Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 153 a 168

Partial Least Squares Regression on Symmetric Positive-Definite Matrices

Regresión de mínimos cuadrados parciales sobre matrices simétricas definidas positiva

Raúl Alberto Pérez[1,a], Graciela González-Farias[2,b]

[1]Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia

[2]Departamento de Probabilidad y Estadística, CIMAT-México Unidad Monterrey, Monterrey Nuevo León, México

Resumen

Recently there has been an increased interest in the analysis of different types of manifold-valued data, which include data from symmetric positive-definite matrices. In many studies of medical cerebral image analysis, a major concern is establishing the association among a set of covariates and the manifold-valued data, which are considered as responses for characterizing the shapes of certain subcortical structures and the differences between them.

The manifold-valued data do not form a vector space, and thus, it is not adequate to apply classical statistical techniques directly, as certain operations on vector spaces are not defined in a general Riemannian manifold. In this article, an application of the partial least squares regression methodology is performed for a setting with a large number of covariates in a euclidean space and one or more responses in a curved manifold, called a Riemannian symmetric space. To apply such a technique, the Riemannian exponential map and the Riemannian logarithmic map are used on a set of symmetric positive-definite matrices, by which the data are transformed into a vector space, where classic statistical techniques can be applied. The methodology is evaluated using a set of simulated data, and the behavior of the technique is analyzed with respect to the principal component regression.

***Palabras clave*:** Matrix theory, Multicollinearity, Regression, Riemann manifold.

[a]Assistant Professor. E-mail: raperez1@unal.edu.co
[b]Associate Professor. E-mail: gmfarias@cimat.mx

Abstract

Recientemente ha habido un aumento en el interés de analizar diferentes tipos de datos variedad-valuados, dentro de los cuáles aparecen los datos de matrices simétricas definidas positivas. En muchos estudios de análisis de imágenes médicas cerebrales, es de interés principal establecer la asociación entre un conjunto de covariables y los datos variedad-valuados que son considerados como respuesta, con el fin de caracterizar las diferencias y formas en ciertas estructuras sub-corticales.

Debido a que los datos variedad-valuados no forman un espacio vectorial, no es adecuado aplicar directamente las técnicas estadísticas clásicas, ya que ciertas operaciones sobre espacio vectoriales no están definidas en una variedad riemanniana general. En este artículo se realiza una aplicación de la metodología de regresión de mínimos cuadrados parciales, para el entorno de un número grande de covariables en un espacio euclídeo y una o varias respuestas que viven una variedad curvada llamada espacio simétrico Riemanniano. Para poder llevar a cabo la aplicación de dicha técnica se utilizan el mapa exponencial Riemanniano y el mapa log Riemanniano sobre el conjunto de matrices simétricas positivas definida, mediante los cuales se transforman los datos a un espacio vectorial en donde se pueden aplicar técnicas estadísticas clásicas. La metodología es evaluada por medio de un conjunto de datos simulados en donde se analiza el comportamiento de la técnica con respecto a la regresión por componentes principales.

***Key words*:** multicolinealidad, regresión, teoría de matrices, variedad Riemanniana.

1. Introduction

In studies of diffusion tensor magnetic resonance imaging (TD-MRI), a diffusion tensor (DT) is calculated in each voxel of an imaging space, which describes the local diffusion of water molecules in various directions over this region of the brain. A sequence of images is used to measure this diffusion. The sequence includes a noise that produces uncertainty in the tensor estimation and in the estimation of certain quantities inherent to water molecules, such as eigenvalues, eigenvectors, the anisotropic fraction rate (FA) and the fiber trajectories, which are constructed based on these last parameters. The diffusion-tensor imaging (DTI) is a powerful tool to quantitatively evaluate the integrity of the anatomic connectivity in the white matter of clinical populations. The methods used for the analysis of DTI at a group level include the statistical analysis of certain invariant measures, such as eigenvalues, eigenvectors or principal directions, the anisotropic fraction, and the average diffusivity, among others. However, these invariant measures do not capture all of the information about the complete DTs, which leads to a decrease in the statistical power of the DTI to detect subtle changes in white matter. Hence, new statistical methods are being developed to fully analyze the DTs as responses and to establish their association with a set of covariates (Li, Zhu, Chen, Ibrahim, An, Lin, Hall & Shen 2009, Zhu, Chen, Ibrahim, Li & Lin 2009, Yuan, Zhu, Lin & Marron 2012). In some of these development the log-euclidean metric has been used with the transformation of the DTs from a non-linear space into their

logarithmic matrices on a Euclidean space. Semi-parametrical models have been proposed to study the relationship between the set of covariates and the DTs as responses. Estimation processes and hypotheses test based on test statistics and resampling methods have been developed to simultaneously evaluate the statistical significance of linear hypotheses throughout large regions of interest (ROI).

An appropriate statistical analysis of DTs is important to understand the normal development of the brain, the neurological bases of neuropsychiatric disorders and the joined effects of environmental and genetic factors on the brain structure and function. In addition, any statistical method for complete diffusion tensors can be directly applied to positive-definitive tension matrices in computational anatomy to understand the variations in shapes of brain-structure imaging (Grenander & Miller 1998, Lepore, Brun, Chou, Chiang, Dutton, Hayashi, Luders, Lopez, Aizenstein, Toga, Becker & Thompson 2008).

Symmetric positive-definite (SPD) matrix-valued data occur in a wide variety of applications, such as DTI for example, where a SPD 3x3 DT, which tracks the effective diffusion of the water molecules in certain brain regions, is estimated at each voxel of an imaging space. Another application of SPD matrix-valued data can be seen in studies on functional magnetic resonance imaging (fMRI), where an SPD covariance matrix is calculated to delineate the functional connectivity between different neuronal assembles involved in the execution of certain complex cognitive tasks or in perception processes (Fingelkurts & Kahkonen 2005). Despite the popularity of SPD matrix-valued data, there are few statistical methods to analyze SPD matrices as response variables in a Riemannian manifold. Data considered as responses with a small number of covariates of interest in a Euclidian space can be found from the following studies in the literature for statistical analysis using regression models of SPD matrices: Batchelor, Moakher, Atkinson, Calamante & Connelly (2005), Pennec, Fillard & Ayache (2006), Schwartzman (2006), Fletcher & Joshi (2007), Barmpoutis, Vemuri, Shepherd & Forder (2007), Zhu et al. (2009) and Kim & Richards (2010). However, because the SPD matrix data do not form a vector space, classical multivariate regression techniques cannot be applied directly to establish the relationship between these types of data and a set of covariates of interest.

In a setting with a large number of covariates with a high multicollinearity presence and few available observations, no regression methods have been previously proposed to study the relationship between such covariates and the response variables of SPD matrices living in non-Euclidian spaces. In this article, partial least squares (PLS) regression is proposed using a strategy of exponential and Riemannian logarithmic maps to transform data into Euclidian spaces. The development of the technique is similar to the scheme for the analysis of SPD matrices data as responses in a classical regression model and in a local polynomial regression model, as proposed in Zhu et al. (2009) and Yuan et al. (2012). The PLS regression model is initially evaluated using a set of simulated data and statistical validation techniques which currently exist, such as cross validation techniques. The behavior of the PLS regression technique is analyzed by comparing it to that of the classic dimension-reduction technique, called principal component (PC) regression.

The article is structured as follows: In Section 2, a brief revision of the existing theory for PC and the PLS regression classical model is outlined. In Section 3, some properties of the Riemannian geometric structure of SPD matrices are reviewed. An outline of the regression models, as well as the estimation methods of their regression coefficients are also presented. In Section 4, our PLS regression model is presented, along with the estimation process used and their application and evaluation on a simulated-data set. In Section 5, conclusions and recommendations for future work are given.

2. Regression Methods

2.1. Classical Regression

We will examine the following data set $\{(\mathbf{x}_i, y_i) \ : \ i = 1, 2, \ldots, n\}$, composed of a response y_i and a $k \times 1$ covariate vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T$, where the response can be continuous, discrete, or qualitative observations, and the covariates can be qualitative or quantitative. A regression model often includes two key elements: A link function $\mu_i(\boldsymbol{\beta}) = E[y|\mathbf{x}_i] = g(\mathbf{x}_i, \boldsymbol{\beta})$ and a residual $\epsilon_i = y_i - \mu_i(\boldsymbol{\beta})$, where $\boldsymbol{\beta}_{q\times 1}$ is a regression-coefficients vector and $g(.\ ,\ .)$: from $\mathbb{R}^k \times \mathbb{R}^q \to \mathbb{R}$, $(\mathbf{x}_i, \boldsymbol{\beta}) \to g(\mathbf{x}_i, \boldsymbol{\beta})$ with $q = k + 1$, can be known or unknown according to the type of model: Parametric, not-parametric or semi-parametric. The parametric regression model can be defined as: $y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + \epsilon_i$, with $g(\mathbf{x}_i, \boldsymbol{\beta})$: Known and $E[\epsilon_i|\mathbf{x}_i] = 0, \forall i = 1, 2, \ldots, n$, where the expectation is taken with respect to the conditional distribution of ϵ given $\mathbf{x}$. The non-parametric model can be defined as $y_i = g(\mathbf{x}_i) + \epsilon_i$, with $g(\mathbf{x}_i)$: Unknown and $E[\epsilon_i|\mathbf{x}_i] = 0$.

For inference on $\boldsymbol{\beta}$ in the parametric case (or on $g(.)$, in the non-parametric case), at least three statistical procedurals are needed. First, an estimation method needs to be developed to calculate the estimate of the coefficients of vector $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$. Second, it needs to be proven that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ and that it has certain asymptotic properties. Third, test statistics need to be developed for testing hypotheses with the form:

$$H_0 \ : \mathbf{H}\boldsymbol{\beta} = \mathbf{b}_0 \quad \text{v.s} \quad H_a \ : \mathbf{H}\boldsymbol{\beta} \neq \mathbf{b}_0$$

where normally $\mathbf{H}_{r\times s}$, $\boldsymbol{\beta}_{s\times 1}$ and $\mathbf{b_0}_{r\times 1}$ are a constant matrix, a regression-coefficients vector and a constant vector respectively.

2.2. Regression in Sub-Spaces of Variables

In many practical situations, the number of variates is much greater than the quantity of available observations in the data set for a regression model, causing the problem of multicollinearity between the predictors. Among the available options for handling this problem are techniques based in explicit or implicit subspaces and the Bayesian approach, which includes additional information about the parameters of the model. In the case of the sub-spaces, the regression is

realized within a feasible space of a lesser dimension. The sub-space may be constructed explicitly with a geometric-type motivation derived from the use of latent variables, or implicitly using regularization techniques to avoid the problem of multicollinearity. A latent variable is a non-observable variable that is inferred from other variables by being directly observed and measured. The introduction of latent variables allows to capture more relevant information about the covariates matrix, denoted by $\mathbf{X}$, or information about the structure of the interaction between $\mathbf{X}$ and the response variables matrix, denoted by $\mathbf{Y}$.

In this approach, latent, non-correlated variables are introduced, denoted by $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_a$ and $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_a$, where a is the number of componets retained. The use of latent variables allows for the factorization of low ranges of the predictor and/or the response matrix, which allows for the adjustment of a linear regression model by least squares upon this set of latent variables.

The vectors loadings $\mathbf{p}_k$ and $\mathbf{q}_k$, with $k = 1, 2, \ldots, a$, generate a-dimensional spaces, where the coefficients $\mathbf{t}_{k\ n\times 1}$ and $\mathbf{u}_{k\ n\times 1}$ are considered as latent variables. Among the approaches based on latent variables are PCR and PLS regression, which are briefly described below.

In PC regression, which was introduced in Massy (1965), latent variates called principal components are obtained out of the correlation matrix $\mathbf{X}$, denoted by $\mathbf{R}$. PC regression avoids the problem of multicollinearity by reducing the dimension of the predictors. The loadings $\{\mathbf{p}_k\}_{k=1}^{a}$ are taken as a-first eigenvectors of the spectral decomposition of $\mathbf{R}$ matrix, and these vectors are the directions that maximize the variance of the principal components. The principal components are defined using the projections of the $\mathbf{X}$'s upon these directions. That is, the ith principal component of $\mathbf{X}$ is defined as $\mathbf{t}_k = \mathbf{X}\mathbf{p}_k$ so that $\mathbf{p}_k$ maximizes the variance of $\mathbf{t}_k$,

$$\max_{\mathbf{p}_k} \langle \mathbf{X}\mathbf{p}_k, \mathbf{X}\mathbf{p}_k \rangle = \max_{\mathbf{p}_k} \mathbf{p}_k^T \mathbf{X}^T \mathbf{X}\mathbf{p}_k$$

with $\mathbf{p}_k^T\mathbf{p}_k = 1$ y $\mathbf{p}_k^T\mathbf{p}_l = 0$, $l < k$. The principal components represent the selection of a new coordinate system obtained when rotating the original system of axes, $X_1, X_2, \ldots, X_p$. All of the loadings or principal directions are then obtained, $\mathbf{P} = [\mathbf{p}_1|\mathbf{p}_2|\cdots|\mathbf{p}_a]_{p\times a}$, as are the projections of the $X_i's$ on $\mathbf{p}_k's$, that is, all of the principal components, $\mathbf{T} = [\mathbf{t}_1|\mathbf{t}_2|\cdots|\mathbf{t}_a]_{n\times a}$, with the restrictions $\langle \mathbf{t}_k, \mathbf{t}_l \rangle = 0$ and $\langle \mathbf{t}_k, \mathbf{t}_k \rangle = Var(\mathbf{t}_k) = \lambda_k$, with λ_k: the eigenvalues associated with the eigenvectors P_k with $\lambda_1 \geq \lambda_2 \geq \ldots, \lambda_a$. A regression model of $\mathbf{Y}$ is then adjusted against the latent variates $\mathbf{T}$. Then, the response for $\mathbf{Y}$-new ones is predicted associated with new observations of the predictors vector. In PC regression, the principal components in the predictor space $\mathbf{X}$'s are used without taking into account the information of the responses $\mathbf{Y}$'s.

PLS regression was introduced in Wold (1975) and applied in the economic and social sciences fields. However, due to the contributions made by his son in Wold, Albano, Dunn, Edlund, Esbensen, Geladi, Hellberg, Johansson, Lindberg & Sjöström (1984), it gained great popularity in the area of chemometrics, where data characterized by many predictor variables with multicollinearity problems and few available observations are analyzed. This happens in many studies of imaging analysis. The PLS regression methodology generalizes and combines

characteristics of Principal Component Analysis (PCA) and Multiple Regression Analysis (MLR). Its demand and evidence has increased and it is being applied in many scientific areas. PLS regression is similar to the canonic correlation analysis (CCA), but instead of maximizing the correlation, it maximizes the covariance between the components. That is, $\mathbf{p}$ and $\mathbf{q}$ directions are found so that

$$\max_{\mathbf{p},\mathbf{q}} \langle \mathbf{Xp}, \mathbf{Yq} \rangle = \max_{\mathbf{p},\mathbf{q}} \mathbf{p}^T \mathbf{X}^T \mathbf{Yq}$$

subject to $\|\mathbf{p}\| = \|\mathbf{q}\| = 1$

In general, the PLS regression is a two-phase process. First, the predictor matrix $\mathbf{X}$ is transformed with the help of the vector of response variables, $\mathbf{Y}$, in a matrix of latent, non-correlated variables $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p)$, called PLS components. This distinguishes it from the PLS regression, in which the components are obtained using only the predictor matrix, $\mathbf{X}$. Second, the estimated regression model is adjusted using the original response vector and the PLS components as predictors, and then, response for $\mathbf{Y}$'new ones associated with future observations of the repetition vector are of predict. A reduction of dimensionality is obtained directly on the PLS components because they are orthogonal, and the number of components necessary for the regression analysis is much lower than the number of original predictors. The process of maximizing the covariance instead of the correlation prevents the possible problem of numeric instability that can appear when using correlation, which is due to the division of covariances by variances that may be too small. The directions of the maximum covariance p and q among the PLS components can be found by the following eigen-decomposition problem:

$$\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{Xp} = \lambda \mathbf{p} \quad \text{and} \quad \mathbf{Y}^T \mathbf{X} \mathbf{X}^T \mathbf{Yq} = \lambda \mathbf{q}$$

with $\|\mathbf{p}\| = \|\mathbf{q}\| = 1$. The latent variates (or PLS components) are calculated by projecting the $\mathbf{X}$ and $\mathbf{Y}$ data in the $\mathbf{p}$ and $\mathbf{q}$ directions, that is, $\mathbf{t} = \mathbf{Xp}$ and $\mathbf{u} = \mathbf{Yq}$ results in all latent components being obtained such that $\mathbf{T} = \mathbf{XP}$ and $\mathbf{U} = \mathbf{YQ}$.

3. Geometrical Structure of $\mathrm{Sym}^+(m)$

A summary will now be given of some of the basic results of (Schwartzman 2006) on the geometric structure of the $\mathrm{Sym}^+(m)$ set as a Riemannian manifold. The $\mathrm{Sym}^+(m)$ space is a sub-manifold of the Euclidian space $\mathrm{Sym}(m)$. Geometrically, the $\mathrm{Sym}^+(m)$ and $\mathrm{Sym}(m)$ spaces are differential manifolds of $m(m+1)/2$ dimensions, and they are homeomorphically related by an exponential and logarithmic transformation matrix. For any matrix $\mathbf{A} \in \mathrm{Sym}(m)$, its exponential matrix is given by $\exp(\mathbf{A}) = \sum_{k=1}^{\infty} \frac{\mathbf{A}^k}{k!} \in \mathrm{Sym}^+(m)$. Reciprocally, for any matrix $\mathbf{S} \in \mathrm{Sym}^+(m)$, there is a $\log(\mathbf{S}) = \mathbf{A} \in \mathrm{Sym}(m)$, such that $\exp(\mathbf{A}) = \mathbf{S}$.

For responses in Euclidian spaces in non-parametric standard regression models, $E[\mathbf{S}|X = x]$ is estimated. However, for responses on a curved space, the conditional expectancy of $\mathbf{S}$, given $x = x$, cannot be defined. For $\mu(x) = E[\mathbf{S}|X = x]$,

a tangent vector is introduced in $\mu(x)$ on $\text{Sym}^+(m)$. For a small scalar $\delta > 0$, the differentiable map $C : (-\delta, \delta) \longrightarrow \text{Sym}^+(m), \quad t \to C(t)$, is considered such that $C(0) = \mu(x)$. A tangent vector in $\mu(x)$ is defined as the derivative of the soft curve $C(t)$, with respect to t, valued at $t = 0$. The set of all tangent vectors in $\mu(x)$ is called the tangent space of $\text{Sym}^+(m)$ in $\mu(x)$, and it is denoted by $T_{\mu(x)}\text{Sym}^+(m)$. This space can be identified by a copy of $\text{Sym}(m)$. The $T_{\mu(x)}\text{Sym}^+(m)$ space is equipped with an internal product $\langle \ . \ , \ . \ \rangle$, called a Riemannian metric, which varies softly from point to point. For example, the Frobenius metric can be used as a Riemannian metric. For a given Riemannian metric, $\langle \mathbf{u} \ , \ \mathbf{v} \rangle$ is calculated for any $\mathbf{u}$ and $\mathbf{v}$ in $T_{\mu(x)}\text{Sym}^+(m)$, and then, the length of the soft curve $C(t) : [t_0, t_1] \longrightarrow \text{Sym}^+(m)$ is calculated, which is equal to: $\|C(t)\| = \int_{t_0}^{t_1} \sqrt{\langle \dot{C}(t), \dot{C}(t) \rangle} dt$, where $\dot{C}(t)$ is the derivative of $C(t)$, with respect to t. A geodesic is a soft curve in $\text{Sym}^+(m)$ with tangent vectors that do not change in length or direction along the curve. For any $\mathbf{u} \in T_{\mu(x)}\text{Sym}^+(m)$, there is a single geodesic, denoted by $\gamma_{\mu(x)}(t; \mathbf{u})$, with a dominion that contains the range $[0, 1]$, such that $\gamma_{\mu(x)}(0; \mathbf{u}) = \mu(x)$ and $\dot{\gamma}_{\mu(x)}(0; \mathbf{u}) = \mathbf{u}$.

The exponential Riemannian map is defined as

$$\text{Exp}_{\mu(x)} : T_{\mu(x)}\text{Sym}^+(m) \longrightarrow \text{Sym}^+(m) \ \ ; \ \mathbf{u} \longrightarrow \text{Exp}_{\mu(x)}(\mathbf{u}) = \gamma_{\mu(x)}(1; \mathbf{u}) \quad (1)$$

The inverse of the exponential Riemannian map, called a Riemannian logarithmic map, is defined as

$$\text{Log}_{\mu(x)} : \text{Sym}^+(m) \longrightarrow \ T_{\mu(x)}\text{Sym}^+(m) \ \ ; \ \mathbf{S} \longrightarrow \text{Log}_{\mu(x)}(\mathbf{S}) = \mathbf{u} \quad (2)$$

such that $\text{Exp}_{\mu(x)}(\mathbf{u}) = \mathbf{S}$. Finally, the shortest distance between 2 points $\mu_1(x)$ and $\mu_2(x)$ in $\text{Sym}^+(m)$, is called the geodesic distance and is denoted by $g(\mu_1(x), \mu_2(x))$, which satisfies

$$d_g^2(\mu_1(x), \mu_2(x)) = \langle \text{Log}_{\mu_1(x)}\mu_2(x), \text{Log}_{\mu_1(x)}\mu_2(x) \rangle = \|\text{Log}_{\mu_1(x)}\mu_2(x)\|_g^2 \quad (3)$$

where $d_g^2(. \ , \ .)$, denoted the geodesic distance.

The residual from $\mathbf{S}$ with respect to $\mu(x)$, denoted by $\varepsilon_\mu(x)$, is defined as $\varepsilon_\mu(x) = \text{Log}_{\mu(x)}\mathbf{S} \in T_{\mu(x)}\text{Sym}^+(m)$. The vectorization of $C = [c_{ij}] \in \text{Sym}(m)$ is defined as $\text{Vecs}(C) = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1m} & c_{22} & \dots & c_{2m} & \dots & c_{mm} \end{bmatrix}^T \in \mathbb{R}^{\frac{m(m+1)}{2}}$. The conditional expectancy of $\mathbf{S}$, given $\mathbf{x} = x$, is defined as the matrix $\mu(x) \in \text{Sym}^+(x)$, such that

$$E[\text{Log}_{\mu(x)}\mathbf{S}|X = x] = E[\varepsilon_\mu(x)|X = x] = \mathbf{0}_{m \times m} \quad (4)$$

where the expectancy is taken component by component with respect to the $m(m+1)$-vector aleatory multivaried $\text{Vecs}[\text{Log}_{\mu(x)}S] \in \mathbb{R}^{\frac{m(m+1)}{2}}$.

3.1. Regression Model for Response Data in $\text{Sym}^+(m)$

Because the DTs are in a non-linear space, it is theoretically and computationally difficult to develop a formal statistical framework that includes estimation

theory and hypothesis tests where by a set of covariates are used to directly predict DTs as responses. With the recently developed log-Euclidian metric Arsigny, Fillard, Pennec & Ayache (2006), DTs can be transformed from non-linear space into logarithmic matrices in a Euclidian space. Zhu et al. (2009) developed a regression model with the log-transformation of the DTs as the response. The model was based on a semi-parametric method, which avoids the specification of parametric distributions for aleatory log-transformed DTs. Inference processes have been proposed for estimating the regression coefficients and test statistics of this model to contrast linear hypotheses of unknown parameters as well as to test processes based on re-sampling methods to simultaneously evaluate the statistical significance of linear hypotheses throughout large ROIs. The procedure for the laying out of the local intrinsic polynomial regression model (RPLI) for SPD matrices as a response is described below, ver Zhu et al. (2009).

The procedure to estimate $\mu(x) = E[\mathbf{S}|X = x_0]$ in the RPLI model will now be described. Because $\mu(x)$ is on a curved space, it cannot be directly expand to $\mu(x)$ in $\mathbf{x} = x_0$ using a Taylor series. Instead, the Riemannian logarithmic map of $\mu(x)$ in $\mu(x_0)$ on the space $T_{\mu(x)}\text{Sym}^+(m)$ is considered, that is, we are considering $\text{Log}_{\mu(x_0)}\mu(x) \in T_{\mu(x)}\text{Sym}^+(m)$. Because $\text{Log}_{\mu(x_0)}\mu(x)$ occupies a different tangent space for each value of $\mathbf{X}$, it can be transported from the common tangent space $T_{I_m}\text{Sym}^+(m)$ through the parallel transport given by:

$$\Phi_{\mu(x_0)} : T_{\mu(x_0)}\text{Sym}^+(m) \longrightarrow T_{I_m}\text{Sym}^+(m); \\ \text{Log}_{\mu(x_0)}\mu(x) \longrightarrow \Phi_{\mu(x_0)}(\text{Log}_{\mu(x_0)}\mu(x)) = Y(x) \quad (5)$$

Its inverse is given by $\text{Log}_{\mu(x_0)}\mu(x) = \Phi^{-1}_{\mu(x_0)}(Y(x)) \in T_{\mu(x_0)}\text{Sym}^+(m)$.

For $\text{Log}_{\mu(x_0)}\mu(x_0) = O_m \in T_{\mu(x_0)}Sym^+(m)$, because $\Phi_{\mu(x_0)}(O_m) = Y(x_0) = O_m$ and because $Y(x)$ y $Y(x_0)$ are in the same tangent space $T_{I_m}\text{Sym}^+(m)$, a Taylor series expansion can be used for $Y(x)$ in x_0. The following is obtained:

$$\text{Log}_{\mu(x_0)}\mu(x) = \Phi^{-1}_{\mu(x_0)}(Y(x)) \approx \Phi^{-1}_{\mu(x_0)}\left(\sum_{k=1}^{k_0} Y^{(k)}(x_0)(x - x_0)^k\right) \quad (6)$$

with k_0 as a whole and $Y^{(k)}$ as the kth derivative of $Y(x)$ with respect to x divided by por $k!$. Equivalently,

$$\mu(x) = \text{Exp}_{\mu(x_0)}\left(\Phi^{-1}_{\mu(x_0)}(Y(x))\right) = \\ \text{Exp}_{\mu(x_0)}\left(\Phi^{-1}_{\mu(x_0)}\left(\sum_{k=1}^{k_0} Y^{(k)}(x_0)(x - x_0)^k\right)\right) = \mu\left(x, \alpha(x_0), k_0\right) \quad (7)$$

where $\alpha(x_0)$-contains all the parameters in $\{\mu(x_0), Y^{(1)}(x_0), \ldots, Y^{(k)}(x_0)\}$.

For a set of vectors in $T_{\mu(x)}\text{Sym}^+(m)$, various Riemannian metrics can be defined. Among these metrics is the log-Euclidian metric, and some of its basic properties will now be reviewed. Notations exp(.) and log(.) are used to represent the exponential and log matrices, respectively; Exp and Log are used to

represent the exponential and logarithmic maps, respectively. The differential of the logarithmic matrix in $\mu(x) \in \text{Sym}^+(m)$ is denoted by $\partial_{\mu(x)}\log.(\mathbf{u})$, which acts on an infinitesimal movement $\mathbf{u} \in T_{\mu(x)}\text{Sym}^+(m)$. The log-Euclidian metric on $\text{Sym}^+(m)$ is defined as:

$$\langle \mathbf{u}, \mathbf{v} \rangle := tr\left[(\partial_{\mu(x)}\log.\mathbf{u})(\partial_{\mu(x)}\log.\mathbf{v})\right] \tag{8}$$

for $\mathbf{u}, \mathbf{v} \in T_{\mu(x)}\text{Sym}^+(m)$.

The geodesic $\gamma_{\mu(x)}(t;\mathbf{u})$-is given by:

$$\gamma_{\mu(x)}(t;\mathbf{u}) := \exp\left[\log(\mu(x)) + t\partial_{\mu(x)}\log.\mathbf{v}\right] \ , \quad \forall t \in \mathbb{R} \tag{9}$$

The differential of the exponential matrix is denoted by $\partial_{\log(\mu(x))}\exp.(\mathbf{A})$, in $\log(\mu(x)) \in \text{Sym}(m) = T_{\mu(x)}\text{Sym}^+(m)$ which acts on an infinitesimal movement $\mathbf{A} \in T_{\log(\mu(x))}\text{Sym}^+(m)$. The exponential and logarithmic Riemannian maps are defined, respectively, as follows: for $\mathbf{S} \in \text{Sym}^+(m)$,

$$\begin{aligned} \text{Exp}_{\mu(x)}(\mathbf{u}) &:= \exp\left[\log(\mu(x)) + \partial_{\mu(x)}\log.(\mathbf{u})\right]; \\ \text{Log}_{\mu(x)}(\mathbf{S}) &:= \partial_{\log(\mu(x))}\exp\left[\log(\mathbf{S}) - \log(\mu(x))\right] \end{aligned} \tag{10}$$

For $\mu(x)$ and $\mathbf{S} \in \text{Sym}^+(m)$, the geodesic distance is given by:

$$d_g^2(\mu(x), \mathbf{S}) := tr\left[(\log\,\mu(x) - \log(\mathbf{S}))^{\otimes 2}\right] \tag{11}$$

with $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ and with $\mathbf{a}$-vector. For two matrices $\mu(x)$ and $\mu(x_0) \in \text{Sym}^+(m)$ and any $\mathbf{u}_{\mu(x_0)} \in T_{\mu(x_0)}\text{Sym}^+(m)$, the parallel transport is defined as follows:

$$\begin{aligned} \Phi_{\mu(x_0)} : T_{\mu(x_0)}\text{Sym}^+(m) &\longrightarrow T_{I_m}\text{Sym}^+(m); \\ \mathbf{u}_{\mu(x_0)} &\longrightarrow \Phi_{\mu(x_0)}(\mathbf{u}_{\mu(x_0)}) := \partial_{\mu(x_0)}\log.\left(U_{\mu(x_0)}\right) \end{aligned}$$

If $\mathbf{u}_{\mu(x_0)} = \text{Log}_{\mu(x_0)}\mu(x) \in T_{\mu(x_0)}\text{Sym}^+(m)$, then

$$Y(x) = \Phi_{\mu(x_0)}\left(\text{Log}_{\mu(x_0)}\mu(x)\right) = \log\,\mu(x) - \log\,\mu(x_0) \tag{12}$$

and $\mu(x) = \exp\left[\log\,\mu(x_0) + Y(x)\right]$.

The residual of $\mathbf{S}$ with respect to $\mu(x)$ is defined as: $\varepsilon_\mu(x) := \log(\mathbf{S}) - \log(\mu(x))$ with $E[\log\mathbf{S}|X = x] = \log\,\mu(x)$. The model RPLI is defined as:

$$\log(\mathbf{S}|x) = \log(\mu(x)) + \varepsilon_\mu(x) \tag{13}$$

with $E[\varepsilon_\mu(x)] = 0$, which indicates that $E[\log\mathbf{S}|X = x] = \log(\mu(x))$.

4. The PLS Regression Model

Suppose we have n DTs, denoted by $\mathbf{T}_i : i = 1, 2, \ldots, n$, obtained from a voxel correspondent with a normalized and especially re-oriented DTI from n subjects. The log-transformation of T_k is then obtained, which is denoted by

$$\mathbf{L}_{T,i} = (L^i_{T_{(1,1)}}, L^i_{T_{(1,2)}}, L^i_{T_{(1,3)}}, L^i_{T_{(2,2)}}, L^i_{T_{(2,3)}}, L^i_{T_{(3,3)}})^T \tag{14}$$

where $L^i_{T_{(j,k)}}$-denotes the (j,k)-element of the logarithm matrix of $\mathbf{T}_k$. For each individual, a set of covariates of interest is observed as well.

In studies of medical images, many demographic or clinical measurements are normally observed for different patients considered in a certain study. The amount of available information is abundant, and there may be problems of linear dependences between the covariates of interest, which generates the problem of multicollinearity. In addition, available data to analyze the information are scarce. For the log-transformed DTs, a linear model is considered, which is given by:

$$\underset{1\times 6}{\mathbf{L}_{T,i}} = \underset{1\times pp\times 6}{\mathbf{x}_i\ \boldsymbol{\beta}} + \underset{1\times 6}{\boldsymbol{\varepsilon}_i}\ , \quad i = 1, 2, \ldots, n \tag{15}$$

or

$$\underset{n\times 6}{\mathbf{L}_T} = \underset{n\times pp\times 6}{\mathbf{X}\ \mathbf{B}} + \underset{n\times 6}{\boldsymbol{\varepsilon}} \tag{16}$$

with $E[\boldsymbol{\varepsilon}|\mathbf{x}] = \mathbf{0}_{n\times p}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}|\mathbf{x}) = \Sigma_{np\times np}$ and where $\mathbf{X}$, $\mathbf{Y}$=$\mathbf{L}$, $\mathbf{B}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\Sigma}$, are matrices representing the covariates, responses, regression coefficients, the model errors and covariance of $\boldsymbol{\varepsilon}|\mathbf{x}$.

Compared to the general lineal model, the model, based on the conditional mean and covariance in equation (16)does not assume any distributional suppositions for the image measurements.

If $\boldsymbol{\theta}_{(6p+21)\times 1}$ is the vector of unknown parameters contained in $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, then to estimate $\boldsymbol{\theta}$, the objective function given by:

$$l_n(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n}\left(\log|\boldsymbol{\Sigma}| + (L_{T,i} - \boldsymbol{\beta}\mathbf{x}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{L}_{T,i} - \boldsymbol{\beta}\mathbf{x}_i)\right) \tag{17}$$

is maximized using the iterative algorithm proposed by Li et al. (2009).

The regression model (16) has been adjusted using existing algorithms for PC and PLS regression, following the steps described in Section 2.2 and taking into account the log-transformations on the original data to transfer them to a Euclidian space.

4.1. Evaluation of the PLS Regression Model with Simulated Data

The behavior of the PLS regression model is evaluated with sets of simulated data, and predicted results are compared with those obtained using the PC technique in the case of a design matrix of full range.

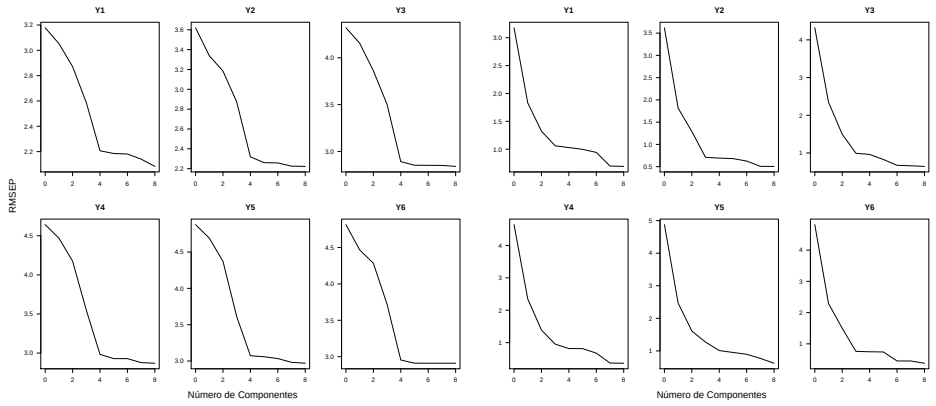
The settings considered to simulate the data are the following. First, a sample of SPD matrices with a size of $n = 20$ with $k = 15$ covariates was generated from a multivariate normal distribution with a mean of zero and a covariance structure given by $\boldsymbol{\Sigma} = 0.6\mathbf{I}_6$. Then, the sample size was increased to $n = 30$, and the number of covariates was increased from $k = 15$ to $k = 40$, with a covariance structure given by $\boldsymbol{\Sigma} = 0.3\mathbf{I}_6 + 0.61_6\mathbf{1}_6^T$, with $\mathbf{1}_6$, a vector of ones. In both settings, the values for the coefficients of beta were used in the matrix given by

$p \times 6$, $\beta_k = [1 + 0.1 \times (k-1)]^T$. The exponential of $\mathbf{\Sigma}$ was calculated to ensure its positive definiteness. Results obtained in each scenario are expounded below.

For the first setting, shown in Table 1, the percentages of variance explained by each of the latent components through PC and PLS regression demonstrate that PC explains more of the variability of $\mathbf{X}$ than PLS regression, which is a typical result. In Table 2, the PLS components explain a higher percentage of the variability of $\mathbf{Y}$ than the PC components; with two components, more than 80% of the variability in $\mathbf{Y}$ and approximately 20% of the variability in $\mathbf{X}$ is explained. Figure 1 shows the graphs of the square root of the prediction middle quadratic error (RMSEP) against the number of components used in the cross validation (CV). Here, it can be observed that in PC, approximately four components would be needed to explain a majority of the variability in the data. However, in PLS regression, three components are needed in most cases. In general, few repetition are shown through this illustration of the repetition results obtained by each method, when compared with the simulation. Figure 2 shows the graphs of the predicted data with the observed responses. A greater precision in the adjustment can be observed when PLS regression is used. For the second setting, Table 3 shows the percentages of variance explained by each of the latent components using PC and PLS regression. Again, PC explains more of the variability of $\mathbf{X}$ than PLS regression. Table 4 shows that the PLS components explain a greater percentage of the variability of $\mathbf{Y}$ than the PLS components. In five components, more than 60% of the variability in $\mathbf{Y}$ and approximately 35% of the variability in $\mathbf{X}$ is explained. Figure 3 shows the graphs for the RMSEP against the number of components. It can be observed that in PC, approximately 7 components would be needed to explain most of the variability of the data, while in PLS regression, five components are needed in most cases. Figure 4 shows the graphs of the predicted data along with the observed values of the responses; a greater precision in the adjustment can be observed when PLS regression is used.

Table 1: Percentages of variance explained by each component.

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8
PC	17.57	15.55	13.59	12.46	11.16	9.16	6.81	4.64
PLS	14.27	9.93	10.16	13.45	12.60	5.75	4.46	7.07



(a) PC regression (b) PLS regression

FIGURE 1: RMSEP versus number of components by PC regression and PLS regression

TABLE 2: Percentages of variance explained cumulated of $\mathbf{X}$ and $\mathbf{Y}$ for the components by PC and PLS regression.

		Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8
PC	X	17.57	33.11	46.70	59.16	70.32	79.48	86.28	90.93
PLS	X	14.27	24.20	34.37	47.82	60.43	66.17	70.64	77.70
PC	Y1	7.69	18.38	33.74	51.79	52.72	52.91	54.64	57.04
PLS	Y1	66.85	82.64	88.85	89.51	90.13	91.21	95.17	95.26
PC	Y2	14.95	22.65	36.98	58.96	60.99	61.09	62.21	62.29
PLS	Y2	74.87	87.30	96.16	96.35	96.46	97.01	98.04	98.05
PC	Y3	7.45	20.12	34.30	55.21	56.38	56.43	56.44	56.77
PLS	Y3	70.51	88.00	94.72	95.05	96.33	97.57	97.67	97.78
PC	Y4	7.30	19.10	41.57	58.71	60.19	60.20	61.57	61.78
PLS	Y4	74.39	91.05	95.78	96.90	96.92	97.87	99.36	99.39
PC	Y5	7.44	19.65	45.13	60.30	60.66	61.30	62.61	62.93
PLS	Y5	74.38	89.10	93.22	95.70	96.19	96.62	97.51	98.38
PC	Y6	13.89	20.83	40.31	62.35	63.45	63.46	63.46	63.47
PLS	Y6	77.35	90.32	97.51	97.60	97.63	99.12	99.12	99.38

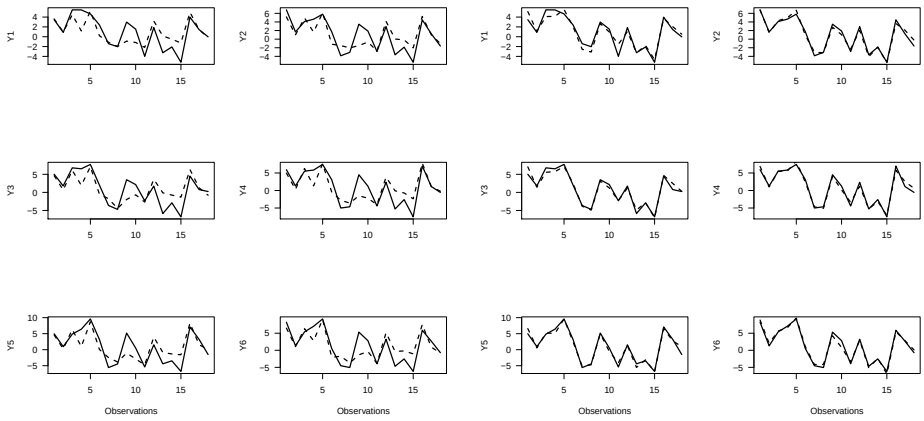
TABLE 3: Percentages of variance explained by each component, 2.

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10
PC	12.81	9.33	8.74	7.42	7.22	6.39	6.33	5.12	4.97	4.44
PLS	10.63	8.65	6.39	5.21	3.85	5.34	4.88	5.36	5.32	5.00

5. Conclusions and Recommendations

A PLS linear regression model is proposed in this article to study the relationship between a large set of covariates of interest in a Euclidian space with a set of response variables in a symmetric Riemannian space. The theory of exponential and Riemannian maps has been used to transform data from a non-Euclidian space into a Euclidian space of symmetrical matrices, where the methodology has been developed. Results indicate support for the proposed methodology as compared to

a technique using regression by major components, as has been observed in classic situations of data analysis in euclidean spaces with matrices of covariates presenting high multicollinearity, or in problems with a low number of observations and many covariates. In future works, we will investigate more realistic models, such as non-linear PLS models for the types of SPD matrix data discussed in this study and other types of manifold-valued data, such as data obtained by geometric representations of objects via medial axial representation (m-rep), orthogonal rotation groups, and other methods. The illustration presented in this article for simulated data favorably sheds light on the results that can be obtained by applying these types of models to real data.

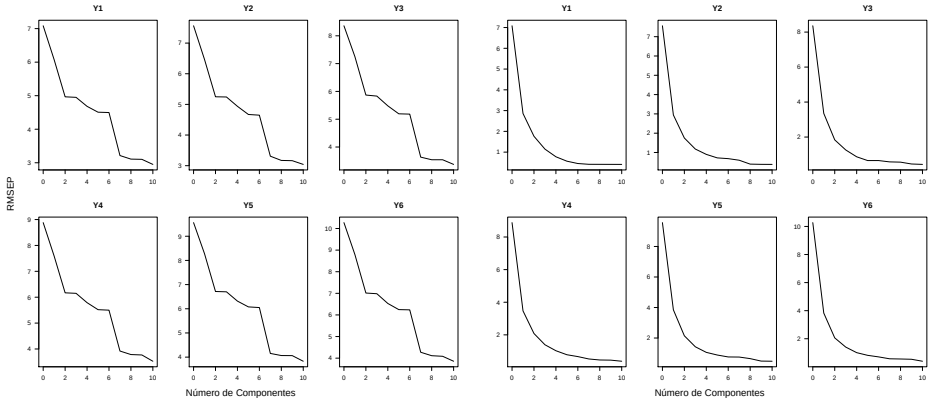


(a) PC regression (b) PLS regression

FIGURE 2: Predicted values with the observables values by PC regression and PLS regression. Solid lines: Observed, dashed lines: Predicted.

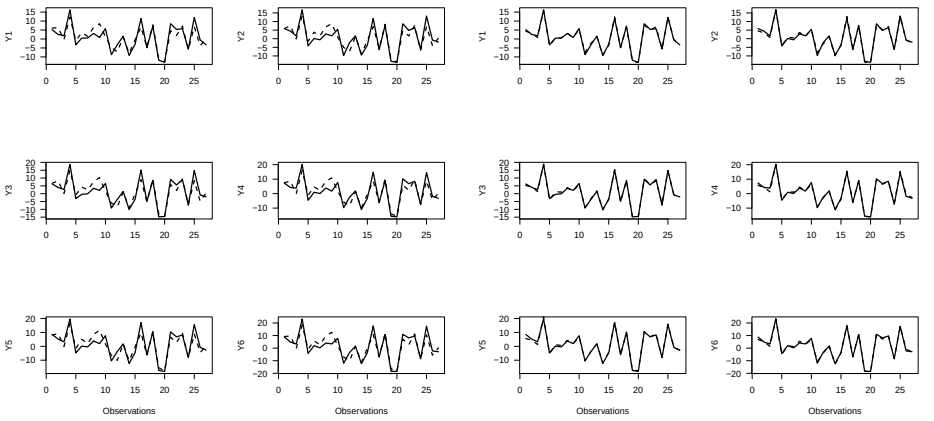
TABLE 4: Percentages of variance explained cumulated of $\mathbf{X}$ and $\mathbf{Y}$ for the components by PC and PLS regression, 2.

		Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10
PC	X	12.81	22.14	30.88	38.30	45.52	51.90	58.23	63.35	68.33	72.77
PLS	X	10.63	19.28	25.67	30.88	34.73	40.07	44.95	50.32	55.64	60.64
PC	Y1	26.52	50.85	51.17	56.31	59.51	59.69	79.40	80.74	80.83	82.65
PLS	Y1	83.70	93.81	97.39	98.80	99.37	99.59	99.66	99.66	99.66	99.67
PC	Y2	26.97	51.87	51.99	57.41	61.87	62.25	80.86	82.42	82.52	83.83
PLS	Y2	84.85	94.65	97.57	98.58	99.09	99.19	99.37	99.72	99.74	99.74
PC	Y3	24.82	50.72	51.34	57.02	61.40	61.56	81.08	82.05	82.05	83.70
PLS	Y3	83.92	95.16	97.72	98.91	99.38	99.38	99.52	99.54	99.70	99.73
PC	Y4	27.00	51.74	52.05	57.50	61.39	61.65	80.51	81.84	81.99	84.23
PLS	Y4	84.74	94.50	97.54	98.67	99.23	99.44	99.66	99.73	99.74	99.81
PC	Y5	25.11	50.70	50.90	56.36	59.61	59.97	81.14	81.93	81.96	83.96
PLS	Y5	83.80	94.97	97.77	98.77	99.14	99.37	99.38	99.54	99.74	99.75
PC	Y6	26.75	53.38	53.80	59.58	63.02	63.15	82.70	83.96	84.18	85.90
PLS	Y6	86.10	95.97	98.12	99.03	99.37	99.53	99.69	99.71	99.73	99.85



(a) PC regression (b) PLS regression

FIGURE 3: RMSEP versus number of components by PC regression and PLS regression, 2.



(a) PC regression (b) PLS regression

FIGURE 4: Predicted values with the observables values by PC regression and PLS regression, 2. Solid lines: Observed, dashed lines: Predicted.

[Recibido: junio de 2012 — Aceptado: mayo de 2013]

References

Arsigny, V., Fillard, P., Pennec, X. & Ayache, N. (2006), 'Log-euclidean metrics for fast and simple calculus on diffusion tensors', *Magnetic Resonance in Medicine,* **56**, 411–421.

Barmpoutis, A., Vemuri, B. C., Shepherd, T. M. & Forder, J. R. (2007), 'Tensor splines for interpolation and approximation of DT-MRI with applications to segmentation of isolated rat hippocampi', *IEEE Transations on Medical Imaging,* **26**, 1537–1546.

Batchelor, P., Moakher, M., Atkinson, D., Calamante, F. & Connelly, A. (2005), 'A rigorous framework for diffusion tensor calculus', *Magnetic Resonance in Medicine,* **53**, 221–225.

Fingelkurts, A. A. & Kahkonen, S. (2005), 'Functional connectivity in the brain - is it an elusive concepts?', *Neuroscience and Biobehavioral Reviews,* **28**, 827–836.

Fletcher, P. T. & Joshi, S. (2007), 'Riemannian geometry for the statistical analysis of diffusion tensor data', *Signal Processing,* **87**, 250–262.

Grenander, U. & Miller, M. I. (1998), 'Computational anatomy: An emerging discipline', *Quarterly of Applied Mathematics,* **56**, 617–694.

Kim, P. T. & Richards, D. S. (2010), 'Deconvolution density estimation on spaces of positive definite symmetric matrices', *IMS Lecture Notes Monograph Series. A Festschrift of Tom Hettmansperger* .

Lepore, N., Brun, C. A., Chou, Y., Chiang, M., Dutton, R. A., Hayashi, K. M., Luders, E., Lopez, O. L., Aizenstein, H. J., Toga, A. W., Becker, J. T. & Thompson, P. M. (2008), 'Generalized tensor-based morphometry of HIV/AIDS using multivariate statistics on deformation tensors', *IEEE Transactions in Medical Imaging,* **27**, 129–141.

Li, Y., Zhu, H., Chen, Y., Ibrahim, J. G., An, H., Lin, W., Hall, C. & Shen, D. (2009), RADTI: Regression analysis of diffusion tensor images, *in* E. Samei & J. Hsieh, eds, 'Progress in Biomedical Optics and Imaging - Proceedings of SPIE', Vol. 7258.

Massy, W. F. (1965), 'Principal components regression in exploratory statistical research', *Journal of the American Statistical Association,* **64**, 234–246.

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

Schwartzman, A. (2006), Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data, PhD thesis, Stanford University.

Wold, H. (1975), 'Soft modeling by latent variables; the non-linear iterative partial least squares approach', *Perspectives in Probability and Statistics,* pp. 1–2.

Wold, S., Albano, C., Dunn, W.J., I., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W. & Sjöström, M. (1984), Multivariate data analysis in chemistry, *in* B. Kowalski, ed., 'Chemometrics', Vol. 138 of *NATO ASI Series*, Springer Netherlands, pp. 17–95.

Yuan, Y., Zhu, H., Lin, W. & Marron, J. S. (2012), 'Local polynomial regression for symmetric positive-definite matrices', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(4), 697–719.

Zhu, H. T., Chen, Y. S., Ibrahim, J. G., Li, Y. M. & Lin, W. L. (2009), 'Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging', *Journal of the American Statistical Association* **104**, 1203–1212.