Geometric Sparsity in High Dimension

by

Daniel N. Kaslovsky

B.A., Colgate University, 2003

M.S., University of Colorado, 2009

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2012

This thesis entitled:
Geometric Sparsity in High Dimension
written by Daniel N. Kaslovsky
has been approved for the Department of Applied Mathematics

_____

Prof. François G. Meyer

_____

Prof. James H. Curry

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Kaslovsky, Daniel N. (Ph.D., Applied Mathematics)

Geometric Sparsity in High Dimension

Thesis directed by Prof. François G. Meyer

While typically complex and high-dimensional, modern data sets often have a concise underlying structure. This thesis explores the sparsity inherent in the geometric structure of many high-dimensional data sets.

Constructing an efficient parametrization of a large data set of points lying close to a smooth manifold in high dimension remains a fundamental problem. One approach, guided by geometry, consists in recovering a local parametrization (a chart) using the local tangent plane. In practice, the data are noisy and the estimation of a low-dimensional tangent plane in high dimension becomes ill posed. Principal component analysis (PCA) is often the tool of choice, as it returns an optimal basis in the case of noise-free samples from a linear subspace. To process noisy data, PCA must be applied locally, at a scale small enough such that the manifold is approximately linear, but at a scale large enough such that structure may be discerned from noise.

We present an approach that uses the geometry of the data to guide our definition of locality, discovering the optimal balance of this noise-curvature trade-off. Using eigenspace perturbation theory, we study the stability of the subspace estimated by PCA as a function of scale, and bound (with high probability) the angle it forms with the true tangent space. By adaptively selecting the scale that minimizes this bound, our analysis reveals the optimal scale for local tangent plane recovery. Additionally, we are able to accurately and efficiently estimate the curvature of the local neighborhood, and we introduce a geometric uncertainty principle quantifying the limits of noise-curvature perturbation for tangent plane recovery. An algorithm for partitioning a noisy data set is then studied, yielding an appropriate scale for practical tangent plane estimation.

Next, we study the interaction of sparsity, scale, and noise from a signal decomposition perspective. Empirical Mode Decomposition is a time-frequency analysis tool for nonstationary

data that adaptively defines modes based on the intrinsic frequency scales of a signal. A novel understanding of the scales at which noise corrupts the otherwise sparse frequency decomposition is presented. The thesis concludes with a discussion of future work, including applications to image processing and the continued development of sparse representation from a geometric perspective.

Dedication

Dedicated to my family, whose unwaivering and unconditional support is the foundation for any accomplishment I may claim.

Acknowledgements

Any list of acknowledgements for this thesis must start with Prof. François Meyer who has advised me for the past four years. The work contained in this thesis is a reflection of the countless hours he has selflessly spent with me. In addition to learning mathematics, I have learned a great deal about character and professionalism from his example. Prof. Meyer has been a role model, mentor, and friend. Thank you.

The department of Applied Mathematics has been a second home (first home?) for the past five years. The list of people to thank is far too long for this format. I would, however, like to acknowledge Prof. James Curry for his support and advice that have come in many forms.

I am grateful to the members of my thesis committee for their time and efforts in seeing this thesis through to completion.

Thank you to Brendt Wohlberg of Los Alamos National Laboratory for hosting me as a graduate research assistant during the summer of 2010 and for continued collaboration.

I have been fortunate to have received financial support during my graduate career from the National Science Foundation (NSF), the NSF-IGERT in Computational Optical Sensing and Imaging (COSI), and the Department of Applied Mathematics. I am grateful this support.

Finally, thank you to many of my fellow graduate students in Applied Mathematics, or more appropriately, my friends. Those who have helped me (on the whiteboard and in life) do not need to be listed here; they know who they are.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Massive data sets are now commonplace in both science and society. For example, various types of data are created or collected by a wide range of scientific disciplines and applications (e.g., genomics, astrophysics, internet and network analysis, physical simulation models), industry (e.g., inventory databases, consumer behavior tracking), and every day social and societal interactions (e.g., web searches, medical records, picture sharing). Ever-increasing computational power and storage capacity both facilitate the creation of and necessitate new algorithms for such data. While typically complex and high-dimensional, modern data sets often have a concise underlying structure. This thesis explores the sparsity inherent in the geometric structure of many high-dimensional data sets.

1.1 Geometric Sparsity

A signal or datum is considered to be sparse if it depends on fewer degrees of freedom than the dimension in which it is observed. This is to say that a signal's information content may be described in a concise manner, often via a suitable transform. Fortunately, sparsity is inherent in a wide range of data. For example, while image pixels are typically nonzero, the majority of an image's wavelet coefficients are very close to zero. As most of the image information (or signal energy) is carried in only a small number of coefficients, images are effectively sparse in the wavelet domain and are therefore compressible (as is done by the JPEG-2000 coding standard). Wavelet sparsity has been an active area of study during recent decades (see [58] for a thorough reference)

and is one particular example of "transform sparsity." More generally, consider a discrete signal $x \in \mathbb{R}^D$ and its coefficients in a transform basis. If x has only $d < D$ large coefficients and its other $(D-d)$ coefficients are small, x is said to be sparse and retaining only those large coefficients yields an accurate and efficient representation.

This traditional notion of sparsity has been the focus of much recent attention and excitement. It has been shown that the sparse signal $x \in \mathbb{R}^D$, with only d nonzero coefficients in some transform basis, can be exactly reconstructed from only $\mathcal{O}(d \log(D/d)) \ll D$ measurements [14, 21]. This idea underlies the exciting new field of compressive sensing and has prompted recent work on the computational aspects of sparse recovery [15, 64, 78]. Other approaches adaptively design overcomplete dictionaries for which each point in a data set may be represented by a small number of dictionary elements, yielding a sparse representation (see [26] and the references therein). State of the art denoising algorithms have been demonstrated using such adaptive sparse techniques [1]. The common theme of such recent advances is the efficiency afforded by the underlying sparsity of many large and complex data sets.

Just as sparsity implies a concise data representation for high-dimensional signals, other low-dimensional data models exist. Consider a collection of sufficiently sampled human vocal signals as a set of points in a high-dimensional vector space. While such signals are band-limited by the sampling process, the frequency range of human speech is in fact far more constrained by physiology. Thus, these points do not occupy all of Fourier space but instead are confined to only a small subset of this domain. As the information content is much smaller than the dimension of the ambient space, there exists an inherent sparse representation for such data.

The few, often unobservable, degrees of freedom inherent in a data set provide a low-dimensional parameterization of the data. Geometrically, this parameterization describes a (sub)-manifold embedded in the ambient space. While measured in high dimension, the data are thought to be confined to this low-dimensional structure, yielding a sparse geometric representation. For example, high contrast optical images and collections of human face images have been shown to organize about low-dimensional manifolds despite each image (consisting of $N \times N$ pixels) being

observed in N^2 dimensions [13, 53]. The geometry of this underlying manifold provides both a concise and information-rich description of the data.

Parameterization of a data set via its geometric structure has been an active area of research over the past decade (see, for example, [7, 18, 77]). Even more recently, work has emerged [3, 4, 5] linking traditional forms of sparsity (described above) with data models such as manifolds, unions of subspaces, and point clouds. In this thesis, we consider the geometry of manifold-valued data. Geometric sparsity therefore expresses the idea that manifold-valued data points observed in high dimension may be well represented using many fewer coordinates than those of the observed ambient space. While several algorithms have been developed to estimate these intrinsic manifold coordinates, the approach taken in this work is to use the coordinates of the local tangent space to the manifold (see the discussion in Section 1.2). In a manner analogous to the linearization of a function, points on a Riemannian manifold $\mathcal{M}$ in a local neighborhood of a reference point x_0 are well represented by the basis vectors spanning the tangent space of $\mathcal{M}$ at x_0 ($T_{x_o}\mathcal{M}$). Figure 1.1 provides an illustration. The dimension, d, of the local tangent space ("tangent plane") indicates the intrinsic dimension of the manifold, and is typically much smaller than the dimension, D, of the ambient space. Viewed through the lens of traditional sparse representation, points in this local neighborhood are well approximated via linear projection onto only a small number of basis vectors.

1.2 Recovery of Manifold Geometry

1.2.1 Global vs. Local

Large data sets of points in high-dimension often lie close to a smooth low-dimensional manifold. A fundamental problem in processing such data sets is the construction of an efficient parameterization that allows for the data to be well represented in fewer dimensions. Such a parameterization may be realized by exploiting the inherent manifold structure of the data. The past decade has seen the introduction of many important algorithms for learning manifold parameteriza-

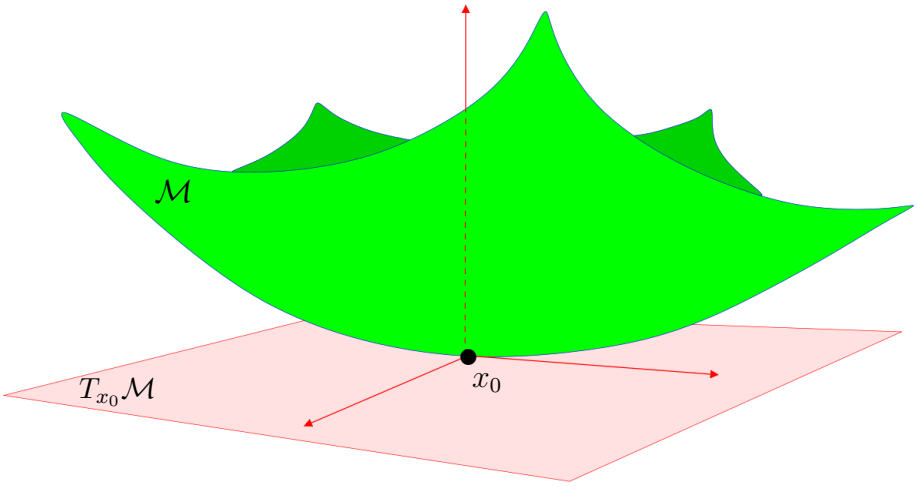


Figure 1.1: The tangent plane $T_{x_0}\mathcal{M}$ provides a linear approximation to $\mathcal{M}$ in a local neighborhood about x_0.

tions ("manifold learning"), including Diffusion Maps [18], Hessian Eigenmaps [22], ISOMAP [77], Laplacian Eigenmaps [7], and Local Linear Embedding [67]. However, recovering the geometry of an underlying manifold from only noisy samples remains an open topic of research.

The case of data sampled from a linear subspace is well studied (see [42, 45, 63], for example). The optimal parameterization is given by principal component analysis (PCA), as the singular value decomposition (SVD) produces the best low-rank approximation for such data. However, most interesting manifold-valued data organize on or near a nonlinear manifold. PCA, by projecting data points onto the linear subspace of best fit, is not optimal in this case, as curvature may only be accommodated by choosing a subspace of dimension higher than that of the manifold. Algorithms designed to process nonlinear data sets typically proceed in one of two directions. One approach is to consider the data globally and produce a nonlinear embedding. Alternatively, the data may be considered in a piecewise-linear fashion and linear methods such as PCA may be applied locally.

Unlike PCA, nonlinear global methods can accommodate the curvature of the manifold. However, global parameterizations are not without important drawbacks. By selecting an embedding according to a global figure-of-merit, such methods sacrifice local accuracy to obtain a global optimum. For example, the embedding produced by Laplacian Eigenmaps [7] is chosen to minimize a cost function that penalizes for distorting local mutual distances over the entire data set. Minimizing such a function guarantees that mutual distances are approximately preserved, but cannot provide an estimate as to the distortion between any given pair of points. Distances may be well preserved in one neighborhood at the cost of distorting those in another. Similarly, the ISOMAP algorithm [77] produces a global embedding that lacks a control of the local approximation error. Such global methods may not scale to accommodate large data sets. By operating on the entire data set as a whole, global methods give rise to extremely large matrices and the computational burden can become prohibitively large (but see [52] for developments addressing such computational issues). These difficulties are not specific to the cited examples but are instead typical of global embeddings.

The approach in this thesis consists in recovering local approximations from neighborhoods of a data set organized about a Riemannian manifold. Considering the data to be piecewise-linear is to say that locally, the underlying manifold is similar to Euclidean space and is therefore well approximated by a linear subspace. In each neighborhood, we may exploit the fact that PCA returns an optimal basis for a linear subspace. Maintaining a geometric perspective, when oriented at the appropriate location on the manifold, this subspace may be thought of as a tangent plane, and its basis provides a chart from the manifold to Euclidean space. Such a tangent plane provides the best linear approximation to the given local neighborhood of the manifold. Each approximation may therefore be constructed in a manner that respects the local geometry of the data in order to provide (often with high probability) low-distortion error bounds on the local scale. The collection of all such charts yields a covering of the data set and an efficient covering will use as few charts as possible. To minimize the number of charts, we may ask that each tangent plane covers as large a neighborhood as allowed by the local geometry.

There have been several versions of localized PCA for tangent plane recovery proposed in the literature. While the need for locality has been acknowledged, a precise treatment of the size of such neighborhoods is often not addressed. The appropriate neighborhood size must be a function of intrinsic (manifold) dimensionality, curvature, and noise level. Despite the fact that these properties may change as different regions of the manifold are explored, locality is often defined via an *a priori* fixed number of neighbors or as the output of an algorithm. For example, before using PCA for dimensionality reduction, Kambhatla and Leen [46] partition data into local regions via vector quantization, and the size of any neighborhood is thus a function of the clustering algorithm's distortion function. The Local Tangent Space Alignment (LTSA) algorithm of Zhang and Zha [83] defines neighborhoods for local PCA using a fixed number of points and tangent spaces are organized into a global coordinate system via affine transformations. Brand [10] proposes a similar method, but defines locally linear neighborhoods by tracking the growth rate of the number of points falling in a ball of increasing radius. The Local Linear Embedding (LLE) algorithm of Roweis and Saul [67] uses neighborhoods of fixed size to construct a global coordinate system

from linear charts. Bengio and Monperrus [8] train a nonlocal manifold "prediction" function by gathering a fixed number of points to span noisy estimates of tangent spaces. Yang [81] introduces a localized version of the classic multidimensional scaling algorithm, covering a data set with local neighborhoods of a fixed size and, noting that the algorithm's performance depends on the definition of these neighborhoods, suggests doing so adaptively. Ohtake and coauthors [65], working in image space rather than feature space, adaptively define some neighborhood parameters while leaving others fixed. Lin and Zha [54] cleverly define neighborhoods for dimensionality estimation, but resort to the PCA of a fixed number of points when building a basis for a local tangent space. Addressing the fundamental issue of neighborhood selection, they note that adding more points may increase stability, but this may come at the price of accuracy.

A main focus of this thesis is an analysis of the optimal neighborhood size, or scale, at which to estimate a local tangent plane in the presence of noise (Chapter 3). As noted above, the optimal scale must reach a balance between the curvature of the manifold and the noise that perturbs the data. The trade-off between noise and curvature is a key aspect of our analysis and a recurring theme of this thesis. The next subsection provides geometric intuition.

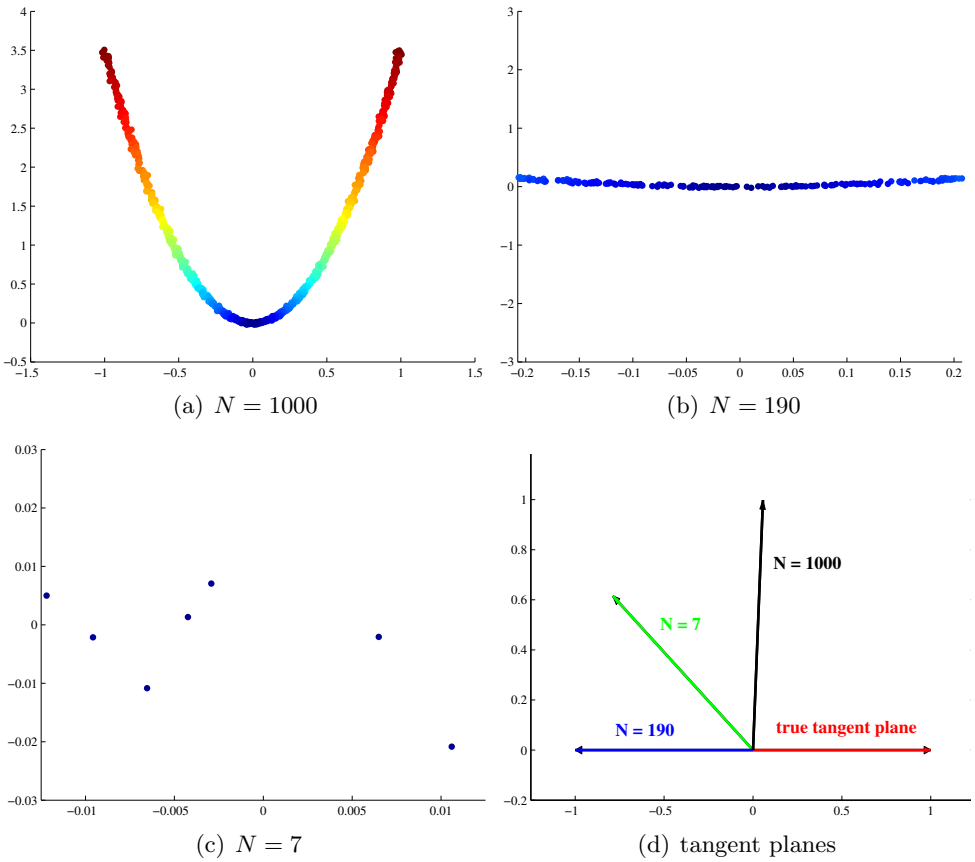
1.2.2 The Noise-Curvature Trade-off

Consider a d-dimensional linear subspace from which N data points have been sampled. Note that the points are observed in the ambient dimension D. Arranging the data points as the columns of a matrix, the top d PCA basis vectors (those associated with the d largest eigenvalues of the sample covariance matrix) provide the best (least-square) approximation to the tangent plane. In fact, given $N = d + 1$ points, we exactly reconstruct the original subspace. Next, given the same linear subspace, let each sample point be perturbed by Gaussian noise in each of its D coordinates. The data points are no longer true samples from the linear subspace, but instead are organized near the subspace. We may recover an approximation to the subspace by again performing PCA. The PCA subspace provides the best possible approximation, and the quality of approximation increases as more points are included. We would therefore wish to include as many points as possible.

Now consider points sampled from the nonlinear d-dimensional manifold $\mathcal{M}$ and the goal is to recover the tangent space of $\mathcal{M}$ at reference point x_0. If we were to proceed in a global manner and perform PCA on the entire data set, curvature would force our linear approximation to use more dimensions than the intrinsic dimensionality of the manifold to capture the nonlinearity of the data. Noting that $\mathcal{M}$ locally resembles d-dimensional Euclidean space $\mathbb{R}^d$, there exists a local scale about x_0 such that the effects of curvature are small (figure 1.2-a and 1.2-b). If we perform PCA at such a small scale, we may recover a good approximation to the linear tangent space of $\mathcal{M}$ at x_0 (figure 1.2-d). In fact, in such a setting, the approximation improves as the scale becomes smaller.

Finally, add noise to these sample points and consider them to be organized near $\mathcal{M}$. As in the linear example, we wish to include as many points as possible to overcome the effects of noise and improve the quality of approximation. However, the curvature of the manifold prevents the inclusion of a large number of points, as we wish to approximate a linear subspace. This linear requirement suggests allowing only a very small radius about x_0, yet at small scales, the sample points are indistinguishable from noise (figure 1.2-c). We therefore seek a balance and assume there exists a scale large enough to be above the noise level, but still small enough to avoid curvature. This scale reveals a linear structure that is sufficiently decoupled from both the noise and the curvature to be well approximated by a tangent plane. Figure 1.2 illustrates this trade-off between noise and curvature.

Remark. This concept of optimal trade-off bears resemblance to that of the bias-variance trade-off in the context of nonparametric density estimation, where a kernel or fixed number of neighbors is used to estimate an unknown underlying density function from local data. A fundamental problem in this context is the selection of the optimal kernel bandwidth or the optimal number of neighbors to be used. The estimator's mean-squared-error (MSE) may be expressed as a sum of the estimator bias and the estimator variance. Noting that the bias becomes large for large bandwidths (or large number of neighbors) while the variance is large for small bandwidths (or small number of neighbors), the MSE is minimized by balancing these two effects. By analogy,



(a) $N = 1000$ (b) $N = 190$

(c) $N = 7$ (d) tangent planes

Figure 1.2: A noisy data set composed of points sampled from a 1-dimensional manifold is presented at different scales (note the change in the scale of the axes between each plot). (a) All $N = 1000$ points are shown and it is clear that curvature prevents an accurate linear approximation. (b) At this scale ($N = 190$) the manifold is nearly linear and an accurate tangent plane estimate may be recovered. (c) At this very small scale ($N = 7$) the data points are indistinguishable from noise and the tangent plane estimate may be oriented in any random direction. (d) The tangent plane estimates are shown at the three demonstrated scales: at the large scale ($N = 1000$) the estimated tangent plane is almost completely orthogonal to the true tangent plane; at the intermediate scale ($N = 190$) the estimated and true tangent planes are aligned; at the small scale ($N = 7$) noise dominates the estimation and orients the tangent plane in a random direction.

curvature and noise contribute to estimator bias and estimator variance, respectively. To estimate the tangent plane with minimal error, we seek the scale that optimally balances these two effects.

1.3 Organization and Original Contribution

The organization of the remainder of this thesis is as follows. Chapter 2 outlines the main theoretical tools used in the subsequent chapters. The analysis of optimal tangent plane recovery in the presence of noise is the subject of Chapter 3. Chapter 4 studies the partitioning of a noisy data set and presents a geometric analysis of an algorithm to find an appropriate scale for practical tangent plane recovery. In Chapter 5, the focus shifts to a study of the interaction of sparsity, scale, and noise from a signal decomposition point of view. Empirical Mode Decomposition (EMD) is an adaptive time-frequency analysis tool for nonstationary data [40]. Rather than producing a decomposition using fixed projections onto the Fourier basis, EMD decomposes a signal into adaptively defined modes representing the intrinsic frequency scales of the signal. In this way, EMD produces a sparse decomposition in the frequency domain. A novel understanding of the scales at which noise interference corrupts this decomposition is developed in this chapter. The thesis concludes in Chapter 6 with a discussion of future work, including applications to image processing and the continued development of sparse representation from a geometric perspective.

The contents of the thesis draw from original analyses and results that are both published and in preparation. The main results on tangent plane estimation (Chapter 3) are found in:

[48] D.N. Kaslovsky and F.G. Meyer. Optimal tangent plane recovery from noisy manifold samples. Submitted to Annals of Statistics, 57 pages, 2011.

[50] D.N. Kaslovsky and F.G. Meyer. Overcoming noise, avoiding curvature: optimal scale selection for tangent plane recovery. In Proceedings of IEEE Conference on Statistical Signal Processing, August, 2012.

The results on partitioning a noisy data set for tangent plane estimation (Chapter 4) are part of a manuscript in preparation:

[49] D.N. Kaslovsky and F.G. Meyer. Estimating local manifold geometry via data partitioning. In preparation, 2012.

Finally, the results on noise corruption of EMD (Chapter 5) are found in:

[47] D.N. Kaslovsky and F.G. Meyer. Noise corruption of Empirical Mode Decomposition and its Effect on Instantaneous Frequency. Advances in Adaptive Data Analysis, 2:373–396, 2010.

Chapter 2

Mathematical Tools

The main mathematical tools used for the analysis in this thesis are reviewed below. Standard references are noted where appropriate.

2.1 Principal Component Analysis

Consider a $D \times N$ data matrix X with its columns holding N points in D dimensions. Given a target dimension $d < D$, Principal Component Analysis (PCA) finds the rank d linear approximation that best represents the data. More precisely, over all possible rank d linear approximations, the PCA approximation retains the maximum amount of variance in the data. PCA, also known as the discrete Karhunen-Loeve transform, is one of the most widely used techniques for dimensionality reduction. As many standard PCA references exist, we follow the description given in [37] and the reader is referred to [43] for an entire text devoted to the topic.

PCA models the data in X with a rank d affine subspace ("hyperplane") given by

$$f(y) = \mu + Qy, \tag{2.1.1}$$

where $\mu \in \mathbb{R}^D$ specifies an offset vector, the $D \times d$ matrix Q has orthonormal columns that span the affine subspace, and $y \in \mathbb{R}^d$ is a vector of coefficients. These parameters of the PCA subspace are chosen to minimize the mean squared error (MSE) of the approximation:

$$\min_{\mu,Q,y} \sum_{i=1}^{N} \|x_i - \mu - Qy_i\|^2, \tag{2.1.2}$$

where $x_i \in \mathbb{R}^D$ is the ith data point (ith column of X). Optimizing for μ and the y_i yields

$$\mu = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$y_i = Q^T(x_i - \overline{x})$$

Then (2.1.2) has the form

$$\min_{Q} \sum_{i=1}^{N} \left\| (x_i - \overline{x}) - QQ^T(x_i - \overline{x}) \right\|^2. \tag{2.1.3}$$

Recognizing that QQ^T is the orthogonal projector onto the d-dimensional subspace spanned by the columns of Q, the solution is given by the singular value decomposition (SVD) of the centered, normalized, data matrix

$$\frac{1}{\sqrt{N}}\widetilde{X} = \frac{1}{\sqrt{N}}(X - \overline{X}),$$

where $\overline{X}$ is the $D \times N$ matrix with N copies of $\overline{x}$ as its columns. Writing the SVD of $\widetilde{X}/\sqrt{N}$ as

$$\frac{1}{\sqrt{N}}\widetilde{X} = U\Sigma V^T, \tag{2.1.4}$$

equation (2.1.3) is minimized by choosing Q to be the first d columns of U.

Note that an equivalent solution to (2.1.3) may be found by instead using the eigendecomposition of the centered sample covariance matrix

$$\frac{1}{N}\widetilde{X}\widetilde{X}^T = \frac{1}{N}(X - \overline{X})(X - \overline{X})^T = U\Lambda U^T,$$

where $\Lambda = \Sigma^2$. The entries of Λ (the eigenvalues of XX^T/N) correspond to the variance in each principal direction (each column of U) and it can be shown that this construction maximizes the variance captured by the d-dimensional linear approximation.

The PCA algorithm therefore consists of the following steps:

(1) Compute the center of the data: $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

(2) Compute the SVD of the centered, normalized, data matrix: $\frac{1}{\sqrt{N}}\widetilde{X} = \frac{1}{\sqrt{N}}(X - \overline{X}) = U\Sigma V^T$. Store the first d columns of U in the matrix Q.

(3) Project each centered data point onto the subspace spanned by the columns of Q and add back the offset $\overline{x}$. Letting Z denote the PCA approximation of the data set,

$$Z = QQ^T(X - \overline{X}) + \overline{X}.$$

2.2 Subspace Perturbation

2.2.1 The Angle Between Subspaces

The geometric concept of the angle between two subspaces provides a measure of similarity (or dissimilarity) for the two subspaces. The principal angles θ_k between subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ of respective dimensions d_1 and d_2 are defined [35] by

$$\cos(\theta_k) = \max_{u \in \mathcal{S}_1} \max_{v \in \mathcal{S}_2} \; u^T v = u_k^T v_k \tag{2.2.1}$$

subject to

$$\|u\| = \|v\| = 1$$

$$u^T u_i = 0, \quad 1 \leq i \leq k-1$$

$$v^T v_i = 0, \quad 1 \leq i \leq k-1.$$

In this thesis we quantify the angle between two subspaces by the largest such principal angle and we will consider $d_1 = d_2$ such that $\mathcal{S}_1$ and $\mathcal{S}_2$ are equidimensional subspaces.

Given a d-dimensional subspace $\mathcal{S}$ spanned by the orthonormal basis $[u_1 \; u_2 \; \ldots \; u_d]$, the unique orthogonal projection onto S is given by $P = UU^T$, where U is the matrix with the u_j as its columns. Let $P_1 = U_1 U_1^T$ and $P_2 = U_2 U_2^T$ be the respective orthogonal projectors onto $\mathcal{S}_1$ and $\mathcal{S}_2$. Then the spectral norm of the difference between the orthogonal projectors yields the sine of the largest principal angle between $\mathcal{S}_1$ and $\mathcal{S}_2$

$$\sin\Theta = \|P_1 - P_2\|_2. \tag{2.2.2}$$

In this equidimensional setting, Θ quantifies the distance between the subspaces. Reference [35] (sections 2.6.3 and 12.4.3) provides more details.

2.2.2 Invariant Subspaces

The following definitions are found in [74] to which the reader is referred for further discussion. A subspace $\mathcal{X}$ is an invariant subspace of a matrix A if $A\mathcal{X} \subset \mathcal{X}$. Let the columns of X form a basis for invariant subspace $\mathcal{X}$. Then there is a unique matrix L such that

$$AX = XL \tag{2.2.3}$$

and L is the representation of A on $\mathcal{X}$ (with respect to the basis X). Further, the eigenvalues of L are eigenvalues of A.

The following theorem (V.1.1 of [74]) provides a characterization of invariant subspaces. The theorem is stated here using similar notation to that of [74]. In particular, $\mathcal{R}(X)$ denotes the column space of X and $\mathcal{R}(X)^{\perp}$ denotes the orthogonal complement of $\mathcal{R}(X)$.

Theorem 1. *Let the columns of X be linearly independent and let the columns of Y span $\mathcal{R}(X)^{\perp}$. Then $\mathcal{R}(X)$ is an invariant subspace of A if and only if*

$$Y^T AX = 0. \tag{2.2.4}$$

Let X and Y be as given above and let $[X\ Y]$ be a unitary matrix. We then have the following representation of A:

$$[X\ Y]^T A[X\ Y] = \begin{bmatrix} X^T AX & X^T AY \\ Y^T AX & Y^T AY \end{bmatrix} = \begin{bmatrix} L_1 & H \\ 0 & L_2 \end{bmatrix} \tag{2.2.5}$$

with

$$L_1 = X^T AX$$

$$L_2 = Y^T AY$$

$$H = X^T AY.$$

Thus L_1 is the representation of A on $\mathcal{X}$ (with respect to X) and the eigenvalues of L_1 are those of A associated with $\mathcal{X}$. Finally, $\mathcal{X}$ is said to be a simple invariant subspace of A if $\lambda(L_1) \cap \lambda(L_2) = \emptyset$, where $\lambda(M)$ denotes the set of eigenvalues of matrix M.

2.2.3 Perturbation of Invariant Subspaces

Consider a matrix A and an invariant subspace of A spanned by the columns of the matrix U_1. Let Δ be a perturbation such that $\widehat{A} = A + \Delta$. We wish to quantify by how much the perturbation Δ has rotated the invariant subspace U_1 of A. More precisely, we wish to bound, in terms of Δ, the angle between U_1 and the corresponding invariant subspace $\widehat{U}_1$ of $\widehat{A}$. The classic results of Davis and Kahan [20] provide bounds on trigonometric functions of this angle. Their theorems rely on two quantities: a residual in the form of the difference between the perturbed matrix $\widehat{A}$ restricted to the subspace U_1 and the representation of A in U_1; and either a spectral gap in A or a spectral gap between the representations of A and Δ in U_1.

We use a theorem due to Stewart (Theorem V.2.7 of [74], see also [73] for a detailed discussion), originally posed as a generalization of the Davis-Kahan $\sin\Theta$ theorem to the non-Hermitian setting. Applying this theorem to the Hermitian matrices $\widehat{A}\widehat{A}^T$, AA^T, and $\Delta\Delta^T$, and using the Frobenius norm ($\|M\|_F = \sqrt{\text{trace } M^T M}$) yields a simplified version that most efficiently facilitates the analysis in the chapters to follow. We now state the theorem in the form in which it is used.

Theorem 2 (Davis & Kahan [20], Stewart [74])**.** *Let $U = [U_1\ U_2]$ be unitary with the columns of U_1 spanning a simple invariant subspace of Hermitian matrix A such that*

$$[U_1\ U_2]^T\ A\ [U_1\ U_2] = \begin{bmatrix} L_1 & H \\ 0 & L_2 \end{bmatrix}. \tag{2.2.6}$$

Given a (Hermitian) perturbation Δ, let $\widehat{A} = A + \Delta$ and $\widehat{U}_1$ be the invariant subspace of $\widehat{A}$ corresponding to U_1. Let P and $\widehat{P}$ be the orthogonal projectors onto U_1 and $\widehat{U}_1$, respectively.

If

$$\delta \ = \ \min|\lambda(L_1) - \lambda(L_2)| - \|U_1^T \Delta U_1\|_F - \|U_2^T \Delta U_2\|_F \ > \ 0 \tag{2.2.7}$$

and

$$\frac{\|U_2^T \Delta U_1\|_F \left(\|H\|_F + \|U_1^T \Delta U_2\|_F\right)}{\delta^2} \ < \ \frac{1}{4} \tag{2.2.8}$$

where $\lambda(M)$ denotes the set of eigenvalues of matrix M, then

$$\|P - \widehat{P}\|_F \ \leq \ 2\sqrt{2}\frac{\|U_2^T \Delta U_1\|_F}{\delta}. \tag{2.2.9}$$

This theorem bounds the sine of the angle between the invariant subspaces spanned by the columns of U_1 and $\widehat{U}_1$. The numerator of (2.2.9) corresponds to the norm of the residual considered by Davis and Kahan and the Frobenius norm provides a simplification of the denominator in the Hermitian setting (see V.3.1 of [74]). Further, for the analysis of the chapters to follow we have $H = 0$ as (2.2.6) will be the eigendecomposition of A. The geometric interpretation of the bound (2.2.9) is a focus of Chapter 3.

2.3 Concentration Inequalities

Concentration inequalities are used to bound random variables, or functions of random variables, about a constant value usually associated with the mean of the distribution. The term "concentration" implies that the probability of such a random variable deviating from this constant decays exponentially with the size of the deviation. Standard results used throughout this thesis are reviewed below and may be found in much greater detail in references such as [59].

Two results on the concentration of Gaussian measure play important roles in the analysis to follow. The Gaussian measure γ_D on $\mathbb{R}^D$ with mean μ and variance σ^2 has density

$$\frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right).$$

Our first concentration result expresses the fact that a Lipschitz function rarely deviates from its mean (or median). A function $f : \mathbb{R}^D \to \mathbb{R}$ is L-Lipschitz (with Lipschitz constant $L > 0$) if

$$|f(x) - f(y)| \leq L\|x-y\| \quad \text{for all } x, y \in \mathbb{R}^D.$$

Theorem 3 (Concentration of Lipschitz functions)**.** *Let $f : \mathbb{R}^D \to \mathbb{R}$ be L-Lipschitz and let M denote the mean or median of f with respect to the standard Gaussian measure on $\mathbb{R}^D$. Then we have*

$$\text{Prob}\left[|f - M| \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2L^2}\right),$$

where Prob *denotes the standard Gaussian measure on $\mathbb{R}^D$.*

Many of the matrices we will analyze will have entries that are Lipschitz functions of Gaussian random variables. This result bounds the entries of such matrices with high probability.

Next, consider a random vector drawn from the $\mathcal{N}\left(0, \sigma^2 I_D\right)$ distribution, where I_D denotes the identity matrix of order D. We may use the concentration of Lipschitz functions to derive a result on the concentration of the norm of such vectors. Begin by noting that the norm $\|x\| = \left(\sum_{i=1}^{D} x_i^2\right)^{1/2}$ is a 1-Lipschitz function. We also have that the random variable $\|x\|/\sigma$ follows a χ distribution with D degrees of freedom. The mean (expectation) of this distribution is given by

$$\mathbb{E}\left[\|x\|/\sigma\right] = \sqrt{2}\frac{\Gamma\left((D+1)/2\right)}{\Gamma\left(D/2\right)}.$$

Therefore, by Theorem 3, we have that

$$\text{Prob}\left[\left|\frac{\|x\|}{\sigma} - \sqrt{2}\frac{\Gamma\left((D+1)/2\right)}{\Gamma\left(D/2\right)}\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2}\right). \tag{2.3.1}$$

Using Stirling's approximation, the mean $\mathbb{E}\left[\|x\|/\sigma\right]$ has the form

$$\sqrt{2}\frac{\Gamma\left((D+1)/2\right)}{\Gamma\left(D/2\right)} = \sqrt{D} + \mathcal{O}\left(\frac{1}{\sqrt{D}}\right). \tag{2.3.2}$$

Thus (2.3.1) expresses the fact that a random vector x drawn from the $\mathcal{N}(0, \sigma^2 I_D)$ distribution has norm $\|x\| = \sigma\sqrt{D}$ with extremely high probability. Note that the $\mathcal{O}\left(\frac{1}{\sqrt{D}}\right)$ term in (2.3.2) will be at least an order of magnitude smaller than $\sqrt{D}$. Further, this term is an artifact of approximating the Gamma function and plays no role in the concentration of the norm. We therefore neglect this small term to arrive at a standard result (see [36], for example):

Theorem 4. *Let S be the set such that*

$$S = \{x \in \mathbb{R}^D : \sqrt{D}(1-\epsilon) \ \leq \ \|x\|/\sigma \ \leq \ \sqrt{D}(1+\epsilon)\}.$$

Then we have

$$\gamma_D\left(S\right) \ > \ 1 - 2e^{\frac{-D\epsilon^2}{2}}.$$

Geometrically, this result implies that points drawn from the $\mathcal{N}(0, \sigma^2 I_D)$ distribution concentrate on a sphere of radius $\sigma\sqrt{D}$. Throughout this thesis, we will assume the noise corrupting a data

set to be drawn from this distribution (although any distribution exhibiting similar concentration may be considered). Thus the noise perturbation has bounded norm over a set with very large measure and we will consider it to be concentrated in a "noise ball" of radius $\sigma\sqrt{D}$. Recognizing this quantity as the radius of the noise ball will prove useful for analysis and provide intuition.

More generally, it will be necessary to bound the norms of matrices whose entries are functions of random variables. Two standard results on the concentration of such functions are applicable. First we state the bounded difference inequality, also known as McDiarmid's Inequality:

Theorem 5 (Bounded Difference Inequality, McDiarmid)**.** *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,x_i' \in A} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \;\leq\; c_i, \quad 1 \leq i \leq n.$$

Then, for all $\epsilon > 0$,

$$\text{Prob}\left[f(X_1, \ldots, X_n) - \mathbb{E}\, f(X_1, \ldots, X_n) \;\geq\; \epsilon\right] \;\leq\; \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

For the special case of the sum of bounded random variables, Hoeffding's Inequality may be used to bound its deviation from its mean.

Theorem 6 (Hoeffding's Inequality)**.** *Let $X_1, \ldots X_n$ be independent random variables satisfying $X_i \in [a_i, b_i]$. Then for all $\epsilon > 0$,*

$$\text{Prob}\left[\left|\sum_{i=1}^n (X_i - \mathbb{E}\, X_i)\right| \geq \epsilon\right] \;\leq\; 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

The authors of [70] utilize these results to bound the deviation of a sample mean (denoted by $\widehat{\mathbb{E}}$) of a random variable from its expectation. We will use this result to bound the norm of random matrices in the analysis of Chapter 3.

Theorem 7 (Shawe-Taylor & Cristianini, [70])**.** *Given N samples of a random variable Y generated independently at random from $\mathcal{Y}$ according to the distribution P_Y, with probability at least $1 - e^{-\eta^2}$ over the choice of the samples, we have*

$$\left\|\mathbb{E}[Y] - \widehat{\mathbb{E}}[Y]\right\|_F \leq \frac{R}{\sqrt{N}}\left(2 + \eta\sqrt{2}\right)$$

where $R = \sup_{supp(P_Y)} \|Y\|_F$ *and* $supp(P_Y)$ *is the support of distribution* P_Y*.*

To apply this result to unbounded Gaussian random variables, we must restrict the analysis to a subset of $\mathbb{R}^D$ for which such random variables have bounded norm. Fortunately, doing so sacrifices very little probability as we have seen that such a set has very large measure.

Theorem 7 allows us to bound the norm of a matrix without bounding each of its entries. It therefore holds with high probability, whereas bounding the norm by simultaneously bounding each entry of a matrix requires a large union bound that sacrifices much probability. However, the deviation in Theorem 7 is controlled by the supremum of the underlying random variable over its support and is therefore not as tight of a result as is possible through other concentration inequalities. For example, in the limit of infinite sampling, one may use the Central Limit Theorem (CLT) and Gaussian tail bounds to show that the deviation is controlled by the variance of the underlying random variables, a quantity that is typically much smaller than the supremum.

Theorem 8 (Central Limit Theorem)**.** *Let* $X_1, \ldots, X_n$ *denote the observations of a random sample from a distribution that has mean* μ *and variance* σ^2*. Then the random variable*

$$Y_n = \left(\sum_{i=1}^{n} X_i - n\mu \right) / \sqrt{n}\sigma$$

converges in distribution to a random variable with a normal distribution $\mathcal{N}(0, 1)$*.*

Theorem 9 (Gaussian Tail Bound)**.** *Let* $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ *be a Gaussian random variable. Then we have*

$$\text{Prob}\left[|Y - \mu| \geq \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

While the analysis presented in Chapter 3 will use the CLT to demonstrate the tightest possible constants, bounds for which the deviation is controlled by the variance may be rigorously derived in the finite-sample setting through Bernstein-type inequalities.

Theorem 10 (Bernstein's Inequality 1)**.** *Let* $X_1, \ldots, X_n$ *be i.i.d. random variables bounded in absolute value by one, with* $\mathbb{E}\, X = 0$ *and* $\text{Var}\, X = \sigma^2$*. Then for every* $\epsilon > 0$

$$\text{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \epsilon\right] \;\leq\; 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}\epsilon}\right).$$

A version of this inequality for unbounded random variables is found in [59].

Theorem 11 (Bernstein's Inequality 2)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables and assume there exist constants v and c satisfying*

$$\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] \leq v$$

and

$$\sum_{i=1}^{n} \mathbb{E}\left[(X_i)_+^k\right] \leq \frac{k!}{2} v c^{k-2}$$

for all integers $k \geq 3$, where $(X)_+$ denotes the positive part of random variable X. Then for any $\epsilon > 0$,

$$\text{Prob}\left[\left|\sum_{i=1}^{n}(X_i - \mathbb{E}\, X_i)\right| \geq \epsilon\right] \;\leq\; 2\exp\left(-\frac{\epsilon^2}{2(v + c\epsilon)}\right).$$

Note that Bernstein's inequality has a form similar to the Gaussian tail bound but is not quite as sharp. Thus Bernstein's inequality yields slightly larger constants in the finite sample setting. For the case of $n \to \infty$, Bernstein's inequality recovers the Gaussian tail bound, as expected from the CLT.

Chapter 3

Optimal Tangent Plane Recovery From Noisy Manifold Samples

3.1 Local Tangent Plane Recovery

3.1.1 Introduction

The selection of the optimal scale, or neighborhood size, for local tangent plane recovery is the key contribution of this chapter. What is novel about our approach is that we use the geometry of the data to guide our definition of locality. On the one hand, a neighborhood must be small enough so that it is approximately linear and avoids curvature. On the other hand, a neighborhood must be be large enough to overcome the effects of noise. We use eigenspace perturbation theory to study the stability of the tangent plane as the size of the neighborhood varies. We bound, with high probability, the angle between the recovered linear subspace and the true tangent plane. In doing so, we are able to adaptively select the neighborhood that minimizes this bound, yielding the best approximate tangent plane. Further, the behavior of this bound demonstrates the non-trivial existence of such an optimal scale. We are also able to accurately and efficiently estimate the curvature of the local neighborhood. Finally, we introduce a geometric uncertainty principle quantifying the limits of noise-curvature perturbation for tangent plane recovery.

Our approach is similar to the analysis presented by Nadler in [63], who studies the finite-sample properties of the PCA spectrum. Through matrix perturbation theory, Nadler examines the angle between the leading finite-sample-PCA eigenvector and that of the leading population-PCA eigenvector. As a linear model is assumed, perturbation results from noise only. Despite this key

difference, the two analyses utilize similar techniques to bound the effects of perturbation on the PCA subspace and our results recover those of Nadler in the curvature-free setting. Nadler also reports that sample-PCA suffers from a sudden "loss of tracking" of the true dominant eigenvector due to a crossover between signal and noise eigenvalues. We demonstrate a similar phenomenon, owing to geometry rather than noise. The present work therefore generalizes the study of Nadler to noisy samples from a nonlinear manifold model.

Other recent related works include that of Singer and Wu [71], who use local PCA to build a tangent plane basis and give an analysis for the neighborhood size to be used in the absence of noise. Using the hybrid linear model, Zhang, *et al.* [82] assume data are samples from a collection of "flats" (affine subspaces) and choose an optimal neighborhood size from which to recover each flat by studying the least squares approximation error in the form of Jones' β-number (see [44] and also [29] in which this idea is used for curve denoising). An analysis of noise and curvature for normal estimation of smooth curves and surfaces in $\mathbb{R}^2$ and $\mathbb{R}^3$ is presented by Mitra, *et al.* [61] with application to computer graphics. We also note the work of Maggioni and coauthors [17], in which multiscale PCA is used to discover the intrinsic dimensionality of a data set.

The chapter is organized as follows. The remainder of this section provides the intuition and assumptions of our approach and introduces the geometric model that is used throughout this work. We frame the problem as one of subspace perturbation in Section 3.2 and study the size of the perturbation as a function of scale in Section 3.3. The selection of the optimal scale is our main result and is presented in Section 3.4, along with the necessary geometric conditions for tangent plane recovery. Numerical results are given in Section 3.5. We conclude with algorithmic considerations and a discussion of future directions in Section 3.6.

3.1.2 Problem Setup

Our goal is to recover the best approximation to a local tangent space of a nonlinear d-dimensional Riemannian manifold $\mathcal{M}$ from noisy samples presented in dimension $D > d$. Working about a reference point x_0, an approximation to the linear tangent space of $\mathcal{M}$ at x_0 is given by the

span of the top d singular vectors of the centered data matrix (where "top" refers to the d singular vectors associated with the d largest singular values). The question becomes: how many neighbors of x_0 should be used (or in how large of a radius about x_0 should we work) to recover the best approximation?

To answer this question, we examine the noise-curvature trade-off. Given noisy samples of a linear subspace, the quality of PCA approximation improves as more points are included. However, the curvature of $\mathcal{M}$ prevents the inclusion of a large number of points. Similarly, there exists a local scale about x_0 such that the effects of curvature are small, as $\mathcal{M}$ locally resembles Euclidean space. This suggests allowing only a very small radius about x_0, yet at small scales, the sample points are indistinguishable from noise. We therefore seek a balance and assume there exists a scale large enough to be above the noise level, but still small enough to avoid curvature. This scale reveals a linear structure that is sufficiently decoupled from both the noise and the curvature to be well approximated by a tangent plane. We note that the concept of noise-curvature trade-off has been a subject of interest for decades in dynamical systems theory [31].

3.1.3 Geometric Data Model

A d-dimensional manifold of codimension 1 may be described locally by the surface $y = f(\ell_1, \ldots, \ell_d)$, where ℓ_i is a coordinate in the tangent plane. After translating the origin, a rotation of the coordinate system can align the coordinate axes with the principal directions associated with the principal curvatures at the given reference point x_0. Aligning the coordinate axes with the plane tangent to $\mathcal{M}$ at x_0 gives a local quadratic approximation to the manifold. Using this choice of coordinates, the manifold may be described locally [34] by the Taylor series of f at the origin x_0:

$$y = f(\ell_1, \ldots, \ell_d) = \frac{1}{2}(\kappa_1 \ell_1^2 + \cdots + \kappa_d \ell_d^2) + o\left(\ell_1^2 + \cdots + \ell_d^2\right), \tag{3.1.1}$$

where $\kappa_1, \ldots, \kappa_d$ are the principal curvatures of $\mathcal{M}$ at x_0. In this coordinate system, x_0 has the form

$$x_0 = [\ell_1 \ \ell_2 \ \cdots \ \ell_d \ f(\ell_1, \ldots, \ell_d)]^T$$

and points in a local neighborhood of x_0 have similar coordinates. Generalizing to a d-dimensional manifold of arbitrary codimension in $\mathbb{R}^D$, there exist $(D-d)$ functions

$$f_i(\ell) = \frac{1}{2}(\kappa_1^{(i)}\ell_1^2 + \cdots + \kappa_d^{(i)}\ell_d^2) + o\left(\ell_1^2 + \cdots + \ell_d^2\right)$$

for $i = (d+1), \ldots, D$, with $\kappa_1^{(i)}, \ldots, \kappa_d^{(i)}$ representing the principal curvatures in codimension i at x_0. Then, given the coordinate system aligned with the principal directions, a point in a neighborhood of x_0 has coordinates $[\ell_1, \ldots, \ell_d, f_{d+1}, \ldots, f_D]$. We truncate this Taylor expansion and use the quadratic approximation

$$f_i(\ell) = \frac{1}{2}(\kappa_1^{(i)}\ell_1^2 + \cdots + \kappa_d^{(i)}\ell_d^2), \tag{3.1.2}$$

$i = (d+1), \ldots, D$, as the local model for our analysis.

Consider now discrete samples from $\mathcal{M}$ that are contaminated with an additive Gaussian noise vector e drawn from the $\mathcal{N}\left(0, \sigma^2 I_D\right)$ distribution. Each sample x is a D-dimensional vector and N such samples may be stored as columns of a matrix $X \in \mathbb{R}^{D\times N}$. The coordinate system above allows the decomposition of x into its linear (tangent plane) component ℓ, its quadratic (curvature) component c, and noise e, three D-dimensional vectors

$$\ell = [\ell_1\ \ell_2\ \cdots\ \ell_d\ \ 0\ \ \cdots\ \ 0]^T \tag{3.1.3}$$

$$c = [0\ \ \cdots\ \ 0\ c_{d+1}\ \cdots\ c_D]^T \tag{3.1.4}$$

$$e = [e_1\ e_2\ \ \cdots\ \ e_D]^T \tag{3.1.5}$$

such that the last $(D-d)$ entries of c are of the form

$$c_i = \frac{1}{2}(\kappa_1^{(i)}\ell_1^2 + \ \ldots\ \ + \kappa_d^{(i)}\ell_d^2). \tag{3.1.6}$$

We may store the N samples of ℓ, c, and e as columns of matrices L, C, E, respectively, such that our data matrix is decomposed as

$$X = L + C + E. \tag{3.1.7}$$

Remark. Of course it is unrealistic for the data to be observed in the described coordinate system. As noted, we may use a rotation to align the coordinate axes with the principal directions

associated with the principal curvatures. Doing so allows us to write (3.1.2) as well as (3.1.7). Because we will ultimately quantify the norm of each matrix using the unitarily-invariant Frobenius norm, this rotation will not affect our analysis. We therefore proceed by assuming that the coordinate axes align with the principal directions.

The true tangent plane we wish to recover is given by the PCA of L. Because we do not have direct access to L, we work with X as a proxy, and instead recover a subspace spanned by the corresponding eigenvectors of XX^T. We will study how close this recovered invariant subspace of XX^T is to the corresponding invariant subspace of LL^T as a function of scale. Throughout this work, scale refers to the number of points N in the local neighborhood within which we perform PCA. Given a fixed density of points, scale may be equivalently quantified as the radius r about the reference point x_0 defining the local neighborhood.

3.2 Perturbation of Invariant Subspaces

Given the decomposition of the data (3.1.7), we have

$$XX^T = LL^T + CC^T + EE^T + LC^T + CL^T + LE^T + EL^T + CE^T + EC^T. \tag{3.2.1}$$

To account for the centering required by PCA, define the sample mean of N realizations of random variable Y as

$$\widehat{\mathbb{E}}[Y] = \frac{1}{N}\sum_{i=1}^{N} Y^{(i)}, \tag{3.2.2}$$

where $Y^{(i)}$ denotes the ith realization. Let the mean of a matrix M be the matrix $\widehat{\mathbb{E}}[M]$ such that each entry of row i is the sample mean of the ith row of M. Let $\widetilde{M}$ denote the centered version of M:

$$\widetilde{M} = M - \widehat{\mathbb{E}}[M]. \tag{3.2.3}$$

Thus we have

$$\widetilde{X}\widetilde{X}^T = \widetilde{L}\widetilde{L}^T + \widetilde{C}\widetilde{C}^T + \widetilde{E}\widetilde{E}^T + \widetilde{L}\widetilde{C}^T + \widetilde{C}\widetilde{L}^T + \widetilde{L}\widetilde{E}^T + \widetilde{E}\widetilde{L}^T + \widetilde{C}\widetilde{E}^T + \widetilde{E}\widetilde{C}^T. \tag{3.2.4}$$

The problem may be posed as a perturbation analysis of invariant subspaces. Rewrite (3.2.1) as

$$\frac{1}{N}\widetilde{X}\widetilde{X}^T = \frac{1}{N}\widetilde{L}\widetilde{L}^T + \Delta, \tag{3.2.5}$$

where

$$\Delta = \frac{1}{N}(\widetilde{C}\widetilde{C}^T + \widetilde{E}\widetilde{E}^T + \widetilde{L}\widetilde{C}^T + \widetilde{C}\widetilde{L}^T + \widetilde{L}\widetilde{E}^T + \widetilde{E}\widetilde{L}^T + \widetilde{C}\widetilde{E}^T + \widetilde{E}\widetilde{C}^T) \tag{3.2.6}$$

is the perturbation that prevents us from working directly with $\widetilde{L}\widetilde{L}^T$. The dominant eigenspace of $\widetilde{X}\widetilde{X}^T$ is therefore a perturbed version of the dominant eigenspace of $\widetilde{L}\widetilde{L}^T$. Seeking to minimize the effect of this perturbation, we look for the scale N^* at which the dominant eigenspace of $\widetilde{X}\widetilde{X}^T$ is closest to that of $\widetilde{L}\widetilde{L}^T$. Before proceeding, we review material on the perturbation of eigenspaces relevant to our analysis. The reader familiar with this topic is invited to skip directly to Theorem 12.

The distance between two subspaces of $\mathbb{R}^D$ can be defined as the spectral norm of the difference between their respective orthogonal projectors [35]. As we will always be considering two equidimensional subspaces, this distance is equal to the sine of the largest principal angle between the subspaces. We state our results in terms of the Frobenius norm as it will provide a simplification of Theorem 12. Then, by the equivalence of norms, we may define the optimal scale N^* as

$$N^* = \arg\min_N \|P - \widehat{P}\|_F, \tag{3.2.7}$$

where P and $\widehat{P}$ are the orthogonal projectors onto the subspaces computed from L and X, respectively. The solution to (3.2.7) is the main goal of this work.

The distance $\|P - \widehat{P}\|_F$ may be bounded by the classic $\sin\Theta$ theorem of Davis and Kahan [20]. We will use a version of this theorem presented by Stewart (Theorem V.2.7 of [74]), modified for our specific purpose. First, we establish some notation, following closely that found in [74]. Consider the eigendecompositions

$$\frac{1}{N}\widetilde{L}\widetilde{L}^T = U\Lambda U^T = [U_1\ U_2]\ \Lambda\ [U_1\ U_2]^T, \tag{3.2.8}$$

$$\frac{1}{N}\widetilde{X}\widetilde{X}^T = \widehat{U}\widehat{\Lambda}\widehat{U}^T = [\widehat{U}_1\ \widehat{U}_2]\ \widehat{\Lambda}\ [\widehat{U}_1\ \widehat{U}_2]^T, \tag{3.2.9}$$

such that the columns of U are the eigenvectors of $\frac{1}{N}\widetilde{L}\widetilde{L}^T$ and the columns of $\widehat{U}$ are the eigenvectors of $\frac{1}{N}\widetilde{X}\widetilde{X}^T$. The columns of U_1 are those eigenvectors associated with the d largest eigenvalues in Λ arranged in descending order. The columns of U_2 are then those eigenvectors associated with the smallest $(D-d)$ eigenvalues, and $\widehat{U}$ is similarly partitioned. The subspace we recover is spanned by the columns of $\widehat{U}_1$ and we wish to have this subspace as close as possible to the tangent space spanned by the columns of U_1. The orthogonal projectors onto the tangent and computed subspaces, P and $\widehat{P}$ respectively, are given by

$$P = U_1 U_1^T \quad \text{and} \quad \widehat{P} = \widehat{U}_1 \widehat{U}_1^T.$$

Define λ_d to be the dth largest eigenvalue of $\frac{1}{N}\widetilde{L}\widetilde{L}^T$, or the last entry on the diagonal of Λ_1. Note that λ_d corresponds to variance in a tangent plane direction.

We are now in position to state the theorem. Note that we have made use of the fact that the columns of U are the eigenvectors of $\widetilde{L}\widetilde{L}^T$, that Λ_1, Λ_2 are Hermitian (diagonal) matrices, and that the Frobenius norm is used to measure distances. The reader is referred to [74] for the theorem in its original form.

Theorem 12. ***(Davis & Kahan [20], Stewart [74])***

Let $\delta = \lambda_d - \left\|U_1^T \Delta U_1\right\|_F - \left\|U_2^T \Delta U_2\right\|_F$ *and consider*

- *(Condition 1)* $\quad \delta > 0$
- *(Condition 2)* $\quad \left\|U_1^T \Delta U_2\right\|_F \left\|U_2^T \Delta U_1\right\|_F < \frac{1}{4}\delta^2.$

Then, provided that conditions 1 and 2 hold,

$$\left\|P - \widehat{P}\right\|_F \leq 2\sqrt{2}\, \frac{\left\|U_2^T \Delta U_1\right\|_F}{\delta}. \tag{3.2.10}$$

The two conditions of the theorem have important geometric interpretations. Informally, condition 1 requires that the linear structure we seek to recover be sufficiently decoupled from both

the noise and curvature (this is consistent with our assumption of the existence of a scale yielding sufficient decoupling). We may consider δ^{-1} to be the condition number for subspace recovery. When δ approaches zero, the condition number becomes large, and bound (3.2.10) loses meaning as we cannot recover an approximating subspace. In Section 3.4 we will see that condition 1 naturally gives rise to an uncertainty principle that quantifies the limits of noise-curvature perturbation for tangent plane recovery. We will also see that the second condition naturally implies that the manifold be sufficiently sampled.

The solution to (3.2.7) is impractical to compute. However, (3.2.10) is a tight bound, as will be demonstrated by the experiments (Section 3.5). Thus, a solution may be approximated by minimizing the right-hand side of (3.2.10). To do so, and to give each quantity in the theorem a geometric interpretation, we must first understand the behavior of the perturbation Δ as a function of the scale parameter N.

3.3 Bounding the Effects of Noise and Curvature

In this section we study the behavior of each term in (3.2.6) as a function of the scale parameter N. First, we provide insight as to their leading order behavior. As explained by Fukunaga [32], estimator bias and estimator variance depend on the Hessian and gradient, respectively, of the function being estimated. Consider the local manifold model (3.1.2). This second order approximation is presented in a coordinate system such that its gradient is zero and its Hessian is a diagonal matrix with the principal curvatures as its entries. We therefore expect perturbation terms associated with variance to tend to zero as the scale parameter N increases. Likewise, we expect pure curvature terms to grow with N. Formal calculations will show that $\frac{1}{N}CC^T$, the term associated purely with curvature, has nonzero expectation that increases with N. Note that while the diagonal entries of $\frac{1}{N}EE^T$ also have nonzero expectation, these terms do not grow with N and are therefore associated with a noise-floor rather than with estimator bias. All other terms in (3.2.6) are zero in expectation, and thus only carry variance. Accordingly, these terms decay as $1/\sqrt{N}$.

3.3.1 Preliminaries

3.3.1.1 Sampling a Linear Subspace

Consider sampling a linear subspace by uniformly sampling points inside $B^d_{x_0}(r)$, the d-dimensional ball of radius r centered at x_0. We drop the dependence on x_0 from our notation for the remainder of this analysis. Because we are sampling from a noise-free linear subspace, the number of points N captured inside $B^d(r)$ is a function of the sampling density ρ:

$$N = \rho v_d r^d, \tag{3.3.1}$$

where v_d is the volume of the d-dimensional unit ball. As we wish to maintain a local analysis, we must enforce that r be small. To make this explicit, denote by r_{max} the largest radius within which the local model (3.1.2) holds and compute the number of points captured in $B^d(r_{max})$:

$$N_{max} = \rho v_d r^d_{max}. \tag{3.3.2}$$

Then rescale (3.3.1) by dividing by (3.3.2) and solve for r:

$$r = r_{max} \left(\frac{N}{N_{max}} \right)^{\frac{1}{d}}. \tag{3.3.3}$$

Remark. Equation (3.3.3) suppresses the dependence on sampling that is captured by the ρv_d term. Note that because r is small, the sampling density ρ may have to be large to allow for large N, as is explicitly seen in equation (3.3.1). The volume of the unit ball, v_d, is very small for even reasonable values of d, further necessitating a large sampling density. The analysis in this section may be performed entirely in the context of equation (3.3.1) provided that r is taken to be small. In doing so, the dependence on sampling density ρ is clear in all steps. We prefer to instead perform the analysis in the context of the rescaled equation (3.3.3), explicitly forcing the analysis to the local scale by considering r to be a fraction of the largest radius allowed by the local model.

3.3.1.2 Notation

In this section and throughout the remainder of this work, we will make use of the following definitions involving the principal curvatures:

$$K_i = \sum_{n=1}^{d} \kappa_n^{(i)}, \tag{3.3.4}$$

$$K = \left(\sum_{i=d+1}^{D} K_i^2 \right)^{\frac{1}{2}}, \tag{3.3.5}$$

$$K_{nn}^{ij} = \sum_{n=1}^{d} \kappa_n^{(i)} \kappa_n^{(j)}, \qquad K_{mn}^{ij} = \sum_{\substack{m,n=1 \\ m \neq n}}^{d} \kappa_m^{(i)} \kappa_n^{(j)}. \tag{3.3.6}$$

The constant K_i quantifies the curvature in codimension i, for $i = (d\ +\ 1), \ldots, D$. Note that given our choice of coordinate system in Section 3.1.2, K_i is the trace of the Hessian in the ith codimension. The overall curvature of our local model is quantified by K and is a natural result of our use of the Frobenius norm. We note that $K_i K_j = K_{nn}^{ij} + K_{mn}^{ij}$.

By the choice of coordinate system, U_2 is the $D \times (D-d)$ matrix whose columns are the last $(D-d)$ columns of I_D, the identity matrix of order D. Due to the specific form of each matrix, we have $U_1^T C$, $C^T U_1$, $U_2^T L$, and $L^T U_2$ are all zero matrices of the appropriate size. Because Δ is a symmetric matrix, we have that $\left\| U_2^T \Delta U_1 \right\|_F = \left\| U_1^T \Delta U_2 \right\|_F$.

Finally, we will work with projections of vector a onto U_1 and U_2, where a takes the form of ℓ, c, or e (equations (3.1.3)–(3.1.5)), and denote such projections by

$$U_p^T a = a_{u_p}, \qquad \text{for } p = \{1, 2\}. \tag{3.3.7}$$

3.3.2 Analysis of Perturbation Terms

We begin by presenting our general strategy for bounding terms of the form $\|U_p^T \frac{1}{N} \widetilde{A} \widetilde{B}^T U_q\|_F$ for $p, q = \{1, 2\}$ where A and B are general matrices of size $D \times N$. The key observation is that

$\frac{1}{N}\widetilde{A}\widetilde{B}^T$ is a sample mean of N outer products of vectors a and b, each sampled from a given distribution:

$$\frac{1}{N}\widetilde{A}\widetilde{B}^T = \widehat{\mathbb{E}}[(a - \widehat{\mathbb{E}}[a])(b - \widehat{\mathbb{E}}[b])^T],$$

where $\widehat{\mathbb{E}}[Y]$ is the sample mean defined in (3.2.2). We therefore expect that $\frac{1}{N}\widetilde{A}\widetilde{B}^T$ will converge toward the centered outer product of a and b.

We will use the following result of Shawe-Taylor and Cristianini [70] to bound, with high probability, the norm of the difference between this sample mean and its expectation,

$$\left\|\mathbb{E}[U_p^T(a - \mathbb{E}[a])(b - \mathbb{E}[b])^T U_q] \;-\; \widehat{\mathbb{E}}[U_p^T(a - \widehat{\mathbb{E}}[a])(b - \widehat{\mathbb{E}}[b])^T U_q]\right\|_F \tag{3.3.8}$$

where $\mathbb{E}[Y]$ is the expectation of the random variable $Y \in \mathcal{Y}$.

Theorem 13. ***(Shawe-Taylor & Cristianini, [70]).*** *Given N samples of a random variable Y generated independently at random from $\mathcal{Y}$ according to the distribution P_Y, with probability at least $1 - e^{-\eta^2}$ over the choice of the samples, we have*

$$\left\|\mathbb{E}[Y] - \widehat{\mathbb{E}}[Y]\right\|_F \leq \frac{R}{\sqrt{N}}\left(2 + \eta\sqrt{2}\right) \tag{3.3.9}$$

where $R = \sup_{supp(P_Y)} \|Y\|_F$ and $supp(P_Y)$ is the support of distribution P_Y.

Remark. With a slight abuse of notation, we note that the "Frobenius norm of a vector" is equivalent to the vector's Euclidean norm, and thus we use $\|\cdot\|_F$ for both matrices and vectors.

Remark. The choice of R in (3.3.9) need not be unique. Our analysis will proceed by using upper bounds for $\|Y\|_F$ which may not be suprema.

Continuing from (3.3.8),

$$\begin{aligned}
&\left\|\mathbb{E}[U_p^T(a - \mathbb{E}[a])(b - \mathbb{E}[b])^T U_q] \;-\; \widehat{\mathbb{E}}[U_p^T(a - \widehat{\mathbb{E}}[a])(b - \widehat{\mathbb{E}}[b])^T U_q]\right\|_F \\
&= \left\|\mathbb{E}[a_{u_p} b_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p} b_{u_q}^T] \;+\; \widehat{\mathbb{E}}[a_{u_p}]\widehat{\mathbb{E}}[b_{u_q}^T] - \mathbb{E}[a_{u_p}]\,\mathbb{E}[b_{u_q}^T]\right\|_F \\
&\leq \left\|\mathbb{E}[a_{u_p} b_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p} b_{u_q}^T]\right\|_F \;+\; \left\|\mathbb{E}[a_{u_p}]\,\mathbb{E}[b_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p}]\widehat{\mathbb{E}}[b_{u_q}^T]\right\|_F .
\end{aligned} \tag{3.3.10}$$

Because $\mathbb{E}[\ell] = 0$ and $\mathbb{E}[e] = 0$, $\mathbb{E}[a_{u_p}]\,\mathbb{E}[b_{u_q}^T]$ is nonzero only for the case $(a = b = c, p = q = 2)$. In this case, $\widehat{\mathbb{E}}[U_p^T(a - \widehat{\mathbb{E}}[a])(b - \widehat{\mathbb{E}}[b])^T U_q] = \widehat{\mathbb{E}}[U_2^T(c - \widehat{\mathbb{E}}[c])(c - \widehat{\mathbb{E}}[c])^T U_2]$ is an empirical covariance matrix. As shown in [70], such a matrix is unchanged when the origin is shifted by a fixed translation. Therefore we may assume that the origin has been shifted to the center of mass of the distribution and we may take $\mathbb{E}[c_{u_2}]$ and $\mathbb{E}[c_{u_2}^T]$ to be zero. Note that we may only do so in the context of this calculation, and in general $\mathbb{E}[c_{u_2}]$ and $\mathbb{E}[c_{u_2}^T]$ are nonzero. Then for all choices of (a, b, p, q), we have $\mathbb{E}[a_{u_p}]\,\mathbb{E}[b_{u_q}^T] = 0$ and the right-hand side of (3.3.10) becomes

$$\begin{aligned} &\left\|\mathbb{E}[a_{u_p} b_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p} b_{u_q}^T]\right\|_F \;+\; \left\|\widehat{\mathbb{E}}[a_{u_p}]\widehat{\mathbb{E}}[b_{u_q}^T]\right\|_F \\ \leq\; &\left\|\mathbb{E}[a_{u_p} b_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p} b_{u_q}^T]\right\|_F \;+\; \left\|\,\mathbb{E}[a_{u_p}] - \widehat{\mathbb{E}}[a_{u_p}]\right\|_F \left\|\mathbb{E}[b_{u_q}^T] - \widehat{\mathbb{E}}[b_{u_q}^T]\right\|_F . \end{aligned} \tag{3.3.11}$$

We now use Theorem 13 to bound each of the three terms in (3.3.11). For this analysis, the random variable Y in Theorem 13 takes one of the following two forms:

$$Y = a_{u_p} b_{u_q}^T \quad \text{or} \quad Y = a_{u_p}$$

for $p, q = \{1, 2\}$. Thus there are two corresponding definitions for R:

$$R_{ab}^{pq} = \sup_{\substack{\text{supp}(P_a) \\ \text{supp}(P_b)}} \|a_{u_p} b_{u_q}^T\|_F \tag{3.3.12}$$

$$R_a^p = \sup_{\text{supp}(P_a)} \|a_{u_p}\|_F \tag{3.3.13}$$

where a and b are sampled according to distributions P_a and P_b, respectively. Directly applying Theorem 13 to each of the three terms in (3.3.11) and using a standard union bound argument yields

$$\begin{aligned} &\left\|\mathbb{E}[a_{u_p} b_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p} b_{u_q}^T]\right\|_F \;+\; \left\|\,\mathbb{E}[a_{u_p}] - \widehat{\mathbb{E}}[a_{u_p}]\right\|_F \left\|\mathbb{E}[b_{u_q}^T] - \widehat{\mathbb{E}}[b_{u_q}^T]\right\|_F \\ \leq\; &\frac{R_{ab}^{pq}}{\sqrt{N}}\left(2 + \eta_{ab}\sqrt{2}\right) \;+\; \frac{R_a^p R_b^q}{N}\left(2 + \eta_a\sqrt{2}\right)\left(2 + \eta_b\sqrt{2}\right) \end{aligned} \tag{3.3.14}$$

with probability greater than

$$1 - e^{-\eta_{ab}^2} - e^{-\eta_a^2} - e^{-\eta_b^2} \tag{3.3.15}$$

over the random sampling of a and b. For the case that $a = b$ we instead simply have the result holding with probability greater than

$$1 - e^{-\eta_{aa}^2} - e^{-\eta_a^2} \tag{3.3.16}$$

over the random sampling of a. The probability constants may be chosen to ensure such an event holds with high probability. For example, in (3.3.15), letting $\eta_{ab} = \eta_a = \eta_b = \eta$, we have probability greater than 0.9451 for $\eta = 2$ and greater than 0.9996 for $\eta = 3$.

Putting it all together, we have that

$$\begin{aligned} &\left| \| \mathbb{E}[U_p^T (a - \mathbb{E}[a])(b - \mathbb{E}[b])^T U_q] \|_F \;-\; \| \widehat{\mathbb{E}}[U_p^T (a - \widehat{\mathbb{E}}[a])(b - \widehat{\mathbb{E}}[b])^T U_q] \|_F \right| \\ &\leq \frac{R_{ab}^{pq}}{\sqrt{N}} \left(2 + \eta_{ab}\sqrt{2}\right) \;+\; \frac{R_a^p R_b^q}{N} \left(2 + \eta_a \sqrt{2}\right) \left(2 + \eta_b \sqrt{2}\right) \end{aligned} \tag{3.3.17}$$

and we may conclude that

$$\left\| U_p^T \left(\frac{1}{N} \widetilde{A} \widetilde{B}^T \right) U_q \right\|_F \;\in\; \Big[\mu - \Gamma,\; \mu + \Gamma \Big], \tag{3.3.18}$$

$$\begin{aligned} \text{where} \quad \mu &= \left\| \mathbb{E}[a_{u_p} b_{u_q}^T] - \mathbb{E}[a_{u_p}] \, \mathbb{E}[b_{u_q}^T] \right\|_F \\ \text{and} \quad \Gamma &= \frac{R_{ab}^{pq}}{\sqrt{N}} \left(2 + \eta_{ab}\sqrt{2}\right) \;+\; \frac{R_a^p R_b^q}{N} \left(2 + \eta_a \sqrt{2}\right) \left(2 + \eta_b \sqrt{2}\right), \end{aligned}$$

with probability greater than

$$\begin{cases} 1 - e^{-\eta_{ab}^2} - e^{-\eta_a^2} - e^{-\eta_b^2} & \text{for } a \neq b \\ 1 - e^{-\eta_{ab}^2} - e^{-\eta_a^2} & \text{for } a = b \end{cases}$$

over the random sampling of a and b.

Before computing the constants R_{ab}^{pq} and R_a^p, we must ensure that either the suprema (3.3.12) and (3.3.13) exist or that finite bounds may be given in place of suprema. Noting that

$$\| a_{u_p} b_{u_q}^T \| \leq \| a_{u_p} \| \| b_{u_q} \|, \tag{3.3.19}$$

it suffices to show that the projections ℓ_{u_p}, c_{u_p}, and e_{u_p} are bounded. The vectors ℓ and c are functions of the coordinates of points drawn uniformly from $B^d(r)$. Therefore their entries are bounded, as are the norms of their projections. The entries of e, while unbounded in general, are Gaussian random variables and are therefore bounded over a set with large measure. Let Ω_e be the set in $\mathbb{R}^D$ for which both

$$\|e_{u_1}\| \leq \sigma\left(\sqrt{d} + \xi_e\sqrt{2}\right) \tag{3.3.20}$$

$$\|e_{u_2}\| \leq \sigma\left(\sqrt{D-d} + \xi_e\sqrt{2}\right) \tag{3.3.21}$$

hold. A formal construction of Ω_e is given in Appendix A.1 and the measure $\gamma_D(\Omega_e)$ of this set is shown to be large,

$$\gamma_D(\Omega_e) > 1 - 2e^{-\xi_e^2}.$$

Therefore, on Ω_e, we may state bounds for R_e^p $(p = 1, 2)$. All results involving projections of the vector e will be given over this set. We apply Theorem 13 by conditioning on $e \in \Omega_e$. If we consider the joint probability of the independent random variables a and e, then we can estimate the probability that the deviation of the outer-product of a_{u_p} and e_{u_q} from its expectation is large. This probability is given by:

$$\begin{aligned}
&\text{Prob}\left[\|\,\mathbb{E}[a_{u_p}e_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p}e_{u_q}^T]\|_F > \frac{R_{ae}^{pq}}{\sqrt{N}}(2+\eta_{ae}\sqrt{2})\right] \\
&= \text{Prob}\left[\|\,\mathbb{E}[a_{u_p}e_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p}e_{u_q}^T]\|_F > \frac{R_{ae}^{pq}}{\sqrt{N}}(2+\eta_{ae}\sqrt{2}) \mid e \in \Omega_e\right]\text{Prob}\left[e \in \Omega_e\right] \\
&+ \text{Prob}\left[\|\,\mathbb{E}[a_{u_p}e_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p}e_{u_q}^T]\|_F > \frac{R_{ae}^{pq}}{\sqrt{N}}(2+\eta_{ae}\sqrt{2}) \mid e \in \overline{\Omega}_e\right]\text{Prob}\left[e \in \overline{\Omega}_e\right] \\
&\leq\ (e^{-\eta_{ae}^2})(1-2e^{-\xi_e^2})\ +\ (1)(2e^{-\xi_e^2}) \\
&=\ e^{-\eta_{ae}^2} + 2e^{-\xi_e^2} - 2e^{-\eta_{ae}^2}e^{-\xi_e^2}
\end{aligned}$$

Thus we have that

$$\left\|\mathbb{E}[a_{u_p}e_{u_q}^T] - \widehat{\mathbb{E}}[a_{u_p}e_{u_q}^T]\right\|_F\ \leq\ \frac{R_{ae}^{pq}}{\sqrt{N}}\left(2+\eta_{ae}\sqrt{2}\right) \tag{3.3.22}$$

with probability greater than $1 - e^{-\eta_{ae}^2} - 2e^{-\xi_e^2} + 2e^{-\eta_{ae}^2}e^{-\xi_e^2}$ over the random sampling of a and random realization of e, and an identical calculation holds when applying the theorem to

$\left\|\mathbb{E}[e_{u_p}] - \widehat{\mathbb{E}}[e_{u_p}]\right\|_F$. The probability given in (3.3.15) can be bounded by

$$\begin{aligned} &1 - e^{-\eta_{ae}^2} - e^{-\eta_a^2} - e^{-\eta_e^2} - 2e^{-\xi_e^2} + 2e^{-\xi_e^2}(e^{-\eta_{ae}^2} + e^{-\eta_e^2}) \\ &> 1 - e^{-\eta_{ae}^2} - e^{-\eta_a^2} - e^{-\eta_e^2} - 2e^{-\xi_e^2} \end{aligned} \tag{3.3.23}$$

and when $(a = b = e)$ the probability given in (3.3.16) can be bounded by

$$\begin{aligned} &1 - e^{-\eta_{ee}^2} - e^{-\eta_e^2} - 2e^{-\xi_e^2} + 2e^{-\xi_e^2}(e^{-\eta_{ee}^2} + e^{-\eta_e^2}) \\ &> 1 - e^{-\eta_{ee}^2} - e^{-\eta_e^2} - 2e^{-\xi_e^2} \end{aligned} \tag{3.3.24}$$

where we have neglected the positive contribution of the product of exponentially small terms.

3.3.2.1 Suprema R_{ab}^{pq} and R_a^p

We now compute bounds for the R_a^p terms. Simple norm calculations give

$$\|\ell_{u_1}\|_F^2 = \sum_{i=1}^{d} \ell_i^2 \leq r_{max}^2 \left(\frac{N}{N_{max}}\right)^{\frac{2}{d}},$$

$$\begin{aligned} \|c_{u_2}\|_F^2 = \sum_{i=d+1}^{D} c_i^2 = \frac{1}{4} \sum_{i=d+1}^{D} \left(\kappa_1^{(i)} \ell_1^2 + \ \ldots \ + \kappa_d^{(i)} \ell_d^2\right)^2 \leq \\ \frac{r_{max}^4}{4} \left(\frac{N}{N_{max}}\right)^{\frac{4}{d}} \sum_{i=d+1}^{D} \left(\sum_{n=1}^{d} \kappa_n^{(i)}\right)^2 = \frac{K^2 r_{max}^4}{4} \left(\frac{N}{N_{max}}\right)^{\frac{4}{d}}, \end{aligned} \tag{3.3.25}$$

where we have assumed all principal curvatures have the same sign. Combining with (3.3.20) and (3.3.21) yields bounds for R_a^p terms, and (3.3.19) may be used to bound the R_{ab}^{pq} terms. The results are listed in Appendix A.3.

Remark. The calculation (3.3.25) for $\|c_{u_2}\|_F$ requires that all principal curvatures have the same sign for the inequality to hold. This requirement will carry through as an assumption in the statement of Theorem 14 (Main Result 1). To avoid this requirement we must work with moments of ℓ_i rather than with norms, specifically when computing CL^T and LC^T terms. While not possible here, it is possible to do so as demonstrated in Main Result 2 (see Section 3.4.3.2). Despite this assumption, it will be seen in Section 3.5 that this current analysis does in fact provide meaningful

results for principal curvatures of mixed signs (except for when $K_i = 0$), indicating that tighter R_{ab}^{pq} bounds are possible. We note that Main Result 2 will require no such assumption and will hold for any value of K_i.

3.3.2.2 Expectations

We are almost in position to define confidence intervals of the form (3.3.18), where $\widetilde{A}$ and $\widetilde{B}$ may be the centered matrices $\widetilde{L}$, $\widetilde{C}$, and $\widetilde{E}$. All that remains is to compute the true expectation term of equation (3.3.17):

$$\left\| \mathbb{E}[U_p^T(a - \mathbb{E}[a])(b - \mathbb{E}[b])^T U_q] \right\|_F = \left\| \mathbb{E}[a_{u_p} b_{u_q}^T] - \mathbb{E}[a_{u_p}]\,\mathbb{E}[b_{u_q}^T] \right\|_F. \tag{3.3.26}$$

As the coordinates of ℓ and c are functions of points sampled uniformly from $B^d(r)$ and $e \sim \mathcal{N}(0, \sigma^2 I_D)$, the expectation terms are zero for $c\ell^T$, ℓe^T, and ce^T. Only the pure curvature (cc^T) and pure noise (ee^T) terms may have nonzero expectations and their calculations are given in Appendices A.2, and A.3. We list here only the results.

Pure Curvature Term:

$$\left\| \mathbb{E}[c_{u_2} c_{u_2}^T] - \mathbb{E}[c_{u_2}]\,\mathbb{E}[c_{u_2}^T] \right\|_F = \tag{3.3.27}$$
$$\frac{r_{max}^4}{2(d+2)^2(d+4)} \left(\frac{N}{N_{max}} \right)^{\frac{4}{d}} \left[\sum_{i=d+1}^{D} \sum_{j=d+1}^{D} \left[(d+1) K_{nn}^{ij} - K_{mn}^{ij} \right]^2 \right]^{\frac{1}{2}}.$$

We note that later, for the purpose of interpretation, we will replace this exact expectation with an upper bound. Using that $\mathbb{E}[\ell_m^2 \ell_n^2] < \mathbb{E}[\ell_n^4]$ and $K_{nn}^{ij} + K_{mn}^{ij} = K_i K_j$, we bound

$$\left\| \mathbb{E}[c_{u_2} c_{u_2}^T] - \mathbb{E}[c_{u_2}]\,\mathbb{E}[c_{u_2}^T] \right\|_F < K^2 \frac{(d+1)}{2(d+2)^2(d+4)} r_{max}^4 \left(\frac{N}{N_{max}} \right)^{\frac{4}{d}}. \tag{3.3.28}$$

The nonzero expectation (3.3.27) and its bound (3.3.28) grow with N and in this way may be thought of as the source of the estimator bias (see the beginning discussion of Section 3.3).

Pure Noise Terms:

$$\left\| \mathbb{E}[e_{u_p} e_{u_q}^T] - \mathbb{E}[e_{u_p}]\,\mathbb{E}[e_{u_q}^T] \right\|_F = \begin{cases} \sigma^2\sqrt{d} & \text{if } (p,q) = (1,1), \\ \sigma^2\sqrt{D-d} & \text{if } (p,q) = (2,2), \\ 0 & \text{if } (p,q) = (2,1). \end{cases}$$

Note that unlike the nonzero expectation of the pure curvature term, the nonzero expectations of the pure noise terms ($p = q$) are constant and do not grow with N. Thus they represent a noise floor rather than a source of estimator bias.

3.3.2.3 Norm Bounds

We may now use the right-hand side of the confidence interval (3.3.18) to bound the size of the perturbation norms. When considering noise terms, recall that we must condition on $e \in \Omega_e$. To aid in interpretation, we recall the rescaled notation $r = r_{max}(N/N_{max})^{1/d}$. As each curvature term c has norm roughly of size Kr^2, we expect $\|\frac{1}{N}CC^T\|_F$ to grow as K^2r^4. Concentration of Gaussian measure indicates that the norm of the noise matrix will have size that depends on the square root of the projection dimension and the variance σ^2. All other terms are zero in expectation and thus we expect $1/\sqrt{N}$ decay. The linear-curvature, linear-noise, and curvature-noise matrices should have norm $Kr^3/\sqrt{N}$, $\sigma r/\sqrt{N}$, and $K\sigma r^2/\sqrt{N}$, respectively. The leading order behavior of all norm bounds may be found in Table 3.1 and the full expressions with associated probabilities are given in Appendix A.3. Note that we may work with either the matrix in question or its transpose when computing the norm and our notation may reflect either choice.

3.4 Optimal Scale Selection and Subspace Recovery

Our main result, a bound on the angle between the recovered and true tangent planes, is formulated in this section. First we use the triangle inequality to bound the norms appearing in Theorem 12. We then inject the perturbation norms computed in Section 3.3 and Appendix A.3 to formulate the main result.

We have:

$$\begin{aligned}
\left\|U_1^T \Delta U_1\right\|_F &\leq 2\left\|U_1^T \frac{1}{N}\widetilde{L}\widetilde{E}^T U_1\right\|_F + \left\|U_1^T \frac{1}{N}\widetilde{E}\widetilde{E}^T U_1\right\|_F \\
\left\|U_2^T \Delta U_2\right\|_F &\leq \left\|U_2^T \frac{1}{N}\widetilde{C}\widetilde{C}^T U_2\right\|_F + 2\left\|U_2^T \frac{1}{N}\widetilde{C}\widetilde{E}^T U_2\right\|_F + \left\|U_2^T \frac{1}{N}\widetilde{E}\widetilde{E}^T U_2\right\|_F \\
\left\|U_2^T \Delta U_1\right\|_F &\leq \left\|U_2^T \frac{1}{N}\widetilde{C}\widetilde{L}^T U_1\right\|_F + \left\|U_2^T \frac{1}{N}\widetilde{E}\widetilde{L}^T U_1\right\|_F + \left\|U_2^T \frac{1}{N}\widetilde{C}\widetilde{E}^T U_1\right\|_F \\
&\quad + \left\|U_2^T \frac{1}{N}\widetilde{E}\widetilde{E}^T U_1\right\|_F .
\end{aligned}$$

As Δ is a symmetric matrix, we also have that $\left\|U_2^T \Delta U_1\right\|_F = \left\|U_1^T \Delta U_2\right\|_F$. Using a standard union bound argument, the bounds for each term hold simultaneously with probability greater than

$$1 - e^{-\eta_{cc}^2} - 3e^{-\eta_{ee}^2} - e^{-\eta_{\ell c}^2} - 2e^{-\eta_{\ell e}^2} - 2e^{-\eta_{ce}^2} - e^{-\eta_{\ell}^2} - e^{-\eta_c^2} - 2e^{-\eta_e^2} - 2e^{-\xi_e^2} \tag{3.4.1}$$

over the joint random selection of the sample points and random realization of the noise. We may pick a constant η and set

$$\eta_{cc} = \eta_{ee} = \eta_{\ell c} = \eta_{\ell e} = \eta_{ce} = \eta_{\ell} = \eta_c = \eta_e = \eta \tag{3.4.2}$$

such that (3.4.1) becomes

$$1 - 13e^{-\eta^2} - 2e^{-\xi_e^2}. \tag{3.4.3}$$

Finally, recall from Theorem 12 that

$$\delta \;=\; \lambda_d - \left\|U_1^T \Delta U_1\right\|_F - \left\|U_2^T \Delta U_2\right\|_F . \tag{3.4.4}$$

In order to bound the size of δ we must compute λ_d, the dth eigenvalue of $\frac{1}{N}\widetilde{L}\widetilde{L}^T$. This matrix is a centered covariance matrix and therefore its dth eigenvalue is the variance in the dth coordinate. From our moment calculations in Appendix A.4, we let

$$\lambda_d \;=\; \mathrm{Var}[\ell_d] \;=\; \frac{r_{max}^2}{d+2}\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}}. \tag{3.4.5}$$

3.4.1 Main Result: Bounding the Angle Between Subspaces

We are now in position to apply Theorem 12 and state our main result. First, define the following constants:

$$\begin{aligned}
\mathcal{K} &= \left[\sum_{i=d+1}^{D}\sum_{j=d+1}^{D}\left[(d+1)K_{nn}^{ij}-K_{mn}^{ij}\right]^2\right]^{\frac{1}{2}}, \\
p_1(\eta) &= \left(2+\eta\sqrt{2}\right)\left(1+\frac{1}{\sqrt{N}}(2+\eta\sqrt{2})\right), \\
p_2(\xi_e) &= \left(\sqrt{d}+\xi_e\sqrt{2}\right), \\
p_3(\xi_e) &= \left(\sqrt{D-d}+\xi_e\sqrt{2}\right), \\
\zeta(K,\sigma,\eta,\xi_e) &= \left[K^2 r_{max}^4\left(\frac{N}{N_{max}}\right)^{\frac{4}{d}}+\sigma p_3(\xi_e)Kr_{max}^2\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}}\right. \\
&\qquad \left.+2\sigma p_2(\xi_e)r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}}+\sigma^2\left(p_2(\xi_e)^2+p_3(\xi_e)^2\right)\right].
\end{aligned}$$

Then we have our main result:

Theorem 14. ***(Main Result 1)***. *Let the following conditions hold:*

- *(Condition 1)* $\delta = \lambda_d - \left\|U_1^T\Delta U_1\right\|_F - \left\|U_2^T\Delta U_2\right\|_F > 0,$
- *(Condition 2)* $\|U_2^T\Delta U_1\|_F < \frac{1}{2}\delta.$

For each $i=(d+1),\ldots,D$, let principal curvatures $\kappa_n^{(i)}$ have the same sign, $n=1,\ldots,d$. Then we have

$$\left\|P-\widehat{P}\right\|_F \leq \frac{2\sqrt{2}\,\frac{p_1(\eta)}{\sqrt{N}}\left[\frac{K}{2}r^3+\sigma p_2(\xi_e)\frac{K}{2}r^2+\sigma p_3(\xi_e)r+\sigma^2 p_2(\xi_e)p_3(\xi_e)\right]}{\frac{r^2}{(d+2)}-\frac{\mathcal{K}r^4}{2(d+2)^2(d+4)}-\sigma^2\left(\sqrt{d}+\sqrt{D-d}\right)-\frac{p_1(\eta)}{\sqrt{N}}\zeta(K,\sigma,\eta,\xi_e)} \tag{3.4.6}$$

with probability greater than $1-13e^{-\eta^2}-2e^{-\xi_e^2}$ over the joint random selection of the sample points and random realization of the noise, where the rescaled notation $r=r_{max}(N/N_{max})^{1/d}$ has been used.

Proof. Applying the norm bounds computed in Appendix A.3 to Theorem 12 and choosing the probability constants as given in (3.4.2) yields the result. □

The optimal scale N^* may be selected as the N for which (3.4.6) is minimized. As will be analyzed in Section 3.4.2.1, (3.4.6) is either monotonically decreasing (for the curvature-free case), monotonically increasing (for the noise-free case), or decreasing at small scales and increasing at large scales (for the general case). We therefore expect a unique minimizer in all cases. Note that the constants η and ξ_e need to be selected to ensure that this bound holds with high probability.

As discussed in Section 3.2, we may interpret δ^{-1} as the condition number for tangent plane recovery, and we may analyze it by bounding (3.4.4) using the bounds for $\|U_1^T \Delta U_1\|_F$ and $\|U_2^T \Delta U_2\|_F$. We note that the denominator in (3.4.6) is a lower bound on δ and we therefore analyze the condition number via this bound. From Main Result 1 (3.4.6), we see that when δ^{-1} is small, we may recover a tight approximation to the true tangent space. Likewise, when δ^{-1} becomes large, the angle between the computed and true subspaces becomes large. The notion of an angle loses meaning as δ^{-1} tends to infinity, and we are unable to recover an approximating subspace.

Condition 1 imposes that the spectrum corresponding to the linear subspace (λ_d) must be well separated from the spectra of the noise ($\|U_1 \Delta U_1\|_F$) and curvature ($\|U_2 \Delta U_2\|_F$) perturbations. In this way, condition 1 quantifies our requirement that there exists a scale such that the linear subspace is sufficiently decoupled from the effects of curvature and noise. When the spectra are not well separated, the angle between the subspaces becomes ill-defined. In this case, the approximating subspace contains an eigenvector corresponding to a direction orthogonal to the true tangent plane. In the language of [63], there is a crossover between the spectrum of the linear subspace and the spectrum of the perturbation, and we will observe a loss of tracking of the true tangent plane. Unlike the result of [63] where the crossover results from noise perturbation, we will demonstrate a crossover due to geometry in Section 3.5. As we will see, condition 1 indeed imposes a geometric requirement for tangent plane recovery.

With condition 1 imposing restrictions on the effects of curvature and noise, condition 2 may be interpreted as a control on sampling. This condition may be satisfied by increasing the sampling density, as such an increase allows N to take on large values. Recall that $r = r_{max} \left(N/N_{max}\right)^{1/d}$. While (N/N_{max}) terms are bounded by one, the numerator of (3.4.6) is composed of terms that

behave as $1/\sqrt{N}$. Thus, provided that the denominator is well-conditioned, condition 2 may be satisfied by allowing for large enough N.

Before numerically demonstrating our main result, we give a practical interpretation of its conditions and implications. In doing so, we quantify the separation needed between the linear structure and the noise and curvature with a geometric uncertainty principle. Then in Section 3.4.3, we present evidence of a tighter main result through the Central Limit Theorem.

3.4.2 Interpreting the Bound

The bound (3.4.6) is difficult to interpret due to many complicated terms. For subspace tracking, the behavior of the bound as a function of scale is of more practical use than the bound's actual value. The scale at which the bound reaches its minimum and the scale(s) at which it becomes ill-conditioned are the quantities of interest. Thus we now analyze our main result using more practical and interpretable, albeit less sharp, bounds. The following bound is not as sharp as Main Result 1 owing mainly to a less precise treatment of the principal curvatures. We replace equation (3.3.27) with (3.3.28) for the expectation of the pure curvature term, and we analyze only leading-order behavior of each term. Neglecting the probability-dependent constants and the contributions of the $\frac{1}{N}LE^T$ and $\frac{1}{N}CE^T$ terms, we may write our main result as:

Interpretable Main Result 1.

$$\left\|P-\widehat{P}\right\|_F \leq \frac{2\sqrt{2}\,\frac{1}{\sqrt{N}}\left[\frac{K}{2}r^3 \;+\; \sigma^2\sqrt{d(D-d)}\right]}{\frac{r^2}{d+2} - K^2r^4\frac{(d+1)}{2(d+2)^2(d+4)} - \sigma^2\left(\sqrt{d}+\sqrt{D-d}\right)} \tag{3.4.7}$$

and we recall that $r = r_{max}(N/N_{max})^{1/d}$.

3.4.2.1 Revisiting the Noise-Curvature Trade-off Through the Condition Number

We may now use this more compact, yet less sharp, form (3.4.7) of our main result to provide some interpretation for the bound (3.4.6). Consider first the denominator. Assume that sampling is sufficiently dense such that N may become large. Doing so ensures condition 2 is met and we

may focus our attention on condition 1 (and our neglect of the $\frac{1}{N}LE^T$ and $\frac{1}{N}CE^T$ terms in the denominator of (3.4.6) is also justified). The denominator of (3.4.7) is of the form

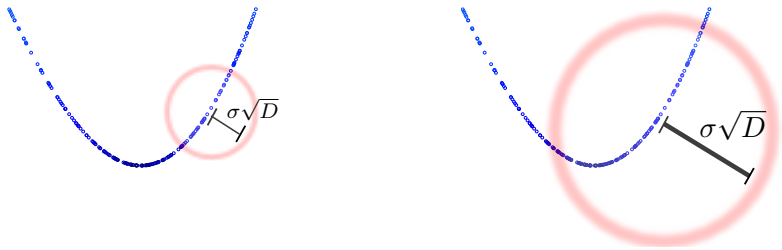
$$\delta \;=\; \frac{r^2}{d+2} - K^2 r^4 \frac{(d+1)}{2(d+2)^2(d+4)} - \sigma^2 \left(\sqrt{d} + \sqrt{D-d}\right). \tag{3.4.8}$$

It is now easy to see that δ quantifies the separation between the linear subspace ($\mathcal{O}(r^2)$) and the perturbation due to the curvature ($\mathcal{O}(K^2r^4)$) and the noise level ($\sigma^2(\sqrt{d} + \sqrt{D-d})$). To approximate the appropriate linear subspace, we must at least have that $\delta > 0$ as required by condition 1, and the approximation improves for larger δ. We note the similarity of this condition to that of equation (2.10) in [63].

The noise-curvature trade-off is now readily apparent. The linear and curvature contributions, controlled by $r = r_{max}(N/N_{max})^{1/d}$, are small for small values of N. Thus for N small, the denominator (3.4.8) is either negative or ill conditioned for most values of σ. This makes intuitive sense as we have not yet encountered much curvature but the linear structure has also not been explored. Therefore the noise dominates the early behavior of this bound and an approximating subspace may not be recovered from noise. As N increases, the conditioning of the denominator improves, and the bound is controlled by the $1/\sqrt{N}$ behavior of the numerator. This again corresponds with our intuition, as the addition of more points serves to overcome the effects of noise as the linear structure is explored. Thus, the bound becomes tighter. Eventually, N becomes large enough such that the curvature contribution approaches the size of the linear contribution, and δ^{-1} becomes large. The $1/\sqrt{N}$ term is overtaken by the ill conditioning of the denominator and the bound is forced to become large. The noise-curvature trade-off, seen analytically here in (3.4.8), will be demonstrated numerically in Section 3.5.

3.4.2.2 Geometric Uncertainty Principle for Subspace Recovery

Imposing condition 1 on (3.4.8) yields a geometric uncertainty principle quantifying the amount of curvature and noise we may tolerate. Solving for the range of scales such that $\delta > 0$, the following condition naturally arises. To recover an approximating subspace, we must have that:



(a) (b)

Figure 3.1: Illustration of the geometric uncertainty principle (3.4.9). For a manifold of fixed curvature K, (a) shows an acceptable noise level such that the geometry of the data remains intact and a tangent plane may be approximated from the noisy data. (b) illustrates a violation of the uncertainty principle as the manifold geometry may be destroyed by the noise. In this case a tangent plane approximation cannot be recovered.

Geometric Uncertainty Principle.

$$K\sigma \ < \ \sqrt{\frac{(d+4)}{2(d+1)(\sqrt{d}+\sqrt{D-d})}}. \tag{3.4.9}$$

By preventing curvature and noise from simultaneously becoming large, this requirement ensures that the geometry of the data is not destroyed by noise. With high probability, the noise concentrates on a sphere with mean curvature $1/\sigma\sqrt{D}$. Intuitively, we expect to require that the curvature of the manifold be less than the curvature of this noise-ball. Recalling the definitions of K_i and K from equations (3.3.4) and (3.3.5), K_i/d is the mean curvature in codimension i. The quadratic mean of the $(D-d)$ mean curvatures is then given by $K/d\sqrt{D-d}$ and we denote this normalized version of curvature as $\overline{K}$. Then (3.4.9) requires that $\overline{K} < \mathcal{O}(1/\sigma d D^{\frac{3}{4}})$. Noting that $d \geq 1 > D^{-\frac{1}{4}}$, the uncertainty principle (3.4.9) indeed may be interpreted as a requirement that the mean curvature of the manifold be less than that of the perturbing noise-ball. Figure 3.1 provides an illustration.

It is important to keep in mind that equation (3.4.9) is computed using the compact bound (3.4.7), and is thus meant for interpretation. For the precise expression represented by (3.4.9), the derivation must start with the full bound (3.4.6).

3.4.3 Towards a Tighter Bound: Chasing the Constants

Thus far we have presented a rigorous analysis bounding the norm of each perturbation term. The analysis captures leading order behavior with high probability by utilizing Theorem 13, but does so at the cost of attaching large constants to each term. Theorem 13, a result derived from the bounded difference and Hoeffding inequalities [70], introduces constants based on suprema of functions of random variables taken over the support of their distributions. Accordingly, each perturbation term is shown to deviate from its expectation by factors larger than constant multiples of its variance.

In this section we use the Central Limit Theorem (CLT) to show that the variance of the perturbation terms controls the deviation from their expectations. Doing so yields tighter bounds for each term and a tighter overall main result. Further, by working with moments of the underlying random variables rather than norms, a more precise treatment of curvature terms is possible, allowing a relaxation of the assumption in Main Result 1 that all principal curvatures in codimension i have the same sign. Despite the fact that our analysis is most often to be applied to sample sizes on the order of $N = 10^5$ or 10^6, we must acknowledge that the sample means with which we work have a Gaussian distribution only in the limit as N tends to infinity. This finite-sample analysis can be made rigorous through the use of Bernstein-type inequalities and concentration of measure (in fact such approaches yield only sightly larger constants). However, we proceed with a CLT-based analysis, treating convergence in distribution as equality, to provide evidence that a tight bound may be achieved and to motivate future analyses that may rigorously yield tighter constants.

3.4.3.1 Central Limit Theorem and Gaussian Tail Bounds

As previously seen, each entry of a matrix $\frac{1}{N}AB^T$ is a sample mean of N i.i.d. random variables. The CLT and a Gaussian tail bound yield a confidence interval for such an entry. Using a union bound to simultaneously control all of the entries of this matrix, we may give an overall confidence interval for the value of its Frobenius norm. While such an analysis yields a tighter

result than that using Theorem 13, it holds with lower probability due to the use of many union bounds.

REMARK. It is important to note that the probability attached to this analysis corresponds to the result serving as an upper bound for the true subspace recovery error. The probability is increased by choosing large values for the constants appearing in the exponential terms. This in turn loosens the result so as to bound any random fluctuation of the true error from above. In many practical applications, we are most interested in tracking the recovery error regardless of whether we guarantee an upper bound. When the need for an upper bound is relaxed, we will demonstrate in Section 3.5 that the leading order behavior of the following analysis tightly tracks the trend of the true error curve.

The analysis proceeds as follows. An entry of matrix $\frac{1}{N}\widetilde{A}\widetilde{B}^T$ has the form

$$\left(\frac{1}{N}\widetilde{A}\widetilde{B}^T\right)_{i,j} = \frac{1}{N}\sum_{k=1}^{N} A_{i,k}B_{j,k} - \widehat{\mathbb{E}}[A_i]\widehat{\mathbb{E}}[B_j], \tag{3.4.10}$$

where A and B represent the matrices L, C, and E from equation (3.1.7). We use the CLT to assert that for the i.i.d. random variables $(A_{i,k}B_{j,k})$ with mean μ and variance σ^2, $k = 1, \ldots, N$, and for large N,

$$\frac{1}{N}\sum_{k=1}^{N} A_{i,k}B_{j,k} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right). \tag{3.4.11}$$

Let $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Then we have

$$\text{Prob}\{|Y - \mu| \geq \epsilon\} \leq \exp\left\{-\frac{\epsilon^2}{2\sigma^2}\right\} \tag{3.4.12}$$

and we may pick $\epsilon \sim \mathcal{O}(\sigma)$ to ensure that $|Y - \mu| < \epsilon$ with high enough probability. Setting $\epsilon = \eta\sigma\sqrt{2}$ gives

$$Y \in \left[\mu - \eta\sigma\sqrt{2},\ \mu + \eta\sigma\sqrt{2}\right]$$

with probability greater than $1 - e^{-\eta^2}$, and we have a confidence interval that depends on the variance of Y and whose probability is controlled by the constant $\eta > 0$.

We compute the size of the entries in Δ and detail the norm bounds in Appendix A.5. The main result is now stated after defining some constants.

3.4.3.2 A Tighter Main Result

Just as we needed notation for the sake of the readability of Main Result 1, we now define new constants for Main Result 2:

$$\mathcal{CC}_{ij} = \left\{ \frac{2\left[(d+1)K^{ij}_{nn} - K^{ij}_{mn}\right]}{(d+2)(d+4)} + \left[\frac{1}{\sqrt{N}} \left(\eta_{CC_1} K^{ij}_{nn} \sqrt{\frac{48(d+1)(4d+17)}{(d+4)^2(d+6)(d+8)}} \right.\right.\right.$$
$$+4\eta_{CC_2} K^{ij}_{mn} \sqrt{\frac{(d^2+5d+3)}{(d+4)^2(d+6)(d+8)}} + \frac{4\eta_C K_i K_j}{(d+2)} \sqrt{\frac{d+1}{d+4}}$$
$$\left.\left.\left. + \frac{1}{\sqrt{N}} \frac{4\eta_C^2 K_i K_j (d+1)}{(d+2)(d+4)} \right) \right] \right\},$$

$$\mathcal{LC}_{ij} = \left[K_j \left(\frac{\eta_{LC_1}\sqrt{3}}{\sqrt{(d+4)(d+6)}} + \frac{\eta_L}{(d+2)} \right) + \kappa_i^{(j)} \frac{\sqrt{3}\left(\eta_{LC_2}\sqrt{5} - \eta_{LC_1}\right)}{\sqrt{(d+4)(d+6)}} \right.$$
$$\left. + K_j \frac{1}{\sqrt{N}} \frac{2\eta_L \eta_C}{(d+2)} \sqrt{\frac{d+1}{d+4}} \right],$$

$$\mathcal{CE}_i = \left[\frac{\eta_{CE}}{\sqrt{d+4}} \sqrt{3K^{ii}_{nn} + K^{ii}_{mn}} + \frac{\eta_E}{\sqrt{d+2}} K_i \left(1 + \frac{2\eta_C}{\sqrt{N}} \sqrt{\frac{d+1}{d+4}} \right) \right],$$

$$\mathcal{EE}(x) = \sigma^2 \sqrt{x} \left[\sqrt{2} \left(\eta_{EE_1}\sqrt{2} + \eta_{EE_2}\sqrt{x-1} \right) + \frac{2}{\sqrt{x}} \eta_E^2 \left(1 + \sqrt{x-1} \right) \right],$$

Additionally, we will need

$$K' = \left(\sum_{i=d+1}^{D} \sum_{j=d+1}^{D} \mathcal{CC}_{ij}^2 \right)^{\frac{1}{2}}, \qquad L' = \sqrt{d(D-d)} \left[\eta_{LE} + \eta_L \eta_E \frac{2}{\sqrt{N}} \right],$$

$$K'' = \left(\sum_{i=1}^{d} \sum_{j=d+1}^{D} \mathcal{LC}_{ij}^2 \right)^{\frac{1}{2}}, \qquad E' = \sqrt{d(D-d)} \left[\eta_{EE_2} + \eta_E^2 \frac{2}{\sqrt{N}} \right],$$

$$K''' = \left(\sum_{i=d+1}^{D} \mathcal{CE}_i^2 \right)^{\frac{1}{2}},$$

and

$$\zeta' = K'''\sigma\sqrt{\frac{2(D-d)}{d+2}}r_{max}^2\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}} + 2L'\sigma d\sqrt{\frac{2}{d+2}}r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}} + \mathcal{EE}(d) + \mathcal{EE}(D-d).$$

After using the triangle inequality in the exact same manner as in the beginning of Section 3.4, combining the norm bounds (computed in Appendix A.5) with Theorem 12 yields a new main result. Assume that conditions 1 and 2 hold. Then using the rescaled notation $r = r_{max}\,(N/N_{max})^{1/d}$, we have:

Main Result 2.

$$\left\|P - \widehat{P}\right\|_F \;\leq\; \frac{2\sqrt{2}\,\frac{1}{\sqrt{N}}\left[\frac{K''r^3}{\sqrt{2(d+2)}} + K'''\sigma r^2\sqrt{\frac{d}{2(d+2)}} + L'\sigma r\sqrt{\frac{2}{d+2}} + 2\sigma^2E'\right]}{\frac{r^2}{(d+2)} - \frac{K'r^4}{4(d+2)} - \sigma^2\left(\sqrt{d}+\sqrt{D-d}\right) - \frac{1}{\sqrt{N}}\zeta'} \tag{3.4.13}$$

with probability greater than

$$1 - (D-d)^2\left[de^{-\eta_{CC_1}^2} + \frac{d(d-1)}{2}e^{-\eta_{CC_2}^2}\right] - De^{-\eta_{EE_1}^2} - \frac{D^2-D}{2}e^{-\eta_{EE_2}^2}$$
$$- d(D-d)\left[(d-1)e^{-\eta_{LC_1}^2} + e^{-\eta_{LC_2}^2}\right] - dDe^{-\eta_{LE}^2} - D(D-d)e^{-\eta_{CE}^2}$$
$$- de^{-\eta_L^2} - de^{-\eta_C^2} - De^{-\eta_E^2}$$

over the joint random selection of the sample points and random realization of the noise.

The comments following Main Result 1 apply here as well. In particular, conditions 1 and 2 have the same interpretation and the denominator δ controls the conditioning of the recovery problem. Further, we note that the leading order behavior of the perturbation norms has not changed. Main Result 2 exhibits the same behavior as Main Result 1, but provides a tighter tracking of the true subspace recovery error, as will be shown in Section 3.5. Thus this analysis gives rise to the same interpretable bound as (3.4.7), up to multiplicative constants. The same geometric uncertainty principle applies as well.

Table 3.1 shows the leading order behavior for each perturbation norm. We note that $r_{max}\,(N/N_{max})^{\frac{1}{d}}$ has been replaced by r and only the leading order in d is shown. This side-by-side comparison reveals the reasons why Main Result 2 is a tighter bound. As suprema terms are replaced by variance terms, the CLT result introduces powers of $1/d$ that reduce the size of many norms. Additionally, the approach of Main Result 2 allows for a more precise treatment of the principal curvatures, most importantly for the CL^T term. This precise treatment allows Main Result 2 to relax the assumption that all principal curvatures in codimension i have the same sign, as required by Main Result 1. Finally, notice that Theorem 13 introduces probability constants of the form $(2+\eta\sqrt{2})$, whereas the CLT introduces probability constants of the form $\eta\sqrt{2}$. Thus the CLT yields tighter bounds.

3.4.4 Consistency with Previously Established Results

In [71], Singer and Wu study local PCA for tangent plane recovery in the absence of noise. The covariance matrix is decomposed following the assumption that for a given r, the number of points in a ball of radius r is large, implying a model with variable density. Fixing a density allows us to translate their results to our model. Then the covariance matrix decomposition yields error terms corresponding to curvature of size $\mathcal{O}(r^4)$ and finite-sample variance of sizes $\mathcal{O}(r^2)/\sqrt{N}$, $\mathcal{O}(r^3)/\sqrt{N}$, and $\mathcal{O}(r^4)/\sqrt{N}$. Recalling that $\|\frac{1}{N}CC^T\|_F \sim \mathcal{O}(r^4)+\mathcal{O}(r^4)/\sqrt{N}$ and $\|\frac{1}{N}CL^T\|_F \sim \mathcal{O}(r^3)/\sqrt{N}$, our analysis recovers the same leading order behavior as reported by Singer and Wu in the noise-free setting.

As previously discussed, Nadler presents a finite-sample PCA analysis in [63] assuming a linear model. Setting curvature terms in Main Results 1 and 2 to zero recovers Nadler's leading order bound on the angle between the finite-sample eigenvector(s) and the true eigenvector(s). In our notation, Nadler reports that, to leading order, the angle is bounded by

$$\sin\theta_{\widehat{U}_1,U_1} \;\lesssim\; \frac{\sigma}{\sqrt{\lambda_d}}\sqrt{\frac{D}{N}} + \mathcal{O}(\sigma^2),$$

where d is taken to be one. We now show that our results recover this leading order behavior.

Norm	**Main Result 1** (top) / **Main Result 2** (bottom)
$\|U_2^T\widetilde{C}\widetilde{C}^TU_2\|_F$	$\frac{r^4}{2d^3}\left[\sum_{i=d+1}^{D}\sum_{j=d+1}^{D}[(d+1)K_{nn}^{ij}-K_{mn}^{ij}]^2\right]^{\frac{1}{2}}$
	$\frac{r^4}{2d^3}\left[\sum_{i=d+1}^{D}\sum_{j=d+1}^{D}[(d+1)K_{nn}^{ij}-K_{mn}^{ij}]^2\right]^{\frac{1}{2}}$
$\|U_1^T\widetilde{E}\widetilde{E}^TU_1\|_F$	$\sigma^2\sqrt{d}$
	$\sigma^2\sqrt{d}$
$\|U_2^T\widetilde{E}\widetilde{E}^TU_2\|_F$	$\sigma^2\sqrt{D-d}$
	$\sigma^2\sqrt{D-d}$
$\|U_2^T\widetilde{E}\widetilde{E}^TU_1\|_F$	$\frac{\sigma^2}{\sqrt{N}}\left[\sqrt{d(D-d)}+\xi_e\sqrt{2}\left(\sqrt{d}+\sqrt{D-d}\right)+2\xi_e^2\right]\left(2+\eta_{ee}\sqrt{2}\right)$
	$\frac{\sigma^2}{\sqrt{N}}\sqrt{2d(D-d)}\eta_{EE_2}$
$\|U_2^T\widetilde{C}\widetilde{L}^TU_1\|_F$	$\frac{1}{\sqrt{N}}\frac{r^3}{2}K\left(2+\eta_{\ell c}\sqrt{2}\right)$
	$\frac{1}{\sqrt{N}}\frac{r^3}{d\sqrt{d}}\sqrt{\frac{3}{2}}\left[\sum_{i=1}^{d}\sum_{j=d+1}^{D}\left(K_j(\eta_{LC_1}+\frac{\eta_L}{\sqrt{3}})+\kappa_i^{(j)}(\eta_{LC_2}\sqrt{5}-\eta_{LC_1})\right)^2\right]^{\frac{1}{2}}$
$\|U_1^T\widetilde{E}\widetilde{L}^TU_1\|_F$	$\frac{\sigma}{\sqrt{N}}r\left(\sqrt{d}+\xi_e\sqrt{2}\right)\left(2+\eta_{\ell e}\sqrt{2}\right)$
	$\frac{\sigma}{\sqrt{N}}r\sqrt{2d}\eta_{LE}$
$\|U_2^T\widetilde{E}\widetilde{L}^TU_1\|_F$	$\frac{\sigma}{\sqrt{N}}r\left(\sqrt{D-d}+\xi_e\sqrt{2}\right)\left(2+\eta_{\ell e}\sqrt{2}\right)$
	$\frac{\sigma}{\sqrt{N}}r\sqrt{2(D-d)}\eta_{LE}$
$\|U_2^T\widetilde{C}\widetilde{E}^TU_1\|_F$	$\frac{\sigma}{\sqrt{N}}\frac{r^2}{2}K\left(\sqrt{d}+\xi_e\sqrt{2}\right)\left(2+\eta_{ce}\sqrt{2}\right)$
	$\frac{\sigma}{\sqrt{N}}r^2\left[\sum_{i=d+1}^{D}\left(\frac{\eta_{CE}}{\sqrt{2(d+4)}}\sqrt{3K_{nn}^{ii}+K_{mn}^{ii}}+\frac{\eta_E}{\sqrt{2(d+2)}}K_i\right)^2\right]^{\frac{1}{2}}$
$\|U_2^T\widetilde{C}\widetilde{E}^TU_2\|_F$	$\frac{\sigma}{\sqrt{N}}\frac{r^2}{2}K\left(\sqrt{D-d}+\xi_e\sqrt{2}\right)\left(2+\eta_{ce}\sqrt{2}\right)$
	$\frac{\sigma}{\sqrt{N}}r^2\sqrt{\frac{D-d}{d}}\left[\sum_{i=d+1}^{D}\left(\frac{\eta_{CE}}{\sqrt{2(d+4)}}\sqrt{3K_{nn}^{ii}+K_{mn}^{ii}}+\frac{\eta_E}{\sqrt{2(d+2)}}K_i\right)^2\right]^{\frac{1}{2}}$

Table 3.1: Comparison of leading order perturbation terms for Main Result 1 (top) and Main Result 2 (bottom). Notationally, $r_{max}\,(N/N_{max})^{\frac{1}{d}}$ has been replaced by r and only leading order d terms are shown.

First, set all curvature terms to zero. Next, assume condition 1 holds such that the denominator δ is sufficiently well conditioned and we may neglect all terms other than λ_d. Using the more compact notation of r in place of $r_{max}\,(N/N_{max})^{\frac{1}{d}}$ and following the approach in [63] of retaining only leading order terms and dropping probability constants, we have

$$\sin\theta_{\widehat{U}_1,U_1} \lesssim \frac{\frac{\sqrt{D-d}}{\sqrt{N}}\left(\sigma r+\sigma^2\sqrt{d}\right)}{\frac{r^2}{(d+2)}} = \frac{\sigma}{r}\frac{(d+2)\sqrt{D-d}}{\sqrt{N}}+\mathcal{O}(\sigma^2).$$

Taking $d=1$ and noticing that $\sqrt{\lambda_d}\sim r$ yields Nadler's bound.

3.5 Numerical Results

In this section we demonstrate that the bounds of Main Results 1 and 2 accurately and efficiently track the true subspace recovery error and may therefore be used to obtain the optimal scale for tangent plane recovery. We then address the case of data sampled from a saddle (such that the principal curvatures are of mixed signs) that brings to light a particular difference between the two main results. We explain and numerically demonstrate the behavior of the true subspace recovery error at large scales and the corresponding lack of subspace tracking and connect this observation to the "crossover phenomenon" reported in [63]. Finally we demonstrate the accurate and efficient recovery of local curvature.

3.5.1 Subspace Tracking and Recovery

We generate a data set sampled from a 3-dimensional manifold embedded in $\mathbb{R}^{20}$ according to the local model (3.1.2). The radius of the local model is set to $r_{max}=1$ and $N=1.25\times10^6$ points are uniformly sampled from the tangent plane. Note that $2^{20}\approx10^6$, and thus this choice of N represents reasonable sampling in $\mathbb{R}^{20}$. Curvature and the standard deviation σ of the added Gaussian noise will be specified in each experiment.

We compare our bounds with the true tangent plane recovery error. The tangent plane at reference point x_0 is computed at each scale N via PCA of the N nearest neighbors of x_0. The true tangent plane recovery error $\|P-\widehat{P}\|_F$ is then computed at each scale. Note that computing the true

error requires N SVDs. A "true bound" is computed by applying Theorem 12 after measuring each perturbation norm directly from the data. While no SVDs are required, this true bound utilizes information that is not practically available, and therefore represents the best possible bound that we can hope to achieve. We will compare the mean of the true error and mean of the true bound over 10 trials (with error bars indicating one standard deviation) to the following three curves:

(1) Main Result 1 holding with probability 0.5 (magenta),

(2) Main Result 2 holding with probability 0.5 (black), and

(3) Main Result 2 with all probability constants set to 1 (green).

$\kappa_i^{(j)}$	$i = 1$	$i = 2$	$i = 3$
$j = 4, \ldots, 6$	3.0000	1.5000	1.5000
$j = 7, \ldots, 20$	1.6351	0.1351	0.1351

Table 3.2: Principal curvatures of the manifold for Figure 3.2-c.

The third curve abandons any guarantee of providing an upper bound in favor of capturing the trend of the true error. This curve represents the case where we may not care to upper bound the error, but instead wish to track the height of the true error curve as closely as possible. We will refer to these curves as "Main Result 1," "Main Result 2," and "Main Result 2-trend."

The results are displayed in figure 3.2. Panel (a) shows the noisy ($\sigma = 0.01$) curvature-free (linear subspace) result. As the only perturbation is due to noise, we expect the error to decay as $1/\sqrt{N}$ as the scale increases. The curves are shown on a logarithmic scale (for the Y-axis) and decrease monotonically, indicating the expected decay. As expected, Main Result 2 (black) is tighter than Main Result 1 (magenta) and both accurately track the behavior of the true error (blue). Main Result 2-trend (green) tightly tracks the same behavior, in this case sitting on top of the true bound (red). Panel (b) shows the results for a noise-free manifold with principal curvatures given in Table

3.2 such that $K = 12.6025$. Notice that three of the codimensions experience high curvature while the others are flatter, giving a tube-like structure to the manifold. In this case, perturbation is due to curvature only and the error increases monotonically (ignoring the slight numerical instability at extremely small scales), as predicted in the discussion of Section 3.4.2.1. Eventually, a scale is reached at which there is too much curvature and the bounds blow up to infinity. This corresponds exactly to where the true error plateaus at its maximum value, representing the fact that the computed subspace is now orthogonal to the true tangent plane. This large scale behavior will be further explained in Section 3.5.3.

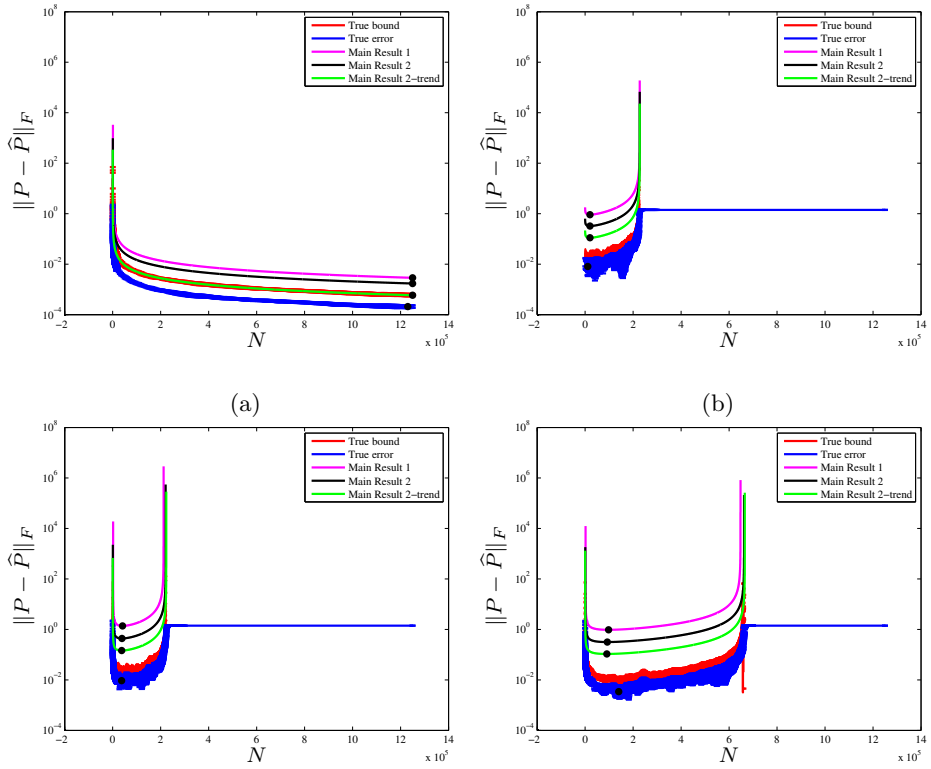


Figure 3.2: Norm of the perturbation: (a) flat manifold with noise, (b) curved (tube-like) manifold with no noise, (c) curved (tube-like) manifold with noise, (d) curved manifold with noise. Black dots indicate minima of the curves. Note the logarithmic scale on the Y-axes. See text for discussion.

Figure 3.2-c shows the results for a noisy ($\sigma = 0.01$) version of the manifold used in panel (b). Note that the true error is large at small scales due to noise and large at large scales due to curvature. At these scales the bounds are accordingly ill conditioned and track the behavior of the true error when well conditioned. Panel (d) shows the results for a manifold again with $K = 12.6025$, but with the principal curvatures equal in all codimensions ($\kappa_i^{(j)} = 1.0189$ for $i = 1, \ldots, 3$ and $j = 4, \ldots, 20$), and noise ($\sigma = 0.01$) is added. We observe the same general behavior as seen in panel (c), but both the true error and the bounds remain well conditioned at larger scales. This is explained by the fact that higher curvature is encountered at smaller scales for the manifold corresponding to panel (c) but is not encountered until larger scales in panel (d).

In all four plots, the bounds accurately track the behavior of the true error. In fact, the curves are shown to be parallel on a logarithmic scale, indicating that they differ only by multiplicative constants. Note also that the true bound (red) tightly tracks the true error (blue), providing evidence that the triangle inequalities used in computing the bounds are reasonably tight. As no matrix decompositions are needed to compute our bounds, we have efficiently tracked the tangent plane recovery error. The black dots in figure 3.2 indicate the minimum of each curve. In general we see agreement of the location at which the minima occur, indicating the scale that will yield the optimal tangent plane approximation. We note that when the location of the bounds' minima do not correspond with the minimum of the true error (such as in panel (d)), the discrepancy occurs at a range of scales for which the true error is quite flat. In fact, in panel (d), the difference between the error at the computed optimal scale and the error at the true optimal scale is on the order of 10^{-2}. Thus the angle between the computed and true tangent planes will be less than half of a degree. For a large data set it is impractical to examine every scale and one would instead most likely use a coarse sampling of scales. The true optimal scale would almost surely be missed by such a coarse sampling scheme. Our analysis indicates that despite missing the true optimum, we may recover a scale that yields an approximation to within a fraction of a degree of the optimum.

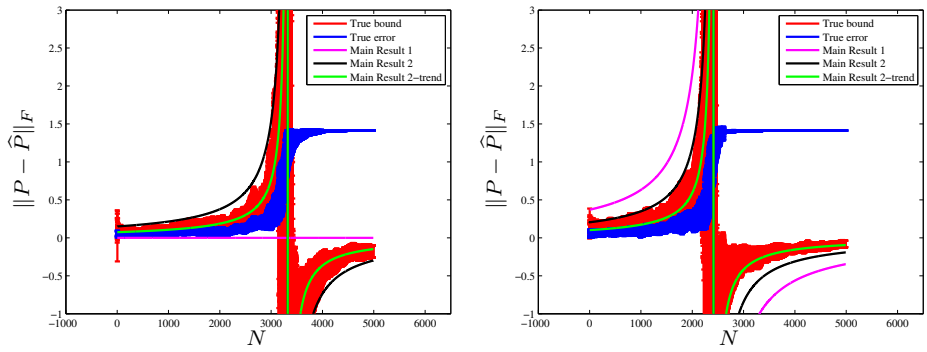


Figure 3.3: A 2-dimensional saddle (noise free) is shown with (a) $K = 0$ and (b) $K = 1$. Note that Main Result 1 is identically zero in (a) but accurately tracks the true error in (b). See text for discussion.

3.5.2 Principal Curvatures of Mixed Signs (Saddle)

As discussed in Section 3.3.2.1, a key difference between Main Results 1 and 2 is the ability of Main Result 2 to properly handle principal curvatures of mixed signs. Main Result 1 requires the assumption that all principal curvatures in codimension i have the same sign and thus cannot properly track the tangent plane recovery error for points sampled from a saddle. This is demonstrated in figure 3.3-a, showing the results for a 2-dimensional noise-free saddle ($d = 2, D = 3$) with principal curvatures $\kappa_1^{(3)} = 3$, and $\kappa_2^{(3)} = -3$. While all other curves behave as expected, the curve corresponding to Main Result 1 is identically zero because $K = 3 - 3 = 0$. Main Result 2, through its use of K_{nn}^{ij} and K_{mn}^{ij}, avoids this problem. Figure 3.3-b shows the results for a 2-dimensional noise-free saddle ($d = 2, D = 3$) with principal curvatures $\kappa_1^{(3)} = 4$, and $\kappa_2^{(3)} = -3$. Despite the fact that the assumption of Main Result 1 is violated (the principal curvatures are of mixed signs), the corresponding curve does in fact track the recovery error because $K = 4 - 3 = 1$ is not identically zero. The fact that the proper behavior is seen despite the violated assumption indicates that a tighter curvature analysis in Section 3.3.2.1 is possible.

3.5.3 Spectral Crossover at Large Scales

Here we examine the inability to track the proper subspace at large scales, which is clearly indicated by the ill conditioning of the bounds and the plateau of the true error at its maximum value in figures 3.2 and 3.3. We demonstrate that this is an effect of curvature.

In [63] it was shown that the PCA of a noisy linear subspace is prone to a "sudden loss of tracking" of the dominant eigenvector. This occurs when an eigenvalue corresponding to noise overtakes an eigenvalue corresponding to signal. In this setting, once the crossover has occurred, the dominant eigenvector may point in any random direction. Consider now our geometric model and let the sample points be noise-free to demonstrate a similar phenomenon owing to geometry rather than noise. Recall that, in this setup, condition 1 requires there be sufficient separation between the spectrum of the linear structure and the spectrum of the curvature. Also recall that δ^{-1} is the corresponding condition number. When δ^{-1} becomes large, there is little separation of the spectra, and a curvature eigenvalue approaches a tangent plane eigenvalue. Once the curvature eigenvalue becomes larger than the tangent plane eigenvalue, the computed eigenspace contains a direction orthogonal to the true subspace. This is seen in figure 3.2-b and figure 3.3 where the bounds blow up to infinity and the true error plateaus at its maximum value indicating orthogonality. As the crossover is due to curvature, an eigenvector in a direction orthogonal to the true tangent plane is introduced into the top d computed eigenvectors. Thus the computed and true tangent planes are orthogonal at large scales.

Numerical evidence of this phenomenon is given in figure 3.4. The eigenvalues (mean over 10 trials) corresponding to the saddle from figure 3.3-b ($d = 2$, $D = 3$, $\kappa_1^{(3)} = 4$, $\kappa_2^{(3)} = -3$) are plotted as a function of scale. At small scales, the two tangent plane eigenvalues (blue and red) dominate the curvature eigenvalue (green) and the subspace recovery error is well conditioned at small scales in figure 3.3-b. Notice that at roughly $N = 2500$, the curvature eigenvalue crosses the two tangent plane eigenvalues. After this scale, the largest (blue) eigenvalue corresponds to curvature but is now included in the computed tangent plane. Thus the computed tangent plane

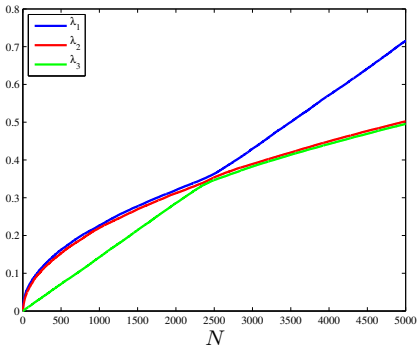


Figure 3.4: The eigenvalues computed from the saddle in figure 3.3b are plotted as a function of scale. Note the crossover between the curvature and tangent plane eigenvalues at roughly $N = 2500$, corresponding to the lack of subspace tracking at the same scale in figure 3.3b.

contains a direction orthogonal to the true tangent plane. This is seen in figure 3.3-b as the true error (blue) plateaus at its maximum value. At this large scale the bounds become ill conditioned (or negative) as condition 1 ($\delta > 0$) is violated. As there is no noise in this example, the crossover phenomenon is similar to that reported in [63], but is the result of curvature at large scales rather than noise.

3.5.4 Recovering Neighborhood Curvature

The expectation of the curvature term in each codimension has the following form:

$$\mathbb{E}[C_i] \;=\; \frac{K_i}{2}\frac{r_{max}^2}{(d+2)}\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}}. \tag{3.5.1}$$

Thus given data in the coordinate system aligned with the principal directions described in Section 3.1.2, we may track the trajectory of the center as a function of scale and compare it to (3.5.1) to recover an estimate of K_i for $i = (d+1), \ldots, D$. Table 3.3 shows that this procedure results in a very accurate recovery of the local curvature at low noise levels, and the recovery becomes worse as the noise increases and the limit of the geometric uncertainty principle (3.4.9) is approached. We note that accuracy improves as N becomes large, as expected by the CLT. Note that using (3.5.1), we recover the individual K_i's from which the overall K is computed (by equation (3.3.5)) and is reported in the tables. While this method does recover each K_i, the individual principal curvatures

$\kappa_n^{(i)}$ are not recovered. As it does not

	$K = 2$	$K = 10$	$K = 20$
$\sigma = 0$	1.9989 ± 0.0038	10.0079 ± 0.0145	20.0047 ± 0.0476
$\sigma = 0.005$	2.0020 ± 0.0032	10.0096 ± 0.0251	19.9903 ± 0.0406
$\sigma = 0.01$	2.0005 ± 0.0048	9.9952 ± 0.0202	20.0051 ± 0.0478
$\sigma = 0.025$	1.9928 ± 0.0044	9.9516 ± 0.0279	19.9104 ± 0.0428
$\sigma = 0.1$	1.8877 ± 0.0705	8.8781 ± 0.0808	17.8829 ± 0.0949

Table 3.3: Estimation of curvature at different noise levels ($d = 5$, $D = 20$, $N = 10^4$). The mean and standard deviation are reported from 10 trials. The estimation is accurate for low levels of noise and loses accuracy as the noise level increases. Note that the individual K_i's are recovered from which the overall K is computed according to equation (3.3.5).

require matrix decompositions and only uses vector addition, this method is computationally efficient. If one is willing to perform N SVDs, this method combined with the analysis of [17] might yield the individual principal curvatures.

While it is unrealistic for data to be observed in the desired coordinate system aligned with the principal directions, tracking the trajectory of the center in each dimension yields the rotation necessary to transform to this coordinate system. Further, tracking the trajectory may yield a clean estimate of the reference point of the local model in the presence of noise. While the noise renders this trajectory unstable at small scales, it is very stable at scales above the noise level. Using the stability of the trajectory at large scales may allow us to extrapolate back and accurately recover the trajectory at small scales, yielding an estimate of the "denoised" reference point.

3.6 Algorithmic Considerations and Future Work

There are several algorithmic issues to be considered in implementing this analysis for optimal tangent plane recovery. Such considerations are topics of our current research and we give a brief

discussion here.

3.6.1 Parameter Recovery

In any practical use of this analysis (and in keeping with its spirit), each of the parameters d, r_{max}, K, and σ must be recovered from the data itself rather than estimated by an *a priori* fixed value.

d. There exist algorithms for estimating the (local) intrinsic dimensionality of a data set. The recent work in [17] presents a multiscale approach to estimate d in a pointwise fashion. Performing an SVD at each scale, d is determined by examining growth rate of the multiscale singular values. It would be interesting to investigate if this approach remains robust if only a coarse exploration of the scales is performed, as it may be possible to reduce the computational cost through an SVD-update scheme. Another scale-based approach is presented in [79] and the problem was studied from a dynamical systems perspective in [31].

$\mathbf{r_{max}}$. The maximum radius for which the local model (3.1.2) is valid may be estimated by a multiscale partitioning of the data set. Partitioning from fine to coarse, regions that produce similar tangent plane estimates at the same scale may be merged. Such an approach is similar to the aggregation process in [52], hierarchical clustering [19], data partitioning to find affine subspaces ("flats") [75], subspace arrangement for homogeneous data subsets [57], and spectral clustering [2].

K. We have demonstrated our ability to recover K given data in the coordinate system described in Section 3.1.2. Additionally we have discussed how tracking the trajectory of the centering may yield both the rotation into the desired coordinate system as well as a clean estimate of the otherwise noisy reference point. The accuracy and stability of such a scheme remains to be tested and it will be interesting to investigate if this may be a path to a simpler recovery of the tangent plane.

It is worth mentioning that while the definitions of K and K_i used in this work arise naturally

from the analysis, they are not the only possible definitions. One could define

$$K_i = \left(\sum_{n=1}^{d} \left(\kappa_n^{(i)} \right)^2 \right)^{\frac{1}{2}}. \tag{3.6.1}$$

This definition does not match the calculation for $\mathbb{E}[C_i]$ but has the advantage of handling negative principal curvatures in a more natural manner. Indeed, we have seen that Main Result 1 does not hold for the case of a saddle with $K_i = 0$, but in fact does hold for a saddle with $K_i \neq 0$, thereby indicating that this bound may hold in a more general context.

σ. There exist many statistical methods for estimating the noise level present in a data set (see, for example, [12, 24]). In [17], the smallest multiscale singular values are used as an estimate for the noise level and a scale-dependent estimate of noise variance is suggested in [29] for curve-denoising.

The parameters of our analysis may not remain constant over the entire data set. It is possible, if not likely, to experience very different sampling densities, noise-levels, curvature and dimensionality as one explores different regions of a data set. This fact increases the need for careful parameter selection and emphasizes the importance of a local analysis. Initial experiments indicate that Main Results 1 and 2 are sensitive to changes in these parameters. For example, over/under estimating K or σ will result in ill conditioning at smaller/larger scales than seen in the true error. In depth experimentation will be necessary to precisely quantify the robustness of the results to parameter perturbation.

3.6.2 Sampling

For a tractable analysis, assumptions about sampling must be made. In this work we have assumed uniform sampling in the tangent plane. This is merely one choice and we have conducted initial experiments uniformly sampling the manifold rather than the tangent plane. Results suggest that for a given radius, sampling the manifold yields a smaller curvature contribution than does sampling the tangent plane. While more rigorous analysis and experimentation is needed, it is clear that consideration must be given to the sampling assumptions of any practical algorithm.

3.6.3 From Tangent Plane Recovery to Data Parameterization

The tangent plane recovered by our approach may not provide the best approximation over the entire neighborhood from which it was derived. Depending on a user-defined error tolerance, a smaller or larger sized neighborhood may be parameterized by the local chart. If high accuracy is required, one might only parameterize a neighborhood of size $N < N^*$ to ensure the accuracy requirement is met. Similarly, if an application requires only modest accuracy, one may be able to parameterize a larger neighborhood than that given by N^*.

Finally, we may wish to use tangent planes recovered from different neighborhoods to construct a covering of a data set. There exist methods for aligning local charts into a global coordinate system (for example [10, 68, 83], to name a few). Care should be taken to define neighborhoods such that a data set may be optimally covered.

Chapter 4

Local Analysis of Global Data

4.1 Approximation of Data and Estimation of Geometry

4.1.1 Local PCA

Chapter 3 provides an analysis for estimating the local tangent plane from noisy samples of a nonlinear manifold. An optimal basis for a linear subspace may be computed from noisy samples via principal component analysis (PCA). Estimating the local tangent plane therefore becomes a problem of finding the proper scale at which to perform local PCA. The scale must be large enough to be above the level of the noise but small enough to avoid the nonlinear curvature of the manifold. In this chapter, we study the connection between local tangent plane estimation and local PCA approximation of a global data set.

The goal of PCA approximation is to concisely describe as much of the variation in the data as possible. In the same way that our tangent plane estimation analysis requires a local data model, PCA-based approximation of a nonlinear data set requires a partitioning of the points into local clusters. For example, see (amongst many others) [33, 46, 83] for local PCA in several contexts, as well as [25] and the references therein. In this chapter, we study the problem of breaking a manifold-valued data set into local partitions. Our goal is not to develop a new algorithm but instead to study the relationship between estimation of geometry and approximation of data. We present a geometric analysis of a generic approximation-based partitioning algorithm and numerically study the correspondence of the returned partitioning to the optimal scale for tangent plane estimation.

Our analysis reveals that the approximation algorithm is guided by the same noise-curvature trade-off that is at the heart of the tangent plane recovery analysis of the previous chapter. Further, we numerically observe that the partitions are of appropriate size for local tangent plane recovery.

The chapter is organized as follows. The remainder of this section describes the partitioning algorithm. Our geometric analysis is presented in Section 4.2 and numerical experiments studying the partitioning of noisy data sets are presented in Section 4.3. We also return to the problem of tangent plane recovery in Section 4.3 with a comparison of the local partitioning and tangent plane estimation. We conclude in Section 4.4 with a discussion of algorithmic considerations for the partitioning of a noisy, manifold-valued, data set.

4.1.2 A Generic Partitioning Algorithm

In [25], Einbeck, Evers, and Bailer-Jones review several approaches for using local PCA on complex data sets. Seeking to divide the data into clusters, an iterative "cluster-wise" algorithm is outlined, combining the K-means clustering algorithm with PCA. Given a partitioning or clustering of a data set, K-means (also known as the Generalized Lloyd algorithm [55]) updates the partitions by reclustering those points closest to the center of mass of each partition. The centers of mass are then recomputed, points are reclustered, and the process continues until convergence. Given an initial partitioning, the cluster-wise PCA algorithm updates the partitioning by replacing the center-of-mass calculation with a local PCA within each cluster. Points are then reassigned to clusters based on proximity to the hyperplane segment defined by the local PCA of each cluster, and the process is iterated in a manner analogous to K-means.

It is well known that the K-means algorithm is very sensitive to the choice of initial partitioning. Eschewing the random partitioning typically used to initialize K-means, the authors propose to augment the cluster-wise PCA algorithm with a recursive partitioning of the data based on a simple criterion [25]. Starting with an initial partition consisting of the entire data set, the partition is split if doing so yields a better PCA approximation. The process is recursively repeated. Formally, the partitioning proceeds as follows [25]:

(1) Given a partition $R^{(q)}$ containing $n^{(q)}$ points in $\mathbb{R}^D$, define $\lambda_j^{(q)}$ $(j = 1, \ldots, D)$ to be the jth largest eigenvalue of the covariance matrix of the data in $R^{(q)}$.

(2) Let $d < D$ be the target dimension for PCA approximation.

(3) Split $R^{(q)}$ at the mean of the partition orthogonally to the first principal component (the eigenvector associated with the largest eigenvalue of the covariance matrix). Denote the two resulting partitions as $R^{(l)}$ and $R^{(r)}$.

(4) Accept the split if

$$\frac{\sum_{j=1}^{d} \lambda_j^{(q)}}{\sum_{j=1}^{D} \lambda_j^{(q)}} \; < \; C \left(\frac{n^{(l)}}{n^{(q)}} \frac{\sum_{j=1}^{d} \lambda_j^{(l)}}{\sum_{j=1}^{D} \lambda_j^{(l)}} + \frac{n^{(r)}}{n^{(q)}} \frac{\sum_{j=1}^{d} \lambda_j^{(r)}}{\sum_{j=1}^{D} \lambda_j^{(r)}} \right) \tag{4.1.1}$$

for some constant $C > 0$. Otherwise reject the split.

(5) Recursively repeat this process for all partitions.

Partitioning a data set according to criterion (4.1.1) is quite intuitive from a statistical approximation perspective. PCA provides the best linear approximation to the data in the sense that it maximizes the variance captured over all possible d-dimension linear approximations. Recall that an eigenvalue of a covariance matrix is the variance in the data captured by the corresponding eigenvector (principal component). Thus the ratio of eigenvalues on the left-hand-side of (4.1.1) is the percent variance captured by the d-dimensional linear PCA approximation to the current partition. The right-hand-side is the weighted sum of percent variance captured by splitting the current partition into two smaller partitions. If splitting the partition captures more variance. it yields a better approximation and should therefore be accepted. If a split fails to provide a better approximation (or fails to provide a significantly better approximation, as controlled by the constant C), the original partition should be retained. Thus (4.1.1) describes a natural measure for the quality of PCA approximation and provides a criterion for partitioning a data set to maximize this quality of approximation.

Einbeck, Evers, and Bailer-Jones combine this partitioning algorithm with the K-means/PCA clustering scheme previously described to avoid sensitivity to initialization. In their algorithm,

partitions are split and remerged in an attempt to avoid the local minima of the K-means clustering. The authors comment that they typically choose the constant C to be one and that it is often helpful to require the algorithm to split the partitions during the first few levels of recursion regardless of whether or not criterion (4.1.1) indicates to do so.

As steps 1-5 outline a generic partitioning scheme, our goal is to analyze its performance from a geometric perspective. Doing so will provide intuition as to the relationship between the approximation-based criterion (4.1.1) and the estimation of geometry. As we explore the main steps of this recursion, we will also comment on implementation details. However, our focus will remain on estimating geometry and we do not aim to describe a complete algorithm in full detail.

4.2 Geometric Analysis

Here we study the generic partitioning algorithm from a geometric perspective. We begin by injecting the local geometric model studied in Chapter 3. To present a self-contained analysis, the local model, notation, and assumptions briefly reviewed here.

4.2.1 Local Model and Preliminaries

The analysis of this section is restricted to a manifold of codimension 1. The generalization to arbitrary codimension follows naturally.

A d-dimensional manifold of codimension 1 may be described locally by the surface $y = f(\ell_1, \ldots, \ell_d)$, where ℓ_i is a coordinate in the tangent plane. Choosing the coordinate system to align with the principal directions associated with the principal curvatures at a given reference point x_0, the manifold may be approximated by its Taylor series about x_0:

$$f(\ell_1, \ldots, \ell_d) = \frac{1}{2}(\kappa_1 \ell_1^2 + \cdots + \kappa_d \ell_d^2) + o(\ell_1^2 + \cdots + \ell_d^2) \tag{4.2.1}$$

where $\kappa_1, \ldots, \kappa_d$ are the principal curvatures of the manifold at x_0. We truncate this Taylor series and work with the local quadratic approximation:

$$f(\ell_1, \ldots, \ell_d) = \frac{1}{2}(\kappa_1 \ell_1^2 + \cdots + \kappa_d \ell_d^2). \tag{4.2.2}$$

In this coordinate system, the point $x_0 \in \mathbb{R}^D$ has the form

$$x_0 = [\ell_1 \ \ell_2 \ \cdots \ell_d \ f(\ell_1, \ldots, \ell_d)]$$

and points a the local neighborhood have similar coordinates.

Consider N discrete samples of a nonlinear d-dimensional Riemannian manifold, observed as points in $\mathbb{R}^D$ in the local coordinate system described above (with $D = d + 1$). Let each point be contaminated with an additive Gaussian noise vector e drawn from the $\mathcal{N}\left(0, \sigma^2 I_D\right)$ distribution. Each sample is a D-dimensional vector and may be stored as columns of a matrix $X \in \mathbb{R}^{D \times N}$. The chosen coordinate system allows the decomposition of a point x into its linear, quadratic, and noise components, ℓ, c, and e, respectively, three D-dimensional vectors

$$\ell = [\ell_1 \ \ell_2 \ \cdots \ \ \ell_d \ \ 0]^T$$

$$c = [0 \ \ 0 \ \ \cdots \ \ 0 \ \ c_D]^T$$

$$e = [e_1 \ e_2 \ \ \cdots \ \ \ e_D]^T$$

where

$$c_D = \frac{1}{2}(\kappa_1 \ell_1^2 + \ \ \ldots \ \ + \kappa_d \ell_d^2).$$

We may store the N samples of ℓ, c, and e as columns of matrices L, C, E, respectively, such that our data matrix is decomposed as

$$X = L + C + E. \tag{4.2.3}$$

The curvature K of the local model is quantified by the sum of the principal curvatures,

$$K = \sum_{n=1}^{d} \kappa_n, \tag{4.2.4}$$

and K^2 may be written as $K^2 = K_{nn} + K_{mn}$, where K_{nn} and K_{mn} are the like-indexed and cross-indexed terms sums,

$$K_{nn} = \sum_{n=1}^{d} \kappa_n^2, \qquad K_{mn} = \sum_{\substack{m,n=1 \\ m \neq n}}^{d} \kappa_m \kappa_n. \tag{4.2.5}$$

Finally, we sample the manifold by uniformly sampling points in the tangent plane falling inside $B^d_{x_0}(r)$, the d dimensional ball of radius r centered at reference point x_0. The scale of the analysis is quantified by the radius r within which we sample the points in the tangent plane. The number of points N captured inside $B^d_{x_0}(r)$ scales as $N \sim r^d$.

4.2.2 Eigenvalue Analysis

Given the decomposition (4.2.3), the sample covariance matrix $\frac{1}{N}XX^T$ has the form

$$\frac{1}{N}XX^T = \frac{1}{N}\left(LL^T + CC^T + EE^T + LC^T + CL^T + EL^T + LE^T + EC^T + CE^T\right). \tag{4.2.6}$$

To ease the analysis and develop geometric intuition, we proceed by neglecting the cross-terms in (4.2.6) and consider the covariance matrix

$$\frac{1}{N}XX^T = \frac{1}{N}\left(LL^T + CC^T + EE^T\right). \tag{4.2.7}$$

Note that the neglected terms are zero in expectation and only those remaining terms have nonzero expectation. We further assume the data to be centered as required for PCA, but do not carry the $\widetilde{X}$, $\widetilde{L}$, $\widetilde{C}$, or $\widetilde{E}$ notation from Chapter 3.

Our goal is to recast the partitioning criterion (4.1.1) in terms of the geometry encoded by L, C, and E. We therefore study the spectra of $\frac{1}{N}LL^T$, $\frac{1}{N}CC^T$, and $\frac{1}{N}EE^T$, before appealing to spectral theory to characterize the spectrum of $\frac{1}{N}XX^T$ in terms of its decomposition.

Let $\lambda_j(A)$ denote the jth largest eigenvalue of matrix A, ordered by magnitude, and recall that the eigenvalues of each covariance matrix correspond to the variance of the underlying random variables. The moments of ℓ_i are calculated in Appendix A.4, and we have that $\mathrm{Var}[\ell_i] = r^2/(d+2)$. Trivially, we also have $\mathrm{Var}[e_i] = \sigma^2$. In Chapter 3, we calculated $\mathbb{E}[c_D] = \frac{Kr^2}{2(d+2)}$, and the variance

calculation for c is as follows:

$$\begin{aligned}
\mathrm{Var}[c_D] &= \mathbb{E}[c_D^2] - (\mathbb{E}[c_D])^2 \\
&= \mathbb{E}\left[\left(\frac{1}{2}(\kappa_1\ell_1^2 + \cdots + \kappa_d\ell_d^2)\right)^2\right] - \frac{K^2r^4}{4(d+2)^2} \\
&= \frac{1}{4}\sum_{n=1}^{d}\kappa_n^2\,\mathbb{E}[\ell_n^4] + \frac{1}{4}\sum_{\substack{m,n=1\\ m\neq n}}^{d}\kappa_m\kappa_n\,\mathbb{E}[\ell_m^2\ell_n^2] - \frac{K^2r^4}{4(d+2)^2} \\
&= \frac{[3K_{nn}+K_{mn}]r^4}{4(d+2)(d+4)} - \frac{[K_{nn}+K_{mn}]r^4}{4(d+2)^2} \\
&= \frac{(d+1)r^4}{2(d+2)^2(d+4)}\left[K^2 - \frac{d+2}{d+1}K_{mn}\right]. \qquad (4.2.8)
\end{aligned}$$

Finally, for ease of interpretation, we neglect the K_{mn} cross-term. The eigenvalues of each matrix have the following form:

$$\lambda_i\left(\frac{1}{N}LL^T\right) = \mathrm{Var}[\ell_i] = \begin{cases} \frac{r^2}{d+2}, & i = 1,\ldots,d \\ 0, & i = D \end{cases} \qquad (4.2.9)$$

$$\lambda_i\left(\frac{1}{N}CC^T\right) = \mathrm{Var}[c_i] = \begin{cases} 0, & i = 1,\ldots,d \\ \frac{(d+1)K^2r^4}{2(d+2)^2(d+4)}, & i = D \end{cases} \qquad (4.2.10)$$

$$\lambda_i\left(\frac{1}{N}EE^T\right) = \mathrm{Var}[e_i] = \sigma^2, \qquad i = 1,\ldots,D. \qquad (4.2.11)$$

The spectrum of $\frac{1}{N}XX^T$ may now be characterized in terms of (4.2.9)–(4.2.11) (see [76] for a review on the eigenvalues of sums of Hermitian matrices). Begin by noting that the matrices in the decomposition (4.2.7) are normal and are therefore diagonalizable by the spectral theorem. Next, we have that the matrices LL^T and CC^T commute, since by definition $L^TC = 0$ and thus $LL^TCC^T = L(L^TC)C^T = 0 = C(L^TC)^TL^T = CC^TLL^T$. Two diagonalizable matrices that commute must be simultaneously diagonalizable (see Theorem 1.3.12 of [38]). It follows that the (properly ordered) eigenvalues of $(LL^T + CC^T)$ are sums of the eigenvalues of LL^T and CC^T. Assume for now that $\min_i \lambda_i\left(\frac{1}{N}LL^T\right) > \max_i \lambda_i\left(\frac{1}{N}CC^T\right)$ (we will elaborate on the validity of this

assumption later). Then, due to the specific forms of L and C, we have

$$\lambda_i\left(\frac{1}{N}(LL^T+CC^T)\right)=\begin{cases}\lambda_i\left(\frac{1}{N}LL^T\right)=\frac{r^2}{d+2}, & i=1,\ldots,d\\ \lambda_i\left(\frac{1}{N}CC^T\right)=\frac{(d+1)K^2r^4}{2(d+2)^2(d+4)}, & i=D.\end{cases} \tag{4.2.12}$$

Finally, we consider the eigenvalues of $\frac{1}{N}(LL^T+CC^T+EE^T)$. Note that the matrices (LL^T+CC^T) and EE^T do not commute. However, the linearity of the matrix trace implies

$$\sum_{i=1}^{D}\lambda_i\left(\frac{1}{N}(LL^T+CC^T+EE^T)\right)=\sum_{i=1}^{D}\lambda_i\left(\frac{1}{N}(LL^T+CC^T)\right)+\sum_{i=1}^{D}\lambda_i\left(\frac{1}{N}EE^T\right). \tag{4.2.13}$$

When only the partial sum from $i=1$ to $d<D$ is needed, equality is replaced by an upper bound (the Ky Fan inequality [76]):

$$\sum_{i=1}^{d}\lambda_i\left(\frac{1}{N}(LL^T+CC^T+EE^T)\right)\leq\sum_{i=1}^{d}\lambda_i\left(\frac{1}{N}(LL^T+CC^T)\right)+\sum_{i=1}^{d}\lambda_i\left(\frac{1}{N}EE^T\right). \tag{4.2.14}$$

Having characterized the eigenvalues of the covariance matrix $\frac{1}{N}XX^T$, we may now proceed with a geometric analysis of the generic partitioning algorithm.

4.2.3 Partitioning and the Noise-Curvature Trade-off

Combining equations (4.2.9)–(4.2.14) and recalling the assumption that the eigenvalues of LL^T are larger than those of CC^T, the ratio of eigenvalues on the left-hand-side of criterion (4.1.1) takes the form

$$\frac{\sum_{i=1}^{d}\lambda_i\left(\frac{1}{N}(LL^T+CC^T+EE^T)\right)}{\sum_{i=1}^{D}\lambda_i\left(\frac{1}{N}(LL^T+CC^T+EE^T)\right)}\ \leq\ \frac{\frac{dr^2}{d+2}+d\sigma^2}{\frac{dr^2}{d+2}+\frac{(d+1)K^2r^4}{2(d+2)^2(d+4)}+D\sigma^2}. \tag{4.2.15}$$

Given the symmetry of the local model (4.2.2), assume that splitting the current partition $R^{(q)}$ creates two new partitions, $R^{(l)}$ and $R^{(r)}$, with $n^{(l)}=n^{(r)}=n^{(q)}/2$. Further assume that the radii of both $R^{(l)}$ and $R^{(r)}$ are exactly half the radius of $R^{(q)}$. The criterion (4.1.1) by which we decide

if $R^{(q)}$ should be split takes the form

$$\begin{aligned}
&\frac{\frac{dr^2}{d+2} + d\sigma^2}{\frac{dr^2}{d+2} + \frac{(d+1)K^2r^4}{2(d+2)^2(d+4)} + D\sigma^2} \\
&\qquad < \; C\left(\frac{1}{2}\,\frac{\frac{d(r/2)^2}{d+2} + d\sigma^2}{\frac{d(r/2)^2}{d+2} + \frac{(d+1)K^2(r/2)^4}{2(d+2)^2(d+4)} + D\sigma^2} + \frac{1}{2}\,\frac{\frac{d(r/2)^2}{d+2} + d\sigma^2}{\frac{d(r/2)^2}{d+2} + \frac{(d+1)K^2(r/2)^4}{2(d+2)^2(d+4)} + D\sigma^2}\right) \\
&\qquad = \; C\left(\frac{\frac{dr^2}{d+2} + 4d\sigma^2}{\frac{dr^2}{d+2} + \frac{(d+1)K^2r^4}{8(d+2)^2(d+4)} + 4D\sigma^2}\right). \qquad (4.2.16)
\end{aligned}$$

Define the constants

$$\alpha_1 = \frac{d}{d+2} = \mathcal{O}(1) \qquad (4.2.17)$$

$$\alpha_2 = \frac{d+1}{2(d+2)^2(d+4)} = \mathcal{O}\left(\frac{1}{2d^2}\right). \qquad (4.2.18)$$

Then the criterion becomes

$$\frac{\alpha_1 r^2 + d\sigma^2}{\alpha_1 r^2 + \alpha_2 K^2 r^4 + D\sigma^2} \;<\; C\left(\frac{\alpha_1 r^2 + 4d\sigma^2}{\alpha_1 r^2 + \frac{\alpha_2 K^2 r^4}{4} + 4D\sigma^2}\right). \qquad (4.2.19)$$

Before proceeding with a full analysis, we pause to build intuition for how geometry guides this criterion. Without loss of generality, momentarily set $r = 1$, $C = 1$, and consider only the leading order behavior of the constants α_1 and α_2. Doing so yields

$$\frac{1 + d\sigma^2}{1 + \frac{K^2}{2d^2} + D\sigma^2} \;<\; \frac{1 + 4d\sigma^2}{1 + \frac{K^2}{8d^2} + 4D\sigma^2}. \qquad (4.2.20)$$

Recall that the geometric uncertainty principle introduced in the previous chapter (equation (3.4.9) of Chapter 3) prevents noise (σ) and curvature (K) from simultaneously taking large values. Assuming that this principle is not violated, we may examine criterion (4.1.1) for the cases when: (1) curvature is large and noise is small enough to neglect; and (2) noise is large and curvature is small enough to neglect.

- For K large, σ small:

  $$\frac{1}{1 + \frac{K^2}{2d^2}} \;<\; \frac{1}{1 + \frac{K^2}{8d^2}},$$

  indicating that the current partition should be split in the presence of curvature, as smaller scales may be explored in the absence of noise.

- For K small, σ large:

$$\frac{1+d\sigma^2}{1+D\sigma^2} \;>\; \frac{1+4d\sigma^2}{1+4D\sigma^2}.$$

  indicating that the current partition should not be split as noise prevents exploring smaller scales. Further, in the absence of curvature it is not necessary to explore small scales.

Returning to (4.2.19), we now rigorously analyze the criterion. To ease the analysis, again set $C = 1$ and we will comment on practical choices for C later. Rearranging and simplifying equation (4.2.19) yields

$$\alpha_1\alpha_2 K^2 r^4 + 5\alpha_2 d K^2\sigma^2 r^2 \;>\; 4\alpha_1(D-d)\sigma^2. \tag{4.2.21}$$

Substituting leading order terms for α_1 and α_2, we have

$$\frac{1}{2}\frac{K^2}{d^2}r^4 + \frac{5}{2}\frac{K^2}{d}\sigma^2 r^2 \;>\; 4(D-d)\sigma^2. \tag{4.2.22}$$

Next, we recognize that K^2/d^2 is the square of the mean curvature $\overline{K} = \frac{1}{d}\sum_{n=1}^{d}\kappa_n$. Rewriting (4.2.22), we have

$$\frac{1}{2}\overline{K}^2 r^4 + \frac{5}{2}\overline{K}^2 d\sigma^2 r^2 \;>\; 4(D-d)\sigma^2, \tag{4.2.23}$$

which yields

$$\frac{1}{2}\overline{K}^2 r^4 \;>\; \frac{4(D-d)\sigma^2}{1+5\frac{d\sigma^2}{r^2}}. \tag{4.2.24}$$

We now give this expression a clear geometric interpretation. Recall that, by concentration of measure [59], $\mathcal{D}$-dimensional Gaussian noise (with each realization drawn from the $\mathcal{N}(0,\sigma^2 I_{\mathcal{D}})$ distribution) concentrates in a ball of radius $\sigma\sqrt{\mathcal{D}}$. Thus, with high probability, the component of the noise contaminating the d-dimensional tangent plane concentrates in a ball of radius $\sigma\sqrt{d}$. We must have that this radius be no larger than r, the radius within which we sample points in the tangent plane. Otherwise, if the perturbing noise ball were to have radius larger than r, there would be no hope of recovering geometric information. To ensure that the geometry remains intact, we require the ratio of these radii to scale with a bounded universal constant C_1

$$\frac{\sigma\sqrt{d}}{r} \propto C_1. \tag{4.2.25}$$

Then the denominator of (4.2.24) has the form $1 + 5C_1^2$ and the inequality can be rewritten as

$$\frac{1}{2}\overline{K}^2 r^4 \;>\; C_2(D-d)\sigma^2, \tag{4.2.26}$$

with constant $C_2 \propto 4/(1+5C_1^2)$.

The geometric interpretation is now clear. The left-hand-side of (4.2.26) scales with the mean curvature of the local model and matches the size of the norm $\|\frac{1}{N}CC^T\|_F^2$ to leading order. When $\overline{K}$ is large the partition should be split. The right-hand-side of (4.2.26) quantifies the size of the out-of-plane noise component, concentrated in a ball of radius $\sigma\sqrt{D-d}$. Note that noise confined to the tangent plane does not hinder its recovery (provided that (4.2.25) holds). The partition should not be split when the out-of-plane noise perturbation is large, as doing so would drive the algorithm to explore scales below the level of the noise. Estimation of local geometry is ill posed in such a case. Thus, the partitioning criterion embodies the noise-curvature trade-off. At large scales where the effect of curvature may be significant, the criterion is met and the data are partitioned. As smaller and smaller scales are explored, the effect of the noise perturbation overtakes the effect of curvature. In this case, the criterion is not satisfied and the data are no longer partitioned.

The criterion may be viewed as a geometric requirement to partition the data until a scale is reached at which the manifold is approximately linear (and thus suitable for PCA) but remains above the noise level, ensuring that its structure is discernible from noise. This criterion and its noise-curvature trade-off is therefore quite similar to the main results of the previous chapter. We numerically explore the similarity between these results in the next section. The assumptions and parameter selection for this analysis are explained before demonstrating numerical results.

4.3 Numerical Experiments

4.3.1 Implementation Details and Assumptions

The authors of [25] note that it can be beneficial to enforce a small number of initial partitionings to begin the algorithm, splitting partitions irrespective of any criteria. Geometrically, we understand that doing so is necessary to validate our assumption on the ordering of eigenvalues.

In the analysis of Section 4.2, we assume that the eigenvalues $\lambda_i\left(\frac{1}{N}LL^T\right)$ are larger in magnitude than the eigenvalues $\lambda_i\left(\frac{1}{N}CC^T\right)$. This assumption may not be valid at the global scale of a data set. However, we can find a local scale at which this assumption is valid. By forcing a fixed number of initial partitionings, the global scales at which curvature eigenvalues are large are avoided.

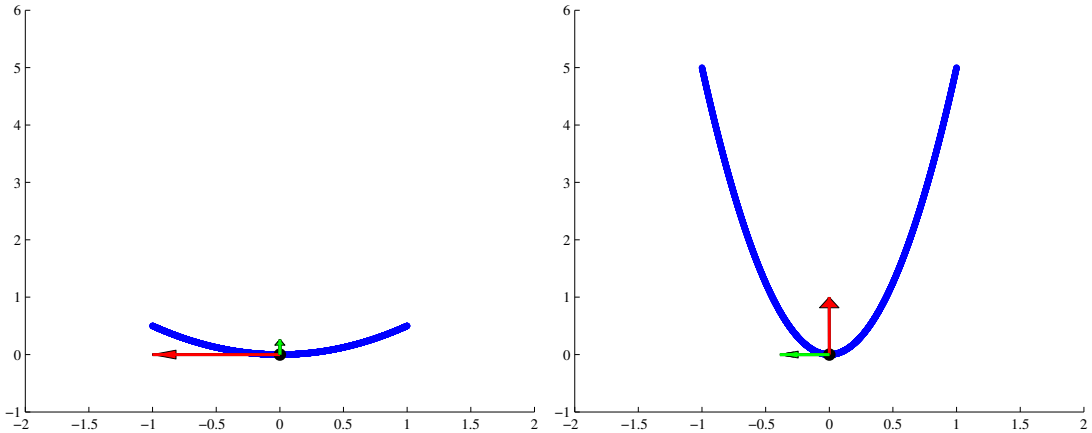
Figure 4.1 illustrates this point. The top row shows a 1-dimensional manifold $y = \frac{1}{2}\kappa x^2$ (blue) with the first (red) and second (green) eigenvectors scaled according to their corresponding eigenvalue. Panel (a) shows the low curvature setting ($\kappa = 1$) such that the first d eigenvalues correspond to the tangent plane. Panel (b) shows the high curvature ($\kappa = 10$) setting in which a curvature eigenvalue is included among the first d largest eigenvalues. To avoid the latter setting, a small number of initial partitionings are enforced. The bottom row shows the manifold in (b) after partitioning (colors indicate partitions). The results of enforcing one and two partitionings are shown respectively in (c) and (d). The assumption that the d largest eigenvalues correspond only to the tangent plane is valid in each partition in (d). In practice we observe that between one and three enforced partitionings are needed.

While enforcing a fixed number of partitions is necessary, doing so may lead to over-partitioning, where very flat (nearly linear) regions of the manifold are unnecessarily split. To avoid over-partitioning and maintain efficiency and geometric integrity, we post-process the data to remerge partitions that should not have been split. The intuition behind the post-processing step is that neighboring (adjacent) partitions whose linear PCA approximations are similarly oriented should not have been split. Therefore, the algorithm to remerge such partitions proceeds as follows:

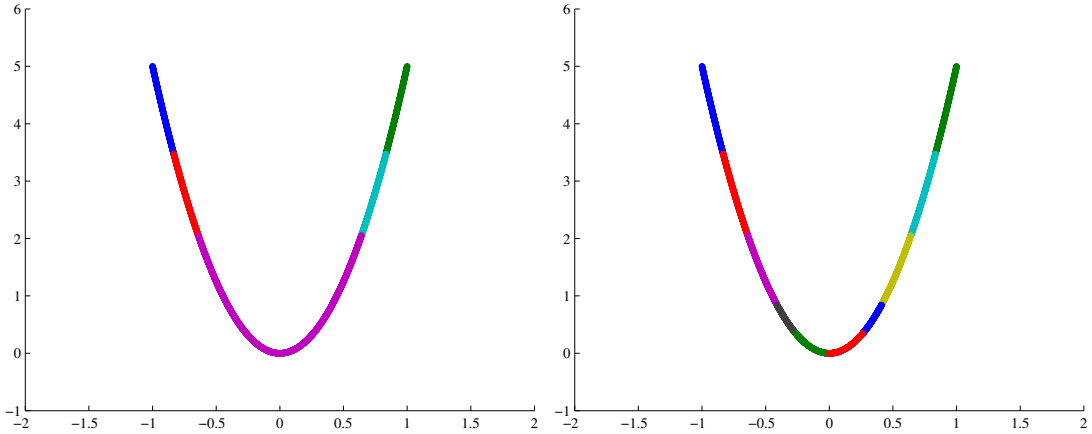
(1) Compute the center (mean) m_q of each partition $R^{(q)}$.

(2) For a given reference partition $R^{(j)}$:

   (a) Select all partitions $R^{(i)}$ with $\|m_j - m_i\|_2 < \tau_1$ as candidate partitions for merging. In practice, we set $\tau_1 = \frac{1}{4}\max_{x,y\in R^{(1)}}\|x - y\|_2$, where $R^{(1)}$ is the initial partition containing the entire data set.



(a) $\kappa = 1$ (b) $\kappa = 10$



(c) one enforced partitioning (d) two enforced partitionings

Figure 4.1: Top: a 1-dimensional manifold $y = \frac{1}{2}\kappa x^2$ (blue) with the first (red) and second (green) eigenvectors scaled according to the corresponding eigenvalue. Bottom: the manifold shown in (b) after partitioning (colors indicate partitions). See text for discussion.

(b) Amongst the candidate partitions, find those whose centers lie in the hyperplane spanned by the first d principal components of partition $R^{(j)}$. Let U_j be the matrix with the first d principal components of partition $R^{(j)}$ as its columns and similarly define U_i for all candidate partitions. Then the center m_i of candidate partition $R^{(i)}$ lies in the hyperplane approximation of $R^{(j)}$ if

$$\frac{1}{2}\left(\left\|\frac{U_j^T(m_j - m_i)}{\|m_j - m_i\|}\right\| + \left\|\frac{U_i^T(m_j - m_i)}{\|m_j - m_i\|}\right\|\right) \;>\; \tau_2. \tag{4.3.1}$$

In practice we set $\tau_2 = 0.98$.

(c) Amongst the candidate partitions, find those whose d-dimensional PCA hyperplane approximations are at an angle less than τ_3 to the d-dimensional PCA hyperplane approximation of $R^{(j)}$. In practice we set $\tau_3 = 2$ degrees in the noise-free case and $\tau_3 = 5$ degrees in the noisy case.

(d) Mark those candidate partitions that fall in the intersection of those selected in steps (b) and (c). These partitions should be merged with $R^{(j)}$.

(e) Repeat for the next reference partition $R^{(j)}$ until all partitions have been explored.

(3) Merge all similarly marked partitions.

We find that this algorithm successfully merges those partitions that should not have been split.

Criterion (4.1.1) requires a choice of the constant C. Note that by choosing $C < 1$, the criterion is biased towards not splitting the partition in question. We find it necessary to choose C slightly less than 1 to prevent gross over-partitioning in the noise-free case (where the error decreases monotonically with the radius).

4.3.2 Partitioning a Data Set

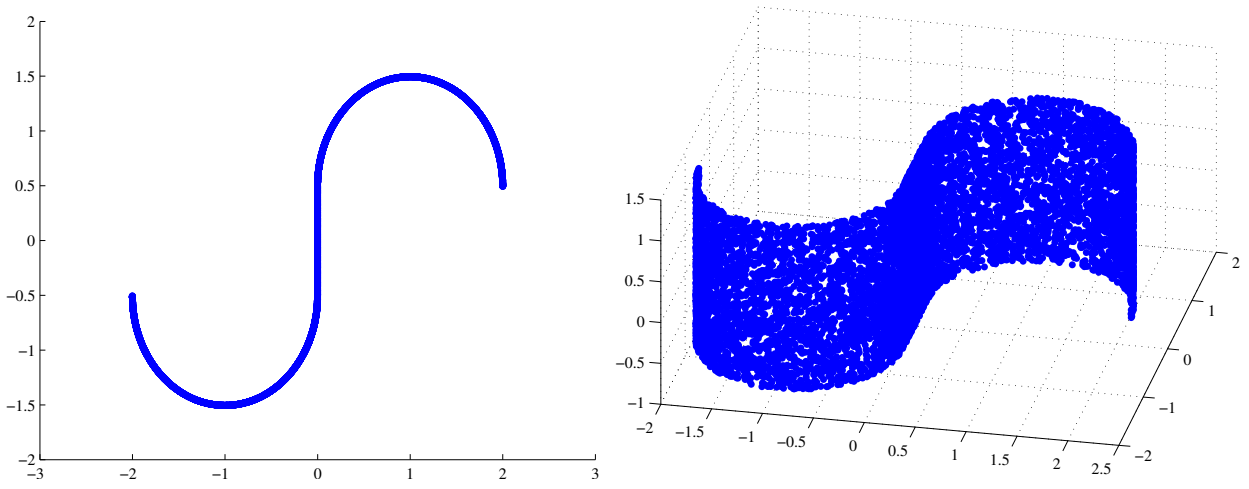
We now numerically explore the partitioning algorithm and make concrete connections to the estimation of geometry. In all experiments, we choose $C = 0.99$, and enforce 2 initial partitionings

when $d = 1$ and 3 initial partitionings when $d = 2$. We choose $(\tau_2, \tau_3) = (0.98, 2)$ for noise-free data and $(\tau_2, \tau_3) = (0.98, 5)$ for noisy data.

We generate the following data sets:

- $(d = 1,\ D = 2)$: Uniformly sample two half-circles of radius 1 and a line segment of length 1 (units are arbitrary). Connect the two half-circles with the line segment as shown in figure 4.2-a.
- $(d = 2,\ D = 3)$: Uniformly sample two half-cylinders of radius 1 and height 2 (units are arbitrary). Uniformly sample a plane with length 1 and height 2 (units are arbitrary). Connect the two half-cylinders with the plane as shown in figure 4.2-b.
- Add realizations of Gaussian noise drawn from the $\mathcal{N}\left(0, \sigma^2 I_D\right)$ distribution, with $\sigma = 0.025$ and $\sigma = 0.1$.

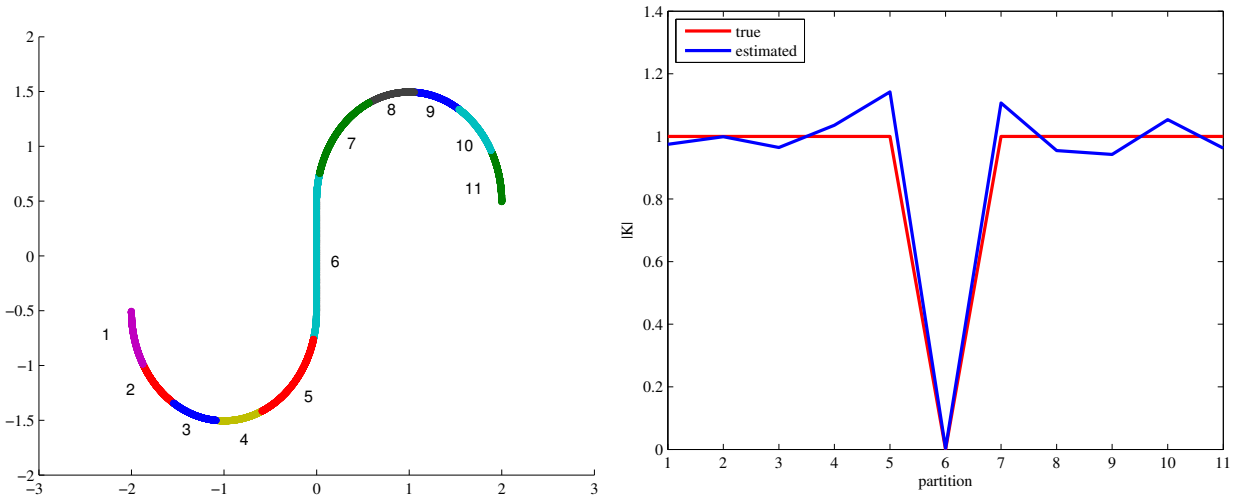
For the experiments in this section, a total of 10,000 points were sampled, with 2,500 points sampled from the line/plane and the remainder split evenly between the two half-circles/cylinders.



(a) $d = 1,\ D = 2$ (b) $d = 2,\ D = 3$

Figure 4.2: The two data sets used in this section.

First we demonstrate the performance of the partitioning algorithm on the noise-free $d = 1$ data set and show that the returned partitions are suitable for estimation of geometry. Figure 4.3-a shows the partitioned data set (colors indicate partitions; repeated but nonadjacent coloring indicates distinct partitions). The two circles have been split into partitions of roughly equal size, each at a scale small enough to appear approximately linear. Importantly, the line segment (partition 6) is not partitioned as it is clearly linear and should not be split. Using the analysis of Chapter 3.5.4, the curvature in each partition is estimated (blue) and compared to the known true value (red) in figure 4.3-b. The estimated curvature closely corresponds with the ground truth, indicating that the scale of the local geometric model (see Section 4.2.1) was found by the partitioning algorithm. While the accurate estimation of curvature is not a goal of the approximation-based criterion as originally posed, these results support our geometric interpretation and analysis of Section 4.2.



(a) partitioned data set

(b) curvature estimate

Figure 4.3: The partitioning of a noise-free data set yields a local scale at which curvature may be accurately estimated.

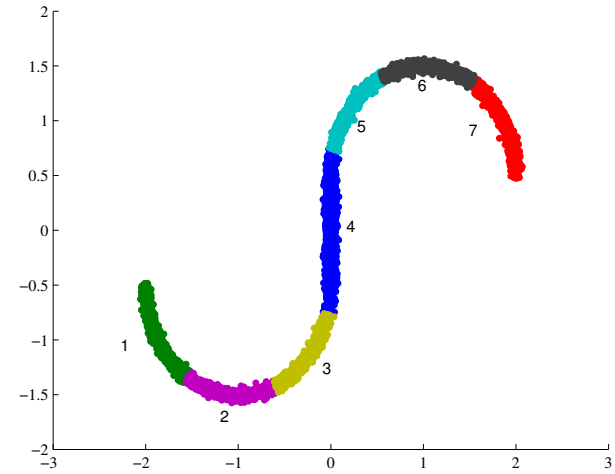
We now contaminate this data set with additive noise ($\sigma = 0.025$ and $\sigma = 0.1$) as indicated above. The performance of the partitioning algorithm in the presence of noise is shown in figure 4.4.

A partitioning similar to that observed in the noise-free setting occurs at low noise-levels (figure 4.4-a). However very few partitions are returned when the noise-level is high, as explained by our geometric analysis (4.2.26) of the partitioning criterion. Figure 4.4-b indicates that only 3 partitions are returned when $\sigma = 0.1$. Comparing this with the 7 returned partitions when $\sigma = 0.025$ and the 11 returned partitions when $\sigma = 0$, it is clear that the criterion (4.2.26) prevents partitioning at small scales in the presence of noise. The estimates of curvature in the presence of noise are shown in figure 4.4-c and 4.4-d. While the noise has corrupted the estimate as expected, the partitioning algorithm was still able to find local scales yielding reasonable estimates.

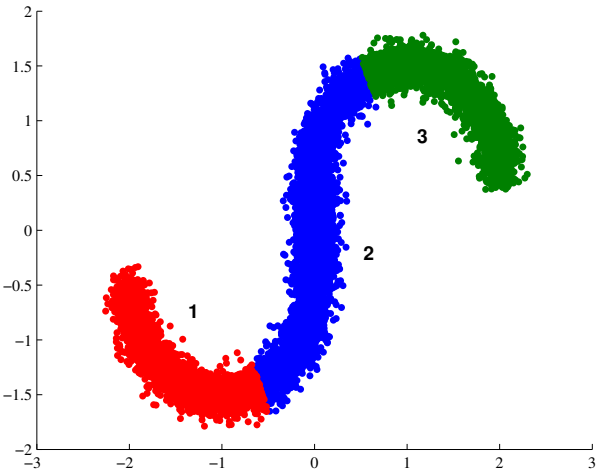
The performance of the partitioning algorithm is now compared with our tangent plane estimation results from Chapter 3. Our goal is to understand how the size of the returned partitions relates to the optimal scale for tangent plane recovery. Consider the $d = 2$ data set (see figure 4.2-b) with noise added as described above. Figures 4.5-a and 4.5-b show the partitioning for $\sigma = 0.025$ and $\sigma = 0.1$, respectively. Note that the nosier data yields fewer partitions. Figure 4.5-a shows that the segment of the data sampled from the (curvature-free) linear plane has not been split apart, while the curvature of the cylinders necessitates partitioning. Figure 4.5-b shows a partitioning in which the linear structure of the plane has only been approximately preserved due to noise.

Two example partitions are highlighted for closer analysis. Partition 1 corresponds to the plane and partition 2 corresponds to a segment of the cylinder. We examine the problem of estimating the tangent plane at the center of each example partition. The analysis in Chapter 3 bounds the angle between the true tangent plane and that computed from local PCA. The analytic bounds for the example partitions at both noise levels are shown in figures 4.5-c–4.5-f. The scale is quantified by the number of points used in the local PCA, and varies from $n = 1$ point up to the entire partition on the x-axis of each plot.

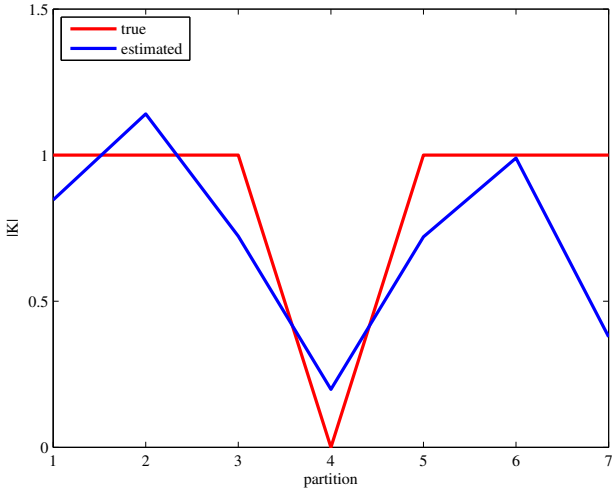
For both partitions in the $\sigma = 0.025$ data set, the bound is ill conditioned at very small scales due to the noise. Because partition 1 is curvature-free, the error (angle) decays as $1/\sqrt{n}$ and is then very close to zero at most scales (figure 4.5-c). Therefore, the entire partition may be used to accurately estimate the local tangent plane. Partition 2 is not curvature-free and we expect the



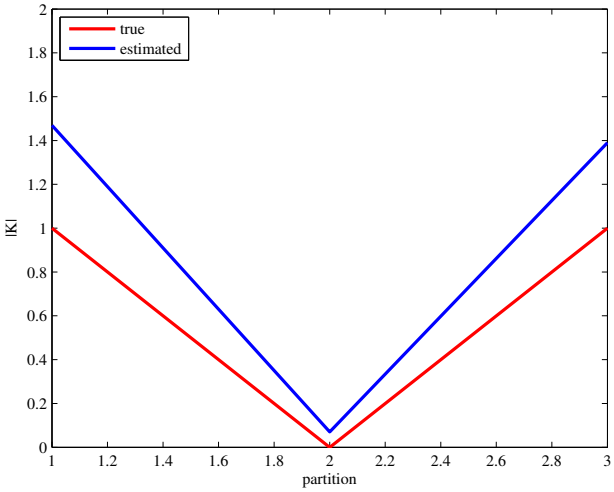
(a) $\sigma = 0.025$



(b) $\sigma = 0.1$

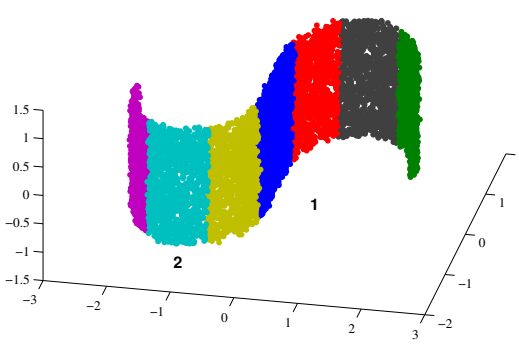


(c) curvature estimate, $\sigma = 0.025$

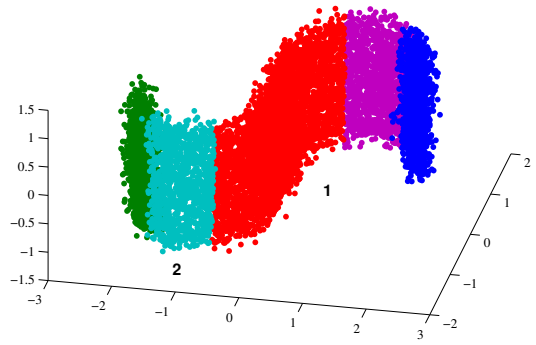


(d) curvature estimate, $\sigma = 0.1$

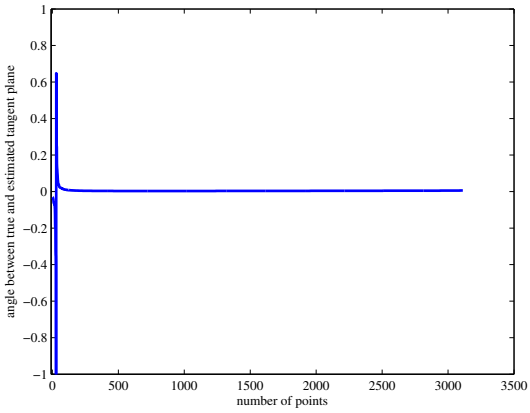
Figure 4.4: Partitioning in the presence of noise yields fewer partitions than in the noise-free case as scales below the noise-level cannot be explored. The partitioning algorithm is still able to find local scales yielding reasonable curvature estimates.



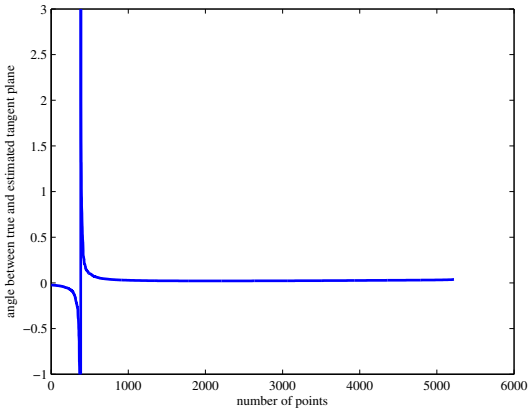
(a) $\sigma = 0.025$



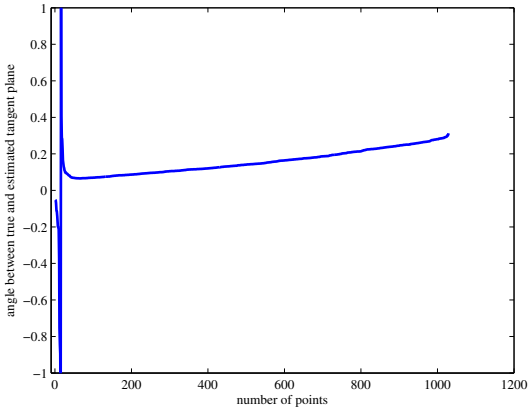
(b) $\sigma = 0.1$



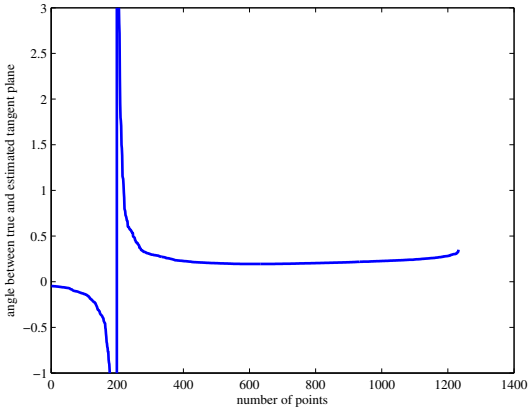
(c) partition 1, $\sigma = 0.025$



(d) partition 1, $\sigma = 0.1$



(e) partition 2, $\sigma = 0.025$



(f) partition 2, $\sigma = 0.1$

Figure 4.5: Tangent plane estimation is studied using the partitions labeled 1 and 2 above in the noisy data sets shown in (a) and (b). Panels (c)–(f) show the bound between the true tangent plane and that computed at various scales, based on the results of Chapter 3.

error to increase as large scales are explored. Figure 4.5-e indicates that the scale corresponding to roughly one-quarter of the size of the partition should be used to recover the best tangent plane estimate. However, the plot also indicates that using the entire partition does not produce a significantly worse estimate than that obtained from the optimal scale. In fact, the difference in error is quite small for large sampling density. Therefore, while the partitioning algorithm has not found the exact optimal scale for tangent plane estimation, this analysis indicates that the size of the partitions corresponds to an appropriate scale at which to accurately estimate geometry.

Similar results are shown for the $\sigma = 0.1$ data set in figures 4.5-d and 4.5-f. As there is significantly more noise in this example, the plots indicate that tangent plane estimation is ill conditioned at larger scales than in the previous example. Once a scale is reached that is large enough to be above the noise-level, the bound exhibits the same behavior as in the low-noise example. The tangent plane estimation error is very flat and close to zero for partition 1 as this partition is approximately linear. Similarly, the error grows at large scales for partition 2 due to curvature. However, note that in this noisy example, the error at the optimal scale is extremely similar to that at the scale of the entire partition. This example demonstrates that in the presence of noise, the partitioning algorithm finds a scale yielding accurate tangent plane estimation.

4.4 Discussion and Future Directions

The analysis of this chapter equips the partitioning algorithm with a geometric interpretation that mimics that noise-curvature trade-off essential to tangent plane estimation. The crucial observation to be gleaned from the numerical experiments in Section 4.3 is that a scale is never reached at which tangent plane estimation is ill conditioned due to curvature. The partitioning algorithm successfully finds local scales such that curvature does not corrupt a tangent plane estimate, even at the scale of the full partition in the presence of noise. Therefore, while our careful analysis in Chapter 3 provides a tool to find the optimal scale for tangent plane recovery, the approximation-based partitioning algorithm analyzed in this chapter yields a practical method to find a near-optimal scale.

Given our analysis, the partitioning scheme may be viewed as a coarse algorithm for breaking apart a large data set in a geometrically meaningful way. Much development is necessary to realize a robust data analysis algorithm. One particular area for further research is the choice of C, which should depend on radius, dimension, and curvature. A more thorough investigation into $C = C(r, d, K)$ is needed to handle data sets with high curvature. Another key future contribution lies in injecting our tangent plane estimation analysis into the partitioning criterion, rather than resorting to an *a posteriori* comparison. We aim to expand and combine these analyses to understand how a more sophisticated and robust criterion can be developed. Such an analysis can be used for proper scale selection in existing local algorithms that typically define locality via a fixed number of neighboring points (see the discussion in Chapter 1.2). Further, our analysis provides an error bound on the local scale, rather than a global minimization of a cost function. Through our analysis of the stability of the local tangent plane estimate, low-distortion guarantees remain tight over an entire data set. Finally, stable recovery of geometry in an algorithmically efficient manner should be particularly useful in applications where the tangent plane has a physical interpretation, as is the often the case in the analysis and numerical simulation of dynamical systems.

Chapter 5

Noise Corruption of Empirical Mode Decomposition and its Effect on Instantaneous Frequency

5.1 Introduction

Empirical Mode Decomposition (EMD) is an analysis tool for nonstationary data introduced by Huang *et al.* [40] in 1998. Nonstationary signals have statistical properties that vary as a function of time and should be analyzed differently than stationary data. Rather than assuming that a signal is a linear combination of predetermined basis functions, the data are instead thought of as a superposition of fast oscillations onto slow oscillations [30]. EMD identifies those oscillations that are intrinsically present in the signal and produces a decomposition using these modes as the expansion basis. We note that throughout this chapter the term "basis" is used in the same sense as used by [40]: the modes of a signal's decomposition do not span a particular space, but provide an expansion for the specific signal. In this way, the basis is data driven and adaptively defined each time a decomposition is performed [30]. EMD has been used for data analysis in a variety of applications including engineering, biomedical, financial and geophysical sciences [39].

In contrast with Fourier analysis, EMD requires no assumptions on its input and is therefore well suited to analyze nonstationary data. Since nonstationarity implies that a signal is not well represented by pure tones, a significant number of harmonics is required to represent a nonstationary signal in the Fourier basis. Energy must be spread across many modes to accommodate deviations from a pure tone. In producing an adaptive decomposition consisting of modes that allow for such deviations, EMD efficiently represents the signal by relaxing the need to explore all frequencies. A

signal is expanded using only a small number of adaptively defined modes.

As EMD is an algorithm and not yet a theoretical tool, its limits must be tested experimentally. Several authors have reported on its performance in the presence of noise [39, 30, 51]. In this chapter, we propose a new understanding of the mechanism that prevents the algorithm from properly estimating the instantaneous frequency of a noisy signal. The chapter is organized as follows. Section 5.2 gives a brief description of the EMD algorithm and demonstrates its use in estimating the instantaneous frequency of a clean signal. The same estimation, performed in the presence of noise, is seen to be problematic in Section 5.3 and the cause is identified. Section 5.4 outlines a new explanation for this poor performance. Finally synthetic seismic data are used in Section 5.5 to extend our study from simple signals to a model for real world data.

5.2 Empirical Mode Decomposition

5.2.1 Algorithm

The goal of Empirical Mode Decomposition is to represent a signal as an expansion of adaptively defined basis functions with well defined frequency localization. Each basis function, called an Intrinsic Mode Function (IMF), should be physically meaningful, representing ideally one frequency (nearly monochromatic). To accomplish this, an IMF is defined as a function for which (1) at any point, the mean of the envelopes defined by local maxima and minima is zero, and (2) the number of extrema and the number of zero crossings differ by at most one [40]. Such a definition attempts to ensure that a meaningful instantaneous frequency can be obtained from each IMF, a process that is defined and detailed in the next subsection, but does not guarantee that each IMF is narrow band [40]. To decompose a signal $x(t)$, the EMD algorithm works as follows [30]:

(1) Interpolate (usually with cubic splines) the local maxima of $x(t)$ to form an envelope. Repeat for the minima.

(2) Compute the mean, $m(t)$, of the two envelopes.

(3) Compute the detail, $d(t)$, by subtracting the mean from the signal, $d(t) = x(t) - m(t)$. Extract the detail as an IMF.

(4) Repeat the iteration on the residual $m(t)$. Continue until the residual is such that no IMF can be extracted and represents the trend.

While the trend does not meet the definition of an IMF, we will refer to it as the final IMF for convenience. Before the detail, $d(t)$, can be considered an IMF, a "sifting" process takes place during which the detail is treated as a new signal and is iterated until a predefined stopping criterion is reached. The purpose of this step is to enforce the definition of an IMF [30]. Ideally, all modes are now nearly monochromatic and can be used to give a meaningful estimate of the signal's instantaneous frequency.

The algorithm can be described in the time-frequency domain as a collection of data-dependent projections. Olhede and Walden [66] formalize this idea by defining projection operators P_{R_j}, not necessarily orthogonal, that project a signal $x(t)$ into regions R_j of the time-frequency plane. The signal may then be written as

$$x(t) = \sum_{j=1}^{K} [P_{R_j} x](t),$$

where K is the number of IMFs produced, with the Kth IMF being the trend. Since each projection gives rise to an IMF, an expansion of the signal is then given by

$$x(t) = \sum_{j=1}^{K} X_j(t),$$

where X_j is the jth IMF.

5.2.2 Estimation of instantaneous frequency

A signal is often characterized in terms of its frequency content. When a signal's statistical properties are shift-invariant in time, it is said to be stationary. As this definition implies, frequency remains constant throughout the signal's duration, and is easily defined as the number of periods per unit time. However, if the signal's frequency varies with time, it is said to be nonstationary,

and this global definition of frequency loses meaning. It is therefore necessary to characterize the frequency content of the signal in a local manner. For example, a chirp with a quadratic phase has frequency that changes linearly from one instant to the next. It is not possible to pinpoint one frequency for the entire chirp. Instead the chirp's frequency is described as a (linear) function of time. It is therefore more useful to characterize such a signal in terms of its instantaneous frequency.

Boashash [9] describes instantaneous frequency (IF) as "a time-varying parameter which defines the location of the signal's spectral peak as it varies with time." He points to seismic, radar, sonar, communication, and biomedical applications as fields where IF is utilized. Two conditions are needed to produce a physically meaningful and well defined instantaneous frequency. The signal must be analytic and it must be narrow band. An analytic signal is produced via the Hilbert transform:

$$[\mathcal{H}x](t) = \frac{1}{\pi} PV \int_{-\infty}^{\infty} \frac{x(t')}{t - t'} dt',$$

where PV denotes the Cauchy principle value. Given a real valued signal, $x(t)$, its analytic representation is then defined as $z(t) = x(t) + i[\mathcal{H}x](t)$. The analytic signal $z(t)$ may be written in the form

$$z(t) = a(t)e^{i\phi(t)},$$

and the instantaneous frequency, $v(t)$, can then be defined [9] in terms of the derivative of the phase $\phi(t)$:

$$v(t) = \frac{1}{2\pi} \frac{d\phi}{dt}.$$

The derivative must be well defined since physically there can only be one instantaneous frequency value $v(t)$ at a given time t. This is ensured by the narrow band condition: the signal must contain nearly one frequency. Further, as detailed by Boashash [9], the Hilbert transform produces a more physically meaningful result the closer its input signal is to being narrow band. However, we wish to work with signals that are much more interesting than those that are monochromatic. This can be achieved by decomposing such a signal into several nearly monochromatic components, each of which provides a well defined, meaningful instantaneous frequency. An overall IF estimate

of a signal x, given its decomposition into K IMFs, is then calculated as a weighted sum of the individual IFs:

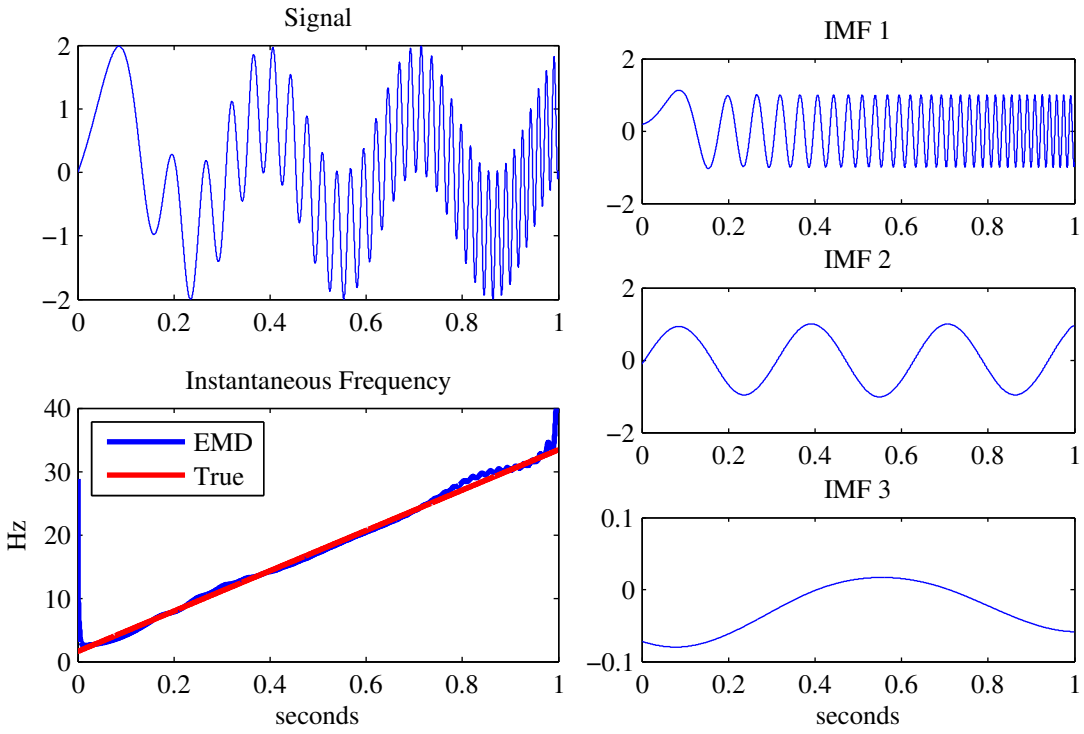
$$IF\big(x(t)\big) = \frac{\sum_{j=1}^{K} A_j^2(t) v_j(t)}{\sum_{j=1}^{K} A_j^2(t)},$$

where $A_j(t)$ and $v_j(t)$ are, respectively, the magnitude and instantaneous frequency of the analytic representation of IMF X_j [66].

To demonstrate the calculation of IF, consider $x(t) = \sin(200t^2) + \sin(20t)$, the superposition of a linear chirp onto a stationary sine wave, on the interval $t \in [0, 1]$ seconds. Figure 5.1a shows the true analytic[1] IF (in red) and the overall IF estimate (in blue) obtained from the IMFs (shown in figure 5.1b) of $x(t)$. We are able to calculate a physically meaningful instantaneous frequency when using the decomposition of a signal in the absence of noise.

Huang *et al.* [41] give a detailed discussion on the shortcomings of this method of IF calculation. In particular, they note that the analytic signal obtained from the Hilbert transform is only physically meaningful if the conditions of the Bedrosian theorem are met. They introduce a normalization scheme that empirically separates the AM and FM components of each IMF, where the AM carries the envelope and the FM is the constant amplitude variation in frequency. The "normalized" FM component of an IMF is guaranteed to satisfy the Bedrosian theorem and is therefore suitable for the Hilbert transform. This process is referred to as the "normalized Hilbert transform." Alternatively, once an IMF has been normalized, Huang *et al.* [41] propose eschewing any Hilbert transform in favor of applying a 90 degree phase shift by means of a direct quadrature. Both methods are demonstrated to be more accurate on clean signals than the standard method presented above. Since the focus of our work is the performance of EMD in the presence of noise, the performance of this normalization scheme on noisy data will be addressed in the next section.

[1] The analytic IF of the superposition of two signals, $x(t) = A_1(t)e^{i\phi_1(t)} + A_2(t)e^{i\phi_2(t)}$, is defined as the average of the individual IFs of each signal only when $|A_1(t)| = |A_2(t)|$ [56]. We note that this condition holds for this example, and we compute the analytic IF accordingly. An example for which the condition does not hold will be encountered in Section 5.5.



(a) IF estimate

(b) IMFs

Figure 5.1: The instantaneous frequency estimate and IMFs of a clean signal.

5.3 Performance in the Presence of Noise

A clean signal can produce a decomposition that lends itself to a meaningful instantaneous frequency estimate. However, as is the case in many applications, data are often contaminated by noise. Decomposing a noisy signal produces both narrow and wide band IMFs. While most of the wide band IMFs contain noise and may be discarded, a small number capture the transition from noise to signal and must be kept. This leads to a corrupted estimate of the instantaneous frequency.

5.3.1 Evidence of a problem

In the previous section the calculation of instantaneous frequency was described. This process is now applied to the same signal contaminated with additive white Gaussian noise such that its SNR is 27dB. Throughout this chapter we use $\text{SNR} = 10\log_{10}\left(\frac{\|x\|_2}{\sigma}\right)\text{dB}$, where σ is the standard deviation of the noise. The result is shown in figure 5.2 and it is clear that a meaningful instantaneous

frequency estimate was not produced.

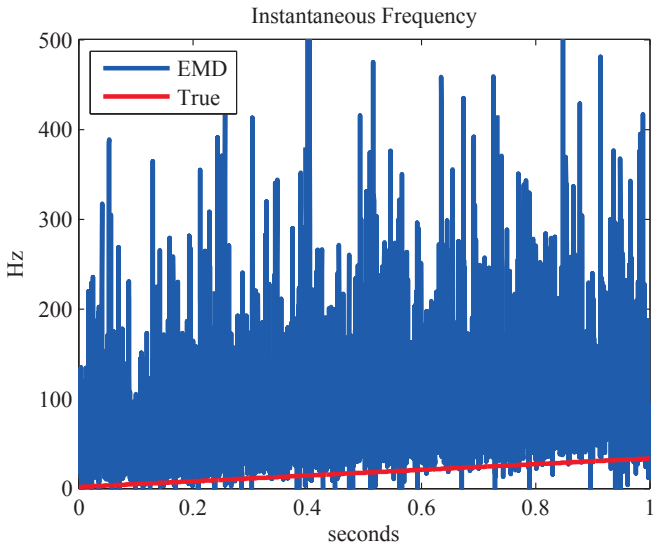


Figure 5.2: The corrupted instantaneous frequency estimate of a noisy signal.

To understand this poor result, recall that a signal's IF is computed as a weighted sum of the IF from each of its IMFs. The analytic representation of each IMF is required and thus each IMF must be narrow band to ensure a meaningful Hilbert transform. Moreover, IF is well defined only in the case of a nearly monochromatic signal. Therefore, for the purpose of computing a meaningful IF, the key feature of the decomposition is that each IMF contains nearly one frequency.

It is important to recall that the definition of an IMF does not guarantee monochromaticity. This is illustrated with a deterministic example. The decomposition of a signal composed of a slow sinusoid with high frequency sinusoids superimposed at each crest and trough is shown in figure 5.3. Despite the fact that this signal was constructed in a completely deterministic manner, its first two IMFs contain both high and low frequencies. Such IMFs are not suitable for the Hilbert transform and will not yield a well defined IF. Wu and Huang [80] use a very similar example, developed independently from our example, to note that a decomposition may give rise to IMFs containing oscillations of drastically different scales. They refer to the creation of such IMFs as

"mode mixing," and introduce the Ensemble EMD (EEMD) to alleviate this issue. We will discuss the performance of EEMD on noisy data in Section 5.4.

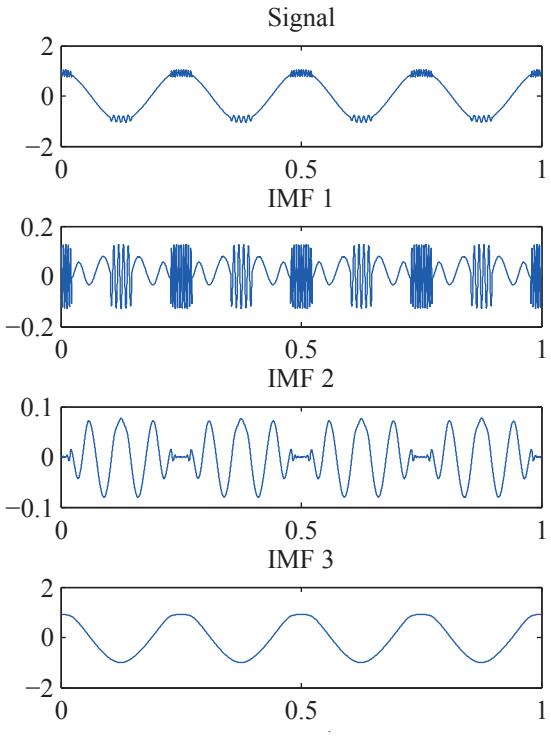


Figure 5.3: IMFs of a deterministic signal. IMFs 1 and 2 contain both high and low frequencies, illustrating that monochromaticity is not guaranteed.

5.3.2 Identifying the culprit

The poor quality IF estimate from a noisy signal can be explained by the creation of wide band IMFs. More precisely, the EMD decomposition of a noisy signal will generate some "noisy" IMFs. As explained below, such noisy IMFs are neither monochromatic signals nor pure noise; rather their Fourier transform is localized over a well defined frequency range. Consequently, such IMFs cannot contribute a well defined IF because noise is wide band by definition. Figure 5.4 shows the decomposition of the noisy example signal. We identify three categories of IMFs:

(1) **Noisy**: IMFs 1-4 are wide band as they clearly contain noise.

(2) **Transition**: IMFs 5-7 contain both signal and noise. These IMFs capture the "transition" from the noise captured in IMFs 1-4 and the monochromatic components extracted as IMFs 8-11.

(3) **Monochromatic**: IMFs 8-11 are nearly monochromatic and yield meaningful IF contributions.

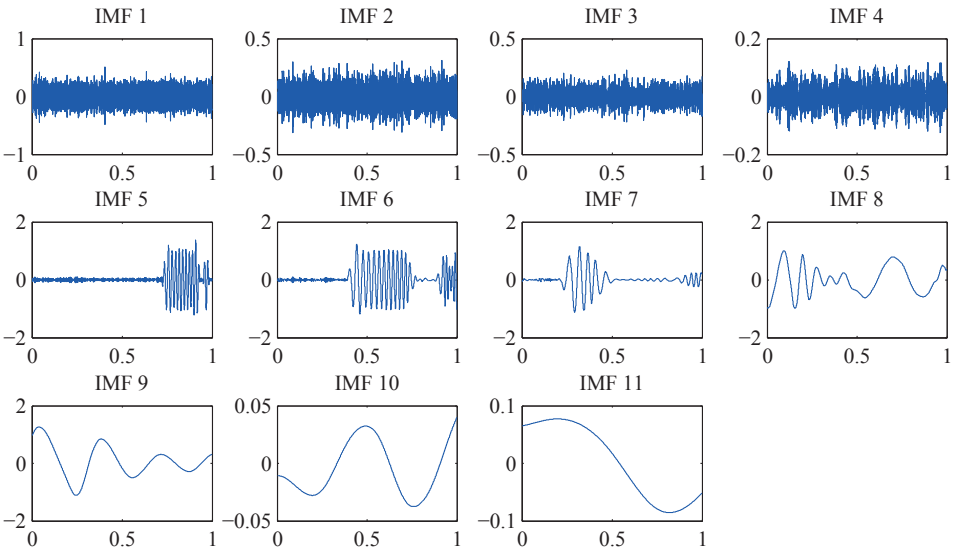
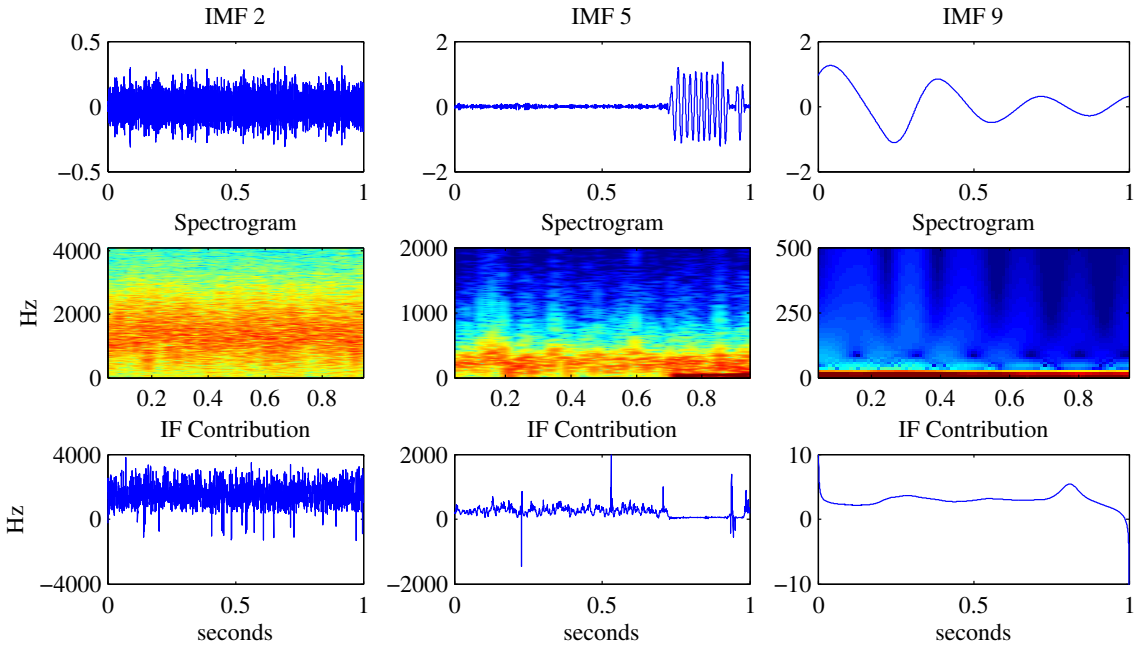


Figure 5.4: IMFs of a noisy signal. IMFs 1-4 capture most of the noise, while IMFs 5-7 represent the transition from noise to signal, and IMFs 8-11 are nearly monochromatic.

To demonstrate the effect of each type of IMF on the overall IF estimate, figure 5.5 highlights an example from each category. IMF 2 (left) is a noisy IMF; IMF 5 (center) contains both signal and noise and is a transition IMF; IMF 9 (right) is nearly monochromatic. Spectrograms[2] are used to illustrate the frequency content that characterizes each IMF. The spectrogram of the noisy mode, IMF 2, shows that it is wide band and therefore yields an IF that is not physically meaningful. In contrast, the nearly monochromatic IMF 9 is seen to be narrow band and contributes a well

[2] Spectrograms are displayed as a log-scale color representation of the power spectral density calculated using a Kaiser window of duration 0.1 seconds with 90% overlap. Red and blue correspond to higher and lower density, respectively, and the scale is uniform within a figure but not necessarily throughout the chapter.

defined IF. Finally, despite its signal content, transition IMF 5 is wide band and cannot contribute a clean IF. The inclusion of IF contributions from wide band IMFs pollutes the overall IF and is responsible for the poor result seen in figure 5.2.



(a) IMF 2 - Noisy (b) IMF 5 - Transition (c) IMF 9 - Monochromatic

Figure 5.5: Characteristic IMFs representing (a) noise, (b) transition from noise to signal, and (c) monochromatic components extracted from a noisy signal.

Since the inclusion of certain IMFs results in a poor IF estimate, it is reasonable that some nonlinear thresholding process would yield better results. Specifically, discarding the IMFs identified as noise will provide a more meaningful IF estimate. In figure 5.6 the IF of the signal shown in figure 5.1a(top) is now computed using only IMFs 5-11. It is important to note that IMF 5 is not discarded because as a transition IMF, it contains both signal and noise. We would like to ignore such an IMF since it will provide poor IF information derived partially from noise, but cannot discard its signal content. Therefore, it must be included and contaminates our overall estimate. The same is true of IMFs 6 and 7. Other thresholding methods may be utilized, including using only those IMFs with energy between specified thresholds [39]. However, to our knowledge, there

is not a clear cut method of thresholding that will produce a faithful IF estimation. While the thresholded estimate in figure 5.6 is an improvement over the previous estimate shown in figure 5.2, the transition IMFs' contribution has left the IF mostly incoherent. The necessary inclusion of transition IMFs is therefore identified as the main problem in estimating the IF in the presence of noise.

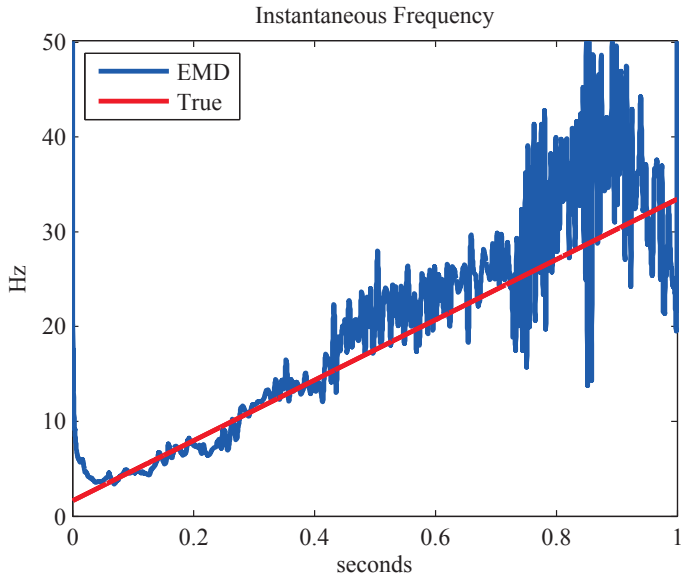
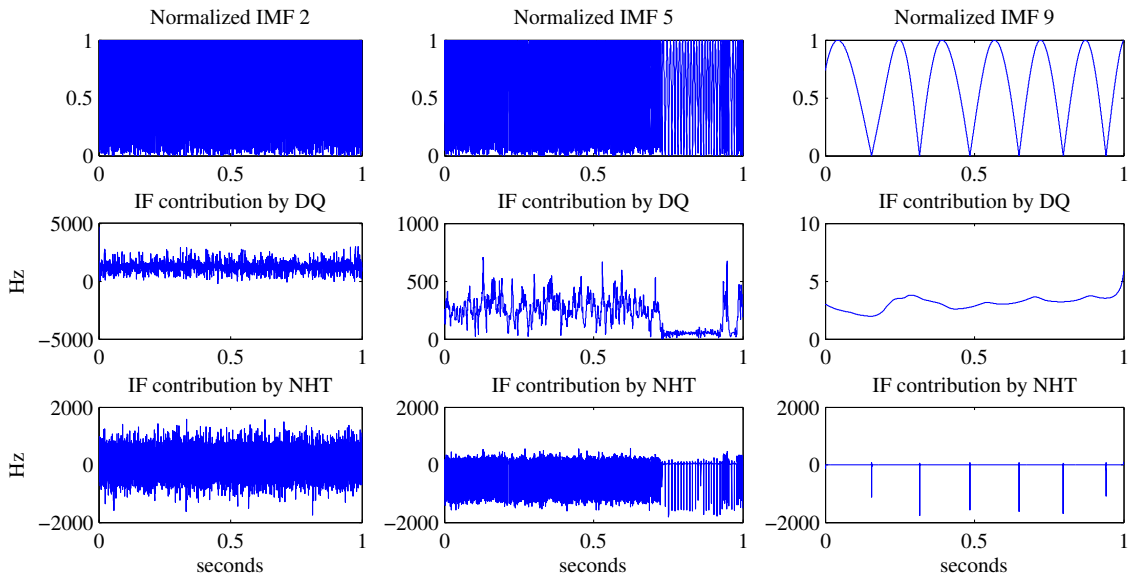


Figure 5.6: Instantaneous frequency estimate using IMFs 5-11. The necessary inclusion of transition IMFs prevents a clean estimation.

It is also reasonable that computing the IF from normalized IMFs [41] (see Section 5.2) might yield cleaner results. However, Huang *et al.* [41] note that the normalized scheme encounters problems when an IMF contains noise and recommend computing the analytic signal with the standard Hilbert transform approach. Figure 5.7 shows the normalized version of the example IMFs from figure 5.5. We observe that we still have (from left to right) a noisy IMF, a transition IMF, and a monochromatic IMF. The IF contribution from each normalized IMF is shown, calculated by direct quadrature (middle) and normalized Hilbert transform (bottom). Just as in the standard unnormalized case, transition IMFs with corrupted IF contributions still exist and their necessary

inclusion will prevent a clean IF estimate (not shown).



(a) IMF 2 - Noisy (b) IMF 5 - Transition (c) IMF 9 - Monochromatic

Figure 5.7: Normalized IMFs of a noisy signal (top), IF contribution from direct quadrature (middle), and IF contribution from normalized Hilbert transform (bottom).

5.4 Analysis of Noisy Decompositions

With an understanding of how transition IMFs pollute the estimation of IF, we address the more fundamental question of why transition IMFs are produced when EMD operates on a noisy signal. To begin, we note the work of Flandrin and Goncalves [30] showing that EMD acts as a filter bank when decomposing pure noise, and add our observation that the boundaries of the frequency bands vary with time. We propose two mechanisms that lead to the creation of transition IMFs:

(1) **Spectral leak** between frequency bands: frequency content of the underlying signal falls within a band treated as noise.

(2) **Phase alignment**: the alignment of the signal with the lowest level of noise present in the band is controlled by the signal's phase.

Spectral leak is mostly a nonstationary condition while the contribution of phase alignment is best seen in the stationary setting.

5.4.1 EMD decomposition of pure white noise

Before returning to the decomposition of a noisy signal, EMD's performance on pure noise is analyzed. Figure 5.8 shows the spectrogram of a realization of white Gaussian noise (zero mean, standard deviation of 0.2). It is not surprising that the spectrogram shows nearly uniform power spectral density since, in principle, the density of such noise should be constant. This specific noise realization will be used in all experiments that follow in this section.

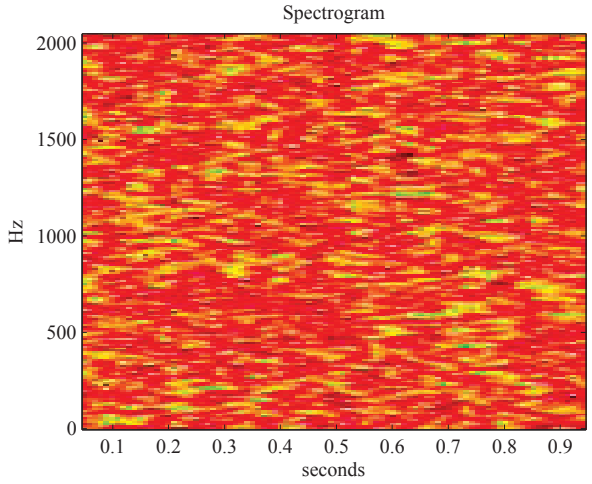


Figure 5.8: Spectrogram of white Gaussian noise used throughout this section.

Flandrin and Goncalves [30] reported that EMD acts as a filter bank when decomposing pure Gaussian noise. By selecting entire frequency bands as IMFs rather than a single frequency, the IMFs are by definition multicomponent. We observe a similar result and note that the boundaries of each band are not straight line cuts through the frequency axis, but instead vary as a function of time. This is clearly seen in the IMFs of the noise as their spectrograms (figure 5.9) show that the borders of the frequency-bands do not resemble straight lines. The spectrograms also reveal that the IMFs provide a nearly dyadic decomposition of the spectrum shown in figure 5.8. Since

the noise is composed of realizations of random variables, we define its mean power spectral density $M_{psd}(t)$ and associated standard deviation $SD_{psd}(t)$ at a given time t as follows:

$$M_{psd}(t) = \sum_{k=0}^{F_s/2} k \cdot P(k,t)$$
$$M^2_{psd}(t) = \sum_{k=0}^{Fs/2} k^2 \cdot P(k,t)$$
$$SD_{psd}(t) = \sqrt{M^2_{psd}(t) - \big(M_{psd}(t)\big)^2}$$

where F_s is the sampling rate and $P(k,t)$ is the normalized power spectral density at frequency k and time t. The plots of the mean power spectral density with error bars representing one standard deviation show that the statistics of the IMFs vary with time (figure 5.10). Some frequency mixing between modes is also observed.

5.4.2 EMD decomposition of a signal corrupted by noise

5.4.2.1 Spectral leak

Kijewski-Correa and Kareem [51] attributed the poor quality of IF estimation in the presence of noise to the empirical nature of the algorithm, leading to a basis derived from the noise. They observed the mixing of the input signal over many IMFs, making it difficult to isolate the clean signal from the noise. We extend this explanation with our observations to explain the extraction of transition IMFs. The process is best understood by considering the noisy signal in the time-frequency plane. The algorithm is operating on projections in this plane, starting with the highest frequency band and adaptively moving down the frequency axis. These projections are not completely orthogonal, and thus there is some frequency mixing in the modes. As EMD tiles down the time-frequency plane, it first extracts pure noise as it has not yet reached the frequency of the signal. While in the pure noise region, EMD behaves as a filter bank, as observed by Flandrin and Goncalves, extracting noise in an almost dyadic manner.

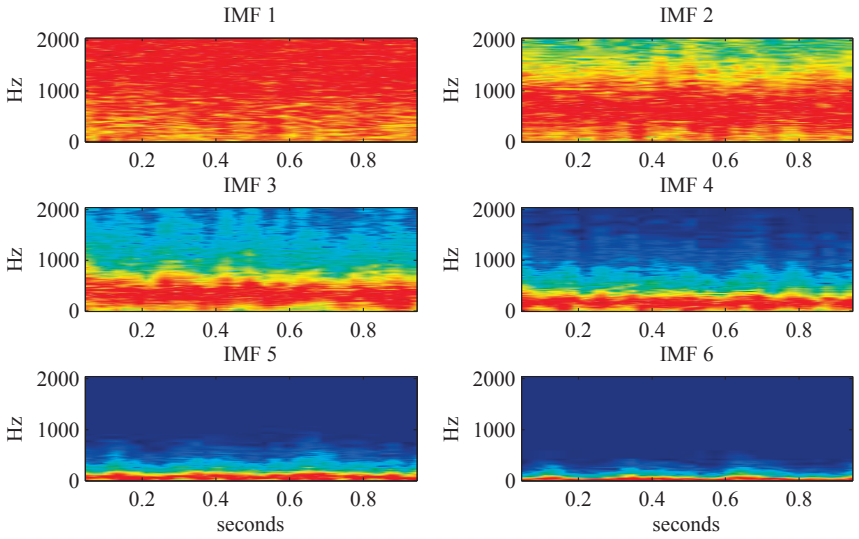


Figure 5.9: Spectrogram of first six IMFs of white Gaussian noise, highlighting EMD's filter bank behavior.

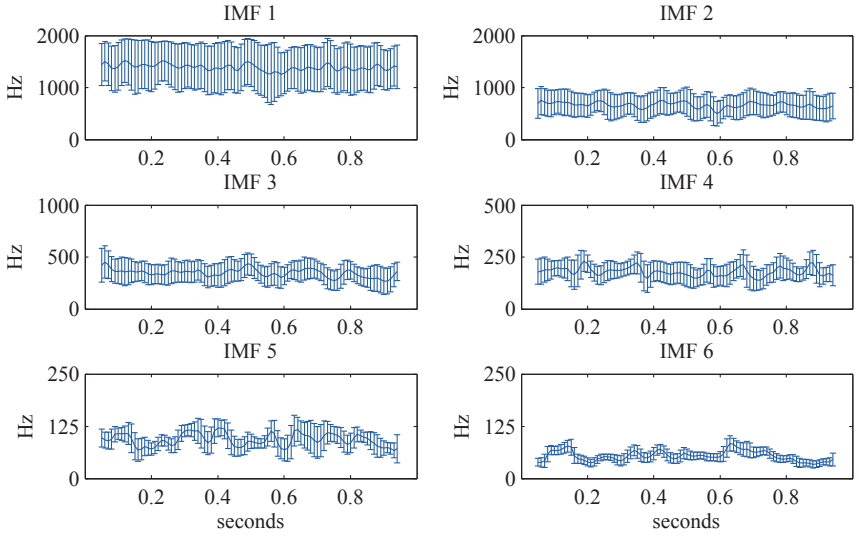


Figure 5.10: Mean (with error bars representing one standard deviation) power spectral density of IMFs extracted from white Gaussian noise. Note the different scales on the frequency axis, clearly indicating an almost dyadic decomposition of the noise spectrum.

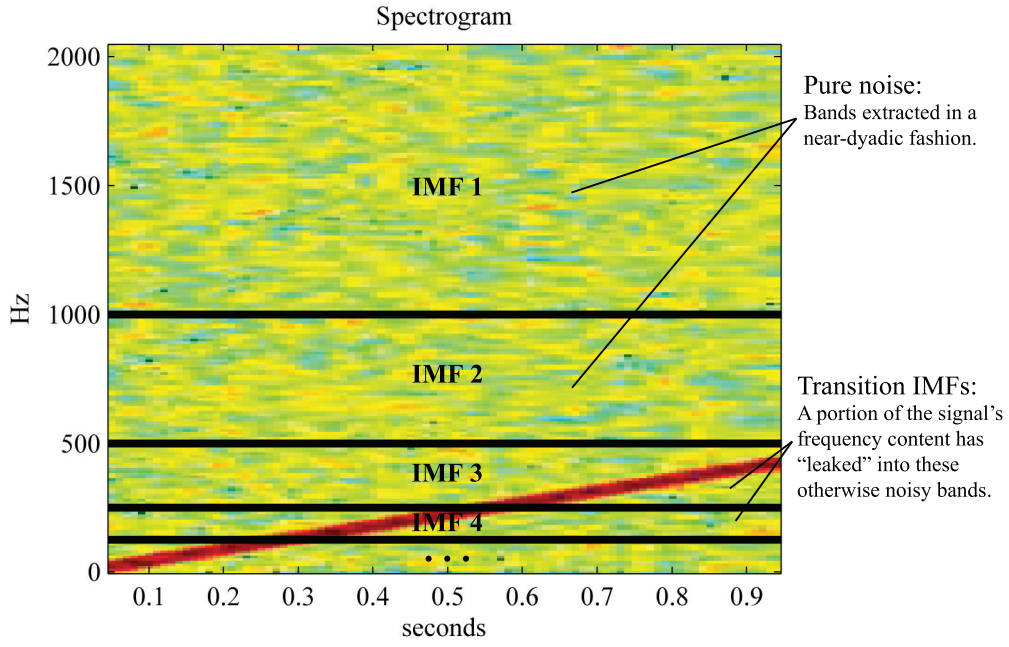


Figure 5.11: A model of EMD's filter bank action shown in the time-frequency plane. Pieces of chirping signal are captured in noisy bands. The bands contributing to IMFs 1-4 are illustrated and the boundaries between the bands are idealized.

A model for this process in the time-frequency plane is provided in figure 5.11. The model shows a spectrogram of the noisy chirp $\sin(2\pi f t^2) + n(t)$, where $f = 225$ Hz, $t \in [0, 1]$, and $n(t)$ is the exact same realization of noise shown in figure 5.8. The boundaries between the bands are idealized, highlighting EMD's filter bank behavior. Noise is removed until a frequency present in the signal matches or exceeds that of the noise. The model demonstrates the situation where a portion of a nonstationary signal leaks into an otherwise noisy band (IMF 3 in this example). In this case, the signal's frequency is high enough to be included in the IMF for only part of its duration. Still behaving in the noise regime, EMD extracts both signal and noise as it cannot distinguish which should be removed. Because of the variation in the boundaries of the identified frequency bands (seen in figures 5.9 and 5.10, not shown in the model), EMD will encounter such a band even when decomposing a stationary signal. This is the general process that leads to the creation of a transition IMF, and will be seen explicitly in the following example.

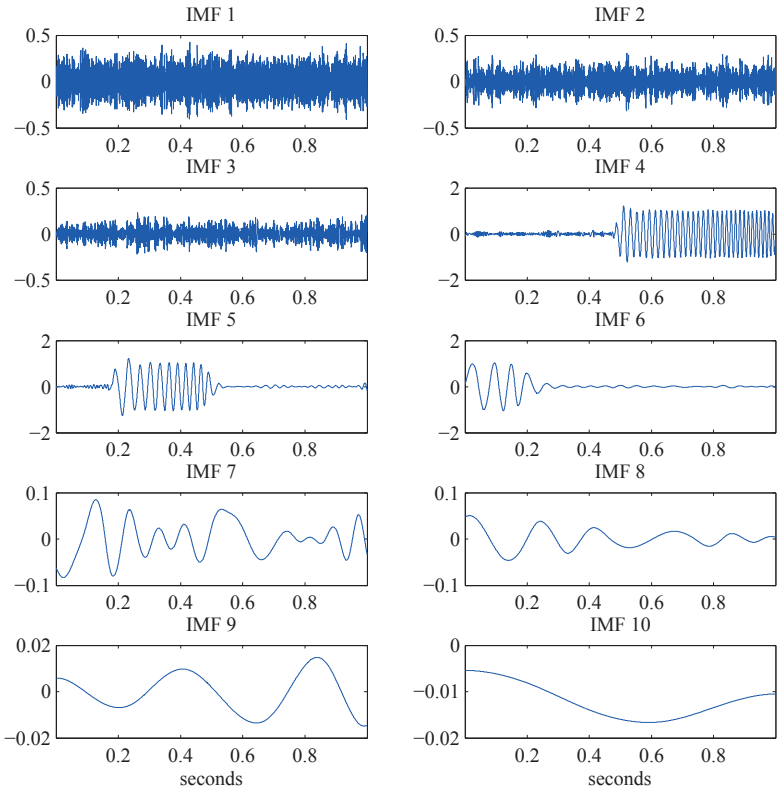


Figure 5.12: Decomposition of a noisy linear chirp. Note the signal content present in the transition IMFs 4-6.

To demonstrate the extraction of transition IMFs, we add the exact same realization of noise shown in figure 5.8 to the linear chirp $\sin[2\pi(35t^2 + 10t)]$. The decomposition of this noisy signal is shown in figure 5.12 and spectrograms of the first six IMFs are shown in figure 5.13. IMFs 1-3 show the filter bank action of EMD. The frequency of the signal is well below that of the noise, and EMD extracts the noise in a nearly dyadic fashion. We note the boundaries of the frequency bands vary with time, as expected. Once IMFs 1-3 have been removed, the next frequency band selected contains both noise and signal as can be seen in the spectrogram of IMF 4 (see figure 5.13). The noise remaining in the residual forces EMD to continue behaving as a filter bank. However, the highest frequency content of the chirp now falls within this band. In removing this band, a portion of the signal is pushed into IMF 4. In this respect, we observe the signal leaking into the noise. IMF 4 will be composed of a mixture of noise and signal: noise for the temporal

locations corresponding to those where the chirp's frequency is too low to be included; signal for the temporal locations where the chirp's frequency reaches into the noise band. Thus a transition IMF is produced, containing signal that has been prematurely removed. Because this portion of signal no longer remains in the residual, it cannot be accounted for in the next IMF. Therefore, subsequent IMFs will be damaged as each is derived from the remaining incomplete residual. This process continues for IMFs 5 and 6, and the portions of the chirp that leak into the empirically defined bands are removed with the noise in a manner similar to IMF 4. We see the formation of transition IMFs is consistent with the model presented in figure 5.11.

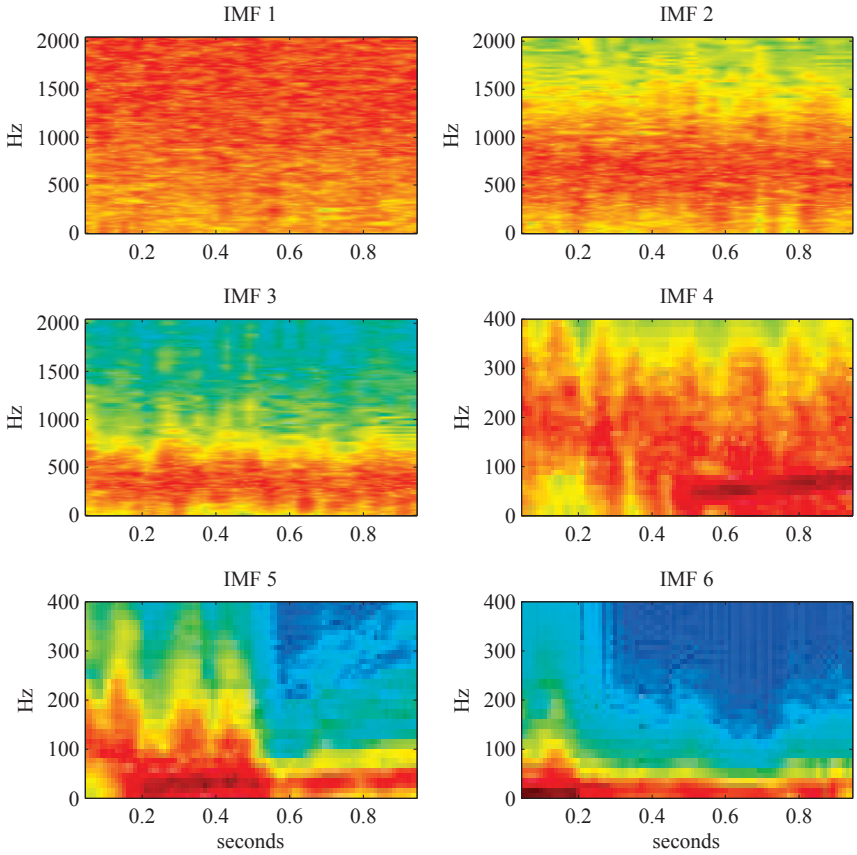


Figure 5.13: Spectrograms of the decomposition of a noisy linear chirp. Transition IMFs 4-6 display the spectral leak of signal into noise. Note the change in scale on the frequency axis.

Spectral leak is similar to the mode mixing observed by Wu and Huang [80]. To resolve

the mode mixing issue, they introduce EEMD to produce IMFs that represent only one scale of oscillation. EEMD cleverly uses noise perturbations to force the algorithm to explore all frequencies while not adding too much noise so as to push the algorithm into the spectral leak regime. Noise is added to the original signal and a standard EMD decomposition is performed. This is repeated with different noise realizations for a fixed number of times. The resulting IMFs from each run are then averaged, producing an "ensemble" result. Wu and Huang demonstrate that this is an effective way of eliminating mode mixing even in signals that contain a mild amount of noise. Our analysis continues this line of thought by examining decompositions of signals with noise of higher levels, as is often encountered in real world data. It is this noise that causes spectral leak between IMFs and presents a different problem than that solved by EEMD. Adding more noise to the already contaminated signal will not produce cleaner results. The realization of the original contaminating noise remains the same over all trials and thus cannot be eliminated through averaging. For these reasons, our analysis is focused on the standard EMD decomposition of noisy signals.

5.4.2.2 Phase alignment

The spectral explanation is not the entire story; the phase of the underlying signal also plays a role in the creation of transition IMFs. We have seen that the boundaries of the frequency bands of noisy IMFs dip lower in some locations and extend higher in others (figure 5.9). We also have observed that the standard deviation of a band's frequency varies with time (figure 5.10). When the energy of the noise is high, the energy of the signal cannot be felt by the algorithm. In this way, we think of the noise as insulating the signal from extraction. However, at a given time, if the energy of the noise is small, EMD may include part of the underlying signal in the current IMF as well. At these time locations, the noise does not insulate the signal from extraction. Thus signal leaks into an otherwise noisy IMF at the locations where the standard deviation is small. This process is illustrated by the model seen in figure 5.14, showing a noisy signal in the time-frequency plane. From 0.5 to 0.6 seconds there is a clear dip in the energy of the noise. In this region, the energy of the signal is exposed and will be extracted into the next IMF. Outside of this region,

the energy of the noise is high and insulates the underlying signal. Here, only the noise will be extracted and the signal will remain untouched. The locations at which signal is extracted into an otherwise noisy IMF will be shown to be phase dependent.

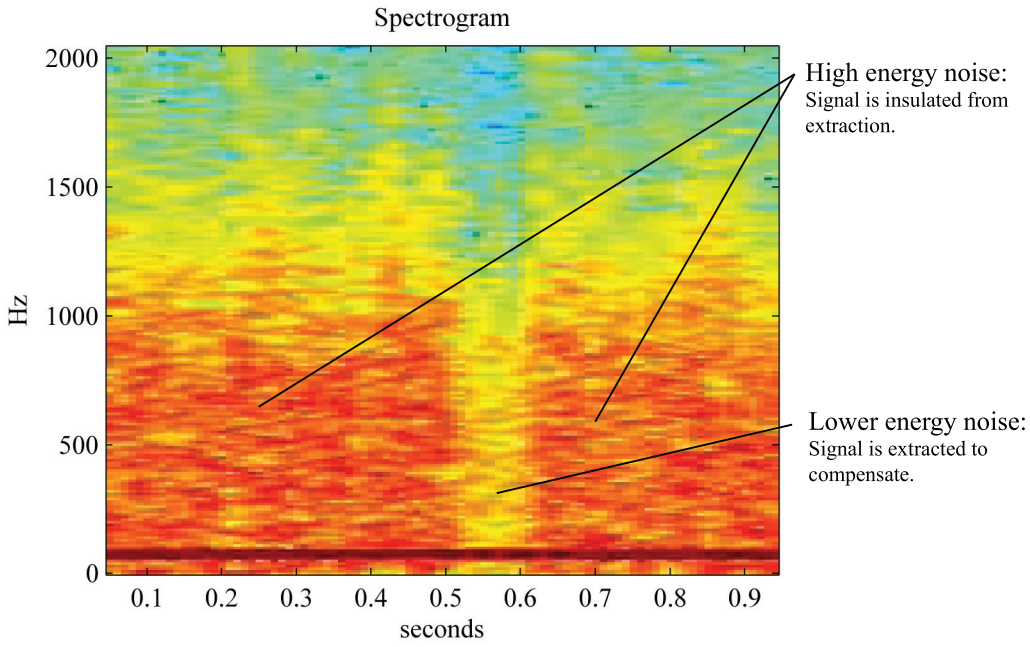
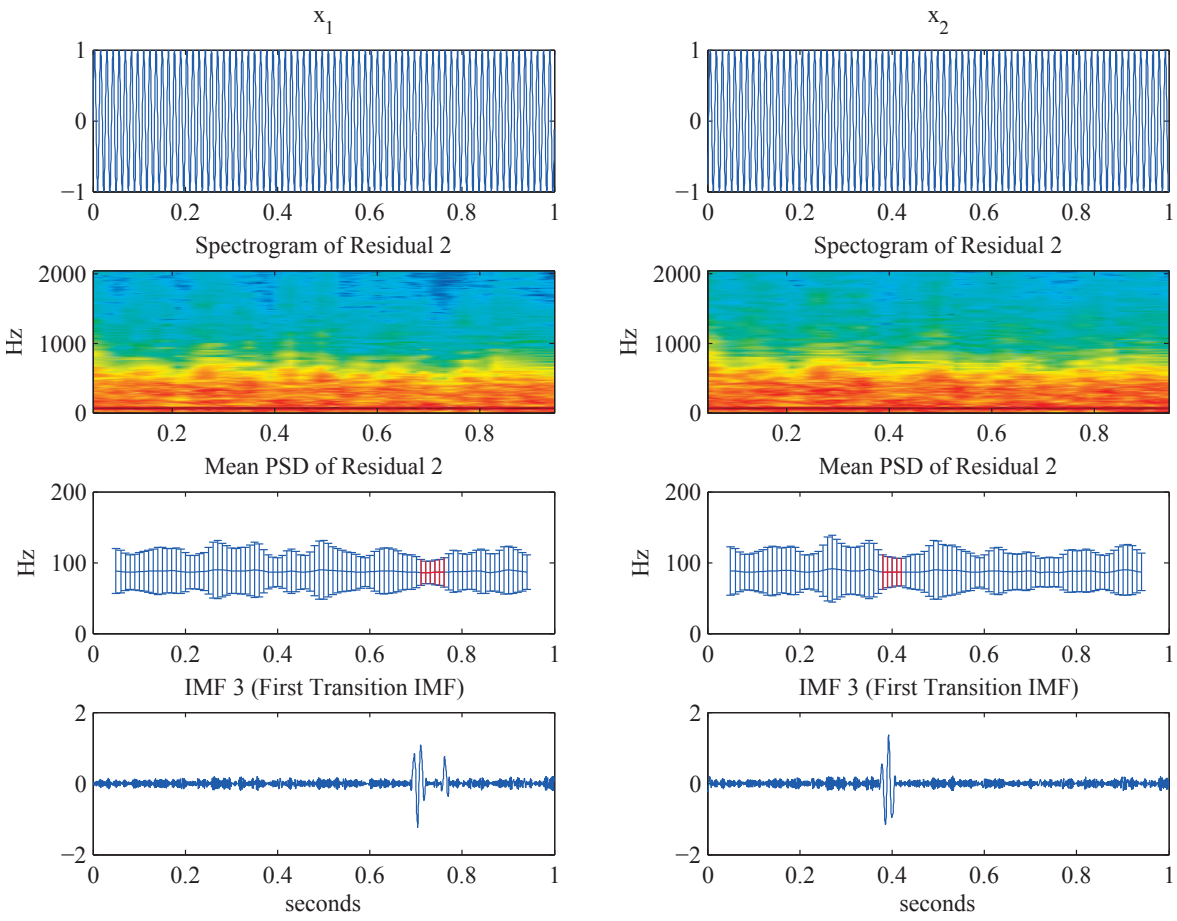


Figure 5.14: A model of a noisy signal in the time-frequency plane. Signal will be extracted in the region corresponding to 0.5-0.6 seconds. Here the energy of the noise is too low to insulate the signal from extraction. Outside of this region, only the energy of the noise will be extracted.

Consider two signals with identical spectral content, differing only by a constant phase factor and contaminated with the same noise realization. For simplicity, we consider two stationary signals. Using a stationary example will limit the effect of spectral leak, as unlike the chirp used in the previous nonstationary case, a signal with one frequency should not have energy spread over many IMFs. Let $f = 75$ Hz and $t \in [0, 1]$ seconds. We examine $x_1 = \sin(2\pi f t)$ and a phase-shifted copy $x_2 = \sin(2\pi f t + .9p)$, where $p = \frac{1}{f}$ is the period of x_1. Because x_1 and x_2 have the same frequency content, we expect that when contaminated with the same noise realization, EMD should produce very similar results. Figure 5.15 shows that the first transition IMFs for each noisy signal



(a) x_1 (b) x_2

Figure 5.15: Two stationary signals with identical spectral content differing only by a phase shift. From top to bottom: the clean signal, spectrograms of the noisy residual from which the first transition IMFs are extracted, mean power spectral density (PSD) of the residual with error bars representing one standard deviation, and the first transition IMFs. The PSD sections highlighted in red correspond to those with the smallest standard deviations and is where signal leaks into the otherwise noisy IMFs.

contain signal in different locations. Examining the residual from which each transition IMF was extracted lends an explanation. The smallest standard deviation in each residual occurs near 0.7 seconds and 0.4 seconds for x_1 and x_2 respectively and is highlighted in red. These time locations correspond exactly with the location of signal content in each transition IMF. At these locations, the level of the noise is too small to insulate the signal from extraction into the current IMF. This process demonstrates that the extraction of transition IMFs is also phase dependent.

In the above example, the first IMF contains pure noise for both signals. Because the exact same noise realization was used to contaminate both signals, one might expect that the first IMF, and thus the first residual, for each signal would be identical. However, as noted above and seen in figure 5.15, the statistics of the residuals are different, showing dips in the energy of the noise at different locations. For a more complete understanding of the demonstrated phase dependence, we consider how the phase of a signal interacts with noise. The interference between the sinusoidal function $x_i(t) = \alpha\cos(\omega t+\beta_i)$ $(i = 1 \text{ or } 2)$ and a realization $n(t)$ of the white noise can be described by the following simple model. We consider $n(t)$ to be a realization of a white noise process sampled at a finite number of samples N. We can decompose $n(t)$ using a finite Fourier transform [11] and the Fourier series expansion can be written as follows:

$$n(t) = \sum_{k=0}^{N-1} \rho_k \cos(2\pi k\frac{t}{N} + \varphi_k)$$

where the $\rho_k \geq 0$ and φ_k are defined by

$$a_k = \rho_k cos\varphi_k, \qquad b_k = -\rho_k sin\varphi_k, \qquad \text{and} \qquad a_0 = 2\rho_0 cos\varphi_0,$$

with

$$a_k = \frac{2}{N}\sum_{t=0}^{N-1} n(t)\cos(2\pi k\frac{t}{N}) \quad (k = 0,\ldots) \qquad \text{and} \qquad b_k = \frac{2}{N}\sum_{t=0}^{N-1} n(t)\sin(2\pi k\frac{t}{N}) \quad (k = 1,\ldots).$$

We now contaminate the signal $x_i(t)$ by adding the noise realization $n(t)$ to $x_i(t)$,

$$x_i(t) + n(t) = \alpha\cos(\omega t + \beta_i) + \sum_{k=0}^{N-1} \rho_k \cos(2\pi k\frac{t}{N} + \varphi_k) \qquad (t = 0, 1, \ldots, N-1).$$

Because the noise is white, we expect the realization of the noise to have a uniform distribution of the energy in the Fourier domain. In other words, we expect that all ρ_k have similar amplitudes.

We now examine under what circumstances the noise will interfere with the signal. First, we assume that the signal amplitude is about the same as the noise level, ($\alpha \approx \rho_{k_0}$). Second, we consider the frequency index of the noise that matches the frequency of the signal, k_0 such that $\omega \approx 2\pi k_0$. At this frequency the noise will interfere with the signal. Formally, we can consider the interaction of the two cosine function,

$$\alpha \cos\left(\omega \frac{t}{N} + \beta_i\right) + \rho_{k_0} \cos\left(2\pi k_0 \frac{t}{N} + \varphi_{k_0}\right) \approx$$
$$2\rho_{k_0} \cos\left(\frac{\omega + 2\pi k_0}{2} \frac{t}{N} + \frac{\beta_i + \varphi_{k_0}}{2}\right) \cos\left(\frac{\omega - 2\pi k_0}{2} \frac{t}{N} + \frac{\beta_i - \varphi_{k_0}}{2}\right).$$

If $\omega \approx 2\pi k_0$, then the function

$$\cos\left(\frac{\omega - 2\pi k_0}{2} \frac{t}{N} + \frac{\beta_i - \varphi_{k_0}}{2}\right)$$

slowly modulates the other cosine function,

$$\rho_{k_0} \cos\left(\frac{\omega + 2\pi k_0}{2} \frac{t}{N} + \frac{\beta_i + \varphi_{k_0}}{2}\right)$$

which still oscillates at the frequency ω since $(\omega + 2\pi k_0)/2 \approx \omega$. The overall amplitude of the slowly varying envelope $\cos((\omega - 2\pi k_0)/2\ t/N + (\beta_i - \varphi_{k_0})/2)$ clearly depends on the phase difference $(\beta_i - \varphi_{k_0})/2$, as is shown in figure 5.15.

We conclude that the exact amount of cancellation created by the interference between the original signal $x_i(t)$ and the noise realization $n(t)$ depends on the phase of the signal $x_i(t)$. We note that this analysis is concerned with one realization of the noise, and is not in contradiction with the fact that the noise statistical properties are translation invariant, since the noise is considered to be stationary.

5.5 EMD Decomposition of Synthetic Seismic Data

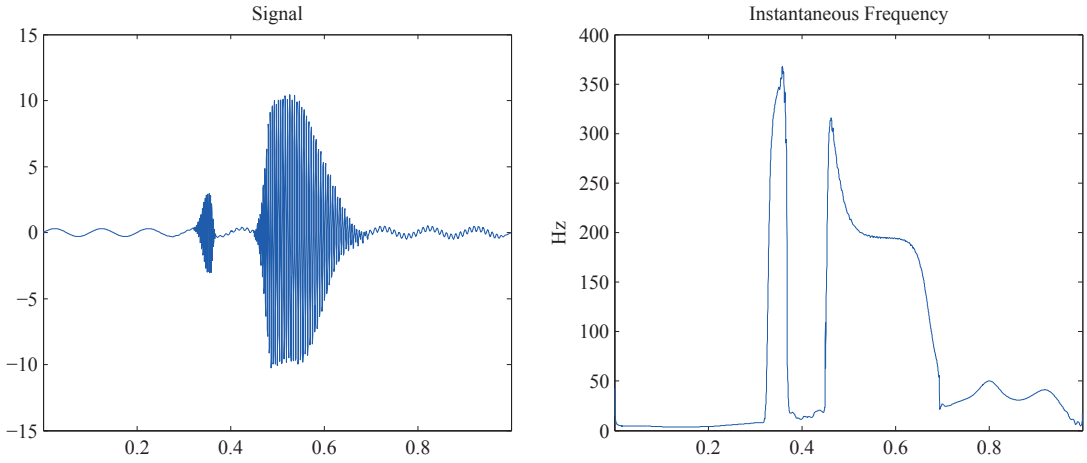
Having demonstrated both the effect and mechanism of noise corruption on simple synthetic examples, we turn our attention to a synthetic seismic signal which will serve as a model for real

world data. The signal was constructed using elementary chirplet wave packets. Such chirplet packets were proposed in [6] to decompose seismograms. Details of the construction are given in the next subsection. Figure 5.16a shows the clean signal that will be considered along with the estimate of its instantaneous frequency[3]. In the absence of noise we observe that the decomposition of the signal yields a physically meaningful IF (figure 5.16b).

To investigate the effect of noise, the same signal is contaminated with additive white Gaussian noise and we consider an SNR of 24dB. The noisy signal is shown in figure 5.17a and it is clear that a meaningful IF was not produced (figure 5.17b). Examining the IMFs of the noisy signal shows that IMF 1 contains noise and IMF 2 represents the transition from noise to signal. It is noted that 91.8% of the signal's total energy is captured in this transition IMF. Eleven IMFs were produced and figure 5.18 shows the first five, capturing 98.6% of the energy. It is clear that to produce a meaningful instantaneous frequency, IMF 1 must be discarded. IMF 2 must be included as it contains almost all of the energy, but will be problematic as it also contains noise. Recomputing the IF (not shown) using all but the first IMF fails to produce a meaningful IF estimate due to the noise present in IMF 2.

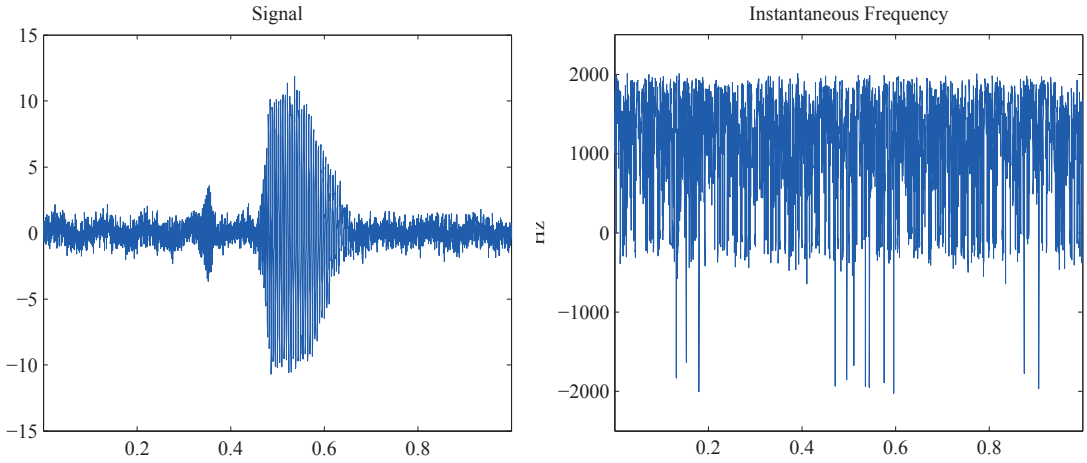
The seismic signal is clearly nonstationary. We therefore expect that the transition IMF was formed due to spectral leak. The IMFs in figure 5.18 indicate that the decomposition indeed followed the process presented in the model for spectral leak (figure 5.11). IMF 1 is pure noise, extracted by EMD operating in the filter bank regime. The spectrogram of IMF 2 shows that EMD continued down the frequency axis in a somewhat dyadic fashion. In principle, IMF 2 would have contained only pure noise, but the frequency content of the signal leaked into the bottom of this frequency band. The spectrograms of IMFs 3 - 5 show that the extraction of signal into the transition IMF damaged all subsequent IMFs.

[3] This synthetic seismic waveform is the result of the superposition of several signals, each with different frequency and amplitude functions. Therefore, the waveform is a multicomponent signal and its analytic IF is not well defined. The IF must be computed numerically (as the weighted sum of the IF from each of its IMFs) as shown in figure 5.16b.



(a) Signal (b) Instantaneous Frequency

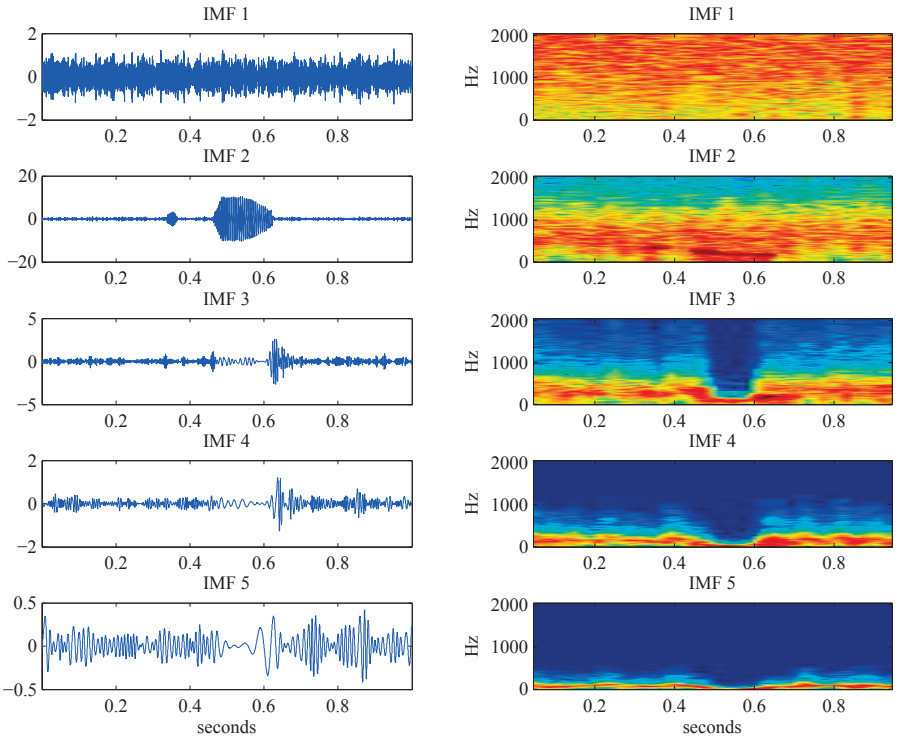
Figure 5.16: Clean seismic signal from which a physically meaningful IF is calculated.



(a) Signal (b) Instantaneous Frequency

Figure 5.17: Noisy seismic signal (SNR = 24dB) from which a physically meaningful IF cannot be calculated.

Finally, there is also evidence of phase dependence. Let the original signal be denoted by x, and consider x_1 and x_2, two phase-shifted copies of x with identical spectral content. Phase shift is accomplished by adding a constant c to the argument of the sine in the wave packet $w_k(t)$ (see Section 5.5.1). The values used for c are 0.9π and 0.3π for x_1 and x_2, respectively. Figure 5.19 shows the transition from noise to signal is captured in IMF 2 for x and x_1. Although subtle, these

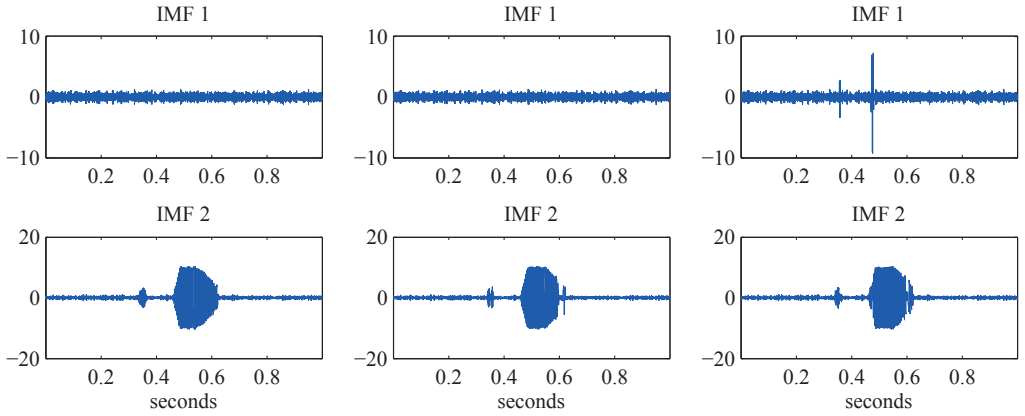


(a) IMFs

(b) Spectrograms of IMFs

Figure 5.18: First five IMFs with spectrograms from the decomposition of the noisy seismic signal. 91.8% of the total energy is captured in transition IMF 2. IMFs 3-5 are damaged by the extraction of signal into IMF 2.

IMFs contain signal at different locations (most easily seen at 0.6 seconds). A more obvious effect is seen in the decomposition of x_2, where the transition begins in IMF 1 instead of IMF 2.



(a) x (b) x_1 (c) x_2

Figure 5.19: First two IMFs of noisy seismic signals differing only by a phase factor. IMF 2 is the transition IMF for x and x_1, while the transition begins in IMF 1 for x_2. The transition IMFs for x and x_1 contain signal content in slightly different locations, most notable at time $t = 0.6$ seconds.

5.5.1 Construction of the seismic waveform

The synthetic seismic waveform, $f(t)$, used in this section is based on the work of Bardainne [6] and is constructed as follows:

Let $f(t) = \sum_{k=1}^{4} a_k w_k\left((t - t_k)/d_k\right), \quad t \in [0, 1]$

- Wave packet $w_k(t) = g(t)\,\sin\left[2\,\pi(f_k + p_k t^{q_k})\,t\right]$
- Envelope $g(t)$ = two Gaussians smoothly glued:

$$\begin{cases} \exp\left[-\left(\frac{c_k(1-l_k)-t}{\frac{1}{2}c_k(1-l_k)}\right)^2\right] & \text{if } 0 < t < c_k(1-l_k) \\ 1 & \text{if } c_k(1-l_k) < t < c_k + (1-c_k)l_k \\ \exp\left[-\left(\frac{c_k+(1-c_k)l_k-t}{\frac{1}{2}(1-c_k)(1-l_k)}\right)^2\right] & \text{if } c_k + (1-c_k)l_k < t < 1 \end{cases}$$

- (f_k, p_k, q_k) control the frequency of the wave packet

- (c_k, l_k) control the boundary between the attack and the silencing of the wave packet.

The parameter values used in this section are shown in table 5.5.1 below.

k	t_k	d_k	f_k	a_k	c_k	l_k	p_k	q_k
1	0	1	10	0.3	0.0	1	0	0
2	0.2	0.8	80	0.2	0.9	0.5	10	1
3	0.32	0.05	300	3	0.7	0.1	2	-1
4	0.45	0.24	195	10	0.2	0.2	-5	10

Table 5.1: Parameters used for constructing the seismic waveform.

5.6 Conclusions

All data analysis tools are susceptible to noise corruption; EMD is not an exception. Despite this reality, EMD has emerged as an effective tool for nonstationary data analysis. Wavelet decompositions, which suffer from similar corruption in the presence of noise, are accompanied by rich theory from which this noise corruption may be studied and understood. A complete theoretical framework for EMD has yet to emerge. Therefore, EMD is best understood through experiments to discover and test its limits. EMD is an effective tool for estimating the IF of a clean signal but provides a poor estimate in the presence of noise. When decomposing a noisy signal, "transition" IMFs are extracted, capturing both noise and signal in the same mode. Such IMFs are problematic as their noise pollutes the IF calculation yet their signal content cannot be ignored. We have demonstrated both the existence of and mechanism by which transition IMFs are created. Specifically, transition IMFs arise from spectral leak between modes and EMD's filter bank behavior in the presence of noise. In addition, the manner in which signal leaks into an otherwise noisy IMF has been shown to be phase dependent. Given this understanding, there is an opportunity to more faithfully estimate instantaneous frequency in the presence of noise. In doing so, care must be taken to treat transition IMFs in a manner that preserves any meaningful physical information, as this is an idea at the core of the development of EMD.

Chapter 6

Conclusion and Future Directions

The work in this thesis is motivated by the goal of constructing an efficient parameterization of a large data set of points lying close to a smooth manifold in high dimension. We have studied the recovery of the local tangent plane from a collection of noisy manifold samples. The tangent plane yields an efficient local parameterization that allows for the data to be well represented in fewer dimensions than those of the ambient space. Such a parameterization therefore yields a sparse representation guided by the geometry of the data.

We have presented a detailed analysis of the optimal scale for tangent plane recovery. Using local PCA, we seek a scale small enough such that the manifold is approximately linear, but a scale large enough such that structure may be discerned from noise. We use eigenspace perturbation theory to study the stability of the subspace estimated by PCA and bound, with high probability, the angle it forms with the true tangent space. The scale that optimally balances the noise-curvature trade-off is identified, yielding the optimal tangent plane estimate.

Local PCA is frequently used for subspace approximation and tangent plane recovery in the manifold learning and data analysis literature. Most often, locality is defined via a fix number of nearest neighbors or a fixed radius about a point. However, estimates that do not consider the curvature of the manifold or noise level of the data are bound to be suboptimal. The analysis in this thesis uses the geometry of the data to guide the definition of locality and thus offers new results for optimal scale selection and tangent plane recovery.

To connect this analysis with practical algorithmic considerations, we have studied a PCA

approximation-based partitioning scheme for noisy data. Our geometric analysis and numerical results indicate that the main loop of this algorithm is driven by a noise-curvature trade-off and therefore recovers a partitioning that provides an appropriate scale for tangent plane estimation. In future work, we plan to combine our analysis of the optimal scale for tangent plane recovery with the partitioning criterion of such an algorithm. Rather than resorting to an *a posteriori* comparison of the returned partitioning to the optimal scale, the algorithm should be guided by our tangent plane stability analysis. Further, as our results yield error bounds on the local scale, a partitioning may be found in a manner that maintains low-distortion guarantees over the entire data set. Such guarantees are desirable not only in this context, but in any data parameterization algorithm.

A particular area of application for the ideas discussed in this thesis is image processing. The manifold geometry of collections of images or image patches ("patch-space") has been a growing area of study over the past decade (see [23, 53], for example). In fact, face images are used as a standard data set with which to benchmark the performance of various parameterization algorithms (see the experiments in [67, 77], amongst many others). As sparse representation algorithms have recently demonstrated state-of-the-art image denoising results [1], the geometric sparsity afforded by the local tangent plane parameterization of patch-space may be used in a similar manner. Projecting noisy points (each point representing an image or image patch) into the optimal estimate of the local tangent plane effectively eliminates noise in all of the out-of-plane dimensions. As the dimensionality of the tangent space (intrinsic dimensionality) is typically much smaller than that of the ambient space, such a projection can eliminate much of the noise. Denoising via the sparsity of this geometric parameterization is similar in spirit to the well-studied idea of wavelet thresholding [58]. A key difference, however, is that thresholding transform coefficients relies on sparsity in a fixed basis, whereas our results can adaptively find the proper geometric parameterization of the data. We are currently studying the performance of image filtering in the tangent plane, comparing the statistics of this method with those of standard image filtering techniques (see the analysis in [60]). Image processing in the tangent plane therefore remains a focus of our current research.

Finally, sparsity has been demonstrated to be a powerful data model that allows for efficient

representation of complex and high-dimensional data sets. The past decade has witnessed many new and exciting results exploiting the sparsity inherent in such data. The work in this thesis has highlighted a geometric notion of sparsity that can be used to extend sparse representation to a wider range of data. The geometric interpretation of standard sparse representation is clear and has been discussed in several contexts [26, 69]. By representing points as linear combinations of a small number of basis elements or dictionary atoms, traditional sparse representation operates under the assumption of samples from a linear subspace or union of linear subspaces. Spectral clustering of the sparse coefficient matrix can then be used to partition a data set into linear clusters (see [2, 27] for example). However, most data sets exhibit curvature and thus cannot be completely characterized by such a "flat" geometric model. Extremely recent results of the past months have presented algorithmic [28] and theoretical [72] considerations of sparse representation in the manifold context. Still, a more detailed analysis incorporating curvature is needed (see [16] for recent related work). The continued development of sparse representation from a geometric perspective is a primary focus of our current work and an important extension of this thesis.

Bibliography

[1] M. Aharon, M. Elad, and A. Bruckstein. KSVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing, 54:4311–4322, 2006.

[2] E. Arias-Castro, G. Chen, and G.Lerman. Spectral clustering based on local linear approximations. Electronic Journal of Statistics, 5:1537–1587, 2011.

[3] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hedge. Model-based compressive sensing. IEEE Transactions on Information Theory, 56:1982–2001, 2010.

[4] R.G. Baraniuk, V. Cevher, and M.B. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: a geometric perspective. IEEE Proceedings Special Issue on Applications of Sparse Representation & Compressive Sensing, 98:959–971, 2010.

[5] R.G. Baraniuk and M.B. Wakin. Random projections of smooth manifolds. Foundations of Computational Mathematics, 9:51–77, 2009.

[6] T. Bardainne. Characterization of seismic waveforms and classification of seismic events using chirplet atomic decomposition. Example from the Lacq gas field (Western Pyrenees, France). Geophysical Journal International, 166:699–718, 2006.

[7] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15:1373–1396, 2003.

[8] Y. Bengio and M. Monperrus. Non-local manifold tangent learning. In Advances in Neural Information Processing Systems 17. MIT Press, 2005.

[9] B. Boashash. Estimating and interpreting the instantaneous frequency of a signal part 1: Fundamentals. Proceedings of the IEEE, 80:520–538, 1992.

[10] M. Brand. Charting a manifold. In Advances in Neural Information Processing Systems 15, pages 961–968. MIT Press, 2003.

[11] D.R. Brillinger. Time Series: Data Analysis and Theory. SIAM, 2001.

[12] D.S. Broomhead and G.P. King. Extracting qualitative dynamics from experimental data. Physica D, 20(2-3):217–236, June 1986.

[13] D.S. Broomhead and M.J. Kirby. The whitney reduction network: a method for computing autoassociative graphs. Neural Computation, 13:2595–2616, 2001.

[14] E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory, 52:489–509, 2006.

[15] E.J. Candes and T. Tao. Decoding by linear programming. IEEE Transactions on Information Theory, 51:4203–4215, 2005.

[16] G. Chen and G. Lerman. Spectral curvature clustering. International Journal of Computer Vision, 81:317–330, 2009.

[17] G. Chen, A.V. Little, M. Maggioni, and L. Rosasco. Some recent advances in multiscale geometric analysis of point clouds. In J. Cohen and A.I. Zayed, editors, Wavelets and Multiscale Analysis: Theory and Applications, pages 199–225. Springer, 2011.

[18] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proceedings of the National Academy of Sciences, 102:7426–7431, 2005.

[19] G. David and A. Averbuch. Hierarchical data organization, clustering and denoising via localized diffusion folders. Applied and Computational Harmonic Analysis, 33:1–23, 2012.

[20] C. Davis and W.M. Kahan. The rotation of eigenvectors by a perturbation III. SIAM Journal on Numerical Analysis, 7:1–46, 1970.

[21] D.L. Donoho. Compressed sensing. IEEE Transactions on Information Theory, 52:1289–1306, 2006.

[22] D.L. Donoho and C. Grimes. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences, 100:5591–5596, 2003.

[23] D.L. Donoho and C. Grimes. Image manifolds which are isometric to euclidean space. Journal of Mathematical Imaging and Vision, 23:5–24, 2005.

[24] D.L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association, 90:1200–1224, 1995.

[25] J. Einbeck, L. Evers, and C. Bailer-Jones. Representing complex data using localized principal components with application to astronomical data. In A. Gorban, B. Kegl, D. Wunch, and A. Zinovyev, editors, Principal Manifolds for Data Visualization and Dimension Reduction, pages 178–201. Springer, Berlin-Heidelberg, 2008.

[26] M. Elad. Sparse and redundant representations: from theory to applications in signal and image processing. Springer, 2010.

[27] E. Elhamifar and R. Vidal. Sparse subspace clustering. In IEEE International Conference on Computer Vision and Pattern Recognition, 2009.

[28] E. Elhamifar and R. Vidal. Sparse manifold clustering and embedding. In Advances in Neural Information Processing Systems 24, pages 55–63, 2011.

[29] M. Feiszli and P.W. Jones. Curve denoising by multiscale singularity detection and geometric shrinkage. Applied and Computational Harmonic Analysis, 31:392–409, 2011.

[30] P. Flandrin and P. Goncalves. The Hilbert spectrum via wavelet projections. International Journal of Wavelets, Multiresolution, and Information Processing, 2:477–496, 2004.

[31] H. Froehling, J.P. Crutchfield, D. Farmer, N.H. Packard, and R. Shaw. On determining the dimension of chaotic flows. Physica D, 3:605–617, 1981.

[32] K. Fukunaga. Statistical Pattern Recognition. Academic Press, 2nd edition, 1990.

[33] K. Fukunaga and D. Olsen. An algorithm for finding intrinsic dimensionality of data. IEEE Transactions on Computers, 20:176–183, 1971.

[34] M. Giaquinta and G. Modica. Mathematical Analysis: An Introduction to Functions of Several Variables. Springer, 2009.

[35] G.H. Golub and C.F. Van Loan. Matrix Computations. JHU Press, 1996.

[36] M. Gromov, M. Katz, P. Pansu, and S. Semmes. Metric Structures for Riemannian and Non-Riemannian Spaces. Birkhauser, 2001.

[37] T. Hastie, R. Tibshirani, and J.H. Friedman. The Elements of Statistical Learning. Springer, 2003.

[38] R.A. Horn and C.R. Johnson. Matrix Analysis. Cambridge University Press, 1985.

[39] N.E. Huang and S.S.P. Shen. Hilbert-Huang Transform and Its Applications. World Scientific, 2005.

[40] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N. C. Yen, C.C. Tung, and H. H. Liu. The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London A, 454:903–995, 1998.

[41] N.E. Huang, Z. Wu, S.R. Long, K.C. Arnold, X. Chen, and K. Blank. On instantaneous frequency. Advances in Adaptive Data Analysis, 1:177–229, 2009.

[42] I.M. Johnstone. On the distribution of the largest eigenvalue in principal component analysis. Annals of Statistics, 29:295–327, 2001.

[43] I.T. Jolliffe. Principal Component Analysis. Springer Verlag, 2nd edition, 2002.

[44] P.W. Jones. Rectifiable sets and the traveling salesman problem. Inventiones Mathematicae, 102:1–15, 1990.

[45] S. Jung and J.S. Marron. Pca consistency in high dimension, low sample size context. Annals of Statistics, 27:4104–4130, 2009.

[46] N. Kambhatla and T.K. Leen. Dimension reduction by local principal component analysis. Neural Computation, 9:1493–1516, 1997.

[47] D.N. Kaslovsky and F.G. Meyer. Noise corruption of Empirical Mode Decomposition and its effect on instantaneous frequency. Advances in Adaptive Data Analysis, 2:373–396, 2010.

[48] D.N. Kaslovsky and F.G. Meyer. Optimal tangent plane recovery from noisy manifold samples. Submitted to Annals of Statistics, pages 1–57, 2011.

[49] D.N. Kaslovsky and F.G. Meyer. Estimating local manifold geometry via data partitioning. In preparation, 2012.

[50] D.N. Kaslovsky and F.G. Meyer. Overcoming noise, avoiding curvature: Optimal scale selection for tangent plane recovery. In Proceedings of IEEE Conference on Statistical Signal Processing, August 2012.

[51] T.L. Kijewski-Correa and A. Kareem. Performance of wavelet transform and Empirical Mode Decomposition in extracting signals embedded in noise. Journal of Engineering Mechanics-ASCE, 133:849–852, 2007.

[52] D. Kushnir, M. Galun, and A. Brandt. Fast multiscale clustering and manifold identification. Pattern Recognition, 39:1876–1891, 2006.

[53] A.B. Lee, K.S. Pedersen, and D. Mumford. The nonlinear statistics of high-contrast patches in natural images. International Journal of Computer Vision, 54:83–103, 2003.

[54] T. Lin and H. Zha. Riemannian manifold learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30:796–809, 2008.

[55] S.P. Lloyd. Least squares quantization in pcm. IEEE Transactions on Information Theory, 28:129–137, 1982.

[56] P.J. Loughlin and B. Tacer. Instantaneous frequency and the conditional mean frequency of a signal. Signal Processing, 60:153–162, 1997.

[57] Y. Ma, A.Y. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. SIAM Review, 50:413–458, 2008.

[58] S. Mallat. A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way. Academic Press, 3rd edition, 2008.

[59] P. Massart. Concentration Inequalities and Model Selection. Springer, 2007.

[60] P. Milanfar. A tour of modern image filtering. IEEE Signal Processing Magazine, To appear.

[61] N.J. Mitra, A. Nguyen, and L. Guibas. Estimating surface normals in noisy point cloud data. International Journal of Computational Geometry and Applications, 14:261–276, 2004.

[62] R.J. Muirhead. Aspects of Multivariate Statistical Theory. Wiley Online Library, 1982.

[63] B. Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. Annals of Statistics, 36:2792–2817, 2008.

[64] D. Needell and J.A. Tropp. CoSaMP: iterative signal recovery from incomplete and inaccurate samples. Applied and Computational Harmonic Analysis, 26:301–321, 2009.

[65] Y. Ohtake, A. Belyaev, and H-P Seidel. A composite approach to meshing scattered data. Graphical Models, 68:255–267, 2006.

[66] S. Olhede and A.T. Walden. Empirical Mode Decomposition as data-driven wavelet-like expansions. Proeedings of the Royal Society of London A, 460:955–975, 2004.

[67] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 290:2323–2326, 2000.

[68] S.T. Roweis, L.K. Saul, and G.E. Hinton. Global coordination of locally linear models. In Advances in Neural Information Processing Systems 14, pages 889–896. MIT Press, 2002.

[69] R. Rubinstein, A.M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. IEEE Proceedings Special Issue on Applications of Sparse Representation & Compressive Sensing, 98:1045–1057, 2010.

[70] J. Shawe-Taylor and N. Cristianini. Estimating the moments of a random vector with applications. In Proceedings of GRETSI 2003 Conference, pages 47–52, 2003.

[71] A. Singer and H.-T. Wu. Vector diffusion maps and the connection laplacian. Communications on Pure and Applied Mathematics, 64:1067–1144, 2012.

[72] M. Soltanolkotabi and E.J. Candes. A geometric analysis of subspace clustering with outliers. http://arxiv.org/abs/1112.4258, 2011.

[73] G. W. Stewart. Matrix Algorithms, Volume II: Eigensystems. SIAM: Society for Industrial and Applied Mathematics, 2001.

[74] G.W. Stewart and J. Sun. Matrix Perturbation Theory. Academic Press, 1990.

[75] A. Szlam. Modifications on k q-flats for supervised learning. Technical Report 08-56, UCLA, 2008.

[76] T. Tao. Topics in Random Matrix Theory. American Mathematical Society, 2012.

[77] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290:2319–2322, 2000.

[78] J. Tropp and A.C. Gilbert. Signal recovery from partial information via orthogonal matching pursuit. IEEE Transactions on Information Theory, 53:4655–4666, 2007.

[79] X. Wang and J.S. Marron. A scale-based approach to finding effective dimensionality in manifold learning. Electronic Journal of Statistics, 2:127–148, 2008.

[80] Z. Wu and N.E. Huang. Ensemble Empirical Mode Decomposition: a noise-assisted data analysis method. Advances in Adaptive Data Analysis, 1:1–41, 2009.

[81] L. Yang. Alignment of overlapping locally scaled patches for multidimensional scaling and dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30:438–450, 2008.

[82] T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Randomized hybrid linear modeling by local best-fit flats. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1927–1934, 2010.

[83] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM Journal on Scientific Computing, 26:313–338, 2004.

Appendix A

Appendix: Optimal Tangent Plane Recovery From Noisy Manifold Samples

A.1 The Set Ω_e

Here the set Ω_e over which projections of the vector e have bounded norm is formally defined. Begin by recalling a standard result on the concentration of the Gaussian measure on $\mathbb{R}^N$, denoted by γ_N. Consider a random vector drawn from the $\mathcal{N}(0, \sigma^2 I_N)$ distribution. By the concentration of Gaussian measure [36], the set

$$S = \{x \in \mathbb{R}^N : \ \sqrt{N}(1-\varepsilon) \ \leq \ \|x\|/\sigma \ \leq \ \sqrt{N}(1+\varepsilon)\}$$

has measure

$$\gamma_N(S) > 1 - 2e^{-\frac{N\varepsilon^2}{2}}.$$

This result states that the Gaussian measure of the set of points in $\mathbb{R}^N$ that concentrate about the sphere of radius $\sigma\sqrt{N}$ is extremely large.

The sets

$$\Omega_1 = \{x \in \mathbb{R}^d : \ \|x\| \leq \sigma\sqrt{d}(1+\varepsilon_1)\}$$
$$\Omega_2 = \{x \in \mathbb{R}^{D-d} : \ \|x\| \leq \sigma\sqrt{D-d}(1+\varepsilon_2)\}$$

have Gaussian measure

$$\gamma_d(\Omega_1) > 1 - e^{\frac{-d\varepsilon_1^2}{2}}$$
$$\gamma_{D-d}(\Omega_2) > 1 - e^{\frac{-(D-d)\varepsilon_1^2}{2}}.$$

We may define sets

$$\Omega_{U_1} = \{x \in \mathbb{R}^D : \ \|U_1^T x\| \leq \sigma\sqrt{d}(1+\varepsilon_1)\}$$
$$\Omega_{U_2} = \{x \in \mathbb{R}^D : \ \|U_2^T x\| \leq \sigma\sqrt{D-d}(1+\varepsilon_2)\}$$

such that Ω_{U_1} and Ω_{U_2} are the preimages of Ω_1 and Ω_2, respectively, in $\mathbb{R}^D$. The Gaussian measures of the sets Ω_1 and Ω_2 are the pushforwards of the Gaussian measure in $\mathbb{R}^D$ by the respective projections U_1^T and U_2^T:

$$\gamma_D(\overline{\Omega}_{U_1}) = \gamma_d(\overline{\Omega}_1) \leq e^{\frac{-d\varepsilon_1^2}{2}}$$
$$\gamma_D(\overline{\Omega}_{U_2}) = \gamma_{D-d}(\overline{\Omega}_2) \leq e^{\frac{-(D-d)\varepsilon_2^2}{2}},$$

where $\overline{\Omega}$ denotes the complement of the set Ω.

Finally, define

$$\Omega_e = \Omega_{U_1} \cap \ \Omega_{U_2}$$

and a standard union bound argument yields

$$\gamma_D(\overline{\Omega}_e) = \gamma_D(\overline{\Omega}_{U_1} \cup \ \overline{\Omega}_{U_2}) = \gamma_D(\overline{\Omega}_{U_1}) + \gamma_D(\overline{\Omega}_{U_2}) \leq e^{\frac{-d\varepsilon_1^2}{2}} + e^{\frac{-(D-d)\varepsilon_2^2}{2}}.$$

Set $\varepsilon_1\sqrt{d/2} = \varepsilon_2\sqrt{(D-d)/2} = \xi_e$. Then both

$$\|U_1^T e\| \leq \sigma\left(\sqrt{d} + \xi_e\sqrt{2}\right)$$
$$\|U_2^T e\| \leq \sigma\left(\sqrt{D-d} + \xi_e\sqrt{2}\right)$$

hold on Ω_e, and

$$\gamma_D(\Omega_e) > 1 - 2e^{-\xi_e^2}.$$

A.2 Suprema and Expectations for Main Result 1

A.2.1 Suprema R_{ab}^{pq} and R_a^p

Listed here are the suprema terms needed for the calculations leading to Main Result 1. The calculation is outlined in Chapter 3.3.2.1.

First we have R_a^p terms:

$$
\begin{aligned}
R_\ell^1 &\le r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}}, \\
R_c^2 &\le \frac{1}{2}Kr_{max}^2\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}}, \\
R_e^1 &\le \sigma\left(\sqrt{d}+\xi_e\sqrt{2}\right) && \text{on } \Omega_e, \\
R_e^2 &\le \sigma\left(\sqrt{D-d}+\xi_e\sqrt{2}\right) && \text{on } \Omega_e.
\end{aligned}
$$

Then using (3.3.19) we may bound the remaining R_{ab}^{pq}:

$$
\begin{aligned}
R_{cc}^{22} &\le \frac{1}{4}K^2r_{max}^4\left(\frac{N}{N_{max}}\right)^{\frac{4}{d}}, \\
R_{c\ell}^{21} &\le \frac{1}{2}Kr_{max}^3\left(\frac{N}{N_{max}}\right)^{\frac{3}{d}}, \\
R_{ee}^{11} &\le \sigma^2\left(\sqrt{d}+\xi_e\sqrt{2}\right)^2 && \text{on } \Omega_e, \\
R_{ee}^{22} &\le \sigma^2\left(\sqrt{D-d}+\xi_e\sqrt{2}\right)^2 && \text{on } \Omega_e, \\
R_{ee}^{21} &\le \sigma^2\left(\sqrt{d}+\xi_e\sqrt{2}\right)\left(\sqrt{D-d}+\xi_e\sqrt{2}\right) && \text{on } \Omega_e, \\
R_{e\ell}^{11} &\le \sigma r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}}\left(\sqrt{d}+\xi_e\sqrt{2}\right) && \text{on } \Omega_e, \\
R_{e\ell}^{21} &\le \sigma r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}}\left(\sqrt{D-d}+\xi_e\sqrt{2}\right) && \text{on } \Omega_e, \\
R_{ce}^{21} &\le \frac{1}{2}K\sigma r_{max}^2\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}}\left(\sqrt{d}+\xi_e\sqrt{2}\right) && \text{on } \Omega_e, \\
R_{ce}^{22} &\le \frac{1}{2}K\sigma r_{max}^2\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}}\left(\sqrt{D-d}+\xi_e\sqrt{2}\right) && \text{on } \Omega_e.
\end{aligned}
$$

A.2.2 Expectations

Here we detail the calculation of the expectations from Chapter 3.3.2.2. Each term is of the form

$$\left\| \mathbb{E}[U_p^T(a - \mathbb{E}[a])(b - \mathbb{E}[b])^T U_q] \right\|_F \;=\; \left\| \mathbb{E}[a_{u_p} b_{u_q}^T] - \mathbb{E}[a_{u_p}]\,\mathbb{E}[b_{u_q}^T] \right\|_F.$$

As only pure curvature (cc^T) and pure noise (ee^T) terms have nonzero expectation, the calculations of all other terms are omitted.

Pure Curvature Term: The expectation of the pure curvature term is computed as follows. Consider first $(c_{u_2} c_{u_2}^T)_{i,j} = c_i c_j$ for $i, j = (d+1), \ldots, D$. Then

$$\begin{aligned}
\mathbb{E}[c_i c_j] \;&=\; \frac{1}{4}\,\mathbb{E}[(\kappa_1^{(i)}\ell_1^2 + \;\ldots\; + \kappa_d^{(i)}\ell_d^2)(\kappa_1^{(j)}\ell_1^2 + \;\ldots\; + \kappa_d^{(j)}\ell_d^2)] \\
&=\; \frac{1}{4} K_{nn}^{ij}\,\mathbb{E}[\ell_n^4] \;+\; \frac{1}{4} K_{mn}^{ij}\,\mathbb{E}[\ell_m^2 \ell_n^2] \\
&=\; \frac{r_{max}^4}{4(d+2)(d+4)} \left(\frac{N}{N_{max}}\right)^{\frac{4}{d}} \left[3K_{nn}^{ij} + K_{mn}^{ij}\right]. \qquad \text{(A.2.1)}
\end{aligned}$$

Next consider $(c_{u_2})_i = c_i$ for $i = (d+1), \ldots, D$. Then

$$\mathbb{E}[c_i]\,\mathbb{E}[c_j] = \frac{r_{max}^4}{4(d+2)^2} \left(\frac{N}{N_{max}}\right)^{\frac{4}{d}} \left[K_{nn}^{ij} + K_{mn}^{ij}\right] \qquad \text{(A.2.2)}$$

and we have

$$\begin{aligned}
&\left\| \mathbb{E}[c_{u_2} c_{u_2}^T] - \mathbb{E}[c_{u_2}]\,\mathbb{E}[c_{u_2}^T] \right\|_F = \\
&\frac{r_{max}^4}{2(d+2)^2(d+4)} \left(\frac{N}{N_{max}}\right)^{\frac{4}{d}} \left[\sum_{i=d+1}^{D} \sum_{j=d+1}^{D} \left[(d+1)K_{nn}^{ij} - K_{mn}^{ij}\right]^2\right]^{\frac{1}{2}}. \qquad \text{(A.2.3)}
\end{aligned}$$

Pure Noise Terms: Using that the entries of e are i.i.d. random variables from the $\mathcal{N}(0, \sigma^2)$ distribution, we have

$$\left\| \mathbb{E}[e_{u_p} e_{u_q}^T] - \mathbb{E}[e_{u_p}]\,\mathbb{E}[e_{u_q}^T] \right\|_F \;=\; \left\| \mathbb{E}[e_{u_p} e_{u_q}^T] \right\|_F =$$

$$\begin{cases} \left(\sum_{i=1}^{d} \sum_{j=1}^{d} \mathbb{E}[e_i e_j]^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^{d} \mathbb{E}[e_i^2]^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{d} \mathbb{E}[e_i e_j]^2 \right)^{\frac{1}{2}} = \sigma^2 \sqrt{d} & \\ & \text{if } (p,q) = (1,1), \\ \left(\sum_{i=d+1}^{D} \sum_{j=d+1}^{D} \mathbb{E}[e_i e_j]^2 \right)^{\frac{1}{2}} = \left(\sum_{i=d+1}^{D} \mathbb{E}[e_i^2]^2 + \sum_{\substack{i,j=d+1 \\ i \neq j}}^{D} \mathbb{E}[e_i e_j]^2 \right)^{\frac{1}{2}} = \sigma^2 \sqrt{D-d} & \\ & \text{if } (p,q) = (2,2), \\ \left(\sum_{i=d+1}^{D} \sum_{j=1}^{d} \mathbb{E}[e_i e_j]^2 \right)^{\frac{1}{2}} = 0 & \text{if } (p,q) = (2,1). \end{cases}$$

A.3 Norm Bounds for Main Result 1

The right-hand side of the confidence interval (3.3.18) from Chapter 3 is used to bound the size of the perturbation norms. When considering noise terms, recall that we must condition on $e \in \Omega_e$. Recalling the rescaled notation $r = r_{max}\,(N/N_{max})^{1/d}$ is helpful for interpretation. Note that we may work with either the matrix in question or its transpose when computing the norm and our notation may reflect either choice.

Curvature

$$\left\| U_2^T \left(\tfrac{1}{N} \widetilde{C} \widetilde{C}^T \right) U_2 \right\|_F \leq$$

$$\begin{aligned} & \frac{r_{max}^4}{2(d+2)^2(d+4)} \left(\frac{N}{N_{max}} \right)^{\frac{4}{d}} \left[\sum_{i=d+1}^{D} \sum_{j=d+1}^{D} \left[(d+1) K_{nn}^{ij} - K_{mn}^{ij} \right]^2 \right]^{\frac{1}{2}} \\ + \;\; & \frac{1}{\sqrt{N}} \frac{K^2}{4} r_{max}^4 \left(\frac{N}{N_{max}} \right)^{\frac{4}{d}} \left[\left(2 + \eta_{cc} \sqrt{2} \right) + \frac{1}{\sqrt{N}} \left(2 + \eta_c \sqrt{2} \right)^2 \right] \end{aligned} \tag{A.3.1}$$

with probability at least $1 - e^{-\eta_{cc}^2} - e^{-\eta_c^2}$ over the random selection of the sample points.

Noise

$$\left\|U_1^T\left(\tfrac{1}{N}\widetilde{E}\widetilde{E}^T\right)U_1\right\|_F \leq$$
$$\sigma^2\sqrt{d} \;+\; \frac{1}{\sqrt{N}}\sigma^2\left(\sqrt{d}+\xi_e\sqrt{2}\right)^2\left[\left(2+\eta_{ee}\sqrt{2}\right)+\frac{1}{\sqrt{N}}\left(2+\eta_e\sqrt{2}\right)^2\right] \tag{A.3.2}$$

with probability at least $1-e^{-\eta_{ee}^2}-e^{-\eta_e^2}-2e^{-\xi_e^2}$ over the random realization of the noise.

$$\left\|U_2^T\left(\frac{1}{N}\widetilde{E}\widetilde{E}^T\right)U_2\right\|_F \leq\; \sigma^2\sqrt{D-d} \;+\; \frac{1}{\sqrt{N}}\sigma^2\left(\sqrt{D-d}+\xi_e\sqrt{2}\right)^2$$
$$\times\left[\left(2+\eta_{ee}\sqrt{2}\right)+\frac{1}{\sqrt{N}}\left(2+\eta_e\sqrt{2}\right)^2\right] \tag{A.3.3}$$

with probability at least $1-e^{-\eta_{ee}^2}-e^{-\eta_e^2}-2e^{-\xi_e^2}$ over the random realization of the noise.

$$\left\|U_2^T\left(\frac{1}{N}\widetilde{E}\widetilde{E}^T\right)U_1\right\|_F \leq\; \frac{1}{\sqrt{N}}\sigma^2\left(\sqrt{d}+\xi_e\sqrt{2}\right)\left(\sqrt{D-d}+\xi_e\sqrt{2}\right)$$
$$\times\left[\left(2+\eta_{ee}\sqrt{2}\right)+\frac{1}{\sqrt{N}}\left(2+\eta_e\sqrt{2}\right)^2\right] \tag{A.3.4}$$

with probability at least $1-e^{-\eta_{ee}^2}-e^{-\eta_e^2}-2e^{-\xi_e^2}$ over the random realization of the noise.

Linear-Curvature Interaction

$$\left\|U_2^T\left(\tfrac{1}{N}\widetilde{C}\widetilde{L}^T\right)U_1\right\|_F \leq$$
$$\frac{1}{\sqrt{N}}\frac{K}{2}r_{max}^3\left(\frac{N}{N_{max}}\right)^{\frac{3}{d}}\left[\left(2+\eta_{\ell c}\sqrt{2}\right)+\frac{1}{\sqrt{N}}\left(2+\eta_\ell\sqrt{2}\right)\left(2+\eta_c\sqrt{2}\right)\right] \tag{A.3.5}$$

with probability at least $1-e^{-\eta_{\ell c}^2}-e^{-\eta_\ell^2}-e^{-\eta_c^2}$ over the random selection of the sample points.

Linear-Noise Interaction

$$\left\|U_1^T\left(\frac{1}{N}\widetilde{E}\widetilde{L}^T\right)U_1\right\|_F \leq\; \frac{1}{\sqrt{N}}\sigma r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}}\left(\sqrt{d}+\xi_e\sqrt{2}\right)$$
$$\times\left[\left(2+\eta_{\ell e}\sqrt{2}\right)+\frac{1}{\sqrt{N}}\left(2+\eta_\ell\sqrt{2}\right)\left(2+\eta_e\sqrt{2}\right)\right] \tag{A.3.6}$$

with probability at least $1 - e^{-\eta_{\ell e}^2} - e^{-\eta_\ell^2} - e^{-\eta_e^2} - 2e^{-\xi_e^2}$ over the joint random selection of the sample points and random realization of the noise.

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{E} \widetilde{L}^T \right) U_1 \right\|_F \leq \frac{1}{\sqrt{N}} \sigma r_{max} \left(\frac{N}{N_{max}} \right)^{\frac{1}{d}} \left(\sqrt{D-d} + \xi_e \sqrt{2} \right) \times \left[\left(2 + \eta_{\ell e} \sqrt{2} \right) + \frac{1}{\sqrt{N}} \left(2 + \eta_\ell \sqrt{2} \right) \left(2 + \eta_e \sqrt{2} \right) \right] \tag{A.3.7}$$

with probability at least $1 - e^{-\eta_{\ell e}^2} - e^{-\eta_\ell^2} - e^{-\eta_e^2} - 2e^{-\xi_e^2}$ over the joint random selection of the sample points and random realization of the noise.

Curvature-Noise Interaction

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{C} \widetilde{E}^T \right) U_1 \right\|_F \leq \frac{1}{\sqrt{N}} \frac{K}{2} \sigma r_{max}^2 \left(\frac{N}{N_{max}} \right)^{\frac{2}{d}} \left(\sqrt{d} + \xi_e \sqrt{2} \right) \times \left[\left(2 + \eta_{ce} \sqrt{2} \right) + \frac{1}{\sqrt{N}} \left(2 + \eta_c \sqrt{2} \right) \left(2 + \eta_e \sqrt{2} \right) \right] \tag{A.3.8}$$

with probability at least $1 - e^{-\eta_{ce}^2} - e^{-\eta_c^2} - e^{-\eta_e^2} - 2e^{-\xi_e^2}$ over the joint random selection of the sample points and random realization of the noise. Note that $\| U_1^T \left(\frac{1}{N} E C^T \right) U_2 \|_F = \| U_2^T \left(\frac{1}{N} C E^T \right) U_1 \|_F$.

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{C} \widetilde{E}^T \right) U_2 \right\|_F \leq \frac{1}{\sqrt{N}} \frac{K}{2} \sigma r_{max}^2 \left(\frac{N}{N_{max}} \right)^{\frac{2}{d}} \left(\sqrt{D-d} + \xi_e \sqrt{2} \right) \times \left[\left(2 + \eta_{ce} \sqrt{2} \right) + \frac{1}{\sqrt{N}} \left(2 + \eta_c \sqrt{2} \right) \left(2 + \eta_e \sqrt{2} \right) \right] \tag{A.3.9}$$

with probability at least $1 - e^{-\eta_{ce}^2} - e^{-\eta_c^2} - e^{-\eta_e^2} - 2e^{-\xi_e^2}$ over the joint random selection of the sample points and random realization of the noise.

A.4 Moment Calculations

The following moment calculations will be used in the confidence interval calculations. Let L_i be the random variable that returns the ith coordinate of a point from $B^d(r)$, randomly chosen according to a uniform distribution. Let $x = [x_1\ x_2\ \ldots\ x_d]$ and compute the following expectations

with respect to μ, the uniform measure on $B^d(r)$:

$$\mathbb{E}[L_i^a] = \int_{B^d(r)} x_i^a \, d\mu(x)$$

$$= \frac{1}{\mathrm{vol}(B^d(r))} \int_{-r}^{r} x_i^a \int\limits_{\substack{x1,\ldots,x_{i-1},x_{i+1},\ldots,x_d \\ \in\ B^{d-1}(\sqrt{r^2-x_i^2})}} dx_1 \ \ldots \ dx_{i-1}\, dx_{i+1} \ \ldots \ dx_d \ dx_i$$

$$= \frac{1}{\mathrm{vol}(B^d(r))} \int_{-r}^{r} x_i^a \ \mathrm{vol}(B^{d-1}(\sqrt{r^2 - x_i^2})) \ dx_i$$

$$= \frac{\mathrm{vol}(B^{d-1}(1))}{\mathrm{vol}(B^d(r))} \int_{-r}^{r} x_i^a \ (r^2 - x_i^2)^{\frac{d-1}{2}} \ dx_i.$$

Similarly,

$$\mathbb{E}[L_i^a L_j^b] = \frac{\mathrm{vol}(B^{d-2}(1))}{\mathrm{vol}(B^d(r))} \int_{-r}^{r} \int_{-\sqrt{r^2-x_i^2}}^{\sqrt{r^2-x_i^2}} x_i^a \, x_j^b \ (r^2 - x_i^2 - x_j^2)^{\frac{d-2}{2}} \ dx_j \, dx_i.$$

Then we compute the following moments:

$\mathbb{E}[L_i] = 0$	$\mathrm{Var}[L_i] = \frac{r^2}{d+2}$
$\mathbb{E}[L_i L_j] = 0$	$\mathrm{Var}[L_i L_j] = \frac{r^4}{(d+2)(d+4)}$
$\mathbb{E}[L_i^2] = \frac{r^2}{d+2}$	$\mathrm{Var}[L_i^2] = \frac{2(d+1)r^4}{(d+2)^2(d+4)}$
$\mathbb{E}[L_i L_j^2] = 0$	$\mathrm{Var}[L_i L_j^2] = \frac{3r^6}{(d+2)(d+4)(d+6)}$
$\mathbb{E}[L_i^3] = 0$	$\mathrm{Var}[L_i^3] = \frac{15r^6}{(d+2)(d+4)(d+6)}$
$\mathbb{E}[L_i^2 L_j^2] = \frac{r^4}{(d+2)(d+4)}$	$\mathrm{Var}[L_i^2 L_j^2] = \frac{8(d^2+5d+3)r^8}{(d+2)^2(d+4)^2(d+6)(d+8)}$
$\mathbb{E}[L_i^4] = \frac{3r^4}{(d+2)(d+4)}$	$\mathrm{Var}[L_i^4] = \frac{24(d+1)(4d+17)r^8}{(d+2)^2(d+4)^2(d+6)(d+8)}$.

A.5 Central Limit Theorem Calculations for Main Result 2

Here we detail the Central Limit Theorem (CLT)-based calculations that are used for Main Result 2. In the following analysis we will write $\frac{1}{N}\sum_{k=1}^{N} Y_k \xrightarrow{d} Y$ meaning that the sum converges in distribution to the random variable Y, and we will then indicate the distribution from which Y is drawn.

A.5.1 Matrix Entries

A.5.1.1 Centering

We first compute the entries of the matrices representing the centering terms $\widehat{\mathbb{E}}[L]$, $\widehat{\mathbb{E}}[C]$, and $\widehat{\mathbb{E}}[E]$.

- Linear

$$\widehat{\mathbb{E}}[L_i] \;=\; \frac{1}{N}\sum_{k=1}^{N} L_{i,k} \;\xrightarrow{d}\; Y \;\in\; [\mu-\Gamma,\ \mu+\Gamma] \tag{A.5.1}$$

$$\mu = 0 \tag{A.5.2}$$

$$\Gamma = \eta_L \frac{1}{\sqrt{N}}\sqrt{\frac{2}{d+2}} r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}}, \tag{A.5.3}$$

with probability greater than $1-e^{-\eta_L^2}$, where $Y \sim \mathcal{N}\left(\mathbb{E}[L_i], \frac{1}{N}\mathrm{Var}[L_i]\right)$.

- Curvature

$$\begin{aligned}
\widehat{\mathbb{E}}[C_i] &= \frac{1}{N}\sum_{k=1}^{N} C_{i,k} \\
&= \frac{1}{2}\left[\kappa_1^{(i)}\frac{1}{N}\sum_{k=1}^{N} L_{1,k}^2 + \cdots + \kappa_d^{(i)}\frac{1}{N}\sum_{k=1}^{N} L_{d,k}^2\right] \\
&\xrightarrow{d} \frac{1}{2}\left[\kappa_1^{(i)} Y_1 + \cdots + \kappa_d^{(i)} Y_d\right] \\
&\in\ [\mu-\Gamma,\ \mu+\Gamma]
\end{aligned} \tag{A.5.4}$$

$$\mu = \frac{K_i}{2}\frac{r_{max}^2}{(d+2)}\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}} \tag{A.5.5}$$

$$\Gamma = \frac{K_i}{2}\frac{r_{max}^2}{(d+2)}\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}}\left(\frac{2\eta_C}{\sqrt{N}}\sqrt{\frac{d+1}{d+4}}\right) \tag{A.5.6}$$

with probability greater than $1 - de^{-\eta_C^2}$, where $Y_j \sim \mathcal{N}\Big(\mathbb{E}[L_i^2], \frac{1}{N}\mathrm{Var}[L_i^2]\Big)$ for $j = 1, \ldots, d$.

- Noise

$$\widehat{\mathbb{E}}[E_i] \;=\; \frac{1}{N}\sum_{k=1}^{N} E_{i,k} \;=\; Y \in \left[-\eta_E\frac{\sigma\sqrt{2}}{\sqrt{N}},\ \eta_E\frac{\sigma\sqrt{2}}{\sqrt{N}}\right] \tag{A.5.7}$$

with probability greater than $1 - e^{-\eta_E^2}$, where $Y \sim \mathcal{N}\left(0, \frac{1}{N}\sigma^2\right)$.

Combining these results, we will use the following centering terms:

- $\widehat{\mathbb{E}}[C_i]\widehat{\mathbb{E}}[C_j] \;\in\; [\mu - \Gamma,\ \mu + \Gamma]$

$$\mu = \frac{K_iK_j}{4(d+2)^2}r_{max}^4\left(\frac{N}{N_{max}}\right)^{\frac{4}{d}} \tag{A.5.8}$$

$$\Gamma = \frac{K_iK_j}{4(d+2)^2}r_{max}^4\left(\frac{N}{N_{max}}\right)^{\frac{4}{d}}\left(\frac{4\eta_C}{\sqrt{N}}\sqrt{\frac{d+1}{d+4}} + \frac{4\eta_C^2}{N}\left(\frac{d+1}{d+4}\right)\right) \tag{A.5.9}$$

with probability $> 1 - de^{-\eta_C^2}$,

- $\widehat{\mathbb{E}}[E_i]\widehat{\mathbb{E}}[E_j] \;\in\; \left[-\eta_E^2\frac{2\sigma^2}{N},\ \eta_E^2\frac{2\sigma^2}{N}\right]$

with probability $> \begin{cases} 1 - 2e^{-\eta_E^2} & (i \neq j) \\ 1 - e^{-\eta_E^2} & (i = j), \end{cases}$

- $\widehat{\mathbb{E}}[L_i]\widehat{\mathbb{E}}[C_j] \;\in\; [\mu - \Gamma,\ \mu + \Gamma]$

$$\mu = 0 \tag{A.5.10}$$

$$\Gamma = \eta_L\frac{1}{\sqrt{N}}\frac{K_j}{\sqrt{2}(d+2)^{\frac{3}{2}}}r_{max}^3\left(\frac{N}{N_{max}}\right)^{\frac{3}{d}}\left[1 + \frac{2\eta_C}{\sqrt{N}}\sqrt{\frac{d+1}{d+4}}\right] \tag{A.5.11}$$

with probability $> 1 - e^{-\eta_L^2} - de^{-\eta_C^2}$,

- $\widehat{\mathbb{E}}[L_i]\widehat{\mathbb{E}}[E_j] \ \in \ [\mu - \Gamma, \ \mu + \Gamma]$

$$\mu = 0 \tag{A.5.12}$$

$$\Gamma = \eta_L \eta_E \frac{1}{N} \frac{2\sigma}{\sqrt{d+2}} r_{max} \left(\frac{N}{N_{max}} \right)^{\frac{1}{d}} \tag{A.5.13}$$

with probability $> 1 - e^{-\eta_L^2} - e^{-\eta_E^2}$,

- $\widehat{\mathbb{E}}[C_i]\widehat{\mathbb{E}}[E_j] \ \in \ [\mu - \Gamma, \ \mu + \Gamma]$

$$\mu = 0 \tag{A.5.14}$$

$$\Gamma = \eta_E \frac{1}{\sqrt{N}} \frac{\sigma K_i}{\sqrt{2}(d+2)} r_{max}^2 \left(\frac{N}{N_{max}} \right)^{\frac{2}{d}} \left[1 + \frac{2\eta_C}{\sqrt{N}} \sqrt{\frac{d+1}{d+4}} \right] \tag{A.5.15}$$

with probability $> 1 - de^{-\eta_C^2} - e^{-\eta_E^2}$.

A.5.1.2 Curvature

The entries of the pure curvature term CC^T are computed as follows. Note that the curvature term is the only one for which the entries grow with N.

$$\begin{aligned}
&\frac{1}{N} \sum_{k=1}^{N} C_{i,k} C_{j,k} \qquad (A.5.16)\\
&= \frac{1}{N} \sum_{k=1}^{N} \frac{1}{2} \left(\kappa_1^{(i)} L_{1,k}^2 + \cdots + \kappa_d^{(i)} L_{d,k}^2 \right) \ \frac{1}{2} \left(\kappa_1^{(j)} L_{1,k}^2 + \cdots + \kappa_d^{(j)} L_{d,k}^2 \right) \\
&= \frac{1}{4} \sum_{n=1}^{d} \kappa_n^{(i)} \kappa_n^{(j)} \frac{1}{N} \sum_{k=1}^{N} L_{n,k}^4 \ + \ \frac{1}{4} \sum_{\substack{m,n=1 \\ m \neq n}}^{d} \kappa_m^{(i)} \kappa_n^{(j)} \frac{1}{N} \sum_{k=1}^{N} L_{m,k}^2 L_{n,k}^2 \\
&\xrightarrow{d} \frac{1}{4} \sum_{n=1}^{d} \kappa_n^{(i)} \kappa_n^{(j)} \ Z_n \ + \ \frac{1}{4} \sum_{\substack{m,n=1 \\ m \neq n}}^{d} \kappa_m^{(i)} \kappa_n^{(j)} \ Y_{mn} \\
&\in \ [\mu - \Gamma, \ \mu + \Gamma]
\end{aligned}$$

$$\mu = \frac{1}{4}\frac{r_{max}^4}{(d+2)(d+4)}\left(\frac{N}{N_{max}}\right)^{\frac{4}{d}}\left[3K_{nn}^{ij} + K_{mn}^{ij}\right] \tag{A.5.17}$$

$$\begin{aligned}\Gamma = &\frac{1}{\sqrt{N}}\frac{r_{max}^4}{2\sqrt{2}(d+2)(d+4)}\left(\frac{N}{N_{max}}\right)^{\frac{4}{d}} \\ &\times\left(\eta_{CC_1}K_{nn}^{ij}\sqrt{\frac{24(d+1)(4d+17)}{(d+6)(d+8)}} + \eta_{CC_2}K_{mn}^{ij}\sqrt{\frac{8(d^2+5d+3)}{(d+6)(d+8)}}\right)\end{aligned} \tag{A.5.18}$$

with probability greater than $1 - de^{-\eta_{CC_1}^2} - \frac{d(d-1)}{2}e^{-\eta_{CC_2}^2}$, where $Z_n \sim \mathcal{N}\Big(\mathbb{E}[L_i^4], \frac{1}{N}\mathrm{Var}[L_i^4]\Big)$ and $Y_{mn} \sim \mathcal{N}\Big(\mathbb{E}[L_i^2L_j^2], \frac{1}{N}\mathrm{Var}[L_i^2L_j^2]\Big)$, for $m, n = 1, \ldots, d$. Subtracting $\widehat{\mathbb{E}}[C_i]\widehat{\mathbb{E}}[C_j]$ from (A.5.16),

$$\left(\frac{1}{N}\widetilde{C}\widetilde{C}^T\right)_{i,j} \in \ [\mu - \Gamma,\ \mu + \Gamma] \tag{A.5.19}$$

$$\mu = \frac{1}{2}\frac{r_{max}^4}{(d+2)^2(d+4)}\left(\frac{N}{N_{max}}\right)^{\frac{4}{d}}\left[(d+1)K_{nn}^{ij} - K_{mn}^{ij}\right] \tag{A.5.20}$$

$$\begin{aligned}\Gamma = &\frac{1}{\sqrt{N}}\frac{r_{max}^4}{4(d+2)}\left(\frac{N}{N_{max}}\right)^{\frac{4}{d}} \\ &\times\left(\eta_{CC_1}K_{nn}^{ij}\sqrt{\frac{48(d+1)(4d+17)}{(d+4)^2(d+6)(d+8)}} + 4\eta_{CC_2}K_{mn}^{ij}\sqrt{\frac{(d^2+5d+3)}{(d+4)^2(d+6)(d+8)}}\right. \\ &\left.+\frac{4\eta_C K_i K_j}{(d+2)}\sqrt{\frac{d+1}{d+4}} + \frac{4\eta_C^2 K_i K_j}{\sqrt{N}}\frac{(d+1)}{(d+2)(d+4)}\right)\end{aligned} \tag{A.5.21}$$

with probability greater than $1 - de^{-\eta_{CC_1}^2} - \frac{d(d-1)}{2}e^{-\eta_{CC_2}^2} - de^{-\eta_C^2}$. Note that only the lower right $(D-d)\times(D-d)$ entries are nonzero.

A.5.1.3 Noise

A diagonal entry of the pure noise matrix EE^T is the square of the norm of a vector in $\mathbb{R}^N$ of Gaussian random variables. We could use the concentration of Gaussian measure to bound this norm (see Section A.1), but obtain a slightly tighter result using the CLT. Note that this norm neither grows nor decays with N, and its leading order term σ^2 represents a noise-floor. An off-diagonal entry is the inner-product between two such vectors, and we therefore expect it to be small. Using a Bernstein-type inequality [59] to bound such entries yields the same leading behavior but with higher order terms, whereas using the CLT does not. Properties of the Wishart distribution could also be used [62]. Note that the off-diagonal terms tend to zero as N grows.

- Diagonal entry ($i = j$)

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N} E_{i,k}E_{i,k} = \ & \frac{\sigma^2}{N}\sum_{k=1}^{N} Y_k \ \xrightarrow{d} \ \sigma^2 Z \\ \in \ & \left[\sigma^2\left(1-\eta_{EE_1}\frac{2}{\sqrt{N}}\right),\ \sigma^2\left(1+\eta_{EE_1}\frac{2}{\sqrt{N}}\right)\right]\end{aligned} \tag{A.5.22}$$

with probability greater than $1-e^{-\eta_{EE_1}^2}$, where $Y_k \sim \chi^2(1)$ so that $\mathbb{E}[Y_k]=1$, $\mathrm{Var}[Y_k]=2$, and $Z \sim \mathcal{N}\left(\mathbb{E}[Y_k], \frac{1}{N}\mathrm{Var}[Y_k]\right)$. Subtracting $\widehat{\mathbb{E}}[E_i]\widehat{\mathbb{E}}[E_j]$ yields

$$\begin{aligned}&\left(\frac{1}{N}\widetilde{E}\widetilde{E}^T\right)_{i,i} \in \\ &\left[\sigma^2\left(1-\frac{2}{\sqrt{N}}\left(\eta_{EE_1}+\frac{1}{\sqrt{N}}\eta_E^2\right)\right),\ \sigma^2\left(1+\frac{2}{\sqrt{N}}\left(\eta_{EE_1}+\frac{1}{\sqrt{N}}\eta_E^2\right)\right)\right]\end{aligned} \tag{A.5.23}$$

with probability greater than $1-e^{-\eta_{EE_1}^2}-e^{-\eta_E^2}$.

- Off-diagonal entry ($i \neq j$)

$$\frac{1}{N}\sum_{k=1}^{N} E_{i,k}E_{j,k} \ \xrightarrow{d} \ Y \ \in \ \left[-\eta_{EE_2}\frac{\sigma^2\sqrt{2}}{\sqrt{N}},\ \eta_{EE_2}\frac{\sigma^2\sqrt{2}}{\sqrt{N}}\right] \tag{A.5.24}$$

with probability greater than $1-e^{-\eta_{EE_2}^2}$, where $Y \sim \mathcal{N}\left(\mathbb{E}[E_iE_j], \frac{1}{N}\mathrm{Var}[E_iE_j]\right)$. Subtracting $\widehat{\mathbb{E}}[E_i]\widehat{\mathbb{E}}[E_j]$ yields

$$\left(\frac{1}{N}\widetilde{E}\widetilde{E}^T\right)_{i,j} \ \in \ \left[-\frac{\sigma^2\sqrt{2}}{\sqrt{N}}\left(\eta_{EE_2}+\eta_E^2\sqrt{\frac{2}{N}}\right),\ \frac{\sigma^2\sqrt{2}}{\sqrt{N}}\left(\eta_{EE_2}+\eta_E^2\sqrt{\frac{2}{N}}\right)\right] \tag{A.5.25}$$

with probability greater than $1-e^{-\eta_{EE_2}^2}-2e^{-\eta_E^2}$.

A.5.1.4 Linear-Curvature Interaction

The entries of the linear-curvature term are computed as follows.

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N} L_{i,k}C_{j,k} \qquad\qquad (A.5.26)\\
&= \frac{1}{N}\sum_{k=1}^{N} L_{i,k}\,\frac{1}{2}\left(\kappa_1^{(j)}L_{1,k}^2+\cdots+\kappa_d^{(j)}L_{d,k}^2\right)\\
&= \frac{1}{2}\left[\kappa_1^{(j)}\frac{1}{N}\sum_{k=1}^{N}L_{i,k}L_{1,k}^2 \;+\; \ldots \;+\; \kappa_i^{(j)}\sum_{k=1}^{N}L_{i,k}^3 \;+\; \ldots \;+\; \kappa_d^{(j)}\frac{1}{N}\sum_{k=1}^{N}L_{i,k}L_{d,k}^2\right]\\
&\xrightarrow{d} \frac{1}{2}\left[\sum_{\substack{n=1\\ n\neq i}}^{d}\kappa_n^{(j)}Y_n \;+\; \kappa_i^{(j)}Z\right]\\
&\in \;[\mu-\Gamma,\;\mu+\Gamma]
\end{aligned}$$

$$\mu = 0 \qquad (A.5.27)$$

$$\begin{aligned}
\Gamma = \frac{1}{\sqrt{N}}\sqrt{\frac{3}{2(d+2)(d+4)(d+6)}}r_{max}^3\left(\frac{N}{N_{max}}\right)^{\frac{3}{d}} \qquad (A.5.28)\\
\times\left[\eta_{LC_1}\sum_{\substack{n=1\\ n\neq i}}^{d}\kappa_n^{(j)} \;+\; \eta_{LC_2}\sqrt{5}\kappa_i^{(j)}\right]
\end{aligned}$$

with probability greater than $1-(d-1)e^{-\eta_{LC_1}^2}-e^{-\eta_{LC_2}^2}$, where $Z\sim\mathcal{N}\Big(\mathbb{E}[L_i^3],\frac{1}{N}\mathrm{Var}[L_i^3]\Big)$, and $Y_n\sim\mathcal{N}\Big(\mathbb{E}[L_iL_j^2],\frac{1}{N}\mathrm{Var}[L_iL_j^2]\Big)$, for $n=1,\ldots,i-1,i+1,\ldots d$, and Subtracting $\widehat{\mathbb{E}}[L_i]\widehat{\mathbb{E}}[C_j]$,

$$\left(\frac{1}{N}\widetilde{L}\widetilde{C}^T\right)_{i,j}\in\;[\mu-\Gamma,\;\mu+\Gamma] \qquad (A.5.29)$$

$$\mu = 0 \qquad (A.5.30)$$

$$\begin{aligned}
\Gamma = \frac{1}{\sqrt{N}}\frac{r_{max}^3}{\sqrt{2(d+2)}}\left(\frac{N}{N_{max}}\right)^{\frac{3}{d}}\left[K_j\left(\frac{\eta_{LC_1}\sqrt{3}}{\sqrt{(d+4)(d+6)}}+\frac{\eta_L}{(d+2)}\right)\right. \qquad (A.5.31)\\
\left.+\;\kappa_i^{(j)}\frac{\sqrt{3}\left(\eta_{LC_2}\sqrt{5}-\eta_{LC_1}\right)}{\sqrt{(d+4)(d+6)}} \;+\; K_j\frac{1}{\sqrt{N}}\frac{2\eta_L\eta_C}{(d+2)}\sqrt{\frac{d+1}{d+4}}\right]
\end{aligned}$$

with probability greater than $1-(d-1)e^{-\eta_{LC_1}^2}-e^{-\eta_{LC_2}^2}-e^{-\eta_L^2}-de^{-\eta_C^2}$.

A.5.1.5 Linear-Noise Interaction

An entry of the linear-noise matrix may be shown to be a Lipschitz function of Gaussian variables on a set with large measure. One may show that on this set, such a function concentrates tightly about its expectation (see [59]). Using the CLT to compute the Lipschitz constant yields the same leading order behavior as directly applying the CLT to the entries, but results in higher order terms as well. Thus we proceed with the usual CLT calculation.

$$\frac{1}{N}\sum_{k=1}^{N} L_{i,k}E_{j,k} \ \xrightarrow{d} \ Y \ \in \tag{A.5.32}$$
$$\left[-\eta_{LE}\frac{\sigma}{\sqrt{N}}\sqrt{\frac{2}{d+2}}r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}}, \ \eta_{LE}\frac{\sigma}{\sqrt{N}}\sqrt{\frac{2}{d+2}}r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}}\right]$$

with probability greater than $1-e^{-\eta_{LE}^2}$, where $Y \sim \mathcal{N}\left(\mathbb{E}[L_iE_j], \frac{1}{N}\mathrm{Var}[L_iE_j]\right)$. Subtracting $\widehat{\mathbb{E}}[L_i]\widehat{\mathbb{E}}[E_j]$,

$$\left(\frac{1}{N}\widetilde{L}\widetilde{E}^T\right)_{i,j} \in \ [\mu-\Gamma, \ \mu+\Gamma] \tag{A.5.33}$$

$$\mu = 0 \tag{A.5.34}$$
$$\Gamma = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{2}{d+2}}r_{max}\left(\frac{N}{N_{max}}\right)^{\frac{1}{d}}\left[\eta_{LE}+\eta_L\eta_E\frac{\sqrt{2}}{\sqrt{N}}\right] \tag{A.5.35}$$

with probability greater than $1-e^{-\eta_{LE}^2}-e^{-\eta_L^2}-e^{-\eta_E^2}$.

A.5.1.6 Curvature-Noise Interaction

The entries of the curvature-noise matrix may be shown to be Lipschitz functions over a large set and the same comment holds as in the linear-noise case. Directly applying the CLT to the entries of this matrix, we have

$$\frac{1}{N}\sum_{k=1}^{N} C_iE_j \ \xrightarrow{d} Y \ \in \ [\mu-\Gamma, \ \mu+\Gamma] \tag{A.5.36}$$

$$\mu = 0 \tag{A.5.37}$$
$$\Gamma = \eta_{CE}\frac{1}{\sqrt{N}}\frac{\sigma r_{max}^2}{\sqrt{2(d+2)(d+4)}}\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}}\sqrt{3K_{nn}^{ii}-K_{mn}^{ii}} \tag{A.5.38}$$

with probability with probability greater than $1 - e^{-\eta_{CE}^2}$, where $Y \sim \mathcal{N}\left(\mathbb{E}[C_i E_j], \frac{1}{N}\mathrm{Var}[C_i E_j]\right)$. Subtracting $\widehat{\mathbb{E}}[C_i]\widehat{\mathbb{E}}[E_j]$,

$$\left(\frac{1}{N}\widetilde{C}\widetilde{E}^T\right)_{i,j} \in \ [\mu - \Gamma,\ \mu + \Gamma] \tag{A.5.39}$$

$$\mu = 0 \tag{A.5.40}$$

$$\begin{aligned}\Gamma = &\frac{\sigma}{\sqrt{N}}\frac{r_{max}^2}{\sqrt{2(d+2)}}\left(\frac{N}{N_{max}}\right)^{\frac{2}{d}} \\ &\times\left[\frac{\eta_{CE}}{\sqrt{d+4}}\sqrt{3K_{nn}^{ii} - K_{mn}^{ii}} + \frac{\eta_E}{\sqrt{d+2}}K_i\left(1 + \frac{2\eta_C}{\sqrt{N}}\sqrt{\frac{d+1}{d+4}}\right)\right]\end{aligned} \tag{A.5.41}$$

with probability greater than $1 - e^{-\eta_{CE}^2} - de^{-\eta_C^2} - e^{-\eta_E^2}$.

A.5.2 Norm Bounds

Recall the following constants, previously defined in Chapter 3.4.3 and restated here for clarity:

$$\begin{aligned}\mathcal{CC}_{ij} = \Bigg\{ &\frac{2\left[(d+1)K_{nn}^{ij} - K_{mn}^{ij}\right]}{(d+2)(d+4)} + \Bigg[\frac{1}{\sqrt{N}}\Bigg(\eta_{CC_1}K_{nn}^{ij}\sqrt{\frac{48(d+1)(4d+17)}{(d+4)^2(d+6)(d+8)}} \\ &+4\eta_{CC_2}K_{mn}^{ij}\sqrt{\frac{(d^2+5d+3)}{(d+4)^2(d+6)(d+8)}} + \frac{4\eta_C K_i K_j}{(d+2)}\sqrt{\frac{d+1}{d+4}} \\ &+\frac{1}{\sqrt{N}}\frac{4\eta_C^2 K_i K_j(d+1)}{(d+2)(d+4)}\Bigg)\Bigg]\Bigg\},\end{aligned}$$

$$\begin{aligned}\mathcal{LC}_{ij} = \Bigg[&K_j\left(\frac{\eta_{LC_1}\sqrt{3}}{\sqrt{(d+4)(d+6)}} + \frac{\eta_L}{(d+2)}\right) + \kappa_i^{(j)}\frac{\sqrt{3}\left(\eta_{LC_2}\sqrt{5} - \eta_{LC_1}\right)}{\sqrt{(d+4)(d+6)}} \\ &+K_j\frac{1}{\sqrt{N}}\frac{2\eta_L\eta_C}{(d+2)}\sqrt{\frac{d+1}{d+4}}\Bigg],\end{aligned}$$

$$\mathcal{CE}_i = \left[\frac{\eta_{CE}}{\sqrt{d+4}}\sqrt{3K_{nn}^{ii} + K_{mn}^{ii}} + \frac{\eta_E}{\sqrt{d+2}}K_i\left(1 + \frac{2\eta_C}{\sqrt{N}}\sqrt{\frac{d+1}{d+4}}\right)\right],$$

$$\mathcal{EE}(x) = \sigma^2\sqrt{x}\left[\sqrt{2}\left(\eta_{EE_1}\sqrt{2} + \eta_{EE_2}\sqrt{x-1}\right) + \frac{2}{\sqrt{x}}\eta_E^2\left(1 + \sqrt{x-1}\right)\right],$$

We now use the confidence intervals computed above to bound each perturbation norm. Note again that we may work with either the matrix in question or its transpose when computing the norm and our notation may reflect either choice.

- **Curvature**

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{C} \widetilde{C}^T \right) U_2 \right\|_F \leq \frac{1}{4} \frac{r_{max}^4}{(d+2)} \left(\frac{N}{N_{max}} \right)^{\frac{4}{d}} \left(\sum_{i=d+1}^{D} \sum_{j=d+1}^{D} \mathcal{C}\mathcal{C}_{ij}^2 \right)^{\frac{1}{2}} \tag{A.5.42}$$

with probability greater than
$1 - (D-d)^2 \left[d e^{-\eta_{CC_1}^2} + \frac{d(d-1)}{2} e^{-\eta_{CC_2}^2} \right] - d e^{-\eta_C^2}$.

- **Noise**

$$\left\| U_1^T \left(\frac{1}{N} \widetilde{E} \widetilde{E}^T \right) U_1 \right\|_F \leq \tag{A.5.43}$$
$$\sigma^2 \sqrt{d} \left[1 + \frac{\sqrt{2}}{\sqrt{N}} \left(\eta_{EE_1} \sqrt{2} + \eta_{EE_2} \sqrt{d-1} \right) + \frac{2}{N} \eta_E^2 \left(1 + \sqrt{d-1} \right) \right]$$

with probability greater than $1 - d e^{-\eta_{EE_1}^2} - \frac{d(d-1)}{2} e^{-\eta_{EE_2}^2} - d e^{-\eta_E^2}$.

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{E} \widetilde{E}^T \right) U_2 \right\|_F \leq \sigma^2 \sqrt{D-d} \tag{A.5.44}$$
$$\times \left[1 + \frac{\sqrt{2}}{\sqrt{N}} \left(\eta_{EE_1} \sqrt{2} + \eta_{EE_2} \sqrt{D-d-1} \right) + \frac{2}{N} \eta_E^2 \left(1 + \sqrt{D-d-1} \right) \right]$$

with probability greater than
$1 - (D-d) e^{-\eta_{EE_1}^2} - \frac{(D-d)(D-d-1)}{2} e^{-\eta_{EE_2}^2} - (D-d) e^{-\eta_E^2}$.

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{E} \widetilde{E}^T \right) U_1 \right\|_F \leq \sigma^2 \sqrt{d(D-d)} \frac{\sqrt{2}}{\sqrt{N}} \left[\eta_{EE_2} + \eta_E^2 \frac{\sqrt{2}}{\sqrt{N}} \right] \tag{A.5.45}$$

with probability greater than $1 - d(D-d) e^{-\eta_{EE_2}^2} - D e^{-\eta_E^2}$.

- **Linear-Curvature Interaction**

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{C} \widetilde{L}^T \right) U_1 \right\|_F \leq \frac{1}{\sqrt{N}} \frac{r_{max}^3}{\sqrt{2(d+2)}} \left(\frac{N}{N_{max}} \right)^{\frac{3}{d}} \left(\sum_{i=1}^{d} \sum_{j=d+1}^{D} \mathcal{LC}_{ij}^2 \right)^{\frac{1}{2}} \tag{A.5.46}$$

with probability greater than

$1 - d(D-d) \left[(d-1) e^{-\eta_{LC_1}^2} + e^{-\eta_{LC_2}^2} \right] - d e^{-\eta_L^2} - d e^{-\eta_C^2}$.

- **Linear-Noise Interaction**

$$\left\| U_1^T \left(\frac{1}{N} \widetilde{E} \widetilde{L}^T \right) U_1 \right\|_F \leq \tag{A.5.47}$$
$$d \frac{\sigma}{\sqrt{N}} \sqrt{\frac{2}{d+2}} r_{max} \left(\frac{N}{N_{max}} \right)^{\frac{1}{d}} \left[\eta_{LE} + \eta_L \eta_E \frac{\sqrt{2}}{\sqrt{N}} \right]$$

with probability greater than $1 - d^2 e^{-\eta_{LE}^2} - d e^{-\eta_L^2} - d e^{-\eta_E^2}$.

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{E} \widetilde{L}^T \right) U_1 \right\|_F \leq \tag{A.5.48}$$
$$\sqrt{d(D-d)} \frac{\sigma}{\sqrt{N}} \sqrt{\frac{2}{d+2}} r_{max} \left(\frac{N}{N_{max}} \right)^{\frac{1}{d}} \left[\eta_{LE} + \eta_L \eta_E \frac{\sqrt{2}}{\sqrt{N}} \right]$$

with probability greater than $1 - d^2 e^{-\eta_{LE}^2} - d e^{-\eta_L^2} - d e^{-\eta_E^2}$.

- **Curvature-Noise Interaction**

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{C} \widetilde{E}^T \right) U_1 \right\|_F \leq \frac{\sigma}{\sqrt{N}} \sqrt{\frac{d}{2(d+2)}} r_{max}^2 \left(\frac{N}{N_{max}} \right)^{\frac{2}{d}} \left(\sum_{i=d+1}^{D} \mathcal{CE}_i^2 \right)^{\frac{1}{2}} \tag{A.5.49}$$

with probability at least $1 - d(D-d)e^{-\eta_{CE}^2} - de^{-\eta_C^2} - de^{-\eta_E^2}$.

$$\left\| U_2^T \left(\frac{1}{N} \widetilde{C} \widetilde{E}^T \right) U_2 \right\|_F \leq \frac{\sigma}{\sqrt{N}} \sqrt{\frac{D-d}{2(d+2)}} r_{max}^2 \left(\frac{N}{N_{max}} \right)^{\frac{2}{d}} \left(\sum_{i=d+1}^{D} \mathcal{CE}_i^2 \right)^{\frac{1}{2}} \tag{A.5.50}$$

with probability greater than

$1 - (D-d)^2 e^{-\eta_{CE}^2} - de^{-\eta_C^2} - (D-d)e^{-\eta_E^2}$.

The norm bounds are combined to yield Main Result 2 (equation (3.4.13) in Chapter 3) and a union bound is used to establish its associated probability.