Statistica Sinica Preprint No: SS-2023-0268	
Title	Intrinsic Minimum Average Variance Estimation for Dimension Reduction with Symmetric Positive Definite Matrices and Beyond
Manuscript ID	SS-2023-0268
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0268
Complete List of Authors	Baiyu Chen, Shuang Dai and Zhou Yu
Corresponding Authors	Zhou Yu
E-mails	zyu@stat.ecnu.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Intrinsic Minimum Average Variance Estimation for Dimension Reduction with Symmetric Positive-Definite Matrices and Beyond

Baiyu Chen, Shuang Dai and Zhou Yu*

East China Normal University

Abstract: In this paper, we estimate the central mean subspace in a dimension reduction problem where the response is a symmetric positive-definite matrix. We propose the intrinsic minimum average variance estimation and the intrinsic outer product gradient method which fully exploit the geometric structure of the Riemannian manifold where the response resides. We present algorithms for our newly developed methods under the log-Euclidean metric and the log-Cholesky metric. The two metrics endow the manifold with a commutative Lie group structure that transforms our manifold model into a Euclidean one and helps us derive the consistency and asymptotic normality of estimators. Our methods are then naturally extended to the case allowing $p = p_n$ to diverge and the case of general Riemannian manifolds. Several simulation studies and an application to the New York taxi network data showcase the superiority of our proposals.

Key words and phrases: Central mean subspace, log-Cholesky metric, log-Euclidean

*Corresponding author

metric, minimum average variance estimation, outer product gradient, symmetric positive-definite matrix.

1. Introduction

For $Y \in R$ and $X \in R^p$, sufficient dimension reduction (SDR) seeks a $p \times d$ matrix B with $d << p$ such that $Y \perp\!\!\!\perp X \mid B^\mathsf{T} X$. The space spanned by the columns of B, denoted by $\mathcal{S}(B)$, is called the SDR subspace. If $\mathcal{S}(B)$ is a subspace of all other SDR subspaces, it is called the central subspace (CS). Popular methods estimating CS include sliced inverse regression (Li, 1991), sliced average variance estimation (Cook and Weisberg, 1991), directional regression (Li and Wang, 2007), semiparametric approaches (Ma and Zhu, 2012, 2013, 2019), among others. Although CS provides a complete picture of the dependency of Y on X, one might be only interested in the conditional mean function for which the dimension reduction assumes

$$Y \perp\!\!\!\perp E(Y \mid X) \mid B^\mathsf{T} X. \tag{1.1}$$

Similar to CS, the central mean subspace (CMS) can be defined as the intersection of all $\mathcal{S}(B)$ with B satisfying (1.1). The minimum average variance estimation (MAVE) and the outer product of gradient (OPG) method (Xia et al., 2022; Xia, 2007) were pioneer tools to estimate CMS.

The above-mentioned dimension reduction methods deal with high-dimensional Euclidean vectors. However, with the rapid development of data collection techniques, non-Euclidean data are encountered frequently and it is necessary to consider dimension reduction for non-Euclidean data. These complex data often reside in a Riemannian manifold or a general metric space whose nonlinear nature disables Euclidean methods. Symmetric positive-definite (SPD) matrices, emerging in numerous scientific applications, serve as a representative of such data. A concrete example is analysis of functional connectivity between brain regions. Such connectivity is often characterized by the covariance (SPD matrices) of blood-oxygen-level dependent signals from different regions (Huettelet al., 2008). Another application is diffusion tensor magnetic resonance imaging (DTI) widely applied in medical imaging for diagnosis. This technique models the shape of diffusion of water molecules in a voxel by an ellipsoid in R^3 and estimates diffusion tensors to describe this ellipsoid. Diffusion tensors are 3×3 SPD matrices with three positive eigenvalues representing the lengths of three principal diameters of the ellipsoid and corresponding eigenvectors implying the directions of three axes. SPD matrices can also be generated by network data. Dubey and Müller (2020) divided the New York city into several zones (nodes) and collected networks (adjacency matrices) describing taxi movements between zones. Finally these adjacency matrices are turned into SPD matrices for later research by matrix exponentiation.

All $m \times m$ SPD matrices form a Riemannian manifold denoted by $\mathrm{Sym}^+(m)$ under some Riemannian metric. Up to now there have been many papers generalizing traditional statistical methods in Euclidean spaces to manifolds or more general metric spaces such as local polynomial regression for SPD matrices (Yuan et al., 2012; Zhu et al., 2009; Cornea et al., 2016), Fréchet regression for random objects (Peterson and Müller, 2019), intrinsic Riemannian functional principal component analysis and functional linear regression (Lin and Yao, 2019), additive model for SPD matrices (Lin et al., 2022), Fréchet SDR for random objects (Ying and Yu, 2022; Zhang et al., 2021), intrinsic Wasserstein correlation analysis (Zhou et al., 2021), single index Fréchet regression (Bhattacharjee and Müller, 2021), autoregressive optimal transport model (Zhu and Müller, 2021) and so on. Among these works, two recent papers are related to non-Euclidean dimension reduction. Ying and Yu (2022) and Zhang et al. (2021) modified several Euclidean dimension reduction methods to accommodate Euclidean X and metric space-valued Y. They incorporated non-Euclidean information in Y by substituting the Euclidean norm $\|Y_i - Y_j\|$ by the geodesic distance $d(Y_i, Y_j)$ or a universal kernel $K(Y_i, Y_j)$. However, when the response lies in a manifold, even though the two methods can be applied, they fail to fully exploit the intrinsic geometry of the manifold and some information contained in the response may be inevitably lost.

In this paper, we focus on dimension reduction (1.1) with Y being SPD matrices. We generalize the state-of-the-art sufficient mean dimension reduction methods MAVE and OPG for the estimation of CMS. The basic idea of our proposal also stems from the local polynomial regression for SPD matrices introduced by Yuan et al. (2012), which replaced the squared distance by the geodesic distance on $\mathrm{Sym}^+(m)$ and performed Taylor expansion after parallel transport to estimate the intrinsic conditional expectation of an SPD response, given a covariate vector X. Yuan et al. (2012) only considered the case where X is a scalar and we here take a step forward to handle the high-dimensional X. We call our method intrinsic MAVE and intrinsic OPG since we do not assume an ambient space surrounding $\mathrm{Sym}^+(m)$ or an isometric embedding into a Euclidean space (Lin and Yao, 2019) during the construction of our models. Furthermore, we generalize our proposals to the situation where the dimension p of the predictor X diverges, i.e., $p = p_n \to \infty$ as $n \to \infty$.

The rest of the paper is organized as follows. Some preliminaries on manifolds are presented in Section 2. Then we introduce our intrinsic dimension reduction proposals for SPD matrices in Section 3, together with asymptotic analysis of our estimators. A cross validation procedure for selecting the structural dimension d is also included in Section 3. Section 4 contains two adaptations of our methods: one is the case allowing $p = p_n \to \infty$ as $n \to \infty$; the other is the formulation of the methods on a general manifold. Simulation studies and an application to the New York taxi network data are presented in Section 5. A discussion in Section 6 finishes this paper.

2. Preliminaries on Manifolds

We first introduce some basic notions for Riemannian manifolds and Lie groups (Tu, 2011; Lang, 1999). Let $\mathcal{M}$ be a simply connected and smooth manifold and $p \in \mathcal{M}$. For a small scalar $\delta > 0$, let $c(t)$ be a continuously differential map from $(-\delta, \delta)$ to $\mathcal{M}$ passing through $c(0) = p$. A tangent vector at p is the derivative of the curve $c(t)$ at $t = 0$. All such tangent vectors at p form a vector space named the tangent space at p, which is denoted by $T_p\mathcal{M}$. The tangent space of $p \in \mathrm{Sym}^+(m)$ is a vector space $\mathrm{Sym}(m)$ consisting of all $m \times m$ symmetric matrices. Each tangent space $T_p\mathcal{M}$ can be endowed with an inner product $\langle \cdot, \cdot \rangle_p$ that varies smoothly with p. The inner products $\{\langle \cdot, \cdot \rangle_p : p \in \mathcal{M}\}$ are collectively denoted by $\langle \cdot, \cdot \rangle$, which is referred to as the Riemannian metric of $\mathcal{M}$. With a Riemannian metric, we can define a distance $d(\cdot,\cdot)$ on $\mathcal{M}$ that turns $\mathcal{M}$ into a metric space. The length of a continuously differentiable curve $c(t):[t_0,t_1]\to\mathcal{M}$ is calculated as $\int_{t_0}^{t_1}\langle c'(t),c'(t)\rangle_{c(t)}^{1/2}\mathrm{d}t$, where $c'(t)$ is the derivative of $c(t)$. And $d(p,q)$ is the infimum of the length over all continuously differentiable curves joining p and q.

A geodesic γ is a curve defined on $[0,\infty)$ such that for each $t\in[0,\infty)$, $\gamma([t,t+\epsilon])$ is the shortest path connecting $\gamma(t)$ and $\gamma(t+\epsilon)$ for sufficiently small $\epsilon>0$. The Riemannian exponential map Exp_p at $p\in\mathcal{M}$ is a function mapping $T_p\mathcal{M}$ into $\mathcal{M}$ and is defined by $\mathrm{Exp}_p(u)=\gamma(1)$ with $\gamma(0)=p$ and $\gamma'(0)=u\in T_p\mathcal{M}$. The inverse of Exp_p, if exists, denoted by Log_p and called the Riemannian logarithm map at p, can be defined as $\mathrm{Log}_p q=u$ for $q\in\mathcal{M}$ such that $\mathrm{Exp}_p u=q$.

A vector field U is a function defined on $\mathcal{M}$ such that $U(p)\in T_p\mathcal{M}$. Given a curve $\gamma(t)$ on $\mathcal{M}$, $t\in I$ for a real interval I, a vector field along γ is a smooth map defined on I such that $U(t)\in T_{\gamma(t)}\mathcal{M}$. We say U is parallel along γ if $\nabla_{\gamma'(t)}U=0$ for all $t\in I$ where ∇ is the Levi-Civita connection on $\mathcal{M}$. In this paper we only focus on parallel vector fields along geodesics. Let $\gamma:[0,1]\to\mathcal{M}$ be a geodesic connecting p and q, and U is a parallel vector field along γ such that $U(0)=u$ and $U(1)=v$. Then the parallel transport of u along γ is denoted as $\phi_p(u)=v$.

When $(\mathcal{M}, \oplus)$ is a group and the group operation $\oplus$ and its inverse are both smooth, $(\mathcal{M}, \oplus)$ is called a Lie group. The tangent space at the identity element e is called a Lie algebra denoted by $\mathfrak{g}$. It consists of left-invariant vector fields U which satisfies $U(p \oplus q) = (DL_p)(U(q))$, where $L_p : q \to p \oplus q$ is the left translation at p and DL_p is the differential of L_p. A Riemannian metric $\langle \cdot, \cdot \rangle$ is called left-invariant if $\langle u, v \rangle_q = \langle DL_p(u), DL_p(v) \rangle_{p \oplus q}$ for all $p, q \in \mathcal{M}$ and $u, v \in T_q\mathcal{M}$. Right invariance can be defined similarly. A metric is bi-invariant if it is both left-invariant and right-invariant. The Lie exponential map, denoted by $\mathfrak{exp}$ is defined by $\mathfrak{exp}(u) = \gamma(1)$ where $\gamma : R \to \mathcal{M}$ is the unique one-parameter subgroup such that $\gamma'(0) = u \in \mathfrak{g}$. Its inverse, if exists, is denoted by $\mathfrak{log}$. Please make a distinction between the Riemannian exponential map "Exp", the Lie exponential map "$\mathfrak{exp}$" and the common matrix exponential operation "exp" which appear frequently in later sections. When $\langle \cdot, \cdot \rangle$ is bi-invariant, $\mathfrak{exp}$ coincides with Exp_e.

3. Methodology

3.1 Minimum Average Variance Estimation Revisited

Let Y and X be respectively R-valued and R^p-valued random variables. Minimum average variance estimation (MAVE) considers a regression-type model for dimension reduction:

$$Y = g(B_0^\top X) + \varepsilon, \tag{3.2}$$

where g is an unknown smooth function, $B_0 = (\beta_1, ..., \beta_d)$ is a $p \times d$ orthogonal matrix ($B_0^\top B_0 = I_d$) with $d < p$ and $E(\varepsilon \mid X) = 0$ almost surely. The aim is to estimate B_0 since $B_0^\top X$ captures all the information provided by X on Y.

The direction B_0 is the solution of

$$\min_B E\{Y - E(Y \mid B^\top X)\}^2, \quad \text{subject to } B^\top B = I. \tag{3.3}$$

Since the conditional variance given $B^\top X$ is

$$\sigma_B^2(B^\top X) = E[\{Y - E(Y \mid B^\top X)\}^2 \mid B^\top X], \tag{3.4}$$

it follows that $E\{Y - E(Y \mid B^\top X)\}^2 = E\sigma_B^2(B^\top X)$ and minimizing (3.3) is equivalent to minimizing $E\sigma_B^2(B^\top X)$, which explains the name "minimum average variance estimation".

Suppose $(X_i, Y_i), i = 1, ..., n$ are samples from (X, Y). Let $g_B(\cdot) = E(Y \mid B^\top X = \cdot)$. For a given X_0, a local linear approximation is

$$E(Y_i \mid B^\top X_i) \approx a + b^\top B^\top (X_i - X_0),$$

where $a = g_B(B^\top X_0)$ and $b = \nabla g_B(B^\top X_0)$.

According to (3.4) and the idea of local linear smoothing, we can estimate $\hat{\sigma}_B^2(B^T X_0)$ by

$$\sum_{i=1}^{n}\{Y_i - E(Y_i \mid B^\intercal X_i)\}^2 w_{i0} \approx \sum_{i=1}^{n}[Y_i - \{a + b^\intercal B^\intercal (X_i - X_0)\}]^2 w_{i0}, \quad (3.5)$$

where $w_{i0} \geq 0$, $(i = 1, ..., n)$ are some weights such that $\sum_{i=1}^{n} w_{i0} = 1$.

Eventually minimizing $E\sigma_B^2(B^\intercal X)$ can be approximated by

$$\min_{B^\intercal B = I} \sum_{j=1}^{n} \hat{\sigma}_B^2(B^\intercal X_j) = \min_{\substack{B^\intercal B=I,\\ a_j, b_j}} \sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i - \{a_j + b_j^\intercal B^\intercal (X_i - X_j)\}]^2 w_{ij},$$

One usually employs $w_{ij} = K_h(X_i - X_j)/\sum_{i=1}^{n} K_h(X_i - X_j)$ and for $u \in R^p$, $K_h(u) = K(u/h)/h^p$ where $K(v_1, ..., v_p) = K_0(v_1^2 + ... + v_p^2)$ with $K_0(\cdot)$ being the univariate density function and $h \in R$ being the bandwidth.

3.2 Intrinsic MAVE and OPG for SPD Matrices

When $X \in R^p$ but $Y \in \mathrm{Sym}^+(m)$ and $(X_i, Y_i), i = 1, ..., n$ are sampled from (X, Y), the first obstacle is the definition of $E(Y \mid B^\intercal X)$. According to Yuan et al. (2012), the intrinsic conditional expectation of Y at $B^\intercal X = B^\intercal x$ is defined as $D(B^\intercal x) \in \mathrm{Sym}^+(m)$ such that

$$E\left\{\mathrm{Log}_{D(B^\intercal x)} Y \mid B^\intercal x\right\} = O_m,$$

where O_m is an $m \times m$ matrix with all elements 0 and the expectation is taken in a component-wise way. From now on we use $D(B^\intercal x)$ instead of $E(Y \mid B^\intercal x)$.

Euclidean operations like addition and subtraction are invalid in $\mathrm{Sym}^+(m)$, so the squared distance in Euclidean MAVE may be substituted by the geodesic distance $d(\cdot,\cdot)$ in $\mathrm{Sym}^+(m)$ and (3.5) is modified to

$$\sum_{i=1}^{n} d^2\{Y_i, D(B^\top X_i)\} w_{i0}. \tag{3.6}$$

Next we want to similarly expand $D(B^\top X_i)$ at a given point $B^\top X_0$. Since $D(B^\top X_i)$ is in the curved space, directly expanding $D(B^\top X_i)$ at $B^\top X_0$ is infeasible. Instead, we first use the Riemannian logarithm map to transform $D(B^\top X_i)$ into $\mathrm{Log}_{D(B^\top X_0)} D(B^\top X_i) \in T_{D(B^\top X_0)}\mathrm{Sym}^+(m)$. Since $\mathrm{Log}_{D(B^\top X_0)} D(B^\top X_i)$ for different X_0 are in different tangent spaces, these tangent vectors are transported from $T_{D(B^\top X_0)}\mathrm{Sym}^+(m)$ to the same tangent space $T_{I_m}\mathrm{Sym}^+(m)$ by using parallel transport given by:

$$\phi_{D(B^\top X_0)} : T_{D(B^\top X_0)}\mathrm{Sym}^+(m) \to T_{I_m}\mathrm{Sym}^+(m),$$

where I_m is the identity matrix. Then $f(B^\top X_i) = \phi_{D(B^\top X_0)} \mathrm{Log}_{D(B^\top X_0)} D(B^\top X_i)$ is a function from R^d to $T_{I_m}\mathrm{Sym}^+(m)$ which is a vector space. We now can expand $f(B^\top X_i)$ at $B^T X_0$ using Taylor series expansion. Considering $f(B^\top X_i)$ is an $m \times m$ symmetric matrix and $B^\top X_0$ is a d-dimensional vector, we differentiate each component of $f(B^\top X_i)$ with respect to $B^\top X_0$ and this leads to

$$\begin{aligned}\mathrm{Log}_{D(B^\top X_0)} D(B^\top X_i) =& \phi^{-1}_{D(B^\top X_0)}\{f(B^\top X_i)\}\\ &\approx \phi^{-1}_{D(B^\top X_0)}\left(C_0\left[I_m \otimes \left\{B^\top (X_i - X_0)\right\}\right]\right),\end{aligned}$$

which gives

$$D(B^\top X_i) \approx \mathrm{Exp}_{D(B^\top X_0)} \circ \phi^{-1}_{D(B^\top X_0)}\left(C_0\left[I_m \otimes \left\{B^\top (X_i - X_0)\right\}\right]\right), \quad (3.7)$$

where only up to first order approximation is considered and $\phi^{-1}_{D(B^\top X_0)}$ is the inverse map of $\phi_{D(B^\top X_0)}$. We write $f\{g(\cdot)\}$ as $f \circ g$ and $\otimes$ is the Kronecker product. In the above expressions, $D(B^\top X_0)$ serves as the 0-order term in Taylor expansion and C_0 is the derivative matrix of $f(B^\top X_i)$ at $B^\top X_0$ with the structure

$$C_0 = \begin{pmatrix} c_{11}^\top(X_0) & \cdots & c_{1m}^\top(X_0) \\ \vdots & \ddots & \vdots \\ c_{m1}^\top(X_0) & \cdots & c_{mm}^\top(X_0) \end{pmatrix}_{m\times md}, \quad (3.8)$$

where $c_{kl}(X_0) = c_{lk}(X_0) \in R^d, k, l = 1, ..., m$. The subscript "0" in C_0 indicates its relation with X_0. Inserting (3.7) into (3.6), we get $\hat{\sigma}^2_B(B^\top X_0)$. Similar to Euclidean MAVE, minimizing $\sum_{j=1}^n \hat{\sigma}^2_B(B^T X_j)$ can be approximated by

$$\min_{\substack{B^T B = I,\\ D(B^\top X_j), C_j}} \sum_{j=1}^n \sum_{i=1}^n d^2\Big\{Y_i, \mathrm{Exp}_{D(B^\top X_j)} \circ \phi^{-1}_{D(B^\top X_j)}\left(C_j\left[I_m \otimes \{B^\top (X_i - X_j)\}\right]\right)\Big\} w_{ij}. \quad (3.9)$$

We call the above formulation intrinsic MAVE (iMAVE) since we derive it without any information of the ambient space. As a by-product of MAVE, the outer product of gradients estimation (OPG) has a similar form to MAVE. Immediately we have intrinsic OPG (iOPG) formulated as:

$$\min_{D(B^\intercal X_j), C_j} \sum_{i=1}^{n} d^2 \left(Y_i, \mathrm{Exp}_{D(B^\intercal X_j)} \circ \phi^{-1}_{D(B^\intercal X_j)} \left[C_j \{ I_m \otimes (X_i - X_j) \} \right] \right) w_{ij}, j = 1, ..., p, \tag{3.10}$$

where $D(B^\intercal X_j) \in \mathrm{Sym}^+(m)$ and C_j is $m \times mp$ in (3.10).

3.3 Algorithms under the log-Euclidean Metric

Our intrinsic models (3.9) and (3.10) can produce estimated $\hat{B}$ once the Riemannian metric in $\mathrm{Sym}^+(m)$ is specified. The choice of the Riemannian metric does have an impact on the complexity of optimization of (3.9) and (3.10). For example, in local linear regression for SPD matrices, Yuan et al. (2012) had to employ an annealing evolutionary stochastic approximation Monte Carlo algorithm to estimate coefficients when $\mathrm{Sym}^+(m)$ is endowed with the affine-invariant metric since the object function under this metric is neither convex nor possesses closed-form solutions. We here circumvent this dilemma by adopting the log-Euclidean metric and the log-Cholesky metric. As shown below, these two metrics not only help us derive our models in a simpler manner, but also pave the way for theoretical analysis.

3.3 Algorithms under the log-Euclidean Metric

The log-Euclidean metric is proposed by Arsigny et al. (2007). The matrix logarithm operation log: $\mathrm{Sym}^+(m) \to \mathrm{Sym}(m)$ and its inverse exp are both deffeomorphisms. Because $\mathrm{Sym}(m)$ has an additive group structure, to obtain a group structure in $\mathrm{Sym}^+(m)$, one can simply transport the additive structure of $\mathrm{Sym}(m)$ to $\mathrm{Sym}^+(m)$. More precisely, for $S_1, S_2 \in \mathrm{Sym}^+(m)$, define an operation $\oplus$ by

$$S_1 \oplus S_2 = \exp\{\log(S_1) + \log(S_2)\}. \tag{3.11}$$

Then $(\mathrm{Sym}^+(m), \oplus)$ is a commutative Lie group whose identity element is the identity matrix. Additionally, the Lie group exponential map $\mathfrak{exp}$ and Lie logarithm map $\mathfrak{log}$ are given by the matrix exponential "exp" and logarithm "log". The geodesic distance between $S_1, S_2 \in \mathrm{Sym}^+(m)$ under the log-Euclidean metric is $d(S_1, S_2) = \|\log S_1 - \log S_2\|_F$ where $\|\cdot\|_F$ is the Frobenius norm. Thus by (3.6), we have

$$d\left\{Y_i, D(B^\top X_i)\right\} = \|\log\{D(B^\top X_i)\} - \log Y_i\|_F.$$

Since $\log\{D(B^\top X_i)\}$ and $\log Y_i$ coincide with $\mathfrak{log}\{D(B^\top X_i)\}$ and $\mathfrak{log} Y_i$ which always reside in $T_{I_m}\mathrm{Sym}^+(m)$, no parallel transportation is needed. Directly expand $\log\{D(B^\top X_i)\}$ by Taylor series expansion and we get iMAVE under the log-Euclidean metric:

$$\min_{\substack{B:B^\top B=I \\ a_j, b_j}} \sum_{j=1}^{n} \sum_{i=1}^{n} w_{ij} \|a_j + b_j \left[I_m \otimes \{B^\top(X_i - X_j)\}\right] - \log Y_i\|_F^2, \tag{3.12}$$

and similarly iOPG under the log-Euclidean metric:

$$\min_{a_j,b_j} \sum_{i=1}^{n} w_{ij} \| a_j + b_j \{ I_m \otimes (X_i - X_j) \} - \log Y_i \|_F^2, \quad j = 1, ..., p. \tag{3.13}$$

where $w_{ij} = K_h(B^\top(X_i - X_j)) / \sum_{i=1}^n K_h(B^\top(X_i - X_j))$, a_j is an $m \times m$ symmetric matrix and b_j in (3.12) and (3.13) has the same structure as $D(B^\top X_j)$ in (3.9) and (3.10). Algorithms for (3.12) and (3.13) resemble classic MAVE and OPG in Xia (2007) and detailed procedures can be found in the supplementary material.

In practice, the structural dimension d in $B_{p\times d}$ is usually unknown and we now propose a cross validation procedure to determine it. Suppose l is the working dimension and d is the true dimension. Define

$$\hat{a}_{l0,j} = \sum_{i=1, i\neq j}^{n} K_{h_l}^{(i,j)} \mathrm{vecs}(\log Y_i) \Big/ \sum_{i=1, i\neq j}^{n} K_{h_l}^{(i,j)},$$

$$\mathrm{CV}(l) = \frac{1}{n} \sum_{j=1}^{n} ||\mathrm{vecs}(\log Y_i) - \hat{a}_{l0,j}||_F^2 \quad (l = 1, ..., p).$$

where $K_{h_l}^{(i,j)} = K_{h_l}(\hat{B}_l^\top (X_i - X_j))$ and $|| \cdot ||_F$ is the Frobenius norm. We then estimate d as $\hat{d} = \arg\min_{1\leq l\leq p} \mathrm{CV}(l)$.

Theorem 1. *Suppose assumptions (A1)-(A3) stated in section 3.4 hold. Then* $\lim_{n\to\infty} P(\hat{d} = d) = 1$.

Theorem 1 shows that the probability of choosing the right dimension tends to 1 as the sample size increases. Overall, the estimation of CMS is two-step: 1) for each $1 \leq l \leq p$, run iMAVE or iOPG to get estimated $\hat{B}_l$ and consequently CV(l); 2) the l with the smallest CV value is the chosen dimension and the corresponding $\hat{B}_l$ provides an eventual estimation of CMS.

The log-Cholesky metric is introduced by Lin (2019) and it shares similar merits to the log-Euclidean metric. When $\mathrm{Sym}^+(m)$ is endowed with the log-Cholesky metric, the geodesic distance between $S_1, S_2 \in \mathrm{Sym}^+(m)$ is $d(S_1, S_2) = \|\mathrm{chol}(L_1) - \mathrm{chol}(L_2)\|_F$. Here L_1, L_2 are Cholesky factors of S_1, S_2 ($L_1 L_1^\intercal = S_1$ such that the diagonal elements of L_1 are positive) and $\mathrm{chol}(L) = \lfloor L \rfloor + \log \mathbb{D}(L)$ where $\lfloor L \rfloor$ is the strict lower triangle part of L and $\mathbb{D}(L)$ the diagonal part of L. For any $S \in \mathrm{Sym}^+(m)$ and its Cholesky factor L, $\mathrm{chol}(L)$ belongs to $T_{I_m}\mathrm{Sym}^+(m)$ and no parallel transport is needed. Consequently substituting $\log(\cdot)$ in the log-Euclidean case for $\mathrm{chol}(\cdot)$ and keeping other things unchanged, we get iMAVE, iOPG under the log-Cholesky metric. Details are omitted.

3.4 Asymptotic Analysis

In this section we assume the structural dimension d in B_0 is given and consider theoretical properties of $\hat{B}$ from iMAVE and iOPG with the log-Euclidean metric. The case of the log-Cholesky metric is much the same.

We assume the model on $(\mathrm{Sym}^+(m), \oplus)$ by

$$Y = g(B_0^\top X) \oplus \varepsilon, \tag{3.14}$$

which is a modification of the Euclidean MAVE (3.2). Recall that with $\oplus$ defined in (3.11), $(\mathrm{Sym}^+(m), \oplus)$ is a commutative Lie group and the log-Euclidean metric is bi-invariant that turns $\mathrm{Sym}^+(m)$ into a Hadamard manifold.

Proposition 1. *Under the log-Euclidean metric, (3.14) is equivalent to*

$$\log Y = \log\{g(B_0^\top X)\} + \log\varepsilon. \tag{3.15}$$

Proposition 1 turns the model (3.14) defined on a Riemannian manifold into a Euclidean model defined in the vector space $T_{I_m}\mathrm{Sym}^+(m)$. Actually (3.15) coincides with the multivariate MAVE introduced in Zhang (2021).

Denote $h(B_0^\top X) = \log\{g(B_0^\top X)\}$. Since $h(B_0^\top X)$ is an $m\times m$ symmetric matrix, denote its (k,l)-th component as $h_{kl}, 1 \le l \le k \le m$. Let $\mu_B(u) = E(X \mid B^\top X = u)$, $w_B(u) = E(XX^\top \mid B^\top X = u)$. We need the following assumptions for (3.15) to prove our theoretical results.

(A1) [Design of X and Y] The density function $f(x)$ of X has bounded second order derivatives; $E|X|^k < \infty$ for some $k > 8$; $E|y_{kl}|^3 < \infty$ for every component y_{kl} in $\log Y, 1 \le l \le k \le m$; the functions $\mu_B(u), \omega_B(u)$

have bounded derivatives w.r.t. u and B for B in a small neighborhood of $B_0 : |B - B_0| \leq \delta$ for some $\delta > 0$.

(A2) [Link function] The link function $h_{kl}(u) = E(y_{kl} \mid B^\top X = u)$ has bounded fourth order derivatives w.r.t. u and B for B in a small neighborhood of B_0.

(A3) [Kernel function] $K_0(u)$ is a univariate symmetric density function with bounded second order derivatives and a compact support.

(A4) [Efficient dimension] The matrix $M_{\text{SPD}} = E\left\{\sum_{k=1}^{m}\sum_{l=1}^{k} h_{kl}^{(1)}(B_0^\top X)h_{kl}^{(1)}(B_0^\top X)^\top\right\}$ has full rank d, where $h_{kl}^{(1)} \in R^d$ is the derivative vector of h_{kl}.

(A5) [Bandwidth] The bandwidth $h_0 = c_1 n^{-r_h}$. For $t \geq 1$, $h_t = \max\{n^{-r_h/2}h_{t-1}, c_2 n^{-r_h'}\}$ where $0 < r_h \leq 1/(p_0 + 6)$, $0 < r_h' \leq 1/(d+3)$, $p_0 = \max\{p, 3\}$ and c_1, c_2 are constants.

The moment requirement on X in (A1) is not strong and we impose a lightly higher order moment condition than second moment for y_{kl} to apply Lemma 6.6 in Xia (2006) in our proof. The quadratic kernel and the Epanechnikov kernel are included in (A3). Intuitively assumption (A4) indicates that the dimension d cannot be further reduced. Assumption (A5) is made to ensure the convergence of algorithms of iMAVE and iOPG. Denote "vec" as the vectorization operator.

Theorem 2. *Under assumptions (A1)-(A6), $\hat{B}_{\rm iMAVE}$ from (3.12) satisfies*

$$\|\hat{B}_{\rm iMAVE}\hat{B}_{\rm iMAVE}^{\mathsf{T}} - B_0B_0^{\mathsf{T}}\|_F = O(h^3 + h\delta_{dh} + \delta_{dh}^2/h + n^{-1/2})$$

in probability as $n \to \infty$, where $\delta_{dh} = (nh^d/\log n)^{-1/2}$. If $h^3 + h\delta_{dh} + \delta_{dh}^2/h = o(n^{-1/2})$, then

$$\sqrt{n}\left\{\mathrm{vec}(\hat{B}_{\rm iMAVE}\hat{B}_{\rm iMAVE}^{\mathsf{T}}B_0) - \mathrm{vec}(B_0)\right\} \xrightarrow{d} N(0, W_{\rm SPD}^{+}\Sigma_{\rm SPD}W_{\rm SPD}^{+}).$$

Theorem 3. *Under assumptions (A1)-(A6), $\hat{B}_{\rm iOPG}$ from (3.13) satisfies*

$$\|\hat{B}_{\rm iOPG}\hat{B}_{\rm iOPG}^{\mathsf{T}} - B_0B_0^{\mathsf{T}}\|_F = O(h^3 + h\delta_{dh} + n^{-1/2})$$

in probability as $n \to \infty$, where $\delta_{dh} = (nh^d/\log n)^{-1/2}$. If $h^3 + h\delta_{dh} = o(n^{-1/2})$, then

$$\sqrt{n}\left\{\mathrm{vec}(\hat{B}_{\rm iOPG}\hat{B}_{\rm iOPG}^{\mathsf{T}}B_0) - \mathrm{vec}(B_0)\right\} \xrightarrow{d} N(0, W_0^{\rm SPD}).$$

The asymptotic covariance matrices of iMAVE and iOPG in Theorem 2 and 3 are detailed in the supplementary material. The above results of consistency and asymptotic normality are consistent with those in Xia et al (2002), Xia (2007) and Zhang (2021) and the proof follows a similar pattern to them as well. Our iMAVE shares the merit of classic MAVE that it can achieve a faster consistency rate even without undersmoothing the nonparametric link function estimator.

4. Adaptations

4.1 When $p = p_n \to \infty$ as $n \to \infty$

In this section we aim to include the diverging-dimensional case allowing $p = p_n \to \infty$ as sample size $n \to \infty$ in our intrinsic MAVE (3.12) and OPG (3.13). Following Cai et al. (2022), the main idea is to utilize the distance correlation (Székely et al., 2007) to define a window of for the local linear regression so that it is able to estimate gradients efficiently.

To make a distinction from the fixed-p-dimensional dimension reduction, we use the superscript $[j]$ in the following notations to indicate the jth component of a p-dimensional vector. Now denote the predictor $X = (X^{[1]}, ..., X^{[p]})^\mathsf{T}$ with diverging dimension $p = p_n \to \infty$ as $n \to \infty$. Recall that $B_0 = (\beta_1, ..., \beta_d)$ is a $p \times d$ orthogonal matrix ($B_0^\mathsf{T} B_0 = I_d$ with $d < p$). For each $k \in \{1, ..., d\}$, write $\beta_k = (\beta_k^{[1]}, ..., \beta_k^{[p]})^\mathsf{T}$. Denote the row vectors in B_0 by $\beta^{[j]} = (\beta_1^{[j]}, ..., \beta_d^{[j]})$, $j = 1, ..., p$. Since B_0 is orthogonal,

$$d = \sum_{k=1}^{d} \|\beta_k\|^2 = \sum_{k=1}^{d} \sum_{j=1}^{p} (\beta_k^{[j]})^2 = \sum_{j=1}^{p} \|\beta^{[j]}\|^2,$$

where $\|\cdot\|$ is the Euclidean norm.

The most significant difference between diverging-dimensional OPG and MAVE and ordinary ones is the choice of bandwidths in the multivariate kernel function $K(\cdot)$. Suppose $(X_i, Y_i), i = 1, ..., n$ are random observations.

Let $K(u)$ be a kernel function on R^p and

$$K_h(u, \alpha) = K(\frac{u^{[1]}}{h^{\alpha_1}}, ..., \frac{u^{[p]}}{h^{\alpha_p}})/h^{|\alpha|},$$

where bandwidth $h = h_n \to 0$, $\alpha = (\alpha_1, ..., \alpha_p)$ and $|\alpha| = \sum_{j=1}^{p} \alpha_j$. The diverging-dimensional intrinsic MAVE under the log-Euclidean metric is

$$\min_{\substack{B:B^\top B=I \\ a_j, b_j}} \sum_{j=1}^{n} \sum_{i=1}^{n} \|\log Y_i - a_j - b_j \left[I_m \otimes \{B^\top (X_i - X_j)\} \right] \|_F^2 K_h(X_i - X_j; \alpha) \tag{4.16}$$

and the diverging-dimensional intrinsic OPG under the same metric is

$$\min_{a_j, b_j} \sum_{i=1}^{n} \|\log Y_i - a_j - b_j \{I_m \otimes (X_i - X_j)\}\|_F^2 K_h(X_i - X_j; \alpha), j = 1, ..., p. \tag{4.17}$$

The indices $\alpha_1, ..., \alpha_p$ for (4.16) and (4.17) are chosen as follows. Define

$$\alpha_j = \mathrm{dCor}(R_j, X^{[j]}), \quad j = 1, ..., p,$$

where R_j is the residual of linear regression of $\log Y$ on $X^{[j]}$ and $\mathrm{dCor}(\cdot, \cdot)$ is the distance correlation coefficient introduced by Székely et al. (2007). As argued in Cai et al. (2022), in the conventional kernel smoothing, $\alpha_j = 1$ is used uniformly for all $j \in \{1, ..., p\}$. However, if $X^{[j]}$ contributes more linearly to the response, α_j should be smaller resulting in a bigger bandwidth for $X^{[j]}$. On the contrary, if $X^{[j]}$ contributes more nonlinearly, α_j should be bigger resulting in a smaller bandwidth for the calculation of partial derivative along $X^{[j]}$. It can be seen that α_j measures the nonlinear dependence between log$Y$ and $X^{[j]}$ and it is 0 if their dependence is either purely linear or is 0. Typically $\alpha = (\alpha_1, ..., \alpha_p)$ in (4.16) and (4.17) is replaced by its estimation $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_p)$ from samples. Since diverging-dimensional intrinsic OPG and MAVE under the log-Cholesky metric is almost the same as the log-Euclidean metric, we only present diverging-dimensional methods under the log-Euclidean metric (4.16), (4.17) and call them DMAVE and DOPG for short. The implementation details of DOPG and DMAVE resemble iOPG and iMAVE except that DOPG is one-step, i.e., B will not be refined by iteration. We in the following present the consistency results of $\hat{B}$ for DMAVE and DOPG.

(B1) Covariate X has a compact support in R^p and the response $y_{kl}, 1 \leq l \leq k \leq m$ is almost surely bounded. Suppose $|\alpha| = \alpha_1 + ... + \alpha_p = o(\text{log}n)$ as $n \to \infty$.

(B2) The kernel function $K(\cdot)$ is bounded with a compact support in R^p and it is Lipschitz continuous, i.e., $|K(u) - K(v)| \leq C\|u - v\|$ for some positive constant C.

(B3) As $n \to \infty$, dimension $p = p_n \to \infty$ and $p_n^2/n \to 0$. The bandwidth $h_n \to 0$ such that $p_n\text{log}n(nh_n^{|\alpha|})$ and $\omega_n = \sum_{\alpha_j \neq 0} h_n^{\alpha_j}\|\beta^{[j]}\| \to 0$.

(B4) The matrix S(X) (detailed in the proof in the supplementary material) is almost surely invertible, and the smallest eigenvalue of $E\{S(X) \mid X\}$ is bounded away from 0 almost surely.

(B5) Assume $\sum_{\alpha_j \neq 0} p_n^2 \log n/(nh_n^{|\alpha|+2\alpha_j}) \to 0$ and $\sum_{\alpha_j \neq 0} p_n \omega_n^2/h_n^{\alpha_j} \to 0$ as $n \to \infty$.

Assumptions (B1), (B3) and (B5) are technically necessary for the consistency and they are justified in Cai et al. (2022). Assumption (B1) is made to ensure there exists $h_n \to 0$ and $nh_n^{|\alpha|} \to \infty$. Assumption (B3) is made for the requirement of kernel regression: $h \to 0$ and a neighbor with diameter h contains diverging number of observations. The Lipschitz condition in (B2) is also satisfied by the Epanechnikov kernel and the quadratic kernel. Assumption (B4) is commonly used in kernel regression.

Theorem 4. *Under Assumptions (B1)-(B5) and (A2), (A4), we have*

$$\hat{B}_{\text{DMAVE}}\hat{B}_{\text{DMAVE}}^{\mathsf{T}} - B_0 B_0^{\mathsf{T}} = O_P(p_n\sigma_n);$$

$$\hat{B}_{\text{DOPG}}\hat{B}_{\text{DOPG}}^{\mathsf{T}} - B_0 B_0^{\mathsf{T}} = O_P(p_n\sigma_n),$$

where $\sigma_n = \{\sum_{\alpha_j \neq 0}(c_n^{[j]})^2 + \sum_{\alpha_j = 0} p_n/n\}^{1/2}$ *with* $c_n^{[j]} = (p_n \log n/nh_n^{|\alpha|+2\alpha_j})^{1/2} + \omega_n^2/h_n^{\alpha_j}$. *Hence, if* $\sum_{\alpha_j \neq 0} p_n^3 \log n/nh_n^{|\alpha|+2\alpha_j} \to 0$ *and* $\sum_{\alpha_j \neq 0} p_n^2(\omega_n^2/h_n^{\alpha_j})^2 \to 0$ *hold,*

$$|\hat{B}_{\text{DMAVE}}\hat{B}_{\text{DMAVE}}^{\mathsf{T}} - B_0 B_0^{\mathsf{T}}| \to 0;$$

$$|\hat{B}_{\text{DOPG}}\hat{B}_{\text{DOPG}}^{\mathsf{T}} - B_0 B_0^{\mathsf{T}}| \to 0,$$

as $n \to \infty$, *where* $|A|$ *represents the largest absolute value of entries in matrix* A.

We only consider consistency of $\hat{B}_{\mathrm{DMAVE}}$ and $\hat{B}_{\mathrm{DOPG}}$ in Theorem 4 and the conclusions coincide with Theorem 1 and 2 in Cai et al. (2022). Based on restrictions in the theorem and assumptions (B3) and (B5), p_n is allowed to diverge at a speed of $o(n^{2/(|\alpha|+4)})$.

4.2 General Riemannian Manifolds

Recall that Proposition 1 transforms (3.14) into the Euclidean model (3.15). However if the chosen metric is not bi-invariant or if the manifold of interest is a general manifold other than $\mathrm{Sym}^+(m)$, one usually cannot derive a Euclidean model by this way. Let $X \in R^p$ and $Y \in \mathcal{M}$ where $(\mathcal{M}, \langle\cdot,\cdot\rangle)$ is a general s-dimensional Riemannian manifold. In this case, as Lin et al. (2022) did, we make the assumption that the model takes the form:

$$\mathrm{Log}_\mu Y = h(B_0^\top X) + \zeta \tag{4.18}$$

where $\mu = \arg\min_{y\in\mathcal{M}} E\{d^2(Y,y)\}$ is the Fréchet mean of Y. Model (4.18) is defined in $T_\mu\mathcal{M}$ and we still aim at estimating B_0. It is obvious that (4.18) coincides with the multivariate MAVE developed by Zhang (2021).

In model (4.18), the only concern is the existence of the Fréchet mean. For general Riemannian manifolds whose sectional curvature is positive, the Fréchet mean may not exist and therefore additional conditions are needed for (4.18).

(C1) The minimizer of the Fréchet function $Ed^2(\cdot, Y)$ exists and is unique.

This is automatically satisfied when $\mathcal{M}$ is $\mathrm{Sym}^+(m)$ equipped with either the log-Euclidean metric or the log-Cholesky metric.

For a subset A of $\mathcal{M}$, A^ϵ denotes the set $\cup_{p\in A} B(p;\epsilon)$ where $B(p;\epsilon)$ is the ball with center p and radius ϵ in $\mathcal{M}$. We use $\mathrm{Im}^{-\epsilon}(\mathrm{Exp}_\mu)$ to denote the set $\mathcal{M} \setminus \{\mathcal{M} \setminus \mathrm{Im}(\mathrm{Exp}_\mu)\}^\epsilon$. In order to define $\mathrm{Log}_{\hat\mu} Y_i$ at least with a dominant probability for a large sample, we assume

(C2) There is some constant $\epsilon_0 > 0$ such that $\mathrm{pr}\{Y \in \mathrm{Im}^{-\epsilon_0}(\mathrm{Exp}_\mu)\}$=1.

The condition (C2) is only needed when $\mathcal{M}$ is not a Hadamard manifold. If (C1) and (C2) are satisfied, (4.18) is well defined.

Next we establish the consistency and asymptotic normality of the iMAVE and iOPG estimators under the general manifold case in model (4.18). We consider a manifold $\mathcal{M}$ that satisfies one of the following conditions:

(M1) $\mathcal{M}$ is a finite-dimensional Hadamard manifold having sectional curvature bounded from below by $\mathfrak{c}_0 < 0$.

(M2) $\mathcal{M}$ is a complete compact Riemannian manifold.

An example satisfying (M1) is $\mathrm{Sym}^+(m)$ endowed with the log-Euclidean metric, the log-Cholesky metric or the affine-invariant metric while the unit sphere serves as an example satisfying (M2).

We have to treat $\phi\mathrm{Log}_{\hat{\mu}}Y_i - \mathrm{Log}_{\mu}Y_i$ during our proof where ϕ is short for $\phi_{\hat{\mu},\mu}$. The method in Lin and Yao (2019) is applied here to write $\phi\mathrm{Log}_{\hat{\mu}}Y_i - \mathrm{Log}_{\mu}Y_i$ as $\{-H_i(\mu)+\Delta_i(\hat{\mu})\}\mathrm{Log}_{\mu}\hat{\mu}$ and the asymptotic normality of $\mathrm{Log}_{\mu}\hat{\mu}$ helps us control the discrepancy between $\mathrm{Log}_{\hat{\mu}}Y_i$ and $\mathrm{Log}_{\mu}Y_i$. Above $\Delta_i(\hat{\mu}) = o_P(1)$ and $H_i(y) = -(\nabla Z_i)(y)$ acting on vector fields U, V by $\langle H_i U, V\rangle(y) = \langle -\nabla_U Z_i, V\rangle(y) = \mathrm{Hess}_y\{d^2(y, Y_i)/2\}(U, V)$. Here Z_i is a vector field with $Z_i(y) = \mathrm{Log}_y Y_i$ and "Hess" denotes the Hessian matrix (Kendall and Le, 2011). To make above reasoning valid, following conditions are needed.

(C3) $\mathcal{M}$ satisfies at least one of the conditions (M1) and (M2).

(C4) For all $y \in \mathcal{M}$, $E\{d^2(y, Y)\} < \infty$.

(C5) For some constant $\mathfrak{c}_1 > 0$, $F(y) - F(\mu) \geq \mathfrak{c}_1 d^2(y, \mu)$ when $d(y, \mu)$ is sufficiently small.

(C6) $\lambda_{\min}\{E(H_t)\} > 0$ where $\lambda_{\min}(\cdot)$ is the smallest eigenvalue of an operator or a matrix.

Conditions (C3)-(C6) are standard assumptions also made by Lin et al. (2022), Kedall and Le (2011) and Lin and Yao (2019). (C4) is analogous to the moment condition in the Euclidean case. (C5) is satisfied for Hadamard manifolds with $c_2 = 1$ according to the lemma S.7 of Lin and Müller (2021). (C6) is made to ensure H_i is invertible.

The skeleton of the theoretical proof in the general manifold case is similar to classic MAVE methods. Thus not only (C1)-(C6), but also standard assumptions made in Section 3.4 are needed here. Terms in model (4.18) are all s-dimensional vectors. Denote the k-th component of h as $h_k, k = 1, ..., s$. Substitute y_k, h_k for y_{kl}, h_{kl} in conditions (A1)-(A5). Replace the matrix M_{SPD} in condition (A4) with $M_0 = E\{h^{(1)}(B_0^\top X)^\top h^{(1)}(B_0^\top X)\}$ where $h^{(1)} = \nabla h(B_0^\top X) \in R^{s\times d}$. Denote the modified conditions as (A1')-(A5'). We can derive results similar to Theorem 2 and 3 under assumptions (A1')-(A5') and (C1)-(C6), which are moved to the supplementary material to avoid duplication.

5. Simulations and Real Data Applications

5.1 Simulation Study I

In simulation I and II, the structural dimension d is assumed as known. We test the performance of our proposed iMAVE with log-Euclidean metric (eu-iMAVE), iOPG with log-Euclidean metric (eu-iOPG), iMAVE with log-Cholesky metric (ch-iMAVE), iOPG with log-Cholesky metric (ch-iOPG), weighted inverse regression ensemble method (WIRE, Ying and Yu (2022)), Fréchet MAVE and Fréchet OPG (fMAVE and fOPG, Zhang et al. (2021)) for SPD matrix-valued responses.

In simulation I, we generate Y similar to Lin et al. (2022). Let the predictors $X_1, X_2, ..., X_p$ be independently and identically sampled from the uniform distribution on $[0, 1]$. Fix μ to be the identity matrix. Set $Y = \mu \oplus w(X_1, ..., X_p) \oplus \zeta$, where $\oplus$ is defined in (3.11) and $w(X_1, ..., X_p) = \mathfrak{exp}\phi_{\mu,e} f(X_1, ..., X_p)$ with the following two settings for f:

I-1: $f(X_1, ..., X_p) = f_{12}(X_1, X_2)$, where $f_{12}(X_1, X_2)$ is an $m \times m$ matrix with (j, l)-entry being $\exp\{-1/|j - l|\} \sin[2\pi\{X_1 + X_2 - 1/(j + l)\}]$;

I-2: $f(X_1, ..., X_p) = \sum_{k=1}^{2} f_k(X_k)$ where $f_k(X_k)$ is an $m \times m$ matrix with (j, l)-entry being $\exp\{-1/|j - l|\} \sin[2\pi\{X_k - 1/(j + l)\}]$.

We set $m = 3$. The random noise ζ is generated according to $\mathfrak{log}\zeta = \sum_{i=1}^{6} Z_j v_j$, where $Z_1, ..., Z_6$ are independently sampled form $N(0, 0.1^2)$ and $v_1, ..., v_6$ are a basis of the tangent space $T_e\mathrm{Sym}^+(m)$. Note that μ is identical with e so $\phi_{\mu,e}$ is just the identity map. We adopt the log-Euclidean metric so that $\mathfrak{exp} = \exp$ and $\mathfrak{log} = \log$. In model I-1, $d = 1$ and $B_0 = (1, 1, 0, ..., 0)^\top/\sqrt{2}$. In model I-2, $d = 2$ and $B_0 = (\beta_1, \beta_2)^\top$, where $\beta_1 = (1, 0, ..., 0)^\top/\sqrt{2}$ and $\beta_2 = (0, 1, 0, ..., 0)^\top/\sqrt{2}$. We take $(p, n) = (20, 200), (20, 500), (40, 200), (40, 500)$ and each combination is replicated for 50 times. The means and standard deviations of the estimation errors $\|\hat{B}\hat{B}^\top - B_0B_0^\top\|_F$ are listed in Table 1 from which we can summarize that in all scenarios except $p = 40$ in model II-2, our methods either with the

Model	(p, n)	WIRE	eu-iOPG	eu-iMAVE	ch-iOPG	ch-iMAVE	fOPG	fMAVE
I-1	(20,200)	1.3944	0.9606	0.9518	0.9367	0.9022	1.1798	1.1798
		±0.0246	±0.6283	±0.6334	±0.6113	±0.6079	±0.0135	±0.0134
	(20,500)	1.3231	0.0345	0.0307	0.0348	0.0318	1.1175	1.1228
		±0.1109	±0.0044	±0.0037	±0.0041	±0.0040	±0.0911	±0.0835
	(40,200)	1.3902	1.3739	1.3418	1.3973	1.3927	1.1790	1.1790
		±0.0262	±0.0446	±0.1154	±0.0235	±0.0181	±0.0102	±0.0101
	(40,500)	1.3850	1.3911	1.3948	0.9505	0.9445	1.1794	1.1794
		±0.0037	±0.0284	±0.0218	±0.6339	±0.6406	±0.0021	±0.0021
I-2	(20,200)	1.4108	0.1842	0.0914	0.5568	0.5433	1.3179	1.3497
		±0.0452	±0.1578	±0.0045	±0.6140	±0.6305	±0.0437	±0.0244
	(20,500)	1.3977	0.0536	0.0487	0.0586	0.0534	1.1876	1.3402
		±0.0184	±0.0015	±0.0029	±0.0018	±0.0035	±0.0313	±0.0104
	(40,200)	1.5266	1.8300	1.7490	1.8891	1.8768	1.3839	1.3692
		±0.0330	±0.1052	±0.1886	±0.1187	±0.1395	±0.0187	±0.0271
	(40,500)	2.1420	1.9718	1.9051	2.0575	2.0429	1.7067	1.8143
		±.0194	±0.1133	±0.1494	±0.0016	±0.0200	±0.0917	±0.1465

Table 1: Mean (± standard deviation) of estimation errors for different methods in model I-1 and I-2.

log-Euclidean metric or the log-Cholesky metric achieve the minimum errors. And it can be expected that results of $p = 40$ can be improved by a larger sample size n.

5.2 Simulation Study II

We in this section consider that $p = p_n$ diverges and test the performance of our newly developed diverging-dimensional methods. Let $\beta_1^\intercal = (1, 1, 0, ..., 0)/\sqrt{2}$, $\beta_2^\intercal = (0, ...0, 1, 1)/\sqrt{2}$. The predictors $X_1, X_2, ..., X_p$ are independent random variables each from the uniform distribution on $[0, 1]$. We generate n i.i.d samples $(X_{1i}, X_{2i}, ..., X_{pi}), i = 1, ..., n$. Let $M(X)$ be matrices specified by the following models:

II-1: $M(X) = \begin{pmatrix} 1 & \rho(X) \\ \rho(X) & 1 \end{pmatrix}, \rho(X) = \{\exp(\beta_1^\intercal X) - 1\}/\{\exp(\beta_1^\intercal X) + 1\};$

II-2: $M(X) = \begin{pmatrix} 1 & \rho_1(X) & \rho_1(X) & \rho_2(X) & \rho_2(X) \\ \rho_1(X) & 1 & \rho_2(X) & \rho_2(X) & \rho_2(X) \\ \rho_1(X) & \rho_2(X) & 1 & \rho_2(X) & \rho_1(X) \\ \rho_2(X) & \rho_2(X) & \rho_2(X) & 1 & \rho_1(X) \\ \rho_2(X) & \rho_2(X) & \rho_1(X) & \rho_1(X) & 1 \end{pmatrix},$

$\rho_1(X) = 0.2\{\exp(\beta_1^\intercal X) - 1\}/\{\exp(\beta_1^\intercal X) + 1\}$ and $\rho_2(X) = 0.2\sin(\beta_2^\intercal X)$.

We generate $Y = \exp\{\log(M(X)) + \sigma Z\}$ where Z has independent $N(0, 1)$ diagonal elements and independent $N(0, 1/2)$ off-diagonal elements.

In model II-1, $m = 2$, $B_0 = \beta_1$ and $d = 1$. In model II-2, $m = 5$, $B_0 = (\beta_1, \beta_2)$ and $d = 2$. Model II-1,II-2 are also considered in Zhang et al. (2021).

We choose the sample size n as 100 and 200, and for each n, we set $p \in \{10/n, n/5, n/2, 4n/5, 1.5n\}$. In every combination of (n, p), we run model II-1 and II-2 for 50 times and examine the mean estimation errors. For the clarity of display, we only plot errors of WIRE, eu-iOPG, DOPG, ch-iOPG and fOPG in Figure 1. It can be seen that first, in most scenarios our intrinsic OPG and diverging-dimensional OPG outperform WIRE and fOPG; second, our methods can produce accurate estimates when $p \leq 4n/5$ and more samples may be needed for better estimates when p is as large as $1.5n$ which is consistent with simulation results of Cai et al. (2022).

For lack of space, simulation studies about the determination of the structural dimension d and about the general manifold case can be found in the supplementary material.

5.3 New York Taxi Network Data

In this section, we apply our methods to the New York Taxi network data. The New York City Taxi and Limousine Commission provides records on pick-up and drop-off dates and times, pick-up and drop-off locations, trip

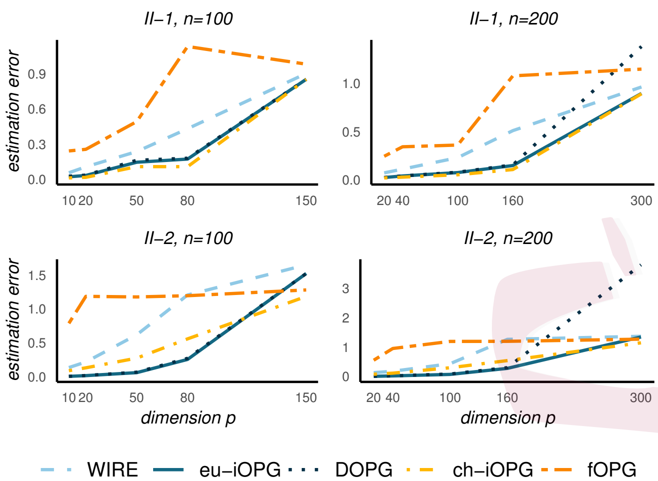

Figure 1: Estimation errors of different methods with dimension $p = n/10$, $n/5$, $n/2$, $4n/5$ and $1.5n$.

distances, itemized fares, payment types and other information for yellow taxis (Tucker et al., 2021). The data are available at `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`

Eventually we collect 1416 3×3 SPD matrices as the realizations of the response describing the intensity of taxi movements between three zones in Manhattan. Additionally we collect 14 predictive variables. Details of data processing and collection can be found in the supplementary material.

We randomly divide the dataset into a training set (991 samples) and a test set (425 samples). On the training set, respectively setting $d = 1, ..., 7$, we apply a cross-validation procedure to calculate CV(d). The result is:

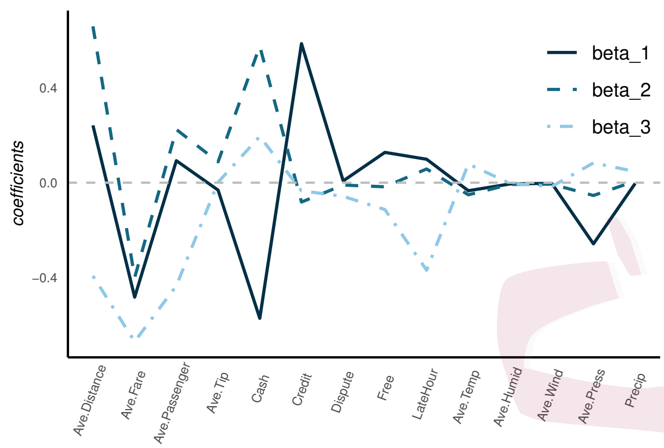

Figure 2: Coefficients of three estimated CMS directions.

0.0430, 0.0283, 0.0257, 0.0626, 0.0834, 0.0687, 0.0612, which suggests that $\hat{d} = 3$ is a reasonable choice. So we apply iMAVE with $d = 3$ again to the training set and get $\hat{B} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ which is plotted in Figure 2.

Coefficients with larger absolute values in Figure 2 indicate more significance of corresponding predictors. The estimated results show that fare amount and type of payment are important covariates, which is consistent with the results of Tucker et al. (2021). Ave.Fare and Ave.Distance are closely related and both of them are significant in the first three directions. Cash and Credit are significant in the first direction, showing that most passengers tend to pay the fare by cash or credit. Another obvious observation is that all the 5 weather variables seem negligible since their coefficients are almost 0 in all of the first three directions. This is reasonable because the weather condition during January and February 2019 was rather stationary, which accounts for the insignificance of weather variables.

To show our dimension reduction methods have further statistical applications, we conduct the regression on our data using the manifold additive model (MAM) introduced by Lin et al. (2022). The MAM is formulated as $Y = \mu \oplus w_1(X_1) \oplus ... \oplus w_q(X_q) \oplus \xi$, where Y is an SPD matrix, μ is the Fréchet mean of Y, each w_k is function mapping X_k into the SPD space, ξ is random noise which has a Fréchet mean corresponding to the group identity element, $X_1, ..., X_q$ are scalar variables and $\oplus$ is the group operation.

We apply MAM to the dimension-reduced training set to get estimated $\hat{\mu}$ and functions $\hat{w}_1$, $\hat{w}_2$ and $\hat{w}_3$. Then we apply the trained MAM to the test set to estimate the response Y. The prediction RMSE on the test set is 0.3220, which shows MAM generates good estimation after processing data with our intrinsic dimension reduction methods and indicates our methods are ready for more applications.

6. Discussion

Further improvements can be expected from our proposed methods. For example, a penalty term can be utilized in combination with our method to get the penalized iMAVE for simultaneous dimension reduction and variable screening. Specifically, a group-LASSO penalty can be considered to improve our method as group-wise iMAVE for sparse ultra-high dimensional dimension reduction with SPD-valued responses.

Supplementary Materials

Contain: 1) algorithms for iOPG and iMAVE; 2) expressions of asymptotic covariance matrices in Theorem 2 and 3; 3) convergence results of iOPG and iMAVE on a general manifold; 4) a simulation study testing the CV procedure of choosing the structural dimension d and a simulation study under the general manifold case; 5) details of data collection and processing in the New York taxi network application; 6) all proofs of theoretical results that appear in this paper.

Acknowledgments

The authors thank the Editor, Associate Editor, and two anonymous reviewers for their constructive feedback on earlier versions of this paper. The authors contributed equally to this paper and are listed in alphabetical order. The research is supported by the National Key R&D Program of China (Grant No. 2023YFA1008700 and 2023YFA1008703), the National Natural Science Foundation of China (Grant No. 12371289), the Basic Research Project of Shanghai Science and Technology Commission (Grant No. 22JC1400800) and the Shanghai Pilot Program for Basic Research (Grant No. TQ20220105).

References

Arsigny, V., P. Fillard, X. Pennec, and N. Ayache (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications 29*, 328–347.

Batchelor, P. G., M. Moakher, D. Atkinson, F. Calamante, and A. Connelly (2004). A rigorous framework for diffusion tensor calculus. *Magnetic Resonance in Medicine 53*, 221–225.

Bhattacharjee, S. and H.-G. Müller (2021). Single index Fréchet regression. arXiv:2108.05437 [stat.ME].

Cai, Z., Y. Xia, and W. Hang (2022). An outer-product-of-gradient approach to dimension reduction and its application to classification in high dimensional space. *Journal of the American Statistical Association 118*(543), 1671–1681.

Chen, Y., Z. Lin, and H.-G. Müller (2020). Wasserstein regression. arXiv:2006.09660 [stat.ME].

Cook, R. D. and B. Li (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics 30*, 455–474.

Cook, R. D. and S. Weisberg (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association 86*, 328–332.

Cornea, E., H. Zhu, P. Kim, and J. G. Ibrahim (2016). Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*, 463–482.

Dubey, P. and H.-G. Müller (2020). Functional models for time-varying random objects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*, 275–327.

Fletcher, P., C. Lu, S. Pizer, and S. Joshi (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging 23*, 995–1005.

Huettel, S. A., A. W. Song, and G. McCarthy (2008). *Functional magnetic resonance imaging*. Sinauer Associates.

Kendall, W. S. and H. Le (2021). Limit theorems for empirical fréchet means of independent and non-identically distributed manifold-valued random variables. *Brazilian Journal of Probability and Statistics 25(3)*, 323–352.

Lang, S. (1999). *Fundamentals of Differential Geometry*. Springer New York.

Li, B. (2018). *Sufficient Dimension Reduction*. Chapman and Hall/CRC.

Li, B. and S. Wang (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association 102*, 997–1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*, 316–327.

Lin, Z. (2019). Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications 40*, 1353–1370.

Lin, Z. and H.-G. Müller (2021). Total variation regularized fréchet regression for metric-space valued data. *The Annals of Statistics 49(6)*, 3510–3533.

Lin, Z., H.-G. Müller, and B. U. Park (2022). Additive models for symmetric positive-definite matrices and Lie groups. *Biometrika*.

Lin, Z. and F. Yao (2019). Intrinsic riemannian functional data analysis. *The Annals of Statistics 47*, 3533–3577.

Ma, Y. and L. Zhu (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association 107*, 168–179.

Ma, Y. and L. Zhu (2013). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics 41*, 250–268.

Ma, Y. and L. Zhu (2019). Semiparametric estimation and inference of variance function with large dimensional covariates. *Statistica Sinica 29*, 567–588.

Pennec, X., P. Fillard, and N. Ayache (2006). A Riemannian framework for tensor computing. *International Journal of Computer Vision 66*, 41–66.

Petersen, A. and H.-G. Müller (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics 47*, 691–719.

Schwartzman, A. (2006). Random ellipsoids and false discovery rates: statistics for diffusion tensor imagining data. *PhD Thesis*, Stanford University.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*, 2769–2794.

Terras, A. (1985). *Harmonic Analysis on Symmetric Spaces and Applications I.* Springer New York.

Tu, L. W. (2011). *An Introduction to Manifolds.* Springer New York.

Tucker, D. C., Y. Wu, and H.-G. Müller (2021). Variable selection for global Fréchet regression. *Journal of the American Statistical Association*, 1–15.

Wang, T., P. Xu, and L. Zhu (2013). Penalized minimum average variance estimation. *Statistica Sinica 23*, 543–569.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory 22*, 1112–1137.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics 35*, 2654–2690.

Xia, Y., H. Tong, W. K. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*, 363–410.

Ying, C. and Z. Yu (2022). Fréchet sufficient dimension reduction for random objects. *Biometrika*.

Yuan, Y., H. Zhu, W. Lin, and J. S. Marron (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*, 697–719.

Zhang, H. (2021). Minimum average variance estimation with group lasso for the multivariate response central mean subspace. *Journal of Multivariate Analysis 184*.

Zhang, Q., L. Xue, and B. Li (2021). Dimension reduction and data visualization for Fréchet regression. arXiv:2110.00467 [stat.ME].

Zhu, C. and H.-G. Müller (2021). Autoregressive optimal transport models. arXiv:2105.05439 [stat.ME].

Zhu, H., Y. Chen, J. G. Ibrahim, Y. Li, C. Hall, and W. Lin (2009). Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association 104*, 1203–1212.

School of Statistics, East China Normal University, Shanghai, China

E-mail: 52214404011@stu.ecnu.edu.cn

E-mail: shuangdai95@163.com

E-mail: zyu@stat.ecnu.edu.cn