Submitted to the Annals of Applied Statistics

VARIABLE SCREENING AND SPATIAL SMOOTHING IN FRÉCHET REGRESSION WITH APPLICATION TO DIFFUSION TENSOR IMAGING

BY LEI YAN[1], XIN ZHANG[1,a] ZHOU LAN[2], DIPANKAR BANDYOPADHYAY[3], YICHAO WU[4], AND FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

[1]*Department of Statistics, Florida State University, Tallahassee, FL,* [a]*henry@stat.fsu.edu*

[2]*Department of Radiology at Brigham and Women's Hospital and Harvard Medical School, Boston, MA*

[3]*Department of Biostatistics, Virginia Commonwealth University, Richmond, VA*

[4]*Department of Mathematics, Statistics, and Computer Science, The University of Illinois, Chicago, IL*

Modern applications in medical imaging often include high-dimensional predictors and spatially dependent responses in the non-Euclidean space. For example, in imaging-genetics studies, our objective is to study the relationship between single-nucleotide polymorphisms (SNPs), a high-dimensional predictor vector, and diffusion tensor imaging (DTI) responses, which are thousands to millions of voxel-wise 3×3 symmetric positive definite (SPD) matrices. In this paper, We develop a fast and pragmatic method of regressing spatially associated random responses on a high-dimensional predictor set. Specifically, we focus on two related problems: fast variable screening of high-dimensional predictors and smoothing techniques for non-Euclidean spatially associated responses. Under a Fréchet regression framework (which handles regression of SPD matrix-variate responses on covariates in Euclidean space), we propose a two-stage approach, where a screening method (using distance covariances in metric spaces) is employed to mitigate high-dimensionality (Stage 1), followed by deriving a closed-form solution that powers elegant smoothing of the spatially associated SPD responses (Stage 2). We investigate the finite-sample properties of our method using synthetic data generated under various settings, and present illustration via analysis of an imaging-genetics (DTI responses with genetic and demographic predictors) dataset, derived from the Alzheimer's Disease Neuroimaging Initiative 2. Code for implementing our proposed method is available in the `GitHub` link: https://github.com/leiyan-ly/Frechet-regression.

1 Introduction Diffusion tensor (magnetic resonance) imaging (DTI), which presents a rotationally-invariant description of the shape of water diffusion (O'Donnell and Westin, 2011), has been used extensively to investigate the microstructure of white matter of the brain (Soares et al., 2013). Aging and pathological processes of the central nervous system can alter the microstructure of affected tissues. Since the diffusion of water within tissues is sensitive to these tissue microstructural changes, this technique has been used to study various neurological disorders, such as Alzheimer's disease (AD; Stebbins and Murphy, 2009; Esrael et al., 2021). Unlike conventional imaging data, DTI generates a 3×3, symmetric, positive definite (SPD) matrix at each voxel, proportional to the covariance matrix of the local three-dimensional (3D) Brownian motion of water molecules (Dryden, Koloydenko and Zhou, 2009). Unlike the isotropic (equal in all directions, with no directionality) diffusion process manifested in water molecules within a cup, microscopic tissue heterogeneity triggers anisotropic diffusion, such as faster (or, slower) diffusion measured parallel (or, perpendicular) to the axons in a white matter fiber tract. The diffusion tensor characterizes this

Keywords and phrases: Diffusion tensor imaging, Fréchet regression, Distance covariance, Sure screening property, Smoothing.

anisotropy, and its decomposition into eigenvectors and eigenvalues provides insight into the orientation and quantity of diffusion, respectively. The three orthogonal eigenvectors establishes the local fiber coordinate system, while the three eigenvalues represent the diffusivity in the direction of each eigenvector. Positive eigenvalues ensure non-negative diffusivity in all directions (negative diffusivity is improbable), and aligns with the intrinsic properties of diffusion processes (Mori, 2007; Alexander et al., 2007). Taken together, the eigenvectors and eigenvalues define an ellipsoid that represents an isosurface of the (Gaussian) diffusion probability (O'Donnell and Westin, 2011).

The motivation for this paper comes from DTI data generated under the Alzheimer's Disease Neuroimaging Initiative 2 (ADNI2) Study, which recorded diffusion tensors at a large number of voxels, along with genetic variations, e.g., single nucleotide polymorphisms (SNPs). Our central objective is to construct a regression framework for this DTI data (from ADNI2) with diffusion tensors $\mathbf{Y}$ as the response collected at the voxel level within a specific region of interest (ROI), and the SNPs and other available demographic information as predictors $\mathbf{X}$. Along the way, the two key clinical goals that we plan to achieve are (a) identifying important predictors (e.g., which SNPs are related to AD status?), and (b) detecting regions of differences in the brain between two groups (e.g., AD versus normal controls), thereby revealing abnormal topological organization (Lo et al., 2010).

The DTI data from ADNI2 presents newer challenges while employing existing statistical methods. To begin with, it contains over half a million SNPs and fewer than 200 subjects. Even after successful preprocessing steps, the number of SNPs from candidate disease genes is usually greater than the sample size, leading to complications in conducting fast and reliable variable selection from a high-dimensional predictor set when the response is a complex object, such as the SPD matrices from DTI. By eliminating irrelevant predictors, we can improve the accuracy and interpretability of statistical models and reduce overfitting (Chowdhury and Turin, 2020). Furthermore, previous studies (Xue, Bowman and Kang, 2018) have revealed that proximally-located/neighboring voxels may exhibit similar neuronal activity and similar disease status (compared to distally-located voxels), and incorporating this spatial dependence generally improves the precision of estimates by borrowing strength from other regions, leading to efficient and valid inference (Spence et al., 2007; Lan et al., 2022). Existing spatial smoothing methods are mostly model-based and designed for real-valued responses. Directly incorporating SPD matrix-valued responses and spatial information in a non-parametric/model-free way is more desirable, within the framework of random object (non-Euclidean) response regression (Petersen and Müller, 2019; Jeon, Park and Van Keilegom, 2021). Finally, the sheer size of voxels in the DTI data calls for the development of scalable algorithms for screening over the predictor domain, and smoothing over the response domain. This motivates our computationally efficient, two-stage method, that addresses these challenges effectively.

In the existing literature, diffusion tensors are routinely converted to fractional anisotropy (FA), a commonly-used scalar quantity (Ennis and Kindlmann, 2006). FA is a projection of the eigenvalues to a univariate response, quantifying the degree of anisotropy in the diffusion process. FA lies between 0 and 1, with higher/lower values implying enhanced/disrupted connectivity between regions. Many variable selection methods are suitable for scalar responses (Fan and Lv, 2010; Graña et al., 2011; Demirhan et al., 2015; Kronlage et al., 2018; Heinze, Wallisch and Dunkler, 2018), and several Bayesian models have been proposed to incorporate spatial information for univariate modeling (Woolrich et al., 2004; Johnson et al., 2013; Reich et al., 2018). However, these univariate mappings lose statistical information contained in the matrix structure as different diffusion tensors could have the same scalar projections. To this end, matrix-variate models have been proposed via parameterizing the diffusion tensors as matrix-variate random distributions (Schwartzman, 2016; Lee and Schwartzman, 2017;

Lan, Reich and Bandyopadhyay, 2021), which has shown promise in capturing the spatial dependence, and detecting disease prevalence at neighboring voxels (Wu et al., 2013; Xue, Bowman and Kang, 2018). Recently, Lan et al. (2022) studied how predictors affect diffusion tensors using geostatistical regression models. Existing methods for modeling DT responses are overtly parametric, relying on the assumed distributions, e.g., Wishart. Moreover, these matrix-variate models are not equipped to screen out inactive predictors, and incorporate spatial information within a unified framework, thereby failing to address the complexities related to modeling the ADNI2 DTI data.

To alleviate the aforementioned challenges, we cast the problem within the Fréchet regression (FR) framework proposed by Petersen and Müller (2019), and propose a two-stage modeling that involves screening, and smoothing. FR is flexible in that it handles complex types of responses (e.g., distribution functions, SPD matrices, and Riemannian manifold-valued data) by associating the (conditional) barycenter (Fréchet mean) of the responses $\mathbf{Y}$ to predictors $\mathbf{X}$ in Euclidean space. FR does not require the response to follow any particular distribution, but only assumes some mild conditions on the geometry of the metric space. Recently, several methods have been developed for random objects valued in metric space. For example, Dubey and Müller (2019) extended analysis of variance for random objects by deriving a central limit theorem for the Fréchet variance, while Qi Zhang and Li (2023) introduced sufficient dimension reduction methods in the context of FR, thereby offering a valuable visual inspection tool for employing FR. Furthermore, Lin, Müller and Park (2023) proposed an additive regression model for SPD matrix-valued response and scalar predictors that capitalizes on the inherent Abelian group structure characterizing the manifold of the SPD space. However, when the number of predictors p is large, it is important to select the subset of predictors that drives the changes in the response. Very recently, Tucker, Wu and Müller (2023) proposed Fréchet ridge selection operator (`FRiSO`), which can perform variable selection by imposing a sparsity-encouraging penalty in the optimization step of FR. Unfortunately, the method is computationally expensive as the number of tuning parameters is proportional to the number of predictors.

Our two-stage method proceeds as follows. Inspired by Li, Zhong and Zhu (2012) and Lyons (2013), we propose a fast, and easy to implement screening method for the high-dimensional predictors in Stage 1, which only requires the distance functions within covariances in the estimation. We also establish the sure screening property (Fan and Lv, 2008) under very general conditions. The proposed screening method is flexible, generalizable, widely applicable to a variety of random object-valued responses, and is thus not limited to SPD matrix responses, or DTI data. In Stage 2, we propose a smoothing method specifically designed to handle spatially associated SPD matrix responses. The method is based on the weighted Fréchet mean, which fits into the FR framework but makes no restrictive model assumptions. We derive a closed-form solution for the smoothed matrix, leading to a highly efficient and scalable algorithm. Our proposal offers several advantages over existing methods, such as (a) direct modeling of the SPD matrix response, bypassing often-restrictive underlying assumptions, such as lognormal (Schwartzman, 2016), or Wishart distributions (Lee and Schwartzman, 2017), and (b) efficient inclusion of spatial information, without relying on specific modeling assumptions.

The rest of this paper is organized as follows. In Section 2, we propose our two-stage model, study the theoretical properties, and outline the implementation algorithm. After evaluating the finite sample properties of our method via simulation studies in Section 3, we illustrate our method via application to the ADNI2 data in Section 4. Finally, Section 5 concludes, alluding to future directions. Additional numerical results and technical proofs are relegated to the Supplementary Materials (Yan et al., 2024).

2 Method We present our two-stage method for modeling DTI data as the responses in regression. In Section 2.1, we propose a flexible and computationally efficient variable selection procedure and study its statistical properties. In Section 2.2, we introduce a smoothing method that accounts for spatial dependence and enhances the efficiency of global Fréchet regression for spatially-associated SPD responses. Finally, Section 2.3 details the implementation algorithm and discusses tuning parameter selection.

2.1 ***Stage I****: Variable screening in regression of object-valued response on high-dimensional predictor* Our proposed general screening procedure is based on distance covariances for random objects data defined in metric spaces of the strong negative type. Although $\mathbf{X} \in \mathbb{R}^p$ in our applications, we state our screening procedure in a more general setting where $\mathbf{X}$ consists of p predictors in metric space. Let $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_p\}$ be predictor objects taking values in $\mathcal{X} = \prod_{k=1}^{p} \mathcal{X}_k$, with $\mathbf{X}_k \in \mathcal{X}_k$ for all $1 \leq k \leq p$, and $\mathbf{Y}$ be response objects taking values in $(\mathcal{Y}, d_{\mathcal{Y}})$. The distance function associated with each $\mathcal{X}_k$ is $d_{\mathcal{X}_k}$ and a product metric $d_{\mathcal{X}}$ can be defined on $\mathcal{X}$, according to Deza and Deza (2009). Our distance covariance (DC) screening method is widely applicable to various types of responses, including but not limited to SPD matrices, probability distributions, and spherical data, all of which are demonstrated in simulations in Section 3.2. We begin with a review of FR and DC in metric spaces in Subsection 2.1.1, leading to development of the variable screening method in Subsection 2.1.2, followed by a discussion of its theoretical properties in Subsection 2.1.3.

2.1.1 Background The pioneering work of Petersen and Müller (2019) introduced FR, a regression tool that can be applied to complex responses in general metric spaces. In FR, $\mathcal{X}$ represents the Euclidean space $\mathbb{R}^p$; $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote population mean and covariance of $\mathbf{X}$, respectively, and $\mathbf{x}, Y$ denote realizations of $\mathbf{X}, \mathbf{Y}$. The key idea of FR is to find the conditional expectation of $\mathbf{Y}$ given $\mathbf{X}$. However, we are not able to directly generalize the classic conditional expectation to this setting, since a metric space does not have an algebraic structure. Based on the definition of the Fréchet mean (Fréchet, 1948),

$$\omega_{\oplus} = \underset{\omega \in \mathcal{Y}}{\operatorname{argmin}} \, \mathrm{E}[d_{\mathcal{Y}}^2(\mathbf{Y}, \omega)]$$

Petersen and Müller (2019) introduced the global FR as

$$\mathbf{Y}(\mathbf{x}) = \underset{\omega \in \mathcal{Y}}{\operatorname{argmin}} \, M(\omega, \mathbf{x}), \quad M(\omega, \mathbf{x}) = \mathrm{E}[s(\mathbf{X}, \mathbf{x}) d_{\mathcal{Y}}^2(\mathbf{Y}, \omega)], \tag{2.1}$$

where, the weight function $s(\mathbf{z}, \mathbf{x}) = 1 + (\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$. The global FR is equivalent to multiple linear regression, if $(\mathcal{Y}, d_{\mathcal{Y}})$ is the Euclidean space. Given a sample $\{(\mathbf{x}_i, Y_i) : i = 1, 2, \ldots, n\}$, an estimator of $\mathbf{Y}(\mathbf{x})$ is given by

$$\widehat{\mathbf{Y}}(\mathbf{x}) = \underset{\omega \in \mathcal{Y}}{\operatorname{argmin}} \sum_{i=1}^{n} [1 + (\mathbf{x} - \overline{\mathbf{x}})^T \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}})] d_{\mathcal{Y}}^2(Y_i, \omega), \tag{2.2}$$

where, $\overline{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the sample mean and sample covariance matrix, respectively.

To conduct variable selection within FR, the aforementioned `FRiSO` (Tucker, Wu and Müller, 2023) initiates individually penalized ridge regression in terms of squared distances, where the estimation involves solving a sequence of constrained optimization problems that are computationally expensive and complex, rendering it infeasible under high-dimensional settings. This motivates our proposal of DC based variable screening.

Lyons (2013) introduced DC defined on general metric spaces. For separable metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, let θ be the probability measure on the product space $\mathcal{X} \times \mathcal{Y}$, and $\mu := \pi_1(\theta), \nu := \pi_2(\theta)$ be, respectively, the marginal probability measures on $\mathcal{X}$ and $\mathcal{Y}$, where π_1

and π_2 are the coordinate projections onto the marginal spaces. For any $k > 0$, a finite signed Borel measure on $(\mathcal{X}, d_{\mathcal{X}})$ is said to have finite k-th moment if $\int d^k_{\mathcal{X}}(x, o)d|\mu|(x) < \infty$, for some $o \in \mathcal{X}$. Once we assume $\pi_1(|\theta|)$ and $\pi_2(|\theta|)$ have finite first moments, we can define the DC $\mathrm{dcov}(\theta)$.

Let $\mathbf{X}, \mathbf{Y}$ have a joint distribution $(\mathbf{X}, \mathbf{Y}) \sim \theta$, with marginals $\mathbf{X} \sim \mu$ and $\mathbf{Y} \sim \nu$. We define $\mathrm{dcov}(\mathbf{X}, \mathbf{Y}) := \mathrm{dcov}(\theta)$. From Jakobsen (2017), we have

(2.3)
$$\mathrm{dcov}(\mathbf{X}, \mathbf{Y}) = \mathrm{E}\, d_{\mathcal{X}}(\mathbf{X}, \mathbf{X}')d_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}') + \mathrm{E}\, d_{\mathcal{X}}(\mathbf{X}, \mathbf{X}')\, \mathrm{E}\, d_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}') - 2\,\mathrm{E}\, d_{\mathcal{X}}(\mathbf{X}, \mathbf{X}')d_{\mathcal{Y}}(\mathbf{Y}, \mathbf{Y}''),$$

where, $(\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')$ and $(\mathbf{X}'', \mathbf{Y}'')$ are mutually independent random objects with the same distribution θ. From the definition, we do not need to assume the form of the conditional measure of $\mathbf{Y}|\mathbf{X}$. Let $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ be a random sample of $(\mathbf{X}, \mathbf{Y})$. The V-statistic estimator of $\mathrm{dcov}(\mathbf{X}, \mathbf{Y})$ is given by

$$\widehat{\mathrm{dcov}}(\mathbf{X}, \mathbf{Y}) = \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3, \tag{2.4}$$

where,

$$\widehat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j)d_{\mathcal{Y}}(Y_i, Y_j)$$

$$\widehat{S}_2 = \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j)\right)\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{\mathcal{Y}}(Y_i, Y_j)\right)$$

$$\widehat{S}_3 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} d_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_l)d_{\mathcal{Y}}(Y_j, Y_l).$$

This estimator is purely data-driven and highly efficient, only requiring computation of distances between observations. Lyons (2013) showed that $\mathrm{dcov}(\mathbf{X}, \mathbf{Y})$ can be used as an indicator of independence for a certain subset of metric spaces.

PROPOSITION 1. *If $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are metric spaces of strong negative type, then $\mathrm{dcov}(\mathbf{X}, \mathbf{Y}) = 0$ is equivalent to $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$.*

More details on DC, including technical reviews on strong negative spaces, are provided in Supplement A. Due to the simplicity and flexibility of distance covariances for measuring statistical independence, we are now ready to introduce our variable screening procedure.

2.1.2 Proposed Variable Screening Under high-dimensional settings, the number of predictors p is one or several orders of magnitude larger than the sample size n. We usually assume that only a small group of predictors is relevant to the response variable. Without specifying any regression model, we define the important variables as follows.

DEFINITION 1. *A set $\mathcal{I} \subseteq \{1, 2, \ldots, p\}$ is called the important predictor set if $\mathcal{I}$ is the smallest set satisfying $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}_{\mathcal{I}^c}|\mathbf{X}_{\mathcal{I}}$, where $\mathbf{X}_{\mathcal{I}^c} = \{\mathbf{X}_k : k \in \mathcal{I}^c\}$ and $\mathbf{X}_{\mathcal{I}} = \{\mathbf{X}_k : k \in \mathcal{I}\}$.*

The primary goal is to estimate the important predictor set $\mathcal{I}$. We define $\omega_k = \mathrm{dcov}(\mathbf{X}_k, \mathbf{Y})$ and $\widehat{\omega}_k = \widehat{\mathrm{dcov}}(\mathbf{X}_k, \mathbf{Y})$. Our procedure utilizes the DC defined in (2.3) to rank the importance of $\mathbf{X}_k$ to $\mathbf{Y}$. We choose DC, since it allows any regression relationship between $\mathbf{X}$ and $\mathbf{Y}$ in the general metric space settings, such as probability density functions, symmetric

positive definite matrices, or Riemannian manifolds. To proceed, we first compute $\widehat{\omega}_k$, and then select the predictors with large $\widehat{\omega}_k$, i.e., an estimator of $\mathcal{I}$ is given by

$$\widehat{\mathcal{I}} = \{k : \widehat{\omega}_k \geq cn^{-\kappa}\},$$

where, c and κ are threshold values defined later. However, in practice, we use the following Lemma 1 with computational complexity $O(pn^2)$ to obtain $\widehat{\omega}_k$, instead of the direct implementation in (2.4), which requires a higher computational cost of $O(pn^3)$. This approach is analogous to the computation method used for distance covariance in Euclidean space (Székely, Rizzo and Bakirov, 2007; Chaudhuri and Hu, 2019). The proof of Lemma 1 appears in Supplement B.

LEMMA 1. *Define* $a_{ij} = d_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j)$, $\bar{a}_{i,\cdot} = 1/n\sum_{j=1}^n a_{ij}$, $\bar{a}_{\cdot,j} = 1/n\sum_{i=1}^n a_{ij}$, $\bar{a}_{\cdot\cdot} = 1/n^2\sum_{i,j=1}^n a_{ij}$, $A_{ij} = a_{ij} - \bar{a}_{i,\cdot} - \bar{a}_{\cdot,j} + \bar{a}_{\cdot\cdot}, i, j = 1, 2, \ldots, n$. *Similarly define* $b_{ij} = d_{\mathcal{Y}}(Y_i, Y_j)$, $\bar{b}_{i,\cdot} = 1/n\sum_{j=1}^n b_{ij}$, $\bar{b}_{\cdot,j} = 1/n\sum_{i=1}^n b_{ij}$, $\bar{b}_{\cdot\cdot} = 1/n^2\sum_{i,j=1}^n b_{ij}$ $B_{ij} = b_{ij} - \bar{b}_{i,\cdot} - \bar{b}_{\cdot,j} + \bar{b}_{\cdot\cdot}, i, j = 1, 2, \ldots, n$. *Then we have* $\widehat{\mathrm{dcov}}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{i,j=1}^n A_{ij}B_{ij}$, *where* $\widehat{\mathrm{dcov}}(\mathbf{X}, \mathbf{Y})$ *is defined in* (2.4).

2.1.3 Theoretical Properties We now obtain the convergence rate of $\widehat{\omega}_k$, and prove that our method has the sure screening property under the following mild assumptions.

ASSUMPTION 1. *There exists a positive constant* s_0 *such that for all* $0 < s \leq s_0$, $\mathrm{E}[\exp(sd^2_{\mathcal{X}_k}(\mathbf{X}_k, o_k))] < \infty$, *for* $k = 1, \ldots, p$ *and* $\mathrm{E}[\exp(sd^2_{\mathcal{Y}}(\mathbf{Y}, o))] < \infty$, *where* $o_k \in \mathcal{X}_k, k = 1, \ldots, p$ *and* $o \in \mathcal{Y}$.

ASSUMPTION 2. *The minimum DC of important predictors satisfies* $\min_{k\in\mathcal{I}}\omega_k \geq 2cn^{-\kappa}$, *where* $c > 0$ *and* $0 \leq \kappa < 0.5$.

ASSUMPTION 3. $(\mathcal{X}_k, d_{\mathcal{X}_k})$ *and* $(\mathcal{Y}, d_{\mathcal{Y}}), k = 1, 2, \ldots, p$ *are metric spaces of strong negative type.*

We make a few remarks about these assumptions. Assumption 1 is a generalization of Condition 1 in Li, Zhong and Zhu (2012), and is satisfied when $\mathbf{X}$ and $\mathbf{Y}$ are bounded, almost surely. We point out that this is a reasonable condition. For FR, the consistency of estimators in Petersen and Müller (2019) relies on boundedness of the predictors and responses, while the variable selection consistency in Tucker, Wu and Müller (2023) also assumes boundedness. If we consider the Euclidean spaces, Assumption 1 holds when $\mathbf{X}$ and $\mathbf{Y}$ are sub-exponential distributions. Assumption 2 requires the DC of the active predictors to be distinguishable from those of the irrelevant variables. This concept is analogous to the irrepresentability assumption used in high-dimensional inference literature, e.g., Assumption (3.2) in Lee, Sun and Taylor (2015), which implies active predictors are not overly well-aligned with the inactive predictors. Similar assumptions are considered by Li, Zhong and Zhu (2012); Fan and Lv (2008) for real-valued predictors and responses, and serve as an identifiability condition ensuring identification of the active set. Likewise, Condition B in Tucker, Wu and Müller (2023) is also about identifiability, which assumes that the global FR will change, if we exclude any important predictor. Assumption 3 guarantees the equivalence of $\mathrm{dcov}(\mathbf{X}_k, \mathbf{Y}) = 0$ and independence of $\mathbf{X}_k$ and $\mathbf{Y}$. For FR, the predictors are in Euclidean space. Thus, metric spaces $(\mathcal{X}_k, d_{\mathcal{X}_k}), k = 1, \ldots, p$ are of strong negative type. A list of metric spaces that satisfy this assumption is provided in Supplement A. Finally, we have the following Theorem, whose proof is presented in Supplement C.

THEOREM 1. *Suppose Assumptions 1 holds. Then, for any* $0 < \gamma < 0.5 - \kappa$, *there exist positive constants* $c_1 > 0$ *and* $c_2 > 0$, *such that*

$$P\left(\max_{1\le k\le p}|\widehat{\omega}_k - \omega_k| \ge cn^{-\kappa}\right) \le O\left(p[\exp(-c_1 n^{1-2(\gamma+\kappa)}) + n\exp(-c_2 n^{\gamma})]\right). \tag{2.5}$$

If we further assume Assumption 2 and 3, then

$$P\left(\mathcal{I} \subseteq \widehat{\mathcal{I}}\right) \ge 1 - O\left(|\mathcal{I}|[\exp(-c_1 n^{1-2(\gamma+\kappa)}) + n\exp(-c_2 n^{\kappa})]\right), \tag{2.6}$$

where, $|\mathcal{I}|$ *is the cardinality of* $\mathcal{I}$.

REMARK 1. *The first part of the theorem, (2.5), does not depend on Assumption 3. Therefore, even if the strong negative type assumption does not hold, our screening method can still be used to identify important variables with large distance covariance with response, hence more relevant. Moreover, if Assumption 3 is satisfied, we obtain a stronger result in the second part (2.6), which ensures that the response is independent of unimportant variables when conditioned on the important variables, as defined in Definition 1.*

From the above theorem, we now obtain the same sure screening property as in Li, Zhong and Zhu (2012), but under a more general setting where $\mathbf{X}$ and $\mathbf{Y}$ are random objects in metric space. To balance the two terms on the right-hand side of (2.5), let $\gamma = (1 - 2\kappa)/3$. Then, $P\left(\max_{1\le k\le p}|\widehat{\omega}_k - \omega_k| \ge cn^{-\kappa}\right) \le O(p\exp(-c_1 n^{(1-2\kappa)/3}))$. Therefore, Theorem 1 indicates our distance covariance screening can handle the non-polynomial dimensionality of order $\log(p) = o(n^{(1-2\kappa)/3})$.

In contrast to the variable selection consistency of `FRiSO` in Tucker, Wu and Müller (2023) established under fixed-dimension predictor space, our theoretical results allow p to diverge and provide theoretical guarantees under ultra-high dimensionality. In addition, we provide a simplified and unified alternative to `FRiSO` variable selection. For varying response metric spaces $(\mathbf{Y}, d_{\mathcal{Y}})$, `FRiSO` requires optimization algorithms of various types to solve the objective functions, which are defined using the distance functions $d_{\mathcal{Y}}$. However, taking advantage of distance covariances, our screening procedure is valid for any response metric space, and bypasses the optimization steps, leading to faster implementation (compared to the `FRiSO`). Simulation results presented in Section 3.2 confirm that the proposed method can be hundreds of thousands of times faster than `FRiSO`.

2.2 ***Stage II****: Locally Spatial Smoothing for SPD matrix-valued Responses* At the conclusion of Stage I, the high-dimensional predictor space is drastically reduced to a small subset of the most desired variables. We are now in a position to fit FR to the SPD matrix-valued responses, and that constitutes Stage II. However, the traditional FR approach would ignore accommodating the (possible) spatial association between the responses at each location, as observed in our motivating DTI data. Hence, we introduce a simple smoothing method that accounts for this spatial dependence, and enhances the efficiency of multiple global FR regression for spatially-associated SPD matrix-valued responses.

Let N be the number of responses, $v \in \{v_1, v_2, \ldots, v_N\}$ be a spatial location, $\mathbf{Y}_v$ and $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$ denote the random object at location v and Euclidean predictor, respectively. Given $\mathbf{X} = \mathbf{x}$, $\mathbf{Y}_v(\mathbf{x})$ is the FR function and $\widehat{\mathbf{Y}}_v(\mathbf{x})$ is the global FR estimator for the response variable at location v. We assume the response space $(\mathcal{Y}, d_{\mathcal{Y}})$ is a space of SPD matrices, accommodating Cholesky distances. A typical DTI dataset includes diffusion tensors from subjects at various voxels and predictors for each subject. Let n be the number of subjects, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ be predictors of subject i, $Y_{i,v}$ be the observation

of subject i at voxel v, and $Y_v = \{Y_{i,v} : i = 1, 2, \ldots, n\}$ be the set of diffusion tensors at voxel v. Let P_1 and P_2 be two SPD matrices. Then, we can write $P_1 = (P_1^{1/2})^T P_1^{1/2}$ and $P_2 = (P_2^{1/2})^T P_2^{1/2}$ using Cholesky decomposition, where $P_1^{1/2}$ and $P_2^{1/2}$ are upper triangular, with positive diagonal elements. The Cholesky decomposition distance between P_1 and P_2 is defined as $\|P_1^{1/2} - P_2^{1/2}\|_F$, where $\|\cdot\|_F$ is the Frobenius norm. Using the relationship between the Frobenius norm and trace, we have

$$d_C(P_1, P_2) = \|P_1^{1/2} - P_2^{1/2}\|_F = \sqrt{\text{tr}\left\{(P_1^{1/2} - P_2^{1/2})^T (P_1^{1/2} - P_2^{1/2})\right\}}.$$

Several distance functions can be used for modeling SPD matrices (Dryden, Koloydenko and Zhou, 2009), including Euclidean distance, log-distance, and Riemannian distance. The choice of the above Cholesky distance, however, is made intentionally, such that both the FR fitting and the proposed spatial smoothing steps have closed-form solutions. This property greatly facilitates the development of efficient algorithms, suitable for large datasets.

We propose a simple non-parametric approach to exploit neighborhood information among the response SPD matrices. Specifically, we calculate the weighted Fréchet mean of $\widehat{\mathbf{Y}}_v(\mathbf{x})$, and its neighborhood. For a location v, let $\mathcal{J} = (j_1, j_2, j_3)$ be the 3-dimensional coordinates of voxel v. The neighborhood is defined as $\mathcal{N}_v = \{\mathcal{I} = (i_1, i_2, i_3) : |i_r - j_r| \leq 1, r = 1, 2, 3, \mathcal{I} \in \mathbf{O}\} \setminus \mathcal{J}$, where $\mathbf{O}$ is the set of coordinates for all voxels. The weights are $w_{\mathcal{J}} = 1, w_{v'} = u$, where $v' \in \mathcal{N}_v$, and $u \geq 0$ is the user-specified weight. Analogous to the definition of the Fréchet mean and FR, the spatially smoothed response is obtained by solving the following optimization problem:

$$\widehat{\mathbf{Y}}_v(\mathbf{x})^{\text{Smooth}} = \underset{\omega \in \mathcal{Y}}{\text{argmin}} \left[d_C^2(\widehat{\mathbf{Y}}_v(\mathbf{x}), \omega) + \sum_{v' \in \mathcal{N}_v} w_{v'} d_C^2(\widehat{\mathbf{Y}}_{v'}(\mathbf{x}), \omega) \right], \tag{2.7}$$

where $\widehat{\mathbf{Y}}_v(\mathbf{x})$ is the fitted value at location v from (2.2). We show in Lemma 2 and 3 below (proofs appear in Supplement B) that $\widehat{\mathbf{Y}}_v(\mathbf{x})$ and $\widehat{\mathbf{Y}}_v(\mathbf{x})^{\text{Smooth}}$ have closed-form solutions.

LEMMA 2. *Let* $\widehat{\mathbf{Y}}_v(\mathbf{x}) = \text{argmin}_{\omega \in \mathcal{Y}} \sum_{i=1}^n r_i(\mathbf{x}) d_C^2(Y_{i,v}, \omega)$, $r_i(\mathbf{x}) = 1 + (\mathbf{x} - \overline{\mathbf{x}})^T \widehat{\mathbf{\Sigma}}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}})$. *We have*

$$\widehat{\mathbf{Y}}_v(\mathbf{x}) = \left(\frac{\sum_{i=1}^n r_i(\mathbf{x}) Y_{i,v}^{1/2}}{\sum_{i=1}^n r_i(\mathbf{x})} \right)^T \left(\frac{\sum_{i=1}^n r_i(\mathbf{x}) Y_{i,v}^{1/2}}{\sum_{i=1}^n r_i(\mathbf{x})} \right).$$

LEMMA 3. *The optimization problem* (2.7) *has an explicit solution given by:*

$$\widehat{\mathbf{Y}}_v(\mathbf{x})^{\textit{Smooth}} = \left(\widetilde{w}_1 \widehat{\mathbf{Y}}_v(\mathbf{x})^{1/2} + \sum_{v' \in \mathcal{N}_v} \widetilde{w}_{v'} \widehat{\mathbf{Y}}_{v'}(\mathbf{x})^{1/2} \right)^T \left(\widetilde{w}_1 \widehat{\mathbf{Y}}_v(\mathbf{x})^{1/2} + \sum_{v' \in \mathcal{N}_v} \widetilde{w}_{v'} \widehat{\mathbf{Y}}_{v'}(\mathbf{x})^{1/2} \right),$$

where $\widehat{\mathbf{Y}}_v(\mathbf{x})^{1/2} = \sum_{i=1}^n r_i(\mathbf{x}) Y_{i,v}^{1/2} / (\sum_{i=1}^n r_i(\mathbf{x}))$, $\widehat{\mathbf{Y}}_{v'}(\mathbf{x})^{1/2} = \sum_{i=1}^n r_i(\mathbf{x}) Y_{i,v'}^{1/2} / (\sum_{i=1}^n r_i(\mathbf{x}))$, $r_i(\mathbf{x}) = 1 + (\mathbf{x} - \overline{\mathbf{x}})^T \widehat{\mathbf{\Sigma}}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}})$, $\widetilde{w}_1 = 1/(1 + \sum_{v' \in \mathcal{N}_v} w_{v'})$, *and* $\widetilde{w}_{v'} = w_{v'}/(1 + \sum_{v' \in \mathcal{N}_v} w_{v'})$.

The definition of neighborhoods is based on the fact that proximal locations are likely to have similar effects, and incorporating them into the smoothing method can improve the accuracy of the analysis, as posited by a variety of neuroimaging studies in the literature (Spence et al., 2007; Min and Mai, 2022; Wu et al., 2013; Xue, Bowman and Kang, 2018;

Lan, Reich and Bandyopadhyay, 2021). Following Min and Mai (2022), our neighborhood $\mathcal{N}_v$ is a $3D$ cube centered at voxel v with side length 3, ensuring that the smoothing method utilizes local and contiguous spatial information. If neighboring locations do not have all the 26 voxels, $\mathcal{N}_v$ contains existing voxels inside the cube. The user-defined weight u allows control of the degree of smoothness. A larger value gives more weight to neighboring locations, while a smaller value places more emphasis on the local mean. If $u = 0$, the smoothing method reduces to the normal FR.

Our method is highly flexible, allowing for easy integration of other definitions of neighborhoods and weights. For example, adaptive neighborhoods can be defined, where the size and shape of the neighborhood are adjusted based on the local spatial structure of the data. Instead of a binary weights strategy, other schemes can also be adopted, such as a Gaussian weighting (Lan et al., 2022) that assigns higher weights to closer neighbors and lower weights to farther neighbors. Despite these variations, the smoothed regression function $\widehat{\mathbf{Y}}_v(\mathbf{x})^{\text{Smooth}}$ retains the same closed-form solution, as shown in Lemma 3. Such flexibility makes our method readily amenable to a variety of applications. Note, $\mathbf{Y}_v(\mathbf{x})^{\text{Smooth}}$ is the weighted Fréchet mean of $\mathbf{Y}_v(\mathbf{x})$ and $\mathbf{Y}_{v'}(\mathbf{x})$, combining information from location v and its neighborhood. If $u = 0$, $\mathbf{Y}_v(\mathbf{x})^{\text{Smooth}} = \widehat{\mathbf{Y}}_v(\mathbf{x})$, implying no smoothness. As we increase u, $\mathbf{Y}_v(\mathbf{x})^{\text{Smooth}}$ becomes smoother across its neighborhood. In practice, we can use cross-validation to tune u. However, we observe that this is redundant when there is an underlying smoothness structure. In the subsequent simulation and data analysis, we observe that the results are not sensitive to u, when u is in a proper range. Following Min and Mai (2022), we suggest to choose $u \in [0.5, 1]$.

2.3 Algorithm and Tuning Parameter Selection Combining the variable screening procedure we developed in Subsection 2.1 and the smoothing method in Subsection 2.2 , we obtain a two-stage method for DTI analyses. The first stage of our method involves variable selection, followed by fitting the global FR on a voxelwise basis. Subsequently, for each voxel, the spatial information is incorporated by averaging its adjacent voxels. We summarize the two-stage method in Algorithm 1.

We discuss the choice of the model size d in Stage I. According to Theorem 1, the ideal estimated set of important predictors $\widehat{\mathcal{I}}$ depends on c and κ. However, the true values of c and κ are unknown. Based on the sparsity assumption that only $o(n)$ predictors are truly related to the response variables, we can select variables with the largest d distance covariance as suggested by Fan and Lv (2008). Therefore, we introduce two rules to determine the model size d in practice. The first choice is $d = \lfloor n/\log(n) \rfloor$, where $\lfloor a \rfloor$ denotes the integer part of a. This strategy is widely used in variable screening literature (Fan and Lv, 2008; Li, Zhong and Zhu, 2012; Cui, Li and Zhong, 2015; Liu, Li and Wu, 2014). The second rule is based on the Bayesian information criterion (BIC), defined as

$$\text{BIC} = \log\Big(\frac{1}{n}\sum_{i=1}^{n} d_{\mathcal{Y}}(\widehat{Y}_i, Y_i)^2\Big) + \frac{d}{n}\log(n),$$

where $\widehat{Y}_i$ is the Fréchet regression estimator using d variables and the squared distance $d^2_{\mathcal{Y}}(\cdot,\cdot)$ is used to resemble the squared errors in the usual linear regression. Specifically, if we use the Euclidean distance, then the BIC reduces to the standard BIC in a linear regression model. The model size with the smallest BIC is selected.

Algorithm 1: Smoothed Global Fréchet Regression with Variable Selection

Input : A DTI dataset $(\mathbf{x}_i, Y_{i,v}), i = 1, \ldots, n, v \in \{v_1, \ldots, v_N\}$, sparsity level d, weights vector $\mathbf{w}_v = \{w_{v'} : v' \in \mathcal{N}_v\}, v \in \{v_1, \ldots, v_N\}$

Output: Smoothed estimations $\widehat{\mathbf{Y}}_v(\mathbf{x})^{\text{Smooth}}, v \in \{v_1, \ldots, v_N\}$

Stage I:
for $v \in \{v_1, v_2, \ldots, v_N\}$ **do**
  compute DC $\widehat{\omega}_k = \widehat{\text{dcov}}(\mathbf{X}_k, \mathbf{Y}_v)$ using Lemma 1;
  select variables with the largest d DC and denote them by $\mathbf{x}_{\mathcal{I}}$;
  fit a global Fréchet regression of $\mathbf{Y}_v$ on $\mathbf{x}_{\mathcal{I}}$ and get the estimate $\widehat{\mathbf{Y}}_v(\mathbf{x}_{\mathcal{I}})$ by Lemma 2;
end
Stage II:
for $v \in \{v_1, v_2, \ldots, v_N\}$ **do**
  compute smoothed estimation $\widehat{\mathbf{Y}}_v(\mathbf{x}_{\mathcal{I}})^{\text{Smooth}}$ by Lemma 3;
end

3 Simulations The efficacy of the proposed two-stage method is assessed by demonstrating its effectiveness on synthetic datasets generated under spatial dependence and with high-dimensional predictors. In Subsection 3.1, we simulate DTI datasets employing the spatial Wishart process model to illustrate how the smoothing step improves prediction when responses are spatially dependent. Next, in Subsection 3.2, we evaluate the efficiency of our screening method on data simulated from three types of FR models, viz, SPD matrices with Cholesky metric, one-dimensional probability distributions with Wasserstein metric, and spherical data with geodesic distance.

3.1 Data generation: Spatial Wishart Process Model For simulating spatially-dependent DTI datasets, we utilize the spatial Wishart process (SWP) model of Lan et al. (2022). Although our method is model-free and does not exploit this specific model information, we observe encouraging performance. Let $\mathbf{x}_i$ denote the predictor vector for subject i, and $Y_{i,v}$ be the corresponding 3×3 SPD matrix (response) for subject $i \in \{1, 2, \ldots, n\}$ at voxel $v \in \{v_1, v_2, \ldots, v_N\}$, following a Wishart distribution $\mathcal{W}(\mathbf{\Sigma}_i(v), m)$. We decompose $Y_{i,v}$ as

$$Y_{i,v} = L_i(v) U_i(v) L_i(v)^T, \tag{3.8}$$

where, $L_i(v)$ is the lower triangular matrix from the Cholesky decomposition of $\mathbf{\Sigma}_i(v)$, and $U_i(v)$ is random Wishart matrix with mean I_3. The Wishart matrix $U_i(v)$ is the residual term and assumed spatially dependent. We use a SWP with degree of freedom m, cross-covariance matrix I, and correlation function $\mathcal{K}(v, v'|\phi) = \exp(-\|v - v'\|_2/\phi)$ to generate $\{U_i(v), v \in 1, 2, \ldots, v_N\}$ for $i \in \{1, 2, \ldots, n\}$. The covariate vector $\mathbf{x}_i$ is now regressed on the lower triangular $L_i(v)$ through its elements as follows:

$$\log l_{ikk}(v) = \mathbf{x}_i^T \beta_{kk}(v), \quad l_{ikl}(v) = \mathbf{x}_i^T \beta_{kl}(v) \quad k > l,$$

where, l_{ikl} is the element at row k column l of $L_i(v)$, $\beta_{kl}(v) = [\beta_{1kl}(v), \beta_{2kl}(v), \ldots, \beta_{pkl}(v)]^T$ is the vector of spatially-correlated coefficients. Define

$$\beta(v) = [\beta_{11}(v)^T, \beta_{22}(v)^T, \beta_{33}(v)^T, \beta_{21}(v)^T, \beta_{31}(v)^T, \beta_{32}(v)^T]$$

following a mean-zero Gaussian process with cross-covariance $\sigma^2 I$ and dependence function $\mathcal{K}(v, v'|\phi) = \exp(-\|v - v'\|_2/\phi)$. We generate DTs on a 20×20 grid with spacing of 1 between adjacent points. Each point represents a voxel and we can assign coordinates to

voxels. Let $j(v) = (j_1(v), j_2(v))$ be the coordinate of v, whose neighborhood is defined as $\Omega_v = \{v' : \|j(v) - j(v')\|_2 \leq 1\}$. We set sample size $n = 200$ and simulate predictor $\mathbf{X} \in \mathbb{R}^p$ using the method described in Subsection 3.2 for continuous predictors, with $\rho = 0.5$. For diagonal and off-diagonal elements of $L_i(v)$, we use variables X_1, X_5, X_{10} and X_{15}, X_{20}, X_{25}. We set degree of freedom $m = 3$ and spatial parameter $\phi = 15$.

We evaluate four methods for predicting matrices and compare the matrix prediction errors on an independent test set of 200 samples. The methods are (a) compute the Cholesky decomposition and run LASSO independently on the 6 elements of the decomposition (LASSO+Cholesky); (b) six linear regressions using variables selected by DC, each corresponding to an element of the Cholesky decomposition (DC+Cholesky); (c) FR using variables selected by DC (DC+Fréchet); and (d) smoothed FR using variables selected by DC (DC+Fréchet+Smooth). The BIC is utilized to determine the model size in DC screening. We measure prediction errors as the squared Cholesky distances between predicted and observed matrices, and assume equal neighbor weights $(= 0.5)$ in the smoothing step. The box plots of mean errors, for $p = 1000$ and 3000, are presented in Figure 1 (a). The plots reveal that all three DC-based methods have smaller prediction errors than LASSO+Cholesky, indicating that DC screening outperformed LASSO. DC+Cholesky and DC+Fréchet have similar performances, while the smoothing step in DC+Fréchet+Smooth further improves the predictions. Compared to $p = 1000$ (left panel), the reduction of prediction errors due to smoothing is more prominent when $p = 3000$ (right panel). We investigate the impact of weights u by comparing the prediction errors for different values of u. The results, presented in Figure 1 (b), indicate that smoothing improves the prediction performance and remains very stable for $u \geq 0.5$. This suggests that fine-tuning u is unnecessary under the smoothness assumption. The default setting $u = 0.5$ is more robust to potential violations of the smoothness assumption. We summarize that the smoothed FR with DC-selected variables provides the best predictive performance among the competing models. These findings highlight the importance of the proposed screening method under high-dimensionality, as well as the benefits of smoothing in accommodating spatial dependence.

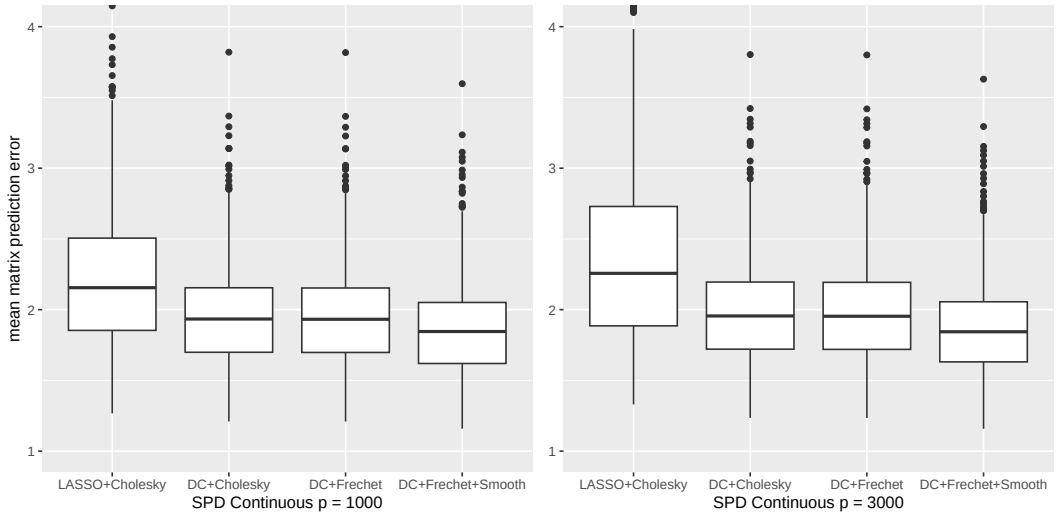
3.2 Data generation: FR models In this subsection, we generate data utilizing the 3 schemes:

Scheme I: SPD matrices with Cholesky metric. Consider 3 different choices of predictors, i.e., continuous, categorical, and mixed, described later. Given predictors $\mathbf{x}$, the regression function is given by

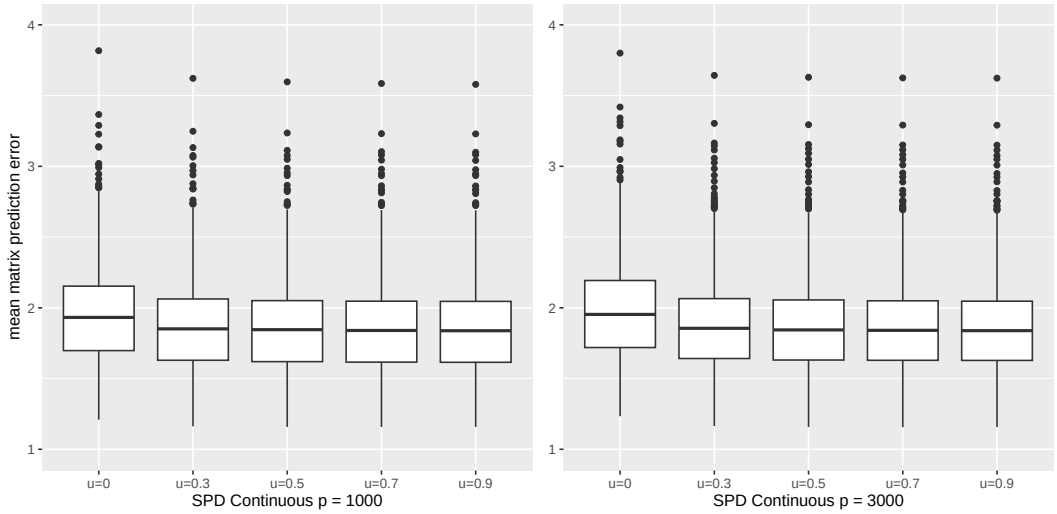
$$\mathrm{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = \mathrm{E}(A)^T \mathrm{E}(A), \tag{3.9}$$

where, $\mathrm{E}(A) = \{\mu_0 + \beta \sum_{i\in\mathcal{I}_1} x_i + \sigma_0 + \gamma \sum_{i\in\mathcal{I}_2} x_i\}I + \{\sigma_0 + \gamma \sum_{i\in\mathcal{I}_2} x_i\}U$, I is the $M \times M$ identity matrix, $U = (U_{ij})$ is an $M \times M$ matrix with $U_{ij} = I_{\{i<j\}}$, and $\mathcal{I}_1, \mathcal{I}_2$ are active predictor sets. Conditional on $\mathbf{X}$, the random response $\mathbf{Y}$ is generated by adding noise as follows: $\mathbf{Y} = A^T A$ where $A = (\mu + \sigma)I + \sigma U$ and with $\mu|\mathbf{X} \sim \mathcal{N}(\mu_0 + \beta \sum_{i\in\mathcal{I}_1} x_i, \nu_1)$ and $\sigma|\mathbf{X} \sim \mathrm{Gamma}((\sigma_0 + \gamma \sum_{i\in\mathcal{I}_2} x_i)^2\nu_2, \nu_2/(\sigma_0 + \gamma \sum_{i\in\mathcal{I}_2} x_i))$ being independently sampled. We set $\mu_0 = 3, \sigma = 3, \beta = 2, \gamma = 3, \nu_1 = 1, \nu_2 = 2$ and $M = 3$. The active predictors are $\mathcal{I}_1 = \{1, 5\}$ and $\mathcal{I}_2 = \{1, 5, 9, 13, 17, 20\}$. Note, when categorical variables are present in predictor $\mathbf{X}$ (Schemes Ia and Ib), one approach to represent them is by using multiple dummy variables. Specifically, if $\mathbf{X}_1$ is a categorical variable, we can construct a dummy vector $\widetilde{\mathbf{X}}_1$. In this context, the marginal utility of the grouped variable $\widetilde{\mathbf{X}}_1$ can be defined as the estimated DC between $\widetilde{\mathbf{X}}_1$ and the response variable $\mathbf{Y}$: $\widehat{\omega}_1 = \widehat{\mathrm{dcov}}(\widetilde{\mathbf{X}}_1, \mathbf{Y})$.

Scheme Ia: Continuous predictors. Here, we let $\mathbf{X}$ be continuous. The predictors are generated following two steps: (1) generate $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_p)^T$ from a multivariate normal distribution with mean zero and covariance matrix $\Sigma_{ij} = \rho^{|i-j|}, \rho = 0.5, i, j = 1, \ldots, p$;



(a) Box-plots of mean matrix prediction errors corresponding to the 4 competing methods (LASSO+Cholesky, DC+Cholesky, DC+Fréchet and DC+Fréchet+Smooth), for $p = 1000$, and $p = 3000$.



(b) Box-plots of mean matrix prediction errors corresponding to different weights u.

FIG 1. *Box plots of mean matrix prediction errors: (a) for different methods and (b) for different choices of weights in method DC+Fréchet+Smooth.*

(2) $\mathbf{X}_i = 2\Phi(\mathbf{Z}_i), i = 1, \ldots, p$, where Φ is the cumulative distribution function of the standard normal distribution.

Scheme Ib: Categorical predictors. Here, we follow the method of Han and Pan (2012) to generate discrete predictor vector $\mathbf{X}$, with elements taking values 0, 1, and 2. Initially, we sample $\mathbf{Z}, \mathbf{Z}'$ from a multivariate normal distribution with mean zero and covariance matrix $\Sigma_{ij} = \rho^{|i-j|}, \rho = 0.5, i, j = 1, \ldots, p$. Then, each element of $\mathbf{Z}$ and $\mathbf{Z}'$ is dichotomized to be 1 with probability uniformly sampled from $(0.2, 0.8)$. We set $\mathbf{X} = \mathbf{Z} + \mathbf{Z}'$.

Scheme Ic: Mixed predictors. Here, the predictors contain both continuous and categorical variables. First, we generate a continuous variable $\mathbf{X}_1 = |\mathbf{Z}|$, where $\mathbf{Z}$ follows the standard normal distribution. A binary variable $\mathbf{X}_2$ is sampled from the Bernoulli distribution with probability 0.5. To generate a 5-level categorical variable $\mathbf{X}_3$, we discretize a

standard normal distribution at 25, 45, 65, and 85 percentiles. Then, we simulate $\mathbf{X}_4$ to $\mathbf{X}_p$ as described in the previous schemes.

Scheme II: Probability distributions with Wasserstein metric. Here, $(\mathcal{Y}, d_{\mathcal{Y}})$ is a one-dimensional probability distribution space with Wasserstein metric. The Wasserstein distance between two probability distributions with cumulative distributions $F(\cdot)$ and $G(\cdot)$ is defined as

$$W(F,G) = \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt}.$$

The predictor vector $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)^T$ is generated following the simulation settings in Tucker, Wu and Müller (2023) in two steps: (1) Generate $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_p)^T$ from a multivariate normal distribution with mean zero and covariance matrix $\Sigma_{ij} = \rho^{|i-j|}, \rho = 0.5, i, j = 1, \ldots, p$; (2) Then, $\mathbf{X}_i = 2\Phi(\mathbf{Z}_i) - 1, i = 1, \ldots, p$, where Φ is the cumulative distribution function of the standard normal distribution. The regression function is defined as

$$\mathrm{E}[\mathbf{Y}(\cdot)|\mathbf{X} = \mathbf{x}] = \mu + \sigma\Phi^{-1}(\cdot), \tag{3.10}$$

where, $\mu|\mathbf{X} \sim \mathcal{N}(\mu_0 + \beta\sum_{i\in\mathcal{I}_1} x_i, \nu_1)$ and $\sigma|\mathbf{X} \sim \mathrm{Gamma}((\sigma_0 + \gamma\sum_{i\in\mathcal{I}_2} x_i)^2/\nu_2, \nu_2/(\sigma_0 + \gamma\sum_{i\in\mathcal{I}_2} x_i))$. The additional parameters are set as $\mu_0 = 0, \sigma_0 = 3, \beta = 3/4, \gamma = 1, \nu_1 = 1$ and $\nu_2 = 0.5$. The important predictors are $\mathcal{I}_1 = \{1, 5, 9, 13, 17, 20\}$ and $\mathcal{I}_2 = \{1\}$. This corresponds to the response distribution, on average, being a normal distribution with parameters linearly on $\mathbf{X}$.

Scheme III: Spherical data. We consider the situation where the random object responses lie in a Riemannian manifold object space. To be precise, let $\mathcal{Y} = S^2$ be the unit sphere in $\mathbb{R}^3$ with geodesic distance $d(\mathbf{Y}, \mathbf{Y}') = \arccos((\mathbf{Y}')^T\mathbf{Y})$. The predictors are generated in the same way as in previous examples; first generate $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_p)^T$ from a multivariate normal distribution with mean zero and covariance matrix $\Sigma_{ij} = \rho^{|i-j|}, \rho = 0.5, i, j = 1, \ldots, p$ and then let $\mathbf{X}_i = \Phi(\mathbf{Z}_i), i = 1, \ldots, p$, where Φ is the cumulative distribution function of the standard normal distribution. The regression function is defined as

$$\mathrm{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = \left(\sqrt{1 - x_5^2}\cos(\pi(x_1 + x_5 + x_9)), \sqrt{1 - x_5^2}\sin(\pi(x_1 + x_5 + x_9)), x_5\right). \tag{3.11}$$

To generate a random sample $(\mathbf{x}_i, \mathbf{Y}_i), i = 1, \ldots, n$, we first sample predictor vector $\mathbf{x}_i$ as above and a bivariate normal random vector $\mathbf{u}_i$ on the tangent space $T_{m_\oplus}(\mathbf{x}_i)\mathcal{Y}$. Let $\|\cdot\|_2$ be the Euclidean norm. The response is generated as

$$Y_i = \mathrm{Exp}_{m_\oplus(\mathbf{x}_i)}(\mathbf{u}_i) = \cos(||\mathbf{u}_i||_2)m_\oplus(\mathbf{x}_i) + \sin(||\mathbf{u}_i||_2)\frac{\mathbf{u}_i}{||\mathbf{u}_i||_2},$$

where Exp_p is the exponential map on the manifold for the tangent plane at point $p \in S^2$. The components of $\mathbf{u}_i$ are generated as independent normal random variables with mean 0 and standard deviation $\sigma_U = 0.2$.

Under high-dimensional settings, we set the sample size $n = 200$ and let the number of covariates p be 1000 and 3000 for all examples. We repeat each experiment 500 times and assess the performance by the following criteria: (1) $\mathcal{S}$, the minimum model size to include all important variables, and we report the 5%, 25%, 50%, 75%, and 95% percentiles of $\mathcal{S}$; (2) $\mathcal{P}_s$, the proportion that an individual active predictor is selected for a given model size d; (3) $\mathcal{P}_a$, the proportion that all important predictors are selected for a given model size d.

When the response is SPD matrices, we compare our method with two competing methods, where we (a) consider the fractional anisotropy (FA, Scheipl, Gertheiss and Greven,

2016), the more popular scalar response in DTI research (quantifying water diffusion direction hypothesized to be associated with white matter integrity and myelination) and use group LASSO to select variables (henceforth, LASSO+FA); and (b) compute the Cholesky decomposition and run group LASSO independently on the 6 elements of the decomposition (henceforth, LASSO+Cholesky). For (a), we compute the FA of a 3×3 SPD matrix using the formula

$$\text{FA} = \sqrt{\frac{1}{2}} \frac{\sqrt{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}},$$

where $\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues. For (b), a variable is considered important, if its coefficient is nonzero in one of the six LASSO estimations.

Tables 1 and 2 present the simulation results for $\mathcal{S}$, $\mathcal{P}_s$ and $\mathcal{P}_a$. For ease of presentation, we use SPD continuous, SPD categorical, and SPD mixed to represent the Schemes Ia, Ib and Ic, respectively. The percentiles in Table 1 imply that the minimum model size to ensure the inclusion of all important variables is small with high probability. Table 2 reveals that the proportions the active predictors selected by DC are close to 1 for varying model size d, which supports the sure screening property of our method. We observe that LASSO+FA has the worst performance, which is not surprising since FA only uses eigenvalue information. On the other hand, LASSO+Cholesky selects all the important variables. However, the average model sizes $\bar{d}$ are very large when the predictors are not continuous. DC can achieve 100% selection rate under much smaller model sizes. In particular, the results from Schemes Ib and Ic (where categorical variables are included) reveal that all active predictors, including the grouped variables, can be selected almost perfectly. Thus, our findings suggest that DC screening is highly efficient for selecting grouped predictors.

Under low-dimensional settings, we compare DC with `FRiSO`. For Scheme III (spherical data), we set the sample size $n = 100$ and the number of predictors $p = 9$ due to the inefficiency of `FRiSO`. The important variables are (X_1, X_5, X_9). For all the other models, we let $n = 200$, $p = 20$. The important variables are $(X_1, X_5, X_9, X_{13}, X_{17}, X_{20})$. The results from the variable selection averaged over 100 datasets are summarized in Table D.1 in Supplement D. While `FRiSO` demonstrated superior performance compared to our method, it comes at the cost of significantly longer computation time, which can be hundreds of thousands of times longer. For example, `FRiSO` needs more than 24,000 seconds to converge while DC gives the results in 0.03 seconds. Furthermore, DC screening is capable of producing comparable outcomes when the model size is slightly larger than the number of active variables. Therefore, DC is very efficient and reliable, while `FRiSO` is a computationally expensive method, whose application to high-dimensional settings is practically impossible.

Overall, our study suggests that the use of DC screening is a highly efficient and reliable technique for solving high-dimensional FR problems. Although `FRiSO` is effective in low-dimensional scenarios, our testing reveals that DC screening provides competitive results at a significantly faster rate. Furthermore, our screening approach can be directly applied to grouped predictors, as shown in Scheme Ib, whereas `FRiSO` is limited in this context.

TABLE 1

Table entries are the 5%, 25%, 50%, 75% and 95% percentiles of $\mathcal{S}$, corresponding to the 5 schemes: SPD continuous, SPD categorical, SPD mixed, Probability: Wasserstein, and Spherical, when $p = 1000$ and 3000. While the spherical scheme comprises 3 active variables, all others consist of 6.

Problem Size	#active variables	Scheme	Percentiles				
			5%	25%	50%	75%	95%
$p = 1000$	6	SPD continuous	6	6	8	9	13
		SPD categorical	6	6	8	12	37
		SPD mixed	6	6	7	11	28
		Probability: Wasserstein	6	7	10	14	42
	3	Spherical	3	3	3	4	5
$p = 3000$	6	SPD continuous	6	6	8	10	15
		SPD categorical	6	7	12	22	90
		SPD mixed	6	6	7	11	28
		Probability: Wasserstein	6	8	11	19	72
	3	Spherical	3	3	3	4	5

TABLE 2

Table entries are the selection rates of each variables $\mathcal{P}_s \times 100$ (denoted by the columns X_1, X_5, X_9, X_{13}, X_{17}, and X_{20}) and selection rates of all active variables $\mathcal{P}_a \times 100$ (denoted by column 'All'), under the 5 data generation schemes, i.e., SPD continuous, SPD categorical, SPD mixed, Probability density (Wasserstein metric), and Spherical, when $p = 1000$ and 3000. The d_1, d_2, d_3 are model sizes when we use DC, and $\bar{d}$ is the average number of selected variables for LASSO+FA and LASSO+Cholesky.

Model	Size/	$p = 1000$							$p = 3000$						
	Method	X_1	X_5	X_9	X_{13}	X_{17}	X_{20}	All	X_1	X_5	X_9	X_{13}	X_{17}	X_{20}	All
SPD Continuous	$d_1 = 6$	99.6	99.6	79.6	73.6	82.6	75.4	28.4	99.4	99.4	78.2	72.6	83.8	76.2	28.0
	$d_2 = 10$	100	100	96.8	94.6	97.4	94.6	83.4	100	100	96.4	93.4	96.6	94.8	81.4
	$d_3 = 15$	100	100	99.8	99.2	99.8	99.6	98.4	100	100	99.0	98.6	99.2	98.4	95.2
	FA $\bar{d} = 169$	99.8	99.2	99.6	100	100	99.8	99.0	100	99.8	100	100	100	99.8	99.6
	Chol $\bar{d} = 33$	100	100	100	100	100	100	100	100	100	100	100	100	100	100
SPD Categorical	$d_1 = 6$	98.8	99.4	77.0	76.0	84.4	72.0	27.4	98.2	99.0	69.8	68.0	75.4	63.2	14.8
	$d_2 = 12$	99.8	100	93.4	93.2	96.2	90.0	75.8	99.8	99.8	87.2	86.2	91.0	82.8	54.4
	$d_3 = 25$	100	100	97.0	97.0	98.8	96.2	89.8	99.8	100	94.0	92.8	96.8	92.0	78.0
	FA $\bar{d} = 180$	90.0	92.2	99.8	99.6	99.6	100	85.2	77.8	80.2	96.2	97.2	97.4	96.2	62.0
	Chol $\bar{d} = 137$	100	100	100	100	100	100	100	100	100	100	100	100	100	100
SPD Mixed	$d_1 = 6$	100	99.2	79.0	79.8	83.0	77.4	33.4	100	98.8	73.4	70.2	75.8	71.6	17.8
	$d_2 = 12$	100	100	95.2	94.0	95.8	94.8	80.8	100	99.6	89.6	88.6	92.4	89.0	64.4
	$d_3 = 25$	100	100	98.8	98.4	98.6	98.6	94.4	100	100	94.8	95.0	97.0	96.4	84.2
	FA $\bar{d} = 162$	96.4	90.8	99.8	99.4	100	99.2	88.4	93.0	81.8	96.6	96.2	97.2	97.6	75.0
	Chol $\bar{d} = 80$	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Probability Density	$d_1 = 6$	100	77.2	76.2	71.6	76.8	72.2	13.6	100	71.6	73.2	67.6	80.0	71.4	11.6
	$d_2 = 15$	100	95.0	96.4	95.2	97.6	93.8	79.4	100	91.2	92.8	92.6	95.0	92.8	68.6
	$d_3 = 30$	100	97.2	99.2	98.0	99.2	98.6	92.6	100	96.2	96.4	95.6	98.0	96.4	83.6
Sphere Data	$d_1 = 3$	84.0	100	84.2				68.6	85.2	100	83.2				69.0
	$d_2 = 4$	95.2	100	94.4				89.6	95.8	100	92.6				88.4
	$d_3 = 5$	99.0	100	98.4				97.4	98.6	100	97.4				96.0

4 Application: ADNI2 Data In this section, we illustrate our method via application to a dataset obtained from the Alzheimer's Disease Neuroimaging Initiative 2 (ADNI2; see https://adni.loni.usc.edu/ for more information). In the ADNI2 baseline scan, there are 169 subjects: 29 are clinically normal (CN), 29 have significant memory concern (SMC), 48 have early mild cognitive impairment (EMCI), 24 have late mild cognitive impairment (LMCI), and 39 have Alzheimer's disease (AD). Our focus is on four brain regions: the left corticospinal tract (CST), left frontopontine tracts (FPT), right CST, and right FPT. This dataset includes both DTI and genetic data. The genetic data comprise SNPs that belong to AD candidate genes. Based on the ADNI2 database, 357 SNPs from 33 targeted genes are obtained after quality control and imputation steps, as described in Szefer et al. (2017). Then, we follow the preprocessing steps as in Greenlaw et al. (2017) to focus on 294 SNPs. Therefore, our 297 predictors are sex, age, group (CN, SMC, EMCI, LMCI, and AD) index, and these 294 SNPs. Primarily, we aim to demonstrate the effectiveness of distance covariance screening. While our approach provides voxel-wise distance covariance screening results, we aim to find the set of most important variables that generates the maximum impact in the prediction of the 3×3 spd responses. Our strategy is to select the top 20 important predictors, voxel-wise, using our approach. Then, after pooling the selections for all voxels, selection frequencies of the 297 predictors will be computed.

Setting the model size $d = 20$ is a conservative choice between the BIC and the commonly used size in the literature, e.g., $\lfloor n/\log(n) \rfloor$. The average BIC selected model size of left CST and left FPT is 6. Given $n = 169$, the commonly used size is 32. Considering the complexity of the brain and the substantial amount of noise in DTI data (Carmichael et al., 2013), using SNPs to predict diffusion tensors is extremely challenging. Therefore, choosing $d = 20$ allows us to avoid excluding important variables (i.e., more conservative than the BIC selected $d = 6$). Additionally, the prediction errors shown in Figure 4 (b) demonstrate that our method is quite robust to the choice of model size.

Since SNPs, group, and sex are categorical variables, it is natural to code them by dummy variables. For SNPs, we use 2 dummy variables to code them since they can take on 3 different values (0, 1, and 2) to represent the number of minor alleles. When comparing prediction errors, we use both 3-level and binary codings. Group is coded by 4 dummy variables. Then we compute the distance covariance between the dummy variables and the diffusion tensor matrices, and select the variables with large distance covariance. For comparison, we also evaluate the performance of two alternative methods, LASSO+Cholesky and LASSO+FA, in selecting relevant variables. These methods serve as benchmarks for assessing the efficacy of our proposed method.

The frequencies of the top 20 most frequently selected genes (expressed as percentages) corresponding to the four regions (left CST, left FPT, right CST and right FPT) are summarized in Tables E.1 - E.4 in Supplement E. We observe that using the DC method, Group and Age are the dominant variables for all regions. Such finding aligns with the conventional understanding of AD (Guerreiro and Bras, 2015). However, the factor Group is not selected by other 2 methods, likely due to information loss when using FA responses, or model misspecification. Furthermore, the underlined SNPs in the frequency tables have been shown to be related to AD and memory retention, as revealed from the SNP database (dbSNP) – a public-domain archive for a broad collection of simple genetic polymorphisms (Kitts and Sherry, 2014). For instance, prior research studies (McCarthy et al., 2012; Hu et al., 2017) have demonstrated the significance of `rs7945931` and `rs4878104` (identified by all three methods in the region right FPT) to AD risk. Furthermore, our method uniquely selected `rs688` and `rs1133174` in the left FPT and CST regions as AD-related SNPs, as revealed in Bai et al. (2016) and Assareh et al. (2014). The bold SNPs (from the columns corresponding to the 2 competing methods) were also selected by our proposed DC method, in addition

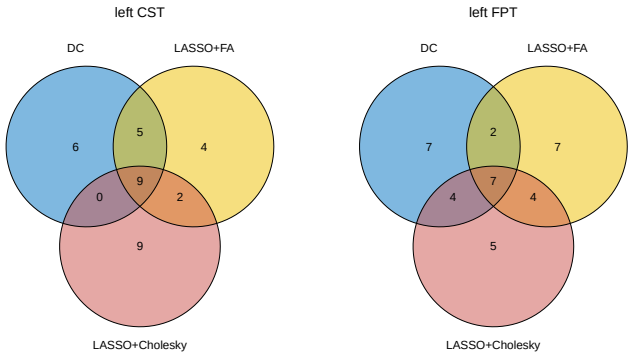
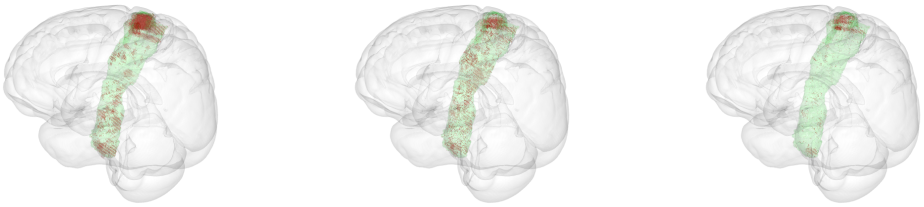
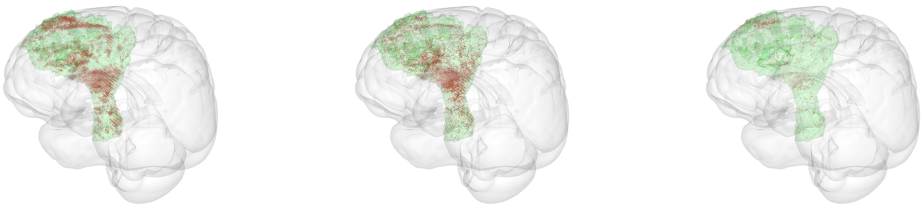


FIG 2. *Venn diagrams of selection frequencies from the 3 competing methods (DC, LASSO+FA, and LASSO+Cholesky), corresponding to the left CST and left FPT regions.*

to several more AD-related SNPs. Figure 2 displays the Venn diagrams of frequency tables, corresponding to the left CST and left FPT regions, implying that half of the most frequent SNPs selected by the competing LASSO+Cholesky and LASSO+FA methods were also detected by our proposed DC method.



(a) left CST | DC (b) left CST | LASSO+Cholesky (c) left CST | LASSO+FA



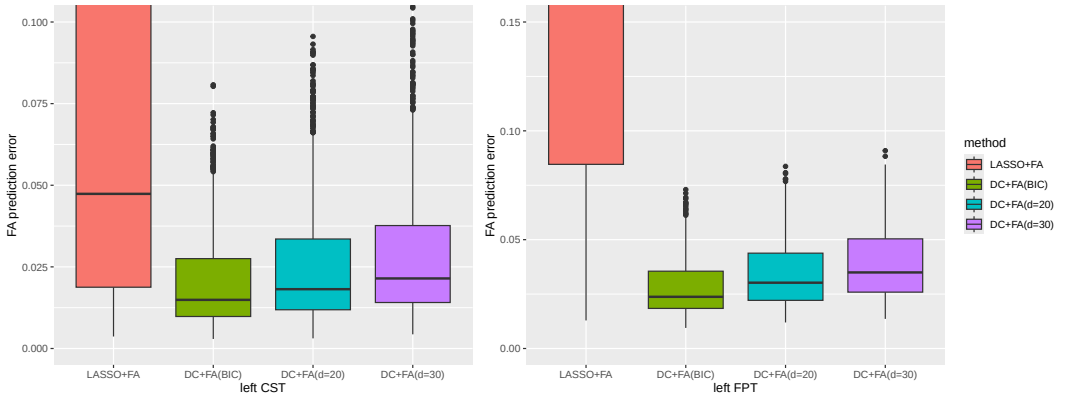
(d) left FPT | DC (e) left FPT | LASSO+Cholesky (f) left FPT | LASSO+FA

FIG 3. *Plots of selected voxels (locations) corresponding to specific SNPs (*`rs3737964` *and* `rs2279339`*) for left CST (upper panel) and left FPT (lower panel), respectively. While the selected voxels appear in red, the brain ROI are colored green. The contiguity measures ρ with a radius of 2 for the regions in the subfigures: (a) 67.7, (b) 33.2, (c) 28.9, (d) 45.7, (e) 36.0, (f) 14.4.*

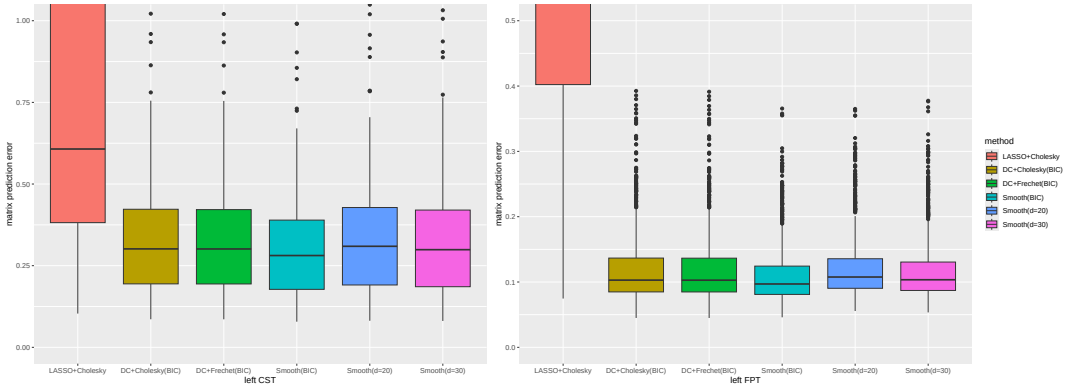
Figure 3 depicts the selected regions (voxels) corresponding to specific SNPs (`rs3737964` for left CST, and `rs2279339` for left FPT), as obtained from the 3 competing selection methods. While the selected voxels appear in red, the brain regions of interests are shown in green, and the entire brain serves as the background. We utilize the contiguity measure as described in Orsi (2023). The contiguity ρ is defined as the average number of selected voxels within a neighborhood with a given radius of each selected voxel. A larger ρ indicates more contiguous regions. The contiguity measures with a radius of 2 for the regions in Figure 3 are: (a) 67.7, (b) 33.2, (c) 28.9, (d) 45.7, (e) 36.0, (f) 14.4. We also compute ρ for two additional radii, with results summarised in Table E.5 in Supplement E. The results indicate that the regions selected via DC exhibit greater contiguity, and are easier to identify compared to those selected by other methods. Furthermore, the DC-selected regions cover most of the regions selected by the other two methods. These regions consist of superficial white matter, which has been known to play an important role in brain's cognitive functions (Kirilina et al., 2020). Figure E.4 (Supplement E) presents plots of selected voxels corresponding to various SNPs, using the DC method. The plot reveals that SNPs may vary corresponding to the distinct active regions, with possible overlaps. This observation implies that certain regions may be influenced by the combined effects of multiple SNPs. The results presented in these figures provide valuable insights into the effectiveness of the proposed method in identifying relevant regions associated with genetic variations, and offer newer opportunities for the discovery of novel genetic associations.

Furthermore, we compare the prediction errors generated employing the two-stage Algorithm 1 to the competing methods. We select two regions, each of size $10 \times 10 \times 10$, from the left CST and left FPT, to compare the out-of-sample testing errors using the leave-one-out cross-validation (LOOCV). There are two reasons for choosing these two regions instead of using all voxels. First, approximately 25% of the voxels in the left CST and 14% of the voxels in the left FPT have a significant portion of zero FA values among all subjects. A zero FA value indicates the diffusion tensor is isotropic, and thus less meaningful in our framework of regressing DTI on SNPs. Second, LOOCV is computationally expensive because we need to fit a non-parametric Fréchet regression model for each voxel and for each leave-one-out sample. Therefore, we select a region with approximately 1000 voxels that do not have zero FA values to access the predictive performance using LOOCV. First, let FA be the prediction target. We consider two methods: (a) LASSO+FA, and (b) linear regression using FA as the response and variables selected by distance covariance (DC+FA). Next, we consider the 3×3 diffusion tensor be the prediction target, and compare four methods: (c) LASSO+Cholesky, (d) DC+Cholesky; (e) DC+Fréchet, and (f) DC+Fréchet+Smooth. For methods (a) and (b), we obtain the predicted diffusion tensor by multiplying the decomposition matrix by its transpose. Then, we compare their prediction errors measured in squared Cholesky distance.

The box plots of LOOCV mean prediction errors are shown in Figure 4. The left and right panels correspond to the region of left CST and left FPT. We have several findings based on the results. First, from upper panels, we observe that DC+FA outperforms LASSO+FA. From the lower panels, we observe DC+Fréchet+Smooth exhibit the best performance, which supports the assumption that voxels in the brain are influenced by their neighboring voxels. However, using FR does not improve the prediction accuracy considerably, when compared to prediction via linear regression of the Cholesky decomposition. One possible reason is that for this dataset, the relationship between the DTI and predictors can be captured through linear regressions of the Cholesky decomposition. The average model sizes for LASSO and DC are 28 and 6 for the region of left CST, and 53 and 5 for the region of left FPT. Even when setting the model size for DC to a comparable number to LASSO, the prediction errors remain significantly smaller than LASSO. Second, we explore the impact of the model size d in DC. Three scenarios are considered: d is selected by BIC, $d = 20$, and $d = 30$. Our method



(a) The third quantiles of LASSO+FA are 0.3 for left CST and 87.2 for left FPT. To better highlight the differences between the methods, the y-axis has been capped at 0.1 for left CST and 0.15 for left FPT.



(b) The third quantiles of LASSO+Cholesky are 6.2 for left CST and 410.7 for left FPT. To better highlight the differences between the methods, the y-axis has been capped at 1 for left CST and 0.5 for left FPT. DC+Fréchet+Smooth is abbreviated as Smooth.

FIG 4. *Boxplots of mean LOOCV prediction errors corresponding to FA prediction targets (upper panel), and* 3×3 *matrix prediction targets (lower panel) measured in Cholesky distances, for* $10 \times 10 \times 10$ *regions of left CST and left FPT. The characters in the parentheses indicate how the model size is selected and the SNPs are coded as categorical variables with three categories.*

shows robust performance across these model sizes, especially when combining DC and Fréchet regression to predict diffusion tensors. In addition, from Figure E.1 in Supplement E coding SNPs as a 3-level variable is better than coding them as binary.

Stability measures are computed to compare the variable selection stability of LASSO and DC. A variable selection algorithm is stable if small changes in data lead to small changes in the chosen variable subset. For each voxel, there are 169 selected variable subsets since we use LOOCV. Therefore, we compute the stability measure for each voxel. Three commonly used measures are Jaccard (Kalousis, Prados and Hilario, 2005), Lustgarten (Lustgarten, Gopalakrishnan and Visweswaran, 2009), and Nogueira (Nogueira, Sechidis and Brown, 2018). The maximum value of these measures is 1, with values close to 0 indicating instability and values near 1 indicating stability. Figure 5 shows box plots of stability indices for voxels. The results indicate DC is much more stable than LASSO, regardless of the measure used. Combining with previous results, DC+Fréchet+Smooth is most desirable since it has high predictive accuracy, high variable selection stability, and a small number of selected variables.

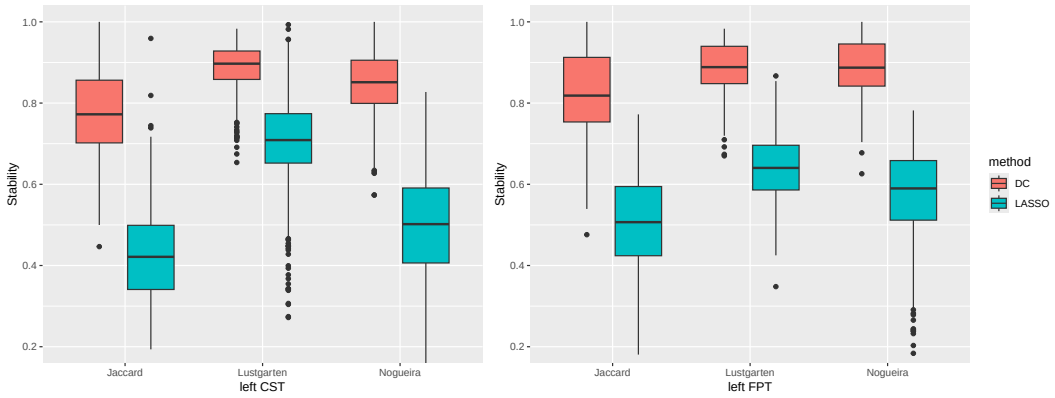
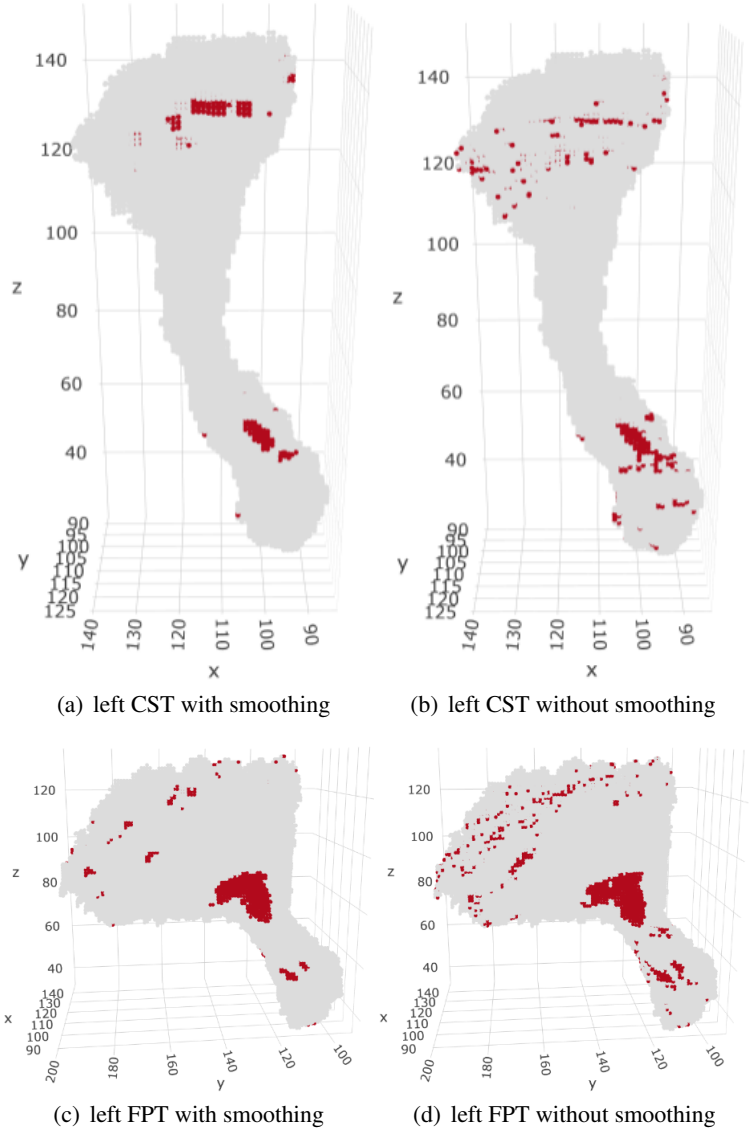


FIG 5. *Boxplots of variable selection stability of LASSO (blue) and DC (red), measured using three different stability indices: Jaccard, Lustgarten, and Nogueira. Values close to 0 indicate instability, while values near 1 indicate stability.*

Following this, we demonstrate how the smoothing method aids us to study the differences between AD and CN groups. First, we select a subject from the CN group, a 73 years old woman. Then, we create an artificial subject with the same predictors as this subject, except that the subject is in the AD group. We now fit global FR using the selected 20 variables, and apply local smoothness with weights $(= 0.5)$ for the neighbors. Figure 6 plots the difference in predictions between the fake AD subject and real CN subject at selected voxels (denoted in red) lying in the left CST (upper panel) and left FPT (lower panel) regions, from employing our proposed method with/without smoothing. The corresponding plots for the right CST and right FPT regions are presented in Figures E.2 and E.3 in Supplement E. We observe that local averaging can highlight the important areas in the brain. Several voxels that appear in the right panel plots (without smoothing) form contiguous regions in the left panel (when smoothing is applied), demonstrating the power of smoothing. Using smoothing, we are able to detect spatially contiguous regions since the small separated regions surrounded by inactive regions will be smoothed by the neighbors, and thus become undetectable.



(a) left CST with smoothing (b) left CST without smoothing

(c) left FPT with smoothing (d) left FPT without smoothing

FIG 6. *Plots of the differences in prediction between a real subject (73 year old woman) in the Control group from the ADNI data and a fake subject in the AD group (with same covariates as the real subject), at selected voxels (denoted in red) in the left CST (upper panel), and left FPT (lower panel) regions, obtained by fitting our method (with/without smoothing).*

5 Conclusion In this paper, we proposed a novel variable screening and spatial smoothing methodology for fitting complicated biomedical matrix-variate responses that arises in DTI research. Based on distance covariances in metric spaces, the implementation of the variable screening method does not require solving optimization problems and model specification, which makes it efficient and flexible. We believe our proposal is the maiden attempt to perform variable screening in a high-dimensional FR framework, and enjoys the promising property of sure screening under mild assumptions. For spatially correlated responses in the form of SPD matrices, we developed a smoothing method that aggregates spatial information in adjacent matrices, leading to a closed-form solution. We empirically demonstrate the efficacy of both techniques (variable screening and smoothing) on both simulated and real

data, and emphasize that our approach is versatile, and generalizable beyond DTI to different types of biomedical data.

The results presented in this paper motivate a number of future directions. From a methodological perspective, our current method selects variables with the largest d distance covariance. Given this ad-hoc choice, it is unclear whether all of the active predictors are selected or not, and it may include many inactive predictors. Therefore, future research pertains to developing a threshold selection procedure. A related interesting problem is establishing theoretical properties of the size of $\widehat{\mathcal{I}}$. From the definition, the size depends on unknown constants c and κ, making it challenging to provide a precise bound within the current theoretical framework. This issue has not been addressed in existing variable screening literature for simpler data structures (Fan and Lv, 2008; Li, Zhong and Zhu, 2012; Cui, Li and Zhong, 2015; Liu, Li and Wu, 2014; Pan et al., 2019). Consequently, new mathematical tools need to be developed to tackle this problem. Additionally, the current variable screening procedure treats each response as independent, while adjacent responses tend to have the same set of important variables, indicating the potential to integrate response dependence into the procedure. Therefore, it is imperative to enhance the strength of the current screening method by accommodating spatial dependence between responses.

From an application perspective, our proposed approach can be applied to DTI data with other representations such as fiber orientation distribution (Zhu, Blumenthal and Lowe, 1997). While the overwhelmingly complex data representations in metric spaces prohibit application of well-utilized statistical modeling tools, the FR framework provides an elegant solution and alleviates the shortcomings. Yet, the computational burden and scalability of these approaches for addressing high-dimensionality bottlenecks (arising in neurological studies) and in deriving meaningful inference require closer examination, to be pursued elsewhere.

Acknowledgments. The authors would like to thank the Editor, Dr. Jie Peng, the Associate Editor and two reviewers for helpful comments. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Funding. Data collection and sharing for this project was funded by the ADNI National Institutes of Health (NIH) Grant U01AG024904, and Department of Defense ADNI award W81XWH-12-2-0012. Lei Yan and Xin Zhang are supported in part by grant DMS-2053697 from the National Science Foundation. Bandyopadhyay acknowledges partial funding support from NIH grants R21DE031879 and R01DE031134. Wu's research was partially supported by the National Science Foundation grant 2152070.

SUPPLEMENTARY MATERIAL

Supplement to "Variable Screening and Spatial Smoothing in Fréchet Regression with Application to Diffusion Tensor Imaging"

The Supplement contains detailed proofs of the Lemmas and Theorems presented in the main paper, and additional Tables and Figures from simulations and the real data application. Computing code along with implementation vignette is available at: https://github.com/leiyan-ly/Frechet-regression.

REFERENCES

ALEXANDER, A. L., LEE, J. E., LAZAR, M. and FIELD, A. S. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics* **4** 316–329.

ASSAREH, A. A., PIGUET, O., LYE, T. C., MATHER, K. A., BROE, G. A., SCHOFIELD, P. R., SACHDEV, P. S. and KWOK, J. B. (2014). Association of SORL1 gene variants with hippocampal and cerebral atrophy and Alzheimer's disease. *Current Alzheimer research* **11** 558–563.

BAI, F., YUAN, Y., SHI, Y. and ZHANG, Z. (2016). Multiple genetic imaging study of the association between cholesterol metabolism and brain functional alterations in individuals with risk factors for Alzheimer's disease. *Oncotarget* **7** 15315.

CARMICHAEL, O., CHEN, J., PAUL, D. and PENG, J. (2013). Diffusion tensor smoothing through weighted Karcher means. *Electronic journal of statistics* **7** 1913.

CHAUDHURI, A. and HU, W. (2019). A fast algorithm for computing distance correlation. *Computational statistics & data analysis* **135** 15–24.

CHOWDHURY, M. Z. I. and TURIN, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family medicine and community health* **8**.

CUI, H., LI, R. and ZHONG, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association* **110** 630–641.

DEMIRHAN, A., NIR, T. M., ZAVALIANGOS-PETROPULU, A., JACK, C. R., WEINER, M. W., BERNSTEIN, M. A., THOMPSON, P. M. and JAHANSHAD, N. (2015). Feature selection improves the accuracy of classifying Alzheimer disease using diffusion tensor images. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* 126–130. IEEE.

DEZA, M. M. and DEZA, E. (2009). *Encyclopedia of Distances* In *Encyclopedia of Distances* 1–583. Springer Berlin Heidelberg, Berlin, Heidelberg.

DRYDEN, I. L., KOLOYDENKO, A. and ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics* **3** 1102–1123.

DUBEY, P. and MÜLLER, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika* **106** 803–821.

ENNIS, D. B. and KINDLMANN, G. (2006). Orthogonal tensor invariants and the analysis of diffusion tensor magnetic resonance images. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **55** 136–146.

ESRAEL, S. M. A. M., HAMED, A. M. M., KHEDR, E. M. and SOLIMAN, R. K. (2021). Application of diffusion tensor imaging in Alzheimer's disease: quantification of white matter microstructural changes. *Egyptian Journal of Radiology and Nuclear Medicine* **52** 1–8.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911.

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20** 101–148.

FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré* **10** 215–310.

GRAÑA, M., TERMENON, M., SAVIO, A., GONZALEZ-PINTO, A., ECHEVESTE, J., PÉREZ, J. and BESGA, A. (2011). Computer aided diagnosis system for Alzheimer disease using brain diffusion tensor imaging features selected by Pearson's correlation. *Neuroscience letters* **502** 225–229.

GREENLAW, K., SZEFER, E., GRAHAM, J., LESPERANCE, M., NATHOO, F. S. and INITIATIVE, A. D. N. (2017). A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics* **33** 2513–2522.

GUERREIRO, R. and BRAS, J. (2015). The age factor in Alzheimer's disease. *Genome medicine* **7** 1–3.

HAN, F. and PAN, W. (2012). A composite likelihood approach to latent multivariate gaussian modeling of snp data with application to genetic association testing. *Biometrics* **68** 307–315.

HEINZE, G., WALLISCH, C. and DUNKLER, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical journal* **60** 431–449.

HU, Y., CHENG, L., ZHANG, Y., BAI, W., ZHOU, W., WANG, T., HAN, Z., ZONG, J., JIN, S., ZHANG, J. et al. (2017). Rs4878104 contributes to Alzheimer's disease risk and regulates DAPK1 gene expression. *Neurological Sciences* **38** 1255–1262.

JAKOBSEN, M. E. (2017). Distance Covariance in Metric Spaces: Non-Parametric Independence Testing in Metric Spaces, Master's Thesis, University of Copenhagen.

JEON, J. M., PARK, B. U. and VAN KEILEGOM, I. (2021). Additive regression for non-Euclidean responses and predictors. *The Annals of Statistics* **49** 2611–2641.

JOHNSON, T. D., LIU, Z., BARTSCH, A. J. and NICHOLS, T. E. (2013). A Bayesian non-parametric Potts model with application to pre-surgical FMRI data. *Statistical methods in medical research* **22** 364–381.

KALOUSIS, A., PRADOS, J. and HILARIO, M. (2005). Stability of feature selection algorithms. In *Fifth IEEE International Conference on Data Mining (ICDM'05)* 8–pp. IEEE.

KIRILINA, E., HELBLING, S., MORAWSKI, M., PINE, K., REIMANN, K., JANKUHN, S., DINSE, J., DEISTUNG, A., REICHENBACH, J. R., TRAMPEL, R. et al. (2020). Superficial white matter imaging: Contrast mechanisms and whole-brain in vivo mapping. *Science Advances* **6** eaaz9281.

KITTS, A. and SHERRY, S. (2014). The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. In *The NCBI Handbook [Internet]* (J. McEntyre and J. Ostell, eds.) 5, 266-290. National Center for Biotechnology Information (US), Bethesda, MD.

KRONLAGE, M., SCHWEHR, V., SCHWARZ, D., GODEL, T., UHLMANN, L., HEILAND, S., BENDSZUS, M. and BÄUMER, P. (2018). Peripheral nerve diffusion tensor imaging (DTI): normal values and demographic determinants in a cohort of 60 healthy individuals. *European radiology* **28** 1801–1808.

LAN, Z., REICH, B. J. and BANDYOPADHYAY, D. (2021). A spatial Bayesian semiparametric mixture model for positive definite matrices with applications in diffusion tensor imaging. *Canadian Journal of Statistics* **49** 129–149.

LAN, Z., REICH, B. J., GUINNESS, J., BANDYOPADHYAY, D., MA, L. and MOELLER, F. G. (2022). Geostatistical modeling of positive-definite matrices: An application to diffusion tensor imaging. *Biometrics* **78** 548–559.

LEE, H. N. and SCHWARTZMAN, A. (2017). Inference for eigenvalues and eigenvectors in exponential families of random symmetric matrices. *Journal of Multivariate Analysis* **162** 152–171.

LEE, J. D., SUN, Y. and TAYLOR, J. E. (2015). On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics* **9** 608–642.

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107** 1129–1139.

LIN, Z., MÜLLER, H.-G. and PARK, B. (2023). Additive models for symmetric positive-definite matrices and Lie groups. *Biometrika* **110** 361–379.

LIU, J., LI, R. and WU, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association* **109** 266–274.

LO, C.-Y., WANG, P.-N., CHOU, K.-H., WANG, J., HE, Y. and LIN, C.-P. (2010). Diffusion tensor tractography reveals abnormal topological organization in structural cortical networks in Alzheimer's disease. *Journal of Neuroscience* **30** 16876–16885.

LUSTGARTEN, J. L., GOPALAKRISHNAN, V. and VISWESWARAN, S. (2009). Measuring stability of feature selection in biomedical datasets. In *AMIA annual symposium proceedings* **2009** 406. American Medical Informatics Association.

LYONS, R. (2013). Distance covariance in metric spaces. *The Annals of Probability* **41** 3284–3305.

MCCARTHY, J. J., SAITH, S., LINNERTZ, C., BURKE, J. R., HULETTE, C. M., WELSH-BOHMER, K. A. and CHIBA-FALEK, O. (2012). The Alzheimer's associated 5' region of the SORL1 gene cis regulates SORL1 transcripts expression. *Neurobiology of aging* **33** 1485–e1.

MIN, K. and MAI, Q. (2022). A general framework for tensor screening through smoothing. *Electronic Journal of Statistics* **16** 451–497.

MORI, S. (2007). *Introduction to diffusion tensor imaging*. Elsevier.

NOGUEIRA, S., SECHIDIS, K. and BROWN, G. (2018). On the stability of feature selection algorithms. *Journal of Machine Learning Research* **18** 1–54.

ORSI, F. (2023). Exploring the role of compactness in path-dependent land-taking processes in Italy. *GeoJournal* **88** 69–87.

O'DONNELL, L. J. and WESTIN, C.-F. (2011). An introduction to diffusion tensor image analysis. *Neurosurgery Clinics* **22** 185–196.

PAN, W., WANG, X., XIAO, W. and ZHU, H. (2019). A generic sure independence screening procedure. *Journal of the American Statistical Association*.

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

QI ZHANG, L. X. and LI, B. (2023). Dimension Reduction for Fréchet Regression. *Journal of the American Statistical Association* **0** 1–15.

REICH, B. J., GUINNESS, J., VANDEKAR, S. N., SHINOHARA, R. T. and STAICU, A.-M. (2018). Fully Bayesian spectral methods for imaging data. *Biometrics* **74** 645–652.

SCHEIPL, F., GERTHEISS, J. and GREVEN, S. (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics* **10** 1455 – 1492.

SCHWARTZMAN, A. (2016). Lognormal distributions and geometric averages of symmetric positive definite matrices. *International Statistical Review* **84** 456–486.

SOARES, J. M., MARQUES, P., ALVES, V. and SOUSA, N. (2013). A hitchhiker's guide to diffusion tensor imaging. *Frontiers in neuroscience* **7** 31.

SPENCE, J. S., CARMACK, P. S., GUNST, R. F., SCHUCANY, W. R., WOODWARD, W. A. and HALEY, R. W. (2007). Accounting for spatial dependence in the analysis of SPECT brain imaging data. *Journal of the American Statistical Association* **102** 464–473.

STEBBINS, G. and MURPHY, C. (2009). Diffusion tensor imaging in Alzheimer's disease and mild cognitive impairment. *Behavioural neurology* **21** 39–49.

SZEFER, E., LU, D., NATHOO, F., BEG, M. F., GRAHAM, J., INITIATIVE, A. D. N. et al. (2017). Multivariate association between single-nucleotide polymorphisms in Alzgene linkage regions and structural changes in the brain: discovery, refinement and validation. *Statistical applications in genetics and molecular biology* **16** 367–386.

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics* **35** 2769–2794.

TUCKER, D. C., WU, Y. and MÜLLER, H.-G. (2023). Variable selection for global Fréchet regression. *Journal of the American Statistical Association* **118** 1023–1037.

WOOLRICH, M. W., JENKINSON, M., BRADY, J. M. and SMITH, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE transactions on medical imaging* **23** 213–231.

WU, G.-R., STRAMAGLIA, S., CHEN, H., LIAO, W. and MARINAZZO, D. (2013). Mapping the voxel-wise effective connectome in resting state fMRI. *PloS one* **8** e73670.

XUE, W., BOWMAN, F. D. and KANG, J. (2018). A Bayesian spatial model to predict disease status using imaging data from various modalities. *Frontiers in Neuroscience* **12** 184.

YAN, L., ZHANG, X., LAN, Z., BANDYOPADHYAY, D. and WU, Y. (2024). Supplement to "Variable Screening and Spatial Smoothing in Fréchet Regression with Application to Diffusion Tensor Imaging". DOI: [provided by typesetter].

ZHU, Y., BLUMENTHAL, W. and LOWE, T. (1997). Determination of non-symmetric 3-D fiber-orientation distribution and average fiber length in short-fiber composites. *Journal of Composite Materials* **31** 1287–1301.