross error in the DTI data.

4.2.2 Model Predictability

We proceed to investigate the predictability of PALMR when compared with existing methods such as FA regression and MGLM. For each of the six slices, we randomly partition our data into 40 training (20 female + 20 male) and 18 testing (7 female + 11 male) data, then train all three methods on each voxel within the white matter region. To test the ability of PALMR in handling gross errors, we consider three different experimental settings: (1) No gross error, where all training data are fully preprocessed as described at the beginning of Section 4.2; (2) 20 percent manual gross error, where for each voxel we randomly select 20 percent of training instances and insert gross error with magnitude $\sigma_g = 5$; (3) 20 percent registration error, where 20 percent of the patients in the training data are randomly selected to undergo an incomplete registration processing. Compared with fully preprocessed data, DTI data with registration error are obtained by skipping the diffeomorphic registration step in DTI-TK. The purpose of

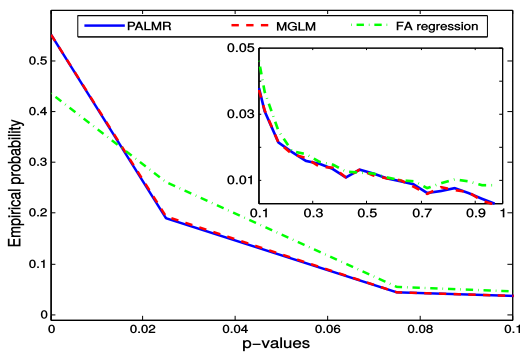


Fig. 8. Distribution of p-values for white matter tensors in all six slices. The inlet plot shows distribution of p-values over range $[0.1, 1]$.

TABLE 1
Median Values of Prediction Errors on All Six Slices of Testing Data

	Metrics	Methods	No gross error	20% manual gross error	20% registration error
Slice $z = 32$	Relative FA error	FA regression	0.9376	1.0414	0.9467
		MGLM	0.3223	0.4349	0.1654
		PALMR	**0.3210**	**0.3409**	**0.1316**
	MSGE	MGLM	0.1475	0.3530	0.1949
		PALMR	**0.1386**	**0.2196**	**0.1508**
Slice $x = 55$	Relative FA error	FA regression	0.9238	1.0362	0.8688
		MGLM	0.3298	0.5089	0.2067
		PALMR	**0.3279**	**0.3682**	**0.1882**
	MSGE	MGLM	0.1606	0.3631	0.3513
		PALMR	**0.1602**	**0.2562**	**0.2915**
Slice $y = 64$	Relative FA error	FA regression	0.8822	1.0136	0.9528
		MGLM	**0.3162**	0.4564	0.1917
		PALMR	0.3166	**0.3665**	**0.1562**
	MSGE	MGLM	0.1687	0.3720	0.2449
		PALMR	**0.1614**	**0.2843**	**0.1906**
Slice $z = 24$	Relative FA error	FA regression	0.8478	1.0066	0.8144
		MGLM	0.3570	0.7342	0.2140
		PALMR	**0.3564**	**0.5081**	**0.1581**
	MSGE	MGLM	0.1227	0.3466	0.2954
		PALMR	**0.1160**	**0.2530**	**0.2445**
Slice $x = 64$	Relative FA error	FA regression	0.9723	1.0526	0.9067
		MGLM	0.2142	0.4053	0.5023
		PALMR	**0.2114**	**0.3318**	**0.4318**
	MSGE	MGLM	0.1646	0.3663	0.2436
		PALMR	**0.1639**	**0.2779**	**0.2226**
Slice $y = 45$	Relative FA error	FA regression	0.9715	1.0695	0.9379
		MGLM	0.3779	0.5976	0.1739
		PALMR	**0.3767**	**0.5319**	**0.1664**
	MSGE	MGLM	0.2162	0.4205	0.2928
		PALMR	**0.2113**	**0.3780**	**0.2593**

We use two metrics, relative FA error and MSGE, to measure the prediction error. The best results in each setting are highlighted in bold.

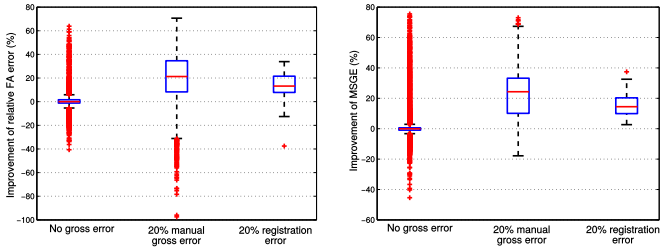
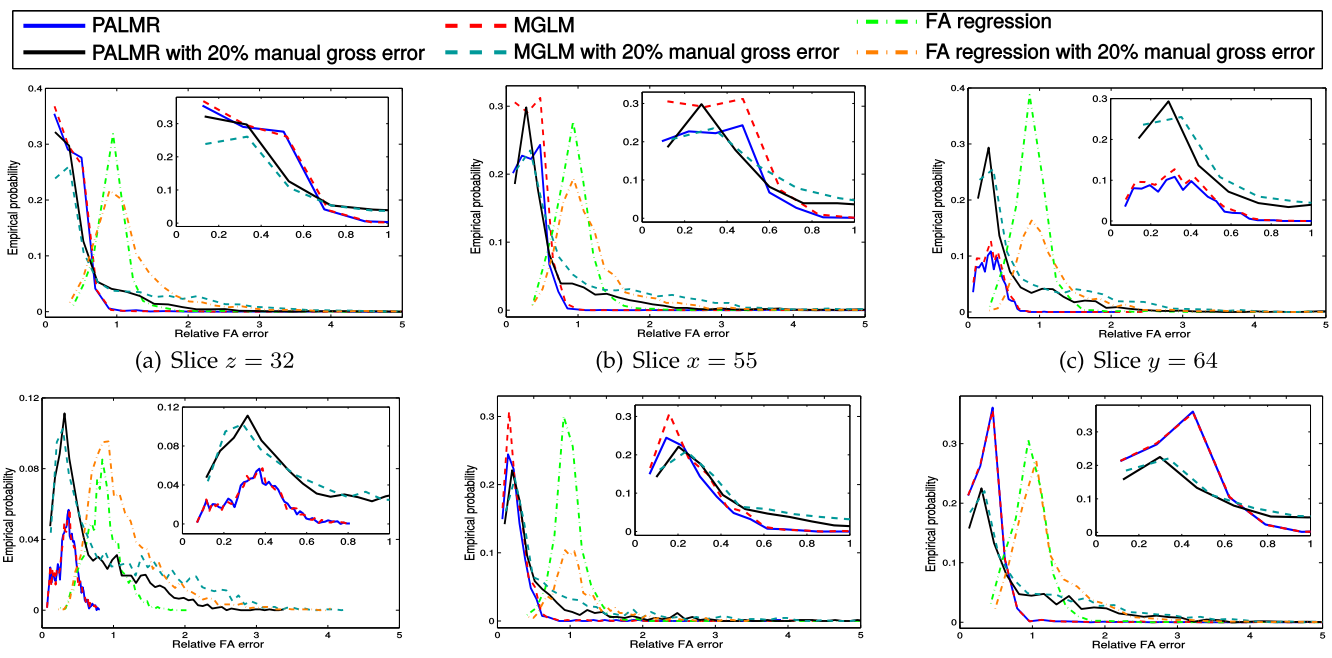


Fig. 9. Performance improvement obtained by PALMR measured with the relative FA error (left) and the MSGE (right). A positive value means that PALMR is better than the best competitor, and a negative value means that PALMR is worse. We first compute the performance improvement of PALMR on each voxel of all six slices to get a percentage value, then put all values under the same metric and experimental setting to plot a box plot.

experimenting on data with registration error is to imitate the realistic scenario that gross error can be caused by improper preprocessing of the data. We should remark that registration error is more challenging to handle than the manual gross error, since its magnitude varies dramatically for different voxels and patients. A heat map of registration error for each slice is provided in Fig. 1 of the supplementary file, available online. In this case, instead of considering all voxels on each slice, we set a threshold value ω and consider those voxels whose minimum registration error is greater than ω. For the first four slices, we set $\omega = 0.7$ and for the last two slices we set $\omega = 0.5$. The three comparison methods are examined on the three types of training data, and for each voxel the experiments are repeated 10 times.

Table 1 provides the median values of prediction errors measured with both relative FA error and MSGE on all voxels and over all six slices. As clearly indicated in Table 1, geodesic regression models again outperform FA regression model, which is to be expected. Moreover, when there is no gross error in the training data, both MGLM and PALMR achieve similar results. This is consistent with the claim that MGLM is a special case of PALMR when there is no gross error. In addition, the '20 percent manual gross error' column shows that when 20 percent of the training data contain gross errors PALMR outperforms MGLM by a large margin. For the challenging case of 20 percent registration error, the last column of Table 1 shows that PALMR is still much better than its competitors. In Fig. 9, we use box plots to demonstrate the performance advantage of PALMR over its competitors. For each metric, the performance improvement is computed as *(error of the best competitor - error of PALMR) / error of the best competitor * 100 percent*. Fig. 9 displays the same results as in Table 1 from a different perspective and with more details. We first compute the performance improvement of PALMR on each voxel of all six slices to get a percentage value, then put all values under the same metric and experimental setting to plot a box plot. Fig. 9 shows that PALMR improves the median prediction error by at least 20 and 15 percent in the case of manual gross error and registration error, respectively.

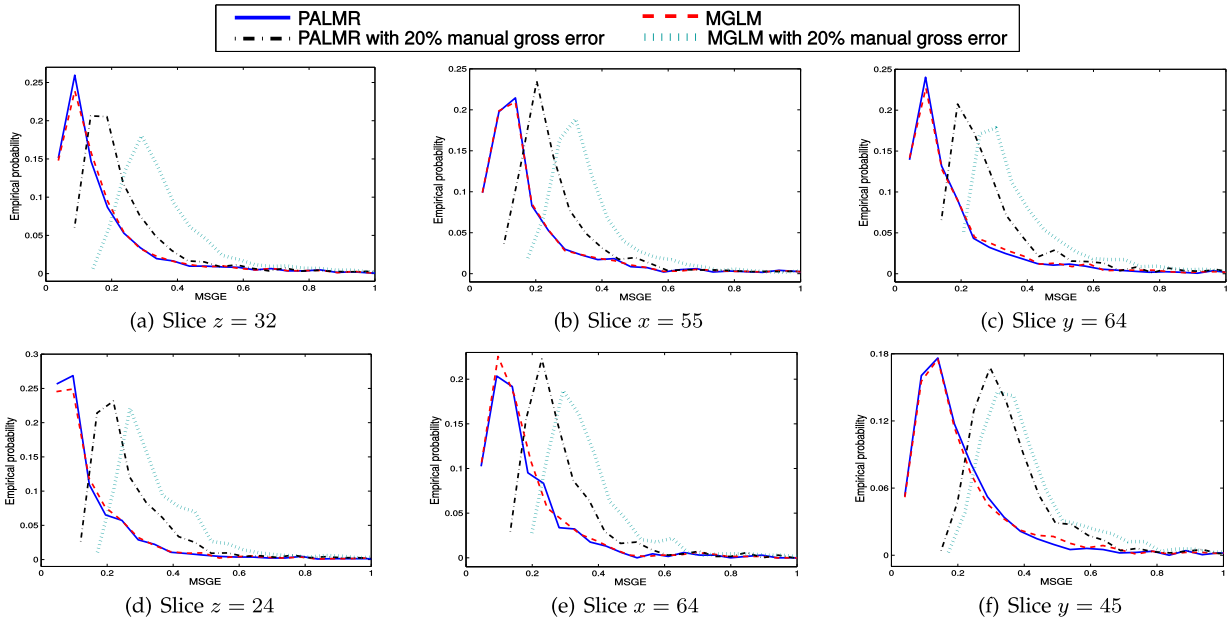
The distribution of prediction errors measured by the relative FA error and the MSGE on each slice is shown in Figs. 10 and 11, respectively. In each plot, the method with corresponding distribution on the left is better than the one with corresponding distribution on the right. From both Figs. 10 and 11, we get similar observation as in Table 1. Moreover, Fig. 11 shows that PALMR is more robust to gross errors than its competitors. In Fig. 2 of the supplementary file, available online, we also show the comparison of prediction errors of MGLM and PALMR on each voxel of all slices. We observe that on most of the voxels PALMR is better than MGLM when gross errors are present. More experimental results on real DTI data are available in Section 5 of the supplementary file, available online.



(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. 10. Distribution of relative FA errors on testing data. The inset figures show zoom-in plots of the prediction errors by MGLM and PALMR over the error interval $[0, 1]$. Better viewed in color.



(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. 11. Distribution of mean squared geodesic errors on testing data for each of the six slices. Better viewed in color.

5 Conclusion and Future Work

This paper focuses on the interesting problem of multivariate regression on manifolds with gross error contamination, where mathematical formulation nevertheless resides in a challenging landscape concerning a nonconvex and nonsmooth optimization on manifolds. A new algorithm, PALMR, is proposed to address this problem and its convergence property is analyzed. Through empirical studies, PALMR is shown to be capable of dealing with the presence of gross error and produces reliable results. For future work, there are several directions to explore. In terms of theoretical study, it remains to investigate the recoverbility of the proposed model, that is, to study conditions under which our model can correctly locate gross errors and recover their magnitude. It is also of interest to analyze the asymptotic behaviour of the resulting estimators. In terms of applications, in addition to age and gender, one may also consider the influence of handedness (i.e., left- or right-handed) on DTI responses. We also plan to apply our framework to different applications including shape analysis and robotics, where the manifolds of interest could be $SO(3)$ and $SE(3)$.

Acknowledgments

The research was carried out when Xudong Shi was an intern in A*STAR.

References

[1] B. C. Davis, T. Fletcher, E. Bullitt, and S. C. Joshi, "Population shape regression from random design data," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 255–266, 2010.

[2] H. Kim, et al., "Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2705–2712.

[3] E. Cornea, et al., "Regression models on Riemannian symmetric spaces," *J. Roy. Statistical Soc.: Series B (Statistical Methodology)*, vol. 79, no. 2, pp. 463–482, 2017.

[4] P. Muralidharan and P. T. Fletcher, "Sasaki metrics for analysis of longitudinal data on manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1027–1034.

[5] J. Hsu, et al., "Gender differences and age-related white matter changes of the human brain: A diffusion tensor imaging study," *NeuroImage*, vol. 39, no. 2, pp. 566–577, 2008.

[6] M. Wu, et al., "Comparison of EPI distortion correction methods in diffusion tensor MRI using a novel framework," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2008, pp. 321–329.

[7] A. Zalesky, "Moderating registration misalignment in voxelwise comparisons of DTI data: A performance evaluation of skeleton projection," *Magn. Resonance Imag.*, vol. 29, no. 1, pp. 111–125, 2011.

[8] M. Bastin, P. Armitage, and I. Marshall, "A theoretical study of the effect of experimental noise on the measurement of anisotropy in diffusion imaging," *Magn. Resonance Imag.*, vol. 16, no. 7, pp. 773–785, 1998.

[9] S. Basu, T. Fletcher, and R. Whitaker, "Rician noise removal in diffusion tensor MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2006, pp. 117–125.

[10] J. Wright and Y. Ma, "Dense error correction via ℓ^1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3540–3560, Jul. 2010.

[11] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[12] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix recovery from errors and erasures," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4324–4337, Jul. 2013.

[13] N. Nguyen and T. Tran, "Robust lasso with missing and grossly corrupted observations," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2036–2058, Apr. 2013.

[14] H. Xu and C. Leng, "Robust multi-task regression with grossly corrupted observations," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 1341–1349.

[15] K. Bhatia, P. Jain, and P. Kar, "Robust regression via hard thresholding," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 721–729.

[16] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, no. 1/2, pp. 459–494, 2014.

[17] P. Basser and D. Jones, "Diffusion-tensor MRI: Theory, experimental design and data analysis – a technical review," *NMR Biomed.*, vol. 15, pp. 456–467, 2002.

[18] A. Srivastava, S. Joshi, W. Mio, and X. Liu, "Statistical shape analysis: Clustering, learning, and testing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 590–602, Apr. 2005.

[19] T. P. Fletcher, C. Lu, S. M. Pizer, and S. C. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 995–1005, Aug. 2004.

[20] J. Hinkle, P. T. Fletcher, and S. C. Joshi, "Intrinsic polynomials for regression on Riemannian manifolds," *J. Math. Imag. Vis.*, vol. 50, no. 1/2, pp. 32–52, 2014.

[21] A. Saxena, J. Driemeyer, and A. Ng, "Learning 3-D object orientation from images," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 794–800.
[22] R. Wang, K. Pulli, and J. Popović, "Real-time enveloping with rotational regression," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. no. 73.
[23] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 728–735.
[24] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *Int. J. Comput. Vis.*, vol. 66, no. 1, pp. 41–66, 2006.
[25] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Free-form region description with second-order pooling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1177–1189, Jun. 2015.
[26] A. Cherian and S. Sra, "Riemannian dictionary learning and sparse coding for positive definite matrices," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 2859–2871, Dec. 2017.
[27] J. Du, A. Goh, S. Kushnarev, and A. Qiu, "Geodesic regression on orientation distribution functions with its application to an aging study," *NeuroImage*, vol. 87, pp. 416–426, 2014.
[28] T. Fletcher, "Geodesic regression and the theory of least squares on Riemannian manifolds," *Int. J. Comput. Vision*, vol. 105, no. 2, pp. 171–185, 2013.
[29] P. Fletcher and S. C. Joshi, "Riemannian geometry for the statistical analysis of diffusion tensor data," *Signal Process.*, vol. 87, no. 2, pp. 250–262, 2007.
[30] X. Pennec, "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements," *J. Math. Imag. Vis.*, vol. 25, no. 1, pp. 127–154, 2006.
[31] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.
[32] M. Banerjee, R. Chakraborty, E. Ofori, M. Okun, D. Viallancourt, and B. Vemuri, "A nonlinear regression technique for manifold valued data with applications to medical image analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4424–4432.
[33] Y. Hong, R. Kwitt, N. Singh, N. Vasconcelos, and M. Niethammer, "Parametric regression on the Grassmannian," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2284–2297, Nov. 2016.
[34] F. Steinke and M. Hein, "Non-parametric regression between manifolds," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 1561–1568.
[35] M. Hein, "Robust nonparametric regression with metric-space valued output," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 718–726.
[36] X. Li, "Compressed sensing and matrix completion with constant proportion of corruptions," *Constructive Approximation*, vol. 37, no. 1, pp. 73–99, 2013.
[37] M. Harandi, C. Sanderson, R. Hartley, and B. Lovell, "Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 216–229.
[38] M. T. Harandi, R. I. Hartley, C. Shen, B. C. Lovell, and C. Sanderson, "Extrinsic methods for coding and dictionary learning on Grassmann manifolds," *Int. J. Comput. Vis.*, vol. 114, no. 2/3, pp. 113–136, 2015.
[39] R. Vemulapalli, J. Pillai, and R. Chellappa, "Kernel learning for extrinsic classification of manifold features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1782–1789.
[40] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Optimizing over radial kernels on compact manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3802–3809.
[41] M. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li, "Expanding the family of Grassmannian kernels: An embedding perspective," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 408–423.
[42] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel methods on riemannian manifolds with Gaussian RBF kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2464–2477, Dec. 2015.
[43] M. Harandi and M. Salzmann, "Riemannian coding and dictionary learning: Kernels to the rescue," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3926–3935.
[44] A. Feragen, F. Lauze, and S. Hauberg, "Geodesic exponential kernels: When curvature and linearity conflict," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3032–3042.
[45] M. P. do Carmo, *Riemannian Geometry*. Basel, Switzerland: Birkhäuser, 1992.
[46] E. A. Papa Quiroz, "An extension of the proximal point algorithm with Bregman distances on Hadamard manifolds," *J. Global Optimization*, vol. 56, no. 1, pp. 43–59, 2013.
[47] J. X. da Cruz Neto, L. L. de Lima, and P. R. Oliveira, "Geodesic algorithms in Riemannian geometry," *Balkan J. Geom. Appl.*, vol. 3, no. 2, pp. 89–100, 1998.
[48] 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. Operations Res.*, vol. 35, no. 2, pp. 438–457, 2010.
[49] J. X. da Cruz Neto, P. R. Oliveira, A. S. Pedro 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.
[50] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton Univ. Press, 2007.
[51] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a Matlab toolbox for optimization on manifolds," *J. Mach. Learn. Res.*, vol. 15, pp. 1455–1459, 2014.
[52] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1999.
[53] W. Huang, K. A. Gallivan, and P.-A. Absil, "A Broyden class of quasi-Newton methods for Riemannian optimization," *SIAM J. Optimization*, vol. 25, no. 3, pp. 1660–1685, 2015.
[54] O. Ferreira and P. Oliveira, "Subgradient algorithm on Riemannian manifolds," *J. Optimization Theory Appl.*, vol. 97, no. 1, pp. 93–104, 1998.
[55] M. Bačák, R. Bergmann, G. Steidl, and A. Weinmann, "A second order non-smooth variational model for restoring manifold-valued images," *SIAM J. Sci. Comput.*, vol. 38, no. 1, pp. A567–A597, 2016.
[56] A. Kovnatsky, K. Glashoff, and M. Bronstein, "MADMM: A generic algorithm for non-smooth optimization on manifolds," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 680–696.
[57] S. Hosseini and A. Uschmajew, "A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds," *SIAM J. Optimization*, vol. 27, no. 1, pp. 173–189, 2017.
[58] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Berlin, Germany: Springer, 1998.
[59] J. J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société Mathématique de France*, vol. 93, pp. 273–299, 1965.
[60] H. Kim, J. Xu, B. Vemuri, and V. Singh, "Manifold-valued Dirichlet processes," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1199–1208.
[61] A. Kheyfets, W. Miller, and G. Newton, "Schild's ladder parallel transport procedure for an arbitrary connection," *Int. J. Theoretical Physics*, vol. 39, no. 12, pp. 2891–2898, 2000.
[62] M. Lorenzi and X. Pennec, "Efficient parallel transport of deformations in time series of images: From Schild's to pole ladder," *J. Math. Imag. Vis.*, vol. 50, no. 1/2', pp. 5–17, 2014.
[63] M. Jenkinson, C. Beckmann, T. Behrens, M. Woolrich, and S. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.
[64] P. Basser, "Inferring microstructural features and the physiological state of tissues from diffusion-weighted images," *NMR Biomed.*, vol. 8, pp. 333–344, 1995.
[65] P. Basser and C. Pierpaoli, "Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI," *J. Magn. Resonance*, vol. 111, no. 3, pp. 209–219, 1996.

Xiaowei Zhang received the PhD degree in applied mathematics from the National University of Singapore, in 2013. He is a senior post doctoral research fellow in the Bioinformatics Institute, A*STAR, Singapore. His current research interests include machine learning and its applications to computer vision, data mining, matrix computations and its applications, and numerical optimization. He is a member of the IEEE.

Xudong Shi received the BS degree from Sichuan University, China, in 2013. He is working toward the PhD degree in the School of Computing, National University of Singapore. His research interests include computer vision, machine learning, and biometrics.

Yu Sun (M'12) received the BEng degree in biomedical engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2005, the PhD degree in electronic, electrical, and system engineering from Loughborough University, Leicestershire, United Kingdom, in 2011, and the joint-PhD degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2012. Since 2012, he has been a postdoctoral research fellow and currently a senior research fellow in the Singapore Institute for Neurotechnology (SINAPSE), National University of Singapore, Singapore. His current research interests include the area of biomedical signal processing, imaging photoplethysmography, and functional/structural neuroimaging processing with particular focus on human connectome. He is the associate editor of the *Medical and Biological Engineering and Computing*. He received the Best Paper Award in the 2013 IEEE International Neurotechnology Consortium Workshop and the Best Poster Award in the 2013 IEEE Life Sciences Grand Challenges Conference. He is a member of the IEEE.

Li Cheng received the PhD degree in computer science from the University of Alberta, Canada. He is a research scientist and group leader in Bioinformatics Institute (BII). Prior to joining BII July of 2010, he worked in Statistical Machine Learning group of NICTA, Australia, TTI-Chicago, and the University of Alberta, Canada. His research expertise is mainly on machine learning and computer vision. He is a senior member of the IEEE.