Spherical Regression Models Using Projective Linear Transformations

Michael Rosenthal, Wei Wu, Eric Klassen & Anuj Srivastava
Accepted author version posted online: 07 Mar 2014. Published online: 22 Dec 2014.

To link to this article: http://dx.doi.org/10.1080/01621459.2014.892881
Spherical Regression Models Using Projective Linear Transformations

Michael ROSENTHAL, Wei WU, Eric KLASSEN, and Anuj SRIVASTAVA

This article studies the problem of modeling relationship between two spherical (or directional) random variables in a regression setup. Here the predictor and the response variables are constrained to be on a unit sphere and, due to this nonlinear condition, the standard Euclidean regression models do not apply. Several past papers have studied this problem, termed spherical regression, by modeling the response variable with a von Mises-Fisher (VMF) density with the mean given by a rotation of the predictor variable. The few papers that go beyond rigid rotations are limited to one- or two-dimensional spheres. This article extends the mean transformations to a larger group—the projective linear group of transformations—on unit spheres of arbitrary dimensions, while keeping the VMF density to model the noise. It develops a Newton–Raphson algorithm on the special linear group for estimating the MLE of regression parameter and establishes its asymptotic properties when the sample-size becomes large. Through a variety of experiments, using data taken from projective shape analysis, cloud tracking, etc., and some simulations, this article demonstrates improvements in the prediction and modeling performance of the proposed framework over previously used models. Supplementary materials for this article are available online.

KEY WORDS: Intrinsic Newton-Raphson; Nonlinear; Spherical data.

1. INTRODUCTION

Many scientific problems require statistical analyses of data that are completely restricted to spherical domains. The statistical analysis of variables under unit norm constraints was pioneered by Mardia and colleagues (1972, 2000), in the context of directional data. Examples of these situations can also be frequently found in image analysis, shape analysis, signal processing, geology, meteorology, bioinformatics, and so on. While most of the work in directional statistics is focused on analyzing individual random variables, for example, imposing probability models—in the form of parametric families—and developing efficient estimators, one often needs to model relationships between two or more directional variables. This problem garnered a lot of interest from statisticians in the 80s and 90s, especially by Chang (1986, 1989) who laid foundations for early developments in this area, followed closely by several others including Downs (2003), Kim (1991), and Rivest (1989). In Presnell, Morrison, and Littell (1998), the authors studied projected linear regression models in situations where only the response variable was directional and one-dimensional.

Due to a growing awareness of geometrical and computational tools there has been a steady increase in formulations of statistical inferences on nonlinear Riemannian manifolds. As an example, the use of statistical inferences on nonlinear manifolds is the centerpiece of advances in shape analysis of landmarks, curves, and surfaces (Dryden and Mardia 1998; Kaziska and Srivastava 2007; Patrangenaru, Liu, and Sugathadasa 2010; Castillo and Colosimo 2011; Srivastava et al. 2011; Kurtek et al. 2012). Although a variety of Riemannian manifolds have been featured in general pattern analysis problems, the most frequent and basic nonlinear domain for formulating such problems is a unit sphere (Jung, Dryden, and Marron 2012). A simple unit-norm constraint on the data, that is, analysis of unit-length vectors, leads to the representation space becoming a unit sphere. In this article we focus on the problem of capturing relationships between random variables that take values on unit spheres of arbitrary dimensions.

The development of regression models involving predictor and response variables that take values on a unit sphere is termed spherical regression. Consider two random variables $x$ and $y$ that are elements of a unit sphere $S^{d-1}$ with $d \geq 2$. We are given $n$ paired-observations of these random variables: $(x_i, y_i) \in S^{d-1} \times S^{d-1}, i = 1, 2, \ldots, n$ and our goal is to find a regression model that relates the predictor variable $x$ to the response variable $y$. The difficulty in solving these problems using traditional methods is that $S^{d-1}$ is nonlinear. If these variables were simple vectors, that is, elements of a Euclidean space, then a common technique is to assume a model of the type: $y = \mu(x) + \epsilon$, where $\epsilon$ is typically a random quantity with zero mean (independent of $x$) and $\mu(x)$ is the conditional mean function that takes a parametric, semiparametric, or a nonparametric form (see, e.g., Seber and Lee 2003). In the current problem, one has to develop the corresponding terms for a spherical domain. There are two important choices to be made before proceeding further: the form of $\mu(x)$ and the density of $\epsilon$. Assuming that $\mu(x)$ takes a parametric form, with the parameter $\theta$, then it is the density of $\epsilon$ which decides the conditional density function $y|x, \theta$.

1.1 Past Literature and Its Limitations

We start by discussing these choices as presented in the current literature.

Conditional Mean Direction Function: One needs to choose a mapping $\mu: S^{d-1} \to S^{d-1}$ to serve as the conditional mean direction function. Similar to the Euclidean case, this mapping can take one of many forms—parametric, semiparametric, or...
is interesting to note that most of the papers (Chang 1986, 1989; Rivest 1989; Kim 1991; Prentice and Mardia 1995) used a rigid rotation for $\mu$, that is, $\mu(x) = Ax$, where $A \in SO(d)$, the special orthogonal group given by $\{ A \in R^{d \times d} | \det(A) = 1, A^T A = I_d \}$ or one of its subsets. (We call it rigid, since it is the same rotation applied to all the data points.) Although it is used commonly, it is very limited in capturing relationships between variables. In fact, it is akin to using a constant addition $y = (x + a) + e$ in the classical Euclidean setup. The cartoon plots in the top row of Figure 1 explain this problem more clearly. In this figure the predictor points $x_i$’s are shown as red crosses and the response points $y_i$’s as shown as yellow squares. In panel (a), where all the yellow points seem to be displaced similarly from the red points, one can use the rotation model with good success. However, in case the response points are dispersed in all directions around the predictor points, as in panel (b), the rotation model will not perform well and will need a richer class of transforms.

The bottom row in Figure 1 shows a few examples of similar situations in real applications. In panel (d) the Gulf Aden data (taken from Chang 1986) are plotted with the red and yellow points, respectively, denoting corresponding points on the Somali and Arabian plate. If we take the red points as predictors and the yellow points as responses, the rigid rotation can characterize this relationship almost perfectly. However, in panel (e) the coastal landmarks associated with the Mid-Atlantic Ridge are on both sides of the ridge. It is not possible to rotate all of the red coastal landmarks to their corresponding yellow ridge landmarks using a single rotation. In panel (f) we present an example from projective shape analysis (Goodall and Mardia 1999; Bryner, Klassen, and Srivastava 2012), where shapes are viewed as curves on a sphere and different projective linear transforms of these curves are deemed equivalent (shape-invariant transformation). These transformations arise due to the nature of light transport in near-field imaging of objects, relative to the camera aperture. That is, the deer shape seen in red and its transformation in yellow are considered equivalent and one is interested in capturing their relationship. Again, it seems difficult to fit the red points to yellow points via a rigid rotation. One needs a larger class of transformations, beyond rotations, to capture this relationship.

A family of transformations, which is bigger than and contains the set of rigid rotation, is the Möbius group, used in Downs (2003) for $d = 3$. This is a six-dimensional group that contains the three-dimensional rotation group as a subgroup and, hence, can capture more diverse relationships than simple rotations. However, this transformation group has not been extended to higher-dimensional spheres in the context of a regression problem. Another possibility is to first flatten the sphere and then apply standard (Euclidean) regression models in the resulting vector space. A flattening is basically a smooth, invertible mapping to a vector space and can be performed, for example, using the inverse exponential mapping or a stereographic projection as shown in Figure 1(c). Although this method can provide general solutions, a potential problem here is that a global flattening of a sphere is a highly nonlinear map and can potentially distort relationships between predictor and response variables. As the figure shows, a predictor-response pair may be close to each other on the sphere but may map to distant points on the flattened space. It seems more natural to develop an approach that uses the geometry of sphere and works directly on the sphere rather than flattening it.

**Noise Distribution:** A majority of past works assume that noise is sampled from a von Mises-Fisher density (VMF):

$$p(e | \mu, \kappa) = C_d(\kappa) \exp (\kappa e^T \mu), \quad C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2 \pi)^{d/2} I_{d/2-1}(\kappa)},$$

with $e, \mu \in S^{d-1}, \kappa > 0$. (1)
Here, $C_\theta$ is the normalizing constant, and $T_\nu$ is the modified Bessel function of the first kind and order $\nu$. The VMF density is symmetric about its mean direction $\mu$. The parameter $\kappa$ measures the concentration of points about $\mu$; $\kappa = 0$ implies a uniform density on $S^{d-1}$ while the limit $\kappa = \infty$ implies a Dirac delta at $\mu$. With this noise model, the conditional probability density of $y$, given $\mu(x)$ and a noise parameter $\kappa$, is given by: $C_\kappa(\kappa) = e^{-n(\kappa|x|)}$. For instance, in the limit $\kappa = 0$, then the variables $x$ and $y$ are independent, and $y$ is uniformly distributed on $S^{d-1}$.

Another interesting point here is that the normalization constant $C_\theta$ does not depend on the mean $\mu(x)$; it depends only on the concentration parameter $\kappa$ which makes the estimation process much simpler. The estimation of mean $\mu$ and concentration $\kappa$ of a VMF from the observed data has been studied in Banerjee et al. (2005), Tanabe et al. (2007), Sra (2012) and other papers referenced therein.

If $\mu(x) = Ax$, where $A$ is a rigid rotation and $e$ follows VMF density, the estimation of $A \in SO(d)$ is given by MacKenzie (1957) and Stephens (1979) as the Procrustes rotation: let $C = E[x y^T] \in \mathbb{R}^{d \times d}$ and use the modified SVD of $C = U \Lambda V^T$ (where $U, V \in SO(d)$ and $\Lambda$ is diagonal with entries $\lambda_1, \ldots, \lambda_d$ satisfying $\lambda_1 \geq \cdots \geq \lambda_d$) to evaluate the optimal rotation as $A = UV^T$. Several papers have studied the asymptotic distribution and other properties of this estimate, including Chang (1986); Shin, Takahara, and Murdoch (2001). The main result on consistency and asymptotic normality come from Chang (1986), albeit using a nonparametric noise model, which showed that $\hat{A}$ is a strongly consistent estimator of $A$, and the inverse-exponential transformation of $A^T \hat{A}$ is asymptotically multivariate normal. This result is based on a condition that $\frac{1}{n} \sum_{i=1}^n x_i x_i^T$ converges to a full rank matrix $\Sigma$. In Shin, Takahara, and Murdoch (2001), it is pointed out that if the rank of $\Sigma$ is at least $d - 1$, then this is in fact a necessary and sufficient condition for strong consistency.

1.2 Our Approach

Our goal in this article is to generalize the transformations between variables to go beyond rigid rotations. One can imagine a number of applications where the relationship between the dependent and the independent variables is too complex to be handled by rotation models. Of course, the most general situation will be to use nonparametric ideas, that is, develop a fully functional map $\mu: S^{d-1} \rightarrow S^{d-1}$ to define the conditional mean direction, by considering the set of all diffeomorphisms of $S^{d-1}$ to itself, but that transformation space is complicated due to infinite dimensionality and is left for future research. In this article we seek a balance between the ability to capture complex relationships and computational efficiency by restricting to a parametric family that is larger than the rotation group. There are several possibilities: the projective linear transformation group, the extension of Möbius group to higher-dimensional sphere, and so on. In this article we will focus on the projective general linear (PGL) group and study its use in defining a parametric class of conditional mean functions for modeling regression between variables on a unit sphere. We will continue using VMF density (Equation (1)) to model the noise perturbations and to specify the likelihood function for parameter estimation. The main contribution of this article is to develop a Newton–Raphson algorithm on an appropriate manifold for parameter estimation and to establish the asymptotic properties of this estimator. In particular, we provide a large sample-size asymptotic distribution of this estimator using the geometry of the underlying representation space introduced later. To illustrate this framework, we provide a number of examples using both simulated and real data, and demonstrate the superiority of this framework over current ideas in spherical regression.

The rest of this article is organized as follows. We describe the use of the projective general linear group in modeling spherical regression and the Newton–Raphson algorithm for estimation of regression parameters in Section 2. The asymptotic analysis of these estimators is performed in Section 3, while experimental evaluations of this model are presented in Section 4. All proofs and additional details of derivations for the Hessian and gradient are provided in a separate online supplementary document. We close the article with a brief discussion in Section 5.

2. PROPOSED REGRESSION FRAMEWORK

The central part of our approach is to use the projective linear transformation, also called PGL transform, for capturing the mean transformation from $x \in S^{d-1}$ to $y \in S^{d-1}$. In this section we introduce the PGL transform, use it in developing a regression model, and derive a maximum-likelihood estimator of regression parameters.

2.1 PGL($d$) Group and Regression Model

The projective linear transformation (PLT) is defined as follows. For any $x \in S^{d-1}$ and $A \in GL(d)$ (the set of $d \times d$ nonsingular matrices), define a transformation: $x \mapsto Ax$, where $|x|$ denotes the 2-norm of a vector. Since the result lies in $S^{d-1}$, this can be viewed as mapping from $S^{d-1}$ to itself. In fact, this mapping defines an action of the group $GL(d)$ on $S^{d-1}$ (please refer to Boothby 2003 for the definition of group action). However, the normalization makes the relation between elements of $GL(d)$ and the resulting mapping many-to-one. Specifically, two different elements of $GL(d)$ can result in the same mapping of $S^{d-1}$ as long as they are scalar multiples of each other. To resolve this issue, define $\Sigma_d = \{A I_d|A \in \mathbb{R}_{+}\}$, where $I_d$ is the $d \times d$ identity matrix. Then, it is easy to see that for any $A \in \Sigma_d$, $(A, x) \equiv Ax = x$, that is, the action of $\Sigma$ is trivial. For any $A \in GL(d)$, we define its orbit $[A] = \{AB|B \in \Sigma_d\}$; all elements of an orbit provide exactly the same transformation of $S^{d-1}$ and, hence, are termed equivalent. Therefore, we will use the full orbits (equivalence classes), rather than individual elements of $GL(d)$, to parameterize projective linear transformation. The set of all such orbits forms a quotient space denoted by $GL(d)/\Sigma_d$. (Since $\Sigma_d$ is a normal subgroup of $GL(d)$, the quotient set is actually a group with the same group operation as $GL(d)$). It is often referred to as the projective general linear group $PGL(d) = GL(d)/\Sigma_d$.

Thus, the regression mapping is parameterized by the elements of $PGL(d)$. However, it is not convenient to work with set-valued parameters, especially for estimation and asymptotic analysis. We can simplify this treatment by selecting a canonical representative of each equivalence class for analysis. Since the variability in an orbit $[A]$ is only due to scaling of matrices, we can designate the unique matrix that has determinant one to be
a representative of \([A]\). The set of such representatives is given by the special linear group: \(SL(d) = \{ A \in \mathbb{R}^{d \times d} \mid \det(A) = 1 \} \subset GL(d)\). The dimension of \(SL(d)\) is \(m = d^2 - 1\) and it contains the rotation group \(SO(d)\). There is a one-to-one correspondence between \(SL(d)\) and \(PGL(d)\) according to \(\frac{A}{(\det(A))^{1/2}} \rightarrow [A]\).

Now that we have specified the transformation space, we can state the regression model. Let \(\{ x_i \mid i \neq j \} \) be an iid random sample from a nondegenerate distribution \(P_x\) on the unit sphere \(S^{d-1}\) \((n > d \geq 2)\). In the context of this article, “nondegenerate” means the following: With probability one, any iid random sample of sample size \(n\) has a subsample of size \(d\) in which any \(d\) elements are linearly independent. This is actually a weak condition since any positive density over \(S^{d-1}\) (e.g., a uniform density or a VMF density) will satisfy this condition.

As stated earlier, we assume that the noise in the observation process comes from a VMF density. Let \(y_i \mid x_i \sim \text{VMF}(\mu(x_i) = \frac{x_i}{\|x_i\|}, \kappa)\). The model parameters are: (1) the transformation matrix \(A\) and (2) the concentration \(\kappa\). So a central question in this approach is: Given independent observations \(\{ (x_i, y_i), i = 1, 2, \ldots, n\}\), how can we estimate \(A\) and \(\kappa\)? While the estimation of \(\kappa\) has been treated by several authors previously, we will focus on finding the MLE of \(A \in SL(d)\) and will show that it is indeed a consistent estimator of \(A\). The likelihood function can be written as

\[
\mathcal{L}(A, \kappa) = p(\{y_i\} \mid \{x_i\}, A, \kappa) = \prod_{i=1}^{n} p(y_i \mid x_i, A, \kappa)
\]

\[
= C_d(\kappa)^n \exp \left( \kappa \sum_{i=1}^{n} y_i^T A x_i - \frac{\|Ax_i\|^2}{\|x_i\|^2} \right).
\]

Thus, the MLE of \(A\) and \(\kappa\) are given by: \(\hat{A}_n, \hat{\kappa}_n = \arg\max_{A \in SL(d), \kappa \in \mathbb{R}_+} \mathcal{L}(A, \kappa)\). Setting \(f_n(A) = \frac{1}{n} \sum_{i=1}^{n} y_i^T A x_i\), we note that the likelihood depends on \(A\) only through \(f_n(A)\), which is free of \(\kappa\). Therefore, we can perform the estimation sequentially, first \(A\) and then \(\kappa\) according to

\[
\hat{A}_n = \arg\max_{A \in SL(d)} f_n(A), \quad \hat{\kappa}_n = \arg\max_{\kappa \in \mathbb{R}_+} \mathcal{L}(\hat{A}_n, \kappa).
\]

We develop these estimators in the next two sections.

2.2 Maximum-Likelihood Estimation of \(A\)

In the past papers where \(A\) is a rotation, its estimation is straightforward using SVD. The estimation on \(SL(d)\), however, does not lend to a similar analytical expression and we will use the Newton–Raphson algorithm for an iterative maximization of \(f_n\). This algorithm will be intrinsic to \(SL(d)\), that is, the iterative updates are restricted to \(SL(d)\), and, therefore, we start by introducing its geometry.

2.2.1 Geometry of the Special Linear Group. The set \(SL(d)\) is a subgroup of \(GL(d)\). The tangent space of \(SL(d)\) at the identity matrix is \(sl(d) = \{ V \in \mathbb{R}^{d \times d} \mid \text{tr}(V) = 0 \}\). At any arbitrary point \(A \in SL(d)\), the tangent space is given by \(T_A(SL(d)) = \{ AV \mid V \in sl(d) \}\). We will endow \(SL(d)\) with the standard Euclidean metric to make it a Riemannian manifold. That is, for any two \(d \times d\) matrices \(V_1, V_2\), the Riemannian metric is defined to be \(\langle V_1, V_2 \rangle = \text{tr}(V_1 V_2^T)\). The gradient of \(f_n\) at a point \(A\) will be defined as an element of \(T_A(SL(d))\), while its Hessian will form a linear operator on \(T_A(SL(d))\), respectively.

The Newton–Raphson iterative updates will be applied in the larger space \(GL(d)\) and the result will be projected back into \(SL(d)\) using \(B \mapsto B/(\det(B)^{1/2})\).

Let \(\tilde{E}_n\) denote the \(d \times d\) matrix with 1 in the \(i, j\)-location and zero everywhere else. Then, we can construct an orthogonal basis for \(T_{\tilde{E}_n}(SL(d))\) using \(\{ \tilde{E}_i \mid i \neq j \} \cup \{ \tilde{E}_i - \tilde{E}_{dd} \mid i = 1, \ldots, d - 1 \}\) under the chosen metric (see, e.g., Li et al. 2009). We will rename these elements as \(\{ E_1, \ldots, E_m \}\), and then using the Gram–Schmidt process (under the chosen metric). We will term this basis set \(E_A \equiv \{ E_1, A, \ldots, E_m, A \}\). Since the codimension of \(SL(d)\) inside \(GL(d)\) is one, the normal space of \(SL(d)\) is one-dimensional, spanned by \(N_A = (A^{-T} A^{-1})\), where for a matrix we use the Frobenius norm \(\|A\| = \sqrt{\langle A, A \rangle}\).

For the asymptotic analysis, we will also need a mapping (in fact, a diffeomorphism) between a neighborhood of some point say \(A \in SL(d)\) and a neighborhood of the origin in \(T_A(SL(d))\). This mapping, \(\psi_A : T_A(SL(d)) \rightarrow SL(d)\), should satisfy the following two properties:

1. For \(0 \in \mathbb{R}^{d \times d}\), \(\psi_A(0) = A\) for all \(A \in SL(d)\), and
2. The shooting vector \(\frac{d}{dt} \psi_A(tV)_{t=0} = V\) for all \(V \in T_A(SL(d))\).

Such a mapping is called a retraction, and the point \(A\) is called the focal point of the mapping. In this article we will use the retraction \(\psi_A(V) = A \expm(A^{-1}V)\) where \(\expm\) denotes the standard matrix exponential. It can be verified that this \(\psi_A\) satisfies the two properties listed above. The inverse of this retraction is \(\psi_A^{-1} : SL(d) \rightarrow T_A(SL(d))\) given by:

\[
\psi_A^{-1}(B) = A \logm(A^{-1} B),
\]

where \(\logm\) denotes the standard matrix logarithm. It can be shown that in a certain neighborhood of \(A\) in \(SL(d)\), \(\psi_A\) is uniquely defined and is a diffeomorphism.

2.2.2 Newton–Raphson (N–R) Procedure on \(SL(d)\). To establish Newton–Raphson algorithm for maximizing the function \(f_n\) on \(SL(d)\), we need two ingredients—the gradient and the Hessian of \(f_n : SL(d) \rightarrow \mathbb{R}\). These quantities need to be defined intrinsically on \(SL(d)\), as follows.

1. Gradient of \(f_n\): The gradient of \(f_n\) at a point \(A \in SL(d)\) is the tangential direction where the increase in the value of \(f_n\) is maximum. It can be obtained by taking the unconstrained gradient, that is, the gradient of \(f_n\) in \(\mathbb{R}^{d \times d}\), and then projecting it in \(T_A(SL(d))\). The projection is performed by subtracting the normal component of this unconstrained gradient. The unconstrained gradient of \(f_n\), when treated as a function from \(\mathbb{R}^{d \times d}\) to \(\mathbb{R}\), can be shown to be

\[
\nabla_A f_n = \left( \frac{1}{n} \sum_{i=1}^{n} y_i x_i^T (x_i^T A x_i)^{-1/2} - A x_i x_i^T y_i y_i^T (x_i^T A x_i)^{-3/2} \right) \in \mathbb{R}^{d \times d}.
\]

Therefore, the intrinsic gradient of \(f_n\) on \(SL(d)\) is

\[
\delta_A f_n = \nabla_A f_n - \langle \nabla_A f_n, N_A \rangle N_A \in T_A(SL(d)).
\]
The tangent $\delta f_n$ can be written via an $m$-vector of coefficients with respect to the basis $E_A$.

2. Hessian of $f_n$: The Hessian $H_A$ of a function $f_n$ at a point $A$ is a linear mapping from the tangent space $T_A(SL(d))$ to itself defined as follows. For any direction $V \in T_A(SL(d))$, $H_A(V)$ represents the directional derivative of $\delta f_n$ in the direction of $V$, expressed as an element of $T_A(SL(d))$. In other words, if we perturb $A$ in the direction of $V$, the Hessian measures the resulting changes in the gradient vector. Since the Hessian is a linear operator on a finite-dimensional space, it can also be expressed as a matrix with respect to an orthonormal basis of $T_A(SL(d))$.

Now we provide the details of this construction. Let $D_A(G)$ denote the directional derivative of $G$, for any smooth vector field $G$ in the larger space $\mathbb{R}^{d \times d}$, that is, $D_V(G) = \lim_{\epsilon \to 0} \frac{G(A + \epsilon V) - G(A)}{\epsilon}$. Since we can view $\delta f_n$ as a vector field on all $\mathbb{R}^{d \times d}$, we can compute its directional derivative, in the direction of $V$, using Equation (3)

$$D_V(\delta f_n) = D_V(\nabla A f_n) - \langle \delta f_n, N_A \rangle N_A - \langle \delta f_n, N_A \rangle D_V(N_A).$$

The Hessian $H_A$ applied to $V$ is then the projection of this directional derivative on the tangent space at $A$. Note that the middle two terms are orthogonal to the tangent space, and will become zero in the projection. Also, since $D_V(N_A) \in T_A(SL(d))$, the last term is already in the tangent space. So, only the first term needs to be projected. Consequently, the Hessian $H_A$ applied to $V$ is given by

$$H_A(V) = D_V(\nabla A f_n) - \langle \delta f_n, N_A \rangle N_A - \langle \delta f_n, N_A \rangle D_V(N_A).$$

For the function $f_n$, defined earlier and the normal vector $N_A$, some of these terms can be derived explicitly as

$$D_V(N_A) = -(A^T)^{-1}V(A^T)^{-1}/\|A^{-1}\| + (A^T)^{-1}[A^{-1}, A^{-1}V A^{-1}]/\|A^{-1}\|^3,$$

$$D_V(\nabla A f_n) = \frac{1}{n} \sum_{i=1}^{n} \left( -y_i x_i^T A^T V x_i + V x_i x_i^T A^T y_i x_i^T + A x_i x_i^T V y_i x_i^T \right)/\|A x_i\|^3 + 3 A x_i x_i^T A^T y_i x_i^T (x_i^T A^T V x_i)/\|A x_i\|^5.$$

Now we can express the Hessian as an $m \times m$ matrix $W$ with respect to the chosen orthonormal basis of $T_A(SL(d))$. Recall that $m = d^2 - 1$ is the dimension of $SL(d)$. This matrix $W$ is computed as $W_{i,j} = \langle H_A(E_{i,A}), E_{j,A} \rangle$, $i, j = 1, 2, \ldots, m$. This $W$, in turn, can be used to evaluate the inverse of Hessian operator as needed in the Newton–Raphson method. Let $v = (v_1, \ldots, v_m)$ denote the coefficients of the intrinsic gradient $\delta f_n$ with respect to $E_A$. Then, define $u = W^{-1}v$ and form $U = \sum_{i=1}^{m} u_i E_{i,A}$ using $u = (u_1, \ldots, u_m)$. This $U$ represents the tangential direction obtained by applying the inverse of Hessian to the gradient direction and provides the direction of update in the Newton–Raphson algorithm.

3. Update on SL(d): To update the current estimate using $A \rightarrow \tilde{A}$, we used an additive update, followed by a projection:

$$\tilde{A} = \frac{B}{(\det(B))^{1/d}}, \text{ where } B = A + \sum_{i=1}^{m} u_i E_{i,A}. \tag{5}$$

Note that we can also use the retraction map $\psi_A$ instead, according to $\tilde{A} = \psi_A(\sum_{i=1}^{m} u_i E_{i,A})$. Since both the updates are identical up to the first order, the final results will be the same.

The full Newton–Raphson algorithm is as follows:

**Algorithm 1** Newton–Raphson Algorithm for Estimating $A$

1. Initialize $A$ using the rigid rotation $U_1 U_2^T$, where $D = U_1 \Phi U_2^T$ is the modified SVD of $D = \sum_{i=1}^{n} y_i x_i^T$ with $U_1, U_2 \in SO(d)$.

2. Compute extrinsic gradient $\nabla_A f_n$ using Equation (2) and the unit normal $N_A = (A^{-1})^T / \|A^{-1}\|$. Then, use Equation (3) to obtain the intrinsic gradient $\delta f_n$.

3. Construct an orthonormal basis $E_A$ for $T_A(SL(d))$ by applying the Gram–Schmidt procedure on the set $\{A E_i\}$.

4. Compute the Hessian matrix $W$ according to $W_{i,j} = \langle H_A(E_{i,A}), E_{j,A} \rangle$ for $E_{i,A}, E_{j,A} \in E_A$.

5. Express the intrinsic gradient using $v_i = \langle \delta f_n, E_{i,A} \rangle$ and compute $u = W^{-1}v$.

6. Compute the update direction as $U = \sum_{i=1}^{m} u_i E_{i,A}$ and perform the update according to Equation (5). If the algorithm has converged then stop, otherwise return to step 2.

2.3 Maximum-Likelihood Estimation of $\kappa$

The problem of estimating the concentration parameter $\kappa$ has been presented previously, including in Schou (1978) and Sra (2012). We will simply repeat these results here for convenience. First, consider the problem where the observations are $z_i \sim \text{id VMF}(\mu, \kappa)$, for $i = 1, \ldots, n$ and let $\|z\|$ be norm of their Euclidean mean, $\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$, considered as vectors in $\mathbb{R}^d$. It can be shown that the 2-norm of the expected value of $\bar{z}$, often called the resultant length, is given by $\rho(\kappa) = \frac{2 \sqrt{\kappa}}{\sqrt{2 - \kappa}}$. Thus, $\kappa$ can be estimated as $\hat{\kappa} = \rho^{-1}(\|\bar{z}\|)$. The inversion of $\rho$ is performed using the truncated Newton approximation (Sra 2012) with the iterations: $\hat{\kappa}_0 = \frac{\|\bar{z}\|}{\sqrt{2 - \|\bar{z}\|}}$, $\hat{\kappa}_{i+1} = \frac{\hat{\kappa}_i - \|\bar{z}\|}{1 - \rho(\hat{\kappa}_i)}$. Next, we return to our problem where $y_i \sim \text{VMF}(\mu(x_i), \kappa)$. To handle this situation, we use the following property.

**Proposition 1.** For any orthogonal matrix $O \in SO(d)$, if $y \sim \text{VMF}(\mu, \kappa)$ then $Oy \sim \text{VMF}(O\mu, \kappa)$.

This implies that a rigid rotation does not change the concentration $\kappa$ of a VMF density. Thus, for each $i = 1, \ldots, n$, we can rotate $y_i$ by a matrix $O_i \in SO(d)$ such that each $\mu(i) = A x_i / \|A x_i\|$ rotates to a fixed point $\mu$, say the north pole. Such a rotation matrix can be calculated using the Rodrigues’ formula (Kanatani 2012), $O_i = I_d + \sin(\theta_i) D_i + (\cos(\theta_i) - 1)(\mu \mu^T + c_i c_i^T)$, where $\theta_i = \cos^{-1}(\mu^T \mu), c_i = \frac{\mu^T \mu - 1}{\|\mu\|^2},$ and $D_i = \mu c_i c_i^T - c_i \mu^T$. The corresponding $y_i$ maps to a new point we will call $\bar{y}_i = O_i y_i \in S^{d-1}$. Now we can estimate $\kappa$ using these
rotated response variables since they are iid random vectors on $S^{d-1}$, as described above.

2.4 An Example of the Estimation Process

In this section we illustrate Algorithm 1 using some simulated examples. To simulate data we first generate $n = 25$ observations of $\mathbf{x}$ from VMF density with mean direction $\mathbf{u} = (1, 1, 1)^T / \sqrt{3}$ and a concentration $\kappa = 6$. Then, using a fixed but arbitrarily selected $A_0$ in $SL(d)$, we form the mean directions for each observation of the response variable as $\mu_i = \frac{A_0 \mu}{\sqrt{\kappa}}$. To proceed further, two cases were studied:

1. **Noiseless case.** Here $y_i = \mu_i$ for all $i$ and we need to estimate $A_0$.

2. **Noisy case:** Here we use the general model $y_i \sim VMF(\mu_i, \kappa)$ with $\kappa > 0$ and some $A_0 \in SL(d)$ to generate the data, and estimate $A_0$ from this data.

We applied the N–R algorithm for estimating $A_0$. The plots in Figure 2 illustrate the convergence of the algorithm. Figure 2(a) and (c) shows the evolution of $\hat{A}_n / |\hat{A}_n|$ during N–R iterations using white lines in the noiseless and noisy cases, respectively. These paths are initialized by a rigid rotation solution, and then iteratively updated until the final prediction (denoted by the blue open circle) is obtained. The $x_i$’s and $y_i$’s are shown using red crosses and yellow squares, respectively. Figure 2(b) and (d) shows the evolution of $f_n(A)$ during N–R iterations in these two cases. The horizontal lines denote the value $f_n(A_0)$, with value in the range $[-1, 1]$. In particular, in the noiseless case, $f_n(A_0) \equiv 1$.

3. STRONG CONSISTENCY AND ASYMPTOTIC NORMALITY

In this section we investigate the large sample-size asymptotics of the MLE of $A$. The setup is as follows. We have $n$ pairs of points $\{(x_i, y_i) \in S^{d-1} \times S^{d-1}, i = 1, 2, \ldots, n\}$, where $x_i$ follows a nondegenerate distribution and $y_i|\mathbf{x}_i \sim VMF(\mu_i, \kappa)$ for some $A_0 \in SL(d)$. Further, assume that $x_i$’s are independent. We have derived the MLE $\hat{A}_n$ by maximizing the function $f_n(A)$ using the Newton–Raphson method. Now we examine the asymptotic properties of $\hat{A}_n$ as $n \to \infty$. We will focus on two specific results: strong consistency (i.e., convergence with probability 1) and asymptotic normality, starting with the former.

3.1 Strong Consistency

First we will show that as $n \to \infty$, the MLE $\hat{A}_n$ converges to $A_0$ with probability 1. That is, $\hat{A}_n \xrightarrow{a.s.} A_0$ as $n \to \infty$. To simplify explanation, we will consider $\hat{A}_n$ as element of the Euclidean space $\mathbb{R}^{d \times d}$ and study the convergence using conventional tools in that vector space.

**Case I:** Noiseless Observations on $\{y_i\}$. To develop some understanding of our approach, we first consider a simpler situation where each $y_i = \frac{A_0 \mu_i}{\sqrt{\kappa}}$, that is, there is no observation noise. This is true when the concentration parameter $\kappa = \infty$. In this case, we can show a stronger result—$A_0$ is the unique maximizer of the likelihood.

**Theorem 1.** For $y_i = \frac{A_0 \mu_i}{\sqrt{\kappa}}$, and $x_i$’s as described above, $A_0$ is the unique maximizer of $f_n$ in $SL(d)$ with probability one.

**Case II:** Noisy Observations on $\{y_i\}$. Now we examine the general case when $y_i|\mathbf{x}_i \sim VMF(\frac{A_0 \mu_i}{\sqrt{\kappa}}, \kappa)$. Due to randomness in data, $A_0$ may no longer be the maximizer of $f_n(A)$ as can be seen in Figure 2(d). Note that $SL(d)$ is not a compact domain, and the MLE $\hat{A}_n$ may not exist in this general case. To address this issue, we limit $A$ to a compact subset $S_M = \{A \in \mathbb{R}^{d \times d} | \det A = 1, |A_{i,j}| \leq M, 1 \leq i, j \leq d\} \subset SL(d)$, where $M$ is a large positive constant (much greater than 1) such that each entry $|(A_0)_{i,j}| < M$ (this upper bound $M$ is used for theoretical purpose only, and can be ignored in practice). As $f_n(A)$ is a continuous function on a compact domain $S_M$, the MLE $\hat{A}_n$ must exist in $S_M$. The strong consistency in this case is stated as follows.

**Theorem 2.** In the setup described above, the maximum likelihood estimate $\hat{A}_n = \arg\max_{A \in S_M} f_n(A)$ is a strongly consistent estimator of $A_0$. That is, $\hat{A}_n \xrightarrow{a.s.} A_0$.

3.2 Asymptotic Normality

The next step is to establish the asymptotic normality of $\hat{A}_n$ and this uses the geometry of $S_M \subset SL(d)$. Since $SL(d)$ is a nonlinear space, the asymptotic analysis is better...
performed in the tangent space at $A_0$, a more convenient domain for defining normal distribution. The mapping from $SL(d)$ to $T_{A_0}(SL(d))$ is given by the inverse retraction map: $\hat{A}_n \mapsto \hat{V}_n = \psi_0^{-1}(\hat{A}_n) = A_0 \log m(A_0^{-1} \hat{A}_n)$. Note that $\psi_0^{-1}(A_0) = 0$. Furthermore, since $E_{\hat{A}_0}$ forms an orthonormal basis of $T_{A_0}(SL(d))$, we can express $\hat{V}_n$ in terms of coefficients with respect to this basis: $\hat{V}_n = \sum_{i=0}^{n-1} \alpha_{i,n} E_{i,\hat{A}_0}$. Thus, the vector of coefficients $\hat{a}_n = [\hat{a}_{1,n}, \ldots, \hat{a}_{m,n}]$ provides a representation of the estimator $\hat{A}_n$ in the Euclidean space $\mathbb{R}^m$.

We have already shown that as $n$ gets large $\hat{A}_n$ converges to $A_0$ (a.s.). Therefore, we can assume that, for a sufficiently small $\epsilon > 0$, $\hat{A}_n$ lies in an $\epsilon$-neighborhood of $A_0$ in $S_M$ with probability one; call it $B_\epsilon(A_0)$. There are several ways of defining this neighborhood but an easy option is: $B_\epsilon(A_0) = \{ A \in S_M | \| A_0^{-1} A - I_d \| \leq \epsilon \}$, with $\| \cdot \|$ denoting the matrix Frobenius norm. Since $B_\epsilon(A_0)$ is a compact set in $S_M \subset SL(d)$, its image under the inverse retraction map, given by $T_{A_0}^{-1} = \psi_0^{-1}(B_\epsilon(A_0))$, is compact in $T_{A_0}(SL(d))$. It follows from Theorem 2 that

$$A_0^{-1} \hat{A}_n \xrightarrow{a.s.} I_d \iff \hat{V}_n = \psi_0^{-1}(\hat{A}_n) \xrightarrow{a.s.} 0 \iff \hat{a}_n \xrightarrow{a.s.} 0.$$

That is, the consistency result for matrices $\hat{A}_n \in S_M$ is identical to that for coefficients $\hat{a}_n \in \mathbb{R}^m$. We will present asymptotic normality using the coefficients since the underlying space is $\mathbb{R}^m$. Define a forward mapping from $\mathbb{R}^m$ to $SL(d)$ by $F$ where $F(\alpha) = \psi_0(\sum_{i=1}^{m} \alpha_i E_{i,\hat{A}_0})$. In particular, we have $F(0) = A_0$. Recall that the likelihood function for a paired observation is $\psi \sim VMF(\mu, \kappa)$ where $\mu_0$ is a fixed vector in $\mathbb{R}^d$. In this model, the conditional mean direction is a constant map and all the variability in $y$ is attributed to noise. This constant can be estimated using $\hat{\mu}_y = \bar{y} / |\bar{y}| \in \mathbb{S}_d^{-1}$, where $\bar{y}$ is the sample mean of $y$’s when viewed as elements of $\mathbb{R}^d$. This, in general, is a poor model but can provide a baseline for comparisons.

2. Rigid Rotation (RR): Here $A$ is restricted to $SO(d)$. Since this model has been used extensively in the past for spherical regression, it is useful to include it in this study.

3. Log Linear Regression (LLR): This is a model where data are mapped to a tangent space on $\mathbb{S}_d^{-1}$ and one performs the standard linear regression in that vector space. Let $\bar{x} \in \mathbb{S}_d^{-1}$ be the point where the tangent space is located. For any point $x \in \mathbb{S}_d^{-1}$, its inverse exponential map at a point $\bar{x}$ is given by $x = \exp^{-1}(\bar{x}) \exp(\bar{x})$, where $\theta = \cos^{-1}(\bar{x}^T x)$. Using this expression, each point $y_i$ and $x_i$ are mapped to the same vector space $T_{\bar{x}}(\mathbb{S}_d^{-1})$ and one solves for a nonlinear regression model $(y_i = Bx_i + \epsilon)$ in that space. Once we have a predicted value in that tangent space, it can be mapped back to $\mathbb{S}_d^{-1}$ using: $x = \cos(\theta)x + \sin(\theta)v$, where $\theta = \|v\|$. As noted earlier, the main problem with this approach is that the relationship between $x_i$ and $y_i$ is potentially distorted in mapping them to a tangent space.

### 3.3 Confidence-Region Estimation

Using the asymptotic distribution of $\sqrt{n}\hat{a}_n$, we can go beyond point estimation of $A_0$ and can, in fact, provide a confidence region estimate as well. We will use the fact that when $\hat{A}_n$ is mapped to the tangent space $T_{A_0}(SL(d))$ using the inverse retraction $\psi_0$, and expressed with respect to an orthonormal basis, then $\sqrt{n}\hat{a}_n$ are asymptotically normal with mean zero and covariance $K = Q^{-1}$. Since the covariance $K$ is not known in general, we will estimate it using bootstrap as follows. Let $\hat{A}_n^b$ denote the $b$th bootstrap replicate of the estimator $\hat{A}_n$, obtained via a resampling of original data. The next step is to select a focal point to map these replicates to a tangent space using inverse retractions. The ideal choice is $A_0$ but that is unknown and being estimated. Instead, we will use $\hat{A}_n$ as a focal point and let $\hat{a}_n^b$ be the retraction from $T_{\hat{A}_n}(SL(d))$ to $SL(d)$. The expression for $\psi_0^{-1}$ is given earlier. Let $\{E_{i,\hat{A}_n}\}$ denote the orthonormal basis of $T_{\hat{A}_n}(SL(d))$ and let $\hat{a}_n^b = \text{trace}(\hat{V}_n \hat{a}_n^b) E_{i,\hat{A}_n}$, where $\hat{V}_n = \psi_0^{-1}(\hat{A}_n)$. In other words, $\hat{a}_n^b \in \mathbb{R}^m$ denotes the $b$th bootstrap replicate of $\hat{a}_n$. We can use the sample covariance matrix of $\{\hat{a}_n^b\}$, denoted by $K_n^b \in \mathbb{R}^{m \times m}$, to estimate $K / n$. To estimate the predictive performance of a model, we denote the leave-one-out cross-validated (LOOCV) mean-squared error simply by $\text{MSE} = \frac{1}{n-1} \sum_{i=1}^{n} |y_i - \hat{y}_i|^2$, where $\hat{y}_i$ denotes the predicted value of $y_i$ using the model fitted without observation $i$. Note that since the data and predicted values are restricted to a unit sphere the MSE is between 0 and 4.

### 4.1 Cloud Formations

In this application, one is concerned with using cloud shape on a day to predict its shape after a one-day time separation. The cloud shapes are represented by manually-selected landmarks on their boundaries on cloud images. The landmarks for the first day form the predictors $x_i$’s and the landmarks for the second day form the response $y_i$’s. Naturally, we find and fix the correspondence between landmarks. The goal is to use $x_i$’s (cloud cover on the first day) to predict $y_i$’s (cloud cover on the second day). We can see that the projective transformation will be advantageous over log-linear models and rigid rotation models since the deformations of cloud formations are generally unstructured. The clouds can become elongated or shortened, and cannot be characterized using rigid rotation. Also, since the clouds cover vast distances, they often have significant curvature to them, which can make the log-linear...
models more susceptible to distortion. In this section, two examples are constructed using landmarks of cloud formations, using the blue marble earth map obtained from the NASA's Visible Earth project (the original cloud images were taken from http://xplanet.sourceforge.net/clouds.php). The predictive performances obtained under different models are given in Table 1. As one can see, in both examples the PLT model performs the best followed by RR, LLR, and FMM.

### 4.2 Mid-Atlantic Ridge

In this experiment, the objective is to characterize the relationship between points near the Mid-Atlantic Ridge and their corresponding coastal landmarks. The landmarks were selected using topological maps of the sea floor (see http://visibleearth.nasa.gov/view.php?id=73826 for more information about the topological maps). There are 70 paired data points selected from both sides of the ridge. The predictor samples $x_i$’s are points that are close to the ridge (denoted by yellow squares) (see Figure 1(e)), and the response samples $y_i$’s are the coastal landmarks (denoted by red crosses). For each location on the Mid-Atlantic Ridge, there are two measurements—one each to the east and west of the ridge—and similar corresponding coastal landmarks located on the eastern and western continents, respectively. Note that it is difficult to characterize this relationship using a rigid rotation because the predictor variable is more spread out than the response variable. Again we estimate the model performances for each model. The results, reported in Table 2, show that the PLT model performs the best.

### 4.3 Vector-Cardiogram

A vector-cardiogram measures the direction and magnitude of electrical forces that are generated by heart actions. The directional aspect of these vectors has important applications in the diagnoses of certain diseases. The dataset, which was used in Downs (2003), consists of vector-cardiogram data from 98 children ages 2–19. Each child in the dataset is measured using two leads systems, namely the Frank system and the Mcfee system. The actual direction of greatest magnitude of the QRS loop has been plotted on the sphere for each corresponding lead system from each child, and the data have been separated into two categories classified by gender (see Table 3). The objective of this experiment is to define a regression model between these two systems, that is, use the directional vector from the Frank system as the predictor $x$ and the directional vector from the Mcfee system as the response $y$.

The prediction performances of different models on this data are presented in Table 3 and, as shown, PLT outperforms the other models. In both cases, boys and girls, the model performances are ranked in descending order by PLT, RR, LLR, and FMM.

### 4.4 Projective Shape Classification Experiment

The final experiment deals with classification of shapes of 2D contours observed under arbitrary PLTs. The PLTs play

---

**Table 1. Prediction performance of paired geographical coordinates from cloud formation images**

<table>
<thead>
<tr>
<th>Date</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 4, 2012</td>
<td>$6.7 \times 10^{-3}$</td>
</tr>
<tr>
<td>September 5, 2012</td>
<td>$36.7 \times 10^{-3}$</td>
</tr>
<tr>
<td>August 29, 2012</td>
<td>$12.8 \times 10^{-3}$</td>
</tr>
<tr>
<td>August 30, 2012</td>
<td>$139.8 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

**Table 2. Prediction performance for Mid-Atlantic Ridge data.**

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLT</td>
<td>$2.3 \times 10^{-3}$</td>
</tr>
<tr>
<td>LLR</td>
<td>$5.3 \times 10^{-3}$</td>
</tr>
<tr>
<td>RR</td>
<td>$6.0 \times 10^{-3}$</td>
</tr>
<tr>
<td>FMM</td>
<td>$92.1 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

---
Table 3. Prediction performance for vector-cardiograms

<table>
<thead>
<tr>
<th></th>
<th>Boys age 2 to 19</th>
<th>Girls age 2 to 19</th>
</tr>
</thead>
<tbody>
<tr>
<td>v = 56</td>
<td>(1.8 \times 10^{-1})</td>
<td>(2.6 \times 10^{-3})</td>
</tr>
<tr>
<td>PLT</td>
<td>1.8 \times 10^{-1}</td>
<td>7.8 \times 10^{-2}</td>
</tr>
<tr>
<td>LLR</td>
<td>2.0 \times 10^{-1}</td>
<td>13.2 \times 10^{-2}</td>
</tr>
<tr>
<td>RR</td>
<td>1.8 \times 10^{-1}</td>
<td>7.9 \times 10^{-2}</td>
</tr>
<tr>
<td>FMM</td>
<td>2.6 \times 10^{-1}</td>
<td>19.4 \times 10^{-2}</td>
</tr>
</tbody>
</table>

Table 4. Classification rates at various levels of concentration for noisy affine distorted images

<table>
<thead>
<tr>
<th>(\kappa)</th>
<th>5</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLT</td>
<td>0.218</td>
<td>0.388</td>
<td>0.66</td>
<td>0.758</td>
<td>0.84</td>
<td>0.854</td>
</tr>
<tr>
<td>RR</td>
<td>0.272</td>
<td>0.344</td>
<td>0.392</td>
<td>0.434</td>
<td>0.432</td>
<td>0.436</td>
</tr>
<tr>
<td>LLR</td>
<td>0.18</td>
<td>0.268</td>
<td>0.326</td>
<td>0.346</td>
<td>0.378</td>
<td>0.402</td>
</tr>
</tbody>
</table>

NOTE: The red points denote the Frank system directions and the yellow points denote the corresponding Mcfee system directions.

Table 4. Classification rates at various levels of concentration for noisy affine distorted images

<table>
<thead>
<tr>
<th>(\kappa)</th>
<th>5</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLT</td>
<td>0.218</td>
<td>0.388</td>
<td>0.66</td>
<td>0.758</td>
<td>0.84</td>
<td>0.854</td>
</tr>
<tr>
<td>RR</td>
<td>0.272</td>
<td>0.344</td>
<td>0.392</td>
<td>0.434</td>
<td>0.432</td>
<td>0.436</td>
</tr>
<tr>
<td>LLR</td>
<td>0.18</td>
<td>0.268</td>
<td>0.326</td>
<td>0.346</td>
<td>0.378</td>
<td>0.402</td>
</tr>
</tbody>
</table>

NOTE: At each concentration level, the model with the highest classification rate is highlighted in bold.

an important role in projective shape analysis which is of particular interest in image-based object recognition (Mardia and Patrangenaru 2005; Kent and Mardia 2012; Bryner, Klassen, and Srivastava 2012). The goal is to use prelabeled training shapes to classify the noisy test shapes using a measure of similarity between them. Since PLTs are considered shape-preserving in this context, the points on shapes are naturally treated as points on \(S^2\). Thus, the PLT regression model and the resulting maximized likelihood is a natural “score” for classifying shapes. We will treat the points on a training shape as \(x_i\)’s and a test shape as \(y_i\)’s, and study the maximum likelihoods associated with each candidate training shape. The training shape class that provides the highest likelihood is assigned to the test shape.

The shape dataset used in this experiment consists of 65 classes of shapes, with 20 shapes in each class (this dataset is taken from http://visionlab.utexas.edu/shape_data.htm). The original shape data are given as points along 2D contours and we convert them into points on \(S^2\) using: \(\phi: \mathbb{R}^2 \mapsto S^2\), according to \(\phi(u) = (u_1, u_2)^T = \left(\frac{(u_1, u_2, 1)^T}{\sqrt{(u_1, u_2, 1)^T(u_1, u_2, 1)}}\right) = \mathbf{x}\). This map is in fact a one-to-one mapping between \(\mathbb{R}^2\) and the upper hemisphere \(S^2_+\). For \(\mathbf{x} = (x_1, x_2, x_3)^T \in S^2_+\) (with \(x_3 > 0\)), it has an inverse: \(\phi^{-1}(\mathbf{x}) = (x_1, x_2, x_3^3)^T = (\frac{x_1, x_2}{x_3})^T = \mathbf{u}\). Then we simulate test shapes by randomly choosing shapes from the data, applying random PLTs to them and using transformed shapes as mean points under the VMF model to generate observations. The random PLT for each realization is obtained as follows: generate coefficients \(\alpha\) from a multivariate normal distribution with mean \(\mathbf{0}\) and covariance \(\frac{1}{n} I_n\), construct a tangent vector at \(T_{E_i}(SL(d))\) using \(V = \sum_{i=1}^m \alpha_i E_i \cdot d\), and map this tangent from \(T_{E_i}(SL(d))\) back to SL(d) using the retraction map \(A_0 = \psi_{E_i}(V)\). This data-generation process is illustrated with a horse shape in Figure 3.

In each replication, we select one out of 1300 contours to generate the test shape and treat the remaining 1299 shapes as training. If the training shape that provides the largest likelihood is in the same class as the test shape, we term this classification successful. This process is repeated 500 times at various concentration \(\kappa\) levels to study the effect of noise. Three models—PLT, LLR, and RR—are evaluated using the classification performance. An example of maximum-likelihood estimation of transformations under different methods is shown in Figure 4. Table 4 summarizes the classification performance for different methods. At all noise levels except one, where the noise level is very high, PLT outperforms the other methods. In large noise situation, the PLT model appears to be over-fitting and RR provides a better performance.

Figure 3. Generation of a test shape. (a) The original shape. (b) Shape mapped to the sphere. (c) A PLT applied to the shape. (d) Noise introduced using VMF model.
5. CONCLUSION

The modeling of regression between spherical variables is an important problem and past works are mainly limited to rigid rotations. While Möbius transformations have been proposed as an extension to rotations, this extension is currently limited to $S^2$. In this article we have introduced PLTs, a family of transformations that is larger than and contains rigid rotations, to model conditional mean directions in the regression model. Keeping the VMF density for modeling the noise, we provide a Newton–Raphson procedure for maximum-likelihood estimation of the model parameters, and demonstrate asymptotic consistency and normality of these estimators. We demonstrate the superiority of our framework using a number of applications dealing with real data.

Further extensions of this framework include expanding to a nonparametric mapping to capture relationships between spherical variables. While the model will be much richer than a parametric family, including the PLT, this will come at an added cost in terms of computation and analysis complexity.

SUPPLEMENTARY MATERIALS

The supplementary material includes detailed derivations of the gradient and Hessian of the log-likelihood on $SL(d)$ for the Newton-Raphson algorithm and the mathematical proofs for the theorems presented in the article.

[Received November 2012. Revised December 2013.]

REFERENCES


