이학석사 학위논문

Robust Geodesic Regression

강건 측지선형회귀

2020년 2월

서울대학교 대학원

통계학과

신 하 영

Abstract

Ha-Young Shin
The Department of Statistics
The Graduate School
Seoul National University

This thesis studies the application of methods for robust regression to data on Riemannian manifolds. Geodesic regression is the generalization of linear regression to a setting where there is a manifold-valued dependent variable and one or more real-valued independent variables. The existing work on geodesic regression uses the sum-of-squared errors to find the solution, but as in the classical Euclidean case, the least-squares method is highly sensitive to outliers. In this paper, we use M-estimators, including the L_1, Huber's and Tukey's bisquare estimators, to perform robust geodesic regression, and describe how to calculate the tuning parameters for the latter two. We also show that, on certain Riemannian manifolds, all M-estimators are also maximum likelihood estimators. We apply these models to synthetic data.

Keywords: Geodesic regression, manifold statistics, M-estimators
Student Number: 2018-20517

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Much work has been done on generalizing classical statistical methods for Euclidean data to manifold-valued data. Examples include principal geodesic analysis ([1]), analogous to principal component analysis, and geodesic regression ([2]), analogous to linear regression.

Many useful types of non-linear data can be modeled as lying on manifolds. Directional data in 3-dimensional space can be modeled on S^2; rotations in $\mathbb{R}^3$ can be represented by unit quaternions lying on S^3. Diffusion in the brain can be modeled by orientation distribution functions which lie on S^∞, which is approximated by S^n for a high value of n. The space of symmetric, positive-definite matrices has many useful applications: In neuroimagery, diffusion tensor imaging data can be modeled as 3×3 SPD matrices ([3], [4]), and in computer vision, covariance matrices, which are SPD matrices, are used in appearance tracking ([5]). For shape analysis, 2-

dimensional shape data can be represented as points on the complex projective space ([2], [6]), and the medial manifolds, $M(n) = (\mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2)^n$, provide models for the shape of organs, such as the hippocampus ([1]).

Geodesic regression, which generalizes linear regression to manifolds, has been studied in recent years ([2], [3], [6]), including geodesic regression that models gross errors in the data ([4]). This thesis explores a new approach to robust geodesic regression that accounts for possible outliers by using M-estimators.

The rest of this thesis is organized as follows. Chapter 2 briefly covers the prerequisite knowledge of differential geometry and chapter 3 reviews some of the existing work on geodesic regression. Chapter 4 applies the theory of M-estimators to manifold-valued data and includes a result showing that M-estimators are equivalent to maximum likelihood estimators on certain manifolds. After presenting results on sythetic data in chapter 5, focusing on the least-squares, median, Huber's and Tukey's bisquare estimators, the paper concludes in chapter 6 with a summary and possible avenues for future research. An appendix explains the details of calculating the cutoff parameter for the Huber and bisquare estimators and provides the steps of the gradient descent algorithm used to find the M-estimators.

Chapter 2

Differential geometry preliminaries

Consider a smooth manifold M. For a point $p \in M$, the tangent space T_pM is the subspace consisting of all vectors tangent to M at p. The elements of the tangent bundle of M, TM, take the form $(p, v) \in M \times T_pM$, so TM is the disjoint union of the tangent spaces of M. A Riemannian manifold M is a smooth manifold with a *Riemannian metric*; that is, a family of inner products on the tangent spaces that smoothly vary with p. This metric can be used to measure lengths on M. A geodesic between two points on M is the shortest length curve on M connecting them (geodesics are straight lines in Euclidean space), and the *geodesic*, or *Riemannian*, *distance* between the two points is the length of this geodesic segment.

A geodesic γ is defined by its initial point, $p = \gamma(0) \in M$, and velocity,

$v = \gamma'(0) \in T_{\gamma(0)}M$. Then the exponential maps, $\text{Exp}_p : T_pM \to M$, are defined by $\text{Exp}_p(v) = \gamma(1)$, and the logarithm maps, Log_p, are the inverses of the exponential maps. The exponential and logarithm maps are analogous to vector addition and subtraction in $\mathbb{R}^d$. If q is in the domain of Log_p, then the Riemannian distance between p and q is defined by $d(p,q) = \|\text{Log}_p(q)\|$. In this paper we will write $\text{Exp}_p(v)$ and $\text{Log}_p(q)$ as $\text{Exp}(p,v)$ and $\text{Log}(p,q)$, respectively.

Take a differentiable curve $\gamma : [a,b] \to M$, not necessarily a geodesic, and a tangent vector $v \in T_{\gamma(a)}M$. The unique vector field X along γ that satisfies $X(a) = v$ and $\nabla_{\gamma'}X = 0$, where ∇ is the Levi-Civita connection, is called the *parallel transport* of v along γ.

Given a family of geodesics, $\{\gamma_s\}$, parametrized by and varying smoothly with respect to $s \in \mathbb{R}$, a Jacobi field is a vector field along the geodesic γ_0 and, intuitively, it describes how the geodesic varies at each point with respect to s:

$$J(t) = \frac{\partial \gamma_s(t)}{\partial t}\Big|_{s=0}. \tag{2.1}$$

Jacobi fields satisfy a second order differential equation called the Jacobi equation. Jacobi fields are important here because they can be used to calculate the derivative of the exponential map.

Chapter 3

Geodesic regression

Given a dependent variable y on a Riemannian manifold M and an independent variable $x \in \mathbb{R}$, the simple geodesic regression model as defined by Fletcher [2] is

$$y = \mathrm{Exp}(\mathrm{Exp}(p, xv), \epsilon). \tag{3.1}$$

where $p \in M, v \in T_pM$ and $\epsilon \in T_{\mathrm{Exp}(p,xv)}M$. Kim et al. [3] extended this model to include multiple independent variables $x^1, ..., x^d \in \mathbb{R}$:

$$y = \mathrm{Exp}(\mathrm{Exp}(p, \sum_{j=1}^{d} x^j v^j), \epsilon), \tag{3.2}$$

with $v^1, ..., v^d \in T_pM$ and ϵ in the tangent space at $\mathrm{Exp}(p, \sum_{j=1}^{d} x^j v^j)$. For convenience, let $V = (v^1, ...v^d)$ and $Vx := \sum_{j=1}^{d} x^j v^j$.

Now given N data points $(x_i, y_i) \in \mathbb{R}^d \times M$, we define the energy E by

$$E(p, v) = \sum_{i=1}^{N} \frac{1}{2} d(\text{Exp}(p, Vx_i), y_i)^2. \tag{3.3}$$

Then the least squares estimator $(\hat{p}, \hat{V}) \in M \times T_p M^d$ is

$$(\hat{p}, \hat{V}) = \underset{(p,V) \in M \times T_p M^d}{\text{argmin}} E(p, V). \tag{3.4}$$

Unlike in the Euclidean case, this estimator is generally found using a gradient descent algorithm as an analytical solution is typically not available. If we fix $V = 0$ in (3.4), the resulting $\hat{p}$ is the (sample) intrinsic, or Karcher, mean, and its corresponding energy is the (sample) Fréchet variance.

Differentiating E with respect to p and each v_j yields

$$\nabla_p E = -\sum_{i=1}^{N} d_p \text{Exp}(p, Vx_i)^\dagger \text{Log}(\hat{y}_i, y_i) \tag{3.5}$$

and

$$\nabla_{v_j} E = -\sum_{i=1}^{N} x_i^j d_{v_j} \text{Exp}(p, Vx_i)^\dagger \text{Log}(\hat{y}_i, y_i) \tag{3.6}$$

for $j = 1, ..., d$. Here † represents the adjoint operator and $\hat{y}_i = \text{Exp}(p, Vx)$. In the case of simple geodesic regression (i.e. $d = 1$) on a Riemannian symmetric space, these operators can be calculated explicitly using Jacobi fields, as in [2]. Generalizing this approach to calculate exact gradients in multiple regression models is non-trivial, but the gradients can be approx-

imated well by

$$\nabla_p E \approx -\sum_{i=1}^{N} \Gamma_{\hat{y}_i \to p} \mathrm{Log}(\hat{y}_i, y_i) \tag{3.7}$$

and

$$\nabla_{v_j} E \approx -\sum_{i=1}^{N} x_i^j \Gamma_{\hat{y}_i \to p} \mathrm{Log}(\hat{y}_i, y_i), \tag{3.8}$$

as described in [3], where $\Gamma_{\hat{y}_i \to p}$ parallel transports tangent vectors from $T_{\hat{y}_i} M$ to $T_p M$ along the connecting geodesic.

Chapter 4

M-estimators on Riemannian manifolds

Consider M-estimators in the case of a univariate dependent variable, i.e. $M = \mathbb{R}$. The standard linear regression model is

$$y = \beta_0 + \beta_1 x^1 + ... + \beta_d x^d \tag{4.1}$$

where β_0 and $\beta = (\beta_1, ..., \beta_d) \in \mathbb{R}^d$ take the roles of p and V respectively.

M-estimators have an associated loss, or objective, function $\rho : \mathbb{R} \to \mathbb{R}$, and its derivative $\psi : \mathbb{R} \to \mathbb{R}$. The objective function should satisfy several properties: 1. $\rho(0) = 0$ and 2. $\rho(t_2) \geq \rho(t_1)$ if $|t_2| > |t_1|$. Then letting

$x_i = (x_i^1, ..., x_i^d)$, the M-estimator would be

$$(\hat{\beta}_0, \hat{\beta}) = \underset{(\beta_0,\beta)}{\operatorname{argmin}} \sum_{i=1}^{N} \rho(y_i - \beta_0 - x_i^T \beta), \tag{4.2}$$

which can be found by solving

$$\sum_{i=1}^{N} x_i \psi(y_i - \beta_0 - x_i^T \beta) = 0. \tag{4.3}$$

Generalizing to the manifold setting, define the energy E_ρ with

$$E_\rho(p, V) = \sum_{i=1}^{N} \rho(d(\mathrm{Exp}(p, Vx_i), y_i)), \tag{4.4}$$

so that the M-estimator is

$$(\hat{p}_\rho, \hat{V}_\rho) = \underset{(p,V)\in M\times T_pM^d}{\operatorname{argmin}} E_\rho(p, V), \tag{4.5}$$

which is a solution to

$$\nabla_p E_\rho = -\sum_{i=1}^{N} d_p \mathrm{Exp}(p, Vx_i)^\dagger \left(\frac{\rho'(\|\mathrm{Log}(\hat{y}_i, y_i)\|)}{\|\mathrm{Log}(\hat{y}_i, y_i)\|} \mathrm{Log}(\hat{y}_i, y_i) \right) = 0 \tag{4.6}$$

and

$$\nabla_{v_j} E_\rho = -\sum_{i=1}^{N} x_i^j d_{v_j} \mathrm{Exp}(p, Vx_i)^\dagger \left(\frac{\rho'(\|\mathrm{Log}(\hat{y}_i, y_i)\|)}{\|\mathrm{Log}(\hat{y}_i, y_i)\|} \mathrm{Log}(\hat{y}_i, y_i) \right) = 0 \tag{4.7}$$

for $j = 1, ..., d$. As in the least-squares case, gradients can either be calculated exactly using Jacobi fields for simple regression, or be approximated using parallel transport for multiple regression:

$$\nabla_p E_\rho \approx -\sum_{i=1}^{N} \Gamma_{\hat{y}_i \to p} \left(\frac{\rho'(\|\mathrm{Log}(\hat{y}_i, y_i)\|}{\|\mathrm{Log}(\hat{y}_i, y_i)\|} \mathrm{Log}(\hat{y}_i, y_i) \right), \tag{4.8}$$

$$\nabla_{v_j} E_\rho \approx -\sum_{i=1}^{N} x_i^j \Gamma_{\hat{y}_i \to p} \left(\frac{\rho'(\|\mathrm{Log}(\hat{y}_i, y_i)\|}{\|\mathrm{Log}(\hat{y}_i, y_i)\|} \mathrm{Log}(\hat{y}_i, y_i) \right). \tag{4.9}$$

In this study we consider the L_1 estimator, Huber's estimator and Tukey's bisquare estimator as robust alternatives to the least squares estimator. The least-squares and L_1 estimators are defined by $\rho_{L_2}(t) = \frac{1}{2}t^2$ and $\rho_{L_1}(t) = |t|$, respectively. Huber's and Tukey's bisquare estimators are defined by the following:

$$\rho_H(t) = \begin{cases} \frac{1}{2}t^2 & t \leq c \\ c|t| - \frac{1}{2}c^2 & t > c, \end{cases} \tag{4.10}$$

$$\rho_B(t) = \begin{cases} \frac{c^2}{6}\left\{1 - \left[1 - (\frac{t}{c})^2\right]^3\right\} & t \leq c \\ \frac{c^2}{6} & t > c, \end{cases} \tag{4.11}$$

or

$$\rho'_H(t) = \begin{cases} t & t \leq c \\ t \cdot sign(c), & \end{cases} \tag{4.12}$$

$$\rho'_B(t) = \begin{cases} t\big[1 - (\frac{t}{c})^2\big]^2 & t \leq c \\ 0 & t > c, \end{cases} \tag{4.13}$$

The calculations involved in determining c in the above functions are very tedious and lengthy. For details, including how the tuning parameters for the Huber and bisquare estimators are incorporated into the gradient descent algorithm, refer to the appendix.

4.1 M-estimators on symmetric spaces

A symmetric space, also called a globally symmetric space, is a connected Riemannian manifold M such that for all $p \in M$, there exists an involutive isometry that fixes p and reverses the geodesics that pass through p. Here, an isometry is a diffeomorphism that preserves the Riemannian distance, and an involutive isometry is an isometry that is its own inverse. We will need the concept of the the diameter of a manifold M: $\text{diam}(M) = \sup_{p_1,p_2 \in M} d(p_1, p_2)$.

Important examples of symmetric spaces are the Euclidean spaces $\mathbb{R}^n$, hyperbolic spaces, the spaces of symmetric, positive-definite matrices and the cylinder $S^1 \times \mathbb{R}$. Examples of symmetric spaces with finite diameter include the spheres S^n, compact Lie groups and Kendall's 2-dimensional shape spaces Σ_2^k, which are equivalent to the complex projective spaces $\mathbb{C}P^{k-2}$.

M-estimators are a generalization of maximum likelihood estimators. For ordinary Euclidean data, some M-estimators, such as the least-squares,

L_1 and Huber's estimators, can be expressed as MLEs under a certain distribution for the errors (Gaussian in the case of least squares); others, including the bisquare estimator, cannot. However, on certain symmetric spaces, it can be shown that all M-estimators are also MLEs.

Proposition 1. *Let M be a symmetric space with finite diameter, $x_1, ..., x_N \in \mathbb{R}^d$ and $y_1, ..., y_N \in M$. Any M-estimator whose objective function satisfies $\rho(e) \geq \rho(0)$ is equivalent to the maximum likelihood estimator of the geodesic regression model with Y conditionally distributed by*

$$p(y|X = x) = f(y; \text{Exp}(p, Vx), b, \rho) \tag{4.14}$$

for any $b > 0$, where

$$f(y; \mu, b, \rho) = \frac{1}{C(\mu, b, \rho))} \exp\left(- \frac{\rho(d(\mu, y))}{b} \right), \tag{4.15}$$

with

$$C(\mu, b, \rho) = \int_M \exp\left(- \frac{\rho(d(\mu, y))}{b} \right) dy. \tag{4.16}$$

In (4.15), b acts as a scale parameter. For example, $\rho_{L_2}(x) = \frac{1}{2}x^2$ and $b = \sigma^2$ for the least-squares estimator, so the estimator is equivalent to the MLE of the geodesic regression model with Gaussian errors as defined by Fletcher (2013) (this is true even for symmetric spaces with infinite diameter).

Proof. Parts of this proof closely follow the proof of Theorem 4 in [2]. First note that the term in (4.16) is finite because M is bounded and $\rho(e) \geq \rho(0)$ for all $e \in \mathbb{R}$, which means that

$$C(\mu, b, \rho) \leq \int_M \exp\left(- \frac{\rho(0)}{b} \right) dy = \exp\left(- \frac{\rho(0)}{b} \right) A < \infty, \tag{4.17}$$

where A is the volume of M; A is finite because the diameter of M is finite. So the function in (4.15) is a well-defined density function.

The log-likelihood of the observations $\{(x_i, y_i)\}_{1,\dots N}$ under the distribution in (4.15) is

$$\sum_{i=1}^{N} \log\{C(\mathrm{Exp}(p, Vx_i), b, \rho)\} - \frac{1}{b} \sum_{i=1}^{N} \rho(d(\mathrm{Exp}(p, Vx_i), y_i)). \tag{4.18}$$

Because M is a symmetric space, it is also a homogeneous space, meaning that for any two points on the manifold, there exists an isometry which maps one to the other. Because the integral in (4.16) depends only on the distance from μ to y, it is invariant to isometries, so the expression is independent of μ. Therefore the first sum in (4.18) is constant with respect to p and V. Comparing the second sum in this equation to (4.5), we see that the parameters $(p, V) \in M \times T_p M^d$ that minimize $E_\rho(p, V)$ also maximize the log-likelihood. □

Another point to note is that the concept of the breakdown point is not

meaningful on compact manifolds as distances between points on the manifold are bounded from above, so observations cannot be moved arbitrarily far away.

Chapter 5

Experiments on S^k

The n-spheres, S^n are useful manifolds with many applications. Here we will test the efficacy of various M-estimators for simple geodesic regression on S^2 and multiple geodesic regression on S^3 using simulated data. The settings used in these simulation experiments are similar, but not identical, to the ones used in [2].

The exponential map for S^n is given by

$$\mathrm{Exp}(p, v) = \cos(\|v\|)p + \sin(\|v\|)\frac{v}{\|v\|} \tag{5.1}$$

for $p \in S^n, v \in T_pS^n$. For $p, q \in S^n, p \neq q$, the logarithmic map is given by

$$\mathrm{Log}(p, q) = \cos^{-1}\Big(\frac{|\langle p, q\rangle|}{\|p\|\|q\|}\Big)\frac{q - \langle p, q\rangle p}{\|q - \langle p, q\rangle p\|}. \tag{5.2}$$

For $p, q \in S^n, p \neq q$, parallel transport of a vector $v \in T_pS^n$ along the

unique minimizing geodesic from p to q is given by

$$\Gamma_{p\to q}(v) = v^{\perp} + \|v^{\top}\|\Big[\cos(\|\mathrm{Log}(p,q)\|)\frac{\mathrm{Log}(p,q)}{\|\mathrm{Log}(p,q)\|} - \sin(\|\mathrm{Log}(p,q)\|)p\Big], \tag{5.3}$$

where

$$v^{\top} = \Big\langle v, \frac{\mathrm{Log}(p,q)}{\|\mathrm{Log}(p,q)\|}\Big\rangle \frac{\mathrm{Log}(p,q)}{\|\mathrm{Log}(p,q)\|},\ v^{\perp} = v - v^{\top} \tag{5.4}$$

the parts of v that are parallel and orthogonal to $\mathrm{Log}(p,q)$, respectively. In the simple regression case, the exact gradients with respect to p and v, calculated using Jacobi fields, are

$$\begin{aligned}\nabla_p E_\rho &= -\sum_{i=1}^{N} d_p \mathrm{Exp}(p, Vx_i)^{\dagger}\left(\frac{\rho'(\|\epsilon_i\|)}{\|\epsilon_i\|}\epsilon_i\right) \\ &= -\sum_{i=1}^{N}\left(\frac{\rho'(\|\epsilon_i\|)}{\|\epsilon_i\|}\right)\left(\cos(\|x_i v\|)\Gamma_{\hat{y}_i\to p}(\epsilon_i)^{\perp} + \Gamma_{\hat{y}_i\to p}(\epsilon_i)^{\top}\right),\end{aligned} \tag{5.5}$$

and

$$\begin{aligned}\nabla_v E_\rho &= -\sum_{i=1}^{N} x_i d_v \mathrm{Exp}(p, Vx_i)^{\dagger}\left(\frac{\rho'(\|\epsilon_i\|)}{\|\epsilon_i\|}\epsilon_i\right) \\ &= -\sum_{i=1}^{N} x_i\left(\frac{\rho'(\|\epsilon_i\|)}{\|\epsilon_i\|}\right)\left(\frac{\sin(\|x_i v\|)}{\|x_i v\|}\Gamma_{\hat{y}_i\to p}(\epsilon_i)^{\perp} + \Gamma_{\hat{y}_i\to p}(\epsilon_i)^{\top}\right).\end{aligned} \tag{5.6}$$

where $\epsilon_i = \mathrm{Log}(\hat{y}_i, y_i)$ and $\Gamma_{\hat{y}_i\to p}(\epsilon_i)^{\perp}, \Gamma_{\hat{y}_i\to p}(\epsilon_i)^{\top}$ are defined similarly to $v^{\perp}, v^{\top}$ in (5.4). For more information on how to calculate the Jacobi fields and use them to derive the exact gradients of exponential maps, see [2].

For both S^2 and S^3, we conducted the experiments in two stages. In

the first stage, we generated training data with a single value of p and V using Gaussian errors and estimated MSE($\hat{p}$), and MSE($\hat{v}_j$), $j = 1, ..., d$, for different sample sizes. The MSEs for the parameters are also their variances assuming consistency. The x_i came from the uniform distribution on $[0, 1]$, and the errors were generated from the isotropic multivariate normal distribution in the tangent space, with $\sigma^2 = (\pi/8)^2$. If the magnitude of an error was greater than π, which means that the generated point is beyond the cut locus, the data was resampled; however this could be ignored as the likelihood of such an event is negligible for small σ. For each k, the experiment was repeated M times. Then the MSEs for the parameters were calculated by

$$\text{MSE}(\hat{p}) \approx \frac{1}{M} \sum_{l=1}^{M} d(\hat{p}_l, p)^2, \tag{5.7}$$

$$\text{MSE}(\hat{v}_j) \approx \frac{1}{M} \sum_{l=1}^{M} \|\Gamma_{\hat{p}_l \to p}(\hat{v}_{j,l}) - v\|^2, \tag{5.8}$$

where $\hat{p}_l, \hat{v}_{j,l}$ are the estimates for p and v_j from the lth trial.

In the second stage, we tested the resistance to outliers by generating 2^6 samples, of which $\widetilde{N} = 2^{\widetilde{k}}$, $\widetilde{k} = 0, ..., 4$, came from an alternative $\widetilde{p}$ and $\widetilde{V}$, while the rest came from the same p, V as in the first stage. The errors were again Gaussian, this time with $\sigma^2 = (\pi/12)^2$. We again calculated the MSEs for the parameters and the test error.

For the simple regression model on S^2, we used the exact gradients that incorporate Jacobi fields. The original $p = (1, 0, 0)$, $v_1 = (0, \pi/4, 0)$ and the alternative $\widetilde{p} = (-1, 0, 0)$, $\widetilde{v}_1 = (0, 0, 0)$. The sample sizes considered

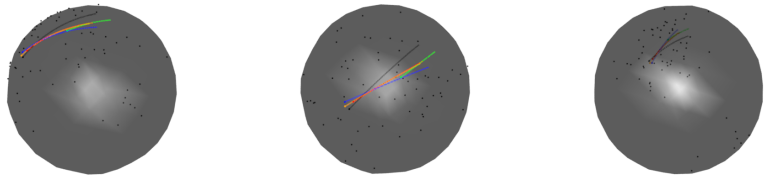

Figure 5.1: An example of a simulation used to test resistance to outliers in the simple regression case on S^2. These images show the results of the same simulation from 3 different angles. Each image shows 5 geodesics from $\gamma(0)$ to $\gamma(1)$: The black one is the true geodesic, the green one is the L_2 solution, the blue one is the L_1 solution, the red one is the Huber solution and the orange one is the bisquare solution. In the first and third images, we can clearly see the cluster of 8 points that were generated from the alternative $\widetilde{p}$ and $\widetilde{V}$ on the bottom right. In the second image, we can clearly see that that least-squares estimator is by far the worst performing of the four estimators by comparing them to the true geodesic.

were $N = 2^k$ for $k = 2, ..., 7$, and $M = 256$. In the S^3 case, we used the approximate gradients in (4.8) and (4.9). The original $p = (1, 0, 0, 0)$, $v_1 = (0, \pi/4, 0, 0)$, $v_2 = (0, 0, 0, -\pi/6)$ and the alternative $\widetilde{p} = (-1, 0, 0, 0)$, $\widetilde{v}_1 = \widetilde{v}_2 = (0, 0, 0, 0)$. The sample sizes considered were $N = 2^k$ for $k = 3, ..., 7$, and $M = 256$.

The results in Figure 5.2 are as expected. The least-squares estimator performs the best for the uncontaminated data, but the Huber and bisquare estimators are almost almost as good as their tuning parameters have been calculated to have an asymptotic efficiency of 95% relative to the least-squares esimator. The L_1 estimator performs somewhat worse. For the contaminated data, the bisquare estimator performs the best and the least-

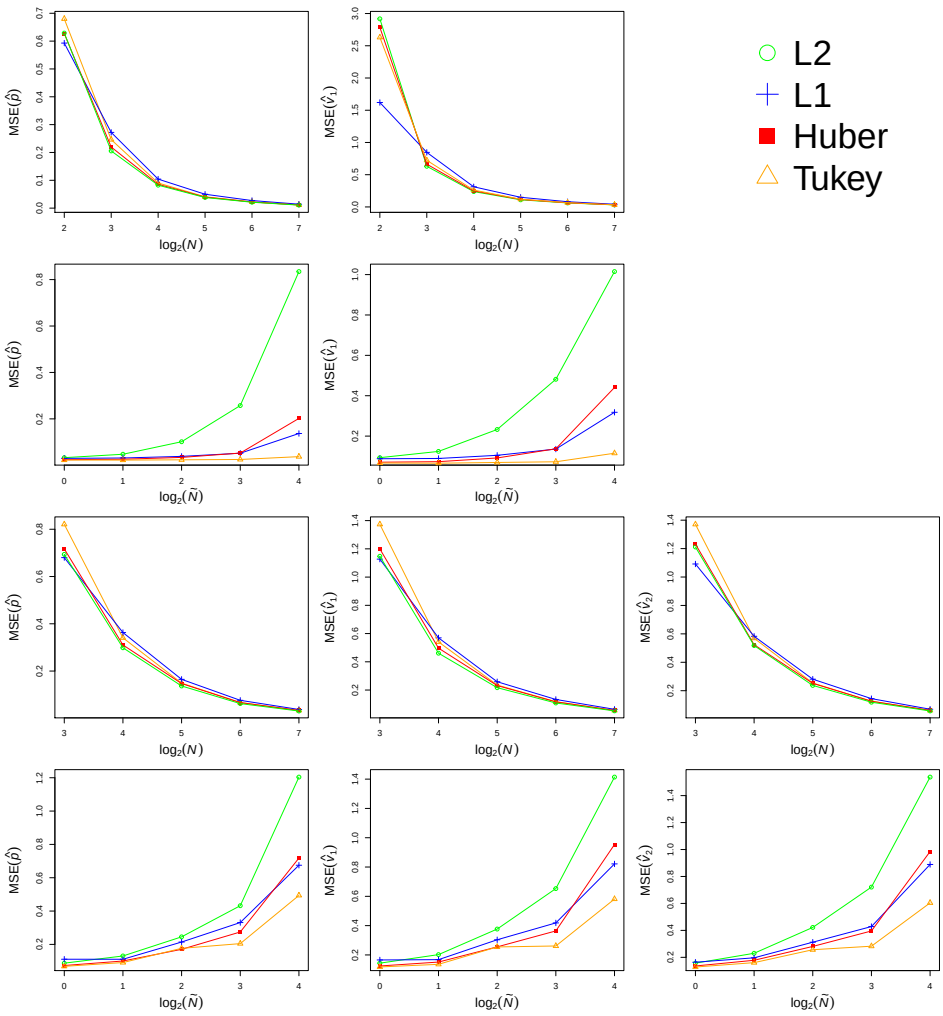

Figure 5.2: Various MSEs estimated from synthetic data. The first two rows show the results on S^2; the second two rows show the results on S^3. The first and third rows show the effect of sample size, N, when there are no outliers; the second and fourth rows show the effect of the number of outliers, $\widetilde{N}$, with sample size fixed at 64.

squares estimator the worst. This disparity increases as the size of the contaminated sample increases.

Chapter 6

Conclusion

In this paper, we have developed methods for robust geodesic regression that can withstand the influence of outliers. These methods adapted M-estimators, including the L_1, Huber and bisquare estimators, to a manifold setting; for the Huber and bisquare estimators, we develop in the appendix a method, using tangent space approximations, for calculating the tuning parameters that ensures efficiency in the case of Gaussian errors while also providing protection against outliers. Finally, these methods were tested on synthetic data. These experiments showed Tukey's bisquare estimator to be particularly effective.

A potentially fruitful avenue for future research is quantile regression on Riemannian manifolds, which would require developing the notion of quantiles for manifold-valued data. One could also explore pseudo-quantiles, such as expectiles and M-quantiles, on manifolds.

Appendix A

A.1 Introduction

For univariate Euclidean data, the tuning parameters for the Huber and bisquare estimators are typically chosen to be $1.345\hat{\sigma}$ and $4.685\hat{\sigma}$, where $\hat{\sigma} = MAD/0.6745$. $MAD = \text{Median}(|e_1|, ..., |e_N|)$ is the median absolute value, where $e_i = y_i - \hat{y}_i$. 0.6745 is chosen because, for $X \sim N(\mu, \sigma^2)$, $Pr(|X - \mu| < 0.6745\sigma) = 1/2$, and the values of 1.345 and 4.685 are chosen so that, given i.i.d $X_i \sim N(\mu, \sigma^2), i = 1, ..., N$, the asymptotic relative efficiency of the sample M-estimator for μ, $\hat{X}$, to the least-squares estimator, the sample mean $\bar{X}$, is 95% (i.e., $\lim_{N\to\infty}[\text{Var}(\bar{X})/\text{Var}(\hat{X})] = 0.95$). By analogy, determining c for the Huber and bisquare estimators on a symmetric space also requires two steps: (1) robustly estimating σ by MAD/f, and (2) finding the multiple of σ that would give an ARE to the sample mean of 95% under a Gaussian distribution.

Recall that the Gaussian distribution on a k-dimensional symmetric

space M has density

$$f(y;\mu,\sigma^2) = \frac{1}{C(\mu,\sigma^2)}\exp\left(-\frac{d(y,\mu)^2}{2\sigma^2}\right), \tag{A.1}$$

where

$$C(\mu,\sigma^2) = \int_M \exp\left(-\frac{d(y,\mu)^2}{2\sigma^2}\right)dy, \tag{A.2}$$

and $C(\mu,\sigma^2)$ is invariant with respect to μ because M is symmetric. Given i.i.d $Y_i, i = 1,...,N$, distributed according to (A.1), we will approximate the M-estimator $\hat{Y}$ for these points on the manifold by $\mathrm{Exp}(\mu,\hat{Y}^*)$, where $\hat{Y}^*$ is the M-estimator for the points $Y_i^* \equiv \mathrm{Log}(\mu,Y_i)$ in the tangent space at μ. As the tangent space is isomorphic to $\mathbb{R}^k$, we will treat these points as belonging to $\mathbb{R}^k$ and consider the Y_i to be distributed according to a isotropic multivariate Gaussian distribution with mean 0 and variance $\sigma^2 I_k$, truncated by the pre-image of M under the logarithmic map at μ, which we will call M^*. That is, after letting $\sigma = 1$ without loss of generality, the density of Y_i^* is given by

$$f(y) = \frac{\phi_k(y)}{Pr(Z_k \in M^*)} \tag{A.3}$$

for $y \in M^* \subset \mathbb{R}^k$, where $Z_k \sim N(0, I_k)$ is a standard k-dimensional Gaussian random variable and $\phi_k = (2\pi)^{-\frac{k}{2}}\exp(-\sum_{j=1}^k (y^j)^2)$ is its density (here y^j denotes the jth coordinate of y, not the jth power of y).

As the values and equations in this appendix are calculated using a

tangent space approximation, when $M^* = \mathbb{R}^k$, they can also be used for robust Euclidean multivariate regression.

Table A.1 gives the values of f, and c_H and c_B, which are the multiples of σ for the Huber and bisquare estimators, respectively, for $k = 1, 2, 3, 4, 5, 6$ when $M^* = \mathbb{R}^k$. We also include the efficiency of the median estimator. These values can be used in other cases, such as when M^* is a k-ball, if σ is small. The coming sections will show how these values are calculated.

Table A.1: f, c_H and c_B for $k = 1, ..., 6$ when $M^* = \mathbb{R}^k$

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
f	0.67449	1.17741	1.53817	1.83213	2.08601	2.31260
c_H	1.34500	1.50114	1.62799	1.73107	1.81202	1.86934
c_B	4.68506	5.12299	5.49025	5.81032	6.09627	6.35622
L_1 efficiency	0.63662	0.78540	0.84883	0.88357	0.90541	0.92039

Note that the efficiency of the L_1 estimator increases with k; when $k = 10$, the efficiency is 0.95131. So in higher dimensions, Huber's estimator becomes unnecessary as the L_1 estimator is already efficient.

The rest of the appendix will make use of the beta function $B(x, y)$, the gamma function $\Gamma(a)$, the lower incomplete gamma function $\gamma(a, z)$, the upper incomplete gamma function $\Gamma(a, z)$, the lower regularized gamma function $P(a, z) = \gamma(a, z)/\Gamma(z)$, the upper regularized gamma function $Q(a, z) = \Gamma(a, z)/\Gamma(a)$, and the inverses of the two regularized gamma functions $P^{-1}(a, z)$ and $Q^{-1}(a, z)$.

A.2 Identities

Before we proceed, we will need four identities related to integrals. We will use the spherical coordinate substitution where $r^2 = \sum_{j=1}^{k}(y^j)^2$, $y^1 = r\sin(\theta_1)...\sin(\theta_{k-2})\sin(\theta_{k-1})$ and $y^j = r\sin(\theta_1)...\sin(\theta_{k-j})\cos(\theta_{k-j+1})$ for $j = 2, ..., k$, so that $dy = dy_1...dy_k = r^{k-1}\sin^{k-2}(\theta_1)...\sin(\theta_{k-2})d\theta_{k-1}...d\theta_1$. Take a function $g : \mathbb{R}^+ \to \mathbb{R}$. Letting $B_R \subset \mathbb{R}^k$ denote the k-ball centered at 0 of radius R,

$$\int_{B_R} g(r)\phi_k(y)dy$$
$$= \int_0^R \int_0^\pi ... \int_0^\pi \int_0^{2\pi} g(r)\frac{1}{(2\pi)^{\frac{k}{2}}}e^{-r^2}r^{k-1}\sin^{k-2}(\theta_1)...$$
$$...\sin(\theta_{k-2})d\theta_{k-1}...d\theta_1 dr$$
$$= \frac{1}{(2\pi)^{\frac{k}{2}}}\Big(\int_0^R g(r)r^{k-1}e^{-r^2}dr\Big)\Big(\int_0^\pi \sin^{k-2}(\theta_1)d\theta_1\Big)...$$
$$...\Big(\int_0^\pi \sin(\theta_{k-2})d\theta_{k-2}\Big)\Big(\int_0^{2\pi} d\theta_{k-1}\Big)$$
$$= \frac{1}{(2\pi)^{\frac{k}{2}}}\Big(\int_0^R g(r)r^{k-1}e^{-r^2}dr\Big)\Big(2\int_0^{\pi/2} \sin^{k-2}(\theta_1)d\theta_1\Big)...$$
$$...\Big(2\int_0^{\pi/2} \sin(\theta_{k-2})d\theta_{k-2}\Big)\Big(4\int_0^{\pi/2} d\theta_{k-1}\Big)$$
$$= \frac{1}{(2\pi)^{\frac{k}{2}}}\Big(\int_0^R g(r)r^{k-1}e^{-r^2}dr\Big)B\Big(\frac{k-1}{2},\frac{1}{2}\Big)...B\Big(\frac{2}{2},\frac{1}{2}\Big)\cdot 2B\Big(\frac{1}{2},\frac{1}{2}\Big)$$
$$= \frac{1}{(2\pi)^{\frac{k}{2}}}\Big(\int_0^R g(r)r^{k-1}e^{-r^2}dr\Big)\frac{\Gamma(\frac{k-1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{k}{2})}...\frac{\Gamma(\frac{2}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{3}{2})}\cdot 2\frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{2}{2})}$$

$$= \frac{1}{(2\pi)^{\frac{k}{2}}}\Big(\int_0^R g(r)r^{k-1}e^{-r^2}dr\Big)\frac{2\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2})}$$
$$= 2^{-\frac{k}{2}}\frac{2}{\Gamma(\frac{k}{2})}\Big(\int_0^R g(r)r^{k-1}e^{-r^2}dr\Big)$$
$$= 2^{-\frac{k}{2}}\cdot\frac{k}{\Gamma(\frac{k+2}{2})}\Big(\int_0^R g(r)r^{k-1}e^{-r^2}dr\Big), \tag{A.4}$$

where we have used $\Gamma(1/2) = \pi^{\frac{1}{2}}, \Gamma(1) = 1$ and $\Gamma(z+1) = z\Gamma(z)$. The next two identities are derived in similar fashion:

$$\int_{B_R} g(r)(y^1)^2\phi_k(y)dy$$
$$= \int_0^R\int_0^\pi ...\int_0^\pi\int_0^{2\pi} g(r)(r\sin(\theta_1)...\sin(\theta_{k-2})\sin(\theta_{k-1}))^2\frac{1}{(2\pi)^{\frac{k}{2}}}e^{-r^2}r^{k-1}$$
$$\sin^{k-2}(\theta_1)...\sin(\theta_{k-2})d\theta_{k-1}...d\theta_1 dr$$
$$= \int_0^R\int_0^\pi ...\int_0^\pi\int_0^{2\pi} g(r)\frac{1}{(2\pi)^{\frac{k}{2}}}e^{-r^2}r^{k+1}\sin^k(\theta_1)...$$
$$...\sin^2(\theta_{k-1})d\theta_{k-1}...d\theta_1 dr$$
$$= \frac{1}{(2\pi)^{\frac{k}{2}}}\Big(\int_0^R g(r)r^{k+1}e^{-r^2}dr\Big)\frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{k+2}{2})}...\frac{\Gamma(\frac{4}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{5}{2})}\cdot 2\frac{\Gamma(\frac{3}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{4}{2})}$$
$$= 2^{-\frac{k}{2}}\cdot\frac{1}{\Gamma(\frac{k+2}{2})}\Big(\int_0^R g(r)r^{k+1}e^{-r^2}dr\Big) \tag{A.5}$$

and

$$\int_{B_R} g(r)y^1y^2\phi_k(y)dy$$

$$
\begin{aligned}
&= \int_0^R \int_0^\pi ... \int_0^\pi \int_0^{2\pi} g(r)(r\sin(\theta_1)...\sin(\theta_{k-2})\sin(\theta_{k-1})) \\
&(r\sin(\theta_1)...\sin(\theta_{k-2})\cos(\theta_{k-1}))\frac{1}{(2\pi)^{\frac{k}{2}}}e^{-r^2}r^{k-1}\sin^{k-2}(\theta_1)... \\
&...\sin(\theta_{k-2})d\theta_{k-1}...d\theta_1 dr \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}}\Big(\int_0^R g(r)r^{k-1}e^{-r^2}dr\Big)\Big(\int_0^\pi \sin^k(\theta_1)d\theta_1\Big)... \\
&...\Big(\int_0^\pi \sin^3(\theta_{k-2})d\theta_{k-2}\Big)\Big(\int_0^{2\pi} \sin(\theta_{k-1})\cos(\theta_{k-1})d\theta_{k-1}\Big) \\
&= 0, \qquad \text{(A.6)}
\end{aligned}
$$

because the last factor in the penultimate line is zero. The final identity uses the substitution $r' = r^2/2$ and $dr = [(r')^{-\frac{1}{2}}/\sqrt{2}]dr'$:

$$
\begin{aligned}
\int_0^R r^m e^{-r^2}dr &= \int_0^{\frac{R^2}{2}} 2^{\frac{m-1}{2}}(r')^{\frac{m-1}{2}}e^{-r'}dr' \\
&= 2^{\frac{m-1}{2}} \cdot \gamma\Big(\frac{m+1}{2}, \frac{R^2}{2}\Big) \\
&= 2^{\frac{m-1}{2}} \cdot \Big[\Gamma\Big(\frac{m+1}{2}\Big) - \Gamma\Big(\frac{m+1}{2}, \frac{R^2}{2}\Big)\Big]. \qquad \text{(A.7)}
\end{aligned}
$$

A.3 Step 1

The first step uses the $MAD = \text{Median}(\|e_1\|, ..., \|e_N\|)$ to find a robust estimate of σ. In the manifold case, $e_i = \text{Log}(\text{Exp}(p, x_i v), y_i)$. For a random variable Y' distributed according to (A.3), the goal is to find a factor f such that $Pr(\|Y'\| < f) = 1/2$. Letting $g(r) = 1$ in (A.4) and $m = k - 1$ in

(A.7),

$$\begin{aligned}
Pr(\|Y'\| < f) &= \frac{1}{P(Z_k \in M')} \int_{B_f} \phi_k(y) dy \\
&= \frac{1}{Pr(Z_k \in M^*)} \cdot 2^{-\frac{k}{2}} \frac{k}{\Gamma(\frac{k+2}{2})} \Big(\int_0^f r^{k-1} e^{-r^2} dr \Big) \\
&= \frac{1}{Pr(Z_k \in M^*)} \cdot 2^{-\frac{k}{2}} \frac{k}{\Gamma(\frac{k+2}{2})} \cdot 2^{\frac{k-2}{2}} \cdot \gamma\Big(\frac{k}{2}, \frac{f^2}{2}\Big) \\
&= \frac{1}{Pr(Z_k \in M^*)} \cdot 2^{-1} \frac{2}{\Gamma(\frac{k}{2})} \cdot \gamma\Big(\frac{k}{2}, \frac{f^2}{2}\Big) \\
&= \frac{1}{Pr(Z_k \in M^*)} \cdot P\Big(\frac{k}{2}, \frac{f^2}{2}\Big) \\
&= \frac{1}{2}
\end{aligned}$$

meaning,

$$f = \sqrt{2P^{-1}\Big(\frac{k}{2}, \frac{1}{2} Pr(Z_k \in M^*)\Big)}. \tag{A.8}$$

In the case of $M = S^k$, a k-sphere of radius r_M, M^* will be a k-ball centred at the origin of radius $R_M = \pi r_M / \sigma$, and so by same argument as above, $Pr(Z_k \in M^*) = P(k/2, R_M^2/2)$ and

$$f = \sqrt{2P^{-1}\Big(\frac{k}{2}, \frac{1}{2} P\Big(\frac{k}{2}, \frac{R_M^2}{2}\Big)\Big)}. \tag{A.9}$$

For other manifolds, such as the manifold of symmetric, positive-definite matrices or Euclidean space, $M^* = \mathbb{R}^k$, so $Pr(Z_k \in M^*) = 1$ and

$$f = \sqrt{2P^{-1}\Big(\frac{k}{2}, \frac{1}{2}\Big)}. \tag{A.10}$$

When $M^* = B_{R_M}$ but σ is small (i.e., R_M is large), the result in (A.8) can be well approximated by (A.10). For example, in our synthetic experiments on S_k in chapter 5, $r_M = 1$ and $\sigma = \pi/8$, so $R_M = 8$, and (A.9) and (A.10) agree to 13 decimal places for both $d = 2$ and 3. Finally, $\hat{\sigma} = MAD/f$.

A.4 Step 2

The next step finds the multiple of σ that gives an ARE to the sample mean of 95% assuming a Gaussian distribution. This step require the four identities (A.4),(A.5),(A.6) and (A.7).

The variance of a manifold-valued random variable W with an intrinsic mean μ_W is defined as

$$V_W \equiv \mathrm{E}(d(\mu_W, W)^2) = \mathrm{E}(\|\mathrm{Log}(\mu_W, W)\|^2). \tag{A.11}$$

If $W^* \equiv \mathrm{Log}(\mu_W, W)$ has an isotropic Gaussian distribution in $\mathbb{R}^d$ i.e. its covariance $\Sigma_W = \sigma_W^2 I_k$ is a multiple of the identity matrix, then

$$\frac{1}{\sigma_W^2}\mathrm{E}(\|\mathrm{Log}(\mu_W, W)\|^2) = \mathrm{E}((W^*)^T\Sigma_W^{-1}W^*) = k \implies V_W = k\sigma_W^2. \tag{A.12}$$

as $(W^*)^T\Sigma_W^{-1}W^* \sim \chi^2(k)$. Then let Y_i and Y_i^*, $i = 1, ..., n$ be as described in the introduction to the appendix, denote their sample intrinsic mean as $\bar{Y}$ and sample M-estimator (whether it is the Huber or bisquare estimator) as $\hat{Y}$, and define $\bar{Y}^* = \text{Log}(\mu, \bar{Y})$ and $\hat{Y}^* = \text{Log}(\mu, \hat{Y})$. Assuming the latter two converge in distribution to $N(0, \sigma_1^2 I_k)$ and $N(0, \sigma_2^2 I_k)$, respectively,

$$ARE(\hat{Y}, \bar{Y}) = \frac{\sigma_1^2}{\sigma_2^2} \tag{A.13}$$

by (A.11) and (A.12).

The covariance matrix of a sample M-estimator can be found using influence functions. For an objective (loss) function $\rho : \mathbb{R} \to \mathbb{R}$, define $\|\rho\| : \mathbb{R}^d \to \mathbb{R}$ by $\|\rho\|(t) = \rho(\|y\|)$. Then define $\psi : \mathbb{R}^k \to \mathbb{R}^k$ by $\psi(y) = \nabla_y \|\rho\|(e)$. Note that this coincides with the definition of ψ as ρ' in the $k = 1$ case. If F is the distribution of e and $T(F)$, the statistical functional at F representing the M-estimator, is the solution to $E_F[\psi(y - T(F))]$, then the influence function at $y_0 \in \mathbb{R}^k$ is defined as

$$IF(y_0; T, F) = E(J_\psi(y - T(F)))^{-1}\psi(y_0 - T(F)), \tag{A.14}$$

where J_ψ denotes the Jacobian matrix of ψ. It is known by the central limit theorem that for the sample M-estimator, $T(\hat{F})$, $\sqrt{N}(T(\hat{F}) - T(F)) \Rightarrow N(0, \int IF(y; T, F)^2 dF(y))$. Since $T(F) = \mu = 0$ in our case, the covariance of the M-estimator is asymptotically given by

$$\Sigma_\psi = \frac{1}{N}(E(J_\psi(y))^{-1})^2 E[\psi(y)\psi(y)^T]. \tag{A.15}$$

A.4.1 Covariance of the sample mean

The covariance of the sample mean $\bar{Y}^* = (1/N)\sum_{i=1}^{N} Y_i^*$ is simply $(1/N)\text{Cov}(Y_1^*) = (1/N)\text{E}[(Y_1^*)(Y_1^*)^T]$. Assuming $M^* = B_{R_M}$ (as, for example, in the case where $M = S^d$), where R_M is as defined in Step 1, and letting f be as defined in (A.3), B_R be as defined in (A.2) and B_R^c be its complement, and assuming $B_c \subset M^*$ (i.e. $c \leq R_M$),

$$\text{E}[(Y_1^*)(Y_1^*)^T]_{12} = \frac{1}{Pr(Z_k \in B_{R_M})} \int_{B_{R_M}} (y^1)(y^2)\phi_k(y)dy = 0 \tag{A.16}$$

by identity (A.5). On the other hand,

$$\begin{aligned}
\text{E}[(Y_1^*)(Y_1^*)^T]_{11} &= \frac{1}{Pr(Z_k \in B_{R_M})} \int_{B_{R_M}} (y^1)^2\phi_k(y)dy \\
&= \frac{1}{P\big(\frac{k}{2}, \frac{R_M^2}{2}\big)} \cdot 2^{-\frac{k}{2}} \cdot \frac{1}{\Gamma(\frac{k+2}{2})}\Big(\int_0^{R_M} r^{k+1}e^{-r^2}dr\Big) \\
&= \frac{1}{P\big(\frac{k}{2}, \frac{R_M^2}{2}\big)} \cdot 2^{-\frac{k}{2}} \cdot \frac{1}{\Gamma(\frac{k+2}{2})} \cdot 2^{\frac{k}{2}} \cdot \gamma\Big(\frac{k+2}{2}, \frac{R_M^2}{2}\Big) \\
&= \frac{P\big(\frac{k+2}{2}, \frac{R_M^2}{2}\big)}{P\big(\frac{k}{2}, \frac{R_M^2}{2}\big)},
\end{aligned} \tag{A.17}$$

where we have again made use of the identities (A.4), (A.5) and (A.7) and $Pr(Z_k \in M^*) = P(d/2, R_M^2/2)$. By symmetry, $\text{E}[(Y_1^*)(Y_1^*)^T]_{jj} = \text{E}[(Y_1^*)(Y_1^*)^T]_{11}$ for $j = 1, ..., k$, and $\text{E}[(Y_1^*)(Y_1^*)^T]_{lj} = \text{E}[(Y_1^*)(Y_1^*)^T]_{12}$ for all $j, l = 1, ..., k$, $l \neq j$, so the covariance of the sample mean is a scalar

multiple of the identity matrix:

$$\frac{1}{N}\mathrm{Cov}(Y_1^*) = \frac{P\left(\frac{k+2}{2}, \frac{R_M^2}{2}\right)}{NP\left(\frac{k}{2}, \frac{R_M^2}{2}\right)} \cdot I_k. \tag{A.18}$$

Note that this is simply $(1/N)I_k$ when $M^* = \mathbb{R}^k$ ($\gamma(a,z) \to \gamma(a)$, so $P(a,z) \to 1$, as $z \to \infty$). For example, this is the case when M is the Euclidean space or the space of symmetric, positive-definite matrices.

A.4.2 Huber's estimator

It easy to show that in the case of the Huber estimator,

$$\psi_H(y) = \begin{cases} y & \|y\| \le c \\ c \cdot \frac{y}{\|y\|} & \|y\| > c \end{cases} \tag{A.19}$$

and

$$J_{\psi_H}(y) = \begin{cases} I_k & \|y\| \le c \\ c\left(\frac{1}{\|y\|} I_k - \frac{1}{\|y\|^3} yy^T\right) & \|y\| > c. \end{cases} \tag{A.20}$$

Consider the first matrix term in (A.15).

$$\begin{aligned} E(J_{\psi_H}(y))_{12} &= -\frac{c}{Pr(Z_k \in B_{R_M})} \int_{B_c^c \cap B_{R_M}} \frac{1}{\|y\|^3}(y^1)(y^2)\phi_k(y)dy \\ &= 0 \end{aligned} \tag{A.21}$$

and

$$
\begin{aligned}
E(J_{\psi_H}(y))_{11} = \frac{1}{Pr(Z_k \in B_{R_M})} \Bigg\{ & \int_{B_c} \phi_k(y) dy \\
& + c \int_{B_c^c \cap B_{R_M}} \frac{1}{\|y\|} \phi_k(y) dy \\
& - c \int_{B_c^c \cap B_{R_M}} \frac{1}{\|y\|^3} (y^1)^2 \phi_k(y) dy \Bigg\} \\
= \frac{1}{P\big(\frac{k}{2}, \frac{R_M^2}{2}\big)} \Bigg\{ & 2^{-\frac{k}{2}} \cdot \frac{k}{\Gamma\big(\frac{k+2}{2}\big)} \Big(\int_0^c r^{k-1} e^{-r^2} dr \Big) \\
& + c \cdot 2^{-\frac{k}{2}} \cdot \frac{k}{\Gamma\big(\frac{k+2}{2}\big)} \Big(\int_c^{R_M} \frac{1}{r} r^{k-1} e^{-r^2} dr \Big) \\
& - c \cdot 2^{-\frac{k}{2}} \cdot \frac{1}{\Gamma\big(\frac{k+2}{2}\big)} \Big(\int_c^{R_M} \frac{1}{r^3} r^{k-1} e^{-r^2} dr \Big) \Bigg\} \\
= \frac{1}{P\big(\frac{k}{2}, \frac{R_M^2}{2}\big)} \cdot \frac{1}{\Gamma\big(\frac{k+2}{2}\big)} \Bigg\{ & 2^{-\frac{k}{2}} \cdot k \cdot 2^{\frac{k-2}{2}} \cdot \gamma\Big(\frac{k}{2}, \frac{c^2}{2}\Big) \\
& + c \cdot 2^{-\frac{k}{2}} k \cdot 2^{\frac{k-3}{2}} \cdot \Big[\gamma\Big(\frac{k-1}{2}, \frac{R_M^2}{2}\Big) - \gamma\Big(\frac{k-1}{2}, \frac{c^2}{2}\Big) \Big] \\
& - c \cdot 2^{-\frac{k}{2}} \cdot 2^{\frac{k-3}{2}} \cdot \Big[\gamma\Big(\frac{k-1}{2}, \frac{R_M^2}{2}\Big) - \gamma\Big(\frac{k-1}{2}, \frac{c^2}{2}\Big) \Big] \Bigg\} \\
= \frac{1}{P\big(\frac{k}{2}, \frac{R_M^2}{2}\big)} \cdot \frac{1}{\Gamma\big(\frac{k+2}{2}\big)} \Bigg\{ & \frac{k}{2} \gamma\Big(\frac{k}{2}, \frac{c^2}{2}\Big) \\
& + 2^{-\frac{3}{2}} c(k-1) \Big[\Gamma\Big(\frac{k-1}{2}, \frac{c^2}{2}\Big) - \Gamma\Big(\frac{k-1}{2}, \frac{R_M^2}{2}\Big) \Big] \Bigg\}, \qquad \text{(A.22)}
\end{aligned}
$$

where we have made use of the same identities as in the previous subsection.

Also, by the same arguments from symmetry as in the previous subsection, the matrix is a scalar multiple of the identity matrix; namely, $E(J_{\psi_H}(y))$ is I_k multiplied by the result of (A.22).

Now consider the second matrix term in (A.15). The non-diagonal terms can again be shown to be zero using identity (A.6) and symmetry, and the diagonal terms can be shown to be equal by symmetry. Then with $\psi_H = (\psi_H^1, ..., \psi_H^k)$ in (A.19),

$$\begin{aligned}
E[\psi_H(y)\psi_H(y)^T]_{11} &= E[(\psi_H^1(y))^2] \\
&= \frac{1}{Pr(Z_k \in B_{R_M})}\Bigg\{ \int_{B_c} (y^1)^2 \phi_k(y) dy \\
&+ c^2 \int_{B_c^c \cap B_{R_M}} \frac{1}{\|y\|^2} (y^1)^2 \phi_k(y) dy \Bigg\} \\
&= \frac{1}{P\left(\frac{k}{2}, \frac{R_M^2}{2}\right)} \Bigg\{ 2^{-\frac{k}{2}} \cdot \frac{1}{\Gamma\left(\frac{k+2}{2}\right)} \Big(\int_0^c r^{k+1} e^{-r^2} dr \Big) \\
&+ c^2 \cdot 2^{-\frac{k}{2}} \cdot \frac{1}{\Gamma\left(\frac{k+2}{2}\right)} \Big(\int_c^{R_M} r^{k-1} e^{-r^2} dr \Big) \Bigg\} \\
&= \frac{1}{P\left(\frac{k}{2}, \frac{R_M^2}{2}\right)} \cdot \frac{1}{\Gamma\left(\frac{k+2}{2}\right)} \Bigg\{ \gamma\Big(\frac{k+2}{2}, \frac{c^2}{2}\Big) \\
&+ \frac{c^2}{2} \Big[\Gamma\Big(\frac{k}{2}, \frac{c^2}{2}\Big) - \Gamma\Big(\frac{k}{2}, \frac{R_M^2}{2}\Big) \Big] \Bigg\}, \qquad \text{(A.23)}
\end{aligned}$$

again using (A.5) and (A.7). Then the matrix $E[\psi_H(y)\psi_H(y)^T]$ is the above expression multipled by I_k, and the variance Σ_ψ in (A.15) can be calculated

using (A.22) and (A.23):

$$\Sigma_{\psi_H} = \frac{E[\psi_H(y)\psi_H(y)^T]_{11}}{N(E(J_{\psi_H}(y))_{11})^2} \cdot I_k. \tag{A.24}$$

Then from (A.13) and (A.18), the ARE to the sample mean is

$$\frac{P\big(\frac{k+2}{2}, \frac{R_M^2}{2}\big)\Big\{\frac{k}{2}\gamma\big(\frac{k}{2}, \frac{c^2}{2}\big) + 2^{-\frac{3}{2}}c(k-1)\Big[\Gamma\big(\frac{k-1}{2}, \frac{c^2}{2}\big) - \Gamma\big(\frac{k-1}{2}, \frac{R_M^2}{2}\big)\Big]\Big\}^2}{P\big(\frac{k}{2}, \frac{R_M^2}{2}\big)^2 \cdot \Gamma\big(\frac{k+2}{2}\big)\Big\{\gamma\big(\frac{k+2}{2}, \frac{c^2}{2}\big) + \frac{c^2}{2}\Big[\Gamma\big(\frac{k}{2}, \frac{c^2}{2}\big) - \Gamma\big(\frac{k}{2}, \frac{R_M^2}{2}\big)\Big]\Big\}}. \tag{A.25}$$

$P(a,z) \to 1$ and $\Gamma\big(a,z) \to 0$ as $z \to \infty$, so if $M^* = \mathbb{R}^k$, which can be conceptualized as letting $R_M \to \infty$, and the Gaussian distribution is not truncated, the ARE becomes

$$\frac{\Big\{\frac{k}{2}\gamma\big(\frac{k}{2}, \frac{c^2}{2}\big) + 2^{-\frac{3}{2}}c(k-1)\Gamma\big(\frac{k-1}{2}, \frac{c^2}{2}\big)\Big\}^2}{\Gamma\big(\frac{k+2}{2}\big)\Big\{\gamma\big(\frac{k+2}{2}, \frac{c^2}{2}\big) + \frac{c^2}{2}\Gamma\big(\frac{k}{2}, \frac{c^2}{2}\big)\Big\}}. \tag{A.26}$$

Finally, we can calculate the derivative of (A.25) or (A.26) with respect to c and use the Newton-Raphson method to find the value of c for which the ARE is 95%. If $M^* = B_{R_M}$ but σ is small (i.e., R_M is large), (A.25) can be well approximated by (A.26). In our synthetic experiments on S_k in chapter 5, $R_M = 8$, and the values of c derived from (A.25) and (A.26) agree to 10 decimal places for both $k = 2$ and 3.

By letting $c \to 0$, we can also use (A.25) and (A.26) to calculate the efficiency of the L_1 estimator. As noted in the introduction to the appendix, in higher dimensions, the L_1 estimator becomes so efficient that Huber's es-

timator is unnecessary; this happens at $k = 10$ in (A.26) for 95% efficiency, and depends on R_M in (A.25).

A.4.3 Bisquare estimator

For the bisquare estimator, it is easy to show that

$$\psi_B(y) = \begin{cases} \left[1 - \left(\frac{\|y\|}{c}\right)^2\right]^2 \cdot y & \|y\| \le c \\ 0 & \|y\| > c, \end{cases} \tag{A.27}$$

and

$$J_{\psi_B}(y) = \begin{cases} \left[1 - \left(\frac{\|y\|^2}{c^2}\right)^2\right]^2 I_k - \frac{4}{c^2}\left[1 - \left(\frac{\|y\|^2}{c^2}\right)^2\right] yy^T & \|y\| \le c \\ 0 & \|y\| > c, \end{cases} \tag{A.28}$$

By similar arguments to the ones used for Huber's estimator,

$$E(J_{\psi_B}(y))_{12} = 0, \tag{A.29}$$

$$\begin{aligned} E(J_{\psi_H}(y))_{11} = \frac{1}{P\left(\frac{k}{2}, \frac{R_M^2}{2}\right)} \cdot \frac{1}{\Gamma\left(\frac{k+2}{2}\right)} \Bigg\{ & \frac{2(k+4)}{c^4}\gamma\Big(\frac{k+4}{2}, \frac{c^2}{2}\Big) \\ & - \frac{2(k+2)}{c^2}\gamma\Big(\frac{k+2}{2}, \frac{c^2}{2}\Big) + \frac{k}{2}\gamma\Big(\frac{k}{2}, \frac{c^2}{2}\Big)\Bigg\}, \end{aligned} \tag{A.30}$$

$$E[\psi_H(y)\psi_H(y)^T]_{12} = 0, \tag{A.31}$$

$$E[\psi_H(y)\psi_H(y)^T]_{11} = \frac{1}{P\left(\frac{k}{2}, \frac{R_M^2}{2}\right)} \cdot \frac{1}{\Gamma\left(\frac{k+2}{2}\right)} \Bigg\{ \gamma\Big(\frac{k+2}{2}, \frac{c^2}{2}\Big) - \frac{8}{c^2}\gamma\Big(\frac{k+4}{2}, \frac{c^2}{2}\Big) + \frac{24}{c^4}\gamma\Big(\frac{k+6}{2}, \frac{c^2}{2}\Big) - \frac{32}{c^6}\gamma\Big(\frac{k+8}{2}, \frac{c^2}{2}\Big) + \frac{16}{c^8}\gamma\Big(\frac{k+10}{2}, \frac{c^2}{2}\Big) \Bigg\}, \tag{A.32}$$

and the variance Σ_ψ in (A.15) can be calculated using (A.30) and (A.32):

$$\Sigma_{\psi_B} = \frac{E[\psi_B(y)\psi_B(y)^T]_{11}}{N(E(J_{\psi_B}(y))_{11})^2} \cdot I_k. \tag{A.33}$$

Then from (A.13) and (A.18), the ARE to the sample mean is

$$\frac{P\left(\frac{k+2}{2}, \frac{R_M^2}{2}\right)\Big\{\frac{2(k+4)}{c^4}\gamma\Big(\frac{k+4}{2}, \frac{c^2}{2}\Big) - \frac{2(k+2)}{c^2}\gamma\Big(\frac{k+2}{2}, \frac{c^2}{2}\Big) + \frac{k}{2}\gamma\Big(\frac{k}{2}, \frac{c^2}{2}\Big)\Big\}^2}{P\left(\frac{k}{2}, \frac{R_M^2}{2}\right)^2 \cdot \Gamma\left(\frac{k+2}{2}\right)\Big\{\gamma\Big(\frac{k+2}{2}, \frac{c^2}{2}\Big) - \frac{8}{c^2}\gamma\Big(\frac{k+4}{2}, \frac{c^2}{2}\Big) + \frac{24}{c^4}\gamma\Big(\frac{k+6}{2}, \frac{c^2}{2}\Big) - \frac{32}{c^6}\gamma\Big(\frac{k+8}{2}, \frac{c^2}{2}\Big) + \frac{16}{c^8}\gamma\Big(\frac{k+10}{2}, \frac{c^2}{2}\Big)\Big\}}, \tag{A.34}$$

and the non-truncated version is

$$\frac{\Big\{\frac{2(k+4)}{c^4}\gamma\Big(\frac{k+4}{2}, \frac{c^2}{2}\Big) - \frac{2(k+2)}{c^2}\gamma\Big(\frac{k+2}{2}, \frac{c^2}{2}\Big) + \frac{k}{2}\gamma\Big(\frac{k}{2}, \frac{c^2}{2}\Big)\Big\}^2}{\Gamma\left(\frac{k+2}{2}\right)\Big\{\gamma\Big(\frac{k+2}{2}, \frac{c^2}{2}\Big) - \frac{8}{c^2}\gamma\Big(\frac{k+4}{2}, \frac{c^2}{2}\Big) + \frac{24}{c^4}\gamma\Big(\frac{k+6}{2}, \frac{c^2}{2}\Big) - \frac{32}{c^6}\gamma\Big(\frac{k+8}{2}, \frac{c^2}{2}\Big) + \frac{16}{c^8}\gamma\Big(\frac{k+10}{2}, \frac{c^2}{2}\Big)\Big\}}. \tag{A.35}$$

Then we can apply the Newton-Raphson method to find the value of c for which the ARE is 95%. Finally, (A.35) approximates (A.34) well when R_M is large: when $R_M = 8$ as in our synthetic experiments, the resulting values

of c for these two expressions agree to 8 decimal places.

A.5 Algorithm

The gradient desent algorithm to find the solution to the geodesic regression problem is presented on the next page. $\nabla_p E_\rho$ and $\nabla_V E_\rho$ in lines 15 and 16 are calculated exactly using Jacobi fields in the case of simple regression and approximately using parallel transport, as in (4.8) and (4.9), in the case of multiple regression; E_ρ is defined as in (4.4).

Note lines 20 and 25, which are marked as optional: For our simulated experiments on $M = S^k$ in Chapter 5, the values can be recalculated using $R_M = \pi/\hat{\sigma}$. As mentioned in previous sections, if R_M is sufficiently large, the difference between (A.25) and (A.26) is negligibile, as is the difference between (A.34) and (A.35). In this case, lines 20 and 25 can be ignored, and in our experiments in Chapter 5, we did so. Similarly, if $M^* = \mathbb{R}^k$, these values should not be recalculated.

Algorithm 1 Gradient descent algorithm for geodesic regression

Input: $x_1, ..., x_N \in \mathbb{R}^d$, $y_1, ..., y_N \in M$ for k-dim'l M and $\rho : \mathbb{R} \to \mathbb{R}^+$.
Output: $p \in M, V \in T_pM^d$
Initialize p, V, λ, λ_{max} and center x.
if $\rho = \rho_H$ or ρ_B **then**
    Calculate f using (A.10).
    for i in 1 to N **do**
        $e_i = \text{Log}(\text{Exp}(p, Vx_i), y_i)$
    end for
    $MAD = \text{Median}(\|e_1\|, ..., \|e_N\|)$
    Calculate c_H or c_B using (A.26) or (A.35).
    $\hat{\sigma} = MAD/f$
    $c = c_H\hat{\sigma}$ or $c = c_H\hat{\sigma}$.
end if
while termination condition **do**
    $p_{new} = \text{Exp}(p, -\lambda\nabla_p E_\rho)$
    $V_{new} = \Gamma_{p \to p_new}(V - \lambda\nabla_V E_\rho)$
    if $E_\rho(p, V) \geq E_\rho(p_{new}, V_{new})$ **then**
        $p = p_{new}$ and $V = V_{new}$
        if $\rho = \rho_H$ or ρ_B **then**
            Recalculate f using (A.9). ▷ Optional
            for i in 1 to N **do**
                $e_i = \text{Log}(\text{Exp}(p, Vx_i), y_i)$
            end for
            $MAD = \text{Median}(\|e_1\|, ..., \|e_N\|)$
            Recalculate c_H or c_B using (A.25) or (A.34). ▷ Optional
            $\hat{\sigma} = MAD/f$
            $c = c_H\hat{\sigma}$ or $c = c_H\hat{\sigma}$
        end if
        $\lambda = \min(\lambda_{max}, 2\lambda)$
    else
        $\lambda = \lambda/2$
    end if
end while

Bibliography

[1] P. T. Fletcher, C. Lu, S. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Transactions on Medical Imaging*, vol. 23, no. 8, p. 995–1005, 2004.

[2] P. T. Fletcher, "Geodesic regression and the theory of least squares on Riemannian manifolds," *International Journal of Computer Vision*, vol. 105, no. 2, pp. 171–185, 2012.

[3] H. J. Kim, N. Adluru, M. D. Collins, M. K. Chung, B. Bendin, S. C. Johnson, R. J. Davidson, and V. Singh, "Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[4] X. Zhang, X. Shi, Y. Sun, and L. Cheng, "Multivariate regression with gross errors on manifold-valued data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, p. 444–458, Jan 2019.

[5] G. Cheng and B. C. Vemuri, "A novel dynamic system in the space of SPD matrices with applications to appearance tracking," *SIAM Journal on Imaging Sciences*, vol. 6, no. 1, p. 592–615, 2013.

[6] E. Cornea and H. Zhu, "Regression models on Riemannian symmetric spaces," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 2, 2017.

국문초록

본 논문은 리만 다양체 위의 자료에 대한 강건 회귀분석 방법을 탐색한다. 측지선형회귀는 선형회귀의 일반화된 방법으로 종속변수가 다양체 값을 가지고 실수 값을 갖는 독립변수가 하나 이상 존재할 때 사용된다. 지금까지의 측지선형회귀에 대한 연구에선 해를 찾기 위해 오차제곱합을 사용해왔다. 하지만 이는 기존의 유클리드 공간에서의 경우와 마찬가지로 이상치에 아주 민감히 반응한다. 본 논문에서는 강건 측지선형회귀분석을 하기 위해 L_1, Huber, Tukey의 bisquare 추정량과 같은 M-추정량을 이용하고, Huber와 Tukey의 방법을 사용하는 경우 조율상수를 어떻게 얻을 수 있는지에 대해 기술한다. 또한 M-추정량은 특정 리만 다양체 위에서 최대가능도추정량이 됨을 보인다. 우리는 이 모형들을 합성자료에 대해 적용한다.

주요어: 측지선형 회귀, 다양체 통계, M-추정량
학 번: 2018-20517