NEW RESULTS IN DIMENSION REDUCTION AND MODEL SELECTION

A Dissertation
Presented to
The Academic Faculty

by

Andrew Korb Smith

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
April 2008

NEW RESULTS IN DIMENSION REDUCTION AND MODEL SELECTION

Approved by:

Professor Xiaoming Huo, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Hongyuan Zha
College of Computing
Georgia Institute of Technology

Professor Alexander Shapiro
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Ming Yuan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Nicoleta Serban
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: 21 March 2008

ACKNOWLEDGEMENTS

First and foremost, great thanks are due to my advisor, Xiaoming Huo. Without his great technical insight and seemingly infinite patience, this thesis would never have been finished. My graduate school experience has surely been much more enjoyable and less stressful because of his guidance.

I would also like to thank my other committee members: Hongyuan Zha, Alexander Shapiro, Nicoleta Serban, and Ming Yuan. Special thanks go to Hongyuan Zha for several very helpful discussions about the material covered in Chapters 2 and 3, and to Alexander Shapiro for his advice which led to substantial improvement in Chapter 4.

My family has been a tremendous source of encouragement throughout the past 5 years. Thanks Mom, Dad, Jonathan, and Julie, for supporting me, even though I was never really able to explain to you just what I've been doing all this time.

Finally, I gratefully acknowledge the generous financial support of the ARCS foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

In a broad sense, this thesis focuses on two aspects of the problem of model selection in applied statistics. In the first part we consider the theoretical properties of two nonlinear dimensionality reduction algorithms, which can be viewed as a first step in the model selection process, since many statistical methods will fail in the presence of very high dimensionality in the covariates. In the second, we develop a new approach to the model selection process by considering convex combinations of a fixed set of model selection criteria.

Dimension reduction plays an extremely important role in many areas of statistical application. In particular, many problems arising in the modern practice of statistics involve many predictor variables, even in some cases more than the number of available observations (most notably in the analysis of microarray data.) Many classical statistical methods become useless in these situations. However, often one has prior reason to believe that, despite the large number of predictor variables available, the structure underlying them is actually much simpler. Donoho and Grimes [6] discuss canonical examples involving sets of images of similar subjects, but at different angles and positions. In these cases, dimension reduction can be a vital tool which greatly expands the number of statistical modelling methodologies available to the analyst, accurately condensing the information contained in the many predictor variables into a much smaller set of reduced predictors.

The area of dimension reduction has a long history, going all the way back to the development of principal components. Nonlinear dimension reduction, however, is a more recently developed generalization of classical methods. We focus on two algorithms which show particular promise in this field, namely Local Tangent Space Alignment (LTSA) and Hessian Locally Linear Embedding (HLLE). Our work in this thesis focuses on the asymptotic properties of these algorithms, and in particular we demonstrate, for the first time, that each algorithm will yield the correct result with a large enough sample size, subject to mild conditions on the underlying low-dimensional structure. Thus we provide a stronger theoretical foundation for the application of these procedures in situations where the sample size is large, and the noise structure is well-understood. In a sense, this fills in a gap between the fairly well-known asymptotic properties of linear dimension reduction methods [1] and the relatively unexplored theoretical aspects of the newer nonlinear methods.

The problem of model selection remains an important and difficult one in almost every area of applied statistics. The fundamental tradeoff between fidelity to the data and complexity of the resulting model is nearly always present, and is almost never easy to resolve. This is perhaps easiest to see in the case of regression, in which the use of n linearly independent predictor variables on a set of n observations will always yield a perfect fit, but such a model would never be useful in practice. Thus, in many applied problems, the analyst must decide to exclude some predictor variables, but this will always entail what seems to be a poorer fit to the data. Deciding how many and which predictors to exclude is thus a crucial step, but there is no method which generally acknowledged to be the best way to determine which predictors are to be kept, and which to be left out.

To that end, many model selection criteria have been proposed. These criteria are functions of a fitted model, most of which involve both the likelihood of the data and a penalty term for model complexity. Many of these criteria have an appealing simplicity – one need only fit all possible regression models (of which there will be 2^p in the case when there are p possible predictor variables available), and among the resulting models, choose the one that minimizes (or maximizes) the corresponding criterion function. The most commonly used criteria are Akaike's An Information Criterion (AIC) and the Bayesian Information Criterion (BIC), and cross validation (CV), though there are many others also to be found in the literature.

Though the model selection problem may seem easier with the aid of these criteria, there is a new problem that arises with their use – it is not at all obvious, or agreed upon, which model selection criteria are best. Rather than the original model selection problem, we are now faced with a model selection criterion selection problem. In Chapter 4, we propose a new simulation method to compare the quality of model selection criteria. The main strength of this method is that it allows one to remain agnostic on the general question of which criterion is "best" overall, while giving useful recommendations specific to the particular problem under consideration.

In general, we can think of most model selection criteria as being in one of 3 categories – those based purely on in-sample fit, such as AIC and BIC, those based on cross-validated in-sample fit, such as PRESS in the case of regression, and those based on true out-of-sample fit, which use a holdout sample not included in the model fitting process as a means to evaluate the fitted models. Intuitively, one expects these 3 types of criteria to yield different information about the fitted models, all of which could potentially be useful. Rather than simply choosing between these criteria, which would necessarily involve ignoring all the other criteria not chosen, an attractive alternative would be to *combine* them in some sensible way, hopefully exploiting all the different types of information available. In Chapter 4, we do just that, generalizing our simulation procedure to allow for the combination of model selection criteria based on a simple idea of allowing convex combinations of ranks. These combinations give additional insight into the structure of the problem, and as we demonstrate in simulation case studies, can sometimes yield substantial improvement over traditional model selection procedures.

CHAPTER II

PERFORMANCE ANALYSIS OF LTSA

2.1 Introduction

Manifold-based dimensionality reduction methods have attracted substantial attention in both the machine learning and statistics communities due to their demonstrated potential. Though many methods have been proposed, little work has been done to analyze the performance of these methods. The main contribution of this chapter is to establish some asymptotic performance properties of a manifold learning algorithm, as well as a demonstration of some of its limitations. The key idea in our analysis is to treat the solutions of manifold learning algorithms as invariant subspaces, and then carry out a matrix perturbation analysis. A common feature of several manifold learning algorithms (e.g., [18, 3, 6, 4, 29]) is that their solutions correspond to invariant subspaces, typically the eigenspace associated with the smallest eigenvalues of a kernel matrix. The exact form of this kernel matrix, of course, depends on the details of the particular algorithm. These subspaces, however, are clearly invariant regardless of the exact form of the matrix involved, because they are spanned by eigenvectors [25, Section I.3.4].

Many efficient ML algorithms have been developed. A partial list of them is: locally linear embedding (LLE) [18], ISOMAP [26], charting [4], local tangent space alignment (LTSA) [29], Laplacian eigenmaps [3], and Hessian eigenmaps [6], etc. LTSA, in particular, enjoys several advantages. First of all, in numerical simulation (e.g., using the tools offered by [27]), we find empirically that LTSA performs among the best of the available algorithms. Second, the solution to each step of the LTSA algorithm is an invariant subspace, which makes analysis of its performance more tractable. Third, the similarity between LTSA and several other ML algorithms (e.g., LLE, Laplacian eigenmaps and Hessian eigenmaps) suggests that our results may generalize. Thus, it is our hope that this performance analysis will provide a theoretical foundation for the application of ML algorithms. Our main theoretical result is Theorem 2.3.8, which is a worst-case upper bound on the angle between the subspaces spanned by the computed coordinates and by the intrinsic parameters.

The rest of the chapter is organized as follows. The problem formulation and background information are presented in Section 2.2. In Section 2.3, perturbation analysis is carried out, and the main theorem is proved. In Section 2.4, more simulation results are presented to illustrate the analytical properties. Some discussion related to existing work in this area is included in Section 2.5. Finally, we present concluding remarks in Section 2.6. Technical proofs are relegated to Appendix A when convenient.

2.2 Problem Statement and Illustration

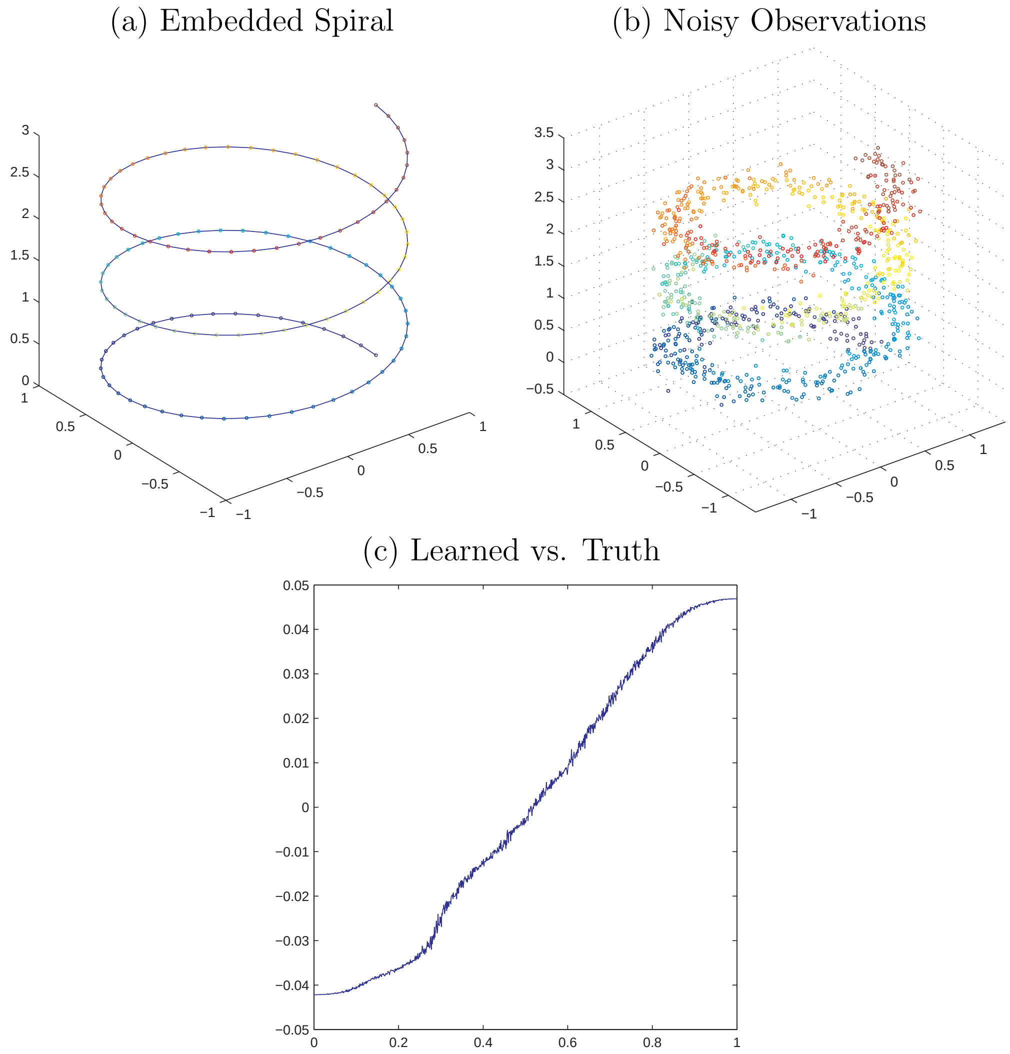
2.2.1 Model

To be more specific, we formulate our DR problem as follows. For a positive integer n, let $y_i \in \mathbb{R}^D, i = 1, 2, \ldots, n$, denote n observations. We assume that there is a mapping $f : \mathbb{R}^d \to \mathbb{R}^D$ which satisfies a set of regularity conditions. In addition, we require another set of (possibly multivariate) values $x_i \in \mathbb{R}^d, d < D, i = 1, 2, \ldots, n$, such that

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\varepsilon_i \in \mathbb{R}^D$ denotes a random error. For example, we may assume $\varepsilon_i \sim N(\vec{0}, \sigma^2 I_D)$; i.e., a multivariate normal distribution with mean zero and variance-covariance proportional to the identity matrix. The central questions of DR are: (1) Can we find a set of low-dimensional vectors such that (1) holds? (2) What kind of regularity conditions should be imposed on f? (3) Is the model well defined? These questions will be answered in the following.

2.2.2 A Pedagogical Example


(a) Embedded Spiral

(b) Noisy Observations

(c) Learned vs. Truth

Figure 1: An illustrative example of LTSA in nonparametric dimension reduction. The straight line pattern in (c) indicates that the underlying parametrization has been approximately recovered.

An illustrative example of DR that makes our formulation more concrete is given in Figure 1. Subfigure (a) shows the true underlying structure of a toy example, a 1-D spiral. The *noiseless* observations are equally spaced points on this spiral. In subfigure (b), 1024 *noisy* observations are generated with multivariate noise satisfying $\varepsilon_i \sim N(\vec{0}, \frac{1}{100}\mathbf{I}_3)$. We then apply LTSA to the noisy observations, using $k = 10$

nearest neighbors. In subfigure (c), the result from LTSA is compared with the true parametrization. When the underlying parameter is faithfully recovered, one should see a straight line, which is observed in subfigure (c).

2.2.3 Regularity and Uniqueness of the Mapping f

If the conditions on the mapping f are too general, the model (1) is not well defined. For example, if the mapping $f(\cdot)$ and point set $\{x_i\}$ satisfy (1), so do $f(A^{-1}(\cdot - b))$ and $\{Ax_i + b\}$, where A is an invertible d by d matrix and b is a d-dimensional vector. As is common in the manifold-learning literature, we adopt the following condition on f.

Condition 2.2.1 (Local Isometry) *The mapping f is locally isometric: For any $\varepsilon > 0$ and x in the domain of f, let $N_\varepsilon(x) = \{z : \|z - x\|_2 < \varepsilon\}$ denote an ε-neighborhood of x using Euclidean distance. We have*

$$\|f(x) - f(x_0)\|_2 = \|x - x_0\|_2 + o(\|x - x_0\|_2).$$

The above condition indicates that in a local sense, f preserves Euclidean distance. Let $J(f; x_0)$ denote the Jacobian of f at x_0. We have $J(f; x_0) \in \mathbb{R}^{D\times d}$, where each column (resp., row) of $J(f; x_0)$ corresponds to a coordinate in the feature (resp., data) space. The above in fact implies the following lemma.

Lemma 2.2.2 *The matrix $J(f; x_0)$ is orthonormal for any x_0, i.e., $J^T(f; x_0)J(f; x_0) = I_d$.*

A reference for this result is Zhang and Zha [29].

Given the previous condition, model (1) is still not uniquely defined. For example, for any d by d orthogonal matrix O and any d-dimensional vector b, if $f(\cdot)$ and $\{x_i\}$ satisfy (1) and Condition 3.2.1, so do $f(O^T(\cdot - b))$ and $\{Ox_i + b\}$. We can force b to be $\vec{0}$ by imposing the condition that $\sum_i x_i = 0$. In DR, we can consider the sets $\{x_i\}$ and $\{Ox_i\}$ "invariant," because one is just a rotation of the other. In fact, the invariance coincides with the concept of "invariant subspace" that will be discussed later.

Condition 2.2.3 (Local Linear Independence Condition) *Let $Y_i \in \mathbb{R}^{D\times k}$, $1 \leq i \leq n$, denote a matrix whose columns are made by the ith observation y_i and its $k-1$ nearest neighbors. We choose $k-1$ neighbors so that the matrix Y_i has k columns. It is generally assumed that $d < k$. For any $1 \leq i \leq n$, the rank of $Y_i\overline{P}_k$ is at least d; in other words, the dth largest singular value of matrix $Y_i\overline{P}_k$ is greater than 0.*

The regularity of the manifold can be determined by the Hessians of the mapping. Rewrite $f(x)$ for $x \in \mathbb{R}^d$ as

$$f(x) = (f_1(x), f_2(x), \ldots, f_D(x))^T.$$

Furthermore, let $x = (x_1, \ldots, x_d)^T$. A Hessian is

$$[H_i(f; x)]_{jk} = \frac{\partial^2 f_i(x)}{\partial x_j \partial x_k},$$

for $1 \leq i \leq D, 1 \leq j, k \leq d$.

The following condition ensures that f is locally smooth. We impose a bound on all the components of the Hessians.

Condition 2.2.4 (Regularity of the Manifold) $|[H_i(f; x)]_{jk}| \leq C_1$ *for all i, j, and k, where $C_1 > 0$ is a prescribed constant.*

2.2.4 Solutions as Invariant Subspaces and a Related Metric

We now give a more detailed discussion of invariant subspaces. Let $\mathcal{R}(X)$ denote the subspace spanned by the columns of X. Recall that $x_i, i = 1, 2, \ldots, n$, are the true low-dimensional representations of the observations. We treat the x_i's as column vectors. Let

$$X = (x_1, x_2, \cdots, x_n)^T;$$

i.e., the ith row of X corresponds to $x_i, 1 \leq i \leq n$. If the set $\{Ox_i\}$, where O is a d by d orthogonal square matrix, forms another solution to the dimension reduction problem, we have

$$(Ox_1, Ox_2, \cdots, Ox_n)^T = XO^T.$$

It is evident that $\mathcal{R}(XO^T) = \mathcal{R}(X)$. This justifies the *invariance* that was mentioned earlier.

The goal of our performance analysis is to answer the following question: Letting $\|\tan(\cdot,\cdot)\|_2$ denote the Euclidean norm of the vector of canonical angles between two invariant subspaces ([25, Section I.5]), and letting X and $\widetilde{X}$ denote the true and estimated parameters, respectively, how do we evaluate $\|\tan(\mathcal{R}(X), \mathcal{R}(\widetilde{X}))\|_2$?

2.2.5 LTSA: Local Tangent Space Alignment

We now review LTSA. There are two main steps in the LTSA algorithm [29].

1. The first step is to compute the local representation on the manifold. Consider a projection matrix $\overline{P}_k = I_k - \frac{1}{k} \cdot \mathbf{1}_k \mathbf{1}_k^T$, where I_k is the k by k identity matrix and $\mathbf{1}_k$ is a k-dimensional column vector of ones. It is easy to verify that $\overline{P}_k = \overline{P}_k \cdot \overline{P}_k$, which is a characteristic of projection matrices.

   We solve the minimization problem:

   $$\min_{\Lambda, V} \|Y_i \overline{P}_k - \Lambda V\|_F,$$

   where $\Lambda \in \mathbb{R}^{D \times d}, V \in \mathbb{R}^{d \times k}$, and $VV^T = I_d$. Let V_i denote optimal V. Then the row vectors of V_i are the d right singular vectors of $Y_i \overline{P}_k$.

(2) The solution to LTSA corresponds to the invariant subspace which is spanned and determined by the eigenvectors associated with the 2nd to the $(d+1)$st

smallest eigenvalues of the matrix

$$(S_1, \ldots, S_n) \begin{pmatrix} \overline{P}_k - V_1^T V_1 & & & \\ & \overline{P}_k - V_2^T V_2 & & \\ & & \ddots & \\ & & & \overline{P}_k - V_n^T V_n \end{pmatrix} (S_1, \ldots, S_n)^T. \tag{2}$$

where $S_i \in \mathbb{R}^{n \times k}$ is a selection matrix such that $Y^T S_i = Y_i$, where $Y = (y_1, y_2, \ldots, y_n)^T$.

As mentioned earlier, the subspace spanned by the eigenvectors associated with the 2nd to the $(d+1)$st eigenvalues of the matrix in (2) is an invariant subspace, which will be analyzed under perturbation.

We have slightly reformulated the original algorithm as presented in [29], in order to simplify the theoretical analysis. The verification of the equivalence is a standard exercise in linear algebra, and it is given in the Appendix of [12].

2.3 Perturbation Analysis

We now carry out a perturbation analysis on the reformulated version of LTSA. There are two steps in our analysis: in the *local* step (Section 2.3.1), we characterize the deviation of the null spaces of the matrices $\overline{P}_k - V_i^T V_i, i = 1, 2, \ldots, n$. In the *global* step (Section 2.3.2), we derive the variation of the null space under global alignment. The detailed calculations are again relegated to Appendix A.

2.3.1 Local Coordinates

Let X be the matrix of true parameters. We define

$$X_i = X^T S_i = (x_1, x_2, \cdots, x_n) S_i;$$

i.e., the columns of X_i are made by x_i and those x_j's that correspond to the $k-1$ nearest neighbors of y_i. We require a bound on the size of the local neighborhoods defined by the X_i's.

Condition 2.3.1 (Universal Bound on the Sizes of Neighborhoods) *For all $i, 1 \leq i \leq n$, we have $\tau_i < \tau$, where τ is a prescribed constant and τ_i is an upper bound on the distance between two columns of X_i: $\tau_i = \max_{x_j, x_k} \|x_j - x_k\|$, where the maximum is taken over all columns of X_i.*

In this chapter, we are interested in the case when $\tau \to 0$.

We will need conditions on the local tangent spaces. Let $d_{\min,i}$ (respectively, $d_{\max,i}$) denote the minimum (respectively, maximum) singular values of $X_i \overline{P}_k$. Let

$$d_{\min} = \min_{1 \leq i \leq n} d_{\min,i},$$

and

$$d_{\max} = \max_{1 \leq i \leq n} d_{\max,i}.$$

We have the following result regarding $d_{\max}$:

Lemma 2.3.2

$$d_{\min} \leq d_{\max} \leq \tau\sqrt{k}. \tag{3}$$

For the proof, see Appendix A.1.1.

Condition 2.3.3 (Local Tangent Space) *There exists a constant $C_2 > 0$, such that*

$$C_2 \cdot \tau \leq d_{\min}. \tag{4}$$

The above can roughly be thought of as requiring that the local dimension of the manifold remain constant (i.e., the manifold has no singularities.)

The following condition defines a global bound on the errors (ε_i).

Condition 2.3.4 (Universal Error Bound) *There exists $\sigma > 0$, such that $\forall i, 1 \leq i \leq n$, we have $\|y_i - f(x_i)\|_\infty < \sigma$. Moreover, we assume $\sigma = o(\tau)$; i.e., we have $\frac{\sigma}{\tau} \to 0$, as $\tau \to 0$.*

It is reasonable to require that the error bound (σ) be smaller than the size of the neighborhood (τ), which is reflected in the above condition. We discuss the necessity of this condition in Section 2.3.3.

Within each neighborhood, we give a perturbation bound between an invariant subspace spanned by the true parametrization and the invariant subspace spanned by the singular vectors of the matrix of noisy observations. Let

$$X_i \overline{P}_k = A_i D_i B_i$$

be the singular value decomposition of the matrix $X_i \overline{P}_k$; here $A_i \in \mathbb{R}^{d \times d}$ is orthogonal ($A_i A_i^T = I_d$), $D_i \in \mathbb{R}^{d \times d}$ is diagonal, and the rows of $B_i \in \mathbb{R}^{d \times k}$ are the right singular vectors corresponding to the largest singular values ($B_i B_i^T = I_d$). It is not hard to verify that

$$B_i = B_i \overline{P}_k. \tag{5}$$

Let $Y_i \overline{P}_k = \widetilde{A}_i \widetilde{D}_i \widetilde{B}_i$ be the singular value decomposition of $Y_i \overline{P}_k$, and assume that this is the "thin" decomposition of rank d. We may think of this as the perturbed version of $J(f; x_i^{(0)}) X_i \overline{P}_k$. The rows of $\widetilde{B}_i$ are the eigenvectors of $(Y_i \overline{P}_k)^T (Y_i \overline{P}_k)$ corresponding to the d largest eigenvalues. Let $\mathcal{R}(B_i^T)$ (respectively, $\mathcal{R}(\widetilde{B}_i^T)$) denote the invariant subspace that is spanned by the columns of matrix B_i^T (respectively, $\widetilde{B}_i^T$).

Theorem 2.3.5 *Given invariant subspaces $\mathcal{R}(B_i^T)$ and $\mathcal{R}(\widetilde{B}_i^T))$ as defined above, we have*

$$\lim_{\tau \to 0} \| \sin(\mathcal{R}(B_i^T), \mathcal{R}(\widetilde{B}_i^T)) \|_2 \leq C_3 \left(\frac{\sigma}{\tau} + C_1 \tau \right),$$

where C_3 is a constant that depends on k, D and C_2.

The proof is presented in Appendix A.1.2. The above gives an upper bound on the deviation of the local invariant subspace in step (1') of the modified LTSA. It will be used later to prove a global result.

2.3.2 Global Alignment

Condition 2.3.6 (No Overuse of One Observation) *There exists a constant* C_4*, such that*

$$\left\| \sum_{i=1}^{n} S_i \right\|_\infty \leq C_4.$$

Note that we must have $C_4 \geq k$. The next condition (Condition 2.3.7) will implicitly give an upper bound on C_4.

Recall that the quantity $\| \sum_{i=1}^{n} S_i \|_\infty$ is the maximum row sum of the absolute values of the entries in $\sum_{i=1}^{n} S_i$. The value of $\| \sum_{i=1}^{n} S_i \|_\infty$ is equal to the maximum number of nearest neighbor subsets to which a single observation belongs.

We will derive an upper bound on the angle between the invariant subspace spanned by the result of LTSA and the space spanned by the true parameters.

Given (5), it can be shown that

$$X_i \overline{P}_k (\overline{P}_k - B_i^T B_i)(X_i \overline{P}_k)^T = 0.$$

Recall $X = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^{n \times d}$. It is not hard to verify that the row vectors of

$$(\mathbf{1}_n, X)^T \tag{6}$$

span the $(d+1)$-dimensional null space of the matrix:

$$(S_1, \ldots, S_n)\overline{P}_k \begin{pmatrix} I - B_1^T B_1 & & & \\ & I - B_2^T B_2 & & \\ & & \ddots & \\ & & & I - B_n^T B_n \end{pmatrix} \overline{P}_k (S_1, \ldots, S_n)^T. \tag{7}$$

Assume that

$$\begin{pmatrix} \frac{\mathbf{1}_n^T}{\sqrt{n}} \\ X^T \\ (X^c)^T \end{pmatrix}$$

is orthogonal, where $X^c \in \mathbb{R}^{n\times(n-1-d)}$. Although in our original problem formulation, we made no assumptions about the x_i's, we can still assume that the columns of X are orthonormal because we can transform any set of x_i's into an orthonormal set by rescaling the columns and multiplying by an orthogonal matrix. Based on the previous paragraph, we have

$$\begin{pmatrix} \frac{\mathbf{1}_n^T}{\sqrt{n}} \\ X^T \\ (X^c)^T \end{pmatrix} M_n \left(\frac{\mathbf{1}_n}{\sqrt{n}}, X, X^c\right) = \begin{pmatrix} \mathbf{0}_{(d+1)\times(d+1)} & \mathbf{0}_{(d+1)\times(n-d-1)} \\ \mathbf{0}_{(n-d-1)\times(d+1)} & L_2 \end{pmatrix} \tag{8}$$

where

$$M_n = (S_1, \dots, S_n)\overline{P}_k \begin{pmatrix} I_k - B_1^T B_1 & & \\ & \ddots & \\ & & I_k - B_n^T B_n \end{pmatrix} \overline{P}_k (S_1, \dots, S_n)^T$$

and

$$L_2 = (X^c)^T M_n X^c.$$

Let $\lambda_{\min}^+$ denote the minimum singular value (i.e., eigenvalue) of L_2. We will need the following condition on $\lambda_{\min}^+$.

Condition 2.3.7 (Appropriateness of Global Dimension) $\lambda_{\min}^+ > 0$ *and* $\lambda_{\min}^+$ *goes to* 0 *at a slower rate than* $\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau$*; i.e., as* $\tau \to 0$*, we have*

$$\frac{\left(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau\right) \cdot \|\sum_{i=1}^n S_i\|_\infty}{\lambda_{\min}^+} \to 0.$$

As discussed in [28], this condition is actually related to the amount of overlap between the nearest neighbor sets.

Theorem 2.3.8 (Main Theorem)

$$\lim_{\tau\to 0} \|\tan(\mathcal{R}(\widetilde{X}), \mathcal{R}(X))\|_2 \le \frac{C_3(\frac{\sigma}{\tau} + C_1\tau) \cdot \|\sum_{i=1}^n S_i\|_\infty}{\lambda_{\min}^+}. \tag{9}$$

As mentioned in Section 2.1, the above theorem gives a worst-case bound on the performance of LTSA. A discussion on when Condition 2.3.7 is satisfied will be long and beyond the scope of this thesis. We leave it to future investigation.

2.3.3 The requirement that $\sigma \to 0$

A natural question to ask, in light of the above analysis, is whether LTSA is still consistent without the restrictive assumption that $\sigma \to 0$. In this section, we discuss a simple example which demonstrates that the answer is, surprisingly, no.

Consider the following model:

$$\begin{aligned} x_i &\sim N(0, \sigma_x^2) \\ \epsilon_i &\sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_\epsilon^2 I_2\right] \\ y_i &= \begin{pmatrix} x_i \\ 0 \end{pmatrix} + \epsilon_i, \end{aligned}$$

where $x_i \perp \epsilon_i$, . It is then easy to see that

$$y_i \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 + \sigma_x^2 & 0 \\ 0 & \sigma_\epsilon^2 \end{pmatrix}\right].$$

Suppose, as usual, that we wish to reconstruct the x_i's from the given y_i's. This is a particularly simple case of dimension reduction, where $D = 2$, $d = 1$, and the data lie near a *linear* manifold. Thus, the entire manifold may be thought of as a single linear patch. In applying LTSA to this model, we may therefore assume that $k = n$, that is, that all points in the data set are neighbors of one another. This implies that $S_i = I_n$ for each i.

Now, in the first step, LTSA will find the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix. It is a standard result [1, Section 13.5] that the space spanned by the leading eigenvector converges to the space spanned by $(1, 0)^T$. Without loss of generality, we may suppose that the eigenvector is chosen so that the first component is positive, and therefore the leading eigenvector converges to $(1,0)^T$. The estimated local coordinates will then be

$$\theta_i = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} y_i = x_i + \epsilon_i^{(1)},$$

where we have denoted the first component of ϵ_i by $\epsilon_i^{(1)}$. Let $\Theta = (\theta_1, \theta_2, \ldots, \theta_n)$ denote the row vector formed by the n estimated local coordinates. We assume that $\|\Theta\|_2 = 1$, that is, that the eigenvector associated with the largest eigenvalue of the sample covariance matrix is normalized. Also note that we have $\Theta \perp \mathbf{1}_n$.

The alignment step is especially simple due the structure of our artificial example. The computed $\widehat{x}_i$'s are given by the eigenvector corresponding to the smallest eigenvalue of

$$(S_1, S_2, \cdots, S_n)\overline{P}_{k\times n} \begin{pmatrix} I_k - \Theta_1^\dagger\Theta_1 & & & \\ & I_k - \Theta_2^\dagger\Theta_2 & & \\ & & \ddots & \\ & & & I_k - \Theta_n^\dagger\Theta_n \end{pmatrix} \overline{P}_{k\times n}(S_1, S_2, \cdots, S_n)^T, \tag{10}$$

The computation is easily simplified, however. As noted above, each S_i as the identity, and the diagonal blocks in the center matrix are all the same. Therefore, the $\widehat{x}$'s can be expressed as the eigenvector corresponding to the second smallest eigenvalue of

$$\overline{P}_n(I_n - \Theta^\dagger\Theta)\overline{P}_n.$$

It is easy to see that the correct eigenvector is proportional to $\Theta^\dagger = \Theta^T$ by noting that

$$(I_n - \Theta^\dagger\Theta)\Theta^\dagger = \Theta^\dagger - \Theta^\dagger\Theta\Theta^\dagger = 0.$$

Therefore, the vector $\Theta^\dagger$ corresponds to the eigenvalue 0 of $\overline{P}_n(I_n - \Theta^\dagger\Theta)\overline{P}_n$. Further, we know that the dimension of the nullspace of $\overline{P}_n(I_n - \Theta^\dagger\Theta)\overline{P}_n$ is exactly $d+1 = 2$, so there can be no other vector in the nullspace except, of course, for $\mathbf{1}_n$. If σ_ϵ^2 is constant, then in general we will have

$$X = (x_1, x_2, \ldots, x_n)^T,$$

but

$$\widetilde{X} = (x_1 + \epsilon_1^{(1)}, x_2 + \epsilon_2^{(1)}, \ldots, x_n + \epsilon_n^{(1)})^T.$$

Now, we consider the angle formed between the two subspaces ($\mathcal{R}(X)$ and $\mathcal{R}(\widetilde{X})$. In this special one-dimensional case, this has a particularly simple form:

$$\angle(\mathcal{R}(X), \mathcal{R}(\widetilde{X})) = \cos^{-1}\left(\frac{X^T\widetilde{X}}{\|X\| \cdot \|\widetilde{X}\|}\right).$$

Supposing that n is sufficiently large, we may use the strong law of large numbers to evaluate the limits of the quantities on the right-hand side. For the numerator of the fraction, we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{X^T\widetilde{X}}{n} &= \frac{1}{n}\lim_{n\to\infty}\sum_{i=1}^{n} x_i \cdot (x_i + \epsilon_i^{(1)}) \\
&\overset{\text{SLLN}}{=} E(x_i \cdot (x_i + \epsilon_i^{(1)})) \\
&= E(x_i^2) + E(x_i) \cdot E(\epsilon_i^{(1)}) \\
&= \sigma_x^2.
\end{aligned}$$

For the denominator, a similar argument shows that

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\|X\|^2 &= \sigma_x^2, \\
\lim_{n\to\infty} \frac{1}{n}\|\widetilde{X}\|^2 &= \sigma_x^2 + \sigma_\epsilon^2.
\end{aligned}$$

Putting these limits together,

$$\begin{aligned}
\lim_{n\to\infty} \angle(\mathcal{R}(X), \mathcal{R}(\widetilde{X})) &= \cos^{-1}\left(\frac{n\sigma_x^2}{\sqrt{n\sigma_x^2} \cdot \sqrt{n(\sigma_x^2 + \sigma_\epsilon^2)}}\right) \\
&= \cos^{-1}\left(\frac{\sigma_x^2}{\sqrt{\sigma_x^4 + \sigma_x^2\sigma_\epsilon^2}}\right).
\end{aligned}$$

If σ_ϵ^2 is constant (i.e., does not have limit 0,) then the argument of $\cos^{-1}$ will not have limit 1, and $\lim_{n\to\infty} \angle(\mathcal{R}(X), \mathcal{R}(\widetilde{X})) \neq 0$. Note that this inconsistency would still apply even if we add the stronger assumption that the distribution of the errors is bounded.

Thus, we see that our assumption that $\sigma \to 0$ is, in fact, necessary to ensure the consistency of LTSA, even in what might be considered the simplest possible case of the dimension reduction problem. While this result may at first seem somewhat counterintuitive, it is less surprising when one considers the fact that the number of unknown parameters (in this case, the x_i's) grows as n increases, so our dimension reduction problem is not analogous to traditional parameter estimation problems such as the classical model

$$y_i = \mu + \epsilon_i$$

with only μ (1 parameter) unknown.

An additional difficulty which arises in the absence of the assumption that $\sigma \to 0$ is the fact that the estimated selection matrices (the $\widehat{S}_i$'s) may not converge to the correct population counterparts (the S_i's). An implicit assumption throughout our analysis is that, at least asymptotically $\widehat{S}_i = S_i$, which is crucial in our derivation of bounds on the deviation of the estimated alignment matrix from the true alignment matrix. In the asymptotically noiseless case, this convergence is automatic, provided that the underlying manifold is not self-intersecting. However, in the asymptotic case with noise, such convergence is not guaranteed and in fact, will not hold in general. This is compounded with the difficulties discussed above related to our toy example.

Considering the problem from a geometric perspective is also illuminating. While it is well-known that we can asymptotically recover the correct local tangent space at each point (at least in our simplified example), the problem occurs in the alignment step. The simple structure of the example makes it easy to see what is going on — we extract the *projection* of y_i onto the local tangent space. However, this of course does not correspond to the generating coordinate x_i in the general case, though it could be construed as a maximum likelihood estimate of the generating coordinate, being the closest point in the transformed parameter space to the actual observation in terms of Euclidean distance. What we can recover, then, is the projection of $f(x_i) + \epsilon_i$ onto the (asymptotically correct) tangent space in each neighborhood, but the original generating coordinate itself is unrecoverable. Figure 2 illustrates this phenomenon.

It would certainly be interesting to know whether f can still be recovered asymptotically if σ is constant, but this question remains open. The analysis of the reconstruction of f is more complicated because LTSA does not compute any function explicitly — an estimated f can only be computed implicitly, for example by polynomial regression of Y on $\widetilde{X}$ as discussed in Section 5 of [29]. An analysis of this situation would involve consideration of the interplay of the errors in $\widetilde{X}$ with the errors in reconstructing the function f via indirect methods based on $\widetilde{X}$. We leave this to future investigation.

A further consequence of this result is that while plots such as those shown in Figure 3 can be useful as rough indicators of LTSA's performance, they are not reliable in a strict sense for determining consistency. Although the relationship between the true and estimated coordinates may appear to be roughly linear, this alone does not imply that the algorithm will asymptotically recover the correct coordinates — the trouble is the "bandwidth" of the graph. If the underlying parameters are truly recovered, the graph must eventually converge to *exactly* a straight line with no dispersion. Such information is difficult to discern from plots of this type.

2.4 Simulations

In the same setting as in Section 2.2.2, if we change the value of σ from $\sigma = 0.1$ to $\sigma = 0.025$ and 0.2, we have Figure 3. Based on our theorem, the smaller the error standard deviation is, the closer the result of LTSA is to the true parametrization. In the case of $\sigma = 0.2$, the result of LTSA breaks down.

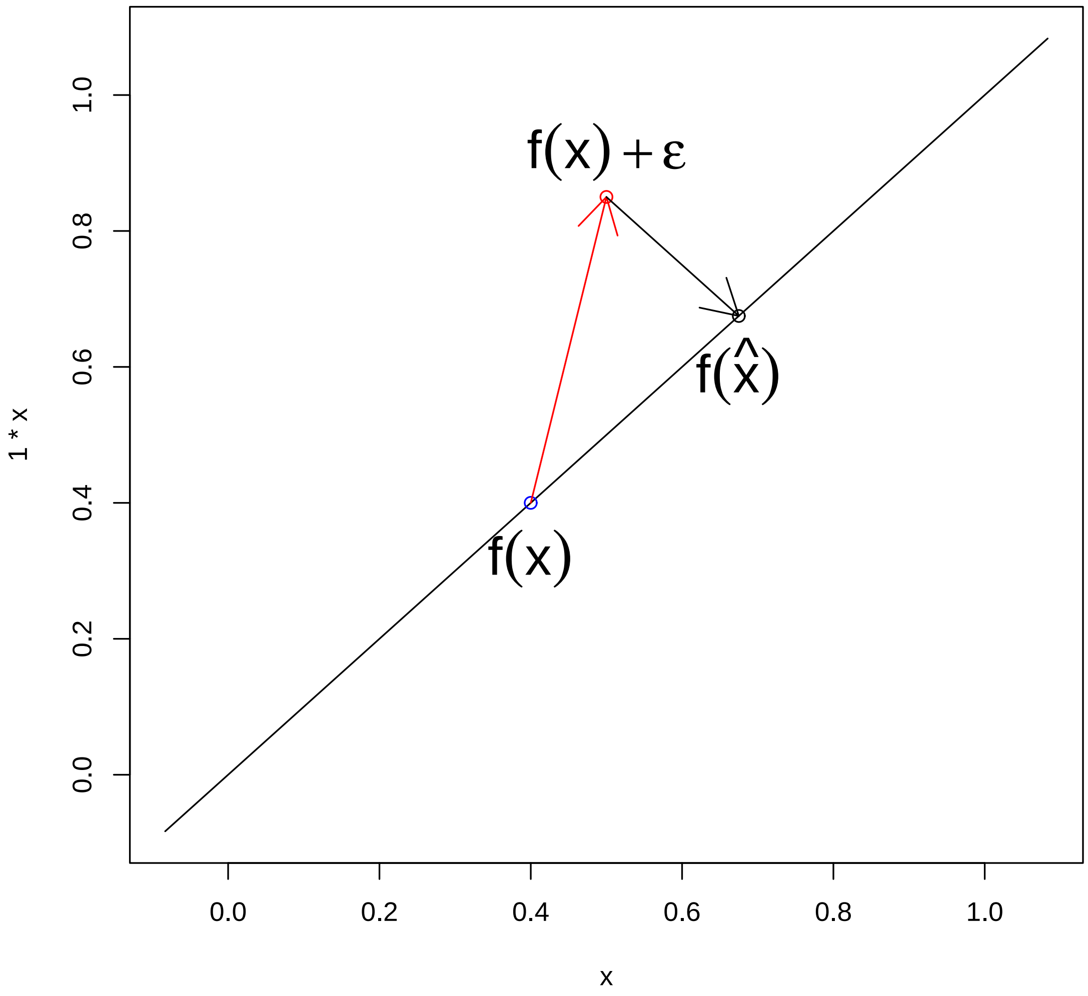
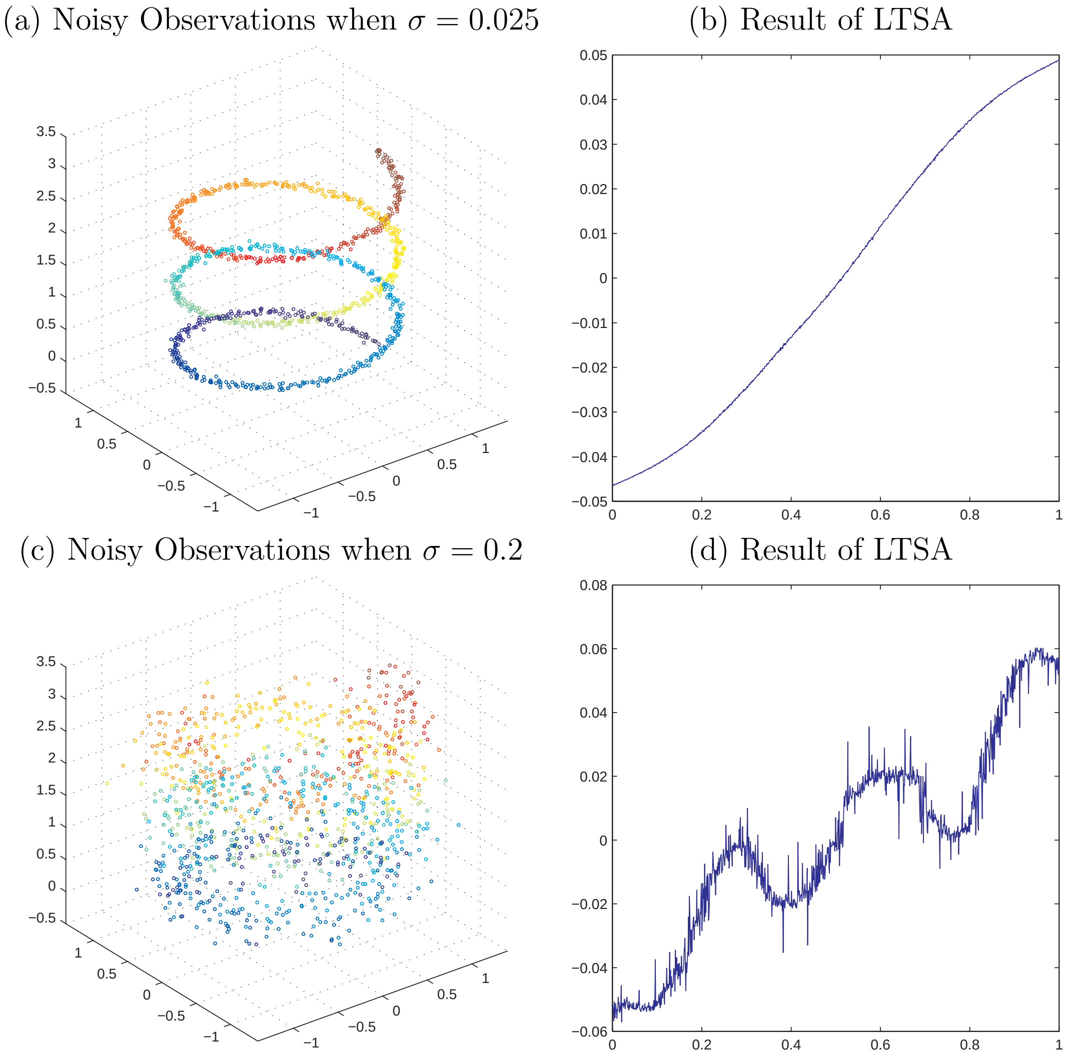


Figure 2: An illustration of why $\sigma \to 0$ is a necessary condition for convergence of LTSA. Though the principal subspace will be estimated correctly, the projection of $f(x_i) + \epsilon_i$ onto the principal subspace is not the same as the underlying coordinate x_i.



(a) Noisy Observations when $\sigma = 0.025$

(b) Result of LTSA

(c) Noisy Observations when $\sigma = 0.2$

(d) Result of LTSA

Figure 3: Reruns of the illustrative example in Section 2.2.2, with different noise standard deviations.

When X and $\widetilde{X}$ are one-dimensional, we have

$$\sqrt{1-[\mathrm{corr}(X,\widetilde{X})]^2} = \|\sin(\mathcal{R}(X),\mathcal{R}(\widetilde{X}))\|_2 \leq \|\tan(\mathcal{R}(X),\mathcal{R}(\widetilde{X}))\|_2,$$

where $\mathrm{corr}(X,\widetilde{X})$ is the correlation coefficient between two vectors. If

$$\|\tan(\mathcal{R}(X),\mathcal{R}(\widetilde{X}))\|_2 \to 0,$$

we have $\mathrm{corr}(X,\widetilde{X}) \to 1$, which corresponds to the consistency.

In Figure 3 (b), when σ is small, we observe a nearly straight line; while in Figure 3 (d), where σ is large, the estimates are drastically different from what they are supposed to be. This phenomenon is consistent with our theory.

2.5 Discussion

To the best of our knowledge, the performance analysis that is based on invariant subspaces is new. Consequently the worst-case upper bound is the first of its kind. There are still open questions to be addressed (Section 2.5.1). In addition to a discussion on the relation of LTSA to existing DR methodologies, we will also address relation with known results as well (Section 2.5.2).

2.5.1 Open Questions

The rate of convergence of $\lambda^+_{\min}$ is determined by the topological structure of f. It is important to estimate this rate of convergence, but this issue has not been addressed here.

We assume that $\tau \to 0$. One can imagine that it is true when the error bound (σ) goes to 0 and when the x_i's are sampled with a sufficient density in the support of f. An open problem is how to derive the rate of convergence of $\tau \to 0$ as a function of the topology of f and the sampling scheme. After doing so, we may be able to decide where our theorem is applicable.

Given a covering scheme, such as choosing the k-nearest neighbors, a verification of $\tau \to 0$ and a derivation of its corresponding rate is an open question, too. The answer to this will depend on the topology of f, which is not covered in this chapter, and the sampling scheme.

2.5.2 Relation to Existing Work

The error analysis in the original LTSA paper is the closest to our result. However, Zhang and Zha [29] do not interpret their solutions as invariant subspaces, and hence their analysis does not yield a worst case bound as we have derived here.

Reviewing the original papers on LLE [18], Laplacian eigenmaps [3], and Hessian eigenmaps [6] reveals that their solutions are subspaces spanned by a specific set of eigenvectors. This naturally suggests that results analogous to ours may be derivable as well for these algorithms. A recent book chapter [11] stresses this point. After deriving corresponding upper bounds, we can establish different proofs of consistency than those presented in these papers.

ISOMAP, another popular manifold learning algorithm, is an exception. Its solution cannot immediately be rendered as an invariant subspace. However, ISOMAP calls for MDS, which can be associated with an invariant subspace; one may derive an analytical result through this route.

2.6 Conclusion

We have derived an upper bound of the distance between two invariant subspaces that are associated with the numerical output of LTSA and an assumed intrinsic parametrization.

CHAPTER III

PERFORMANCE ANALYSIS OF HLLE

3.1 Introduction

In [6], the authors present a new nonlinear dimensionality reduction algorithm. Along with it, they present an intriguing Theorem which, intuitively, suggests that their algorithm is consistent—that is, with a sufficiently large sample, the algorithm can recover the underlying parameters up to an isometry. However, the Theorem does not actually establish this property. It is a statement about a functional in the continuum, which involves unknown quantities, while the algorithm forms a discrete estimate of this functional based on the sample data points. Thus, in order to establish rigorously the consistency of this method, several issues of convergence need to be investigated. In this chapter, we hope to fill in this theoretical gap and show that the estimated quantities used in the algorithm converge to their counterparts in the continuous manifold.

This chapter makes several contributions. First, our results give new understanding of the asymptotic properties of the HLLE algorithm. To our knowledge, this is the first time that the consistency of the algorithm has been proven. The proof also yields insight into the factors that affect the performance of the algorithm, and the implications of various geometric properties of the underlying manifold on the ability of HLLE and similar algorithms to recover the manifold structure. Second, we propose a modified estimator of the Hessian matrix and demonstrate that it results in a small improvement in performance in terms of Procrustes error [22]. If this small improvement in performance is viewed as significant, then obviously the contribution is important. If the improvement is judged to be insignificant, then we have provided stronger theoretical support for the existing methodology — i.e., in this case, HLLE performs almost as well as our modified estimator, which has many optimality properties due to the fact that it is a least-squares estimate. Finally, we provide new insight into the connections between HLLE and LTSA, another manifold learning algorithm originally proposed in [29].

The rest of the chapter is organized as follows. First, we review notation and preliminaries in Section 3.2. In Section 3.3, we derive a modified version of the estimator of the Hessian matrix, which differs from the one originally proposed in [6]. This estimator is shown to be the least-squares estimate of the Hessian. We then investigate the asymptotic properties of both of these estimators in Section 3.4, eventually showing that they yield the same asymptotic result. In Section 3.4, we also investigate the convergence of the various quantities estimated in the algorithm, which ultimately results in a proof that HLLE is consistent. We then demonstrate the improvement of our modified estimator, as well as the convergence of both the modified and original estimators, in simulations with toy examples in Section 3.5. The connections between HLLE and LTSA are explored in Section 3.6. We then conclude and summarize possible avenues for future investigation in Section 3.7.

3.2 Preliminaries

3.2.1 Problem Statement

The problem is formulated as follows: We are given a set of N observations in $\mathbb{R}^D$: $\{y_1, y_2, \ldots y_N\}$. We assume that the points lie on or near a lower-dimensional manifold $\mathcal{M}$, of dimension d $(< D)$, and that the points are sampled with respect to some continuous probability measure m on $\mathcal{M}$. We further assume that the manifold is smooth, and can be represented as a differentiable function of the parameters: $g : \mathbb{R}^d \rightarrow \mathcal{M} \subset \mathbb{R}^D$. The model can then be summarized as:

$$y_i = g(\theta_i) + \epsilon_i, \tag{11}$$

where θ_i denotes the intrinsic parameter vector of the ith observation. Our problem is to recover the set $\{\theta_i \in \mathbb{R}^d : i = 1, 2, \ldots, N\}$ such that the residuals are minimized in some sense. The standard interpretation is that we are finding the low-dimensional parameters or coordinates which "generate" the y's on the manifold. In order to make the problem identifiable, we also assume that g is locally isometric. Even with this restriction, the solution is only unique up to an isometry. This problem of ambiguity will be a recurring one throughout the chapter.

3.2.2 Review of HLLE algorithm

We now give a brief review of the HLLE algorithm itself, adapted from the recipe given in [6]. We assume that d is known. (If not, it can be estimated from the data by examining the singular value decompositions of the neighborhoods defined below, and finding a "knee" in the spectrum, in a manner analogous to the procedure commonly used in Principal Components Analysis.) The algorithm further requires one tuning parameter, k, the number of nearest neighbors from which to construct the local neighborhoods. We change the notation somewhat from that used in [6] in order to simplify the presentation of our results.

1. *Identify Neighbors.* For each point y_i, identify the $k-1$ nearest neighbors in terms of Euclidean distance in $\mathbb{R}^D$. Denote the set of indices of the nearest neighbors of y_i by $\mathcal{N}_i$, and let Y_i be the matrix formed by taking the neighbors as its rows. For each i, form a selection matrix $\widehat{S}_i$ such that $\widehat{S}_i Y = Y_i$, where Y is the matrix of all the data points. Note that $S_i \in \mathbb{R}^{k \times N}$ is formed from any permutation of the rows $\{e_{i_1}^T, e_{i_2}^T, \ldots, e_{i_k}^T\}$, where e_j is a vector of zeros with a one in the jth position, and $i_1, \ldots, i_k$ are the indices of the nearest neighbors of y_i. Finally, center Y_i by assigning $Y_i = Y_i - \overline{Y}_i$, that is, center Y_i such that each column has mean zero.

2. *Obtain Tangent Coordinates.* Perform a singular value decomposition of Y_i:

$Y_i = \widehat{U}_i \widehat{D}_i \widehat{V}_i$, where the hat notation is used to denote an estimated quantity.

3. *Develop Hessian Estimator.* Form a matrix $\widehat{X}_i$ as follows: for $d = 2$,

$$\widehat{X}_i = \begin{pmatrix} \mathbf{1} & \widehat{U}_1 & \widehat{U}_2 & \widehat{U}_1^2 & \widehat{U}_2^2 & \widehat{U}_1 \times \widehat{U}_2 \end{pmatrix}$$

i.e., in general, $\widehat{X}_i$ contains a column of ones, the original d columns of $\widehat{U}_i$, and all second order terms (squares and cross-products of all columns.) Then consider the QR factorization of $\widehat{X}_i$:

$$\widehat{X}_i = \widehat{Q}_i \widehat{R}_i$$

Now define $\widehat{H}_i$ by taking the last $\frac{d(d+1)}{2}$ columns of $\widehat{Q}_i$ and transposing.

4. *Develop Quadratic Form.* Form the matrix $\widehat{\mathcal{H}}$ as follows: define $\widehat{S} = \begin{pmatrix} \widehat{S}_1^T & \widehat{S}_2^T & \dots & \widehat{S}_N^T \end{pmatrix}$, and let

$$\widehat{\mathcal{H}} = \widehat{S} \begin{pmatrix} \widehat{H}_1^T \widehat{H}_1 & & & 0 \\ & \widehat{H}_2^T \widehat{H}_2 & & \\ & & \ddots & \\ 0 & & & \widehat{H}_N^T \widehat{H}_N \end{pmatrix} \widehat{S}^T$$

5. *Find Approximate Nullspace.* Perform an eigendecomposition of $\widehat{\mathcal{H}}$, and take the eigenvectors corresponding through the 2nd through $(d+1)$st smallest eigenvalues. These are the embedding coordinates.

3.2.3 Assumptions

In order to proceed with a perturbation-based proof, we will need conditions similar to those assumed in Chapter 2. We summarize the important ones here.

Condition 3.2.1 (Local Isometry) *The mapping g is locally isometric: For any $\varepsilon > 0$ and any x in the domain of g, let $N_\varepsilon(x) = \{y : \|y - x\|_2 < \varepsilon\}$ denote an*

ε-neighborhood of x using Euclidean distance. We have

$$\|g(x) - g(x_0)\|_2 = \|x - x_0\|_2 + o(\|x - x_0\|),$$

for any x_0 in the domain of g and $x \in N_\varepsilon(x_0)$.

Condition 3.2.2 (Local Linear Independence Condition) *Let $\overline{P}_k = I_k - \mathbf{1}_k \mathbf{1}_k^T$. For any $1 \leq i \leq N$, the rank of $Y_i \overline{P}_k$ is at least d; in other words, the dth largest singular value of $Y_i \overline{P}_k$ is greater than 0.*

Condition 3.2.3 (Regularity of the Manifold) $|[H_i(g;x)]_{jk}| \leq C_1$ *for all i, j, and k, where $C_1 > 0$ is a prescribed constant, and where $H_i(g;x)$ denotes the ordinary Hessian of the ith component function of g.*

Condition 3.2.4 (Universal Bound on the Sizes of Neighborhoods) *For all i, $1 \leq i \leq N$, we have $\tau_i < \tau$, where τ is a prescribed constant and τ_i is defined as* $\max\{\|y_j - \overline{y}_i\| : j \in \mathcal{N}_i\}$.

Condition 3.2.5 (Local Tangent Space) *There exists a constant $C_2 > 0$, such that*

$$C_2 \cdot \tau \leq d_{\min}, \tag{12}$$

where $d_{\min} \overset{def}{=} \min_{i,j} [\widehat{D}_i]_{jj}$.

3.3 Modified Hessian Estimator

3.3.1 Construction of the Estimator

For notational simplicity, in this section we will assume that i is fixed, and suppress the hat notation and i subscripts and let the singular value decomposition of Y_i be $Y_i = UDV$. Let U^* and V^* be the matrices formed by the first d columns (rows) of U (V), respectively, and let D^* be the principal submatrix of D formed from the first

d rows and columns. Consider the optimization problem

$$\begin{aligned} \min_{\substack{\Theta \in \mathbb{R}^{k \times d} \\ Z \in \mathbb{R}^{d \times D}}} \quad & \|Y_i - \Theta Z\|_F \\ s.t. \quad & ZZ^T = I_d \end{aligned}$$

It is an elementary property of the singular value decomposition that U^*D^* is the optimal Θ and V^* is the optimal Z. The columns of U^*D^* form the local tangent coordinates of the k nearest neighbors. We will form the X matrix described above using the columns of U^*D^*, rather than U^*, as used in [6]. We will discuss the consequences of this difference later.

Now, we consider the second-order Taylor expansion of a smooth function f:

$$f(x) \approx f(0) + J^T x + \frac{1}{2} x^T H x,$$

where $J \in \mathbb{R}^{d \times 1} = J_f^{(tan)}(0)$ is the Jacobian, and $H = (H_f^{(tan)}(0))_{ij}$ is the Hessian of f at m in a tangent coordinate system. See [6] for more details. Recall that the vectorization of a matrix $A = \{a_{ij}\} \in \mathbb{R}^{n \times n}$ is defined as

$$\text{Vec}(A) = (a_{11}, \ldots, a_{n1}, a_{12}, \ldots, a_{n2}, \ldots, a_{1n}, \ldots, a_{nn})^T;$$

i.e., the vector formed by stacking the columns of A. Applying this definition to H in the above equation, we have

$$\begin{aligned} f(x) &\approx f(0) + x^T J + \frac{1}{2}[\text{Vec}(xx^T)]^T \text{Vec}(H) \\ &= \left[1, x^T, \frac{1}{2}(\text{Vec}(xx^T))^T\right] \begin{bmatrix} f(0) \\ J \\ \text{Vec}(H) \end{bmatrix}. \end{aligned}$$

This is a linear regression function. Now, consider a vector $f(Y_i)$ such that $f(Y_i)_j = f(y_{i_j})$ for $j \in \mathcal{N}_i$, $1 \leq i \leq N$. Consider a model matrix X formed according to the recipe above — e.g., if $d = 2$, we have:

$$X = \left(\mathbf{1}_k \quad (U^*D^*)_1 \quad (U^*D^*)_2 \quad (U^*D^*)_1^2 \quad (U^*D^*)_2^2 \quad (U^*D^*)_1 \times (U^*D^*)_2 \right)$$

For later convenience, let X_3 denote the submatrix formed from the last $d(d+1)/2$ columns of X, and correspondingly, let H_3 denote a vector made by the $d(d+1)/2$ unique entries in the Hessian. We have

$$f(Y_i) \approx X \begin{bmatrix} 1 \\ c \\ H_3 \end{bmatrix}.$$

Suppose we apply the modified Gram-Schmidt procedure applied up to the $d+1$st column of X. We have

$$X = (Q_1, M_2) \begin{pmatrix} R_{11} & R_{12} \\ \mathbf{0}_{d(d+1)/2 \times (d+1)} & I_{d(d+1)/2} \end{pmatrix},$$

where $R_{11} \in \mathbb{R}^{(d+1)\times(d+1)}$ is upper triangular and $R_{12} \in \mathbb{R}^{(d+1)\times d(d+1)/2}$ is arbitrary. It is a basic property of the modified Gram-Schmidt procedure [7] that

$$M_2^T Q_1 = \mathbf{0}_{d(d+1)/2 \times (d+1)}. \tag{13}$$

Hence, a least-squares estimate of H_3 is given by

$$\begin{aligned} M_2^T f(Y_i) &= M_2^T X \begin{bmatrix} 1 \\ c \\ H_3 \end{bmatrix} \\ &= (\mathbf{0}_{d(d+1)/2 \times (d+1)}, M_2^T M_2) \cdot \begin{bmatrix} 1 \\ c \\ H_3 \end{bmatrix} \\ &= M_2^T M_2 H_3. \end{aligned}$$

Hence, we have $\widehat{H}_3 = (M_2^T M_2)^\dagger M_2^T z$, where $(M_2^T M_2)^\dagger$ denotes the pseudo-inverse of the symmetric matrix $M_2^T M_2$. Consequently, the matrix H_i that is defined in Step 4 of HLLE has the following form in our notation:

$$H_i = (M_2^T M_2)^\dagger M_2^T. \tag{14}$$

Note that in the original HLLE paper, it is implicitly assumed that $k \geq (d+1)(d+2)/2$, and that X has full column rank. These assumptions are not required in order to compute the estimate given in (14).

3.3.2 Differences from the Original Algorithm

We explain the differences in two steps:

D-1. In [6], the columns of U^*, rather than those of U^*D^*, are used to generate X. The resulting X matrix therefore differs from ours by a diagonal matrix $\widetilde{D}$, whose diagonal entries are the reciprocals of the squares and cross-products of the diagonal entries of D.

D-2. In [6], the QR-decomposition of X is computed, and H_i is then defined to be the transpose of the last $d(d+1)/2$ columns of Q. Our H_i also involves a factor related to the R matrix in the QR-decomposition. The end result is that our Hessian estimator matrix differs by a factor of $\widetilde{D}\widetilde{R}_{22}^{-1}$, which is not, in general, proportional to the identity matrix, and is therefore not ignorable. Details are given below.

Proof of D-1. Suppose $\widetilde{X} = (\mathbf{1}_k, U, \widetilde{X}_3)$, where $\mathbf{1}_k$ and U have been defined, and the columns of $\widetilde{X}_3$ are made by the squares and cross-products of the columns of U (instead of U^*D^* as for X_3). It is evident that $\widetilde{X}_3 = X_3\widetilde{D}$. Recall in the QR-decomposition, we have $(\mathbf{1}_k, U^*D^*) = Q_1R_{11}$; hence we have

$$(\mathbf{1}_k, U^*) = Q_1 R_{11} \begin{pmatrix} 1 & \\ & D^{-1} \end{pmatrix} = Q_1\widetilde{R}_{11}.$$

Let us assume an incomplete QR-decomposition stops at the $(d+1)$st column of $\widetilde{X}$:

$$(\mathbf{1}_k, U^*, \widetilde{X}_3) = (Q_1, \widetilde{M_2}) \begin{pmatrix} \widetilde{R}_{11} & \widetilde{R}_{12} \\ & I_{d(d+1)/2} \end{pmatrix}.$$

The above immediately leads to

$$\widetilde{X}_3 = Q_1\widetilde{R}_{12} + \widetilde{M}_2. \tag{15}$$

From the above, on one hand, we have

$$\widetilde{R}_{12} \overset{(13),(15)}{=} Q_1^T\widetilde{X}_3; \tag{16}$$

On the other hand, we have

$$\widetilde{M}_2 \overset{(15)}{=} \widetilde{X}_3 - Q_1\widetilde{R}_{12} \overset{(16)}{=} (I - Q_1Q_1^T)X_3\widetilde{D} = M_2\widetilde{D}. \tag{17}$$

In the last equality, it is not hard to verify that $M_2 = (I - Q_1Q_1^T)X_3$. One can easily see that

$$\widetilde{H}_i = \widetilde{D}^{-1}H_i,$$

as stated in D-1. In Section 3.5, we will see that $\widetilde{D}$ in general does not have common diagonal entries, and is therefore not proportional to the identity matrix.

Proof of D-2. Suppose $\widetilde{M}_2$ is of full column rank. Consider the QR-decomposition:

$$\widetilde{M}_2 = Q_2\widetilde{R}_2. \tag{18}$$

Note that Q_2^T is the Hessian estimator used in [6]. Due to (17), a QR-decomposition of M_2 gives the same Q matrix. We have

$$\begin{aligned} H_i &\overset{(14)}{=} (M_2^TM_2)^\dagger M_2^T \\ &\overset{(17),(18)}{=} ((Q_2\widetilde{R}_2\widetilde{D}^{-1})^T(Q_2\widetilde{R}_2\widetilde{D}^{-1}))^\dagger(Q_2\widetilde{R}_2\widetilde{D}^{-1})^T \\ &= \widetilde{D}\widetilde{R}_2^{-1}Q_2^T. \end{aligned}$$

Again, as we will demonstrate in Section 3.5, $\widetilde{D}\widetilde{R}_2^{-1}$ in general is not ignorable.

3.4 Perturbation Analysis

3.4.1 Perturbation of X

By Theorem 2.3.5, we have

Fact 3.4.1

$$\left\|\tan\left(\mathcal{R}(\widehat{U}_i),\mathcal{R}\left(U_i\right)\right)\right\| \leq C_3 \cdot \left(\frac{\sigma}{\tau}+\frac{C_1\cdot\tau}{2}\right), \tag{19}$$

where we define U_i as the "true" tangent coordinates of a point in $\mathcal{N}_i$, the neighborhood of the sample point y_i, i.e., the columns of U_i form an orthonormal basis of the span of $J_g^T(\overline{y}_i)$. In the language of [6], U_i would be formed from the set $\{\theta_j-\bar{\theta}_i : j\in\mathcal{N}_i\}$, where θ_j represents the underlying parameter ($\in\mathbb{R}^d$), and $\bar{\theta}_i$ is defined analogously to $\bar{Y}_i$: $\bar{\theta}_i=\mathrm{Ave}\{\theta_j : j\in\mathcal{N}_i\}$. Here $C_3\overset{\text{def}}{=}\frac{8k\sqrt{D}}{C_2^2}$, and $\sigma\overset{\text{def}}{=}\max_i\|\epsilon_i\|$.

Roughly, Fact 3.4.1 tells us that, for each i, the range spaces of $\widehat{U}$ and U are "close." However, there is still the problem of rotational (and possibly reflectional) ambiguity since $\widehat{U}$ and U may be expressed in different bases — that is, $U-\widehat{U}$ may still be far from 0 despite the fact that U and $\widehat{U}$ have nearly the same range spaces. As an extreme example, any two nonsingular matrices in $\mathbb{R}^{n\times n}$ have the same range space ($\mathbb{R}^n$), yet they obviously need not be "close" to one another in their individual entries. To resolve this difficulty, we apply a theorem related to the CS decomposition. Let Σ and Γ denote diagonal matrices formed by the sines and cosines, respectively, of the canonical angles between $\mathcal{R}(\widehat{U})$ and $\mathcal{R}(U)$. We have, by Theorem I.5.2 of [25]:

Lemma 3.4.2 *There exists an orthogonal matrix Q such that if we define*

$$W=Q^T\begin{pmatrix}\Gamma & -\Sigma & 0\\ \Sigma & \Gamma & 0\\ 0 & 0 & I\end{pmatrix}Q,$$

then:

$$W^TW=WW^T=I,$$

$$W\mathcal{R}(\widehat{U})=\mathcal{R}(U),$$

$$\|I-W\|_2=2\sin\left(\frac{\theta_1}{2}\right),$$

where θ_1 is the largest canonical angle between $\mathcal{R}(\widehat{U})$ and $\mathcal{R}(U)$.

W is called the *direct rotation* that maps $\mathcal{R}(\widehat{U})$ onto $\mathcal{R}(U)$. Notice that, by Fact 3.4.1, $\theta_1 \rightarrow 0$, and thus $W \rightarrow I$, if we assume that $\tau \rightarrow 0$ and $\frac{\sigma}{\tau} \rightarrow 0$ (cf. Conditions 2.3.4 and 2.3.7.

Now, we have that the range spaces of $W\widehat{U}$ and U are identical. Still, this by no means implies that $W\widehat{U} = U$, as discussed above. As noted in [6], however, we are only concerned with the functional $\mathcal{H}(f) = \int_{\mathcal{M}} \|H_f^{(tan)}(m)\|_F^2 dm$. Since $W\widehat{U}$ and U are both orthogonal and have the same range space, it follows that

$$W\widehat{U} = UV$$

for some orthogonal matrix V. If we define H to be the Hessian formed by using the columns of $W\widehat{U}$, and H' to be the Hessian formed by using the columns of U, then we have, directly by the definition of the Hessian,

$$H' = VHV^T$$

Though these matrices may differ in their entries, we have

$$\begin{aligned} \|H'\|_F^2 &= \|VHV^T\|_F^2 \\ &= \mathrm{tr}((VHV^T)^T(VHV^T)) \\ &= \mathrm{tr}(VH^THV^T) \\ &= \mathrm{tr}(V^TVH^TH) \\ &= \mathrm{tr}(H^TH) \\ &= \|H\|_F^2 \end{aligned}$$

The second equality is the elementary characterization of the Frobenius norm: $\|A\|_F^2 = \mathrm{tr}(A^TA)$. The fourth follows from the property of the trace operator that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. Therefore, since the HLLE algorithm only uses an estimate of $\mathcal{H}(f)$, which is a function of $\|H\|_F$, and no other property of H, we may assume WLOG that $W\widehat{U} = U$. We are now ready to state the main result of this section.

Theorem 3.4.3 *Define $\Delta X = X - \widehat{X}$. We have:*

$$\|\Delta X\| \leq C_4 \cdot ((\frac{\sigma}{\tau} + \tau) + (\frac{\sigma}{\tau} + \tau)^2) \tag{20}$$

for some constant C_4.

Proof. Consider the matrix consisting of a column of ones, each column of $\widehat{U}$, and all cross-products of columns of $\widehat{U}$, as required to form the estimated Hessian functional. We have:

$$\begin{aligned} \|\widehat{U} - U\| &= \|\widehat{U} - W\widehat{U}\| \\ &= \|(I - W)\widehat{U}\| \\ &= \|I - W\| \\ &= \left|\sin\left(\frac{\theta_1}{2}\right)\right| \\ &\leq \left|\tan\left(\frac{\theta_1}{2}\right)\right| \\ &\overset{(19)}{\leq} C_3 \cdot \left(\frac{\sigma}{\tau} + \frac{C_1 \cdot \tau}{2}\right) \end{aligned} \tag{21}$$

Now, to bound the errors of the cross-product terms, we write

$$I - W = E,$$

where $\|E\| \leq C_3 \cdot (\frac{\sigma}{\tau} + \frac{1}{2}C_1 \cdot \tau)$, as shown above. Trivially, for each entry of E, say e_{ij},

$$|e_{ij}| \leq C_3 \cdot \left(\frac{\sigma}{\tau} + \frac{1}{2}C_1 \cdot \tau\right).$$

Now, each cross-product of columns of $\widehat{U}$ will consist of k terms of the form $[\widehat{U}]_{ia} \cdot [\widehat{U}]_{ib}$, $1 \leq a, b \leq d$, where $[\widehat{U}]_{ia}$ denotes the entry in the ith row and ath column of $\widehat{U}$. By the above, we have, for the cross-product of columns a and b:

$$\begin{aligned}
&\left\|\sum_{i=1}^{k}[\widehat{U}]_{ia}\cdot[\widehat{U}]_{ib}-[W\cdot\widehat{U}]_{ia}\cdot[W\cdot\widehat{U}]_{ib}\right\| \\
\leq\ &\left\|\sum_{i=1}^{k}(|[W\cdot\widehat{U}]_{ia}|+|e_{ia}|)\cdot(|[W\cdot\widehat{U}]_{ib}|+|e_{ib}|)-|[W\cdot\widehat{U}]_{ia}|\cdot|[W\cdot\widehat{U}]_{ib}|\right\| \\
\leq\ &\left\|\sum_{i=1}^{k}|e_{ia}|\cdot|[W\cdot\widehat{U}]_{ib}|+|e_{ib}|\cdot|[W\cdot\widehat{U}]_{ia}|+|e_{ia}\cdot e_{ib}|\right\| \\
\leq\ &k(2\|E\|+\|E\|^2) \qquad (22)
\end{aligned}$$

where the last inequality follows because $W\cdot\widehat{U}$ has orthonormal columns, implying that each entry has absolute value less than or equal to 1. Putting all of this together, we have a total of $d(d+1)/2$ cross-product terms, and the original d columns of $\widehat{U}$. In our notation,

$$\Delta X = X-\widehat{X} = \begin{pmatrix} \mathbf{0} & U-\widehat{U} & U\times U-\widehat{U}\times\widehat{U} \end{pmatrix}$$

Above, we have shown that

$$\|U-\widehat{U}\| \overset{(21)}{\leq} \|E\| \qquad \text{and} \qquad \|U\times U-\widehat{U}\times\widehat{U}\| \overset{(22)}{\leq} \frac{kd(d+1)}{2}(2\|E\|+\|E\|^2)$$

Now, it is easy to see that

$$\|\Delta X\|_F = \|U-\widehat{U}\|_F + \|U\times U-\widehat{U}\times\widehat{U}\|_F$$

and (20) follows. □

3.4.2 Perturbation of Hessian Estimator Matrix

The next step is the key difference between our modified version and the original version of the algorithm. At this step, the original HLLE requires that we compute the Gram-Schmidt orthonormalization process on $\widehat{X}_i$, and then extract the last $d(d+1)/2$ columns and transpose. In the modified version, we compute the QR-factorization of

$\widehat{X}_i$, and obtain

$$\widehat{X}_i = \begin{pmatrix} \widehat{Q}_1 & \widehat{Q}_2 \end{pmatrix} \begin{pmatrix} \widehat{R}_{11} & \widehat{R}_{12} \\ 0 & \widehat{R}_{22} \end{pmatrix}$$

where $\widehat{Q}_1 \in \mathbb{R}^{k \times d+1}, \widehat{Q}_2 \in \mathbb{R}^{k \times d(d+1)/2}, \widehat{R}_{11} \in \mathbb{R}^{d+1 \times d+1}, \widehat{R}_{22} \in \mathbb{R}^{d(d+1)/2 \times d(d+1)/2}$. Using this notation, the original estimate is

$$\widehat{H}_i^{orig} = \widehat{Q}_2^T$$

while the modified one, the least-squares estimator, is

$$\widehat{H}_i^{mod} = \widetilde{D} \widehat{R}_{22}^{-1} \widehat{Q}_2^T$$

Thus, we will need a perturbation bound on the QR factorization of a matrix.

Let

$$X_i = Q_i R_i$$

denote the usual QR decomposition of X_i, and let $\kappa(X)$ denote the condition number of X, i.e.,

$$\kappa(X) = \|X\| \cdot \|X^\dagger\|.$$

In order to apply a perturbation bound on the QR decomposition of a matrix, we will need to impose the following Condition.

Condition 3.4.4 (Bound for $\kappa(\widehat{X}_i)$) *For all i, $\kappa(\widehat{X}_i) \leq C_5$ for some constant C_5.*

This condition is rather awkward, and is a manifestation of the "numerical difficulty" alluded to in [6]. Note that we cannot bound the condition number (or even guarantee full-column rank) based on the orthogonality of the $\widehat{U}$ alone - there are pathological cases in which the columns of $\widehat{U}$ are orthogonal, yet their cross-products or squares are still collinear, or arbitrarily ill-conditioned. For example, consider the following case when $k = 7$ and $d = 2$:

$$U = \begin{pmatrix} U_1 & U_2 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 0 \\ 0 & 1/\sqrt{2} \\ -1/\sqrt{2} & 0 \\ 0 & -1/\sqrt{2} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Following the given recipe for the construction of X, we get:

$$X = \begin{pmatrix} 1/\sqrt{7} & 1/\sqrt{2} & 0 & 1/2 & 0 & 0 \\ 1/\sqrt{7} & 0 & 1/\sqrt{2} & 0 & 1/2 & 0 \\ 1/\sqrt{7} & -1/\sqrt{2} & 0 & 1/2 & 0 & 0 \\ 1/\sqrt{7} & 0 & -1/\sqrt{2} & 0 & 1/2 & 0 \\ 1/\sqrt{7} & 0 & 0 & 0 & 0 & 0 \\ 1/\sqrt{7} & 0 & 0 & 0 & 0 & 0 \\ 1/\sqrt{7} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Clearly, X in this case does not have full column rank even though U does. In fact, it is even possible that $\text{rank}(X) = d + 1$, the rank of the original matrix U from which the columns of X are constructed. Of course, this is a pathological example which is unlikely to arise in practical situations, but analysis of this situation in the general case is made difficult by the nonlinear relationships between the columns and their cross-products. However, even in practice we find that sometimes the X matrices do become ill-conditioned, and this can lead to substantial distortion of the results. Obviously, this distortion only affects points which count the badly conditioned X_i as a nearest neighbor, but this can still skew the results substantially. From a theoretical point of view, we know that, with every column we add to the X matrix, the condition number must increase (see, e.g., [25, Theorem IV.4.2] for more details). Further, the number of columns in X is a quadratic function of d. Thus, we expect this problem to become more and more prevalent as d increases. We also remark here that, as d increases, the parameter k must also increase in order to ensure that X has more rows than columns – i.e., we must have $k > 1 + d + \frac{d(d+1)}{2}$, or else X cannot have full column rank. This is also an illustration of the need for increasingly large sample sizes as d grows. We will further explore the effect of d on the condition number of X in Section 3.5.

Another possible way of going about deriving perturbation bounds would be to use our alternate formulation given in Section 3.3, which involves the pseudo-inverse (which always exists), rather than the inverse $(R_i^{22})^{-1}$, the existence of which is, of course, not guaranteed. However, while our modified procedure always allows us to *compute* the least-squares estimate, it is not of much use in the analysis of the estimate under perturbation. This is because the perturbation theory for the pseudo-inverse is complicated by the fact that the pseudo-inverse behaves very badly under perturbations which change the rank of the matrix (i.e., $\|(X + E)^\dagger - X^\dagger\|$ may be arbitrarily large, even if $\|E\|$ is small, when the ranks of X and $X + E$ differ.) Thus, we essentially have the same problem if we pursue this alternative route — we would need the condition that $\text{rank}(X_i) = \text{rank}(\widehat{X}_i)$ for all i in order to bound $\|\widehat{X}^\dagger - X^\dagger\|$, which could not be guaranteed if each X_i were not of full column rank. It seems that we cannot avoid Condition 3.4.4.

Condition 3.4.4, together with Theorem 3.1 of [24], implies the following.

Theorem 3.4.5 *For each i, $1 \leq i \leq n$ there exist an upper-triangular matrix ΔR_i*

and an orthogonal matrix ΔQ_i *such that:*

$$\begin{aligned} \widehat{X}_i &= (Q_i + \Delta Q_i)(R_i + \Delta R_i), \\ (Q_i + \Delta Q_i)^T (Q_i + \Delta Q_i) &= I, \\ \|\Delta Q_i\| &\leq \frac{3\kappa(X_i)\frac{\|\Delta X_i\|}{\|X_i\|}}{1 - 2\kappa(X_i)\frac{\|\Delta X_i\|}{\|X_i\|}}, \qquad (23) \\ \|\Delta R_i\| &\leq (d+1)(d+2)(2+\sqrt{2})\kappa(R_i)\|\Delta X_i\|. \qquad (24) \end{aligned}$$

Notice that Condition 3.4.4, together with Theorem 3.4.3, guarantees that $\|\Delta Q_i\|$ and $\|\Delta R_i\|$ go to 0. Consider the difference $\|\widehat{H}_i - H_i\|$, where H_i denotes the corresponding Hessian formed using the true (unknown) local coordinates U. It is easy to see that

$$\|\widehat{Q}_2 - Q_2\| \leq \|\widehat{Q}_i - Q_i\| \qquad (25)$$

and that

$$\|\widehat{R}_{22} - R_{22}\| \leq \|\widehat{R}_i - R_i\| \qquad (26)$$

Also, by the standard perturbation theory for inverses [25], we have

$$\|(\widehat{R}_{22})^{-1} - (R_{22})^{-1}\| \leq \|(\widehat{R}_{22})^{-1}\| \cdot \|(R_{22})^{-1}(\widehat{R}_{22} - R_{22})\| \qquad (27)$$

Notice here that we have not explicitly addressed the issue of the existence of $\widehat{R}_{22}^{-1}$. However, Condition 3.4.4 guarantees that each X is of full column rank, which implies the existence of R_{22}^{-1}. Further, we have $\widehat{X}_i \to X_i$ for each i as N grows. (26) implies that $\widehat{R}_{22} \to R_{22}$. It is easy to see that, for sufficiently large N, $\widehat{R}_{22}^{-1}$ must exist, and the given bound for the difference of the inverses holds.

As a final preparation to derive a bound on $\|\Delta H\| \overset{\text{def}}{=} \|\widehat{H} - H\|$, we must investigate the behavior of the diagonal matrix $\widetilde{D}$ defined in Section 3.3.2. This discussion is deferred to the Appendix.

Now we are ready to bound $\|\Delta H_i\|$. For notational simplicity, we will drop the superscripts corresponding to the partitions of the matrices and the subscript corresponding the original index of the observation, and let $\Delta Q = \widehat{Q}_2 - Q_2$, $\Delta R^{-1} =$

$(\widehat{R}_{22})^{-1} - (R_{22})^{-1}$, and $\Delta D = \widetilde{D}_{\widetilde{Y}} - \widetilde{D}_{\widetilde{\Theta}}$. Putting together the above results, we have:

$$\begin{aligned}
\|\Delta H_i\| &= \|\widehat{H}_i - H_i\| \\
&= \|\widetilde{D}_{\widetilde{Y}} \widehat{R}^{-1} \widehat{Q} - \widetilde{D}_{\widetilde{\Theta}} R^{-1} Q\| \\
&= \|(\widetilde{D}_{\widetilde{\Theta}} + \Delta D)(R^{-1} + \Delta R^{-1})(Q + \Delta Q) - \widetilde{D}_{\widetilde{Y}} R^{-1} Q\| && (28) \\
&= \|\Delta D R^{-1} Q + \widetilde{D}_{\widetilde{\Theta}} \Delta R^{-1} Q + \widetilde{D}_{\widetilde{\Theta}} R^{-1} \Delta Q\| + \text{higher-order terms} && (29) \\
& && (30)
\end{aligned}$$

3.4.3 Perturbation of the Quadratic Form

We now consider

$$\widehat{\mathcal{H}}_N = \begin{pmatrix} \widehat{S}_1 & \widehat{S}_2 & \cdots & \widehat{S}_N \end{pmatrix} \begin{pmatrix} \widehat{H}_1^T \widehat{H}_1 & & & 0 \\ & \widehat{H}_2^T \widehat{H}_2 & & \\ & & \ddots & \\ 0 & & & \widehat{H}_N^T \widehat{H}_N \end{pmatrix} \begin{pmatrix} \widehat{S}_1^T \\ \widehat{S}_2^T \\ \vdots \\ \widehat{S}_N^T \end{pmatrix} \quad (31)$$

Recall our assumption that, asymptotically, the nearest neighbors in the feature space are the same as those in the parameter space. Thus we may assume that

$$\lim_{N \to \infty} \widehat{S}_i = S_i$$

for each i. Before we can bound the deviation of the estimated quadratic form from the true one, we will need one condition about the S_i matrices.

Condition 3.4.6 (No Overuse of One Observation) *There exists a constant C_8, such that*

$$\left\| \sum_{i=1}^{N} S_i \right\|_\infty \leq C_8 \quad (32)$$

This means that, for any N, a particular observation y_i can only appear in the nearest-neighbor set of a *bounded* number of observations. We are now prepared to bound the difference of the two quadratic forms. Using the above bounds, we have:

$$
\begin{aligned}
\|\Delta\mathcal{H}\| &\overset{\text{def}}{=} \|\widehat{\mathcal{H}}_N - \mathcal{H}\| \\
&= \left\| \sum_{i=1}^{N} S_i (H_i^T H_i - \widehat{H}_i^T \widehat{H}_i) S_i^T \right\| \\
&= \left\| \sum_{i=1}^{N} S_i (H_i^T H_i - (H_i + \Delta H_i)(H_i + \Delta H_i)^T) S_i^T \right\| \\
&= \left\| \sum_{i=1}^{N} S_i (\Delta H_i \cdot H_i^T + H_i \cdot \Delta H_i^T + \Delta H_i \cdot \Delta H_i^T) S_i^T \right\| \\
&\leq \left\| \sum_{i=1}^{N} S_i \right\|_{\infty}^{2} \cdot \|\Delta H_i \cdot H_i^T + H_i \cdot \Delta H_i^T + \Delta H_i \cdot \Delta H_i^T)\| \\
&\overset{(32)}{\leq} C_1 \cdot C_8^2 \cdot (2\|\Delta H_i\| + \|\Delta H_i\|^2) \qquad (33)
\end{aligned}
$$

where $\|\Delta H_i\|$ is bounded in (29). The key is equation (32), which bounds the size of each term in the diagonal matrix in (31).

3.4.4 Perturbation of the Nullspaces

Recall that the embedding coordinates are given by the eigenvectors corresponding to the 2nd through $d+1$st smallest eigenvalues of the above quadratic form. We must have one more condition in order to apply Theorem V.2.7 from [25]. Consider the eigenvalues of $\widehat{\mathcal{H}}_N$, arranged in increasing order:

$$\mathcal{L}(\widehat{\mathcal{H}}_N) = \lambda_1, \cdots, \lambda_d, \lambda_{d+1}, \lambda_{d+2}, \cdots, \lambda_N$$

In order to apply the standard perturbation theorem on invariant subspaces, we need the 0 eigenvalue associated with the nullspace to be well separated from all the other eigenvalues of $\mathcal{H}$. Now, we know by Theorem 1 in [6] that $\mathcal{H}$ has a $d+1$-dimensional nullspace, and therefore the first $d+1$ eigenvalues are all equal to 0. Thus, we will require that all eigenvalues beyond the first $d+1$ be bounded away from 0, which will ensure that no other (spurious) eigenvectors get mixed up with the nullspace computed in the final step of HLLE. More precisely, we will impose the following condition:

Condition 3.4.7 (Separation of Eigenvalues) *Let $\mathcal{L}(\mathcal{H})$ be the spectrum of $\mathcal{H}$ as defined above. Then, for every N, we have*

$$\lambda_{d+2} \geq \epsilon$$

for some $\epsilon > 0$ which does not depend on N.

Viewed from a matrix algebra perspective, this is a difficult problem. There seem to be few results in the literature on sequences of matrices which grow in dimension as n increases, and standard perturbation theory appears to be of little help. From the functional point of view, it can be interpreted as follows: Suppose we have a basis $\{f_1, f_2, \ldots\}$ of the space of C^2 functions on $\mathcal{M}$. Theorem 1 of [6] tells us that there is some choice of basis such that $\mathcal{H}(f_1) = \mathcal{H}(f_2) = \ldots = \mathcal{H}(f_{d+1}) = 0$, and that for $i > d+1$, $\mathcal{H}(f_i) > 0$. Condition 3.4.7, in this context, asserts that

$$\inf\{\mathcal{H}(f_i) : i > d+1\} = \epsilon > 0.$$

Theorem 3.4.8 *For sufficiently large N,*

$$\| \tan[\mathcal{N}(\mathcal{H}), \widehat{\mathcal{N}}(\mathcal{H} + \Delta\mathcal{H})]\| \leq 2 \cdot \frac{\|\Delta\mathcal{H}\|}{\epsilon - 2\|\Delta\mathcal{H}\|},$$

where we have used the notation $\widehat{\mathcal{N}}$ to denote the estimated null space consisting of the eigenvectors corresponding to the $d+1$ smallest eigenvalues.

Proof. Let the spectral decomposition of $\mathcal{H}$ be

$$\mathcal{H} = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Lambda_1 & \\ & \Lambda_2 \end{pmatrix} \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix}$$

where the columns of U_1 span $\mathcal{N}(\mathcal{H})$ and the columns of U_2 span $\mathcal{R}(\mathcal{H})$. Define

$$E = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix} = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \Delta\mathcal{H} \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix}$$

Since U is orthogonal, we have that $\|E\| = \|\Delta\mathcal{H}\|$, and by (33) we have that $\|\Delta\mathcal{H}\| \to 0$. Theorem V.2.7 in [25] yields the desired result. □

This Theorem tells us, essentially, that HLLE is consistent. Since the tangents of the canonical angles between the subspaces spanned by the true and estimated coordinates go to zero, they are "equivalent" in the sense that one is a rotation of the other. This means that, with a sufficiently large sample, HLLE can recover the structure of the underlying parameter set up to an isometry.

3.4.5 Convergence of Estimated Nullspaces

We have shown that the estimated Hessians converge to true Hessian formed when the isometric coordinates are used to form the X matrix. We now must show the asymptotic equivalence of the estimated null-space to the solutions of the equation $\int_{\mathcal{M}} \|H_f^{(tan)}(m)\|_F^2 dm = 0$. First, we will need a Lemma.

Lemma 3.4.9 *Let $f : \mathcal{M} \to \mathbb{R}$ be a C^2 function defined on the articulation manifold. For every i, $\lim_{N\to\infty} \widehat{H}_{f,N}^{mod}(\overline{y}_i) = H_f(\overline{y}_i)$*

Proof. Consider the Taylor expansion of f:

$$f(y_{i_j}) = f(\overline{y}_i) + J_f(\overline{y}_i)(y_{i_j} - \overline{y}_i) + \frac{1}{2}(y_{i_j} - \overline{y}_i)^T H_f(\overline{y}_i)(y_{i_j} - \overline{y}_i) + O(\|y_{i_j} - \overline{y}_i)\|^3)$$

Recall that we have defined τ such that

$$\|y_{i_j} - \overline{y}_i\| \leq \tau$$

whenever y_{i_j} is one of the k nearest neighbors of y_i. Thus, it follows that

$$\lim_{N\to\infty} \left\| \sum_{j=1}^{k} (f(y_{i_j}) - f(\overline{y}_i) - J_f(\overline{y}_i)(y_{i_j} - \overline{y}_i) - \frac{1}{2}(y_{i_j} - \overline{y}_i)^T H_f(\overline{y}_i)(y_{i_j} - \overline{y}_i)) \right\|^2 \leq C_6 \cdot \tau^6 \tag{34}$$

for some constant C_6. Recall the discussion on the least-squares estimation of the Hessian matrix in Section 3.3. What we have shown is that there exist a vector $J_f(\overline{y}_i)$ and a matrix $H_f(\overline{y}_i)$ such that the sum of squared errors in Equation 34 approaches zero, by our assumption that $\tau \to 0$. Hence, since the sum of squared errors is always positive, we have that $J_f(\overline{y}_i)$ and $H_f(\overline{y}_i)$ asymptotically give the minimum possible sum of squared errors. The only issue that remains is the uniqueness. It is clear from the construction of the least-squares estimator, however, that it is unique whenever the X matrix formed by y_i and its nearest neighbors has full column rank (cf. Condition 3.4.4), since the pseudo-inverse is unique for a matrix of full rank. Therefore, since $\widehat{H}_{f,N}^{mod}$ is the least-squares estimate as shown in Section 3.3, the result follows. □

Theorem 3.4.10 *Suppose the data points $\{y_1, y_2, \cdots, y_N\}$ are sampled from a continuous probability measure strictly positive everywhere in the interior of the manifold $\mathcal{M}$. Then for any C^2 function $f : \mathcal{M} \to \mathbb{R}$,*

$$\lim_{N\to\infty} \widehat{\mathcal{H}}_N^{mod} \cdot (f(y_1), f(y_2), \ldots, f(y_N))^T = 0$$

if and only if

$$\int_{\mathcal{M}} \|H_f^{(tan)}(m)\|_F^2 dm = 0$$

Proof.

1. $\Leftarrow$ Suppose $\int_{\mathcal{M}} \|H_f^{(tan)}(m)\|_F^2 dm = 0$. It follows that $H_f^{(tan)} = 0$ m-a.e., since $\|H_f^{(tan)}(m)\|_F^2$ is a nonnegative function. Therefore, for each i, $\lim_{N\to\infty}(\widehat{H}_i)^T \widehat{H}_i = 0$, which implies that $\lim_{N\to\infty} \widehat{\mathcal{H}}_N^{mod} = 0$.

2. $\Rightarrow$ Suppose $\lim_{N\to\infty} \mathcal{H}_N = 0$, and suppose that $\int_{\mathcal{M}} \|H_f^{(tan)}(m)\|_F^2 dm \neq 0$. It follows that $\|H_f^{(tan)}\| > 0$ on some ball in M, say, $B_\epsilon(z)$. To derive a contradiction, notice that we have assumed that the observations are sampled with

respect to a density that is positive everywhere in M - thus,

$$P\{x_i \in B_\epsilon(z) > 0,\, i = 1, 2, \ldots, N\}$$

and therefore,

$$\lim_{N\to\infty} P\{x_i \in (B_\epsilon(z))^C,\, i = 1, 2, \ldots, N\} = 0$$

But if some x_i is in $B_\epsilon(z)$, then we have by Lemma 3.4.9 that $\lim_{N\to\infty} \widehat{H}_i \neq 0$, and thus, $\lim_{N\to\infty} \widehat{\mathcal{H}}_N \neq 0$, contradicting the hypothesis. □

With Theorem 3.4.10, we can now link our results with Theorem 1 of [6]. Informally, this result can be viewed as stating that $\lim_{N\to\infty} \widehat{\mathcal{N}}(\widehat{\mathcal{H}}^{mod}) = \mathcal{N}(\mathcal{H})$, where $\widehat{\mathcal{N}}$ denotes the estimated nullspace formed by the eigenvectors corresponding to the 2nd through $d+1$st smallest eigenvalues.

3.4.6 Comparison between Original and Least-Squares Estimators of the Hessian

We have carried out a perturbation analysis of the least-squares version of the Hessian estimator. However, we have also shown that the actual least-squares estimator of the Hessian differs from the original, so we have still not answered the question of how the original estimator behaves asymptotically. However, the following simple result shows that the original and least-squares estimators are asymptotically equivalent.

Theorem 3.4.11 (Equivalent Nullspaces) *Let $\widehat{\mathcal{H}}$ and $\widehat{\mathcal{H}}^{mod}$ denote the original and modified versions of the estimated Hessian functional, respectively. That is:*

$$\widehat{\mathcal{H}} = \widehat{S} \begin{pmatrix} \widehat{Q}_1\widehat{Q}_1^T & & & 0 \\ & \widehat{Q}_2\widehat{Q}_2^T & & \\ & & \ddots & \\ 0 & & & \widehat{Q}_N\widehat{Q}_N^T \end{pmatrix} \widehat{S}^T$$

and

$$\widehat{\mathcal{H}}^{mod} = \widehat{S} \begin{pmatrix} \widehat{Q}_1\widehat{R}_1^{-T}\widetilde{D}_1^2\widehat{R}_1^{-1}\widehat{Q}_1^T & & & 0 \\ & \widehat{Q}\widehat{R}_2^{-T}\widetilde{D}_2^2\widehat{R}_2^{-1}\widehat{Q}^T & & \\ & & \ddots & \\ 0 & & & \widehat{Q}_N\widehat{R}_N^{-T}\widetilde{D}_N^2\widehat{R}_N^{-1}\widehat{Q}_N^T \end{pmatrix} \widehat{S}^T.$$

Then $\mathcal{N}(\widehat{\mathcal{H}}) = \mathcal{N}(\widehat{\mathcal{H}}^{mod})$.

Proof. To simplify notation, define

$$\begin{aligned} A_i &= \widehat{Q}_i^T\widehat{S}_i^T, \\ B_i &= \widehat{R}_i^{-T}\widetilde{D}_i^T\widetilde{D}_i\widehat{R}_i^{-1}. \end{aligned}$$

Suppose $x \in \mathcal{N}(\widehat{\mathcal{H}})$. By direct multiplication, we can see that this implies

$$A_i^T A_i x = 0$$

for $i = 1, 2, \ldots, N$. But then

$$\begin{aligned} & x^T A_i^T A_i x = 0 \\ \Rightarrow \quad & A_i x = 0 \\ \Rightarrow \quad & A_i^T B_i A_i x = 0 \\ \Rightarrow \quad & x \in \mathcal{N}(\widehat{\mathcal{H}}^{mod}) \end{aligned}$$

Suppose $x \in \mathcal{N}(\widehat{\mathcal{H}}^{mod})$. Then $A_i^T B_i A_i x = 0$ for $i = 1, 2, \ldots, N$. Now by hypothesis, $\widetilde{D}_i$ and R_i are nonsingular, which implies that B_i is positive definite. Recall that, for any positive definite matrix B_i, we have

$$B_i = B_i^{1/2} B_i^{1/2} \tag{35}$$

for some positive definite $B_i^{1/2}$, that is, positive definite matrices have positive definite square roots. Also recall that, by the positive-definiteness of $B_i^{1/2}$, we have that

$$B_i^{1/2} x = 0 \Leftrightarrow x = 0. \tag{36}$$

Putting together these basic results , we have, for all i:

$$\begin{aligned}
& A_i^T B_i A_i x = 0 \\
\Rightarrow \quad & x^T A_i^T B_i A_i x = 0 \\
\overset{(35)}{\Rightarrow} \quad & x^T A_i^T B_i^{1/2} B_i^{1/2} A_i x = 0 \\
\Rightarrow \quad & B_i^{1/2} A_i x = 0 \\
\overset{(36)}{\Rightarrow} \quad & A_i x = 0 \\
\Rightarrow \quad & x \in \mathcal{N}(\widehat{\mathcal{H}}).
\end{aligned}$$

□

We have shown that both the original Hessian estimator and the least-squares estimator have the same null space. Thus, although we will see in Section 3.5 that the original Hessian estimator differs from the least-squares estimator in finite samples, asymptotically we should expect the results from the two estimators to be the same. In fact, Theorem 3.4.11 shows that we can left-multiply the original estimator by *any* nonsingular matrix, and still expect the same asymptotic result to hold. The natural question to ask, then, is the magnitude of the difference between these estimators in finite samples. We investigate this, as well as questions regarding the numerical stability of the least-squares estimation procedure, in some detail in Section 3.5.

3.5 Simulations

3.5.1 Comparison of Ordinary and Least-Squares HLLE

We compare performance using the canonical S-Curve example in Figure 4. We generate 482 points in a 2-dimensional grid, which we then embed into $\mathbb{R}^3$, as shown in Figure 5, and run both the original and least-squares algorithms to attempt to recover the underlying 2D structure. To explore further the effect of multiplying Q_i by a nonsingular matrix to form each local estimate $\widehat{H}_i$, we also consider a "randomized" HLLE algorithm, in which Q_i is multiplied by a randomly generated (nonsingular)

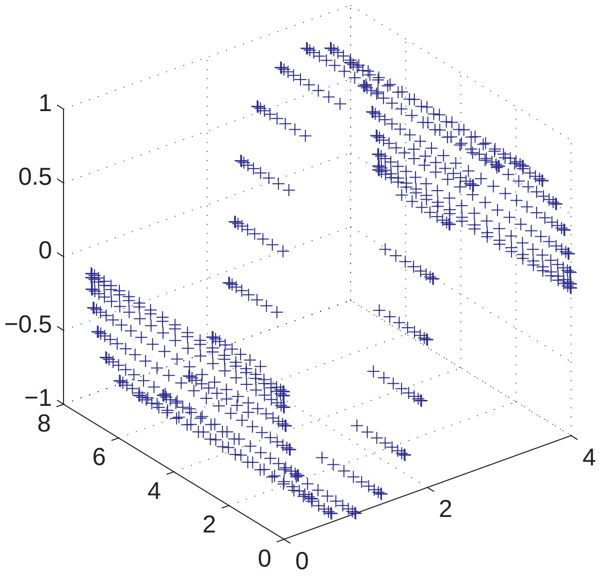


Figure 4: The 3-dimensional S-Curve data, which were used as input to the 3 different algorithms.

matrix. Notice that all 3 algorithms all appear to recover the underlying structure quite nicely — none of them suffer from any noticeable distortion. The fact that the Randomized HLLE algorithm appears to rotate the results is, of course, of no significance, since the rotated square is still isometric to the original.

As a quantitative measure of the accuracy of recovery, we use Procrustes statistics [22] to compute the optimal matching of each set of recovered coordinates to the original isometric coordinates. This is necessary because, as has been noted, the recovery of the original isometric coordinates is only possible up to an isometry (i.e., rotation and/or reflection). Also, since eigenvectors are always scaled to have norm one, the original and recovered coordinates may also have different scales. Thus, comparing the sets of coordinates directly is not a meaningful measure of the "error" in recovering the low-dimensional structure. Notice that all three recovered coordinate sets have scales which differ from that of the original data set, and the randomized HLLE and original HLLE results are oriented differently than the original coordinates. A Procrustes matching computes the optimal rescaling, rotation, and reflection of

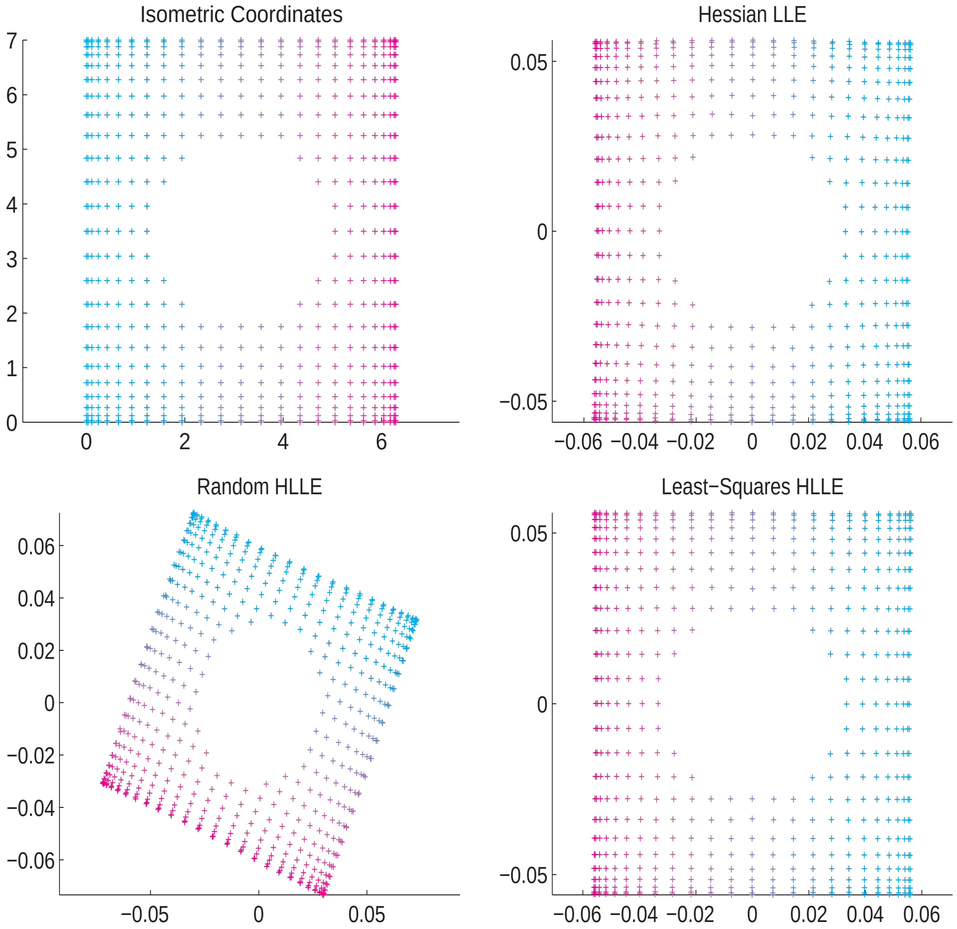


Figure 5: A comparison of the performance of the original, modified, and "randomized" HLLE algorithms. The Procrustes SSE for the 3 algorithms are .003, .002976, and .003086, respectively.

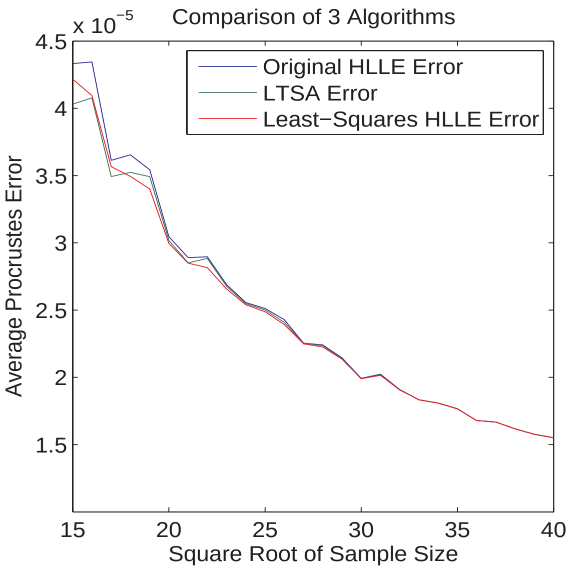


Figure 6: The results of Regular HLLE, Least-Squares HLLE, and LTSA, as a function of sample size.

the sets of points, and then computes the SSE of these matched coordinate sets. Note that the modified least-squares estimator does, in fact, have the smallest overall error (measured by the Procrustes statistic) as we might expect, given the well-known optimality properties of least-squares estimators. However, the difference is obviously quite small, suggesting that the asymptotic result we expect from Theorem 3.4.11 holds approximately even with relatively small sample sizes, as used here.

To demonstrate the actual rate of convergence, we simulate data sets of increasing size, sampled from the same manifold as used in Figure 4. We then use each of 3 different algorithms to recover the underlying 2-dimensional coordinates. The results (given in Figure 6) demonstrate our theorems — first, note that with increasing sample size, the original and least-squares HLLE seem to be equivalent. Second, the rate of convergence agrees with our result — in this case $d = 2$, so we expect the error to decrease at the rate $\frac{1}{\sqrt{N}}$, which is essentially what we observe.

3.5.2 Effect of Parameters on Performance

We now investigate the effect of d and k on $\kappa(X)$. We know from the pathological example presented in Section 3.4.2 that it is possible for X not to be of full column rank, and it is easy to see that by perturbing the entries in the pathological example, we could generate matrices which are of full column rank, but are arbitrarily badly conditioned. However, this would never occur in practice if the original data points were chosen with respect to some continuous probability measure on the manifold $\mathcal{M}$. In this experiment, we generate "random" orthogonal columns by computing the QR-factorization of a matrix whose entries are independent standard normal random variables. We then follow our recipe to compute the $\widehat{X}$ matrices for each resulting set of "random" orthonormal vectors, and then compute $\kappa(\widehat{X}_i)$. The results are given in Figures 7 and 8.

We can see that the results are essentially what we expect — increasing d, the intrinsic dimension of the manifold, dramatically increases the average condition number of the resulting X matrices, which indicates potential difficulty in the computation of $\widehat{R}_{22}^{-1}$. This result is explained by the discussion following Condition 3.4.4. On the other hand, k does not appear to have a significant effect on $\kappa(\widehat{X})$. This is also expected, since increasing k does not increase the size of $\widehat{X}^T\widehat{X}$, and there is therefore no reason to expect that $\kappa(\widehat{X})$ will increase as a result. Our results here indicate that, despite the fact that least-squares HLLE performs slightly better than ordinary HLLE in our toy examples, it still may not be practical to use the least-squares procedure if d is larger than, say, 3 due to the numerical instability inherent in the least-squares procedure. One possible solution would be to choose the Hessian estimator *adaptively* for each neighborhood — first, a maximum threshold for κ can be chosen. $\kappa(X_i)$ can then be estimated for each i, and if it is below the threshold, the least-squares estimator can be computed, and if it is above the threshold, then the ordinary estimator can be used instead, reducing the sensitivity of the estimator to numerical error.

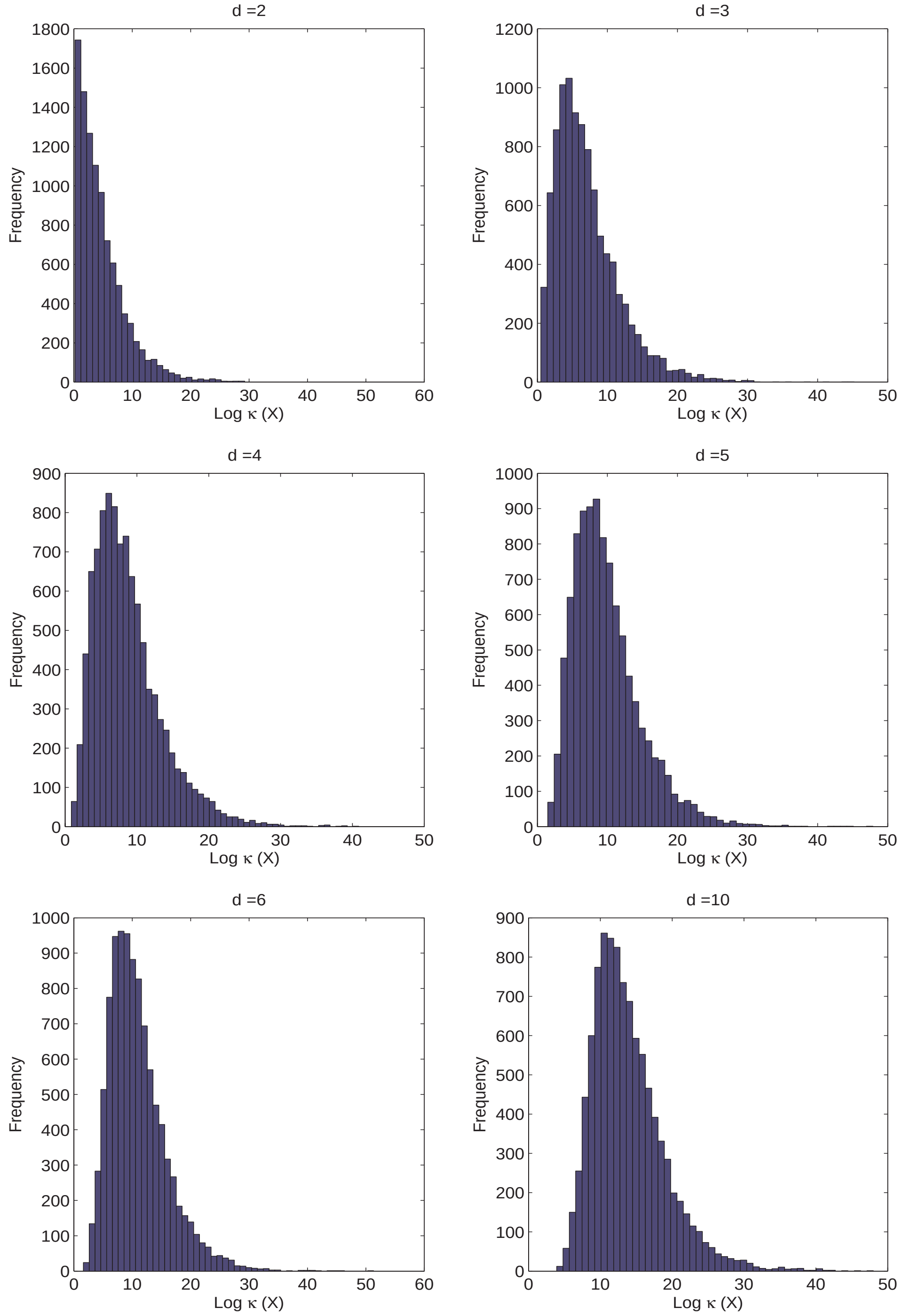


Figure 7: The effect of the parameter d on $\kappa(X)$. For these histograms we fixed $k = 100$, and varied d.

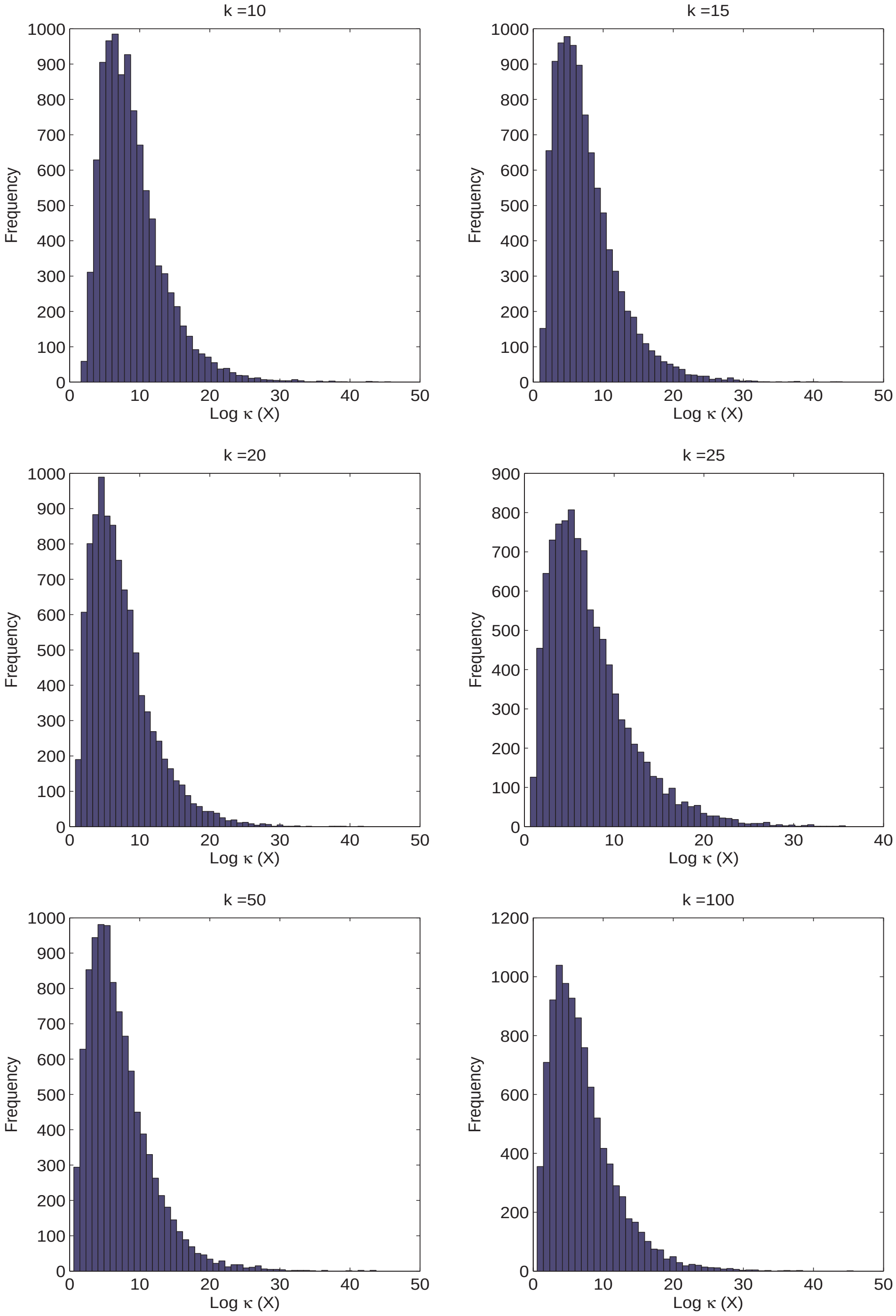


Figure 8: The effect of the parameter k on $\kappa(X)$. For these histograms we fixed $d = 3$, and varied k.

3.6 Comparison of LTSA and HLLE

Since we have now established the consistency of both LTSA and HLLE as long as certain regularity conditions are imposed on the underlying manifold, it seems natural to wonder if the two algorithms are, in some sense, equivalent. In this Section, we investigate this question and show that there is, in fact, a strong similarity between the two, despite the considerable differences in their actual implementations. Consider again the Taylor expansion of a C^2 function $f : \mathcal{M} \to \mathbb{R}$ at a sample point:

$$f(y_{i_j}) = f(\overline{y}_i) + J_f(\overline{y}_i)(y_{i_j} - \overline{y}_i) + \frac{1}{2}(y_{i_j} - \overline{y}_i)^T H_f(\overline{y}_i)(y_{i_j} - \overline{y}_i) + O(\|y_{i_j} - \overline{y}_i\|^3) \quad (37)$$

Now, recall that LTSA, in its second step, finds the null space of the matrix (see [29] and Chapter 2 for details)

$$\widehat{S}\overline{P}_k \begin{pmatrix} I - \widehat{U}_1\widehat{U}_1^T & & & 0 \\ & I - \widehat{U}_2\widehat{U}_2^T & & \\ & & \ddots & \\ 0 & & & I - \widehat{U}_N\widehat{U}_N^T \end{pmatrix} \overline{P}_k\widehat{S}^T$$

while HLLE finds the null space of

$$\widehat{S} \begin{pmatrix} \widehat{Q}_1\widehat{Q}_1^T & & & 0 \\ & \widehat{Q}_2\widehat{Q}_2^T & & \\ & & \ddots & \\ 0 & & & \widehat{Q}_N\widehat{Q}_N^T \end{pmatrix} S^T$$

The connection between the two may be viewed as follows: First, recall the Theorem proven in [6]: The functional

$$\mathcal{H}(f) = \int_{\mathcal{M}} \|H_f(m)\|_F^2 dm \quad (38)$$

has a $d + 1$-dimensional nullspace, consisting of the constant function and the d isometric coordinate functions. LTSA and HLLE are both ways of finding functions which are well-approximated by their estimated derivatives given by the optimal d-dimensional approximation over each neighborhood — the difference is only in the way they define “well-approximated.”

In view of Fact 3.4.1, we may assume that

$$U_i U_i^T \approx J_g(\overline{y}_i) J_g^T(\overline{y}_i) \tag{39}$$

by noting that these two matrices are projections, and therefore are functions only of the column space of the two matrices, and are invariant with respect to the bases chosen. To see this, suppose that the columns of two matrices, say A and $B \in \mathbb{R}^{m \times n}$, form orthonormal bases of the same subspace. Then we have $A = BV$ for some orthogonal $V \in \mathbb{R}^{n \times n}$. Then, by definition, the projection onto the column space of A is given by $AA^T = BV(BV)^T = BVV^TB^T = BB^T$, which is the projector onto the column space of B.

Suppose that $f_k : \mathcal{M} \to \mathbb{R}, k = 1, 2, \ldots, d$ are the global coordinate functions of the data points, and let $F = (f_1, f_2, \ldots f_d)^T$. Then, because the f_k are functions of θ, the underlying parameters, it is easy to see that $\mathcal{R}(J_F^T(\overline{y}_i)) \subset \mathcal{R}(J_g(\overline{y}_i))$. On the other hand, $J_F^T(\overline{y}_i)$ clearly has rank d, as does $J_g(\overline{y}_i)$. Thus, $\mathcal{R}(J_F^T(\overline{y}_i)) = \mathcal{R}(J_g(\overline{y}_i))$. Further, F must be a locally *linear* function of θ. Therefore, we have

$$\begin{aligned} \|(I - J_F^T(\overline{y}_i)J_F(\overline{y}_i))g(F(Y_i))\| &= \|(I - J_g(\overline{y}_i)J_g^T(\overline{y}_i))g(F(Y_i))\| \\ &\approx \|(I - \widehat{U}_i\widehat{U}^T)g(F(Y_i))\| \\ &= O(\tau^2) \end{aligned} \tag{40}$$

where $f(Y_i) \overset{\text{def}}{=} (f(y_{i_1}), f(y_{i_2}), \ldots f(y_{i_n}))^T$. Thus, $(I - \widehat{U}_i\widehat{U}_i^T)f(Y_i)$ may be viewed as the approximate error of the first-order Taylor expansion of f, using $\widehat{U}_i$ as an approximation to $J_g^T(\overline{y}_i)$.

Meanwhile, HLLE minimizes $\widehat{Q}_i\widehat{Q}_i^T$, which, as we have seen, is a (somewhat crude) estimate of $\|H_f(\overline{y}_i)\|_F^2$. To see the connection with (37), notice that, if we set $\tau_{min} \overset{\text{def}}{=}$

$\min_{i,j}\{\|y_{i_j} - \overline{y}_i\|\}$, we have

$$\tau_{min}^2 \|H_f(\overline{y}_i)\| \leq \|(y_{i_j} - \overline{y}_i)^T H_f(\overline{y}_i)(y_{i_j} - \overline{y}_i)\|_F \leq \tau^2 \|H_f(\overline{y}_i)\| \tag{41}$$

so HLLE seeks functions which minimize the second term in (37). In this sense, LTSA may be interpreted as seeking d orthogonal scalar functions which minimize

$$\sum_{l=1}^{d}\sum_{i=1}^{N}\sum_{j=1}^{k} \|f_l(y_{i_j}) - (f_l(\overline{y}_i) + J_{f_l}(\overline{y}_i)(y_{i_j} - \overline{y}_i))\|_2^2$$

while HLLE seeks d orthogonal scalar functions which minimize

$$\sum_{l=1}^{d}\sum_{i=1}^{N}\sum_{j=1}^{k} \|(y_{i_j} - \overline{y}_i)^T H_{f_l}(\overline{y}_i)(y_{i_j} - \overline{y}_i)\|_F^2$$

(see [29] for more details on the interpretation of LTSA as an optimization problem.) Essentially, then, recalling (37) and (41), we see that the two are just different ways of exploiting a Taylor expansion by assuming that the observations are smooth functions of the underlying parameters. The difference is simply that LTSA seeks functions for which the first-order Taylor approximation is most accurate, while HLLE seeks to minimize the second term in the Taylor expansion. Asymptotically, of course, these are equivalent since the second term dominates the remainder as $\tau \to 0$. We expect, therefore, that the difference between the results of the two algorithms, after allowing for a possible rotation and reflection, is of $O(\tau^3)$. Notice that this is consistent with Figure 6, in which we observe that the Procrustes error for both LTSA and HLLE converge to zero, and seem to do so at nearly the same rate, since the error curves coincide almost exactly for larger values of N. However, the above explanation leads us to expect that the two might differ more substantially if the underlying manifold has large third- and higher-order derivatives at least at some points.

It is also illuminating to view the connection between the two from a matrix algebra perspective. It seems natural to view the diagonal blocks in the alignment matrix used in HLLE as a quadratic form of the estimated Hessian, but they can also be viewed as projections, just like the blocks in the alignment matrix of LTSA. Let us consider again the QR decomposition of X, as constructed in HLLE:

$$X = \left(\begin{array}{cc} Q_1 & Q_2 \end{array} \right) R,$$

where $Q_1 \in \mathbb{R}^{k\times(d+1)}$ consists of a (normalized) column of ones and the original d local coordinate functions. HLLE then takes $Q = Q_2$. The ith diagonal term of the alignment matrix is therefore the projection onto the column space of the $\frac{d(d+1)}{2}$ columns representing the second-order terms. But this can be regarded as a subspace of the orthogonal complement of Q_1 — since Q_1 has k columns, the dimension of its orthogonal complement is $k-(d+1)$, while HLLE projects only onto a $\frac{d(d+1)}{2}$-dimensional subspace. It is easy to see that $\mathcal{R}(Q_2)$ is a subspace of $Q_1^\perp$ since the Gram-Schmidt procedure (or, equivalently, the QR factorization) ensures that all columns of Q are orthogonal. Thus, LTSA projects directly onto $Q_1^\perp$, while HLLE explicitly constructs a $\frac{d(d+1)}{2}$ dimensional subspace of $Q_1^\perp$ and projects onto this subspace. Supposing k is large enough, we can regard the extra columns that are ignored by HLLE (i.e., those in $Q_1^\perp \cap Q_2^\perp$) as estimates of the third- and higher-order terms in the Taylor expansion.

What, then, should we make of the differences between the two algorithms? From a computational perspective, LTSA is the clear winner. It only requires the computation of the pseudo-inverse of the left-singular vectors of each local singular value decomposition, and leads to a sparse eigenvalue problem. HLLE, on the other hand, requires the comparatively difficult computation of both the second-order matrix of cross products and its QR-factorization at every neighborhood. In practice, LTSA is far faster (in our particular simulations, about an order of magnitude faster). From a purely statistical perspective, however, there is no clear winner. The importance of the higher-order terms in the Taylor expansion seems to be specific to each particular application. If we have reason to suspect that higher-order derivatives may be large, then LTSA may offer a significant improvement. However, if we anticipate that the data may be explained by a simple curve (in particular, if we suspect *a priori* that the underlying manifold may be represented as a function with only first- and second-order terms in the parameters), then LTSA may be more sensitive to noise in the data, while HLLE would be relatively more stable. We suspect, therefore, that neither algorithm will strictly dominate the other in terms of performance — the choice of which algorithm is preferable will depend on the particular problem under consideration.

3.7 Conclusions and Future Research

We have established the asymptotic consistency of the original and least-squares versions of the HLLE algorithm. However, our understanding of these algorithms and their performance is still far from perfect. The following is a partial list of questions that we have not addressed.

1. Analysis of HLLE's performance in the finite-sample case is an interesting (and seemingly very challenging) problem. The key difficulty is that we have no guarantee that the S_i matrices will coincide with the "true" selection matrices in the underlying parameter space. This problem may be especially prevalent if the manifold is nearly self-intersecting. The consequences of this disparity are not currently well-understood. Though simulation suggests that the estimated S_i's converge fairly quickly to their true counterparts, rigorous analysis of this situation has been elusive.

2. The issue of bounding the condition numbers of the $\widehat{X}$ matrices seems to be unique to this particular algorithm in dimension reduction (e.g., it is not an issue in LTSA). It would be interesting to know under what conditions we can guarantee that each $\widehat{X}_i$ will at least be nonsingular, and from a numerical perspective, it would be useful to know if we can somehow bound $\kappa(\widehat{X}_i)$. In particular, having singular (or close to singular) X matrices could affect both

performance and numerical stability in practice, and as N increases, it seems possible that at least some of the $\widehat{X}_i$'s will be ill-conditioned.

3. Condition 3.4.7 also seems difficult to verify. Conditions under which it always holds, as well as the effects on performance if it is violated, would be useful contributions.

CHAPTER IV

COMBINED MODEL SELECTION CRITERIA

4.1 Introduction

Model selection remains, despite the considerable progress that has been made in this area, a fundamental and very challenging problem in virtually every area of applied statistics. A multitude of model selection criteria have been proposed in the literature to address this problem. Broadly, one can think of these criteria as being of three different types: some measure in-sample fit only, possibly with a penalty imposed on the number of parameters in the model. Many popular criteria, such as AIC, BIC, and Mallows' Cp fall into this class. Others use cross-validation to estimate out-of-sample prediction performance. The most common example from this class is PRESS in linear regression. Finally, others use a true holdout sample, not used in the construction of the model itself, as a measure of out-of-sample prediction error. One example of this type is MAD, or median absolute deviation, sometimes used in time-series analysis.

While all of the above types of criteria can certainly be useful, a natural question to ask is whether one can formulate a criterion which simultaneously considers more than one of these types of goodness of fit. No commonly-used criterion, for example, considers both in-sample fit and cross-validation error. In this chapter we propose a new procedure which generates a new class of combined model selection criteria. This procedure allows the analyst to combine, for example, the benefits of good in-sample fit as measured by criteria such as AIC or BIC, and also good out-of-sample prediction performance into a single criterion.

This new procedure generates a very large (in fact, infinite) new class of criteria, and therefore the question of how to compare different criteria is very important. In the model selection literature, criteria are typically compared on the basis of the proportion of simulated data sets in which the criterion chooses the known correct model from a set of candidates. We propose a generalization of this procedure based on ranks of criterion values which, we argue, is a more realistic measure of a criterion's usefulness in an applied context. The traditional method of comparing criteria turns out to be a special case of our more general comparison methodology.

Combining these two contributions, then, our main result is an algorithm which, as we show, can be proven to find the optimal combination of a fixed set of criteria, either using the traditional definition of optimality as above or a more general definition which we discuss below. Since the straightforward use of a single criterion is a special case of our combined criteria, our algorithm is a true generalization of the traditional model selection procedure in the sense that the optimal combined criterion can be no worse than any of the original criteria.

The rest of this chapter is organized as follows: In section 4.2 we propose our method to combine existing selection criteria via a simple ranking procedure. Section 4.3 then discusses a generalization of the traditional method of comparing criteria, and presents an algorithm to select an optimal combined criterion. In section 4.4 we present simulation results from our algorithm, focusing on the two special cases of ARIMA and linear regression models. We discuss the theory behind our algorithm, and in particular prove that the algorithm can find the optimal combined criterion, in section 4.5. Computational details are presented in section 4.6. Section 4.7 presents a discussion of inferential issues involved in our algorithm, and proves the ϵ-optimality of the solutions produced. In section 4.8 we discuss the role that prior distributions play in our algorithm. Finally, in section 4.9 we conclude and present possible topics for future research in this area.

4.2 Combining Model Selection Criteria

We will assume throughout that we are working with a fixed set of candidate models

$$\mathcal{M} = \{M_1, M_2, \ldots, M_s\},$$

where s denotes the cardinality of $\mathcal{M}$.

Let X denote the matrix of covariates, which we assume to be fixed, and let the response be denoted by $\mathcal{M}$. Associated with each model $M_i \in \mathcal{M}$, we assume that there is an equation

$$\mathbf{y} = g_i(\mathbf{X}, \boldsymbol{\theta}(M_i)) + \boldsymbol{\epsilon},$$

where $\boldsymbol{\theta}(M_i)$ denotes the parameter space associated with M_i.

If we fix a particular model selection criterion, say BIC for the sake of example, then the model selection problem becomes an optimization problem of the form

$$\min_{\{i=1,2,\ldots,s\}} BIC(M_i).$$

If we view the optimization from this perspective, combining several MSCs is problematic due to their different scales. For example, considering the sum

$$BIC(M_i) + MAD(M_i)$$

is meaningless because BIC is computed based on log-likelihoods, while MAD is on the same absolute scale as the original observations. Thus, even if one allows a linear combination of BIC and MAD, choosing appropriate constant multipliers to make the combination meaningful is a difficult task. The chief problem with this simple approach, then, is *scaling.*

However, a simple modification to the above procedure can make the linear combination meaningful. Rather than viewing the original problem as minimizing the absolute BIC value, we can view it as minimizing the *rank* of the BIC value of M_i among the set of BIC values of all the models in $\mathcal{M}$. Considering the rank of a model's MSC value rather than the absolute MSC value itself has the advantage of automatically putting all MSCs on the same scale. If we have several MSCs, say $MSC_1, MSC_2, \ldots, MSC_k$, we can then form a meaningful combination of them

$$MSC_{\boldsymbol{\alpha}}(M_i) = \alpha_1 \cdot R_{MSC_1}(M_i) + \alpha_2 \cdot R_{MSC_2}(M_i) \ldots + \alpha_k \cdot R_{MSC_k}(M_i)$$

for any vector of convex coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)^T$ and where $R_{MSC_j}(M_i)$ denotes the rank of $MSC_j(M_i)$ among the set $\{MSC_j(M_1), MSC_j(M_2), \ldots, MSC_j(M_s)\}$. Note that the values of $MSC_{(1,0,0)}$ are not the same as those of the original criterion MSC_1 due to the rank transform, though the ordering of the values is the same. Notice that it is not necessary to consider linear combinations with positive weights other than convex combinations, since any linear combination with positive weights amounts to a re-scaling of a convex combination — that is, it is simply a convex combination multiplied by a constant. It is easy to see that multiplying the objective function by a constant does not change the optimal solution, and therefore the scaling of the set of linear coefficients is irrelevant. Linear combinations involving negative weights are possible in principle, but are quite unintuitive since we expect all model selection criteria to be of some value in distinguishing the true model from some of the false ones. This intuition is confirmed by our simulations below. Notice that this way of resolving the scaling problem noted above crucially depends on the specification of the set of candidate models — in effect we are using the criterion values of all the models in $\mathcal{M}$ as a way of imposing a scale. As a preview of the potential effectiveness of this method, we present a simple artificial example in Table 1 below. We present many more simulation results in Section 4.4.

However, the above method of considering ranks is not the only way to make different criteria have similar scale. Another simple alternative is to *standardize* the values of each criterion. This yields another class of combined criteria, which are very much related to, but not identical to, those produced by ranking. As we show below, however, these criteria tend to be less effective than those produced by ranking. Why

Table 1: Example of the potential utility of combined criteria: No individual criterion selects the correct model (Model 2), but the sum of ranks does.

		AIC	BIC	PRESS	sum
	Model 1	1	3	4	8
(*)	Model 2	2	2	2	6
	Model 3	3	4	1	8
	Model 4	4	1	3	8

this is the case remains an open question. We expect it is due to the distributions of the criterion values themselves – standardizing the values forces a common mean and variance, but does *not* guarantee any particular type of distribution, while ranking always produces values with exactly the same distribution. In the sequel we will focus primarily on the approach employing ranks, but we give some simulation results for the standardizing method in section 4.4.

Armed with these new methods of constructing combined MSCs, we are now faced with the problem of finding the optimal convex combination of existing MSCs. In order to do so, one must specify what exactly "optimal" means with respect to an MSC. We address this question in the next section, and then in section 4.4 we propose an algorithm for computing the optimal criterion.

4.3 Comparing MSCs

In the model selection literature, criteria are often compared in a similar way – see, e.g., [19], [8]. Typically, for a fixed candidate model set $\mathcal{M}$, some data sets are simulated and all models in $\mathcal{M}$ are fitted to the resulting data. Different criteria are then compared on the basis of how often each criterion chooses the correct model – that is, how often the model with the optimal criterion value is in fact the true model chosen in the simulation. In this section, we propose a much more general framework in which to compare criteria, which we argue can be more useful in applied situations.

In order to determine what makes one criterion better than another, one should consider the applied context in which criteria are used. An idealistic approach to model selection may consist of a procedure such as the following: Choose a model selection criterion (say, BIC), fit all the feasible models in $\mathcal{M}$ to the data, and calculate the BIC for each model. Then select the model with the minimum BIC. The underlying assumption, of course, is that if the model selection criterion is well-designed, then the true model ought to be the one selected by the criterion. If one adopts this approach, then the existing method of comparison described above seems quite natural.

In practice, however, this approach is rather naive, as it amounts to the analyst allowing the criterion to completely *dictate* model choice, rather than to guide it. An experienced analyst would always use other considerations such as diagnostic plots along with criterion values in order to construct a sensible model. A more realistic procedure would be to consider several models with small BIC, and then use other non-quantitative information, such as residual plots, to choose among these top few candidates. If one uses this procedure, however, the comparison method above becomes less meaningful – we are not particularly interested in the probability that our criterion chooses the true model, but rather the probability that the criterion includes the true model in the top few choices. It is this general model selection strategy which motivates our method to compare criteria.

The key new idea to our approach is to consider ranks. To illustrate, we need a bit more notation. Suppose we have fixed k "basis" MSCs and the candidate model set $\mathcal{M}$. With a vector $\boldsymbol{\alpha}$ of convex coefficients (i.e., $\boldsymbol{\alpha} \geq 0, \sum_{i=0}^{k} \boldsymbol{\alpha}_i = 1$), let us define

$$R_{\boldsymbol{\alpha}}(M_i) = \text{rank}_{MSC_{\boldsymbol{\alpha}}}(M_i),$$

where the rank is taken with respect to the set $\{MSC_{\boldsymbol{\alpha}}(M_1), MSC_{\boldsymbol{\alpha}}(M_2), \ldots, MSC_{\boldsymbol{\alpha}}(M_s)\}$. With this notation, the traditional approach to comparing criteria described above corresponds to calculating $P\{R_{\boldsymbol{\alpha}}(M^*) = 1\}$, where M^* is the true model, for different values of $\boldsymbol{\alpha}$. The best criterion would then be the one with highest such value.

However, defining the ranks as above allow us to compare criteria in a much more general way. Rather than considering only the binary outcome $R_{\boldsymbol{\alpha}}(M^*) = 1$, we can regard $R_{\boldsymbol{\alpha}}(M^*)$ as a random variable and consider any functional of its distribution. Indeed, there are many other functionals besides $P\{R_{\boldsymbol{\alpha}}(M^*) = 1\}$ which may be meaningful. As suggested above, one useful alternative would be $P\{R_{\boldsymbol{\alpha}}(M^*) > c\}$, where c is some specified constant. It is natural to think of c as a maximum number of models the analyst is willing to consider "by hand" – that is, the number of candidates suggested by the criterion from which the analyst is willing to choose based on information other than the criterion itself. Of course, c might vary depending on the type of problem. For example, model selection in time series is notoriously difficult using diagnostic plots – often one has considerable trouble distinguishing between a simple AR(1) and an MA(1) process based on plots of the autocovariance function. For other types of models such as linear regression, there are more diagnostics at our disposal, and c may correspondingly be larger in the hopes that our chance of finding the true model will accordingly be better. Other functionals such as the mean and median can also be used.

In general, we can define an arbitrary functional T of the empirical distribution of $R_{\boldsymbol{\alpha}}$, and formulate our optimization problem as

$$\min_{\boldsymbol{\alpha}} T(\hat{F}(R_{\boldsymbol{\alpha}}(M^*))).$$

We have implicitly assumed that $R_{\boldsymbol{\alpha}}(M^*)$ may be treated as a random variable. This is most naturally interpreted in a Bayesian context in which we assume a full probability model for the data. Such a framework would consist of

1. $\pi_{\mathcal{M}}$, a prior distribution on the set of candidate models

2. $\pi_{\boldsymbol{\Theta}_{\mathcal{M}}}$, a prior on the parameter space associated with each model in $\mathcal{M}$

3. $f(\boldsymbol{\epsilon})$, an assumed distribution on the errors of the model

With all of these ingredients, we can now formulate our main algorithm, listed as

Algorithm 1.

```
Data: $\pi_{\mathcal{M}}$, $\pi_{\mathbf{\Theta}_{\mathcal{M}}}$, $f(\boldsymbol{\epsilon})$, $\mathcal{M}$, $\mathbf{X}$, sample size $N$
Result: A vector of ranks of the true model
for $i = 1 : N$ do
    Choose $j$, the index of the true model, from $\pi_{\mathcal{M}}$;
    Choose $\boldsymbol{\theta}(M_j)$ from $\pi_{\mathbf{\Theta}(M_j)}$;
    Simulate errors from $f(\boldsymbol{\epsilon})$;
    Set $\mathbf{y} = g_j(\mathbf{X}, \boldsymbol{\theta}(M_j)) + \boldsymbol{\epsilon}$;
    Fit all models in $\mathcal{M}$ to $\mathbf{y}$;
    Compute the matrix $\mathbf{B}$ of ranked basis criterion values;
    For each $\boldsymbol{\alpha}$, compute the rank of the $j$th element of the vector $\mathbf{B}\boldsymbol{\alpha}$.
end
```

Algorithm 1: Optimal combination of model selection criteria

4.4 Results

It is a well-known fact that BIC is the only *consistent* model selection criterion – that is, as the sample size grows, BIC is guaranteed to choose the correct model. Thus, by considering combinations of BIC with other criteria, we are in effect asking to what extent this asymptotic result holds in particular finite samples. In this section we explore this question in two applied contexts – regression models and ARIMA models.

4.4.1 Regression Models

Variable selection in linear regression is perhaps the oldest and most-studied model selection problem in statistics. Here we give a few examples of our algorithm applied to linear regression problems.

For the interesting case, the covariate matrix was generated as 4 independent standard normals, each of length 40, which was then right-multiplied by another random matrix to induce correlation among the predictors. The correlation matrix is given in Table 2. $\mathcal{M}$ was simply the set of all 2^4 subsets of predictors from the matrix. The prior $\pi_{\mathcal{M}}$ was specified indirectly by giving randomly assigned weights to each

Model	BIC	Cp	Adj.Rsq
y ∼ 1	264.049	3.072	0.000
y ∼ X1	268.527	4.995	0.010
y ∼ X2	268.413	4.880	0.009
y ∼ X3	265.864	2.341	−0.018
(*) y ∼ X4	265.399	1.886	−0.023
y ∼ X1 + X2	272.637	6.548	0.016
y ∼ X1 + X3	267.835	1.838	−0.034
y ∼ X1 + X4	269.247	3.199	−0.019
y ∼ X2 + X3	268.816	2.781	−0.024
y ∼ X2 + X4	269.694	3.633	−0.015
y ∼ X3 + X4	269.891	3.825	−0.013
y ∼ X1 + X2 + X3	271.722	3.203	−0.030
y ∼ X1 + X2 + X4	273.800	5.198	−0.009
y ∼ X1 + X3 + X4	272.358	3.808	−0.024
y ∼ X2 + X3 + X4	273.334	4.746	−0.014
y ∼ X1 + X2 + X3 + X4	276.062	5.000	−0.022

BIC	Cp	Adj.Rsq		$\boldsymbol{\alpha} = (0.78, 0, 0.22)$		Ranked combined values
1	5	13		3.64		3
6	13	15		7.98		8
5	12	14		6.98		6
3	3	8		4.10		4
(*) 2	2	5		2.66		1
13	16	16		13.66		14
4	1	1		3.34		2
8	6	7	→	7.78	→	7
7	4	3		6.12		5
9	8	9		9.00		9
10	10	11		10.22		11
11	7	2		9.02		10
15	15	12		14.34		16
12	9	4		10.24		12
14	11	10		13.12		13
16	14	6		13.80		15

Figure 9: An illustration of Algorithm 1. The top panel shows the raw values for each of the 16 models in $\mathcal{M}$. The bottom panel illustrates the sequence of transformations – rank by column, combine the columns using convex coefficients, re-rank the resulting values.

Table 2: Correlation matrix of predictor variables used to generate Figure.

	X1	X2	X3	X4
X1	1.000	0.504	0.774	−0.545
X2	0.504	1.000	0.421	−0.036
X3	0.774	0.421	1.000	−0.854
X4	−0.545	−0.036	−0.854	1.000

model size, and the weight for each model size was split equally among all models of that size. $\pi_{\boldsymbol{\Theta}(\mathcal{M})}$ and $f(\boldsymbol{\epsilon})$ were both standard normal distributions. The grid was generated using a width of 0.02, and the basis MSCs were BIC, Mallows Cp, and adjusted R^2. We did 2 runs of 2000 simulated data sets each: on the first run, we used combined ranked criterion values, and in the second we used combined standardized criterion values. The results are displayed in Figure 10. Note that the improvement observed by allowing combinations of these MSCs is actually unexpected, since all 3 criteria are based only on penalized in-sample fit. Nevertheless, we find the results quite interesting due to the substantial improvement of the combined criteria. Also interesting is the high sensitivity of the response surface near the optimum. Indeed, the optimum is quite close to the BIC corner, but the objective value varies greatly in this small region.

It is also instructive to compare the results of the ranked values versus the standardized values. In general, the two response surfaces look very similar, as one might expect, due to the very high correlation between the raw values and the ranked values of random vectors in general. Furthermore, the optimal points are very close to each other on the two surfaces. The primary difference is the moderately worse performance of the standardized values when considering the functional $P(R_{\boldsymbol{\alpha}}(M^*) > 1)$, noting that the optimal value using standardized values is nearly 76%, while for ranked values it is under 74%. We saw this phenomenon in several similar experiments (results not shown here,) and expect that it is true in general. For this reason, we focus only on ranked values in the sequel.

4.4.2 ARIMA Models

Some example simulation results are given in Figures 11 and 12. Figure 11 is fairly typical – in particular, the optimal point for each summary function is either at or near one of the corners, indicating that combined criteria are of little use in this case. Further, the different summary functions all behave similarly. Figure 12, on the other hand, has a more interesting structure. Especially noteworthy is the fact that the optimum point for the function $P\{R_{\boldsymbol{\alpha}}(M^*) > 1\}$ is not near any of the 3 corners, indicating that the convex combinations of criteria do better than any individual criterion in this case. Further, the improvement is quite substantial – over 5%. This example also illustrates the importance of considering which functional to consider, since the optimal criterion varies considerably depending on which function is chosen.

Interestingly, the only difference in the parameters used to generate Figures 11 and 12 is the specification of $\pi_{\mathcal{M}}$. Figure 11 used a uniform prior over all models, and Figure 12 used a prior giving weight only to models of size 3. Both figures were based on 500 simulated data sets of size 100, and $\mathcal{M}$ was defined for both to be all ARIMA models up to order $(2, 1, 2)$. In both cases, all parameters were generated as independent $U[-1, 1]$ random variables, subject to the restriction that the resulting model be stationary, and the errors were independent standard normals.

We caution against drawing general conclusions based on the example results shown here. The behavior of the response function depends on a variety of other factors, such as the structure of the matrix of regressors, and the distribution of the error terms, in an extremely complex way which has not been fully explored. The results shown here are intended only to serve as examples, and should not be used to infer the relative merits of particular criteria or combinations of criteria for any specific problem. Indeed, we feel that one of the biggest advantages of our procedure

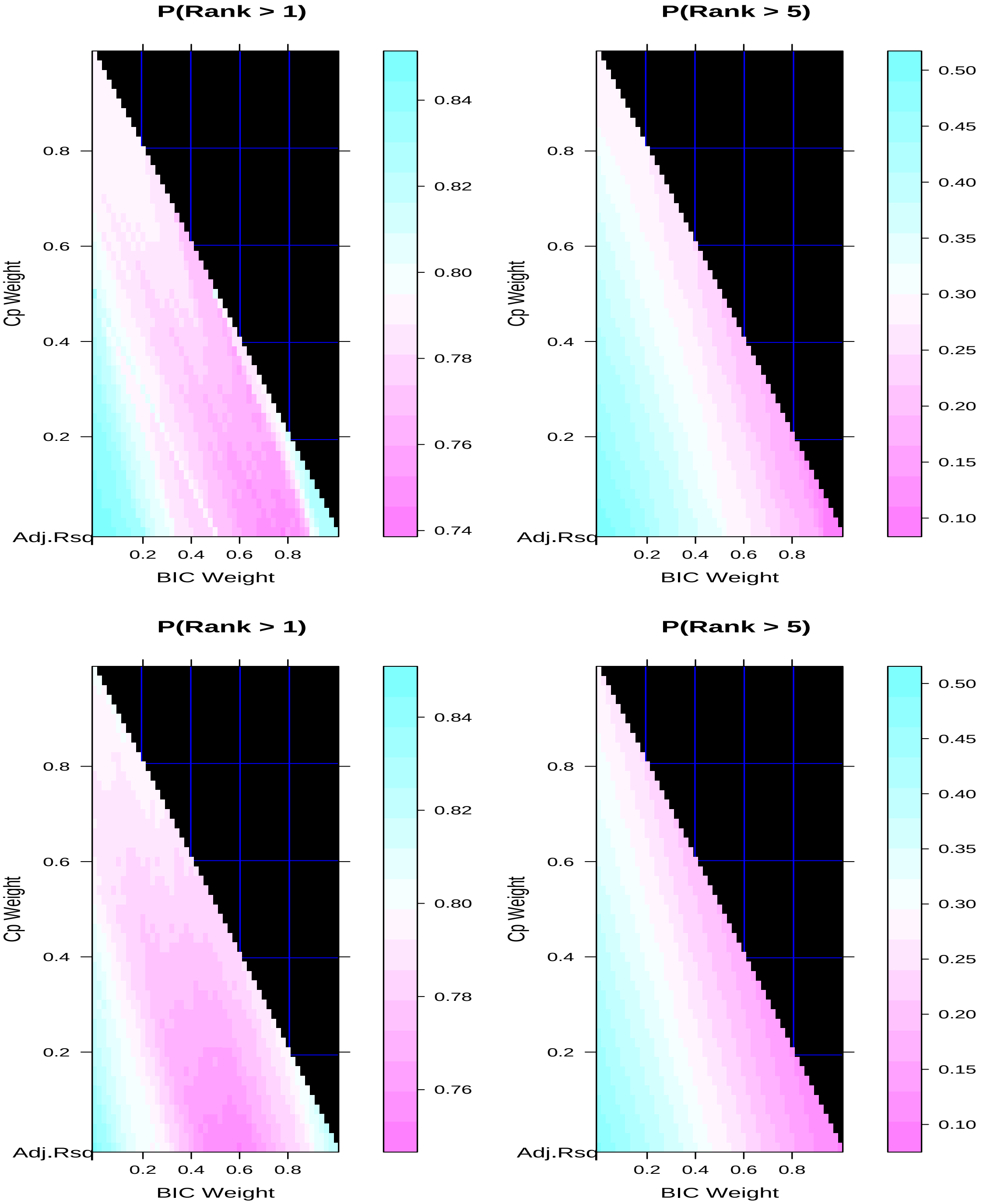


Figure 10: Example of regression in which combined criteria perform better than basis criteria. Top panels: Use of combined ranked basis criterion values. Bottom panels: Use of combined standardized basis criterion values. Note the slightly inferior performance of the standardized criteria. Figures are best viewed on a color display.

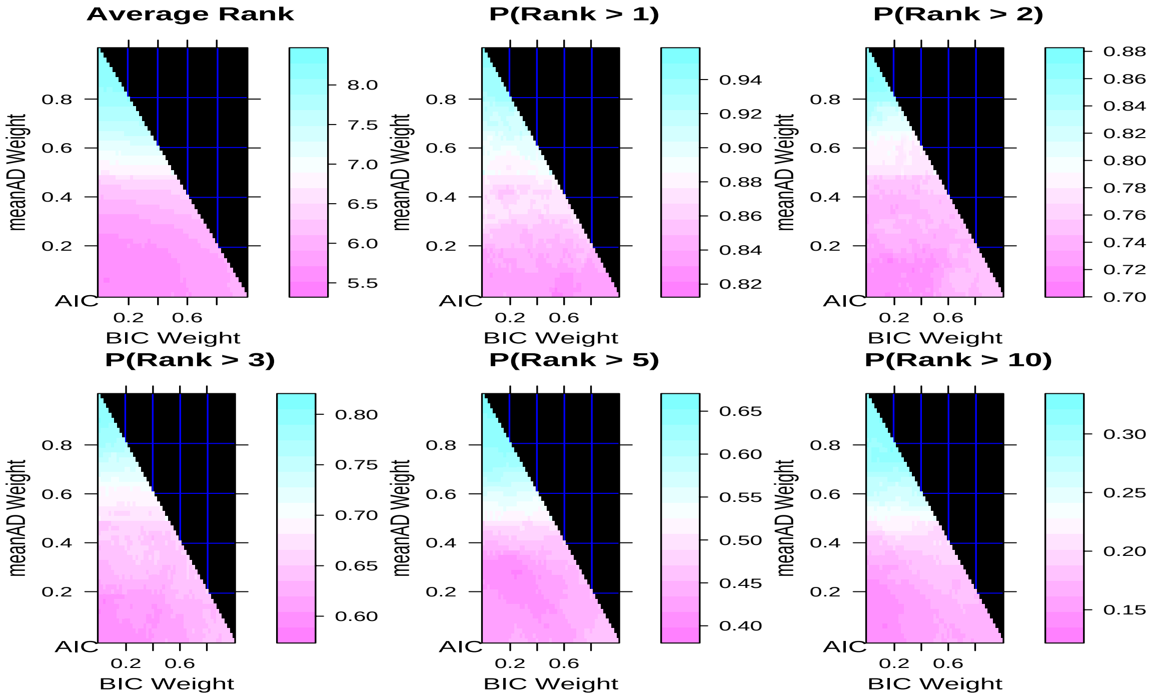


Figure 11: A fairly typical result for ARIMA models, in which the combined criteria appear to be of little value.

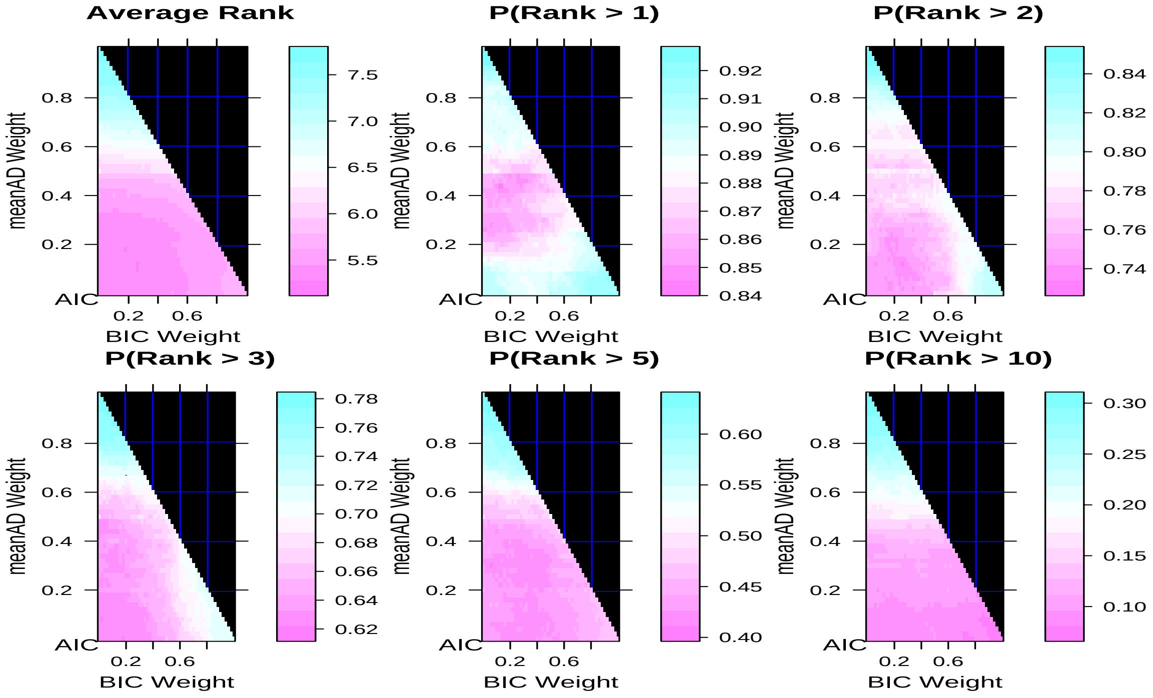


Figure 12: A more interesting result for ARIMA models, illustrating the potential utility of combined criteria.

is that it is easy to apply it to any problem, and in particular we need not rely on general recommendations. Rather, by the same simulation technique used to generate the example figures, we can obtain results tailored to the specific problem under consideration.

4.5 Theoretical Considerations

In this section we explore the theory behind our algorithm. In particular, we investigate the behavior of the functionals such as $\mathbb{E}[R_{MSC_i}(M^*)]$, which are the primary functions of interest. Our results show that these functionals are piecewise constant, discontinuous functions of $\boldsymbol{\alpha}$, and thus in a sense they justify our consideration of only a discrete set of values of $\boldsymbol{\alpha}$ in Algorithm 1. To simplify the discussion, we will operate under the assumption that there are never any ties in the set of values of an individual MSC, e.g., there are never ties in the set $\{BIC(M_1), BIC(M_2), \dots, BIC(M_s)\}$. This assumption is generally justified because most MSCs involve log-likelihoods, and are thus continuous. Further, define $\mathbf{B}$ as a random matrix formed by taking the vectors $[R_{MSC_j}(M_1), R_{MSC_j}(M_2), \dots, R_{MSC_j}(M_s)]^T, j = 1, \dots, k$ as its columns. Thus, by the above assumption, each column of $\mathbf{B}$ is a permutation of the integers 1 through s. We have the following theorem which illustrates the role of the convex coefficients $\boldsymbol{\alpha}$ in the distribution of $\mathbf{B}$.

Theorem 4.5.1 *Let $\boldsymbol{\alpha}$ be a vector with all nonnegative entries, such that $\sum_{i=1}^{k} \alpha_i = 1$. Suppose that, for all possible values of the matrix-valued random variable $\mathbf{B}$, the values in the vector $\mathbf{B}\boldsymbol{\alpha}$ are all distinct. Then there exists $\epsilon > 0$ such that*

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\| < \epsilon \Rightarrow \mathit{rank}(\mathbf{B}\boldsymbol{\alpha}) = \mathit{rank}(\mathbf{B}\boldsymbol{\alpha}')$$

Proof. First, note that by elementary linear algebra,

$$\|\mathbf{B}\|_2 \leq \|\mathbf{B}\|_F = \left(\sum_{i,j} \mathbf{B}_{ij}^2\right)^{\frac{1}{2}} = \left(k \cdot \sum_{i=1}^{s} i^2\right)^{\frac{1}{2}} = \sqrt{\frac{k \cdot s \cdot (s+1) \cdot (2s+1)}{6}}$$

Now, we have

$$\begin{aligned}
\|\mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\boldsymbol{\alpha}'\|_\infty &\leq \|\mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\boldsymbol{\alpha}'\|_2 \\
&\leq \|\mathbf{B}\|_2 \cdot \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2 \\
&\leq \sqrt{\frac{k \cdot s \cdot (s+1) \cdot (2s+1)}{6}} \cdot \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2
\end{aligned}$$

Thus, by choosing $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2$ sufficiently small, we can make $\|\mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\boldsymbol{\alpha}'\|_\infty$ as small as desired. Now, let $c = \min_{\mathbf{B}, i \neq j} |(\mathbf{B}\boldsymbol{\alpha})_i - (\mathbf{B}\boldsymbol{\alpha})_j|$, where the subscript is used to denote the index of a component of a vector. Note that $c > 0$ since by assumption $(\mathbf{B}\boldsymbol{\alpha})_i \neq (\mathbf{B}\boldsymbol{\alpha})_j$ for any $\mathbf{B}$ when $i \neq j$, and there are only a finite number of possible values of $\mathbf{B}$ since each column must be a permutation of $[1, \ldots, s]^T$. Suppose that we have chosen $\boldsymbol{\alpha}'$ such that

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2 \leq \frac{c \cdot \sqrt{6}}{2\sqrt{k \cdot s \cdot (s+1) \cdot (2s+1)}}.$$

Then, for each i, $|(\mathbf{B}\boldsymbol{\alpha})_i - (\mathbf{B}\boldsymbol{\alpha}')_i| \leq \|\mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\boldsymbol{\alpha}'\|_\infty \leq \frac{c}{2}$, and it is easy to see that $\text{rank}(\mathbf{B}\boldsymbol{\alpha}) = \text{rank}(\mathbf{B}\boldsymbol{\alpha}')$. □

Theorem 4.5.1 naturally leads to the question of which values of $\boldsymbol{\alpha}$ satisfy the condition that there are no ties in $\mathbf{B}\boldsymbol{\alpha}$ for any $\mathbf{B}$.

Theorem 4.5.2 *Let k denote the number of base MSCs. Then the set $A = \{\boldsymbol{\alpha} : \mathbf{B}\boldsymbol{\alpha}$ has at least 1 tie for some $\mathbf{B}\}$ has $(k-1)$-dimensional Lebesgue measure 0.*

Proof: For there to be a tie in $\mathbf{B}\boldsymbol{\alpha}$, the equation

$$B_{i1}\alpha_1 + B_{i2}\alpha_2 + \ldots + B_{ik}\alpha_k = B_{j1}\alpha_1 + B_{j2}\alpha_2 + \ldots + B_{jk}\alpha_k \tag{42}$$

must hold for some integers $i \neq j$. Now, since each column of $\mathbf{B}$ has distinct entries, we know that $B_{mi} \neq B_{ni}$ when $m \neq n$. Thus we can rewrite (42) as

$$n_1 \cdot \alpha_1 + n_2 \cdot \alpha_2 + \ldots n_{k-1}\alpha_{k-1} + n_k \cdot \alpha_k = 0$$

where the n_i are all nonzero integers. Since the vector $\boldsymbol{\alpha}$ is assumed to be a vector of convex coefficients, we can use the condition $\sum_{i=1}^{k} \alpha_i = 1$, to write

$$(n_1 - n_k)\alpha_1 + (n_2 - n_k)\alpha_2 + \ldots + (n_{k-1} - n_k)\alpha_{k-1} = -n_k$$

This is simply a linear equation in the $k-1$ variables $\alpha_1, \ldots \alpha_{k-1}$, and it is well-known that the set of solutions has dimension at most $k-2$, and therefore has $(k-1)$-dimensional Lebesgue measure 0. Finally, note that since by assumption, each column of $\mathbf{B}$ is a permutation of the integers 1 through s, there are only a finite number of possible values for $\mathbf{B}$. The result follows. □

Since the set of convex combinations of k criteria is a $k-1$ dimensional set, Theorem 4.5.2 tells us that for almost all convex combinations, there can be no ties in $\mathbf{B}\boldsymbol{\alpha}$, and thus Theorem 4.5.1 applies. Thus, these two theorems together give us a qualitative description of the behavior of the response as a function of the convex coefficients — it is a locally flat step function. This runs contrary to the intuition one might develop from looking at the simulation results presented in Section 4.4.

If one is willing to disregard the "small" set of convex coefficients which may result in ties in $\mathbf{B}\boldsymbol{\alpha}$, then Theorems 4.5.2 and 4.5.1 together imply that we need only check a finite number of values for $\boldsymbol{\alpha}$ in order to find the optimum for any particular functional – it is easy to see that if we test a grid of small enough mesh, at least one point in the grid must lie in the optimum region.

4.6 Computational Considerations

The computation involved in our algorithm can be quite demanding, both in terms of storage and computing requirements. The computation required depends primarily on the size of $\mathcal{M}$, and the number of grid points chosen (which itself is a function of the grid size and the number of MSCs being combined.) There is a natural tradeoff when choosing the size of the grid — a smaller grid size provides better insight into the behavior of the cost functional as a function of the MSC weights, yet it can result in an exponential increase in both computation and storage requirements.

The relative amount of computation required for the fitting of all models in $\mathcal{M}$ and for the calculation of all points on the grid depends on the type of model being fitted. For example, regression models are comparatively very easy to fit, and therefore in our simulations with regression models, most of the computation time is spent on the calculations for the grid. On the other hand, fitting ARIMA models requires a costly iterative maximization procedure, so the model fitting part of the algorithm typically dominates in ARIMA simulations. Deciding on an appropriate grid size and number of iterations, then, very much depends on the specifics of the problem.

Storage can also become an important issue, particularly when the number of MSCs to be compared is large and the corresponding grid of convex combinations becomes large in size. One key special feature to exploit is the structure of the summary functions we wish to compute for each convex combination. Essentially, the result of Algorithm 1 will be a vector of ranks for each combined MSC:

$$R_{\boldsymbol{\alpha}}(M_i^*), i = 1, 2, \ldots, N$$

If the summary functions can be computed dynamically, then we need not actually store this vector. Rather, we can update the summary function value and discard the individual ranks. Thankfully, this special structure occurs in many simple summary functions one might wish to use, including $\mathbb{E}[R_{\boldsymbol{\alpha}}(M^*)], \mathrm{Var}[R_{\boldsymbol{\alpha}}(M^*)], P\{R_{\boldsymbol{\alpha}}(M^*) > k\}$ for fixed k.

In choosing the number of base MSCs to use, there is of course a tradeoff between computation and storage requirements and the possibility of discovering more powerful combinations. Our experience indicates that using more than 3 or occasionally 4 MSCs is usually unhelpful. Further, it is also worth considering the similarities between MSCs. For example, many criteria are of the form

$$\text{loglik} - \lambda p$$

where p is the number of parameters in a fitted models. There is often little to be gained by considering many criteria of this same form. It is more likely that combinations of different *types* of criteria will be useful – e.g., an in-sample criterion of the form described above combined with an out-of-sample criterion based on a holdout sample. In our examples, we have always chosen 3 as the number of base MSCs, both to keep computational and storage requirements reasonable and also to allow easier graphical interpretation of the results.

4.7 Inference

In light of the simulation results presented above, there are several natural questions to ask:

- Is the observed improvement in performance by the combined criteria *significantly* better than the performance of the corresponding original criteria? That is, could the observed improvement be attributed to sampling error?
- Can we guarantee that the chosen combination is (nearly) optimal?

4.7.1 Hypothesis Testing

In the above simulation results, the empirical minimum of the response surface lies close to, but not exactly on, one of the corners of the domain, which naturally leads to the question of hypothesis testing: Is the observed minimum of the surface significantly smaller than a pre-specified value, or smaller than the values observed at the corners (which correspond to traditional model selection criteria)? This type of hypothesis allows one to determine if it is statistically worthwhile to consider the possibility of using our combined MSCs, rather than simple MSCs.

There are several possible approaches to this problem, each of which has some benefits and some drawbacks. The first is to regard the problem as a simple hypothesis test. Associated with each MSC we have a vector

$$(R_{\boldsymbol{\alpha}}(M_1^*), R_{\boldsymbol{\alpha}}(M_2^*), \ldots, R_{\boldsymbol{\alpha}}(M_N^*))^T,$$

where the subscript denotes the iteration number. This vector is a sample from the population of ranks of true models, under the priors $\pi_{\mathcal{M}}$ and $\pi_{\boldsymbol{\Theta}(\mathcal{M})}$, and assuming the set of feasible models is $\mathcal{M}$. We are then free to apply any of a number of nonparametric tests directly to two pre-specified MSCs, say MSC_i and MSC_j . For example, if the cost function is $\mathbb{E}[R_{\boldsymbol{\alpha}}(M^*)]$, we can directly compare the means of the two vectors by a simple paired t-test. Recall that we have required that the set of feasible models $\mathcal{M}$ be finite, which guarantees that the observations $R_{\boldsymbol{\alpha}}(M_l^*)$ are bounded for all $\boldsymbol{\alpha}, l$, which ensures that the populations involved have finite variances, and hence the CLT is always applicable, provided the sample sizes are sufficiently large. For comparing proportions of observations which exceed a fixed threshold k, a simple 2-sample proportion test can be used. For comparing the medians of two samples, there are simple nonparametric tests available. In particular, if we wish to test the hypothesis that the median of the sample differences is 0, we can use a sign test.

The difficulty with this approach is that it is invalid in the context of our original question, which is to test the observed *minimum* of the response surface against the original MSCs from which the combined MSCs were constructed. This is a case of the well-known problem of data snooping — in paying attention only to the minimum of the surface, we are implicitly testing many hypotheses simultaneously, which inflates the Type I error rate. The usual remedy to this problem of using the Bonferroni method or studentized range distribution to adjust $\boldsymbol{\alpha}$ is worthless here due to the overwhelmingly large number of hypotheses being tested, which is usually in the hundreds or thousands. An easy though statistically inefficient solution is to split the sample into two parts. Using only the first part, we locate the minimum of the response surface, corresponding to, say, $MSC_{\boldsymbol{\alpha}}$ for some particular value of $\boldsymbol{\alpha}$. We then compare the distribution of ranks corresponding to $MSC_{\boldsymbol{\alpha}}$ to the control MSCs, typically the original MSCs such as AIC and BIC, using only the observations from the second part of the sample. Since the two parts of the sample are independent and we have pre-specified the hypotheses to be tested using the second part of the sample, we are *not* implicitly testing many hypotheses in using this procedure, so a significant result may be interpreted with more confidence. Though the power of such hypothesis tests are decreased due to the smaller sample size which we may use without "cheating," we always have the option of simulating more observations. The only limitation is the computation involved in generating such observations.

4.7.2 Sample Size Estimation

To simplify notation, in this section we will abbreviate $R_{\boldsymbol{\alpha}}(M^*)$ as $R_{\boldsymbol{\alpha}}$. If the cost functional has the special form

$$T(F(R_{\boldsymbol{\alpha}})) = \mathbb{E}[G(\mathbf{B}, \boldsymbol{\alpha})]$$

for some function G, then the Algorithm 1 is actually a special case of the Sample Average Approximation method described in [14]. For the remainder of this section, we restrict our attention to this special case, noting that the functionals $\mathbb{E}[R_{\boldsymbol{\alpha}}]$ and $P\{R_{\boldsymbol{\alpha}} > k\} = \mathbb{E}[I\{R_{\boldsymbol{\alpha}} > k\}]$ are both expectations of functions of $\mathbf{B}$ and $\boldsymbol{\alpha}$.

Adopting the notation of [14], let $\mathcal{S}$ denote the set of convex coefficients being tested. Further, let

$$\begin{aligned} v^* &= \min_{\boldsymbol{\alpha}} \mathbb{E}[G(\mathbf{B}, \boldsymbol{\alpha})] \\ \mathcal{S}^* &= \{\boldsymbol{\alpha} : \mathbb{E}[G(\mathbf{B}, \boldsymbol{\alpha})] = v^*\} \\ \mathcal{S}^\epsilon &= \{\boldsymbol{\alpha} : \mathbb{E}[G(\mathbf{B}, \boldsymbol{\alpha})] \leq v^* + \epsilon\} \end{aligned}$$

and define sample counterparts $\hat{v}^*$, $\hat{\mathcal{S}}^*$, and $\hat{\mathcal{S}}^\epsilon$, where the expectation is replaced by sample average in the corresponding definitions. If we specify a type I error rate p and a positive number $\delta \in [0, \epsilon)$, then by Equation (2.23) in [14], we have that $\hat{\mathcal{S}}^\delta \subset \mathcal{S}^\epsilon$ with probability at least $1 - p$ if

$$N \geq \frac{\sigma^2_{\max}}{(\epsilon - \delta)^2} \cdot \log\left(\frac{|\mathcal{S}|}{p}\right), \tag{43}$$

where $\sigma^2_{\max} = \max_{\boldsymbol{\alpha} \in \mathcal{S} \setminus \mathcal{S}^\epsilon, \boldsymbol{\alpha}' \in \mathcal{S}^*} \operatorname{Var}[G(\mathbf{B}, \boldsymbol{\alpha}) - G(\mathbf{B}, \boldsymbol{\alpha}')]$. The most important feature of Equation (43) is that it depends only logarithmically on the size of the set of coefficients being considered, mitigating the exponential increase in the number of convex coefficients accompanying the addition of a basis MSC or a decrease in the mesh of the grid. We now derive a simple bound on the size of $\sigma^2_{\max}$.

Theorem 4.7.1 *1. If $G(\mathbf{B}, \boldsymbol{\alpha}) = R_{\boldsymbol{\alpha}}$, then $\sigma^2_{\max} \leq (s-1)^2$*

2. If $G(\mathbf{B}, \boldsymbol{\alpha}) = I\{R_{\boldsymbol{\alpha}} > k\}$ for some constant k, then $\sigma^2_{\max} \leq 1$.

Proof. For any $\boldsymbol{\alpha}$, the random variable $R_{\boldsymbol{\alpha}}$ can take only the values $\{1, 2, \ldots, s\}$. Therefore, for any $\boldsymbol{\alpha}, \boldsymbol{\alpha}'$,

$$R_{\boldsymbol{\alpha}} - R_{\boldsymbol{\alpha}'} \in \{1-s, 2-s, \ldots, -1, 0, 1, \ldots, s-2, s-1\}.$$

Now, we have

$$\begin{aligned} \operatorname{Var}[R_{\boldsymbol{\alpha}} - R_{\boldsymbol{\alpha}'}] &= \mathbb{E}[(R_{\boldsymbol{\alpha}} - R_{\boldsymbol{\alpha}'})^2] - \mathbb{E}[(R_{\boldsymbol{\alpha}} - R_{\boldsymbol{\alpha}'})]^2 \\ &\leq \mathbb{E}[(R_{\boldsymbol{\alpha}} - R_{\boldsymbol{\alpha}'})^2] \\ &\leq (s-1)^2 \end{aligned}$$

where the last inequality follows since $|R_{\boldsymbol{\alpha}} - R_{\boldsymbol{\alpha}'}| \leq s - 1$ w.p. 1. This establishes 1; 2 follows similarly. □

We note that, unfortunately, the bounds given in Theorem 4.7.1 are very weak. This is due to the fact that, for any values of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$, we expect $\operatorname{cor}(R_{\boldsymbol{\alpha}}, R_{\boldsymbol{\alpha}'})$ to be positive. Indeed, it has been noted (see, e.g., [19]) that all good criteria are typically highly positively correlated, and our combined criteria would of course be no exception to this general rule. A large positive correlation between the criteria would greatly reduce the variances computed in the proof above, and correspondingly give a lower bound for $\sigma^2_{\max}$. However, the correlation between criteria is merely an empirical fact, and is difficult to establish rigorously except in a few special cases.

Typically in practice, a more effective way to ensure the ϵ-optimality of the solution is to generate subsamples sequentially and use an optimality gap estimator, as discussed in [14] and [21], stopping when the estimate of the optimality gap is smaller than some specified threshold. This amounts to estimating the gap This approach can also be quite conservative, as is demonstrated in the numerical simulations in [14], but the sample sizes required are generally much smaller than those given by the theoretical bounds above.

4.8 The role of the Prior Distributions

As is frequently the case when a statistical method involves the use of prior distributions, a difficulty one must resolve is how to choose an appropriate prior. In our case, we have 2 prior distributions to consider, $\pi_{\mathcal{M}}$ and $\pi_{\mathbf{\Theta}(\mathcal{M})}$. In the case of π_M, as mentioned above, often a prior with simple structure is sufficient. For example, one might choose a prior on the *order* of the true model based on the effect sparsity principle, and, conditioning on the order, assume that all models of that order are equally likely. Of course, knowledge about the specific problem should also be incorporated whenever possible.

The prior $\pi_{\mathbf{\Theta}(\mathcal{M})}$ is somewhat more difficult to specify in a principled way. First, we must note that the parameters $\Theta(M_i)$ are actually *nuisance* parameters, since we are not actually interested in doing any inference on them. However, it is our empirical finding that the specification of this prior distribution typically has little effect on the *shape* of the resulting response surface, although it may cause a shift depending on the problem. It seems that, at least in this context, knowledge of the particular distribution from which the parameters are assumed to be drawn is not especially important. Nevertheless, we still recommend adopting a common strategy in Bayesian modelling, which is to repeat an experiment with several different prior distributions in order to assess the sensitivity of the problem to the modelling assumptions. Perhaps more interesting, however, is the observation that the response surface seems to be rather sensitive to the magnitude of the parameters, as we observe that the minimum of the surface shifts from the BIC corner to the PRESS corner as we increase the variance of the prior distribution. Thus, it seems that the only truly important aspect of the specification of $\pi_{\boldsymbol{\Theta}(\mathcal{M})}$ is the order of magnitude of the variance, while the particular shape of distribution chosen has little effect. Here, one should rely on domain-level knowledge of the problem under consideration. We also recommend standardizing the predictors to ensure that the estimated regression coefficients always have a common scale.

4.9 Discussion

We have implemented Algorithm 1 as an R package which is freely available for download on the R Project website.

There is still much investigation left to be done in this area. In particular, we have only begun to explore the effects of the many factors that may affect the optimal MSC — the size of the data set, the distributions of the parameters, the correlation structure of the covariates, etc. In particular, a characterization of situations in which combined criteria outperform traditional criteria would be a very useful advance.

Our method, as described, is entirely empirical. In general, we expect analytical results will be very difficult to derive, due to the complexity of the distributions involved. However, we expect that in certain simple, special cases, analytical results may be feasible. It would be interesting to compare the empirical simulation results from our method with these mathematical results. One example of a simple special case would be regression in which the predictor matrix is constrained to be orthogonal.

Further, there are many other MSCs which may be incorporated into the convex combinations. We chose the ones used here mainly due to their readily available software implementations, and the desire to limit the number of base MSCs to 3 in order to make the results easy to visualize. Nevertheless, there is no a priori reason to exclude or favor certain MSCs over others, as our procedure should place weight only on MSCs which have proven power to discriminate between good and poor models.

It would also be useful if there were more efficient procedures for handling the implicit multiple comparisons issue discussed in Section 4.7. This would allow us to decide whether a result is statistically significant using less computation. We anticipate that the key lies in exploiting the fact that all good MSCs, and combinations thereof, are typically highly correlated with each other. Taking advantage of this fact would likely reduce the Type I error rate in the hypothesis tests.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

A.1 Proofs

A.1.1 Proof of Lemma 2.3.2

The first inequality in the lemma is obvious. For the second inequality, we have

$$\begin{aligned} d_{\max,i} = \|X_i \overline{P}_k\|_2 &= \|(X_i - x_0 \cdot \mathbf{1}_k^T)\overline{P}_k\|_2 \\ &\leq \|X_i - x_0 \cdot \mathbf{1}_k^T\|_2 && (44) \\ &\leq \sqrt{k} \cdot \max_{j \in P_i} \|x_j - x_0\|_2 && (45) \\ &\leq \sqrt{k} \cdot \tau. \end{aligned}$$

Taking the maximum over i on both sides, we obtain the second inequality.

In the above, inequality (44) is true because in general, for two matrices A and B, we have $\|AB\|_2 \leq \|A\|_2 \cdot \|B\|_2$ (Stewart and Sun [25, page 69]). The inequality (45) is also standard linear algebra ([25, page 71]).

A.1.2 Proof of Theorem 2.3.5

The following two equations will be used:

$$Y_i \overline{P}_k = (Y_i - f(x_i^{(0)}) \cdot \mathbf{1}_k^T)\overline{P}_k \tag{46}$$

and

$$X_i \overline{P}_k = (X_i - x_i^{(0)} \cdot \mathbf{1}_k^T)\overline{P}_k. \tag{47}$$

They can easily be verified by recalling the definition of $\overline{P}_k$.

To exploit the local isometry, we consider the Taylor expansion at $x_i^{(0)}$. It is not hard to verify the following: for $j \in P_i$, $1 \leq i \leq n$,

$$\begin{aligned}
& \|y_j - f(x_i^{(0)}) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_\infty \\
\leq \; & \|y_j - f(x_j)\|_\infty + \|f(x_j) - f(x_i^{(0)}) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_\infty \\
\leq \; & \sigma + \frac{1}{2} C_1 \|x_j - x_i^{(0)}\|_2^2 + O(\|x_j - x_i^{(0)}\|_2^3) \\
\leq \; & \sigma + \frac{1}{2} C_1 \tau^2.
\end{aligned}$$

Note that in the last step, we dropped an $O(\tau^3)$ term because we are only interested in the case when $\tau \to 0$, in which case the quadratic term dominates.

Let $E_i = Y_i \overline{P}_k - J(f; x_i^{(0)}) X_i \overline{P}_k$. Note that $E_i \in \mathbb{R}^{D \times k}$. We have the following upper bound for $\|E_i\|_2$:

$$\begin{aligned}
\|E_i\|_2 &= \|Y_i \overline{P}_k - J(f; x_i^{(0)}) X_i \overline{P}_k\|_2 \\
&\leq \sqrt{k} \cdot \sup_{j \in P_i} \|(y_j - f(x_i^{(0)})) \cdot \overline{P}_k - J(f; x_i^{(0)})(x_j - x_i^{(0)}) \cdot \overline{P}_k\|_2 \\
&\leq \sqrt{k} \cdot \sup_{j} \|(y_j - f(x_i^{(0)})) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_2 \\
&\leq \sqrt{kD} \cdot \sup_{j} \|(y_j - f(x_i^{(0)})) - J(f; x_i^{(0)})(x_j - x_i^{(0)})\|_\infty \\
&\leq \sqrt{kD} \cdot [\sigma + \frac{1}{2} C_1 \tau^2]. \qquad (48)
\end{aligned}$$

In the above, the first and third inequalities are standard linear algebra, the second inequality is due to the fact that $\overline{P}_k$ is a projection matrix.

We now wish to derive a bound on the angle between the subspaces spanned by the right singular vectors associated with the d largest singular values of $J(f; x_i^{(0)}) X_i \overline{P}_k$ and by those of $Y_i \overline{P}_k$. To this end, define the following two quantities:

$$\begin{aligned}
R &= J(f; x_i^{(0)}) X_i \overline{P}_k \widetilde{B}_i - \widetilde{A}_i \widetilde{D}_i \\
S &= \overline{P}_k X_i^T J^T(f; x_i^{(0)}) \widetilde{A}_i - \widetilde{B}_i \widetilde{D}_i
\end{aligned}$$

By substituting the identity $E_i = Y_i \overline{P}_k - J(f; x_i^{(0)}) X_i \overline{P}_k$, it is easy to see that $\|R\|_2 \leq \|E_i\|_2$ and $\|S\|_2 \leq \|E_i\|_2$. Finally, consider the smallest singular value of $Y_i \overline{P}_k$. We have

$$\begin{aligned}
\sigma_{\min}(Y_i\overline{P}_k) &= \sigma_{\min}(J(f;x_i^{(0)})X_i\overline{P}_k + E_i) \\
&\geq \sigma_{\min}(J(f;x_i^{(0)})X_i\overline{P}_k) - \sigma_{\max}(E_i) \\
&\overset{(48)}{\geq} C_2 \cdot \tau - \sqrt{kD}\left[\sigma + \frac{1}{2}C_1\tau^2\right].
\end{aligned}$$

We can now apply Theorem V.4.4 in [25], and conclude

$$\begin{aligned}
\|\sin((\mathcal{R}(B_i^T), \mathcal{R}(\widetilde{B}_i^T)))\|_2 &\leq \frac{\|E_i\|_2}{\sigma_{\min}(Y_i\overline{P}_k)} \\
&\leq \frac{\sqrt{kD}\cdot[\sigma + \frac{1}{2}C_1\tau^2]}{C_2\cdot\tau - \sqrt{kD}\cdot[\sigma + \frac{1}{2}C_1\tau^2]}
\end{aligned}$$

If we ignore higher-order terms, we can take $C_3 = \frac{\sqrt{kD}}{C_2}$, and the theorem is established.

A.1.3 Proof of Theorem 2.3.8

Now we consider the step of global alignment. Recall that the columns of $(\mathbf{1}_n, X)$, where X is defined in (6), are eigenvectors associated with the zero eigenvalue of (7).

First, similar to M_n, define $\widetilde{M_n}$ as

$$\widetilde{M_n} = (S_1, \ldots, S_n)\overline{P}_{k\cdot n}\begin{pmatrix} I_k - \widetilde{B}_1^T\widetilde{B}_1 & & \\ & \ddots & \\ & & I_k - \widetilde{B}_n^T\widetilde{B}_n \end{pmatrix}\overline{P}_{k\cdot n}(S_1, \ldots, S_n)^T,$$

where $\widetilde{B}_i$ is defined right before Theorem 2.3.5.

We now consider $\|M_n - \widetilde{M_n}\|_2$, the norm of the difference between the alignment matrices formed from the true and estimated local coordinates. This is equivalent to

$$\left\|\sum_{i=1}^n S_i\cdot\overline{P}_k\left(\widetilde{B}_i^T\widetilde{B}_i - B_i^TB_i\right)\overline{P}_k\cdot S_i^T\right\|_2.$$

Theorem 2.3.5 and [25, Theorem I.5.5] together imply that

$$\|\widetilde{B}_i^T\widetilde{B}_i - B_i^TB_i\|_2 \leq C_3\left(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau\right), i = 1, \ldots, n.$$

Now, since $M_n - \widetilde{M}_n$ is symmetric, we have $\|M_n - \widetilde{M}_n\|_2 \leq \|M_n - \widetilde{M}_n\|_1$. By Condition 2.3.6, each column of $M_n - \widetilde{M}_n$ will be the sum of at most C_4 terms, each of which is a column of one of the matrices $B_i^T B_i - \widetilde{B}_i^T \widetilde{B}_i$. Therefore, we have

$$\|M_n - \widetilde{M}_n\|_2 \leq C_4 \cdot C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right)$$

To verify the conditions of Theorem V.2.7 in [25], consider

$$\begin{pmatrix} \frac{\mathbf{1}_n^T}{\sqrt{n}} \\ X^T \\ (X^c)^T \end{pmatrix} \left(M_n - \widetilde{M}_n \right) \left(\frac{\mathbf{1}_n}{\sqrt{n}}, X, X^c \right) = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}, \tag{49}$$

where $E_{11} \in \mathbb{R}^{(d+1)\times(d+1)}, E_{12} \in \mathbb{R}^{(d+1)\times(n-d-1)}, E_{21} \in \mathbb{R}^{(n-d-1)\times(d+1)}$, and $E_{22} \in \mathbb{R}^{(n-d-1)\times(n-d-1)}$. Since we have assumed that

$$\begin{pmatrix} \frac{\mathbf{1}_n^T}{\sqrt{n}} \\ X^T \\ (X^c)^T \end{pmatrix}$$

is orthogonal, and since an upper bound on the spectral norm of a matrix is also an upper bound on the spectral norm of any submatrix, we have $\|E_{11}\|_2 \leq C_4 \cdot C_3 \left(\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau \right)$, and similarly for all the other blocks.

It now easily follows that we can apply Theorem V.2.7 in [25], and therefore

$$\begin{aligned} \left\| \tan(\mathcal{R}(\widetilde{X}), \mathcal{R}(X)) \right\|_2 &\leq \frac{2\|E_{12}\|_2}{\lambda_{\min}^+ - \|E_{11}\|_2 - \|E_{22}\|_2} \\ &\leq \frac{C_4 \cdot C_3 \left(\frac{\sigma}{\tau} + C_1 \tau \right)}{\lambda_{\min}^+ - 2C_4 \cdot C_3 \left(\frac{\sigma}{\tau} + C_1 \tau \right)} \end{aligned}$$

A.2 Preliminary Results on $\lambda_{\min}^+$

In this Appendix, we discuss a possible modification to the LTSA algorithm which will provably converge, supposing that the conjectured relationship between k, $\lambda_{\min}^+$, and τ in Zha's notes on LTSA and Biharmonic Eigenvalue Problems is true in general.

From these notes, we have the following result for the special case when the x_i are sample on a uniform grid:

$$\lambda^+_{\min} \approx C(k) \cdot \nu^+_{\min}(\Delta^2) \cdot \tau^4, \tag{50}$$

where $\nu^+_{\min}(\Delta^2)$ is a constant, and $C(k) \approx k^5$. Throughout the current version of our paper, we assume that k is constant. However, allowing k to be a function of the sample size N, say

$$k = N^\alpha,$$

where $\alpha \in [0, 1)$ allows us to control the asymptotic behavior of $\lambda^+_{\min}$ along with the convergence of the estimated alignment matrix to the true alignment matrix.

Consider our original bound on the angle between the true coordinates and the estimated coordinates:

$$\lim_{\tau \to 0} \| \tan(\mathcal{R}(\widetilde{X}), \mathcal{R}(X))\|_2 \leq \frac{C_3(\frac{\sigma}{\tau} + C_1\tau) \cdot \| \sum_{i=1}^n S_i\|_\infty}{\lambda^+_{\min}}$$

Now, set $k = N^\alpha$, where $\alpha \in [0, 1)$ is an exponent, the value of which we can decide later. We must be careful in disregarding constants, since they may involve k. We have from the original paper that $C_3 = \frac{\sqrt{kD}}{C_2}$. C_1 and C_2 are fundamental constants not involving k. Further, it is easy to see that $\| \sum_{i=1}^n S_i\|_\infty$ is $O(k)$ - since each point has k neighbors, the maximum number of neighborhoods to which a point belongs is of the same order as k.

Now, we can use a simple heuristic to estimate the order of τ, the neighborhood size. For example, suppose we fix ϵ and consider ϵ-neighborhoods. For simplicity, assume that the parameter space is the unit hypercube $[0, 1]^d$, where d is the intrinsic dimension. The law of large numbers tells us that

$$k \approx \epsilon^d \cdot N.$$

Thus we can approximate τ as

$$\tau \approx O(N^{\frac{\alpha-1}{d}})$$

Plugging all this in to the original equation and dropping the constants, we get

$$\lim_{\tau \to 0} \| \tan(\mathcal{R}(\widetilde{X}), \mathcal{R}(X)) \|_2 \leq \frac{N^{\frac{\alpha-1}{d}} \cdot N^{\frac{3\alpha}{2}}}{\lambda_{\min}^{+}}$$

If we conjecture that the relationship in (50) holds in general (i.e., the generating coordinates can follow a more general distribution rather than only lying in a uniform grid), then we have

$$\lim_{\tau \to 0} \| \tan(\mathcal{R}(\widetilde{X}), \mathcal{R}(X)) \|_2 \leq \frac{N^{\frac{\alpha-1}{d}} \cdot N^{\frac{\alpha}{2}} \cdot N^{\alpha}}{N^{5\alpha} \cdot N^{4 \cdot \frac{\alpha-1}{d}}}$$

Now the exponent is a function only of α and the constant d. We can try to solve for α such that the convergence is as fast as possible. Simplifying the exponents, we get

$$\lim_{\tau \to 0} \| \tan(\mathcal{R}(\widetilde{X}), \mathcal{R}(X)) \|_2 \leq N^{\frac{-7\alpha}{2} - 3(\frac{\alpha-1}{d})}$$

As a function of α restricted to the interval $[0, 1)$, there is no minimum — the exponent decreases with α, and we should choose α close to 1.

However, in the proof of the convergence of LTSA, it is assumed that the errors in the local step converge to 0. This error is given by

$$\| \sin(\mathcal{R}(B_i^T), \mathcal{R}(\widetilde{B}_i^T)) \|_2 \leq \frac{\sqrt{kD} \cdot [\sigma + \frac{1}{2} C_1 \tau^2]}{C_2 \cdot \tau - \sqrt{kD} \cdot [\sigma + \frac{1}{2} C_1 \tau^2]}$$

Thus, our choice of α is restricted by the fact that the RHS of this equation must still converge to 0. Disregarding constants and writing this as a function of N, we get

$$\frac{N^{\frac{\alpha}{2}} \cdot N^{\frac{2\alpha-2}{d}}}{N^{\frac{\alpha-1}{d}} - N^{\frac{\alpha}{2}} \cdot N^{\frac{2\alpha-2}{d}}}$$

This quantity converges to 0 as $N \to \infty$ if and only if we have

$$\begin{aligned} \frac{\alpha}{2} + \frac{2\alpha - 2}{d} &< \frac{\alpha - 1}{d} \\ d \cdot \alpha + 4\alpha - 4 &< 2\alpha - 2 \\ \alpha &< \frac{2}{d+2} \end{aligned}$$

Table 3: Convergence rates for a few values of the underlying dimension d.

d	1	2	3	4	5
Optimal α	0.66	0.5	0.4	0.33	0.29
Convergence rate	-1.33	-1	-0.8	-0.66	-0.57

Note that this bound is strictly less than 1 for all positive integers d, so our possible choices of α are restricted further.

By the reasoning above, we want the exponent to be as large as possible. Further, it is easy to see that for all d, choosing an exponent roughly equal to $\frac{2}{d+2}$ will always yield a bound converging to 0. The following table gives the optimal exponents for selected values of d along with the convergence rate of $\lim_{\tau \to 0} \| \tan(\mathcal{R}(\widetilde{X}), \mathcal{R}(X)) \|_2$. In general, using the optimal value of α, the convergence rate will be roughly $N^{\frac{-4}{d+2}}$.

We present some numerical experiments to illustrate the above results. In these examples, we simulate coordinates $\in \mathbb{R}$ according to a uniform distribution on the interval $[\frac{\pi}{5}, 2\pi]$, and transform the coordinates via the function

$$g(x) = (x \cdot \cos(x), x \cdot \sin(x)).$$

The resulting points lie on a spiral as illustrated in Figure 13. In Figure 14, we illustrate the effect of choosing different values of k as the sample size increases. The response in this figure is $|\mathrm{corr}(X, \widetilde{X})|$. We observe an interesting effect in these experiments – for fixed k, as the sample size N increases, the performance eventually deteriorates and becomes highly erratic. On the other hand, as expected from the above analysis, with $k = \sqrt{N}$, the performance is very good for all large values of N. This is due to the eigenvalue mixing suggested in [28] – as N grows, $\lambda^+_{\min}$ gets small, allowing the possibility, depending on the randomly generated coordinates, that the eigenvalues of the alignment matrix could switch order, thus causing a spurious eigenvector corresponding to $\lambda^+_{\min}$ to be recovered instead of the correct null space vector.

Associated with each fixed value of k, there seems to be a threshold value of N

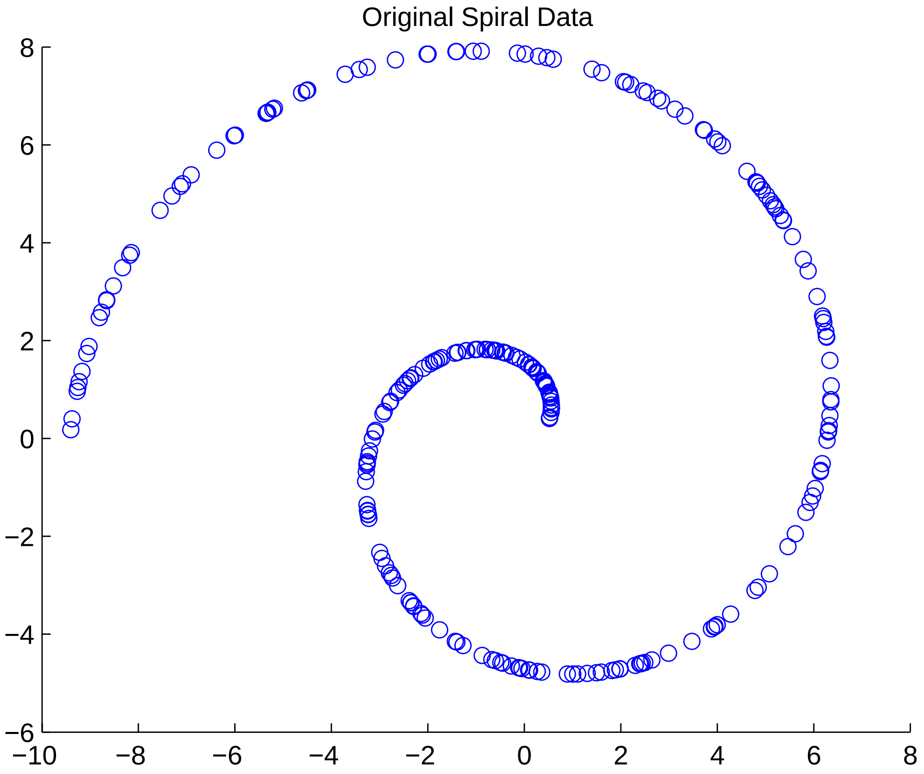


Figure 13: Illustration of the manifold used for the examples in Figure 14

above which the performance degrades. This value increases with k, though perhaps at the cost of worse performance for small N. However, we expect from the above analysis that, regardless of the value chosen, the performance will eventually become unacceptable for any fixed k.

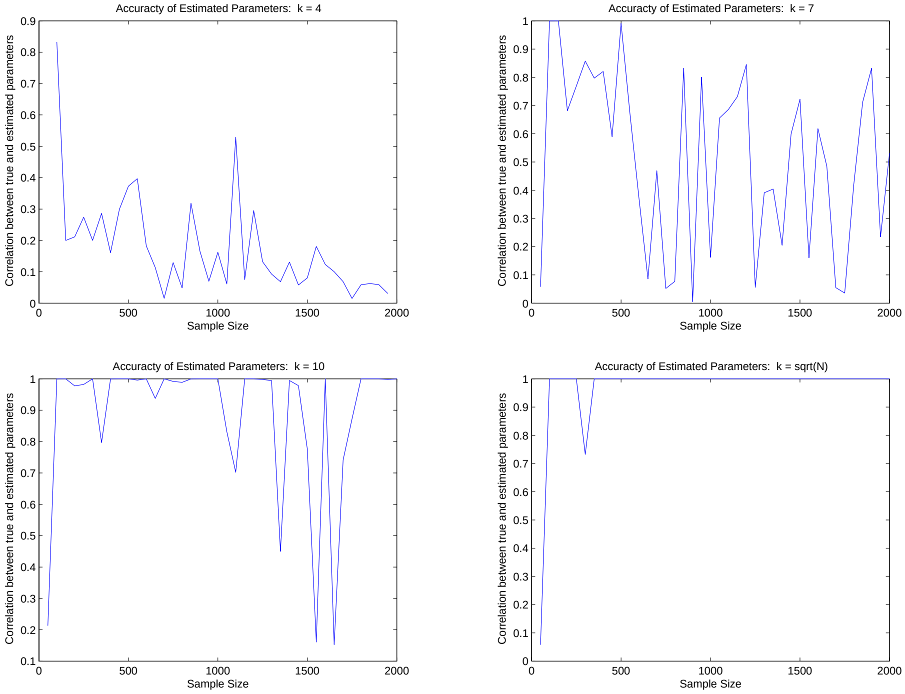


Figure 14: Illustration of the effect of increasing k. $k = 4, 7, 10, \sqrt{N}$ in the four panels, respectively.

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

B.1 The equivalence of the Hessian estimator under orthogonal transforms

Recall that $U \in \mathbb{R}^{k\times d}$ is made by d orthonormal column vectors, where k is the number of nearest neighbors and d is the intrinsic dimension of the underlying manifold. In the HLLE algorithm, the following matrix is considered:

$$[\mathbf{1}_k, U, F(U)],$$

where $F(U) \in \mathbb{R}^{k\times d(d+1)/2}$ is made by self-products and cross-products of columns of U.

To be more specific, let $U_\times$ define a matrix whose $[(i-1)d+j]$th column ($1 \leq i, j \leq d$) is $(U_{1i}U_{1j}, U_{2i}U_{2j}, \ldots, U_{ki}U_{kj})^T$, where U_{ab} is the (a,b) entry of U. One may notice that in $U_\times$, for $i \neq j$, columns $(i-1)d+j$ and $(j-1)d+i$ are identical. Letting S be a selection matrix that will eliminate these repeated columns, we can write

$$F(U) = U_\times \cdot S.$$

For later convenience, we define a reverse selection matrix S^R, such that

$$F(U) \cdot S^R = U_\times.$$

Suppose we have a QR-decomposition:

$$[\mathbf{1}_k, U, F(U)] = (Q_1, Q_2)\begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix},$$

where $Q_1 \in \mathbb{R}^{k\times(d+1)}$ and $Q_2 \in \mathbb{R}^{k\times d(d+1)/2}$ have orthonormal columns—$Q_1^T Q_1 = I_{d+1}$ and $Q_2^T Q_2 = I_{d(d+1)/2}$—and R_{11} and R_{22} are upper triangular. In the HLLE algorithm, one takes $H^i = Q_2^T$ and the "core" matrix is $(H^i)^T H^i = Q_2 Q_2^T$.

Consider a right orthogonal transform of U: $U^* = U \cdot O$, where $O \in \mathbb{R}^{d\times d}$ is an orthogonal matrix ($O^T O = OO^T = I_d$). Evidently, the Hessian multiplier matrix in HLLE is $H^{i*} = (Q_2^*)^T$, where Q_2^* is based on the QR decomposition of $[\mathbf{1}_k, U^*, F(U^*)]$, associated with the columns corresponding to the submatrix $F(U^*)$. We have the following result.

Theorem B.1.1 *We have*

$$(H^{i*})^T(H^{i*}) = (H^i)^T(H^i),$$

or equivalently

$$Q_2 Q_2^T = Q_2^*(Q_2^*)^T.$$

Proof. Similar to the definition of $U_\times$ let $U_\times^*$ be the corresponding matrix for U^*. We can easily verify that

$$U_\times^* = U_\times \cdot (O \otimes O),$$

where $O \otimes O$ is the Kronecker product [15]. Consequently, we have

$$\begin{aligned} F(U^*) &= U_\times^* \cdot S \\ &= U_\times \cdot (O \otimes O) \cdot S \\ &= F(U) \cdot S^R \cdot (O \otimes O) \cdot S. \end{aligned}$$

Hence

$$\begin{aligned} [\mathbf{1}_k, U^*, F(U^*)] &= [\mathbf{1}_k, U \cdot O, F(U) \cdot S^R \cdot (O \otimes O) \cdot S] \\ &= [\mathbf{1}_k, U, F(U)] \begin{bmatrix} 1 & & \\ & O & \\ & & S^R \cdot (O \otimes O) \cdot S \end{bmatrix} \\ &= (Q_1, Q_2) \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix} \begin{pmatrix} O_2 & \\ & S^R \cdot (O \otimes O) \cdot S \end{pmatrix}, \end{aligned}$$

where O_2 =diag$(1, O)$, which is another orthogonal matrix. Suppose we have QR-decompositions:

$$\begin{aligned} R_{11} O_2 &= Q_3 R_4, \\ R_{22} \cdot S^R \cdot (O \otimes O) \cdot S &= Q_4 R_5, \end{aligned}$$

where Q_3 and Q_4 have orthonormal columns and R_4 and R_5 are upper triangular. Hence we have

$$[\mathbf{1}_k, U^*, F(U^*)] = (Q_1 Q_3, Q_2 Q_4) \begin{pmatrix} R_4 & Q_3^T R_{12} \cdot S^R \cdot (O \otimes O) \cdot S \\ & R_5 \end{pmatrix}.$$

Due to the uniqueness of the QR-decomposition, we have

$$Q_2^* = Q_2 Q_4;$$

Consequently, we have

$$Q_2^* (Q_2^*)^T = Q_2 Q_4 Q_4^T Q_2^T = Q_2 Q_2^T.$$

B.2 The Asymptotic Behavior of $\widetilde{D}_i$

In this appendix we discuss some technical details regarding the asymptotic behavior of the diagonal matrices $\widetilde{D}_i$. As an introduction to the problem, consider the typical term of this matrix, $\frac{1}{\hat{d}_i \hat{d}_j}$. Recall that, by Condition 3.2.4, we have

$$\|y_{i_j} - \overline{y}_i\| \le \tau, j = 1, \ldots, k, i = 1, \ldots, N.$$

It is then easy to show (Lemma 2.3.2) that

$$C_2 \tau \le d_{\min} \le d_{\max} \le \tau \sqrt{k}. \tag{51}$$

We therefore see that, as τ goes to zero, the diagonal entries of $\widetilde{D}_i$ grow without bound. This complicates the analysis of the deviations $\widetilde{D}_i - D_i^*$, where D^* is the diagonal matrix formed from the reciprocals of the cross-products and squares of the singular values of $J_g^T(\overline{y}_i J_g)(\overline{y}_i)Y_i$, i.e., the "population" singular values at $\overline{y}_i$. This problem can be resolved, however, by rescaling the data in a suitable way.

Specifically, let us define a rescaled neighborhood as follows:

$$\widetilde{Y}_i \overset{\mathrm{def}}{=} \frac{Y_i}{\tau}$$

This rescaling essentially keeps the local neighborhoods of roughly constant size, rather than shrinking with τ. First, we note that rescaling each neighborhood does not affect the ultimate result of the algorithm. To see this, note that if $Y_i = U_i D_i V_i^T$ is the singular value decomposition of Y_i, then it is easy to show

$$\widetilde{Y}_i = U_i \left(\frac{1}{\tau} D_i\right) V_i^T \tag{52}$$

is the singular value decomposition of $\widetilde{Y}_i$. Therefore the left singular vectors are unchanged by this transformation, and so by its construction, the matrix X_i defined in Step 3 of HLLE will also remain unchanged. Thus, if we consider the Hessian estimator $\widehat{H} = \widetilde{D} R_{22}^{-1} Q_2^T$, the only factor affected by the rescaling is $\widetilde{D}$. It is clear from (52), however, that $\widetilde{D}$ is changed only by a scalar multiple.

To analyze $\widetilde{D}_i$ under perturbation, we must consider Θ_i, the matrix formed by taking the embedding coordinates of the k nearest neighbors of y_i as its rows (hence $\Theta_i \in \mathbb{R}^{k \times d}$.) Analogous to $\widetilde{Y}_i$, we define $\widetilde{\Theta}_i = \frac{\Theta_i}{\tau}$. Since the columns of $J_g(y_i)$ are assumed to be orthonormal (Condition 3.2.1), the singular values of $\widetilde{\Theta}_i$ are the same as those of $J_g(y_i)\widetilde{\Theta}_i^T$. Further, we have (see the proof of Theorem 2.3.5)

$$\|E_i\| \overset{\mathrm{def}}{=} \|\widetilde{Y}_i - J_g(y_i)\widetilde{\Theta}_i\| \leq \sqrt{kD} \cdot (\frac{\sigma}{\tau} + \frac{1}{2} C_1 \tau). \tag{53}$$

Summarizing the above results, we have the following bounds on the singular values of $\widetilde{Y}_i$ and $\widetilde{\Theta}_i$:

$$\begin{aligned} \min \sigma_{\widetilde{Y}_i} &\geq C_2 - \sqrt{kD} \cdot (\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau) && (54) \\ \max \sigma_{\widetilde{Y}_i} &\leq \sqrt{k} + \sqrt{kD} \cdot (\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau) && (55) \\ \min \sigma_{\widetilde{\Theta}_i} &\geq C_2 && (56) \\ \max \sigma_{\widetilde{\Theta}_i} &\leq \sqrt{k} && (57) \end{aligned}$$

If we let σ_j and σ_k denote any two singular values of $\widetilde{\Theta}_i$, and $\widetilde{\sigma}_j$ and $\widetilde{\sigma}_k$ denote any singular values of $\widetilde{Y}_i$ ($1 \leq j, k \leq d$), then each term of $\widetilde{D}_{\widetilde{Y}_i} - \widetilde{D}_{\widetilde{\Theta}_i}$ will be of the form

$$\frac{1}{\sigma_j \sigma_k} - \frac{1}{\widetilde{\sigma}_j \widetilde{\sigma}_k}.$$

The standard perturbation theory for inverses gives

$$\left| \frac{1}{\sigma_j \sigma_k} - \frac{1}{\widetilde{\sigma}_j \widetilde{\sigma}_k} \right| \leq (\sigma_j \sigma_k)^{-1} \cdot (\widetilde{\sigma}_j \widetilde{\sigma}_k)^{-1} \cdot |\sigma_j \sigma_k - \widetilde{\sigma}_j \widetilde{\sigma}_k|.$$

Now, using the bounds on σ_j and $\widetilde{\sigma}_j$ given above, and letting $\epsilon = \sqrt{kD} \cdot (\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau)$, we have

$$\begin{aligned} (\sigma_j \sigma_k)^{-1} &\overset{(56)}{\leq} \frac{1}{C_2^2} \\ (\widetilde{\sigma}_j \widetilde{\sigma}_k)^{-1} &\overset{(54)}{\leq} \frac{1}{(C_2 - \sqrt{kD} \cdot (\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau))^2}. \end{aligned}$$

Further, we have

$$\begin{aligned} |\sigma_j \sigma_k - \widetilde{\sigma}_j \widetilde{\sigma}_k| &\overset{(53)}{\leq} |\sigma_j \sigma_k - (\sigma_j + \epsilon)(\sigma_k + \epsilon)| \\ &= |\sigma_j \epsilon + \sigma_k \epsilon + \epsilon^2| \\ &\overset{(57)}{\leq} 2 \cdot \sqrt{k}\epsilon + \epsilon^2. \end{aligned}$$

Putting these results together, and dropping the higher-order term ϵ^2 which is asymptotically dominated, we have

$$\left| \frac{1}{\sigma_j \sigma_k} - \frac{1}{\widetilde{\sigma}_j \widetilde{\sigma}_k} \right| \leq \frac{2k\sqrt{D}(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau)}{C_2^2(C_2 - \sqrt{kD}(\frac{\sigma}{\tau} + \frac{1}{2}C_1\tau))^2}.$$

□

REFERENCES

[1] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*. Hoboken: John Wiley and Sons, 3rd ed., 2003.

[2] Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., and van der Vorst, H., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. Philadelphia: Society for Industrial and Applied Mathematics, 2000.

[3] Belkin, M. and Niyogi, P., "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[4] Brand, M., "Charting a manifold," in *Neural Information Processing Systems*, vol. 15, Mitsubishi Electric Research Labs, MIT Press, March 2003.

[5] Cook, R. D. and Ni, L., "Sufficient dimension reduction via inverse regression: a minimum discrepancy approach," *Journal of the American Statistical Association*, vol. 100, pp. 410–428, June 2005.

[6] Donoho, D. L. and Grimes, C. E., "Hessian eigenmaps: New locally linear embeddimg techniques for high-dimensional data," *Proceedings of the National Academy of Arts and Sciences*, vol. 100, pp. 5591–5596, 2003.

[7] Golub, G. H. and van Loan, C. F., *Matrix computations*. Baltimore: Johns Hopkins University Press, 3rd ed., 1996.

[8] Hansen, M. H. and Yu, B., "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[9] Huo, X. and Chen, J., "Local linear projection (LLP)," in *First IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, (Raleigh, NC), October 2003. http://www.gensips.gatech.edu/proceedings/.

[10] Huo, X. and Chen, J., "Detecting the presence of an inhomogeneous region in a homogeneous background: taking advantages of the underlying geometry via manifolds," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, (Montreal Quebec, CANADA), May 2004.

[11] Huo, X., Ni, X. S., and Smith, A. K., "A survey of manifold-based learning methods," in *Recent Advances in Data Mining of Enterprise Data* (Liao, T. W. and Triantaphyllou, E., eds.), pp. 691–745, Singapore: World Scientific, 2007.

[12] Huo, X. and Smith, A. K., "Performance analysis of a manifold learning algorithm in dimension reduction," tech. rep., Georgia Institute of Technology, March 2006. http://www2.isye.gatech.edu/statistics/papers/06-06.pdf.

[13] Ji, X. and Zha, H., "Sensor positioning in wireless ad-hoc sensor networks with multidimensional scaling," in *Proceedings of IEEE INFOCOM*, pp. 2652–2661, 2004.

[14] Kleywegt, A. J., Shapiro, A., and Homem-De-Mello, T., "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2001.

[15] Lancaster and Tismenetsky, *The theory of matrices, with applications.* New York: Academic Press, 1985.

[16] Li, B., Zha, H., and Chiaromonte, F., "Contour regression: a general approach to dimension reduction," *Annals of Statistics*, vol. 33, pp. 1580–1616, August 2005.

[17] Mira Bernstein, Vin de Silva, J. C. L. and Tenenbaum, J. B., "Graph approximations to geodesics on embedded manifolds." 2000.

[18] Roweis, S. T. and Saul, L. K., "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[19] Rust, R. T., Simester, D., Brodie, R. J., and Nilikant, V., "Model selection criteria: an investigation of relative accuracy, posterior probabilities, and combination of criteria.," *Management Science*, vol. 41, no. 2, pp. 322–333, 1995.

[20] Saul, L. K. and Roweis, S. T., "An introduction to locally linear embedding." http://www.cs.toronto.edu/∼roweis/lle/publications.html (Date accessed: March 2005).

[21] Shapiro, A. and Ruszczynski, A., "Lectures on stochastic programming." http://www2.isye.gatech.edu/∼ashapiro/download.php?Down=book (Date accessed: March 2008).

[22] Sibson, R., "Studies in the robustness of multidimensional scaling: procrustes statistics," *J. Roy. Statist. Soc. Ser. B*, vol. 40, no. 2, pp. 234–238, 1978.

[23] Sibson, R., "Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling," *J. Roy. Statist. Soc. Ser. B*, vol. 41, no. 2, pp. 217–229, 1979.

[24] Stewart, G. W., "Perturbation bounds for the QR factorization of a matrix," *SIAM Journal on Numerical Analysis*, vol. 14, no. 3, pp. 509–518, 1977.

[25] Stewart, G. W. and Sun, J.-G., *Matrix Perturbation Theory.* Boston, MA: Academic Press, 1990.

[26] Tenenbaum, J. B., de Silva, V., and Langford, J. C., "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[27] Wittman, T., "MANIfold learning Matlab demo." URL: http://www.math.umn.edu/∼wittman/mani/index.html (Date accessed: March 2008).

[28] Zha, H. and Zhang, Z., "Spectral analysis of alignment in manifold learning," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[29] Zhang, Z. and Zha, H., "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM Journal of Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.