

Estimating the Intrinsic Dimension of  
High-Dimensional Data Sets:  
A Multiscale, Geometric Approach

by

Anna V. Little

Department of Mathematics  
Duke University

Date: \_\_\_\_\_

Approved:

---

Mauro Maggioni, Advisor

---

Thomas Beale

---

James Nolen

---

Jonathan Mattingly

Dissertation submitted in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy in the Department of Mathematics  
in the Graduate School of Duke University  
2011

ABSTRACT  
(Applied mathematics)

Estimating the Intrinsic Dimension of  
High-Dimensional Data Sets:  
A Multiscale, Geometric Approach

by

Anna V. Little

Department of Mathematics  
Duke University

Date: \_\_\_\_\_

Approved:

---

Mauro Maggioni, Advisor

---

Thomas Beale

---

James Nolen

---

Jonathan Mattingly

An abstract of a dissertation submitted in partial fulfillment of the requirements for  
the degree of Doctor of Philosophy in the Department of Mathematics  
in the Graduate School of Duke University  
2011

Copyright © 2011 by Anna V. Little  
All rights reserved except the rights granted by the  
Creative Commons Attribution-Noncommercial Licence

# Abstract

This work deals with the problem of estimating the intrinsic dimension of noisy, high-dimensional point clouds. A general class of sets which are locally well-approximated by  $k$  dimensional planes but which are embedded in a  $D \gg k$  dimensional Euclidean space are considered. Assuming one has samples from such a set, possibly corrupted by high-dimensional noise, if the data is linear the dimension can be recovered using PCA. However, when the data is non-linear, PCA fails, overestimating the intrinsic dimension. A multiscale version of PCA is thus introduced which is robust to small sample size, noise, and non-linearities in the data.

For Trevor, whose constant friendship and support made this work possible.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>List of Tables</b>	<b>x</b>
<b>List of Figures</b>	<b>xi</b>
<b>List of Abbreviations and Symbols</b>	<b>xii</b>
<b>Acknowledgements</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Notation and Setting</b>	<b>3</b>
2.1 Notation . . . . .	3
2.2 Setting . . . . .	5
<b>3 Background</b>	<b>9</b>
3.1 Intrinsic Dimension Estimation . . . . .	9
3.1.1 Notions of dimensionality . . . . .	10
3.1.2 Estimators based on the correlation integral . . . . .	11
3.1.3 Nearest neighbor estimators . . . . .	12
3.1.4 Estimators based on local PCA . . . . .	13
3.1.5 Estimators based on Bayesian methods . . . . .	15
3.2 Results from Random Matrix Theory . . . . .	16
3.2.1 Largest eigenvalue of Wigner matrices . . . . .	16
3.2.2 Extreme eigenvalues of covariance matrices . . . . .	18

3.2.3	Approximation of covariance matrices . . . . .	20
3.3	Intersection of Geometric Measure Theory & Harmonic Analysis . . .	21
<b>4</b>	<b>Motivating Examples</b>	<b>25</b>
4.1	Multiscale SSV's of a Curve . . . . .	25
4.2	Multiscale SSV's of the Sphere . . . . .	27
4.2.1	Set-up . . . . .	27
4.2.2	Tangent SSV's . . . . .	29
4.2.3	Curvature SSV . . . . .	32
4.3	Multiscale SSV's of Co-dimension One Manifold . . . . .	35
4.3.1	Moments of the unit sphere . . . . .	35
4.3.2	Moments of the unit ball . . . . .	39
4.3.3	Calculation for general co-dimension one manifold . . . . .	40
<b>5</b>	<b>Assumptions and Main Results</b>	<b>45</b>
5.1	Assumptions . . . . .	45
5.1.1	Assumptions on geometry . . . . .	45
5.1.2	Assumptions on noise . . . . .	47
5.1.3	Observations . . . . .	47
5.2	Main Results . . . . .	48
<b>6</b>	<b>Analysis of Simplified Model</b>	<b>55</b>
6.1	$\mathbf{P}_1$ : Geometric Cross-terms . . . . .	56
6.2	$\mathbf{P}_2$ : Tangent and Normal Sampling . . . . .	57
6.3	$\mathbf{P}_3$ : Tangential and Normal Noise . . . . .	57
6.4	$\mathbf{P}_4$ : Noisy Cross-terms . . . . .	60
6.5	Largest Gap . . . . .	62

<b>7</b>	<b>Recentering and Noise</b>	<b>67</b>
7.1	Recentering . . . . .	70
7.1.1	Comparing $\tilde{X}_{n,\tilde{z},r_\sigma}$ and $\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2$ . . . . .	70
7.1.2	Comparing $\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2$ and $\tilde{X}_{n,z,r_{2\sigma}}$ . . . . .	72
7.2	The Effect of Noise . . . . .	74
7.2.1	Comparing $\widetilde{X}_{n,\tilde{z},r} = (\widetilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r}) \cup I$ with $\widetilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r}$ . . . . .	75
7.2.2	Comparing $\widetilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r} = (\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2) \setminus Q_1$ with $\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2$ . . . . .	79
7.2.3	Comparing $\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2$ and $\tilde{X}_{n,\tilde{z},r_\sigma}$ . . . . .	82
7.3	Summary . . . . .	85
<b>8</b>	<b>Algorithm</b>	<b>87</b>
8.1	Pseudo-code . . . . .	87
8.2	Computational Complexity . . . . .	89
8.3	Numerical Experiments . . . . .	90
8.3.1	Manifold data . . . . .	90
8.3.2	Collections of manifolds . . . . .	95
8.3.3	Real-world data . . . . .	97
<b>9</b>	<b>Extensions and Applications</b>	<b>100</b>
9.1	Bi-Lipschitz Perturbations . . . . .	100
9.2	Dimensionality Reduction Algorithms . . . . .	102
<b>10</b>	<b>Conclusion</b>	<b>106</b>
<b>A</b>	<b>Perturbation Lemmas</b>	<b>108</b>
<b>B</b>	<b>Concentration Inequalities</b>	<b>114</b>
<b>C</b>	<b>Inequalities for Covariance Operators</b>	<b>118</b>
<b>D</b>	<b>Norms of Random Matrices</b>	<b>123</b>
<b>E</b>	<b>Outlier Removal</b>	<b>128</b>

**Bibliography**

**135**

**Biography**

**141**

# List of Tables

6.1	High probability events in a simplified model. . . . .	58
6.2	Definition of the $P_i$ 's, which bound the $\mathbf{P}_i$ perturbations. . . . .	61
6.3	Bounding of the $P_i$ 's. . . . .	64
7.1	High probability events: recentering and noise. . . . .	73
8.1	ID estimates for various manifolds with no noise. . . . .	94
8.2	ID estimation for MNIST database of hand written digits. . . . .	97

# List of Figures

2.1	Multiscale SSV's of $\mathbb{S}^9$ . . . . .	7
4.1	Rescaled multiscale SSV's of the sphere: $k = 2, 4, 8, 16, 32$ . . . . .	35
7.1	Set intersections. . . . .	69
8.1	Pseudo-code for MSVD algorithm. . . . .	88
8.2	Manifold data sets: cube. . . . .	91
8.3	Manifold data sets: sphere. . . . .	92
8.4	Manifold data sets: S-shaped manifold and Meyer's staircase. . . . .	93
8.5	Two challenging examples . . . . .	95
8.6	ID estimation for collections of manifolds. . . . .	96
8.7	ID estimation for real-world data sets. . . . .	99

# List of Abbreviations and Symbols

## Symbols

$\ \cdot\ $	The usual Euclidean norm when applied to a vector; the spectral norm when applied to a matrix.
$\mathbb{S}_R^k$	The $k$ -dimensional sphere of radius $R$ in $\mathbb{R}^{k+1}$ , centered at the origin; $\mathbb{S}^k = \mathbb{S}_1^k$ .
$\mathbb{B}_R^k$	Closed Euclidean ball of radius $R$ in $\mathbb{R}^k$ , centered at the origin; $\mathbb{B}^k = \mathbb{B}_1^k$ .

## Abbreviations

w.h.p.	With high probability.
i.i.d.	Independent, identically distributed.
r.v.	Random variable.
ID	Intrinsic dimension.
SSV	Squared singular value.
MSVD	Multiscale singular value decomposition.
PCA	Principal component analysis.

# Acknowledgements

I would like to thank Mauro Maggioni for his insightful guidance; also Lorenzo Rosasco, Yoon-Mo Jung, Guangliang Chen, and Jason Lee for their contributions to this research. I am also grateful to the National Science Foundation and the Office of Naval Research for their support under grants NSF DMS 0650413 and ONR N00014-07-1-0625.

On a more personal note, I would like to thank my family: my husband Trevor for his love and support, and for making me smile during stressful times; my parents Bob and Vicki, for their confidence in me, inspiring me to dream, and their unconditional love and acceptance; my sister Rachel and brother-in-law Michael, for their friendship and all the fun times spent together; my in-laws Jeri and Gloria, second parents to me in every way; my siblings-in-law Marcus and Kelsey, for their friendship and support; and my grandparents Louis, Jean, and LaVerne, who were as proud of me as any grandparents can be.

I would also like to acknowledge the support of my friends and mentors: Brad and Sarah, for countless fun times spent together; Laura, for being a great friend through undergrad and grad school; Dr. Spivey and the other math faculty at Samford, for inspiring me to pursue math and mentoring me; and all my friends in the Duke math department, especially Sarah, Aaron, Shishi, Rachel, Kash, Tiffany, Miles, and Liz, for being a great community and knowing how to have fun.

# Introduction

This dissertation deals with the problem of estimating the intrinsic dimension of noisy, high-dimensional point clouds, as modern data sets often consist of samples taking values in a high-dimensional Euclidean space yet containing an underlying structure that is low-dimensional. Assuming one has samples from a set of intrinsic dimension  $k$  (meaning the set is locally well approximated by  $k$ -dimensional planes) embedded in  $\mathbb{R}^D$  with  $D \gg k$ , the goal is to estimate  $k$  given the samples, which are generally corrupted by high-dimensional noise.

Intrinsic dimension estimation is important in numerous applications, including statistics, economics, molecular dynamics, genomics, finance, and machine learning. Experts in these various areas commonly need to analyze extremely high-dimensional data sets with limited sample size. Many dimensionality reduction algorithms have been developed to compute low-dimensional representations of these high-dimensional data sets; the idea is to map the data from  $\mathbb{R}^D$  to  $\mathbb{R}^K$ , where again  $K \ll D$ , in such a way that the pairwise distances between the points are preserved. These low-dimensional representations are vital in exploratory data analysis, enabling the visualization of complicated data and the discovery of intrinsic patterns

and structure. However, all of these techniques require the user to specify  $K$ , the dimension of the Euclidean space into which the data will be mapped. Choosing  $K$  too small may result in the loss of significant information, whereas choosing  $K$  too large may obscure the underlying structure. Knowing the intrinsic dimension of the data enables the users of dimensionality reduction algorithms to select this parameter smartly and thus obtain meaningful low-dimensional representations of the data.

Accurately estimating the intrinsic dimension (ID) is thus extremely useful in data analysis, but there are many obstructions. When the data is nonlinear, classical linear methods for ID estimation such as principal component analysis (PCA) consistently overestimate the dimension. The presence of noise often obscures the low-dimensional structure, and small sample size may yield an incomplete picture of the data. To reliably estimate the intrinsic dimension in real-world data sets, it is thus critical to develop techniques robust to noise, small samples size, and curvature in the data. This dissertation introduces a multiscale, geometric approach to this problem, in which local dimension estimates are obtained by the analysis of the eigenvalues of multiscale covariance matrices. A new algorithm for intrinsic dimension estimation is also discussed, which requires a number of local samples essentially linear in the intrinsic dimension.

The results in this dissertation were obtained by joint work with M. Maggioni, who designed and implemented the MSVD algorithm in matlab and ran the numerical experiments described in Chap. 8. L. Rosasco also made important contributions to this research, including the compilation of results in Appendix C.

# 2

## Notation and Setting

The first section of this chapter defines the notation used throughout this dissertation; after clarifying the notation, the setting of interest is described. A multiscale version of PCA is proposed for estimating intrinsic dimension, and the constraints on the range of informative scales due to curvature, noise, and sample size are discussed.

### 2.1 Notation

Define the following random variables and related quantities:

- $X$  A random variable in  $\mathbb{R}^D$  with density  $\mu_X$ .
- $\mathcal{M}$  The support of  $\mu_X$ ; a special case is when  $\mathcal{M}$  is a compact Riemannian manifold.
- $N$  A random variable in  $\mathbb{R}^D$  representing noise; for example,  $N \sim \sigma\mathcal{N}(0, I_D)$ .
- $\tilde{X}$   $X + N$ ; a random variable in  $\mathbb{R}^D$  (noisy version of  $X$ ).

Fix a center  $z \in \mathbb{R}^D$  and scale  $r$ , and define the following random variables and related quantities:

- $B_z(r)$   $\{x \in \mathbb{R}^D : \|x - z\| \leq r\}$ ; closed Euclidean ball centered at  $z$  of radius  $r$ .

$X_{z,r}$	$[X \mid X \in B_z(r)]$ ; the random variable $X$ conditioned on taking values in $B_z(r)$ .
$\tilde{X}_{z,r}$	$[X + N \mid X \in B_z(r)]$ ; the random variable $X + N$ conditioned on $X$ taking values in $B_z(r)$ .
$\widetilde{X}_{z,r}$	$[X + N \mid X + N \in B_z(r)]$ ; the random variable $X + N$ conditioned on taking values in $B_z(r)$ .

Now consider drawing  $n$  samples  $x_1, \dots, x_n$  from the distribution of  $X$  and  $n$  samples  $\eta_1, \dots, \eta_n$  from the distribution of  $N$ ; define the following empirical quantities related to the above:

$X_n$	Denotes both the set $\{x_i\}_{i=1}^n$ of samples and the $n$ by $D$ matrix whose rows consist of the samples.
$N_n$	Denotes both the set $\{\eta_i\}_{i=1}^n$ and the associated matrix.
$\tilde{X}_n$	Denotes both the set $\{x_i + \eta_i\}_{i=1}^n$ and the associated matrix.
$X_{n,z,r}$	$X_n \cap B_z(r)$ ; elements are i.i.d. samples of $X_{z,r}$ .
$\tilde{X}_{n,z,r}$	$(X_n \cap B_z(r)) + N_{\{i: x_i \in B_z(r)\}}$ ; elements are i.i.d. samples of $\tilde{X}_{z,r}$ .
$\widetilde{X}_{n,z,r}$	$(X_n + N_n) \cap B_z(r)$ ; elements are i.i.d. samples of $\widetilde{X}_{z,r}$ .

Define the true and empirical covariance matrices of  $X$ :

$\text{cov}(X)$	$\mathbb{E}[(X - \mathbb{E}[X]) \otimes (X - \mathbb{E}[X])]$ ; the covariance matrix of $X$ .
$\text{cov}(X_n)$	$\frac{1}{n} \sum_{i=1}^n (x_i - \mathbb{E}_n[X]) \otimes (x_i - \mathbb{E}_n[X])$ , where $\mathbb{E}_n[X] = \frac{1}{n} \sum_{i=1}^n x_i$ ; the empirical covariance of $X$ given $n$ samples $x_1 \dots, x_n$ .

Define the following notation for the eigenvalues of the above:

$\lambda_i^2(\text{cov}(X))$	The eigenvalues of $\text{cov}(X)$ , sorted in decreasing order; equivalently, the squared singular values (SSV's) of $X$ .
$\Delta_i(\text{cov}(X))$	$\lambda_i^2(\text{cov}(X)) - \lambda_{i+1}^2(\text{cov}(X))$ for $i = 1, \dots, D - 1$ ; $\lambda_D^2(\text{cov}(X))$ for $i = D$ ; the gaps in the eigenvalues of the covariance.
$\Delta_{\max}$	$\max_{i=1, \dots, D} \Delta_i$ ; the largest gap in the eigenvalues.

$$\lambda_{i,z,r}^2 \quad \lambda_i^2(\text{cov}(X_{z,r})); \text{ the eigenvalues of } \text{cov}(X_{z,r}) \text{ (SSV's of } X_{z,r}\text{).}$$

$$\tilde{\lambda}_{i,z,r}^2 \quad \lambda_i^2(\text{cov}(\widetilde{X_{n,\tilde{z},r}})); \text{ the eigenvalues of } \text{cov}(\widetilde{X_{n,\tilde{z},r}}) \text{ (SSV's of } \widetilde{X_{n,\tilde{z},r}}\text{),}$$

where  $\tilde{z} = z + \eta_z$ , the noisy version of  $z$ .

## 2.2 Setting

Fix a center  $z \in \mathcal{M}$  and consider the following question: what is the dimension of  $\mathcal{M}$  at  $z$ ? What is meant by “intrinsic dimension” at  $z$  is more clearly defined in Chap. 5, but essentially “intrinsic dimension” at  $z$  is defined as the smallest integer  $k$  such that in a local neighborhood of  $z$ ,  $\mathcal{M}$  is well-approximated by a  $k$ -dimensional plane.

The classical technique in dimension estimation is principal component analysis (PCA). Given a random variable  $X \in \mathbb{R}^D$ , PCA estimates the intrinsic dimension by counting how many large eigenvalues  $\text{cov}(X)$  has; thus if  $\lambda_1^2 \geq \dots \geq \lambda_k^2 \gg \lambda_{k+1}^2 \geq \dots \geq \lambda_D^2$ , where  $\lambda_i^2 = \lambda_i^2(\text{cov}(X))$ , PCA estimates the dimension to be  $k$ . Unfortunately, this technique always fails on nonlinear data. Consider for example the 1-dimensional unit sphere  $\mathbb{S}^1$  embedded in any dimension  $\mathbb{R}^D$  via the natural embedding by the first two coordinates. If  $X$  is a random variable uniformly distributed on  $\mathbb{S}^1$ , then  $\text{cov}(X)$  has exactly two nonzero eigenvalues, each equal to  $\frac{1}{2}$ . Because  $\mathbb{S}^1$ , though one-dimensional, curves into a second dimension, global PCA fails to detect the one-dimensional structure. In fact, PCA may overestimate the dimension by an arbitrary amount; consider a 1-dimensional curve that spirals into more and more dimensions. Despite this shortcoming, PCA is attractive because of its simplicity and low sample size requirements: for a  $k$ -dimensional plane, only  $O(k \log k)$  samples are needed for  $\text{cov}(X_n)$  to be an accurate approximation of  $\text{cov}(X)$ .

In fact, the failure of PCA on nonlinear data is due to the fact that it is applied globally instead of locally. To obtain a local dimension estimate at a point  $z \in \mathcal{M}$ , one can simply perform PCA on  $X_{z,r}$ , the local neighborhood of  $z$  of radius  $r$ . For small  $r$  and smooth  $\mathcal{M}$ ,  $X_{z,r}$  is very nearly linear, and PCA will give the correct

dimension estimate as long as one has  $O(k \log k)$  samples in  $X_{z,r}$ .

This is a neat solution, but it raises a further question: what is the correct choice of  $r$ , the size of the local neighborhood? One encounters several constraints on an appropriate choice of  $r$ :

- **Curvature** If  $r$  is chosen too large, then the data will no longer appear linear, and PCA will overestimate the dimension because it also detects the dimensions into which the data curves. One must thus choose  $r$  small enough, as defined by some measure of the amount of curvature in the data.
- **Sample size** If  $r$  is chosen too small, however, then one could fail to have  $O(k \log k)$  samples in  $X_{z,r}$ , and PCA will underestimate the dimension due to lack of samples. Therefore  $r$  must be chosen large enough to ensure that there are at least  $O(k \log k)$  samples in  $X_{z,r}$ .
- **Noise** If  $r$  is chosen too small relative to the size of the noise, the  $k$ -dimensional structure will not be detectable. One cannot compute  $\lambda_{i,z,r}^2$ , but only the noisy  $\tilde{\lambda}_{i,z,r}^2$ , and so  $r$  must be chosen large enough so that the noise is not dominant.

Thus the simple solution of performing PCA locally is actually quite complex, as the question of how local is appropriate depends on factors that vary from data set to data set, and defining an algorithm that can select an appropriate scale given only the noisy samples  $\tilde{X}_n$  is non-trivial. As mentioned in Chap. 3, Fukunaga and Olsen (1971) develop the idea of using PCA locally to estimate intrinsic dimension, but without much success compared to competing volume-based techniques. However, the situation is made more tenable by the application of a multiscale, instead of a fixed scale, version of PCA. It is straight forward to show that for “nice” datasets, the  $\lambda_{i,z,r}^2$  which correspond to a vector lying in the tangent plane to  $\mathcal{M}$  at  $z$  will grow like  $r^2$ , while the  $\lambda_{i,z,r}^2$  which correspond to a direction normal to the tangent plane

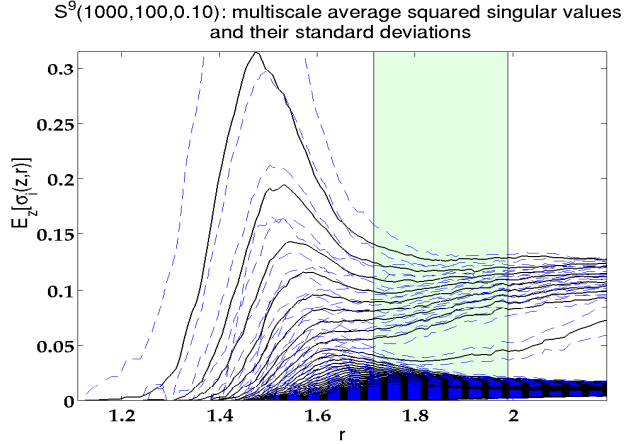


FIGURE 2.1: Plot of  $\mathbb{E}_z[\lambda_{i,z,r}^2]$  (the SSV's averaged over the samples), as a function of  $r$ , for 1000 noisy samples ( $\sigma = .1$ ) of  $\mathbb{S}^9$ ; the dotted lines indicate standard deviation.

and into which the data is curving, will grow like  $r^4$  (see the examples in Chap. 4). Furthermore, the remaining  $\lambda_{i,z,r}^2$ , which correspond to directions normal to all of  $\mathcal{M}$ , will be identically zero.

The  $\lambda_{i,z,r}^2$  are not computable, but for noise not too large, the  $\tilde{\lambda}_{i,z,r}^2$  will exhibit the same behavior: the  $\tilde{\lambda}_{i,z,r}^2$  due to curvature will grow quadratically with respect to the  $\tilde{\lambda}_{i,z,r}^2$  corresponding to an intrinsic dimension, which, assuming  $\mathcal{M}$  has dimension  $k$  at  $z$ , will cause  $\tilde{\lambda}_{k,z,r}^2 - \tilde{\lambda}_{k+1,z,r}^2$  to be the largest gap in the eigenvalues of the covariance for a large range of scales. The remaining  $\tilde{\lambda}_{i,z,r}^2$ , which are only nonzero because of noise, will level off to the noise variance  $\sigma^2$  after small scales. See Fig. 2.1 for an example of how the  $\tilde{\lambda}_{i,z,r}^2$  (averaged over  $z$ ) grow as a function of scale for  $\mathbb{S}^9$ . From this plot it seems clear that the top  $k = 9$  SSV's correspond to an intrinsic dimension,  $\tilde{\lambda}_{10,z,r}^2$  is due to curvature, and the remaining SSV's are due to noise. These observations are the foundation of the MSVD (multiscale singular value decomposition) algorithm for dimension estimation described in Chap. 8.

The remainder of this dissertation is organized as follows: Chap. 3 reviews relevant background in intrinsic dimension estimation, random matrix theory, and the

intersection of geometric measure theory with harmonic analysis. Chap. 4 calculates the multiscale SSV's  $\lambda_{i,z,r}^2$  for some important examples; these examples illustrate why the assumptions made in Chap. 5 are natural for many data sets. Chap. 5 states the main results; the main theorem addresses for which values of  $r$   $\tilde{\lambda}_{k,z,r}^2 - \tilde{\lambda}_{k+1,z,r}^2$  is the largest gap with high probability, that is, in which range of scales PCA on  $\widetilde{X_{n,\bar{z},r}}$  will correctly estimate the intrinsic dimension. The main result, given in Theorem 11, is a non-asymptotic statement, giving the probabilities exactly for a fixed  $k, n, D$ , although asymptotic statements for various regimes (e.g.  $n \rightarrow \infty$ ,  $D \rightarrow \infty$ , and both  $n, D \rightarrow \infty$  with  $\frac{n}{D} \rightarrow \gamma$  for some constant  $\gamma$ ) are derived as corollaries. Chap.'s 6 and 7 are dedicated to the proof of the main theorem. In Chap. 6, results are proved for a simplified model, in which  $z$  is known exactly, and local neighborhoods are taken before adding noise. Chap. 7 then shows that, up to a small change in scale, the SSV's of the simplified model are in fact close to those which are actually computable. Chap. 8 then describes the MSVD algorithm for dimension estimation in some detail, summarizes numerical experiments run on both manifold and real-world data sets, and compares the performance of the MSVD algorithm with that of competing algorithms.

# 3

## Background

This chapter reviews relevant background in intrinsic dimension estimation, the primary topic of this dissertation, and also in random matrix theory, as random matrix results are needed to analyze the behavior of various covariance matrices throughout this work. Furthermore, some results in the intersection of geometric measure theory with harmonic analysis are recounted, as these ideas inspired the multiscale, geometric approach taken in this dissertation.

### 3.1 Intrinsic Dimension Estimation

Intrinsic dimension (ID) estimation techniques fall into two categories: global and local. Global techniques are concerned with estimating the dimension of the entire data set (assuming that the data set has the same dimension throughout), while local techniques estimate the dimension of a data point from its local neighborhood (see Camastra and Vinciarelli (2002), who provide a nice review of ID estimation techniques). In this section, various notions of dimensionality are explored and dimension estimators discussed, including estimators based on the correlation integral, nearest neighbor estimators, estimators based on local PCA, and Bayesian estimators.

### 3.1.1 Notions of dimensionality

The question of how to define dimension of course arises, and various definitions are useful for different methods. Recall the definition of the  $d$ -dimensional Hausdorff measure of a set  $A \subseteq \mathbb{R}^D$ , which can be found in David and Semmes (1993):

$$H^d(A) = \lim_{\delta \rightarrow 0} \inf_{\substack{\cup_i E_i \supseteq A \\ \text{diam}(E_i) \leq \delta}} \left( \sum_i \text{diam}(E_i)^d \right).$$

The infimum is taken over all sequences of sets  $E_i \subseteq \mathbb{R}^D$  with diameter bounded by  $\delta$  whose union covers  $A$ ; if  $H^d(A) = C$  for some finite, positive constant, then  $A$  has Hausdorff dimension  $d$ .

The box-counting dimension is a simplified version of Hausdorff dimension; the box-counting dimension  $d_B$  of a data set  $\Omega$  is defined as:

$$d_B = \lim_{r \rightarrow 0} \frac{\ln(v(r))}{\ln(\frac{1}{r})},$$

where  $v(r)$  is the minimal number of boxes of size  $r$  needed to cover  $\Omega$ . In practice, however, the box-counting dimension can only be computed for low-dimensional sets because of the computational complexity, which is exponential in the dimension.

A good approximation to the box-counting dimension is the correlation dimension; if one has  $N$  i.i.d. samples  $x_1, \dots, x_N$  from some domain  $\Omega \subseteq \mathbb{R}^n$ , then the correlation integral is defined as:

$$C(r) := \lim_{N \rightarrow \infty} \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N \mathbf{1}_{(\|x_j - x_i\| \leq r)},$$

and the corresponding correlation dimension is:

$$d_C = \lim_{r \rightarrow 0} \frac{\ln(C(r))}{\ln(r)}.$$

### 3.1.2 Estimators based on the correlation integral

Estimating the above by plugging in a small  $r$  is sensitive to the choice of  $r$ ; one can, however, plot  $\ln(r)$  versus  $\ln(C(r))$  and estimate the slope of the linear portion of the plot. This is the approach used by Grassberger and Procaccia, see Grassberger and Procaccia (1983), known as the GP algorithm. To obtain the intrinsic dimension, however, one needs  $N > 10^{\frac{d_C}{2}}$  samples, which is prohibitive even for  $d_C$  small. Camastra and Vinciarelli (2002) develop a fractal method which is a modification of the GP algorithm and is designed to be more accurate on small sample size. Basically, they assume that the true dimension  $d$  can be determined from the relationship between the correlation dimension  $d_C$  and sample size  $N$  (a distinct reference curve is generated for each fixed sample size  $N$ ; for small  $N$ , there is a much larger difference between  $d$  and  $d_C$ ). Takens (1985) analyzes the convergence of the GP algorithm and suggests modifications which reduce the expected error.

Levina and Bickel (2005) estimate the ID of a data set by applying the maximum likelihood principle to the distances between data points. The number of samples landing in a ball around a data point  $x$  is modeled as a Poisson process.

Extending the above maximum likelihood technique, Haro et al. (2008) develop a translated Poisson mixture model for clustering sample points drawn from a stratification of manifolds and for estimating the density and dimension of each manifold in the stratification. The user must input the number of manifolds, the number of nearest neighbors to consider, the noise level, and a parameter regulating the trade off between spatial considerations and the log likelihood of the Poisson model. The model incorporates noise and non-uniform densities, but requires the user to specify a number of parameters.

### 3.1.3 Nearest neighbor estimators

Closely related to techniques based on the correlation integral are nearest neighbor techniques. All nearest neighbor methods are based on the following observation: given  $n$  samples  $X_n = \{x_1, \dots, x_n\}$  drawn according to some unknown density  $p(x)$  in  $\mathbb{R}^D$  that generates a subset of intrinsic dimension  $d$ ,

$$\frac{k}{n} \approx p(x) V_d R_k(x)^d,$$

where  $V_d$  is the volume of the  $d$ -dimensional unit sphere and  $R_k(x)$  is the distance of  $x$  to its  $k^{\text{th}}$  nearest neighbor (Levina and Bickel (2005)).

Pettis et al. (1979) show that  $\log(\mathbb{E}[\bar{R}_k])$  is approximately linear in  $\log k$  with slope  $\frac{1}{d}$ , where  $\bar{R}_k = \frac{1}{n} \sum_{i=1}^n R_k(x_i)$  is the average distance of a point to its  $k^{\text{th}}$  nearest neighbor. Thus using the sample average  $\bar{R}_k$  to approximate  $\mathbb{E}[\bar{R}_k]$ , an estimator of the intrinsic dimension  $d$  is calculated using linear regression.

More recently, Costa and Hero (2004) estimate the intrinsic dimension of a manifold using a global method based on minimal spanning trees of geodesic graphs (GMST). A similarity matrix based on the geodesic distances between all points is constructed and then a minimal spanning subgraph is computed; the intrinsic dimension is then estimated from the subgraph using a method-of-moments technique. More precisely, the  $k$ -nearest neighbor ( $k$ -NN) algorithm they propose involves the computation of the total edge length of  $k$ -NN graphs. Given  $n$  samples  $X_n = \{x_1, \dots, x_n\}$ , and defining  $\mathcal{N}_{k,i}$  to be the set of  $k$  nearest neighbors of  $x_i$ , the total edge length of the  $k$ -NN graph is then given by:

$$L_{\gamma,k}(X_n) := \sum_{i=1}^n \sum_{x \in \mathcal{N}_{k,i}} |x - x_i|^\gamma,$$

where  $\gamma$  is a parameter controlling locality. Assuming the samples are drawn from a bounded density supported on a compact,  $d$ -dimensional Riemannian sub-manifold

$(\mathcal{M}, g)$  of  $\mathbb{R}^D$ , Costa and Hero (2004) prove that with probability 1

$$\lim_{n \rightarrow \infty} \frac{L_{\gamma,k}(X_n)}{n^{\frac{d'-\gamma}{d'}}} = \begin{cases} \infty & d' < d \\ c & d' = d \\ 0 & d' > d \end{cases}$$

Thus

$$L_{\gamma,k}(X_n) = n^{\frac{d-\gamma}{d}} c + \epsilon_n,$$

where  $\epsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ , and so viewing  $L_{\gamma,k}(X_n)$  as a function of the sample size  $n$ , the dimension  $d$  is estimated using a non-linear least squares procedure. By applying this technique in local neighborhoods, Costa et al. (2005) extend the method to give local dimension estimates.

The  $k$ -NN algorithm exhibits a large negative bias due to under-sampling. For example, in high dimensions the  $d$ -dimensional unit cube appears  $d - 1$  dimensional, as almost all of the samples concentrate along the boundary. Carter et al. (2007) compensate for this negative bias by giving more weight to interior points, thus improving the performance of the  $k$ -NN algorithm. Carter and Hero (2008) further improve performance of local ID estimation by applying a post-processing of the  $k$ -NN algorithm, which forces the dimension to be constant in local neighborhoods, thus smoothing the neighborhoods and alleviating the negative impact of boundary points.

In all of these nearest neighbor techniques, the user must specify the number of nearest neighbors to consider, thus fixing the scale. The choice is often guided by  $n$ , but extensive experimentation may be required to optimize results, which tend to be quite sensitive to the choice of this parameter value.

### 3.1.4 Estimators based on local PCA

So far, all of the methods discussed have been volume-based techniques, i.e. methods based on counting the number of points landing in a ball around a sample point  $x$ .

Another class of estimators are based on computing the eigenvalues of sample covariance matrices. Given  $n$  mean-zero samples  $\{x_1, \dots, x_n\}$  in  $\mathbb{R}^D$ , define a (centered) data matrix whose rows constitute the samples:

$$X_n = \frac{1}{\sqrt{n}} \begin{bmatrix} -x_1- \\ -x_2- \\ \dots\dots \\ -x_n- \end{bmatrix}.$$

Defining  $C_n = X_n^T X_n$  (the  $D$  by  $D$  empirical covariance matrix), one then computes the eigenvalues of  $C_n$ :  $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_D^2$ , and the intrinsic dimension of the data is estimated by counting the number of “large” eigenvalues. This technique is known as principal component analysis (PCA), but when applied globally, PCA notoriously overestimates the intrinsic dimension of non-linear data. For example, the  $d$ -dimensional sphere appears  $d + 1$  dimensional.

Fukunaga and Olsen (1971) develop the idea of applying PCA locally: the data is divided into small subregions, and PCA is applied locally within each subregion to estimate the intrinsic dimension. The main difference between their method and the one presented here is that their estimates are based on a fixed scale (to be determined interactively), while the MSVD method examines the local squared singular values across a range of scales and estimates the dimension based on how the SSV’s grow as a function of scale. Bruske and Sommer (1998) reduce the computational complexity of the method from cubic to linear in the ambient dimension  $D$ , and improve its performance in the presence of noise.

Verveer and Duin (1995) compare the merits of the local PCA method of Fukunaga and Olsen (1971) with the nearest neighbor method of Pettis et al. (1979). They report that the nearest neighbor method underestimates the intrinsic dimension (especially when the intrinsic dimension is large), which is not the case for the local PCA method. However, they observed that the local PCA method was slower, required

a larger sample size, and was more sensitive to noise. They suggest improvements to both of these methods: for local PCA, they suggest a method of determining appropriate thresholds for the local eigenvalues, and for the nearest neighbor method, they suggest considering more than two neighbors for more accurate results. They acknowledge that for both of these methods choosing the correct scale is difficult. Contrary to the method of Fukunaga and Olsen (1971), the MSVD algorithm is very robust both to noise and small sample size, as the intrinsic dimension is determined from larger scales than were previously appropriate. Also, the algorithm determines the correct range of scales to consider independently; this task is not left to the user.

### *3.1.5 Estimators based on Bayesian methods*

Certain Bayesian methods have also been proposed to estimate the intrinsic dimension of data. Assuming that data points are sampled from a low-dimensional subspace of some high-dimensional Euclidean space, Chen et al. (2010b) model the data as a mixture of Gaussians, fitting the data to the mixture model using nonparametric Bayesian techniques. Their algorithm determines both the number of clusters in the model and the dimension of the clusters, which are all constrained to have the same rank. The idea is that a compact  $d$ -dimensional Riemannian manifold can be covered by a finite number of  $d$ -dimensional balls, and so they seek to cover the data using a finite number of  $d$ -dimensional Gaussian ellipsoids. The algorithm seeks to minimize both the number of clusters and their intrinsic dimension by adjusting the relevant posterior log-probabilities. The authors explore the compressive sensing applications of this method, in particular reconstructing signals on manifolds.

Chen et al. (2010a) apply this Bayesian framework to compute a low-dimensional embedding of the data. Again, the data is fit to a Gaussian mixture model, which is used to map the data into a low-dimensional space. Although pairwise distances between points from different components are not in general preserved, the map has

nice invertibility properties (as the data is essentially modeled as a collection of linear components), and each cluster in the mixture may have a different dimensionality, with the overall intrinsic dimension of the data estimated as the average intrinsic dimension of the clusters. One application the authors explore is the synthesis of time dependent data.

## 3.2 Results from Random Matrix Theory

This section reviews classic results in random matrix theory, which will be needed for precisely estimating various covariance matrices which arise in the PCA-based ID estimation process proposed in this work. Included are results on the largest eigenvalue of Wigner matrices, the extreme eigenvalues of covariance matrices, and the approximation of covariance matrices.

### 3.2.1 Largest eigenvalue of Wigner matrices

This section reviews classical results on the largest eigenvalue of both real and complex random symmetric matrices. Ledoux (2005) provides a summary of these results.

The Gaussian Orthogonal Ensemble of degree  $n$  ( $\text{GOE}_n$ ) may be defined as the real, symmetric,  $n$  by  $n$  matrices  $X$  satisfying  $X_{ij} \sim \mathcal{N}(0, \sigma^2)$  for  $i > j$  and  $X_{ii} \sim \mathcal{N}(0, 2\sigma^2)$ ; this definition gives rise to a probability measure  $P_n$  on the space of  $n$  by  $n$  real-valued matrices. The ensemble is called GOE because it is invariant under the action of orthogonal matrices, that is, if  $O$  is any orthogonal  $n$  by  $n$  matrix and  $\mathcal{A}$  is any measurable set of  $n$  by  $n$  real matrices, then  $P_n(\mathcal{A}) = P_n(O\mathcal{A}O^T)$ .

Similarly, the Gaussian Unitary Ensemble of degree  $n$  ( $\text{GUE}_n$ ) may be defined as the complex, Hermitian,  $n$  by  $n$  matrices  $X$  satisfying  $X_{ij} \sim \mathcal{N}(0, \frac{1}{2}\sigma^2) + i\mathcal{N}(0, \frac{1}{2}\sigma^2)$  for  $i > j$  and  $X_{ii} \sim \mathcal{N}(0, \sigma^2)$ . This gives rise to a probability measure  $P_n$  on the space of  $n$  by  $n$  complex-valued matrices which is invariant under the action of unitary matrices, that is,  $P_n(\mathcal{A}) = P_n(U\mathcal{A}U^*)$  for any unitary matrix  $U$  and measurable

set  $\mathcal{A}$ . More generally, if  $X$  has the same structure as the GOE or GUE, except its entries are drawn from a parent distribution other than Gaussian, it is called a Wigner matrix.

Wigner (1957) proves the semicircle law for both the GUE and GOE: with the normalization  $\sigma^2 = \frac{1}{4n}$ , the distribution of the eigenvalues approaches the semicircle on  $(-1, 1)$  as  $n \rightarrow \infty$ . This result can be extended to Wigner matrices satisfying the same moment assumptions. The eigenvalues having a semicircle distribution, however, does not necessarily imply anything about the convergence of the largest eigenvalue. Assuming the parent distribution has mean zero and finite fourth moment, again under the normalization  $\sigma^2 = \frac{1}{4n}$ , the largest eigenvalue of any Wigner matrix does in fact converge to 1 almost surely:

$$\lambda_{\max}(X) \rightarrow 1.$$

Furthermore, for  $X \in \text{GUE}_n$ , Forrester (1993) and Tracy and Widom (1994) establish the rate of convergence, showing that the distribution of  $n^{\frac{2}{3}}(\lambda_{\max}(X) - 1)$  converges to a Tracy-Widom distribution, which will be denoted  $F_{\text{GUE}}$ . Note that this result can be written as:

$$\frac{n\lambda_{\max}(X) - n}{n^{\frac{1}{3}}} \rightarrow F_{\text{GUE}},$$

that is, the deviation of  $n\lambda_{\max}(X)$  about its mean is  $O(\text{mean}^{\frac{1}{3}})$ . This is in contrast to Central Limit Theorem type results: if  $Y_i \sim \mathcal{N}(0, 1)$  for  $i = 1, \dots, n$  are i.i.d. and  $S_n = \sum_{i=1}^n Y_i$ , then the law of  $\frac{S_n - n}{n^{\frac{1}{2}}}$  converges to a standard normal distribution, that is, the deviation of  $S_n$  about its mean is  $O(\text{mean}^{\frac{1}{2}})$ . Similar results hold when  $X \in \text{GOE}_n$ .

### 3.2.2 Extreme eigenvalues of covariance matrices

This section reviews results regarding the extreme eigenvalues of covariance matrices. Let  $X$  be an  $n$  by  $D$  matrix whose entries are i.i.d. draws from some parent distribution. The first results dealing with the extreme eigenvalues of  $X^T X$  are derived for  $X$ 's having i.i.d. standard normal entries; in this special case,  $X^T X$  is called a null Wishart matrix.

Assuming  $\frac{n}{D} := \gamma \geq 1$ , Marčenko and Pastur (1967) extend the semicircle law to null Wishart matrices; in particular, they show that as  $n \rightarrow \infty$ , the distribution of the eigenvalues of  $X^T X/n$  converges almost surely to the density:

$$g(t) = \frac{\gamma}{2\pi t} \sqrt{(b-t)(t-a)} \quad \text{for } a \leq t \leq b, \quad \text{where } a = (1 - \gamma^{-\frac{1}{2}})^2, \quad b = (1 + \gamma^{-\frac{1}{2}})^2.$$

Observe that for  $\gamma$  small, the eigenvalues are more spread out, but as  $\gamma$  increases, the eigenvalues concentrate around 1.

Geman (1980) shows that the largest eigenvalue,  $\lambda_{\max}(\frac{1}{n}X^T X)$ , converges almost surely to the right edge of the “semicircle,” that is,

$$\lambda_{\max}(\frac{1}{n}X^T X) \rightarrow (1 + \gamma^{-\frac{1}{2}})^2.$$

Yin et al. (1988) furthermore show that this occurs not just for null Wisharts, but precisely when the parent distribution has mean zero, unit variance, and finite fourth moment. Similar results can be derived for the smallest eigenvalue. Silverstein (1985) shows that for null Wisharts,  $\lambda_{\min}(\frac{1}{n}X^T X)$  converges to the left edge of the “semicircle” almost surely:

$$\lambda_{\min}(\frac{1}{n}X^T X) \rightarrow (1 - \gamma^{-\frac{1}{2}})^2,$$

and Bai and Yin (1993) extend this to all distributions with mean zero, unit variance, and finite fourth moment.

However, the above tells us nothing about the rate of convergence of  $\lambda_{\max}(\frac{1}{n}X^T X)$  or  $\lambda_{\min}(\frac{1}{n}X^T X)$ . Johnstone (2001) shows that if one defines

$$\mu_{n,D} = (\sqrt{n-1} + \sqrt{D})^2, \quad \sigma_{n,D} = (\sqrt{n-1} + \sqrt{D}) \left( \frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{D}} \right)^{\frac{1}{3}},$$

then

$$\text{Law} \left( \frac{\lambda_{\max}(X^T X) - \mu_{n,D}}{\sigma_{n,D}} \right) \rightarrow F_{GOE},$$

where  $F_{GOE}$  is the Tracy-Widom distribution which appears as the limiting distribution of the Gaussian Orthogonal Ensemble. Observe that  $\mu_{n,D} = O(n)$ , while  $\sigma_{n,D} = O(n^{\frac{1}{3}})$ , so that, just as for Wigner matrices, the deviation of  $\lambda_{\max}(X^T X)$  from its mean is  $O(\text{mean}^{\frac{1}{3}})$ ; again, one has a tighter concentration around the mean than in the Central Limit Theorem.

All of the above are examples of asymptotic results; what is most useful in applications, however, are non-asymptotic statements; that is, for a fixed, finite  $n$  and  $D$ , what can be said about the eigenvalues and with what probability? Non-asymptotic random matrix theory results are built upon techniques developed in geometric functional analysis. For many applications, one cannot assume that  $X$  has i.i.d. entries; rather, one has  $n$  i.i.d. draws from some distribution in  $\mathbb{R}^D$  and  $X$  is the  $n$  by  $D$  matrix whose rows consist of the samples. In this section, two non-asymptotic results compiled by Vershynin (2010b) are recalled. The following definitions are needed:

**Definition 1.** *A random vector  $X \in \mathbb{R}^D$  is isotropic if  $\mathbb{E}[X \otimes X] = I_D$ . Equivalently,  $X$  has mean zero and identity covariance.*

**Definition 2.** *A random vector  $X \in \mathbb{R}^D$  is subgaussian if  $\langle X, \theta \rangle$  is a subgaussian random variable for every  $\theta \in \mathbb{S}^{D-1}$ ; furthermore,  $\|X\|_{\psi_2} = \sup_{\theta \in \mathbb{S}^{D-1}} \|\langle X, \theta \rangle\|_{\psi_2}$ . This is an extension of the definition of one-dimensional subgaussian random variables given in Appendix B.*

The first result is for random matrices  $X$  with subgaussian rows, Gaussian random matrices being a special case:

**Theorem 3.** *Let  $X$  be an  $n$  by  $D$  matrix whose rows are independent, subgaussian, isotropic random vectors in  $\mathbb{R}^D$ , and let  $\gamma = \frac{n}{D}$ . Then for any  $t \geq 0$ :*

$$\left(1 - C\gamma^{-\frac{1}{2}} - \frac{t}{\sqrt{n}}\right)^2 \leq \lambda_{\min}\left(\frac{1}{n}X^T X\right) \leq \lambda_{\max}\left(\frac{1}{n}X^T X\right) \leq \left(1 + C\gamma^{-\frac{1}{2}} + \frac{t}{\sqrt{n}}\right)^2$$

with probability at least  $1 - 2e^{-ct^2}$ , where  $C$  and  $c$  are constants depending only on the subgaussian norm of the rows.

Vershynin proves the above using a covering argument, concentration of measure results (see for example Talagrand (1995)), and Gordon's theorem for Gaussian matrices (Gordon (1984), Gordon (1985), Gordon (1992)).

The second result allows for the rows of  $X$  to have a heavy-tailed distribution.

**Theorem 4.** *Let  $X$  be an  $n$  by  $D$  matrix whose rows  $X_i$  are independent, subgaussian, isotropic random vectors in  $\mathbb{R}^D$  satisfying  $\|X_i\|_2 \leq M$  almost surely for all  $i$ . Then for any  $t \geq 0$ :*

$$\left(1 - \frac{tM}{\sqrt{n}}\right)^2 \leq \lambda_{\min}\left(\frac{1}{n}X^T X\right) \leq \lambda_{\max}\left(\frac{1}{n}X^T X\right) \leq \left(1 + \frac{tM}{\sqrt{n}}\right)^2$$

with probability at least  $1 - 2ne^{-ct^2}$ , where  $c$  is an absolute constant.

Vershynin proves the above using a non-commutative Bernstein inequality developed in Tropp (2010). Rudelson (1999) pioneers much of the work on the covariance matrices of heavy-tailed distributions, where only finite variance is assumed.

### 3.2.3 Approximation of covariance matrices

The results listed in Thm.'s 3 and 4 are relevant to an important question in statistics: how many samples  $n$  from a (centered) distribution in  $\mathbb{R}^D$  are needed before  $\frac{1}{n}X^T X$

(where again  $X$  is the  $n$  by  $D$  matrix whose rows consist of the samples) is close to the covariance matrix of the random vector? Rudelson (1999) shows that, under the minimal assumption of finite second moments, one needs  $n = O(D \log D)$ . In fact, for an arbitrary distribution one cannot do better than this; consider the random vector that is uniform on  $\{\sqrt{D}e_1, \dots, \sqrt{D}e_D\}$  and then appropriately centered, where  $\{e_1, \dots, e_D\}$  is the standard basis of  $\mathbb{R}^D$ . To obtain an accurate approximation, one must sample each of these vectors at least once, and the problem reduces to the coupon collector problem, thus requiring  $O(D \log D)$  samples.

However, when the random vector has a subgaussian distribution, Thm. 3 can be used to show that only  $n = O(D)$  samples are required. Furthermore, Adamczak et al. (2010) show that  $n = O(D)$  is also sufficient for subexponential random vectors. Vershynin (2010a) conjectures that this is in fact the case for any distribution with finite fourth moments, and he proves this up to an iterated log factor.

### 3.3 Intersection of Geometric Measure Theory & Harmonic Analysis

Many of the ideas presented in this dissertation were inspired by results in the intersection of geometric measure theory and harmonic analysis; this section highlights these connections. The following definitions, which can be found in David and Semmes (1993), are needed.

**Definition 5.** *A set  $E \in \mathbb{R}^D$  is regular with dimension  $k$  ( $k$ -regular) if it is closed and there exists a positive constant  $C$  such that for every  $x \in E$  and  $r > 0$ :*

$$C^{-1}r^k \leq H^k(E \cap B_x(r)) \leq Cr^k ,$$

where  $H^k$  is  $k$ -dimensional Hausdorff measure.

**Definition 6.** *A set  $E \in \mathbb{R}^D$  is rectifiable with dimension  $k$  ( $k$ -rectifiable) if it is contained (up to a set of measure zero) in the image of a countable union of Lipschitz*

mappings of  $\mathbb{R}^k$  into  $\mathbb{R}^D$ , that is

$$H^k \left( E \setminus \bigcup_i f_i(\mathbb{R}^k) \right) = 0,$$

where  $f_i : \mathbb{R}^k \rightarrow \mathbb{R}^D$  is Lipschitz, the  $\{f_i\}$  form a countable sequence, and  $H^k$  is  $k$ -dimensional Hausdorff measure.

When  $E$  is measurable and  $H^k(E) < \infty$ ,  $k$ -rectifiability means that for almost every  $x \in E$ , there is an approximate tangent  $k$ -plane to  $E$  at  $x$ . Rectifiability is a very useful and robust notion, but it is not quantifiable. In general, one would like some measure of how well  $E$  is locally approximated by  $k$ -planes.

The first connections between geometric measure theory and harmonic analysis are explored by Jones (1990), who poses the following questions: given a bounded set  $E \subseteq \mathbb{R}^2$ , is  $E$  contained in a rectifiable curve  $\Gamma$ ? ( $\Gamma$  is a rectifiable curve if it is the image of a finite interval under a single Lipschitz mapping.) If so, what is the shortest possible length of such a rectifiable curve? When  $E$  is finite, finding a curve  $\Gamma \supseteq E$  such that  $H^1(\Gamma)$  is minimal is equivalent to the Traveling Salesman Problem (TSP), which seeks to determine the shortest path visiting all points in a finite subset of  $\mathbb{R}^2$ .

To answer the above questions, Jones introduces a measure of how much  $E$  deviates from locally looking like a straight line. If  $Q$  is a dyadic square in  $\mathbb{R}^2$ , he defines:

$$\beta_E(Q) := \inf_L \left( \sup_{y \in E \cap 3Q} \frac{d(y, L)}{l(Q)} \right),$$

where the infimum is taken over all lines  $L$ ,  $d(y, L)$  is the distance from  $y$  to the line  $L$ , and  $l(Q)$  is the side-length of  $Q$ . Note that one always has  $0 \leq \beta_E(Q) \leq 3$ , so  $\beta_E(Q)$  is a scale-invariant measure of how close  $E$  is to a straight line in the vicinity of  $Q$  at scale  $l(Q)$ ; these numbers are called the Jones'  $\beta$ -numbers. Furthermore, the

following quantity is defined to measure how well  $E$  is approximated by lines across scales:

$$\beta^2(E) := \sum_Q \beta_E^2(Q)l(Q),$$

where the sum is taken over all dyadic squares. Jones proves that if  $\beta^2(E) < \infty$ , then  $E$  is contained in a rectifiable curve. Furthermore, the shortest rectifiable curve containing  $E$  has length equivalent to  $\text{diam}(E) + \beta^2(E)$ . Jones thus introduces a geometric approach to the TSP; Jones (1991) highlights the connections between the TSP and harmonic analysis.

These notions can be extended to higher dimensions by defining more general  $\beta$ -numbers. For a  $k$ -regular set  $E \in \mathbb{R}^D$  let

$$\beta_q(x, r) := \inf_P \left( r^{-k} \int_{E \cap B_x(r)} \left( \frac{d(y, P)}{r} \right)^q dy \right)^{\frac{1}{q}} \quad \text{for } 1 \leq q < \infty$$

$$\beta_q(x, r) := \inf_P \left( \sup_{y \in E \cap B_x(r)} \frac{d(y, P)}{r} \right) \quad \text{for } q = \infty$$

where the infimum is taken over all  $k$ -planes  $P$  and  $d(y, P)$  is the distance from  $y$  to the plane  $P$ .  $\beta_q(x, r)$  measures how well  $E$  is approximated (in  $L^q$ ) by a  $k$ -plane at location  $x$  and scale  $r$ ; when  $q = \infty$  and  $k = 1$ , one recovers the continuous version of the original Jones'  $\beta$ -numbers. If  $E$  is a  $k$ -regular set and there exists some constant  $C > 0$  such that

$$\int_0^R \int_{B_z(R)} \beta_1^2(x, r) \frac{dx dr}{r} \leq CR^k \quad (3.1)$$

for every  $z \in E$  and  $R > 0$ , then  $E$  is called uniformly rectifiable. This is a quantitative notion of rectifiability, quantified by the size of the constant  $C$ . David (1991) derives many equivalent conditions to (3.1) which highlight the connections of geometric measure theory with harmonic analysis.

The ideas presented in this work are related to the above: a range of scales (varying in space, that is, with  $z \in E$ ) at which the data is well approximated by a  $k$ -dimensional plane is determined. However, to compute  $\beta$ -numbers one must first specify a dimension; the multiscale quantities that are proposed here are computed without knowledge of the dimension for the purpose of inferring that dimension.

## Motivating Examples

Before stating and proving the main results, the multiscale squared singular values (SSV's) of some basic data sets are computed. These examples illustrate how the tangent singular values grow linearly with respect to scale, whereas the singular values due to curvature grow quadratically with respect to scale, an observation foundational to the MSVD intrinsic dimension estimator. First of all, the simplest case of a 1-dimensional curve is considered; then the SSV's of the  $k$ -dimensional sphere  $\mathbb{S}^k$  are computed, and their dependence on the intrinsic dimension  $k$  made explicit. Finally, the more general case of a co-dimension one, compact Riemannian manifold embedded in  $\mathbb{R}^D$  is considered.

### 4.1 Multiscale SSV's of a Curve

Let  $\mathcal{M}$  be a curve and let  $\mu_X$  be uniform measure on  $\mathcal{M}$ ; for simplicity suppose that the curve is the arc of some circle of radius  $R$ . Fix a center  $z \in \mathcal{M}$  and consider the SSV's of  $X_{z,r}$ . Let  $v_1$  be the unit tangent vector to  $\mathcal{M}$  at  $z$ ; this is the singular vector giving the tangent SSV  $\lambda_{1,z,r}^2$ , and let  $v_2$  be the unit normal vector to  $\mathcal{M}$  at  $z$ , which is the singular vector giving the curvature SSV  $\lambda_{2,z,r}^2$ .  $\lambda_{1,z,r}^2$  and  $\lambda_{2,z,r}^2$  are

simply the variances of the data in the  $v_1$  and  $v_2$  directions respectively.

Let  $d\theta$  be the angular measure of the osculating circle, and orient the circle so that  $\theta(z) = \frac{\pi}{2}$ ; let  $\theta_1$  and  $\theta_2$  be the endpoints of the arc  $\mathcal{M} \cap B_z(r)$ , and let  $l$  be the arclength. Since points are uniformly distributed on  $\mathcal{M}$ :

$$dp = \frac{R}{l} d\theta,$$

the probability measure of the random variable  $X_{z,r}$ . Let  $P_{v_1}(\theta)$  be the projection of a point at angle  $\theta$  onto  $v_1$ , and let  $\bar{v}_1 = \int_{\theta_1}^{\theta_2} P_{v_1}(\theta) dp$  be the mean of the projected points; similarly for  $P_{v_2}(\theta)$  and  $\bar{v}_2$ . One has:

$$\lambda_{1,z,r}^2 = \text{Var}[v_1 \text{ direction}] = \int_{\theta_1}^{\theta_2} (P_{v_1}(\theta) - \bar{v}_1)^2 \frac{R}{l} d\theta.$$

Since  $\bar{v}_1 = 0$  and  $P_{v_1}(\theta) = R \cos \theta$ :

$$\begin{aligned} \lambda_{1,z,r}^2 &= \int_{\theta_1}^{\theta_2} R^2 \cos^2 \theta \frac{R}{l} d\theta \\ &= \int_{\frac{\pi}{2} - \frac{l}{2R}}^{\frac{\pi}{2} + \frac{l}{2R}} R^2 \cos^2 \theta \frac{R}{l} d\theta \\ &= \frac{R^2(l - R \sin(\frac{l}{R}))}{2l}. \end{aligned}$$

Now consider  $\lambda_{2,z,r}^2$ .

$$\lambda_{2,z,r}^2 = \text{Var}[v_2 \text{ direction}] = \int_{\theta_1}^{\theta_2} (P_{v_2}(\theta) - \bar{v}_2)^2 \frac{R}{l} d\theta.$$

Now since  $P_{v_2}(\theta) = R(1 - \sin \theta)$ :

$$\begin{aligned} \bar{v}_2 &= \int_{\theta_1}^{\theta_2} R(1 - \sin \theta) \frac{R}{l} d\theta \\ &= \frac{R^2}{l} \int_{\frac{\pi}{2} - \frac{l}{2R}}^{\frac{\pi}{2} + \frac{l}{2R}} (1 - \sin \theta) d\theta. \end{aligned}$$

Thus:

$$\begin{aligned}
\lambda_{2,z,r}^2 &= \int_{\theta_1}^{\theta_2} (R(1 - \sin \theta) - \bar{v}_2)^2 \frac{R}{l} d\theta \\
&= \int_{\frac{\pi}{2} - \frac{l}{2R}}^{\frac{\pi}{2} + \frac{l}{2R}} (R(1 - \sin \theta) - \bar{v}_2)^2 \frac{R}{l} d\theta \\
&= \frac{R^2(l^2 - 4R^2 + 4R^2 \cos(\frac{l}{R}) + lR \sin(\frac{l}{R}))}{2l^2}.
\end{aligned}$$

Now  $l = 4R \arcsin(\frac{r}{2R})$ ; plugging this into the expressions for  $\lambda_{1,z,r}^2$  and  $\lambda_{2,z,r}^2$  and computing the Taylor series about  $r = 0$  yields:

$$\begin{aligned}
\lambda_{1,z,r}^2 &= \frac{r^2}{3} - \frac{7r^4}{180R^2} + O(r^6) \\
\lambda_{2,z,r}^2 &= \frac{r^4}{45R^2} - \frac{r^6}{1890R^4} + O(r^8).
\end{aligned} \tag{4.1}$$

If  $\mathcal{M}$  is the unit sphere  $\mathbb{S}^1$ , one obtains:

$$\begin{aligned}
\lambda_{1,z,r}^2 &= \frac{r^2}{3} - \frac{7r^4}{180} + O(r^6) \\
\lambda_{2,z,r}^2 &= \frac{r^4}{45} - \frac{r^6}{1890} + O(r^8).
\end{aligned} \tag{4.2}$$

## 4.2 Multiscale SSV's of the Sphere

### 4.2.1 Set-up

Let  $\mathbb{S}_R^k$  be the  $k$ -dimensional sphere of radius  $R$  embedded in  $\mathbb{R}^{k+1}$  via the natural embedding, with  $\mathbb{S}^k = \mathbb{S}_1^k$ . Let  $\mathcal{M} = \mathbb{S}^k$  and let the density  $\mu_X$  be uniform on  $\mathcal{M}$ . Fixing  $z$  to be the north pole, let  $V_r^k = \mathbb{S}^k \cap B_z(r)$  denote the spherical cap, so that  $X_{z,r}$  is a random variable uniformly distributed on  $V_r^k$ . The SSV's of  $X_{z,r}$ ,  $\{\lambda_{i,z,r}^2\}_{i=1}^{k+1}$ , are computed as a function of  $r$  and how the coefficients depend on  $k$  is determined. Throughout this section, let  $|\mathbb{S}_R^k|$  denote  $H^k(\mathbb{S}_R^k)$  and  $|V_r^k|$  denote  $H^k(V_r^k)$ , where  $H^k$  is  $k$ -dimensional Hausdorff measure on  $\mathbb{R}^{k+1}$  and Hausdorff measure is normalized so that  $H^k$  on  $\mathbb{R}^k$  corresponds to Lebesgue measure on  $\mathbb{R}^k$ .

Let  $\theta_0$  be the angle that the origin makes with  $z$  and any point on the boundary of  $V_r^k$ ; one has

$$\theta_0 = 2 \arcsin\left(\frac{r}{2}\right). \quad (4.3)$$

Non-normalized uniform measure on the cap is given by  $d\mathbb{S}^k = |\mathbb{S}_{\sin\theta}^{k-1}| d\theta$   
 $= |\mathbb{S}^{k-1}| \sin^{k-1} \theta d\theta$  (the latter equality valid for  $0 \leq \theta \leq \pi$ ), where  $\theta$  measures the angle formed with the  $x_{k+1}$ -axis. To obtain the density of  $X_{z,r}$ , normalize by the area of  $V_r^k$ :

$$d\mu_{X_{z,r}} = \frac{d\mathbb{S}^k}{|V_r^k|} = \frac{|\mathbb{S}_{\sin\theta}^{k-1}|}{|V_r^k|} d\theta,$$

where

$$\begin{aligned} |V_r^k| &= \int_0^{\theta_0} |\mathbb{S}_{\sin\theta}^{k-1}| d\theta \\ &= |\mathbb{S}^{k-1}| \int_0^{\theta_0} \sin^{k-1} \theta d\theta. \end{aligned}$$

### 4.2.2 Tangent SSV's

Let  $\lambda_{i,\theta_0}(k, R)$  be  $i$ -th singular value of the spherical cap of angle  $\theta_0$  of  $\mathbb{S}_R^{k+1}$ , so that  $\lambda_{i,z,r}^2 = \lambda_{i,\theta_0}^2(k, 1)$ . For  $1 \leq i \leq k$ ,  $\mathbb{E}[x_i] = 0$ . Thus:

$$\begin{aligned}
\lambda_{1,\theta_0}^2(k, 1) &= \frac{1}{|V_r^k|} \int_{V_r^k} x_1^2 d\mathbb{S}^k \\
&= \frac{1}{|V_r^k|} \int_0^{\theta_0} x_1^2 |\mathbb{S}_{\sin \theta}^{k-1}| d\theta. \\
&= \frac{1}{|V_r^k|} \int_0^{\theta_0} \left( \int_{\mathbb{S}_{\sin \theta}^{k-1}} x_1^2 d\mathbb{S}_{\sin \theta}^{k-1} \right) d\theta. \\
&= \frac{1}{|V_r^k|} \int_0^{\theta_0} \left( \frac{1}{|\mathbb{S}_{\sin \theta}^{k-1}|} \int_{\mathbb{S}_{\sin \theta}^{k-1}} x_1^2 d\mathbb{S}_{\sin \theta}^{k-1} \right) |\mathbb{S}_{\sin \theta}^{k-1}| d\theta. \\
&= \frac{|\mathbb{S}^{k-1}|}{|V_r^k|} \int_0^{\theta_0} \lambda_{1,\pi}^2(k-1, \sin \theta) \sin^{k-1} \theta d\theta. \\
&= \frac{|\mathbb{S}^{k-1}|}{|V_r^k|} \int_0^{\theta_0} \sin^2 \theta \lambda_{1,\pi}^2(k-1, 1) \sin^{k-1} \theta d\theta. \\
&= \frac{|\mathbb{S}^{k-1}|}{|V_r^k|} \left( \int_0^{\theta_0} \sin^{k+1} \theta d\theta \right) \lambda_{1,\pi}^2(k-1, 1) \\
&= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^{k-1}| \int_0^{\theta_0} \sin^{k-1} \theta d\theta} \left( \int_0^{\theta_0} \sin^{k+1} \theta d\theta \right) \lambda_{1,\pi}^2(k-1, 1) \\
&= \left( \frac{\int_0^{\theta_0} \sin^{k+1} \theta d\theta}{\int_0^{\theta_0} \sin^{k-1} \theta d\theta} \right) \lambda_{1,\pi}^2(k-1, 1).
\end{aligned}$$

Thus the singular values of the caps can be evaluated in terms of the singular values of the entire sphere by the following recursion formula:

$$\lambda_{1,\theta_0}^2(k, 1) = \left( \frac{\int_0^{\theta_0} \sin^{k+1} \theta d\theta}{\int_0^{\theta_0} \sin^{k-1} \theta d\theta} \right) \lambda_{1,\pi}^2(k-1, 1).$$

Note that this can also be written as

$$\lambda_{1,\theta_0}^2(k, 1) = \frac{|\mathbb{S}^{k-1}| \cdot |V_r^{k+2}|}{|\mathbb{S}^{k+1}| \cdot |V_r^k|} \lambda_{1,\pi}^2(k-1, 1).$$

By the above, note that  $\lambda_{1,\pi}^2(k-1, 1)$  must satisfy

$$\lambda_{1,\pi}^2(k, 1) = \frac{|\mathbb{S}^{k-1}| \cdot |\mathbb{S}^{k+2}|}{|\mathbb{S}^{k+1}| \cdot |\mathbb{S}^k|} \lambda_{1,\pi}^2(k-1, 1).$$

Now

$$\frac{|\mathbb{S}^{k-1}| \cdot |\mathbb{S}^{k+2}|}{|\mathbb{S}^{k+1}| \cdot |\mathbb{S}^k|} = \frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{k+3}{2})} = 1 - \frac{1}{k+1},$$

and so to satisfy the recursion  $\lambda_{1,\pi}^2(k, 1) = (1 - \frac{1}{k+1})\lambda_{1,\pi}^2(k-1, 1)$ , one must have

$$\lambda_{1,\pi}^2(k, 1) = \frac{1}{k+1}.$$

Thus:

$$\lambda_{1,z,r}^2 = \lambda_{1,\theta_0}^2(k, 1) = \frac{1}{k} \left( \frac{\int_0^{\theta_0} \sin^{k+1} \theta \, d\theta}{\int_0^{\theta_0} \sin^{k-1} \theta \, d\theta} \right). \quad (4.4)$$

By symmetry, all of the other tangent singular values are identical.

One can now obtain a Taylor expansion for  $\lambda_{1,z,r}^2 = t_2\theta_0^2 + t_4\theta_0^4 + O(\theta_0^6)$ :

$$\left( \int_0^{\theta_0} \sin^{k-1} \theta \, d\theta \right) (t_2\theta_0^2 + t_4\theta_0^4 + O(\theta_0^6)) = \frac{1}{k} \int_0^{\theta_0} \sin^{k+1} \theta \, d\theta. \quad (4.5)$$

Writing the Taylor series expansions for the above integrals about  $\theta_0 = 0$ :

$$\begin{aligned}
\int_0^{\theta_0} \sin^k \theta \, d\theta &= \int_0^{\theta_0} \left( \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} + \dots \right)^k d\theta \\
&= \int_0^{\theta_0} \theta^k + \binom{k}{1} \theta^{k-1} \left( -\frac{\theta^3}{3!} \right) + \binom{k}{2} \theta^{k-2} \left( -\frac{\theta^3}{3!} \right)^2 \\
&\quad + \binom{k}{1} \theta^{k-1} \left( -\frac{\theta^5}{5!} \right) + O(\theta^{k+6}) \, d\theta \\
&= \int_0^{\theta_0} \theta^k + a_k \theta^{k+2} + b_k \theta^{k+4} + O(\theta^{k+6}) \, d\theta \\
&= \frac{1}{k+1} \theta_0^{k+1} + \frac{a_k}{k+3} \theta_0^{k+3} + \frac{b_k}{k+5} \theta_0^{k+5} + O(\theta_0^{k+7})
\end{aligned} \tag{4.6}$$

where  $a_k = -\frac{k}{6}$  and  $b_k = \frac{k(k-1)}{72} + \frac{k}{120}$ .

Plugging the above expansions for the sine integrals into (4.5):

$$\begin{aligned}
&\left[ \frac{1}{k} \theta_0^k + \frac{a_{k-1}}{k+2} \theta_0^{k+2} + O(\theta_0^{k+4}) \right] (t_2 \theta_0^2 + t_4 \theta_0^4 + O(\theta_0^6)) \\
&= \left[ \frac{1}{k+2} \theta_0^{k+2} + \frac{a_{k+1}}{k+4} \theta_0^{k+4} + O(\theta_0^{k+6}) \right],
\end{aligned}$$

so that

$$\frac{1}{k^2} \theta_0^{2+k} + \left( \frac{t_4}{k} + \frac{t_2 a_{k-1}}{k+2} \right) \theta_0^{k+4} + O(\theta_0^{k+6}) = \frac{1}{k(k+2)} \theta_0^{k+2} + \frac{a_{k+1}}{k(k+4)} \theta_0^{k+4} + O(\theta_0^{k+6}).$$

Setting coefficients of  $\theta_0^{k+2}$  equal one obtains  $\frac{t_2}{k} = \frac{1}{k(k+2)}$ , so  $t_2 = \frac{1}{k+2}$ . Setting coefficients of  $\theta_0^{k+4}$  equal one obtains  $\frac{t_4}{k} + \frac{a_{k-1}}{(k+2)^2} = \frac{a_{k+1}}{k(k+4)}$ . Plugging in for  $a_k$  and solving gives  $t_4 = -\frac{2+k(k+6)}{3(k+2)^2(k+4)}$ . Thus:

$$\lambda_{1,z,r}^2 = \frac{1}{k+2} \theta_0^2 - \frac{2+k(k+6)}{3(k+2)^2(k+4)} \theta_0^4 + O(\theta_0^6).$$

In terms of the scale  $r$ , using (4.3), one obtains:

$$\lambda_{i,z,r}^2 = \frac{1}{k+2} r^2 + O(r^4), \quad i = 1, \dots, k. \tag{4.7}$$

### 4.2.3 Curvature SSV

One has:

$$\begin{aligned}
\lambda_{k+1,z,r}^2 &= \int_{V_r^k} x_{k+1}^2 d\mu_{X_{z,r}} - \mathbb{E}[x_{k+1}]^2 \\
&= \int_{V_r^k} (1 - x_1^2 - x_2^2 \cdots - x_k^2) d\mu_{X_{z,r}} - \mathbb{E}[x_{k+1}]^2 \\
&= 1 - \lambda_{1,z,r}^2 - \lambda_{2,z,r}^2 - \cdots - \lambda_{k,z,r}^2 - \mathbb{E}[x_{k+1}]^2 \\
&= 1 - k\lambda_{1,z,r}^2 - \mathbb{E}[x_{k+1}]^2.
\end{aligned}$$

Now

$$\begin{aligned}
\mathbb{E}[x_{k+1}] &= \int_{V_r^k} x_{k+1} d\mu_{X_{z,r}} \\
&= \frac{1}{|V_r^k|} \int_0^{\theta_0} x_{k+1} |\mathbb{S}^{k-1}| \sin^{k-1} \theta d\theta \\
&= \frac{|\mathbb{S}^{k-1}|}{|V_r^k|} \int_0^{\theta_0} \cos \theta \sin^{k-1} \theta d\theta \\
&= \frac{|\mathbb{S}^{k-1}|}{|V_r^k|} \int_0^{\sin \theta_0} u^{k-1} du \\
&= \frac{|\mathbb{S}^{k-1}|}{|V_r^k|} \left( \frac{1}{k} \sin^k \theta_0 \right) \\
&= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^{k-1}| \int_0^{\theta_0} \sin^{k-1} \theta d\theta} \left( \frac{1}{k} \sin^k \theta_0 \right) \\
&= \frac{1}{k} \left( \frac{\sin^k \theta_0}{\int_0^{\theta_0} \sin^{k-1} \theta d\theta} \right).
\end{aligned}$$

Plugging the above expression for  $\mathbb{E}[x_{k+1}]$  as well as the expansion for  $\lambda_{1,z,r}^2$  given in (4.4) into the equation  $\lambda_{k+1,z,r}^2 = 1 - k\lambda_{1,z,r}^2 - \mathbb{E}[x_{k+1}]^2$  gives

$$\lambda_{k+1,z,r}^2 = 1 - \frac{\int_0^{\theta_0} \sin^{k+1} \theta d\theta}{\int_0^{\theta_0} \sin^{k-1} \theta d\theta} - \frac{1}{k^2} \frac{\sin^{2k} \theta_0}{\left( \int_0^{\theta_0} \sin^{k-1} \theta d\theta \right)^2}. \quad (4.8)$$

Multiplying:

$$\begin{aligned} \left( \int_0^{\theta_0} \sin^{k-1} \theta \, d\theta \right)^2 \lambda_{k+1}^2 &= \left( \int_0^{\theta_0} \sin^{k-1} \theta \, d\theta \right)^2 \\ &\quad - \left( \int_0^{\theta_0} \sin^{k-1} \theta \, d\theta \right) \left( \int_0^{\theta_0} \sin^{k+1} \theta \, d\theta \right) - \frac{1}{k^2} \sin^{2k} \theta_0. \end{aligned} \quad (4.9)$$

Again, the above integrals are expanded about  $\theta_0 = 0$ . From (4.6), one gets:

$$\int_0^{\theta_0} \sin^k \theta \, d\theta = \frac{1}{k+1} \theta_0^{k+1} + \frac{a_k}{k+3} \theta_0^{k+3} + \frac{b_k}{k+5} \theta_0^{k+5} + O(\theta_0^{k+7}),$$

where  $a_k = -\frac{k}{6}$  and  $b_k = \frac{k(k-1)}{72} + \frac{k}{120}$ .

Now assume that  $\lambda_{k+1,z,r}^2 = c_4 \theta_0^4 + c_6 \theta_0^6 + O(\theta_0^8)$ , and let LHS (resp. RHS) denote the left (resp. right) hand side of (4.9). Plugging into (4.9):

$$\begin{aligned} \text{LHS} &= \left[ \frac{1}{k} \theta_0^k + \frac{a_{k-1}}{k+2} \theta_0^{k+2} + \frac{b_{k-1}}{k+4} \theta_0^{k+4} + O(\theta^{k+6}) \right]^2 \cdot (c_4 \theta_0^4 + c_6 \theta_0^6 + O(\theta_0^8)) \\ &= \left[ \frac{1}{k^2} \theta_0^{2k} + \frac{2a_{k-1}}{k(k+2)} \theta_0^{2k+2} + \left( \frac{2b_{k-1}}{k(k+4)} + \frac{a_{k-1}^2}{(k+2)^2} \right) \theta_0^{2k+4} + O(\theta_0^{2k+6}) \right] \\ &\quad \cdot (c_4 \theta_0^4 + c_6 \theta_0^6 + O(\theta_0^8)) \\ &= \frac{c_4}{k^2} \theta_0^{2k+4} + O(\theta_0^{2k+6}). \end{aligned}$$

$$\begin{aligned} \text{RHS} &= \left[ \frac{1}{k} \theta_0^k + \frac{a_{k-1}}{k+2} \theta_0^{k+2} + \frac{b_{k-1}}{k+4} \theta_0^{k+4} + O(\theta^{k+6}) \right]^2 \\ &\quad - \left[ \frac{1}{k} \theta_0^k + \frac{a_{k-1}}{k+2} \theta_0^{k+2} + \frac{b_{k-1}}{k+4} \theta_0^{k+4} + O(\theta^{k+6}) \right] \cdot \left[ \frac{1}{k+2} \theta_0^{k+2} + \frac{a_{k+1}}{k+4} \theta_0^{k+4} + O(\theta^{k+6}) \right] \\ &\quad - \frac{1}{k^2} (\theta_0 - \frac{\theta_0^3}{3!} + \frac{\theta_0^5}{5!} + O(\theta_0^7))^{2k} \end{aligned}$$

$$\begin{aligned}
&= \left[ \frac{1}{k^2} \theta_0^{2k} + \frac{2a_{k-1}}{k(k+2)} \theta_0^{2k+2} + \left( \frac{2b_{k-1}}{k(k+4)} + \frac{a_{k-1}^2}{(k+2)^2} \right) \theta_0^{2k+4} + O(\theta_0^{2k+6}) \right] \\
&\quad - \left[ \frac{1}{k(k+2)} \theta_0^{2k+2} + \left( \frac{a_{k+1}}{k(k+4)} + \frac{a_{k-1}}{(k+2)^2} \right) \theta_0^{2k+4} + O(\theta_0^{2k+6}) \right] \\
&\quad - \frac{1}{k^2} \left[ \theta_0^{2k} + \binom{2k}{1} \theta_0^{2k-1} \left( -\frac{\theta_0^3}{3!} \right) + \binom{2k}{2} \theta_0^{2k-2} \left( -\frac{\theta_0^3}{3!} \right)^2 + \binom{2k}{1} \theta_0^{2k-1} \left( -\frac{\theta_0^5}{5!} \right) + O(\theta_0^{2k+6}) \right] \\
&= \left[ \frac{2a_{k-1}}{k(k+2)} - \frac{1}{k(k+2)} + \frac{1}{3k} \right] \theta_0^{2k+2} \\
&\quad + \left[ \frac{a_{k-1}^2}{(k+2)^2} + \frac{2b_{k-1}}{k(k+4)} - \frac{a_{k+1}}{k(k+4)} - \frac{a_{k-1}}{(k+2)^2} - \frac{(2k-1)}{36k} - \frac{1}{60k} \right] \theta_0^{2k+4} + O(\theta_0^{2k+6}).
\end{aligned}$$

Now observe that  $\frac{2a_{k-1}}{k(k+2)} - \frac{1}{k(k+2)} + \frac{1}{3k} = -\frac{(k-1)}{3k(k+2)} - \frac{3}{3k(k+2)} + \frac{(k+2)}{3k(k+2)} = 0$ , as expected; so that the highest power on both sides is  $O(\theta_0^{2k+4})$ . Setting the coefficients equal to each other, one obtains:

$$\begin{aligned}
\frac{c_4}{k^2} &= \frac{a_{k-1}^2}{(k+2)^2} + \frac{2b_{k-1}}{k(k+4)} - \frac{a_{k+1}}{k(k+4)} - \frac{a_{k-1}}{(k+2)^2} - \frac{(2k-1)}{36k} - \frac{1}{60k} \\
c_4 &= \frac{k^2(k-1)^2}{36(k+2)^2} + \frac{2k}{k+4} \left( \frac{(k-1)(k-2)}{72} + \frac{(k-1)}{120} \right) \\
&\quad + \frac{k(k+1)}{6(k+4)} + \frac{k^2(k-1)}{6(k+2)^2} - \frac{k(2k-1)}{36} - \frac{k}{60}.
\end{aligned}$$

Simplifying:

$$c_4 = \frac{k}{(k+2)^2(k+4)}.$$

Thus

$$\lambda_{k+1,z,r}^2 = \frac{k}{(k+2)^2(k+4)} \theta_0^4 + O(\theta_0^6).$$

In terms of the scale  $r$ , using (4.3), one obtains

$$\lambda_{k+1,z,r}^2 = \frac{k}{(k+2)^2(k+4)} r^4 + O(r^6). \quad (4.10)$$

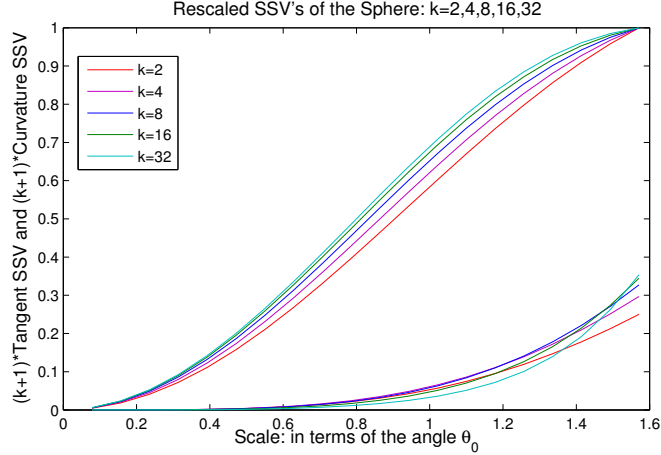


FIGURE 4.1: Plot of  $(k+1)\lambda_{1,z,r}^2 = (k+1)\lambda_{k,z,r}^2$  and  $(k+1)\lambda_{k+1,z,r}^2$  as a function of  $\theta_0 = 2 \arcsin(\frac{r}{2})$  for  $\mathbb{S}^k$ ,  $k = 2, 4, 8, 16, 32$ .

Observe that when  $k = 1$ , (4.7) and (4.10) give the same expansions for  $\mathbb{S}^1$  as those in (4.2). Fig. 4.1 shows a plot of both  $\lambda_{1,z,r}^2 = \lambda_{k,z,r}^2$  and  $\lambda_{k+1,z,r}^2$  (computed using (4.4) and (4.8)), the tangent and curvature multiscale SSV's of  $\mathbb{S}^k$ , scaled by  $(k+1)$  for  $k = 2, 4, 8, 16, 32$ . Observe that when rescaled, the SSV's for different  $k$  are remarkable similar, although the curvature SSV's for large  $k$  exhibit greater steepness. Also note that the  $k$ -th gap dominates until very large scales; the plot shows that the  $k$ -th gap is still the largest at  $\theta_0 = \frac{\pi}{2}$ , which corresponds to  $r = \sqrt{2}$ , independently of  $k$ .

### 4.3 Multiscale SSV's of Co-dimension One Manifold

Before computing the multiscale SSV's of a co-dimension one manifold, the second and fourth order moments of both  $\mathbb{S}^k$  and  $\mathbb{B}^k$  will be computed, as they will be needed in the manifold calculation.

#### 4.3.1 Moments of the unit sphere

Let  $X = (X_1, \dots, X_{k+1})$  be a random variable distributed uniformly on  $\mathbb{S}^k$ , and let  $\xi_i^2(\mathbb{S}^k) = \mathbb{E}[X_i^2]$  and  $\xi_{i,j}^4(\mathbb{S}^k) = \mathbb{E}[X_i^2 X_j^2]$  for  $i, j = 1, \dots, k+1$ . Consider slicing the

sphere perpendicular to the  $x_1$ -axis and let  $\theta$  be the angle (originating at the origin) that the  $x_1$ -axis makes with any point on the slice. One obtains:

$$d\mu_X = \frac{1}{|\mathbb{S}^k|} |\mathbb{S}_{\sin\theta}^{k-1}| d\theta,$$

where  $\mu_X$  is uniform measure on the sphere. By symmetry, it is sufficient to compute only  $\xi_1^2$ ,  $\xi_1^4$ , and  $\xi_{1,2}^4$ . One obtains:

$$\begin{aligned} \xi_1^2(\mathbb{S}^k) &= \frac{1}{|\mathbb{S}^k|} \int_0^\pi \cos^2 \theta |\mathbb{S}^{k-1}| \sin^{k-1} \theta d\theta \\ &= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \int_0^\pi \cos^2 \theta \sin^{k-1} \theta d\theta \\ &= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \left( \int_0^\pi \sin^{k-1} \theta d\theta - \int_0^\pi \sin^{k+1} \theta d\theta \right) \\ &= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \left( \frac{|\mathbb{S}^k|}{|\mathbb{S}^{k-1}|} - \frac{|\mathbb{S}^{k+2}|}{|\mathbb{S}^{k+1}|} \right) = 1 - \frac{|\mathbb{S}^{k-1}| \cdot |\mathbb{S}^{k+2}|}{|\mathbb{S}^k| \cdot |\mathbb{S}^{k+1}|} \\ &= 1 - \frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{k+3}{2})} = 1 - \frac{k}{k+1} = \frac{1}{k+1}. \end{aligned}$$

The fourth order moments are given by:

$$\begin{aligned}
\xi_{1,1}^4(\mathbb{S}^k) &= \frac{1}{|\mathbb{S}^k|} \int_0^\pi \cos^4 \theta |\mathbb{S}^{k-1}| \sin^{k-1} \theta \, d\theta \\
&= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \int_0^\pi (1 - 2 \sin^2 \theta + \sin^4 \theta) \sin^{k-1} \theta \, d\theta \\
&= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \left( \int_0^\pi \sin^{k-1} \theta \, d\theta - 2 \int_0^\pi \sin^{k+1} \theta \, d\theta + \int_0^\pi \sin^{k+3} \theta \, d\theta \right) \\
&= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \left( \frac{|\mathbb{S}^k|}{|\mathbb{S}^{k-1}|} - 2 \frac{|\mathbb{S}^{k+2}|}{|\mathbb{S}^{k+1}|} + \frac{|\mathbb{S}^{k+4}|}{|\mathbb{S}^{k+3}|} \right) \\
&= 1 - 2 \frac{|\mathbb{S}^{k-1}| \cdot |\mathbb{S}^{k+2}|}{|\mathbb{S}^k| \cdot |\mathbb{S}^{k+1}|} + \frac{|\mathbb{S}^{k-1}| \cdot |\mathbb{S}^{k+4}|}{|\mathbb{S}^k| \cdot |\mathbb{S}^{k+3}|} \\
&= 1 - 2 \frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{k+3}{2})} + \frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+4}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{k+5}{2})} \\
&= 1 - 2 \frac{k}{k+1} + \frac{k(k+2)}{(k+3)(k+1)} = \frac{3}{(k+1)(k+3)}.
\end{aligned}$$

To find  $\xi_{1,2}^4(\mathbb{S}^k)$ , again slice  $\mathbb{S}^k$  perpendicular to the  $x_1$ -axis, and find  $\mathbb{E}[X_2^2]$  within

each slice:

$$\begin{aligned}
\xi_{1,2}^4(\mathbb{S}^k) &= \frac{1}{|\mathbb{S}^k|} \int_0^\pi \cos^2 \theta \mathbb{E}[X_2^2 \text{ of } \mathbb{S}_{\sin \theta}^{k-1}] |\mathbb{S}^{k-1}| \sin^{k-1} \theta \, d\theta \\
&= \frac{1}{|\mathbb{S}^k|} \int_0^\pi \cos^2 \theta \sin^2 \theta \xi_2^2(\mathbb{S}^{k-1}) |\mathbb{S}^{k-1}| \sin^{k-1} \theta \, d\theta \\
&= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \int_0^\pi \cos^2 \theta \frac{\sin^2 \theta}{k} \sin^{k-1} \theta \, d\theta \\
&= \frac{1}{k} \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \int_0^\pi (\sin^2 \theta - \sin^4 \theta) \sin^{k-1} \theta \, d\theta \\
&= \frac{1}{k} \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \left( \int_0^\pi \sin^{k+1} \theta \, d\theta - \int_0^\pi \sin^{k+3} \theta \, d\theta \right) \\
&= \frac{1}{k} \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^k|} \left( \frac{|\mathbb{S}^{k+2}|}{|\mathbb{S}^{k+1}|} - \frac{|\mathbb{S}^{k+4}|}{|\mathbb{S}^{k+3}|} \right) \\
&= \frac{1}{k} \left( \frac{|\mathbb{S}^{k-1}| \cdot |\mathbb{S}^{k+2}|}{|\mathbb{S}^k| \cdot |\mathbb{S}^{k+1}|} - \frac{|\mathbb{S}^{k-1}| \cdot |\mathbb{S}^{k+4}|}{|\mathbb{S}^k| \cdot |\mathbb{S}^{k+3}|} \right) \\
&= \frac{1}{k} \left( \frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{k+3}{2})} - \frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+4}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{k+5}{2})} \right) \\
&= \frac{1}{k} \left( \frac{k}{k+1} - \frac{k(k+2)}{(k+1)(k+3)} \right) \\
&= \frac{1}{(k+1)(k+3)}.
\end{aligned}$$

In summary:

$$\xi_i^2(\mathbb{S}^k) = \frac{1}{k+1} \quad , \quad \xi_{i,j}^4(\mathbb{S}^k) = \frac{1+2\delta_{i,j}}{(k+1)(k+3)} \quad i, j = 1, \dots, k+1. \quad (4.11)$$

Observe that because uniform measure on the sphere is very similar to Gaussian measure, the above are approximately the moments of a Gaussian. If  $X_1, \dots, X_{k+1}$  are i.i.d.  $N(0, \sigma^2)$  with  $\sigma^2 = \frac{1}{k+1}$ , then one has:

$$\mathbb{E}[X_i^2] = \frac{1}{k+1} \quad , \quad \mathbb{E}[X_i^2 X_j^2] = \frac{1+2\delta_{i,j}}{(k+1)^2} \quad , \quad i, j = 1, \dots, k+1. \quad (4.12)$$

### 4.3.2 Moments of the unit ball

Let  $X = (X_1, \dots, X_k)$  be a random variable distributed uniformly on  $\mathbb{B}^k$ , and let  $\xi_i^2(\mathbb{B}^k)$  and  $\xi_{i,j}^4(\mathbb{B}^k)$  for  $i, j = 1, \dots, k$  denote the second and fourth order moments as before. In polar coordinates one has:

$$d\mu_X = \frac{1}{|\mathbb{B}^k|} |\mathbb{S}_r^{k-1}| dr.$$

The moments of the ball are easily calculated from the moments of the sphere as follows:

$$\begin{aligned} \xi_i^2(\mathbb{B}^k) &= \frac{1}{|\mathbb{B}^k|} \int_0^1 \mathbb{E}[X_i^2 \text{ of } \mathbb{S}_r^{k-1} | \mathbb{S}_r^{k-1}|] dr \\ &= \frac{1}{|\mathbb{B}^k|} \int_0^1 r^2 \xi_1^2(\mathbb{S}^{k-1}) |\mathbb{S}_r^{k-1}| dr \\ &= \frac{|\mathbb{S}^{k-1}|}{k \cdot |\mathbb{B}^k|} \int_0^1 r^{k+1} dr \\ &= \frac{1}{k+2}, \end{aligned}$$

since  $\frac{|\mathbb{S}^{k-1}|}{|\mathbb{B}^k|} = k$ . Similarly:

$$\begin{aligned} \xi_{i,j}^4(\mathbb{B}^k) &= \frac{1}{|\mathbb{B}^k|} \int_0^1 \mathbb{E}[X_i^2 X_j^2 \text{ of } \mathbb{S}_r^{k-1} | \mathbb{S}_r^{k-1}|] dr \\ &= \frac{1}{|\mathbb{B}^k|} \int_0^1 r^4 \xi_{i,j}^4(\mathbb{S}^{k-1}) |\mathbb{S}_r^{k-1}| dr \\ &= \frac{|\mathbb{S}^{k-1}|}{|\mathbb{B}^k|} \frac{1 + 2\delta_{i,j}}{k(k+2)} \int_0^1 r^{k+3} dr \\ &= \frac{1 + 2\delta_{i,j}}{(k+2)(k+4)}. \end{aligned}$$

In summary:

$$\xi_i^2(\mathbb{B}^k) = \frac{1}{k+2} \quad , \quad \xi_{i,j}^4(\mathbb{B}^k) = \frac{1 + 2\delta_{i,j}}{(k+2)(k+4)}, \quad i, j = 1, \dots, k. \quad (4.13)$$

### 4.3.3 Calculation for general co-dimension one manifold

Let  $\mathcal{M}$  be a  $k$ -dimensional manifold embedded in  $\mathbb{R}^{k+1}$ , that is,  $\mathcal{M}$  is a hypersurface in  $\mathbb{R}^{k+1}$ . Fix a point  $z \in \mathcal{M}$ . Let  $\kappa_1, \dots, \kappa_k$  be the principal curvatures of  $\mathcal{M}$  at  $z$ .

Then there exists a choice of coordinates  $(x_1, x_2, \dots, x_k, y)$ , centered at  $z$ , such that locally  $\mathcal{M}$  is given by  $y = f(x)$ , where

$$f(x) = \frac{1}{2}(\kappa_1 x_1^2 + \dots + \kappa_k x_k^2) + O(\|x\|^3), \quad (4.14)$$

that is, the second order Taylor expansion of  $f$  is quadratic with coefficients given by the principal curvatures (see Lee (1997)). The Jacobian of the map  $(x_1, \dots, x_k) \rightarrow (x_1, \dots, x_k, f(x))$  is the  $k + 1$  by  $k$  matrix:

$$J = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \kappa_1 x_1 & \kappa_2 x_2 & \dots & \kappa_k x_k \end{bmatrix}.$$

Thus the surface area element on the surface is given by

$$\begin{aligned} dA &= \sqrt{|\det(J^T J)|} dx_1 \dots dx_k = \sqrt{1 + \kappa_1^2 x_1^2 + \dots + \kappa_k^2 x_k^2} dx_1 \dots dx_k \\ &= \left( 1 + \frac{1}{2} \sum_{i=1}^k \kappa_i^2 x_i^2 + O(\|x\|^4) \right) dx_1 \dots dx_k. \end{aligned}$$

The multiscale SSV's  $\{\lambda_{i,z,r}^2\}_{i=1}^{k+1}$  of  $X_{0,r} = \mathcal{M} \cap B_z(r)$  are now computed (up to leading order terms). First of all, observe that

$$X_{0,r} = \{(x, f(x)) : \|(x, f(x))\| \leq r\}.$$

For small  $x$  and curvature not too large, this set is well approximated by the set

$$\hat{X}_{0,r} := \{(x, f(x)) : \|x\| \leq r\}.$$

A straightforward calculation shows that:

$$\hat{X}_{0,r\sqrt{1-\frac{1}{4}\kappa_{\max}^2 r^2}} \subseteq X_{0,r} \subseteq \hat{X}_{0,r\sqrt{1-\frac{1}{4}\kappa_{\min}^2 r^2}}. \quad (4.15)$$

Also using (4.13):

$$\begin{aligned} \text{vol}(\hat{X}_{0,r}) &= \int_{\mathbb{B}_r^k} \left( 1 + \frac{1}{2} \sum_{i=1}^k \kappa_i^2 x_i^2 \right) dx_1 \dots dx_k \\ &\leq \int_{\mathbb{B}_r^k} \left( 1 + \frac{\kappa_{\max}^2}{2} \sum_{i=1}^k x_i^2 \right) dx_1 \dots dx_k \\ &= |\mathbb{B}_r^k| + \frac{\kappa_{\max}^2}{2} k |\mathbb{B}_r^k| \xi_1^2(\mathbb{B}_r^k) \\ &= |\mathbb{B}_r^k| \left( 1 + \frac{k \xi_1^2(\mathbb{B}_r^k)}{2} \kappa_{\max}^2 r^2 \right) \\ &= |\mathbb{B}_r^k| \left( 1 + \frac{k}{2(k+2)} \kappa_{\max}^2 r^2 \right). \end{aligned} \quad (4.16)$$

Similarly:

$$\text{vol}(\hat{X}_{0,r}) \geq |\mathbb{B}_r^k| \left( 1 + \frac{k}{2(k+2)} \kappa_{\min}^2 r^2 \right). \quad (4.17)$$

Let  $r_{\kappa_{\min}} := r\sqrt{1 - \frac{1}{4}\kappa_{\min}^2 r^2}$  and  $r_{\kappa_{\max}} := r\sqrt{1 - \frac{1}{4}\kappa_{\max}^2 r^2}$ . Combining (4.15), (4.16),

and (4.17) gives:

$$\begin{aligned}
\text{vol}(X_{0,r}) &\leq \text{vol}(\hat{X}_{0,r\kappa_{\min}}) \leq |\mathbb{B}_r^k| \left(1 - \frac{1}{4}\kappa_{\min}^2 r^2\right)^{\frac{k}{2}} \left(1 + \frac{k}{2(k+2)}\kappa_{\max}^2 r^2 \left(1 - \frac{1}{4}\kappa_{\min}^2 r^2\right)\right) \\
&= |\mathbb{B}_r^k| \left(1 - \frac{k}{8}\kappa_{\min}^2 r^2 + O(r^4)\right) \left(1 + \frac{k}{2(k+2)}\kappa_{\max}^2 r^2 + O(r^4)\right) \\
&\leq |\mathbb{B}_r^k| \left(1 + \frac{k}{2(k+2)}\kappa_{\max}^2 r^2\right) \left(\frac{1 + \frac{k}{2(k+2)}\kappa_{\max}^2 r^2 - \frac{k}{8}\kappa_{\min}^2 r^2 + O(r^4)}{1 + \frac{k}{2(k+2)}\kappa_{\min}^2 r^2}\right) \\
&\leq |\mathbb{B}_r^k| \left(1 + \frac{k}{2(k+2)}\kappa_{\max}^2 r^2\right) (1 - O(r^2)) \\
&\leq |\mathbb{B}_r^k| \left(1 + \frac{k}{2(k+2)}\kappa_{\min}^2 r^2\right) (1 + O(r^2)) (1 - O(r^2)) \\
&\leq \text{vol}(\hat{X}_{0,r}) (1 \pm O(r^2)) .
\end{aligned}$$

Since  $\text{vol}(X_{0,r}) \leq \text{vol}(\hat{X}_{0,r})$ ,  $\frac{\text{vol}(X_{0,r})}{\text{vol}(\hat{X}_{0,r})} \leq 1 - O(r^2)$ . Similarly, one can show that  $\frac{\text{vol}(\hat{X}_{0,r})}{\text{vol}(X_{0,r})} \leq 1 + O(r^2)$ . Thus in the following computations,  $X_{0,r}$  is approximated by  $\hat{X}_{0,r}$ , as the leading order terms of the multiscale SSV's will not be affected;  $\text{vol}(X_{0,r})$  is also approximated by  $|\mathbb{B}_r^k|$ , as (4.16) and (4.17) show the leading order terms will be unaffected. Thus up to higher order terms, for  $i = 1, \dots, k$  using (4.13) one obtains:

$$\begin{aligned}
\lambda_{i,0,r}^2 &= \frac{1}{|\mathbb{B}_r^k|} \int_{\mathbb{B}_r^k} x_i^2 \left(1 + \frac{1}{2} \sum_{j=1}^k \kappa_j^2 x_j^2\right) dx_1 \dots dx_k \\
&= \xi_i^2(\mathbb{B}_r^k) + \frac{1}{2} \sum_{j=1}^k \kappa_j^2 \xi_{i,j}^4(\mathbb{B}_r^k) \\
&= \frac{r^2}{k+2} + \frac{1}{2} \sum_{j=1}^k \kappa_j^2 \frac{1 + 2\delta_{i,j}}{(k+2)(k+4)} r^4 = \frac{r^2}{k+2} + O(r^4) .
\end{aligned}$$

$\lambda_{k+1,0,r}^2$  is also computed using (4.13):

$$\begin{aligned}
\lambda_{k+1,0,r}^2 &= \frac{1}{|\mathbb{B}_r^k|} \int_{\mathbb{B}_r^k} \frac{1}{4} \left( \sum_{l=1}^k \kappa_l x_l^2 \right)^2 dA - \left( \frac{1}{|\mathbb{B}_r^k|} \int_{\mathbb{B}_r^k} \frac{1}{2} \left( \sum_{l=1}^k \kappa_l x_l^2 \right) dA \right)^2 \\
&= \frac{1}{4} \sum_{l=1}^k \sum_{j=1}^k \kappa_l \kappa_j (\xi_{l,j}^4(\mathbb{B}^k) - \xi_l^2(\mathbb{B}^k) \xi_j^2(\mathbb{B}^k)) r^4 + O(r^6) \\
&= \frac{1}{4} \left( \frac{1 + 2\delta_{l,j}}{(k+2)(k+4)} - \frac{1}{(k+2)^2} \right) \sum_{l=1}^k \sum_{j=1}^k \kappa_l \kappa_j + O(r^6) \\
&= \frac{1}{(k+2)^2(k+4)} \left( \frac{k+1}{2} \sum_{l=1}^k \kappa_l^2 - \sum_{l=1}^k \sum_{j<l}^k \kappa_l \kappa_j \right) r^4 + O(r^6).
\end{aligned}$$

In summary:

$$\begin{aligned}
\lambda_{i,0,r}^2 &= \frac{r^2}{k+2} + O(r^4) \quad \text{for } i = 1, \dots, k \\
\lambda_{k+1,0,r}^2 &= \frac{1}{(k+2)^2(k+4)} \left[ \frac{k+1}{2} \sum_{l=1}^k \kappa_l^2 - \sum_{l=1}^k \sum_{j<l}^k \kappa_l \kappa_j \right] r^4 + O(r^6).
\end{aligned} \tag{4.18}$$

Observe that in the case of  $\mathbb{S}^k$  centered at  $(0, 0, \dots, 1)$ , the Taylor expansion for points close to the origin is  $f(x) = \frac{1}{2}(x_1^2 + \dots + x_k^2)$ , so that  $\kappa_i = 1$  for  $i = 1, \dots, k$ , and (4.18) reduces to (4.7) and (4.10). Furthermore, observe that if  $\max_i \kappa_i = O(1)$  (all  $\kappa_i$  independent of  $k$ ), then  $\lambda_{k+1,0,r}^2 \lesssim \frac{1}{k} r^4$ , because the term in brackets in (4.18) is at most  $O(k^2)$ ; if in addition all the  $\kappa_i$  have the same magnitude and  $O(k)$  of them have the same sign, then the term in brackets is  $O(k)$ , and  $\lambda_{k+1,0,r}^2 \lesssim \frac{1}{k^2} r^4$ , as is the case for the sphere.

**Example 7.** If  $\mathcal{M}$  is a hyperbolic paraboloid with  $\kappa_1, \dots, \kappa_{k-1} = 1$  and  $\kappa_k = -1$ , then

$$\lambda_{k+1,0,r}^2(f) = \frac{(3k-2)}{(k+2)^2(k+4)} r^4 \sim \frac{3}{k^2} r^4.$$

If  $\mathcal{M}$  is a hyperbolic paraboloid with  $\kappa_1, \dots, \kappa_{\frac{k}{2}} = 1$  and  $\kappa_{\frac{k}{2}+1}, \dots, \kappa_k = -1$ , then

$$\lambda_{k+1,0,r}^2(f) = \frac{k(k+1)}{2(k+2)^2(k+4)} r^4 \sim \frac{1}{2k} r^4.$$

## Assumptions and Main Results

### 5.1 Assumptions

The examples in Chap. 4 motivate the following assumptions, which fall into two categories, assumptions on geometry and assumptions on the noise.

#### 5.1.1 Assumptions on geometry

Fix  $z \in \mathcal{M}$  and assume that  $\mathcal{M}$  has intrinsic dimension  $k$  at  $z$  in a certain range of scales. More precisely, assume that there exists a projection operator  $P^{(z,r)}$  onto a  $k$ -dimensional affine subspace such that if one defines

$$X_{z,r}^{\parallel} := P^{(z,r)}(X_{z,r}) \quad , \quad X_{z,r}^{\perp} := (I - P^{(z,r)})X_{z,r}$$

then the following assumptions hold for all  $r \in (R_{\min}, R_{\max})$  and some choice of positive parameters  $\lambda_{\min}, \lambda_{\max}, \delta, \kappa$ :

$$\lambda_i^2(\text{cov}(X_{z,r}^{\parallel})) \subseteq k^{-1}[\lambda_{\min}^2, \lambda_{\max}^2]r^2, \quad \text{with} \quad \max_{i < k} \Delta_i(\text{cov}(X_{z,r}^{\parallel})) \leq k^{-1}\delta^2r^2 \quad (5.1)$$

$$\|x^{\perp}\|^2 \leq k^{-1}\kappa^2r^4, \quad \forall x^{\perp} \in X_{z,r}^{\perp} \quad (\text{implying } \lambda_i^2(\text{cov}(X_{z,r}^{\perp})) \leq k^{-1}\kappa^2r^4).$$

Note that  $X_{z,r}^{\parallel}$  is an approximate tangent plane to  $\mathcal{M}$  at  $z$ , although it passes through  $\mathbb{E}[X_{z,r}]$  rather than  $z$ . At least for manifolds, as illustrated in Chap. 4, the eigenvalues of  $\text{cov}(X_{z,r}^{\parallel})$  scale like  $k^{-1}$ , which motivates the scaling above;  $\lambda_{\min}$  and  $\lambda_{\max}$  are parameters determined by the elongation of  $X_{z,r}$  when it is projected onto a local approximating plane. From (4.18), one sees that at least for the co-dimension one manifold case, the eigenvalues of  $\text{cov}(X_{z,r}^{\perp})$  scale at least like  $k^{-1}$ , and this motivates the scaling of the eigenvalues of  $\text{cov}(X_{z,r}^{\perp})$ ; the parameter  $\kappa$  measures the amount of (extrinsic) curvature in the data.

Furthermore, several assumption relating to volume growth are made. For any  $\tilde{z} \in \mathbb{R}^D$ , let  $v_{\tilde{z}}(\rho)$  be the function defined by

$$\mu_X(B_{\tilde{z}}(r)) := v_{\tilde{z}}(\rho) \mu_{\mathbb{R}^k}(\mathbb{B}^k) \rho^k, \text{ where } \rho = \sqrt{r^2 - d(\tilde{z}, \mathcal{M})^2};$$

here  $d(\tilde{z}, \mathcal{M})$  is the distance from  $\tilde{z}$  to a closest point (not necessarily unique) on  $\mathcal{M}$  and  $\mu_{\mathbb{R}^k}$  is  $k$ -dimensional Lebesgue measure. Assume that for all  $r \in (R_{\min}, R_{\max})$  and  $\tilde{z} \in \mathbb{R}^D$  satisfying  $\|\tilde{z} - z\| \leq R_{\max}$ , and some choice of positive parameters  $v_{\min}, v_{\max}$ :

$$\begin{aligned} v_{\tilde{z}}(r) &\in [v_{\min}, v_{\max}] \\ \frac{v_{\tilde{z}}(r+h)}{v_{\tilde{z}}(r)} &\leq \left(1 + \frac{h}{r}\right)^k \quad \forall h \in (0, r) \quad \text{and} \quad \frac{v_{\tilde{z}}(r)}{v_z(r)} \leq \left(1 + \frac{\|z - \tilde{z}\|}{r}\right) \end{aligned} \quad (5.2)$$

$$\sum_{i=1}^{\infty} e^{-i^2} \mu_X((B_z(r + (i+1)\phi)) \setminus B_z(r + i\phi)) \leq C_{k, \frac{\phi}{r}} r^k,$$

where  $C_{k, \frac{\phi}{r}}$  is a constant depending only on  $k$  and the ratio between the width of the annuli  $\phi$  and scale  $r$ . When  $R_{\min} = 0$  and  $\mu_X$  is distributed according to the natural volume measure of  $\mathcal{M}$ , that is,  $\mu_X(E) = H^k(\mathcal{M} \cap E)/H^k(\mathcal{M})$  where  $H^k$  is  $k$ -dimensional Hausdorff measure, the first condition implies that  $\mathcal{M}$ , at least locally, is a  $k$ -dimensional regular set, as defined in David and Semmes (1993). The second

condition guarantees smoothness of  $\mu_X(B_z(r))$  in both  $z$  and  $r$ . The last condition prevents  $\mathcal{M}$  from “wrapping around itself too tightly,” that is, from repeatedly coming close to self-intersection.

### 5.1.2 Assumptions on noise

The following assumptions on the distribution of the noise are made:

- (i)  $N$  has mean 0 and a radially symmetric distribution and is independent of  $X$ ;
- (ii)  $N$  has independent, variance  $\sigma^2$ , subgaussian coordinates with subgaussian moment bounded by  $\sigma$ .

### 5.1.3 Observations

The above assumptions are completely local. In particular, the parameters  $\lambda_{\min}$ ,  $\lambda_{\max}$ ,  $\delta$ ,  $\kappa$ ,  $v_{\min}$ ,  $v_{\max}$ ,  $R_{\min}$ ,  $R_{\max}$  may be chosen differently for different  $z \in \mathcal{M}$ .

$X_{z,r}^{\parallel}$  and  $X_{z,r}^{\perp}$  are not assumed to be independent and in general they are not. Furthermore, realizations of both  $X_{z,r}^{\parallel}$  and  $X_{z,r}^{\perp}$  are not assumed to have independent entries.

The noise observations  $\{\eta_i\}_{i=1}^n$  actually do not need to be i.i.d. draws: each  $\eta_i$  could be drawn from a distinct noise distribution  $N_i$ , as long as each  $N_i$  satisfies the assumptions in Sec. 5.1.2. Also, it is sufficient for the subgaussian moment to be bounded by  $C\sigma$  for some absolute constant  $C$ , as long as the slightly stronger volume growth condition  $\sum_{i=1}^{\infty} e^{-\left(\frac{i}{C}\right)^2} \mu_X((B_z(r + (i + 1)\phi)) \setminus B_z(r + i\phi)) \leq C_{k, \frac{\phi}{r}} r^k$  holds; only the absolute constants in the following results are affected.

For a fixed  $z$ , the above assumptions may be satisfied for different  $k$ 's in different scale ranges  $(R_{\min}, R_{\max})$ . This formulation makes the question of intrinsic dimension well-posed because it is allowed to depend on scale.

An important special case is compact,  $k$ -dimensional Riemannian manifolds isometrically embedded in  $\mathbb{R}^D$ ; for detecting intrinsic dimension at the finest scales,

one can always choose  $R_{\min} = 0$ .  $\mu_X$  is generally taken to be the natural volume measure of  $\mathcal{M}$ , normalized to be a probability measure, although  $\mu_X$  could also be a measure absolutely continuous with respect to the natural volume measure, with bounded Radon-Nykodym derivative.

One can also consider a collection of  $k$ -dimensional manifolds: the above assumptions will be satisfied as long as  $z$  is not close to the intersection of two or more manifolds. For  $z$  in a collection of  $s$  manifolds  $\mathcal{M}_1 \cup \mathcal{M}_2 \cup \dots \cup \mathcal{M}_s$  of different dimensionalities  $k_1, k_2, \dots, k_s$ , if  $z \in \mathcal{M}_i$  and  $z$  is not too close to an intersection, the above assumptions are satisfied with  $k = k_i$ . In fact, one could classify the points according to intrinsic dimension, thus learning the different manifolds.

## 5.2 Main Results

The main theorem gives a range of scales (dependent upon noise, curvature, and sample size) in which PCA on  $\widetilde{X_{n,\tilde{z},r}}$  will correctly estimate the intrinsic dimension  $k$ . As a first step,  $\tilde{X}_{n,z,r}$  is analyzed. Recall from Sec. 2.1 that:

$$\tilde{X}_{n,z,r} = (X_n \cap B_z(r)) + N_{\{i: x_i \in B_z(r)\}}, \quad \widetilde{X_{n,\tilde{z},r}} = (X_n + N_n) \cap B_{\tilde{z}}(r).$$

Thus  $\tilde{X}_{n,z,r}$  is obtained by first intersecting with a local ball and then adding noise, while  $\widetilde{X_{n,\tilde{z},r}}$  is obtained by first adding noise and then intersecting with a local ball.  $\tilde{X}_{n,z,r}$  is easier to analyze because the center  $z$  is a point on  $\mathcal{M}$ , its cardinality is independent of the noise, and the distribution of the samples is exactly  $X_{z,r} + N$ . However, this object is not computable: one does not have access to  $z$ , but only to the noisy version  $\tilde{z}$  which no longer lies on  $\mathcal{M}$ . Furthermore, even if an oracle disclosed  $z$ ,  $\tilde{X}_{n,z,r}$  would still not be computable, because one has no way of knowing which points were in  $B_z(r)$  before they were corrupted by noise.

What is computable is  $\widetilde{X_{n,\tilde{z},r}}$ ; one can always intersect the noisy samples with

local balls, centered at some noisy sample. This object is more difficult to analyze, however, because it has a noisy center  $\tilde{z}$  off of  $\mathcal{M}$ , its cardinality depends upon the noise, and its elements have a complicated distribution, which cannot be written as the sum of two independent distributions.

Thus as a first step, results are derived for  $\tilde{X}_{n,z,r}$ ; Prop. 8 gives the range of scales in which PCA on  $\tilde{X}_{n,z,r}$  will correctly estimate the dimensionality. It is then shown that, up to a small change in scale,  $\tilde{X}_{n,z,r}$  is in fact close to  $\widetilde{X_{n,\tilde{z},r}}$  with high probability; more precisely, Prop. 10 shows that  $\|\text{cov}(\widetilde{X_{n,\tilde{z},r}}) - \text{cov}(\tilde{X}_{n,z,r_{2\sigma}})\|$  is small, where  $r_{2\sigma} = r^2 - 2\sigma^2 D$ . One is thus able to use  $\tilde{X}_{n,z,r_{2\sigma}}$  as a simplified model of the computable quantity  $\widetilde{X_{n,\tilde{z},r}}$ . The main theorem, Thm. 11, is then proved by combining Prop. 8 with Prop. 10 to obtain a range of scales where PCA on  $\widetilde{X_{n,\tilde{z},r}}$  will correctly estimate the intrinsic dimension with high probability.

In the following results, let

$$\begin{aligned} \sigma_0 &:= \sigma\sqrt{D} \quad , \quad r_{2\sigma}^2 := r^2 - 2\sigma_0^2 \quad , \quad \xi := \frac{\sigma_0}{r} \\ \hat{R}_{\min} &:= R_{\min}(1 + C\xi) \quad , \quad \hat{R}_{\max} := R_{\max}(1 - C\xi) \end{aligned} \quad (5.3)$$

where  $C$  is some absolute constant. First, the simplified model is analyzed:

**Proposition 8** (Analysis of  $\text{cov}(\tilde{X}_{n,z,r_{2\sigma}})$ ). *Let the assumptions in Section 5.1 and the notation of 5.3 hold and let*

$$\epsilon = \epsilon_{t,r_{2\sigma}} := \frac{6t^2}{\lambda_{\max}} \sqrt{\frac{2k \log k}{\mu_X(B_z(r_{2\sigma}))n}}.$$

*Then for any  $z \in \mathcal{M}$  and  $r_{2\sigma} \in (R_{\min}, R_{\max})$ , if  $n \geq \frac{72t^4 k \log k}{\lambda_{\max}^2 \mu_X(B_z(r_{2\sigma}))}$  and  $r_{2\sigma}$  satisfies*

$$\frac{\sigma_0}{\sqrt{1 - \frac{\delta^2}{\lambda_{\min}^2} - \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon}} \left( \frac{\frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon}{\sqrt{1 - \frac{\delta^2}{\lambda_{\min}^2} - \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon}} \vee \frac{\sqrt{k}}{\lambda_{\min} \sqrt{D}} \right) \lesssim r_{2\sigma} \lesssim \frac{\lambda_{\min}}{\kappa} \left( \frac{\lambda_{\min}}{\lambda_{\max}} \right) \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon \right),$$

then  $\Delta_k(\text{cov}(\tilde{X}_{n,z,r_{2\sigma}}))$  is the largest gap of  $\text{cov}(\tilde{X}_{n,z,r_{2\sigma}})$  with probability  $1 - Ce^{-ct^2}$  for  $t \geq C$ , where  $c$  and  $C$  are universal constants. Furthermore, under the same conditions and with the same probability:

$$\begin{aligned} & \|\text{cov}(X_{z,r_{2\sigma}}) - \text{cov}(\tilde{X}_{n,z,r_{2\sigma}})\| \\ & \lesssim \frac{\kappa^2 r_{2\sigma}^4}{k} + \theta_\epsilon + \sigma^2 + \left( \frac{\lambda_{\max} \kappa r_{2\sigma}^3}{k} + v_\epsilon \right) \left[ \frac{\lambda_{\max} \kappa r_{2\sigma}^3 + kv_\epsilon}{\lambda_{\min}^2 r_{2\sigma}^2 - C(\kappa^2 r_{2\sigma}^4 + k\theta_\epsilon)} \wedge 1 \right] \end{aligned}$$

where

$$\begin{aligned} \theta_\epsilon &= \epsilon \left[ \frac{\lambda_{\max}^2 r_{2\sigma}^2}{k} + \sigma_0^2 \frac{\lambda_{\max}}{\sqrt{k}} \left( \frac{1}{\sqrt{D}} + \frac{\lambda_{\max}}{\sqrt{k}} \epsilon \right) \right] \\ v_\epsilon &= \epsilon \left[ \sigma_0 \frac{\lambda_{\max}^2 r_{2\sigma}}{k} + \sigma_0 \frac{\kappa r_{2\sigma}^2}{\sqrt{k}} \left( \frac{1}{\epsilon_{t,r_{2\sigma}} \sqrt{D}} \wedge \frac{\lambda_{\max}}{\sqrt{k}} \right) + \sigma_0^2 \frac{\lambda_{\max}}{\sqrt{Dk}} \right]. \end{aligned}$$

*Proof.* See Chapter 6. □

**Remark 9.** The above illustrates that one of course needs  $\delta < \lambda_{\min}$  to obtain a non-empty range; it is in fact sufficient (but non-optimal) under niceness assumptions on the noise, curvature, and sample size, to obtain a non-empty range for  $\lambda_{\max}^2 > \frac{1}{2} \lambda_{\min}^2$ . For example, assuming  $\lambda_{\min}^2 \geq \frac{2}{3} \lambda_{\max}^2$  (with  $\lesssim$  indicating the same constants as in Prop. 8), the range becomes:

$$\frac{\sigma_0}{\sqrt{\frac{1}{2} - \frac{3}{2} \epsilon_{t,r_{2\sigma}}}} \left( \frac{3}{2} \frac{\epsilon_{t,r_{2\sigma}}}{\sqrt{\frac{1}{2} - \frac{3}{2} \epsilon_{t,r_{2\sigma}}}} \vee \frac{\sqrt{k}}{\lambda_{\min} \sqrt{D}} \right) \lesssim r_{2\sigma} \lesssim \frac{\lambda_{\min}}{\kappa} \sqrt{\frac{2}{3}} \left( \frac{1}{2} - \frac{3}{2} \epsilon_{t,r_{2\sigma}} \right).$$

The following result shows that the simplified model  $\tilde{X}_{n,z,r_{2\sigma}}$  is a good approximation of the computable quantity  $\widetilde{X}_{n,\bar{z},r}$ .

**Proposition 10** (Recentering and Noise). *Using the notation in 5.3, assume*

$\sqrt{\frac{\log D}{D}} \leq \xi \leq \frac{1}{\sqrt{3}} - \tau$  for some  $\tau > 0$ ,  $n \geq t^2 / \mu_X(B_z(r_{2\sigma}))$ , and  $D \gtrsim k^3$ . Then for

$r_{2\sigma} \in (\hat{R}_{\min}, \hat{R}_{\max})$  and any  $1 \leq s^2 \leq \sqrt{D} \left(1 \wedge \frac{c}{\xi^2 k}\right)$ ,

$$\|\text{cov}(\widetilde{X_{n,\bar{z},r}}) - \text{cov}(\tilde{X}_{n,z,r_{2\sigma}})\| \lesssim s^2 \left( \beta \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n} \right) r^2,$$

with probability at least  $1 - Ce^{-cs^2 \min\{(\beta\mu_X(B_z(r_{2\sigma}))n)\vee 1, s^2\}} - Ce^{-ct^2}$ , where

$\beta = \frac{\xi k}{\sqrt{D}} \left(1 + s^2 \xi + (\sigma \sqrt{D} \vee 1) \sqrt{k \log \frac{1}{1-2\xi^2} + \log \frac{D}{\xi^2 k^2}}\right)$  and  $c$  and  $C$  are universal constants.

*Proof.* See Chapter 7. □

Finally, Prop. 8 and Prop. 10 are combined to yield the main result.

**Theorem 11** (Main Result). *Let the assumptions in Sec. 5.1 and the notation of (5.3) hold and suppose  $D \gtrsim k^3$ ,  $\xi \in \{0\} \cup [\sqrt{\frac{\log D}{D}}, \frac{1}{\sqrt{3}} - \tau]$  for some  $\tau > 0$ . Define:*

$$\epsilon_{t,r_{2\sigma}} := \frac{6t^2}{\lambda_{\max}} \sqrt{\frac{2k \log k}{\mu_X(B_z(r_{2\sigma}))n}}$$

$$\beta_s := \frac{s^2 \xi k^2}{\lambda_{\min}^2 \sqrt{D}} \left(1 + s^2 \xi + (\sigma_0 \vee 1) \sqrt{k \log \frac{1}{1-2\xi^2} + \log \frac{D}{\xi^2 k^2}}\right)$$

$$f = f(\lambda_{\min}, \lambda_{\max}, \delta, \epsilon_{t,r_{2\sigma}}, \beta_s, \xi) := 1 - \frac{\delta^2}{\lambda_{\min}^2} - \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon_{t,r_{2\sigma}} - C \left( \beta_s \vee \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon_{t,r_{2\sigma}}^2 \right).$$

Then for any  $z \in \mathcal{M}$ ,  $r_{2\sigma} \in (\hat{R}_{\min}, \hat{R}_{\max})$ ,  $1 \leq s^2 \leq \sqrt{D} \left(1 \wedge \frac{c}{\xi^2 k}\right)$ ,  $s \vee C \leq t$ ,  $\epsilon_{t,r_{2\sigma}} \in [0, 1]$ , if:

$$\frac{\sigma_0}{\sqrt{f}} \left( \frac{\lambda_{\max}^2 \epsilon_{t,r_{2\sigma}}}{\lambda_{\min}^2 \sqrt{f}} \vee \frac{\sqrt{k}}{\lambda_{\min} \sqrt{D}} \right) \lesssim r_{2\sigma} \lesssim \frac{\lambda_{\min}}{\kappa} \left( \frac{\lambda_{\min}}{\lambda_{\max}} \right) f, \quad (5.4)$$

then  $\Delta_k(\text{cov}(\widetilde{X_{n,\bar{z},r}}))$  is the largest gap of  $\text{cov}(\widetilde{X_{n,\bar{z},r}})$  with probability at least

$1 - Ce^{-ct^2} - Ce^{-c \min\{(\lambda_{\min}^2 k^{-1} \beta_s \mu_X(B_z(r_{2\sigma}))n)\vee s^2, s^4\}}$ , where  $c$  and  $C$  are universal constants.

*Proof.* The proof follows from modifying the proof of Prop. 8 to include one final perturbation, the perturbation given in Prop. 10, which is denoted as  $P_5$ . By Prop. 10, for  $s \leq t$ ,

$$P_5 \lesssim \left( \frac{\lambda_{\min}^2}{k} \beta_s \vee \frac{t^2}{\mu_X(B_z(r_{2\sigma}))n} \right) r_{2\sigma}^2 \leq \left( \beta_s \vee \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon_{t,r_{2\sigma}}^2 \right) \frac{\lambda_{\min}^2 r_{2\sigma}^2}{k},$$

where  $\beta_s := \frac{s^2 \xi k^2}{\lambda_{\min}^2 \sqrt{D}} \left( 1 + s^2 \xi + (\sigma_0 \vee 1) \sqrt{k \log \frac{1}{1-2\xi^2} + \log \frac{D}{\xi^2 k^2}} \right)$  and  $\epsilon_{t,r_{2\sigma}}$  is as defined in Thm. 11, with probability at least  $1 - C e^{-ct^2} - C e^{-c \min\{(\lambda_{\min}^2 k^{-1} \beta_s \mu_X(B_z(r_{2\sigma}))n) \vee s^2, s^4\}}$ . Modifying the proof of Prop. 8, one obtains that under the conditions of Prop.'s 8 and 10, if  $r_{2\sigma}$  satisfies:

$$\begin{aligned} & \frac{\lambda_{\min}^2 r_{2\sigma}^2}{k} \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon_{t,r_{2\sigma}} - C \left( \beta_s \vee \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon_{t,r_{2\sigma}}^2 \right) \right) \\ & \gtrsim \underbrace{\frac{\kappa^2 r_{2\sigma}^4}{k} + \frac{\lambda_{\max} \kappa r_{2\sigma}^3}{k}}_{\text{curvature}} + \underbrace{\frac{\lambda_{\max} r_{2\sigma}}{\sqrt{k}} \sigma_0 \frac{\epsilon_{t,r_{2\sigma}} \lambda_{\max}}{\sqrt{k}} + \sigma_0^2 \frac{\epsilon_{t,r_{2\sigma}}^2 \lambda_{\max}^2}{k}}_{\text{noise}} + \frac{\sigma_0^2}{D}, \end{aligned}$$

then  $\Delta_k(\text{cov}(\widetilde{X_{n,\tilde{z},r}})) = \Delta_{\max}(\text{cov}(\widetilde{X_{n,\tilde{z},r}}))$  with probability at least  $1 - C e^{-ct^2} - C e^{-c \min\{(\lambda_{\min}^2 k^{-1} \beta_s \mu_X(B_z(r_{2\sigma}))n) \vee s^2, s^4\}}$ . Solving the above for  $r_{2\sigma}$  completes the proof.  $\square$

In the above result, if  $X_{z,r}^{\parallel}$  were assumed to have a subexponential distribution with moment  $\lambda_{\max} r$ , by the results in Adamczak et al. (2010) the  $\log k$  factor in the definition of  $\epsilon_{t,r}$  could be removed.

It is insightful to consider what Thm. 11 gives in various asymptotic regimes. The first corollary considers a regime classic in statistics: the sample size  $n$  going to infinity.

**Corollary 12** ( $n \rightarrow \infty$ ). *Under the conditions of Thm. 11, if  $n$  is large enough and*

$r_{2\sigma}$  satisfies

$$\frac{\sigma_0}{\sqrt{1 - \frac{\delta^2}{\lambda_{\min}^2} - C\beta_s}} \left( \frac{\sqrt{k}}{\lambda_{\min}\sqrt{D}} \right) \lesssim r_{2\sigma} \lesssim \frac{\lambda_{\min}}{\kappa} \left( \frac{\lambda_{\min}}{\lambda_{\max}} \right) \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - C\beta_s \right), \quad (5.5)$$

then  $\Delta_k(\text{cov}(\widetilde{X_{n,\bar{z},r}}))$  is the largest gap of  $\text{cov}(\widetilde{X_{n,\bar{z},r}})$  with probability at least  $1 - Ce^{-cs^4}$ , for any  $1 \leq s^2 \leq \sqrt{D} \left( 1 \wedge \frac{c}{\xi^2 k} \right)$ .

*Proof.* Choose  $t^2 = n^\alpha$  for any  $\alpha < \frac{1}{2}$ ; then  $\epsilon_{t,r_{2\sigma}} \rightarrow 0$ ,  $Ce^{-ct^2} \rightarrow 0$ , and  $\min\{(\lambda_{\min}^2 k^{-1} \beta_s \mu_X(B_z(r_{2\sigma}))) \vee s^2, s^4\} = s^4$ .  $\square$

The second corollary considers the case when the sample size  $n$  is fixed but the ambient dimension  $D$  goes to infinity; this is a regime that is becoming increasingly relevant in modern data analysis, for example in genetics, where relatively few samples are available, but each sample has a very large dimension.

**Corollary 13** ( $D \rightarrow \infty$ ). *Under the conditions of Thm. 11, if  $D$  is large enough and  $r_{2\sigma}$  satisfies*

$$\frac{\sigma_0}{f} \left( \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \right) \epsilon_{t,r_{2\sigma}} \lesssim r_{2\sigma} \lesssim \frac{\lambda_{\min}}{\kappa} \left( \frac{\lambda_{\min}}{\lambda_{\max}} \right) f \quad (5.6)$$

for  $f = 1 - \frac{\delta^2}{\lambda_{\min}^2} - \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon_{t,r_{2\sigma}} \left( 1 + C \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \epsilon_{t,r_{2\sigma}} \right)$ , then  $\Delta_k(\text{cov}(\widetilde{X_{n,\bar{z},r}}))$  is the largest gap of  $\text{cov}(\widetilde{X_{n,\bar{z},r}})$  with probability at least  $1 - Ce^{-ct^2}$ , for any  $t \geq C$ .

*Proof.* Choose  $s = t$ .  $\square$

Finally, the third corollary considers the classic random matrix theory regime, in which both  $n$  and  $D$  go to infinity in a fixed ratio  $\gamma$ .

**Corollary 14** ( $n, D \rightarrow \infty, \frac{n}{D} \rightarrow \gamma$ ). *Under the conditions of Thm. 11, if  $n, D$  are large enough and  $r_{2\sigma}$  satisfies*

$$\frac{\sigma_0}{\sqrt{1 - \frac{\delta^2}{\lambda_{\min}^2}}} \left( \frac{\sqrt{k}}{\lambda_{\min}\sqrt{D}} \right) \lesssim r_{2\sigma} \lesssim \frac{\lambda_{\min}}{\kappa} \left( \frac{\lambda_{\min}}{\lambda_{\max}} \right) \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} \right), \quad (5.7)$$

then  $\Delta_k(\text{cov}(\widetilde{X_{n,\bar{z},r}}))$  is the largest gap of  $\text{cov}(\widetilde{X_{n,\bar{z},r}})$  with probability at least  $1 - Ce^{-cn^\alpha} - Ce^{-cD^\alpha}$  for any  $\alpha < \frac{1}{2}$ .

*Proof.* Choose  $t^2 = n^\alpha$ , any  $\alpha < \frac{1}{2}$ , and  $s^2 = D^{\frac{\alpha}{2}}$ ; then  $\epsilon_{t,r_{2\sigma}} \sim n^{\alpha-\frac{1}{2}}$  as  $n \rightarrow \infty$ ,  $\beta_s \sim D^{\alpha-\frac{1}{2}}$  as  $D \rightarrow \infty$ , and  $\min\{(\lambda_{\min}^2 k^{-1} \beta_s \mu_X(B_z(r_{2\sigma})))n \vee s^2, s^4\} \sim \min\{D^{\alpha-\frac{1}{2}}n \wedge D^{\frac{\alpha}{2}}, D^\alpha\} = D^\alpha$ .  $\square$

**Remark 15.** *Thm. 11 essentially says that if  $D$  is large,  $\delta$  is not too large, and  $\lambda_{\min}$  and  $\lambda_{\max}$  are close and order 1, then if  $n \gtrsim O\left(\frac{k \log k}{\mu_X(B_z(r_{2\sigma}))}\right)$ ,  $\Delta_k(\text{cov}(\widetilde{X_{n,\bar{z},r}}))$  is the largest gap w.h.p. for  $r$  in the range:*

$$\sigma\sqrt{D} \lesssim r \lesssim \frac{\lambda_{\min}}{\kappa}.$$

The upper bound  $\frac{\lambda_{\min}}{\kappa}$  is based on the curvature, while the lower bound  $\sigma\sqrt{D}$  is caused by the noise level. Because  $\mu_X(B_z(r_{2\sigma}))$  represents the fraction of samples that land in the local cell of interest, the sampling requirement is simply that one has at least  $O(k \log k)$  points in the local cell. Note that to obtain a non-empty range,  $\sigma$  must scale with  $\sqrt{D}$ .

# 6

## Analysis of Simplified Model

This chapter presents the proof of Prop. 8.

*Proof of Prop. 8.* It is shown that the eigenvalues of the “simplified model”  $\text{cov}(\tilde{X}_{n,z,r})$  are close to the eigenvalues of  $\text{cov}(X_{z,r})$ , and a range of scales in which PCA on  $\tilde{X}_{n,z,r}$  correctly estimates intrinsic dimension, that is, the range where  $\Delta_k(\text{cov}(\tilde{X}_{n,z,r}))$  is the largest gap in the eigenvalues with high probability, is computed.

Fix a center  $z \in \mathcal{M}$ ; all the quantities in this chapter depend on  $z$ , but for brevity it will be dropped from the notation. Before all else, condition upon the number of samples that land in  $X_r$ , and denote this quantity  $n_r$ ;  $n_r$  is the cardinality of both  $X_{n,r}$  and  $\tilde{X}_{n,r}$ , and  $\mathbb{E}[n_r] = \mu_X(B_z(r))n$ . At the end of the proof this conditioning is removed.

To avoid working in the ambient dimension as much as possible, the perturbation from  $\text{cov}(X_r)$  to  $\text{cov}(\tilde{X}_{n,r})$  is decomposed into a series of four perturbations  $\{\mathbf{P}_i\}_{i=1}^4$ :

$$\begin{aligned}
\text{cov}(X_r) &= \begin{bmatrix} \text{cov}(X_r^{\parallel}) & \text{cov}(X_r^{\parallel}, X_r^{\perp}) \\ \text{cov}(X_r^{\perp}, X_r^{\parallel}) & \text{cov}(X_r^{\perp}) \end{bmatrix} \xrightarrow[\text{Wielandt's Lemma}]{\mathbf{P}_1} \begin{bmatrix} \text{cov}(X_r^{\parallel}) & 0 \\ 0 & \text{cov}(X_r^{\perp}) \end{bmatrix} \\
&\xrightarrow[\text{Sampling}]{\mathbf{P}_2} \begin{bmatrix} \text{cov}(X_{n,r}^{\parallel}) & 0 \\ 0 & \text{cov}(X_{n,r}^{\perp}) \end{bmatrix} \xrightarrow[\text{Diagonal noise}]{\mathbf{P}_3} \begin{bmatrix} \text{cov}(\tilde{X}_{n,r}^{\parallel}) & 0 \\ 0 & \text{cov}(\tilde{X}_{n,r}^{\perp}) \end{bmatrix} \\
&\xrightarrow[\text{Wielandt's Lemma}]{\mathbf{P}_4} \begin{bmatrix} \text{cov}(\tilde{X}_{n,r}^{\parallel}) & \text{cov}(\tilde{X}_{n,r}^{\parallel}, \tilde{X}_{n,r}^{\perp}) \\ \text{cov}(\tilde{X}_{n,r}^{\perp}, \tilde{X}_{n,r}^{\parallel}) & \text{cov}(\tilde{X}_{n,r}^{\perp}) \end{bmatrix} = \text{cov}(\tilde{X}_{n,r}).
\end{aligned}$$

$\mathbf{P}_1$  is the diagonalization step,  $\mathbf{P}_2$  considers the effects of sampling,  $\mathbf{P}_3$  considers the effects of noise, and  $\mathbf{P}_4$  is the un-diagonalization step. Without loss of generality, it is assumed that  $P^{(r)} = \langle \{e_i\}_{i=1}^k \rangle$ , and with some abuse of notation  $\text{cov}(X_r^{\parallel})$  is viewed as a  $k \times k$  matrix instead of a  $D \times D$  matrix; similarly,  $\text{cov}(X_r^{\perp})$  is viewed as a  $d$  by  $d$  matrix, with  $d = D - k$ . The eigenvalues (sorted in decreasing order) of the above 5 matrices are denoted by:

$$\begin{aligned}
\{\lambda_1^2, \dots, \lambda_D^2\} &\xrightarrow{\mathbf{P}_1} \{(\lambda_1^{\parallel})^2, \dots, (\lambda_k^{\parallel})^2, (\lambda_{k+1}^{\perp})^2, \dots, (\lambda_D^{\perp})^2\} \\
&\xrightarrow{\mathbf{P}_2} \{(\lambda_{n,r,1}^{\parallel})^2, \dots, (\lambda_{n,r,k}^{\parallel})^2, (\lambda_{n,r,k+1}^{\perp})^2, \dots, (\lambda_{n,r,D}^{\perp})^2\} \\
&\xrightarrow{\mathbf{P}_3} \{(\tilde{\lambda}_{n,r,1}^{\parallel})^2, \dots, (\tilde{\lambda}_{n,r,k}^{\parallel})^2, (\tilde{\lambda}_{n,r,k+1}^{\perp})^2, \dots, (\tilde{\lambda}_{n,r,D}^{\perp})^2\} \\
&\xrightarrow{\mathbf{P}_4} \{\tilde{\lambda}_{n,r,1}^2, \dots, \tilde{\lambda}_{n,r,D}^2\}.
\end{aligned}$$

A note on notation: let  $\bar{X}_n := \sqrt{\text{cov}(X_n)}$ ; it is the samples centered by the empirical mean.

## 6.1 $\mathbf{P}_1$ : Geometric Cross-terms

Although the following result is unnecessary to prove Prop. 8, it is included to show that, at least for scales that are not too large, the assumptions which are made on

the spectra of  $\text{cov}(X_r^{\parallel})$  and  $\text{cov}(X_r^{\perp})$  are equivalent to assumptions on the spectrum of  $\text{cov}(X_r)$ .

**Lemma 16.** For  $r \in (R_{\min}, R_{\max})$ :

$$0 \leq \lambda_i^2 - (\lambda_i^{\parallel})^2 \leq \frac{\frac{\kappa^2}{k} \lambda_{\max}^2}{\lambda_{\min}^2 - \kappa^2 r^2} r^4 \wedge \frac{\lambda_{\max} \kappa}{k} r^3 \quad \text{for } 1 \leq i \leq k$$

$$0 \leq (\lambda_i^{\perp})^2 - \lambda_i^2 \leq \frac{\frac{\kappa^2}{k} \lambda_{\max}^2}{\lambda_{\min}^2 - \kappa^2 r^2} r^4 \wedge \frac{\lambda_{\max} \kappa}{k} r^3 \quad \text{for } k+1 \leq i \leq D.$$

The proof is a direct application of Wielandt's Lemma 24. Note that as long as  $r \lesssim \frac{\lambda_{\min}}{\kappa}$ , the perturbation is  $O(r^4)$ . For example, if  $r \leq \frac{\lambda_{\min}}{2\kappa}$ , then the perturbation is bounded by  $\frac{4}{3} \frac{\lambda_{\max}^2}{\lambda_{\min}^2} k^{-1} \kappa^2 r^4$ .

## 6.2 $\mathbf{P}_2$ : Tangent and Normal Sampling

By Theorem 40, one has on an event  $\Omega_{t_1}$  having high probability (see Table 6.1)

$$\|\text{cov}(X_{n,r}^{\parallel}) - \text{cov}(X_r^{\parallel})\| \leq \frac{\lambda_{\max}^2 r^2}{k} \sqrt{\frac{k \log k}{\lambda_{\max}^2 n_r} t_1} + \frac{r^2 \log k}{n_r} t_1^2 =: P_2^{\parallel},$$

for  $t_1 \geq C$ , some universal constant. In particular, on  $\Omega_{t_1}$

$$\frac{1}{\sqrt{n_r}} \|\bar{X}_{n,r}^{\parallel}\| \leq \sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}}.$$

By the boundedness assumption on  $X_{z,r}^{\perp}$ , one also has:

$$\|\text{cov}(X_{n,r}^{\perp})\| \leq \frac{4\kappa^2}{k} r^4, \quad \frac{1}{\sqrt{n_r}} \|\bar{X}_{n,r}^{\perp}\| \leq \frac{2\kappa}{\sqrt{k}} r^2.$$

## 6.3 $\mathbf{P}_3$ : Tangential and Normal Noise

Let  $N_{n,r}$  denote the noise vectors corresponding to points in  $X_{n,r}$ . In this section the noise is decomposed into tangent and normal components:  $N_{n,r} = \sigma N_{n,r}^{\parallel} + \sigma N_{n,r}^{\perp}$ , where  $N_{n,r}^{\parallel}$  is viewed as an  $n_r$  by  $k$  matrix and  $N_{n,r}^{\perp}$  is view as an  $n_r$  by  $D-k$  matrix.

Table 6.1: Events, their definitions, and lower bounds on their probabilities, as they are used in the proof of Prop. 8; here  $c$  and  $C$  are universal constants.

$\Omega$	Description	$\mathbb{P}$	Conditions	From
$\Omega_{t_1}$	$\ \text{cov}(X_{n,r}^{\parallel}) - \text{cov}(X_r^{\parallel})\ $ $\leq \frac{\lambda_{\max}^2 r^2}{k} \sqrt{\frac{k \log k}{\lambda_{\max}^2 n_r}} t_1 + \frac{r^2 \log k}{n_r} t_1^2$	$1 - 3e^{-ct_1^2}$	$t_1 \geq C$	40
$\Omega_{t_3}$	$\frac{1}{n_r} \ (\bar{X}_{n,r}^{\parallel})^T \bar{N}_{n,r}^{\parallel}\ $ $\leq \frac{1}{\sqrt{n_r}} \ (\bar{X}_{n,r}^{\parallel})^T\  \left( \sqrt{\frac{k}{n_r}} t_3 + \frac{\sqrt{k}}{n_r} t_3^2 \right)$	$1 - 4e^{-ct_3^2}$	$t_3 \geq C$	47,38
$\Omega_{t_4}$	$\ \text{cov}(N_{n,r}^{\parallel}) - I_k\ $ $\leq \sqrt{\frac{k}{n_r}} t_4 \left( 1 + \sqrt{\frac{k}{n_r}} t_4^3 \right)$	$1 - 4e^{-ct_4^2}$	$t_4 \geq C$	42
$\Omega_{t_5}$	$\frac{1}{n_r} \ (\bar{X}_{n,r}^{\perp})^T \bar{N}_{n,r}^{\perp}\ $ $\leq \frac{1}{\sqrt{n_r}} \ (\bar{X}_{n,r}^{\perp})^T\  \left( \sqrt{\frac{D}{n_r}} t_5 + \frac{\sqrt{D}}{n_r} t_5^2 \right)$	$1 - 4e^{-ct_5^2}$	$t_5 \geq C$	47,38
$\Omega_{t_{6,1}}$	$\ \text{cov}(N_{n,r}^{\perp}) - I_D\ $ $\leq \sqrt{\frac{D}{n_r}} t_{6,1} \left( 1 + \sqrt{\frac{D}{n_r}} t_{6,1}^3 \right)$	$1 - 4e^{-ct_{6,1}^2}$	$t_{6,1} \geq C$	42
$\Omega_{t_{6,2}}$	$\frac{1}{n_r} \ \bar{N}_{n,r}^{\perp}\ ^2$ $\leq 1 + \frac{D}{n_r} t_{6,2}^2 \left( 1 + \frac{t_{6,2}^2}{n_r} \right)$	$1 - 4e^{-ct_{6,1}^2}$	$t_{6,2} \geq C$	47,38
$\Omega_{t_7}$	$\frac{1}{n_r} \ (\bar{X}_{n,r}^{\parallel})^T \bar{N}_{n,r}^{\perp}\ $ $\leq \frac{1}{\sqrt{n_r}} \ (\bar{X}_{n,r}^{\parallel})^T\  \left( \frac{\sqrt{k} + \sqrt{D}}{\sqrt{n_r}} t_7 + \frac{\sqrt{D}}{n_r} t_7^2 \right)$	$1 - 4e^{-ct_7^2}$	$t_7 \geq C$	47,38
$\Omega_{t_8}$	$\frac{1}{n_r} \ (\bar{N}_{n,r}^{\parallel})^T \bar{X}_{n,r}^{\perp}\ $ $\leq \frac{1}{\sqrt{n_r}} \ (\bar{N}_{n,r}^{\parallel})^T\  \left( \frac{\sqrt{k} + \sqrt{D}}{\sqrt{n_r}} t_8 + \frac{\sqrt{D}}{n_r} t_8^2 \right)$	$1 - 4e^{-ct_8^2}$	$t_8 \geq C$	47, 38
$\Omega_{t_9}$	$\frac{1}{n_r} \ (\bar{N}_{n,r}^{\parallel})^T \bar{N}_{n,r}^{\perp}\ $ $\leq \left( \frac{\sqrt{k} + \sqrt{D}}{\sqrt{n_r}} \right) t_9^2 + \frac{\sqrt{kD}}{n_r} t_9^4$	$1 - ce^{-ct_9^2}$	$t_9 \geq C$	47
$\Omega_{t_{10}}$	$n_r \geq \frac{1}{2} \mu_X(B_z(r))n$	$1 - e^{-\frac{1}{8}t_{10}^2}$	$n \geq \frac{t_{10}^2}{\mu_X(B_z(r))}$	36

First consider the perturbation  $\text{cov}(X_{n,r}^{\parallel}) \rightarrow \text{cov}(X_{n,r}^{\parallel} + \sigma N_{n,r}^{\parallel})$ . Since  $\mathbb{E}[\text{cov}(N_{n,r}^{\parallel})] = I_k$ ,  $I_k$  is subtracted and the following quantity estimated:

$$\|\text{cov}(X_{n,r}^{\parallel} + \sigma N_{n,r}^{\parallel}) - \text{cov}(X_{n,r}^{\parallel}) - \sigma^2 I_k\| \leq \frac{2\sigma}{n_r} \|(\bar{X}_{n,r}^{\parallel})^T \bar{N}_{n,r}^{\parallel}\| + \sigma^2 \|\text{cov}(N_{n,r}^{\parallel}) - \text{cov}(N^{\parallel})\|.$$

On  $\Omega_{t_1}$ , by Proposition 38 and 47,

$$\frac{1}{n_r} \|(\bar{X}_{n,r}^{\parallel})^T \bar{N}_{n,r}^{\parallel}\| \leq \sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}} \left( \sqrt{\frac{k}{n_r}} t_3 + \frac{\sqrt{k}}{n_r} t_3^2 \right)$$

for  $t_3 \geq C$  on the high probability event  $\Omega_{t_1} \cap \Omega_{t_3}$ . Moreover, by Theorem 42

$$\|\text{cov}(N_{n,r}^{\parallel}) - I_k\| \leq \sqrt{\frac{k}{n_r}} t_4 \left( 1 + \sqrt{\frac{k}{n_r}} t_4^3 \right)$$

on an event  $\Omega_{t_4}$  of high probability, for  $t_4 \geq C$ . In particular, on  $\Omega_{t_4}$  one has

$$\frac{\|\bar{N}_{n,r}^{\parallel}\|}{\sqrt{n_r}} \leq \sqrt{1 + \sqrt{\frac{k}{n_r}} t_4 \left( 1 + \sqrt{\frac{k}{n_r}} t_4^3 \right)}. \text{ On } \Omega_{t_1} \cap \Omega_{t_3} \cap \Omega_{t_4}, \text{ one has}$$

$$\begin{aligned} & \|\text{cov}(X_{n,r}^{\parallel} + \sigma N_{n,r}^{\parallel}) - \text{cov}(X_{n,r}^{\parallel}) - \sigma^2 I_k\| \\ & \leq 2\sigma \sqrt{\frac{\lambda_{\max}^2 r^2}{k}} + P_2^{\parallel} \left( \sqrt{\frac{k}{n_r}} t_3 + \frac{\sqrt{k}}{n_r} t_3^2 \right) + \sigma^2 \sqrt{\frac{k}{n_r}} t_4 \left( 1 + \sqrt{\frac{k}{n_r}} t_4^3 \right) =: P_3^{\parallel}, \end{aligned}$$

and therefore

$$(\tilde{\lambda}_{n_r,i}^{\parallel})^2 \in (\lambda_{n_r,i}^{\parallel})^2 + \sigma^2 + [-P_3^{\parallel}, +P_3^{\parallel}].$$

Now consider the perturbation  $\text{cov}(X_{n,r}^{\perp}) \rightarrow \text{cov}(X_{n,r}^{\perp} + \sigma N_{n,r}^{\perp})$ . By Lemma 37 and Propositions 38 and 47

$$\frac{1}{n_r} \|\bar{X}_{n,r}^{\perp T} \bar{N}_{n,r}^{\perp}\| \leq \frac{2\kappa}{\sqrt{k}} r^2 \left( \sqrt{\frac{D}{n_r}} t_5 + \frac{\sqrt{D}}{n_r} t_5^2 \right)$$

for  $t_5 \geq C$ , on an event  $\Omega_{t_5}$  having high probability. When  $n_r \geq D$ , the following is used:

$$\|\text{cov}(X_{n,r}^{\perp} + \sigma N_{n,r}^{\perp}) - \text{cov}(X_{n,r}^{\perp}) - \sigma^2 I_{D-k}\| \leq \frac{2\sigma}{n_r} \|\bar{X}_{n,r}^{\perp T} \bar{N}_{n,r}^{\perp}\| + \sigma^2 \|\text{cov}(N_{n,r}^{\perp}) - I_{D-k}\|.$$

$\text{cov}(N_{n,r}^{\perp})$  is estimated by Theorem 42:

$$\|\text{cov}(N_{n,r}^{\perp}) - I_{D-k}\| \leq \sqrt{\frac{D}{n_r}} t_{6,1} \left( 1 + \sqrt{\frac{D}{n_r}} t_{6,1}^3 \right)$$

for  $t_{6,1} \geq C$ , on an event  $\Omega_{t_{6,1}}$  having high probability. Thus on  $\Omega_{t_2} \cap \Omega_{t_5} \cap \Omega_{t_{6,1}}$ :

$$\begin{aligned} & \|\text{cov}(X_{n,r}^{\perp} + \sigma N_{n,r}^{\perp}) - \text{cov}(X_{n,r}^{\perp}) - \sigma^2 I_{D-k}\| \\ & \leq \frac{4\sigma\kappa}{\sqrt{k}} r^2 \left( \sqrt{\frac{D}{n_r}} t_5 + \frac{\sqrt{D}}{n_r} t_5^2 \right) + \sigma^2 \sqrt{\frac{D}{n_r}} t_{6,1} \left( 1 + \sqrt{\frac{D}{n_r}} t_{6,1}^3 \right) \\ & =: P_3^{\perp} \cdot \mathbf{1}_{(n_r \geq D)} \end{aligned}$$

and therefore

$$(\tilde{\lambda}_{n_r,i}^\perp)^2 \in (\lambda_{n_r,i}^\perp)^2 + \sigma^2 + P_3^\perp \cdot \mathbf{1}_{(n_r \geq D)} \cdot [-1, 1].$$

When  $n_r < D$ , the following is used:

$$\|\text{cov}(X_{n_r}^\perp + \sigma N_{n_r}^\perp) - \text{cov}(X_{n_r}^\perp)\| \leq \frac{2\sigma}{n_r} \|\overline{X}_{n_r}^\perp{}^T \overline{N}_{n_r}^\perp\| + \sigma^2 \|\text{cov}(N_{n_r}^\perp)\|.$$

Now for  $t_{6,2} \geq C$ , on an event  $\Omega_{t_{6,2}}$  of high probability, one can bound  $\text{cov}(N_{n_r}^\perp)$  by

$$\|\text{cov}(N_{n_r}^\perp)\| \leq \left( \frac{1}{\sqrt{n_r}} \|\overline{N}_{n_r}^\perp\| \right)^2 \leq \left( 1 + \frac{D}{n_r} \right) t_{6,2}^2 + \frac{D}{n_r^2} t_{6,2}^4 = 1 + \frac{D}{n_r} t_{6,2}^2 \left( 1 + \frac{t_{6,2}^2}{n_r} \right).$$

Thus on  $\Omega_{t_2} \cap \Omega_{t_5} \cap \Omega_{t_{6,2}}$ :

$$\begin{aligned} & \|\text{cov}(X_{n_r}^\perp + \sigma N_{n_r}^\perp) - \text{cov}(X_{n_r}^\perp)\| \\ & \leq \frac{4\sigma\kappa}{\sqrt{k}} r^2 \left( \sqrt{\frac{D}{n_r}} t_5 + \frac{\sqrt{D}}{n_r} t_5^2 \right) + \sigma^2 \left( 1 + \frac{D}{n_r} t_{6,2}^2 \left( 1 + \frac{t_{6,2}^2}{n_r} \right) \right) \\ & =: P_3^\perp \cdot \mathbf{1}_{(n_r < D)} + \sigma^2 \end{aligned}$$

and therefore

$$(\tilde{\lambda}_{n_r,i}^\perp)^2 \in (\lambda_{n_r,i}^\perp)^2 + (P_3^\perp \cdot \mathbf{1}_{(n_r < D)} + \sigma^2) \cdot [-1, 1].$$

Letting  $\Omega_{t_6} = \Omega_{t_{6,1}} \mathbf{1}_{(n_r \geq D)} + \Omega_{t_{6,2}} \mathbf{1}_{(n_r < D)}$ , on the high probability event  $\Omega_{t_5} \cap \Omega_{t_6}$  one has that

$$(\tilde{\lambda}_{n_r,i}^\perp)^2 \in (\lambda_{n_r,i}^\perp)^2 + \sigma^2 \cdot \mathbf{1}_{(n_r \geq D)} + (P_3^\perp + \sigma^2 \cdot \mathbf{1}_{(n_r < D)}) \cdot [-1, 1].$$

#### 6.4 $\mathbf{P}_4$ : Noisy Cross-terms

Assuming that  $(\tilde{\lambda}_{n_r,k}^\parallel)^2 > (\tilde{\lambda}_{n_r,k+1}^\perp)^2$ , by Wielandt's Lemma 24,  $(\tilde{\lambda}_{n_r,i}^\parallel)^2 < \tilde{\lambda}_{n_r,i}^2$  for  $i = 1, \dots, k$  and  $(\tilde{\lambda}_{n_r,i}^\perp)^2 > \tilde{\lambda}_{n_r,i}^2$  for  $i = k+1, \dots, D$ . Moreover, again by Wielandt's

Table 6.2: Definition of the  $P_i$ 's, which are used to bound the  $\mathbf{P}_i$  perturbations.

Object	Definition
$P_2^{\parallel}$	$\frac{\lambda_{\max}^2 r^2}{k} \sqrt{\frac{k \log k}{\lambda_{\max}^2 n_r} t_1 + \frac{r^2 \log k}{n_r} t_1^2}$
$\sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}}$	$\frac{\lambda_{\max} r}{\sqrt{k}} \sqrt{1 + \sqrt{\frac{k \log k}{\lambda_{\max}^2 n_r} t_1 + \frac{k \log k}{\lambda_{\max}^2 n_r} t_1^2}}$
$P_3^{\parallel}$	$2\sigma \sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}} \left( \sqrt{\frac{k}{n_r}} t_3 + \frac{\sqrt{k}}{n_r} t_3^2 \right) + \sigma^2 \sqrt{\frac{k}{n_r}} t_4 \left( 1 + \sqrt{\frac{k}{n_r}} t_4^3 \right)$
$P_3^{\perp} \cdot \mathbf{1}_{(n_r \geq D)}$	$\frac{4\sigma\kappa}{\sqrt{k}} r^2 \left( \sqrt{\frac{D}{n_r}} t_5 + \frac{\sqrt{D}}{n_r} t_5^2 \right) + \sigma^2 \sqrt{\frac{D}{n_r}} t_{6,1} \left( 1 + \sqrt{\frac{D}{n_r}} t_{6,1}^3 \right)$
$P_3^{\perp} \cdot \mathbf{1}_{(n_r < D)}$	$\frac{4\sigma\kappa}{\sqrt{k}} r^2 \left( \sqrt{\frac{D}{n_r}} t_5 + \frac{\sqrt{D}}{n_r} t_5^2 \right) + \sigma^2 \frac{D}{n_r} t_{6,2}^2 \left( 1 + \frac{t_{6,2}^2}{n_r} \right)$
$P_4$	$\frac{2\kappa}{\sqrt{k}} r^2 \sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}} + \sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}} \sigma \left( \frac{\sqrt{k} + \sqrt{D}}{\sqrt{n_r}} t_7 + \frac{\sqrt{D}}{n_r} t_7^2 \right)$ $+ \frac{2\sigma\kappa}{\sqrt{k}} r^2 \min \left[ \sqrt{1 + \sqrt{\frac{k}{n_r}} t_4 \left( 1 + \sqrt{\frac{k}{n_r}} t_4^3 \right)}, \left( \frac{\sqrt{k} + \sqrt{D}}{\sqrt{n_r}} t_8 + \frac{\sqrt{D}}{n_r} t_8^2 \right) \right]$ $+ \sigma^2 \left( \left( \frac{\sqrt{k} + \sqrt{D}}{\sqrt{n_r}} \right) t_9^2 + \frac{\sqrt{kD}}{n_r} t_9^4 \right)$

lemma, the size of each perturbation is bounded by  $\|B\| \wedge \frac{\|B\|^2}{\Delta}$ , where  $\Delta = (\tilde{\lambda}_{n_r, k}^{\parallel})^2 - (\tilde{\lambda}_{n_r, k+1}^{\perp})^2$ , and

$$\begin{aligned} B &= \text{cov}(X_{n,r}^{\parallel} + \sigma N_{n,r}^{\parallel}, X_{n,r}^{\perp} + \sigma N_{n,r}^{\perp}) \\ &= \frac{1}{n_r} \bar{X}_{n,r}^{\parallel T} \bar{X}_{n,r}^{\perp} + \frac{\sigma}{n_r} \bar{X}_{n,r}^{\parallel T} \bar{N}_{n,r}^{\perp} + \frac{\sigma^2}{n_r} \bar{N}_{n,r}^{\parallel T} \bar{N}_{n,r}^{\perp} + \frac{\sigma}{n_r} \bar{N}_{n,r}^{\parallel T} \bar{X}_{n,r}^{\perp}. \end{aligned}$$

Since  $\bar{X}_{n,r}^{\parallel}$  and  $\bar{X}_{n,r}^{\perp}$  are not necessarily independent, on  $\Omega_{t_1}$  the trivial bound

$$\left\| \frac{1}{n_r} \bar{X}_{n,r}^{\parallel T} \bar{X}_{n,r}^{\perp} \right\| \leq \frac{2\kappa}{\sqrt{k}} r^2 \sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}}$$

is used, which holds w.h.p.; by Proposition 38 and 47, on  $\Omega_{t_1}$

$$\frac{1}{n_r} \left\| \bar{X}_{n,r}^{\parallel T} \bar{N}_{n,r}^{\perp} \right\| \leq \sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}} \left( \left( \sqrt{\frac{k}{n_r}} + \sqrt{\frac{D}{n_r}} \right) t_7 + \frac{\sqrt{D}}{n_r} t_7^2 \right)$$

for  $t_7 \geq C$  on an event  $\Omega_{t_7}$  of high probability. When  $D \gg n_r$ , the following bound is used on  $\Omega_{t_4}$ :

$$\frac{1}{n_r} \left\| \bar{N}_{n,r}^{\parallel T} \bar{X}_{n,r}^{\perp} \right\| \leq \frac{2\kappa}{\sqrt{k}} r^2 \sqrt{1 + \sqrt{\frac{k}{n_r}} t_4 \left( 1 + \sqrt{\frac{k}{n_r}} t_4^3 \right)}.$$

When  $n_r \gg D$ , by Propositions 38 and 47, on  $\Omega_{t_2}$

$$\frac{1}{n_r} \left\| \overline{N}_{n,r}^{\parallel} \text{ }^T \overline{X}_{n,r}^{\perp} \right\| \leq \left( \left( \sqrt{\frac{k}{n_r}} + \sqrt{\frac{D}{n_r}} \right) t_8 + \frac{\sqrt{D}}{n_r} t_8^2 \right) \frac{2\kappa}{\sqrt{k}} r^2$$

for  $t_8 \geq C$ , on an event  $\Omega_{t_8}$  having high probability. Finally, by Lemma 37 and Propositions 38 and 45

$$\frac{1}{n_r} \left\| \overline{N}_{n,r}^{\parallel} \text{ }^T \overline{N}_{n,r}^{\perp} \right\| \leq \left( \sqrt{\frac{k}{n_r}} + \sqrt{\frac{D}{n_r}} \right) t_9^2 + \frac{\sqrt{kD}}{n_r} t_9^4$$

for  $t_9 \geq C$  on an event  $\Omega_{t_9}$  having high probability. Summarizing, on the high probability event  $\cap_{i \in \{1,4,7,8\}} \Omega_{t_i}$ :

$$\begin{aligned} \|B\| &\leq \frac{2\kappa}{\sqrt{k}} r^2 \sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}} + \sigma \sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}} \left( \left( \sqrt{\frac{k}{n_r}} + \sqrt{\frac{D}{n_r}} \right) t_7 + \frac{\sqrt{D}}{n_r} t_7^2 \right) \\ &\quad + \sigma \frac{2\kappa}{\sqrt{k}} r^2 \left[ \sqrt{1 + \sqrt{\frac{k}{n_r}} t_4 \left( 1 + \sqrt{\frac{k}{n_r}} t_4^3 \right)} \wedge \left( \left( \sqrt{\frac{k}{n_r}} + \sqrt{\frac{D}{n_r}} \right) t_8 + \frac{\sqrt{D}}{n_r} t_8^2 \right) \right] \\ &\quad + \sigma^2 \left( \left( \sqrt{\frac{k}{n_r}} + \sqrt{\frac{D}{n_r}} \right) t_9^2 + \frac{\sqrt{kD}}{n_r} t_9^4 \right) =: P_4. \end{aligned}$$

It is shown below that  $\Delta_k = (\tilde{\lambda}_{n_r,k}^{\parallel})^2 - (\tilde{\lambda}_{n_r,k+1}^{\perp})^2 \geq \frac{r^2 \lambda_{\min}^2}{k} - \frac{4\kappa^2}{k} r^4 - P_2^{\parallel} - P_3^{\parallel} - P_3^{\perp}$ .

## 6.5 Largest Gap

Let  $\tilde{\Delta}_i = \tilde{\lambda}_{n_r,i}^2 - \tilde{\lambda}_{n_r,i+1}^2$  for  $i = 1, \dots, D-1$ ,  $\tilde{\Delta}_D = \tilde{\lambda}_{n_r,D}^2$ . The goal is to lower bound the probability that  $\tilde{\Delta}_k = \tilde{\Delta}_{\max} := \max_{i=1, \dots, D} \tilde{\Delta}_i$ . For  $1 \leq i < k$ :

$$\begin{aligned} \tilde{\Delta}_i &= \tilde{\lambda}_{n_r,i}^2 - \tilde{\lambda}_{n_r,i+1}^2 \leq (\tilde{\lambda}_{n_r,i}^{\parallel})^2 - (\tilde{\lambda}_{n_r,i+1}^{\parallel})^2 + P_4 \\ &\leq (\lambda_{n_r,i}^{\parallel})^2 - (\lambda_{n_r,i+1}^{\parallel})^2 + 2P_3^{\parallel} + P_4 \leq (\lambda_i^{\parallel})^2 - (\lambda_{i+1}^{\parallel})^2 + 2P_2^{\parallel} + 2P_3^{\parallel} + P_4 \\ &\leq \frac{\delta^2 r^2}{k} + 2P_2^{\parallel} + 2P_3^{\parallel} + P_4. \end{aligned}$$

For  $i = k$ :

$$\begin{aligned}
\tilde{\Delta}_k &= \tilde{\lambda}_{n_r, k}^2 - \tilde{\lambda}_{n_r, k+1}^2 \geq (\tilde{\lambda}_{n_r, k}^{\parallel})^2 - (\tilde{\lambda}_{n_r, k+1}^{\perp})^2 \\
&\geq (\lambda_{n_r, k}^{\parallel})^2 + \sigma^2 - P_3^{\parallel} - (\lambda_{n_r, k+1}^{\perp})^2 - \sigma^2 \mathbf{1}_{(n_r \geq D)} - (P_3^{\perp} + \sigma^2 \mathbf{1}_{(n_r < D)}) \\
&\geq (\lambda_k^{\parallel})^2 - \frac{4\kappa^2}{k} r^4 + \sigma^2 - \sigma^2 - P_2^{\parallel} - P_3^{\parallel} - P_3^{\perp} \\
&\geq \frac{r^2 \lambda_{\min}^2}{k} - \frac{4\kappa^2}{k} r^4 - P_2^{\parallel} - P_3^{\parallel} - P_3^{\perp}.
\end{aligned}$$

For  $k < i \leq D$ :

$$\tilde{\Delta}_i \leq \tilde{\lambda}_{n_r, k+1}^2 \leq (\tilde{\lambda}_{n_r, k+1}^{\perp})^2 \leq (\lambda_{n_r, k+1}^{\perp})^2 + P_3^{\perp} + \sigma^2 \leq \frac{4\kappa^2}{k} r^4 + P_3^{\perp} + \sigma^2.$$

Therefore, in order for  $\tilde{\Delta}_k$  to be the largest gap, one has the sufficient condition:

$$\begin{aligned}
\frac{r^2 \lambda_{\min}^2}{k} - \frac{4\kappa^2}{k} r^4 - P_2^{\parallel} - P_3^{\parallel} - P_3^{\perp} \\
\geq \left( \frac{\delta^2 r^2}{k} + 2P_2^{\parallel} + 2P_3^{\parallel} + P_4 \right) \vee \left( \frac{4\kappa^2}{k} r^4 + P_3^{\perp} + \sigma^2 \right). \tag{6.1}
\end{aligned}$$

Observe that (6.1) implies  $(\tilde{\lambda}_{n_r, k}^{\parallel})^2 > (\tilde{\lambda}_{n_r, k+1}^{\perp})^2$ , which was assumed in  $P_4$  in order to apply Wielandt's inequality.

The terms in Table 6.2 are now bounded in order to determine a range for  $r$  that guarantees  $\tilde{\Delta}_k = \tilde{\Delta}_{\max}$  with high probability. Assuming  $n_r \geq t^4 \frac{k \log k}{\lambda_{\max}^2}$ , where  $t = \max_i t_i$ , and letting  $\tilde{\epsilon}_t := t^2 \frac{\sqrt{k \log k}}{\lambda_{\max} \sqrt{n_r}}$ ,  $\hat{\epsilon}_t := \frac{\tilde{\epsilon}_t \lambda_{\max}}{\sqrt{k}}$ ,  $\sigma := \frac{\sigma_0}{\sqrt{D}}$ , one easily obtains the bounds in Table 6.3.

From Table 6.3 and Equation 6.1, one obtains that it is sufficient for:

$$\frac{\lambda_{\min}^2 r^2}{k} \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - 6 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t \right) \gtrsim \underbrace{\frac{\kappa^2 r^4}{k} + \frac{\lambda_{\max} \kappa r^3}{k}}_{\text{curvature}} + \underbrace{\frac{\lambda_{\max} r}{\sqrt{k}} \sigma_0 \hat{\epsilon}_t + \sigma_0^2 \tilde{\epsilon}_t^2 + \frac{\sigma_0^2}{D}}_{\text{noise}}.$$

Table 6.3: Bounding of the  $P_i$ 's; assume that  $n_r \geq t^4 \frac{k \log k}{\lambda_{\max}^2}$  and let  $\tilde{\epsilon}_t := t^2 \frac{\sqrt{k \log k}}{\lambda_{\max} \sqrt{n_r}}$ ,  $\hat{\epsilon}_t := \frac{\tilde{\epsilon}_t \lambda_{\max}}{\sqrt{k}}$ ,  $\sigma := \frac{\sigma_0}{\sqrt{D}}$ .

Object	Bound
$P_2^{\parallel}$	$2 \frac{\lambda_{\max}^2 r^2}{k} \tilde{\epsilon}_t$
$\sqrt{\frac{\lambda_{\max}^2 r^2}{k} + P_2^{\parallel}}$	$\sqrt{3} \frac{\lambda_{\max} r}{\sqrt{k}}$
$P_3^{\parallel}$	$4\sqrt{3} \frac{\sigma_0}{\sqrt{D}} \frac{\lambda_{\max} r}{\sqrt{k}} \tilde{\epsilon}_t + 2 \frac{\sigma_0^2}{D} \tilde{\epsilon}_t$
$P_3^{\perp} \cdot \mathbf{1}_{(n_r \geq D)}$	$8\sigma_0 \frac{\kappa r^2}{\sqrt{k}} \hat{\epsilon}_t + \sigma_0^2 \hat{\epsilon}_t \left( \frac{1}{\sqrt{D}} + \hat{\epsilon}_t \right)$
$P_3^{\perp} \cdot \mathbf{1}_{(n_r < D)}$	$8\sigma_0 \frac{\kappa r^2}{\sqrt{k}} \hat{\epsilon}_t + 2\sigma_0^2 \hat{\epsilon}_t^2$
$P_4$	$2\sqrt{3} \frac{\lambda_{\max} \kappa r^3}{k} + 3\sqrt{3} \frac{\lambda_{\max} r}{\sqrt{k}} \sigma_0 \hat{\epsilon}_t + 2\sqrt{3} \frac{\kappa r^2}{\sqrt{k}} \sigma_0 \left( \frac{1}{\sqrt{D}} \wedge \sqrt{3} \hat{\epsilon}_t \right) + 3 \frac{\sigma_0^2}{\sqrt{D}} \hat{\epsilon}_t$

From the curvature terms, one gets the following upper bounds for  $r$ :

$$r \lesssim \frac{\lambda_{\min}}{\kappa} \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - 6 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t \right)^{\frac{1}{2}} \quad (6.2)$$

$$r \lesssim \frac{\lambda_{\min}}{\kappa} \left( \frac{\lambda_{\min}}{\lambda_{\max}} \right) \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - 6 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t \right) \quad (6.3)$$

Note that 6.3 implies 6.2. From the noise terms, one gets the following lower bounds:

$$r \gtrsim \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \sigma_0 \tilde{\epsilon}_t \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - 6 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t \right)^{-1} \quad (6.4)$$

$$r \gtrsim \frac{\lambda_{\max}}{\lambda_{\min}} \sigma_0 \tilde{\epsilon}_t \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - 6 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t \right)^{-\frac{1}{2}} \quad (6.5)$$

$$r \gtrsim \frac{\sigma_0}{\lambda_{\min}} \frac{\sqrt{k}}{\sqrt{D}} \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - 6 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t \right)^{-\frac{1}{2}} \quad (6.6)$$

Note that 6.4 implies 6.5. Thus one obtains the following range of scales for  $r$ :

$$\frac{\sigma_0}{\left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - 6 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t \right)^{\frac{1}{2}}} \left( \frac{\frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t}{\left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - 6 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t \right)^{\frac{1}{2}}} \vee \frac{\sqrt{k}}{\lambda_{\min} \sqrt{D}} \right) \lesssim r \lesssim \frac{\lambda_{\min}}{\kappa} \left( \frac{\lambda_{\min}}{\lambda_{\max}} \right) \left( 1 - \frac{\delta^2}{\lambda_{\min}^2} - 6 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \tilde{\epsilon}_t \right). \quad (6.7)$$

Finally, the conditioning on  $n_r$  is removed. A Chernoff bound (see Thm. 36) is used to bound (w.h.p.) how many points land in a ball of radius  $r$ . Since  $\mathbb{E}[n_r] =$

$\mu_X(B_z(r))n$ , letting  $\Omega_{t_{10}} = \{n_r > \frac{1}{2}\mu_X(B_z(r))n\}$ , one obtains

$$\mathbb{P}(\Omega_{t_{10}}) = \mathbb{P}\left(n_r > \frac{1}{2}\mu_X(B_z(r))n\right) \geq 1 - e^{-\frac{1}{8}t_{10}^2}$$

for  $n \geq t_{10}^2(\mu_X(B_z(r)))^{-1}$ .

Thus if for  $\epsilon = \epsilon_{t,r} := \frac{t^2}{\lambda_{\max}} \sqrt{\frac{2k \log k}{\mu_X(B_z(r))n}} \in [0, 1]$ ,  $r$  satisfies

$$\frac{\sigma_0}{\left(1 - \frac{\delta^2}{\lambda_{\min}^2} - 6\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\epsilon\right)^{\frac{1}{2}}} \left( \frac{\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\epsilon}{\left(1 - \frac{\delta^2}{\lambda_{\min}^2} - 6\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\epsilon\right)^{\frac{1}{2}}} \vee \frac{\sqrt{k}}{\lambda_{\min}\sqrt{D}} \right) \lesssim r \lesssim \frac{\lambda_{\min}}{\kappa} \left(\frac{\lambda_{\min}}{\lambda_{\max}}\right) \left(1 - \frac{\delta^2}{\lambda_{\min}^2} - 6\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\epsilon\right) \quad (6.8)$$

then Equ. 6.7 and the sampling condition of  $\Omega_{t_{10}}$  both hold, and  $\Delta_k(\text{cov}(\tilde{X}_{n,z,r})) = \Delta_{\max}(\text{cov}(\tilde{X}_{n,z,r}))$  on  $\bigcap_{i=1}^{10} \Omega_{t_i}$ , which has probability at least  $1 - Ce^{-ct^2}$ , for  $t \geq C$ , where  $c$  and  $C$  are universal constants. Furthermore:

$$\begin{aligned} \|\text{cov}(X_{z,r}) - \text{cov}(\tilde{X}_{n,z,r})\| &\leq P_1 + \left(P_2^{\parallel} \vee \frac{5\kappa^2 r^4}{k}\right) + \sigma^2 + (P_3^{\parallel} \vee P_3^{\perp}) \\ &\quad + \left(P_4 \wedge P_4^2 \left(\frac{\lambda_{\min}^2 r^2}{k} - \frac{4\kappa^2 r^4}{k} - P_2^{\parallel} - P_3^{\parallel} - P_3^{\perp}\right)^{-1}\right) \\ &\lesssim \frac{\kappa^2 r^4}{k} + \theta_{\epsilon} + \sigma^2 + \left(\frac{\lambda_{\max}\kappa r^3}{k} + v_{\epsilon}\right) \left[\frac{\lambda_{\max}\kappa r^3 + kv_{\epsilon}}{\lambda_{\min}^2 r^2 - C(\kappa^2 r^4 + k\theta_{\epsilon})} \wedge 1\right] \end{aligned}$$

where

$$\begin{aligned} \theta_{\epsilon} &= \epsilon_{t,r} \left[ \frac{\lambda_{\max}^2 r^2}{k} + \sigma_0^2 \frac{\lambda_{\max}}{\sqrt{k}} \left( \frac{1}{\sqrt{D}} + \frac{\lambda_{\max}}{\sqrt{k}} \epsilon_{t,r} \right) \right] \\ v_{\epsilon} &= \epsilon_{t,r} \left[ \sigma_0 \frac{\lambda_{\max} r}{k} + \sigma_0 \frac{\kappa r^2}{\sqrt{k}} \left( \frac{1}{\epsilon_{t,r}\sqrt{D}} \wedge \frac{\lambda_{\max}}{\sqrt{k}} \right) + \sigma_0^2 \frac{\lambda_{\max}}{\sqrt{Dk}} \right]. \end{aligned}$$

**Remark 17.** Note that for  $\theta_{\epsilon} \vee \sigma^2 \leq O(r^4)$  and  $v_{\epsilon} \leq O(r^3)$ , one has an  $O(r^4)$  perturbation precisely when  $\frac{\lambda_{\min}^2}{k}r^2$  is large relative to  $\frac{\lambda_{\max}\kappa}{k}r^3 + \frac{\kappa^2}{k}r^4 + \theta_{\epsilon} + v_{\epsilon}$ ; solving the equation:

$$\frac{\lambda_{\min}^2}{k}r^2 \gtrsim \frac{\lambda_{\max}\kappa}{k}r^3 + \frac{\kappa^2}{k}r^4 + \theta_{\epsilon} + v_{\epsilon}$$

gives the range of scales in (6.8). For noise not too large and  $n$  large enough, the perturbation is bounded by  $C \left( \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \right) \frac{\kappa^2}{k} r^4$  for some absolute constant  $C$  w.h.p..

□

## Recentering and Noise

This chapter presents the proof of Prop. 10. Recall the following definitions:

$$\begin{aligned} \tilde{X}_{n,z,r} &:= \{x_i + \eta_i : x_i \in B_z(r)\} \quad , & \widetilde{X}_{n,\tilde{z},r} &:= \{x_i + \eta_i \in B_{z+\eta_z}(r)\} , \\ r_\sigma^2 &:= r^2 - \sigma^2 D \quad , & r_{2\sigma}^2 &:= r^2 - 2\sigma^2 D \quad , & \xi &:= \frac{\sigma\sqrt{D}}{r} \quad , & d &:= D - k. \end{aligned}$$

**Remark 18.** *By Appendix E, which addresses how the covariance matrices are perturbed when a small fraction of the points (the “outliers”) are removed by thresholding the noise, one may assume all the noise vectors,  $\{\eta_i\}_{i=1}^n$ , are bounded by  $\sigma^2 D(1 + c \frac{(\ln n \wedge \ln D)}{\sqrt{D}}) \approx \sigma^2 D$ , where  $c$  is an absolute constant. Thus by a tiny reduction in the number of samples, one may assume i.i.d., bounded noise vectors. This work does not address how this outlier detection step is actually applied.*

The first step in the proof of Prop. 10 is to show that the sets  $\tilde{X}_{n,z,r_{2\sigma}}$  and  $\tilde{X}_{n,\tilde{z},r_\sigma}$  are close: the set of points within distance  $r_{2\sigma}$  of  $z \in \mathcal{M}$  is roughly equivalent to the set of points within distance  $r_\sigma$  of a noisy center  $\tilde{z} = z + \eta_z \notin \mathcal{M}$ ; thus by a change in scale, one can move from a center  $z \in \mathcal{M}$  to a noisy center  $\tilde{z} \notin \mathcal{M}$ . More precisely, to show that  $\tilde{X}_{n,z,r_{2\sigma}}$  and  $\tilde{X}_{n,\tilde{z},r_\sigma}$  are close, the following perturbations are shown to

be small:

$$\tilde{X}_{n,\tilde{z},r_\sigma} \rightarrow \tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2 \rightarrow \tilde{X}_{n,z,r_{2\sigma}},$$

where

$$\begin{aligned} A_1 &:= \tilde{X}_{n,z,\sqrt{r_{2\sigma}^2-q}} \cap \tilde{X}_{n,\tilde{z},r_\sigma}^c = \{\tilde{x}_i : \|x_i - z\| < \sqrt{r_{2\sigma}^2 - q} \text{ and } \|x_i - \tilde{z}\| > r_\sigma\} \\ A_2 &:= \tilde{X}_{n,z,\sqrt{r_{2\sigma}^2+q}}^c \cap \tilde{X}_{n,\tilde{z},r_\sigma} = \{\tilde{x}_i : \|x_i - z\| > \sqrt{r_{2\sigma}^2 + q} \text{ and } \|x_i - \tilde{z}\| < r_\sigma\} \end{aligned} \quad (7.1)$$

$$q := s^2\sigma^2\sqrt{D} + 4t_0\sigma\left(r_\sigma + \frac{2\kappa}{\sqrt{k}}r_\sigma^2\right),$$

with  $s^2, t_0$  parameters to be chosen later. The first perturbation is shown to be small by showing that  $A_1$  and  $A_2$  are small relative to  $\tilde{X}_{n,\tilde{z},r_\sigma}$ ; the second perturbation is shown to be small by showing that  $\tilde{X}_{n,z,\sqrt{r_{2\sigma}^2+q}} \setminus \tilde{X}_{n,z,\sqrt{r_{2\sigma}^2-q}}$ , which contains the set where  $\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2$  and  $\tilde{X}_{n,z,r_{2\sigma}}$  differ, is small relative to  $\tilde{X}_{n,\tilde{z},r_\sigma}$ . Lemma 26 then gives that  $\|\text{cov}(\tilde{X}_{n,z,r_{2\sigma}}) - \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma})\|$  is small.

The second step is to show that the sets  $\widetilde{X_{n,\tilde{z},r}}$  and  $\tilde{X}_{n,\tilde{z},r_\sigma}$  are close: the set of noisy points that were within distance  $r_\sigma$  of  $\tilde{z}$  before they were corrupted by noise is roughly equivalent to the set of noisy points within distance  $r$  of  $\tilde{z}$ , i.e. up to a change in scale (the size of the intersecting ball). In other words, intersecting with a ball and then adding noise is equivalent to adding noise and then intersecting with a ball of slightly different radius. More precisely, to show  $\widetilde{X_{n,\tilde{z},r}}$  and  $\tilde{X}_{n,\tilde{z},r_\sigma}$  are close, the following perturbations are bounded:

$$\widetilde{X_{n,\tilde{z},r}} = (\widetilde{X_{n,\tilde{z},r}} \cap \tilde{X}_{n,\tilde{z},r}) \cup I \rightarrow \widetilde{X_{n,\tilde{z},r}} \cap \tilde{X}_{n,\tilde{z},r} = (\tilde{X}_{n,\tilde{z},r_\sigma} \setminus Q_1) \cup Q_2 \rightarrow \tilde{X}_{n,\tilde{z},r_\sigma},$$

where

$$\begin{aligned} I &:= \widetilde{X_{n,\tilde{z},r}} \cap \tilde{X}_{n,\tilde{z},r}^c = \{\tilde{x}_i : \|\tilde{x}_i - \tilde{z}\| < r \text{ and } \|x_i - \tilde{z}\| > r\} \\ Q_1 &:= \tilde{X}_{n,\tilde{z},r_\sigma}^c \cap \widetilde{X_{n,\tilde{z},r}} = \{\tilde{x}_i : \|x_i - \tilde{z}\| \in [\sigma\sqrt{d}, r_\sigma] \text{ and } \|\tilde{x}_i - \tilde{z}\| > r\} \\ Q_2 &:= \tilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r_\sigma}^c \cap \widetilde{X_{n,\tilde{z},r}} = \{\tilde{x}_i : \|x_i - \tilde{z}\| \in [r_\sigma, r] \text{ and } \|\tilde{x}_i - \tilde{z}\| < r\}. \end{aligned} \quad (7.2)$$

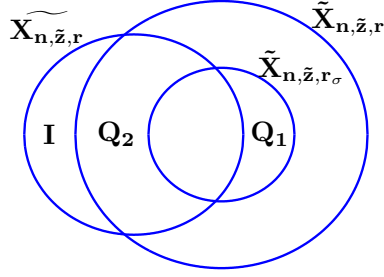


FIGURE 7.1: Set intersections.

Note that the first perturbation is small if  $I$  is small relative to  $\widetilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r}$ , and the second perturbation is small if both  $Q_1$  and  $Q_2$  are small relative to  $\widetilde{X}_{n,\tilde{z},r_\sigma}$ . Once this is established, Lemma 26 allows one to conclude that  $\|\text{cov}(\widetilde{X}_{n,\tilde{z},r}) - \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma})\|$  is small. Table 7.1 keeps track of how large all of the above perturbations are and with what probabilities. It is assumed throughout this section that  $n \geq t^2/\mu_X(B_z(r_{2\sigma}))$ , the smallest set which appears. Throughout this chapter, condition on the event that  $\tilde{z} = z + \eta_z$  is not an “outlier”:

$$\Omega_{s,0} := \{ \omega : \left| \|\eta_z(\omega)\|^2 - \sigma^2 D \right| \leq s^2 \sigma^2 \sqrt{D} \}, \quad (7.3)$$

which has probability at least  $1 - 2e^{-cs^4}$  for  $s^2 \leq \sqrt{D}$ . On  $\Omega_{s,0}$ , the volume growth assumptions imply that the expected cardinalities of  $\tilde{X}_{n,z,r_{2\sigma}}$  and  $\tilde{X}_{n,\tilde{z},r_\sigma}$  are essentially equivalent, and this fact is frequently used. More precisely, one can show that

$$c_1 \mu_X(B_z(r_{2\sigma})) \leq \mu_X(B_{\tilde{z}}(r_\sigma)) \leq c_2 \mu_X(B_z(r_{2\sigma})), \quad (7.4)$$

where

$$c_1 = \left( 1 - \frac{s^2 \xi^2}{\sqrt{D}(1 - 2\xi^2)} \right)^k \left( 1 - \left( 1 + \frac{s^2}{\sqrt{D}} \right)^{\frac{1}{2}} \left( \frac{\xi^2}{1 - 2\xi^2} \right)^{\frac{1}{2}} \right)$$

$$c_2 = \left( 1 + \frac{13s^2 \xi^2}{\sqrt{D}(1 - 2\xi^2)} \right)^k \left( 1 + \left( 1 + \frac{s^2}{\sqrt{D}} \right)^{\frac{1}{2}} \left( \frac{\xi^2}{1 - 2\xi^2} \right)^{\frac{1}{2}} \right)$$

with probability at least  $1 - Ce^{-cs^4}$  for  $s^2 \leq \sqrt{D}$ . It is assumed throughout this chapter that  $s^2 \leq \sqrt{D} \left(1 \wedge \frac{c}{\xi^{2k}}\right)$  and  $\xi$  bounded away from  $\frac{1}{\sqrt{2}}$  (in fact Sec. 7.2 imposes the stronger condition that  $\xi$  is bounded away from  $\frac{1}{\sqrt{3}}$ ), so that  $c_1$  and  $c_2$  are bounded by universal, order 1 constants independent of  $k$ , and  $\mu_X(B_{\tilde{z}}(r_\sigma))$  and  $\mu_X(B_z(r_{2\sigma}))$  are equivalent w.h.p..

The following inequalities are also used, which follow from the volume growth assumptions:

$$\begin{aligned} \frac{\mu_X(B_z(r+h)) - \mu_X(B_z(r))}{\mu_X(B_z(r))} &\leq \left(\frac{r+h}{r}\right)^{2k} - 1 \leq \frac{2kh}{r} + O(h^2) \\ \frac{\mu_X(B_z(r+h)) - \mu_X(B_z(r))}{\mu_X(B_z(r+h))} &\leq 1 - \left(\frac{r}{r+h}\right)^{2k} \leq \frac{2kh}{r+h} + O(h^2). \end{aligned} \quad (7.5)$$

## 7.1 Recentering

### 7.1.1 Comparing $\tilde{X}_{n,\tilde{z},r_\sigma}$ and $\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2$

Let  $q, A_1, A_2$  be as in (7.1). The expected cardinality of these sets relative to  $\tilde{X}_{n,\tilde{z},r_\sigma}$  is computed.

Suppose  $x_i \in B_z(\sqrt{r_{2\sigma}^2 - q})$ . Working on  $X_{z,\sqrt{r_{2\sigma}^2 - q}}$  and using the associated projection  $P^{(z,\sqrt{r_{2\sigma}^2 - q})}$  as in the assumptions in Sec. 5.1, one may write  $x_i - z = x_i^\parallel + x_i^\perp$  and  $z - \tilde{z} = \eta_z^\parallel + \eta_z^\perp$ . One has

$$\|x_i - \tilde{z}\|^2 = \|x_i - z\|^2 + \|\eta_z\|^2 + 2\langle x_i^\parallel, \eta_z^\parallel \rangle + 2\langle x_i^\perp, \eta_z^\perp \rangle$$

and, in expectation,  $\mathbb{E}_{\eta_z} \|x_i - \tilde{z}\|^2 = \mathbb{E}_{\eta_z} \|x_i - z\|^2 + \sigma^2 D$ . For a fixed  $x_i^\parallel$  and  $x_i^\perp$ , the subgaussian condition on the noise gives

$$\mathbb{P}_{\eta_z} \left( |\langle x_i^\parallel, \eta_z^\parallel \rangle| > t_0 \sigma r_\sigma \right) \leq 2e^{-ct_0^2} \quad , \quad \mathbb{P}_{\eta_z} \left( |\langle x_i^\perp, \eta_z^\perp \rangle| > t_0 \frac{\sigma \kappa}{\sqrt{k}} r_\sigma^2 \right) \leq 2e^{-ct_0^2} \quad (7.6)$$

for some absolute constant  $c$ . Thus if one defines

$$\Omega_{t_0} := \left\{ \omega : |\langle x_i^\parallel, \eta_z^\parallel(\omega) \rangle| \leq t_0 \sigma r_\sigma \text{ and } |\langle x_i^\perp, \eta_z^\perp(\omega) \rangle| \leq t_0 \frac{\sigma \kappa}{\sqrt{k}} r_\sigma^2 \right\}, \quad (7.7)$$

then on an event  $\Omega_{t_0}$  with probability at least  $1 - (4e^{-ct_0^2})$

$$\begin{aligned} \|x_i - \tilde{z}\|^2 &\leq \|x_i - z\|^2 + \sigma^2 D + s^2 \sigma^2 \sqrt{D} + 2t_0 \sigma \left( r_\sigma + \frac{\kappa}{\sqrt{k}} r_\sigma^2 \right) \\ &\leq r_{2\sigma}^2 - q + \sigma^2 D + q = r_{2\sigma}^2 + \sigma^2 D = r_\sigma^2, \end{aligned}$$

i.e.  $x_i \in B_{\tilde{z}}(r_\sigma)$ . Thus:

$$\begin{aligned} \mathbb{E}[|A_1|] &= \sum_{i=1}^n \mathbb{P} \left( \|x_i - z\| < \sqrt{r_{2\sigma}^2 - q} \text{ and } \|x_i - \tilde{z}\| > r_\sigma \right) \\ &= \sum_{i=1}^n \mathbb{P} \left( \|x_i - \tilde{z}\| > r_\sigma \mid \|x_i - z\| < \sqrt{r_{2\sigma}^2 - q} \right) \cdot \mathbb{P} \left( \|x_i - z\| < \sqrt{r_{2\sigma}^2 - q} \right) \\ &= 4e^{-ct_0^2} \sum_{i=1}^n \mathbb{P} \left( \|x_i - z\| < \sqrt{r_{2\sigma}^2 - q} \right) \\ &= 4e^{-ct_0^2} \mu_X \left( B_z \left( \sqrt{r_{2\sigma}^2 - q} \right) \right) n \lesssim 4e^{-ct_0^2} \mu_X(B_{\tilde{z}}(r_\sigma)) n. \end{aligned}$$

Now suppose  $x_i \in B_{\tilde{z}}(r_\sigma)$ ; on  $\Omega_{s,0}$ ,  $x_i \in B_z(r_\sigma + \sigma\sqrt{D}(1 + s^2 D^{-\frac{1}{2}}))$ ; thus working now on  $X_{z, r_\sigma + \sigma\sqrt{D}(1 + s^2 D^{-\frac{1}{2}})}$  and using the associated projection  $P^{(z, r_\sigma + \sigma\sqrt{D}(1 + s^2 D^{-\frac{1}{2}}))}$ , one may write  $x_i - z = x_i^\parallel + x_i^\perp$  and  $z - \tilde{z} = \eta_z^\parallel + \eta_z^\perp$  (note that  $\eta_z^\parallel$  and  $\eta_z^\perp$  are not the same vectors as above). 7.6 holds with  $r_\sigma + (\sigma^2 D + s^2 \sigma^2 \sqrt{D})^{\frac{1}{2}}$  replacing  $r_\sigma$ . Assuming  $\xi = \frac{\sigma\sqrt{D}}{r} < \frac{1}{\sqrt{2}}$  and  $s^2 \leq \sqrt{D}$ ,  $r_\sigma + (\sigma^2 D + s^2 \sigma^2 \sqrt{D})^{\frac{1}{2}} < 2r_\sigma$  and one obtains as before that on an event, again denoted by  $\Omega_{t_0}$ , of probability at least  $1 - (4e^{-ct_0^2})$ :

$$\begin{aligned} \|x_i - z\|^2 &= \|x_i - \tilde{z}\|^2 - \|\eta_z\|^2 - 2\langle x_i^\parallel, \eta_z^\parallel \rangle - 2\langle x_i^\perp, \eta_z^\perp \rangle \\ &\leq r_\sigma^2 - \sigma^2 D + s^2 \sigma^2 \sqrt{D} + 4t_0 \sigma \left( r_\sigma + \frac{2\kappa}{\sqrt{k}} r_\sigma^2 \right) = r_{2\sigma}^2 + q, \end{aligned}$$

implying  $x_i \in B_z(\sqrt{r_{2\sigma}^2 + q})$ . Thus:

$$\begin{aligned}
\mathbb{E}[|A_2|] &= \sum_{i=1}^n \mathbb{P} \left( \|x_i - z\| > \sqrt{r_{2\sigma}^2 + q} \text{ and } \|x_i - \tilde{z}\| < r_\sigma \right) \\
&= \sum_{i=1}^n \mathbb{P} \left( \|x_i - z\| > \sqrt{r_{2\sigma}^2 + q} \mid \|x_i - \tilde{z}\| < r_\sigma \right) \cdot \mathbb{P}(\|x_i - \tilde{z}\| < r_\sigma) \\
&\leq 4e^{-ct_0^2} \sum_{i=1}^n \mathbb{P}(\|x_i - \tilde{z}\| < r_\sigma) = 4e^{-ct_0^2} \mu_X(B_{\tilde{z}}(r_\sigma))n.
\end{aligned}$$

Thus, since  $\mu_X(B_{\tilde{z}}(r_\sigma))n \lesssim \mu_X(B_z(r_{2\sigma}))n$ , choosing  $t_0 = \sqrt{c^{-1} \log(Dk^{-2}\xi^{-2})}$ , by Lemma 26,

$$\|\text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma}) - \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2)\| \lesssim s^2 \left( \frac{\xi^2 k^2}{D} \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n} \right) (r_\sigma^2 + \sigma^2 D),$$

with probability given in Table 7.1.

### 7.1.2 Comparing $\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2$ and $\tilde{X}_{n,z,r_{2\sigma}}$

Next  $\|\text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2) - \text{cov}(\tilde{X}_{n,z,r_{2\sigma}})\|$  is shown to be small. One has:

$$\begin{aligned}
\tilde{X}_{n,z,\sqrt{r_{2\sigma}^2 - q}} &\subseteq \tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2 \subseteq \tilde{X}_{n,z,\sqrt{r_{2\sigma}^2 + q}} \\
\tilde{X}_{n,z,\sqrt{r_{2\sigma}^2 - q}} &\subseteq \tilde{X}_{n,z,r_{2\sigma}} \subseteq \tilde{X}_{n,z,\sqrt{r_{2\sigma}^2 + q}}.
\end{aligned}$$

Thus the set where  $\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2$  and  $\tilde{X}_{n,z,r_{2\sigma}}$  differ is contained in

$\tilde{X}_{n,z,\sqrt{r_{2\sigma}^2 + q}} \setminus \tilde{X}_{n,z,\sqrt{r_{2\sigma}^2 - q}}$ . Letting  $r_{2\sigma,q\pm} = \sqrt{r_{2\sigma}^2 \pm q}$ , for  $\frac{q}{r_{2\sigma}^2} \lesssim \frac{1}{k}$ , which holds as long

as  $D \gtrsim k^2 \xi^2$  and  $s^2 \lesssim \frac{\sqrt{D}}{k\xi^2}$ , disregarding higher order terms, and using (7.5), one obtains:

$$\begin{aligned}
\mu_X(B_z(r_{2\sigma,q+}) \setminus B_z(r_{2\sigma,q-}))n &= \mu_X(B_z(r_{2\sigma,q-}))n \left( \frac{\mu_X(B_z(r_{2\sigma,q+})) - \mu_X(B_z(r_{2\sigma,q-}))}{\mu_X(B_z(r_{2\sigma,q-}))} \right) \\
&\leq \mu_X(B_z(r_{2\sigma}))n \left( \left( \frac{r_{2\sigma}^2 + q}{r_{2\sigma}^2 - q} \right)^k - 1 \right) = \mu_X(B_z(r_{2\sigma}))n \left( \left( 1 + \frac{2q}{r_{2\sigma}^2 - q} \right)^k - 1 \right) \\
&= \mu_X(B_z(r_{2\sigma}))n \left( \frac{2qk}{r_{2\sigma}^2 - q} \right) = \mu_X(B_z(r_{2\sigma}))n \frac{2qk}{r_{2\sigma}^2}.
\end{aligned}$$

Table 7.1: High probability events and their definitions;  $\mathbb{P}(\Omega_{s,i}) \geq 1 - 2e^{-\frac{1}{3}s^2(\delta_i \mathbb{E}[nr_{2\sigma}] \vee 1)} - e^{-\frac{1}{8}t^2}$ . Here  $\sigma_0 = \sigma\sqrt{D}$  and  $C_\xi = \frac{1}{1-2\xi^2}$ . Assumptions:  $1 \leq s^2 \leq \sqrt{D} \left(1 \wedge \frac{c}{\xi^2 k}\right)$  for some absolute constant  $c$ ,  $n \geq t^2/\mu_X(B_z(r_{2\sigma}))$ ,  $D \gtrsim k^3$ ,  $\xi \in \{\{0\} \cup [\sqrt{\frac{\log D}{D}}, \frac{1}{\sqrt{3}} - \tau]\}$  for some  $\tau > 0$ , and  $\Omega_{s,0}$  (defined in (7.3)) holds. Recall that  $\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2 \setminus Q_1 = \widetilde{X_{n,\tilde{z},r}}$ , so that  $\text{cov}(\tilde{X}_{n,z,r_{2\sigma}})$  and  $\text{cov}(\widetilde{X_{n,\tilde{z},r}})$  are close when all of the  $\delta_i$  are small; each  $\delta_i$  may be replaced with an upper bound, in particular for each  $\delta_i$  one may substitute  $\delta = \max_i \delta_i$ .  $\Omega_{s,1}$  and  $\Omega_{s,2}$  are from recentering;  $\Omega_{s,3}, \Omega_{s,4}, \Omega_{s,5}$  from noise.

$\Omega_{s,i}$	Event	Def of $\delta_i$
$\Omega_{s,1}$	$\ \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma}) - \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2)\ $ $\lesssim s^2 \left(\delta_1 \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n}\right) r_\sigma^2$	$\frac{\xi^2 k^2}{D}$
$\Omega_{s,2}$	$\ \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2) - \text{cov}(\tilde{X}_{n,z,r_{2\sigma}})\ $ $\lesssim s^2 \left(\delta_2 \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n}\right) r_\sigma^2$	$\frac{\xi k}{\sqrt{D}} \left(s^2 \xi + \sqrt{\log\left(\frac{D}{\xi^2 k^2}\right)}\right)$
$\Omega_{s,3}$	$\ \text{cov}(\widetilde{X_{n,\tilde{z},r}}) - \text{cov}(\widetilde{X_{n,\tilde{z},r}} \cap \tilde{X}_{n,\tilde{z},r})\ $ $\lesssim s^2 \left(\delta_3 \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n}\right) r^2$	$e^{-c\xi^2 D} ((k\xi \wedge 1) \vee C_{k,\xi,v_{\min}})$
$\Omega_{s,4}$	$\ \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2 \setminus Q_1) - \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2)\ $ $\lesssim s^2 \left(\delta_4 \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n}\right) r^2$	$\frac{\xi k}{\sqrt{D}} \left(1 + (\sigma_0 + 1) \sqrt{\log\left(\frac{D}{\xi^2 k^2}\right)}\right)$
$\Omega_{s,5}$	$\ \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2) - \text{cov}(\tilde{X}_{n,z,r_{2\sigma}})\ $ $\lesssim s^2 \left(\delta_5 \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n}\right) r^2$	$\frac{\xi k}{\sqrt{D}} \left(1 + (\sigma_0 \vee 1) \sqrt{k \log C_\xi + \log \frac{D}{\xi^2 k^2}}\right)$

By Lemma 26 on  $\Omega_{s,2}$ :

$$\|\text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup A_1 \setminus A_2) - \text{cov}(\tilde{X}_{n,z,r_{2\sigma}})\| \lesssim s^2 \left(\frac{qk}{r_{2\sigma}^2} \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n}\right) (r_{2\sigma}^2 + \sigma^2 D),$$

with probability given in Table 7.1. Combining the two perturbations, one obtains

$$\begin{aligned} & \|\text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma}) - \text{cov}(\tilde{X}_{n,z,r_{2\sigma}})\| \\ & \lesssim s^2 \left( \left( \frac{\xi^2 k^2}{D} + \frac{s^2 k \sigma^2 \sqrt{D}}{r_{2\sigma}^2} + k \sigma \sqrt{\log\left(\frac{D}{\xi^2 k^2}\right)} \left( \frac{\sqrt{C_\xi}}{r_{2\sigma}} + \frac{C_\xi \kappa}{\sqrt{k}} \right) \right) \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n} \right) r_\sigma^2, \\ & \lesssim s^2 \left( \left( \frac{\xi^2 k^2}{D} + \frac{s^2 \xi^2 k}{\sqrt{D}} + \frac{\xi k \sqrt{\log\left(\frac{D}{\xi^2 k^2}\right)}}{\sqrt{D}} \right) \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n} \right) r_\sigma^2, \end{aligned}$$

where  $C_\xi = \frac{1-\xi^2}{1-2\xi^2}$ ; here it is assumed that  $\frac{1}{1-2\xi^2}$  is bounded by a universal constant and that  $r \lesssim \frac{\sqrt{k}}{\kappa}$ .

## 7.2 The Effect of Noise

This section shows that  $\text{cov}(\widetilde{X}_{n,\tilde{z},r}) - \text{cov}(\tilde{X}_{n,\tilde{z},r\sigma})$  is small by showing each of the following perturbations is small:

$$\widetilde{X}_{n,\tilde{z},r} = (\widetilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r}) \cup I \rightarrow \widetilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r} = (\tilde{X}_{n,\tilde{z},r\sigma} \setminus Q_1) \cup Q_2 \rightarrow \tilde{X}_{n,\tilde{z},r\sigma},$$

allowing one to conclude that the eigenvalues of the associated covariance matrices are also close.

Let  $\tilde{z}_{\mathcal{M}}$  be a closest point to  $\tilde{z}$  on  $\mathcal{M}$  ( $\tilde{z}_{\mathcal{M}} \in \mathcal{M}$ ,  $\|\tilde{z} - \tilde{z}_{\mathcal{M}}\| = \inf_{z \in \mathcal{M}} \|\tilde{z} - z\|$ ). Also let:

$$f_{\tilde{z}}(r) := \sqrt{r^2 + \|\tilde{z} - \tilde{z}_{\mathcal{M}}\|^2} \quad , \quad s_{\tilde{z}}(r) := \sqrt{r^2 - \|\tilde{z} - \tilde{z}_{\mathcal{M}}\|^2} \quad , \quad |A^D| = H^D(A),$$

where  $H^D$  is  $D$ -dimensional Hausdorff measure. Note that with no curvature,  $\|\tilde{z} - \tilde{z}_{\mathcal{M}}\| \in \sigma^2 d(1 \pm \frac{s^2}{\sqrt{d}})$  with probability  $1 - 2e^{-cs^2(s^2 \wedge \sqrt{d})}$ , since  $\tilde{z} - \tilde{z}_{\mathcal{M}}$  is a random vector in  $\mathbb{R}^d$  with i.i.d., mean zero, variance  $\sigma^2$ , subgaussian moment bounded by  $\sigma$  coordinates. Because at scales as fine as  $O(\sigma\sqrt{k})$ , the data is essentially linear (the normal component being bounded by  $O(\kappa^2\sigma^4k)$ , which is negligible compared with  $\sigma^2d$ ), curvature can be ignored, as the dominant terms in the following computations are unchanged. For simplicity  $\|\tilde{z} - \tilde{z}_{\mathcal{M}}\|$  is approximated by  $\sigma^2d$ , although in reality one only has  $\|\tilde{z} - \tilde{z}_{\mathcal{M}}\|$  highly concentrated around  $\sigma^2d$ . This approximation, however, only affects some of the constants in the following computations as long as  $\xi^2$  is bounded away from  $\frac{1}{(1+\frac{s^2}{\sqrt{d}})^2}$ , where  $s$  is the probability parameter appearing in Prop.

10. Thus for simplicity in the following computations  $\|\tilde{z} - \tilde{z}_{\mathcal{M}}\|$  is approximated by  $\sigma^2d$  and  $\xi$  is assumed to be bounded away from  $\frac{1}{\sqrt{3}}$ .

The volume growth conditions from (5.2) imply:

$$\begin{aligned} \frac{\mu_X(B_{\tilde{z}}(f_{\tilde{z}}(r+h))) - \mu_X(B_{\tilde{z}}(f_{\tilde{z}}(r)))}{\mu_X(B_{\tilde{z}}(f_{\tilde{z}}(r)))} &\leq \left(\frac{r+h}{r}\right)^{2k} - 1 \leq \frac{2kh}{r} + O(h^2) \\ \frac{\mu_X(B_{\tilde{z}}(f_{\tilde{z}}(r+h))) - \mu_X(B_{\tilde{z}}(r))}{\mu_X(B_{\tilde{z}}(f_{\tilde{z}}(r+h)))} &\leq 1 - \left(\frac{r}{r+h}\right)^{2k} \leq \frac{2kh}{r+h} + O(h^2) \end{aligned} \quad (7.8)$$

which can be thought of as an extension of (7.5) for  $\tilde{z} \notin \mathcal{M}$ .

For  $0 \leq \theta \leq \pi$ , define  $V_\theta^{D-1}$  to be the spherical cap of  $\mathbb{S}^{D-1}$  given by the angle  $\theta$ .

**Lemma 19.** *The function  $h(\theta) := |V_\theta^{D-1}|/|\mathbb{S}^{D-1}|$  satisfies the following properties:*

1.  $0 = h(0) \leq h(\theta) \leq h(\pi) = 1$  for every  $0 \leq \theta \leq \pi$ .
2.  $h(\theta)$  is strictly increasing.
3. If  $\theta = \frac{\pi}{2} - t$  for any  $0 \leq t \leq \frac{\pi}{2}$ ,  $h(\theta) \leq e^{-\frac{1}{2}t^2(D-2)}$ .

For a proof of these facts, see Lec. 19 of Barvinok (2005). Now define:

$$\theta_0(R, \|\eta\|) := \arccos\left(\frac{R^2 + \|\eta\|^2 - r^2}{2R\|\eta\|}\right). \quad (7.9)$$

If a point  $x$  is at distance  $R$  from  $\tilde{z}$ , this is the angle subsumed by the spherical cap  $(x + \|\eta\| \cdot \mathbb{S}^D) \cap B_{\tilde{z}}(r)$ , so that  $h(\theta_0(R, \|\eta\|))$  gives the probability that the point will be in  $B_{\tilde{z}}(r)$  after a noise vector  $\eta$  is added. A simple computation shows that  $\theta_0$  is decreasing in  $\|\eta\|$  for  $R^2 < \|\eta\|^2 + r^2$  and decreasing in  $R$  for  $R^2 > \|\eta\|^2 - r^2$ .

7.2.1 Comparing  $\widetilde{X_{n,\tilde{z},r}} = (\widetilde{X_{n,\tilde{z},r}} \cap \tilde{X}_{n,\tilde{z},r}) \cup I$  with  $\widetilde{X_{n,\tilde{z},r}} \cap \tilde{X}_{n,\tilde{z},r}$

One has  $\widetilde{X_{n,\tilde{z},r}} = (\widetilde{X_{n,\tilde{z},r}} \cap \tilde{X}_{n,\tilde{z},r}) \cup I$ , where  $I$  is the set of points that enter  $B_{\tilde{z}}(r)$  when noise is added, as defined in (7.2). Partition the set  $I$  into  $I_1$ , the points coming into the ball from a distance less than  $r + \sigma\sqrt{D}$ , and  $I_2$ , points coming into the ball from a distance larger than  $r + \sigma\sqrt{D}$ :

$$I_1 = \{\tilde{x}_i : \|\tilde{x}_i - \tilde{z}\| < r \text{ and } \|x_i - \tilde{z}\| \in [r, r + \sigma\sqrt{D}]\}$$

$$I_2 = \{\tilde{x}_i : \|\tilde{x}_i - \tilde{z}\| < r \text{ and } \|x_i - \tilde{z}\| \geq r + \sigma\sqrt{D}\}$$

Because the size of the noise is highly concentrated around  $\sigma\sqrt{D}$ ,  $I_1$  constitutes the majority of  $I$ . In order to apply Lemma 26 one needs to determine the size of  $I$  relative to  $|\widetilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r}|$ ; since the points in  $B_{\tilde{z}}(r_\sigma)$  have a probability larger than  $1/2$  of staying in  $B_{\tilde{z}}(r)$  when noise is added:

$$\mathbb{E}[|\widetilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r}|] \geq \mathbb{E}[|\widetilde{X}_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r_\sigma}|] \geq \frac{1}{2}\mathbb{E}[|\tilde{X}_{n,\tilde{z},r_\sigma}|] = \frac{1}{2}\mu_X(B_{\tilde{z}}(r_\sigma))n,$$

it will be enough to compute the expected cardinalities of both  $I_1$  and  $I_2$  relative to  $\mu_X(B_{\tilde{z}}(r_\sigma))n$ .

$\mathbb{E}[|I_1|]$  is estimated first:

$$\begin{aligned} \mathbb{E}[|I_1|] &= \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \text{ and } \|x_i - \tilde{z}\| \in [r, r + \sigma\sqrt{D}]) \\ &= \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\| \in [r, r + \sigma\sqrt{D}]) \cdot \mathbb{P}(\|x_i - \tilde{z}\| \in [r, r + \sigma\sqrt{D}]) \end{aligned} \tag{7.10}$$

Now for each sample  $i$ , define the events  $\Omega_{t_1,i}$  (having probability at least  $1 - 2e^{-c(t_1^2 \wedge t_1 \sqrt{D})}$ ) and  $\Omega_{2,i}$ , as follows:

$$\begin{aligned} \Omega_{t_1,i} &:= \left\{ \omega : \left| \|\eta_{x_i}(\omega)\|^2 - \sigma^2 D \right| \leq t_1 \sigma^2 \sqrt{D} \right\} \\ \Omega_{2,i} &:= \Omega_{t_1,i} \cap \left\{ \omega : \left| \|x_i(\omega) - \tilde{z}(\omega)\| \right| \in [r, r + \sigma\sqrt{D}] \right\}. \end{aligned}$$

Then:

$$\begin{aligned}
& \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\| \in [r, r + \sigma\sqrt{D}]) \\
&= \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\| \in [r, r + \sigma\sqrt{D}], \Omega_{t_1, i}) \cdot \mathbb{P}(\Omega_{t_1, i}) \\
&\quad + \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\| \in [r, r + \sigma\sqrt{D}], \Omega_{t_1, i}^C) \cdot \mathbb{P}(\Omega_{t_1, i}^C) \\
&\leq \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \Omega_{2, i}) + \mathbb{P}(\Omega_{t_1, i}^C) = \mathbb{E}[\mathbf{1}_{\|\tilde{x}_i - \tilde{z}\| < r} \mid \Omega_{2, i}] + \mathbb{P}(\Omega_{t_1, i}^C) \\
&= \left( \frac{\int_{\Omega_{2, i}} \mathbf{1}_{\|\tilde{x}_i - \tilde{z}\| < r} dP}{\mathbb{P}(\Omega_{2, i})} \right) + \mathbb{P}(\Omega_{t_1, i}^C) \\
&= \frac{1}{\mathbb{P}(\Omega_{2, i})} \left( \int_{\Omega_{2, i}} \mathbb{E}[\mathbf{1}_{\|\tilde{x}_i - \tilde{z}\| < r} \mid \|x_i - \tilde{z}\|, \|\eta_{x_i}\|] dP \right) + \mathbb{P}(\Omega_{t_1, i}^C) \\
&\quad \text{(By the definition of conditional expectation, because } \Omega_{2, i} \text{ is measurable} \\
&\quad \text{w.r.t the sigma-algebra generated by the r.v.'s } \|x_i - \tilde{z}\|, \|\eta_{x_i}\|) \\
&= \frac{1}{\mathbb{P}(\Omega_{2, i})} \left( \int_{\Omega_{2, i}} \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\|, \|\eta_{x_i}\|) dP \right) + \mathbb{P}(\Omega_{t_1, i}^C) \\
&= \frac{1}{\mathbb{P}(\Omega_{2, i})} \left( \int_{\Omega_{2, i}} h(\theta_0(\|x_i - \tilde{z}\|, \|\eta_{x_i}\|)) dP \right) + \mathbb{P}(\Omega_{t_1, i}^C) \\
&\leq \frac{1}{\mathbb{P}(\Omega_{2, i})} \left( h(\theta_0(r, \sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}}))) \int_{\Omega_{2, i}} dP \right) + \mathbb{P}(\Omega_{t_1, i}^C) \\
&= h(\theta_0(r, \sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}}))) + 2e^{-c(t_1^2 \wedge t_1 \sqrt{D})}
\end{aligned}$$

The above uses that on  $\Omega_{2, i}$ ,  $\theta_0(\|x_i - \tilde{z}\|, \|\eta_{x_i}\|) \leq \theta_0(r, \sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}}))$ . Also:

$$\sum_{i=1}^n \mathbb{P}(\|x_i - \tilde{z}\| \in [r, r + \sigma\sqrt{D}]) = \mu_X \left( B_{\tilde{z}}(r + \sigma\sqrt{D}) \setminus B_{\tilde{z}}(r) \right) n.$$

One has  $\cos(\theta_0(r, \sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}}))) = \frac{\sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}})}{2r}$ ; thus for  $\theta_0 := \frac{\pi}{2} - t$ , one obtains  $\sin(t) = \frac{\sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}})}{2r}$ , and so  $t = \arcsin\left(\frac{\sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}})}{2r}\right) \geq \frac{\sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}})}{2r}$  and  $h(\theta_0(r, \sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}}))) \leq e^{-\frac{1}{2}t^2(D-2)} \leq e^{-\frac{1}{8}\left(\frac{\sigma\sqrt{D}(1 - t_1 D^{-\frac{1}{2}})}{r}\right)^2 (D-2)}$ .

Plugging the above calculations into (7.10):

$$\begin{aligned}
& \mathbb{E}[|I_1|] (\mu_X(B_{\tilde{z}}(r_\sigma))n)^{-1} \\
& \leq \left( h(\theta_0(r, \sigma\sqrt{D}(1-t_1D^{-\frac{1}{2}}))) + 2e^{-c(t_1^2 \wedge t_1\sqrt{D})} \right) \frac{\mu_X(B_{\tilde{z}}(r + \sigma\sqrt{D}) \setminus B_{\tilde{z}}(r))}{\mu_X(B_{\tilde{z}}(r_\sigma))} \\
& \leq \left( e^{-\frac{1}{8} \left( \frac{\sigma\sqrt{D}(1-t_1D^{-\frac{1}{2}})}{r} \right)^2 (D-2)} + 2e^{-c(t_1^2 \wedge t_1\sqrt{D})} \right) \left( \frac{s_{\tilde{z}}(r + \sigma\sqrt{D})^{2k} - s_{\tilde{z}}(r)^{2k}}{s_{\tilde{z}}(r_\sigma)^{2k}} \right) \\
& \leq \left( e^{-c\xi^2(1-t_1D^{-\frac{1}{2}})^2D} + 2e^{-c(t_1^2 \wedge t_1\sqrt{D})} \right) \left( \frac{(1+2\xi)^k - (1-\xi^2)^k}{(1-2\xi^2)^k} \right) \\
& \leq \frac{1}{2} e^{-c\xi^2D} \frac{(1+\xi)^{2k}}{(1-2\xi^2)^k} \left( 1 - \left( 1 - \frac{2\xi}{1+\xi} \right)^k \right).
\end{aligned}$$

Here  $t_1 = \xi\sqrt{D}$ ; if  $\xi < \frac{1}{2k}$ , one can upper bound the last term by  $2\xi k$  and the first term by a constant to obtain  $\mu_X(B_{\tilde{z}}(r_\sigma))n e^{-c\xi^2D} \xi k$ ; otherwise, upper bound the last term by 1 to obtain:  $\mu_X(B_{\tilde{z}}(r_\sigma))n e^{-c\xi^2D + 2k \log\left(\frac{1+\xi}{\sqrt{1-2\xi^2}}\right)} \lesssim \mu_X(B_{\tilde{z}}(r_\sigma))n e^{-c\xi^2D}$ , as soon as  $k^3 \lesssim D$  and  $\frac{1}{1-2\xi^2}$  is bounded by a universal constant.

$|I_2|$  is now shown to be small in expectation. For  $x \sim X$  and  $\eta_x \sim N$ , let:

$$\begin{aligned}
A_i & := \mu_X \left( B_{\tilde{z}}(r + (i+1)\sigma\sqrt{D}) \setminus B_{\tilde{z}}(r + i\sigma\sqrt{D}) \right) \\
p_i & := \mathbb{P}(\|x - \tilde{z}\| \in r + [i\sigma\sqrt{D}, (i+1)\sigma\sqrt{D}] \text{ and } \|x + \eta_x - \tilde{z}\| < r).
\end{aligned}$$

The condition in (5.2) implies  $\sum_{i=1}^{\infty} e^{-i^2} \mu_X(A_i) \leq C_{k,\xi} r^k$ . One has

$\mathbb{E}[|I_2|] \leq \sum_{i=1}^{\infty} p_i \mu_X(A_i) n$ , so an upper bound on  $p_i$  is needed. Note that for a point to enter the ball, two independent conditions must be met: the noise vector must be large enough, i.e.  $\|\eta\| \geq i\sigma\sqrt{D}$ , and the noise vector must be pointed in the right direction. The subgaussian condition on the noise gives  $\mathbb{P}(\|\eta\| \geq i\sigma\sqrt{D}) \leq 2e^{-i^2}$ . To upper bound the probability that the noise is pointed in the right direction, let  $x$  be a point at distance  $r + i\sigma\sqrt{D}$  from  $\tilde{z}$ ; let  $\theta_0$  be the angle formed by the line segment

connecting  $x$  and  $\tilde{z}$  and a tangent line to  $\mathbb{S}_r^{D-1}(\tilde{z})$  ( $\mathbb{S}_r^{D-1}$  centered at  $\tilde{z}$ ) passing through  $x$ . The angle satisfies  $\sin(\theta_0) = \frac{r}{r+i\sigma\sqrt{D}}$ . Note that the probability it is pointing in the appropriate direction is upper bounded by  $|V_{\theta_0}^{D-1}|/|\mathbb{S}^{D-1}|$ . Letting  $t = \frac{\pi}{2} - \theta_0$ , one obtains  $t = \arccos\left(\frac{r}{r+i\sigma\sqrt{D}}\right) \geq \frac{\pi}{2}\left(1 - \frac{r}{r+i\sigma\sqrt{D}}\right) = \frac{\pi}{2}\left(\frac{i\sigma\sqrt{D}}{r+i\sigma\sqrt{D}}\right)$ . By concentration of measure the probability of pointing in the right direction is bounded by  $e^{-c\left(\frac{i\xi}{1+i\xi}\right)^2 D}$  and therefore  $p_i \leq e^{-i^2 - c\left(\frac{i\xi}{1+i\xi}\right)^2 D}$ . Finally:

$$\begin{aligned} \frac{\mathbb{E}[|I_2|]}{\mu_X(B_{\tilde{z}}(r_\sigma))n} &\leq \frac{n}{\mu_X(B_{\tilde{z}}(r_\sigma))n} \sum_{i=1}^{\infty} p_i \mu_X(A_i) \leq \frac{1}{\mu_X(B_{\tilde{z}}(r_\sigma))} e^{-c\left(\frac{\xi}{1+\xi}\right)^2 D} \sum_{i=1}^{\infty} e^{-i^2} \mu_X(A_i) \\ &\leq \frac{e^{-c\left(\frac{\xi}{1+\xi}\right)^2 D} C_{k,\xi} r^k}{\mu_X(B_{\tilde{z}}(r_\sigma))} \leq \frac{e^{-c\left(\frac{\xi}{1+\xi}\right)^2 D} C_{k,\xi}}{v_{\min} \mu_{\mathbb{R}^k}(\mathbb{B}^k)} \left(\frac{r}{\sqrt{r^2 - \sigma^2(d+D)}}\right)^k \\ &\leq e^{-c\left(\frac{\xi}{1+\xi}\right)^2 D} C_{k,\xi,v_{\min}}. \end{aligned}$$

Thus applying Lemma 26, one has (assuming  $k^3 \lesssim D$  and  $\frac{1}{1-2\xi^2}$  bounded),

$$\begin{aligned} &\|\text{cov}(\widetilde{X_{n,\tilde{z},r}}) - \text{cov}(\widetilde{X_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r}})\| \\ &\lesssim s^2 \left( e^{-c\xi^2 D} ((k\xi \wedge 1) \vee C_{k,\xi,v_{\min}}) \vee \frac{1}{\mu_X(B_{\tilde{z}}(r_\sigma))n} \right) (r^2 + \sigma^2 D), \end{aligned}$$

with probability given in Table 7.1.

*7.2.2 Comparing  $\widetilde{X_{n,\tilde{z},r} \cap \tilde{X}_{n,\tilde{z},r}} = (\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2) \setminus Q_1$  with  $\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2$*

Let  $Q_1$  be the set of points in  $B_{\tilde{z}}(r_\sigma)$  that leave the ball when noise is added:

$$Q_1 := \{\tilde{x}_i : \|x_i - \tilde{z}\| \in [\sigma\sqrt{d}, r_\sigma] \text{ and } \|\tilde{x}_i - \tilde{z}\| > r\}$$

and let  $Q_2$  be the set of points in  $B_{\tilde{z}}(r) \setminus B_{\tilde{z}}(r_\sigma)$  that stay in  $B_{\tilde{z}}(r)$  when noise is added:

$$Q_2 := \{\tilde{x}_i : \|x_i - \tilde{z}\| \in [r_\sigma, r] \text{ and } \|\tilde{x}_i - \tilde{z}\| < r\}.$$

First, the size of  $Q_1$  is estimated:

$$\begin{aligned}\mathbb{E}[|Q_1|] &= \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| > r \text{ and } \|x_i - \tilde{z}\| \in [\sigma\sqrt{d}, r_\sigma]) \\ &= \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| > r \mid \|x_i - \tilde{z}\| \in [\sigma\sqrt{d}, f_{\tilde{z}}(\rho_{\epsilon_-})]) \cdot \mathbb{P}(\|x_i - \tilde{z}\| \in [\sigma\sqrt{d}, f_{\tilde{z}}(\rho_{\epsilon_-})])\end{aligned}\tag{7.11}$$

$$+ \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| > r \mid \|x_i - \tilde{z}\| \in [f_{\tilde{z}}(\rho_{\epsilon_-}), r_\sigma]) \cdot \mathbb{P}(\|x_i - \tilde{z}\| \in [f_{\tilde{z}}(\rho_{\epsilon_-}), r_\sigma]),\tag{7.12}$$

for any  $f_{\tilde{z}}(\rho_{\epsilon_-}) \in [\sigma\sqrt{d}, r_\sigma]$ . If a point is at distance  $R$  from  $\tilde{z}$ , the probability of exiting when noise vector  $\eta$  is added is given by  $h(\pi - \theta_0(R, \|\eta\|))$ . Note that  $\pi - \theta_0(R, \|\eta\|)$  is increasing in both  $R$  and  $\|\eta\|$ . By an identical argument to that in 7.2.1, one obtains:

$$\begin{aligned}\mathbb{P}(\|\tilde{x}_i - \tilde{z}\| > r \mid \|x_i - \tilde{z}\| \in [\sigma\sqrt{d}, f_{\tilde{z}}(\rho_{\epsilon_-})]) \\ \leq h(\pi - \theta_0(f_{\tilde{z}}(\rho_{\epsilon_-}), \sigma\sqrt{D}(1 + t_2D^{-\frac{1}{2}}))) + 2e^{-c(t_2^2 \wedge t_2\sqrt{D})}.\end{aligned}$$

Choose  $\rho_{\epsilon_-}$  so that  $\pi - \theta_0(f_{\tilde{z}}(\rho_{\epsilon_-}), \sigma\sqrt{D}(1 + t_2D^{-\frac{1}{2}})) = \frac{\pi}{2} - \epsilon$  and solve

$$\cos\left(\frac{\pi}{2} + \epsilon\right) \sim -\epsilon = \frac{r^2 - f_{\tilde{z}}(\rho_{\epsilon_-})^2 - \sigma^2D(1 + t_2D^{-\frac{1}{2}})^2}{-2f_{\tilde{z}}(\rho_{\epsilon_-})\sigma\sqrt{D}(1 + t_2D^{-\frac{1}{2}})}$$

for  $\rho_{\epsilon_-}$ . Assuming  $\rho_{\epsilon_-}$  has the asymptotic expansion  $\rho_{\epsilon_-} = \rho_0 + \epsilon\rho_1 + \epsilon^2\rho_2 + \dots$  and that  $\epsilon \rightarrow 0$  as  $D \rightarrow \infty$ , one can show as  $\epsilon \rightarrow 0$ :

$$\begin{aligned}\rho_{\epsilon_-} &\sim \sqrt{s_{\tilde{z}}(r_\sigma)^2 - \sigma^2t_2(2\sqrt{D} + t_2)} - \sigma\sqrt{D}(1 + t_2D^{-\frac{1}{2}}) \sqrt{\frac{r_\sigma^2 - \sigma^2t_2(2\sqrt{D} + t_2)}{s_{\tilde{z}}(r_\sigma)^2 - \sigma^2t_2(2\sqrt{D} + t_2)}} \epsilon \\ &\sim \sqrt{s_{\tilde{z}}(r_\sigma)^2 - \sigma^2t_2(2\sqrt{D} + t_2)} - \sigma\sqrt{D}(1 + t_2D^{-\frac{1}{2}}) \frac{r_\sigma}{s_{\tilde{z}}(r_\sigma)} \left(1 + \frac{t_2\xi^2C_\xi}{\sqrt{D}}\right) \epsilon \\ &\sim s_{\tilde{z}}(r_\sigma) - \frac{1}{2}\sigma^2t_2(2\sqrt{D} + t_2) - \sigma\sqrt{D} \left(1 + \frac{t_2(1 + \xi^2C_\xi)}{\sqrt{D}}\right) \frac{r_\sigma}{s_{\tilde{z}}(r_\sigma)} \epsilon.\end{aligned}$$

Thus for the above  $\rho_{\epsilon_-}$ ,  $h(\pi - \theta_0(f_{\bar{z}}(\rho_{\epsilon_-}), \sigma\sqrt{D}(1 + t_2D^{-\frac{1}{2}}))) \leq e^{-\frac{1}{2}\epsilon^2(D-2)}$ , and one can bound (7.11) as follows:

$$\begin{aligned} & \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| > r \mid \|x_i - \tilde{z}\| \in [\sigma\sqrt{d}, f_{\bar{z}}(\rho_{\epsilon_-})]) \cdot \mathbb{P}(\|x_i - \tilde{z}\| \in [\sigma\sqrt{d}, f_{\bar{z}}(\rho_{\epsilon_-})]) \\ & \leq \left( e^{-\frac{1}{2}\epsilon^2(D-2)} + 2e^{-c(t_2^2 \wedge t_2\sqrt{D})} \right) \mu_X(B_{\bar{z}}(f_{\bar{z}}(\rho_{\epsilon_-})) \setminus B_{\bar{z}}(\sigma\sqrt{d}))n \\ & \leq \left( e^{-\frac{1}{2}\epsilon^2(D-2)} + 2e^{-c(t_2^2 \wedge t_2\sqrt{D})} \right) \mu_X(B_{\bar{z}}(r_\sigma))n \end{aligned}$$

(7.12) is also bounded as follows (up to lower order terms):

$$\begin{aligned} & \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| > r \mid \|x_i - \tilde{z}\| \in [f_{\bar{z}}(\rho_{\epsilon_-}), r_\sigma]) \cdot \mathbb{P}(\|x_i - \tilde{z}\| \in [f_{\bar{z}}(\rho_{\epsilon_-}), r_\sigma]) \\ & \leq \mu_X(B_{\bar{z}}(r_\sigma) \setminus B_{\bar{z}}(f_{\bar{z}}(\rho_{\epsilon_-})))n \\ & = \mu_X(B_{\bar{z}}(r_\sigma))n \left( \frac{\mu_X(B_{\bar{z}}(r_\sigma)) - \mu_X(B_{\bar{z}}(f_{\bar{z}}(\rho_{\epsilon_-})))}{\mu_X(B_{\bar{z}}(r_\sigma))} \right) \\ & \leq \mu_X(B_{\bar{z}}(r_\sigma))n \frac{2k(s_{\bar{z}}(r_\sigma) - \rho_{\epsilon_-})}{s_{\bar{z}}(r_\sigma)} \\ & \leq \mu_X(B_{\bar{z}}(r_\sigma))n \left( \frac{k\sigma^2 t_2(2\sqrt{D} + t_2)}{s_{\bar{z}}(r_\sigma)} + \frac{2k\epsilon\sigma\sqrt{D}(1 + \frac{t_2(1+\xi^2 C_\xi)}{\sqrt{D}})r_\sigma}{s_{\bar{z}}(r_\sigma)^2} \right) \\ & \lesssim \mu_X(B_{\bar{z}}(r_\sigma))n \left( \frac{k\sigma t_2 \xi}{\sqrt{1-2\xi^2}} + k\epsilon \frac{\xi(1 + t_2(1 + \xi^2 C_\xi)D^{-\frac{1}{2}})}{1-2\xi^2} \right). \end{aligned}$$

Combining the above bounds for (7.11) and (7.12), one obtains:

$$\begin{aligned} \mathbb{E}[|Q_1|] & \lesssim \mu_X(B_{\bar{z}}(r_\sigma))n \\ & \cdot \left( e^{-\frac{1}{2}\epsilon^2(D-2)} + 2e^{-c(t_2^2 \wedge t_2\sqrt{D})} + \frac{k\sigma t_2 \xi}{\sqrt{1-2\xi^2}} + k\epsilon \frac{\xi(1 + t_2(1 + \xi^2 C_\xi)D^{-\frac{1}{2}})}{1-2\xi^2} \right) \end{aligned}$$

where  $C_\xi \leq \frac{1}{1-2\xi^2}$ . Letting  $\epsilon = \sqrt{\frac{2\log(\frac{\sqrt{D}}{\xi k})}{D}}$  and  $t_2 = c^{-\frac{1}{2}}\epsilon\sqrt{D}$ , and assuming  $\frac{1}{1-2\xi^2}$  is bounded by a universal constant, one obtains

$\mathbb{E}[|Q_1|] \lesssim \mu_X(B_{\tilde{z}}(r_\sigma))n \frac{\xi k}{\sqrt{D}} \left(1 + (\sigma\sqrt{D} + 1)\sqrt{\log\left(\frac{D}{\xi^2 k^2}\right)}\right)$ , and thus by Lemma 26:

$$\begin{aligned} & \|\text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2) - \text{cov}(\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2 \setminus Q_1)\| \\ & \lesssim s^2 \left( \frac{\xi k}{\sqrt{D}} \left(1 + (\sigma\sqrt{D} + 1)\sqrt{\log\left(\frac{D}{\xi^2 k^2}\right)}\right) \vee \frac{1}{\mu_X(B_{\tilde{z}}(r_{2\sigma}))n} \right) (r^2 + \sigma^2 D), \end{aligned}$$

with probability given in Table 7.1.

### 7.2.3 Comparing $\tilde{X}_{n,\tilde{z},r_\sigma} \cup Q_2$ and $\tilde{X}_{n,\tilde{z},r_\sigma}$

Finally,  $|Q_2|$  is estimated:

$$\begin{aligned} \mathbb{E}[|Q_2|] &= \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \text{ and } \|x_i - \tilde{z}\| \in [r_\sigma, r]) \\ &= \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\| \in [r_\sigma, f_{\tilde{z}}(\rho_{\epsilon_+})]) \cdot \mathbb{P}(\|x_i - \tilde{z}\| \in [r_\sigma, f_{\tilde{z}}(\rho_{\epsilon_+})]) \end{aligned} \tag{7.13}$$

$$+ \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\| \in [f_{\tilde{z}}(\rho_{\epsilon_+}), r]) \cdot \mathbb{P}(\|x_i - \tilde{z}\| \in [f_{\tilde{z}}(\rho_{\epsilon_+}), r]), \tag{7.14}$$

for any  $f_{\tilde{z}}(\rho_{\epsilon_+}) \in [r_\sigma, r]$ . For a point at distance  $R \in [r_\sigma, r]$  from  $\tilde{z}$ , the probability of not exiting when noise vector  $\eta$  is added is  $h(\theta_0(R, \|\eta\|))$ . Note that  $h(\theta_0(R, \|\eta\|))$  is decreasing in both  $R$  and  $\|\eta\|$ . By a similar argument to that in 7.2.1:

$$\begin{aligned} & \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\| \in [f_{\tilde{z}}(\rho_{\epsilon_+}), r]) \\ & \leq h(\theta_0(f_{\tilde{z}}(\rho_{\epsilon_+}), \sigma\sqrt{D}(1 - t_3 D^{-\frac{1}{2}}))) + 2e^{-c(t_3^2 \wedge t_3 \sqrt{D})}. \end{aligned}$$

$\rho_{\epsilon_+}$  is chosen so that  $\theta_0(f_{\tilde{z}}(\rho_{\epsilon_+}), \sigma\sqrt{D}(1 - t_3 D^{-\frac{1}{2}})) \leq \frac{\pi}{2} - \epsilon$ ; setting

$\theta_0(f_{\tilde{z}}(\rho_\epsilon), \sigma\sqrt{D}(1 - t_3 D^{-\frac{1}{2}})) := \frac{\pi}{2} - \epsilon$  one obtains

$$\cos\left(\frac{\pi}{2} - \epsilon\right) \approx \epsilon = \frac{r^2 - f_{\tilde{z}}(\rho_\epsilon)^2 - \sigma^2 D(1 - t_3 D^{-\frac{1}{2}})^2}{-2f_{\tilde{z}}(\rho_\epsilon)\sigma\sqrt{D}(1 - t_3 D^{-\frac{1}{2}})}.$$

Assuming  $\rho_\epsilon$  has the asymptotic expansion  $\rho_\epsilon = \rho_0 + \epsilon\rho_1 + \epsilon^2\rho_2 + \dots$ , one obtains as before as  $\epsilon \rightarrow 0$ :

$$\begin{aligned} \rho_\epsilon &\sim \sqrt{s_{\tilde{z}}(r_\sigma)^2 + \sigma^2 t_3(2\sqrt{D} - t_3)} + \sigma\sqrt{D}(1 - t_3 D^{-\frac{1}{2}}) \sqrt{\frac{r_\sigma^2 + \sigma^2 t_3(2\sqrt{D} - t_3)}{s_{\tilde{z}}(r_\sigma)^2 + \sigma^2 t_3(2\sqrt{D} - t_3)}} \epsilon \\ &< \sqrt{s_{\tilde{z}}(r_\sigma)^2 + \sigma^2 t_3(2\sqrt{D} - t_3)} + \sigma\sqrt{D}(1 - t_3 D^{-\frac{1}{2}}) \frac{r_\sigma}{s_{\tilde{z}}(r_\sigma)} \epsilon := \rho_{\epsilon_+}. \end{aligned}$$

Thus  $\theta_0(f_{\tilde{z}}(\rho_{\epsilon_+}), \sigma\sqrt{D}(1 - t_3 D^{-\frac{1}{2}})) \leq \frac{\pi}{2} - \epsilon$  and  $h(\theta_0(f_{\tilde{z}}(\rho_{\epsilon_+}), \sigma\sqrt{D}(1 - t_3 D^{-\frac{1}{2}}))) \leq e^{-\frac{1}{2}\epsilon^2(D-2)}$ . (7.13) is bounded as follows:

$$\begin{aligned} &\sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\| \in [r_\sigma, f_{\tilde{z}}(\rho_{\epsilon_+})]) \cdot \mathbb{P}(\|x_i - \tilde{z}\| \in [r_\sigma, f_{\tilde{z}}(\rho_{\epsilon_+})]) \\ &\leq \mu_X(B_{\tilde{z}}(f_{\tilde{z}}(\rho_{\epsilon_+})) \setminus B_{\tilde{z}}(r_\sigma)) n \\ &\leq \mu_X(B_{\tilde{z}}(r_\sigma)) n \left( \frac{\mu_X(B_{\tilde{z}}(f_{\tilde{z}}(\rho_{\epsilon_+}))) - \mu_X(B_{\tilde{z}}(r_\sigma))}{\mu_X(B_{\tilde{z}}(r_\sigma))} \right) \\ &\leq \mu_X(B_{\tilde{z}}(r_\sigma)) n \left( \frac{2k(\rho_{\epsilon_+} - s_{\tilde{z}}(r_\sigma))}{s_{\tilde{z}}(r_\sigma)} \right) \\ &\leq \mu_X(B_{\tilde{z}}(r_\sigma)) n \left( \frac{k\sigma^2 t_3(2\sqrt{D} - t_3)}{s_{\tilde{z}}(r_\sigma)} + \frac{2k\epsilon\sigma\sqrt{D}(1 - t_3 D^{-\frac{1}{2}})r_\sigma}{s_{\tilde{z}}(r_\sigma)^2} \right) \\ &\leq \mu_X(B_{\tilde{z}}(r_\sigma)) n \left( \frac{2k\sigma t_3 \xi}{\sqrt{1 - 2\xi^2}} + 2k\epsilon \frac{\xi(1 - t_3 D^{-\frac{1}{2}})}{1 - 2\xi^2} \right). \end{aligned}$$

(7.14) is also bounded by:

$$\begin{aligned}
& \sum_{i=1}^n \mathbb{P}(\|\tilde{x}_i - \tilde{z}\| < r \mid \|x_i - \tilde{z}\| \in [f_{\tilde{z}}(\rho_{\epsilon_+}), r]) \cdot \mathbb{P}(\|x_i - \tilde{z}\| \in [f_{\tilde{z}}(\rho_{\epsilon_+}), r]) \\
& \leq \left( e^{-\frac{1}{2}\epsilon^2(D-2)} + 2e^{-c(t_3^2 \wedge t_3 \sqrt{D})} \right) (\mu_X B_{\tilde{z}}(r) \setminus B_{\tilde{z}}(f_{\tilde{z}}(\rho_{\epsilon_+}))) n \\
& \leq \mu_X(B_{\tilde{z}}(r_\sigma)) n \left( e^{-\frac{1}{2}\epsilon^2(D-2)} + 2e^{-c(t_3^2 \wedge t_3 \sqrt{D})} \right) \left( \frac{\mu_X(B_{\tilde{z}}(r)) - \mu_X(B_{\tilde{z}}(r_\sigma))}{\mu_X(B_{\tilde{z}}(r_\sigma))} \right) \\
& \leq \mu_X(B_{\tilde{z}}(r_\sigma)) n \left( e^{-\frac{1}{2}\epsilon^2(D-2)} + 2e^{-c(t_3^2 \wedge t_3 \sqrt{D})} \right) \left( \left( \frac{s_{\tilde{z}}(r)}{s_{\tilde{z}}(r_\sigma)} \right)^{2k} - 1 \right) \\
& \leq \mu_X(B_{\tilde{z}}(r_\sigma)) n \left( e^{-\frac{1}{2}\epsilon^2(D-2)} + 2e^{-c(t_3^2 \wedge t_3 \sqrt{D})} \right) \frac{1}{(1-2\xi^2)^k}.
\end{aligned}$$

Combining the above bounds for (7.13) and (7.14):

$$\begin{aligned}
\mathbb{E}[|Q_2|] & \leq \mu_X(B_{\tilde{z}}(r_\sigma)) n \\
& \cdot \left( \frac{2k\sigma t_3 \xi}{\sqrt{1-2\xi^2}} + 2k\epsilon \frac{\xi(1-t_3 D^{-\frac{1}{2}})}{1-2\xi^2} + \left( e^{-\frac{1}{2}\epsilon^2(D-2)} + 2e^{-c(t_3^2 \wedge t_3 \sqrt{D})} \right) \frac{1}{(1-2\xi^2)^k} \right).
\end{aligned}$$

Choosing  $\epsilon = \sqrt{\frac{2 \log((1-2\xi^2)^{-k} \frac{\sqrt{D}}{\xi k})}{D}}$  and  $t_3 = c^{-\frac{1}{2}} \epsilon \sqrt{D}$ , one obtains

$\left( e^{-\frac{1}{2}\epsilon^2(D-2)} + 2e^{-c(t_3^2 \wedge t_3 \sqrt{D})} \right) (1-2\xi^2)^{-k} \lesssim \frac{\xi k}{\sqrt{D}}$ . Letting  $C_\xi = \frac{1}{1-2\xi^2}$ , one gets:

$$\begin{aligned}
\mathbb{E}[|Q_2|] & \lesssim \mu_X(B_{\tilde{z}}(r_\sigma)) n \frac{\xi k}{\sqrt{D}} \left( 1 + C_\xi (\sigma \sqrt{D} + 1) \sqrt{k \log C_\xi + \log\left(\frac{D}{\xi^2 k^2}\right)} \right) \\
& \lesssim \mu_X(B_{\tilde{z}}(r_\sigma)) n \frac{\xi k}{\sqrt{D}} \left( 1 + (\sigma \sqrt{D} + 1) \sqrt{k \log\left(\frac{1}{1-2\xi^2}\right) + \log\left(\frac{D}{\xi^2 k^2}\right)} \right),
\end{aligned}$$

for  $C_\xi$  bounded by a universal constant. Thus from Lemma 26:

$$\begin{aligned}
& \|\text{cov}(\tilde{X}_{n, \tilde{z}, r_\sigma} \cup Q_2) - \text{cov}(\tilde{X}_{n, \tilde{z}, r_\sigma})\| \\
& \lesssim \left( \frac{s^2 \xi k}{\sqrt{D}} \left( 1 + (\sigma \sqrt{D} + 1) \sqrt{k \log\left(\frac{1}{1-2\xi^2}\right) + \log\left(\frac{D}{\xi^2 k^2}\right)} \right) \vee \frac{1}{\mu_X(B_{\tilde{z}}(r_{2\sigma})) n} \right) (r^2 + \sigma^2 D),
\end{aligned}$$

with probability given in Table 7.1.

### 7.3 Summary

Finally, all of these perturbation results are combined. From Table 7.1, one sees that  $\delta_5$  dominates  $\delta_1, \delta_3, \delta_4$ ;  $\delta_2$  competes with  $\delta_5$  for worst term, but:

$$\delta_2 \leq \frac{\xi k}{\sqrt{D}} \left( s^2 \xi + \sqrt{\log\left(\frac{D}{\xi^2 k^2}\right)} \right), \quad \delta_5 \leq \frac{\xi k}{\sqrt{D}} \left( 1 + (\sigma\sqrt{D} + 1) \sqrt{k \log C_\xi + \log\left(\frac{D}{\xi^2 k^2}\right)} \right),$$

where  $C_\xi = \frac{1}{1-2\xi^2}$ . Thus  $\max_i \delta_i \leq \frac{\xi k}{\sqrt{D}} \left( 1 + s^2 \xi + (\sigma\sqrt{D} \vee 1) \sqrt{k \log C_\xi + \log\left(\frac{D}{\xi^2 k^2}\right)} \right)$ . Replacing each  $\delta_i$ , for  $1 \leq i \leq 5$ , by  $\max_i \delta_i$  (on  $\Omega_{s,i}$ ,  $\delta_i$  may always be replaced by an upper bound), one obtains that for  $k^3 \lesssim D$ ,  $1 \leq s^2 \leq \sqrt{D} \left( 1 \wedge \frac{c}{\xi^2 k} \right)$ ,  $n \geq t^2 / \mu_X(B_z(r_{2\sigma}))$ ,  $\xi \in \{0\} \cup \left[ \sqrt{\frac{\log D}{D}}, \frac{1}{\sqrt{3}} - \tau \right]$  for some  $\tau > 0$ , and conditioned on  $\Omega_{s,0}$ :

$$\|\text{cov}(\widetilde{X_{n,\bar{z},r}}) - \text{cov}(\tilde{X}_{n,z,r_{2\sigma}})\| \lesssim s^2 \left( \delta \vee \frac{1}{\mu_X(B_z(r_{2\sigma}))n} \right) r^2, \quad (7.15)$$

with probability at least  $1 - Ce^{-cs^2[\delta\mu_X(B_z(r_{2\sigma}))n \vee 1]} - Ce^{-ct^2}$ , where

$$\delta = \frac{\xi k}{\sqrt{D}} \left( 1 + s^2 \xi + (\sigma\sqrt{D} \vee 1) \sqrt{k \log C_\xi + \log\left(\frac{D}{\xi^2 k^2}\right)} \right). \text{ Removing the conditioning on } \Omega_{s,0}, (7.15) \text{ holds with probability at least } 1 - Ce^{-cs^2 \min\{(\delta\mu_X(B_z(r_{2\sigma}))n) \vee 1, s^2\}} - Ce^{-ct^2}.$$

Finally, one must determine the restrictions on  $r_{2\sigma}$  in terms of  $R_{\min}, R_{\max}$ , which determine the range where the volume growth and covariance estimation assumptions given in (5.1) and (5.2) hold.

Assumptions related to volume growth:

- Sec. 7.1 assumed  $r_{2\sigma}^2 \in [R_{\min}^2 + q, R_{\max}^2 - q]$ , where

$$q = \frac{C\xi}{\sqrt{D}} \left( \xi s^2 + \sqrt{\log(Dk^{-2}\xi^{-2})} \right) r_{2\sigma}^2.$$

- Sec. 7.2 assumed  $r_{2\sigma} \in [R_{\min}, R_{\max} - \sigma_0]$ .

Assumptions related to covariance estimates:

- Sec. 7.1 assumed  $r_{2\sigma}^2 - q \in [R_{\min}^2, R_{\max}^2]$ .

- Sec. 7.1 assumed  $r_\sigma + \sigma_0 \left(1 + \frac{s^2}{\sqrt{D}}\right) \in [R_{\min}, R_{\max}]$ .

A straightforward calculation shows that is thus sufficient for:

$$\begin{aligned} r_{2\sigma} &> \hat{R}_{\min} := R_{\min} \left( 1 + \frac{C\xi}{\sqrt{D}} \left( \xi s^2 + \sqrt{\log \frac{D}{k^2 \xi^2}} \right) \right)^{\frac{1}{2}} \\ r_{2\sigma} &< \hat{R}_{\max} := R_{\max} \left( 1 - C\xi \left( 1 + \frac{s^2}{\sqrt{D}} \right) \right) \end{aligned} \tag{7.16}$$

Since  $s^2 \leq \sqrt{D}$  and  $\xi^2 \geq \frac{\log D}{D}$ , it is sufficient for:

$$R_{\min}(1 + C\xi) < r_{2\sigma} < R_{\max}(1 - C\xi).$$

**Remark 20.** *Observe that as  $D \rightarrow \infty$ , (7.16) in fact gives the sufficient condition  $R_{\min} < r_{2\sigma} < R_{\max}(1 - C\xi)$ .*

## Algorithm

## 8.1 Pseudo-code

The algorithm estimates the intrinsic dimension of the data as follows: for every  $\tilde{z} \in \tilde{X}_n$ ,  $r \leq \text{diam}(\tilde{X}_n)$ , and  $1 \leq i \leq D$ ,  $\tilde{\lambda}_{i,z,r}^2$  is computed. One then identifies the largest noise squared singular value, and discards the noise SSV's as well as scales that are too small, where the noise dominates. One also discards scales that are too large, where the top SSV,  $\tilde{\lambda}_{1,z,r}^2$ , ceases to increase. As illustrated in Chap. 4, for nice data sets the SSV's corresponding to an intrinsic dimension ("tangent SSV's") grow like  $r^2$ , whereas the SSV's arising from the data curving into more dimensions ("curvature SSV's") grow like  $r^4$ ; that is, for  $r$  not too large, the curvature SSV's are quadratic with respect to the tangent SSV's. Thus on the restricted range of scales, using a least squares regression, one classifies the non-noise SSV's as follows: if the  $\tilde{\lambda}_{i,z,r}^2$  grows linearly, then it is a tangent SSV; if the  $\tilde{\lambda}_{i,z,r}^2$  grows quadratically, then it is a curvature SSV. The intrinsic dimension at  $\tilde{z}$ , denoted  $\hat{k}$ , is then estimated by counting the number of  $\tilde{\lambda}_{i,z,r}^2$  that were classified as tangent SSV's; the algorithm also returns the range of scales where  $\tilde{\Delta}_{\hat{k},z,r}$  is the largest gap, denoted  $(\hat{R}_{\min}, \hat{R}_{\max})$ .

```

 $[\hat{k}, \hat{R}_{\min}, \hat{R}_{\max}] = \text{EstDimMSVD}(\tilde{X}_n, \tilde{z}, K_{\max})$ 
// Input:
//  $\tilde{X}_n$  : an  $n \times D$  set of noisy samples
//  $\tilde{z}$  : a point in  $\tilde{X}_n$ 
//  $K_{\max}$  : upper bound on the intrinsic dimension  $k$ 
// Output:
//  $\hat{k}$  : estimated intrinsic dimension at  $\tilde{z}$ 
//  $(\hat{R}_{\min}, \hat{R}_{\max})$  : estimated interval of good scales

 $\{\hat{k}_1, \tilde{\lambda}_{\hat{k}_1+1, z, r}^2\} \leftarrow \text{FindLargestNoiseSingularValue}(\tilde{X}_n, \tilde{z}, \text{Nets})$ 
 $\hat{R}_{\min} \leftarrow \text{Smallest scale for which } \tilde{\lambda}_{\hat{k}_1+1, z, r}^2 \text{ is decreasing and } |B_z(\hat{R}_{\min})| \gtrsim K_{\max} \log K_{\max}$ 
 $\hat{R}_{\max} \leftarrow \text{Largest scale } > \hat{R}_{\min} \text{ for which } \tilde{\lambda}_{1, z, r}^2 \text{ is non-increasing}$ 
 $\hat{k} \leftarrow \text{Largest } i \text{ such that:}$ 

- for  $r \in (\hat{R}_{\min}, \hat{R}_{\max})$ ,  $\tilde{\lambda}_{i, z, r}^2$  is linear and  $\tilde{\lambda}_{i+1, z, r}^2$  is quadratic in  $r$ , and
- $\Delta_{i, z, r}$  is largest gap for  $r$  in a large fraction of  $(\hat{R}_{\min}, \hat{R}_{\max})$

 $(\hat{R}_{\min}, \hat{R}_{\max}) \leftarrow \text{Largest interval in which } \Delta_{\hat{k}, z, r} \text{ is the largest gap}$ 

```

FIGURE 8.1: Pseudo-code for the ID estimator based on multiscale SVD.

To obtain a global estimate of the intrinsic dimension, simply estimate the intrinsic dimension at each point, and then either average or take a majority vote over the points. However, such a global estimate is not meaningful for many data sets, such as data sets arising from samples of a stratification of manifolds of different dimensions, as well as many real-world data sets, where the intrinsic dimension may also vary throughout the data.

**Remark 21.** *Despite the fact that at small scales the tangent SSV's grow like  $r^2$ , at larger scales the growth is actually very close to linear; a curve that starts off as a quadratic must eventually “level off” as  $r$  approaches  $\text{diam}(\tilde{X}_n)$ , and there is thus an inflection point near which ones sees linear growth; in this region where the tangent SSV's are linear, the curvature SSV's are still convex; because of noise and sampling constraints, one is generally required to work at these larger scales, which is why*

*the algorithm tests for linear vs. quadratic growth instead of quadratic vs. quartic growth.*

The algorithm performs a sub-sampling in space and scale to obtain discretized versions of the  $\tilde{\lambda}_{i,z,r}^2$ . First of all, the scale is discretized; the  $\tilde{\lambda}_{i,z,r}^2$ , which are continuous functions of  $r$ , will be computed only for  $\{r_j\}_{j=0}^J$ . For each scale  $r_j$ , one then computes an  $r_j$ -net, which will be denoted  $\Gamma_j$ , of  $\tilde{X}_n$ .  $\Gamma$  is an  $r$ -net of a metric space  $X$  if every  $x \in X$  is contained in the closed ball  $B_z(r)$  for some  $z \in \Gamma$  and for every distinct  $z_1, z_2 \in \Gamma$ , the distance between  $z_1$  and  $z_2$  is at least  $\frac{r}{\alpha}$ , some  $\alpha \geq 1$  (Har-Peled and Mendel (2006)). How to efficiently construct the multiscale nets  $\Gamma_j$  has been the subject of much investigation (Har-Peled and Mendel (2006); Beygelzimer et al. (2006)). One then computes  $\tilde{\lambda}_{i,z,r}^2$  for every  $r \in r_j, \tilde{z} \in \Gamma_j, i \leq I_j$ , where  $I_j = \min(D, n_{r_j}, K)$  and  $K$  is a user-specified parameter. Note that one still obtains an estimate of the intrinsic dimension *at every point*: given an  $\tilde{x} \in \tilde{X}_n$ , at each scale  $j$ ,  $\tilde{x}$  is associated with the  $\tilde{z}_{\tilde{x},j} \in \Gamma_j$  that is closest to it, and  $\tilde{\lambda}_{i,x,r_j}^2$  is approximated by  $\tilde{\lambda}_{i,\tilde{z}_{\tilde{x},j},r_j}^2$ . To increase the stability of these pointwise estimates (the construction of the  $\Gamma_j$  being a random process), one can run the algorithm multiples times and average the results.

## 8.2 Computational Complexity

The construction of multiscale nets is a preprocessing of the data that allows for efficient nearest neighbor queries; this data structure can be computed in time  $O(D \cdot 2^{ck} n \log n)$  and enables performing a nearest neighbor query in time  $O(D \cdot 2^{ck} \log n)$ , where  $c$  is a universal constant; let  $C_{nn} = O(2^{ck} \log n)$  denote the (normalized) cost of performing a nearest neighbor query after this preprocessing. To compute the intrinsic dimension at every point in a data set  $\tilde{X}_n$  using the sub-sampling in space and scale described above, one must compute  $\tilde{\lambda}_{i,z,r}^2$  for each  $\tilde{z} \in \Gamma_j$  at every scale  $j$ .

For a fixed scale  $j$  and  $\tilde{z} \in \Gamma_j$ , this involves:

1. Finding the points within distance  $r_j$  of  $\tilde{z}$ .
2. Computing an SVD of these points.

For running time, this will cost:

1.  $(|B_{\tilde{z}}(r_j)| + 1) \cdot D \cdot C_{nn}$ : one must perform  $|B_{\tilde{z}}(r_j)|$  nearest neighbor queries before the next query will return a point that is too far from  $\tilde{z}$ .
2.  $O(|B_{\tilde{z}}(r_j)| \cdot D \cdot I_j)$ : using randomized algorithms, an approximate  $k - SVD$  of an  $n$  by  $D$  matrix can be constructed in time  $O(nDk)$ , see for example Rokhlin et al. (2009).

Since there are  $O(n/|B_{\tilde{z}}(r_j)|)$  points in  $\Gamma_j$ , at each scale one incurs a cost of  $O(n \cdot D \cdot (C_{nn} + I_j))$ ; note that one always has  $I_j \leq \min\{n, D, K\} := I$ . Choosing  $J = \log n$  and summing over scales yields a cost of  $O(n \log n \cdot D \cdot (C_{nn} + I)) = O(n \log n \cdot D \cdot (2^{ck} n \log n + \min\{n, D, K\}))$ .

This algorithm was implemented in matlab by M. Maggioni and the code may be accessed at [www.math.duke.edu/~mauro](http://www.math.duke.edu/~mauro).

## 8.3 Numerical Experiments

This section describes the results of numerical experiments testing the performance of the MSVD algorithm on both manifold and real-world data sets and comparing its performance with competing algorithms for dimension estimation. M. Maggioni ran these experiments and compiled the results.

### 8.3.1 Manifold data

The algorithm was run on numerous manifold data sets, including the  $k$ -dimensional unit cube  $\mathbb{Q}^k$ , the  $k$ -dimensional unit sphere  $\mathbb{S}^k$ , the 2-dimensional  $S$ -shaped manifold

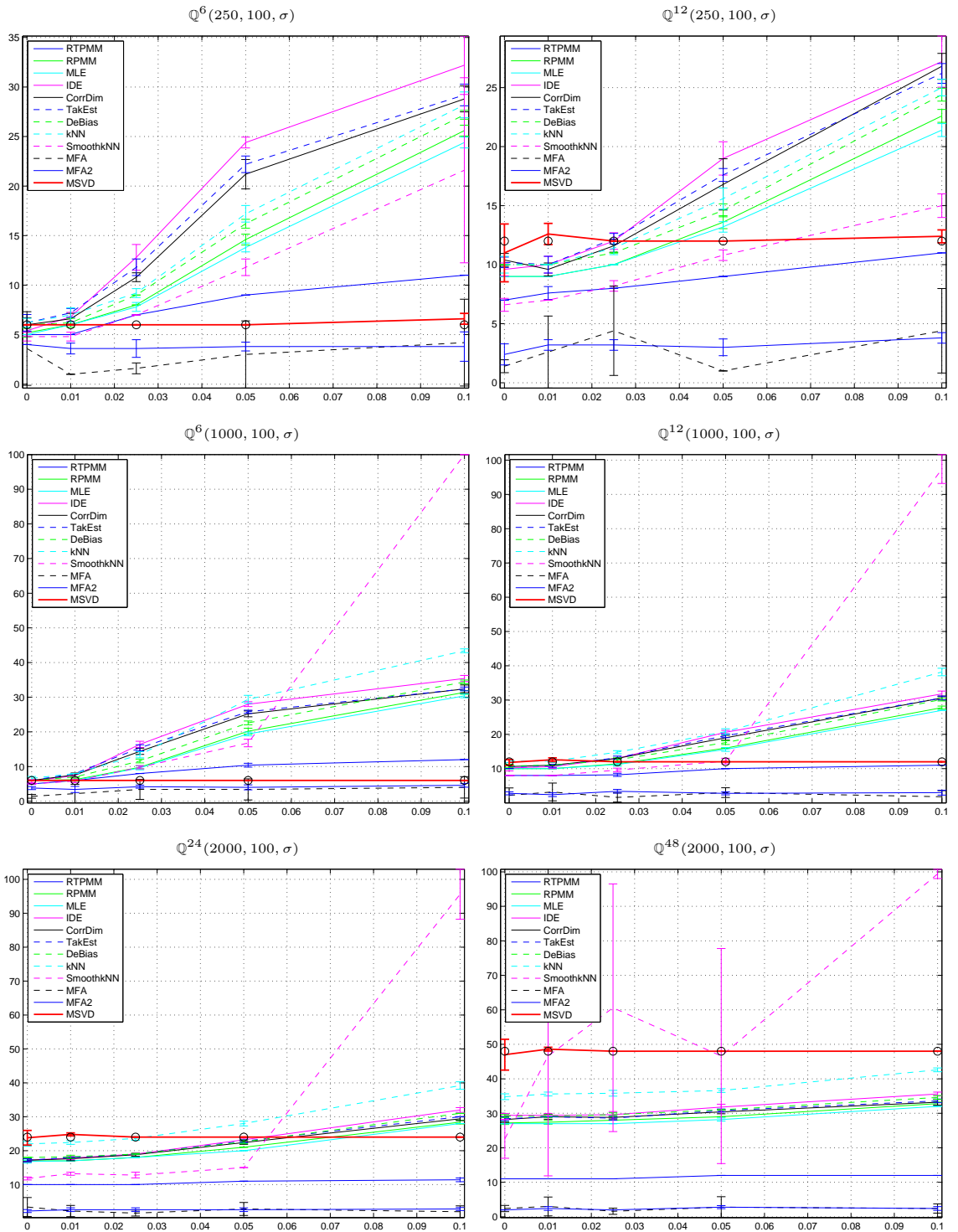


FIGURE 8.2: Manifold data sets: cube.

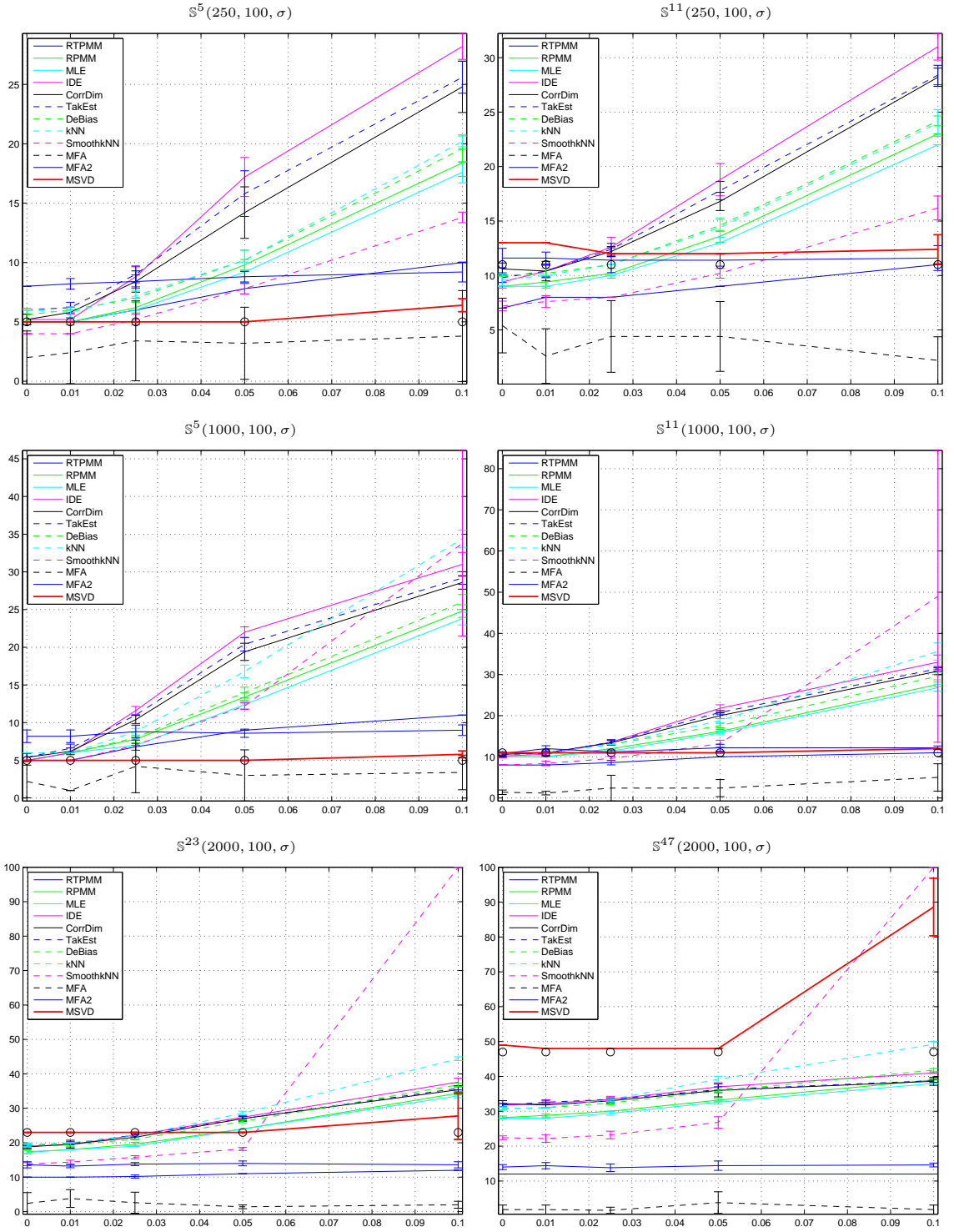


FIGURE 8.3: Manifold data sets: sphere. Note that the MSVD algorithm may fail if the noise is too high (e.g.  $\mathbb{S}^{47}(2000, 100, \sigma)$ ) or the intrinsic dimension is large relative to the sample size (e.g.  $\mathbb{S}^{11}(250, 100, \sigma)$ ).

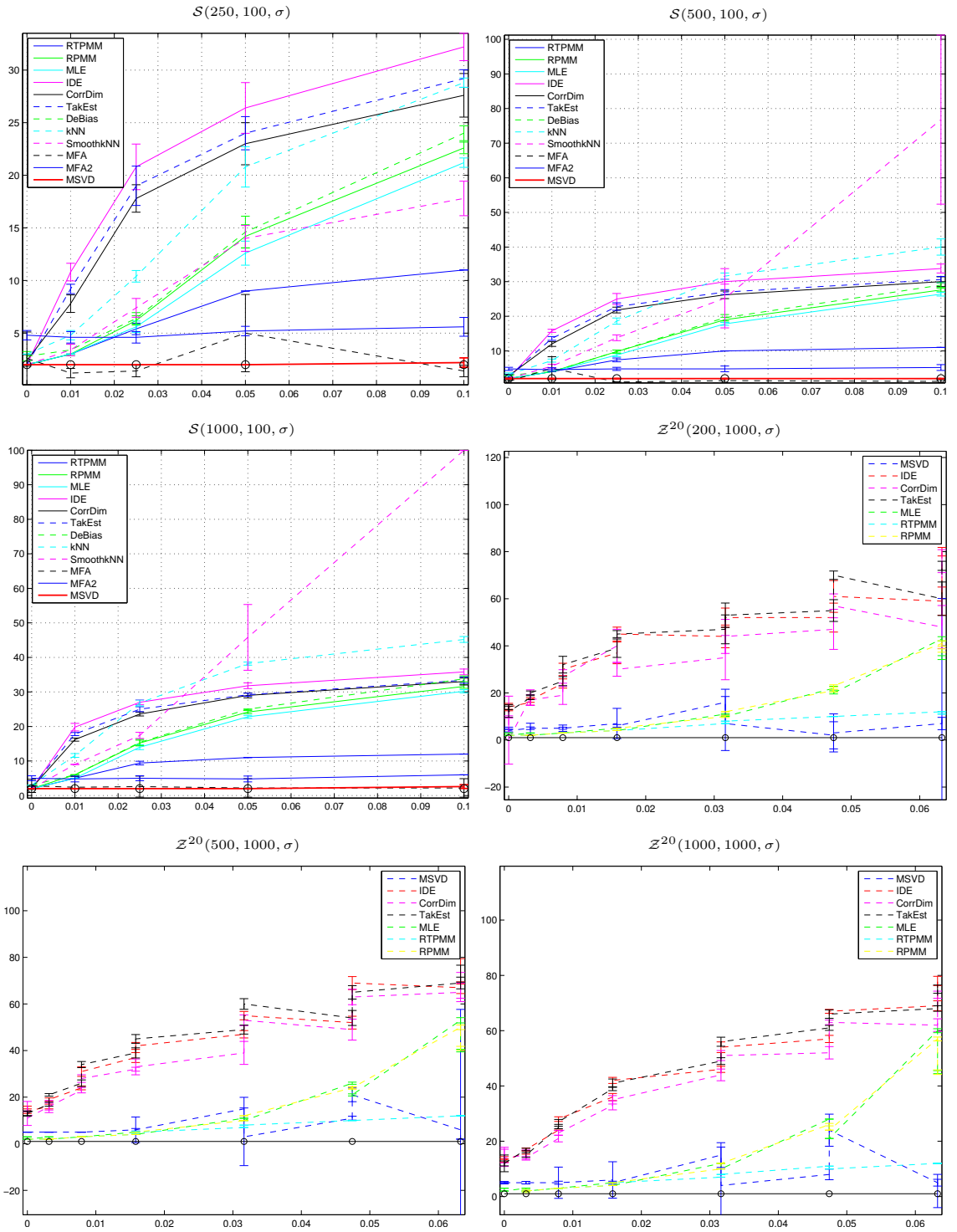


FIGURE 8.4: Manifold data sets: S-shaped manifold  $\mathcal{S}$  and Meyer's staircase  $\mathcal{Z}^{20}$ .

Table 8.1: Dimension estimates for various manifolds;  $n = 1000$ ,  $D = 100$ , and  $\sigma = 0$ .

	RTPMM	RPMM	MLE	IDE	CorrDim	TakEst	DeBias	kNN	SmoothkNN	MFA	MFA2	MSVD
$\mathbb{Q}^6$	5	5	5	<b>6</b>	5	5	<b>6</b>	<b>6</b>	4	1	4	<b>6</b>
$\mathbb{Q}^{12}$	7	9	9	10	10	10	10	<b>12</b>	7	1	3	<b>12</b>
$\mathbb{Q}^{24}$	9	16	16	17	17	17	17	20	11	1	2	<b>24</b>
$\mathbb{Q}^{48}$	11	26	25	29	28	28	28	32	19	1	2	<b>48</b>
$\mathbb{S}^5$	4	<b>5</b>	<b>5</b>	<b>5</b>	<b>5</b>	<b>5</b>	<b>5</b>	<b>5</b>	4	1	9	<b>5</b>
$\mathbb{S}^{11}$	7	9	9	10	10	10	10	10	8	1	12	<b>11</b>
$\mathbb{S}^{23}$	10	17	16	18	18	18	18	18	13	1	14	<b>23</b>
$\mathbb{S}^{47}$	11	27	26	31	30	31	29	29	21	1	14	48

$\mathcal{S}$  (the product of an  $S$ -shaped curve with the unit interval), and the 1-dimensional Meyer’s staircase  $\mathcal{Z}^{20}$  constructed with width parameter 20.<sup>1</sup> Each of these manifolds was embedded in  $\mathbb{R}^D$  according to the natural embedding via the first  $K$  coordinates, where  $K = k, k + 1, 3, D$  for  $\mathbb{Q}^k, \mathbb{S}^k, \mathcal{S}, \mathcal{Z}^{20}$  respectively;  $n$  uniform samples  $x_1, \dots, x_n$  were drawn, and  $D$ -dimensional Gaussian noise of variance  $\sigma^2$  was added to obtain noisy samples  $\tilde{x}_1, \dots, \tilde{x}_n$ , that is,  $\tilde{x}_i = x_i + \eta_i$ , where  $\eta_i \sim \sigma\mathcal{N}(0, I_D)$ . Finally, a random rotation was applied to the data. Throughout this section,  $\mathbb{Q}^k(n, D, \sigma)$  denotes the noisy samples of  $\mathbb{Q}^k$  obtained via the above procedure; similarly for  $\mathbb{S}^k(n, D, \sigma)$ ,  $\mathcal{S}(n, D, \sigma)$ , and  $\mathcal{Z}^{20}(n, D, \sigma)$ . The following values of the parameters  $k, n, D, \sigma$  were considered:

- $k = 6, 12, 24, 48$  for  $\mathbb{Q}^k$ ;  $k = 5, 11, 23, 47$  for  $\mathbb{S}^k$
- $n = 250, 500, 1000, 2000$  for  $\mathbb{S}^k, \mathbb{Q}^k, \mathcal{S}$ ;  $n = 200, 500, 1000, 2000$  for  $\mathcal{Z}^{20}$
- $D = 100$  for  $\mathbb{S}^k, \mathbb{Q}^k, \mathcal{S}$ ;  $D = 1000$  for  $\mathcal{Z}^{20}$

---

<sup>1</sup> Meyer’s staircase is a one-dimensional curve which can be constructed as follows: fix a width parameter  $w < D$ , which will dictate how much curvature is present, and consider the indicator functions  $\{x_i(t) = \mathbf{1}_{[0,w]}(t - i)\}_{i=0}^{D-1}$  on the interval  $[0, D]$ , where the input is taken to be modulo  $D$ , so that each indicator function has width  $w$ . The image of the map  $f : t \rightarrow (x_0(t), x_1(t), \dots, x_{D-1}(t))$  is a set of  $D$  distinct points on the sphere in  $\mathbb{R}^D$  of radius  $\sqrt{w}$ . By linear interpolation or by smoothing out the indicator functions, one gets a one-dimensional curve in  $\mathbb{R}^D$ , spiraling into all  $D$  dimensions such that  $f(t_1)$  and  $f(t_2)$  are orthogonal whenever  $|t_2 - t_1| > w$ .

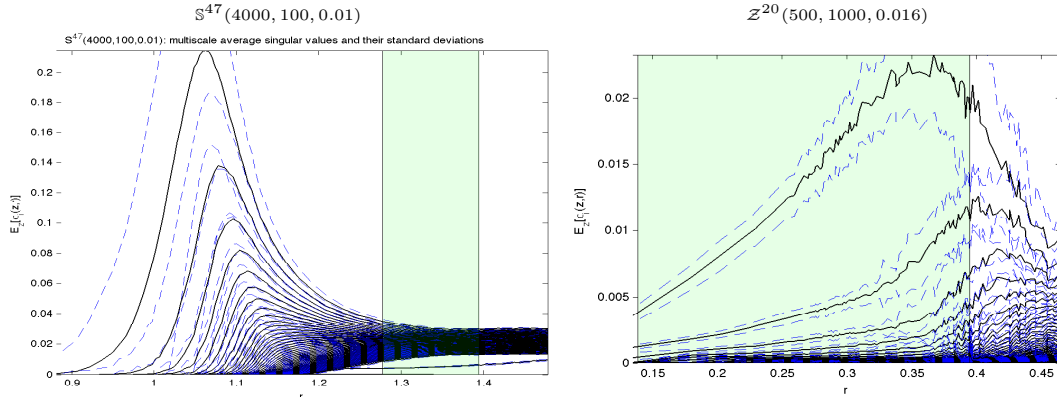


FIGURE 8.5: Two challenging examples:  $\mathbb{S}^{47}(4000, 100, 0.01)$  (left) has large intrinsic dimension which is difficult to detect automatically, although the curvature SSV is clearly distinguishable from the tangent SSV's;  $\mathbb{Z}^{20}(500, 1000, 0.016)$  (right) is clearly one-dimensional at small scales, while at large scales it has the covariance structure of a sphere.

- $\sigma = 0, 0.01, 0.025, 0.05, 0.1$ .

For each combination of the above parameters, the MSVD and competing algorithms were run five times, and the results were averaged and rounded to the nearest integral dimension. The MSVD algorithm was compared with the following competing algorithms: “Debiasing” Carter et al. (2007), “RPMM” Haro et al. (2008), “MLE” Levina and Bickel (2005), “kNN” Costa and Hero (2004), “SmoothKNN” Carter and Hero (2008), “MFA” Chen et al. (2010b), “MFA2” Chen et al. (2010a). Figures 8.2, 8.3, 8.4 and Table 8.1 give a sample of the results. Figure 8.5 discusses two data sets that are challenging to the MSVD algorithm.

### 8.3.2 Collections of manifolds

The algorithm was also run on several data sets consisting of a collection of manifolds of different dimensionalities, for example, points sampled from the union of a sphere and a line segment and also the union of a tight, noisy 1-dimensional spiral with a noisy 2-dimensional plane, an example given in Haro et al. (2008); see Fig. 8.6. By estimating the dimensionality at each point in the data set, the MSVD algorithm

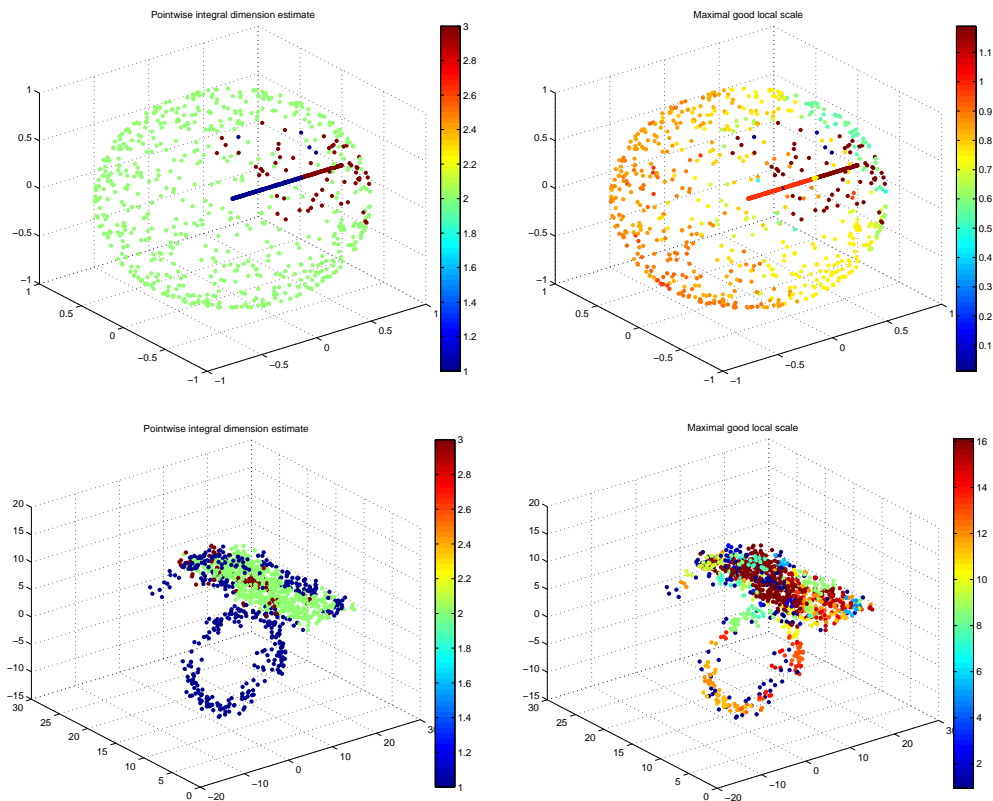


FIGURE 8.6: Pointwise intrinsic dimension (left) and pointwise good range of scales (right) estimated by the MSVD algorithm for the union of a sphere and line segment (top) and the union of a noisy 1-dimensional spiral with a noisy 2-dimensional plane (bottom).

is able to separate the points on the 2-dimensional sphere from the points on the 1-dimensional line segment. Note that points near the intersection of the sphere and the line segment are estimated as 3-dimensional. For the spiral and plane example, 77% of the points on the spiral are estimated to have dimension less than 2 while 86% of points on the plane are estimated to have dimension at least 2, yielding an overall classification accuracy rate of 86%; the state-of-art classification algorithm given in Haro et al. (2008) yields a much higher accuracy rate of 97%, but requires the user to input the number of clusters and also relies on each cluster being smooth, obtaining better estimates by smoothing across neighborhoods. These examples illustrates the

Table 8.2: Estimated intrinsic dimension of each handwritten digit as returned by the MSVD, IDE (smoothed Grassberger-Procaccia estimator, Hein and Audibert (2005)), and HRVQ (high rate vector quantization method, Raginsky and Lazebnik (2005)) algorithms respectively.

Digit	0	1	2	3	4	5	6	7	8	9
<b>MSVD</b>	2	2	3	2	2	2	2	2	3	3
<b>IDE</b>	11	7	13	13	12	11	10	13	11	11
<b>HRVQ</b>	16	7	21	20	17	20	16	16	20	17

usefulness of the MSVD algorithm in problems of classification.

### 8.3.3 Real-world data

The MSVD algorithm was also run on multiple real world data sets, including the MNIST database of handwritten digits, the ISOMAP faces database, the face video database, the CBCL faces database, and the Science News dataset.

The MNIST dataset (available at <http://yann.lecun.com/exdb/mnist>) is a collection of several thousand 28 by 28 pixel greyscale images of the digits 0 through 9, handwritten by different individuals; thus the images of each handwritten digit contain a large amount of variability yet share a similar underlying structure. The MSVD algorithm, treating the dataset as a point cloud in  $\mathbb{R}^{28 \times 28}$ , seeks to detect the number of degrees of freedom in the images of each handwritten digit, see Table 8.2.

The ISOMAP faces database (<http://isomap.stanford.edu/dataset.html>) consists of 698 64 by 64 pixel images of faces. The MSVD algorithm estimates the intrinsic dimension to be  $k = 2$ , see Fig. 8.7; competing algorithms (Kegl (2002), Fan et al. (2009), Hein and Audibert (2005), Farahmand et al. (2007), Levina and Bickel (2005)) estimate  $k \in [3, 4.65]$ . The face video database (available at <http://www.cs.toronto.edu/~roweis/data.html>) consists of 1965 20 by 28 pixel images. The MSVD algorithm again estimates an intrinsic dimension of  $k = 2$ , while Raginsky and Lazebnik (2005) estimate  $k \in [4.25, 8.30]$ . A data set with previously unanalyzed ID was also considered: the CBCL faces database (available at

<http://cbcl.mit.edu>), which consists of 472 19 by 19 pixel images of faces; once again, the average intrinsic dimension was estimated to be  $k = 2$ .

The Science News dataset (see Priebe et al. (2004); Coifman and Maggioni (2005)) was constructed as follows: a large number of articles relating to various scientific subfields were collected, certain key words were selected, and the number of times each key word appeared in each article was recorded. Given  $D$  keywords, each article is represented as a vector in  $\mathbb{R}^D$  with coordinates given by the word counts (this dataset was constructed from  $n = 1161$  articles and  $D = 1153$  keywords). The MSVD algorithm estimates an average intrinsic dimension of  $k = 9$ , see Fig. 8.7; in this context, “intrinsic dimension” can be thought of as the number of “independent” scientific disciplines, as measured by the vocabulary prominent in each discipline.

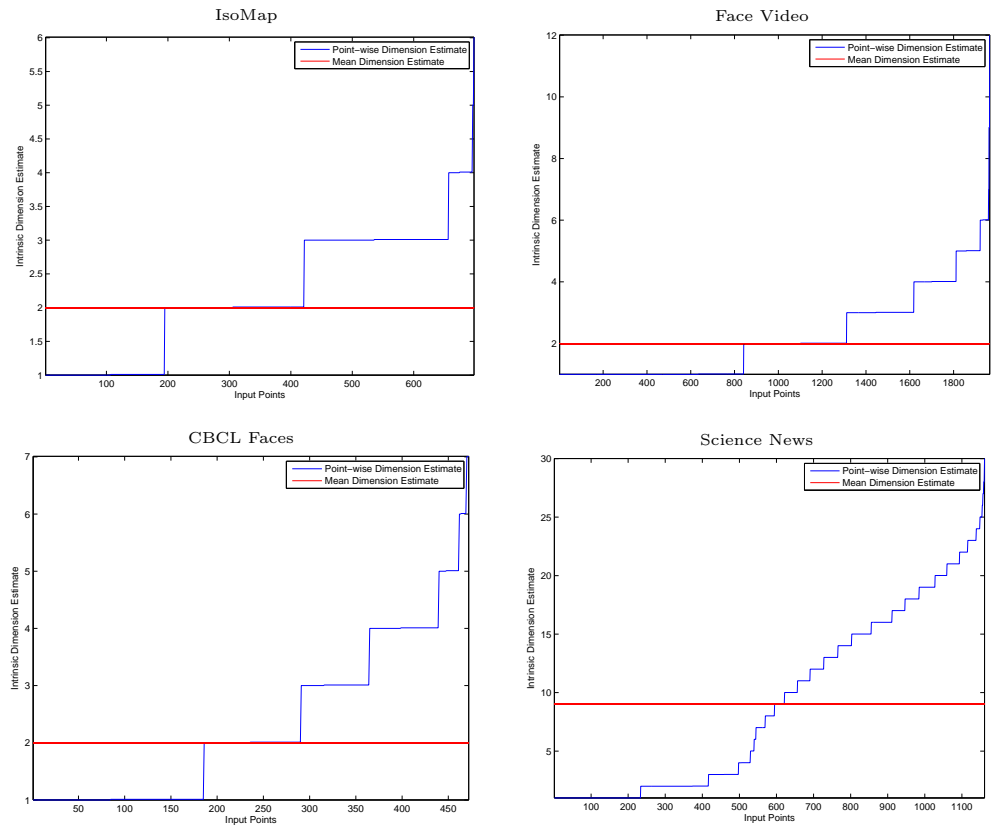


FIGURE 8.7: Estimated intrinsic dimension for the ISOMAP faces database (top left), the face video database (top right), the CBCL faces database (bottom left) and the Science News dataset (bottom right); the blue line shows the estimated intrinsic dimension as a function of point index, while the red line shows the average (across points) estimated intrinsic dimension.

## Extensions and Applications

### 9.1 Bi-Lipschitz Perturbations

The following lemma bounds the perturbation of the covariance under a bi-Lipschitz map. Further analysis is needed to see precisely how Theorem 11 might be extended to this case.

**Lemma 22** (Bi-Lipschitz perturbations). *Let  $X_n := \{x_1, x_2, \dots, x_n\}$  be  $n$  (centered) points in  $\mathbb{R}^D$  with  $\|x_i\| \leq r$  for  $1 \leq i \leq n$  and let  $\Phi$  be an  $\epsilon$  bi-Lipschitz map on these points:*

$$(1 - \epsilon)\|x_i - x_j\|^2 \leq \|\Phi x_i - \Phi x_j\|^2 \leq (1 + \epsilon)\|x_i - x_j\|^2. \quad (9.1)$$

*Then:*

$$|\lambda_i^2(\text{cov}(X_n)) - \lambda_i^2(\text{cov}(\Phi X_n))| \leq 2\epsilon M^2,$$

*where  $\Phi X_n := \{\Phi x_1, \Phi x_2, \dots, \Phi x_n\}$ .*

In addition to denoting a set, let  $X_n$  (resp.  $\Phi X_n$ ) also denote the  $n$  by  $D$  matrix whose  $i^{\text{th}}$  row is the point  $x_i$  (resp.  $\Phi x_i$ ); note that since the  $x_i$  have mean zero so do the  $\Phi x_i$ . Now  $\text{cov}(X_n) = \frac{1}{n} X_n^T X_n$  has the same non-zero eigenvalues

as  $\frac{1}{n}X_nX_n^T$ , similarly for  $\text{cov}(\Phi X_n)$  and  $\frac{1}{n}\Phi X_n(\Phi X_n)^T$ , and  $(\frac{1}{n}X_n^T X_n)_{i,j} = \frac{1}{n}\langle x_i, x_j \rangle$  while  $(\frac{1}{n}\Phi X_n(\Phi X_n)^T)_{i,j} = \frac{1}{n}\langle \Phi x_i, \Phi x_j \rangle$ .

Let  $\mathcal{D} = \{x_i - x_j : x_i, x_j \in X_n\}$  be the set of all differences between the points. For any  $x_i - x_j \in \mathcal{D}$ , one has:

$$\begin{aligned} \langle \Phi^T \Phi(x_i - y_j), x_i - y_j \rangle &:= \langle (I + E)(x_i - y_j), x_i - y_j \rangle \\ &= \|x_i - y_j\|^2 + \langle E(x_i - y_j), x_i - y_j \rangle \\ &= \|x_i - y_j\|^2 \left( 1 + \frac{\langle E(x_i - y_j), x_i - y_j \rangle}{\|x_i - y_j\|^2} \right). \end{aligned}$$

The bi-Lipschitz condition thus gives:

$$\frac{|\langle E(x_i - y_j), x_i - y_j \rangle|}{\|x_i - y_j\|^2} \leq \epsilon \quad \text{for all } x_i - y_j \in \mathcal{D}.$$

Because  $E$  is symmetric,  $\|E|_{\mathcal{D}}\| = \max_{x_i - y_j \in \mathcal{D}} \frac{|\langle E(x_i - y_j), x_i - y_j \rangle|}{\|x_i - y_j\|^2} \leq \epsilon$  by the above.

Thus on the set  $\mathcal{D}$ ,  $\Phi^T \Phi = I + E$  for some  $\|E|_{\mathcal{D}}\| \leq \epsilon$ , that is,  $\Phi^T \Phi$  is close to the identity on the set  $\mathcal{D}$ . This fact is now used to show that  $\langle \Phi x_1, \Phi x_2 \rangle$  is close to  $\langle x_1, x_2 \rangle$ .

Observe that because the points have mean zero,  $x_1$  can be written as  $\frac{1}{n} \sum_{i=1}^n (x_1 - x_i)$ . Thus:

$$\begin{aligned} \langle \Phi^T \Phi x_1, x_2 \rangle &= \left\langle \frac{1}{n} \sum_{i=1}^n \Phi^T \Phi(x_1 - x_i), x_2 \right\rangle \\ &= \left\langle \frac{1}{n} \sum_{i=1}^n (I + E)(x_1 - x_i), x_2 \right\rangle \\ &= \left\langle \frac{1}{n} \sum_{i=1}^n (x_1 - x_i), x_2 \right\rangle + \frac{1}{n} \sum_{i=1}^n \langle E(x_1 - x_i), x_2 \rangle \\ &= \langle x_1, x_2 \rangle + \frac{1}{n} \sum_{i=1}^n \langle E(x_1 - x_i), x_2 \rangle. \end{aligned}$$

Since  $|\frac{1}{n} \sum_{i=1}^n \langle E(x_1 - x_i), x_2 \rangle| \leq \|E|_{\mathcal{D}}\|(2M)(M) = 2\epsilon M^2$ , one has:

$$\frac{1}{n}(\Phi X_n)(\Phi X_n)^T = \frac{1}{n}X_n X_n^T + \frac{1}{n}R,$$

where  $|R_{i,j}| \leq 2\epsilon M^2$ .

Because  $\|R\|_2 \leq n\|R\|_{\max} \leq 2\epsilon M^2 n$  and  $\text{cov}(X_n)$  (resp.  $\text{cov}(\Phi X_n)$ ) has the same non-zero eigenvalues as  $\frac{1}{n}X_n X_n^T$  (resp.  $\frac{1}{n}\Phi X_n(\Phi X_n)^T$ ), one obtains:

$$|\lambda_i^2(\text{cov}(X_n)) - \lambda_i^2(\text{cov}(\Phi X_n))| \leq 2\epsilon M^2.$$

**Example 23** (Johnson and Lindenstrauss (1984)). *Consider taking  $\Phi$  to be a multiple of a random projection. In particular, let  $P : \mathbb{R}^D \rightarrow \mathbb{R}^d$  be a projection onto a random (in the sense of Johnson and Lindenstrauss (1984))  $d$ -dimensional subspace of  $\mathbb{R}^D$ , and let  $\Phi = \sqrt{\frac{D}{d}}P$ . Then if*

$$d \geq \frac{4 \log n + \log(\frac{4}{\delta^2})}{\epsilon^2},$$

*equation (9.1) will be satisfied with probability larger than  $1 - \delta$ .*

If the main results were shown to hold under bi-Lipschitz perturbations, the application of random projections could be useful in reducing the ambient dimension and thus the computational complexity of the MSVD algorithm. It seems that at the very least, one would need  $\epsilon = O(k^{-1})$ , which would reduce the ambient dimension to  $d = O(k^2 \log n)$ ; however, to obtain a similar range of scales as in Theorem 11,  $\epsilon$  would also need to be small relative to the scale  $r$  and curvature  $\kappa$ .

## 9.2 Dimensionality Reduction Algorithms

One of the main applications of the MSVD dimension estimator is to improve the performance of dimensionality reduction algorithms. Given a low-dimensional set

embedded in a high-dimensional space  $\mathbb{R}^D$ , dimensionality reduction algorithms seek to compute a map  $\Phi : \mathbb{R}^D \rightarrow \mathbb{R}^K$ , with  $K \ll D$ , that preserves the “structure” of the original set in some sense.

Principal component analysis (PCA), by projecting the data in the directions of greatest variance, has long been used to find linear low-dimensional maps but fails on non-linear data. Numerous non-linear dimensionality reduction techniques have thus been developed including kernel PCA (Schlkopf et al. (1998)), Laplacian Eigenmaps (Belkin and Niyogi (2001)), Hessian Eigenmaps (Donoho and Grimes (2003)), ISOMAP (Tenenbaum (1998)), and Locally Linear Embeddings (Roweis and Saul (2000)), to name just a few. See Das et al. (2006) for an example of ISOMAP being used in molecular dynamics to discover reaction coordinates which capture the essence of how a protein folds and unfolds.

To illustrate such techniques, the popular dimensionality reduction algorithm ISOMAP (Tenenbaum (1998)) is briefly described. Given  $n$  data points  $x_1, x_2, \dots, x_n \in \mathbb{R}^D$ , where  $D$  is large, ISOMAP computes a low-dimensional embedding of the data as follows:

1. A nearest neighbor graph  $G$  is computed; for example

$$G_{ij} = \begin{cases} \|x_i - x_j\| & \text{for } \|x_i - x_j\| < \delta \\ 0 & \text{for } \|x_i - x_j\| \geq \delta \end{cases},$$

where  $\delta$  is some pre-specified parameter.

2. The matrix of geodesic distances  $D$  is defined ( $D_{ij}$  is the shortest path between  $x_i$  and  $x_j$  in the graph  $G$ ).
3. The matrix  $D$  is centered to obtain a new matrix  $T = -\frac{1}{2}HSH^T$ , where  $S_{ij} = D_{ij}^2$  and  $H = \delta_{ij} - \frac{1}{n}$  is a centering matrix.

4. The eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$  of  $T$  and corresponding eigenvectors  $v_1, v_2, \dots, v_n$  are computed.
5. A low-dimensional embedding is constructed by mapping  $x_i \rightarrow \Phi_K(x_i) = (\sqrt{\lambda_1}v_1^i, \sqrt{\lambda_2}v_2^i, \dots, \sqrt{\lambda_K}v_K^i)$  for some  $K \ll \min(D, n)$ , where again  $K$  is a user-specified parameter and  $v_j^i$  is the  $i^{\text{th}}$  coordinate of  $v_j$ .

ISOMAP thus computes a map  $\Phi_K : \mathbb{R}^D \rightarrow \mathbb{R}^K$  so that the pairwise Euclidean distances (in  $\mathbb{R}^K$ ) are as close as possible to the original geodesic distances. However, the user must specify the parameter  $K$ , and without any knowledge of the intrinsic dimension, this is a challenging task. A residual for  $\Phi_K$  can be computed as follows:

$$\text{resVar}_K := 1 - \frac{\sum_{x_i, x_j} d_{\mathcal{M}}(x_i, x_j) \cdot \|\Phi_K(x_i) - \Phi_K(x_j)\|}{\left(\sum_{x_i, x_j} d_{\mathcal{M}}(x_i, x_j)^2\right)^{\frac{1}{2}} \left(\sum_{x_i, x_j} \|\Phi_K(x_i) - \Phi_K(x_j)\|^2\right)^{\frac{1}{2}}} \in [0, 1],$$

where  $d_{\mathcal{M}}(x_i, x_j) = D_{ij}$  is geodesic distance on  $\mathcal{M}$ . In practice this residual is often treated as a spectrum:  $K$  is selected so that  $\text{resVar}_{K-1} - \text{resVar}_K$  is the “largest gap” in the residuals.  $\Phi_K$  is in a sense the map giving an optimal balance between preserving geodesic distances and representing the data in as few dimensions as possible. This procedure is often used to infer that the intrinsic dimension of the data is  $K$ ; however, rigorous proofs that the intrinsic dimension is accurately detected are lacking and numerical experiments in fact indicate otherwise, see Little et al. (2011a).

In general there are no guarantees that dimensionality reduction algorithms will correctly estimate intrinsic dimension; this has motivated a renewed interest in developing algorithms for ID estimation with provable guarantees. Furthermore, estimating intrinsic dimension via the MSVD algorithm prior to the application of a dimensionality reduction algorithm could serve as a useful preprocessing step, providing guidance to the user in the selection of a meaningful  $K$ .

Another useful application would be to first apply the heat kernel maps described

in Jones et al. (2008, 2010), which give an  $\epsilon$  bi-Lipschitz mapping of “large” (depending upon the choice of  $\epsilon$ ) portions of  $\mathcal{M}$  into Euclidean space. These maps are independent of the embedding of  $\mathcal{M}$  in  $\mathbb{R}^D$  and are designed to flatten the data as much as possible. Once these maps have been computed, one could run the MSVD algorithm on the flattened pieces to estimate intrinsic dimension, which would allow one to take  $R_{\max}$  large.

# 10

## Conclusion

This dissertation presents a novel approach for estimating intrinsic dimension of noisy, high-dimensional point clouds using a multiscale version of PCA. It requires a number of local samples essentially linear in the intrinsic dimension and is highly robust to noise, as illustrated in numerous numerical experiments.

The main result gives a range of scales, determined by the three main constraints to ID estimation: curvature, noise, and sample size, in which PCA on a local neighborhood of the data detects the correct intrinsic dimension with high probability. If the curvature and noise are not too large, this interval is in fact sizable. This result is proved by first analyzing a simplified but uncomputable model, in which local neighborhoods are calculated before being corrupted by noise. It is then shown that, up to a change in scale, the multiscale SSV's of the simplified model are in fact close to the noisy SSV's which are actually computable. The main result is a non-asymptotic statement giving the success probability for a fixed, finite ambient dimension and sample size, although asymptotic statements are derived as corollaries. The proof requires the extensive use of results from random matrix theory and also concentration of measure in high dimensions.

Several specific examples are considered, including a one-dimensional curve, the unit sphere  $\mathbb{S}^k$ , and arbitrary co-dimension one manifolds. The multiscale SSV's of these sets are computed, and the scalings they exhibit partially motivate the assumptions made in Sec. 5.1 to prove the main result. For a smooth manifold, the tangent eigenvalues of the covariance grow quadratically whereas eigenvalues due to curvature grow quartically with respect to scale. The MSVD algorithm thus estimates the intrinsic dimension  $k$  by determining for which  $\hat{k}$   $\tilde{\lambda}_{k,z,r}^2$  grows linearly,  $\tilde{\lambda}_{k+1,z,r}^2$  grows quadratically, and  $\Delta_{\hat{k}}(\widetilde{\text{cov}}(X_{n,\hat{z},r}))$  is the largest gap in the eigenvalues for a large range of scales.

The MSVD algorithm has been tested on numerous data sets, including manifold data, collections of manifolds, and real-world data sets, and its performance compared with competing algorithms. In general, the MSVD algorithm was much more reliable in the presence of large noise. Furthermore, unlike most competing algorithms, where the user must select the correct scale through extensive experimentation, it infers the correct range of scales automatically from the noisy samples. Although applying PCA locally to estimate intrinsic dimension is not a novel idea, the MSVD algorithm is unique in being a multiscale rather than a fixed scale approach. It shows much potential for improving the performance of dimensionality reduction algorithms and for aiding in exploratory data analysis and problems of classification.

# Appendix A

## Perturbation Lemmas

The following lemma is due to Wielandt (1967) and is used in the proof of Prop. 8.

**Lemma 24** (Wieland's Inequality). *Let  $X \in \mathbb{R}^{D \times D}$  be a symmetric, positive definite matrix with the block structure*

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where  $A \in \mathbb{R}^{k \times k}$  and  $C \in \mathbb{R}^{(D-k) \times (D-k)}$  are both positive definite, and let  $\lambda_i(X)$  denote the  $i$ -th eigenvalue of the operator  $X$ , sorted in decreasing order. Then if  $\Delta := \lambda_k(A) - \lambda_1(C) > 0$ ,

$$0 \leq \lambda_i(X) - \lambda_i(A) \leq \|B\| \wedge \frac{\|B\|^2}{\Delta} \quad \text{for } i = 1, \dots, k \quad (\text{A.1})$$

$$0 \leq \lambda_{i-k}(C) - \lambda_i(X) \leq \|B\| \wedge \frac{\|B\|^2}{\Delta} \quad \text{for } i = k + 1, \dots, D. \quad (\text{A.2})$$

The next lemma shows that the empirical covariance matrix of a bounded random variable is not greatly perturbed by the removal of a small number of points. Note that  $E$ , the small subset of points which is removed, can be any subset, and can

be constructed in a deterministic fashion after the random realization of samples, allowing one to remove the “bad” points.

**Lemma 25.** *Let  $X$  be a r.v. in  $\mathbb{R}^D$ , with  $\|X\| \leq M$  almost surely, and let  $\epsilon > 0$ . Let  $x_1, \dots, x_n$  be i.i.d. draws from  $X$ , and let  $E \subseteq X_n := \{x_1, \dots, x_n\}$  have cardinality at most  $\lfloor \epsilon \cdot n \rfloor$ . Then*

$$\|\text{cov}(X_n) - \text{cov}(X_n \setminus E)\| \leq 4\epsilon(4 + \epsilon)M^2 \quad (\text{A.3})$$

*Proof.* Let  $n = n_1 + n_2$ , where  $n_2 = \epsilon n = |E|$ . Let  $\mu_{n_1+n_2} = \frac{1}{n} \sum_{i=1}^n x_i$  (empirical mean of  $X_n$ ),  $\mu_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$  (empirical mean of  $X_n \setminus E$ ),  $\mu_{n_2} = \frac{1}{n_2} \sum_{i=n_1+1}^n x_i$  (empirical

mean of  $E$ ). Then:

$$\text{cov}(X_n) - \text{cov}(X_n \setminus E)$$

$$\begin{aligned}
&= \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} (x_i - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1+n_2}) - \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1}) \otimes (x_i - \mu_{n_1}) \\
&= \left( \frac{n_1}{n_1 + n_2} \right) \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1+n_2}) \\
&\quad + \left( \frac{n_2}{n_1 + n_2} \right) \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1+n_2}) - \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1}) \otimes (x_i - \mu_{n_1}) \\
&= \left( 1 - \frac{n_2}{n_1 + n_2} \right) \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1+n_2}) \\
&\quad + \left( \frac{n_2}{n_1 + n_2} \right) \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1+n_2}) - \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1}) \otimes (x_i - \mu_{n_1}) \\
&= \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1+n_2}) - \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1}) \otimes (x_i - \mu_{n_1}) \quad \} := A \\
&\quad - \left( \frac{n_2}{n_1 + n_2} \right) \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1+n_2}) \quad \} := B \\
&\quad + \left( \frac{n_2}{n_1 + n_2} \right) \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1+n_2}) \quad \} := C
\end{aligned}$$

For both  $B$  and  $C$ :

$$\|B\| \leq \epsilon(2M)^2 \quad , \quad \|C\| \leq \epsilon(2M)^2,$$

using  $\|(x_i - \mu_{n_1+n_2})\| \leq 2M$ . For A:

$$\begin{aligned}
A &= \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1} + \mu_{n_1} - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1} + \mu_{n_1} - \mu_{n_1+n_2}) \\
&\quad - \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1}) \otimes (x_i - \mu_{n_1}) \\
&= \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \mu_{n_1}) \otimes (\mu_{n_1} - \mu_{n_1+n_2}) + \frac{1}{n_1} \sum_{i=1}^{n_1} (\mu_{n_1} - \mu_{n_1+n_2}) \otimes (x_i - \mu_{n_1}) \\
&\quad + \frac{1}{n_1} \sum_{i=1}^{n_1} (\mu_{n_1} - \mu_{n_1+n_2}) \otimes (\mu_{n_1} - \mu_{n_1+n_2}).
\end{aligned}$$

Thus:

$$\|A\| \leq 4M \|\mu_{n_1} - \mu_{n_1+n_2}\| + \|\mu_{n_1} - \mu_{n_1+n_2}\|^2.$$

Now observe that:

$$\mu_{n_1+n_2} = \frac{n_1}{n_1+n_2} \mu_{n_1} + \frac{n_2}{n_1+n_2} \mu_{n_2}.$$

Thus:

$$\begin{aligned}
\mu_{n_1+n_2} - \mu_{n_1} &= \frac{n_1}{n_1+n_2} \mu_{n_1} + \frac{n_2}{n_1+n_2} \mu_{n_2} - \mu_{n_1} \\
&= \left(1 - \frac{n_2}{n_1+n_2}\right) \mu_{n_1} + \frac{n_2}{n_1+n_2} \mu_{n_2} - \mu_{n_1} \\
&= \frac{n_2}{n_1+n_2} (\mu_{n_2} - \mu_{n_1}),
\end{aligned}$$

so that  $\|\mu_{n_1+n_2} - \mu_{n_1}\| \leq \epsilon(2M)$ . Thus:

$$\|A\| \leq 4M(2M\epsilon) + (2M\epsilon)^2 = 8M^2\epsilon + 4M^2\epsilon^2.$$

Combining with the previous result, one obtains:

$$\|\text{cov}(X_n) - \text{cov}(X_n \setminus E)\| \leq 16\epsilon M^2 + 4M^2\epsilon^2.$$

□

For  $E \not\subseteq X_n$ , observe that an equivalent result will hold for  $\|\text{cov}(X_n) - \text{cov}(X_n \cup E)\|$ .

The above lemma is used to prove the following result, which is used extensively in the proof of Prop. 10.

**Lemma 26** (Covariance Perturbation, Chernoff Version). *Let  $X_n := \{x_1, \dots, x_n\}$  be  $n$  i.i.d. random samples from a density  $\mu_X$  in  $\mathbb{R}^D$ . Let  $A, B$  be two ( $\mu_X$ -measurable) sets, with  $B \subseteq A$ ,  $\mu_X(B) \leq \delta\mu_X(A)$ , and  $A$  bounded by  $M$ ; consider the random variables  $n_A = |X_n \cap A| =: |A_n|$  and  $n_B = |X_n \cap B| =: |B_n|$ . Then for  $s \geq 1$ ,  $n \geq t^2/\mu_X(A)$ :*

$$\mathbb{P}\left(n_B \leq 4s^2 \left(\delta \vee \frac{1}{\mu_X(A)n}\right) n_A\right) \geq 1 - e^{-\frac{1}{8}s^2 t^2} - 2e^{-\frac{1}{3}s^2(\delta\mu_X(A)n \vee 1)},$$

and

$$\|\text{cov}(A_n) - \text{cov}(A_n \setminus B_n)\| \lesssim s^2 \left(\delta \vee \frac{1}{\mu_X(A)n}\right) M^2 \quad (\text{A.4})$$

with the same probability and conditions. The same conclusion, with  $M^2$  replaced by  $M^2 + \sigma^2 D$ , holds if  $A_n$  (resp.  $B_n$ ) is replaced by  $\tilde{A}_n := \{x_i + \eta_i : x_i \in A_n\}$  (resp.  $\tilde{B}_n$ ), where the  $\eta_i$  are i.i.d. and satisfy  $\|\eta_i\| \leq \sigma\sqrt{D}$ .

*Proof.* By Chernoff's inequality in Thm. 36:

$$\mathbb{P}\left(n_A \leq \frac{1}{2}\mu_X(A)n\right) \leq e^{-\frac{1}{8}t^2} \quad (\text{A.5})$$

for  $n \geq \frac{t^2}{\mu_X(A)}$ ; denote this event  $\Omega_t$ . Now  $n_B$  is  $\text{Bin}(n, \mu_X(B))$ ; let  $\tilde{n}_B$  be

$\text{Bin}(n, \delta\mu_X(A) \vee \frac{1}{n})$ . Then on  $\Omega_t$ :

$$\begin{aligned}
& \mathbb{P}\left(n_B > 2(1+s^2)\left(\delta \vee \frac{1}{\mu_X(A)n}\right)n_A\right) \\
& \leq \mathbb{P}\left(n_B > (1+s^2)\left(\delta \vee \frac{1}{\mu_X(A)n}\right)\mu_X(A)n\right) \\
& \leq \mathbb{P}\left(\tilde{n}_B > (1+s^2)\left(\delta\mu_X(A) \vee \frac{1}{n}\right)n\right) \\
& = \mathbb{P}\left(\tilde{n}_B > (1+s^2)\mathbb{E}[\tilde{n}_B]\right) \leq e^{-\frac{s^2}{3}(\delta\mu_X(A)n \vee 1)}
\end{aligned}$$

for any  $s \geq 1$ , the last line also following from a Chernoff inequality. Thus for any  $s \geq 1$ ,  $n \geq t^2/\mu_X(A)$ :

$$n_B \leq 4s^2\left(\delta \vee \frac{1}{\mu_X(A)n}\right)n_A$$

with probability at least  $1 - e^{-\frac{1}{8}t^2} - 2e^{-\frac{1}{3}s^2(\delta\mu_X(A)n \vee 1)}$ .

The case when noise is added follows in a similar fashion. □

Note that the same bound also holds for  $\|\text{cov}(A_n) - \text{cov}(A_n \cup B_n)\|$ .

# Appendix B

## Concentration Inequalities

In this chapter, some important definitions and concentration inequalities are reviewed, which are needed in the proof of the main results.

**Definition 27.** *A random variable  $X \in \mathbb{R}$  is called subgaussian if one of the following three equivalent conditions is satisfied:*

1.  $\mathbb{P}(|X| > t) \leq 2e^{-(t/M)^2}$
2.  $(\mathbb{E}|X|^p)^{\frac{1}{p}} \leq C_1 M \sqrt{p}$
3.  $\mathbb{E} e^{C_2 X^2/M^2} \leq e;$

here  $C_1, C_2$  are universal constants and  $M$  is called the subgaussian moment of  $X$ .

**Definition 28.** *A random variable  $X \in \mathbb{R}$  is called subexponential if one of the following three equivalent conditions is satisfied:*

1.  $\mathbb{P}(|X| > t) \leq 2e^{-(t/M)}$
2.  $(\mathbb{E}|X|^p)^{\frac{1}{p}} \leq C_1 M p$

$$3. \mathbb{E} e^{C_2 X/M} \leq e;$$

here  $C_1, C_2$  are universal constants and  $M$  is called the subexponential moment of  $X$ .

**Remark 29.** *The choice of 2 in the first condition of Definitions 27 and 28 is arbitrary: any positive constant may be chosen and the subgaussian moment is unchanged up to an absolute constant; using  $e$  is convenient in many cases.*

Subgaussian and subexponential random variables are special cases of Orlicz- $\alpha$  random variables.

**Definition 30.** *A random variable  $X \in \mathbb{R}$  belongs to the Orlicz space  $\psi_\alpha$  if*

$$\|X\|_{\psi_\alpha} := \inf\{c : \mathbb{E} e^{(\frac{|X|}{c})^\alpha} < 2\} < \infty.$$

Thus the subgaussian moment and Orlicz-2 norm of a random variable are the same up to an absolute constant; similarly for the subexponential moment and Orlicz-1 norm.

**Proposition 31** (Sums of Subgaussian Random Variables). *Let  $X_1, \dots, X_n$  be  $n$  independent, mean zero, subgaussian random variables and let  $K = \max_i \|X_i\|_{\psi_2}$ . Then for any  $a = (a_1, \dots, a_n) \in \mathbb{R}^n$  and  $t \geq 0$ :*

$$\mathbb{P}(|\sum_{i=1}^n a_i X_i| > t) \leq e^{1 - \frac{ct^2}{K^2 \|a\|_2^2}},$$

where  $c$  is an absolute constant. When  $a$  is a unit vector:

$$\mathbb{P}(|\sum_{i=1}^n a_i X_i| > t) \leq e^{1 - \frac{ct^2}{K^2}},$$

so that the subgaussian moment of  $\sum_{i=1}^n a_i X_i$  is bounded by  $K$ , up to an absolute constant.

**Proposition 32** (Sums of Subexponential Random Variables). *Let  $X_1, \dots, X_n$  be  $n$  independent, mean zero, subexponential random variables and let  $K = \max_i \|X_i\|_{\psi_1}$ . Then for any  $a = (a_1, \dots, a_n) \in \mathbb{R}^n$  and  $t \geq 0$ :*

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \leq 2e^{-c\left(\frac{t^2}{K^2\|a\|_2^2} \wedge \frac{t}{K\|a\|_\infty}\right)}.$$

When  $a = (1, \dots, 1)$ , one obtains:

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > tK\sqrt{n}\right) \leq 2e^{-ct(t \wedge \sqrt{n})}.$$

For proofs that the conditions in Defs. 27 and 28 are equivalent and for a proof of Props. 31 and 32 see Vershynin (2010b). Observe that a random variable  $X$  is subgaussian if and only if  $X^2$  is subexponential, so that  $\|X\|_{\psi_2}^2$  and  $\|X^2\|_{\psi_1}$  are equivalent.

The following theorem is derived from Pinelis (1994, 1999):

**Theorem 33** (Pinelis Inequality). *Let  $X_1, \dots, X_n$  be mean zero, independent random variables taking values in a Hilbert space  $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ . Furthermore, suppose that for  $i = 1, \dots, n$ :*

$$\mathbb{E}\|X_i\|_{\mathcal{H}}^p \leq \frac{p!}{2} b^2 L^{p-2}, \tag{B.1}$$

for some choice of positive constants  $L$  and  $b$ . Then for all  $t > 0$

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_{\mathcal{H}} \geq \frac{L}{n}t^2 + \frac{\sqrt{2}b}{\sqrt{n}}t\right) \leq 2e^{-t^2}. \tag{B.2}$$

If  $\|X_i\|_{\mathcal{H}} \leq M$  for  $i = 1, \dots, n$ , then one obtains:

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_{\mathcal{H}} \geq \frac{\sqrt{2}M}{\sqrt{n}}t\right) \leq 2e^{-t^2}. \tag{B.3}$$

**Corollary 34.** *Let  $X_1, X_2, \dots, X_n$  be a sequence of independent, mean zero random variables in  $\mathbb{R}^D$  such that  $\| \|X_i\| \|_{\psi_2} \leq B$  for  $1 \leq i \leq n$ . Then for any  $t \geq C_1$*

$$\mathbb{P} \left( \left\| \frac{1}{n} \sum_{i=1}^n X_i \right\| \geq \frac{C_2 B}{\sqrt{n}} t^2 \right) \leq 2e^{-t^2}, \quad (\text{B.4})$$

where  $C_1, C_2$  are absolute constants.

**Remark 35.** *Although it is not needed for the proofs in this work, the above actually holds with  $\| \cdot \|_{\psi_1}$  replacing  $\| \cdot \|_{\psi_2}$ , and this gives a sharper statement.*

Finally, the following result is due to Chernoff (1952); see Hagerup and Rub (1990) for a nice exposition.

**Theorem 36** (Chernoff's Inequality). *Let  $X_1, \dots, X_n$  be a sequence of independent r.v.'s taking values in  $\{0, 1\}$  and let  $S_n = \sum_{i=1}^n X_i$ . Then*

$$\begin{aligned} \mathbb{P} ( S_n \geq (1 + \epsilon) \mathbb{E}[S_n] ) &\leq e^{-\frac{\epsilon^2 \mathbb{E}[S_n]}{(2+\epsilon)\sqrt{3}}} && \text{for any } \epsilon > 0 \\ \mathbb{P} ( S_n \leq (1 - \epsilon) \mathbb{E}[S_n] ) &\leq e^{-\frac{\epsilon^2 \mathbb{E}[S_n]}{2}} && \text{for } 0 \leq \epsilon \leq 1. \end{aligned}$$

# Appendix C

## Inequalities for Covariance Operators

This section gives several concentration results for empirical covariance and cross-covariance operators; these results were compiled by L. Rosasco, see Little et al. (2011a).

If  $Y$  and  $X$  are two random vectors in  $\mathbb{R}^D$  (not necessarily independent) with mean  $\mathbb{E}[Y]$  and  $\mathbb{E}[X]$ , then let:

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X]) \otimes (Y - \mathbb{E}[Y])].$$

Let  $(Y_1, X_1), \dots, (Y_n, X_n)$  be  $n$  identical and independent copies of  $(Y, X)$  and let  $(y_1, x_1), \dots, (y_n, x_n)$  be a random draw.  $\mathbb{E}_n$  denotes the expectation with respect to the empirical measure  $\frac{1}{n} \sum_{i=1}^n \delta_{(y_i, x_i)}$ . Let:

$$\text{cov}(X_n, Y_n) = \mathbb{E}_n[(X - \mathbb{E}_n[X]) \otimes (Y - \mathbb{E}_n[Y])] = \frac{1}{n} \sum_{i=1}^n (x_i - \mathbb{E}_n[X]) \otimes (y_i - \mathbb{E}_n[Y]).$$

If  $X = Y$ , then  $\text{cov}(Y) = \text{cov}(X, Y)$  and  $\text{cov}(Y_n) = \text{cov}(X_n, Y_n)$ .

This appendix derives concentration properties of the empirical covariance and cross-covariance operators under different assumptions on  $Y$  and  $X$ . In particular, the

following three cases are considered: 1) bounded random vectors, 2) subgaussian random vectors, and 3) random vectors whose norms belong to the  $\psi_2$  Orlicz space. One has the following relationship between truly centered and empirically centered covariance operators.

**Lemma 37.** *The following inequality holds:*

$$\|\text{cov}(X_n, Y_n)\| \leq \left\| \frac{1}{n} \sum_{i=1}^n (x_i - \mathbb{E}[X]) \otimes (y_i - \mathbb{E}[Y]) \right\| + \|\mathbb{E}[X] - \mathbb{E}_n[X]\| \cdot \|\mathbb{E}[Y] - \mathbb{E}_n[Y]\|.$$

*In particular:*

$$\begin{aligned} \|\text{cov}(X_n, Y_n) - \text{cov}(X, Y)\| &\leq \left\| \frac{1}{n} \sum_{i=1}^n (x_i - \mathbb{E}[X]) \otimes (y_i - \mathbb{E}[Y]) - \text{cov}(X, Y) \right\| \\ &\quad + \|\mathbb{E}[X] - \mathbb{E}_n[X]\| \cdot \|\mathbb{E}[Y] - \mathbb{E}_n[Y]\| \end{aligned}$$

and

$$\|\text{cov}(Y_n) - \text{cov}(Y)\| \leq \left\| \frac{1}{n} \sum_{i=1}^n (y_i - \mathbb{E}[Y]) \otimes (y_i - \mathbb{E}[Y]) - \text{cov}(Y) \right\| + \|\mathbb{E}[Y] - \mathbb{E}_n[Y]\|^2.$$

*The same bounds hold if the operator norm is replaced with the Frobenius norm.*

*Proof.* The inequality follows from the definition of empirical cross covariance, in fact, by adding and subtracting  $\mathbb{E}_n[X]$  and  $\mathbb{E}_n[Y]$ :

$$\begin{aligned} \text{cov}(X_n, Y_n) &= \mathbb{E}_n[X - \mathbb{E}[X] + \mathbb{E}[X] - \mathbb{E}_n[X]] \otimes [Y - \mathbb{E}[Y] + \mathbb{E}[Y] - \mathbb{E}_n[Y]] \\ &= \mathbb{E}_n[(X - \mathbb{E}[X]) \otimes (Y - \mathbb{E}[Y])] + \mathbb{E}_n[(\mathbb{E}[X] - \mathbb{E}_n[X]) \otimes (\mathbb{E}[Y] - \mathbb{E}_n[Y])] \\ &\quad + \mathbb{E}_n[(X - \mathbb{E}[X]) \otimes (\mathbb{E}[Y] - \mathbb{E}_n[Y])] + \mathbb{E}_n[(\mathbb{E}[X] - \mathbb{E}_n[X]) \otimes (Y - \mathbb{E}[Y])] \\ &= \mathbb{E}_n[(X - \mathbb{E}[X]) \otimes (Y - \mathbb{E}[Y])] - (\mathbb{E}[X] - \mathbb{E}_n[X]) \otimes (\mathbb{E}[Y] - \mathbb{E}_n[Y]). \end{aligned}$$

The proof is finished by taking the norm of both sides, using the triangle inequality, and recalling that  $\forall u, v \in \mathbb{R}^D$  one has  $\|u \otimes v\| \leq \|u \otimes v\|_F = \|u\| \cdot \|v\|$ .  $\square$

The deviation of the means can be easily controlled using standard concentration inequalities for vector valued random variables:

**Proposition 38.** *Let  $Y$  be a random vector in  $\mathbb{R}^D$ . The following concentration inequalities around the mean hold for any  $t > 0$ :*

$$\mathbb{P}(\|\mathbb{E}(Y) - \mathbb{E}_n(Y)\| \geq \epsilon(t, n)) \leq 2e^{-t^2}, \quad (\text{C.1})$$

where for some absolute constant  $C$

(i) If  $\|Y\| \leq M$  almost surely,  $\epsilon(t, n) = \frac{CMt}{\sqrt{n}}$ ;

(ii) If  $\langle Y, \theta \rangle$  is centered and subgaussian with moment  $M$  for every  $\theta \in \mathbb{S}^{D-1}$ ,  
 $\epsilon(t, n) = \frac{CM\sqrt{Dt^2}}{\sqrt{n}}$ ;

(iii) If  $\|Z\|_{\psi_2} \leq M$ , with  $Z = \|Y\|$ ,  $\epsilon(t, n) = \frac{CMt^2}{\sqrt{n}}$ .

*Proof.* (i) follows from applying (B.3) to the random vector  $Y' = Y - m$ , which satisfies  $\|Y'\| \leq 2M$ ,  $\mathbb{E}Y' = 0$ , and  $\mathbb{E}_n Y' = m_n - m$ .

(ii) and (iii) are proved similarly; only the proof of (iii) is presented here. Let  $Z = \|Y\|$ ,  $Y' = Y - m$  and  $Z' = \|Y'\|$ . First it is shown that

$$\|Z\|_{\psi_2} \leq B \implies \|Z'\|_{\psi_2} \leq 2B. \quad (\text{C.2})$$

Note that  $\|m\|^2 = \|\mathbb{E}Y\|^2 \leq \mathbb{E}[\|Y\|^2]$  since  $\|m\|^2 = \langle \mathbb{E}Y, \mathbb{E}Y \rangle = \mathbb{E}\mathbb{E}\langle Y, \bar{Y} \rangle \leq \mathbb{E}[\|Y\|^2]$ . Then for  $p \geq 2$

$$\mathbb{E}e^{\frac{\|Y'\|^2}{c^2}} = \sum_{p=0}^{\infty} \frac{\mathbb{E}[\|Y'\|^{2p}]}{p!C^{2p}} \leq \sum_{p=0}^{\infty} 2^{2p+1} \frac{\mathbb{E}[\|Y\|^{2p}]}{p!C^{2p}} = \mathbb{E}e^{\frac{4\|Y\|^2}{c^2}} \leq 2$$

if one takes  $C^2 = 4B^2$ , and the claim is proved. Then  $\mathbb{E}[\|Y'\|^p] \leq 1/2p!b^2L^{p-2}$  with  $b = 4\sqrt{2}B$  and  $L = 4B$ . The proof is finished by applying (B.2) to  $Y'$  since  $\mathbb{E}[Y'] = 0$ ,  $\mathbb{E}_n[Y'] = m_n - m$ , and  $Lt^2/n + \sqrt{2}bt/\sqrt{n} \leq 16Bt^2/\sqrt{n}$  for  $t > 1$ .  $\square$

Next some results on the concentration of covariance matrices (Vershynin (2010b), Rudelson and Vershynin (2007), Vershynin (2010a)) are given, starting with a result for covariance matrices of mean zero, *bounded* random vectors:

**Theorem 39** (Vershynin (2010b)). *Let  $Y_1, \dots, Y_n$  be independent, centered random vectors in  $\mathbb{R}^D$  with common covariance matrix  $\text{cov}(Y)$ . Suppose that  $\|Y_i\| \leq M$  for  $1 \leq i \leq n$ . Then for  $t \geq C$  and  $\delta = \frac{tM\sqrt{\log D}}{\sqrt{n}}$ ,*

$$\mathbb{P} \left( \left\| \frac{1}{n} \sum_{i=1}^n Y_i \otimes Y_i - \text{cov}(Y) \right\| \leq \|\text{cov}(Y)\|^{\frac{1}{2}} \delta \vee \delta^2 \right) \geq 1 - e^{-c \log(D)t^2},$$

where  $C$  and  $c$  are universal constants.

Combining Theorem 39 with Lemma 37, one obtains a result for empirically centered covariances. Prop. 38 is used to estimate the concentration of the empirical mean, which implies that  $\|\mathbb{E}[Y] - \mathbb{E}_n[Y]\|^2 \leq CM^2t^2/n$  with probability at least  $1 - 2e^{-t^2}$ .

**Theorem 40** (Empirically centered covariance for bounded r.v.). *Let  $Y_1, \dots, Y_n$  be independent random vectors in  $\mathbb{R}^D$  with common covariance matrix  $\text{cov}(Y)$ . Suppose that  $\|Y_i\| \leq M$  for  $1 \leq i \leq n$ . Then for  $t \geq C$ ,*

$$\mathbb{P} \left( \|\text{cov}(Y) - \text{cov}(Y_n)\| > \|\text{cov}(Y)\|^{\frac{1}{2}} \frac{tM\sqrt{\log D}}{\sqrt{n}} + \frac{t^2 M^2 \log D}{n} \right) \leq 3e^{-ct^2},$$

where  $C$  and  $c$  are universal constants.

The following result gives the covariance concentration for mean zero, subgaussian random vectors:

**Theorem 41** (Vershynin (2010b)). *Let  $Y_1, \dots, Y_n$  be independent, centered random vectors in  $\mathbb{R}^D$  with common covariance matrix  $\text{cov}(Y)$ . Suppose that  $\sup_{\theta \in \mathbb{S}^{D-1}} \|\langle Y_i, \theta \rangle\|_{\psi_2} \leq M$  for  $1 \leq i \leq n$ . Then for any  $t \geq 0$ :*

$$\mathbb{P} \left( \left\| \frac{1}{n} \sum_{i=1}^n Y_i \otimes Y_i - \text{cov}(Y) \right\| \leq (\delta \vee \delta^2) \|\text{cov}(Y)\| \right) \geq 1 - 2e^{-ct^2}$$

where

$$\delta = \frac{C \|\text{cov}(Y)^{-1}\| M^2 \sqrt{D}}{\sqrt{n}} + \frac{t}{\sqrt{n}},$$

and  $c$  and  $C$  are universal constants.

Again, combining Theorem 41 with Lemma 37 and Prop. 38 one obtains the following result for empirically centered covariances.

**Theorem 42** (Empirically centered covariance for subgaussian r.v.). *Let  $Y_1, \dots, Y_n$  be independent random vectors in  $\mathbb{R}^D$  with identity covariance. Suppose that  $\sup_{\theta \in \mathbb{S}^{D-1}} \|\langle Y_i, \theta \rangle\|_{\psi_2} \leq M$  for  $1 \leq i \leq n$ . Then for any  $t \geq CM$ :*

$$\mathbb{P} \left( \|\text{cov}(Y_n) - I\| \leq M^2 \left( \frac{\sqrt{D}}{\sqrt{n}} t + \frac{D}{n} t^4 \right) \right) \geq 1 - 4e^{-ct^2},$$

where  $c$  and  $C$  are universal constants.

**Remark 43.** *The  $\log(D)$  term was necessary in 40.*

The following holds for random vectors, possibly in infinite dimensions, satisfying an Orlicz-2 condition:

**Theorem 44.** *If  $Y$  is a random vector in  $\mathbb{R}^D$  and  $Z = \|Y\|$  satisfies  $\|Z\|_{\psi_2} \leq B$ , then for any  $t > 1$  and some universal constant  $C$ :*

$$\mathbb{P} \left( \|\text{cov}(Y) - \text{cov}(Y_n)\|_{\text{F}} > CB^2 \left( \frac{t^2}{\sqrt{n}} + \frac{t^4}{n} \right) \right) \leq 4e^{-t^2}.$$

*Proof.* Note that from Prop. 38, case (ii),  $\|\mathbb{E}[Y] - \mathbb{E}_n[Y]\|^2 \leq CB^2 t^4/n$  with probability at least  $1 - 2e^{-t^2}$ . Moreover one can control  $\|\mathbb{E}_n[(Y - \mathbb{E}[Y]) \otimes (Y - \mathbb{E}[Y])] - \text{cov}(Y)\|$ , studying the uncentered covariance operator of the centered random vector  $Y' = Y - m$ , since  $\|Z'\|_{\psi_2} \leq \sqrt{2}\|Z\|_{\psi_2}$ , where  $Z' = \|Y'\|$ ,  $Z = \|Y\|$ . Corollary 34 to Pinelis' inequality then yields the desired result.  $\square$

# Appendix D

## Norms of Random Matrices

The following result is useful in bounding the norm of the empirical cross-covariance of two independent random vectors of differing dimensionalities.

**Proposition 45** (Norm of product of random matrices). *Let  $k \leq d$ , and let  $N_1 \in \mathbb{R}^{n \times k}$ ,  $N_2 \in \mathbb{R}^{n \times d}$  have i.i.d. subgaussian entries with mean 0 and subgaussian moment 1. Let  $\alpha = 2c^{-1} \log 6$ , where  $c$  is the constant given in Proposition 32. Then:*

$$\mathbb{P} \left( \frac{1}{n} \|N_1^T N_2\| > \frac{\sqrt{k} + \sqrt{d}}{\sqrt{n}} t \right) \leq \begin{cases} ce^{-\frac{c}{32}(\sqrt{k} + \sqrt{d})^2 t^2}, & 4\sqrt{\frac{\alpha}{1 + \sqrt{\frac{d}{k}}}} \leq t \leq \frac{4\sqrt{n}}{\sqrt{k} + \sqrt{d}} \\ ce^{-\frac{c}{8}\sqrt{n}(\sqrt{k} + \sqrt{d})t}, & t \geq \alpha\sqrt{\frac{k}{n}} \vee \frac{4\sqrt{n}}{\sqrt{k} + \sqrt{d}}. \end{cases} \quad (\text{D.1})$$

The i.i.d. entries assumption of the matrices may be replaced with the following, with an almost identical proof:

1.  $N_1$  and  $N_2$  are independent;
2. the rows of  $N_1$  are independent and for every row  $n_1$ ,  $\langle n_1, \theta \rangle$  is subgaussian with moment 1 for every  $\theta \in \mathbb{S}^{k-1}$ ;
3. the rows of  $N_2$  are independent and for every row  $n_2$ ,  $\langle n_2, \theta \rangle$  is subgaussian with moment 1 for every  $\theta \in \mathbb{S}^{d-1}$ .

**Remark 46.** Observe that  $N_1^T N_2$  grows only like  $\sqrt{n}$ , whereas  $N_1^T N_1$  grows like  $n$ ; the independence of  $N_1$  and  $N_2$  allows for cancellations which keep the norm relatively small.

*Proof of Proposition 45.* Let  $\mathcal{N}^d$  be an  $\epsilon_1$ -net for  $\mathbb{S}^{d-1}$  and let  $\mathcal{N}^k$  be an  $\epsilon_2$ -net for  $\mathbb{S}^{k-1}$ . Observe that by a standard volume estimate one can choose the nets so that  $|\mathcal{N}^d| \leq (3/\epsilon_1)^d$  and  $|\mathcal{N}^k| \leq (3/\epsilon_2)^k$ . Clearly:

$$\|N_1^T N_2\| = \max_{x \in \mathbb{S}^{d-1}, y \in \mathbb{S}^{k-1}} \langle N_1^T N_2 x, y \rangle \leq (1 - \epsilon_1)^{-1} (1 - \epsilon_2)^{-1} \max_{x \in \mathcal{N}^d, y \in \mathcal{N}^k} \langle N_1^T N_2 x, y \rangle.$$

Therefore:

$$\begin{aligned} \mathbb{P}(\|N_1^T N_2\| > t) &\leq \mathbb{P}\left(\max_{x \in \mathcal{N}^d, y \in \mathcal{N}^k} |\langle N_1^T N_2 x, y \rangle| > t(1 - \epsilon_1)(1 - \epsilon_2)\right) \\ &\leq \sum_{x \in \mathcal{N}^d, y \in \mathcal{N}^k} \mathbb{P}\left(|\langle N_1^T N_2 x, y \rangle| > t(1 - \epsilon_1)(1 - \epsilon_2)\right) \\ &\leq \left(\frac{3}{\epsilon_1}\right)^d \left(\frac{3}{\epsilon_2}\right)^k \mathbb{P}\left(|\langle N_2 x, N_1 y \rangle| > t(1 - \epsilon_1)(1 - \epsilon_2)\right). \end{aligned}$$

Restrict  $\epsilon_1 \in (0, \frac{1}{2}]$  and choose  $\epsilon_2$  to satisfy  $(\frac{3}{\epsilon_2})^{\sqrt{k}} = (\frac{3}{\epsilon_1})^{\sqrt{d}}$ . Because  $d \geq k$ ,  $\epsilon_2 \leq \epsilon_1$ , and so  $\epsilon_2 \in (0, \frac{1}{2}]$  as well. Thus:

$$\mathbb{P}(\|N_1^T N_2\| > t) \leq \left(\frac{3}{\epsilon_2}\right)^{\sqrt{k}(\sqrt{d} + \sqrt{k})} \mathbb{P}\left(|\langle N_2 x, N_1 y \rangle| > \frac{t}{4}\right).$$

Since the entries of  $N_1, N_2$  are i.i.d. subgaussian, and  $x, y$  have  $L^2$ -norm 1,  $N_2 x$  has independent subgaussian entries and so does  $N_1 y$ , with the same subgaussian moments as the entries of  $N_1$  and  $N_2$  respectively. Also, clearly  $N_2 x$  and  $N_1 y$  are independent, so that  $\langle N_2 x, N_1 y \rangle$  is the sum of  $n$  independent subexponential random variables (since the product of two subgaussian random variables is subexponential), and therefore Proposition 32, with  $x_i = (N_2 x)_i (N_1 y)_i$  and  $a_i = 1$ , implies

$\mathbb{P}(|\langle N_2x, N_1y \rangle| > \bar{t}) \leq ce^{-c\frac{\bar{t}}{n}(\bar{t} \wedge n)}$ . Letting  $\bar{t} = \frac{t}{4}$ :

$$\mathbb{P}(\|N_1^T N_2\| > t) \leq \left(\frac{3}{\epsilon_2}\right)^{\sqrt{k}(\sqrt{d}+\sqrt{k})} ce^{-\frac{ct}{4n}(\frac{t}{4} \wedge n)} \leq ce^{-\frac{ct}{8n}(\frac{t}{4} \wedge n)}$$

as soon as  $\sqrt{k}(\sqrt{d} + \sqrt{k}) \log(\frac{3}{\epsilon_2}) \leq \frac{ct}{8n}(\frac{t}{4} \wedge n)$ .

**Case 1:** Assume  $t \leq 4n$ . Then for  $\sqrt{k}(\sqrt{d} + \sqrt{k}) \log(\frac{3}{\epsilon_2}) \leq \frac{ct^2}{32n}$ ,  $\mathbb{P}(\|N_1^T N_2\| > t) \leq ce^{-\frac{ct^2}{32n}}$  by the above. For  $t = \sqrt{\frac{32}{c}n\sqrt{k}(\sqrt{d} + \sqrt{k}) \log(\frac{3}{\epsilon_2})}$  one obtains

$$\mathbb{P}\left(\|N_1^T N_2\| > \sqrt{\frac{32}{c}n\sqrt{k}(\sqrt{d} + \sqrt{k}) \log(\frac{3}{\epsilon_2})}\right) \leq ce^{-\sqrt{k}(\sqrt{d}+\sqrt{k}) \log(\frac{3}{\epsilon_2})}.$$

Letting  $\bar{t}\sqrt{n}(\sqrt{d} + \sqrt{k}) = \sqrt{\frac{32}{c}n\sqrt{k}(\sqrt{d} + \sqrt{k}) \log(\frac{3}{\epsilon_2})}$ :

$$\mathbb{P}(\|N_1^T N_2\| > \bar{t}\sqrt{n}(\sqrt{d} + \sqrt{k})) \leq ce^{-\frac{c}{32}(\sqrt{k}+\sqrt{d})^2\bar{t}^2}.$$

One has the following restrictions on  $\bar{t}$ :  $\bar{t} \leq \frac{4\sqrt{n}}{\sqrt{k}+\sqrt{d}}$  (to ensure that  $t \leq 4n$ ) and  $\bar{t} \geq 4\sqrt{2c^{-1} \log 6(\sqrt{dk^{-1}} + 1)^{-\frac{1}{2}}}$  (since  $\log(\frac{3}{\epsilon_2}) \geq \log 6$ ).

**Case 2:** Assume  $t \geq 4n$ . Then  $\mathbb{P}(\|N_1^T N_2\| > t) \leq ce^{-\frac{ct}{8}}$  if  $\sqrt{k}(\sqrt{d} + \sqrt{k}) \log(\frac{3}{\epsilon_2}) \leq \frac{ct}{8}$ .

Choosing  $t = \frac{8}{c}\sqrt{k}(\sqrt{d} + \sqrt{k}) \log(\frac{3}{\epsilon_2})$  gives:

$$\mathbb{P}\left(\|N_1^T N_2\| > \frac{8}{c}\sqrt{k}(\sqrt{d} + \sqrt{k}) \log(\frac{3}{\epsilon_2})\right) \leq ce^{-\sqrt{k}(\sqrt{d}+\sqrt{k}) \log(\frac{3}{\epsilon_2})}.$$

Letting  $\bar{t}\sqrt{n}(\sqrt{d} + \sqrt{k}) = \frac{8}{c}\sqrt{k}(\sqrt{d} + \sqrt{k}) \log(\frac{3}{\epsilon_2})$ :

$$\mathbb{P}(\|N_1^T N_2\| > \bar{t}\sqrt{n}(\sqrt{d} + \sqrt{k})) \leq ce^{-\frac{c}{8}\sqrt{n}(\sqrt{k}+\sqrt{d})\bar{t}}.$$

The restrictions on  $\bar{t}$  are:  $\bar{t} \geq \frac{4\sqrt{n}}{\sqrt{k}+\sqrt{d}}$  (to ensure that  $t \geq 4n$ ) and  $\bar{t} \geq 8c^{-1} \log 6\sqrt{kn^{-1}}$  (since  $\log(\frac{3}{\epsilon_2}) \geq \log 6$ ).  $\square$

The following proposition bounds the norm of a product of random matrices when only one of the matrices is assumed to be subgaussian; it shows that the norm of the product is essentially independent of the inner dimension.

**Proposition 47.** *Let  $B \in \mathbb{R}^{k \times n}$  and  $A \in \mathbb{R}^{n \times d}$ , where  $d \geq k$ , with  $A$  and  $B$  independent random matrices. Also suppose that  $A$  has independent rows  $a_1, \dots, a_n$  and that for all  $i$ ,  $\langle a_i, \theta \rangle$  is centered and subgaussian with moment  $M$  for every  $\theta \in \mathbb{S}^{d-1}$ .*

*Then for  $t^2 \geq 32 \log 6 \frac{1}{1 + \sqrt{kd-1}}$*

$$\mathbb{P} \left( \|BA\| > \|B\|(\sqrt{d} + \sqrt{k})t \right) \leq 2e^{-\frac{(\sqrt{d} + \sqrt{k})^2}{32M^2} t^2}. \quad (\text{D.2})$$

*In particular, when  $B = I_n$  one obtains for  $t^2 \geq 32 \log 6 \frac{1}{1 + \sqrt{nd-1}}$*

$$\mathbb{P} \left( \|A\| > (\sqrt{d} + \sqrt{n})t \right) \leq 2e^{-\frac{(\sqrt{d} + \sqrt{n})^2}{32M^2} t^2}, \quad (\text{D.3})$$

*which may be simplified, when  $d \geq n$ , to*

$$\mathbb{P} \left( \|A\| > \sqrt{d}t \right) \leq 2e^{-\frac{d}{128M^2} t^2}, \quad (\text{D.4})$$

*for  $t^2 \geq 32 \log 6$ .*

*Proof.* When  $B$  is deterministic, (D.2) can be proved with a covering argument just as in the proof of Prop. 45, however in this case one obtains a subgaussian tail because  $\langle Ax, B^T y \rangle$ , where again  $x \in \mathbb{S}^{d-1}$  and  $y \in \mathbb{S}^{k-1}$ , is the sum of  $n$  subgaussian random variables instead of  $n$  subexponential random variables. (D.4) can be found in Rudelson and Vershynin (2009) (see Prop. 2.3); for sharp bounds on the expectation of  $\|BA\|$  (deterministic  $B$  and random  $A$ ), see Vershynin (2008).

The results can be extended to random  $B$  by the following argument:

$$\begin{aligned}
\mathbb{P}(\|BA\| > t\|B\|(\sqrt{d} + \sqrt{k})) &= \mathbb{E}_{A,B} [ \mathbf{1}_{\|BA\| > t\|B\|(\sqrt{d} + \sqrt{k})} ] \\
&= \mathbb{E}_B [ \mathbb{E}_{A,B} [ \mathbf{1}_{\|BA\| > t\|B\|(\sqrt{d} + \sqrt{k})} \mid B ] ] \\
&= \mathbb{E}_B [ \mathbb{E}_A [ \mathbf{1}_{\|BA\| > t\|B\|(\sqrt{d} + \sqrt{k})} \mid B ] ] \\
&\leq \mathbb{E}_B [ 2e^{-\frac{(\sqrt{d} + \sqrt{k})^2}{32M^2}t^2} ] \\
&= 2e^{-\frac{(\sqrt{d} + \sqrt{k})^2}{32M^2}t^2} .
\end{aligned}$$

The above uses the fact that the density  $p_{A|B} = p_A$ , which does not hold when the matrices are not independent. □

# Appendix E

## Outlier Removal

**Lemma 48.** *Let  $N \in \mathbb{R}^D$  be a random variable with mean zero, independent coordinates with subgaussian moment bounded by  $\sigma$ , and let  $\eta_1, \eta_2, \dots, \eta_n$  be  $n$  i.i.d. samples.*

*Let  $X \in \mathbb{R}^D$  be a random variable bounded by  $M$ , and let  $x_1, \dots, x_n$  be  $n$  i.i.d. samples. Define  $E_\theta$  to be the samples where  $\|\eta_i\|^2 > \theta$  and let  $\epsilon$  be the probability of exceeding threshold:*

$$E_\theta = \{x_i + \eta_i : \|\eta_i\|^2 > \theta\} \quad , \quad \mathbb{P}(\|\eta_i\|^2 > \theta) := \epsilon .$$

*Then for  $\epsilon \geq \frac{1}{n}$ :*

$$\|\text{cov}(X_n + N_n) - \text{cov}((X_n + N_n) \setminus E_\theta)\| \lesssim t^2 \left( \frac{t}{n} \vee 1 \right) \epsilon \left( M^2 + \sigma^2 D \left( 1 + \frac{\ln(1/\epsilon)}{\sqrt{D}} \right) \right) ,$$

*with probability at least  $1 - Ce^{-ct}$ , where  $\text{cov}$  denotes the empirical covariance with respect to the empirical mean and  $c, C$  are absolute constants.*

**Corollary 49.** *Choosing  $\epsilon = \frac{1}{n} \vee \frac{1}{D}$  and replacing  $t$  with  $t^2$ , one obtains that*

$\theta \leq \sigma^2 D \left(1 + c_1 \frac{(\ln n \wedge \ln D)}{\sqrt{D}}\right)$  for some absolute constant  $c_1$  and

$$\|\text{cov}(X_n + N_n) - \text{cov}((X_n + N_n) \setminus E_\theta)\| \lesssim t^4 \left(\frac{t^2}{n} \vee 1\right) \left(\frac{1}{n} \vee \frac{1}{D}\right) (M^2 + \sigma^2 D),$$

with probability at least  $1 - Ce^{-ct^2}$ , where  $c, C$  are absolute constants.

*Proof.* Let  $n = n_1 + n_2$ , where  $n_2 = |E_\theta|$ , so that  $\mathbb{E}[|E_\theta|] = \epsilon n$ . Let  $\mu_n = \frac{1}{n} \sum_{i=1}^n \eta_i$ ,

$\mu_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} \eta_i$ ,  $\mu_{n_2} = \frac{1}{n_2} \sum_{i=n_1+1}^n \eta_i$ . Similarly, let  $\nu_n, \nu_{n_1}, \nu_{n_2}$  be the equivalent means

for the  $x_i$ , so that  $\mu_n + \nu_n$  is the empirical mean of  $X_n + N_n$ ,  $\mu_{n_1} + \nu_{n_1}$  is the empirical

mean of  $(X_n + N_n) \setminus E_\theta$ , and  $\mu_{n_2} + \nu_{n_2}$  is the empirical mean of  $E_\theta$ . Then:

$$\begin{aligned}
& \text{cov}(X_n + N_n) - \text{cov}((X_n + N_n) \setminus E_\theta) \\
&= \frac{1}{n} \sum_{i=1}^n (x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) - \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \otimes (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \\
&= \left(\frac{n_1}{n}\right) \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) \\
&\quad + \left(\frac{n_2}{n}\right) \frac{1}{n_2} \sum_{i=n_1+1}^n (x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) \\
&\quad - \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \otimes (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \\
&= \left(1 - \frac{n_2}{n}\right) \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) \\
&\quad + \left(\frac{n_2}{n}\right) \frac{1}{n_2} \sum_{i=n_1+1}^n (x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) \\
&\quad - \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \otimes (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \\
&= \frac{1}{n_1} \sum_{i=1}^{n_1} ((x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) - (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \otimes (x_i + \eta_i - \nu_{n_1} - \mu_{n_1})) \quad \} A \\
&\quad - \left(\frac{n_2}{n}\right) \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) \quad \} B \\
&\quad + \left(\frac{n_2}{n}\right) \frac{1}{n_2} \sum_{i=n_1+1}^n (x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) \quad \} C
\end{aligned}$$

Some high probability events will be defined which will provide bounds for the above terms  $A, B, C$ ; but first, a few observations are needed. The sets  $(X_n + N_n) \setminus E_\theta$  and  $E_\theta$  contain independent samples. Furthermore, by Prop. 32

$$\mathbb{P}(\| |\eta_i|^2 - \sigma^2 D \| > t\sigma^2\sqrt{D}) \leq 2e^{-c(t^2 \wedge t\sqrt{D})} \leq e^{1-ct}. \quad (\text{E.1})$$

Thus  $\| |\eta_i|^2 - \sigma^2 D \|_{\psi_1} \leq c\sigma^2\sqrt{D}$ .

**Remark 50.** *One almost has  $\| |\eta_i|^2 - \sigma^2 D \|_{\psi_2} \leq c\sigma^2\sqrt{D}$ ; if the  $\eta_i$  had identical Gaussian coordinates than this would hold asymptotically by the Central Limit Theorem. Note that one always has  $\| |\eta_i|^2 \|_{\psi_2} \leq c\sigma^2 D$ , but this is not sharp.*

The following calculation shows how the subexponential moment changes if one conditions on  $\|\eta_i\|^2 > \theta$ .

$$\begin{aligned} \mathbb{P}(|\|\eta_i\|^2 - \sigma^2 D| > t \mid \|\eta_i\|^2 > \theta) &= \frac{\mathbb{P}(|\|\eta_i\|^2 - \sigma^2 D| > t \text{ AND } \|\eta_i\|^2 > \theta)}{\mathbb{P}(\|\eta_i\|^2 > \theta)} \wedge 1 \\ &\leq \frac{\mathbb{P}(|\|\eta_i\|^2 - \sigma^2 D| > t)}{\mathbb{P}(\|\eta_i\|^2 > \theta)} \wedge 1 \leq \frac{1}{\epsilon} e^{1-t/c\sigma^2\sqrt{D}} \wedge 1 \\ &= e^{1+\ln(\frac{1}{\epsilon})-t/c\sigma^2\sqrt{D}} \wedge 1 \leq e^{1-t/c\sigma^2\sqrt{D}(1+\ln(\frac{1}{\epsilon}))}, \end{aligned}$$

because  $e^{a-t} \wedge 1 \leq e^{1-t/a}$  for  $a \geq 1$ . Thus conditioned on  $\|\eta_i\|^2 > \theta$ :

$$\|\|\eta_i\|^2 - \sigma^2 D\|_{\psi_1} \leq c\sigma^2\sqrt{D}(1 + \ln(\epsilon^{-1})).$$

A Chernoff bound shows that  $|E_\theta|$  is close to its expectation (see Theorem 36):

$$\mathbb{P}(\Omega_{t_1}^C) := \mathbb{P}(|E_\theta| \geq (1+t_1)n \cdot \epsilon) \leq e^{-\frac{t_1^2 n \epsilon}{(2+t_1)\sqrt{3}}}.$$

Observe that if  $\epsilon = 2e^{-c(s^2 \wedge s\sqrt{D})}$ , then  $\theta \leq \sigma^2 D + s\sigma^2\sqrt{D}$  by E.1; it follows that  $\theta \leq \sigma^2 D + c\sigma^2\sqrt{D}\ln(1/\epsilon)$  for some absolute constant  $c$ .

One also has the following concentration of the mean from Prop. 38:

$$\mathbb{P}(\Omega_{t_2}^C) := \mathbb{P}(\|\mu_n\| \geq \frac{c\sigma\sqrt{D}}{\sqrt{n}}t_2) \leq 2e^{-t_2}. \quad (\text{E.2})$$

One also needs to estimate the norm of the mean of  $E_\theta$ ,  $\|\mu_{n_2}\|$ . On  $E_\theta$ , assuming  $\theta \geq \sigma^2 D$ :

$$\|\|\eta_i\| - \sigma\sqrt{D}\|_{\psi_2}^2 \leq \|\sqrt{\|\eta_i\|^2 - \sigma^2 D}\|_{\psi_2}^2 \leq c\|\|\eta_i\|^2 - \sigma^2 D\|_{\psi_1} \leq c\sigma^2\sqrt{D}(1+\ln(1/\epsilon)).$$

Thus since  $\|\eta_i\| - \sigma\sqrt{D}$  (conditioned on  $\|\eta_i\|^2 \geq \theta$ ) is subgaussian with the above moment, one obtains the following concentration of the mean from Prop. 38:

$$\mathbb{P}\left(\left|\frac{1}{n_2} \sum_{i=1+n_1}^n \|\eta_i\| - \sigma\sqrt{D} - \mathbb{E}[\|\eta_i\| - \sigma\sqrt{D}]\right| \geq \frac{c\sigma D^{\frac{1}{4}}\sqrt{1+\ln(1/\epsilon)}}{\sqrt{n_2}}t_3\right) \leq 2e^{-t_3},$$

so that

$$\|\mu_{n_2}\| \leq \frac{1}{n_2} \sum_{i=1+n_1}^n \|\eta_i\| \leq \sigma\sqrt{D} + \mathbb{E}[|\|\eta_i\| - \sigma\sqrt{D}|] + \frac{c\sigma D^{\frac{1}{4}} \sqrt{1 + \ln(1/\epsilon)}}{\sqrt{n_2}} t_3.$$

Since, again on  $E_\theta$ , using Jensen's inequality:

$$\begin{aligned} \mathbb{E}[|\|\eta_i\| - \sigma\sqrt{D}|] &\leq \mathbb{E}[\sqrt{|\|\eta_i\|^2 - \sigma^2 D|}] \leq (\mathbb{E}[|\|\eta_i\|^2 - \sigma^2 D|])^{\frac{1}{2}} \\ &\leq c \|\|\eta_i\|^2 - \sigma^2 D\|_{\psi_1}^{\frac{1}{2}} \leq c\sigma D^{\frac{1}{4}} \sqrt{1 + \ln(1/\epsilon)}, \end{aligned}$$

one obtains:

$$\mathbb{P}(\Omega_{t_3}^C) := \mathbb{P}\left(\|\mu_{n_2}\| \geq \sigma\sqrt{D} + c\sigma D^{\frac{1}{4}} \sqrt{1 + \ln(1/\epsilon)}(t_3 + 1)\right) \leq 2e^{-t_3}.$$

One also has, again by applying Prop. 32:

$$\mathbb{P}(\Omega_{t_4}^C) := \mathbb{P}\left(\left|\frac{1}{n_2} \sum_{i=1+n_1}^n \|\eta_i\|^2 - \sigma^2 D\right| > t_4 \sigma^2 \sqrt{D} (1 + \ln \frac{1}{\epsilon})\right) \leq 2e^{-cn_2(t_4^2 \wedge t_4)}.$$

Finally, each of the  $A, B, C$  terms is bounded on w.h.p. events. On  $\Omega_{t_2}$  and  $\Omega_{t_1}$ :

$$\begin{aligned} \|B\| &= \frac{n_2}{n} \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) \right\| \\ &\leq \frac{n_2}{n} \cdot \frac{1}{n_1} \sum_{i=1}^{n_1} \|x_i + \eta_i - \nu_n - \mu_n\|^2 \\ &\leq \frac{(1 + t_1)n \cdot \epsilon}{n} \cdot \frac{4}{n_1} \sum_{i=1}^{n_1} (\|x_i\|^2 + \|\nu_n\|^2 + \|\eta_i\|^2 + \|\mu_n\|^2) \\ &\leq \epsilon 4(1 + t_1) \left(2M^2 + \theta + \frac{c\sigma^2 D}{n} t_2^2\right). \end{aligned}$$

On  $\Omega_{t_2}$ ,  $\Omega_{t_1}$ , and  $\Omega_{t_4}$ :

$$\begin{aligned}
\|C\| &= \frac{n_2}{n} \left\| \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (x_i + \eta_i - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_n - \mu_n) \right\| \\
&\leq \frac{(1+t_1)n \cdot \epsilon}{n} \cdot \frac{4}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (\|x_i\|^2 + \|\nu_n\|^2 + \|\eta_i\|^2 + \|\mu_n\|^2) \\
&\leq \epsilon 4(1+t_1) \left( 2M^2 + \sigma^2 D + t_4 \sigma^2 \sqrt{D} (1 + \ln(1/\epsilon)) + \frac{c\sigma^2 D}{n} t_2^2 \right)
\end{aligned}$$

For A:

$$\begin{aligned}
A &= \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_{n_1} - \mu_{n_1} + \nu_{n_1} + \mu_{n_1} - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_{n_1} - \mu_{n_1} + \nu_{n_1} + \mu_{n_1} - \nu_n - \mu_n) \\
&\quad - \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \otimes (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \\
&= \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \otimes (\nu_{n_1} + \mu_{n_1} - \nu_n - \mu_n) \\
&\quad + \frac{1}{n_1} \sum_{i=1}^{n_1} (\nu_{n_1} + \mu_{n_1} - \nu_n - \mu_n) \otimes (x_i + \eta_i - \nu_{n_1} - \mu_{n_1}) \\
&\quad + \frac{1}{n_1} \sum_{i=1}^{n_1} (\nu_{n_1} + \mu_{n_1} - \nu_n - \mu_n) \otimes (\nu_{n_1} + \mu_{n_1} - \nu_n - \mu_n).
\end{aligned}$$

Thus

$$\|A\| \leq 4(M + \sqrt{\theta}) \|\nu_{n_1} + \mu_{n_1} - \nu_n - \mu_n\| + \|\nu_{n_1} + \mu_{n_1} - \nu_n - \mu_n\|^2.$$

Now observe that

$$\nu_n + \mu_n = \frac{n_1}{n} (\nu_{n_1} + \mu_{n_1}) + \frac{n_2}{n} (\nu_{n_2} + \mu_{n_2}).$$

Thus:

$$\begin{aligned}
\nu_n + \mu_n - \nu_{n_1} - \mu_{n_1} &= \frac{n_1}{n} (\nu_{n_1} + \mu_{n_1}) + \frac{n_2}{n} (\nu_{n_2} + \mu_{n_2}) - (\nu_{n_1} + \mu_{n_1}) \\
&= \left(1 - \frac{n_2}{n}\right) (\nu_{n_1} + \mu_{n_1}) + \frac{n_2}{n} (\nu_{n_2} + \mu_{n_2}) - (\nu_{n_1} + \mu_{n_1}) \\
&= \frac{n_2}{n} (\nu_{n_2} + \mu_{n_2} - \nu_{n_1} - \mu_{n_1}).
\end{aligned}$$

Therefore on the high probability events  $\Omega_{t_1}$  and  $\Omega_{t_3}$ ,

$$\begin{aligned} \|\nu_n + \mu_n - \nu_{n_1} - \mu_{n_1}\| &\leq \epsilon(1+t_1)(2M + \|\mu_{n_2}\| + \sqrt{\theta}) \\ &\leq \epsilon(1+t_1) \left( 2M + \sigma\sqrt{D} + c\sigma D^{\frac{1}{4}}\sqrt{1 + \ln(1/\epsilon)}(t_3 + 1) + \sqrt{\theta} \right), \end{aligned}$$

and one obtains

$$\|A\| \leq 4(M + \sqrt{\theta})\epsilon(1+t_1) \left( 2M + \sigma\sqrt{D} + c\sigma D^{\frac{1}{4}}\sqrt{1 + \ln \epsilon^{-1}}(t_3 + 1) + \sqrt{\theta} \right) + O(\epsilon^2 \ln \epsilon^{-1}).$$

Thus if one thresholds at  $\theta$  where  $\mathbb{P}(\|\eta_i\|^2 \geq \theta) = \epsilon$ , on  $\bigcap_{i=1}^4 \Omega_{t_i}$ :

$$\begin{aligned} &\|\text{cov}(X_n + N_n) - \text{cov}(X_n + N_n \setminus E_\theta)\| \\ &\leq \epsilon 4(1+t_1) \left( 2M^2 + \theta + \frac{c\sigma^2 D}{n} t_2^2 \right) \\ &\quad + \epsilon 4(1+t_1) \left( 2M^2 + \sigma^2 D + t_4 \sigma^2 \sqrt{D}(1 + \ln(1/\epsilon)) + \frac{c\sigma^2 D}{n} t_2^2 \right) \\ &\quad + \epsilon 4(M + \sqrt{\theta})(1+t_1) \left( 2M + \sigma\sqrt{D} + c\sigma D^{\frac{1}{4}}\sqrt{1 + \ln(1/\epsilon)}(t_3 + 1) + \sqrt{\theta} \right) \\ &\leq C_{t_1, t_2, t_3, t_4} \epsilon \left( M^2 + \sigma^2 D \left( 1 + \frac{\ln(1/\epsilon)}{\sqrt{D}} \right) \right), \end{aligned}$$

where  $\mathbb{P}(\bigcap_{i=1}^4 \Omega_{t_i}) \geq 1 - C_1 e^{-ct}$  and  $C_{t_1, t_2, t_3, t_4} \leq C_2 t^2 (\frac{t}{n} \vee 1)$  for  $t = \max_i \{t_i\}$  and  $c, C_1, C_2$  universal constants. Here one uses the fact that  $\theta \leq \sigma^2 D \left( 1 + \frac{c \ln(1/\epsilon)}{\sqrt{D}} \right)$ , as previously observed.  $\square$

# Bibliography

- Adamczak, R., Litvak, A., Pajor, A., and Tomczak-Jaegermann, N. (2010), “Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles,” *J. Amer. Math. Soc.*, 23, 535–561.
- Bai, Z. D. and Yin, Y. Q. (1993), “Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix,” *The Annals of Probability*, 21, pp. 1275–1294.
- Barvinok, A. (2005), “Measure Concentration,” lecture notes, University of Michigan.
- Belkin, M. and Niyogi, P. (2001), “Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering,” in *Advances in Neural Information Processing Systems 14*, pp. 585–591, MIT Press.
- Beygelzimer, A., Kakade, S., and Langford, J. (2006), “Cover trees for nearest neighbor,” in *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pp. 97–104, New York, NY, USA, ACM.
- Bruske, J. and Sommer, G. (1998), “Intrinsic Dimensionality Estimation With Optimally Topology Preserving Maps,” *IEEE Trans. Computer*, 20, 572–575.
- Camastra, F. and Vinciarelli, A. (2002), “Estimating the intrinsic dimension of data with a fractal-based method,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24, 1404 – 1407.
- Carter, K. and Hero, A. (2008), “Variance reduction with neighborhood smoothing for local intrinsic dimension estimation,” *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3917–3920.
- Carter, K., Hero, A. O., and Raich, R. (2007), “De-Biasing for Intrinsic Dimension Estimation,” *Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on*, pp. 601–605.
- Chen, G., Little, A., Maggioni, M., and Rosasco, L. (2011), “Some recent advances in multiscale geometric analysis of point clouds,” in *Wavelets and Multiscale Analysis:*

- Theory and Applications*, eds. J. Cohen and A. I. Zayed, Applied and Numerical Harmonic Analysis, Springer Verlag.
- Chen, H., Silva, J., Dunson, D., and Carin, L. (2010a), “Hierarchical Bayesian Embeddings for Analysis and Synthesis of Dynamic Data,” *submitted*.
- Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., and Carin, L. (2010b), “Compressive Sensing on Manifolds Using a Nonparametric Mixture of Factor Analyzers: Algorithm and Performance Bounds,” *IEEE Trans. Signal Processing*.
- Chernoff, H. (1952), “A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations,” *Ann. Stat.*, 23, 493–507.
- Coifman, R. and Maggioni, M. (2005), “Multiscale data analysis with diffusion wavelets,” in *Proc. SIAM Bioinf. Workshop*, MinneapolisTech. Rep. YALE/DCS/TR-1335.
- Costa, J. and Hero, A. (2004), “Geodesic entropic graphs for dimension and entropy estimation in manifold learning,” *Signal Processing, IEEE Transactions on*, 52, 2210–2221.
- Costa, J., Girotra, A., and Hero, A. (2005), “Estimating Local Intrinsic Dimension with k-Nearest Neighbor Graphs,” in *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, pp. 417–422.
- Das, Moll, Stamati, Kavraki, and Clementi (2006), “Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction,” *P.N.A.S.*, 103.
- David, G. (1991), *Wavelets and Singular Integrals on Curves and Surfaces*, Springer-Verlag.
- David, G. and Semmes, S. (1993), *Analysis of and on uniformly rectifiable sets*, vol. 38 of *Mathematical Surveys and Monographs*, American Mathematical Society, Providence, RI.
- Donoho, D. and Grimes, C. (2003), “Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data,” *Proceedings of the National Academy of Sciences*, 100, 5591–5596.
- Fan, M., Qiao, H., and Zhang, B. (2009), “Intrinsic dimension estimation of manifolds by incising balls,” *Pattern Recogn.*, 42, 780–787.
- Farahmand, A. M., Szepesvári, C., and Audibert, J.-Y. (2007), “Manifold-adaptive dimension estimation.” in *Proceedings of the 24th international conference on Machine learning*, pp. 265–272.

- Forrester, P. J. (1993), “The spectrum edge of random matrix ensembles,” *Nuclear Physics B*, 402, 709 – 728.
- Fukunaga, K. and Olsen, D. (1971), “An Algorithm for Finding Intrinsic Dimensionality of Data,” *Computers, IEEE Transactions on*, C-20, 176 – 183.
- Geman, S. (1980), “A Limit Theorem for the Norm of Random Matrices,” *The Annals of Probability*, 8, pp. 252–261.
- Gordon, Y. (1984), “On Dvoretzky’s theorem and extensions of Slepian’s lemma,” in *Israel seminar on geometrical aspects of functional analysis (1983/84), II*, Tel Aviv, Tel Aviv University.
- Gordon, Y. (1985), “Some inequalities for Gaussian processes and applications,” *Israel Journal of Mathematics*, 50, 265–289.
- Gordon, Y. (1992), “Majorization of Gaussian processes and geometric applications,” *Probab. Theory Related Fields*, 91, 251–267.
- Grassberger, P. and Procaccia, I. (1983), “Characterization of Strange Attractors,” *Phys. Rev. Lett.*, 50, 346–349.
- Hagerup, T. and Rub, C. (1990), “A guided tour of chernoff bounds,” *Information Processing Letters*, 33, 305 – 308.
- Har-Peled, S. and Mendel, M. (2006), “Fast Construction of Nets in Low-Dimensional Metrics and Their Applications,” *SIAM J. Comput.*, 35, 1148–1184.
- Haro, G., Randall, G., and Sapiro, G. (2008), “Translated Poisson Mixture Model for Stratification Learning,” *Int. J. Comput. Vision*, 80, 358–374.
- Hein, M. and Audibert, Y. (2005), “Intrinsic Dimensionality Estimation of Submanifolds in Euclidean space,” in *ICML Bonn*, ed. S. W. De Raedt, L., pp. 289 – 296.
- Johnson, W. and Lindenstrauss, J. (1984), “Extension of Lipschitz maps into a Hilbert space,” *Contemp. Math.*, 26, 189–206.
- Johnstone, I. M. (2001), “On the Distribution of the Largest Eigenvalue in Principal Components Analysis,” *The Annals of Statistics*, 29, 295–327.
- Jones, P., Maggioni, M., and Schul, R. (2008), “Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels,” *Proc. Nat. Acad. Sci.*, 105, 1803–1808.
- Jones, P., Maggioni, M., and Schul, R. (2010), “Universal local manifold parametrizations via heat kernels and eigenfunctions of the Laplacian,” *Ann. Acad. Scient. Fen.*, 35, 1–44, <http://arxiv.org/abs/0709.1975>.

- Jones, P. W. (1990), “Rectifiable sets and the traveling salesman problem,” *Invent. Math.*, 102, 1–15.
- Jones, P. W. (1991), “The traveling salesman problem and harmonic analysis,” *Publ. Mat.*, 35, 259–267, Conference on Mathematical Analysis (El Escorial, 1989).
- Kegl, B. (2002), “Intrinsic Dimension Estimation Using Packing Numbers,” in *Advances in Neural Information Processing Systems 14*, pp. 681–688.
- Ledoux, M. (2005), “Deviation Inequalities on Largest Eigenvalues,” *GAFSA Seminar Notes*.
- Lee, J. (1997), *Riemannian manifolds: An introduction to curvature*, Springer.
- Levina, E. and Bickel, P. J. (2005), “Maximum Likelihood Estimation of Intrinsic Dimension,” in *Advances in Neural Information Processing Systems 17*, eds. L. K. Saul, Y. Weiss, and L. Bottou, pp. 777–784, MIT Press, Cambridge, MA.
- Little, Lee, Jung, and Maggioni (2009a), “Estimation of Intrinsic Dimensionality of Samples from Noisy Low-dimensional Manifolds in High Dimensions with Multi-scale SVD,” *Proc. S.S.P.*
- Little, Lee, Jung, and Maggioni (2009b), “Multiscale Estimation of Intrinsic Dimensionality of Data Sets,” *Proc. A.A.A.I.*
- Little, A., Maggioni, M., and Rosasco, L. (2011a), “Multiscale geometric methods for data sets I: Estimation of Intrinsic Dimension,” *in preparation*.
- Little, A., Maggioni, M., and Rosasco, L. (2011b), “Multiscale Geometric Methods for Estimating Intrinsic Dimension,” in *Proc. SampTA*, to appear.
- Marčenko, V. A. and Pastur, L. A. (1967), “Distribution of eigenvalues for some sets of random matrices,” *Mathematics of the USSR-Sbornik*, 1, 457–483.
- Pettis, K. W., Bailey, T. A., Jain, A. K., and Dubes, R. C. (1979), “An Intrinsic Dimensionality Estimator from Near-Neighbor Information,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1, 25–37.
- Pinelis, I. (1994), “Optimum bounds for the distributions of martingales in Banach spaces,” *Ann. Probab.*, 22, 1679–1706.
- Pinelis, I. (1999), “Correction: “Optimum bounds for the distributions of martingales in Banach spaces” [Ann. Probab. **22** (1994), no. 4, 1679–1706; MR1331198 (96b:60010)],” *Ann. Probab.*, 27, 2119.
- Priebe, C., Marchette, D., Park, Y., Wegman, E., Solka, J., Socolinsky, D., Karakos, D., Church, K., Guglielmi, R., Coifman, R., Link, D., Healy, D., Jacobs, M., and Tsao, A. (2004), “Iterative denoising for cross-corpus discovery,” in *COMPSTAT*.

- Raginsky, M. and Lazebnik, S. (2005), “Estimation of intrinsic dimensionality using high-rate vector quantization,” *Proc. NIPS*, pp. 1105–1112.
- Rokhlin, V., Szlam, A., and Tygert, M. (2009), “A randomized algorithm for principal component analysis,” *SIAM Jour. Mat. Anal. Appl.*, 31, 1100.
- Roweis, S. T. and Saul, L. K. (2000), “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, 290, 2323–2326.
- Rudelson, M. (1999), “Random Vectors in the Isotropic Position,” *Journal of Functional Analysis*, 164, 60 – 72.
- Rudelson, M. and Vershynin, R. (2007), “Sampling from large matrices: an approach through geometric functional analysis,” *Journal of the ACM*.
- Rudelson, M. and Vershynin, R. (2009), “The smallest singular value of a random rectangular matrix,” *Communications on Pure and Applied Mathematics*, 62, 1707–1739.
- Schlkopf, B., Smola, A., and Mller, K.-R. (1998), “Nonlinear Component Analysis as a Kernel Eigenvalue Problem,” *Neural Computation*, 10, 1299–1319.
- Silverstein, J. W. (1985), “The Smallest Eigenvalue of a Large Dimensional Wishart Matrix,” *Ann. Probab.*, 13, 1364–1368.
- Takens, F. (1985), “On the numerical determination of the dimension of an attractor,” in *Dynamical systems and bifurcations (Groningen, 1984)*, vol. 1125 of *Lecture Notes in Math.*, pp. 99–106, Springer, Berlin.
- Talagrand, M. (1995), “Concentration of measure and isoperimetric inequalities in product spaces,” *Publications Mathematiques de L’IHS*, 81, 73–205, 10.1007/BF02699376.
- Tenenbaum, J. B. (1998), “Mapping a manifold of perceptual observations,” in *Advances in Neural Information Processing Systems 10*, pp. 682–688, MIT Press.
- Tracy, C. A. and Widom, H. (1994), “Level-Spacing Distributions and the Airy Kernel,” *Commun. Math. Phys.*, 159, 151.
- Tropp, J. A. (2010), “User-friendly tail bounds for matrix martingales,” submitted.
- Vershynin, R. (2008), “Spectral norm of products of random and deterministic matrices,” *Probability Theory and Related Fields*.
- Vershynin, R. (2010a), “How close is the sample covariance matrix to the actual covariance matrix?” submitted.

- Vershynin, R. (2010b), “Introduction to the non-asymptotic analysis of random matrices,” Tutorial, University of Michigan.
- Verveer, P. and Duin, R. (1995), “An evaluation of intrinsic dimensionality estimators,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17, 81–86.
- Wielandt, H. (1967), *Topics in the Analytic Theory of Matrices*, Univ. Wisconsin Press, Madison.
- Wigner, E. P. (1957), “Characteristics Vectors of Bordered Matrices with Infinite Dimensions II,” *The Annals of Mathematics*, 65, pp. 203–207.
- Yin, Y. Q., Bai, Z. D., and Krishnaiah, P. R. (1988), “On the limit of the largest eigenvalue of the large dimensional sample covariance matrix,” *Probability Theory and Related Fields*, 78, 509–521, 10.1007/BF00353874.

# Biography

Anna Victoria Little, formerly Anna Victoria McCuiston, was born in Fort Rucker, Alabama on October 31, 1983. She obtained a B.S. in mathematics from Samford University in 2005 and a Ph.D. in mathematics from Duke University in 2011. She is co-author of Little et al. (2009a), Little et al. (2009b), Chen et al. (2011), Little et al. (2011b), and Little et al. (2011a).